INTERNATIONAL UNIVERSITY
UNINETTUNO

FACULTY OF ECONOMICS
DEGREE IN FINANCIAL MANAGEMENT

FIELD OF STUDY
Machine Learning

Dynamic Price Discrimination for SaaS: Maximizing Sales with Machine Learning

SUPERVISOR
Prof. Luigi Laura

STUDENT
Mattia Italiano

ACADEMIC YEAR 2023-24

Contents

Summary

This thesis, titled ”Dynamic Price Discrimination for SaaS: Maximizing Sales with Machine Learning,” explores the fundamental role of adaptive pricing strategies in the Software as a Service (SaaS) industry, with a particular focus on Micro SaaS companies. In an increasingly dynamic market, where scalability and adaptability are essential, traditional static pricing models often fail to meet the evolving customer needs and competitive pressures. The study investigates how the integration of price discrimination theories with advanced Machine Learning (ML) algorithms can help SaaS companies optimize sales, enhance revenue stability, and improve customer satisfaction through personalized, real-time pricing solutions.

The research begins by contextualizing the shift from the traditional pursuit of ”Unicorn” startups—valued at over a billion dollars—to the emerging trend of Micro SaaS ventures. These smaller entities prioritize efficiency, low overhead costs, and targeting niche markets to achieve profitability within narrower margins. Unlike larger SaaS providers, Micro SaaS companies face unique challenges that require agile and responsive pricing strategies to remain resilient against market fluctuations. The thesis posits that ML-driven dynamic price discrimination can more accurately capture demand elasticity across diverse customer segments, thereby improving both profitability and stability for these businesses.

An in-depth examination of the theoretical foundations of dynamic pricing in the SaaS context is presented. This includes an exploration of price discrimination theories—first-degree (individualized pricing), second-degree (usage or volume-based pricing), and third-degree (demographic-based pricing)—and their applicability to SaaS models. The study emphasizes the importance of understanding customer segments and demand flexibility, highlighting how ML algorithms can dynamically apply these price discrimination strategies. By customizing prices based on individual willingness to pay and adapting to ongoing market demand changes, SaaS companies can more effectively manage revenue streams and customer retention.

The thesis’s methodological core delves into the integration of ML approaches for data analysis and model development. Key algorithms, such as K-Means and DBSCAN clustering techniques, are employed to identify user groups with similar usage patterns and price sensitivities. Time series forecasting models like ARIMA and Holt-Winters are used for demand prediction, allowing for more accurate and responsive price adjustments. The importance of data preprocessing—including cleaning, normalization, and handling missing values—is emphasized to ensure the accuracy and reliability of ML models. Additionally, advanced techniques such as sentiment analysis and Natural Language Processing (NLP) are incorporated to analyze user feedback and refine the model’s predictive capabilities.

An empirical simulation using a synthetic dataset is conducted to demonstrate the practical application of the dynamic pricing model in a Micro SaaS environment. The simulation involves customer segmentation based on their willingness to pay and the prediction of optimal price levels that maximize revenue without compromising customer satisfaction. By implementing

tiered pricing strategies, the model adapts to changes in user behavior, enabling the company to effectively respond to churn rates, user acquisition, and engagement metrics. Scenarios involving add-ons and upselling strategies are also tested to further refine pricing strategies and explore additional revenue opportunities.

The thesis concludes with a discussion of the results, highlighting both the strategic benefits and limitations of adopting dynamic pricing models for Micro SaaS companies. The findings suggest that ML-driven price discrimination can achieve price equilibrium and enhance profitability by offering personalized experiences to customers. However, challenges such as data dependency and market volatility are acknowledged, indicating areas where further research and development are needed. The study suggests that future advancements could include the integration of adaptive learning models capable of reducing inefficiencies and responding more swiftly to market dynamics.

Overall, readers of this thesis will gain comprehensive insights into the intersection of economic pricing theories and practical ML methodologies within the SaaS industry. The work provides a valuable framework for SaaS companies—particularly for Micro SaaS ventures—to leverage data-driven dynamic pricing strategies as a means of achieving sustainable growth and maintaining a competitive advantage in a rapidly evolving market landscape. By understanding and applying the principles outlined in this study, companies can better navigate the complexities of demand elasticity, customer segmentation, and revenue optimization in the digital age.

1 Introduction

In the thriving market of *Software as a Service* (SaaS), the ability to optimize pricing has become a critical success factor. This becomes even more evident when focusing more closely on the ecosystem of micro businesses that populate the *startup* system; an environment now subject to a structural change both in paradigms and in corporate culture. Although the success stories of *Unicorns*[1] capture the shared ambitions of dreamy *startuppers*[2], the reality of digital markets shows a stronger tendency toward the creation of a digital body composed of thousands of small enterprises, rather than the allocation of primacy to these "*giants*" [1].

The size and structure of these micro-enterprises make it less crucial to reach large numbers in terms of user base and profit. However, although the effort shifts away from the time-driven race to reach the milestone of 100 million users [2], profit and user base variables remain critical for the long-term survival of the business. Achieving an adequate value threshold for stakeholders is created with a lower net profit compared to other big tech companies. Nevertheless, the small margin that separates success from failure makes business risk extremely sensitive to market volatility and, as a result, less flexible to deviations from initial forecasts.

Traditional pricing strategies, based on static and predetermined models, are no longer sufficient. The fundamental question during the strategic planning of these companies becomes: what price should be applied to their service? The answer is far from trivial, considering the sensitivity previously described. In the profit equation, costs appear mostly as *exogenous variables*, leaving price as the true discriminating factor along with the number of customers reached. Therefore, it becomes essential to adopt a more dynamic and adaptive pricing model, capable of responding quickly to complexity by effectively identifying the various equilibrium points with the demand curves of different potential customer groups.

This thesis aims to provide an answer to this fundamental question. The study of organizational theories related to different degrees of *price discrimination* represents the flexibility sought by the system, solving its complexity with a proposal of models and *Machine Learning* algorithms. Throughout this Introduction, the relevance of the context and the structure on which this research is built will be explained in greater detail.

1.1 Research Context and Importance

At the beginning of this introductory section, Unicorns were celebrated as symbols of success in the digital economy, serving as benchmarks for the aspirations of many entrepreneurs. However, the new market reality is a clear indicator of the structural shift mentioned earlier: the era of Unicorns is in decline due to the significant reduction in the presence of *Crossover Investors* [3],

[1] Venture capitalist Aileen Lee defines "Unicorn" as any startup valued at over one billion dollars

[2] A neologism to describe those passionate about creating and developing startups

[3] Institutional investors in the equities market who are also active in private sectors.

an event that stimulates a shift towards new corporate cultural needs, more focused on frugality and efficiency. Risk is managed by emphasizing optimization to compensate for the lack of large initial capital, while communication strategies tied to selling a dream (and securing subsequent funding) take a backseat [3].

The future will once again be different from what we know, and once more, markets will adapt faster than our *liquid society*[4]. Understanding the relevance of these issues and the tools available is the only strategic weapon to combat uncertainty.

1.1.1 Introduction to the SaaS Industry and Its Economic Relevance

So far, the term SaaS has only been defined based on its broadest meaning: *Software as a Service*. However, a more comprehensive introduction is necessary, starting by positioning this business model within the broader industry of *Cloud Computing*[5], which also includes other service frameworks such as *Platform as a Service* (PaaS) and *Infrastructure as a Service* (IaaS). Compared to these two *business frameworks*, through which companies can release their products to users, SaaS stands out for its greater ease of use by users and a lower level of technical expertise required to utilize its services [4].

The relevance and extent of Cloud Computing, in general, is undeniable. Just consider the numbers from a study conducted by *Eurostat*[6], where between 2021 and 2023, the use of *Cloud* services by European companies saw an increase of 4.2

The impact of this growth is also evident in market trend forecasts and the evaluations of service provider companies. Narrowing the focus of Cloud Computing to the SaaS sector, the rise of remote work and the need for services accessible from anywhere—in addition to the greater flexibility of development, maintenance, and scalability of this type of software—leads to growth in both demand and supply, representing an exponential increase well illustrated in projection graphs [1]. Supporting these studies are figures that help to understand the market size in quantitative terms. In 2023, the SaaS market was valued at USD 273.55 billion, with growth projections ranging from USD 317.55 billion in 2024 to USD 1.228.87 billion in 2032. These projections represent a *CAGR*[7] of 18.4

However, this thesis focuses on a dimension not represented by the giants of Cloud Computing such as *Google, Amazon, Microsoft*, or other SaaS providers like *Salesforce, Shopify*, or *Zoom*. What constitutes the hidden levels of this iceberg are the so-called *Micro SaaS*, that is, small-scale SaaS businesses focused on niche markets, addressing the very specific needs of users [5]. These Micro SaaS companies are usually run by small teams or even by *solopreneurs*[8], whose profits, though minimal due to the narrow niches, are sufficient to meet earn-

[4]According to Z. Bauman, the modern society where change is the only certainty.

[5]The provision of computing services, such as software, databases, servers, and networks, via an Internet connection.

[6]Statistical office of the European Union

[7]Compound Annual Growth Rate.

[8]Entrepreneurs who, on their own, develop and manage their business without the help of other employees.

ings goals with minimal business risk, low barriers to entry, and unprecedented autonomy in development.

In fact, never before in human history has anyone with access to the Internet and a device been able to acquire the information, skills, and platforms to build their own ventures. Programming frameworks reduce the complexity of development, Cloud Computing IaaS environments allow easy hosting for software market launches, and the real gamble is no longer on technical knowledge or infrastructure management capabilities, but rather on the entrepreneurial, communicative, and financial abilities of those with enough intuition to identify a shared problem and offer the right solution before (or better than) others. Consequently, the thriving environment described above offers anyone with the desire the potential to earn online from their creative insights.

Another incentive to pursue such a path is the instability of the job market for *Software Engineers*, which, following the significant post-COVID 19 layoffs of *FAANG*[9] companies—and, by extension, a global scenario lacking new hires—has led many Software Engineers to invent their own work, capitalizing on the skills they have acquired for their own gain.

1.1.2 Importance of Price Discrimination for the Growth of Micro SaaS Businesses

The importance and relevance of the Micro SaaS market go hand in hand with the need for profit, which this thesis seeks to address with its solutions. Given the niche size and the resulting low probability of attracting large numbers of users, volatility and income instability can harm the profits and sustainability of the projects that support the members of these small teams.

In corporate finance, business profit is calculated as revenue minus costs. In a Micro SaaS, costs are inherently low and stable, approximating fixed costs if we exclude the potential for *scaling*[10] and, therefore, the need to increase the Cloud Computing space required to host one's infrastructure. In fact, the staff and salaries are limited to the small team members, operational costs for maintenance and customer support are also low (the real investment is made in the pre-launch phase in software development and marketing communication), while each product sold, regardless of quantity, does not increase costs as in manufacturing production chains. Once created, the product does not require further development costs for each unit sold, making marginal costs nearly zero. Similarly, small companies do not require the management of physical *assets* (subject to depreciation over time) or significant (often entirely absent) *liabilities* such as debt for financing. The investment, more than monetary, is in terms of time from the development team and its opportunity cost. Consequently, costs are limited to purchasing hosting space in the Cloud, paying salaries, and paying taxes in the relevant country. Given these approximate fixed costs, the real variable lies in revenue, calculated as the quantity sold multiplied by the sale price. Since the quantity is limited by the niche chosen by the Micro SaaS, the true revenue differentiator is the price chosen for the products.

[9] Acronym for the tech giants Facebook, Amazon, Apple, Netflix, Google

[10] The ability to manage resources as user base grows at minimal cost.

In microeconomics, it is clear that the intersection of supply and demand generates an equilibrium (ideally Pareto-efficient) that relates the price to the quantity demanded at that price. However, as often happens in studies, the most general models composed for teaching purposes fail to adequately represent market complexity, and where small variations have the most impact, as just described, even the slightest approximation can significantly deviate from established goals and increase risk. The consumer demand curve for a company is usually depicted as the sum of individual demands from each consumer. However, despite the factors that influence demand, such as the price of the good, the number of consumers, or the prices of related goods, it is also true that demand varies according to more demographic characteristics such as consumer income, tastes, preferences, and expectations. Where it is possible to distinguish between categories of different preferences or characteristics, it is possible to break down demand into more similar demand groups, creating two or more functions with different coordinates and supply intersections.

By visually representing what was just said, two demand curves that intersect a supply curve at one specific and unique price for both correspond to two different quantities but contribute to the creation of a *Deadweight Loss*, the loss of total surplus (both consumer and producer), representing a failure in market efficiency. Although the elasticity of the two demands affects this process, dedicating different supply solutions by treating each curve as an independent market can help reduce this loss and translate into consumer advantage on one hand, and higher profit for the company applying these distinctions on the other. In the economic field of *industrial organization*, the dynamism brought by *price discrimination* ensures that distinguishing groups and adopting perfectly tailored solutions for them reduces loss for the benefit of total surplus.

Adopting these types of price discrimination in SaaS, and especially in Micro SaaS, allows structural inefficiencies due to approximation to be addressed, reducing these intrinsic losses and maximizing prices for more predictable profits less subject to the inevitable fluctuations of the financial system.

1.1.3 Technological Advancements in AI and ML Models, NLP Supporting the Economy

Adding complexity to forecasting and computing systems is a strategy that brings with it a dichotomy of greater control at the expense of speed, tending to slow down the reactivity and implementation with which a company addresses changes, ultimately creating fertile ground for damages and risks generated by calculation inaccuracies. Just as the development of new technologies such as increasingly powerful hardware and faster Internet has allowed Cloud Computing to grow in its importance and adoption [4], so too can the rest of technological advancement help simplify the dynamic implementation of these price discrimination systems.

In particular, this refers to the progress and results achieved in the field of *Artificial Intelligence*. Although enthusiasm for this field follows the evolution of *Large Language Models*[11]

[11] AI models trained on large amounts of text to generate and understand our natural language.

(LLM) tools, the field of Artificial Intelligence is much broader and, in terms of topics, levels, and longevity, extends far beyond the enthusiasm surrounding LLM, which has attracted billions of dollars in investment [6]. Therefore, it becomes necessary to distinguish between various fields of study, grouping under the broader AI umbrella the subsets of *Machine Learning*, *Deep Learning*, and *Generative AI* [7] to best represent the methodology followed in this research.

AI includes all techniques and technologies that enable computers to emulate human behavior by allowing them to learn, make decisions, and recognize patterns to solve complex problems similar to human intelligence. Machine Learning is a subset of the broader AI ecosystem, using advanced algorithms to find patterns in large amounts of data, using supervised or unsupervised training algorithms to create models that can learn and adapt. Deep Learning, in turn, is a subset of Machine Learning that uses neural networks to identify high-level features from raw data, simulating the processes humans use to perceive and understand the world. Finally, Generative AI, discussed in the context of LLM, is that subset of Deep Learning whose models are tasked with creating content such as text, images, and code based on a provided input.

There is currently strong interest in AI, with many investments made for its development and a race among tech giants to develop proprietary technologies or make billion-dollar investments [8]. There is high demand for AI solutions both from companies and consumers. Yet, while on the surface it may seem that interest in AI is primarily focused on Generative AI, the attention drawn to the field has enabled more and more companies to evaluate the implementation of Machine Learning systems or other AI subsets to develop models that support business processes and decision-making. Data Scientist roles are increasingly in demand to train ML models to optimize and make business processes more efficient based on the large amount of data that companies collect daily and/or have accumulated over the years. Machine Learning, combined with other subsets like *Reinforcement Learning*, *Computer Vision*, or *Natural Language Processing (NLP)*, is increasingly sought after for building RPA (Robotic Process Automation) systems that can streamline business tasks and reduce both the time and human resources required for otherwise automatable tasks.

In this ecosystem, the choice proposed in this thesis to use data collected from users' browsing of these SaaS platforms to train Machine Learning models capable of predicting new users' behavior appears to be the most suitable approach for balancing the complexity of financial calculation with dynamic and precise solutions aimed at maximizing profits.

1.2 Thesis Objectives and Structure

The objectives of this thesis become clearer once the system has been described and the problem identified. In the following sections, the research question will be formulated, along with an overview of the structure that will guide the response to this question.

1.2.1 Formulation of the Research Question

By combining the three elements of the SaaS industry, Price Discrimination, and Machine Learning, the issue related to the volatility and scale of Micro SaaS companies, in particular, can be reformulated as a problem of maximizing sales. Indeed, if profit is the central focus for the survival of these businesses, then sales become the pivot around which optimization must be sought. The optimal price can be determined through price discrimination techniques, leveraging data collected from these software solutions to train Machine Learning models. This approach allows for dynamic, tailored solutions specifically suited to consumer groups.

The research question, therefore, is to analyze the impact and validity of this approach to determine if it can effectively address the problem identified within the context outlined in this chapter.

1.2.2 Overview of the Chapters and Contents of the Thesis

After introducing the main actors and context of the research in this introductory chapter, the thesis will proceed with a theoretical and formal exploration of the techniques and formulas required to develop the model on which subsequent feasibility studies related to the research question will be based, ultimately yielding empirical solutions.

In the second chapter, the technical characteristics of the SaaS market will be examined, alongside price discrimination theory and how the latter can be applied to the former due to its specific attributes.

The third chapter will focus on Machine Learning in the context of pricing, along with the algorithms and techniques necessary to build an accurate and suitable model capable of predicting future market and consumer behavior.

Chapter four will cover the actual construction of the model, detailing the techniques, considerations, and values that serve as the foundation for the discussion of results and conclusions presented in chapter five.

2 Dynamic Pricing Discrimination in SaaS, Theoretical Foundations

The objective of this second section is to formalize the analysis of the economic problem posed in the research question. While the introduction served to frame a specific market need and its relevance in the modern economic landscape, it is equally true that simplifying concepts into models more suitable for academic study contrasts sharply with the complexity of markets and the resulting inefficiencies in the real world.

Avoiding the repetition of what has already been described in the previous chapter, the following paragraphs will leverage that initial knowledge to lay the foundation for a more rigorous exploration, both in mathematical terms and economic theory. Specifically, microeconomic theories, industrial organization, and *business management* will be used to define the peculiar nature of the SaaS business model, exploiting these concepts to study cost structures, equilibrium conditions, and, finally, the essential metrics for long-term growth and sustainability.

Subsequently, the deeper analysis of price discrimination theories will take center stage in the discussion, prompting reflection on the effectiveness of different pricing models, with a preference for adopting a more complex and customized system tailored to consumer needs. The behavioral response in the study of supply and demand will contribute to the formulation of various ethical and economic questions which, framed in the context of customer relations and sales forecasts, will later be a key point for price selection in the SaaS domain.

Finally, the intersection of price discrimination theories with the specific nature of the SaaS model will serve as the *trait d'union* for the concluding analysis of this chapter regarding the adoption of pricing structures best suited to maximizing profits in these companies. The complexity of the final responses will be essential to understanding the importance of *Machine Learning* and its predictive algorithms, which will be examined in the following chapters.

2.1 Characteristics of the SaaS Business Model

Although the distinction between the two terms encapsulated in the recurring reference to the *SaaS business model* might be considered implicit due to the reasoning already employed, it is important to distinguish the two different elements contained within this topic. SaaS, as Software as a Service, is nothing more than a company's product—a software product defined by a distribution model that allows the consumer to access it via the internet rather than requiring a physical installation. This ensures the possibility of access through web or mobile browsers, as well as the opportunity to benefit from real-time software updates and maintenance.

SaaS, as a product, is a necessary but not sufficient element in defining a SaaS company, which acts as the provider of the services just described[9]. A SaaS company is, in fact, a type of business that focuses on creating, developing, distributing, and maintaining a proprietary SaaS

product. This corporate structure revolves around the peculiarities of the SaaS product market, influencing sales, marketing, and customer service roles within the company. The benefits for a SaaS company, compared to traditional businesses, are numerous, such as access to the global market and the ability to scale at reduced costs with an agile and flexible structure.

As a result, SaaS companies can vary in size and may be part of a more diversified structure within a larger organization. A SaaS product can also be developed by a company whose core business lies in a different market (such as manufacturing), thus varying the corporate organization and internal financial and accounting dynamics. However, since the main objective of this thesis is to focus on the peculiarities that distinguish SaaS companies from other business models, all hybrid formulations will be omitted in favor of the premises that characterize the structures most susceptible to the conditions analyzed in this research.

The *business model* of SaaS companies is highly dynamic, flexible, and peculiar. It is part of a recent and evolving economic context, as it is deeply tied to and limited by technological innovations. Supporting this point is the recent phenomenon of the emergence of numerous SaaS companies due to discoveries in the field of AI that have revolutionized the landscape of LLMs. All these new technologies have addressed both old and new problems, contributing to the growth of the Micro SaaS ecosystem through solopreneurs and small teams that have launched their creative solutions, while larger companies have exploited these discoveries and services within their products, driving growth through increased sales.

But how does all of this translate into the shared model of both small and large SaaS companies?

2.1.1 Definition and Characteristics of the Business Model

Similar to the process of defining a *set* in mathematical terms, the SaaS business model distinguishes itself from other models by grouping all those structures characterized by specific elements in terms of organization, costs, sales, customer management, and sensitivity to growth values. It is not advisable to rely solely on natural language for a clear and complete definition; rather, the most effective approach in these cases is to define the SaaS business model through the characteristics of its main components.

Why does a SaaS company come into existence?[10] The question is no different from asking about the nature of a generic enterprise[12]. A SaaS, like any other business, is certainly born from a desire for profit and wealth creation. Even when referring to Micro SaaS, composed mainly of solopreneurs or small teams, the ultimate tool for maximizing the value (both economic and social) sought by stakeholders within its ecosystem remains efficiency. Profit is the goal, but efficiency is the tool that makes it possible to maximize that profit. Whatever the personal reasons for embarking on a value creation journey through a SaaS model, the fundamental metrics to follow are those related to efficiency and the reduction of inefficiencies. In

[12] A famous question arising from the 1937 works of English economist *Ronald Coase*.

this definition, there is no room for subjective goals, strategies, and plans: whether focusing on a niche of a few users to maintain a small scale, or aiming for *blitzscaling*[13] like a Unicorn, efficiency in pursuing these goals remains a critical sustainability factor. Not all SaaS companies must aspire to be startups that impact the world[14], yet growth remains a fundamental metric for survival because it is aimed at efficiency. Without growth, there is stagnation, but in economics, if you are not moving forward, you are not standing still: you are falling backward until failure.

According to a McKinsey study[11], it is not enough for companies to achieve positive growth to avoid the risk of failure. To avoid this risk, the growth rate needed for business survival must exceed 20%, although only by surpassing 60% can success be predicted with greater certainty. In fact, historical data in this field shows a strong correlation between growth rates and market value. Only 28% of companies manage to exceed USD 100 million in value, while only 3% reach the billion-dollar mark, and a mere 0.6% surpass USD 4 billion. Value is highly susceptible to both shareholder ROI[15] and the growth rate, especially in the early stages of a company's life. Growth becomes such an important parameter that exceeding 20% makes cost structure a negligible variable for survival, although greater efficiency in cost management certainly helps to increase the growth rate itself.

In the SaaS company context, the mantra of *fast shipping*[16] is often repeated, with smaller projects aiming for software development and launch within a seven-day timeframe. Adopting this *modus operandi* has many advantages: the time and resources spent limit the opportunity cost of the project, and it quickly becomes clear whether an idea is successful in terms of funding or customer interest. However, the short timeframe risks leading to a lack of commitment to the project, which in turn can result in a failure to actively manage and grow the SaaS. Although this thesis frequently refers to the Micro SaaS environment, which is more susceptible to the research topics, it is important to clarify here that it will never refer to those projects left to run on their own, where the solopreneur seeks only additional income without investing their time. Indeed, achieving a high degree of efficiency is correlated with an active commitment to growth, a commitment that goes far beyond software creation and is often driven more by dynamic decisions in sales, marketing, and customer service. Therefore, excluding this subgroup of models is not intended to assign a lower value in terms of utility, financial effectiveness, or ethics compared to other Micro SaaS environments[12]: the difference lies exclusively in the assumptions, which, while conforming to the field, diverge on the premise of actively seeking sustainable growth and long-term value creation.

[13] According to Hoffman and Yeh, it is the method of rapid scaling adopted by tech startups that prioritizes speed over efficiency.

[14] According to Dan Norris, a writer on startups, a business is not a startup unless it has the capacity to impact the global market and—therefore—change the world.

[15] Return on Investment.

[16] Delivering software to market in the shortest time possible.

2.1.2 Financial Metrics and Key Business Functions

Financial metrics follow the business model and define how various business functions are allocated and financed. In the case of Micro SaaS, the solopreneur often has to wear multiple hats, acting as Developer, Graphic Designer, Sales, Marketing, and Customer Service all at once. However, even in more structured contexts, each of these functions operates within increasingly blurred boundaries, making it difficult to clearly define areas of action and divide responsibilities.

The internet provides the opportunity for faster, more flexible, and autonomous information management, offering potential SaaS customers the means to conduct their own research and make informed purchasing decisions. The customer navigates their own decision-making journey, interacting with the sales department, exploring marketing content, reading customer service reviews left by other consumers, assessing the product's UI/UX[17], or comparing the software and prices with competitors. The goals become common and overlapping, requiring coordinated management as a result. This united front is both the result and the reason why SaaS financial metrics impact all components of the company's structure.

Before delving into the complexity of pricing structures in the following paragraphs, we can accept— for the purposes of presenting metrics—one of the simplified versions of the SaaS business model, which identifies SaaS products as online software whose users access the services by paying a recurring monthly (or annual) fee for an agreed-upon period. The recurring payment of a set amount greatly differentiates the SaaS business from the traditional business model, where, taking the software industry as an example, one would purchase a physical CD of software in exchange for long-term use but without the benefit of recurring updates. Conversely, SaaS products are not purchased in physical form, establishing ownership, but rather offer an always-updated experience accessible anywhere as long as the access fee is paid. Based on this definition, the primary challenge for a SaaS company converges on a very specific direction: the customer relationship. The ability to continuously acquire new users and keep them engaged for as long as possible must align with the key goals of profit maximization and company growth. Such efforts must necessarily be clear, measurable, and directly influenced by the company's planned actions.

Below are the main financial metrics[13] managed through the company's *Financial Planning and Analysis*[18] (FP&A) operations. These are all metrics based on the crucial integration of subscription-based sales, the value perceived by consumers, and the churn rate[19] as quantitative data that allows for the representation of customer acquisition and retention fluctuations. Additionally, they are observable and predictable metrics in real-time, which is why they will later play a decisive role in choosing the Machine Learning algorithms to be adopted for profit maximization calculations.

[17] User Interface and User Experience, which contribute to the product's aesthetics and user experience.

[18] Budgeting, Accounting, Financial Forecasting, and Financial Data Analysis

[19] The annual or monthly percentage representing the number of users who stop paying for the service.

Growth and Economic Metrics

1. *Recurring Revenues* (RR): These are the recurring revenues of the SaaS, obtained by multiplying the number of subscribers by the subscription cost.

$$RR = Number\ of\ Subscribers \times Cost\ of\ Subscription \tag{1}$$

   This metric provides a first representation of the company's profitability, although more specific and useful parameters for planning are described below. Referring to recurring payments (subscriptions), one-time payments are excluded from the calculation as they are useful for accounting but unreliable for future forecasts.

2. *Monthly Recurring Revenues* (MRR): These are RR calculated on a monthly basis, taking the subscriber data of a given month and multiplying them by the average revenue per user. This calculation is an index that represents a specific point in time, capturing a snapshot of the MRR of a specific number of users without considering variations caused by customer acquisition or churn. It is also a data point that can be used to calculate monthly growth in quantitative terms in the *Revenue Growth* formula.

3. *Annual Recurring Revenues* (ARR): These are RR adjusted for the long term. While MRR can provide a dynamic and timely indication in the short term, on an annual basis, other fluctuations and contractual dynamics need to be considered. In addition to recurring revenues, it is necessary to calculate any *upselling*[20] or downgrades, as well as churn rates and renewals at a discounted price compared to the original.

$$\begin{aligned} ARR = &(RR\ Subscriptions + RR\ Add-ons + RR\ Upselling) - \\ &(Revenues\ lost\ from\ Downgrades + Cancellations + Renewal\ Discount) \end{aligned} \tag{2}$$

   Calculated annually, this can also be adapted for quarterly, half-yearly, or multi-year calculations.

4. *Expansion Monthly Recurring Revenue* (EMRR): A specific calculation of MRR based on the expansion rate, reporting only the additional MRR due to upselling or add-ons from existing customers. It is a metric related to the company's ability to generate revenue from already acquired customers without additional acquisition costs.

5. *Average Revenue per User* (ARPU): This is the average revenue obtained by dividing total revenues by the number of the company's customers. The average can be representative of more or less general data, for example, considering only a specific segment of the user base rather than focusing on its entirety.

[20] Selling more expensive subscriptions to existing customers.

6. *Revenue Growth* (RG): This is the growth rate obtained by considering the percentage delta between revenues in period *t* and *t-1*. For example, if in a given month the company bills 20

$$GR = \frac{Revenues_t - Revenues_{t-1}}{Revenues_{t-1}} \times 100\% \tag{3}$$

Specifically, recurring revenues represented by MRR or ARR can be used instead of a more general notation to communicate information related to the historical data of subscriptions.

7. *Gross Margin* (GM): This is used to measure a company's profitability in percentage terms, using revenue data net of the cost of goods sold (COGS), all divided by revenues. In particular, COGS refers to direct production costs, which, in the SaaS sector, typically include costs associated with *hosting*, software development, and customer support.

$$GM = \frac{Revenues - COGS}{Revenues} \times 100\% \tag{4}$$

This data measures if and to what extent the company can cover its expenses. The higher the value of this indicator, the more profitable the company becomes. Since the numerator is always smaller than the denominator, as it is reduced by the COGS, lowering costs or making them less impactful relative to growing revenues increases the business's profitability.

8. *Customer Lifetime Value* (LTV): This measures the expected total value that a user will generate over the course of their relationship with the company. It is the long-term value brought by the customer, an indicator that helps identify the most profitable customer segments. Indeed, *Customer Concentration*[21] for the customers with the most significant impact on a company's earnings belongs to the segment with the highest LTV, whose variation in quantitative terms causes greater fluctuations than smaller segments. This indicator is calculated by multiplying ARPU by the average duration of the relationship with the customer, typically both expressed in monthly terms.

$$LTV = ARPU \times CLV \tag{5}$$

Or more generally:

$$LTV = Average\ value\ of\ a\ transaction \times Average\ number\ of\ transactions \times CLV \tag{6}$$

As mentioned, this metric becomes one of the fundamental metrics for identifying the most profitable customer segments for the company and, consequently, also the areas where improvements can be made to increase LTV values in less profitable segments, or

[21] The concentration of customer groups in the impact on revenues.

even to abandon areas with little impact on the company's profits.

Acquisition and Retention Metrics

1. *Customer Acquisition Cost* (CAC): The cost of marketing, advertising, and sales operations to acquire a customer. It is calculated by dividing the total expenses invested in these acquisition operations by the number of customers obtained.

$$CAC = \frac{total\ expenses}{number\ of\ customers} \tag{7}$$

   It is essential to keep this as low as possible to make the acquisition process more efficient, though some factors outside the company's direct control may raise values, such as the entry of new competitors into the market.

2. *Annual Contract Value* (ACV): The value of the annual contract (the monthly value is referred to as MCV) purchased by the customer at the time of acquisition. It is useful for calculating the *ratio* between CAC and ACV to understand the surplus between acquisition costs and the value obtained. It differs from general profit as it does not include other fixed and variable costs related to managing the company's infrastructure due to the addition of a new customer.

3. *CAC Payback Period* (CACPP): The time required to achieve a positive cash flow after the initial costs spent on customer acquisition. This helps calculate when the investment begins to yield profits to the company, thanks to monthly payments discounted for the opportunity cost interest rate. This metric allows the analysis of process efficiency and financial parameters related to the liquidity necessary for growth and survival over time between investments and profits[14].

$$CACPP = \frac{CAC}{ARPU \times Gross\ Margin} \tag{8}$$

   The lower the value, the more profitable the company is considered. Ideally, the result should be less than one year to avoid the need for excessively high initial capital investments. A common technique to address this type of expense is to require payment in advance, thus adjusting the cash flow. An example of these approaches is SaaS plans that offer a slightly discounted price but require full annual payment upfront at that monthly price.

4. *Magic Number Ratio*: This is an indicator that relates CAC and LTV, whose ratio is used to measure the sustainability of the business. If the LTV:CAC ratio exceeds 3:1, it means

that the adopted pricing and cost structure formula helps the company grow over time, making the allocation of resources between perceived customer value and the costs of acquiring them efficient. Together with the Payback Period, these factors influence the economic value for investors.

5. *Churn Rate*: This indicator refers to the customer loss rate. The percentage expressed by the Churn Rate represents the percentage of customers lost over a specific period, calculated by subtracting the total number of customers at the end of the period from the total number of customers at the beginning, excluding new acquisitions. For example, 10% on an annual basis means that the company loses 10% of acquired customers over the course of a year, or if the monthly rate is 3%, the total customer base lost over the year will amount to 36%. Hence the importance of keeping this rate as low as possible, responding to the information it provides by identifying the root cause of these losses and adopting strategies to improve customer satisfaction.

6. *Revenue Churn Rate*: This is the rate of revenue loss due to customer churn. It indicates the impact of the Churn Rate on revenues, i.e., how much the customer loss rate quantitatively affects total revenues. Since it is always calculated as a percentage, a value of 10% would mean that total revenues for the period under consideration decreased by 10% solely due to non-paying customers compared to the previous period.

7. *Daily Active Users* (DAU): This metric does not directly represent dynamics related to costs, profits, or losses but remains essential for forecasting future revenues. Tracking unique daily or monthly active users (MAU) helps perceive the value that each user attributes to the service in response to their needs. While DAU and MAU metrics differ only in the calculation time span, having many unique monthly users but few daily users weighs more in the engagement calculation. High engagement results in a longer-lasting relationship between the user and the company. Making decisions that increase these indicators leads to a greater ability to engage in *upselling* or *cross-selling*[22], and consequently, to an increase in revenues. Assigning a score to engagement with customized KPIs is a *best practice* recommended for every SaaS company.

This list of metrics, while not exhaustive of the financial complexity of a business but adequately detailed for the dynamics addressed by this thesis, represents the starting point for the calculation of costs, revenues, and pricing decisions to be addressed in the following sections. However, before moving on to subsequent considerations regarding the SaaS company, it is worth noting how many metrics revolve around the deeply relational nature of this type of business with its customers. Intercepting and predicting the behavioral responses of the target users means addressing the acquisition, engagement, and retention challenges faced by this business model, thus extending its profitability over time[15].

[22]Selling additional products without increasing the main subscription price as occurs with upselling.

2.1.3 Background and Premises of the Analysis

To facilitate the process of analyzing costs, revenues, and price decisions—the foundation for the subsequent application of the Machine Learning model—it is essential to clarify the assumptions of the paradigm to be used. Economics is not a discipline of certainties; there are many complexities to consider, as well as dynamic variables to monitor, fluctuations, and contingencies that depend on the specific company. Consumer psychology plays a large role, as do behavioral responses to both internal and external variables, along with short- and long-term considerations. Therefore, this paragraph aims to clarify the model under review for subsequent analyses: it presents an extension of the characteristics already outlined, aimed at simplifying the aforementioned complexities and in line with the thesis's objective. The aim is not to study a specific case with real historical data but rather to apply a rigorous theoretical framework to a modern solution aligned with optimization needs relevant to the market.

Market Context

In line with what was explained earlier, we will use the Micro SaaS market as a model: this type of market does not exhibit characteristics of full efficiency, making it impossible to configure it within the paradigm of a perfectly competitive market[16]. Companies within this market have a certain degree of price influence, creating their own small monopolistic-like situations through strong differentiation of the services and attributes they offer. At the same time, this market cannot be defined as strictly monopolistic, as the barriers to entry are minimal due to the availability of technologies, cost reductions, and the accessibility of knowledge and development. This makes virtually anyone with an internet connection a potential competitor, provided they are targeting the same niche.

The consequences of this competitive-monopolistic dichotomy support the theoretical foundation of monopolistic competition, which forms the basis for subsequent analyses. This framework is consistent with other statistics presented in describing the SaaS market, emphasizing the importance of profits tied to growth rates. In fact, the structure of a monopolistic competition[17] market exhibits two distinct behaviors in the short and long run, with competition having greater influence in the latter scenario, leading to a downward shift in the demand curve. Adjustment mechanisms require adding value to the offered product by enhancing tangible attributes (new services, new technologies) or intangible ones (market presence, psychological value, brand prestige). Maximizing profits and growth rates makes it possible to invest in research and development to counter this phenomenon.

In this specific case, we consider a Micro SaaS market of the *Business to Customer*[23] (B2C) type, with a sales model based on monthly subscriptions. Key economic metrics such as MRR, CAC, LTV, *magic number*, and churn rate play an essential role in determining costs and rev-

[23] A sales model aimed at individual customers, not businesses.

enues.

It is also worth reiterating that, in line with the academic nature of this thesis and avoiding strategic deep dives into specific corporate financial management, costs, cumulative market demand, and revenues will be studied *ceteris paribus*[24], even though real-world revenue generation is more driver-based, where revenue as an output is influenced by CAC and the conversion rates of SG&A marketing and sales expenses, not vice versa. For the same reason, assumptions about financing will be treated as exogenous variables: we assume that, due to the nature of Micro SaaS, there is no need for external funding, relying instead on the founder's bootstrapping ability and the capacity to sustain a payback period or cover sales and marketing expenses without jeopardizing the company's financial health.

SaaS Product Type

The product in question is a service offered for a monthly subscription fee, targeting a niche group of users. To make the analysis more concrete, we can hypothesize that this company has developed a service to automate bureaucratic and organizational processes typically faced by a specific subset of *freelancers*. This subset of the buyer persona could include those earning between $50 and $150 per hour. Adding more hypothetical characteristics, we could estimate that the potential customers in the global market number 50,000 freelancers worldwide, but the company can only aim for a 30% market share due to factors such as competition (e.g., loyalty to other brands or existing subscriptions to similar services), inefficiencies in marketing campaigns that prevent reaching the entire target audience, language barriers, missed integrations with specific processes used by clients, or less-valued attributes.

Another simplification for ease of calculation is the static nature of the market share percentage. In reality, this follows the demand curve, making subscription to the product dependent on other variables such as product novelty, peer network effects, and the dynamic behavior described by Roger's S-curve of product adoption, which through logistic regression outlines a trend composed of the four phases of the product life cycle.

With these premises established, we can proceed with the analysis of the cost structure, demand function, equilibrium, and pricing.

2.1.4 Costs Structure of a SaaS

The reference to key financial metrics for managing SaaS companies is linked to the specific market context, the relationship with customers, and the importance of growth as a forecast metric for future sustainability. However, it is also clear that the managerial operations aimed at customer acquisition and retention require fertile ground that fosters the monetization of the relationships established with this resource investment: determining the appropriate price

[24] Keeping other economic factors constant.

that balances supply and demand is, therefore, a critical factor that finds its essence in the analysis of costs and revenues. Yet, the microeconomic structures of these companies take on particular characteristics due to the specific market context, making them optimized in relation to the metrics analyzed earlier and differing from decisions made by more traditional companies regarding market equilibrium.

The definitions of costs do not vary from those studied in other production contexts: the distinction between direct and indirect costs, fixed or variable, as well as total, marginal, and average costs, remains the same. However, in the dynamic world of SaaS, the way these costs are conceived and allocated—along with their impact on the financial health of the company—varies based on the central concept to which they refer: the costs associated with *Cloud Computing* and the consequences of using this technological infrastructure.

Indeed, analyzing the distinctive architecture, the service offering relies on several key elements: the *software*, the ability to *access it online*, the *space needed* for user data structures and databases, the *domain* of the website, and the *facilities* on which the product is built. Starting with the software itself, depending on its size and complexity, it must account for various development and maintenance costs over time. Additionally, it is essential to include the cost of any third-party services (such as APIs[25]) used within the offered SaaS product. The software needs to be hosted on a cloud platform accessible via the web, resulting in costs for the resources required to purchase space for uploading the site code, the algorithmic logic governing it, or the space needed for the user data database. These services are provided by *Cloud Computing* services mentioned earlier: *PaaS* (Platform as a Service) and *IaaS* (Infrastructure as a Service), which, upon subscribing to one of their plans, guarantee the conditions for the online presence of the SaaS software. The subscription for PaaS hosting (used for hardware and software resources) and IaaS (responsible for managing the database and backend logic[26]) is the fundamental pillar around which the cost structure for a SaaS company revolves.

Fixed Hosting Costs

The cost of the space needed to host the SaaS product infrastructure in the *cloud* varies based on the *size* and *performance*, i.e., the costs of hosting services are divided based on the computational capacity and memory of the infrastructure relative to user traffic and data handled by the SaaS product. Excluding the option of purchasing discounted shared computational resources (which are not suitable for the development of a professional and high-performance business), the cost structure related to technology is divided into fixed and variable costs, structured in fixed memory band tiers. For the theoretical purpose of this analysis, moving away from calculations in *Gigabytes* and using the number of users as the unit of measurement, proportional

[25] *Application Program Interfaces*: software interfaces that allow interaction between different programs, even external to the company, with fees based on the resources used.

[26] The backend includes the part of the software that operates behind the scenes, handling the interaction logic with data structures visible in the frontend, the software interface accessible to the user.

to the average memory used by each in usage, it is possible to calculate and visualize a cost structure characterized by a stepwise pattern: using purely arbitrary data for a Micro SaaS, the reference graph for fixed hosting costs[18] (domain, security certificates, space for frontend, backend, database, possible cache optimizations or automations) describes an increase in its cost in a non-linear manner but subject to the memory band tiers needed to support a specific number of users.

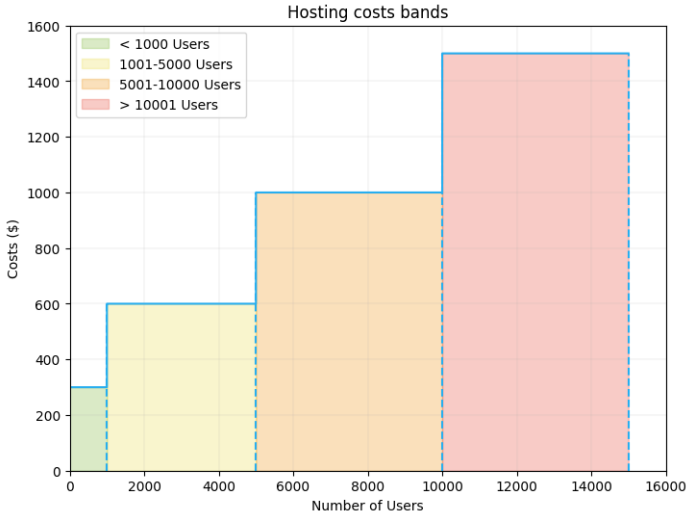


Figure 1: Monthly hosting fixed cost by number of users

2.1.5 Demand curve

The higher the profit, the greater the contribution to the wealth of stakeholders provided by the company. The formula for profit is simple but effective:

$$\pi = p \times q - TC \tag{9}$$

That is, profit π equals revenues (p×q) minus total costs (TC = fixed costs + variable costs). Having already analyzed the costs, it remains to further explore the trend of revenues in a SaaS company, where the quantity doesn't pertain to the product itself but rather to the number of customers, while the price is the average paid to access the service.

Considering the two key elements, price and number of users, we will first analyze customer-related considerations before diving into pricing decisions. Indeed, the two elements are deeply interconnected, influencing each other based on the limitations and flexibility found in both. An example of this is the number of users, the initial topic for analyzing revenues.

In the introductory chapter, the SaaS world was described as a business model that finds its essence in solving real problems. A SaaS company, even before developing the software, is subject to a validation test aimed at recognizing its effectiveness in solving existing problems within a specific environment. This environment can be broad or niche, as previously discussed in the distinction between SaaS and Micro SaaS. These considerations, however, do not impact

solely the structure and size of the company: the number of actors involved in a given environment becomes a fundamental element for understanding the customer base that can be predicted through acquisition processes.

A very large environment may offer greater growth opportunities, although it may also be more competitive. In this case, the potential users will be sufficiently numerous to foresee a revenue system with lower costs (also due to the threat of competition) while still having a large enough user base to more easily cover fixed expenses. The turning point of web 2.0[27] was precisely this: social networks capable of offering their platforms for free by leveraging the sheer number of reachable users and the ability to connect them with other businesses interested in reaching them. A SaaS company with many thousands, if not millions, of potential customers can plan its financial goals in a completely different way from a Micro SaaS with a small user base within a restricted niche. The greater the level of discrimination (understood as identifying the *buyer persona* in highly specific individuals compared to the rest of the population), the more attention must be paid to price efficiency, financial metrics, and cost structure, as these are extremely sensitive to changes in the number of users.

Losing one user out of a hundred available has much more impact than losing one out of a million potential customers. Although managing a smaller user base involves — as previously seen — lower total costs and — as will be discussed later — a lower Break Even Point compared to SaaS with broader potential, finding a competitive price that aligns with the value the customer assigns to the SaaS product is crucial for maximizing revenues with the limited pool of customers that can be reached.

In microeconomics[19], market equilibrium is the point of intersection between the consumers' demand curve (the sum of individual curves of each buyer) and the company's supply curve for a good or service. The supply curve is constrained by costs, the target profit, competition, and the type of market. However, focusing on the demand curve, it represents the customer's willingness to purchase the product at a given price. Consequently, the relationship between the price of a product and the quantity demanded is inversely proportional: as the price decreases, the quantity of the good purchased increases, while an increase in price corresponds to a lower demand. The goal of a company is to sell as much as possible, but revenues are derived from the price/quantity product, meaning that lowering the price but selling more products can yield the same result as selling fewer but at a higher price. All the microeconomic assumptions presented refer to consumer rationality, which implies perfect knowledge of their demand aimed at maximizing purchasing utility; however, psychological and social factors also come into play, which will not be explored for now.

Therefore, the cumulative demand curve, while useful, proves to be extremely complex to estimate. It is subject to all the variables listed, the dynamism of the market, and is difficult to predict despite historical data: the market is constantly evolving, and the influence of other companies involved generates continuous shifts that, in turn, lead to other dynamics. One ex-

[27] an evolution of the internet focused on user-generated content (social media)

ample is churn rate, which removes a user and their demand from the cumulative demand for the service offered by a SaaS for an indefinite time (often permanently), potentially adding them to the demand of another competitor.

Excluding the many uncertainties and relying on the assumptions described in the previous paragraph, a good approximation for studying demand can be based on the example of calculating ROI for the customer. Indeed, the price one is willing to pay for the service is nothing more than the value each customer assigns to that service: it is said that a company does not assign a price to the product but to the customer. The assumptions identified the buyer persona as a freelancer with an hourly rate between \$50 and \$150. After studying this product, it is estimated that the automation offered saves an average of ten hours per month, and the customers' willingness to pay is about 10% - 30% of this ROI, calculated as $ROI = 50 \times 10 = \$500$ and $ROI = 150 \times 10 = \$1500$, i.e., $ROI = hourly; rate \times hours; saved = \$500$. This results in a price range between \$50 and \$450, which, however, reflects too wide a fluctuation in a competitive and price-sensitive context in relation to the product's features.

In fact, to adopt a structure representative of these characteristics, one can think of willingness to pay as the value represented by the Berry, Levinsohn, and Pakes model presented in their paper "Automobile Prices in Market Equilibrium"[20], where the utility of consumer i for product j is given by:

$$U_{ij} = x_j\beta_i - \alpha_i p_i + \xi_j + \epsilon_{ij} \tag{10}$$

With x_j representing the observable characteristics of product j, β_i the consumer i's sensitivity to these characteristics minus the consumer's sensitivity α_i to the price of product p_j, adding the unobservable characteristics of the product ξ_j (such as preferences, brand value, social value), and an idiosyncratic error term ϵ_{ij} that captures unobservable preferences; and, therefore, not precisely quantifiable.

These preferences can be normally distributed around an average, meaning that a demand curve can be exemplified with a price preference among market users that can be represented by a Gaussian bell curve where, symmetrically, preferences are distributed around a mean with a certain degree of deviation. The function is as follows:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{11}$$

With σ^2 equivalent to the square of the standard deviation σ (deviation from the mean) and μ representing the mean around which to generate this symmetry, with the peculiarity that the mean, mode, and median are equal. Based on the quantities used as examples, the mean price preferred by consumers can be found at \$75 with a deviation of \$25, quite high but justified by the target with a wide difference in hourly rates and consequent ROI. As a result, 68% of users are willing to pay in the price range between \$50 and \$100 (one standard deviation), while increasing the range between \$25 and \$125 (two standard deviations) includes 95% of users,

reaching 99.7% between \$0 and \$150, with a small percentage of 0.3% willing to pay over \$150. However, the resulting graph and the projection of this behavior onto a demand curve show that a symmetrical distribution model is not well suited to representing market phenomena. First of all, the elasticity of the curve is smoother than expected in reality, where there is much greater price sensitivity in areas beyond the deviation from the mean, justified by higher price sensitivity compared to the evaluation of observable and unobservable attributes. Moreover, although the price decreases, there is a strongly inelastic portion of the curve below \$50, as, despite having a higher surplus, consumers who value this service more are not keen on a significant price reduction due to a perceived lower response value. This also leads to a reduction in the user base willing to purchase, representing a function akin to Veblen goods[28][21], where the lower the price, the fewer people are willing to buy. In this case, a low price is viewed with suspicion, representing either low product quality and/or long-term survival difficulties (making it inefficient to learn to use this product). Certainly, there is a new trend reversal near zero because the free investment of resources reduces the perceived risk. Finally, the curve's inelasticity also represents the product adoption life cycle, which grows faster as more users can be acquired, leveraging not only marketing channels but also the network effect. To approximate this be-

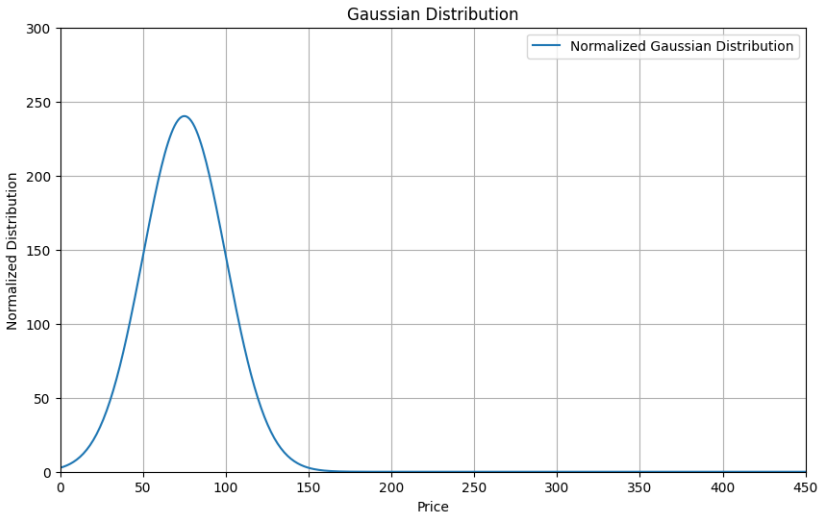


Figure 2: Gaussian Bell

havior (omitting Veblen variations and those near zero since revenues would be too low, and thus no company would be willing to sustain a loss-making business), a logistic function can be used, described by this formula:

$$D(p) = \frac{L}{1 + e^{-k(p-p_0)}} \tag{12}$$

L is the maximum number of reachable users, while Q(P) is the quantity willing to purchase at a given price P. The element k, on the other hand, is a parameter that influences how quickly

[28]Luxury goods.

the curve rises: this S-shaped trend represents more inelastic beginnings and ends (prices that are too high or too low do not significantly change the number of customers available for purchase) while having more elastic behavior near a closer price range. However, since the logistic function follows an S-shaped sigmoid pattern[22], the inverse formula is necessary to represent P(n).

$$P(n) = p_0 + \frac{1}{k}\ln(\frac{L}{n} - 1) \tag{13}$$

By finding the inverse formula to determine the price as a function of the number of users, the curve can approximate a cumulative demand curve in these scenarios, which will serve as a reference for subsequent analyses. For market equilibrium, reference is usually made, as

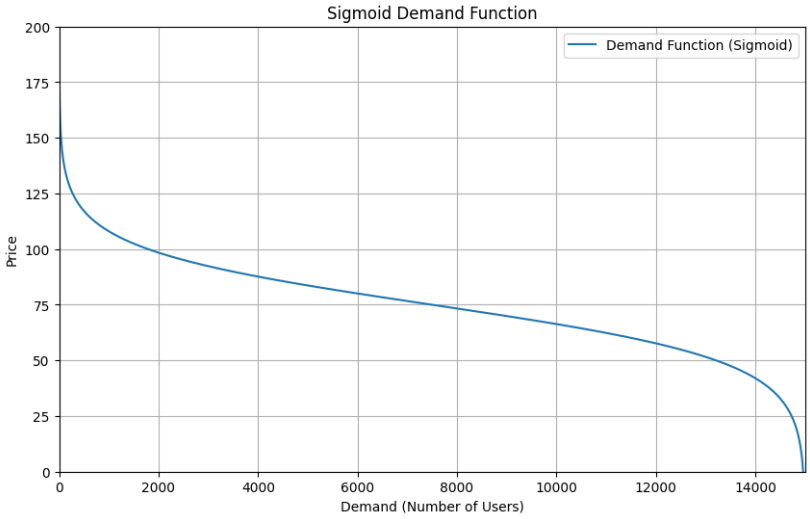


Figure 3: Logistic function for the Demand

expressed, to the supply curve. However, this occurs in perfectly competitive markets, while the context evaluated falls within the standards of monopolistic competition. In this case, thanks to the demand curve, it is possible to estimate marginal revenues and find the optimal number of users at the point where these intersect with marginal costs. Projecting the optimal number of users onto the demand curve also allows the optimal price to be found.

2.1.6 Optimum pricing

Having already obtained the equation for the demand curve (the inverse of the price from the logistic function) and the marginal cost curve, the next step is to estimate the marginal revenue function in terms of demand, i.e., the difference in revenue with respect to changes in quantity. By calculating instant by instant to plot the curve, we can use the definition of the derivative to obtain the following formula:

$$MR = \frac{\delta}{\delta n}(P(n) \times n) \tag{14}$$

Solving this, recalling that $P(n)$ is the demand curve described by the inverse of the logistic function, we obtain:

$$MR = P(n) - \frac{1}{k}(\frac{n}{Q(n) - n} + 1) \tag{15}$$

The curve follows the same pattern as the demand curve, with the y-axis (price) intercept at the same point but growing at a slower rate, intersecting the x-axis (number of users) at a lower number of users than the maximum market slice.

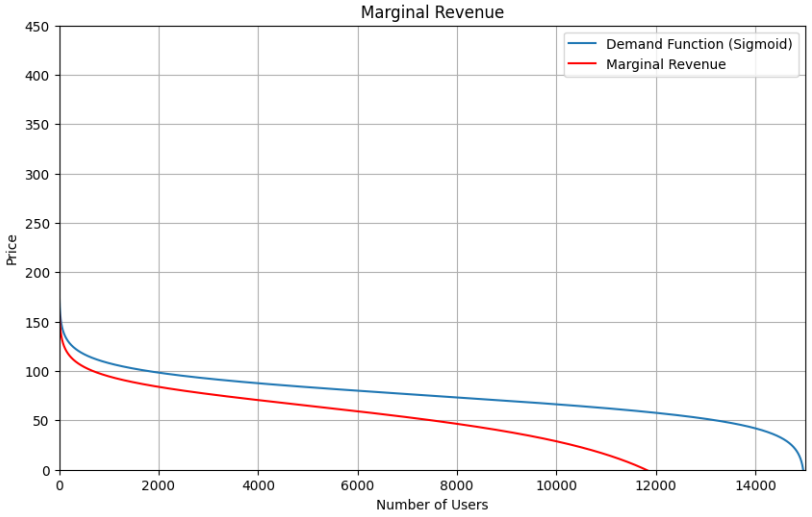


Figure 4: Marginal Revenue curve

To find the point where MC = MR, the marginal cost function can also be plotted on the same graph, along with the average total cost curve for profit calculation.

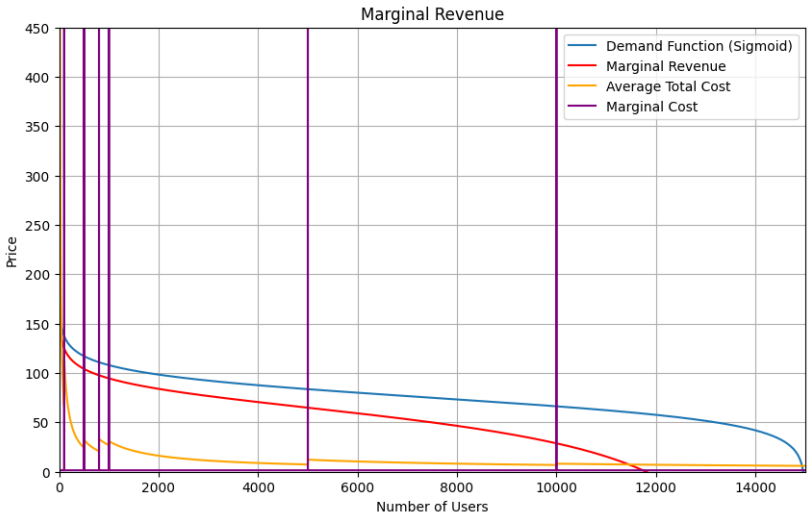


Figure 5: Marginal Revenue and Marginal Costs

It is immediately noticeable that, due to large increments dictated by the stepwise structure of fixed costs, the marginal costs intersect the marginal revenue curve multiple times, making the MC = MR relation valid at several points. However, the ultimate goal is to identify the point

where the area under the demand curve (price $\times$ number of users) is maximized. The price, as anticipated, is related to the coordinates of the projection of the optimal number of users found at the market optimum where MR = MC on the demand curve. This leads to the choice of the last intersection point, yielding a price of \$58.99 for 11,738 users. Given an average total cost at that point of around \$7.50, and rounding the price to \$59.99 due to the psychological influence of the number .99 on prices, we can determine a profit per user of approximately \$52.50. This results in a profit percentage of over 700% relative to costs, which aligns with the exponential growth needs required for survival in this type of market. It is important to note that this situation of monopolistic competition is only valid in the short term; over time, increasing competition lowers the demand curve closer to the average total cost curve, reducing profits and shifting the producer surplus to the consumer surplus. These profits and growth rates will serve for investment in research and development of new technologies to re-differentiate the SaaS product, repositioning it within the monopolistic competition market in the horizon of a new short-term period.

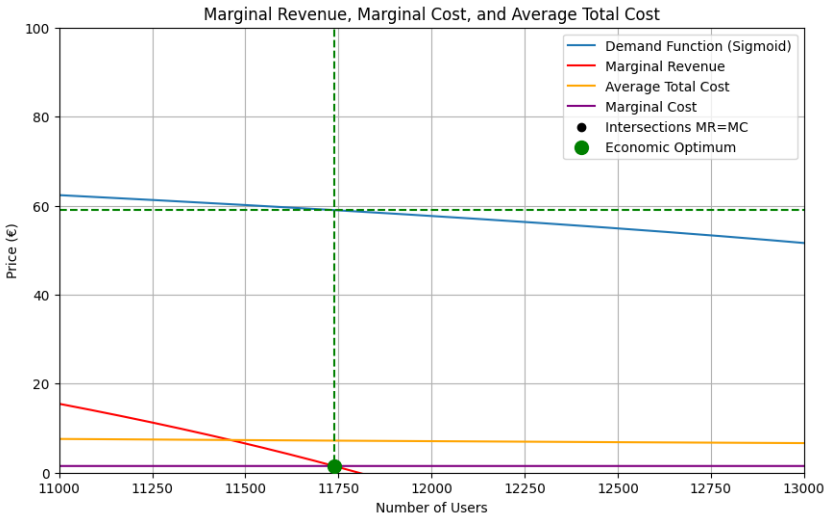


Figure 6: Market Equilibrium

At this point, having found the price, to complete the analysis, we can determine the Break-Even Point (BEP), where total costs equal total revenues, i.e., where the number of users — with the numbers used — is 198 users, corresponding to the point of intersection between the total cost and revenue curves, TC = TR. It should be noted that this analysis is *ceteris paribus* and not exhaustive of the real complexity, where MRR, LTV , and churn rate dynamically modify these figures.

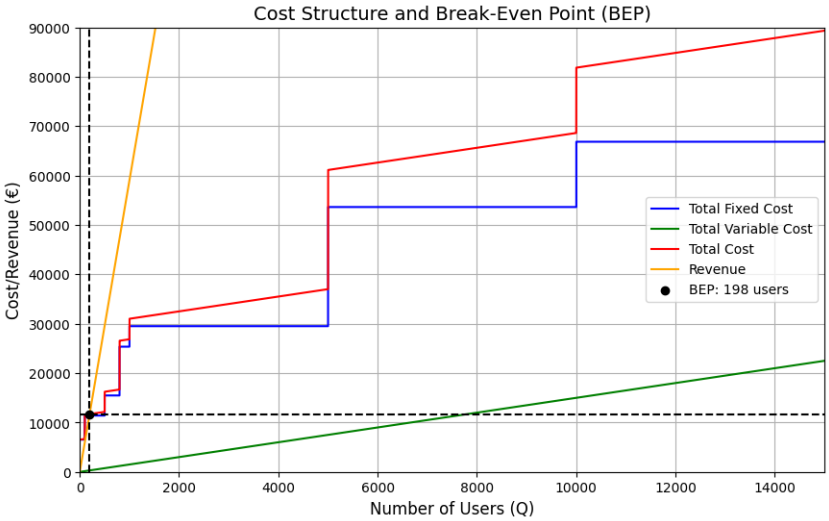


Figure 7: Break-Even Point

2.1.7 Challenges and Pricing models

While this analysis may make the expectations for this type of market appear highly optimistic, it is necessary to emphasize and further explore the challenges of a theoretical and academic approach. Before reaching such a market equilibrium, it is assumed that the company has achieved a stable condition, which typically takes years of success and growth.

The SaaS market presents several distinct points, which can be approximated to a simple model like the one presented but require prompt and active action from the company's executives in both planning and recurring monitoring. After all, this thesis aims to study inefficiencies, assigning them such importance that measures are taken to maximize profit as a precautionary tool against structural fluctuations in the system.

Even though variable costs are negligible, fixed costs make it essential to adopt a cautious approach, especially near level shifts. It is necessary to refer to the capital needed to cover costs during the business life cycle between acquisition expenses and the CAC payback period. Especially in the Micro SaaS model analyzed, where the solopreneur tends to bootstrap, covering these expenses becomes a critical factor. This type of structure, while presenting strong economies of scale with very low average total costs, depends heavily on demand, which is highly volatile.

This volatility is due to several factors related to acquisition and churn rates: the lack of entry barriers, technological evolution, competitor behavior, varying demand curve elasticity, and negative user responses due to administrative errors in customer service, marketing, or communication in general. These factors continuously affect both the demand curve and the market size and share. Moreover, the efficiency of the marketing structure can lead to more or less rapid user growth, and especially, non-linear growth. Simultaneously, competition or poor customer service can raise churn levels, causing user abandonment, most of whom will not be

reacquired, reducing market size. Additionally, the Funnel process[29] involves several variables, more or less directly controlled by the company, which can lead to inefficiencies that further alter the demand curve. Referring to Daniel Kahneman's studies on psychology and the real alternatives to the rational postulates of economic theories aimed at describing the behavior of the *Homo Aeconomicus*, in such cases, mathematical models only contribute to giving us the allure of the illusion of control.

Moreover, thinking about short- and long-term competition structure helps crystallize the concept of growth. In the long term, with technological evolution and the entry of new players into the competitive market, it has already been mentioned that the demand curve steadily approaches the level of average total costs. This also depends on the fact that such changes follow a simple S-curve behavior, indicating that the SaaS product offered is no longer desired. Therefore, it is essential to invest in research and innovation, bringing — if not changes — new additions to the product, exploiting economies of scope to reduce costs for otherwise distinct SaaS services.

All these challenges and difficulties, though not represented in the analyzed graphs, further underscore the theoretical foundation for the company's efforts in optimization and efficiency. In the SaaS space, inefficiencies related to customer relations and pricing choices have a greater impact than an unoptimized cost structure. Thinking about pricing models that do not rely solely on an MRR subscription helps manage and maximize both user engagement and profit.

Pricing Models

Setting a price for the customer, not the product. There is often reference to the Pareto distribution, which states that 80% of output comes from 20% of input. In this specific study context, one could say that 80% of profits may come from 20% of users.

This statement is not necessarily true, neither in percentage nor structural terms. In fact, it can be seen as a condition to aim for: finding the 20% of users on whom to allocate the company's SG&A costs would be an efficient way to maximize the ROI of economic efforts. It is certainly complicated to readjust the company's architecture, even considering the importance of the demand curve in its price sensitivity. However, the premise of monopolistic competition is precisely the ability to influence prices by the company. Consequently, by segmenting demand and formulating two different cumulative demands, it is possible to apply price discrimination with the structures that will be presented later in this chapter.

Yet, price discrimination is not the only way to increase profit potential through more complex structures. Various strategic solutions exist, aimed at specific cases derived from goals determined on precise KPIs. Just think of the psychological impact of prices, to which we are all now accustomed. Prices ending in 7, or .99, have a greater persuasive effect than other fig-

[29]The SaaS funnel represents the customer acquisition process, from service discovery to conversion into paying users, optimizing each stage to maximize conversions and reduce churn.

ures that we unconsciously perceive as more expensive. There are also graphic solutions on websites, such as placing the desired selling price in the middle of other options or highlighting it with brighter colors, lifting it upward, placing it next to two unreachable solutions, or validating it with labels such as "most chosen" to trigger our social animal biases. However, it is not just psychology and marketing that drive the assignment of greater value. It is said that the right price is one at which users complain about the excessive expense but are still willing to pay to use the service. It is not enough to keep the price above costs; it must be brought as close as possible to the maximum value the consumer assigns to it. However, this is not easy due to informational asymmetries that prevent the producer from having complete knowledge of where to set prices to maximize profits.

A *Pricing Model* refers to the model by which a business (in this case, SaaS) determines a structure according to which customers pay the value they receive from the company. As mentioned earlier, it is a fundamental element for companies, as it determines profit, the payback period, and the initial capital required to cover fixed costs before reaching the Break-Even Point. Additionally, other results related to customer-driven strategic planning are added. Below are some solutions to the problem by proposing targeted interventions for specific situations[23].

1. *Flat-rate*: This is the subscription system used so far in analyzing costs and monthly recurring revenues at a given price. That price is an integral part of the flat-rate model, meaning the monthly fee (subscription) paid to access the service. It is a standard, non-diversified fee, whose determination lies in finding the price/user ratio that maximizes demand as seen earlier.

2. *Lifetime access*: With subscription models, recurring monthly revenues are typically adopted. However, the impact of the CAC Payback Period at the start or in times of growth is so high that it encourages alternative solutions. One might be to show the monthly price but require an immediate annual payment to have the funds necessary to cover fixed costs upfront. Another is the *lifetime access* technique, offering a fixed sum that is convenient in the long run to gain more immediate capital. It is calculated based on LTV but is often a temporary solution, with costs amortized through the acquisition of customers gained from the capital obtained through this strategy.

3. *Freemium*: Freemium refers to a solution where customer acquisition is not necessarily compensated through funnel processes, but rather with a call to action aimed at registering for the service by offering free features. This can take different forms, such as a trial period (usually linked to entering a card to reduce the friction to monetization), a limited number of uses to "spend" whenever desired, or completely free access aimed at engagement and loyalty, potentially leading to a future upgrade. This concept lies in the aforementioned Pareto distribution, where 20% of paying users cover the costs of the

freemium users and generate profit. These structures require sustainable cost management.

4. *Tier-Pricing*: Unlike Flat-rate, although this is also a subscription model, there are different tiers, levels of spending depending on the features and higher benefits obtained by those willing to pay a premium. Higher tiers correspond to more services, faster speeds, support, and shared usage: all attributes that differentiate a single product thanks to features available to those willing to pay more due to the higher value attributed to the product. This model relates to the theory of price discrimination but is based on an arbitrary assumption of the various levels available on the demand curve, leaving the customer to choose whether to adhere to the fee or not. Compared to the flat-rate model, it certainly maximizes profit more, effectively replacing the consumer surplus of customers with a higher positive deviation from the flat-tier with the surplus of the SaaS producing company.

5. *Per-user*: A pricing type associated with paying a higher fee if a *workspace* wants to add another user's collaboration. In this way, the price increases each time a new user is added. This system is often paired with freemium solutions for the first customer, then transitioning to a paid plan for each collaborator. This system capitalizes on user engagement and the network effect of growing users through the free plan, then monetizes companies (with much greater price sensitivity than individual users) whose teams choose to collaborate together.

6. *Usage-based*: This is a *pay as you go* framework, where no recurring fee is charged. Instead, payment is made upfront for a service use equivalent to the customer's usage choices. Access to the service may last a month, a few days, or a year: MRR is not controlled, but by studying user behavior, an increase in other metrics (such as LTV) may be found due to multiple uses compared to the estimated monthly usage.

7. *Add-ons*: These are components or additional functionalities that can be purchased or activated to extend the base software's capabilities. These add-ons allow users to customize the SaaS service to their specific needs without having to change the entire platform, such as integration additions, premium services, or larger storage. This strategy is often used by offering add-ons as optional services at an additional cost to the base SaaS subscription plan, allowing users to pay only for the features they actually need.

The models presented are just a few of the common pricing models adopted by companies and are not intended to be exhaustive of the multiple possibilities available to a SaaS company. On the contrary, different structures are implemented depending on the problems the company faces[24]. For instance, by implementing various subsets of models simultaneously, a different structure can be achieved compared to adopting one individually. Often, the choice is not between one model or another, but on how to balance them together to maximize profit.

Moreover, temporary techniques can be applied, such as *price skimming*, where a higher price than value is applied at the launch of a new product to guarantee *early access*, monetize innovation, or leverage competitive differentiation. A price that then returns to expected values, following the sigmoid curve that describes the product lifecycle, already explained earlier. Conversely, another technique is *price penetration*, where a lower, more competitive price is adopted to attract the maximum number of customers and exploit the *bandwagon* effect, where demand increases simply because others are adopting the service.

In general, four main pricing strategies can be identified: value-based, competition-based, cost-plus (applying a target percentage above costs), and dynamic pricing. In the SaaS context and with the pricing tools just described, the priority is to increase the product's value for customers. However, thanks to Machine Learning and the techniques that will be explored further in this thesis, it will be possible to think of dynamic price assignment through user data collection. Dynamic pricing is often viewed with a critical eye, but when used appropriately in the context of service differentiation based on preferences, it can instead increase satisfaction and engagement by giving customers the freedom to pay the right value they themselves assign to the offered SaaS product.

Indeed, it is estimated that only 39% of companies set their prices based on such in-depth analyses, while 27% follow their intuition and 24% base their decisions on competitor pricing[25]. Analyzing and planning the effects of pricing strategies — making them flexible in response to feedback received — increases process and resource allocation efficiency to foster growth. To make informed decisions, it's important to conduct a total cost analysis, recognize the value different customer segments are willing to pay for the offered service, and conduct market research on the competition. Additionally, it's crucial to effectively communicate both differentiation and quality in attributes, as well as any price or policy changes, to maintain customer loyalty and reduce churn.

2.1.8 Data Management and Forecasting

It is essential to review metrics using the historical data collected by the company. By analyzing costs, prices, and profit variables, the importance of adopting a data management system becomes apparent with respect to the information that can be gathered from users and their behaviors in response to company communications, external environment, price changes, and attributes. Building a database of information and managing it properly to track analyses in support of decision-making means undertaking strategies and decisions driven by data, not just intuition. Certainly, one cannot expect historical data to lead to a certain prediction of future results, nor is it easy to recognize all the patterns and causal links that led a certain input to produce a certain output. However, reducing inefficiencies also means reducing risk, and in this, historical and real-time company data become essential tools for achieving the goal.

Managing data, protecting it, maintaining it, and hiring staff capable of extracting the necessary information comes with a significant cost to add to total costs. However, the ROI in return

has value in terms of profit and long-term sustainability: all metrics benefit, as determining their value provides the information needed to dynamically and timely adjust the company to market demands, avoiding errors and economic losses due to informational asymmetries (*revenue leakage*). Indeed, there is a need for a quick, automated response that is based on historical data but updated in real-time, especially in cases of analyzing the development of new strategies or understanding the impact of decisions, such as launching a new product, the lifecycle of an offer, changing prices, and *A/B testing* of these same topics.

The data in question is related to customer information that the business collects through direct surveys, reviews, feedback, or automations related to customer behavior (for example, using *cookies*). These data are collected precisely to better understand the effect of marketing strategies, increase sales, produce positively evaluated services by users, or improve customer service and retention rates. It is therefore important how they are processed and interpreted, as well as validated to provide useful and relevant information. Examples of information include personal contact details, demographic, psychological, and preference data.

These fields are necessary to segment users into groups as accurately as possible, with similar demand curves and different needs. Ultimately, these data serve to understand how to divide and differentiate offers (such as discounts), products, KPIs, communication, and prices.

2.2 Price Discrimination Theory

Typically, when discussing prices, it is done with a linear perspective on the price-quantity relationship. However, in reality, pricing is influenced by multiple variables, which can result in different prices depending on the volume purchased or specific agreements on bundled purchases. This ability to deviate from the economic inertia of natural efficiency (understood as Pareto optimality for a single price) is granted by a company's market power and, consequently, its ability to set its own prices. The inefficiency that allows for this market power is monopolistic behavior, as studied in the theories of *Dupuit* and *Cournot*[26]. However, other variations, such as monopolistic competition, also allow short-term price-setting power and the possibility of adopting non-linear or non-uniform pricing strategies, as seen in *price discrimination*.

Referring to the etymology of the word discrimination, the economic meaning attributed to differential pricing practices becomes clear: it is a process aimed at distinguishing and separating different market segments and consumers. Specifically, because in monopolistic equilibrium the company sets the price and demand dictates the quantity, the various components of demand can be used to segment groups of consumers with similar characteristics, extracting a maximized price-quantity ratio from them, which is more advantageous than choosing a more general price. The characteristics referred to may include demographics, income, psychological factors, or other elements that generally influence the utility expressed in the Berry, Levinsohn, and Pakes equation (10). As a result, the primary factor driving price discrimination is the ability to gather relevant information in order to identify each consumer's willingness to pay.

This differentiation, however, is bound to be imperfect as it is subject to sociological, psychological, and dynamic factors, translating the segmentation attempt into a probabilistic approximation[27]. Moreover, price discrimination is recognized by consumers themselves, and their behavioral responses may vary, sometimes leading to outright opposition when they realize they are paying more for the same service than others. For this reason, when applying differential pricing, it is crucial to understand the impact on customers and their engagement or loyalty. Clear and effective communication is essential, as is offering a distinction in the product's attributes to justify the price differences. Furthermore, price discrimination is often applied not only based on the personal characteristics of the consumer, placing them in a specific segment, but also based on market differences, such as the timing of purchase (e.g., high or low season for vacations) or geographic differences between regions or countries.

Given the discriminatory implications (even if legally permissible), some level of arbitrage must be accepted, allowing consumers to place themselves in a more favorable segment and thus save money, regardless of their individual characteristics. For example, one might purchase a product in a market where it costs less or patiently wait for sales or periods when prices drop. This logic is important for preserving consumer utility, as they recognize the price paid as fair value. This concept is closely related to the notions of risk and opportunity cost, akin to how interest rates compensate a saver for deferring the use of their money. In this case, the consumer's efforts to obtain a lower price in a price discrimination system are rewarded by the surplus gained from the price difference. However, despite the system's built-in balance, companies invest resources to determine whether or how much they should reduce or counteract this arbitrage.

2.2.1 First-degree, second-degree, and third-degree price discrimination

The profitability condition behind this differentiation in pricing models is based on the economic assumption that customers who value the service more are willing to pay a higher price compared to the uniform price applied across the entire market. To achieve this segmentation goal, specific socio-economic conditions must be in place, such as the legality of this type of distinction and the market power that allows the company to influence both the price and the consumer's willingness to pay. The necessary conditions for this phenomenon, however, are further influenced by the amount of information the company possesses and the degree of arbitrage and resale power available to the consumer. In fact, if a customer who pays a lower price is able to resell the purchased good at a higher price to another customer whose demand is greater, and if the resale price is lower than what the company charges the higher willingness-to-pay customer, then doing business with the company would lose its appeal, and market demand would be entirely disrupted.

The varying degrees of balance between the amount of information gathered and the impossibility of arbitrage or resale divide price discrimination into different levels of implementation: for instance, perfect knowledge of each customer's willingness to pay would result in a perfect

transfer of consumer surplus to the producer, while an almost complete absence of information requires entirely different measures. This subsection will explore the three degrees of price discrimination and the measures adopted according to the scenario[28].

First-degree discrimination

First-degree discrimination refers to perfect price discrimination, where each customer knows and communicates their exact willingness to pay (WTP) for a service to the company. In this case, resource allocation is efficient, with no deadweight loss or inefficiency. However, this form of discrimination is the most extreme of the three and is difficult to implement in practice, remaining mostly theoretical. No consumer, and consequently no company, has a perfect understanding of every individual's WTP, making this perfect discrimination nearly impossible in real-world markets.

Nonetheless, it is important to study this model to better understand pricing dynamics and the ideal outcome that can be aimed for with the other two degrees. By capturing the entire consumer surplus, the company maximizes profit as if the producer could hold an auction with all market participants, allocating each good to the highest bidder. In reality, a framework similar to this can be observed in auction markets or luxury and highly specialized services, where the seller can negotiate based on the client's willingness to pay for highly personalized services.

While this model is extremely beneficial to companies, it generates significant dissatisfaction among customers, who may perceive the distribution of benefits as unfair, potentially leading to behavioral responses like withdrawal from purchasing. Furthermore, this model requires a total ban on arbitrage and resale, along with an absence of substitute goods or general competition.

A more realistic variation of this perfect model involves the adoption of a *lump-sum fee*, where a fixed fee guarantees access to the transaction. The objective of first-degree price discrimination lies in the company's desire to capture all of the consumer surplus and replace it with profit. By applying a lump-sum fee, a portion of the surplus is extracted through a fixed charge, in addition to revenue generated by the traditional formula of $price \times quantity$.

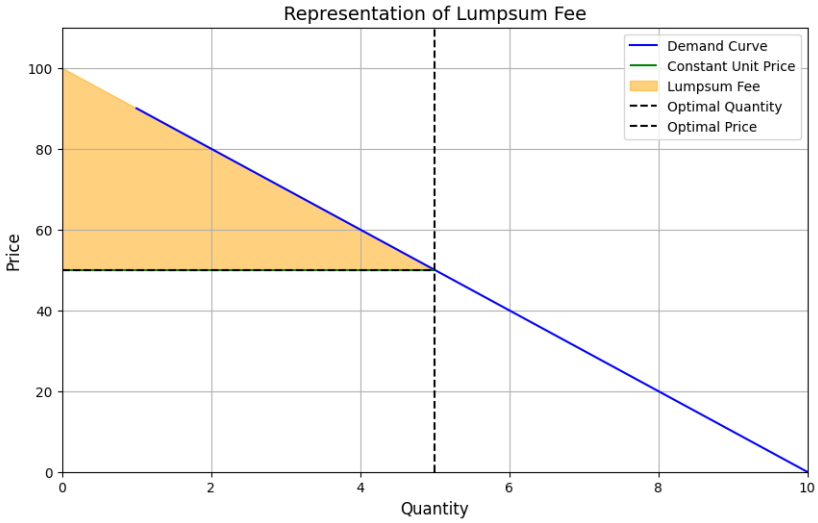


Figure 8: Lump-sum Fee

Second-degree discrimination

Second-degree discrimination, on the other hand, can be defined as a form of self-selection, where customers choose from options provided by the company. Prices vary based on the quantity consumed or the bundle of goods or services purchased, and the customer selects the solution that best fits their needs and willingness to pay.

Unlike first-degree discrimination, prices remain uniform but are based on choices between different values. Since the company lacks knowledge of each consumer's WTP, it identifies different segments by analyzing demand separately and assigning a price to each segment. However, the price offering is static, and the decision to adhere is left to the consumer. This reduces the perceived fairness issues discussed earlier, but at the same time, it is less profitable for the company, which cannot fully capture the consumer surplus. This is, of course, provided the price reflects the quality for which consumers are willing to pay more.

For example, in the transportation market, second-degree discrimination is implemented by selling first-class, business, or economy tickets. If those who are willing to pay significantly more for a premium class find that the benefits do not justify the price (such as having the same comforts as standard economy), the consequences may be similar to the perceived unfairness of first-degree discrimination.

In this case, the lump-sum fee theory can also be applied, as companies seek to maximize tariff payments relative to willingness to pay and demand curves. By identifying two types of consumers and their respective demands (with one being higher than the other), a company can reduce the surplus of the higher-demand segment, which would otherwise exist with a uniform tariff. The company offers two different pricing plans, one with a higher fixed tariff and lower variable price, and the other with the opposite structure. This model is called a two-part tariff, consisting of two distinct pricing components: $P = tariff + price \times quantity$. Examples

include businesses that charge a subscription fee for service access and then add a variable price based on the quantity purchased. To work effectively and avoid a segment purchasing the other's plan, companies must stimulate adherence based on specific segment preferences, encouraging higher usage consumers to pay the higher tariff balanced by a lower variable price (capturing what would otherwise be excessive consumer surplus), or the reverse, with higher variable prices for those expected to consume less.

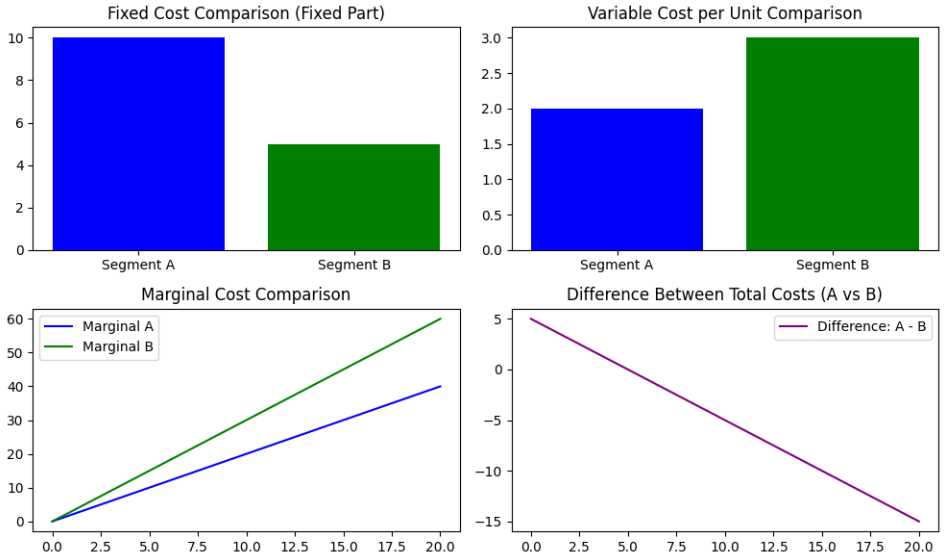


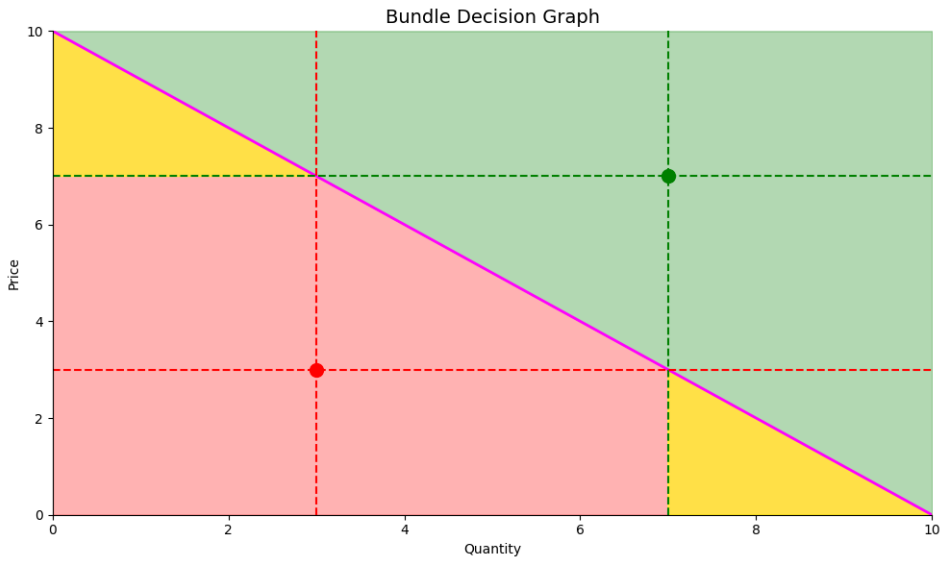
Figure 9: Two-part tariff

There are also various types of second-degree discrimination associated with *tie-in sales*, where the purchase of one good is contingent on purchasing another. In this case, the company can combine products or services, charging more to those willing to spend on both, albeit at a discounted rate, extracting surplus that would otherwise not be gained (by encouraging the customer to purchase more than one good). There are two types of tie-in sales:

1. *Bundling*: A practice of grouping two or more products into a single package sold at a combined price. This can be further divided into pure bundling, where products can only be purchased together, and mixed bundling, where products can be purchased individually or as a package.

2. *Requirements tie-in*: In this case, a low-priced good is linked to the obligation to purchase a related product at a higher price. This allows the company to extract surplus from customers who use the product more by discriminating on the price of the consumable.

By employing these tie-in sales practices, companies gain an additional way to differentiate prices according to consumers' WTP. However, the choice to accept or reject such pricing is still left to the customer, unlike first-degree discrimination, where the company imposes the price based on perfect knowledge.



Green section: accepts the bundle, red section: declines. Yellow triangle at the top purchases only good x, bottom yellow triangle purchases only good y

Figure 10: Bundle decision graph

Third-degree price discrimination

In the case of third-degree price discrimination, the company does not possess detailed information about each individual consumer's willingness to pay. However, it can segment customers into groups based on observable characteristics such as age, geographic location, or socioeconomic status. Once these distinct demand groups are identified, the company applies differentiated pricing to maximize profits by reducing the surplus in each group.

Price discrimination of this kind can only occur in imperfectly competitive markets, where the company has sufficient market power to set prices. The optimal point is reached when $MC = MR$ for each group. Given that the company has a single cost structure, we can assume that marginal costs are the same across different segments: $MC_1 = MC_2$. Therefore, equilibrium is reached when $MR_1 = MC$ and $MR_2 = MC$.

Considering two distinct demand curves, the profit equation can be represented as:

$$\pi = [P_1(Q_1) - MC]Q_1 + [P_2(Q_2) - MC]Q_2 \tag{16}$$

Since $MC_1 = MC_2$, we can rewrite the marginal revenue equation as a function of elasticity:

$$MR = P(1 + \frac{1}{\epsilon_d}) \tag{17}$$

From this, it follows that we can relate MR_1 to MR_2, as both are equal to MC, albeit with

different quantities:

$$P_1(1+\frac{1}{\epsilon_1}) = P_2(1+\frac{1}{\epsilon_2}) \tag{18}$$

Where a higher elasticity in absolute terms implies a lower price, and vice versa:

$$|\epsilon_2| > |\epsilon_1| \implies P_2 < P_1 \tag{19}$$

Finally, the *markup* (i.e., the gross profit margin that the company adds to the total costs of the product) applied by the company will be greater for consumers with less elastic demand since they are less sensitive to price changes:

$$\frac{P_1 - MC}{P_1} = -\frac{1}{\epsilon_d} \tag{20}$$

When implementing this type of discrimination, the company can increase prices only for the less elastic demand segment to avoid losing a significant number of customers in the more price-sensitive segment.

2.2.2 Profit maximization through differentiated pricing

As outlined in the definition of price discrimination and explored through the three different degrees of differentiation, the core objective of differentiated pricing is profit maximization. This goal is achieved by shifting consumer surplus toward the producer. This transfer occurs by reducing the gap between the consumer's willingness to pay and the price set by the company. For example, in third-degree price discrimination, it becomes clear that segmenting the market into groups with similar WTP allows the company to propose different prices, bringing each price closer to the value perceived by each segment, thus reducing their surplus[29].

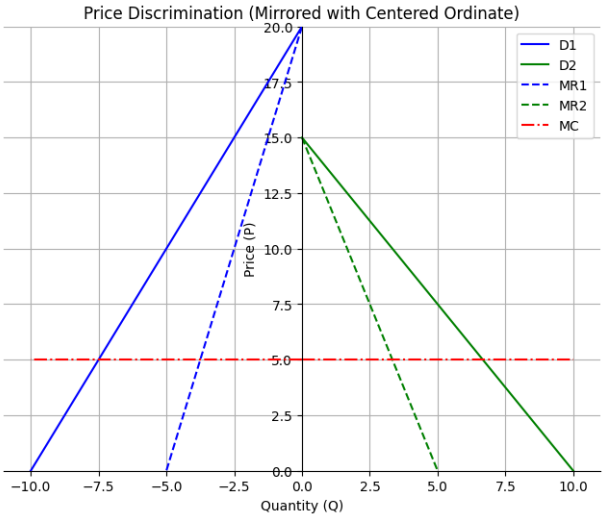


Figure 11: Third Degree Price Discrimination

By comparing marginal revenues with marginal costs and projecting onto demand curves,

the optimal price is determined in accordance with varying elasticities, leading to different prices for different segments.

Although the methods vary, each pricing strategy contributes to reducing consumer surplus in favor of the company. This means that these tools can be combined into broader strategies derived from a balanced approach. For example, to maximize profit, companies may use second-degree price discrimination to establish price levels or bundles, segmenting demand according to different willingness to pay (WTP). In addition, they may offer variable consumption options or apply a two-part pricing model based on these analyses.

The true challenge lies in ethical, communication, and administrative aspects. A company must maintain a good reputation to remain viable in the long term. Overuse of price discrimination can lead to feelings of injustice[30] among customers, potentially resulting in reduced revenues. Therefore, it is important to consider the long-term implications of differentiated pricing strategies. While such strategies may yield higher profits in the short term, they can influence brand perception and customer loyalty. Practices perceived as unfair or manipulative may damage the company's reputation, causing a loss of market share. Consequently, companies must develop strategies that are not only economically efficient but also ethically sustainable and transparent to consumers.

Moreover, there is a trade-off between efficiency and complexity: implementing and managing too many pricing strategies increases room for error, introduces uncertainty during the purchase process, and becomes more difficult (and costly) to manage and predict. Companies must invest in better technology, rapid and effective decision-making processes, and specialized personnel to manage these complexities. Even large companies, despite having the resources and the need for efficiency, often prefer to focus on one or a few diverse pricing strategies to fully exploit their potential.

2.2.3 Future Prospectives and Technology

Predicting the future is complex, both in economic and socio-political terms. Various factors can influence and alter a company's decision to adopt differentiated pricing strategies.

Ethical considerations, as discussed earlier, are one such factor that must be taken into account. While maximizing profits is a positive economic driver because it stimulates investment and production, which benefits society and fosters systematic growth, eroding consumer surplus can have negative effects on consumer responses. However, there are also positive ethical implications: making a product or service accessible to lower-income consumers because wealthier consumers pay higher prices can be seen as an effective redistribution mechanism. Without price discrimination, as in perfect competition, lower-income groups might be excluded from accessing the company's products or services. Thus, differentiated pricing, in an economic sense, can also represent personalization and inclusion, offering a broader and more accommodating system for varying needs and utilities.

From a technical standpoint, the adoption of advanced technologies is critical for imple-

menting differentiated pricing strategies efficiently. Machine learning algorithms can analyze large volumes of heterogeneous data, identifying patterns and trends that inform pricing decisions. Geolocation can tailor offers based on a consumer's geographic location, accounting for factors such as local purchasing power, competition, and cultural preferences. Big Data plays a crucial role in real-time information processing, enabling companies to respond quickly to market changes. The increasing speed of data transmission, thanks to advanced network infrastructures like 5G, facilitates information exchange and strategy synchronization across different channels and platforms. This level of digitization enables real-time dynamism in pricing strategies, enhancing a company's agility and responsiveness.

This advanced technological environment fosters production and innovation because the additional economic surplus can be reinvested into research and development, product improvements, and expansion into new markets. However, it also introduces significant risks. Regulatory unpredictability is one of the main sources of uncertainty. Privacy laws, data protection regulations, and competition laws may limit the use of certain pricing practices or restrict access to crucial information. Additionally, ethical concerns about algorithm-driven decisions can lead to regulatory interventions or reputational damage. To address these risks, companies must develop strong compliance strategies and adopt an ethical approach to using technology. This includes transparency in pricing practices, adherence to data protection regulations, and the adoption of ethical principles in artificial intelligence, such as avoiding algorithmic bias and ensuring fairness in automated decisions.

The evolution of business models is another critical aspect of future prospects. Companies must be ready to adapt to a rapidly changing environment characterized by new technologies, shifting consumer preferences, and new forms of competition. Innovative business models, such as digital platforms, the circular economy, or SaaS services, offer new opportunities but also require rethinking traditional strategies. Investing in digital skills, adopting agile methodologies, and fostering a corporate culture of innovation are key elements for remaining competitive. Partnerships with other businesses, tech startups, or research institutions can accelerate the adoption of new technologies and provide access to specialized expertise.

2.3 Application to the SaaS Context

The synergy between the concepts of price discrimination and the SaaS business model perfectly aligns with both's most valuable characteristics. On the one hand, the SaaS model requires structures that can ensure the highest possible profit to sustain growth and mitigate the market's volatility and sensitivity. On the other hand, the complexity of price discrimination is made feasible and simplified by the ability to track necessary differentiation information through digital behavior monitoring, thanks to technologies such as machine learning.

The integration of these two fields merges methodology with the practicality of financial and economic needs, allowing machine learning tools to provide the desired insights.

2.3.1 Importance of pricing applied to the business type

Revenue management is a fundamental component for the sustainability of a SaaS business. Adopting price discrimination models optimizes the critical factors already described in the introduction to this type of business. It enhances efficiency in areas such as revenue maximization, increased customer loyalty, optimization of LVC (Lifetime Value of Customer), reduction of market saturation risk, and continuous innovation.

The power of differentiation techniques provides effective tools to influence these metrics. The ability to segment the market and propose differentiated pricing helps accumulate scalable growth without losing any revenue—profits that can be reinvested in innovation.

2.3.2 Differentiated pricing strategies: review

In section 2.1.7, various common pricing options were presented, many of which are already forms of price discrimination.

Tier-based solutions can be structured under the second-degree discrimination model, where users self-select a tier to access more or less premium services based on the functionalities they wish to purchase. In SaaS, pricing communication must clearly convey the value and quality of these premium features to justify the price increase compared to basic services. With low marginal costs and overall product provision costs, adding these premium services increases total profits significantly without substantially raising the cost base, resulting in a significant difference in MRR (Monthly Recurring Revenue) or ARR (Annual Recurring Revenue).

Implementing multiple solutions simultaneously is also viable and can lead to greater efficiencies, as long as complexity is managed and kept within a positive ROI. For example, a freemium model is an excellent way to allocate resources through price division. The entire market interested in the product can access it, regardless of their demand elasticity. In addition to economic benefits and higher conversion and engagement probability, the influx of new users can bring significant value in terms of modern information currency: having access to extensive data allows companies to build better analysis models to structure more targeted offerings. Likewise, selecting a basic tier can contribute to segmenting and defining demand curves, applying third-degree price discrimination based on demographics or customizing offers, bundles, or behavior-based pricing using website usage data.

Identifying users within a specific segment creates a shared value increase mechanism between the customer and the SaaS. This is especially true when data collection enables the creation of models capable of dynamically automating these decisions.

2.3.3 Digital Infrastructure and Data Analysis

The technological foundation of the SaaS product allows precise user behavior tracking. Since the service is exclusively internet-based, it becomes easy to administer surveys, analyze behaviors, reviews, purchasing tendencies, and historical data, enriching the company's databases

with vast amounts of information. Properly structuring and analyzing these data based on the required metrics significantly reduces risks and information asymmetry.

Being a digital-based product also means that real-time forecasts can be applied, and automatic or testing systems, such as A/B testing, can be deployed to gather outcomes of different differentiated strategies. These data can train a machine learning model to recognize new, untracked users and place them into a complex network of segments dynamically and immediately, based on their behavior while using the software.

For this reason, in the next chapter, the machine learning techniques and algorithms useful for fulfilling this role will be presented, formalizing the theoretical knowledge that will be used to build the model in response to the research question.

3 Machine Learning Approaches and Data Analysis

From the study of the SaaS business model and the exploration of price discrimination, the importance and necessity of managing large volumes of data and addressing complexities flexibly and dynamically become clear. Technological advancements enable us to gather and model these data at such speed that we can track them in real time, while computational capabilities of computers, replacing human mathematical skills, allow a better approach to complexities with reduced margins of error—provided there is clarity about what one seeks and how to find it.

Machine Learning (ML) can be summarized precisely in this way: mathematical, probabilistic, and statistical algorithms that guide a computer to find solutions and answers independently, rather than by explicitly programming exact instructions. Having a computer capable of autonomously and rapidly producing complex results, predicting the behavior of new inputs never before analyzed, enables a more fluid and dynamic decision-making process, if not directly automatable: speed and accuracy have become essential in such a competitive and evolving world.

In this chapter, the technical specifications of data management, analysis, and Machine Learning algorithms that will be useful in constructing the dynamic price discrimination model proposed in the research question will be introduced and analyzed.

3.1 Introduction to Machine Learning in the Context of Pricing

It is important to clarify that machines do not learn in the human sense[31]. What Machine Learning (ML) truly does is identify mathematical formulas that generate the expected output based on the provided inputs. It is a problem-solving process that, through a large amount of data, builds a statistical model to reproduce results by detecting patterns during training. In contrast, when humans learn about the real world, they can reach the same conclusions even when altering initial conditions or situations. However, changing the inputs given to a computer results in a drastically different outcome from Machine Learning models. Thus, the term ”Machine Learning” can be considered a strategic marketing move by IBM, but it accurately reflects the binary approach of attempting to replicate human-like thinking.

Although Machine Learning has already been classified as a subset of artificial intelligence, this distinction alone is insufficient to fully grasp the potential of ML algorithms in the realm of Revenue Management. It is necessary to delve deeper into its nature to study its approaches and theoretical foundations, which will help build strategic planning for solving pricing problems. If Machine Learning is essentially about rigorously solving specific problems, then every formulation of a different problem will require an algorithm tailored to its unique nature.

3.1.1 General Introduction to ML

There are various types of learning: supervised, semi-supervised, unsupervised, and reinforcement learning. In this thesis, algorithms from each of these fields will be presented, as it will be necessary to construct models capable of both identifying hidden patterns in business data and recognizing known characteristics to effectively manage the segments on which to apply differentiated pricing strategies.

Supervised learning is an ML approach where the model is trained with a set of input data and a corresponding set of labeled outputs. Labeling is typically manual or guided by matching clear input data with certain derived outputs. Training such a model, for example, using linear regressions, decision trees, or Support Vector Machines (SVM), allows it to predict future outcomes based on identified inputs.

In a supervised system, the training dataset collects labeled examples $\{(x_i, y_i)\}_{i=1}^{N}$ where x_i represents the feature vector, and y_i represents the corresponding label. In unsupervised learning, x_i remains the feature vector, but the dataset contains unlabeled examples $\{x_i\}_{i=1}^{N}$. This means unsupervised learning aims to create a model that transforms the feature vector x into another vector or value suitable for solving a practical problem. Later in this chapter, clustering algorithms will be discussed for segmenting various market demands—these clustering algorithms are classified as unsupervised models, transforming the feature vector x into the cluster ID for each x_i.

Semi-supervised models contain both labeled and unlabeled elements in their training datasets, with unlabeled data typically much more abundant. Labeled elements are crucial for the goal of feature vector labeling, such as converting user characteristics into predefined labeled categories. Adding unlabeled features is essential for the algorithm to understand hidden patterns invisible to the human eye, ultimately improving the model's results. The combination of these two models is necessary to understand the objectives and nature of the subsequent analyses.

Reinforcement learning, on the other hand, is a type of learning where an agent interacts with a dynamic environment, making decisions in response to various situations and receiving feedback in the form of rewards or penalties. The goal is to maximize total rewards over time. This method is particularly useful in dynamic and complex contexts where decisions must be continuously adapted based on market reactions.

Machine Learning is therefore the field of AI focused on data learning: the more ML models are exposed to data, the better and more precise their performance. Data acts as fuel for the learning system and plays a central role in model development. Accumulating large quantities of data enhances the effectiveness of the models, but quantity alone is not enough to improve performance—data quality, precision, completeness, and clarity are critical to communicate meaningful information that models use for predictions. This is why the initial stages of the *data pipeline*[30] are so crucial, including data ingestion, collection, cleaning, and transformation

[30] A structured and sequential process for handling, transforming, and analyzing data

into formats suitable for model training.

Finally, before diving into the study of models, it is important to understand the significance of model training in terms of quality and evaluation. As will be further explored, ML algorithms are driven by parameters and variables that can be modified and are determined during the development phase. These parameters define how the model operates and processes the data during training, meaning design errors can render a model ineffective. Overfitting occurs when a model learns the training data too well, failing to generalize patterns for predicting new data, while underfitting refers to a model's inability to even capture the patterns present in the training data. Evaluating a model's performance is therefore essential to benefit from its use: measuring performance quantitatively allows for empirical selection of initial parameters, assessing them based on accuracy, mean squared error (MSE), precision, recall, and other relevant metrics for the specific algorithm employed, including business metrics.

3.1.2 Overview of Machine Learning Techniques for Dynamic Pricing

Dynamic pricing, defined as pricing based on real-time data gathered from a SaaS product, is made possible through the interaction of business economic units with ML algorithms. However, dynamic pricing in the SaaS context differs from the dynamic pricing found in industries offering "exclusive" products—where the use of a product prevents others from using it, such as books sold on Amazon for thousands of euros or flight tickets with fluctuating prices based on demand. These industries can assign products to those who value them the most.

In SaaS, dynamic pricing must balance the goals of maximizing profit for the company and ensuring fairness for the user. SaaS products do not have characteristics of exclusivity since they can be sold and scaled indefinitely. There is no resource scarcity, and any price variation must be justified by the addition or removal of services. The perception of unfairness can increase churn rates, which directly contradicts profit-maximization goals.

Although machine learning models are trained on historical datasets, the real objective is to use them to make predictions on new, unseen data. This is feasible because inputs placed near each other in the feature space[31] tend to produce similar outcomes. In other words, if new real-time data is classified and positioned near elements for which the model has already learned the correct output, it can estimate the outcome with a certain probability. This closeness in the data space allows the model to predict the output of unseen data with varying degrees of probability.

To optimize dynamic pricing using machine learning algorithms, which will be discussed in detail in subsequent sections, various techniques are grouped into the following subsets:

1. *K-means Clustering:* Used to segment customers based on observable characteristics such as income, usage behavior, and churn propensity. This unsupervised algorithm identifies homogeneous groups of customers with similar behaviors, enabling the creation of

[31] In ML, the feature space is a multidimensional space where input data is mapped based on its features or attributes.

personalized pricing strategies for each segment. In a SaaS context, customers can be grouped into segments with different price sensitivities, facilitating dynamic price discrimination.

2. *Inverse Logistic Regression:* Applied to estimate the demand curve and model the relationship between price and demand. This approach is useful for predicting how price variations impact the probability of purchase, helping define optimal price tiers for each customer segment.

3. *Conjoint Analysis:* A tool used to determine which product attributes (such as premium features or additional support) are most relevant to each customer segment[32]. Conjoint analysis helps understand which factors influence the customer's perception of value the most, supporting price variation justification and upselling strategies.

4. *Semi-supervised Models:* Used to automatically label customers into predefined segments. These models leverage both labeled and unlabeled data to improve the accuracy of customer classification in real time. As new data is gathered through A/B testing and conjoint analysis, the model becomes increasingly precise in predicting customer behavior.

5. *ARIMA/Holt-Winters:* Time-series models used to predict key SaaS business metrics such as Annual Recurring Revenue (ARR), churn rate, and Customer Lifetime Value (CLV). These forecasting models help monitor trends in customer behavior and optimize dynamic price thresholds based on demand fluctuations or churn risks.

Another important link between machine learning and dynamic pricing is the feature engineering process, which involves extracting and transforming the most relevant variables from the available data. Data is the fuel that drives the machine learning engine—a well-executed feature engineering process can significantly improve the model's effectiveness, enabling a more accurate representation of user behavior and, consequently, more precise predictions of optimal prices. The creation of features can be enhanced by dimensionality reduction techniques such as Principal Component Analysis (PCA), which eliminates redundancies and identifies the most influential features. For this reason, integrating external data into dynamic pricing models is crucial: economic trends, competitor pricing, and industry trends directly affect costs, customer willingness to pay, and the technological innovations demanded by the market.

3.1.3 Analysis of Challenges and Opportunities in Applying ML for Price Discrimination

Applying machine learning for real-time price discrimination in SaaS offers a wide range of opportunities but also presents significant challenges. One of the primary issues is scalability over time: while initially requiring minimal resources, as the user base grows, models must process data from thousands of users simultaneously and in real time. This necessitates the adoption of

distributed systems such as cloud computing, capable of handling parallelized inference operations, which incurs costs beyond those already considered. Reducing latency is crucial—every pricing decision must be made within milliseconds. Techniques like *model compression*[32] can help improve performance by reducing computational complexity, while *caching*[33] allows for storing pre-calculated results for frequent scenarios, further speeding up the process.

To maintain model performance over time, continuous monitoring systems are essential. Models are susceptible to *model drift*[34], which occurs when customer behavior or market conditions change. Monitoring should include not only technical metrics, such as mean squared error, but also business metrics, such as revenue improvement or conversion rates. Implementing *automated retraining* strategies allows models to be quickly re-adapted when necessary to counter diminishing efficacy.

As the user base grows, company structure and decision-making processes also become more complex. If a company decides to adopt internal dynamic pricing systems, these models must interact in real time with business platforms via APIs. The models often need to work alongside business rules (e.g., predefined price limits) to ensure decisions comply with operational constraints. A continuous feedback mechanism, both automated and human-driven, is equally essential—customer response data to dynamic pricing must be reintegrated into the model to improve future performance. Key opportunities arise from using A/B testing and multivariate optimization[35]. Machine learning enables testing large-scale price combinations, while reinforcement learning allows for improving real-time pricing decisions, adapting them based on user feedback.

Lastly, future opportunities are promising: emerging technologies like deep learning provide new tools to analyze complex user behaviors and predict their willingness to pay. Federated learning[36] allows for training models without directly collecting user data, complying with privacy regulations. These innovations, along with approaches like AutoML, offer the potential to further improve the accuracy and scalability of real-time dynamic pricing models.

3.2 Data and its Analysis

Data is the core of all models, indispensable and important, although not sufficient by itself to lead to meaningful conclusions. Data provides a wealth of information from which ML

[32]Technique to reduce the complexity and size of ML models without compromising performance. It is used to accelerate inference and reduce computational resource consumption.

[33]Temporary storage of data or results to speed up future operations, especially in repetitive contexts or with recurring inputs.

[34]Phenomenon in which the accuracy and effectiveness of an ML model deteriorate over time due to changes in data, user behavior, or market conditions.

[35]A technique that allows for the optimization of multiple variables or parameters simultaneously, balancing potentially conflicting metrics (e.g., profit, retention, and customer satisfaction) to achieve the best overall business performance.

[36]A decentralized approach to training ML models. It does not require transferring raw data from client devices to global servers, thus enhancing privacy.

models, or mathematical, statistical, and probabilistic models in general, extract meaning and further insights useful for decision making. However, one cannot exist without the other: as complexity increases, raw data alone can provide a historical description, but it cannot explain why certain events occurred. Conversely, it is impossible to identify patterns and predict trends based on new future inputs when there is not enough (or adequate) information and data to derive such insights.

Data on its own is not enough, and depending on who manipulates and interprets it, it can lead to different, even biased, results despite the objective premises data analysis promises. However, the importance of data does not lie solely in its existence but in how it is collected, prepared, and analyzed. An incomplete or poorly managed data collection can lead to incorrect conclusions and ineffective models, even with advanced Machine Learning techniques. The data interpretation process requires a deep understanding of both the business context and the relationships between the variables considered. Without proper care in the data preprocessing stage, such as cleaning and processing, the model may suffer from bias or significant errors.

Moreover, the relevance and accuracy of data play a crucial role: a model built on unrepresentative or outdated data can lead to erroneous decisions. For example, in dynamic pricing systems, the use of outdated historical data may reveal past trends that are no longer relevant and ignore crucial information about current user behavior or market fluctuations. An effective Machine Learning model must therefore constantly adapt, relying on real-time data to remain relevant and useful in decision-making.

On the other hand, even with a perfectly curated dataset, it is essential to understand that analysis methodologies are never entirely immune to interpretative errors or inherent biases. The initial assumptions made during the model design phase, such as the choice of variables and analysis techniques, can influence the final results. Some algorithms, especially the more complex and less interpretable ones, like deep neural networks, can return accurate predictions but are difficult to explain, creating a potential risk to the transparency and reliability of automated decisions.

For this reason, the process of analysis and modeling should never be considered static or definitive. It is essential to continuously monitor and reevaluate models, adapting them to new data and market conditions. This cyclical approach, known as *model retraining*, ensures that decisions based on Machine Learning remain valid over time, reducing the impact of errors due to unforeseen changes in data or the business context.

Finally, it is important to remember that the collection and interpretation of data must always be accompanied by a clear understanding of business goals. The proper use of data and models is not just a technical matter but also requires an aligned business strategy and solid governance to ensure that the conclusions drawn are genuinely useful for supporting business decisions and avoiding the pitfalls of *data-driven biases*.

3.2.1 Data Pipeline

In the process of analysis and modeling, the data pipeline plays a fundamental role in ensuring the quality and efficiency of the entire workflow. A well-structured data pipeline allows for the collection, transformation, and availability of data in an automated and continuous manner, providing a solid foundation for subsequent analysis and modeling phases. The main stages of the pipeline are described below, but before addressing them one by one, it is worth pausing to define what a pipeline actually is.

A pipeline can be defined as a method by which raw data is ingested from multiple sources, transformed, and then brought into a *data store* such as a *data lake* or *data warehouse* to be analyzed[33]. Along this path, data is processed with transformation processes such as filtering, aggregation, and masking, ensuring the data is appropriately integrated and standardized for more accurate analysis. This is even more important when dealing with data destined for relational databases, where tables communicate with each other and need to find information with predefined characteristics.

These processes are managed by dedicated professionals such as *data scientists* or *data engineers*, meaning that as the volume of data and the size of the company and user base increase, the cost of maintaining a rigorous data-driven decision-making process also increases. However, the benefits in terms of efficiency and optimization through ML models are directly correlated. Additionally, the uses and functions of a properly managed pipeline are numerous and not limited to obtaining clean and organized data: many *use cases*, especially on the business side, exploit this information.

1. *Exploratory Data Analysis* (EDA): used to investigate and analyze the dataset to summarize its main characteristics, often using data visualization methods with dashboards and graphs. This process is useful for understanding the type of resources available to the company and for manipulating them to obtain the answers sought to strategic business questions. It is also useful for discovering patterns, identifying anomalies, and testing assumptions or hypotheses, often with the help of ML models.

2. *Data Visualization*: the process by which data is represented using various types of charts (bar, pie, infographics, etc.), with possible animations, interactions, and everything that generally helps to represent the complexity of the data in a simple and immediate way, favoring a better understanding of relationships and results to lead to *data-driven decision making*.

3. *Machine Learning*: once data is ready and standardized, with a known structure and easily accessible information, ML models can be trained with significant data, leading to results with fewer errors and anomalies.

If these are some of the many functions that benefit from proper pipeline management, it is also important to briefly present the steps that compose it, from collection to distribution. The

data pipeline can be as simple or as complex as needed, but it can generally be summarized in three main components:

1. *Data Ingestion*: Data can be collected from various sources (ingestion points) across the web, from both structured and unstructured data sources. In the SaaS context, these sources include the SaaS platform itself, tracking user behaviors, choices, and traffic, cookies, external resources such as competitors or markets (via APIs or *web scraping*[37]), surveys, emails, customer service reports, reviews, CRM platforms, web analytics, or social media tools, etc. These data are considered raw, making validation processes to check consistency and accuracy, as well as subsequent cleaning and standardization processes, essential.

2. *Data Transformation*: After raw data is collected, it needs to be transformed to become *business-ready*. The goal is to clean, merge, and optimize the data in preparation for subsequent analyses aimed at decision-making. Various data cleaning methods exist, as well as standardizations, both in metrics (e.g., a specific date format) and structure (e.g., formats such as JSON[38]). The transformation components are numerous, such as augmentation, filtering, grouping, aggregation, sorting, validation, and verification. Once transformed, the data is ready for use.

3. *Data Storage*: This final component of the data pipeline refers to how the transformed data is distributed. This involves choosing where the processed data is made available for analysis and use. Typically, this means storing them in storage systems such as data lakes or warehouses so that they can be accessed by data scientists or analysts, but this data can also be made directly accessible to applications, SaaS platforms, Machine Learning platforms, or other applications as API endpoints.

These stages ensure proper management of data, keeping it up-to-date and ready for use. This provides all the necessary information for decision-making, allowing companies to build analyses or ML models needed: every company must properly manage a pipeline like this to stay dynamic, agile, and flexible and to respond to market volatility and changes.

3.2.2 Data Preprocessing Techniques Overview

When developing a machine learning algorithm, it is necessary to remove noisy data, missing values, and non-standard formats. It has already been highlighted that one of the phases of the data pipeline is Data Transformation, where these steps are carried out to make the data business-ready. However, the data pipeline is a broader concept compared to preprocessing, which can be part of the Machine Learning Pipeline, discussed later. In fact, data preprocessing

[37] A method of extracting information, e.g., from a website's HTML, to store it in a database.

[38] A textual format for structuring data that transforms complex structures into a format easily readable by machines and humans.

can be configured within many different data analyses, whether for data to be stored or for data specifically sought for analysis in a model, as in this thesis. Without precise and valuable information, any analysis will be flawed and unusable.

Several characteristics can make data unusable or devoid of information, and this section outlines the various methods and techniques applied during data preprocessing to correct these erroneous characteristics[34]. The approach is linear, or rather progressive: treating a certain type of problem is possible only if a previous problem has been solved, so it proceeds step-by-step in an orderly path from raw data to standardized and usable data.

The first step in this direction is handling missing values. It's possible that, when collecting data, certain fields related to the same ID may be empty or missing, resulting in a lack of information that could render all other data collected under that ID unusable. You can choose whether to completely remove that row of data (whether it's a database, CSV, or other formats), or arbitrarily assign values based on the rest of the data, using an average or median. However, it is risky to apply an average/median without considering other factors first, as there may be outliers (anomalous data) that are significantly larger or smaller than the median, leading to an excessive deviation that makes these extreme values less meaningful for the analysis of more realistic probabilities.

The next step after handling missing values is managing outliers or anomalous values. These represent data that deviates significantly from the rest of the dataset. Outliers can negatively influence predictive models by distorting statistical analyses, averages, and other centralized measures. One technique for handling outliers is "winsorization," which limits extreme values to the highest or lowest percentiles, making them less influential. Another technique is replacing outliers with the median, as the median is less sensitive to extremes than the mean.

After handling outliers, the process moves to "feature scaling" or data normalization. Variables in a dataset often have different scales, making interpretation or comparison difficult. Normalization through techniques like the "Standard Scaler" (which brings data to a distribution with a mean of 0 and a standard deviation of 1) or the "Min-Max Scaler" (which scales values within a predetermined range, typically between 0 and 1) is essential, especially in models that depend on distances between data points, such as K-nearest neighbors or clustering algorithms.

A critical aspect of preprocessing is handling multicollinearity, which occurs when there are strong correlations between independent variables. This phenomenon can negatively affect regression models by increasing the variance of coefficients and making it difficult to interpret the importance of variables. To address this problem, correlated features can be removed, or the "variance inflation factor (VIF)" can be used to measure how much an independent variable is influenced by others.

Finally, there is the process of "feature encoding," which is crucial for transforming categorical variables into a numerical format usable by machine learning algorithms. The two main techniques are "label encoding," which assigns a unique integer to each category, and "one-hot

encoding," which creates a new binary column for each category, indicating its presence with 1 or 0.

It's important to note that these steps are not strictly executed once and then processed: the various phases may repeat to bring the manipulation to higher levels of standardization and cleaning. For example, it was mentioned that averaging data during the handling of missing data may vary depending on outliers. The averaging process can be readjusted after cleaning out anomalous values, allowing for the recalculation of the average without the obstacle of extreme values distorting the approximation.

3.2.3 Statistical, Mathematical, and Probabilistic Tools

Mathematical, statistical, and probabilistic tools play a key role during the data processing phase[35]. However, the benefits of these fields of study extend to machine learning models and all types of data interpretation analysis. This is because, fundamentally, every machine learning model is guided by principles derived from these disciplines, transforming abstract concepts into concrete tools for understanding data complexity.

Mathematics primarily provides the structural basis for the models. Every algorithm—whether it's linear regression, decision trees, neural networks, or support vector machines—is founded on mathematical equations that describe the relationships between variables. For example, a simple linear regression model is based on the equation $y = mx + c$, where y represents the dependent variable, x is the independent variable, m is the slope (which measures the incline of the line), and c is the intercept. These mathematical components serve to model relationships between variables and make predictions, but without a grasp of basic mathematics, interpreting these models becomes impossible.

Statistics, on the other hand, focuses on understanding and analyzing data, allowing us to precisely define trends, distributions, and correlations between variables. Descriptive statistics, for instance, help summarize data using measures of central tendency (mean, median, mode) and dispersion (variance, standard deviation), which are essential for understanding the behavior of the dataset. Inferential statistics, on the other hand, allow us to make deductions and predictions about an entire population based on a sample, using concepts such as confidence intervals and hypothesis testing. These tools are crucial for verifying whether our assumptions about the data are valid.

Probability theory plays a critical role in understanding uncertainty in models. Every prediction made by a machine learning model is never completely certain but rather an estimate based on probabilities. This concept is central, for instance, in Bayesian models, where the probability of an event is continuously updated in light of new evidence (via Bayes' theorem). Even in less explicitly probabilistic models, like neural networks, uncertainty remains a key factor: final results can be accompanied by a probability indicating how confident the model is in its prediction. Decision trees, when evaluating choices at each node, use probabilistic metrics like entropy or information gain to determine which variable is most useful for making a correct

prediction.

The importance of these mathematical and probabilistic tools extends beyond data analysis and model construction. They also help us evaluate model performance through metrics such as accuracy, precision, recall, and the ROC curve. These measures derive directly from statistical and probabilistic concepts, allowing us to compare different models and select the one that best fits the problem at hand.

It is crucial to understand that choosing and implementing a machine learning algorithm is not a simple mechanical application. Every model has underlying assumptions, often derived from statistical concepts, that must be understood in order to use the algorithm correctly. For instance, in linear regression models, it's assumed that there is a linear relationship between the dependent and independent variables and that the errors are normally distributed. If these assumptions are not met, the model may provide inaccurate or even misleading results.

3.3 Predictive Models for Price Optimization

For price optimization, after giving an overview of what machine learning is and the importance of data, it's crucial to clarify which ML algorithms can be used.

Algorithms are widely discussed, both in everyday life and in development and analysis. In this very paragraph, the importance of applying ML algorithms for the development of the thesis has been briefly mentioned, but what exactly is an algorithm? An algorithm is nothing more than a method, procedure, or rule—in other words, a process that solves a specific problem through a finite number of elementary steps[36]. It takes its name from the Arab mathematician Al-Khwarizmi[39], whose texts presented the mathematical procedures translated by Fibonacci. These mathematical resolutions—essentially procedures to obtain a calculated result—are now called algorithms, as they were referred to with the phrase "*dixit algoritzmi*" (Al-Khwarizmi said so).

These finite steps can be as simple as solving a division on paper or as complex as managing a financial market or AI-based LLMs. In this thesis, algorithms, through a known and developed process by their creators (but more often based on well-established statistical and probabilistic properties), can extract from large datasets the necessary coefficients, weights, and patterns needed to understand the properties and behaviors leading to various possible outcomes. For instance, if certain behaviors from users lead to churn, it is possible to predict a user's churn before it occurs by recognizing behavior patterns similar to the training set data on which these algorithms were trained to produce models.

Algorithms, therefore, become the mathematical foundation for predicting results. They require certain parameters, and in the course of this paragraph, we will explore data and feature engineering designed to identify the best parameters and data so that the model—through

[39] An Arab mathematician from the 9th century whose writings introduced the Indian numeral system, including zero.

training—can successfully complete its task, such as clustering or forecasting.

Before looking at the algorithms with a focus on price discrimination, we can start by analyzing the pipeline that characterizes the progress of an ML project, and then explore each approach in depth.

3.3.1 Machine Learning Pipeline

A Machine Learning Pipeline is a set of interconnected steps that automate and simplify the process of building, training, evaluating, and deploying machine learning models. This tool is fundamental for managing the complexity of models and ensuring precise and scalable solutions. The modular flow of a pipeline allows each phase of the process to be broken down into separate blocks, simplifying management and optimization[37]. For instance, data preprocessing, feature selection, and model training can be developed and tested independently, reducing the risk of errors and enabling rapid iterations.

A significant advantage of pipelines is reproducibility. By defining the steps and parameters within the pipeline, the process can be recreated exactly at any time. This is essential for achieving consistent results, especially in production environments. Additionally, pipelines improve efficiency by automating critical steps such as data cleaning and transformation, saving time and resources. As the complexity of data or models increases, pipelines can be easily scaled without needing to reconfigure every step from scratch.

1. *Preprocessing*: During this phase, data is cleaned by removing missing values, encoding categorical variables, and scaling numerical features. These steps are fundamental to ensuring the model can work with consistent and structured data.

2. *Feature Engineering*: Once preprocessing is completed, the next step is feature engineering, which represents the core of model optimization. Here, the most relevant variables are selected, or if necessary, new features are created to improve the model's predictive capability. This step requires a deep understanding of the domain and can make the difference between a high-performing model and an ineffective one.

3. *Model Selection*: After preparing the data, model selection begins. Depending on the type of problem (classification, regression, clustering, etc.) and the dataset's characteristics, one or more algorithms are chosen. This is where hyperparameter optimization plays a key role, allowing the model to be tailored to the problem's specifics.

4. *Model Training*: Once the model is selected, the next step is training, where the model learns from the training data. Here, the model identifies hidden patterns and relationships in the data, which it will use to make predictions. This step is critical to the model's success, but it requires time and computational resources, especially for complex models or large datasets.

5. *Model Evaluation*: After training, the model is evaluated using a set of metrics that vary depending on the problem being addressed. Metrics such as accuracy, precision, recall, or F1-score help assess how well the model is making predictions. In some cases, techniques like cross-validation are used to ensure the model generalizes well to unseen data.

6. *Deployment*: Once validated, the model is ready for deployment. This step involves integrating the model into production environments where it can make predictions on new data. In some cases, the pipeline also includes creating API endpoints, which allow external systems to interact with the model to obtain real-time results.

7. *Monitoring and Maintenance*: The work doesn't stop after deployment. Once in production, the model must be continuously monitored to ensure it functions correctly over time. As data and external conditions change, it may be necessary to retrain or update the model to maintain high performance.

Beyond modularity and efficiency, pipelines also facilitate collaboration among teams. Each step in the flow is well documented, allowing data scientists and engineers to work in synergy without needing to revise the entire process. Changes can be tracked and versioned, ensuring accurate management of the model's lifecycle. A Machine Learning pipeline not only automates and streamlines the process but also improves scalability, reproducibility, and collaboration, transforming model development into an orderly and reliable workflow.

3.3.2 Clustering, K-Means and DBSCAN

Until now, much has been said about clustering algorithms, without specifying the technical implications of this approach. Beyond its common literal meaning, in the world of machine learning, the term clustering is used to define those learning problems that assign labels to examples by leveraging unlabelled datasets. It falls within the realm of unsupervised learning, where, since the datasets are completely unlabelled, finding the optimal choice becomes more complicated compared to supervised learning categorizations, like the random forest algorithm. In fact, in unsupervised learning, there is a variety of clustering algorithms, and it's challenging to understand which is the best for each use case. For the purposes of this thesis, and referring to the most commonly used algorithms, it is important to delve into two algorithms with two different approaches: k-means and DBSCAN[38].

The k-means algorithm requires specifying a parameter k, which represents the number of clusters (or classes). Then, k random feature vectors called centroids are assigned within the feature space. When computing the distance from each example x in the dataset to these centroids (using metrics like Euclidean distance), each example is assigned to a specific centroid, namely the nearest one. In this way, by choosing the number of centroids, all the different examples will be classified as belonging to a given group: the one closest to that particular centroid. It is an iterative algorithm that continues until the clustering objective is achieved.

However, it is precisely this characteristic of arbitrarily assigning the number k that can be limiting, unlike what happens in the DBSCAN algorithm. Before delving into this other approach, it is important to explore how to determine the optimal number of clusters, especially for this type of algorithm that requires defining the number in advance. One can certainly understand it intuitively without "over-engineering" it. In fact, it is often possible to grasp, logically or structurally, the different groups into which various datasets are divided, while other times, an appropriate data analysis can reveal the relationships within the data with graphs that highlight the presence of well-defined clusters. However, in this case, a more formal method is preferred: the elbow method.

This method uses a graphical strategy to show the optimal choice of cluster number. The elbow graph shows the within-cluster-sum-of-square (WCSS) values on the y-axis corresponding to the different values of K (on the x-axis). Where, graphically, the curve increases its slope, forming an "elbow shape," that is the optimal number of clusters to choose for the k-means algorithm. It works because it calculates this WCSS, which is the total of the squared distances between data points and their cluster center. There are certainly more precise algorithms, like the silhouette score, which determines whether there are large gaps between each sample and all other samples within the same cluster or across different clusters. However, later on in this thesis, the elbow method will be used as it clearly shows a fair number of clusters, also aligned with the marketing and economic needs of the SaaS business.

Finally, it is also worth examining DBSCAN, which, unlike k-means, is not a centroid-based algorithm. In fact, DBSCAN is a density-based algorithm, meaning that instead of guessing how many clusters are needed, this algorithm requires defining two types of hyperparameters: ϵ and n DBSCAN begins by selecting an arbitrary point in the dataset and checks whether it is a core point by looking at its ϵ-neighborhood. If it is a core point, a cluster is formed, and DBSCAN recursively expands the cluster by checking the ϵ-neighborhood of each core point's neighbors. This process continues until no more points can be added to the cluster. If a point is not a core point and not a neighbor of a core point, it is marked as noise (although it may later be reassigned to a different cluster if it falls within the neighborhood of another core point).

One of the strengths of DBSCAN is its ability to automatically detect the number of clusters, as opposed to k-means, where this number must be predefined. It also performs well in discovering clusters of varying shapes, such as elongated or irregular structures, which k-means cannot handle as effectively.

However, DBSCAN does have some limitations. Its performance can suffer when the data density varies significantly across the dataset because the fixed value of ϵ may not work well in regions of different densities. Additionally, determining the optimal values of ϵ and n can be challenging, especially when dealing with real-world datasets where the structure is not well known in advance.

3.3.3 Forecasting with Linear and Logistic Regression

After identifying the different clusters, given a dataset, it becomes possible to leverage its components to solve prediction problems. These problems are known as Regression problems, which, unlike classification problems, involve predicting a real-valued label (called a target) given an unlabelled example. This regression problem is addressed by a regression learning algorithm—an algorithm that takes a collection of labelled examples as inputs and produces a model. This model then takes new, unlabelled examples as inputs and returns an output based on the patterns discovered during training[39].

There are two main types of regression: linear and logistic. Linear regression follows a Gaussian normal distribution, represented by a continuous line on a graph, while logistic regression follows a binomial distribution, returning a probabilistic classification between zero and one. As can be observed, the term "regression" is inherited from statistical theory, though it is essentially a classification algorithm. Nevertheless, both linear and logistic regression models are widely used and will be applied within this thesis as well. They are powerful algorithms for building predictive models and serve as foundational models for constructing further decision-making systems.

Linear regression is a foundational regression algorithm designed to make predictions based on the linear relationship between input features and a real-valued target. When we have a collection of labeled examples, $\{(x_i, y_i)\}_{i=1}^{N}$, where each x_i is a D-dimensional vector of features and y_i is a real-valued target, linear regression aims to approximate y_i as a weighted sum of the features in x_i, adding a constant term for flexibility.

The model for linear regression can be represented as:

$$f_{w,b}(x) = w \cdot x + b$$

where w is a D-dimensional vector of weights (parameters), and b is a bias term. The model learns to adjust w and b so that predictions closely match the real values y in the dataset.

To find the optimal w and b, linear regression minimizes the difference between the predicted values and the actual target values. This difference, called the *loss*, is typically measured using the *mean squared error (MSE)*:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (f_{w,b}(x_i) - y_i)^2$$

Minimizing this error ensures that the model's predictions are as close as possible to the real values in the dataset, effectively finding the line (or hyperplane for higher-dimensional data) that best fits the data.

The optimization process to minimize MSE uses *gradient descent*, which iteratively adjusts w and b by calculating the gradient of MSE with respect to these parameters. This approach is computationally efficient, allowing linear regression to scale well for large datasets.

Linear regression is favoured for its simplicity and interpretability, providing straightforward insights into the relationship between variables. It also has a low tendency to overfit, which means it performs reliably even on unseen data. Thus, linear regression becomes a robust choice in machine learning tasks focused on predictive modeling and will serve as a foundational approach in this thesis for constructing further predictive and decision-making models.

Logistic regression is a foundational classification algorithm, despite its name, which suggests it's a regression method. The name originates from statistics, as the mathematical formulation of logistic regression is similar to that of linear regression. In logistic regression, we apply it to a binary classification problem, though it can be extended to multiclass cases as well.

To frame the problem, in logistic regression we want to model y_i as a function of x_i, where y_i takes on binary values (e.g., 0 or 1). A linear combination, such as $w \cdot x_i + b$, spans from minus infinity to plus infinity, which doesn't directly map to binary values. Instead, we use a *sigmoid* function, also known as the logistic function, which outputs values in the interval (0, 1). This function is defined as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

where e is the base of the natural logarithm. This function gives us a probability, enabling us to classify x as positive if $f(x)$ is closer to 1, and negative if closer to 0. We adjust the linear model to fit within the logistic function:

$$f_{w,b}(x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

where w is a vector of weights and b is the bias term. By interpreting $f_{w,b}(x)$ as the probability of y being 1 given x, we obtain a probabilistic model suitable for classification.

To find the optimal values of w and b, logistic regression maximizes the *likelihood* of the observed data instead of minimizing a squared error as in linear regression. The likelihood function represents the probability of observing the training set labels under the model, defined as:

$$L_{w,b} = \prod_{i=1}^{N} f_{w,b}(x_i)^{y_i} (1 - f_{w,b}(x_i))^{1-y_i}$$

This expression can seem complex, but it essentially means we use $f_{w,b}(x_i)$ when $y_i = 1$ and $1 - f_{w,b}(x_i)$ when $y_i = 0$.

In practice, to simplify calculations, we maximize the *log-likelihood*, which is the natural logarithm of the likelihood function:

$$\text{Log}L_{w,b} = \sum_{i=1}^{N} [y_i \ln(f_{w,b}(x_i)) + (1 - y_i) \ln(1 - f_{w,b}(x_i))]$$

Since the logarithm function is monotonically increasing, maximizing the log-likelihood is

equivalent to maximizing the likelihood itself.

Unlike linear regression, there's no closed-form solution for this optimization problem. Instead, we rely on *gradient descent* or similar iterative optimization methods to adjust w and b until we find the optimal parameters, allowing logistic regression to classify data accurately.

3.3.4 Time series Algorithms: ARIMA and Holt-Winter

Sometimes, data inherited from various pipelines holds its importance depending on the temporal information it incorporates and conveys. Indeed, although certain absolute values can be used to predict a response, it is not always the value itself that determines a specific outcome, but rather the trend by which that value is reached and any seasonal patterns that justify it. For example, low but increasing numbers describe a growth trend and potentially an increase in the measured value or metric. At the same time, much higher values that are declining might represent more negative outcomes than the previous case. Moreover, predicting and managing seasonality is essential for obtaining a forecast that is free from confusing elements: a negative trend occurring during a seasonal period could simply be symptomatic and structural rather than purely negative for calculating the metric.

Later in the thesis, the use of two specific models will be considered to address these types of issues: ARIMA, a time series algorithm that combines an autoregressive component with a moving average component, and Holt-Winters, an algorithm that divides the time series into different parts to detect seasonality and then reassembles these parts to reveal trends. An ARIMA forecast may sometimes appear rather "plain" as it often results in a straight line without evident fluctuations. This outcome is, however, the optimal fit for the underlying mathematical model. Conversely, Holt-Winters produces forecasts that tend to more closely resemble the original data, particularly when there is a marked seasonal pattern (e.g., daily, weekly, or monthly). This quality can be helpful when interpreting or presenting the model, yet it may also introduce a degree of deception, as it visually aligns more closely with observed data but may not always improve predictive accuracy. In general, ARIMA is an excellent starting point. However, if there is reason to believe in a repeating pattern within the data (often based on knowledge of the actual processes driving it), Holt-Winters might provide better results by capturing these cyclic variations more effectively. ARIMA, or *Autoregressive Integrated Moving Average*, is a forecasting model that provides insights into time series data, especially when there's autocorrelation, meaning that data points are interdependent over time. The model consists of three main components—*autoregression* (AR), *integration* (I), and *moving average* (MA)—which together capture different aspects of the time-dependent behavior in the series.

To begin with, ARIMA models are defined by three parameters, (p, d, q), which determine the nature of the model[40]. The *autoregressive* (AR) component captures the relationship between the current value and its previous values. An autoregressive model of order p represents the value y_t as a weighted sum of p past observations:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t$$

where c is a constant, ϕ_i are autoregressive parameters, and ϵ_t represents the residual error, or "white noise." The *integration* (I) term, indicated by d, reflects the number of differences taken to make a non-stationary time series stationary. In other words, differencing helps to eliminate trends and seasonality by examining the change between consecutive values. For example, a first difference, $y_t - y_{t-1}$, stabilizes the mean by removing trends in the data. The *moving average* (MA) component uses past forecast errors to smooth the time series. An MA model of order q represents y_t as a function of past error terms, where θ_i are the moving average parameters:

$$y_t = c + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

Bringing these components together, an ARIMA model with parameters (p, d, q) combines the autoregressive term, differencing (integration), and moving average term to produce a model for y_t that accounts for the structure within the time series. The general form of ARIMA can thus be written as:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

where ϕ_i and θ_j are model parameters, and the model incorporates differencing based on the order d. ARIMA is particularly useful for time series without clear seasonal patterns, as it models underlying trends and periodicities in data. When seasonal patterns are evident, the model can be extended to *Seasonal ARIMA* (SARIMA), adding additional parameters to handle periodic trends. ARIMA's main strength lies in its ability to produce stable, long-term forecasts, often yielding a straight-line forecast that reflects the best statistical fit for the data, even if it may appear overly simplistic.

While ARIMA is highly effective for non-seasonal data, when the series displays significant seasonality, the *Holt-Winters* model, also known as *Triple Exponential Smoothing*, becomes a valuable alternative. This model decomposes a time series into three main components—*level*, *trend*, and *seasonality*—and is particularly suited for capturing patterns that repeat over fixed periods (e.g., daily, weekly, or monthly).

The Holt-Winters model can be represented in two variations: *additive* and *multiplicative*, depending on the nature of the seasonal patterns[41]. Additive models are suitable for series with seasonality that remains relatively constant over time, while multiplicative models are more appropriate when the amplitude of seasonality increases or decreases with the level of the series. In its basic form, the *additive Holt-Winters* model is defined as follows:

1. *Level equation*: The level l_t at time t is updated by smoothing the current observation y_t, adjusting for both the trend b_{t-1} and seasonality s_{t-m}:

$$l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1})$$

where α is the smoothing parameter for the level, and m represents the seasonal period.

2. *Trend equation*: The trend b_t represents the change in the level over time and is updated as:

$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}$$

where β is the smoothing parameter for the trend.

3. *Seasonality equation*: The seasonal component s_t captures repeating patterns and is updated as:

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m}$$

where γ is the smoothing parameter for seasonality. The forecast at a future time $t + h$ is then given by:

$$\hat{y}_{t+h} = l_t + hb_t + s_{t+h-m(k+1)}$$

where k represents the number of completed seasonal cycles within the forecast horizon. For data where the seasonal pattern is multiplicative, the model adapts by multiplying rather than adding the seasonal component, as seen in the multiplicative form:

$$l_t = \alpha\frac{y_t}{s_{t-m}} + (1 - \alpha)(l_{t-1} + b_{t-1})$$

Holt-Winters is particularly valuable when seasonal patterns are expected to continue and the goal is to capture both trend and cyclic patterns accurately. By breaking down the series into these distinct components, the Holt-Winters model provides a forecast that not only aligns closely with past seasonality but also projects how the trend and seasonal patterns might evolve, making it well-suited for more dynamic, interpretable forecasts.

3.4 Sentiment Analysis and NLP

Natural Language Processing, or *NLP*, is a field within artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language[42]. Through NLP, machines can process large volumes of textual data by identifying patterns, extracting information, and performing tasks such as translation, summarization, and language generation. NLP leverages various linguistic and statistical methods to parse and comprehend language in a way that allows for effective interaction between humans and computers.

Within NLP, *sentiment analysis* is a specialized area dedicated to identifying the emotional tone or sentiment expressed within text[43]. This process involves determining whether a piece of text, such as a review or social media post, conveys positive, negative, or neutral sentiments. By analyzing patterns in word choice, syntax, and context, sentiment analysis allows us to assess the underlying emotions and attitudes within a dataset, giving insight into how users feel about a product, topic, or event.

The significance of sentiment analysis in AI-driven prediction lies in its ability to enrich decision-making models with nuanced human reactions and opinions. By quantifying emotions and attitudes, sentiment analysis transforms subjective text data into structured information that can be used to predict trends, detect shifts in public opinion, and support marketing or customer service strategies. Integrating sentiment analysis into AI models thus allows for predictions that consider both quantitative data and qualitative insights from user feedback.

While both NLP and machine learning involve pattern recognition, they diverge in their foundational approaches. NLP specifically focuses on language structure and semantic understanding, relying heavily on linguistics and symbolic processing to interpret language nuances. In contrast, traditional machine learning (ML) operates primarily on numerical data, identifying patterns through statistical models without an inherent understanding of language or context. Therefore, while ML can learn from structured datasets, NLP adds the layer necessary for machines to grasp the complexities of human language, making it essential for tasks that involve interpreting text.

However, the rise of *large language models* (LLMs) has the potential to shift this landscape, as they may render traditional NLP techniques and sentiment analysis frameworks less critical. LLMs, with their vast neural network architectures and extensive training on diverse language data, can generate human-like text and respond to queries with a sophisticated understanding of context and sentiment. By using LLMs, it becomes possible to bypass traditional NLP pipelines, as these models can perform a wide range of language tasks, from sentiment classification to language generation, with minimal additional training. This adaptability allows LLMs to replace multiple specialized NLP components, simplifying workflows and enabling more flexible and comprehensive text analysis in AI systems.

3.4.1 Introduction to Natural Language Processing for User Reviews and Feedback Analysis

Natural Language Processing, or *NLP*, plays a crucial role in extracting meaningful insights from user-generated content, particularly for understanding engagement levels through reviews, forum interactions, and other feedback. By analyzing textual data, NLP techniques can quantify user satisfaction, identify recurring themes, and reveal engagement patterns, providing structured insights that can be stored in a database for continuous monitoring and improvement of the service.

In the context of user reviews, NLP techniques like *sentiment analysis* and *topic model-*

ing allow us to categorize feedback by emotional tone (e.g., positive, negative, neutral) and to identify the specific aspects of the service users discuss most frequently. For example, a user's review indicating satisfaction with the "ease of use" or dissatisfaction with "response time" can be flagged and categorized accordingly. Storing these insights enables tracking engagement metrics over time, such as trends in satisfaction with specific features or common pain points. Each processed review is represented as a set of structured values in the database, including *sentiment scores*, *topics mentioned*, and *overall engagement indicators*, which collectively serve as a valuable metric for evaluating the user experience.

Forum interactions provide another dimension of user engagement data if this kind of service has been adopted by a SaaS. Here, NLP techniques like *entity recognition* and *frequency analysis* can detect high-engagement topics, influential users, and common concerns within community discussions. For instance, entity recognition might identify specific product features or service aspects discussed repeatedly, while frequency analysis can highlight popular or controversial topics. By storing engagement metrics—such as the number of posts, keywords associated with engagement, and the overall sentiment of user interactions—in a database, it becomes possible to gauge not only individual user satisfaction but also the overall sentiment of the community.

To implement this process, an automated NLP pipeline can be developed to continuously analyze new reviews and forum posts. Key steps in this pipeline might include *text preprocessing* (such as tokenization and normalization), *sentiment scoring*, and *topic extraction*. Once processed, each piece of feedback is assigned a series of engagement metrics (e.g., sentiment score, sentiment trends over time, key topics, and response rates) and stored in a structured database. This database serves as a foundation for monitoring user engagement trends and provides actionable insights, allowing for timely responses to issues and the enhancement of user satisfaction with the service.

A sentiment analysis score provides a valuable metric for monitoring user engagement by quantifying the emotional tone within user feedback. This score can serve as an early indicator of a user's satisfaction or dissatisfaction, enabling the prediction of churn likelihood. By identifying users with consistently low sentiment scores, it becomes possible to proactively adjust pricing or offer incentives aimed at reducing churn risk through targeted price discrimination, thus potentially lowering their probability of leaving the service.

However, relying solely on sentiment analysis can introduce limitations, as users might intentionally manipulate feedback to achieve discounts or lower prices. For instance, a user could submit negative reviews or express dissatisfaction in hopes of triggering favorable pricing adjustments. To mitigate this, sentiment analysis should be complemented with additional metrics, such as *usage patterns*, *response to prior discounts* (e.g., seasonal promotions like Black Friday), and *interaction frequency*. By integrating these metrics alongside sentiment scores, a more holistic and robust model of user engagement can be constructed, reducing the risk of manipulation and enhancing the accuracy of pricing strategies tailored to retain users effectively.

3.4.2 Building Sentiment Analysis Models to Influence Pricing Decisions

Building a sentiment analysis model to support pricing decisions involves training an algorithm to categorize user feedback into structured sentiment scores, which can then inform targeted pricing strategies. In this hands-on approach, we use supervised machine learning, leveraging a dataset of labeled user reviews or interactions that reflect varying sentiment levels (e.g., positive, negative, neutral). Through training, the model learns to associate specific textual features with sentiment labels, enabling it to classify new feedback accordingly.

1. *Text Preprocessing and Feature Extraction*:
   To prepare textual data for sentiment analysis, we first preprocess the text, a step that involves tokenizing the text into individual words or phrases, removing stop words (common words with limited meaning, such as "the" or "and"), and applying stemming or lemmatization to reduce words to their base forms. We may also employ n-grams to capture sequences of words that often convey sentiment, such as "highly recommended" or "very disappointed." The processed text is then transformed into numerical features. Common methods for this include *Bag of Words (BoW)* and *Term Frequency-Inverse Document Frequency (TF-IDF)*:

   $$\text{TF-IDF} = \text{TF}_{i,j} \cdot \log\left(\frac{N}{\text{DF}_i}\right)$$

   where $\text{TF}_{i,j}$ is the term frequency of word i in document j, N is the total number of documents, and DF_i is the document frequency of word i. TF-IDF weights terms that are frequent in a particular document but rare across all documents, helping capture significant words associated with sentiment.

2. *Model Selection and Training*:
   For sentiment analysis, a variety of algorithms can be employed, each with distinct strengths depending on the complexity of the text and available resources. A commonly used algorithm is *Naive Bayes*, which applies Bayes' Theorem to calculate the probability of a sentiment class given the observed words:

   $$P(\text{sentiment}|\text{text}) = \frac{P(\text{text}|\text{sentiment}) \cdot P(\text{sentiment})}{P(\text{text})}$$

   This model is computationally efficient and interpretable, making it suitable for baseline sentiment analysis. For more nuanced models, *Support Vector Machines (SVM)* or *Neural Networks* like Recurrent Neural Networks (RNNs) can be used to capture context and complex dependencies in the text. The model is trained by minimizing a loss function that quantifies the error between predicted and actual sentiment labels. For example, in

the case of binary sentiment classification, we can use *binary cross-entropy loss*:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right)$$

where y_i represents the true sentiment label, and $\hat{y}_i$ is the predicted probability of that label.

3. *Calculating Sentiment Scores and Integrating with Pricing Models*:
   Once the model is trained, it can classify new reviews and interactions, assigning a sentiment score to each. This score, typically on a scale from -1 (negative) to +1 (positive), quantifies the user's satisfaction and can be stored as a feature in the user's engagement profile. By tracking these sentiment scores over time, we obtain a dynamic view of user satisfaction that can trigger pricing adjustments. To integrate sentiment analysis with pricing, a simple price adjustment model might employ a rule-based system, for example:

   $$\text{Price}_{\text{new}} = \text{Price}_{\text{base}} \times (1 - \alpha \cdot \text{sentiment score})$$

   where α is a sensitivity factor that determines how much influence sentiment has on the pricing strategy. In this way, users with lower sentiment scores may be offered discounts to reduce churn risk, while highly satisfied users maintain the base price.

4. *Avoiding Manipulation and Enhancing Robustness*:
   To reduce the risk of manipulation, we can integrate additional metrics like *usage frequency*, *historical discount response*, and *seasonal purchasing patterns* alongside sentiment scores. For instance, a weighted combination approach could be applied:

   $$\text{Engagement Score} = \beta_1 \cdot \text{sentiment score} + \beta_2 \cdot \text{usage frequency} + \beta_3 \cdot \text{discount response}$$

   where β_1, β_2, and β_3 are weights assigned to each metric. This multidimensional engagement score helps ensure that pricing decisions are based on a balanced view of user activity, minimizing the effect of sentiment manipulation while providing a more accurate assessment of churn risk.

At the end of this process, it is possible to obtain a comprehensive *Engagement Score* for each user, built by combining the sentiment analysis score with key metrics such as usage frequency, response to previous discounts, and seasonal patterns. This score provides an overall measure of user engagement and churn probability.

With this *Engagement Score*, it becomes possible to implement a targeted *price discrimination* strategy, offering personalized incentives or pricing adjustments to reduce the likelihood of churn for at-risk users. This system enables dynamic monitoring of user satisfaction and proactive interventions to enhance retention.

4 Model Development and Simulation

In this chapter, after spending time outlining the theoretical framework, the focus will shift to the practical construction of the models necessary for investigating the research question. Indeed, considering the structure of a SaaS product, the challenges, and the key metrics required by companies in the short and long term, and having deeply examined industrial, microeconomic theory, monopolistic competition markets, and price differentiation through the application of price discrimination, this chapter will embody all this theoretical foundation in a practical project through the development of Machine Learning algorithms.

Mathematically and statistically, the models will follow the theory approached in the previous chapter. Specifically, the results will be made visible through computational calculation, driven by programming. The chosen language is *Python*, a programming language that is easily understandable (given its syntax, which is similar to English) but whose commands and libraries are particularly effective in the field of Machine Learning, making it the most used language in this field and in real-world markets.

Therefore, this thesis aims to present a practical model that can be effectively used in the study of the SaaS domain.

4.1 Introduction

Recalling the research question, how can Machine Learning be employed to find appropriate price discrimination techniques for a SaaS business model? With a clear understanding of these three macro concepts and the intersections they develop among themselves, the practical task of identifying the models that most effectively support business decision-making and answer the initial question becomes paramount.

The strategy to be implemented involves comparing the profit outputs of a static pricing model with a single price to, *ceteris paribus*, a differentiated pricing model. Specifically, given the lack of access to real company data, the specific strategy for this thesis will be the construction of a multi-step model. This model will first address the issue of creating a *synthetic database*[40] with single pricing, then use ML algorithms for analysis, create a second database with differentiated pricing, and finally conduct a comparative analysis between SaaS metrics. Only then will dynamic user tracking be applied, where a user, based on their behavior and choices, will be assigned to a particular cluster with corresponding pricing aimed at maximizing company profits.

All of this analysis, from the creation of the first database to the final dynamic assignment simulation, must rely on quantitative interpretations and needs. To maintain continuity with the previous analyses, the narrative device of the Micro SaaS provider offering freelancing

[40] A database with data not taken from real measurements but artificially constructed using statistical or random algorithms.

services, described in the second chapter, will be used again. Given the cost structures, pricing considerations, and the impact of economic units, the transition to a differentiated pricing model driven by ML models, as described in this introduction, will feel more natural.

4.1.1 Preview of Chapter Content and How It Connects to Previous Chapters

Initially, in the second section of this chapter, the methods for creating the first reference synthetic database, which will be useful not only for developing the actual model but also for creating the subsequent differentiated database, will be outlined. After presenting the description of the fields to be examined in the form of dataset columns, the methods for populating the fields for all simulated users will be clarified. Once the database is obtained, it will be possible to analyze the outcomes and the metrics necessary for business sustainability, providing a benchmark for the following chapters in the discussion of the impact of using algorithms to identify the best price discrimination strategies.

For this reason, in the third section of the chapter, Machine Learning models will be employed to determine the pricing tiers to be applied based on the identified clusters and profit forecasts.

4.2 Data, Preprocessing and Analysis

For the purpose of this thesis, as no real data from actual databases is available, the technical expedient of a *Synthetic Database* will be employed. A Synthetic Database is a database made up of non-real data, synthetically created from specific development instructions. Following the example of the SaaS product where assumptions about cost structure and revenues were determined, and as that will be the basis for studying the differences in profit between the application of price discrimination and static pricing, a synthetic database will be constructed to follow a certain logic and probability based on those same assumptions.

Indeed, the main challenge of relying on non-real data is the potential inability to illustrate the full benefits of the proposed research. Additionally, even if randomness is not fully embraced, pre-defining possible consumer groups during the development phase may mean already knowing the business needs without employing Machine Learning algorithms. However, the answer to such concerns remains positive and does not conflict with the objectives: having data that is not difficult to interpret in terms of patterns does not invalidate or contradict the use of ML algorithms applied later. On the contrary, it may provide further proof of how these models can correctly interpret the information that emerges from a data collection process.

In proceeding with the creation of the database, probabilistic controls will be explicitly applied to the otherwise random content generation. Additionally, the columns of the data relevant to the analysis will be defined to make the subsequent data analyses feasible.

4.2.1 Description of Synthetic Dataset Used for Simulation and Its Characteristics

This synthetic database will be developed in a single table specific to the calculations and models. Although it is clear that a SaaS company would also have operational tables such as those related to authentication (containing user email, password, etc.) or user profiles (name, surname, nationality, etc.), in the specific case addressed in this thesis, only some of these extended details are useful for defining the models and economic analysis concerning price discrimination.

For this reason, the columns of these tables will be listed below, with their respective values explained. Before presenting the columns, it is important to note that at the beginning of the analysis, before price discrimination is implemented, the company starts with a subscription plan (and therefore, recurrent revenue) tied to a single fixed monthly fee of $59.99. Consequently, at least initially, it is not possible to identify price sensitivity, decisions linked to specific choices, discounts, or bilateral actions between user and company to maximize purchase utility. Later, this data will be expanded to support price discrimination implementation.

The first synthetic database will be constructed with the following columns:

1. *UserID*: A unique identifier for each SaaS product user. This ID will link the user to other internal database tables (e.g., profile or credentials). It will be generated in a progressive and unique manner.

2. *Hourly Rate*: The hourly rate of the freelance user. As the target range is between $50/h and $150/h, this variable is essential for identifying the *Willingness to Pay* (WTP) and determining significant clusters.

3. *Nation*: The country of residence of the user, collected during registration. This information is crucial for analyzing socioeconomic differences and price sensitivity related to specific geographic areas.

4. *Age*: The user's age. This parameter may influence the propensity for technology adoption, user behavior, and willingness to pay for the service.

5. *SubscriptionDate*: The subscription date in *unix time*, providing a numerical reference to calculate the months since registration and correlate usage over time. It is essential for tracking user engagement trends.

6. *MonthlyDailyLogin*: Number of unique login days per month. This field will be populated with an array containing data for the last six months. It is crucial for tracking user engagement, revealing trends that may indicate increased churn risk or deeper engagement.

7. *MonthlyHours*: Number of hours worked monthly on the platform, tracked in an array for the last six months. This data helps understand the perceived value of the service by the user.

8. *MonthlyProjectsCompleted*: Number of projects completed each month, also tracked in a six-element array. More completed projects suggest active use and a higher return on investment for freelancers, a factor closely tied to willingness to pay.

9. *TicketsOpened*: Number of support tickets opened by the user. This data can either be interpreted as a sign of issues, which might increase churn risk, or as a sign of engagement if the user interacts actively with technical support.

10. *AvgTicketScore*: The average score given by the user for the resolution of opened tickets. Dissatisfied users are more likely to churn or be less willing to pay a premium for the service.

11. *SurveyScore*: A score derived from periodic satisfaction surveys administered to users. This data provides direct insight into user engagement and satisfaction.

12. *NLPScore*: A score derived from Natural Language Processing (NLP) analysis of any written interactions by the user (e.g., forums, feedback). Monitoring the tone of messages can reveal satisfaction or dissatisfaction levels that may influence retention.

13. *Churn*: A binary variable indicating whether a user has abandoned the service (1 = churn, 0 = active). This variable is essential for monitoring churn factors and building predictive models.

14. *Tenure*: The number of months the user has spent on the platform. Tenure can be correlated with loyalty and user value, and provide indications of churn likelihood.

In this way, by considering these initial sample columns, it will be possible to conduct statistical and probabilistic analyses on this database, as well as build initial clustering models and examine demand and pricing for each group. This will provide a foundation for exploring price discrimination details and anticipated profit variations.

UserID	HRate	Nation	Age	SDate	MDailyLog	MHours
456	110	Italy	30	18/09/2024	[12,14,15,20,15]	[20,32,36,45,27]

MProjComp	Tickets	AvgTScore	SScore	NLPScore	Churn	Tenure
[1,2,2,3,2]	[null,2,null,1,null]	[null,4,null,5,null]	5	0.8	0	14

Table 1: Example row from synthetic dataset with UserID 456

This is an example of a table with the relevant columns for subsequent analyses. Once these are considered, they can be populated by developing functions based on specific logical considerations to adhere to the non-random assumptions outlined at the beginning.

4.2.2 Populate the Synthetic Dataset

While the *UserID* is generated in a simple progressive and unique manner (starting from one and advancing to the number of desired users), the same linear decision-making process cannot be applied to the other fields in the dataset. The rationale on which the case study analysis of this thesis is based requires that the dataset accurately represent users and their expected behaviors: hence, the data and quantities must be distributed according to the elements composing the hypothesized demand curve, emphasizing that the willingness to pay is derived from a combination of sensitivity to price (also expressed by economic availability), observable attributes, and intangible factors (with a certain margin of error).

At the same time, the objective is to create an artificial base on which to conduct analyses and modifications. Therefore, the rigor and complexity required in constructing the synthetic database will be balanced with a degree of simplification, approximation, and randomness to best represent reality. While many correlations between columns and their values could be explored, this section will focus on the necessary complexities to highlight the potential of the implemented ML model.

For example, the distribution of users across the pricing bands cannot be left to pure chance; instead, it is reasonable to follow the distribution derived from the demand curve established in Chapter Two. Specifically, the demand curve was modeled as a sigmoid, which means using a logistic distribution to represent user allocation across bands, with a higher probability of clustering near the mean (which was determined to be $75) and lower likelihoods of extreme high or low values. This reflects the fact that users with very low hourly rates likely cannot afford a product with a significantly higher value relative to the ROI they receive, while users with very high rates may have different needs compared to those targeted by the SaaS product. Furthermore, lower price bands are more common than higher ones, representing a labor market where earning more is less frequent. The logistic distribution, compared to the normal distribution, tends to spread values further from the median, making it more suitable for identifying potential clusters (given the lack of real data).

$$f(x) = \frac{e^{\frac{-x-\mu}{s}}}{s(1+e^{\frac{-x-\mu}{s}})^2} \tag{21}$$

This logistic distribution equation helps to determine how many users fall within each pricing band, translating this detail into percentages of the user population under consideration, thanks to the *cumulative distribution function* (CDF): 0.05% in Band 1, then respectively 1.05%, 6.48%, 30.17%, 54.67%, and 7.58% for Bands 2 through 6. It is reasonable to assume that the largest portion of users will be in Band 5, as the hypothesized willingness to pay is $75, which was calculated based on a 10%-30% ROI on the saved hours. In the specific case of this synthetic database, 13,000 users will be created to provide enough variation (active or churned users), and these percentages will apply to this number, resulting in 7, 136, 843, 3,922, 7,106,

and 986 users in Bands 1 through 6, respectively. The sum, of course, equals 13,000.

Regarding the users' countries of residence, given their importance due to the cost of living, average salary, and tax implications (important for determining WTP), a small set of countries will be used, with each assigned a weighting for each pricing band. This means that while some countries typically have higher taxes or higher average wages, a percentage of outliers is allowed since, as in reality, residency is not the only factor determining these economic metrics. The same *modus operandi* is applied to user age, with the higher price bands more likely to fall within older age groups to reflect experience. To achieve this, a weighted random distribution based on band membership is created. Arbitrary ranges for each band are established, within which a random value is more likely to be selected. For instance, in the second band ($120/h-$150/h), the 35-60 age range is given a higher weighting to reflect the greater likelihood of a 40-year-old having accumulated enough experience to raise their hourly rate, compared to an 18-year-old. Conversely, as the bands move toward lower values, the distinction becomes less important since both a 20-year-old and a 50-year-old could have lower hourly rates, though the likelihood of the user being younger is still slightly higher.

For the subscription date, it is assumed that the SaaS product launched on January 1, 2024, with the final subscription date set for December 31, 2025. Typically, over a long period, user growth follows a sigmoid curve, but since this period is limited to two years, it is reasonable to consider only growth. However, considering the usual growth pattern for a SaaS product, its objectives, and saturation point, it makes sense to use a logistic function to determine how many users subscribed on a given date. Thus, assuming fewer subscriptions early on due to initial marketing efforts, followed by increasing subscriptions due to network effects, until the market saturates, it is expected that 13,000 users will have subscribed by December 31, 2025, following this "S-shaped" growth curve.

The number of logins, completed projects, hours worked, tickets opened, and average ticket scores calculated monthly all follow the same *ratio*: they tend to grow month over month before stabilizing or decreasing if user engagement declines and churn risk increases. Outliers and seasonality factors must also be considered. Churn, however, can also occur earlier, as low initial usage that never increases over time can lead to churn (low perceived value). This increases the likelihood of canceling the subscription early, rather than in the following months. This introduces three key factors: band, churn, and period. While time (a marker of seasonal presence) is derived from the subscription date, and band affects ROI, churn still needs to be incorporated and distributed among the database rows.

It is important to note that the churn rate is calculated by dividing the lost customers by the total users at the beginning of the period, then multiplying by 100 to obtain the percentage. A realistic churn rate for SaaS can range from 5% to 25%, so an annual churn rate of 10%-15% with a monthly rate between 1% and 3% is reasonable for this simulation. This rate can be distributed based on seasonality or exponential growth in subscriptions; the more users there are, driven by network effects, the more likely it is to lose disinterested users. However, a

different approach will be used to make the dataset more varied by simulating more realistic behavior. Each user will have a unique churn probability that changes based on engagement, seasonality, and band. This creates a dynamic, non-predetermined churn, with a cumulative distribution: the probability can increase or decrease depending on seasonal trends or project success. Since these are synthetic data, we cannot directly refer to projects, hours worked, or other variables that still need to be defined, as churn is also influenced by these factors. Instead, we can rely on other aspects:

1. *Band*: Since each user's perceived value and WTP are tied to their ROI, higher bands will tend to have a higher WTP than lower ones. As a result, the higher the perceived value, the lower the propensity to churn. Band 6, with an hourly rate below $50/h, has a higher propensity to churn compared to Bands 4 and 5, which, in turn, have a higher propensity to churn than the first three bands.

2. *Seasonality*: Vacation periods, arbitrarily set for the months of December, August, and April, increase the probability of churn, although to a lesser degree than band. This is based on the fact that less use of the product leads to a lower willingness to invest in it.

3. *Time Passing*: The first month, during which the services are still unfamiliar and no time has been invested in integrating or saving data, has the highest churn rate, which decreases over time. In fact, as time passes, the only variable that increases churn is decreased engagement due to new products entering the market, lack of innovation, problems, bugs, negative customer service experiences, or external factors such as personal issues, fewer clients for freelancers, or economic system problems.

The cumulative churn distribution can be calculated as follows:

$$P_{churn}(t) = P_{initial} + \sum_{i=-1}^{t} (\Delta P_{seasonal} + \Delta P_{engagement} + \Delta P_{band}) \tag{22}$$

Remembering that individual probability translates into collective probability, we can calculate between 1% and 3% for each user based on their band and other monthly factors that may increase or decrease the churn percentage. For each month, the relative probability is calculated, generating a random value. If it is lower than the probabilistic value, churn occurs, returning the reference date. If it is higher, the user moves on to the next month. Depending on engagement, worked hours, completed projects, and tickets, engagement can rise or fall, which in turn influences (through engagement) an increase or decrease in churn probability. For the purposes of populating the dataset, the history of these values will be fixed up to six months prior, rendering earlier time trends irrelevant for future churn forecasting.

Finally, it is assumed that the survey is sent to users who have been subscribed for more than three months. The score is calculated based on engagement, as is the NLP score (evaluated between 0 and 1). Once all fields are populated and this schema is replicated for each user, a

dataset of 4,999 users is created, providing a sample for the thesis. This dataset can be used to apply ML algorithms, saved externally as an SQLite database (to avoid having random data generated each time).

4.2.3 Preprocessing and Exploratory Data Analysis

In the case of a database already established by the pipeline, the *Exploratory Data Analysis* (EDA) phase would require time spent analyzing the formats and information within the various columns. However, since the dataset was created from scratch using synthetic data with the algorithms just defined, this step has already been extensively covered in the preceding paragraphs. What is now useful is to analyze the results obtained in terms of quantities, distributions, and relevant metrics. Before analyzing the data, preprocessing operations are conducted to address missing values (e.g., for *avg_ticket_score*, survey, and churn), and the array entries are restructured in a proper format. Missing values (in cases where tenure is less than six months) will be counted as NaN. With these necessary steps completed, the dataset can be analyzed using EDA techniques.

Preliminary results based on numerical statistics can then be stated, following calculations on quantities and statistical elements:

1. *Hourly Rate*: ranges from 30 $/h to 250 $/h, with an average of approximately 74 and a median of 75.
2. *Age*: ranges from 18 to 70 years, with an average around 36.
3. *Survey Score*: average of 4.32 on a scale from 1 to 5.
4. *NLP Score*: average NLP score of 0.55, with values ranging from close to 0 to almost 1.
5. *Churn*: approximately 10.5% of users have churned.
6. *Tenure*: average duration of about 12 months, with a maximum of 24 months.

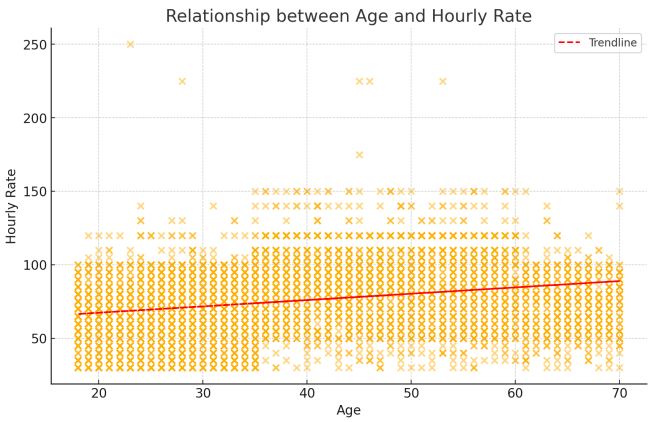


Figure 12: Relationship between Age and Hourly Rate

These details accurately reflect the distribution applied during the construction of the database, providing the final results for what was randomly assigned within a range of values. In the graph, it is evident that there is a direct proportionality between the increase in age and the increase in hourly rate, with an average age of 36 years and 75 $/h.

Moreover, by analyzing other demographic metrics, we can also examine the distribution of nationalities in relation to the hourly rate, illustrating another correlation factor between users' geolocation and willingness to pay in relation to income, taxation, and other variables that influence sensitivity to prices applied to SaaS services.

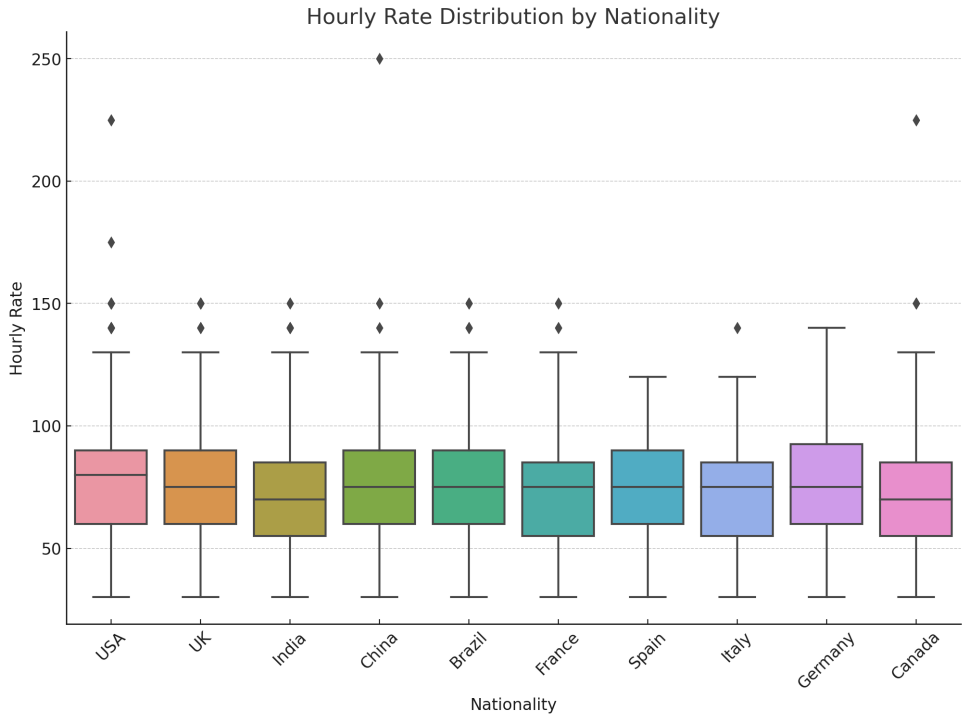


Figure 13: Relationship between Nationality and Hourly Rate

However, more useful than demographic details—though these will later assist in building more specific clusters—is the analysis of SaaS metrics. In fact, analyzing the current quantities concerning the values of economic units of the business model is helpful in comparing single pricing with differentiated pricing, as well as in achieving the final goal of price maximization.

4.2.4 Insights on SaaS Metrics

In particular, metrics related to recurrent revenues (if monthly, MRR), Average Revenue Per User (ARPU), gross margin, LTV, and churn are fundamental for comparing the profitability differences of various pricing models. Other important metrics, such as CAC and the magic number, are also essential but depend on a well-defined cost structure for user acquisition costs. This study, however, focuses on revenues to standardize the profit differences. Notably, the fixed price for the service subscription is set at $59.99 for all users:

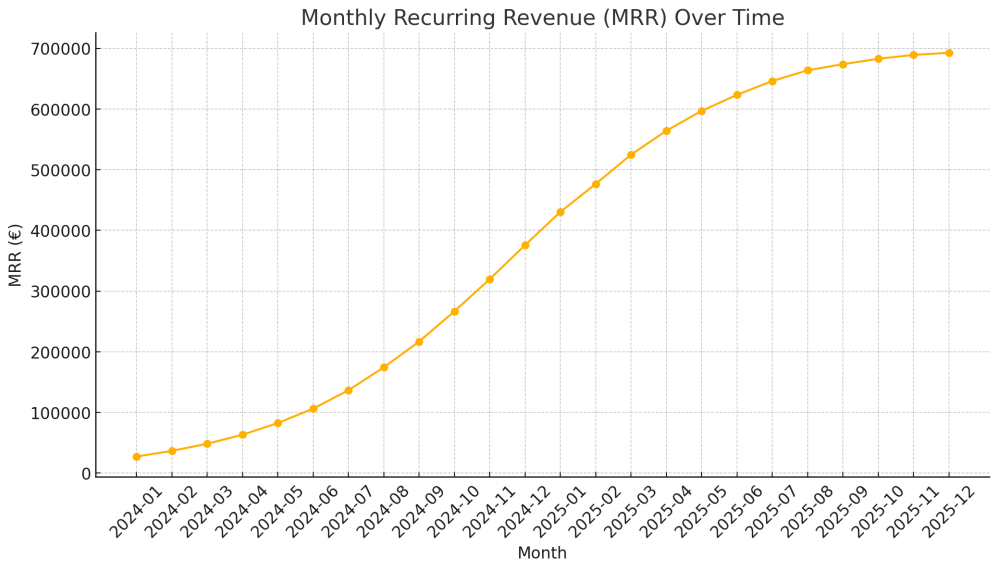


Figure 14: Relationship between Nationality and Hourly Rate

1. *MRR*: Examining the MRR, the final month's estimate is $697,743.69. The growth closely follows the sigmoid curve that represents the monthly acquisition of the new customer base, adjusted for the churn rate and multiplied by the monthly service price.

2. *Churn Rate*: The churn rate aligns with the expectations of an annual churn rate typical for a SaaS at 10.5%, equivalent to a monthly churn rate of 0.88%.

3. *ARPU*: The Average Revenue Per User is easily calculated, as the current single pricing is $59.99 monthly. This factor will eventually change with the possibility of price differentiation, one of the reasons a higher profit is expected with tiered pricing.

4. *Gross Margin*: Assuming an average cost per user at the threshold of 13,000 users (minus churned users) at $7.5, the gross margin is approximately 87.5%, calculated by the ratio of COGS to MRR.

5. *LTV*: The Lifetime Value, for now, can only be based on the observed 24 months. With an average tenure of 11.69 months, multiplying this average tenure by the ARPU gives a lifetime value of approximately $701.47.

These analyzed metrics, as previously mentioned, will be revisited later to compare the differences with the metrics obtained after price differentiation, providing insight into the impact of tiered pricing on company revenues.

4.3 Customer Segmentation

It is now essential to use this data to develop a pricing model that applies price discrimination based on similarities between users. Since it is not possible to perfectly know each individual's willingness to pay, and because managing individual differences in pricing would likely

decrease engagement due to perceived unfairness, the most suitable model for price discrimination is tier pricing: a strategy that differentiates users' needs and the value they assign to the product by adding features from basic to premium, thereby justifying price differences.

Two key elements are necessary: microeconomic theories of organizational strategy to find, within monopolistic competition, the optimizations of quantity and price based on different demand curves, and clustering. Using machine learning algorithms to cluster various types of users becomes feasible thanks to the data collected in the database. By analyzing the right features, it is possible to group similar users, and their demand curve becomes more specific to their willingness to pay. Maximizing profits becomes feasible when, without increasing the churn risk, user surplus—especially among those willing to pay more than the fixed price—can be reduced, thus increasing the company's surplus.

Therefore, after gathering the data, the next step is understanding how to use it to identify relevant features that will help find meaningful clusters.

4.3.1 Application of Clustering Techniques to Segment Customers Based on Their Characteristics

As previously described, unsupervised clustering is generally accomplished through two types of algorithms: k-means and DBSCAN. DBSCAN is more suitable when the number of clusters is unknown a priori; in cases where features overlap, are complex, or are not well-defined, this algorithm can extract more information that would not be easily deduced from the start. K-means, on the other hand, is better aligned with the goal of tier pricing.

While maximizing sales with many clusters may be theoretically possible, psychological factors discussed earlier are strong opponents to the perfect rationality expected when trying to find optimal prices for maximization. Having many clusters, and consequently many prices and additional products to justify price increases, can lead to issues. These include increased production costs for multiple meaningful add-ons and the complexity of navigating a funnel where customers face too many options, ultimately reducing conversion rates. Consider the role of emotion in purchasing decisions: guiding customers toward rational behavior instead of instinctive action can significantly lower the conversion rate and the effectiveness of a call to action.

Furthermore, psychological pricing strategies (such as odd numbers ending in seven or .99) or graphical tactics (highlighting a choice as the most advantageous, labeling it as the most chosen, or positioning it between two less attractive options) may not be as effective when there are too many options. While having many choices may help avoid surplus loss and maximize profits, in practice, this approach may lead to reduced profits due to missed conversions, increased costs, or higher churn rates.

K-means is also preferable for the purposes of this thesis to demonstrate a less complex, descriptive analysis. In real-world revenue management, real data, contingent situations, and specific needs could dictate the choice of one algorithm over another. However, for this thesis,

using k-means is sufficiently significant to emphasize the importance of a data-driven approach in pricing decisions, especially in dynamic pricing.

The process of finding clusters requires characteristics related to price sensitivity, usage factors, engagement, and demographic information such as nationality and age. After training the model, it can be used to predict which cluster a new user belongs to, even if no similar user with the same attributes existed before. For this, specific columns of the database will be used, including demographic features like age and nationality, combined with churn, NLP score, survey score, tickets opened, average ticket score, and willingness to pay. Although there is no specific column for willingness to pay (since asking each user to estimate the maximum price they would pay would be biased and unreliable), it can be derived from other columns, with some used as proxies in the absence of a direct measure.

4.3.2 Estimation of the Willingness to Pay

Willingness to pay (WTP) is one of the fundamental features for identifying different clusters within the current demand curve. As discussed in Chapter 2, the demand curve is, in a sense, the sum of each individual's utility as defined by the Berry, Levinsohn, and Pakes (BLP) formula: each person, based on tangible and intangible characteristics, price sensitivity, and idiosyncratic error, can be placed within a willingness to pay range. Thus, when analyzing the willingness to pay of each user in the dataset (identified by a unique user ID), it can be derived from observable characteristics, intangible attributes, and the calculation of relevant coefficients, including those for price and error, described by the residual of the calculation model. After identifying the necessary data for the estimate, linear regression can be used to find the parameters needed to estimate this factor.

Recalling equation (10), WTP according to the BLP model is defined as:

$$WTP = \beta X + \alpha P + \xi + \epsilon$$

Where βX is the coefficient of the observable characteristics multiplied by the product characteristics, αP is the coefficient of price sensitivity multiplied by the price (59.99$), while ξ represents intangible characteristics and ϵ represents the error.

Starting with observable characteristics, even though the value of certain features, such as service speed, third-party integrations, and other factors important to the user, cannot be quantified directly, it is possible to derive them from parameters like monthly logins and hours worked. The more valuable and important the product, the more frequently it will be used. This differs from more intangible elements such as personal preferences, emotional engagement, fashion effects, UI appeal, or anything that does not depend on rational, measurable use. To calculate the parameter ξ, it can be inferred from dataset variables like NLP score and survey results. Moreover, with a single price of $59.99 and the hourly rate of each user, it's also possible to estimate the sensitivity to price along with error. However, recalling that information

about daily logins and hours is stored as an array with inherent time value, we need to consider using methods such as ARIMA or Holt-Winters. For the purposes of the thesis, it is sufficient to consider these as values of variance, mean, and trend, calculating each of these for the columns of daily logins and monthly hours. Given this, and reporting the coefficients as progressive beta values, we have the following formula for the willingness to pay:

$$\begin{aligned} WTP = \beta_0 + \beta_1 \times avg\_monthly\_hours + \beta_2 \times login\_trend \\ + \beta_3 \times variance\_monthly\_hours + \beta_4 \times nlp\_score \\ + \beta_5 \times survey\_score + \beta_6 \times hourly\_rate + \varepsilon \end{aligned} \tag{23}$$

It was decided to represent the coefficients as β_n to highlight the future possibility of estimating the willingness to pay of each cluster using regression formulas, such as linear regression specifically for WTP, or logistic regression to estimate the probability of churn for each cluster, as follows:

$$P(\text{churn} = 1) = \frac{1}{1 + e^{-WTP}} \tag{24}$$

Indeed, it was decided to use these values of mean, variance, and trend within the calculation of the K-means, while the estimation of the likelihood of churn and the WTP will be calculated later with linear and logistic regression for each cluster to establish an appropriate tier pricing division. For the initial calculation of clusters, the necessary information will be passed to the algorithm to identify patterns and differences, allowing these unsupervised models to generate the segmentation results on which subsequent considerations will be based.

4.3.3 Clusterization Process and Results

To summarize, the elements required for the K-means characteristics include Hourly rate, average monthly daily logins and hours, login trend, monthly hours variance, NLP score, survey score, churn, age, and tenure. Characteristics related to nationality, as well as the addition of other parameters, were omitted to avoid the risk of overfitting. After standardizing the scale of these various inputs with different numerical bases by processing them with a normalization algorithm, it becomes necessary to determine k, or the predefined groups we aim to recognize within the dataset.

To achieve this, referring to the clustering theory in Chapter 3, we rely on the elbow method, which determines the optimal number of clusters in an analysis like K-means by calculating the sum of squared distances between the data points and the centroid of each cluster (known as inertia). By calculating this over a range of possible clusters, a graph of the distances relative to the number of clusters is plotted. Where an ”elbow” can be identified, indicating a slower decrease after a certain number of clusters, this point represents the optimal choice to be used in K-means.

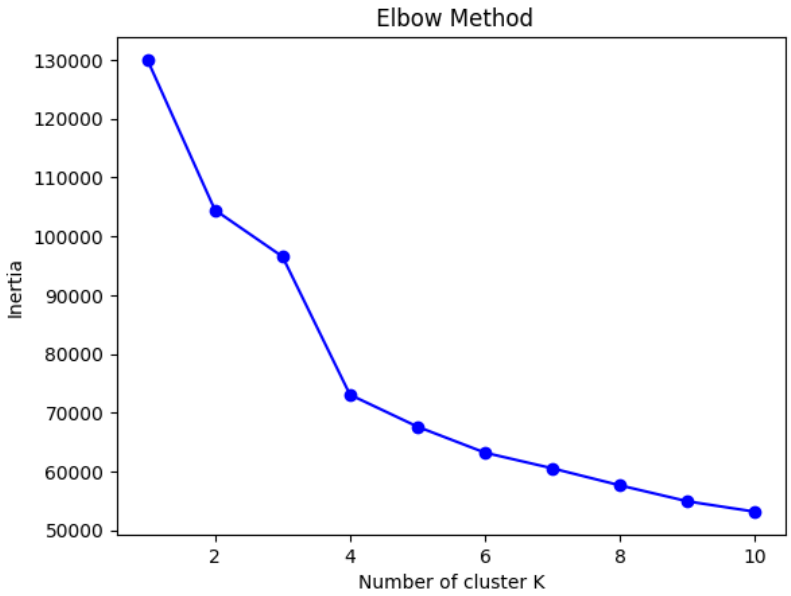


Figure 15: Elbow Method for K-Means Clustering

The elbow method suggests that the optimal clustering choice is four. This implies that—always in probabilistic terms—this division into four clusters is sufficiently significant to derive four groups with different needs, where users belonging to the same group share a good balance between internal variance and parsimony, without dividing into too many clusters. Thus, the K-means algorithm is applied with the aforementioned parameters and four clusters, yielding the following result:

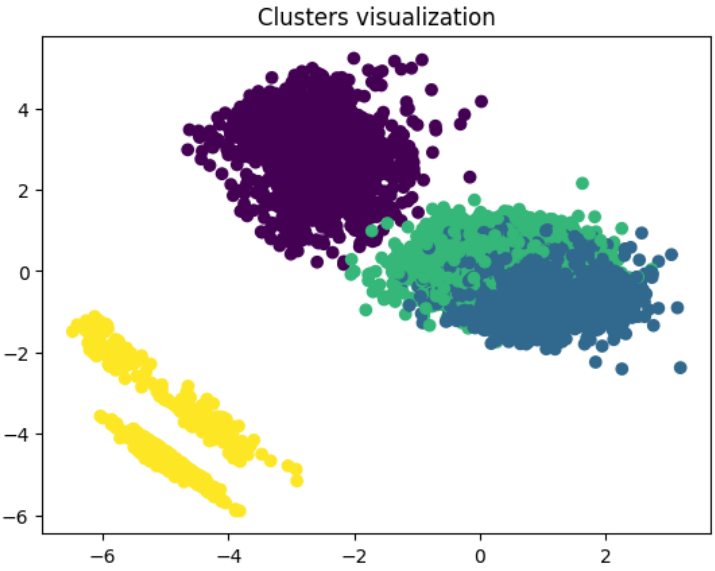


Figure 16: Visualization of Clusters with K-Means algorithm

Following the application of the K-means algorithm with four clusters, each cluster presents distinct characteristics, as detailed below:

- **Cluster 0**: This cluster shows an average *hourly rate* of approximately 54.44 and an

average *age* of 33.5 years. The *survey score* and *NLP score* are relatively low compared to other clusters, at around 2.68 and 0.47, respectively. Additionally, all users in this cluster have churned (churn rate of 1.0), with a short *tenure* of about 3.6 months. The *average monthly hours* and *average monthly daily logins* are high, indicating initial engagement, but the declining *login trend* (-0.0125) and moderate *variance in monthly hours* suggest a potential loss of interest over time.

- **Cluster 1**: Users in this cluster have the highest average *hourly rate* at 88.97 and an average *age* of 48.7 years, making them the oldest group. They exhibit very high *survey scores* (4.50) and *NLP scores* (0.55), indicating high satisfaction and engagement. Notably, there is no churn observed in this cluster, and users show a long *tenure* of about 13.2 months. Both *average monthly hours* and *average monthly daily logins* are high, with a positive *login trend* (0.0637) and the highest *variance in monthly hours* (12.23), suggesting a steady and high-value user base.

- **Cluster 2**: This cluster has an average *hourly rate* of 68.32 and a younger average *age* of 27.6 years. Similar to Cluster 1, users in this cluster report high *survey scores* (4.50) and *NLP scores* (0.56). There is also no churn observed in this group, with a *tenure* of approximately 12.9 months. The *average monthly hours* and *average monthly daily logins* are high, but with a slightly negative *login trend* (-0.0164) and a moderate *variance in monthly hours* (11.98), suggesting stable but slightly fluctuating usage patterns.

- **Cluster 3**: This cluster represents users with an average *hourly rate* of 69.52 and an average *age* of 35.7 years. The *survey score* (4.53) and *NLP score* (0.53) are high, indicating user satisfaction. However, there is a 15.9% churn rate in this cluster, with a relatively short *tenure* of around 3.5 months. The *average monthly hours* (1.34) and *average monthly daily logins* (1.27) are notably low, with a stable *login trend* of 0 and no variance in monthly hours, suggesting minimal engagement.

Based on these results, we can propose a tiered pricing model:

- **Basic Tier**: Suitable for Cluster 0, where initial engagement is high, but users churn quickly. A mid-range price could potentially capture these users before they churn.

- **Standard Tier**: Targeting Cluster 3, characterized by low engagement and moderate satisfaction. The pricing should be kept low to encourage retention and increase usage.

- **Premium Tier**: Targeted at Cluster 2, with stable, high engagement and no churn. These users demonstrate high satisfaction and stable usage, making them suitable for a higher price.

- **Elite Tier**: Best suited for Cluster 1, with the highest hourly rate, age, satisfaction, and engagement levels, and no churn. A premium price can be applied, as these users show a high willingness to pay and value the service significantly.

This tiered pricing structure leverages each cluster's unique characteristics to maximize revenue while catering to the specific needs and engagement levels of different user groups. However, to estimate a specific price for each, as pointed out before, it is necessary to apply some more rigorous calculations.

4.3.4 Tier Pricing Estimation

Having established the calculation of the WTP for each cluster in the form of formula (23), it is possible to use linear regression to study the β_n coefficients and measure the result through the chosen proxies. It is not possible to find an exact price since the algorithm cannot be trained with each user's WTP. However, in the absence of this information, it is feasible to determine the coefficients to provide an estimate based on the single price of \$59.99 previously used. Specifically, the goal is to estimate the willingness to pay, which was initially evaluated at around 10% of each user's 10 hours of work according to their hourly rate. Since the average of the demand curve centers around 75 \$/h, the ROI becomes $750 \times 8\%$, approximately yielding the original \$59.99.

However, assuming some variation and a margin of error, as well as a reduced sensitivity to spending for higher hourly rate bands, each user's WTP was calculated within a range of percentages to estimate the coefficients via linear regression. Additionally, analyzing the variance during the preprocessing phase reveals that both the survey score and NLP score have little impact on the model's information as they do not vary significantly between users. Removing these from the formula and training the linear regression model, the following coefficients are obtained:

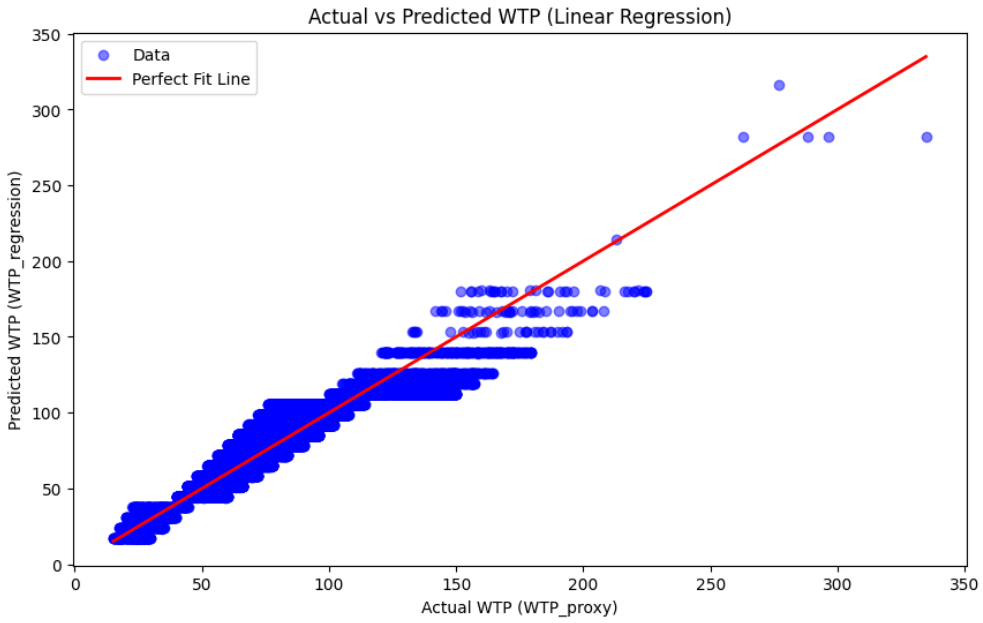


Figure 17: WTP Linear Regression

$$\begin{aligned} \text{WTP} = 77.397 + 27.609 \times \text{hourly\_rate} - 0.0667 \times \text{avg\_monthly\_hours} \\ +0.0812 \times \text{avg\_monthly\_daily\_logins} - 0.0092 \times \text{login\_trend} \\ -0.0771 \times \text{variance\_monthly\_hours} + \varepsilon \end{aligned} \tag{25}$$

Using the average values obtained in the clustering phase with the K-means algorithm, we can feed each cluster into the model to find the average WTP value for each group, thus hypothesizing a price to be assigned to each cluster. To validate this approach, it is reasonable

Cluster	Average WTP
0	\$50.54
1	\$97.40
2	\$69.36
3	\$71.00

Table 2: Average WTP per Cluster

to use the economic approach of setting $MR = MC$ in a monopolistic competition market, as theorized earlier. This entails representing a demand curve for each cluster, using formula (13) and setting marginal costs equal to the respective marginal revenue curves from formula (14). However, two key factors need to be highlighted: first, although each cluster has its own number of users, the total number of users across clusters remains 13,000. Therefore, the MC and Average Cost are now represented by a line at the value assumed in these conditions (at 13,000 users) with MC = \$2 and Average Cost = \$7.5.

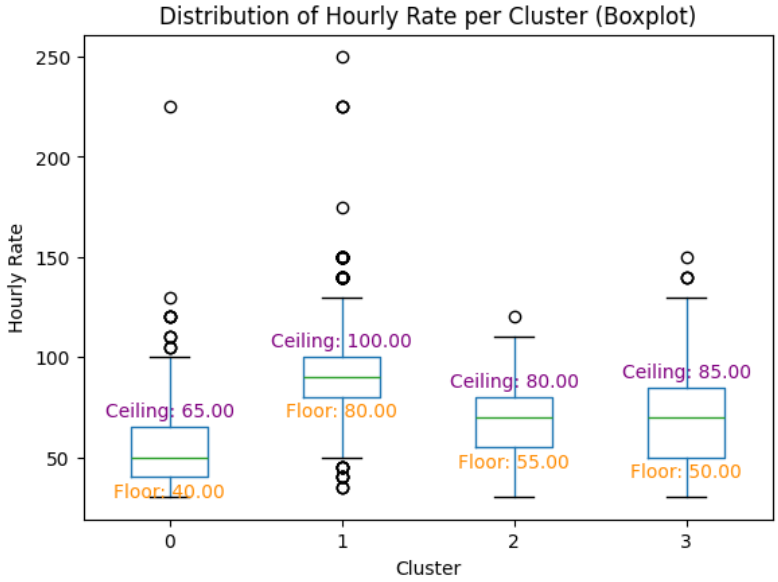


Figure 18: Hourly rate distribution

To further refine the demand curve structure, we can examine the BoxPlot of the four clusters after removing outliers: without the outliers, each group's hourly rate ranges from a minimum

of 30 $/h to a maximum of 150+ $/h, with different medians—specifically, 50, 90, 70, and 70—across clusters, as shown in figure (18).

This allows us to build a demand curve centered around the median, from which the marginal revenue curve can be derived. By intersecting this curve with the marginal costs, the optimal pricing and optimal quantity can be determined.

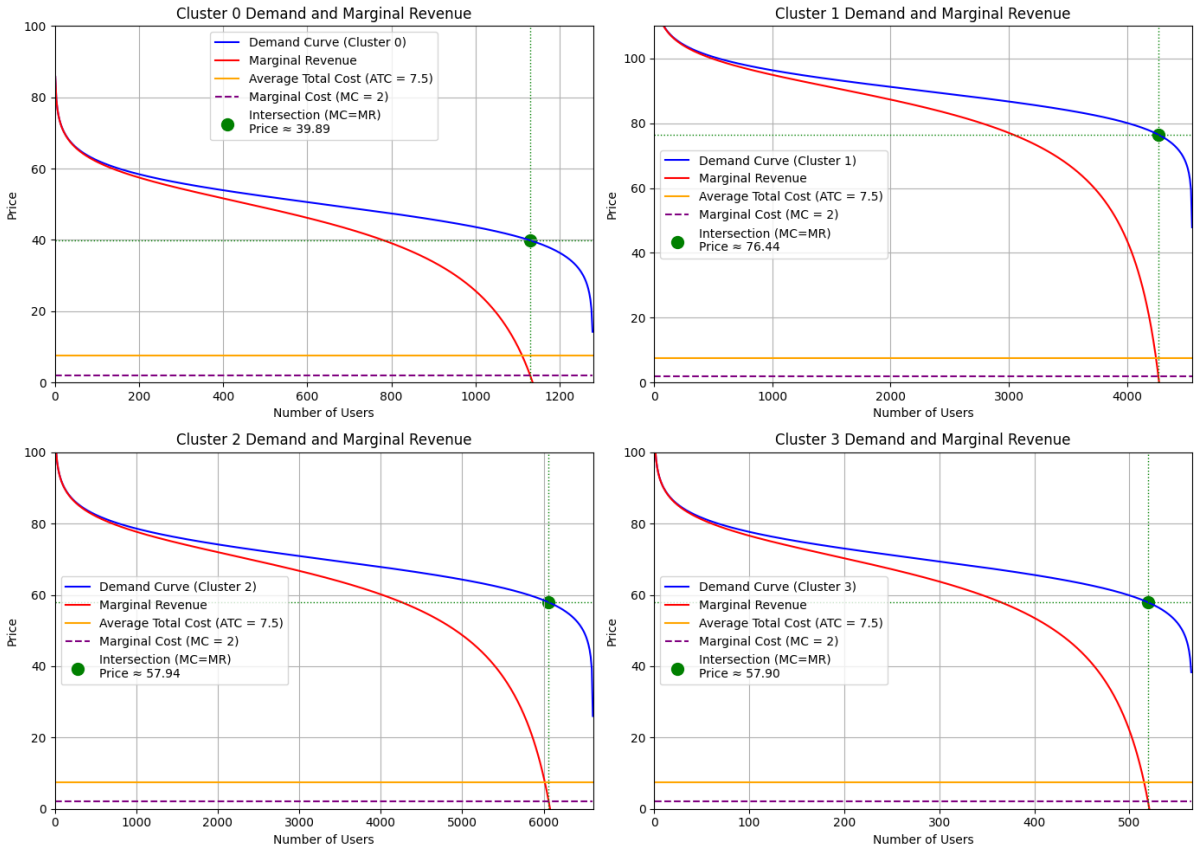


Figure 19: Optimal Pricing per Cluster

This approach highlights why it is necessary to adopt this method, given certain challenges that arise. Firstly, having only the linear regression result is not sufficient to understand the feasible price levels. In comparing with the monopolistic competition model, the price differences are significant (from $50.54, $97.40, $69.36, $71.00 to $39.86, $76.44, $57.94, $57.94). Given the price sensitivity and potential competition from other SaaS providers, selecting a price that maximizes profits without increasing churn risk is essential. However, relying solely on this method would mean treating this model in a limited way: if, for example, a price of $39.86 is set for 1800 users in cluster 0, it does not imply the other 400 users are lost. Instead, they might choose a lower/higher-priced cluster, thereby increasing the number of users in that tier.

The key takeaway from this analysis is the flexibility available in setting the price. Although clusters with a lower WTP (like cluster 0) may have greater price elasticity than clusters with higher hourly rates and incomes (like cluster 1), an A/B testing strategy with initially lower prices and leveraging psychological effects on consumers can be pursued. Based on this, the following tier pricing can be implemented:

- **Basic Tier**: $39.99, targeting cluster 0 with a lower average hourly rate and higher churn propensity.
- **Standard Tier**: $59.99, the original price, with renewal discounts available. Cluster 3 has moderate churn risk but not so high as to warrant a further reduction.
- **Premium Tier**: $67.99, targeted at cluster 2, which has lower churn probability, although the average hourly rate is similar to cluster 3. This setup may make the $59.99 option appear more attractive, mitigating price sensitivity.
- **Elite Tier**: $79.99, the top tier for the cluster 1 with the highest hourly rate. Given their lower price sensitivity and higher income, a higher price can be justified by offering more specific services.

It is important to note that each tier price difference is justified by additional specific services, such as enhanced customer support or advanced technology offerings. This segmentation should represent added value; otherwise, each cluster will gravitate toward the Basic Tier. Thus, understanding user needs and investing in innovation and technology development is essential. Although the costs will be amortized and shared among users over time, they need to be anticipated initially to support cluster segmentation.

4.3.5 Results and Comparison Analysis

Given these results, how do the previously conducted analyses vary with single pricing? Comparing the SaaS metrics analyzed earlier during the EDA phase allows us to determine whether this type of model can indeed lead to price maximization as theorized. This idea is based on the concept that if the willingness to pay of each user exceeds the price charged, there is a portion of consumer surplus that can be converted into producer surplus by raising the price without inducing churn. It is not feasible to predict the churn rate in advance, but greater flexibility potentially reduces churn risk and enables ongoing actions *in itinere* to mitigate this risk. Therefore, although the churn rate derived from the fixed price of $59.99 might decrease with the adoption of tiered pricing, this analysis phase assumes the same churn for previously churned users, comparing *ceteris paribus* with the last month recorded in the database.

1. *MRR*: In the last month, net of previously churned users, an MRR of $697,743.69 was recorded with 11,631 users. By identifying cluster membership and excluding churned users, the MRR would be $841,595.69, higher than the initial figure. This approach is also intuitive when considering cluster composition: since cluster 0 consists of users who churned over time, excluding them from the demand curve's user base raises the price where $MC = MR$, thereby increasing revenues and profits. Additionally, the 1279 users in cluster 0 would not be lost if reducing the price to $39.99 were effective in reducing

churn: even retaining 50% of these users would further boost revenues by $51,147.21, a significant amount.

2. *Churn Rate*: As outlined, the churn rate remains unchanged for the current analysis phase. This can be recalculated in the future when tiered pricing adjustments provide more data.

3. *ARPU*: The Average Revenue Per User now reflects different values across user clusters. Dividing total revenues by active users across clusters, it rises from $59.99 to $72.36, indicating a 20.62% increase.

4. *Gross Margin*: With COGS fixed at $7.5 per user and the adjusted MRR, gross margin rises from 87.5% to 89.64%, reflecting increased profitability after variable costs per user.

5. *LTV*: Lifetime Value, while maintaining the average tenure of 11.69 months, now increases due to the higher ARPU, resulting in $845.89 compared to the previous $701.47.

Comparing these metrics confirms the initial hypothesis that price differentiation enhances SaaS business revenues (and thus, with unchanged costs, profit). However, there are some potential challenges to consider. Issues such as increased costs or rising churn could arise. If the costs of creating added value to justify higher prices are substantial enough to reduce profit despite increased revenues, tiered pricing might not be suitable. Nonetheless, it is essential to analyze both short- and long-term impacts, as development costs are typically amortized over time. If, for instance, this temporary decrease in profit enables access to a broader audience or allows for a recovery of the investment during the payback period, maximizing future profits, a temporary trade-off may be acceptable.

Regarding churn risk, increased prices may affect retention if competitors offer comparable services at lower prices. Therefore, it is crucial to compare tier pricing with competitors' pricing and to emphasize the importance of A/B testing with a smaller user group to understand the response to pricing changes before implementing them across the entire user base.

Ultimately, the analysis highlights the effectiveness of differentiated pricing based on various clusters: having distinct user groups with varied WTP necessitates price differentiation to efficiently capture consumer surplus.

4.4 Dynamic Application of the Model

Up to this point, we have discussed how, based on current data, it is possible to aim for profit maximization by differentiating prices using machine learning to cluster various types of users and applying price discrimination theories. However, due to the capabilities of advanced technology, computational speed, and the entirely online nature of SaaS, even more effective methodologies exist for increasing revenue. This is achievable because we can relatively easily track each user's behaviors, understand trends, and constantly categorize them into clusters

that represent their unique attributes. In fact, by establishing various parameters to monitor, although a user may have chosen a particular pricing tier at the time of subscription, it is possible to counter churn or, conversely, increase their lifetime value by encouraging additional spending.

This monitoring applies to both new and existing users; the real strength of this approach lies in collecting meaningful data and regularly training machine learning models by tracking user responses. Practically, as demonstrated thus far, collecting and processing a dataset to train a model using machine learning algorithms enables the use of that model for forecasting data not available during the training phase. For instance, if a new or existing user exhibits unique values among those analyzed (such as survey score, NLP score, average monthly hours, etc.), feeding this data into the model allows for an effective estimation of their WTP and reduces the likelihood of churn through the offer of discounts, add-ons, or alternative subscription plans.

In the following sections, some key considerations will be analyzed for the adoption of a dynamic price discrimination process. This automated system—without requiring ad hoc human intervention and decision-making—can independently act on each user by tracking their behaviors and consistently offering the optimal pricing solution. This enables the company to cover the user's WTP effectively while minimizing losses due to inefficiencies.

4.4.1 Simulation on Changing Behaviors

To maintain a sustainable pricing strategy, it is essential to simulate and analyze potential changes in user behavior over time. This simulation can be performed by setting up scenarios where key metrics—such as login frequency, time spent on the platform, and user satisfaction scores—change dynamically. These changes can be analyzed to understand their potential impact on user retention, engagement, and overall revenue.

For example, by adjusting user engagement parameters in the model (such as reducing monthly login frequencies or adjusting session durations), the model can predict how these behavioral changes affect willingness to pay (WTP) and churn rates across different user clusters. Machine learning algorithms, such as decision trees or logistic regression, can help identify thresholds where certain behaviors may predict churn or a reduction in subscription level. The goal is to simulate possible outcomes for scenarios where user behaviors deviate from their current patterns and to adapt the pricing model to accommodate these changes.

In a SaaS environment, this tracking is simplified, as actions unrelated to purchasing decisions can still contribute to forecasting based on model parameters. For instance, in the case of the synthetic database used, values within each column are tracked as user behavior within the service, not directly set by users (such as declarations of hourly rate, age, or nationality). Logins, time spent, and open tickets are all metrics that contribute to clustering users within groups with specific willingness to pay. Therefore, changes in usage time, ticket activity, or service scores may alter a user's cluster assignment, resulting in pricing adjustments that more accurately reflect their WTP, which is determined by engagement levels in addition to observable

attributes and price sensitivity based on purchasing power. As previously noted, engagement significantly influences both WTP and churn values.

Additionally, with the introduction of tier pricing, these behaviors, choices, and churn probabilities can be analyzed with additional variables. For example, as further discussed, the addition of add-on products can be tracked by monitoring click counts on pages related to certain services, sensitivity to discounts during sales periods (such as Black Friday), as well as the content of user messages—not only in terms of sentiment analysis but also in content itself—to identify demands for additional services, resolve doubts, or provide targeted incentives for both marketing and sales.

Marketing becomes another area that can be developed through the tracking of such dynamic content. The more choices and behaviors tracked, and the more outcomes known, the more it becomes possible to outline a common profile to target with various offers and services, as well as the most appropriate engagement and marketing communications. Consequently, multiple sub-levels of clustering can be recognized:

- *Clustering for Tier Pricing*: The primary level, previously analyzed, determines the tier pricing levels to be offered. These clusters should not be too numerous, as discussed, as it is essential to limit users' options to specific offer levels to maximize acquisition opportunities through the challenging acquisition funnel in any SaaS.

- *Clustering for Groups of Similar Users*: A second, broader level of clustering can be implemented to identify different behaviors within the same primary cluster that, while sharing the same pricing tier, may exhibit different churn or engagement levels. The creation of these narrower subgroups is facilitated by the wealth of obtained data: for some subgroups, offering a discount may be effective, whereas for others in the same primary cluster, it may not be necessary. The flexibility to understand the varying patterns of similar users within clusters allows for better anticipation of their needs, making the pricing discrimination process more efficient in response to these needs.

- *Individual Clustering*: The ultimate expression of clustering divides each individual based on unique attributes. Although implementing this level is challenging (due to the absence of perfect knowledge of each user's WTP), it can be developed in specific areas, such as customizable add-on features. For instance, users in the same primary and secondary clusters may differ in personal preferences. This could apply to integrations with an additional third-party service, for which a personalized offer may be proposed. Continuous monitoring of user behavior in this context is essential for presenting targeted purchase offers, automating the decision to withdraw or persist with the offer based on purchasing probabilities.

The dynamic clustering and personalized pricing system can therefore increase in complexity and effectiveness based on the SaaS business structure and services offered. It is crucial

to emphasize that monitoring each user's behavior within the SaaS, in addition to adapting to the type of product offered, is essential for an efficient pricing management system. The algorithm used to track specific metrics automates the decision-making process, but it requires a well-defined logic that avoids being overly invasive, thereby preventing unfairness in pricing management, while also capitalizing on multiple revenue opportunities.

4.4.2 Subscription of a New User

A newly registered user, unlike an existing one within the system, lacks the same amount of data necessary to accurately position them within a specific cluster. On the one hand, this increases the risk of inefficiency in the retention process and the associated goal of maximizing their potential Lifetime Value (LTV). On the other hand, however, it provides an opportunity to leverage real-time responsiveness in the delivery of services, which can help minimize inefficiencies.

Upon registration, the user makes certain choices that are indicative of their likely cluster, to which they may have been accurately directed through targeted Customer Acquisition Cost (CAC) efforts by the marketing department. Continuing with the example of the SaaS model explored in this thesis, the registration process alone provides several factors that can help predict cluster affiliation (e.g., using linear regression or decision tree models) and their own WTP. Parameters such as nationality, age, hourly rate, and the choice of subscription plan within tiered pricing options support the initial classification of the user into an *initial cluster*.

From this point onward, as the user interacts with the SaaS platform, each action allows for further refinement of their cluster assignment, maximizing profit by offering options that enhance their utility. The fewer data points available, the greater the risk of missing early indicators of churn, limiting proactive retention responses. Thus, rather than losing a customer, it is preferable to generate slightly lower revenue if it supports the potential for longer-term retention.

The ultimate goal is to implement continuous clustering recalibration, whereby each significant interaction fine-tunes the model's understanding of the user. By using advanced analytics and machine learning models like clustering or logistic regression, companies can make real-time adjustments to a user's LTV calculation, tailoring both retention and revenue-maximizing strategies more precisely.

Onboarding a new user in a SaaS platform is a critical moment for long-term success. An effective onboarding experience is not one-size-fits-all but rather adapted to the user's initial cluster designation. Each cluster, based on its characteristics, requires tailored onboarding strategies to maximize its users' engagement and minimize the likelihood of churn. For example, a user assigned to a low-WTP cluster may benefit from an onboarding experience that focuses on the platform's essential features, providing guidance on how these features can solve specific pain points. High-WTP users, on the other hand, may require a more advanced onboarding approach, emphasizing exclusive features or add-ons that align with their professional needs. In either case, the onboarding program serves not only as an introduction to the platform but as

a powerful tool for increasing immediate engagement and setting usage patterns that maximize LTV over time.

Once the user is fully registered and engaging with the platform, behavioral data can further refine their cluster position, proactively identifying churn risks. Factors like declining usage frequency, decreased engagement with premium features, or low customer satisfaction scores can indicate a user's potential churn likelihood. Machine learning models, particularly those trained on time-series data, are well-suited to track these behaviors and send alerts for early intervention. This proactive approach allows the SaaS provider to implement retention strategies tailored to the specific risks associated with each cluster. For instance, if a user from a high-WTP cluster shows signs of declining engagement, it may be appropriate to offer exclusive content, additional features, or personalized support. Conversely, for users in low-WTP clusters, offering targeted discounts or adjusting the pricing of certain add-ons can provide an incentive to retain their business.

4.4.3 Reducing Churn Probability by Acting on Engagement

Previously, we discussed using the WTP formula (23) to determine, via logistic regression, the formula (24) for calculating churn probability. With differentiated data available, it becomes possible to assign a dedicated churn probability alongside clustering by estimating the likelihood of churn for users with similar characteristics to those of WTP. Especially for SaaS businesses where survival depends on recurrent revenues, having a tool that monitors churn probability with each behavioral change—and reacts automatically when a certain threshold is exceeded—becomes essential for maximizing profits by directly extending the Lifetime Value (LTV) through an increase in tenure duration.

Each behavior contributes to increasing or decreasing the probability of churn due to different engagement factors that shape each user's perception of the product's value, which is dynamic and variable. Keeping users engaged with the platform can be achieved through various means:

One method is to encourage extended usage periods through the strategy of annual payment. Offering a discount on an annual plan compared to a month-to-month plan not only aids in better managing the company's cash flow and payback period, but also ensures that users will engage with the service for at least a year, increasing minimum tenure. This extended tenure allows for greater data accumulation and enables precise responses to actions that signal churn risk.

Another, non-exclusive, approach involves leveraging engagement through discounts, trials, and personalized offers tailored to the user's needs. Demonstrating the benefits of the services offered and their impact on individual utility is the most effective way to encourage the user to engage with the product. The more consistent and essential the usage becomes, the more the perceived value and willingness to pay increase (barring any external issues affecting purchasing power).

A third, complementary approach involves employing gamification. Human biases toward

competition, achievement, and goal attainment provide the emotional background for fostering engagement with a SaaS platform. Allowing users to view their achievements through badges, leaderboards, or visual representations of SaaS usage levels (such as a progression bar or sound notification) enables tracking engagement and encourages users motivated by progression to reach higher levels of usage[44]. Without resorting to manipulative tactics, which only have destructive long-term effects, a win-win condition can be created, allowing users to recognize the product's value while enabling the company to minimize churn risk due to engagement inefficiencies.

Churn is directly linked to each user's LTV: while differentiated pricing boosts revenues, retaining a stable ARPU for as long as possible extends the SaaS business's lifespan and enhances the so-called magic number—the ratio of Customer Acquisition Cost (CAC) to LTV—allowing for efficient investment in product promotion and user acquisition resources.

4.4.4 Implementing Add-ons

Previously, we discussed the introduction of add-ons in relation to the potential for dynamic responses to changing user behaviors. Also known as plug-ins, add-ons are software components (in this case, SaaS) that add functionalities to an existing program. These add-ons are distinct from the features that differentiate one pricing plan from another and can be purchased individually to customize the user experience. Sometimes, they may be purely aesthetic (such as graphic customization of the service) or add practical features, such as integration with third-party software, additional storage, or artificial intelligence tools for data management.

In the SaaS landscape, differentiation does not stop with single and tiered pricing. There are various ways to increase the user base, conversion potential, engagement, and retention. As discussed in Chapter 2, there may be freemium plans tied to resource usage or, as in this case, add-ons that enhance the perceived value of the software, either through discounted access to additional features or by offering valuable functionalities at full price. Add-ons can be used in both ways, with the primary goal of increasing the user's Lifetime Value (LTV).

Based on dynamic variations in behavior, it is possible to track users' needs by analyzing catalog navigation patterns and the level of interest shown in specific add-ons over others. A recommendation system based on other users within the same cluster can further enhance this strategy. Once again, SaaS infrastructure allows companies to evaluate whether developing such services aligns with the software's architecture and whether their implementation adds sufficient value to justify the development costs. However, if a service analysis shows that offering such solutions is feasible, it can contribute to both profit maximization (by increasing revenue through additional purchases beyond the subscription, thereby raising ARPU) and the extension of LTV through enhanced engagement.

4.4.5 Obstacles to the Implementation

Although a range of possibilities for employing a proactive, dynamic, and automated approach has been outlined, it is not always possible to overcome all obstacles related to implementing such a strategy. Technical issues, regulatory compliance, data privacy management, cookie policies, and data quality can hinder the efficiency of these actions. At the data level—necessary as the ”fuel” for machine learning models—regulatory constraints, and even user feedback regarding perceived unfairness due to pricing discrimination can present challenges in building an effective monitoring system.

For instance, countries in the European Union, among others, impose stricter regulations compared to the USA, both in terms of AI-driven product development[45] and cookie/data management as per the GDPR[41]. This results in limited tracking options if users do not consent to cookies and imposes restrictions on retaining certain types of information. In countries like China, SaaS providers face additional structural challenges: for example, they must host their domains and databases on Chinese servers[46], potentially leading to increased costs and complexities that may impact profitability despite user acquisition. Tracking user behavior in these areas becomes more stringent, reducing the efficiency of clustering and dynamic pricing offerings.

In addition to regulatory and structural issues, there are also technical difficulties in identifying suitable machine learning algorithms and models to monitor specific behaviors. Additionally, challenges related to maintaining positive user engagement and the perceived fairness of pricing discrimination can arise. As discussed in Chapter 2, effective price discrimination requires that consumers are unable to circumvent assigned pricing. Beyond the risk that informed users may churn due to perceived inequality, the ability to take actions that might lower prices poses a challenge.

For instance, while it is crucial to track behaviors to determine appropriate pricing, there is also the risk that these behaviors become recognizable and manipulable by users. Certain economic scenarios allow for price variations as trade-offs between the product's current and future value—for instance, paying a higher price now versus waiting for a seasonal discount, as seen with Black Friday sales. However, at other times, users' deliberate price manipulation can undermine the architecture of clustering and price discrimination algorithms.

A low NLP score might, for example, influence the likelihood of increasing churn rates, potentially prompting discount offers. In this case, a cost-conscious user—aiming to reduce the company's revenue intake—could engage in forums or leave reviews that negatively impact sentiment analysis. Such behavior may not truly reflect the user's actual engagement level and could artificially inflate their perceived risk of churn. To mitigate this, it would be beneficial to assign different weights to various characteristics; for instance, using a product infrequently over a prolonged period is more indicative of engagement than an intentionally negative NLP

[41] General Data Protection Regulation (GDPR) is a European Union (EU) law that protects the privacy and security of personal data.

score. Indeed, it is unlikely that a user would forego using a useful, already-purchased service solely to secure a lower price.

It is thus essential to consider these relationships when designing a dynamic pricing system. Some manipulations are challenging to counteract—for example, when a full-paying user seeks a discount code or feigns churn to receive a retention offer at a reduced price. However, by structuring these challenges adequately, the system can become more efficient where human behavior surpasses simple clustering calculations.

Other obstacles may directly affect the predictive capabilities of regression or decision-making models. For example, when the user base is small and underfitting is a risk, reinforcement learning models or scenario analysis (on varying behavior scenarios) can be employed to support the training of models synthetically. In this context, real data should be weighted more heavily than synthetic data, which nonetheless contributes valuable insights into churn likelihood and purchasing behaviors for each user.

5 Discussion of Results and Conclusion

This thesis aimed to analyze the benefits of price discrimination within the SaaS sector through the use of algorithms and the computational power of Machine Learning. Technological tools and computational techniques make it possible not only to handle the complexity of economic analyses and decisions but also to automate responses in real time. Through the application of these algorithms and techniques, it has been concluded that economic inefficiencies driven by information asymmetry, which separates the consumer from the producer, can be partially mitigated by capturing information that, while not explicitly defined, can be collected through the interplay of multiple characteristics.

As introduced in the initial chapter, the SaaS field is increasingly relevant and growing. Compared to other business types, it is more affected by market volatility. The intangibility of the products and services offered makes it difficult to clearly define consumer value and their resulting willingness to pay. However, it is precisely this structure, thanks to its fully online nature, that enables the precise capture of essential information required for calculating costs, revenues, KPIs, and appropriate methods of price discrimination.

In Chapter 2, the industrial economic theory underlying the SaaS business model and its cost/revenue structure was discussed, identifying the market type as monopolistic competition. In the short term, this model assumes certain monopoly characteristics despite the presence of competition from other companies differentiated by products, technology, and innovation. Furthermore, by describing the different types of price discrimination, it was possible to understand how these assumptions can be applied in the SaaS context and which conditions are beneficial for further investigation through machine learning algorithms.

The third chapter was dedicated to this purpose, with an in-depth focus on statistical techniques and machine learning approaches relevant to the initial research question. These were subsequently used in the analysis and model development in Chapter 4, where, through the application of a synthetic dataset, it was possible to quantitatively compare metrics and assess the revenue benefit brought by pricing differentiation versus a single-price strategy.

It was agreed that, by tracking data and clustering, it is possible to segment different market demands and identify an optimal price for each, transferring excessive consumer surplus into producer surplus. This approach maximizes revenues and, consequently, profits, as hypothesized in the research question.

In these concluding sections, the strengths and limitations of this approach will be objectively discussed, along with a summary of related questions that, due to academic constraints, were not addressed.

5.1 Interpretation of Simulation Results

As highlighted in the previous chapter, the results of applying clustering models for revenue and profit growth show a clear advantage when compared to the metrics analyzed under a single pricing strategy. Referring back to the quantitative data, the increase in revenue—accounting for a decrease in churn rate—was significant enough to justify the added complexity of a tiered pricing model. This approach, with an equal number of users, enabled a substantial improvement in overall gains.

5.1.1 Discussion of Empirical Results Obtained from the Simulation

The economic implications of this result are substantial and will be explored in the following section. Here, however, the focus is on the empirical validity of the study's findings. It is important to note that these results are derived from a synthetically constructed database rather than real-world data extracted from actual user behavior. Nevertheless, the empirical value of this study remains significant, as it models a typical SaaS environment and focuses solely on the numerical insights of metric changes while excluding derived impacts on metrics such as churn and costs.

Indeed, by excluding secondary responses to changes—which are difficult to predict due to the influence of competitors and the broader user ecosystem—the benefits of adopting differentiated pricing become clear. This approach allows for increased revenue opportunities even in saturated markets, as it captures the existence of different demand curves within the aggregate demand. Additionally, the value of using machine learning algorithms as both a simplifier and enhancer of clustering and regression calculations is confirmed within the analytical approach taken in this study. Machine learning plays a fundamental role in initially identifying different pricing tiers and later automating dynamic pricing and offer decisions based on user behavior.

Without leveraging the computational power of modern technology to implement statistical concepts through machine learning, much of the potential reactivity and efficiency available to a SaaS business is lost. Machine learning enables the identification of patterns that would otherwise be concealed from human judgment, potentially through unsupervised models like neural networks, principal component analysis (PCA), or deep learning methods more broadly.

5.1.2 Comparison with Theoretical Expectations and Practical Implications

As introduced at the beginning of the conclusion, the expectations regarding the adoption of price discrimination supported by machine learning have been met in the practical analysis and implications obtained. Initially, the question was whether the trade-off of increased pricing complexity could lead to higher revenues by maximizing output. The results confirm that implementing tier-specific pricing, following clustering analysis, has the potential to increase profit—even to the point of maximization—by favoring differentiated pricing over uniform pric-

ing.

The company's approach of targeting the consumer's excessive surplus directly highlights the structural inefficiency within this type of market, where imperfect knowledge of the consumer's willingness to pay (WTP) generates inefficiencies. These inefficiencies prevent the full extraction of profit above the relatively low marginal costs typical of the SaaS market.

5.2 Economic and Strategic Implications

The economic implications of this research are numerous within the field of reference. Initially, it was discussed how small differences, especially in more niche markets like those of Micro SaaS, can be the decisive factor between value creation and, consequently, the long-term sustainability of a company versus bankruptcy.

For the economic environment, ensuring that value is produced is critical to general growth, as this is only achievable when a company can sustain itself over time. This ability allows value to be provided to users and wealth to be generated for the broader stakeholders within the ecosystem. Consistent growth, along with increased revenues, enables investment in future innovations, further reinforcing this cycle.

5.2.1 Reflection on Economic Implications for Small SaaS Companies

In particular, reference can be made to the impact on Micro SaaS. It has already been highlighted that raising revenues while maintaining user quantities is made feasible through price discrimination. This is an extremely valuable property for a Micro SaaS where, due to its niche market and the rapid potential for market saturation, an alternative approach to simply expanding the user base is required to sustain growth over time.

Additionally, considering the monopolistic competition phase, it is important to remember that this applies primarily in the short term, while, in the long term, competition lowers the demand curve due to the increasing homogeneity of competing offerings, which brings the willingness to pay (WTP) closer to marginal costs. Therefore, increasing revenue before reaching this point—thanks to price discrimination—enables the accumulation of resources to invest in the development of features that further differentiate the business from the competitive ecosystem surrounding it.

This approach maintains a monopolistic competition structure in the short term, allowing the Micro SaaS to capitalize on this favorable structure for a more extended period. Furthermore, the research can result in the creation of new products, expanding the user base and available market share. This ensures the company's growth not only over time but also in terms of employees involved, thus increasing value for society and the economic environment in terms of wealth redistribution and overall well-being.

5.2.2 Strategic Recommendations Based on Research Findings

The findings from this research support a series of targeted strategic recommendations aimed at enhancing revenue potential and sustaining user engagement for SaaS companies. By implementing focused adjustments and streamlining actions into core strategies, companies can maximize the benefits of a differentiated pricing model.

- *Prioritize a Tiered Pricing Structure with Value-Based Add-Ons:* Introducing a tiered pricing structure, alongside targeted add-ons, can better address diverse user segments with varied willingness to pay. Offering tailored value within each tier—such as premium features or extended services—allows users to see the advantage in moving up tiers or adding components that improve their experience, leading to higher revenue per user without the need for aggressive upscaling.

- *Leverage Predictive Analytics for Dynamic Adjustments:* Utilizing predictive analytics to monitor user behavior and engagement in real-time can guide pricing adjustments based on evolving needs. Predictive insights allow the model to identify user churn risks, highlight engagement opportunities, and optimize pricing in response to trends. By making data-driven adjustments, companies can tailor pricing to reflect actual user value perception, which reinforces user satisfaction and retention.

- *Enhance User Engagement to Drive Retention and Value Perception:* A focused engagement strategy, starting with personalized onboarding and followed by relevant feature suggestions, creates immediate and lasting user value. When users understand and experience product benefits early on, they tend to remain engaged, perceive a higher value in the service, and, ultimately, increase their lifetime value. Proactively responding to user behavior also helps keep them involved, while strategies like gamification and periodic updates can sustain engagement over time.

- *Incorporate A/B Testing and Pilot Adjustments for Optimal Pricing Decisions:* Testing pricing changes on smaller user groups allows companies to refine their approach based on real-time feedback before scaling adjustments. A/B testing different price points, feature bundles, or add-on offerings enables insights into user preferences and responses, helping SaaS providers reduce risks and adjust prices based on verified user reactions.

- *Maintain Transparency and Compliance to Foster Long-Term User Trust:* Adhering to privacy and data compliance (e.g., GDPR, CCPA) and practicing transparency in pricing policies establishes trust with users. This clarity minimizes perceived price manipulation, which can lead to churn. Transparent communication around data use and personalized pricing further reinforces a fair and user-focused approach, especially when implementing a differentiated pricing structure.

Through these strategic actions, SaaS companies can leverage a data-driven, user-centric pricing model to maximize revenue and encourage long-term loyalty. This balanced approach—by merging actionable insights, compliance, and dynamic engagement—enhances the company's competitive position, enabling sustainable growth and added value for users and stakeholders alike.

5.3 Limitations and Future Development

The limitations of this research stem from the academic nature of the study. Economic practice, unlike theory, cannot be fully captured in simplified models that often overlook the contingent complexities of real-world scenarios. Therefore, it is not possible to distill every case within the SaaS domain into a universal principle.

The approaches examined in this research may not suit all business models, particularly those with unique cost structures, variations in market payment capacity, or distinct price sensitivities due to differing levels of competition. However, the implications drawn from the findings serve as an adequate foundation for evaluating the feasibility of applying these models to a business's specific circumstances.

In the following section, we analyze several final considerations regarding the limitations of this study and the model applied.

5.3.1 Simplistic Model

The current model relies on a synthetic dataset constructed based on theoretical assumptions about customer segmentation and pricing. In a sense, the findings from the EDA and clustering analyses are directly influenced by the initial construction of this dataset. Working with data from real users could yield unexpected results, potentially diverging from the initial assumptions and leading to both quantitative and strategic shifts. While the benefits of certain algorithms remain valuable, it is essential to understand the outcomes from studying a real-world dataset to effectively select exploratory analyses, define clustering, and establish appropriate tiered pricing levels.

Seasonality is another critical aspect, as are trends identified by time series analysis methods commonly used in real-world settings, such as ARIMA and Holt-Winters models. Conducting a more in-depth study of time series allows for better understanding of the factors driving churn and engagement, thereby supporting a more accurate and effective approach to dynamic pricing.

Moreover, as implied but now highlighted, the influence of external market conditions cannot be overlooked. Economic events, regulatory changes, shifts in market demand, technological advances, and competitor activity can dynamically alter results and users' willingness to pay (WTP). This further complicates the task of identifying optimal pricing strategies that effectively maximize profits while minimizing churn and capturing user surplus.

Addressing these limitations in future studies will allow for a more comprehensive understanding of differentiated pricing in SaaS. By refining these models, incorporating actual user data, and adapting to diverse market dynamics, future developments can expand the applicability of these findings, ultimately creating a more robust and actionable framework for revenue optimization in SaaS environments.

5.3.2 Examples of More Advanced Developments

The model presented in this research can be therefore expanded to accommodate a range of sophisticated advancements, addressing some limitations observed in the application of pricing discrimination for SaaS. By enhancing the model through advanced data integration, deeper predictive analysis, and refined user segmentation, it is possible to develop a more adaptive and nuanced pricing strategy that aligns closely with real-world conditions. These improvements aim to elevate the model's responsiveness to user behavior, increase pricing accuracy, and ultimately maximize revenue while reducing churn.

For example, integrating real-time data tracking into the model could provide a more immediate understanding of user engagement metrics and responsiveness to pricing adjustments. Such a real-time update system would allow the model to react dynamically to changes in user behavior and preferences, enabling timely adjustments to both pricing tiers and retention strategies.

In addition, advanced machine learning techniques, such as neural networks or Principal Component Analysis (PCA), could reveal latent patterns within user data, enabling a more granular segmentation that captures subtle differences between users. These techniques would support a more precise approach to personalized engagement, potentially increasing user satisfaction and willingness to pay (WTP) for additional features or tiers.

Moreover, implementing a robust testing and validation framework, such as A/B testing and scenario analysis, would allow for an empirical assessment of pricing strategies, comparing different approaches to user retention and revenue outcomes. This kind of structured testing would improve the decision-making process, allowing for optimized dynamic pricing that aligns with user behavior trends and market conditions.

Another critical enhancement is *constant fine-tuning* of the model parameters to ensure accuracy in response to shifts in user behaviors and external market factors. Fine-tuning entails periodically updating the model with new user data and adjusting algorithm weights based on recent trends in user engagement, WTP, and churn rates. This ongoing adjustment allows the model to learn from new patterns and anomalies, optimizing its performance over time and adapting the pricing strategies to ever-evolving user and market dynamics.

Lastly, integrating external market data on competitor pricing and demand trends would enable the model to adapt its pricing strategy to the broader competitive landscape. By contextualizing user data with market conditions, the model can enhance its pricing precision, making it better suited to changes in user WTP in response to external factors. This capability would

also support targeted engagement initiatives, where pricing is adjusted based on anticipated shifts in user demand or market competition.

In conclusion, this thesis aims to introduce a topic and establish a foundation for research that can later be expanded and specialized within more complex frameworks. These frameworks would be based on data reflecting membership in structures that are more interdependent on data and the factors influencing them.

References

[1] Statista. *Software as a Service - Worldwide*. Accessed: 2024-09-17. 2024. URL: https://www.statista.com/outlook/tmo/public-cloud/software-as-a-service/worldwide.

[2] Harvard Business Review. *What Big Companies Can Learn from the Success of the Unicorns*. Accessed: 2024-09-17. 2016. URL: https://hbr.org/2016/03/what-big-companies-can-learn-from-the-success-of-the-unicorns.

[3] The Economist. *The Age of the Unicorn Is Over*. Accessed: 2024-09-16. 2024. URL: https://www.economist.com/business/2024/02/22/the-age-of-the-unicorn-is-over.

[4] Shivam Gupta, Milind Sathye, and Karminder Singh. "Software as a Service (SaaS) Cloud Computing: An Empirical Investigation on University Students' Perception". In: *ResearchGate* (2021). URL: https://www.researchgate.net/publication/351401872_Software_as_a_Service_SaaS_Cloud_Computing_An_Empirical_Investigation_on_University_Students'_Perception.

[5] Skywinds. *Is Micro SaaS Blowing Up in 2025?* Accessed: 2024-09-23. 2024. URL: https://medium.com/@dpskywinds/is-micro-saas-blowing-up-in-2025-d97d45e16250#:~:text=Statista%20data%20provides%20a%20comprehensive,reach%20%24295.08%20billion%20by%202025..

[6] Standars and Poors. "451 Research: 2024 Trends in Data, AI and Analytics". In: *Standars and Poors* (2024). URL: https://pages.marketintelligence.spglobal.com/CIQPro-2024-Top-TMT-Trends-OutlookReportDataAIAnalytics.html?utm_medium=cpc%5C&utm_source=google%5C&utm_campaign=CIQ_Solutions_TMT_2024_Trends_Search_Google%5C&utm_term=ai%5C&predictions%5C&utm_content=688577911386%5C&gclid=Cj0KCQjwxsm3BhDrARIsAMtVz6OQOAf4DT6fS4Ti1wbpw3RQqduhlAa-1Wxf_tgKZ7v78BHgz3QHupEaAth3EALw_wcB.

[7] Dr Lily Popova Zhuhadar. *Unraveling AI Complexity - A Comparative View of AI, Machine Learning, Deep Learning, and Generative AI.* Accessed: 2024-09-24. 2023. URL: https://commons.wikimedia.org/wiki/File:Unraveling_AI_Complexity_-_A_Comparative_View_of_AI,_Machine_Learning,_Deep_Learning,_and_Generative_AI.jpg.

[8] The Economist. *What happened to the artificial-intelligence investment boom?* Accessed: 2024-09-24. 2024. URL: https://www.economist.com/finance-and-economics/2024/01/07/what-happened-to-the-artificial-intelligence-investment-boom?utm_medium=cpc.adword.pd%5C&

utm_source=google%5C&ppccampaignID=18156330227%5C&ppcadID=%5C&utm_campaign=a.22brand_pmax%5C&utm_content=conversion.direct-response.anonymous%5C&gad_source=1%5C&gclid=Cj0KCQjwxsm3BhDrAR
osZmu07uDPkQaAncAEALw_wcB%5C&gclsrc=aw.ds.

[9] Mohsen Attaran and Jeremy Woods. "Cloud Computing Technology: A viable Option for Small and Medium-Sized Businesses". In: *ResearchGate* (2018). Accessed: 2024-09-24. URL: https://www.researchgate.net/publication/327756016_Cloud_Computing_Technology_A_viable_Option_for_Small_and_Medium-Sized_Businesses#fullTextFileContent.

[10] J. O'Sullivan. *Why do firms exist? Why are some activities directed by market forces and others by firms?* Accessed: 2024-09-24. Sept. 2017. URL: https://www.economist.com/the-economist-explains/2017/09/18/why-do-firms-exist.

[11] McKinsey and Company. *Grow Fast or Die Slow*. Accessed: 2024-09-24. 2014. URL: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/grow-fast-or-die-slow.

[12] Anthony Witherspoon. *10 Top Micro SaaS Examples: Building Profitable Apps for Success*. Accessed: 2024-09-24, 19 min read. June 2024. URL: https://www.saasalliance.io/10-top-micro-saas-examples-building-profitable-apps-for-success/#what-is-micro-saas.

[13] Austin Hammer. *Understanding and Optimizing Unit Economics in SaaS: A Guide for Founders*. Accessed: 2024-09-24. Nov. 2023. URL: https://softwareequity.com/blog/saas-unit-economics.

[14] Tomasz Tunguz. *The Importance of Payback Period for SaaS Startups*. Accessed: 2024-09-24. Sept. 2015. URL: https://tomtunguz.com/payback_period_cash/.

[15] David Skok. *SaaS Metrics 2.0 – A Guide to Measuring and Improving what Matters*. Accessed: 2024-10-28. 2019. URL: https://www.forentrepreneurs.com/saas-metrics-2/.

[16] Dan Ma and Robert J. Kauffman. "Competition Between Software-as-a-Service Vendors". In: *IEEE* (2014). Accessed: 2024-09-24. URL: https://ieeexplore.ieee.org/abstract/document/6857369.

[17] Jeffrey Church and Roger Ware. *Industrial Organization: A Strategic Approach*. Boston, MA: McGraw-Hill, 2000.

[18] Amazon Web Services. *AWS Pricing Calculator*. Accessed: 2024-09-24. 2024. URL: https://calculator.aws/#/.

[19] Jeffrey M. Perloff. *Microeconomics: Theory and Applications with Calculus*. 4th. Boston, MA: Pearson, 2014.

[20] Steven Berry, James Levinsohn, and Ariel Pakes. "Automobile Prices in Market Equilibrium". In: *Econometrica* 63.4 (July 1995), pp. 841–890. URL: http://links.jstor.org/sici?sici=0012-9682%28199507%2963%3A4%3C841%3AAPIME%3E2.0.CO%3B2-U.

[21] Tomasz Tunguz. *The Obscure Economic Idea Behind SaaS Pricing Challenges*. Accessed: 2024-09-24. Oct. 2014. URL: https://tomtunguz.com/obscure-economic-concept-behind-saas-pricing-challenges/.

[22] Dr. Ken Fordyce. *Some Basics on the Value of S Curves and Market Adoption of a New Product*. Accessed: 2024-09-24. Apr. 2020. URL: https://blog.arkieva.com/basics-on-s-curves/#:~:text=S%20Curve%20Logistics%20Equation&text=S(x)%20=%20(1,the%20formula%20I%20use%2C%20where.

[23] Zan Zhang. "Competitive Pricing Strategies for Software and SaaS Products". In: *Journal of Business Research* (2020). Accessed: 2024-09-24. URL: https://www.sciencedirect.com/science/article/abs/pii/S0378720620303050.

[24] Tom Mohr. *Scaling the Revenue Engine — Chapter 9: Pricing and Packaging Precise pricing powers profit.* Accessed: 2024-09-24. Mar. 2017. URL: https://medium.com/ceoquest/scaling-the-revenue-engine-chapter-9-pricing-and-packaging-dbbe3153216a#:~:text=Demand%20curves%20should%20be%20estimated,of%20a%20defined%20quantity%20tier)..

[25] Kyle Poyar. *Pricing Insights from 2,200 SaaS Companies*. Accessed: 2024-10-28. Jan. 2021. URL: https://openviewpartners.com/blog/saas-pricing-insights/.

[26] F. D. Merritt. "Review: [Untitled]". In: *Journal of Political Economy* 6.3 (June 1898). Review of *Researches into the Mathematical Principles of the Theory of Wealth* by Augustin Cournot, translated by Nathaniel T. Bacon, pp. 426–430. URL: https://www.jstor.org/stable/1819059.

[27] Daniel Kahneman and Vernon Smith. "Foundations of Behavioral and Experimental Economics". In: (2002). Accessed: 2024-10-28. URL: https://www.nobelprize.org/uploads/2018/06/advanced-economicsciences2002.pdf.

[28] Nicolas Boccard. *Industrial Organization: A Contract Based Approach*. Dec. 2010.

[29] Leonardo Becchetti and Stefano Zamagni. "Non-competitive Markets and Elements of Game Theory". In: *The Microeconomics of Wellbeing and Sustainability* (2020). Accessed: 2024-10-28. URL: https://www.sciencedirect.com/topics/economics-econometrics-and-finance/price-discrimination.

[30] Anita Ramasastry. *Websites That Charge Different Customers Different Prices: Is Their "Price Customization" Illegal? Should It Be?* Accessed: 2024-09-24. June 2005. URL: https://supreme.findlaw.com/legal-commentary/websites-that-charge-different-customers-different-prices.html.

[31] Andriy Burkov. *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019.

[32] Olivier Chapelle and Zaïd Harchaoui. "A Machine Learning Approach to Conjoint Analysis". In: *Advances in Neural Information Processing Systems* 17 (2004). Accessed: 2024-10-28. URL: https://papers.nips.cc/paper_files/paper/2004/hash/4bbdcc0e821637155ac4217bdab70d2e-Abstract.html.

[33] IBM. *What is a data pipeline?* Accessed: 2024-10-28. 2024. URL: https://www.ibm.com/topics/data-pipeline.

[34] Srivignesh Rajan. *Data Preprocessing Pipeline in Machine Learning: A Walkthrough in Python using Kaggle House Price Prediction Data*. Published in The Startup, Accessed: 2024-10-28. July 2020. URL: https://medium.com/swlh/data-preprocessing-and-data-modeling-for-kaggle-house-price-prediction-data-in-python-c04055ded258.

[35] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer Series in Statistics. Springer, 2009.

[36] Luigi Laura. *Breve e Universale Storia degli Algoritmi*. LUISS University Press, 2019.

[37] IBM. *What is a machine learning pipeline?* Accessed: 2024-10-28. 2024. URL: https://www.ibm.com/topics/machine-learning-pipeline.

[38] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[39] Yao-Liang Yu. *An Introduction to Machine Learning*. Lecture notes for CS480/680, School of Computer Science, University of Waterloo. Dec. 2021. URL: mailto:yaoliang.yu@uwaterloo.ca.

[40] Machinelearningplus.com. *ARIMA Model – Complete Guide to Time Series Forecasting in Python*. Accessed: 2024-10-28. 2024. URL: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/.

[41] Lawrence May. *Comparing Holt-Winters Exponential Smoothing and ARIMA Models for Time Series Analysis*. Accessed: 2024-10-28. Oct. 2021. URL: https://medium.com/@lawrence.may/comparing-holt-winters-exponential-smoothing-and-arima-models-for-time-series-analysis-659d6f7738c1.

[42] Dan Jurafsky and James H. Martin. *Speech and Language Processing*. 3rd (draft). August 20, 2024 release. 2024. URL: https://web.stanford.edu/~jurafsky/slp3/.

[43] IBM. *What is Sentiment Analysis?* Accessed: 2024-10-28. 2024. URL: https://www.ibm.com/topics/sentiment-analysis.

[44] Yu-kai Chou. *Actionable Gamification: Beyond Points, Badges, and Leaderboards*. Yu-kai Chou, 2016.

[45] European Parliament. *EU AI Act: First Regulation on Artificial Intelligence*. Published: 08-06-2023, Last updated: 18-06-2024. June 2023. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[46] AppInChina.com. *Launching Software As A Service (SaaS) In China*. Accessed: 2024-10-28. 2024. URL: https://appinchina.co/services/other/saas/#:~:text=Chinese%20law%20states%20that%20data,that%20physically%20reside%20in%20China..