COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

Application of Data Science and Predictive Models for Churn Prevention:

Optimizing Customer Retention

Submitted By:

Marta Vara
Student ID: 202011615

Submitted To:

Prof. Alvaro Otero
Department of Economic and Business Sciences

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of International Relations and Business Analytics

April 10, 2025

DECLARATION

This project report titled " **Application of Data Science and Predictive Models for Optimal Churn Prevention: Understanding Diverse Causes and Proposing Personalized Actions through Customer Segmentation** " is the result of my research, except as cited in the references. I have conducted this project under the supervision of **Alvaro Otero**. This project partially fulfills the requirements for the degree awarded as Bachelor of International Relations and Business Analytics [2024] session at University Pontificia d Comillas, Madrid, Spain.

Submitted By:

Marta Vara Rodríguez
ID: 202011615
Department of Economic and Business Sciences
Universidad Pontificia de Comillas

Supervised By:

Prof. Álvaro Otero
Professor at Universidad Pontificia de Comillas
Department of Madrid Culinary Campus (MACC)
Universidad Pontificia de Comillas

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, **Álvaro Otero**, for his continuous support, valuable guidance, and constant encouragement throughout the development of this thesis. His insights and feedback have been essential to completing this work.

I am deeply thankful to my parents, family, and friends, whose unwavering support and motivation have accompanied me during this academic journey.

Finally, I would like to extend my appreciation to all the professors at **Universidad Pontificia de Comillas** for shaping my academic path and providing me with a solid foundation in *Business Analytics*. Their dedication and the analytical tools they taught me have been crucial to the successful completion of this project.

RESUMEN

En un entorno digital cada vez más competitivo, la pérdida de clientes (*churn*) representa uno de los mayores retos para los negocios basados en suscripción, especialmente dentro del ecosistema SaaS (*Software as a Service*). Lejos de ser una simple preocupación operativa, el churn amenaza directamente la sostenibilidad del negocio al erosionar los ingresos recurrentes mensuales (MRR) y aumentar la dependencia de costosas estrategias de adquisición.

Este Trabajo de Fin de Grado presenta un marco analítico integrado para la predicción y prevención del churn, enmarcado dentro del modelo de **Revenue Operations (RevOps)** — un enfoque estratégico que alinea los equipos de marketing, ventas y atención al cliente a lo largo de todo el ciclo de vida del cliente. RevOps trasciende el enfoque tradicional centrado únicamente en adquisición, poniendo el foco en la retención, expansión y maximización del valor del cliente (LTV). En este contexto, el **Bowtie Funnel** (Figura 1.1) emerge como un marco clave para integrar las métricas pre y post-venta, asegurando que el crecimiento de clientes no se vea afectado por elevadas tasas de churn.

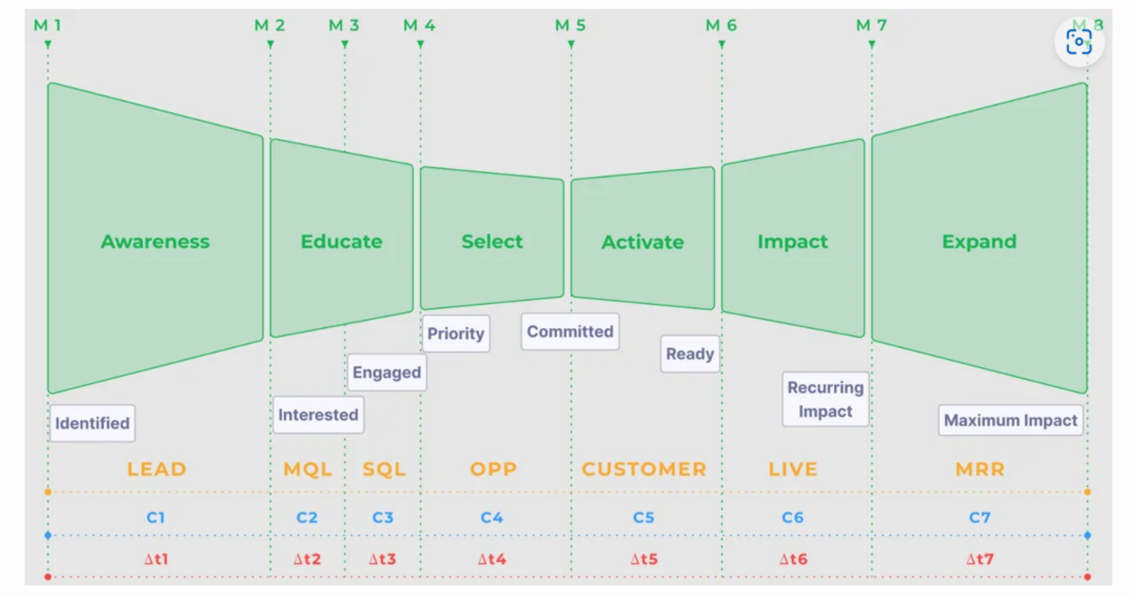


Figure 1: Modelo Bowtie Funnel en RevOps [1]

Asimismo, se presta especial atención a la economía SaaS, caracterizada por una rentabilidad diferida, altos costes de adquisición de clientes (CAC) y una elevada exposición al riesgo de churn. El estudio pone de relieve cómo métricas como MRR, LTV y Retención Neta de Ingresos (NDR) funcionan como barreras financieras en los modelos de negocio SaaS, y cómo el churn erosiona directamente estas métricas, amenazando el crecimiento sostenible.

Para ofrecer una visión clara del proceso de investigación, la Figura 2 ilustra las principales fases desarrolladas en este estudio.

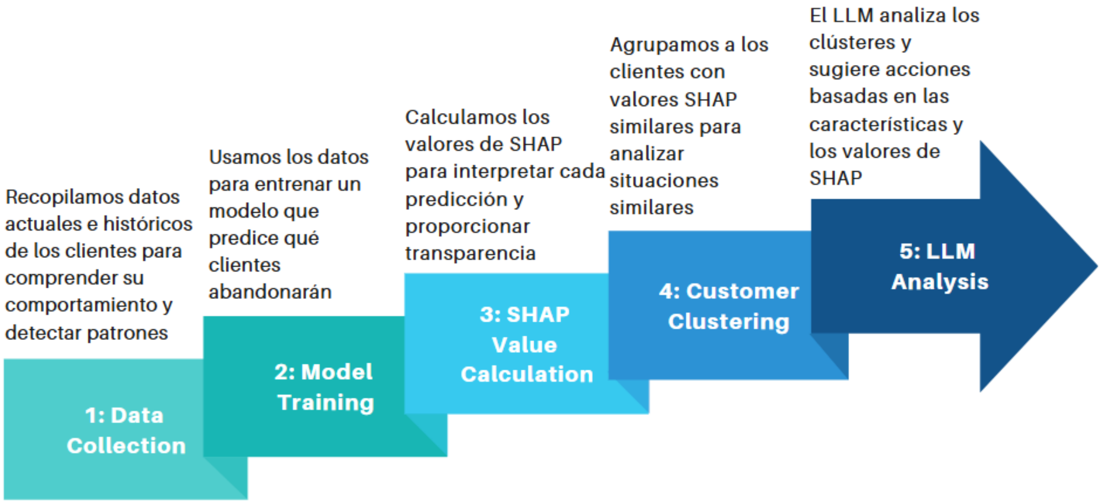

Recopilamos datos actuales e históricos de los clientes para comprender su comportamiento y detectar patrones
1: Data Collection
Usamos los datos para entrenar un modelo que predice qué clientes abandonarán
2: Model Training
Calculamos los valores de SHAP para interpretar cada predicción y proporcionar transparencia
3: SHAP Value Calculation
Agrupamos a los clientes con valores SHAP similares para analizar situaciones similares
4: Customer Clustering
El LLM analiza los clústeres y sugiere acciones basadas en las características y los valores de SHAP
5: LLM Analysis

Figure 2: Flujograma del Proyecto

La metodología se compone de cinco grandes fases:

1. **Recopilación de Datos:** Recolección de transacciones de clientes y registros históricos para analizar patrones de churn.
2. **Entrenamiento del Modelo:** Aplicación de técnicas de aprendizaje supervisado para construir modelos predictivos de churn.
3. **Cálculo de Valores SHAP:** Cálculo de valores SHAP para interpretar la importancia de las variables y la transparencia del modelo.
4. **Segmentación de Clientes:** Aplicación de técnicas de clustering para agrupar clientes en función de similitudes derivadas de SHAP.
5. **Análisis con LLM:** Utilización de un modelo de lenguaje avanzado (LLM) para analizar los clusters y sugerir estrategias de retención.

En primer lugar, tras un exhaustivo Análisis Exploratorio de Datos (EDA), se desarrolló un **modelo de Regresión Logística** que obtuvo los siguientes resultados en la predicción de churn:

- **Precisión Balanceada:** 68.41% ± 0.0020
- **F1 Macro Score:** 60.72%
- **F1 Weighted Score:** 71.44% ± 0.0011
- **Precisión para la Clase Churn (1):** 32.12%
- **Recall para la Clase Churn (1):** 69.19%

Estos resultados muestran que, si bien el modelo captura una proporción razonable de clientes que abandonan (recall = 69.19%), su precisión (32.12%) es baja.

Real / Predicho	0	1	Recall
0	27,000 (55.38%)	12,921 (26.5%)	67.63%
1	2,722 (5.58%)	6,113 (12.54%)	69.19%
Precisión	90.84%	32.12%	67.92%

Table 1: Matriz de Confusión con Precisión y Recall

Posteriormente, se aplicó una técnica de **Clustering** utilizando embeddings de árboles de decisión combinados con **UMAP** para reducción de dimensionalidad. La Figura 6.1 muestra la segmentación de clientes resultante.

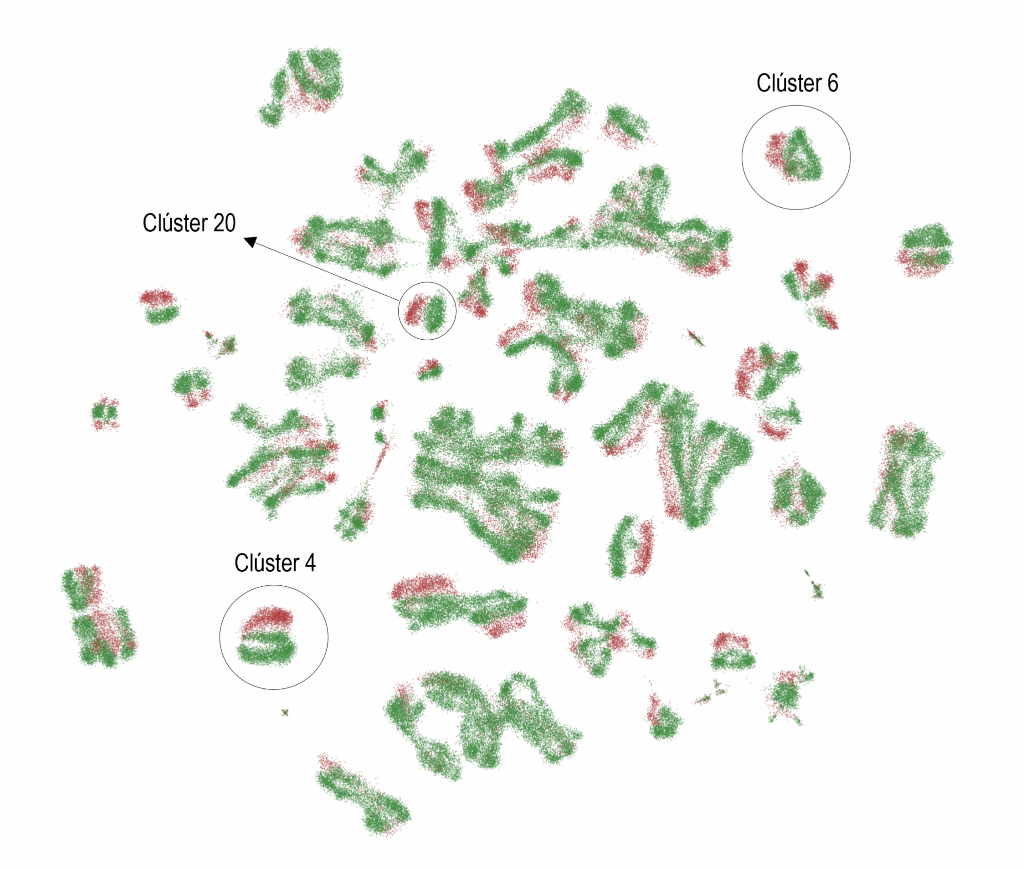


Figure 3: Visualización del Clustering: Distribución de Clientes por Segmentos

Como resultado, surgieron tres clusters especialmente propensos al churn:

Cluster	Arquetipo de Churn	Características Clave
4	Usuarios Premium Pasivos	Alto gasto, baja interacción, fricción en pagos
6	Multidispositivo Bajo Valor	Uso intensivo de dispositivos pero bajo gasto e interacción
20	Casual Soporte-Dependientes	Muchas incidencias de soporte, baja personalización del producto

Table 2: Clusters con Alto Churn: Arquetipos y Características Clave

Finalmente, aprovechando la explicabilidad de los valores **SHAP** y la interpretabilidad de los clusters, se integró un **Modelo de Lenguaje (LLM)** para generar automáticamente estrategias de retención personalizadas por segmento. Algunos ejemplos de estas intervenciones son:

- Automatización de métodos de pago para reducir fricciones (Cluster 4).
- Lanzamiento de campañas de activación temprana (Cluster 6).
- Optimización del soporte técnico mediante herramientas asistidas por IA (Cluster 20).

En definitiva, este proyecto demuestra que integrar la filosofía RevOps con machine learning, clustering y el uso de LLM permite construir un marco escalable y accionable para la mitigación del churn. Este enfoque estratégico transforma datos brutos de clientes en palancas de retención poderosas, incrementando el LTV y garantizando un crecimiento sostenible en negocios basados en suscripción.

Palabras clave: Churn de clientes, RevOps, Bowtie Funnel, SaaS, Predicción de Churn, Clustering, SHAP, LLM, Retención de Clientes, Analítica de Datos.

ABSTRACT

In an increasingly competitive digital environment, customer churn represents one of the most critical challenges for subscription-based businesses, particularly within the SaaS (Software as a Service) ecosystem. Far from being a simple operational concern, churn directly threatens business sustainability by eroding Monthly Recurring Revenue (MRR) and increasing the dependency on costly acquisition strategies.

This Final Degree Project presents an integrated analytical framework for churn prediction and prevention, embedded within the **Revenue Operations (RevOps)** model — a strategic approach that aligns marketing, sales, and customer success teams throughout the entire customer lifecycle. RevOps transcends the traditional focus on acquisition by emphasizing retention, expansion, and customer lifetime value (LTV) maximization. In this context, the **Bowtie Funnel** (Figure 1.1) emerges as a key framework for integrating pre-sale and post-sale metrics, ensuring that customer growth is not undermined by high churn rates.

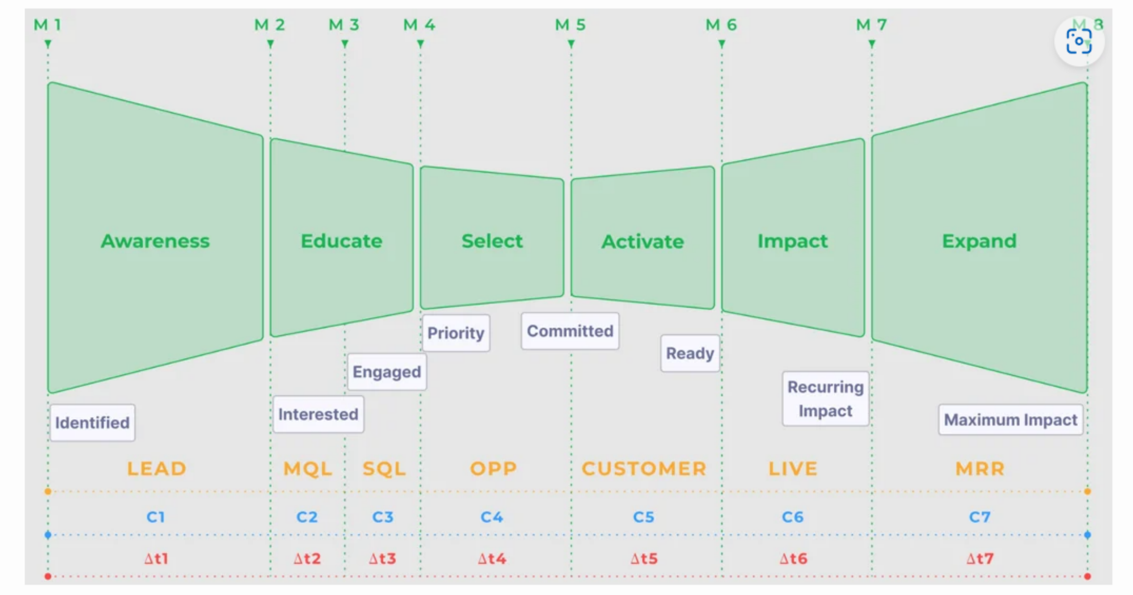


Figure 4: The Bowtie Funnel Model in RevOps [1]

Furthermore, special attention is given to SaaS economics, characterized by delayed profitability, high Customer Acquisition Costs (CAC), and exposure to churn risk. The study highlights how metrics such as Monthly Recurring Revenue (MRR), LTV, and Net Dollar Retention (NDR) operate as financial guardrails within SaaS business models, and how churn directly erodes these metrics, threatening sustainable growth.

To provide a clear overview of the research process, Figure 1.2 illustrates the key steps undertaken in this study.

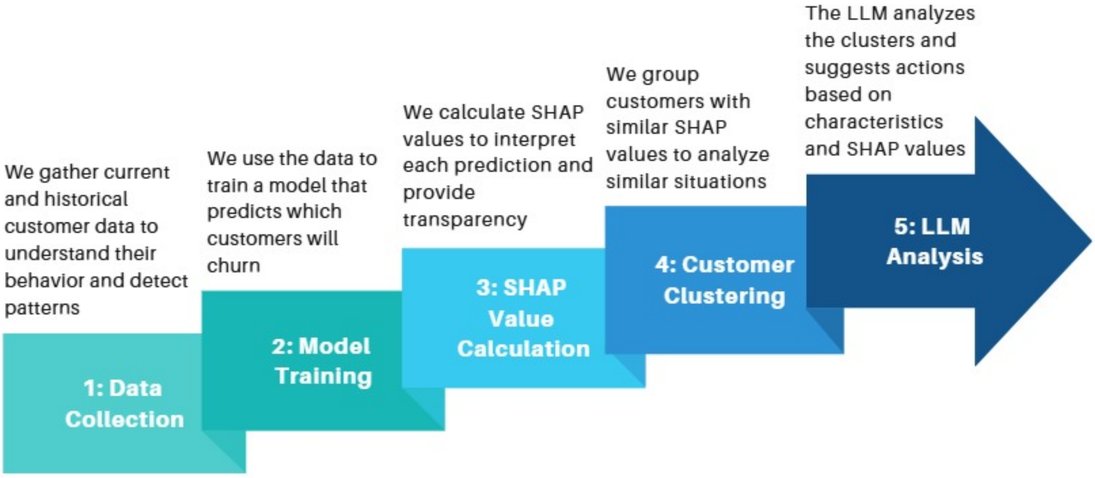

We gather current and historical customer data to understand their behavior and detect patterns
1: Data Collection
We use the data to train a model that predicts which customers will churn
2: Model Training
We calculate SHAP values to interpret each prediction and provide transparency
3: SHAP Value Calculation
We group customers with similar SHAP values to analyze similar situations
4: Customer Clustering
The LLM analyzes the clusters and suggests actions based on characteristics and SHAP values
5: LLM Analysis

Figure 5: Thesis Flow Chart

The methodology consists of five major phases:

1. **Data Collection:** Gathering customer transaction data and historical records to analyze churn patterns.
2. **Model Training:** Implementing supervised learning techniques to build predictive models for churn classification.
3. **SHAP Value Calculation:** Computing SHAP values to interpret feature importance and model transparency.
4. **Customer Clustering:** Applying clustering techniques to group customers based on SHAP-derived similarities.
5. **LLM Analysis:** Utilizing an advanced language model to analyze clusters and suggest retention strategies.

Firstly, after an extensive Exploratory Data Analysis (EDA), a **Logistic Regression model** was developed, achieving the following results in churn prediction:

- **Balanced Accuracy:** 68.41% ± 0.0020
- **F1 Macro Score:** 60.72%
- **F1 Weighted Score:** 71.44% ± 0.0011
- **Precision for Churn Class (1):** 32.12%
- **Recall for Churn Class (1):** 69.19%

These metrics highlight that while the model captures a reasonable proportion of churned customers (recall = 69.19%), its precision (32.12%) remains low.

Actual / Predicted	0	1	Recall
0	27,000 (55.38%)	12,921 (26.5%)	67.63%
1	2,722 (5.58%)	6,113 (12.54%)	69.19%
Precision	90.84%	32.12%	67.92%

Table 3: Confusion Matrix with Precision and Recall

Subsequently, a **Clustering** approach was applied using decision tree embeddings combined with **UMAP** for dimensionality reduction. The following figure (Figure 6.1) illustrates the resulting customer segmentation.

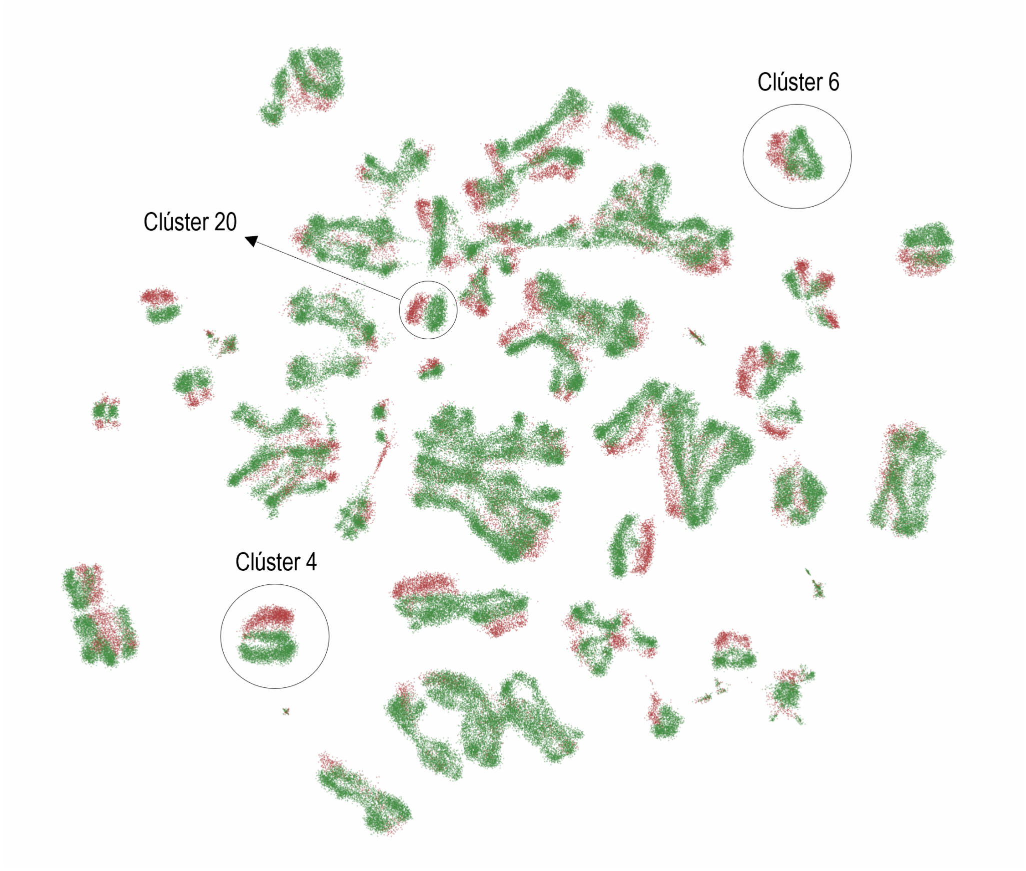


Figure 6: Clustering Visualization: Distribution of Customers by Segments

As a result, three clusters emerged as particularly churn-prone:

Cluster	**Churn Archetype**	**Key Characteristics**
4	Passive Premium Users	High spending, low engagement, payment friction
6	Multi-Device Low Value	High device usage but low spending and interaction
20	Support-Driven Casuals	Frequent support tickets, low product personalization

Table 4: High-Churn Clusters: Archetypes and Key Characteristics

Finally, leveraging the explainability provided by **SHAP** values and the interpretability of clusters, a **Large Language Model (LLM)** was integrated to automatically generate tailored retention strategies per segment. Examples of these interventions include:

- Automating payment methods to reduce friction (Cluster 4).
- Launching early activation campaigns to drive engagement (Cluster 6).
- Optimizing technical support with AI-assisted tools (Cluster 20).

Overall, this project demonstrates that integrating RevOps philosophy with machine learning, clustering, and LLM-driven insights offers a scalable and actionable framework for churn mitigation. This strategic approach transforms raw customer data into powerful retention levers, enhancing customer lifetime value (LTV) and ensuring sustainable growth for subscription-based businesses.

Keywords: Customer churn, RevOps, Bowtie Funnel, SaaS, Churn Prediction, Clustering, SHAP, LLM, Customer Retention, Data Analytics.

Contents

List of Figures

List of Tables

1 Introduction

In today's competitive digital landscape, retaining existing customers has become as critical as acquiring new ones. Customer churn—the phenomenon of users discontinuing their relationship with a company—poses a direct threat to sustainable revenue growth and operational efficiency. As subscription-based models and recurring revenue streams gain prominence, businesses must prioritize long-term engagement, customer satisfaction, and proactive retention strategies.

This thesis explores the strategic importance of churn prediction within the broader context of Revenue Operations (RevOps). By combining predictive analytics, behavioral segmentation, and large language models (LLMs), the thesis aims to uncover patterns of customer disengagement and generate tailored interventions. The goal is not only to identify users likely to churn, but to understand why they leave and how to retain them through actionable, data-driven insights. This cross-functional approach, grounded in real-world data and advanced analytics, reflects the growing need for businesses to make customer-centric decisions at scale.

1.1 Literature Review

The literature on customer retention and churn management has evolved significantly over the past decade. As organizations move from product-centric to customer-centric models, the need to manage the entire customer lifecycle—from acquisition to renewal and expansion—has intensified. This has given rise to integrated frameworks like Revenue Operations (RevOps), which consolidate marketing, sales, and customer success into a unified strategy for revenue optimization.

This section examines key academic and industry contributions related to RevOps, with a specific focus on how it reshapes the approach to churn. We review foundational theories around lifecycle management, the emergence of predictive analytics in customer behavior modeling, and the critical role of data in enabling timely and personalized retention strategies. Special attention is given to the Bowtie Funnel model as a guiding framework, along with the analytical tools and methodologies used to identify and reduce churn in modern organizations.

1.1.1 Understanding Revenue Operations (RevOps)

Revenue Operations (RevOps) has emerged as a strategic response to the fragmentation between go-to-market teams in modern enterprises. Traditionally, marketing, sales, and customer success departments operated in silos, each with separate goals and metrics. RevOps integrates these functions into a single operational model designed to align incentives, streamline processes, and drive predictable revenue growth.

This subsection provides an overview of the RevOps concept, highlighting its relevance in managing the full customer journey—from lead generation to post-sale engagement. We explore how RevOps facilitates collaboration, improves visibility across the funnel, and enables data-driven decision-making at every touchpoint. In the context of this thesis, RevOps serves as the structural backbone through which churn prediction and intervention strategies are implemented, ensuring that insights are actionable across all stages of the customer lifecycle.

1.1.1.1 Definition and Importance of RevOps

Revenue Operations (RevOps) is a strategic approach aimed at aligning an organization's marketing, sales, and customer success teams to optimize revenue growth and operational efficiency. By breaking down traditional silos between these departments, RevOps fosters a unified strategy that ensures seamless collaboration across the entire customer lifecycle, from lead acquisition to retention and expansion [1].

The importance of RevOps lies in its ability to enhance efficiency, improve customer experience, and drive predictable revenue growth. As businesses increasingly focus on customer-centric strategies, RevOps has emerged as a critical framework for achieving sustainable success in highly competitive markets. Organizations that implement RevOps effectively often report improved forecasting, reduced churn, and higher customer lifetime value [4].

1.1.1.2 The Bowtie Funnel Model in RevOps

The Bowtie Funnel model is a central concept within RevOps that provides a comprehensive view of the customer journey, extending beyond the traditional linear sales funnel. Unlike conventional models that primarily emphasize customer acquisition, the Bowtie Funnel integrates both acquisition and retention into a unified framework, highlighting the critical role of customer success and recurring revenue [1].

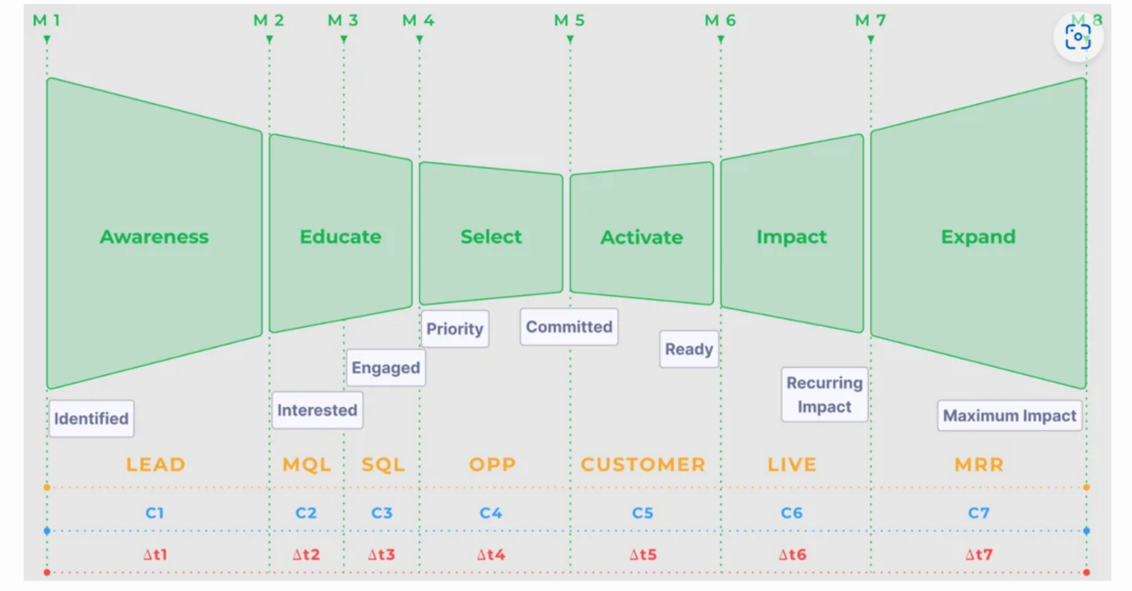


Figure 1.1: The Bowtie Funnel Model in RevOps [1]

As shown in Figure 1.1, the Bowtie Funnel encompasses multiple stages of the customer lifecycle:

- **Awareness to Activation:** Focuses on converting leads into engaged customers through effective marketing and sales strategies.
- **Impact to Expansion:** Emphasizes customer success, recurring revenue, and maximizing customer impact.

This thesis focuses specifically on the **churn prediction** and **actionability insights** within the Bowtie Funnel. By leveraging predictive modeling, we aim to identify customers at risk of churn and provide actionable insights to address the root causes. These actions will be tailored to specific segments and clusters, ensuring targeted interventions that improve customer retention and satisfaction.

Churn Prediction: The predictive model will analyze customer behavior to anticipate churn likelihood at different stages of the funnel. This includes identifying early indicators of disengagement during the transition from *Live* to *MRR* stages.

Actionability Insights: The insights derived from the model will be used to develop segment-specific strategies. For example:

- Addressing dissatisfaction in high-value customers in the *Impact* stage.
- Enhancing engagement for new customers in the *Activation* stage.

By integrating RevOps principles and the Bowtie Funnel, this study provides a structured approach to understanding churn and implementing effective retention strategies.

1.1.2 Data Analysis in Churn Management

As RevOps expands the focus from customer acquisition to the entire customer lifecycle, managing customer churn becomes a critical component of sustained revenue growth. In the Bowtie Funnel model of RevOps, the stages beyond the initial sale (onboarding, value delivery, renewal, expansion, and advocacy) are as important as the pre-sale stages [1]. This represents a strategic shift from an acquisition-centric approach to a retention-centric approach, emphasizing long-term customer value. Low churn is now recognized as being just as significant an indicator of success as high acquisition rates [5]. In fact, churn directly negates growth – “Growth equals acquisition minus churn”, as Gold (2020) succinctly observed. Consequently, companies are increasingly prioritizing data-driven churn management, understanding that reducing customer attrition can translate into proportional gains in growth [5].

Research has long underscored the financial logic of this shift: increasing customer retention rates by as little as 5% can boost profits by 25% to 95% [6]. Moreover, retaining existing customers is generally much less costly than acquiring new ones [7].

Thus, organizations are investing in analytics to identify why customers leave and how to proactively prevent it. In the RevOps context, this means leveraging unified data (from marketing, sales, and customer success touchpoints) to glean insights across the entire “bowtie” customer journey and drive retention-focused strategies.

1.1.2.1 Role of Data in Identifying Churn Patterns

Data plays a crucial role in detecting early warning signs of customer churn. By analyzing behavioral and engagement metrics—such as reduced login frequency, lower product usage, and decreased feature adoption—companies can identify patterns that typically precede cancellations [5]. According to Gold (2020), churn is fundamentally driven by a lack of perceived value, and this becomes evident through usage data: customers who engage more with a product tend to churn significantly less [5].

One of the most effective methods to surface churn patterns is *cohort analysis*, which segments users by tenure or behavior and tracks their retention over time. Gold’s research demonstrates that low-usage cohorts consistently exhibit the highest churn rates, while high-usage cohorts show strong retention [5]. There is often a usage threshold that marks the point where additional engagement leads to reduced churn risk. Identifying this threshold enables companies to define what ”healthy usage” looks like and target low-engagement users with retention efforts. Ultimately, transforming raw engagement data into actionable insights allows firms to intervene proactively and reduce customer attrition.

Beyond usage frequency, data analysis enables companies to uncover deeper churn-related patterns by segmenting users based on demographic, behavioral, or firmographic attributes. For example, small-business clients may churn more than enterprise clients due to budget limitations or business instability, which highlights the need for segment-specific retention strategies [8].

This segmentation approach—often referred to as micro-segmentation—helps businesses design tailored initiatives, such as providing extra support to low-usage users or industry-specific success programs. As Wizdo (2020) notes, applying analytics to identify at-risk groups and engaging in thoughtful, personalized outreach can reduce churn significantly [9].

Importantly, these efforts typically target voluntary churn (customer-driven exits) rather than involuntary churn (due to external events like payment failures), with analytics focusing on early indicators such as declining engagement or satisfaction scores [8].

Overall, the literature stresses that data analysis transforms churn management from a reactive process into a proactive one—allowing RevOps teams to act before the customer disengages entirely. Through methods like customer segmentation and behavioral monitoring, businesses can prioritize retention in the most vulnerable areas of the funnel [5].

1.1.2.2 Tools and Techniques for Data-Driven Churn Management

To operationalize churn insights, companies rely on a combination of performance metrics and advanced analytical tools. Standard indicators such as the *customer churn rate* and *revenue churn rate* provide a foundational view of retention across time and customer segments [8]. However, descriptive statistics are only a starting point.

To move beyond reactive reporting, many organizations adopt **predictive modeling** and **machine learning** techniques. Algorithms like logistic regression, random forests, gradient boosting, and neural networks are trained on historical data to distinguish between churned and retained users. These models generate *propensity scores*, allowing teams to prioritize at-risk customers for timely intervention. In industries like telecommunications, contract type, tenure, and monthly charges have been consistently identified as key churn predictors [10].

These tools are increasingly supported by *explainable AI* techniques such as SHAP values, which provide transparency into model decisions by highlighting the most influential variables per customer. Combined with visual dashboards, they allow RevOps and Customer Success teams to monitor risk in real time and take action accordingly.

Alongside modeling, many companies implement Customer Success Management (CSM) platforms like **Gainsight**, **ChurnZero**, or **Totango**. These platforms integrate CRM data, product usage logs, support tickets, and surveys to calculate *Customer Health Scores*—composite indicators of customer stability. When scores fall below a threshold, these systems can trigger automated workflows such as re-engagement emails, alerts for success agents, or escalation protocols [8].

Together, these capabilities shift churn management from reactive firefighting to a proactive, data-informed retention strategy. They empower RevOps teams to intervene before disengagement becomes irreversible—preserving long-term customer value.

Cohort analysis is another foundational technique. It allows teams to track groups of customers—such as those onboarded in a given quarter—and observe retention trends at 1, 3, or 6-month intervals [1]. This method helps assess the long-term impact of onboarding, product changes, or customer success initiatives.

These insights tie directly into the **Bowtie Funnel model**, which emphasizes optimizing the full customer lifecycle—not just acquisition. Metrics like time-to-first-value, product activation rate, and renewal rate become key performance indicators across the post-sale journey [1].

However, analytics alone is insufficient. Effective RevOps teams operationalize insights through structured action plans or “playbooks” that automate intervention when risk conditions are met. For instance, a drop in feature usage or a surge in support tickets might trigger targeted workflows—such as sending personalized tutorials or routing the issue to senior support agents [1].

Statti (2021) highlights that high-performing RevOps teams combine health scores, engagement monitoring, and proactive playbooks to address churn risk before it materializes [1]. CSM tools are increasingly capable of executing these plays—sending alerts, triggering outreach campaigns, or assigning follow-ups—ensuring that no churn signal is missed.

In this context, CSM platforms act as operational backbones. They consolidate data, automate decision rules, and standardize retention processes based on real-time customer health [8]. This makes churn prevention scalable and systematic.

Beyond individual interventions, churn analytics also fosters cross-departmental alignment. Marketing gains clarity on expectation mismatches, product teams identify features that correlate with long-term engagement, and sales teams learn which customer profiles offer the highest lifetime value [1]. In this way, churn insights evolve into a shared strategic asset.

Ultimately, the literature agrees that reducing churn is not simply a defensive strategy but a key growth lever. Studies show that improving retention yields disproportionately high returns and is significantly more cost-effective than acquiring new customers [5, 9]. In this sense, data-driven churn management becomes central to the RevOps mandate—sealing the right side of the Bowtie Funnel with the same strategic intensity applied to acquisition.

1.2 Methodology

This study employs a structured methodological approach to predict customer churn through machine learning and later segment customers using clustering techniques. The methodology is divided into several key phases: data preprocessing and exploratory analysis, supervised learning for churn prediction, and unsupervised learning for customer segmentation.

To achieve the study’s objectives, a quantitative approach is adopted, leveraging data analytics techniques to process large datasets and derive meaningful insights through statistical analysis and predictive modeling. This approach is particularly suited for identifying behavioral patterns that influence churn, enabling data-driven decision-making. The study utilizes two primary analytical tools: **Python** and **Graphext**. Python is used for data exploration, cleaning, and preliminary analysis, offering robust capabilities through libraries such as `pandas`, `NumPy`, and `scikit-learn` for feature engineering and predictive modeling. Graphext facilitates model creation, iteration, and deployment, incorporating advanced visualization tools and clustering algorithms that enhance interpretability and interactive exploration. The integration of these tools ensures a comprehensive analytical framework, ultimately improving model accuracy and practical applicability. **Additionally, Graphext was used to create variable groups and generate tags, supporting a more structured analysis and simplifying the identification of customer segments.**

The first step involves conducting an **Exploratory Data Analysis (EDA)** to gain a detailed understanding of the dataset. Given that the data originates from transactional databases similar to those used by large enterprises such as for exmaple: Netflix, EDA is essential for identifying missing values, outliers, and inconsistencies. Descriptive statistics, visualizations, and correlation matrices help evaluate data distributions and feature interactions. Both univariate and multivariate analyses allow us to detect significant patterns that may impact model performance.

A rigorous **data preprocessing phase** follows to ensure accuracy and consistency. Missing values are addressed using statistical imputation techniques, while outliers are treated based on Interquartile Range (IQR) and Z-score thresholds. Continuous variables undergo normalization or standardization to maintain uniformity, and categorical variables are transformed using one-hot or ordinal encoding for compatibility with machine learning models.

To enhance predictive performance, **feature engineering** is applied, converting raw data into meaningful attributes that improve interpretability. Key transformations include the segmentation of numerical features, such as grouping customers into "New," "Standard," and "Loyal" categories based on tenure, and defining spending patterns as "Low," "Medium," or "High" based on quartiles. Additionally, derived features, such as the ratio of support tickets to tenure, help capture dissatisfaction trends that may correlate with churn. One of some techniques that could be applied to reduce dimensionality and prevent overfitting, is **Principal Component Analysis (PCA)**, which is are employed to eliminate redundant information while preserving predictive power. However, in this specific case PCA will not be useful for our analysis, as there are no highly correlated variables.

The study applies **supervised learning** to predict churn, formulated as a **binary classification problem** where customers are categorized as either churned ($churn = 1$) or retained ($churn = 0$). Several machine learning models are evaluated, considering their ability to handle class imbalance, interpretability, and predictive accuracy. Logistic regression provides a transparent baseline model, while ensemble methods like `ExtraTreesClassifier` and `RandomForestClassifier` improve classification stability. The study also explores `CatBoostClassifier`, a gradient boosting algorithm optimized for categorical data, ensuring robust performance. Model evaluation is based on **accuracy, balanced accuracy, recall, and AUC-ROC** to assess predictive effectiveness and minimize false negatives.

Following churn prediction, **unsupervised learning techniques** are implemented for customer segmentation. Rather than grouping customers solely based on raw features, the clustering process integrates **SHAP (SHapley Additive exPlanations) values**, which offer a transparent interpretation of model decisions. This approach ensures that clusters are formed based on meaningful drivers of churn, rather than arbitrary feature similarities. By leveraging SHAP, the study aims to uncover distinct customer segments with similar churn risk factors, allowing for more targeted retention strategies.

To validate model performance, **Stratified K-Fold Cross-Validation** is applied, ensuring that training and testing sets maintain class balance. This method prevents overfitting and enhances the model's generalizability. Performance is assessed using multiple metrics, including accuracy, balanced accuracy, recall, and AUC-ROC. These measures provide a comprehensive evaluation of the model's ability to identify churners while minimizing misclassification errors.

The final stage involves integrating clustering results with a **Large Language Model (LLM)** to analyze customer segments and generate actionable insights. This process enables the design of personalized retention strategies, optimizing customer engagement based on data-driven recommendations. Future work will explore hyperparameter tuning and feature selection to refine model performance further, ensuring continuous improvement in predictive accuracy and segmentation effectiveness.

1.3 Objectives

The primary objective of this thesis is to analyze customer churn from a strategic perspective, highlighting its profound impact on business sustainability and long-term profitability. Rather than focusing solely on user acquisition and community growth, this work emphasizes the importance of customer retention as a cost-effective and value-generating pillar for digital businesses.

Specifically, this thesis aims to:

- **Quantify the impact of churn on business performance**, including revenue loss, customer lifetime value (CLV) reduction, and increased acquisition costs.

- **Identify the key behavioral, transactional, and engagement patterns that lead to churn**, using advanced segmentation and predictive modeling techniques.

- **Evaluate and visualize churn probability distributions**, in order to anticipate which customer profiles are most likely to churn.

- **Develop customer typologies** that go beyond a binary churn vs. non-churn comparison by identifying and characterizing customer segments—particularly within churn-heavy clusters—to uncover distinct patterns of usage, satisfaction, payment behavior, and content preferences.

- **Propose tailored retention strategies** for different customer segments, focusing on reducing friction, improving satisfaction, and increasing engagement through proactive intervention. These strategies will be *rooted in the specific churn drivers identified within each cluster*, ensuring that each action is tailored to the underlying causes of disengagement rather than applying generic retention tactics.

Ultimately, this analysis seeks to shift the strategic conversation away from constant expansion and toward intelligent customer lifecycle management. By understanding not just who churns, but *why*, businesses can implement data-driven initiatives that strengthen user retention, improve customer experience, and ensure long-term growth from within.

1.4 Thesis flowchart

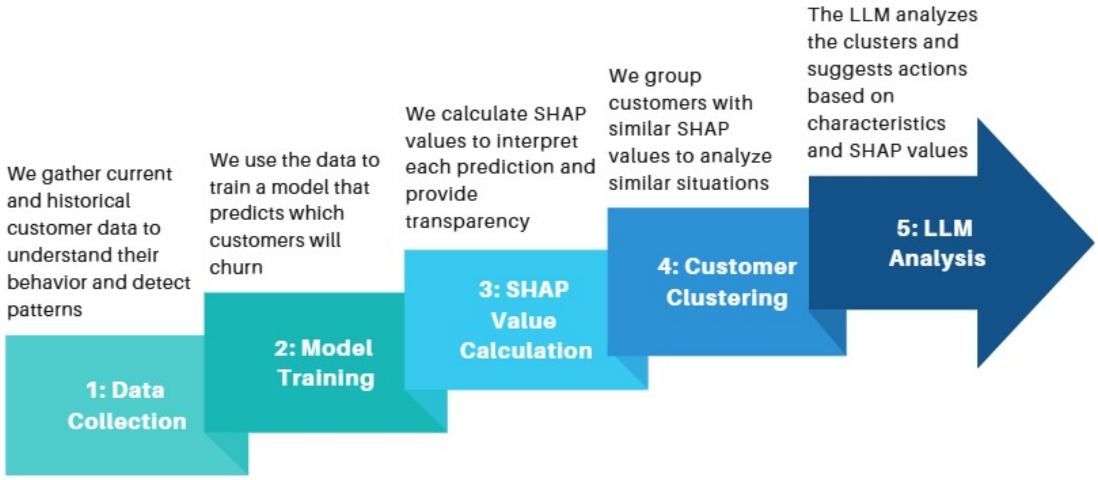

We gather current and historical customer data to understand their behavior and detect patterns
1: Data Collection
We use the data to train a model that predicts which customers will churn
2: Model Training
We calculate SHAP values to interpret each prediction and provide transparency
3: SHAP Value Calculation
We group customers with similar SHAP values to analyze similar situations
4: Customer Clustering
The LLM analyzes the clusters and suggests actions based on characteristics and SHAP values
5: LLM Analysis

Figure 1.2: Thesis Flow Chart

To provide a clear overview of the research process, Figure 1.2 illustrates the key steps undertaken in this study. The methodology consists of five major phases:

1. **Data Collection:** Gathering customer transaction data and historical records to analyze churn patterns.

2. **Model Training:** Implementing supervised learning techniques to build predictive models for churn classification.

3. **SHAP Value Calculation:** Computing SHAP values to interpret feature importance and model transparency.

4. **Customer Clustering:** Applying clustering techniques to group customers based on SHAP-derived similarities.

5. **LLM Analysis:** Utilizing an advanced language model to analyze clusters and suggest retention strategies.

2 Concept of Churn

Understanding customer churn—often referred to simply as "churn"—is crucial for businesses striving for sustainability and growth. This chapter delves into the multifaceted concept of churn, exploring its definitions, types, and the strategic responses that businesses can deploy to mitigate its impacts. As businesses increasingly recognize the significant cost advantages of retaining existing customers over acquiring new ones, the importance of effectively managing churn has come to the forefront of strategic business planning.

Churn represents a critical metric for gauging customer retention and loyalty, reflecting the rate at which customers discontinue their association with a service or product. Beyond its basic definition, churn can manifest in various forms, each presenting unique challenges and requiring tailored strategies for effective management. This chapter will examine both customer and employee churn, providing insights into their distinct dynamics and the underlying reasons for their occurrence.

By integrating theoretical knowledge with practical applications, this chapter aims to equip readers with a comprehensive understanding of churn. It will discuss the direct and indirect impacts of churn on business performance, emphasizing the need for robust predictive tools and strategies that not only identify potential churn but also address the root causes to enhance customer engagement and retention. Through this detailed exploration, we aim to underscore the strategic and financial importance of minimizing churn and fostering lasting customer relationships.

2.1 Definition of Churn

Historically, businesses have focused on acquiring new clients, often overlooking the importance of retaining existing ones. However, as the value of loyal customers has become clearer, particularly due to their greater profitability and opportunities for cross-selling [11], the concept of customer retention, or "churn," has gained significant attention. Over the past few decades, this shift has led to the development of new strategies aimed at reducing churn and enhancing customer loyalty. The emphasis has moved from simply expanding the customer base to also ensuring that current clients remain engaged and satisfied.

In his book “Fighting Churn with Data,” Carl Gold defines churn as the event when a customer ceases to use a service or cancels a subscription [5]. However, the concept of churn extends far beyond just a customer leaving a service provider. It encompasses a broader range of scenarios, including customers discontinuing their relationship with a business, employees resigning from their jobs, members withdrawing from organizations, or the termination of any ongoing subscription or engagement. Churn, therefore, is a critical indicator of attrition in various contexts, reflecting a break in the continuity of relationships, whether with customers, employees, or members [12].

Patrick Campbell describes churn as the "silent killer" of businesses, underscoring its detrimental impact on growth and sustainability [2]. Companies that fail to address churn early in their operations often find themselves expending significant resources just to maintain their current position, let alone achieve growth [2]. This highlights the importance of early detection and intervention strategies to combat churn effectively and maintain a competitive edge in the market.

Competitive businesses are constantly seeking ways to increase revenue and strengthen customer relationships, which has driven the adoption of new methods and technologies

[10]. This shift from an acquisition-focused strategy to one centered on retention marks a turning point in how companies approach growth. It reflects the growing recognition that retaining existing customers is more cost-effective than acquiring new ones.

As a result, businesses are increasingly focused on improving customer churn prediction and developing retention strategies, particularly targeting customers who offer a higher return on investment [10]. This new perspective not only aims to prevent customer loss but also to maximize the lifetime value of each customer, acknowledging that it is more efficient and effective to keep customers satisfied than to constantly seek replacements for those who leave.

This emphasis on churn prevention has driven the development of advanced predictive analytics and loyalty strategies, fundamentally transforming how companies interact with customers and manage their operations.

2.1.1 Types of Churn

There are two main types of churn: customer churn and employee churn, each of which impacts companies differently. Each type requires a unique approach for effective management, as well as different strategies for prevention and retention [12]. In this thesis, we will focus on customer churn.

Customer churn is the opposite of customer retention. It refers to the proportion of customers who discontinue their subscriptions or services with a company during a specific time frame [13]. This is critical for businesses like SaaS providers, which rely heavily on recurring revenue, as we will discuss later in this thesis. Companies must identify why customers leave and apply tactics to reduce churn over time [12].

When talking about customer churn, we can divide it into two main categories: *intentional* and *involuntary* churn.

1) Intentional churn occurs when a customer voluntarily stops using a service because it no longer meets their needs. The customer may switch to another product offering better financial plans, improved service quality, or both [10]. In these cases, companies can retain customers by offering attractive plans or ensuring high-quality service. However, each case must be analyzed individually, as theoretical strategies may not always align with practical outcomes.

If we look in more detail, there are two subcategories within intentional churn: *deliberate* and *incidental* churn.

- **Deliberate churn** happens when customers leave for a competitor due to dissatisfaction with the current product or service. This dissatisfaction may arise from technical issues such as low service quality, outdated offerings, negative customer service experiences, inadequate coverage, or financial reasons such as high costs or uncompetitive pricing [10].
- **Incidental churn** occurs when customers discontinue the service due to external or situational changes rather than dissatisfaction. This includes scenarios such as no longer needing the service, moving to a location where it is unavailable, changing jobs that reduce relevance, or experiencing financial constraints that make it unaffordable [10].

2) Involuntary churn happens when the business discontinues providing service due to unpaid bills or breach of terms and conditions.

The main objective of churn prediction is to foresee intentional churn, as involuntary churn—such as when customers breach contracts or fail to pay— is already recognized by the business. Incidental churn, which includes unforeseen life changes like moving or changing jobs, represents a small fraction and is difficult to predict. Additionally, anticipating incidental churn is often not useful since retaining such customers involves factors beyond the company's control and unrelated to their offerings [10]. Due to the diverse nature of customer expectations and needs, building loyalty with every customer is not feasible. Dissatisfaction typically arises from multiple issues rather than a single problem, leading to inevitable churn. Consequently, businesses concentrate on retention strategies by improving their products and implementing barriers to prevent customers from switching [10]. Churn management focuses on retaining valuable customers and boosting loyalty, though long-term customers are not always inherently loyal. True loyalty means sticking with a provider despite more attractive offers from competitors. Companies need to actively nurture this loyalty, as loyal customers are often the most profitable [11].

2.1.2 Predictive Challenges of Churn

While churn prediction is a widely used technique in subscription-based businesses, relying solely on predictive models does not guarantee effective retention. To achieve a long-term and reliable reduction in churn, it is essential to go beyond identifying who might leave — companies must understand **why** users churn and what actions can be taken to prevent it [5].

Predicting churn is not always effective for two main reasons:

- **First**, predicting the risk of churn alone does not significantly assist in implementing effective interventions to reduce it. This is because churn-reduction efforts are rarely a one-size-fits-all solution and need to be tailored based on factors beyond just the likelihood of churn [5].
    - While churn-risk prediction can be a valuable metric for identifying customers for one-on-one interventions by Customer Success teams, it is still just one of many variables that should be considered [5].
    - This may seem discouraging, but simply deploying an AI system that excels in churn prediction is not enough. If a churn prediction model is delivered without providing actionable insights, the business will struggle to utilize it effectively [5].
- **Second**, churn is inherently difficult to predict with high accuracy, even with advanced machine learning models. This is largely due to external factors influencing the timing of churn, as well as the subjective nature of customer utility and satisfaction [5].
    - Factors such as subjectivity, incomplete data, and external influences all contribute to the complexity of churn prediction, making it a challenging task to achieve high precision [5].

Churn, by its nature, presents a significant challenge for any scalable SaaS business. It acts as a critical barrier to sustained success and growth, emphasizing the necessity of minimizing churn rates wherever possible. As such, businesses must focus on not only predicting churn but also developing comprehensive strategies that directly address its underlying causes and reduce its impact [2].

2.2 The Strategic and Financial Impact of Customer Churn

Customer churn presents a substantial challenge to businesses, especially in the B2B sector, where losing even a single high-value customer can have significant financial repercussions. Churn not only reduces revenue but also increases the cost of acquiring new customers, making retention strategies essential. Research shows that retaining customers is more cost-effective than acquiring new ones, as loyal customers often stay with a company longer and are less susceptible to switching to competitors [14]. Studies have demonstrated that acquiring new customers can be five to twenty-five times more expensive than retaining current ones. Additionally, a mere 5% increase in customer retention can boost profits by up to 95%, depending on the industry [15]. This underscores the necessity of prioritizing customer retention to maximize profitability.

However, even seemingly ”good” churn rates can pose significant long-term challenges. For example, a net churn rate below 5% might initially appear manageable for a SaaS company, as it implies retaining 95% of customers annually. In reality, a 5% churn rate means losing half of the customer base each year. This requires working four times harder—twice as hard to replace lost customers and additional effort to grow the customer base and Monthly Recurring Revenue (MRR) [2]. As illustrated in Table 2.1, this example demonstrates how the compounding effects of churn make addressing it early imperative for sustainable growth:

Churn Rate	**Original Customers**	**Year-End Customers**
5%	100	54
4%	100	61
3%	100	69
2%	100	78
1%	100	89
0.5%	100	94

Table 2.1: Year-End Customers by Churn Rate [2]

To better understand how churn affects different types of businesses, it is essential to contextualize what constitutes a **standard** churn rate by industry. Churn dynamics vary based on the nature of the product, customer expectations, and contract structures. Table 2.2 summarizes churn rate benchmarks across several major verticals:

Industry / Segment	**Typical Churn Rate**
Telecom (Telecommunications) – e.g. mobile/cable providers	˜1–2% per month
E-commerce Subscriptions – subscription boxes & services	˜5–15% per month
Digital Media & Streaming – online media, OTT video/music	˜5–8% per month
SaaS (B2B) – business software subscriptions	˜3–5% per month
SaaS (B2C) – consumer software subscriptions	˜4–6% per month

Table 2.2: Average churn rate benchmarks by industry. B2B = business-to-business, B2C = business-to-consumer.

2.2.1 Industry Breakdown

Telecommunications (Telco) — Lowest Churn
Telecom services typically experience the lowest churn rates among subscription businesses, averaging 1–2% per month [16]. This reflects approximately 10–20% annually, although top-tier providers often report postpaid churn under 1% monthly in mature markets [17]. Switching barriers — such as bundled services, contract penalties, and technical inconvenience — contribute to high retention [16]. Telcos were also early adopters of churn prediction systems and loyalty programs. However, deregulated or highly competitive markets may see increased churn, with some cases reaching 31% annually during market disruptions [17].

E-commerce Subscription Services — High Churn
E-commerce subscriptions show the highest volatility, with churn rates ranging between 5–15% per month [16]. Curated boxes (e.g., fashion, food, beauty) can exceed 10–12% due to novelty fatigue and discretionary spending patterns [18]. In contrast, utilitarian programs like *"subscribe and save"* services retain customers better, averaging ∼5% churn [18]. Businesses mitigate churn by offering personalized experiences, skip/pause options, and loyalty incentives to maintain engagement [16].

Digital Media & Streaming — Content-Driven Churn
Streaming services (Netflix, Disney+, Spotify) experience moderate churn, around 5–8% monthly [16]. A 2022 study found that 37% of OTT users in the U.S. churned within six months [18]. Churn is driven by content cycles — users join to watch specific shows, then leave once interest wanes. Younger consumers churn more frequently, attracted by promotions or exclusive content [18]. High-performing platforms reduce churn to ∼2% through large content libraries, personalized recommendations, and bundling offers [16].

B2B SaaS — Business-Oriented Software
Business-facing software products exhibit relatively low churn, averaging 3–5% monthly when including both voluntary and involuntary churn [19]. Voluntary churn alone may stay near 3–4%, with best-in-class enterprise products reaching annual churn rates under 10% [17]. Longer sales cycles, mission-critical integrations, and annual contracts improve stickiness and reduce cancellations.

B2C SaaS — Consumer Software
Consumer-oriented SaaS offerings face churn between 4–6% per month due to more impulsive purchases, freemium pricing, and monthly billing structures [20]. Payment failures also add to involuntary churn. Successful B2C firms combat this by deploying frequent product updates, personalization, and community engagement to retain users [16, 18].

To sum up, churn benchmarks vary significantly across industries. Telecom has the lowest churn due to essential services and switching barriers, while e-commerce and digital media see high volatility due to discretionary value and episodic engagement. B2B SaaS outperforms B2C in retention thanks to contractual depth and product stickiness. Understanding these benchmarks is critical for tailoring churn reduction strategies to each industry's operational realities.

2.2.2 Churn Management and Data Utilization

In today's digital landscape, customers have easy access to alternatives, making churn management increasingly important. Companies are leveraging data analytics to predict and mitigate churn by identifying at-risk customers and implementing personalized retention strategies. While data mining techniques have shown potential for churn prediction, their use in B2B contexts remains underutilized [21, 22]. Research by Neslin et al. (2006) and Lemmens & Gupta (2013) illustrates that targeted retention efforts based on customer data can significantly enhance profitability [15, 23]. By personalizing customer interactions, businesses can reduce the risk of losing clients in competitive markets where switching providers is easier than ever [11].

A retention cohort curve (Figure 2.1) is one of the most effective tools for visualizing churn and retention. Even a churn rate of 3% results in losing over a quarter of customers within a year, while a 1% churn rate still leads to over 10% customer loss annually [2].

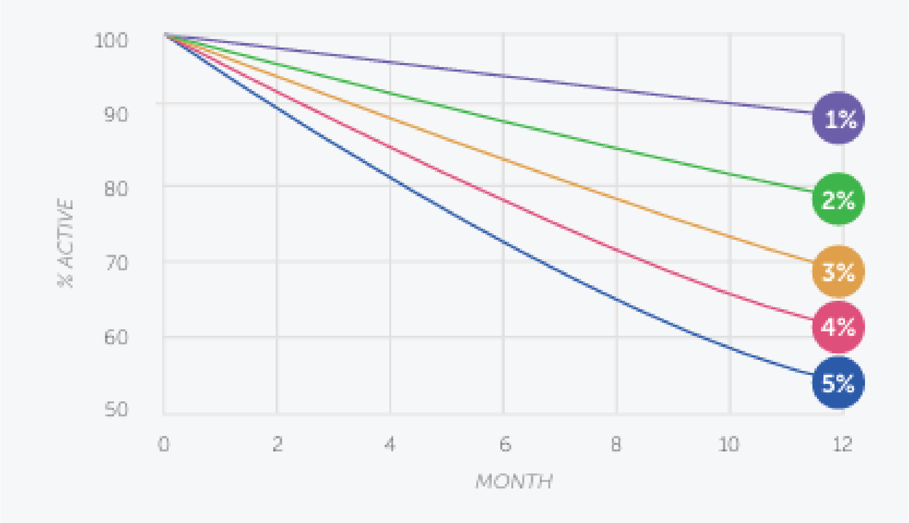


Figure 2.1: Retention Curve by Churn Rate [2]

2.2.3 Factors Contributing to Churn

Churn is driven by a combination of factors that extend beyond product quality and service performance. In B2B settings, where transaction values are typically higher, the importance of trust-based relationships becomes even more pronounced. However, product quality alone is no longer sufficient to guarantee customer loyalty, especially in markets where technology is becoming increasingly commoditized [11]. To reduce churn, businesses must adopt a customer-centric approach that emphasizes personalized engagement, value-driven offers, and proactive retention strategies [11].

A crucial nuance in understanding churn lies in differentiating between "good churn" and "bad churn," as illustrated in Figure 2.2. "Good churn" refers to customers who were never a good fit for the product or service to begin with. These leads—often attracted by broad or misaligned marketing strategies—are not representative of the Ideal Customer Profile (ICP), and their departure should be seen as a natural, even desirable outcome**. Retaining them would only increase costs and dilute product focus.

However, if a company consistently attracts the wrong type of customer, it can accumulate high levels of good churn. **This pattern acts as a strategic signal to marketing and customer acquisition teams**: the targeting and messaging need refinement to better align with the ICP. In other words, persistent good churn is not just a retention issue—it's an acquisition issue.

On the other hand, "bad churn" occurs when customers that do fit the ICP decide to leave, typically due to dissatisfaction, lack of engagement, or unfulfilled expectations. This form of churn is more damaging and must be actively minimized through improved onboarding, value delivery, and relationship management [2].

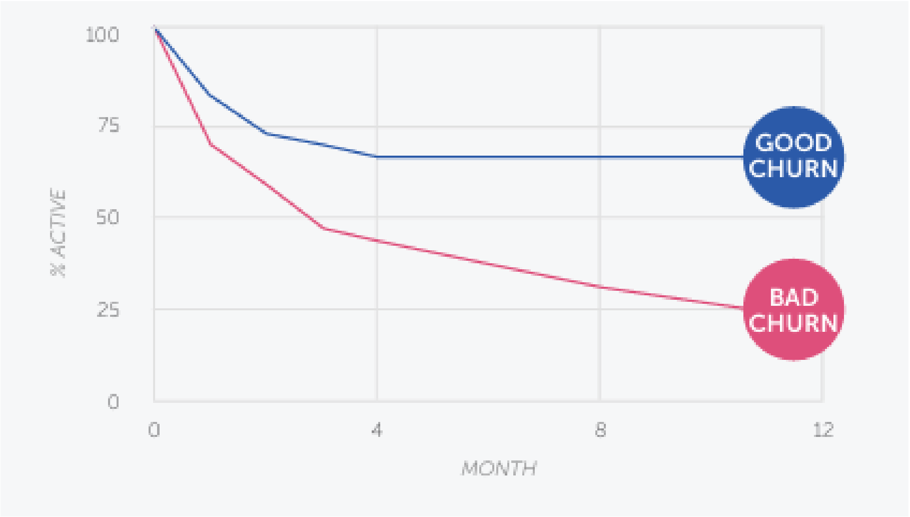


Figure 2.2: Good Churn vs. Bad Churn [2]

2.2.4 Interventions that Reduce Churn

According to Carl Gold in his book *Fighting Churn with Data*, companies typically implement **five key strategies** to minimize churn [5]:

- **Product enhancement**—Product managers, engineers (for software), or content creators (for media) play a crucial role in reducing churn by improving product features or content. This includes introducing new features or repackaging existing ones to enhance the value and enjoyment customers derive from the product. Product improvement is the most direct and primary method of reducing churn.

- **Engagement campaigns**—Marketers use mass communications to engage customers, directing them toward the most popular content or features, which helps in retaining subscribers.
- **Personalized customer interactions**—Customer success teams and support representatives work one-on-one with customers to ensure product adoption and provide assistance when needed. They often reach out proactively to customers who may require help before they request it. These teams are also responsible for onboarding new customers to ensure they take full advantage of the product.
- **Adjusting pricing strategies**—Sales teams can play a role in preventing churn, particularly for paid services, by offering pricing adjustments or modifying subscription terms. This process may involve providing customers with less expensive subscription options as a way to retain them.
- **Targeting new acquisitions**—Rather than focusing solely on retaining existing customers, companies may seek out new customers who are more likely to remain loyal. This strategy is less direct in addressing churn and is limited by the availability of new customers from preferred acquisition channels.

These interventions and service modifications are essential to achieving lower churn rates and longer customer retention. It is important to note that the effectiveness of these strategies often depends on the specific type of subscription service being offered [5].

3 Types of Services and Metrics

This chapter explores the various types of services found in the subscription economy and the key performance metrics associated with them. It begins by distinguishing between different subscription models—primarily Software as a Service (SaaS), media subscriptions, and other recurring offerings such as membership programs and subscription boxes. It also examines non-subscription models, including transactional services, in-app purchases, and ad-supported platforms, highlighting their respective revenue mechanisms. Understanding the operational and financial structure of each model is essential for identifying how churn manifests in different contexts. The chapter concludes with an in-depth analysis of the core metrics used to evaluate customer value and retention, such as Lifetime Value (LTV), Lifetime Revenue (LTR), and Monthly Recurring Revenue (MRR) churn. These metrics form the analytical backbone for understanding business sustainability and for developing effective churn mitigation strategies across diverse service models.

3.1 Subscription Models

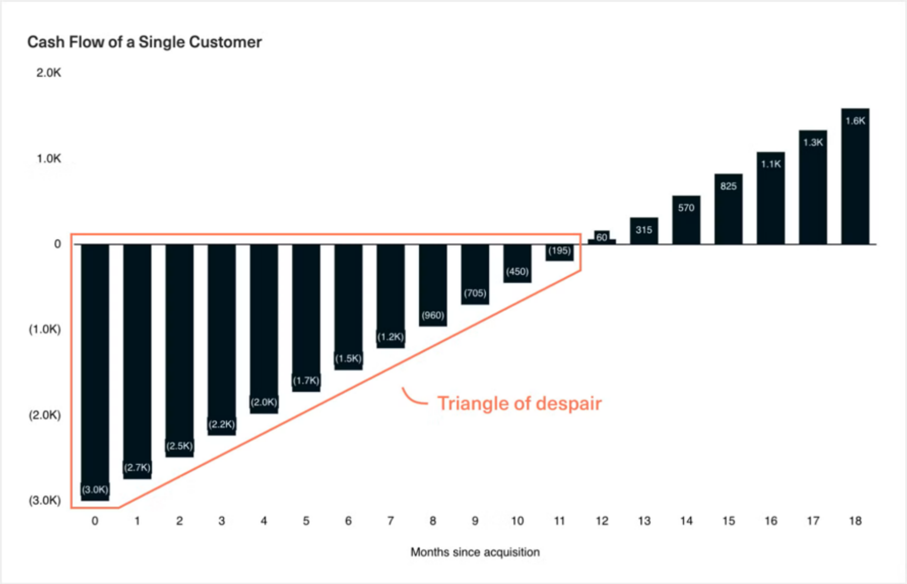
This section delves into the subscription-based business models that dominate the digital economy, with a particular focus on Software as a Service (SaaS). It differentiates between B2B and B2C models, highlighting how each handles acquisition, retention, and monetization strategies. The discussion includes freemium approaches, which combine free and paid tiers to attract and convert users over time. By examining the financial implications and operational challenges of recurring revenue structures—such as delayed profitability and churn exposure—this section lays the groundwork for understanding how subscription models function and scale.

3.1.1 SaaS (Software as a Service)

Software as a Service (SaaS) companies generate revenue primarily by charging recurring subscription fees for access to their software platforms [24]. Unlike traditional software models, which rely on one-time purchases, SaaS businesses operate on a subscription-based revenue model, typically charging customers monthly or annually. The structure of this business model fundamentally alters the path to profitability, as companies often incur significant upfront costs to acquire customers. These expenses include marketing, sales salaries, and customer onboarding, which can take several months to offset. For SaaS companies, it usually takes up to twelve months to break even on a single customer due to these recurring fees being the only revenue source [24].

This business model carries significant risks due to customer churn. Churn rates between 1% and 5% are considered typical for SaaS companies, though in some cases, churn can exceed 15%, which can drastically affect revenue streams [25]. This high churn rate highlights the importance of retaining customers over the long term, as the recurring nature of payments means that the loss of customers results in lost future revenues [3].

One concept often used to describe the economic struggle of a growing SaaS company is the “Triangle of Despair.” This refers to the financial period when a company is unprofitable due to high acquisition costs and insufficient recurring revenue to cover those costs. Essentially, SaaS businesses remain unprofitable until the cumulative payments from each customer exceed the initial acquisition expenses [3]. This dynamic is exacerbated by rapid growth, as faster customer acquisition leads to accumulating acquisition costs, which further stresses cash flow [3].



The Triangle of Despair represents the time it takes to break even on the upfront cost of acquiring a customer.

Figure 3.1: The Triangle of Despair: A SaaS company's financial struggle before profitability [3].

Despite these challenges, the SaaS model offers potential for substantial long-term profitability. Once the acquisition costs are recovered, customer retention becomes highly profitable. By converting initial sales into long-term customer relationships, SaaS companies can enjoy recurring revenue streams that provide a stable financial base for reinvestment and growth [3]. The ability to continuously update products ensures they remain relevant, further encouraging long-term customer retention [3].

3.1.1.1 B2B SaaS

Subscription services tailored for businesses, known as business-to-business (B2B) products, constitute a significant segment of the market. The advent of Salesforce, which launched the first cloud-based customer relationship management (CRM) system in the 2000s, marked the beginning of a substantial expansion in this sector. Today, nearly all new software solutions for businesses are offered as Software as a Service (SaaS). Additionally, many existing on-premises software products are transitioning to cloud-based models, driven by the efficiency of cloud deployment and updates [5].

B2B subscriptions encompass a wide array of categories, including CRM systems, enterprise resource planning (ERP) tools, subscription business management (SBM), human resource management (HR), and support issue tracking systems, among others. These services are designed to enhance operational efficiency and foster collaboration between businesses. Unlike business-to-consumer (B2C) transactions, B2B services focus on meeting the specific needs of other businesses, including offerings like software solutions and consulting services [14].

In the context of B2B exchanges, making a sale is often seen as a milestone in a broader effort to establish and maintain long-term relationships with clients. Understanding the factors that influence a customer's decision to remain with a supplier is crucial. While product quality has traditionally been viewed as a cornerstone of customer loyalty, its significance has diminished as technology becomes increasingly commoditized. Instead, factors such as the perceived quality of service and effective customer bonding techniques employed by suppliers are critical in cultivating trust and commitment in B2B marketing relationships [26].

Moreover, many SaaS companies depend on a diverse range of other SaaS products to support various operational aspects, utilizing in-house software engineers primarily for developing unique features of their offerings. Consequently, most modern SaaS companies leverage a subscription model for a variety of applications, enabling them to focus on their core competencies while relying on other services to manage noncore operations effectively [5].

3.1.1.2 B2C SaaS

B2C SaaS companies focus on delivering software directly to consumers, emphasizing large-scale customer acquisition and streamlined service delivery to handle high volumes at a low cost [25]. These companies face challenges with customer retention, as users can easily cancel subscriptions, making customer retention a strategic priority. Freemium models and free trials attract users, aiming to convert them into paying customers through engaging experiences and targeted marketing [27].

Typically, B2C SaaS companies benefit from high gross margins, as there are few direct costs beyond initial acquisition and minimal customer service expenses, often managed through self-service resources like FAQs [28]. Monthly billing, however, increases churn risks, as customers can quickly end subscriptions; this makes churn prediction models crucial. Customer Acquisition Cost (CAC) and Customer Lifetime Value (CLTV) are essential metrics, with profitable models converting free users into paying customers by optimizing onboarding and engagement processes[27].

Due to significant up-front customer acquisition costs and monthly billing cycles, these companies often experience a “Working Capital Trough,” a financial strain from spending heavily on acquisition without immediate returns[28]. Offering annual billing options is one tactic used to ease capital flow issues, as it increases up-front payments. Additionally, customer segmentation and behavioral analysis help personalize engagement, driving targeted retention strategies and improving churn prediction[27].

Dropbox exemplifies the B2C SaaS approach by leveraging a freemium model for cloud storage, prioritizing low-cost, high-volume acquisition. By offering a free option alongside paid plans, Dropbox uses familiarity to drive user adoption, retaining customers by aligning features with user needs and using data insights to enhance the user experience and predict retention [28] [27].

3.1.1.3 Freemium SaaS

A freemium offering refers to any subscription service that includes both a free version and a paid, or premium, level of service. In some cases, the free version may be time-limited, allowing users to ”try before they buy.” In other instances, users may be able to access the free service indefinitely. Another common variation is having a free version that contains advertisements, while the paid version is ad-free [5].

Regarding churn, freemium services function similarly to those without a free tier, but there are two distinct types of churn: churn from the premium service and churn from the free level of service. The free tier is analyzed using techniques described for non-subscription, activity-based churn analysis. Additionally, there is the transition from the free service to the paid level, known as free trial conversion, which can be analyzed using the same techniques as churn [5].

3.1.2 Other Subscription Services

Beyond software, subscription models have expanded into diverse sectors such as media, retail, and education. This subsection explores these non-SaaS categories, including media subscriptions, curated subscription boxes, and membership-based services. These models often rely on content or physical goods delivery, which introduces unique churn challenges and retention dynamics. The analysis focuses on the consumer experience, engagement strategies, and how businesses adapt to shifting customer expectations in highly competitive environments.

3.1.2.1 Media Subscriptions

Media subscription services encompass a variety of offerings that extend beyond traditional Software as a Service (SaaS) models, focusing primarily on content consumption rather than software functionality. These services include online newspapers, digital magazines, and other platforms that charge users a recurring fee for access to content that was previously available for free. The subscription economy has emerged as a response to the decline of advertising revenue models, prompting media companies to seek sustainable business strategies in an increasingly digital landscape. To mitigate churn, or the loss of subscribers, these media services can implement strategies such as personalized content recommendations, exclusive content creation, and flexible subscription plans that cater to diverse user needs. Additionally, fostering community engagement through interactive features can enhance user loyalty and reduce cancellation rates, ensuring a more stable subscriber base in a competitive environment [29].

3.1.2.2 Subscription Boxes

Subscription box services have gained significant traction by offering curated selections of products delivered directly to consumers on a regular basis. These services span various categories, including beauty products, meal kits, apparel, and more. The appeal lies in the convenience, personalization, and discovery experience they provide to subscribers.

According to Andonova et al., the subscription box industry can be understood through the "four Cs": Categories, Consumer benefits, Customers, and Competitive landscape. This framework helps in analyzing the diverse offerings and the value proposition of subscription boxes. For instance, beauty subscription boxes like Birchbox provide consumers with personalized samples of beauty products, enhancing the discovery of new items tailored to individual preferences.

However, the industry faces challenges such as high customer churn rates and the need for continuous value delivery to retain subscribers. Critical success factors identified include effective customer engagement strategies, maintaining product quality and variety, and creating a unique unboxing experience that delights customers. Understanding these elements is crucial for businesses aiming to succeed in the competitive subscription box market.

In summary, subscription boxes have revolutionized the way consumers access and experience products, offering a blend of convenience and personalization. Academic insights shed light on the dynamics of this industry, providing guidance for businesses to navigate challenges and leverage opportunities effectively.r [30].

3.1.2.3 Membership-Based Services

Membership-based services, such as gyms and online courses, operate on a subscription model where users pay a recurring fee to access specific services or content. This model has gained popularity due to its ability to provide personalized experiences tailored to individual needs and preferences. For instance, gyms may offer various classes and training programs, while online courses can adapt to the learner's pace and interests.

The subscription model fosters customer loyalty over time, as users often become more satisfied with the service as they learn to utilize it effectively. This satisfaction can lead to a lower churn rate, meaning customers are less likely to cancel their subscriptions after an initial period, especially as they develop habits around the service.

Moreover, the growth of digital technology and the impact of the Covid-19 pandemic have accelerated the demand for such services, as more people seek flexible and accessible options for fitness and education. Overall, membership-based services leverage the subscription model to create stable, predictable revenue streams while enhancing customer engagement and satisfaction[31].

3.2 Non-Subscription Models

While subscription models have gained popularity, many successful businesses continue to operate under non-subscription revenue structures. This section examines alternative monetization strategies such as transactional services, in-app purchases, and ad-supported models. These approaches offer greater flexibility for customers and can be more suitable in contexts where usage is infrequent or customer preferences vary widely. By understanding how non-subscription models operate, we can better contextualize churn as a broader behavioral indicator, even outside of recurring billing contexts.

3.2.1 Transactional Services

Transactional service models operate on a pay-per-use basis, allowing customers to access specific products or services without the obligation of recurring payments. These models offer high flexibility and are widely used in both digital and physical retail environments. In his article "Non-subscription revenue as a source of income," Richard Gedye explores these strategies within academic publishing, highlighting models such as Pay-Per-View and licensing for subsidiary markets [32]. However, transactional models extend far beyond publishing and are foundational to some of the world's most successful businesses.

Retail giants such as **Zara** and **IKEA** exemplify traditional transactional models. Their revenue is generated through one-off purchases of physical goods, without the need for ongoing subscriptions. These companies focus on optimizing product turnover, shopping experience, and inventory strategies rather than customer retention through recurring billing [33, 34].

Amazon, although offering subscription-based services like Amazon Prime, primarily operates on a transactional basis for non-Prime users. Customers make one-time purchases of goods or digital products (e.g., books, electronics) without needing to commit to a subscription. In this sense, Amazon straddles both worlds—offering a hybrid model that blends transactional flexibility with subscription-based loyalty [35].

Transactional models are particularly attractive to consumers who prioritize flexibility and control over their spending. They are also well-suited to industries with infrequent or irregular purchase cycles, such as home furnishing (e.g., IKEA), fashion retail (e.g., Zara), or general e-commerce (e.g., non-Prime Amazon) [33, 34, 35].

These examples highlight how transactional revenue strategies remain highly relevant and profitable, especially when supported by strong brand equity, efficient logistics, and customer-centric experiences.

3.2.2 In-App Purchase Models

In-app purchase models differ from freemium subscription services in that they do not necessarily require a recurring payment model. Instead, users interact with a base application—often free—and have the option to make one-time purchases within the app to access specific features, virtual goods, or content. Unlike freemium SaaS, where the goal is to convert users into recurring subscribers, in-app purchases rely heavily on microtransactions to generate revenue. These models are commonly employed in gaming applications, where users pay for items such as extra lives, power-ups, or aesthetic customizations, as well as in platforms offering pay-as-you-go features.

While both freemium and in-app models aim to attract users with a free entry point, the revenue streams differ. Freemium SaaS focuses on long-term customer retention through premium upgrades, often requiring subscription fees. In contrast, in-app purchase models generate immediate income by providing on-demand access to premium content without obligating users to a recurring plan. According to Gedye (2002), this model allows content providers to monetize their offerings incrementally, encouraging frequent but low-commitment transactions

This flexibility is particularly appealing to users who prefer not to engage in subscription services but are willing to make occasional purchases, thereby widening the potential market for businesses leveraging this revenue strategy [36].

3.2.3 Ad-Supported Models

According to their thesis findings of Ralf Schimmer, Kai Karin Geschuhn, and Andreas Vogler, ad-supported models in academic publishing offer a promising approach to providing free access to scholarly content while generating revenue through advertisements. This model aims to reduce the financial burden on libraries and institutions by allowing them to redirect funds typically spent on subscriptions to cover publication costs, thereby facilitating a transition to open access. Although still in its nascent stages and not yet widely implemented globally, this approach has the potential to enhance transparency and sustainability in scholarly communication. By reallocating existing subscription budgets, libraries can foster a more equitable system that promotes broader dissemination of research without imposing additional costs on users. This shift could redefine the roles of publishers and librarians, creating a collaborative environment that supports both open access and the financial viability of publishing [37].

3.3 Key Metrics for Combating Churn

To effectively manage churn, it is not enough to understand business models—companies must also measure performance through targeted metrics. This section introduces the core indicators used in churn analysis and retention planning. Metrics such as Lifetime Value (LTV), Lifetime Revenue (LTR), and Monthly Recurring Revenue (MRR) Churn provide a financial lens through which customer behavior and business health can be evaluated. These metrics are especially crucial in SaaS and other recurring-revenue models where long-term profitability depends on sustained customer relationships.

3.3.1 Key SaaS Metrics: LTV, LTR, and MRR Churn, NDR

Understanding and monitoring key performance metrics is essential for evaluating the health and sustainability of SaaS businesses. This section introduces four critical metrics—Lifetime Value (LTV), Lifetime Revenue (LTR), Monthly Recurring Revenue (MRR) Churn, and Net Dollar Retention (NDR)—which collectively provide a comprehensive view of customer profitability, revenue predictability, and retention performance. These indicators not only inform strategic decisions around customer acquisition and engagement but also help identify churn patterns and revenue growth opportunities within the existing customer base. By leveraging these metrics, SaaS companies can enhance their ability to drive long-term value, optimize financial efficiency, and build scalable subscription models.

3.3.1.1 Lifetime Value (LTV)

The concept of *Lifetime Value* (LTV) extends beyond simple revenue estimation. While Lifetime Revenue (LTR) calculates the total revenue a customer generates during their relationship with the company, LTV provides a more comprehensive view by incorporating the costs of servicing that customer, known as the Cost of Goods Sold (COGS). The formula for LTV is:

$$LTV = LTR \times \text{Gross Margin} = LTR - COGS$$

LTV is crucial for understanding a company's long-term profitability. It quantifies the net revenue a business can expect after covering the costs associated with customer service. For SaaS companies, where recurring revenue is the foundation, monitoring LTV ensures sustainable growth. Comparing LTV with Customer Acquisition Cost (CAC) helps assess whether the company is on a profitable trajectory or experiencing unsustainable negative cash flow, often referred to as the "Triangle of Despair" [3].

A high LTV suggests strong customer loyalty and product value, while a low LTV may indicate poor customer retention, weak product-market fit, or high operational costs. Optimizing LTV involves increasing LTR and reducing COGS. This can be achieved by improving operational efficiency, negotiating better vendor contracts, and reducing cloud or customer support costs, all while maintaining service quality to prevent churn.

3.3.1.2 Lifetime Revenue (LTR)

Lifetime Revenue (LTR) represents the total cumulative revenue a business can expect from a single customer throughout their relationship with the company. It serves as a foundational metric for estimating customer value and plays a critical role in the calculation of Lifetime Value (LTV). LTR is particularly relevant for SaaS businesses, where revenue is realized over time rather than through one-time purchases. The longer a customer remains subscribed, the higher their contribution to total revenue.

LTR can be calculated using the churn rate and the Average Revenue Per Account (ARPA), following the formula:

$$\text{LTR} = \left(\frac{1}{\text{Churn Rate}}\right) \times \text{ARPA}$$

For instance, consider a Netflix subscriber who remains active for 50 years, paying a monthly fee of $12. Over this period, the total revenue generated by the customer would be:

$$\text{LTR} = 50 \times 12 \times 12 = \$7,200$$

However, LTR alone does not provide a complete picture. The revenue generated from a customer must also be analyzed in relation to acquisition costs and ongoing service expenses to ensure profitability. A high LTR suggests strong customer retention and product engagement, whereas a low LTR may indicate churn issues or inadequate monetization strategies.

An essential component in the LTR formula is the Average Revenue Per Account (ARPA), which measures the average revenue generated by each customer over a given period, typically monthly or annually. ARPA serves as a foundational metric for assessing revenue streams across different customer segments and is defined as:

$$\text{ARPA} = \frac{\text{Annual Recurring Revenue (ARR)}}{\text{Total Customer Accounts}}$$

By segmenting ARPA across customer cohorts, businesses can identify revenue distribution patterns and tailor retention strategies accordingly. A rising ARPA suggests successful upselling or improved customer engagement, while a declining ARPA may indicate pricing inefficiencies or customer dissatisfaction.

3.3.1.3 Monthly Recurring Revenue (MRR) Churn

MRR Churn is a key SaaS metric that measures revenue loss due to customer cancellations or delinquent payments. It provides critical insights into revenue retention efficiency and areas that require improvement. Understanding the composition of MRR Churn helps businesses refine their customer engagement and payment strategies.

Revenue loss in MRR Churn primarily stems from two components: active cancellations and delinquent cancellations. **Active cancellations** occur when customers voluntarily decide to end their subscriptions, often due to dissatisfaction, lack of engagement, or better alternatives. On the other hand, **delinquent cancellations** result from failed payment methods, where customers churn unintentionally due to expired credit cards or insufficient funds.

To quantify the total impact of MRR Churn, we use the following formula:

$$\text{MRR Churn} = \sum \text{MRR Cancellation} + \sum \text{MRR Delinquent}$$

For instance, if a company loses $1,000 in MRR during a given month, comprising $800 from active cancellations and $200 from delinquent accounts, then:

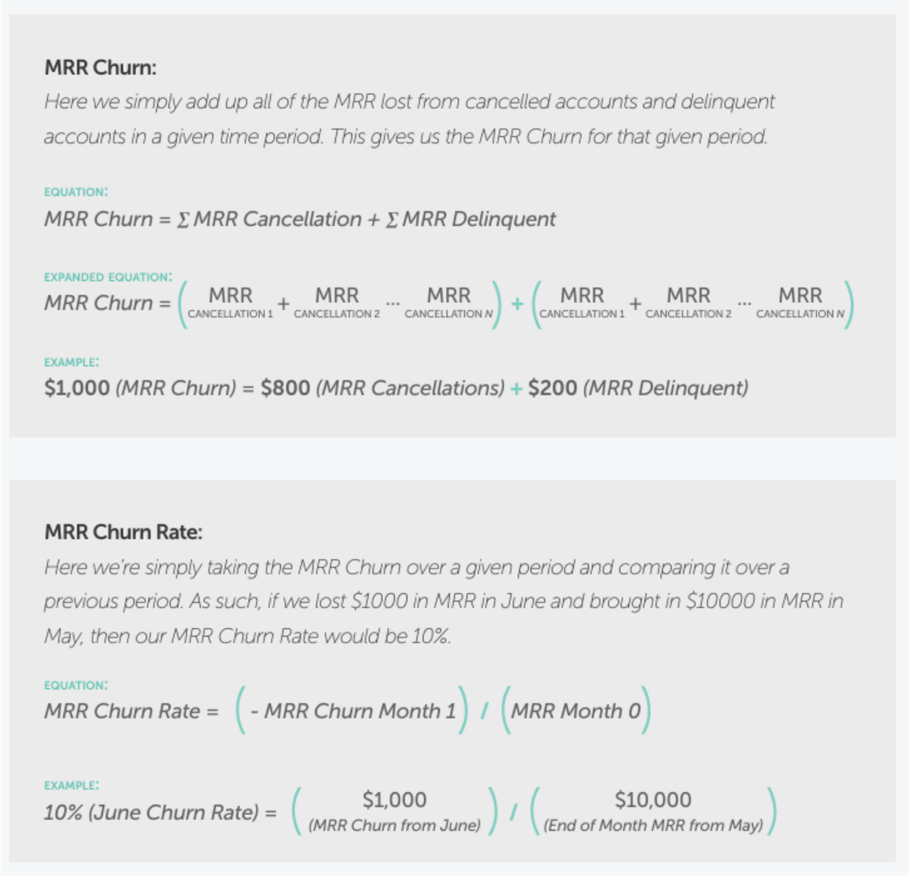
$$\text{MRR Churn} = 800 + 200 = 1,000$$

Beyond absolute revenue loss, it is also important to measure the MRR Churn Rate, which expresses churn as a percentage of total MRR at the beginning of the month. This metric helps track revenue retention trends over time:

$$\text{MRR Churn Rate} = \frac{-\text{MRR Churn Month 1}}{\text{MRR Month 0}}$$

For example, if the company starts with an MRR of $10,000 and loses $1,000 during the month, the churn rate would be:

$$\text{MRR Churn Rate} = \frac{-1,000}{10,000} = 10\%$$



MRR Churn:

Here we simply add up all of the MRR lost from cancelled accounts and delinquent accounts in a given time period. This gives us the MRR Churn for that given period.

EQUATION:

MRR Churn = ∑ MRR Cancellation + ∑ MRR Delinquent

EXPANDED EQUATION:

$$\textit{MRR Churn} = \left(\text{MRR}_{\text{CANCELLATION 1}} + \text{MRR}_{\text{CANCELLATION 2}} \ldots \text{MRR}_{\text{CANCELLATION } N}\right) + \left(\text{MRR}_{\text{CANCELLATION 1}} + \text{MRR}_{\text{CANCELLATION 2}} \ldots \text{MRR}_{\text{CANCELLATION } N}\right)$$

EXAMPLE:

$1,000 *(MRR Churn)* = **$800** *(MRR Cancellations)* + **$200** *(MRR Delinquent)*

MRR Churn Rate:

Here we're simply taking the MRR Churn over a given period and comparing it over a previous period. As such, if we lost $1000 in MRR in June and brought in $10000 in MRR in May, then our MRR Churn Rate would be 10%.

EQUATION:

MRR Churn Rate = (- MRR Churn Month 1) / (MRR Month 0)

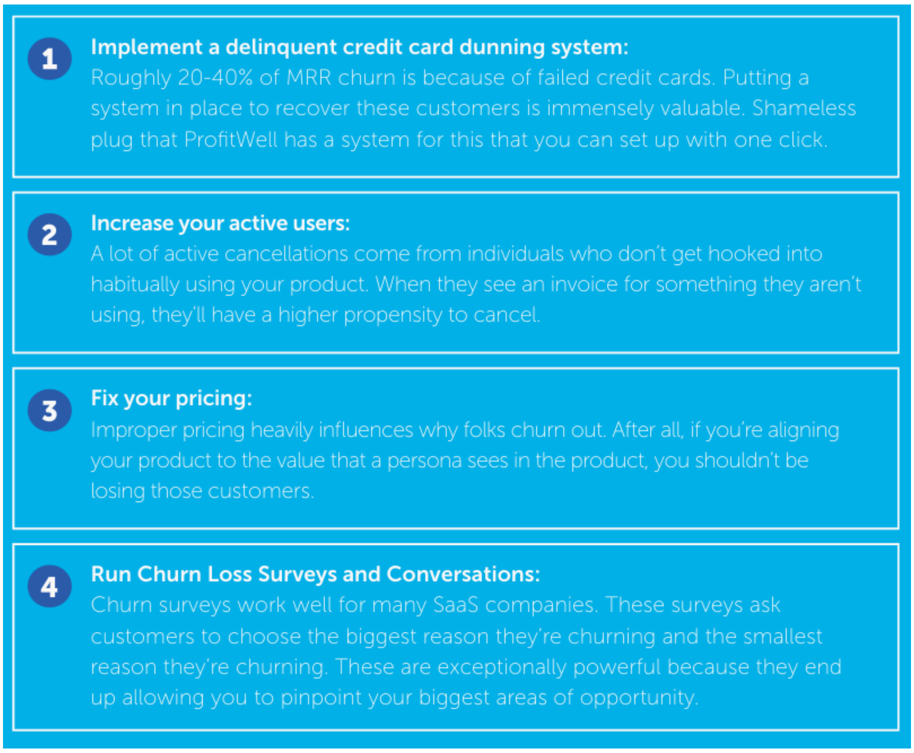
EXAMPLE:

10% (June Churn Rate) = ($1,000 *(MRR Churn from June)*) / ($10,000 *(End of Month MRR from May)*)

Figure 3.2: Formula for Calculating MRR Churn and MRR Churn Rate [2]

MRR Churn is a crucial indicator of business health, impacting both product and financial strategy. A high churn rate may signal product deficiencies, such as missing features or poor customer satisfaction, which must be addressed to enhance retention. From a financial perspective, tracking MRR Churn enables businesses to forecast revenue trends, optimize cash flow, and make informed budgeting decisions.

To mitigate MRR Churn, companies can implement targeted strategies that focus on retention, proactive engagement, and payment recovery. These strategies may include improving customer support, offering personalized retention incentives, optimizing billing processes, and implementing churn prediction models.



1 **Implement a delinquent credit card dunning system:**
Roughly 20-40% of MRR churn is because of failed credit cards. Putting a system in place to recover these customers is immensely valuable. Shameless plug that ProfitWell has a system for this that you can set up with one click.

2 **Increase your active users:**
A lot of active cancellations come from individuals who don't get hooked into habitually using your product. When they see an invoice for something they aren't using, they'll have a higher propensity to cancel.

3 **Fix your pricing:**
Improper pricing heavily influences why folks churn out. After all, if you're aligning your product to the value that a persona sees in the product, you shouldn't be losing those customers.

4 **Run Churn Loss Surveys and Conversations:**
Churn surveys work well for many SaaS companies. These surveys ask customers to choose the biggest reason they're churning and the smallest reason they're churning. These are exceptionally powerful because they end up allowing you to pinpoint your biggest areas of opportunity.

Figure 3.3: Strategies to Reduce MRR Churn [2]

Continuous monitoring and optimization of MRR Churn are essential for sustaining recurring revenue, improving customer satisfaction, and ensuring long-term business profitability. By addressing the root causes of churn and refining customer engagement strategies, SaaS businesses can drive sustainable growth and financial resilience.

3.3.2 Net Dollar Retention (NDR) as a Key SaaS Metric for Combating Churn

Net Dollar Retention (NDR) is a recurring revenue metric that indicates the percentage of revenue retained from an existing customer base over a defined period of time [38]. In other words, it measures how revenue from current customers evolves, considering both gains from upsells and cross-sells, and losses due to reduced usage or cancellations [39].

This metric, also referred to as Net Revenue Retention (NRR) or Net Expansion Rate, is expressed as a percentage [39]. An NDR value above 100% implies that revenue from existing customers has grown overall—meaning expansions outweigh churn—while an NDR below 100% indicates that the company is losing net revenue from its existing customer base [39].

Because NDR incorporates multiple factors that affect recurring revenue (such as upgrades, downgrades, and cancellations), it is considered one of the most comprehensive indicators of customer base health and retention performance in SaaS business models [38].

The standard formula for calculating NDR is:

$$\text{NDR} = \left(\frac{\text{Beginning MRR} + \text{Expansion} - \text{Contraction} - \text{Churn}}{\text{Beginning MRR}} \right) \times 100\%$$

Where:

- **Beginning MRR** refers to the Monthly Recurring Revenue at the start of the period from existing customers.
- **Expansion** includes additional revenue from the same customers (e.g., upsells or cross-sells).
- **Contraction** is the revenue lost from customers who downgraded to a lower-priced plan or reduced their usage.
- **Churn** refers to revenue lost from customers who completely cancelled their subscriptions.

Step-by-step example: Suppose a company starts a year with $1,000,000 in Annual Recurring Revenue (ARR) from its current customers [40]. During the year:

- It earns $200,000 in expansion revenue.
- It loses $100,000 in revenue from churned customers.
- It has no downgrades or contractions.

Then:

$$\text{NDR} = \left(\frac{1,000,000 + 200,000 - 0 - 100,000}{1,000,000} \right) \times 100 = 110\%$$

This means the company experienced net revenue growth of 10% from its existing customer base, reflecting strong retention and effective upselling strategies [40]. A high NDR like this is often associated with high-performing SaaS companies and is a key driver of scalable, sustainable growth [38, 39].

In conclusion, NDR captures the full picture of customer value retention over time, accounting for both positive (expansion) and negative (contraction and churn) revenue movements. For SaaS businesses, it is a critical performance indicator that guides decisions related to customer success, product development, and revenue forecasting [39, 41].

3.3.3 Metrics for Non-Subscription Models

While this study focuses primarily on SaaS models, many principles and metrics also apply to other types of subscription-based businesses, such as digital media, subscription boxes, or online education platforms. In these cases, metrics such as **average customer tenure**, **product adoption rate**, or **engagement frequency** may be more prominent than pure MRR-based indicators. These metrics will not be the focus of our empirical analysis, but they provide valuable context for understanding broader churn dynamics in the subscription economy.

4 Data Preparation

This chapter outlines the steps undertaken to prepare the dataset for predictive modeling. Effective data preparation ensures both the accuracy and reliability of churn prediction models, particularly when dealing with large-scale transactional data from a digital streaming platform. The process encompasses data cleaning, transformation, normalization, and feature engineering, with a dual approach that combines Python-based techniques and Graphext's automated preprocessing pipeline. By systematically preparing and refining the data, we ensure that the resulting models are trained on clean, consistent, and representative inputs, ultimately improving their predictive power and interpretability.

4.1 Description of the Transactional Database

The dataset utilized in this thesis originates from a digital streaming platform specializing in movies and TV series. It comprises **21 variables** and approximately **245,000 observations**, each representing a unique customer. The primary goal is to predict which customers are likely to "churn" in the near future, enabling relevant departments to take proactive measures by targeting these individuals with appropriate action plans.

To achieve accurate predictions, it was crucial to structure the data effectively, ensuring that categorical and numerical features were well-prepared for machine learning models. Graphext facilitated this process by providing an automated pipeline for preprocessing, which included feature encoding, scaling, and missing value imputation, ensuring that the dataset was optimized for predictive analysis.

With a dataset structured in this manner, we gain a comprehensive understanding of customer behavior, preferences, and engagement with the platform. This structured approach enables the development of robust predictive models aimed at identifying potential churners with higher precision.

4.2 Data Cleaning and Transformation

Data cleaning and transformation are foundational tasks in any machine learning workflow. In this section, we describe the sequential steps taken to ensure the integrity and consistency of the dataset. This includes identifying and treating missing values, detecting and handling outliers, and normalizing both numerical and categorical variables to standardize input features. We also detail how Graphext was used alongside traditional Python tools to automate and enhance these processes. The goal of this phase is to minimize noise, reduce bias, and prepare the dataset for optimal performance in predictive churn modeling.

4.2.1 Identification of Missing Data

The initial step in preparing the data involved identifying any missing or incomplete entries. While the dataset was generally complete, containing 243,787 non-null entries across all 21 columns, additional verification was performed using Graphext. The platform automatically identified and handled any missing data by applying statistical imputation techniques such as mean, median, or mode replacements.

This step was critical to ensure that no biases or inconsistencies were introduced due to incomplete records. The combination of manual Python-based verification and Graphext's automated processes ensured a **fully complete dataset**, ready for model development.

4.2.2 Handling Outliers

To assess the presence of outliers, boxplots were generated using Python. This visualization allowed us to identify extreme values in key numerical variables. While most attributes such as **AccountAge**, **MonthlyCharges**, and **ViewingHoursPerWeek** displayed relatively uniform distributions, **TotalCharges** exhibited a long-tailed distribution, suggesting the presence of extreme values.

These outliers were further examined using Graphext, which provided an additional layer of validation. Rather than removing these outliers outright, the data was **standardized using StandardScaler**, ensuring that extreme values did not disproportionately influence the predictive model. This combined approach, leveraging both Python-based visualization and Graphext's automated transformations, provides a clearer understanding of customer behaviors while maintaining the integrity of the dataset.

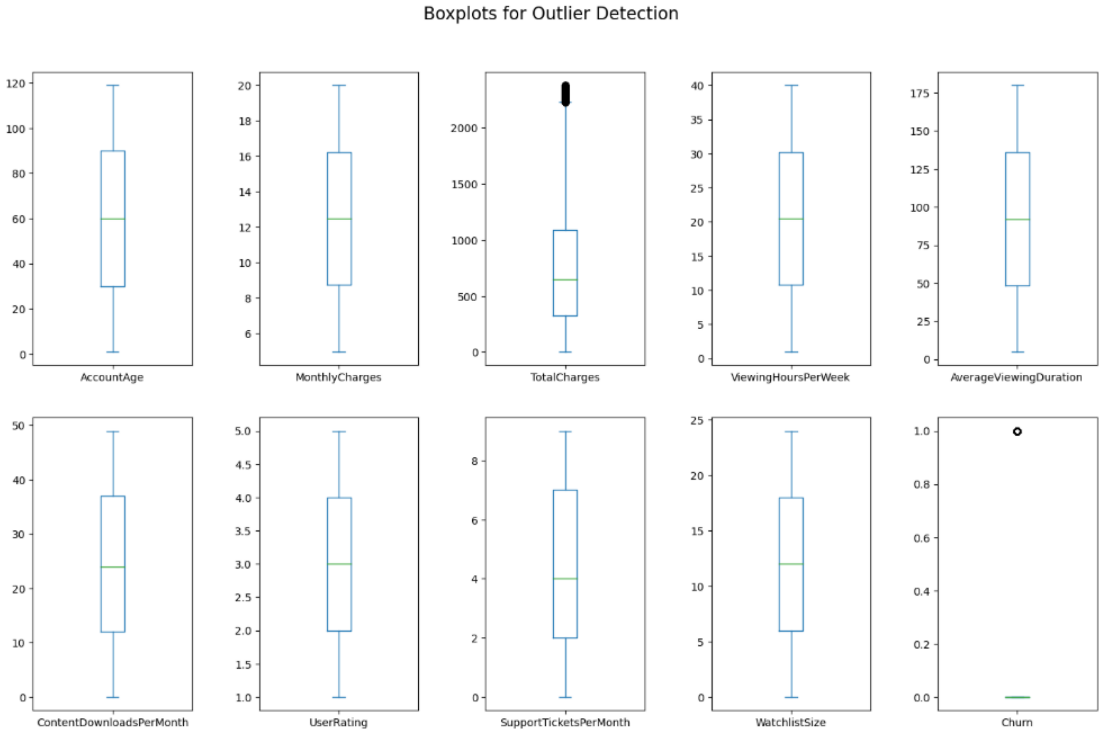


Figure 4.1: Boxplot Analysis for Outlier Detection

Understanding these outliers is crucial, as they often represent **significant customer segments**, such as premium users or long-term subscribers, whose behaviors are valuable for strategic decision-making.

4.2.3 Data Normalization

After handling missing values and outliers, numerical and categorical variables were **normalized** to ensure consistency across the dataset. This step was essential for improving the performance and interpretability of our predictive model.

Numerical Variables: Standardization was applied using the **StandardScaler** transformation, which adjusted variables such as **AccountAge, MonthlyCharges, ViewingHoursPerWeek, and TotalCharges** to have a mean of zero and a standard deviation of one. This prevented variables with larger numerical ranges from dominating model predictions.

Categorical Variables: To handle categorical features, Graphext applied **OneHotEncoding** to variables such as **SubscriptionType, PaymentMethod, and ContentType**, converting them into binary columns suitable for machine learning algorithms. Additionally, **Euclidean Distance Scaling** was applied to features like **GenrePreference and DeviceRegistered**, allowing categorical values to be represented based on similarity distances.

Binary and Multi-Label Features: Attributes like **SupportTicketsIndicator** and multi-category labels such as **customer segmentation and viewer segmentation** were encoded using **MultiLabelBinarizer** to preserve their categorical relationships within the dataset.

Graphext's **ColumnTransformer** and **FeatureUnion** automatically structured these transformations into a cohesive preprocessing pipeline. By **combining Python visualizations with Graphext's automated processing**, we ensured a robust preprocessing framework that balances manual insights with advanced machine learning optimizations.

This comprehensive preprocessing strategy, integrating both Python-based and Graphext-driven approaches, guarantees that the dataset is fully optimized for predictive modeling, minimizing bias and enhancing accuracy.

4.2.4 Feature Engineering and Segmentation

In this subsection, we explore how segmenting key variables enhances our understanding of customer behavior and provides a robust basis for building predictive models. To achieve this, we transitioned from Python-based data cleaning to Graphext, which offers advanced visualizations and an intuitive interface for creating segmented categories. This segmentation process focuses on key customer behaviors, such as account age, viewing habits, and demographic groupings. Below, we explain the rationale behind each segmentation and its contribution to our analysis.

4.2.4.1 Segmentation of Account Age

The variable **AccountAge** was divided into three segments: *new customers*, *standard customers*, and *old customers.* This segmentation is based on specific time intervals:

- *New Customers*: Account age less than 24 months.
- *Standard Customers*: Account age between 24 and 60 months.
- *Old Customers*: Account age greater than 60 months.

By analyzing the mean **ViewingHoursPerWeek**, **AverageViewingDuration**, and **ContentDownloadsPerMonth** for each segment, we gained insights into how customer tenure affects their engagement with the platform. The results reveal that viewing habits and download frequencies remain relatively stable across customer age groups, suggesting consistent user engagement over time.

4.2.4.2 Segmentation of Viewing Hours per Week

To better understand customer engagement, the variable **ViewingHoursPerWeek** was segmented into:

- *Less Watchers*: Below the 25th percentile of viewing hours.
- *Watchers*: Between the 25th and 75th percentiles.
- *Hard Watchers*: Above the 75th percentile.

This segmentation allows us to analyze differences in behavior between casual viewers and highly engaged users. For example, while *hard watchers* have significantly higher weekly viewing hours, their **AverageViewingDuration** remains comparable to other segments. This finding implies that increased viewing time might reflect more frequent, but shorter, sessions rather than extended continuous viewing.

4.2.4.3 Segmentation by Age and Gender through Parental Control

Using the **ParentalControl** and **Gender** variables, we created four demographic segments:

- *Child Male*
- *Child Female*
- *Adult Male*
- *Adult Female*

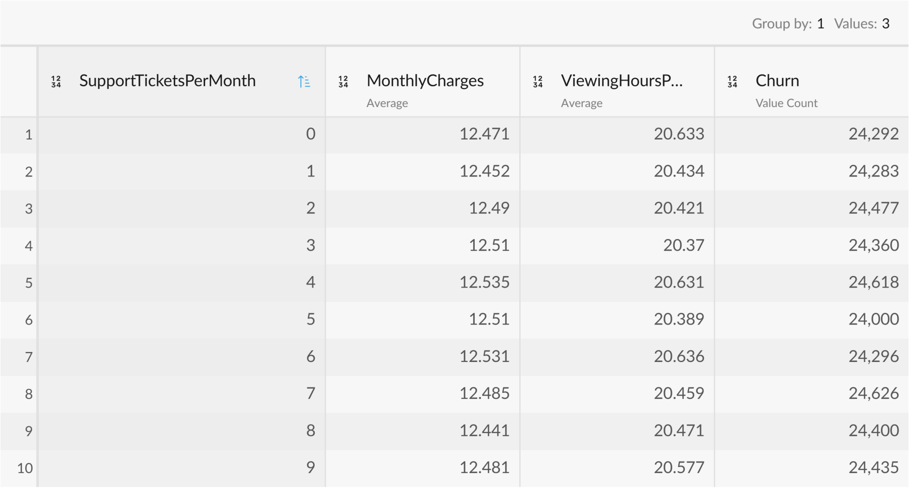
This segmentation helps uncover behavioral differences across demographic groups. For example, *Child Females* had slightly higher **ContentDownloadsPerMonth** compared to *Child Males*, while *Adult Males* exhibited higher **AverageViewingDuration**. These patterns provide valuable insights for tailoring content recommendations and marketing strategies to specific demographic groups.

4.2.4.4 New Column Creation: `NewSupportTicketsPerMonth`

We have created a new binary categorical variable called `NewSupportTicketsPerMonth`, based on the number of support tickets raised by each customer (`SupportTicketsPerMonth`). The goal is to segment customers into two groups based on their support ticket frequency.

- Customers with fewer than 6 support tickets per month are assigned a value of **0** in the new column. This group represents customers with low or moderate support needs.
- Customers with 6 or more support tickets per month are assigned a value of **1**. This group represents high-support users who may require additional resources or attention from the customer support team.

This segmentation allows for better analysis and targeted strategies, such as prioritizing high-support customers for improved service or identifying potential issues that might lead to churn.



Group by: 1 Values: 3

	SupportTicketsPerMonth	MonthlyCharges Average	ViewingHoursP... Average	Churn Value Count
1	0	12.471	20.633	24,292
2	1	12.452	20.434	24,283
3	2	12.49	20.421	24,477
4	3	12.51	20.37	24,360
5	4	12.535	20.631	24,618
6	5	12.51	20.389	24,000
7	6	12.531	20.636	24,296
8	7	12.485	20.459	24,626
9	8	12.441	20.471	24,400
10	9	12.481	20.577	24,435

Figure 4.2: Summary Table: `NewSupportTicketsPerMonth` Segmentation

4.2.4.5 Importance of Segmentation

Segmenting these features provides several advantages:

1. **Improved Feature Engineering**: Segmented variables create new categorical features that may enhance model performance by capturing non-linear relationships.

2. **Actionable Insights**: The ability to group customers by behavior and demographics allows targeted interventions, such as personalized content recommendations or retention strategies.

3. **Model Interpretability**: By creating clear segments, we improve the interpretability of the models, enabling stakeholders to understand how different customer segments contribute to churn prediction.

These segmentations represent a strategic step toward leveraging the dataset for predictive modeling while ensuring that customer diversity and varying engagement levels are adequately captured. Through Graphext, we were able to efficiently implement these segments and visualize their impact, enhancing our overall analysis.

4.3 Analysis of Data Characteristics

Before applying machine learning algorithms, it is essential to conduct a thorough exploratory analysis of the dataset. This section focuses on examining the statistical properties and behavioral patterns within the data, including distribution shapes, skewness, and variable correlations. We analyze both numerical and categorical features, assess the class distribution of the target variable, and evaluate feature relationships using Pearson correlation matrices. These insights provide a foundation for informed feature selection and guide the next stages of model development by highlighting variables with potential predictive relevance.

4.3.1 Analysis of Numerical and Categorical Variables

To gain a comprehensive understanding of the dataset, both numerical and categorical variables were analyzed.

Numerical Variables: Visualizations of numerical distributions, including histograms combined with kernel density estimates (KDE), revealed distinct patterns. For instance, **TotalCharges** showed positive skewness, while other variables had more balanced distributions. These distribution insights provide critical information for preprocessing and model-building.

Categorical Variables: Categorical variables were analyzed to understand customer preferences. Frequency distributions were visualized for variables such as **SubscriptionType**, **PaymentMethod**, and **ContentType**. Key observations include:

- **SubscriptionType**: Balanced distribution among Premium, Basic, and Standard plans, enabling unbiased analysis.
- **PaymentMethod**: Similar distribution across methods like Credit Card, Electronic Check, Mailed Check, and Bank Transfer.
- **PaperlessBilling** and **MultiDeviceAccess**: Nearly equal split between Yes and No responses, reflecting diverse preferences.
- **DeviceRegistered** and **GenrePreference**: Balanced distribution across device types (e.g., Mobile, TV) and genres (e.g., Sci-Fi, Comedy), showcasing customer diversity.

However, the target variable **Churn** demonstrated a significant class imbalance, with 82% of customers classified as non-churn (class 0) and only 18% as churn (class 1). This imbalance presents a challenge and will require specific handling during the modeling phase to ensure robust and unbiased predictions.

4.3.2 Analysis of Distributions and Skewness

Further analysis was conducted on the numerical variables to assess their distribution characteristics and symmetry. Skewness and kurtosis values were computed to quantitatively evaluate the shape of the distributions. While most variables showed minimal skewness, the extbfChurn variable revealed a skewness of 1.65, indicating an imbalance in churned versus non-churned customers.

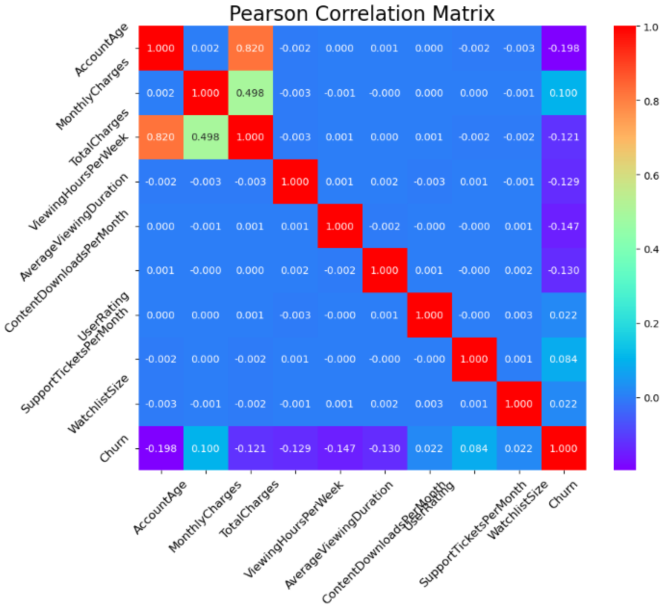
Statistical checks for normality were also performed. For a variable to follow a normal distribution, skewness should be close to 0, and kurtosis should approximate -3. Deviations from these benchmarks indicate non-normal distributions. For example, positively skewed variables have longer tails on the right, while high kurtosis indicates more peaked or flatter distributions.

These metrics are vital for identifying potential transformation requirements, such as log-scaling, to align variable distributions with assumptions underlying specific machine learning models. Distributions with high skewness or kurtosis may affect model performance if left untreated. As part of this step, visual inspections using histograms and KDE plots reinforced these statistical findings, ensuring robust preprocessing for the next stages of analysis.

4.3.3 Pearson Correlation for Numeric Variables

To understand the relationships between numerical variables, a Pearson correlation analysis was conducted. This statistical measure quantifies the linear relationship between pairs of numerical variables, offering insights into how strongly they are related.

Pearson Correlation Matrix: The correlation matrix for numerical variables reveals key relationships, highlighting variables with high positive or negative correlations. Visualized as a heatmap, it facilitates quick identification of strong correlations, such as the high correlation between **TotalCharges** and **AccountAge**, indicating that customers with older accounts tend to have higher total charges.



	AccountAge	MonthlyCharges	TotalCharges	ViewingHoursPerWeek	AverageViewingDuration	ContentDownloadsPerMonth	UserRating	SupportTicketsPerMonth	WatchlistSize	Churn
AccountAge	1.000	0.002	0.820	-0.002	0.000	0.001	0.000	-0.002	-0.003	-0.198
MonthlyCharges	0.002	1.000	0.498	-0.003	-0.001	-0.000	0.000	0.000	-0.001	0.100
TotalCharges	0.820	0.498	1.000	-0.003	0.001	0.000	0.001	-0.002	-0.002	-0.121
ViewingHoursPerWeek	-0.002	-0.003	-0.003	1.000	0.001	0.002	-0.003	0.001	-0.001	-0.129
AverageViewingDuration	0.000	-0.001	0.001	0.001	1.000	-0.002	-0.000	-0.000	0.001	-0.147
ContentDownloadsPerMonth	0.001	-0.000	0.000	0.002	-0.002	1.000	0.001	-0.000	0.002	-0.130
UserRating	0.000	0.000	0.001	-0.003	-0.000	0.001	1.000	-0.000	0.003	0.022
SupportTicketsPerMonth	-0.002	0.000	-0.002	0.001	-0.000	-0.000	-0.000	1.000	0.001	0.084
WatchlistSize	-0.003	-0.001	-0.002	-0.001	0.001	0.002	0.003	0.001	1.000	0.022
Churn	-0.198	0.100	-0.121	-0.129	-0.147	-0.130	0.022	0.084	0.022	1.000

Figure 4.3: Pearson Correlation Matrix for Numerical Variables

Correlation of Target Variable with Numeric Variables: Additionally, the correlation between the target variable, **Churn**, and other numeric variables was analyzed. This helps identify which features may hold predictive power for churn. For instance, variables such as **AccountAge** showed a moderate negative correlation with churn, indicating that customers with older accounts are less likely to churn. Similarly, **MonthlyCharges** exhibited a positive correlation, suggesting that customers with higher monthly charges may be more prone to churn.

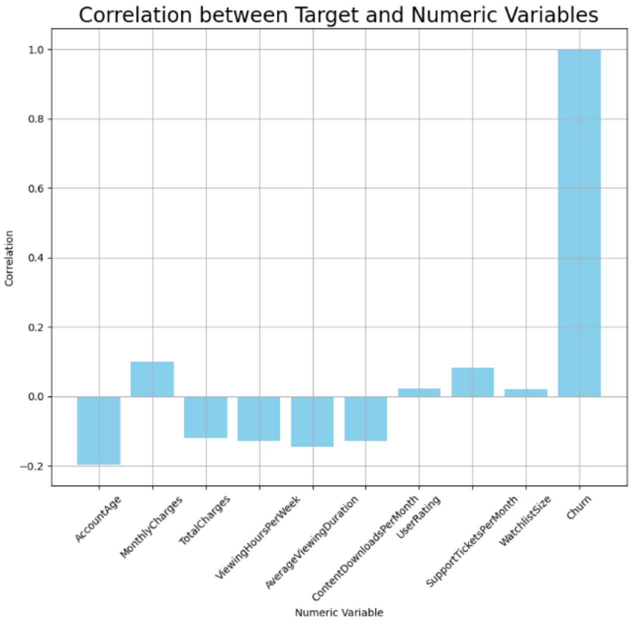


Figure 4.4: Correlation between Target Variable (Churn) and Numeric Variables

The bar plot illustrating the correlation coefficients between **Churn** and each numeric variable offers a visual representation of these relationships, aiding in identifying the most influential predictors for the target variable, in these cases, we observed there are few strong correlations. This analysis provides a solid basis for selecting features and refining predictive models to achieve optimal performance.

5 Model Evaluation

This chapter evaluates the performance of various machine learning models in predicting customer churn, with a particular focus on interpretability, predictive reliability, and business applicability. Starting with a logistic regression model enhanced with elastic net regularization, we prioritize transparency in model decision-making while addressing the challenges of feature multicollinearity and class imbalance. A robust validation strategy is implemented to ensure generalizability, and a suite of evaluation metrics—including precision, recall, F1-score, and AUC-ROC—is used to measure effectiveness in the context of imbalanced data. Special attention is given to the trade-off between recall and precision, where identifying potential churners (even at the cost of some false positives) is considered more valuable than misclassifying loyal customers. The chapter also delves into the interpretability of model outputs through feature importance and behavioral segmentation analyses, helping to identify which customer traits most strongly influence churn. Finally, we compare alternative models such as Random Forest, CatBoost, and ExtraTrees, ultimately selecting logistic regression for its balance between performance and clarity. These insights lay the groundwork for actionable churn mitigation strategies and personalized retention initiatives in the subsequent chapters.

5.1 Model Selection & Training

For this study, we selected a **Logistic Regression** model with *elastic net regularization* as the primary algorithm for predicting customer churn. This choice was based on the need for an interpretable model that balances between L1 (Lasso) and L2 (Ridge) regularization to enhance feature selection and reduce multicollinearity.

The model was implemented using `scikit-learn` version 1.4.2 and trained on a processed dataset containing **21 features** after normalization, encoding, and segmentation. The primary objective was to classify whether a customer would churn (1) or remain active (0).

Key aspects of the model configuration:

- **Regularization:** Elastic Net (combining Lasso and Ridge penalties).
- **Hyperparameters:** Tuning of the regularization parameter was built-in.
- **Solver:** SAGA optimizer, suitable for large-scale datasets.
- **Cross-validation:** 5-fold StratifiedKFold.
- **Class Weight:** Balanced to address class imbalance.

This model setup ensured that we accounted for both feature selection and robust generalization.

5.2 Model Validation Strategy

To evaluate the model's performance, we adopted a **5-fold StratifiedKFold** validation strategy. This approach ensures that each fold maintains the same proportion of churned and non-churned customers, preventing any bias due to class imbalance.

- **Training-Testing Split:** Data was divided into *80% training* and *20% testing.*

- **Cross-validation:** Applied 5-fold Stratified K-Fold during hyperparameter tuning.

This validation technique provided a reliable estimate of the model's generalization performance.

5.3 Performance Metrics & Results

The model achieved an overall **accuracy of 67.92%** on both training and test sets. While accuracy provides a general measure of performance, given the class imbalance (82% non-churn vs. 18% churn), additional metrics were analyzed:

- **Balanced Accuracy:** 68.41% ± 0.0020
- **F1 Macro Score:** 60.72% (macro-averaged across both classes)
- **F1 Weighted Score:** 71.44% ± 0.0011
- **Precision for Churn Class (1):** 32.12%
- **Recall for Churn Class (1):** 69.19%

These metrics highlight that while the model captures a reasonable proportion of churned customers (recall = 69.19%), its precision (32.12%) remains low. This means that although many churn cases are detected, there is a high number of false positives.

Actual / Predicted	**0**	**1**	**Recall**
0	27,000 (55.38%)	12,921 (26.5%)	67.63%
1	2,722 (5.58%)	6,113 (12.54%)	69.19%
Precision	90.84%	32.12%	67.92%

Table 5.1: Confusion Matrix with Precision and Recall

These results raise an important trade-off between **precision** and **recall** in churn prediction. The model's high recall of 69.19% for the churn class indicates that it successfully identifies most customers who are likely to churn. However, the relatively low precision of 32.12% reveals that a large proportion of these predicted churners are actually false positives—customers incorrectly classified as likely to churn.

While this may seem inefficient, it is often acceptable—and even desirable—in churn prevention contexts. Engaging a non-churning customer with a retention campaign may involve low cost (e.g., a message or a small incentive), whereas failing to identify a real churner could result in the permanent loss of revenue. Therefore, in this case, **recall is strategically more important than precision**, as it aligns with the goal of minimizing missed churn cases, even if that comes at the expense of some false alarms. Ultimately, this trade-off reflects a deliberate prioritization of customer retention over campaign efficiency.

To further assess the predictive performance of our model, we analyzed the **Precision-Recall (PR) Curve** and the **Receiver Operating Characteristic (ROC) Curve**, as illustrated in Figures 5.1 and 5.2.

5.3.1 Precision-Recall Curve

The **Precision-Recall (PR) Curve** illustrates the trade-off between **precision** (the proportion of correctly identified churners among all predicted churners) and **recall** (the proportion of actual churners correctly identified). As shown in Figure 5.1, we observe a gradual decline in precision as recall increases, which highlights the challenge of maintaining a balance between identifying churners correctly and minimizing the misclassification of non-churners. This behavior is expected, as higher recall typically leads to an increase in false positives, reducing precision.

Importantly, the model achieved an **average precision of 0.6345**, which is considered acceptable given the class imbalance (only 18% churn cases). In highly imbalanced scenarios, the PR curve is often more informative than the ROC curve, as it focuses specifically on the performance for the minority class (churn). A model with a precision over 0.6 in this context shows that it can correctly identify a substantial portion of churners while limiting the number of unnecessary interventions. However, this value also suggests that further optimization—particularly in feature engineering or threshold selection—may improve precision without sacrificing recall.

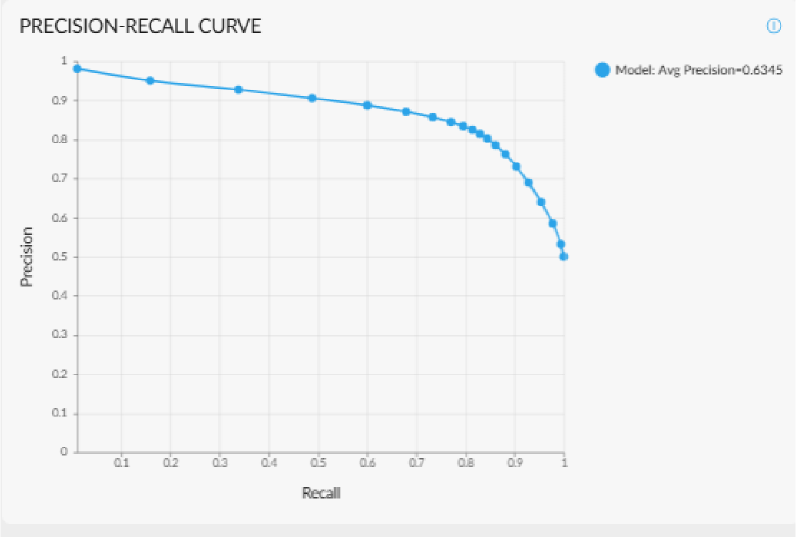


Figure 5.1: Precision-Recall Curve: Illustrating the trade-off between precision and recall at different probability thresholds. Avg Precision = 0.6345.

5.3.2 Receiver Operating Characteristic (ROC) Curve

The **Receiver Operating Characteristic (ROC) Curve** evaluates the model's ability to distinguish between churners and non-churners by plotting the **true positive rate (TPR)** against the **false positive rate (FPR)** at various threshold settings. Figure 5.2 shows that the model achieves an **Area Under the Curve (AUC)** of **0.7181**, which indicates a moderate level of discriminative performance.

An AUC of 0.5 represents random guessing, while a score of 1.0 corresponds to perfect classification. Therefore, a value of 0.7181 suggests that the model performs significantly better than random chance, although it also highlights that there is room for improvement. This performance is typical of early-stage models trained on imbalanced datasets. Enhancing model calibration, experimenting with alternative algorithms, or incorporating additional behavioral features could help improve the model's ability to correctly rank positive instances higher than negative ones.

Overall, the ROC curve confirms the model's general ability to differentiate between classes, but when dealing with churn prediction—where class imbalance is present—the PR curve remains more indicative of real-world utility.

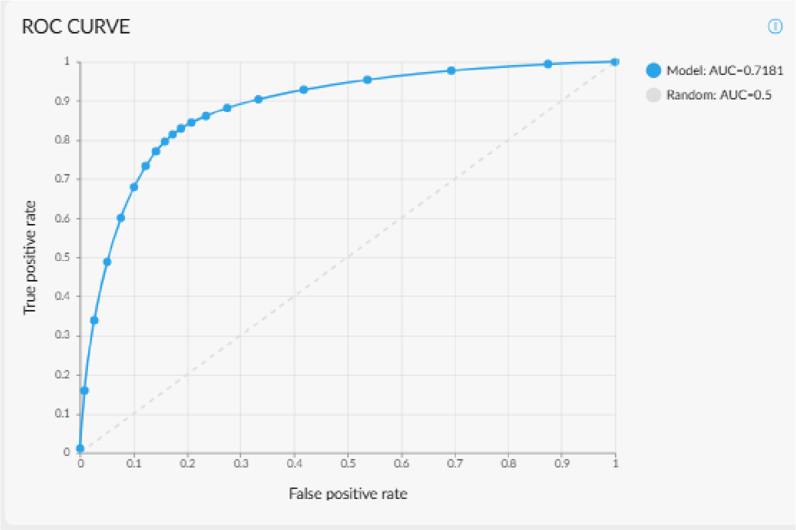


Figure 5.2: Receiver Operating Characteristic (ROC) Curve: Showing the trade-off between true positive rate and false positive rate. AUC = 0.7181.

5.4 Feature Importance Analysis

To interpret the model, we analyzed feature importance using the coefficients from the logistic regression model. In this context, **feature importance is derived from the absolute value of each feature's coefficient**, which reflects the strength and direction of its association with the probability of churn. A higher absolute coefficient indicates a stronger impact on the model's prediction.

The most influential features in predicting churn include:

- **AccountAge:** Customers with longer account history are less likely to churn.
- **AverageViewingDuration:** Longer session duration correlates with customer retention.
- **ContentDownloadsPerMonth:** High download activity is linked to lower churn.
- **MonthlyCharges:** Higher charges are positively associated with churn.
- **SubscriptionType (Premium):** Premium users exhibit different churn behaviors compared to other plans.
- **Payment Method (Credit Card, Electronic Check):** Specific payment methods are correlated with higher churn risk.
- **ViewingHoursPerWeek:** Lower engagement hours increase the likelihood of churn.

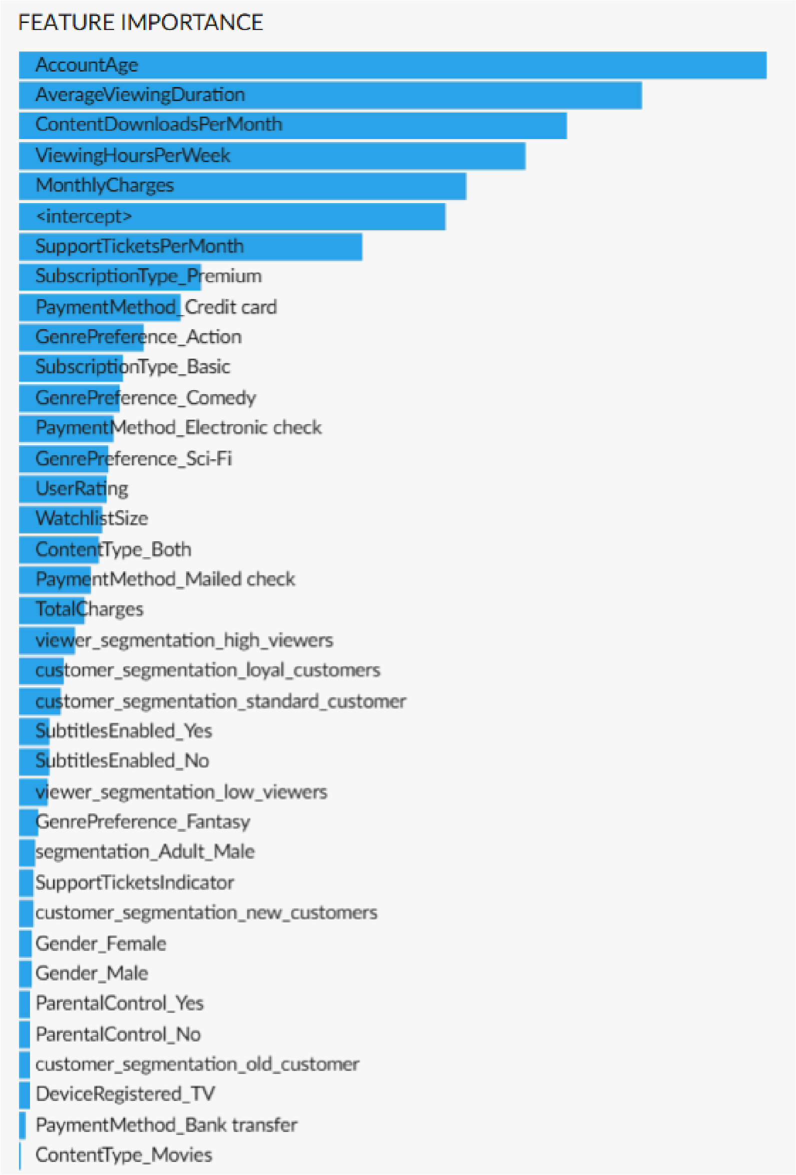


Figure 5.3: Feature Importance Ranking

The interpretation of feature importance provides valuable insights for business strategy. For instance, knowing that **MonthlyCharges** and **PaymentMethod** are positively associated with churn suggests that pricing models and billing experiences may need to be reevaluated. Additionally, high-impact behavioral features such as **AverageViewingDuration** and **ContentDownloadsPerMonth** reinforce the importance of engagement as a protective factor against churn.

These findings empower marketing and retention teams to design more effective interventions. For example, customers with declining engagement or shorter viewing sessions could be targeted with personalized recommendations or re-engagement campaigns. Likewise, payment method segmentation could help tailor messaging or incentives based on risk profiles. Ultimately, this feature-level understanding bridges the gap between model predictions and actionable business decisions.

5.4.1 Impact of Key Features on Churn Prediction

To understand the drivers of churn in greater depth, we conducted a segmentation analysis by visualizing customer behavior and characteristics across various key dimensions. For clarity and structure, this section is organized into thematic subsections, each highlighting specific behavioral patterns and their relationship to churn.

5.4.1.1 Churn by Age, Gender, and Viewing Hours

Age-Gender Group	Viewing Intensity	Churn Rate
Adult Male	Low Viewers	25.74%
Child Male	Low Viewers	24.75%
Adult Female	Low Viewers	24.65%
Child Female	Low Viewers	24.30%
Adult Male	Moderate Viewers	18.33%
Adult Female	Moderate Viewers	17.85%
Child Male	Moderate Viewers	17.72%
Child Female	Moderate Viewers	17.41%
Adult Male	High Viewers	12.44%
Child Male	High Viewers	12.17%
Child Female	High Viewers	11.74%
Adult Female	High Viewers	11.57%

Table 5.2: Detailed churn rates segmented by age, gender, and viewing intensity

The table above provides a detailed breakdown of churn behavior across different demographic and engagement segments. We observe that:

- **Viewing intensity is the primary driver of churn.** While churn rates appear across different age and gender groups, the most significant differentiator is the level of engagement with the platform. Customers classified as *low viewers*—regardless of age or gender—consistently show the highest churn rates, exceeding 24%. In contrast, *high viewers* show significantly lower churn across all demographic segments. This suggests that age and gender have minimal influence compared to behavioral engagement, making viewing activity a more actionable metric for churn prevention.
- **Moderate churn** (orange) is associated with *moderate viewers*, across all age and gender categories, with rates between 17.4% and 18.3%.
- **Low churn** (green) occurs in users with *high engagement*, where all groups stay below 12.5%.

5.4.1.2 Churn by Payment Method and Paperless Billing

Payment Method	Paperless Billing Off	Paperless Billing On
Credit Card	16.34%	**16.17%**
Bank Transfer	17.64%	17.48%
Electronic Check	19.83%	**19.76%**
Mailed Check	18.02%	18.30%

Table 5.3: Churn Rate by Payment Method and Paperless Billing Status

This table explores how payment method and paperless billing adoption jointly influence churn behavior. Several meaningful insights emerge:

- **Electronic check users exhibit the highest churn rates** across both billing scenarios — nearing 20%. This supports the hypothesis that manual or friction-prone payment methods are associated with elevated churn risk. These users may be more likely to experience payment failures or lapses, contributing to involuntary churn.
- **Credit card users consistently demonstrate the lowest churn**, with rates around 16%. This suggests that automated and seamless payment processes contribute positively to user retention. The slight improvement under paperless billing (16.17% vs. 16.34%) reinforces this.
- **Bank transfer and mailed check users fall in the middle**, showing moderate churn levels (approximately 17.5–18.3%). While these methods are less prone to failure than electronic checks, they still require user action or depend on external processing times, possibly creating friction points.
- **Paperless billing, on its own, has minimal impact on churn**. The differences between users who opt in versus those who do not are marginal across all payment methods. This suggests that while operationally convenient, paperless billing is not a significant driver of user retention.

Overall, this analysis highlights that the **choice of payment method is more influential than billing format** in determining churn. Platforms aiming to reduce churn should prioritize promoting secure, automatic payment methods—particularly credit cards and bank transfers—to streamline the subscription experience and reduce service interruptions.

5.4.1.3 Churn by Genre Preference

Genre	Users (Count)	User Rating (Avg)	Churn Rate
Action	48,690	2.998	**16.59%**
Comedy	49,060	3.002	19.34%
Drama	48,744	3.004	17.87%
Fantasy	48,955	3.005	17.67%
Sci-Fi	48,338	3.004	19.14%

Table 5.4: Churn Rate, User Rating, and Genre Popularity

This table examines how users' genre preferences relate to churn behavior. A few key insights emerge:

- **Action content is linked to the lowest churn rate** (16.59%), despite having the lowest user rating (2.998), suggesting that action enthusiasts may have stronger habitual engagement or brand loyalty, even if they rate content less favorably.
- **Comedy and Sci-Fi have the highest churn rates**, above 19%. While their average ratings are similar to other genres (around 3.00), this could indicate that these genres may lack freshness or sustained appeal over time, leading to disengagement.
- **Fantasy and Drama** show more favorable retention patterns, with churn rates under 18%, making them relatively stable content pillars.
- Importantly, user ratings across all genres remain close to neutral-to-positive (between 2.99 and 3.01), indicating that satisfaction alone does not directly drive retention — usage frequency and genre engagement likely play a larger role.

These insights highlight the importance of aligning content strategy not only with popularity but also with churn potential. Reinforcing successful genres like Action, while refreshing and innovating within high-churn categories like Comedy and Sci-Fi, could improve user retention over time.

5.4.1.4 Churn by Content Type and Engagement Level

Content Type	Viewing Segment	Watchlist Size (Avg)	Churn Rate
TV Shows	Low Viewers	12.089	24.14%
Both	Moderate Viewers	12.069	18.46%
Movies	High Viewers	12.058	11.84%
Movies	Low Viewers	12.046	24.37%
Both	High Viewers	12.035	12.43%
Both	Low Viewers	12.015	**26.06%**
TV Shows	High Viewers	11.984	**11.67%**
Movies	Moderate Viewers	11.969	17.45%
TV Shows	Moderate Viewers	11.960	17.57%

Table 5.5: Detailed churn rates by content type, engagement level, and average watchlist size

This table provides a granular breakdown of churn based on users' content preferences and engagement levels:

- The **highest churn** is observed among users who are low-intensity viewers across all content types — especially those who watch both TV and movies at low frequency (26.06%).
- The **lowest churn** appears in users who are high-intensity viewers of TV shows (11.67%) and movies (11.84%).
- Users who consume both types of content and are high viewers also show strong retention (12.43%), reinforcing the importance of content variety and depth of engagement.
- **Watchlist size** remains relatively stable across segments (averaging around 12 titles), indicating that viewing frequency, rather than the volume of saved content, may be a stronger churn predictor.

These insights suggest that promoting broader content exploration and sustaining high engagement levels can significantly reduce churn across viewer profiles.

5.4.2 ARPA and User Satisfaction by Churn Status

Churn Status	**Average User Rating**	**Average Total Charges (ARPA)**
Non-Churned	2.991	$780.40
Churned	**3.057**	$616.74

Table 5.6: User Satisfaction and ARPA by Churn Status

To evaluate long-term user value and perception, we compare **ARPA (Average Revenue Per Account)** and **user satisfaction scores** segmented by churn status. ARPA is a key financial metric that represents the average amount of revenue generated per customer over time. It provides insight into the economic contribution of retained versus lost users.

As shown in Table 5.6, non-churned users generate significantly more revenue on average ($780.40) compared to churned users ($616.74), reinforcing the importance of customer retention from a profitability standpoint.

Interestingly, despite their lower ARPA, churned users report a slightly higher average rating (3.057) than those who remain on the platform (2.991). This counterintuitive result suggests that **user satisfaction is not the sole determinant of churn**. Factors such as payment friction, lack of personalized engagement, or unmet expectations about value may still drive users to leave—even if their experience was rated positively.

These findings underscore the complexity of churn behavior: businesses must consider a broader set of indicators beyond satisfaction scores when designing retention strategies.

To deepen the ARPA analysis, we segmented users by `AccountAge` category and `SubscriptionType`. The table below reveals how revenue and churn vary across these segments:

AccountAge Group	Subscription Type	Average Total Charges (ARPA)	Churn Rate
loyal_customers	Basic	$1,276.11	10.36%
loyal_customers	Premium	$1,274.22	8.07%
loyal_customers	Standard	$1,272.53	9.97%
old_customer	Standard	$925.66	14.04%
old_customer	Premium	$924.09	12.84%
old_customer	Basic	$923.93	15.19%
standard_customer	Basic	$474.99	24.73%
standard_customer	Premium	$473.57	20.51%
standard_customer	Standard	$473.02	23.00%
new_customers	Basic	$81.92	33.68%
new_customers	Premium	$80.83	28.35%
new_customers	Standard	$80.63	31.51%

Table 5.7: ARPA and Churn Rate by Customer Age Group and Subscription Type

The results highlight several critical patterns:

- **Loyal customers** across all plans generate the highest revenue (ARPA ∼ $1270) and have the lowest churn (under 10%), making them the most valuable and stable segment.

- **Old customers** show moderate ARPA values and moderate churn levels, especially in the Basic plan, which may require light retention efforts.

- **Standard customers** have significantly lower ARPA ($470 range) and much higher churn rates (above 20%), indicating vulnerability. Targeted engagement strategies are recommended here.

- **New customers** are the highest risk group—despite generating very low ARPA (below $82), they churn at rates exceeding 30%. These users may not yet perceive value, requiring immediate onboarding improvements or segmentation filtering to avoid attracting non-ideal leads.

This segmentation-driven ARPA analysis enables more strategic prioritization. Specifically, it reveals that some of the users most worth retaining (e.g., “standard customers – Premium plan”) may still be at risk, while others (e.g., “new customers – Basic”) may not be worth retaining unless their early experience improves substantially.

5.4.3 Financial Indicators and Churn Behavior

Churn Status	Monthly Charges (Avg)	Total Charges (Avg)
Non-Churned	$62.58	$780.40
Churned	**$66.34**	$616.74

Table 5.8: Monthly and Total Charges by Churn Status

In Table 5.8, we explore how churned and non-churned users differ in terms of their monthly and cumulative spending behavior.

Although churned users pay slightly more per month ($66.34 vs. $62.58), their total spending is notably lower ($616.74 vs. $780.40). This suggests that they tend to exit the platform earlier in the customer journey, potentially due to unmet expectations or an inability to establish habitual engagement.

The combination of **high short-term cost and low long-term value** may reflect user frustration, affordability concerns, or a perceived mismatch between price and content offering. Identifying users with similar early high-spend, low-engagement patterns could allow for proactive interventions—such as targeted discounts, onboarding improvements, or personalized content suggestions.

Overall, these financial insights reinforce the multifactorial nature of churn: **maximizing lifetime value requires not only high satisfaction, but also sustained engagement, pricing alignment, and operational ease**.

5.5 Performance of Other Predictive Models

To further evaluate the predictive capabilities of different algorithms in the context of churn prediction, we trained and compared four models: **Logistic Regression**, **ExtraTreesClassifier**, **CatBoostClassifier**, and **RandomForestClassifier**. The selection of models was based on their ability to handle class imbalance, interpretability, and overall predictive power. The following table summarizes their key performance metrics:

Table 5.9: Performance Comparison of Predictive Models

Model	**Accuracy**	**Balanced Accuracy**	**Recall (macro)**	**ROC AUC**
Logistic Regression	67.92%	68.41%	68.45%	75.03%
ExtraTreesClassifier	80.31%	57.27%	52.72%	63.97%
CatBoostClassifier	64.33%	66.30%	66.30%	70.47%
RandomForestClassifier	81.35%	51.99%	51.99%	64.94%

5.5.1 Analysis of Logistic Regression

The logistic regression model, which utilizes *ElasticNet* regularization, demonstrates a strong balance between interpretability and predictive performance. Despite having a lower accuracy (67.92%) than ExtraTrees and Random Forest, it achieves the highest recall among all models (68.45%) and an AUC-ROC of 75.03%. These metrics indicate that the model is well-suited for capturing true churn cases while maintaining an adequate ranking capability.

The use of *Stratified K-Fold Cross-Validation* ensures that class imbalance does not significantly affect the model's performance. Additionally, the application of *ElasticNet* regularization helps mitigate overfitting by selecting only the most relevant features. Given its strong recall and AUC-ROC, logistic regression is a robust choice for churn prediction, especially when maximizing the correct identification of at-risk customers is the priority.

5.5.2 Analysis of ExtraTreesClassifier

ExtraTreesClassifier achieves the highest accuracy among the models at 80.31%, but this comes at the expense of balanced classification. The balanced accuracy is significantly lower (57.27%), and recall is the weakest (52.72%), which means that while the model correctly classifies the majority class (non-churners), it struggles to capture the minority class (churners).

The model exhibits characteristics of overfitting, as evidenced by its poor generalization capability on recall and AUC-ROC (63.97%). Despite its strong feature importance analysis and non-linearity handling, it is not suitable for this problem, where identifying churners is critical. Consequently, ExtraTreesClassifier is discarded from further consideration.

5.5.3 Analysis of CatBoostClassifier

The CatBoostClassifier presents a compromise between the previous models. It achieves an accuracy of 64.33%, lower than ExtraTrees and Random Forest but comparable to logistic regression. However, its balanced accuracy (66.30%) and recall (66.30%) are significantly higher than ExtraTreesClassifier and RandomForestClassifier, suggesting that it generalizes better across both classes.

Additionally, its AUC-ROC score of 70.47% indicates a reasonable ability to rank churners correctly. The model leverages gradient boosting with categorical variable support, which enhances its performance in complex datasets. Although it does not outperform logistic regression in recall or AUC-ROC, it remains a viable alternative for further tuning.

5.5.4 Analysis of RandomForestClassifier

RandomForestClassifier achieves the highest overall accuracy (81.35%), outperforming all other models in this metric. However, similar to ExtraTreesClassifier, it struggles with balanced classification. Its **balanced accuracy (51.99%) and recall (51.99%)** are significantly lower than those of Logistic Regression and CatBoostClassifier, indicating that it favors the majority class and fails to correctly capture churners.

The **AUC-ROC score of 64.94%** suggests a moderate ranking ability, but it remains inferior to both Logistic Regression and CatBoost. Furthermore, the model is computationally intensive due to the ensemble nature of decision trees, which may be a limiting factor for scalability.

Although RandomForestClassifier performs well in accuracy, its poor recall and balanced accuracy make it unsuitable for a problem where correctly identifying churners is essential. As a result, it is not considered a viable candidate for final deployment.

5.5.5 Final Model Selection

Given the goal of maximizing churn identification while maintaining model interpretability and generalization, **logistic regression is selected as the final model**. Its superior recall and AUC-ROC make it the most suitable choice, ensuring that a higher proportion of actual churners are correctly classified.

Future steps will involve refining the logistic regression model by further optimizing hyperparameters and exploring feature selection techniques to enhance performance. Additionally, a clustering approach will be implemented to segment customers based on behavioral similarities, facilitating more personalized intervention strategies.

6 Clustering Analysis

6.1 Why Perform Clustering?

While the predictive model provided valuable insights into churn probability, it does not inherently group customers into distinct behavioral segments. Clustering allows us to segment customers based on shared characteristics, helping businesses tailor targeted retention strategies.

The clustering analysis serves three main purposes:

- **Understanding High-Risk Segments:** Identifying groups with higher churn probabilities and the factors contributing to their attrition.
- **Optimizing Customer Engagement:** Detecting patterns in customer viewing habits, payment preferences, and subscription behaviors to improve retention.
- **Feature-Driven Decision Making:** Highlighting key attributes that define each segment, enabling personalized marketing, pricing strategies and others strategies.

6.2 Clustering Methodology

To perform clustering, we utilized **(Uniform Manifold Approximation and Projection)**for dimensionality reduction. Once the customer data was embedded into a lower-dimensional space, we applied a clustering algorithm using a Euclidean distance metric to segment users into meaningful groups.

Unlike traditional methods such as k-means, which requires predefining the number of clusters, our approach allows natural groupings to emerge based on similarities in customer attributes.

6.2.1 Clustering Parameters

The following figure (Figure 6.1) illustrates the resulting customer segmentation after applying the clustering algorithm. In this visualization, data points have been colored according to the binary variable `Churn`, where red indicates churned users and green indicates non-churned users.

This color-coding allows us to visually assess which clusters have a higher or lower concentration of churned customers, helping to identify those clusters that are most relevant for churn analysis. The formation of clusters was based on user behavior, transactional patterns, and engagement metrics, using a dimensionality reduction technique to enable two-dimensional representation. The following parameters where used for the creation of the clusters:

- **Dimensionality Reduction:** UMAP with a Hamming metric and a minimum distance of 0.8, ensuring that local structures are preserved while reducing noise.
- **Embedding Features:** Key variables such as PaymentMethod, PaperlessBilling, DeviceRegistered, UserRating, and ViewingHoursPerWeek were selected to improve clustering accuracy.

- **Clustering Approach:** The embeddings obtained from UMAP were used to cluster customers based on a Euclidean distance metric.
- **Classification Model Integration:** A trained ExtraTreesClassifier was used to generate embeddings before clustering, ensuring that the structure of the clusters aligns with churn probability.

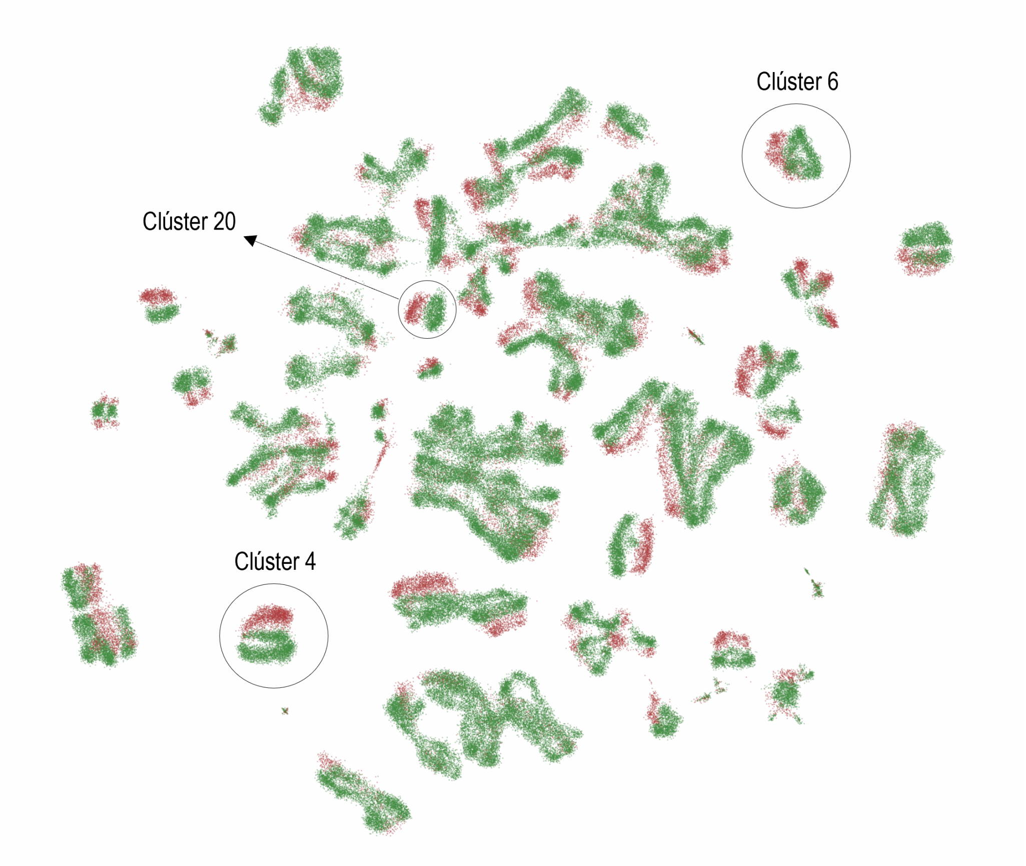


Figure 6.1: Clustering Visualization: Distribution of Customers by Segments

6.3 Clusters Outcomes

6.3.1 Cluster 4: Passive Single-Device Users with Payment Friction

Cluster 4 is characterized by a significantly high concentration of churned users, making it a critical segment for understanding disengagement patterns. These customers share a number of behavioral and transactional traits that increase their risk of leaving the platform.

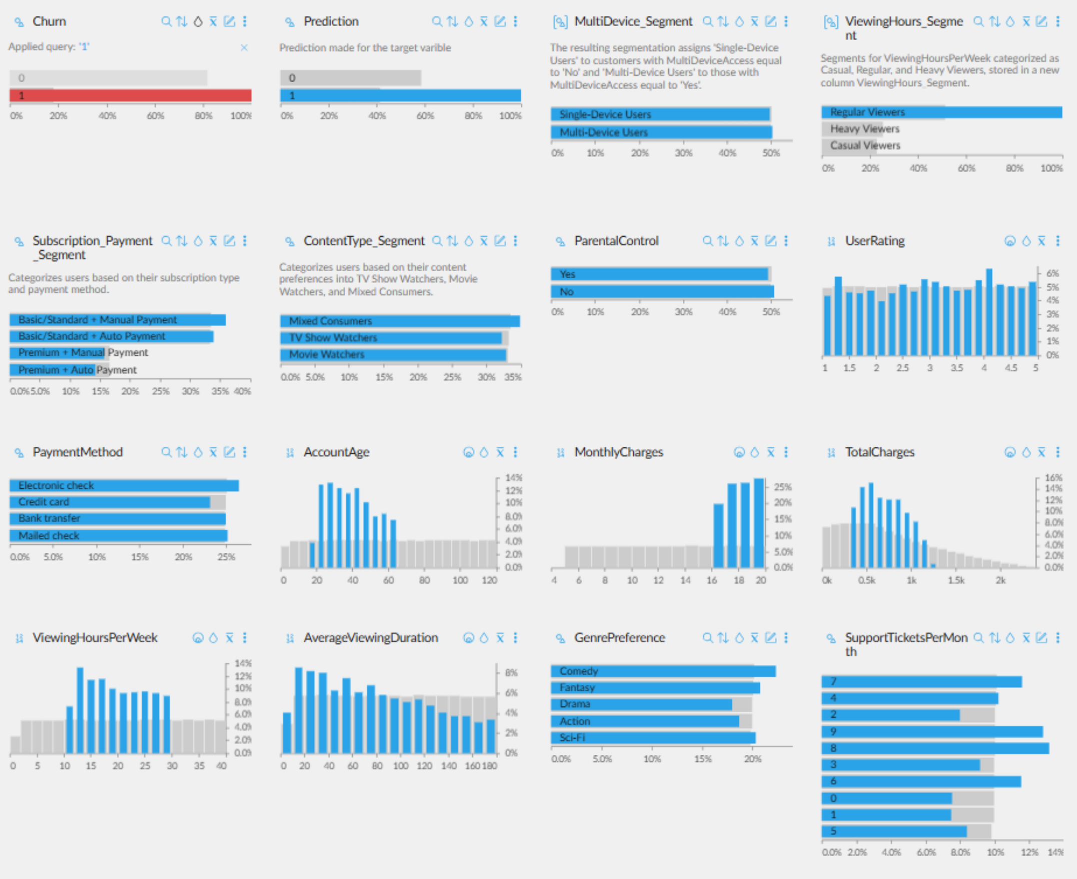
The most defining feature of churned users in this cluster is their **limited engagement**. As seen in Figure 6.2, they tend to consume content from a *single device*, display *moderate viewing intensity*, and have *average watchlist sizes*, which together suggest a more passive interaction with the platform. Their **average viewing durations** are relatively short, and they download fewer pieces of content per month—signaling a lack of immersion in the platform's offerings.

Interestingly, these users are not low-value customers. Many are **high monthly spenders** and show medium to high **total accumulated charges** (see also Figure 6.2), implying strong initial monetization potential. Despite this, they churn early in their customer lifecycle, before consistent usage patterns or loyalty can be established (see `AccountAge` in Figure 6.2).

Payment friction emerges as a major driver of churn in this segment. A large portion of users rely on *manual payment methods* such as electronic checks—methods known for their susceptibility to errors and service interruptions. These payment frictions often result in involuntary churn, as shown in Figure 6.2 (`PaymentMethod` and `Subscription_Payment_Segment`).

Additional red flags include a higher incidence of **support tickets**, pointing to unresolved service issues or dissatisfaction (Figure 6.3). While content preferences (e.g., comedy, fantasy, and sci-fi) are similar to those of loyal users, churners interact with this content less frequently and with less depth (Figure 6.3, `GenrePreference`).

The model outputs in Figure 6.4 confirm the risk profile of this cluster: most users had churn probabilities near or above 0.70 and were correctly classified as churned, demonstrating consistent behavioral patterns that make them easy to identify—but potentially preventable with targeted interventions.



Applied query: '1'

Prediction made for the target varible

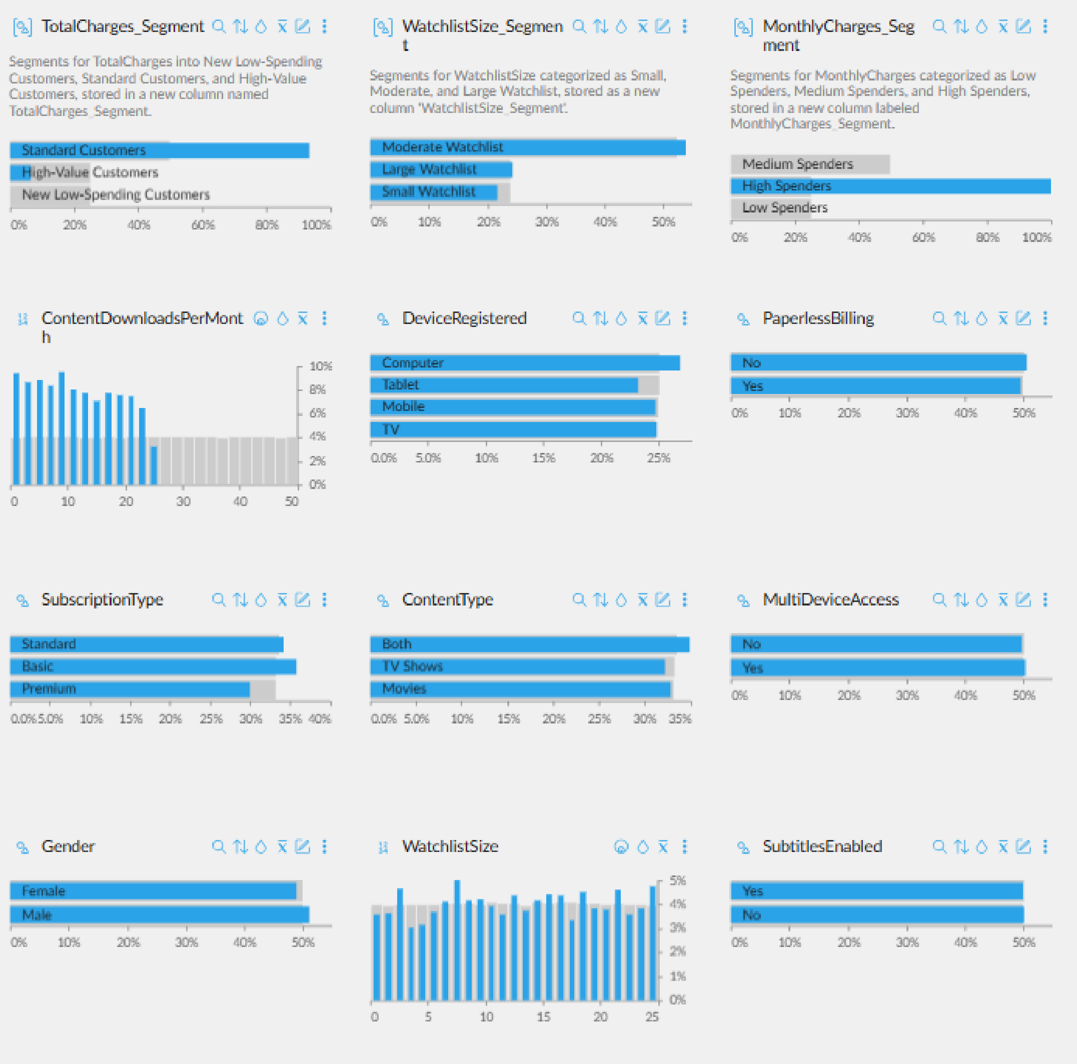
The resulting segmentation assigns 'Single-Device Users' to customers with MultiDeviceAccess equal to 'No' and 'Multi-Device Users' to those with MultiDeviceAccess equal to 'Yes'.

Segments for ViewingHoursPerWeek categorized as Casual, Regular, and Heavy Viewers, stored in a new column ViewingHours_Segment.

Categorizes users based on their subscription type and payment method.

Categorizes users based on their content preferences into TV Show Watchers, Movie Watchers, and Mixed Consumers.

Figure 6.2: Platform access, spending behavior, and payment methods of churned users in Cluster 4



TotalCharges_Segment

Segments for TotalCharges into New Low-Spending Customers, Standard Customers, and High-Value Customers, stored in a new column named TotalCharges_Segment.

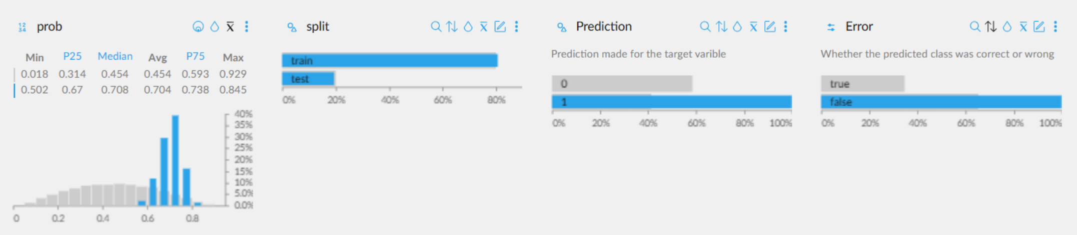
WatchlistSize_Segment

Segments for WatchlistSize categorized as Small, Moderate, and Large Watchlist, stored as a new column 'WatchlistSize_Segment'.

MonthlyCharges_Segment

Segments for MonthlyCharges categorized as Low Spenders, Medium Spenders, and High Spenders, stored in a new column labeled MonthlyCharges_Segment.

Figure 6.3: Content preferences and support interaction of churned users in Cluster 4



prob

Min	P25	Median	Avg	P75	Max
0.018	0.314	0.454	0.454	0.593	0.929
0.502	0.67	0.708	0.704	0.738	0.845

Prediction

Prediction made for the target varible

Error

Whether the predicted class was correct or wrong

Figure 6.4: Churn probability distribution and model classification metrics for Cluster 4 churned users

Summary: Key Traits of Churned Users in Cluster 4

- **Engagement:** Regular viewers with short average viewing durations and low content download activity.
- **Platform Usage:** Single-device users with moderate watchlist sizes.

- **Spending Behavior:** High monthly charges and substantial total charges — valuable but underretained customers.
- **Payment Method:** Predominantly manual (e.g., electronic check), increasing risk of failed transactions.
- **Account Age:** Churn tends to happen early in the user journey — limited time to build habit or loyalty.
- **Support Issues:** Higher support ticket volume, indicating frustration or unresolved problems.
- **Content Preference:** Favor comedy, fantasy, and sci-fi genres, but with less consistent usage.
- **Churn Prediction:** High churn probability and accurate model classification suggest strong signal clarity.

These findings point to clear improvement areas: smoother onboarding, proactive customer support, and reducing payment friction may help retain these otherwise valuable users.

6.3.2 Cluster 6: Low-Spending Multi-Device Users with Broad Preferences

Cluster 6 is characterized by a substantial share of churned users, many of whom demonstrate patterns of light engagement and payment friction. While these customers exhibit positive traits such as cross-device accessibility and genre diversity, their overall behavior suggests limited platform immersion and weak financial commitment.

A defining trait of churned users in this cluster is their **low engagement despite multi-device access**. As shown in Figure 6.5, most users access the platform from multiple devices (`MultiDevice_Segment`), yet their usage remains shallow. They are regular but not heavy viewers, maintain moderate watchlist sizes (`ViewingHours_Segment` and `WatchlistSize_Segment`, Figure 6.6), and download content less frequently.

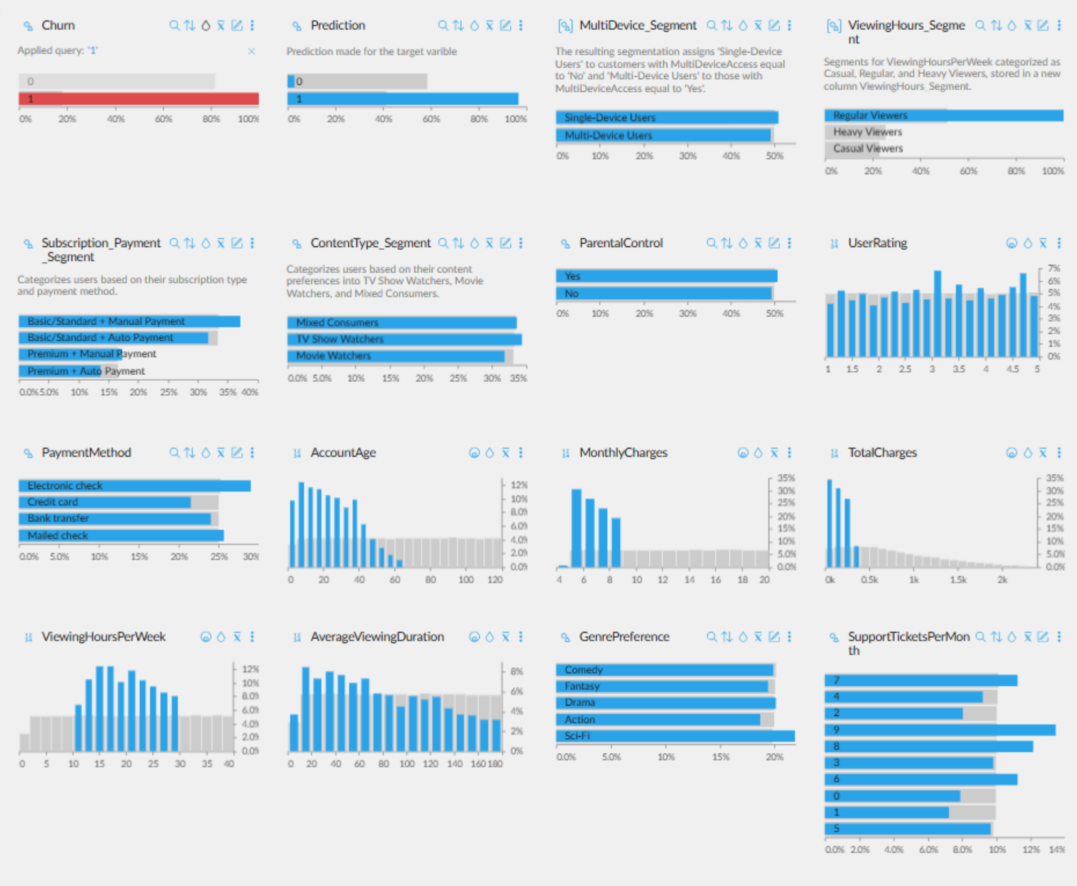
These users are also **low financial contributors**. Figures 6.6 show low `MonthlyCharges` and `TotalCharges`, placing them in the low-spender category. Despite being newer to the platform (`AccountAge`, Figure 6.5), they already exhibit high churn risk.

Payment friction is another major churn trigger. As depicted in Figure 6.5, churned users prefer manual payment methods—especially electronic checks (`Subscription_Payment_Segment` and `PaymentMethod`)—which are prone to failure or delays, contributing to involuntary churn.

Support interactions are high in this group, suggesting unresolved issues or dissatisfaction (`SupportTicketsPerMonth`, Figure 6.6). Moreover, lower `AverageViewingDuration`, fewer `ContentDownloadsPerMonth`, and neutral `UserRating` values further reinforce their weak engagement.

Content preferences, as seen in `GenrePreference` and `ContentType_Segment` (Figure 6.6), are broad, spanning genres like comedy, fantasy, and sci-fi across both movies and TV. Yet, this breadth does not translate into depth of use.

The model's predictions, shown in Figure 6.7, validate this risk profile. Churn probability scores center around 0.62, and most users were accurately predicted as churners, indicating high signal clarity.



Applied query: '1'

Prediction made for the target varible

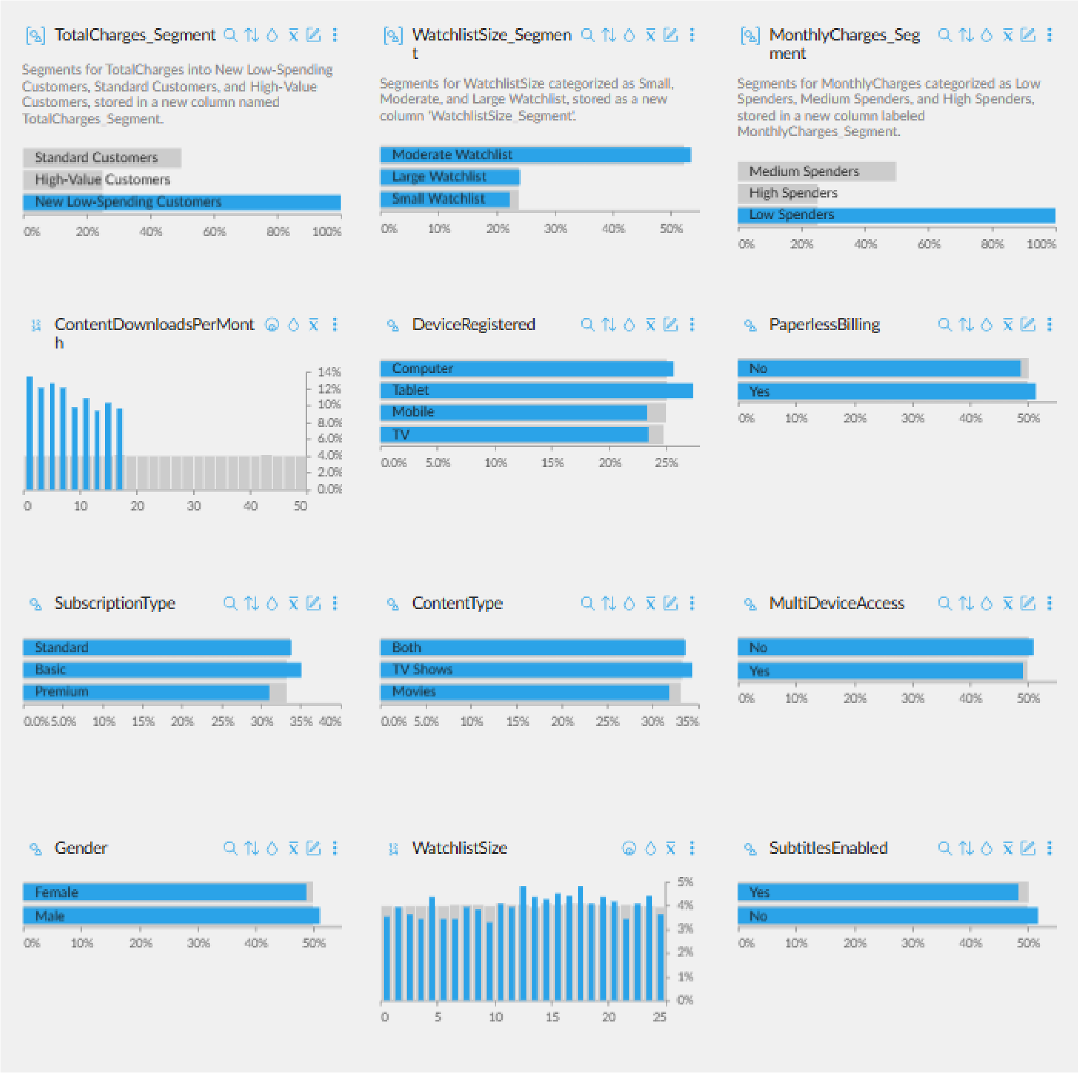
The resulting segmentation assigns 'Single-Device Users' to customers with MultiDeviceAccess equal to 'No' and 'Multi-Device Users' to those with MultiDeviceAccess equal to 'Yes'.

Segments for ViewingHoursPerWeek categorized as Casual, Regular, and Heavy Viewers, stored in a new column ViewingHours_Segment.

Categorizes users based on their subscription type and payment method.

Categorizes users based on their content preferences into TV Show Watchers, Movie Watchers, and Mixed Consumers.

Figure 6.5: User segmentation, device usage, ratings, and payment behavior of churned users in Cluster 6

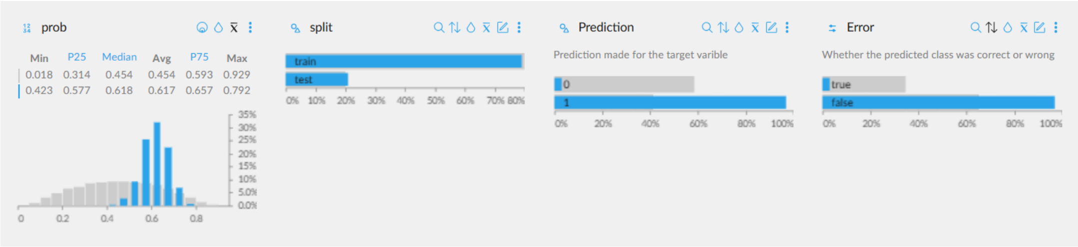


Segments for TotalCharges into New Low-Spending Customers, Standard Customers, and High-Value Customers, stored in a new column named TotalCharges_Segment.

Segments for WatchlistSize categorized as Small, Moderate, and Large Watchlist, stored as a new column 'WatchlistSize_Segment'.

Segments for MonthlyCharges categorized as Low Spenders, Medium Spenders, and High Spenders, stored in a new column labeled MonthlyCharges_Segment.

Figure 6.6: Spending levels, content engagement, and satisfaction indicators of churned users in Cluster 6



Min	P25	Median	Avg	P75	Max
0.018	0.314	0.454	0.454	0.593	0.929
0.423	0.577	0.618	0.617	0.657	0.792

Prediction made for the target varible

Whether the predicted class was correct or wrong

Figure 6.7: Churn probability distribution and model predictions for churned users in Cluster 6

Summary: Key Traits of Churned Users in Cluster 6

- **Engagement:** Regular viewers with short average viewing duration and low content downloads.

- **Platform Usage:** Multi-device users with limited interaction depth.
- **Spending Behavior:** Low monthly and total charges—indicating low financial commitment.
- **Payment Method:** Manual, especially electronic checks—raising churn risk through friction.
- **Account Age:** Typically newer users who churn early.
- **Support Issues:** High support ticket volume—signaling unresolved problems or dissatisfaction.
- **Content Preference:** Broad genre interest, but low engagement across them.
- **Churn Prediction:** Median churn score around 0.62 with high prediction accuracy.

These insights highlight the need for early engagement interventions, frictionless payment systems, and proactive support resolution for retaining users in this segment.

6.3.3 Cluster 20: Friction-Prone Casual Users with High Support Demands

Cluster 20 is characterized by a high concentration of churned users with limited financial value but frequent service interactions and high friction. Although some exhibit multi-device access and broad content preferences, their weak platform engagement and billing issues make them highly vulnerable to churn. This section explores the behavioral differences between churned and non-churned users in this segment and identifies actionable traits for reducing attrition.

6.3.3.1 Key Features of Churned Users

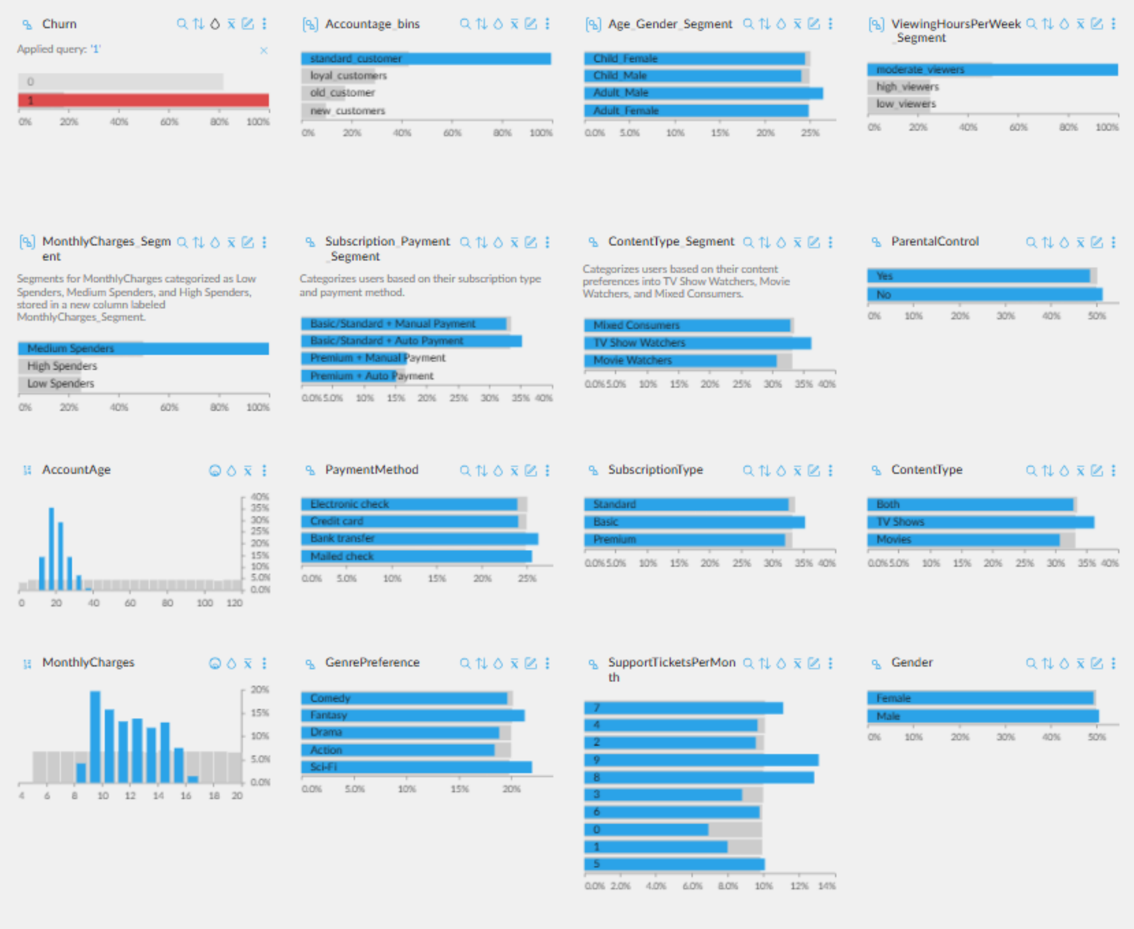
Churned users in Cluster 20 exhibit weak platform commitment, minimal financial contribution, and frequent service interactions. A majority are **casual or regular viewers** with **low-to-moderate weekly engagement**, as shown in Figure 6.8 (`ViewingHours_Segment`, `ViewingHoursPerWeek`). **Average viewing duration** is also low (Figure 6.8, `AverageViewingDuration`).

Most access the platform via **multiple devices** (`MultiDevice_Segment`), with registrations across TVs, computers, and mobile (Figure 6.9, `DeviceRegistered`), but this breadth does not translate into sustained usage.

Financially, they are **low spenders** and new customers, with **monthly charges clustered around $12–$14** and **low total accumulated charges** (`MonthlyCharges`, `TotalCharges`). Manual payment reliance—particularly **electronic checks**—is high (`PaymentMethod`), while **auto-payment adoption remains low** (`Subscription_Payment_Segment`), increasing involuntary churn risk.

Most are **new users** with less than 20 days of activity (`AccountAge`), and while they prefer **TV Shows and Mixed Content** (`ContentType_Segment`) and genres like **Sci-fi and Comedy**, no strong affinity is evident. **User ratings** are widely dispersed and skew low (`UserRating`), while **support demand is high** (`SupportTicketsPerMonth`), suggesting dissatisfaction.

Personalization features like `SubtitlesEnabled` and `ParentalControl` show **low adoption**, pointing to a generic user experience.

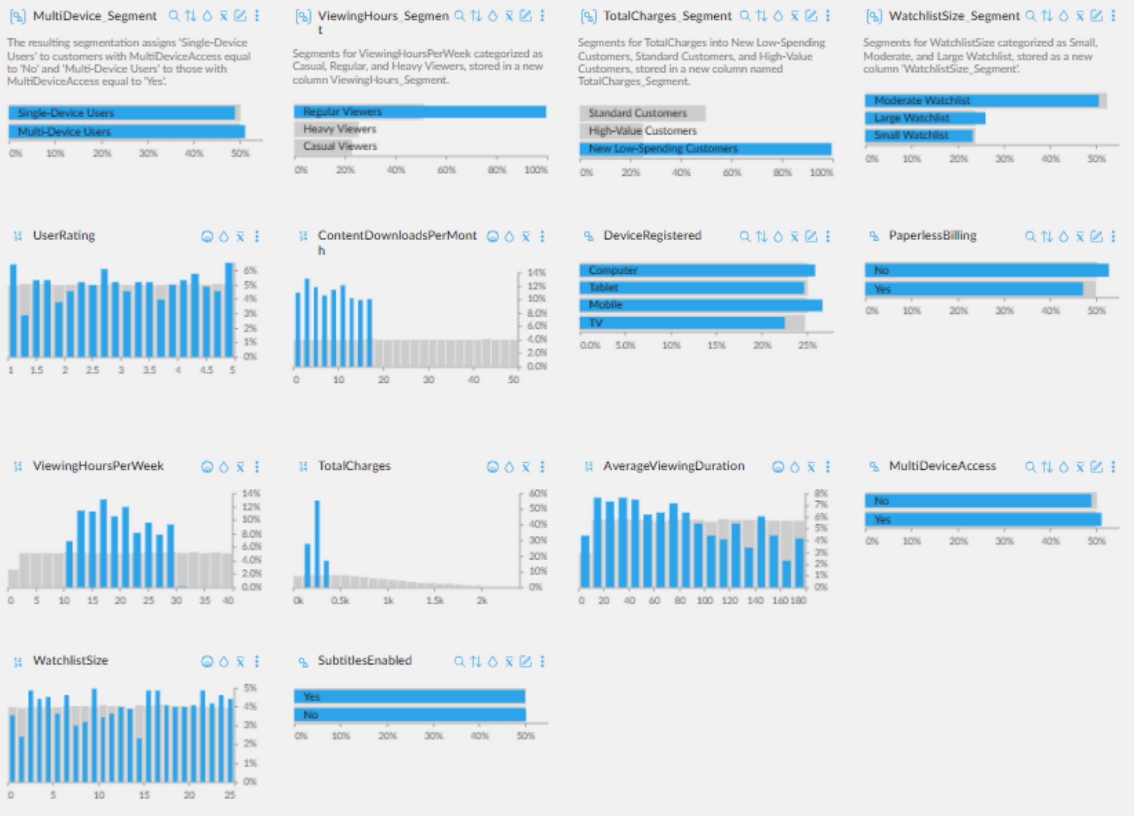


Segments for MonthlyCharges categorized as Low Spenders, Medium Spenders, and High Spenders, stored in a new column labeled MonthlyCharges_Segment.

Categorizes users based on their subscription type and payment method.

Categorizes users based on their content preferences into TV Show Watchers, Movie Watchers, and Mixed Consumers.

Figure 6.8: Behavioral, engagement, and demographic features of churned users in Cluster 20



MultiDevice_Segment

The resulting segmentation assigns 'Single-Device Users' to customers with MultiDeviceAccess equal to 'No' and 'Multi-Device Users' to those with MultiDeviceAccess equal to 'Yes'.

ViewingHours_Segment

Segments for ViewingHoursPerWeek categorized as Casual, Regular, and Heavy Viewers, stored in a new column ViewingHours_Segment.

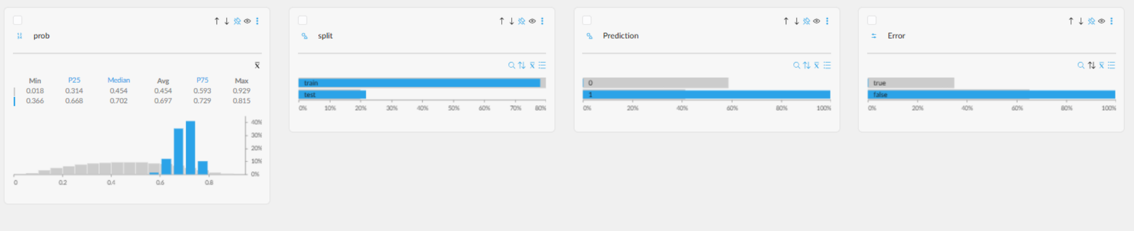
TotalCharges_Segment

Segments for TotalCharges into New Low-Spending Customers, Standard Customers, and High-Value Customers, stored in a new column named TotalCharges_Segment.

WatchlistSize_Segment

Segments for WatchlistSize categorized as Small, Moderate, and Large Watchlist, stored as a new column 'WatchlistSize_Segment'.

Figure 6.9: Content preferences, account data, and satisfaction metrics for churned users in Cluster 20



prob

Min	P25	Median	Avg	P75	Max
0.018	0.314	0.454	0.454	0.593	0.929
0.366	0.668	0.702	0.697	0.729	0.815

Figure 6.10: Prediction probabilities and model evaluation metrics for Cluster 20 churners

Summary: Key Traits of Churned Users in Cluster 20

- **Engagement:** Casual users with low viewing duration and moderate-to-low content interaction.
- **Platform Usage:** Multi-device access but lacking in sustained usage habits.
- **Spending Behavior:** Low monthly and total charges, primarily new customers.
- **Payment Method:** Manual payment dominance—especially electronic checks—leading to higher friction.
- **Account Age:** Short tenure, typically under 20 days.

- **Support Issues:** High volume of support tickets, suggesting dissatisfaction.
- **Personalization:** Low use of features like subtitles or parental control.
- **Churn Prediction:** High predicted churn probabilities with accurate model classification.

7 Large Language Model (LLM) Integration

The final stage of this thesis introduces the integration of clustering results with a Large Language Model (LLM) to derive **automated, segment-specific actionable insights**. This step builds upon the predictive and unsupervised models developed in earlier phases, combining their outputs with the generative reasoning power of modern LLMs.

The process begins with a historical and current database of customer records. Using historical labeled data—where churn outcomes are known—we train a churn prediction model capable of estimating the probability that a customer will leave the platform. This model is then applied on a rolling basis (e.g., weekly or monthly) to the entire active customer base, scoring each user with an up-to-date probability of churn.

From the set of customers identified as **high churn risk**, we apply clustering techniques to group users into behavioral segments. These segments capture different churn patterns, such as low engagement, payment friction, or poor onboarding. Each cluster is characterized by its dominant traits (e.g., content preferences, account age, billing method) and associated average churn probability.

This structured segment information is then passed to an LLM via an automated API integration (e.g., OpenAI, DeepSeek, or other providers). The LLM receives a structured input describing each cluster and is prompted to generate:

- Key reasons for churn specific to that segment
- Personalized, data-driven retention strategies
- Practical business recommendations tailored to operational teams

By combining predictive analytics, unsupervised segmentation, and generative AI, this methodology enables a fully automated pipeline that not only identifies who is likely to churn—but also explains why and suggests what to do about it. The result is a scalable insight-generation engine that can enhance customer retention efforts with precision and speed.

7.1 Large Language Model (LLM) to Analyze Customer Segments and Generate Actionable Insights

This section leverages the behavioral findings from Clusters 4, 6, and 20 to propose targeted strategic actions aimed at reducing churn and enhancing customer retention. Based on the key distinctions between churned and non-churned users, we outline tailored initiatives that address friction points, engagement deficits, and structural weaknesses in the subscription model.

7.1.1 Cluster 4: Passive Single-Device Users with Payment Friction

Cluster 4 is composed of high-spending users who churn early due to payment barriers and weak engagement. These customers show strong revenue potential but exit before loyalty can be developed.

Recommended Actions:

- **Implement default auto-payment onboarding:** Introduce auto-payment as the default method during signup, with incentives for keeping it active.
- **Guided onboarding workflows:** Launch interactive product tours, tutorials, or personalized welcome emails to stimulate early engagement.
- **Single-device reactivation campaigns:** Identify single-device users and offer device syncing tips or cross-device incentives.
- **High-risk trigger system:** Monitor support ticket frequency and payment failures to trigger real-time retention playbooks.

7.1.2 Cluster 6: Low-Spending Multi-Device Users with Broad Preferences

Users in Cluster 6 show solid multi-device access but minimal financial contribution. While churned users are often new and disengaged, non-churners demonstrate stronger usage habits despite low spending.

Recommended Actions:

- **Early value communication:** Push milestone-based messaging during the first 30 days (e.g., "You've watched 5 shows this week!") to build habits.
- **Auto-payment nudging:** Encourage transition from manual to auto-payment methods with discounts or bonus content.
- **Feature discovery campaigns:** Expose underutilized personalization features (e.g., subtitles, parental controls) to increase perceived value.
- **Freemium to upsell pathway:** Develop targeted messaging for engaged users offering light premium upgrades (e.g., "HD + offline viewing").

7.1.3 Cluster 20: Friction-Prone Casual Users with High Support Demands

Cluster 20 includes high-churn, low-engagement users with frequent support issues. While their revenue is modest, the operational cost of supporting them is significant.

Recommended Actions:

- **Support ticket triage automation:** Deploy AI-based ticket categorization to resolve common issues quickly and reduce support overhead.
- **Low-engagement retention loops:** Trigger reactivation sequences when usage drops, offering shortcuts to resume favorite content or explore new genres.
- **Optional usage limits or freemium segmentation:** Consider moving high-support, low-engagement users to a lighter plan or freemium tier with capped support access.
- **Exit surveys + winback campaigns:** Automate surveys at cancellation to diagnose issues and customize winback offers based on the reason.

Cross-Cluster Opportunities: Across all segments, frictionless payment, better onboarding, and engagement tracking emerge as universal levers to reduce churn. Integrating real-time churn signals into the CRM and enabling adaptive playbooks for each cluster can optimize both retention efficiency and customer satisfaction.

8 Conclusions

This thesis set out to analyze customer churn from a strategic perspective, leveraging advanced analytics tools to anticipate its impact and design personalized interventions. Throughout the project, the core objectives have been fulfilled: quantifying the financial implications of churn, identifying behavioral patterns that contribute to it, segmenting users into meaningful clusters, and proposing actionable retention strategies.

One of the major challenges encountered was the nature of the dataset itself. Although it contained a wealth of information on both current and historical users, it proved difficult to isolate decisive patterns that could clearly explain churn behavior. This highlights the multifactorial nature of churn — driven not only by measurable factors like payment method or content consumption, but also by more subjective aspects such as perceived value or user experience.

Moreover, part of the complexity of the analysis stems from the intrinsic challenge of working with anonymized, synthetic, or partially limited datasets in an academic research context. Unlike real business environments — where customer data is richer, more granular, and integrated across touchpoints — public datasets often lack the operational depth required for fully accurate predictive modeling. This represents a natural limitation of the project, but it also reinforces its main contribution: the design and validation of a scalable analytical methodology that, when applied over real customer data in corporate environments, becomes not only highly differential but vastly superior to the current status quo of churn management.

Despite these constraints, the development of the project has showcased the transformative potential of emerging technologies when applied to data analytics. Through the use of machine learning models, clustering algorithms, and integration with Large Language Models (LLMs), we demonstrated how businesses can extract actionable intelligence from large and complex datasets. Specifically:

- A predictive model was developed to identify users with high churn probability.
- Churn-prone users were segmented into behaviorally distinct clusters, enabling more targeted retention strategies.
- The segmentation output was fed into LLMs via API to generate contextualized, automated recommendations to reduce churn.

These insights are not only relevant for subscription-based SaaS platforms but are broadly applicable across diverse business models. Any organization with access to customer data can leverage these analytical tools to better understand its user base, anticipate churn risks, and design personalized interventions.

In conclusion, while the project encountered challenges in identifying clear-cut behavioral signals, it demonstrated that the combination of data analytics and emerging technologies empowers businesses to make smarter, more strategic decisions. This approach opens the door to a future in which retention, customer satisfaction, and business sustainability are driven by data-informed action — and where churn prevention evolves from reactive firefighting to a predictive, proactive, and scalable process.

A Appendix

A.1 Exploratory Data Analysis in Python

This appendix presents the Python scripts used during the exploratory data analysis (EDA) phase. It includes dataset loading, quality checks, visualization of distributions, outlier detection, and correlation analysis.

A.1.1 Initial Data Load and Summary

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

train_data_path = "train.csv"
train_data = pd.read_csv(train_data_path)

print(train_data.head())

def show_info(train_data):
    print("*****␣NAN␣*****")
    print(train_data.isnull().sum())
    print("*****␣Duplicated␣*****")
    print(train_data.duplicated().sum())
    print("*****␣NUNIQUE␣*****")
    print(train_data.nunique())
    print("*****␣INFO␣*****")
    print(train_data.info())
    print("*****␣COLUMNS␣*****")
    print(train_data.columns)
    print("*****␣DESCRIPTIVE␣STATISTICS␣*****")
    print(train_data.describe().T)

print(show_info(train_data))
```

A.1.2 Missing Values and Outlier Detection

```
def missing_data_table(df):
    missing_counts = df.isnull().sum()
    missing_percentage = (missing_counts / len(df)) * 100
    missing_table = pd.DataFrame({
        "Missing␣Values": missing_counts,
        "Percentage": missing_percentage
    }).sort_values(by="Missing␣Values", ascending=False)
    return missing_table

def plot_outliers(df, numeric_columns):
    df[numeric_columns].plot(kind="box", subplots=True, figsize=(15, 10),
                             layout=(2, 5), sharex=False, sharey=False)
    plt.suptitle("Boxplots␣for␣Outlier␣Detection", fontsize=16)
    plt.tight_layout(pad=3.0)
    plt.show()

numeric_columns = train_data.select_dtypes(include=['float64', 'int64']).
    columns
```

```
plot_outliers(train_data, numeric_columns)
```

A.1.3 Distribution and Correlation Analysis

```
def num_var_anal(dataframe, num_cols):
    num_plots = len(num_cols)
    rows = (num_plots // 3) + (1 if num_plots % 3 != 0 else 0)

    plt.figure(figsize=(18, rows * 5))
    for i, col in enumerate(num_cols, 1):
        plt.subplot(rows, 3, i)
        sns.histplot(data=dataframe, x=col, bins=20, kde=True,
                     color="skyblue", edgecolor="black")
        plt.title(f"Distribution␣of␣{col}", fontsize=14)
        plt.grid(visible=True, linestyle='--', linewidth=0.5)
    plt.tight_layout()
    plt.show()

numeric_columns = train_data.select_dtypes(include=['float64', 'int64']).
    columns
num_var_anal(train_data, numeric_columns)

# Skewness and kurtosis
print("Skewness:\n", train_data[numeric_columns].skew())
print("Kurtosis:\n", train_data[numeric_columns].kurtosis())

# Pearson correlation
def pearson_corr(dataframe, num_cols, plot=False):
    print(dataframe[num_cols].corr())
    if plot:
        plt.figure(figsize=[10, 8])
        sns.heatmap(data=dataframe[num_cols].corr(), annot=True, fmt=".3f",
                    cmap="rainbow")
        plt.xticks(rotation=45)
        plt.yticks(rotation=45)
        plt.title("Pearson␣Correlation␣Matrix", fontsize=20)
        plt.show()

num_var = ["AccountAge", "MonthlyCharges", "TotalCharges", "
    ViewingHoursPerWeek",
           "AverageViewingDuration", "ContentDownloadsPerMonth", "
               UserRating",
           "SupportTicketsPerMonth", "WatchlistSize", "Churn"]

pearson_corr(train_data, num_var, plot=True)

# Correlation with target
def analyze_correlation_with_target(dataframe, num_cols, target):
    correlations = {col: dataframe[target].corr(dataframe[col]) for col in
        num_cols}
    for col, corr in correlations.items():
        print(f"Correlation␣between␣{target}␣and␣{col}:␣{corr:.3f}")

    plt.figure(figsize=(10, 8))
    plt.bar(correlations.keys(), correlations.values(), color='skyblue')
    plt.title('Correlation␣between␣Target␣and␣Numeric␣Variables')
    plt.xticks(rotation=45)
    plt.grid()
    plt.show()
```

```
analyze_correlation_with_target(train_data, num_var, "Churn")
```

A.2 Modeling Code in Graphext

This appendix presents the complete preprocessing pipeline and model configuration used in the study. The implementation focuses on customer segmentation, feature engineering, model training, and churn prediction. The scripts were executed using Graphext.

A.2.1 Customer Segmentation and Feature Engineering

```
## Segment customers by account age
segment_rows(ds, {
    "new_customers": "AccountAge:<13",
    "standard_customer": "AccountAge:>=13 AND <64",
    "old_customer": "AccountAge:>=64 AND <85",
    "loyal_customers": "AccountAge:>85"
}) => (ds.customer_segmentation)

configure_column_metadata(ds.customer_segmentation, {
    "label": "Accountage_bins"
})

configure_column_visibility(ds.customer_segmentation, {
    "graph": "pinned"
})

## Segment customers by parental control and gender
segment_rows(ds[["Gender", "ParentalControl"]], {
    "Child_Male": "(ParentalControl: Yes) AND (Gender: Male)",
    "Child_Female": "(ParentalControl: Yes) AND (Gender: Female)",
    "Adult_Female": "(ParentalControl: No) AND (Gender: Female)",
    "Adult_Male": "(ParentalControl: No) AND (Gender: Male)"
}) => (ds.segmentation)

configure_column_metadata(ds.segmentation, {
    "label": "Age_Gender_Segment"
})

configure_column_visibility(ds.segmentation, {
    "graph": "pinned"
})

## Segment viewers by weekly viewing hours
segment_rows(ds, {
    "low_viewers": "ViewingHoursPerWeek:<P25",
    "moderate_viewers": "ViewingHoursPerWeek:>=P25 AND <P75",
    "high_viewers": "ViewingHoursPerWeek:>=P75"
}) => (ds.viewer_segmentation)

configure_column_metadata(ds.viewer_segmentation, {
    "label": "ViewingHoursPerWeek_Segment"
})

configure_column_visibility(ds.viewer_segmentation, {
    "graph": "pinned"
})

## Create binary feature for support ticket volume
cast(ds.SupportTicketsPerMonth, {
    "type": "number"
}) => (ds.SupportTicketsPerMonth)

derive_column(ds, {
    "script": "const supportTickets = row.SupportTicketsPerMonth;\nreturn
        supportTickets != null && supportTickets >= 6 ? 1 : 0;",
    "type": "number"
}) => (ds.SupportTicketsIndicator)

configure_column_metadata(ds.SupportTicketsIndicator, {
    "label": "New_SupportTicketsIndicator"
})

cast(ds.SupportTicketsIndicator, {
    "type": "boolean"
}) => (ds.SupportTicketsIndicator)
```

A.2.2 Tagging Columns and Dataset Configuration

```
configure_tagged_columns(ds[["SupportTicketsIndicator", "Churn", ...]], {
    "Target": ["Churn"],
    "Factors": ["AccountAge", "MonthlyCharges", ..., "
        SupportTicketsIndicator"]
})

configure_category_colors(ds.customer_segmentation, {
    "old_customer": "#356d9d",
    "new_customers": "#7daf9d",
    "standard_customer": "#4b9269",
    "loyal_customers": "#453b7b"
})

configure_category_colors(ds.viewer_segmentation, {
    "high_viewers": "#5fdb5f",
    "moderate_viewers": "#e38f46",
    "low_viewers": "#d73c3d"
})
\end{lstlisting

\vspace{0.5 cm}

\subsection{Model Training and Evaluation}

\begin{lstlisting}
train_classification(ds[["Churn", ...]], {
    "target": "Churn",
    "model": "LogisticRegression",
    "params": {
        "max_iter": 1000,
        "scoring": "accuracy",
        "positive_class": "1"
    },
    "validate": {
        "n_splits": 5,
        "metrics": ["accuracy"]
    }
}) => (ds.gx_prediction, "model-e6")

test_classification(ds[["Churn", ...]], "model-e6", {
    "refit": true,
    "split": {
        "test_size": 0.2
    },
    "target": "Churn",
    "positive_class": "1"
}) => (ds.gx_prediction, ds.prob, ds.Error, ds.split)

configure_column_metadata(ds.gx_prediction, {
    "label": "Prediction",
    "description": "Prediction made for the target variable"
})

configure_column_metadata(ds.Error, {
    "label": "Error",
    "description": "Whether the predicted class was correct or wrong"
})
```

A.2.3 Additional Aggregations and Summary Statistics

```
group_by(ds, {
    "aggregations": [
        {"name": "AGGREGATED_COLUMN-AccountAge-AVG", "on": "AccountAge", "
            type": "AVG"},
        {"name": "AGGREGATED_COLUMN-MonthlyCharges-AVG", "on": "
            MonthlyCharges", "type": "AVG"},
        {"name": "AGGREGATED_COLUMN-SupportTicketsIndicator-COUNT", "on": "
            SupportTicketsIndicator", "type": "COUNT"}
    ],
    "by": ["Churn"]
}) -> (shap_values_model_)

configure_column_metadata(shap_values_model_["AGGREGATED_COLUMN-AccountAge-
    AVG"], {
    "label": "Average of AccountAge"
})

configure_column_metadata(shap_values_model_["AGGREGATED_COLUMN-
    MonthlyCharges-AVG"], {
    "label": "Average of MonthlyCharges"
})

configure_column_metadata(shap_values_model_["AGGREGATED_COLUMN-
    SupportTicketsIndicator-COUNT"], {
    "label": "Count of High Support Users"
})
```

A.3 Clustering Process in Graphext

This section describes the clustering methodology used to group customers with similar churn-related behavior. The process includes training a tree-based model, generating embeddings, applying UMAP for dimensionality reduction, and clustering based on those embeddings.

```
train_classification(ds[["MultiDevice_Segment", "ViewingHours_Segment",
"TotalCharges_Segment", "WatchlistSize_Segment", "MonthlyCharges_Segment",
"viewer_segmentation", "segmentation", "customer_segmentation",
"Subscription_Payment_Segment", "ContentType_Segment", "ParentalControl",
"UserRating", "ContentDownloadsPerMonth", "DeviceRegistered",
"PaperlessBilling", "PaymentMethod", "Churn"]],
{
    "model": "ExtraTreesClassifier",
    "target": "Churn"
}) => (ds.layout_prediction, "ds-layout-model-tXTQ")

embed_with_trees(ds[["MultiDevice_Segment", "ViewingHours_Segment",
"TotalCharges_Segment", "WatchlistSize_Segment", "MonthlyCharges_Segment",
"viewer_segmentation", "segmentation", "customer_segmentation",
"Subscription_Payment_Segment", "ContentType_Segment", "ParentalControl",
"UserRating", "ContentDownloadsPerMonth", "DeviceRegistered",
"PaperlessBilling", "PaymentMethod"]],
"ds-layout-model-tXTQ") => (ds.embedding)

layout_dataset(ds[["embedding"]], {
    "metric": "hamming",
    "min_dist": 0.8
}) -> (ds.x, ds.y)

cluster_embeddings(ds.embedding, {
    "metric": "euclidean",
    "reduce": {
        "algorithm": "umap",
        "metric": "hamming"
    }
}) -> (ds.cluster)

configure_column_visibility(ds.cluster, {
    "graph": "pinned"
})
```

Declaración de Uso de Herramientas de IA Generativa en el Trabajo Fin de Grado

Por la presente, yo, Marta Vara, estudiante del Grado en E6 Analytics de la Universidad Pontificia Comillas, al presentar mi Trabajo Fin de Grado titulado:

"Application of Data Science and Predictive Models for Churn Prevention: Optimizing Customer Retention"

declaro que he utilizado herramientas de Inteligencia Artificial Generativa (IAG), en concreto ChatGPT, únicamente en el contexto de las actividades descritas a continuación:

- Asistente en la comprensión y explicación de conceptos técnicos complejos, relacionados con Business Analytics, Ciencia de Datos y modelos predictivos.
- Apoyo en el entendimiento, análisis y depuración de código en lenguaje Python desarrollado por mí, basado en los materiales, datasets y recursos proporcionados por la Universidad.
- Asistencia en la mejora de redacción, estructura y claridad de determinados apartados del trabajo, siempre manteniendo la autoría, el análisis y las ideas originales propias.
- Traducción de fragmentos del trabajo del español al inglés y viceversa, respetando siempre el contenido y sentido original.
- Revisión y sugerencias para la estructuración y presentación de visualizaciones, tablas y gráficos generados por mí.
- Apoyo en la búsqueda preliminar y localización de artículos académicos, papers y fuentes bibliográficas relevantes, siendo siempre mi responsabilidad la revisión, validación, selección y correcta citación de dichas fuentes.

Afirmo que toda la información, análisis, desarrollo de código, modelos predictivos y contenido presentados en este trabajo son producto de mi esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes. He incluido las referencias adecuadas en el TFG y he explicitado en el propio documento las situaciones concretas en las que se ha utilizado ChatGPT u otras herramientas similares de IAG.

Soy plenamente consciente de las implicaciones académicas y éticas de presentar un trabajo no original, y acepto las consecuencias derivadas de cualquier incumplimiento de esta declaración.

Fecha: 10/04/2025 Firma: Marta Vara

Bibliography

[1] RevPartners. The bowtie funnel: Rethinking revenue operations. *RevPartners Blog*, 2023. Accessed: 2024-11-19.

[2] Patrick Campbell. *The Comprehensive Guide to Churn*. ProfitWell, 2024. Accessed: 2024-11-19.

[3] Bobby Pinero, Sam Rasmussen, and Chris Burgner. The guide to saas metrics. `https://equals.com/guides/saas-metrics/`. Accessed: 2024-10-08.

[4] RevGenius. All you need to know about revops. *RevGenius Magazine*, 2023. Accessed: 2024-11-19.

[5] Carl Gold. *Fighting Churn with Data: The science and strategy of customer retention*. Manning, 2020.

[6] SurveySensum. How nps impacts revenue: The complete guide, 2024. Accessed: 2025-04-02.

[7] Sandra Elizabeth Mena-Clerque and Jorge Aníbal Mena-Clerque. Saas churn calculation: What are different types of churn? `https://userpilot.com/blog/saas-churn-calculation/`, 2024. Accessed: 2024-10-08.

[8] Team Gainsight. Understanding churn and customer segmentation. `https://www.churned.io/knowledge-base/what-is-churn`, 2024. Accessed: 2024-10-08.

[9] Lori Wizdo. Microsegmentation: The missing piece in your customer obsession puzzle. `https://www.forrester.com/blogs/microsegementation-the-missing-piece-in-your-customer-obsession-puzzle`, 2020. Accessed: 2025-04-03.

[10] Awais Manzoor, M Atif Qureshi, Etain Kidney, and Luca Longo. A review on machine learning methods for customer churn prediction and recommendations for business practitioners. *IEEE Access*, 12:70434–70463, 2024.

[11] Miguel A P M Lejeune. Measuring the impact of data mining on churn management. *Internet Res.*, 11(5):375–387, December 2001.

[12] Team Churned. Understanding churn, metrics and how to solve it. `https://www.churned.io/knowledge-base/what-is-churn`. Accessed: 2024-10-8.

[13] Sandra Elizabeth Mena-Clerque and Jorge Aníbal Mena-Clerque. SaaS churn calculation: What are different types of churn? `https://userpilot.com/blog/saas-churn-calculation/`, September 2024. Accessed: 2024-10-8.

[14] Gordon Liu, Yantai Chen, and Wai Wai Ko. The influence of marketing exploitation and exploration on business-to-business small and medium-sized enterprises' pioneering orientation. *Ind. Mark. Manag.*, 117:131–147, February 2024.

[15] Aurélie Lemmens and Sunil Gupta. Managing churn to maximize profits. *Mark. Sci.*, 39(5):956–973, September 2020.

[16] Khushi Lunkad. Average churn rates for subscription services: Data from 8 industries, 2024. Consultado el 5 de abril de 2025.

[17] CustomerGauge. Average churn rate by industry, 2024. Accessed: 2025-04-05.

[18] Churnfree. Average churn rate for subscription services in 2024, 2024. Consultado el 5 de abril de 2025.

[19] Jennifer Clark. Why is the average churn rate for subscription services so high?, 2023. Consultado el 5 de abril de 2025.

[20] Recurly. Churn rate benchmarks, 2024. Consultado el 5 de abril de 2025.

[21] Roland T Rust, V Kumar, and Rajkumar Venkatesan. Will the frog change into a prince? predicting future customer profitability. *Int. J. Res. Mark.*, 28(4):281–294, December 2011.

[22] Rajkumar Venkatesan and V Kumar. A customer lifetime value framework for customer selection and resource allocation strategy. *J. Mark.*, 68(4):106–125, October 2004.

[23] Scott A Neslin, Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H Mason. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *J. Mark. Res.*, 43(2):204–211, May 2006.

[24] Yizhe Ge, Shan He, Jingyue Xiong, and Donald E Brown. Customer churn analysis for a software-as-a-service company. In *2017 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, April 2017.

[25] José Saias, Luís Rato, and Teresa Gonçalves. An approach to churn prediction for cloud services recommendation and user retention. *Information (Basel)*, 13(5):227, April 2022.

[26] Spiros P Gounaris. Trust and commitment influences on customer retention: insights from business-to-business services. *J. Bus. Res.*, 58(2):126–140, February 2005.

[27] Olha Kurinna. B2C SaaS explained - definition, meaning, and examples. `https://www.apptension.com/blog-posts/b2c-saas`. Accessed: 2024-10-30.

[28] Eric Mersch. Managing performance at business-to-consumer (B2C) SaaS companies. `https://flgpartners.com/managing-performance-at-business-to-consumer-b2c-saas-companies/`, October 2022. Accessed: 2024-10-30.

[29] Alenka Lena Klopčič, Jana Hojnik, Štefan Bojnec, and Drago Papler. Global transition to the subscription economy: Literature review on BusinessModel changes in the media landscape. *Manag. Glob. Transit.*, 18(4):323–348, December 2020.

[30] Yana Andonova, Nwamaka A Anaza, and Delancy H S Bennett. Riding the subscription box wave: Understanding the landscape, challenges, and critical success factors of the subscription box industry. *Bus. Horiz.*, February 2021.

[31] Renaissance of the subscription model - C WorldWide asset management. `https://cworldwide.com/insights-news/item/?id=3467`. Accessed: 2025-1-8.

[32] Richard Gedye. Non-subscription revenue as a source of income. *Journal Sales and Marketing Director, Oxford University Press*, 2002. ©Richard Gedye 2002.

[33] Tomasz Makowski. Zara's case study - the strategy of the fast fashion pioneer. *ResearchGate*, 2022. Accessed: 2025-04-05.

[34] IIDE Indian Institute of Digital Education. Business model of ikea – how ikea makes money?, 2024. Accessed: 2025-04-05.

[35] Dave Chaffey. Amazon case study – how amazon makes money, 2023. Accessed: 2025-04-05.

[36] Martin Sanitra and Ziwei Jiang. How to make the freemium subscription-based business model sustainable in a long term? Master's thesis, Copenhagen Business School, Department of Digitalization, May 2019. Academic Supervisor: Xiao Xiao, Associate Professor.

[37] Ralf Schimmer, Kai Karin Geschuhn, and Andreas Vogler. Disrupting the subscription journals' business model for the necessary large-scale transformation to open access. *ScienceOpen Res.*, 0(0), June 2015.

[38] IBM. ¿qué es la retención de clientes?, 2024. Consultado el 6 de abril de 2025.

[39] Stripe. Explicación de la retención neta de dólares (ndr), 2023. Consultado el 6 de abril de 2025.

[40] Maxio. What is net dollar retention, and how can you improve yours?, 2024. Consultado el 6 de abril de 2025.

[41] Marybeth D'Souza and Rohan Gupta. The growth of ndr as a measure of enterprise saas performance, 2021. Consultado el 6 de abril de 2025.