Norwegian School of Economics
Bergen, Fall 2023

Predictive Analysis for Customer Churn in the Credit Card Industry

How do various customer demographic, transactional, and behavioral features influence churn rates in the credit card industry? - A study of applying machine learning techniques to the multifaceted aspects of customer churn.

Oskar Sylte Garnes & Oscar Hexeberg Staveli

Supervisor: Ivan Belik

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

ABSTRACT

This thesis provides a predictive analysis of how various features influence churn rates, based on the dataset from a credit card company. Our research employs binary logistic regression and boosted random forest models, analyzing the features to be consistently linked to customer churn in the credit card industry. We found the following features to be the most impactful: total transaction count, total revolving balance, and number of other products/services with the same bank. Our analysis also revealed that differences in gender played a significant role in churn rates, with distinct trends observed in the churn behavior of male and female customers.

We tested and analyzed five different hypotheses of customer behavior in the dataset. Out of these, we successfully proved the following:

- **Hypothesis 1:** Low credit card usage rates are predictive of customer churn.
- **Hypothesis 2:** A reduction in credit card usage is highly indicative of customer churn.
- **Hypothesis 3:** Customers with a greater number of banking services/products with the same bank are less likely to churn.
- **Hypothesis 4:** Customers with higher months of inactivity, or a higher number of contacts made to the bank are more likely to churn.

Furthermore, the findings of the hypotheses were utilized to develop strategies that effectively address the identified factors, which could lead to improved customer retention in other companies in the credit card industry.

This comprehensive thesis did not contradict any previous literature on customer churn. While some findings aligned with prior research, others provided new insights, especially regarding the influence of multi-product banking relationships and customer engagement levels on churn rates.

The thesis succeeded to convincingly prove some effects and features of churn. However, due to the dynamic and multifaceted aspects of customer churn, the exact causes of what leads to churn remains unknown, emphasizing the need for future studies of factors influencing and causing customer churn.

FOREWORD AND ACKNOWLEDGEMENTS

This master's thesis marks the end of our education at the Norwegian School of Economics. We would like to express our gratitude to Ivan Belik for his guidance in our endeavors. His insights and experience have helped us in structuring the thesis, asking important questions, and giving helpful suggestions throughout the thesis' process.

We also want to express our gratitude to our friends and family for their unwavering support during the entire process of writing.

Furthermore, we extend our appreciation to the faculty and staff at the Norwegian School of Economics for their dedicated teaching and mentorship, which have greatly contributed to our academic and personal growth. The knowledge and skills we have acquired during our time here have been fundamental in guiding our research and shaping our perspectives.

Norwegian School of Economics

Bergen, Fall 2023

Oskar Sylte Garnes

Oscar Hexeberg Staveli

TABLE OF CONTENTS

1. INTRODUCTION

1.1 PROBLEM DESCRIPTION

Churn prediction has gained significant attention within several large markets such as insurance, credit management, banking, telecommunications, SaaS (Software as a Service) for the past 20 years. Attrition, and customer retention which are close synonyms have been around even longer. In recent years, machine learning has helped researchers as a tool to improve models by iterating and letting the computer run through thousands of models to select the most optimal ones. All of this has improved the landscape of churn prediction, the open source making it both viable and easily available for businesses.

Prior research on customer churn has primarily revolved around predicting which customers are at risk. However, there remains a gap in understanding *why* these customers are at risk, determining their retainability, and identifying what incentives effectively reduce churn.

The purpose of this thesis is to add to these gaps in the literature, by studying the factors influencing customer churn using machine learning techniques and using this information to suggest strategies to effectively reduce the overall churn rate and increase customer lifetime value.

Defining Churn

When defining a "churner", it must be relevant to what business you are predicting churning for. One approach is to establish a minimum threshold that a customer must exceed, which could be based on criteria such as the number of orders, order value, or annual activity. However, it is crucial to consider the customer's historical activity. For instance, if the threshold for churn is set at an annual order value of $500 and a customer who usually spends $10 000 per year suddenly drops to only spending $1 000, it is evident they have churned even though their reduced spending still exceeds the $500 threshold.

Previous definitions of churn include Van den Poel and Larivière (2004)'s definition: "someone who closed all his accounts. A customer with no activity.". Buckinx and Van den Poel (2005) "A partial defector is someone with the frequency of purchases below the average and the ratio of the standard deviation of the interpurchase time to the mean interpurchase time above the average.".

To define customer churn, we also need to define customer retention. To keep our study comparable to that of Ascarza, we are adopting their definition of retention (Ascarza, et al., 2017): “First, the central idea that customer retention is continuity – the customer continues to interact with the firm. Second, that customer retention is a form of customer behavior – a behavior that firms intend to manage. Accordingly, we propose that “Customer retention is the customer continuing to transact with the firm.”

Using the definition of customer retention, in which customer churn refers to the opposite of customer retention, it has provided us with a distinct measure of customer churn:

Customer churn is the discontinuity in the customer’s interaction with the firm. It represents a specific form of customer behavior where the customer ceases to transact with the firm. Therefore, the customer churn can be defined as the customer ceasing to transact with the firm. As cited by Ascarza and Gupta, the word “transact” is used due to its broad applicability in describing various forms of interactions between customers and firms. This term not only covers monetary exchanges like subscription fees or online purchases but also extends to non-monetary engagements, such as the usage of free services in digital environments (an e-mail account, social media account). Hence, it effectively encompasses both contractual and non-contractual relationships.

There are several other ways of defining customer churn as is evident by previous research, but simply put, they all involve a customer’s journey ending or slowing down in some way. One could utilize metrics such as subscription cancellations, non-renewals, or decreased usage to gauge churn, depending on the business model and industry.

1.2 Objectives

Extensive research has been conducted on customer churn, highlighting the significance of predicting and understanding customer churn in various industries, including credit card services. Importantly, calculating the duration for which an individual remains a customer is crucial in determining Customer Lifetime Value and Customer Equity. In the credit card sector, churn is primarily identified by account cancellations.

Our thesis aims to address issues in churn management by analyzing the factors driving churn, while merging academic research with practical approaches, providing new perspectives and future research directions in customer churn management. This holds significant potential for

practical application, particularly in the strategic targeting of the most appropriate customer segments, the timing of marketing campaigns, and the selection of effective incentives.

Key proposals of our thesis include:

- Using machine learning techniques to identify causes of customer churn. This framework will focus on a detailed data analysis of the factors leading to customer attrition, based on transactional customer data from a company in the credit card industry.
- Optimal strategies to prevent churn. Based on the insights garnered from our comprehensive churn analysis framework, we aim to formulate and detail optimal strategies for preventing customer churn in the credit card industry
- The design of churn intervention campaigns. This part of the thesis will delve into the specific design elements of effective churn intervention campaigns, merged with our findings and other previous academic research.
- Suggesting future research directions that can further enhance customer churn research and churn management strategies.
- The central theme of our work is the need for a broad and flexible approach to churn management. This includes leveraging transactional data in the credit card industry to analyze the factors of churn, while considering the design and impact of individual churn intervention campaigns and understanding how these efforts align with a company's overall marketing strategy.

1.3 HYPOTHESES

To address the key proposals, we first present five hypotheses which we believe are most pertinent in addressing the core questions of customer churn in the credit card industry. The hypotheses are crafted with the intention of understanding and addressing phenomena observed in the dataset, particularly focusing on customer behavior.

- **Hypothesis 1**: Low credit card usage rates are predictive of churn.

A possible reason for why a customer is leaving is that they do not use the card enough. This can be due to forgetfulness, a habit of using a different card, or that the customer is not aware of the perks of using the card. Even so, the business should investigate card usage both due to income from fees, as well as potential customer churn.

- **Hypothesis 2:** Reduction in credit card usage is highly indicative of customer churn.

This hypothesis posits that there may be a relationship between the number of transaction counts at one time against another and the likelihood of customers churning. A reduction in usage can happen because of changes in life situation, dissatisfaction, change of service provider and more. Importantly, attempting to understand whether churn likelihood decreases usage, or if decreased usage increases churn should be considered.

- **Hypothesis 3:** Customers with a greater number of banking services/products with the bank are less likely to exhibit attrition.

This hypothesis posits that customers who utilize a wider range of banking services and products may be more invested in their relationship with the bank, and hence less likely to leave. These customers may also demonstrate greater loyalty to the company.

- **Hypothesis 4:** Customers with higher months of inactivity, or a higher number of contacts made to the bank are more likely to churn.

This hypothesis suggests that prolonged inactivity or frequent contact with the bank may be indicative of higher churn risk. Both prolonged inactivity and frequent contact can be results of dissatisfaction. If the customer faced a single problem, a lower number of contacts would make sense if it were fixed promptly. Hence, frequent contact can also be due to prolonged problems.

- **Hypothesis 5:** Different age groups exhibit distinct patterns in credit card usage, which in turn impacts customer churn.

This hypothesis suggests that age may be a significant factor influencing credit card usage patterns, such as credit limits, utilization ratios, and perhaps other aspects of credit behavior. For businesses, it can be beneficial to get insights into spending habits of various age groups.

2. BACKGROUND

"…most companies are a lot better at prospecting for new customers than maintaining their customer list. As far as I'm concerned, customer maintenance is imperative to doing business. …Someone once said I behaved as if every IBM customer were on the verge of leaving and that I'd do anything to keep them from bolting."

-Buck Rodgers, IBM

This chapter presents important themes in research on customer retention, churn, and customer relationship management strategies. In the first part, we introduce early research on customer retention. Next, behavioral indicators, customer satisfaction and predictive analytics in context of customer churn will be explained. Finally, ethical and regulatory considerations will be addressed.

2.1 CLIENT RETENTION IN THE CREDIT CARD INDUSTRY

Customer retention has emerged as a significant area of research since the mid-1990s. According to Ang & Buttle, a paper by Dawkins & Reicheld in 1990 "reported that a 5 percent increase in customer retention generated an increase in customer net present value of between 25 percent and 95 percent across a wide range of business environments" (Ang & Buttle, 2004). In the early 2000s companies became more aware of the real value of their existing customer base. Hence, the popularity of customer relationship management grew. Customer attrition, or "churn" is what customer retention strategies seek to reduce (Van den Poel & Larivière, 2001) (Ang & Buttle, 2004), and within the broader scope of customer relationship management concepts, customer churn is of central importance, particularly in terms of customer equity. Additionally, it's a key factor in the customer lifetime value framework, underscoring its role in comprehending and optimizing the long-term value of customer relationships (Ascarza, et al., 2017).

Another report on how large the impact of churners was put forth by Bain in 1990 stating that a reduction in customer defection/attrition "can boost profits from 25% to 85%." (DeSouza, 1992). Similarly, Jamal & Bucklin refer to Gupta et al. that increasing retention 1 % increases firm value by 5 %. A reduction in churn by 5 % has shown to double profits in some industries (Jamal & Bucklin, 2006). The reasons for the increase in profits include a reduction in the cost of marketing and new customer acquisition costs, an increase in customers from existing contacts through referrals and further purchases made by long-term customers. (Van den Poel

& Larivière, 2001; DeSouza, 1992) As an illustrative example of the negative effects the loss of customers might have, a report from J.D. Power and Associates in 1980 stated customers of General Motors, Ford and Chrysler "recorded three times the number of problems with cars 90 days out of the showroom" compared to Japanese cars. 10 years later this decreased to only 25% more problems, but by then the Japanese market had risen eight points (DeSouza, 1992). Eight points in the automobile market is a significant amount.

This example considers a larger scope than what traditional business strategies might encompass. However, identifying and addressing factors contributing to customer attrition is essential. With advanced data and analytic methods, business managers can explore new and specific hypotheses for customer churn. Shini Renjith analyzed the following reasons for customer churn ((Renjith, 2015), shown in Figure 1.

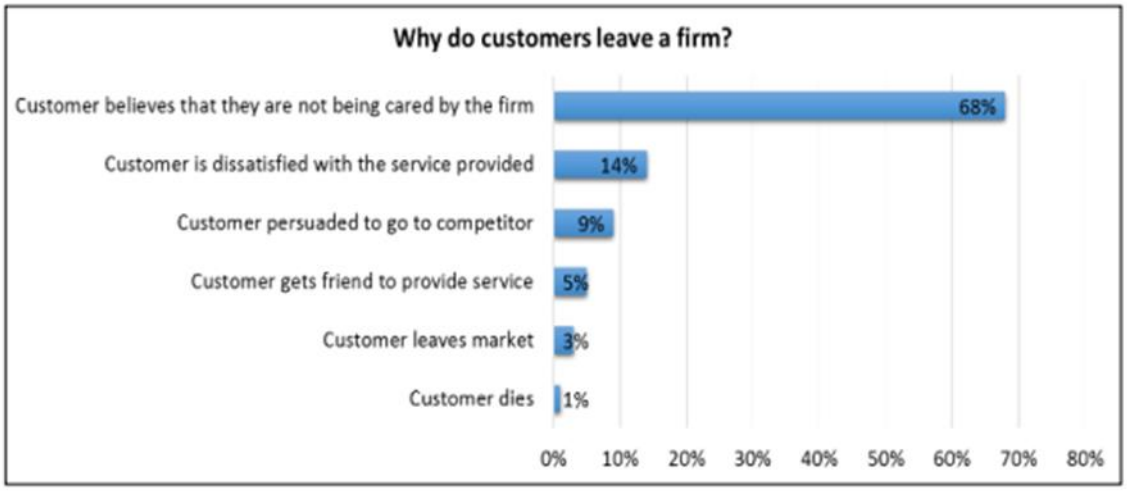


Figure 1 – Reasons for customer attrition.

It is relatively straightforward to understand customers defecting from a company selling defective cars. However, understanding why a customer reduces their use or changes their service provider is more complex.

Prior to retention strategies with advanced data analytics, the focus had to shift to retention strategies over customer acquisition. John W. Gamble, in a 1988 article for Marketing News discussed how service marketing mostly focused on getting new customers. He raises a critical question: "Does your firm have a comprehensive program designed specifically for enhancing relationships with existing customers?" (Gamble, 1998), where he proceeds to present seven questions to consider as a business owner. He illustrates the importance by providing an

example of a health care firm deploying a strategy based on the answers, successfully. In short, he urges the business to find the critical links between them and the customer, what the customer expects from the business and the service, to define key performance indicators (KPIs) specific to the business, to understand competitor's perception of the business, and why accounts are closed. The health care firm saved $60 000 in profits over 24 months. Additionally, a decrease in attrition of 46 % was recorded, saving the company another $400 000 each year in marketing costs to replace said accounts with new ones (Gamble, 1998).

DeSouza emphasizes the importance of understanding the "why" for customers who churn, before creating a retention plan. Managers might make assumptions, but their efforts to create actionable strategies will be better guided with a clearer understanding of the problems customers face. His article does, supporting Gamble's work, urge businesses to create KPIs that are reasonable to track in relation to the business but also to identify specific business activities that lead to customer attrition, for example through mediums like interviewing former customers. This way managers can get insights into whether customers churn because of price, product, market, service, technology or the organization (DeSouza, 1992).

In 2006, Neslin, Gupta et al. reviewed the predictive accuracy of churn prediction models, using techniques such as logistic regression and boosted decision tree models. In the study, the researchers hosted a tournament, where the findings suggested that boosted decision trees was among the more efficient methods for predicting churn. This highlighted the importance of choosing the right method when creating predictive models. Additionally, the study regarded the "staying power" of the models, referring to how long a model stayed useful for the business before it should be reviewed. Their findings indicated that the model should work for three months (Neslin, Gupta, Kamakura, Lu, & Mason, 2006).

Later, in 2016 Ascarza et al. conducted a review aiming to improve customer retention management. Among the key insights of this paper was the emphasis on restructuring the characterization of churning customers. The prior standard was a binary categorization of the customer as either attrited or existing, while Ascarza et al. proposed categorizing customers as "at risk", yet not necessarily as targets for retention campaigns. The study assesses the tradeoff between targeting a customer at risk or not, as well as tradeoffs between proactive and reactive retention campaigns. Finally, the authors suggest focus areas for further enhancement of retention management. Ascarza et al. studied improvements in identifying customers at risk of

churning, and made suggestions as to additional data businesses could include for improving the models. Currently, the models are effective in predicting customer churn. However, they lack the ability to explain the reasons behind churn, which is the next issue to address. Predictors of customer churn, such as demographics and financial aspects of the customer relationship, provide limited insight into the reasons behind the churn. Additionally, there is discussion about whom to target with retention campaigns. Early methods assessed a customer's risk of churning and implemented retention strategies at a certain threshold. However, it is also important to consider factors such as customer profitability and the likelihood of a positive response to the campaign (Ascarza, et al., 2017).

2.2 Customer Satisfaction as a Predictor of Churn

In all business sectors, brand and category expectations are a vital part of customer satisfaction, which in turn can affect client retention (Gupta & Stewart, 1996). This suggests incorporating data on customer satisfaction to model churn prediction enables improved accuracy and understanding of customer defection. However, not all businesses have access to historical data on customer satisfaction. Ascarza et al. stated, in a paper from 2017, that usage trends and contact can be used as predictors with basis in customer satisfaction (Ascarza, et al., 2017).

According to Ascarza et al., social connectivity has a clear negative effect on churn, likely a result of influence from friends and family. Services with higher social connectivity experience a higher likelihood of retention. For instance, in the context of an online gaming platform, or a social network, if friends switch provider, especially a highly connected user, the likelihood of churn among their contacts increases (Ascarza, et al., 2017).

Addressing the topic of complaints, Ron Zemke stated in an article from 1990 that businesses that respond to complaints retain 30 % more customers than those who do not. Additionally, he states 25 % of customers are unhappy, but only 4 % complain. Although a brief article, the critical impact of effective responses to complaints is underlined. Furthermore, improvement on complaint handling can provide insights and lead to retention, from simple measures such as a phone call. (Zemke, 1990; DeSouza, 1992)

In the discussion on customer satisfaction, Ascarza et al. suggests behavioral indicators such as transaction frequency, payment behaviors, interactions with customer service can be potent in predicting churn. Understanding and mapping of behavioral patterns can effectively assist

in designing retention management plans, and determining which customers to target (Ascarza, et al., 2017).

2.3 Predictive Analytics

Researcher in customer churn has long utilized hazard models, and from Ascarza et al. it seems that boosted decision trees are among the more effective models for finding predictors of churn (Ascarza, et al., 2017; Glady, Baesens, & Croux, 2006). Determining which customer to target is crucial. In the context of prediction outcomes, the model will predict either attrition or retention based on a customer's data. The prediction can be correct or incorrect. This raises another critical consideration for the business. Which is more costly? The optimal choice for the business might vary, and what the management decides can affect future profits and retention campaign results. Ultimately, business managers must consider potential harmful effects of their priorities. (Ascarza, et al., 2017).

Achieving an optimal balance is crucial, although choosing a suitable model can save the business thousands of dollars in comparison to a less capable model (Neslin, Gupta, Kamakura, Lu, & Mason, 2006). The ability to design retention strategies using precise predictions and with clear definitions of which customers to target increases the likelihood of cost savings and increased future profits. (Neslin, Gupta, Kamakura, Lu, & Mason, 2006; DeSouza, 1992; Ascarza, et al., 2017).

In addition to developing a precise and balanced model, Jamal & Bucklin proposes the incorporation of heterogeneity in the customer base into the hazard models. (Jamal & Bucklin, 2006) In their report, Jamal & Bucklin discuss the possibility of segmentation, utilizing differences between individuals, and the predictive assets that can potentially be revealed in specific traits. This puts additional pressure on churn prediction modelers and retention campaign designers as they must consider the business, market, behavior of customers in different segments and effectiveness of campaigns towards the different segments. (Jamal & Bucklin, 2006)

An additional observation from Jamal & Bucklin's paper is their conclusion that conventional demographic variables such as age, gender, marital status and housing does not seem to be predictive of churn. They report that this is consistent with "long recognized limitations of demographic variables in predicting differences in consumer purchase behavior." (Jamal & Bucklin, 2006) However, it is important to note that this is a study on a direct-to-home satellite

TV provider. Other sectors, such as the credit card market, could potentially yield differing results.

2.4 Organizational Strategy

Transitioning from prediction models, this section aims to explore the strategies developed within the organization using predictive analytics. Previous sections clearly describe the importance of reducing churn, as well as estimating how impactful it can be. In addition to simple profits lost from churning customers in the form of fees, credit card companies can lose other important aspects connected to the customer equity of the firm. This includes but is not limited to network effects, brand recognition, future customer referrals, and consistent data from loyal customers. (Ascarza, et al., 2017)

In the study from Glady et al., the researchers advance beyond the simplistic binary classification of customer churn. They incorporate ideas from Ascarza et al. about the different types of customers. For instance, a customer can be "walking dead" where they cease transaction with the firm without informing them, or a hybrid version where the customer might formally or informally note the business of their departure. In the context of the credit card industry, a decline in transaction count can be an indicator of one of these categories of customers, which can be a focal point for intriguing analysis and potentially strategic targeting (Glady, Baesens, & Croux, 2006; Ascarza, et al., 2017).

Translating insights from analytics into actionable strategies can be challenging, however it is crucial for improved decision-making (Gamble, 1998; Neslin, Gupta, Kamakura, Lu, & Mason, 2006). As DeSouza claims, understanding why customers cancel your service is key to creating targeted strategies (DeSouza, 1992). The critical question arises: How can managers integrate these insights with data from a single company? Gaining insights into whether customers are switching to a different company or completely canceling the service presents a significant challenge.

While improving a model can yield improved results, attempts at improving the model will eventually lead to diminishing returns. Understanding the link between the predictive characteristics and customer retention is the key. Based on what predictive characteristics are found, the market the business operates in, the position of the company in relation to competitors and the brand profile, the causality of churn can vary greatly. (Ascarza, et al., 2017) A theoretical model with 100 % accuracy, where the management gains complete insights into

who is at risk of churning, would still require the managers to understand the “why” to create strategies and actionable plans.

An article from the Journal of Marketing Research by Lambrecht and Skiera in 2006, illustrates the necessity of the “why” through their work on flat-rate versus pay-per-use bias. Although the study is outside the business model of credit cards, it is closely linked to strategic choices especially within the topic of churn. The flat-rate bias, the that customers tend to pay more to ensure they cover a monthly cost to avoid unpleasant surprises should they exceed their spending limit. A concern was that the increased price customers pay due to overspending on a flat rate subscription would lead to higher churn, however this did not seem to be the case. Over time, the choice of using a flat rate for the business did not significantly increase churn, and instead increased short- and long-term profits. (Lambrecht & Skiera, 2006) The insights from their article highlights why understanding customer behavior improves the design of retention strategies.

Another early study by Engwall highlights another important aspect of moving from data to action. This study revolves around the response time of organizations, where the conclusion is “the main impression … is that prompt responders are more dependent on the market than slow responders” (Engwall, 1976). Since shorter response time, particularly with quality responses, is correlated with higher customer satisfaction which again is correlated with lower churn, this might be a useful insight for businesses when creating strategies. (Hennig-Thurau & Klee, 1997) However, more importantly might be the understanding that churn could stem from the organization or the market dynamics themselves. (Engwall, 1976)

Going back to Ascarza et al. they propose a framework for managing customer retention. This framework includes data and methods for predicting churn, and then suggests designing individual campaigns and coordinating multiple campaigns. Additionally, they suggest integrating retention efforts with marketing strategy. Regarding the campaigns, Ascarza et al. suggests distinguishing between reactive campaigns where customers have already left, and proactive+ campaigns, targeting the customers that are predicted to churn. (Ascarza, et al., 2017)

This is where Jamal & Bucklin’s paper on improving the diagnosis and prediction of customer churn provides interesting conclusions. According to them, the importance lies in finding useful triggers that managers can influence. Furthermore, the difference in customer response

to retention campaigns should be considered. On improvement, Jamal & Bucklin suggest incorporating information on how the customer was acquired in the model. Acquisition data enables the model to consider changing service conditions for individual customers as a predictor. For example, a customer who is acquired through a tempting offer then later closes their account when the price is adjusted. (Jamal & Bucklin, 2006).

In a news article on a churn management system called ChurnManager, line of business manager for decision support systems Rick Kehoe claims that gaining insights into why customers churn provides the staff with information to improve their overall performance, while the organization can improve their customer service (Meyers, 1996). Involving the staff is the important insight from his statement. How will the implementation and success of a retention campaign affect the organization? There is consensus that reducing churn increases profits, both through saved profits from defecting customers, and through future earnings from referrals (Ascarza, et al., 2017; Van den Poel & Larivière, 2001; Ang & Buttle, 2004; Ahmad & Buttle, 2002; Neslin, Gupta, Kamakura, Lu, & Mason, 2006).

Finally, enhancing customer retention management is linked to better customer relationships. This involves shifting from a product-focused approach to a customer-centered one, recognizing that customers inherently boost the return on investment (ROI) the longer they remain with the company. Effective churn management plays a crucial role in this process, as it enables managers to allocate business resources more efficiently (Ascarza, et al., 2017).

2.5 Ethical and Regulatory Considerations

The processing and analysis of customer data in the credit card industry are pivotal activities. However, these activities are subject to a range of ethical and regulatory considerations that must be rigorously adhered to. This subchapter delves into the key aspects of these considerations, illuminating the critical balance between data utility and the protection of individual rights and privacy.

The foremost consideration is adherence to data privacy and protection laws such as the GDPR, CCPA, and other local regulations. These laws provide a framework for the lawful collection, processing, storage, and sharing of personal data. They are designed to protect the rights of individuals while enabling organizations to utilize data responsibly. Non-compliance not only poses legal risks but can also lead to significant reputational damage.

Protecting customer identity is crucial. Anonymization and pseudonymization techniques are pre-employed to ensure that individuals cannot be directly or indirectly identified from the dataset. This practice is essential for maintaining privacy and reducing the risk of harm in the event of a data breach.

Ethical data practice dictates that only the data necessary for the specific analysis should be collected. This minimization reduces the scope for misuse and unauthorized access, thereby safeguarding individual privacy.

Thomas Reid, a prominent philosopher within the commonsense philosophy dating back to the 18th and 19th century, although with roots back to Aristoteles, comments on the inert knowledge humans have of what is good. A rooted sense of what is right, or an in-built moral compass. Amongst this, he gives an example of the golden rule, stating it as one of the "self-evident duties to others; the other duty is that people should act to benefit the society of which they are a part." (Burton, 2008). Burton argues that commonsense morality is a topic that could provide value to businesses that consider the questions raised by the philosophy, such as how the philosophy can affect business decisions (Burton, 2008).

Exploring the topic of commonsense morality, another ethical concern can be considered, the social responsibility of the business. As a credit card provider, certain customers might be prone to mishandling the credit card they are issued, placing themselves in a difficult financial position. Depending on where the business is operating, different laws and directions can influence how the business operates or is allowed to operate.

The authors of this paper note a limitation in their information regarding the geographic origin of the business data, preventing them from addressing region-specific regulations. Nevertheless, should the strategies discussed adversely affect vulnerable customer segments, the authors advice a thorough ethical consideration, regardless of the regional regulations.

3. DATA

3.1 DATA COLLECTION

For our research study, we have exclusively utilized secondary data from https://leaps.analyttica.com (accessed on 10th October 2023). Our purpose in doing so was to make it reproducible, while maintaining a clear framework that would allow for consistency. Employing a standardized dataset for whence others can reproduce the results under the same conditions was an important aspect to establish credibility. This will ensure subsequent research to be comparable to our findings. The selection of this dataset was based on a thorough literature review and preliminary data exploration.

The data collected represents 10 127 customers in the credit card industry, with a total of 1 627 churned customers. The customers have one of four credit card types: blue, silver, gold or platinum. When the customers cancel their credit card or change to a different bank, they are categorized as attrited customers. As a result, attrited, or "churned" customers lead to a decline in the bank's profits. This has prompted banking professionals to develop an early-warning system to identify potential churn and non-churn customers. Such a system would alert the bank's management, enabling them to engage with customers likely to churn. By enhancing their services, they aim to ensure customer satisfaction and retain their clientele.

Variables

It's imperative to establish a clear understanding of the variables that form the foundation of our study. These variables, of which one is dependent and 19 are independent, align with our research objectives and provide a comprehensive view of the factors influencing customer behavior. By defining each variable and its role, we aim to set the stage for the subsequent analytical processes and discussions.

To ensure clarity and precision in our analysis, we've provided detailed descriptions of each variable below. This will aid in understanding the nuances and significance of each factor as it relates to customer behavior. Understanding the dynamics of these variables is crucial, not only for the robustness of our study but also for its applicability in real-world scenarios. The interplay between these variables can offer insights into the multifaceted nature of customer behavior.

The dependent variable:

- **Attrition_Flag:** (Customer Churn) This binary variable indicates whether a customer has left the service or stayed. A value of '1' denotes that the customer has churned, while '0' signifies retention.

Independent variables:

- **Customer_Age:** The customer's age in years.
- **Gender:** The customer's gender, where M=Male and F=Female.
- **Dependent_count:** The number of dependents.
- **Education_Level:** Educational background of the account holder (Unknown, uneducated, high school, college, graduate, post-graduate).
- **Marital_Status:** The marital status of the customer (Unknown, single, married, divorced).
- **Income_Category:** The annual income category of the account holder (<$40K, $40K - $60K, $60K - $80K, $80K - $120K, $120K +).
- **Card_Category:** The type of credit card (Blue, silver, gold, platinum).
- **Months_on_book:** The duration of the customer relationship with the bank in months.
- **Total_Relationship_Count:** The total number of bank products held by the customer.
- **Months_Inactive_12_mon:** The number of months inactive in the last 12 months. Specifies the duration for which the customer has not used the credit card.
- **Contacts_Count_12_mon:** The number of times the customer has contacted the bank in the last 12 months.
- **Credit_Limit:** The credit limit on the credit card. The maximum amount a credit account can be charged.
- **Total_Revolving_Bal:** The total revolving balance on the credit card. This is the amount that can be carried from one month to the next with interest. It represents the portion of the credit that has been utilized and not paid off in full.
- **Avg_Open_To_Buy:** The average open to buy credit line of the last 12 months. It refers to the amount of available and unused credit, which could have been used for purchases.
- **Total_Amt_Chng_Q4_Q1:** The variable represents the percentage wise change in transaction amount from the first quarter to the last quarter. The transaction amount of Q4 is divided by the transaction amount of Q1.
- **Total_Trans_Amt:** The average total transaction amount of all the transactions in the last 12 months.

- **Total_Trans_Ct:** The total count of transactions the last 12 months.
- **Total_Ct_Chng_Q4_Q1:** The total percentage wise change in the number of transactions from Q4 to Q1. The number of transactions in Q4 is divided by the number of transactions in Q1.
- **Avg_Utilization_Ratio:** The average percentage of which their credit card has been utilized. The total revolving balance on the credit card is divided by the total credit limit on the card.

3.2 DESCRIPTIVE STATISTICS

In this section, we present a concise overview of the descriptive statistics derived from our dataset, shedding light on the nuances of our customer data and its inherent characteristics. By examining these basic statistical measures, we can glean initial insights, identify potential anomalies, and set the stage for more intricate analyses.

Table 1

Table I. Data Description for Credit Card Customers Dataset

Table 1 presents descriptive statistics of the customer demographic, financial status, and service engagement variables analyzed in this study. The table reports the following values: number of observations for each variable (N), mean value (Mean), and standard error (SD). On the right-hand side, distribution values are reported for the 10th, 50th, and 90th percentile. The dataset includes comprehensive data of credit card customers, capturing essential details and behavior indicators. The demographic, financial, and service engagement variables are reported as is and are not winsorized nor trimmed in any other way.

Variable	N	Mean	SD	Distribution		
				10th	50th	90th
Independent Variable						
Attrition_Flag	10127	0.1606	0.3672	0	0	1
Factors:						
Demographic Variables						
Customer_Age	10127	46.33	8.02	36	46	57

Gender	10127	0.53	0.50	0	1	1
Dependent_Count	10127	2.35	1.30	1	2	4
Education_Level_1	10127	0.146	0.354	0	0	1
Education_Level_2	10127	0.198	0.399	0	0	1
Education_Level_3	10127	0.100	0.300	0	0	0.4
Education_Level_4	10127	0.308	0.462	0	0	1
Education_Level_5	10127	0.051	0.220	0	0	0
Education_Level_6	10127	0.044	0.206	0	0	0
Marital_Status_1	10127	0.389	0.487	0	0	1
Marital_Status_2	10127	0.463	0.499	0	0	1
Marital_Status_3	10127	0.074	0.262	0	0	0
Income_Category_1	10127	0.177	0.381	0	0	1
Income_Category_2	10127	0.138	0.345	0	0	1
Income_Category_3	10127	0.152	0.359	0	0	1
Income_Category_4	10127	0.072	0.258	0	0	0
Income_Category_5	10127	0.110	0.313	0	0	1
Product and Service Variables						
Card_Category_1	10127	0.055	0.228	0	0	0
Card_Category_2	10127	0.011	0.106	0	0	0
Card_Category_3	10127	0.002	0.044	0	0	0
Total_Relationship_Count	10127	3.81	1.55	2	4	6
Months_on_book	10127	35.93	7.99	26	36	46
Months_Inactive_12_mon	10127	2.34	1.01	1	2	3
Contacts_Count_12_mon	10127	2.45	1.11	1	2	4
Financial Variables						
Total_Revolving_Bal	10127	1163	814.99	0	1276	2228
Avg_Open_To_Buy	10127	7469	9091	683	3473	21965
Total_Trans_Amt	10127	4404	3397	1501	3899	8212
Total_Trans_Ct	10127	64.85	23.47	33	67	92

Total_Ct_Chng_Q4_Q1	10127	0.71	0.24	0.45	0.70	0.95
Total_Amt_Chng_Q4_Q1	10127	0.76	0.22	0.53	0.74	0.99
Avg_Utilization_Ratio	10127	0.274	0.275	0	0.176	0.707
Credit_Limit	10127	8632	9089	1762	4549	23400

From table 1, we can gauge the complete descriptive statistics of each variable and distinct factor. The table provides information on various demographics, product and service, and financial variables. Reviewing the variables, without having trimmed the tails, the median values (50th percentile) vary from the mean values for many variables. For instance, considering the Customer_Age variable, the mean age is 46.33, while the median age is 46 - both values are closely aligned. However, for some other variables like Total_Revolving_Bal, the mean and median values differ considerably.

Addressing the distinction between variables, there is a noticeable difference between demographic factors such as Education_Level and Income_Category. For example, the distribution among different education levels is varied, with Education_Level_4 having a mean value of 0.308, which is the highest among all education levels. Similarly, marital statuses also show diverse distributions.

Furthermore, when examining the financial variables, some show significant dispersion. Consider the Total_Revolving_Bal, which has a mean of 1163, while wider spread up to 2228 as its 90th percentile. Similarly, Credit_Limit has a mean of 8632, but its distribution goes up to 23400 at the 90th percentile. This suggests that some customers have a considerably higher credit limit compared to others.

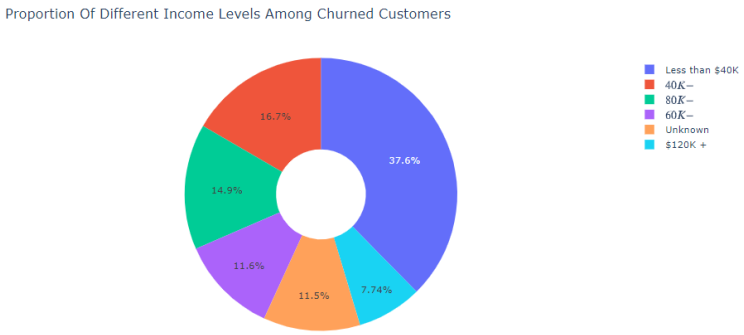


Figure 2 – Proportion of income levels among churned customers

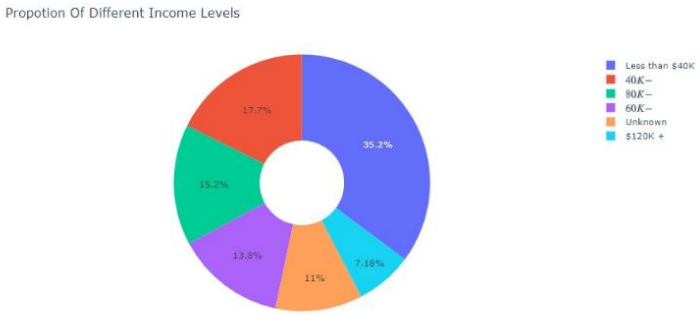


Figure 3- Proportions of income levels

By comparing the visualizations of cancellations count by annual earnings, Figure 2 and Figure 3 show a fairly similar distribution. From these visual representations, we cannot derive any significant insights regarding the differences in cancellation patterns among different earning groups. However, it's noteworthy that the majority (35.2 %) of card holders fall into the group earning less than $40K annually. This may suggest other factors, beyond earnings, play a more dominant role in influencing cancellations.

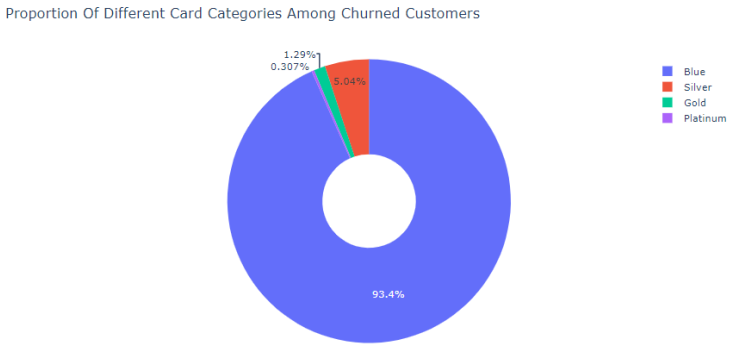


Figure 4 - Proportion of card types among churned customers

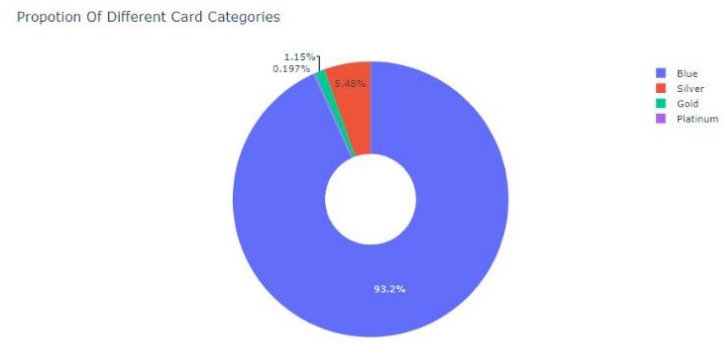


Figure 5 - Proportion of card types

The cancellation count based on card type illustrated in Figure 5 - Proportion of card types and Figure 5 is close to equal within each card category, suggesting that the type of card might not be a major determinant in the decision to cancel. The platinum card holders consist of 20 customers, too small to determine a larger churn rate. The cancellation rate appears to be evenly distributed across the various card categories, suggesting that other factors, such as education

level, marital status or gender, might have a more significant influence on a customer's choice to discontinue their card.

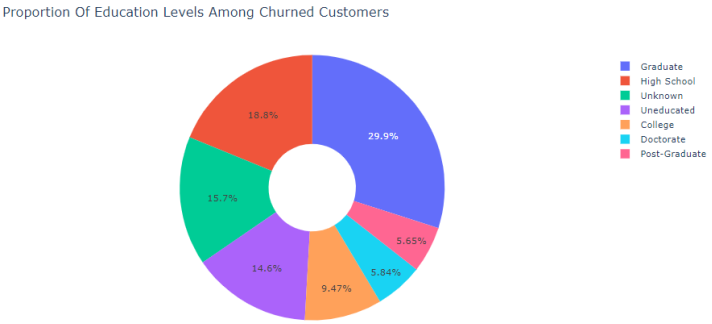


Figure 7 - Proportion of education among churned customers

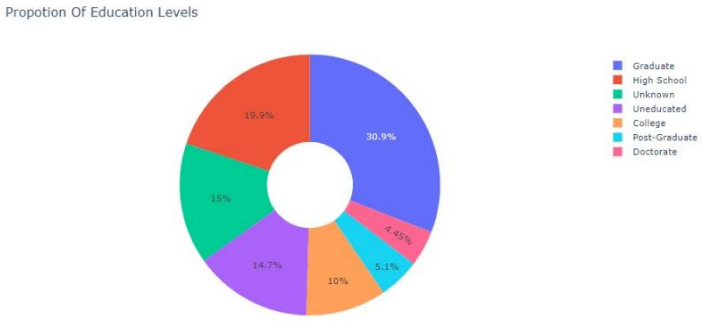


Figure 6 - Proportion of education

Figure 6 - Proportion of education and Figure 7 shows the cancellation count by education level compared to the proportion of education levels. There is no stark difference in the cancellation patterns across the board. However, there is a slightly lower rate of graduates and high schoolers churning, while doctorates and post-graduates skew slightly higher.

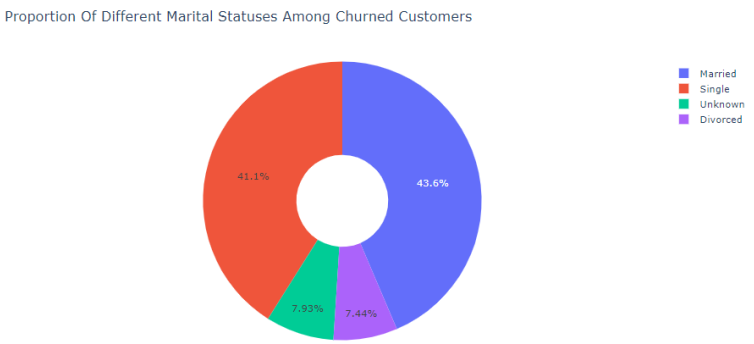


Figure 8 - Proportion of marital status among churned customers

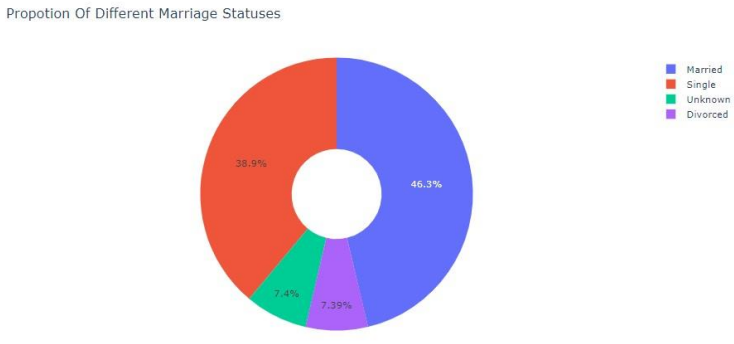


Figure 9 - Proportion of marital status

Comparing the cancellations by marital status to the proportion of different marital statuses, there are no large disparities evident, as shown in Figure 8 and Figure 9. This indicates that marital status might not be a primary factor influencing the decision to cancel. Both single and married individuals, as well as those from other marital categories, seem to have comparable cancellation rates relative to their representation in the dataset.

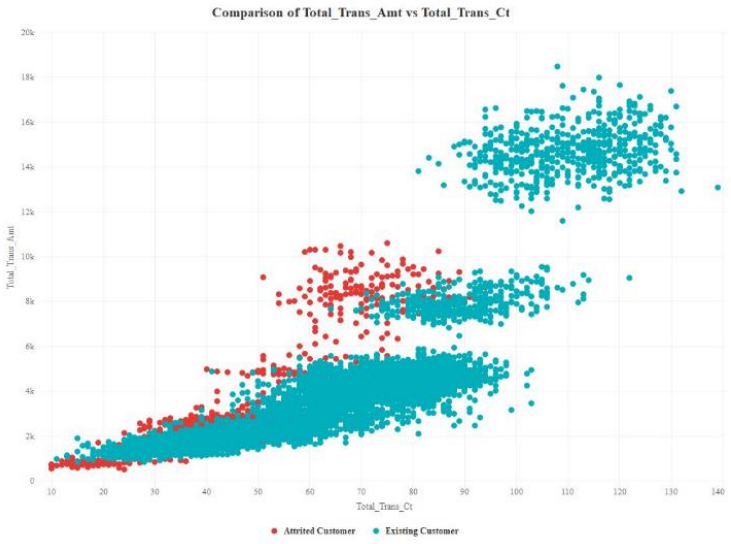


Figure 11 - Comparison of total transaction amount and total transaction count by churn

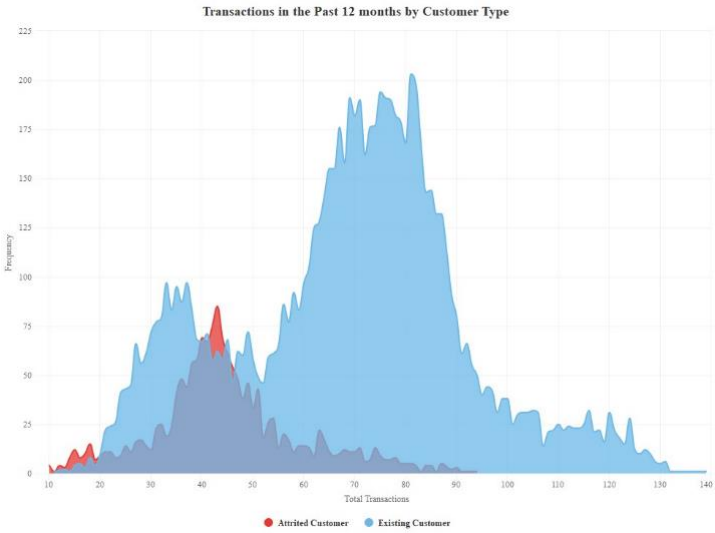


Figure 10 - Total transaction count by churn

Comparing the total transactions of churned and existing customers, shown in Figure 10 and Figure 11, there is a significant difference in the groups. The transaction amounts for churned customers skews towards the lower end, indicating that they were possibly not as engaged or satisfied with the services or products offered. On the other hand, existing customers demonstrate higher transaction amounts, suggesting stronger loyalty and satisfaction. Similarly, the transaction counts also show a stark contrast. Churned customers generally have fewer transactions, which could be a precursor to their eventual departure. Existing customers, with more frequent transactions, underline their consistent engagement with the platform or service. This disparity emphasizes the need to closely monitor customer behavior and engagement to identify potential churn indicators early on and take preventive measures.

Another apparent difference between churning and existing customer is the overall revolving balance, illustrated in Figure 12. Churned customers typically have a lower revolving balance compared to their existing counterparts. This could be indicative of their decreased reliance on or satisfaction with the financial platform or service.

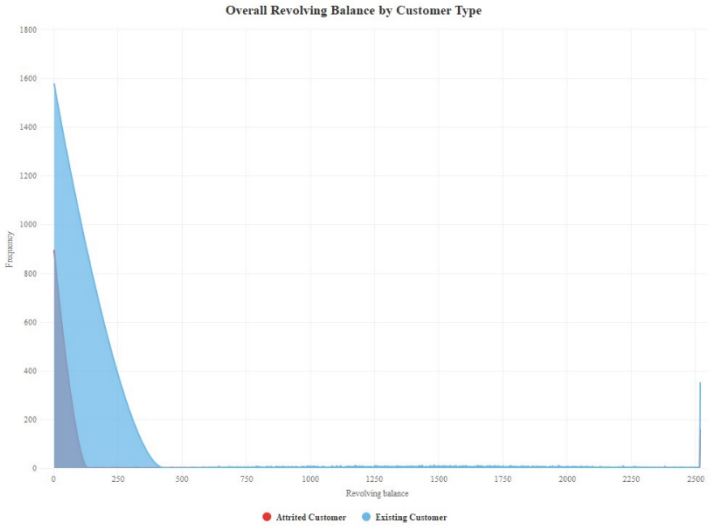


Figure 12 - Overall revolving balance by churn

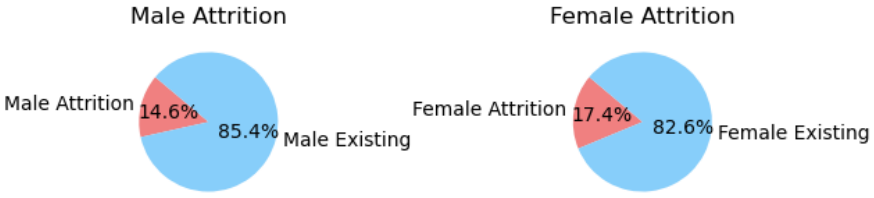


Figure 13 – Churn rates for males and females in the dataset.

The observed differences in churn rates between genders appear to be significant. Specifically, males exhibit a churn rate of 14.62 %, compared to females who show a higher rate of 17.36 %. This discrepancy suggests that females tend to churn at a higher rate than males in this business context. A visual illustration of this is provided in Figure 13 – Churn rates for males and females in the dataset..

3.3 Correlations Between Variables

As an initial investigation of the relationship between the variables, we can apply a correlation matrix. The correlation matrix is used to measure the relation between linearly connected quantitative factors, and it is vital to remember that the correlation indicates relation rather than causation. This is because there might be another element at work that provides a more accurate explanation for the pattern (Damghani, Welch, O'Malley, & Knights, 2013).

The results in the correlation matrix reveal some persistent relationships between our prior independent variables and the customer churn, though the validity of the correlation matrix is debatable given the relatively low correlation values for some variables. Out of the correlation matrix, these are among the most important findings:

1. **Total_Relationship_Count & Attrition_Flag (-0.15)**

A slight positive correlation here might suggest that customers with multiple relationships with the bank (e.g., checking account, mortgage, credit card, loan) are slightly less likely to churn. This could be because of loyalty to the business making them buy more services, speaking to satisfaction. It might also be cumbersome to change banks if the customer is already invested with several products.

2. **Avg_Utilization_Ratio & Gender (0.26)**

A correlation between the average utilization ratio of credit and gender can be influenced by various socio-economic factors. Based on this data, females, on average, tend to use more of their available credit. This could be due to factors like income disparities, spending habits, or behavioral differences that vary between genders.

3. **Credit_Limit & Gender (-0.42)**

The negative correlation may suggest that females typically have a lower credit limit compared to males. This might be reflective of underlying biases, or it could be a result of differences in income, credit history, or risk assessments by the bank. It could also be from customer choice if the bank allows customers to choose from a range themselves.

4. **Total_Trans_Ct & Total_Relationship_Count (-0.24)**

The slight negative correlation suggests that the more transaction counts a customer has, the less likely they are to have a higher number of relationships with the business, or the other way around. Customers with many relationships might feel secure and hence transact less frequently, or something else might cause this correlation.

5. **Total_Trans_Ct & Attrition_Flag (-0.37)**

The negative correlation between transaction count and attrition flag suggests as transaction count goes up, the attrition flag tends to gravitate towards zero, which means less likelihood of

attrition. A higher usage count could suggest the card is necessary for the customer, and they are satisfied with the current state of terms of use.

6. **Attrition_Flag & Months_Inactive_12_mon (0.15)**

The slight positive correlation here could indicate that the longer a customer remains inactive, the more likely they are to leave the bank entirely. This does sound reasonable as high inactivity could speak to a low necessity of the credit card for the customer. Inactivity can also suggest dissatisfaction or transition to other service providers.

7. **Contacts_Count_12_mon & Attrition_Flag (0.2)**

The number of times a customer contacts the bank in a year can be indicative of their relationship with the bank. A positive correlation of 0.2 suggests that as the frequency of customer-initiated contact increases, attrition slightly increases. Frequent contact might signify issues or concerns that the customer faces, leading to potential dissatisfaction. It is also plausible that there are customers who are not contacting the bank, potentially facing unresolved issues, and are therefore silently drifting away or considering leaving without increasing the correlation between attrition and contact counts.

In all these cases, we need to consider other variables and external factors, as correlation does not imply causation.

3.4 VIF Matrix

To tackle the issue of multicollinearity in models that are ill equipped to handle it natively, we utilize the variance inflation factor (VIF) tool. VIF evaluates variables based on the extent to which collinear independent variables increases the variance of estimated coefficients (Craney & Surles, 2002). In practice, VIF conducts a regression analysis for each independent variable as a dependent variable to assess how well the other independent variables predict the specific variable. A higher predictive accuracy result in a higher VIF score.

Typically, VIF values less than or equal to 10 or 50 are considered cutoff points, though Craney & Surles note that there are no formal cutoff values. Hence, we have chosen to use 10 as our threshold for deciding when to exclude a variable for high multicollinearity. (Craney & Surles, 2002). In this thesis, VIF is employed for feature selection. Despite the data being well-structured, it reveals that a few variables can be removed.

From the matrix in Table 4, it is evident that Credit_Limit, Total_Revolving_Bal, and Avg_Open_To_Buy all exhibit infinite values in the VIF analysis. None of the other variables in the matrix display unusually high values. Based on the analysis, we chose to remove three variables where models that do not handle multicollinearity well are deployed. Total_Revolving_Bal and Avg_Open_To_Buy were eliminated to reassess and observe the impact on Credit_Limit. Additionally, CLIENTNUM was removed due to its irrelevance to our analysis, it could be used instead of indexing, but we found it to be redundant. Table 5 includes results from the second VIF calculation, which shows no significant indices following these adjustments.

Table 4 – VIF matrix first run.

	Feature	VIF
1	CLIENTNUM	1.049802
2	Customer_Age	2.761348
3	Gender	3.484005
4	Dependent_count	1.045985
5	Months_on_book	2.782836
6	Total_Relationship_Count	1.158865
7	Months_Inactive_12_mon	1.012555
8	Contacts_Count_12_mon	1.038949
9	Credit_Limit	inf
10	Total_Revolving_Bal	inf
11	Avg_Open_To_Buy	inf
12	Total_Amt_Chng_Q4_Q1	1.196290
13	Total_Trans_Amt	3.331803
14	Total_Trans_Ct	3.117018
15	Total_Ct_Chng_Q4_Q1	1.210496
16	Avg_Utilization_Ratio	3.003598
17	Education_Level_1	1.692865
18	Education_Level_2	1.866049
19	Education_Level_3	1.504336
20	Education_Level_4	2.117916
21	Education_Level_5	1.274807
22	Education_Level_6	1.243038
23	Marital_Status_1	3.844168
24	Marital_Status_2	3.933578
25	Marital_Status_3	1.857568
26	Income_Category_1	1.504177
27	Income_Category_2	2.716521
28	Income_Category_3	3.130763
29	Income_Category_4	2.275432
30	Income_Category_5	1.275085
31	Card_Category_1	1.401314
32	Card_Category_2	1.124002
33	Card_Category_3	1.028991

Table 5 – VIF matrix after changes.

	Feature	VIF
1	Customer_Age	2.691841
2	Gender	3.475553
3	Dependent_count	1.045914
4	Months_on_book	2.660305
5	Total_Relationship_Count	1.158678
6	Months_Inactive_12_mon	1.011157
7	Contacts_Count_12_mon	1.038560
8	Credit_Limit	2.577626
9	Total_Amt_Chng_Q4_Q1	1.195847
10	Total_Trans_Amt	3.293006
11	Total_Trans_Ct	3.103895
12	Total_Ct_Chng_Q4_Q1	1.209664
13	Avg_Utilization_Ratio	1.350267
14	Education_Level_1	1.692400
15	Education_Level_2	1.865662
16	Education_Level_3	1.504309
17	Education_Level_4	2.117555
18	Education_Level_5	1.274716
19	Education_Level_6	1.242669
20	Marital_Status_1	3.842695
21	Marital_Status_2	3.932959
22	Marital_Status_3	1.857561
23	Income_Category_1	1.500571
24	Income_Category_2	2.713811
25	Income_Category_3	3.128580
26	Income_Category_4	2.274978
27	Income_Category_5	1.267181
28	Card_Category_1	1.399147
29	Card_Category_2	1.123597
30	Card_Category_3	1.028625

4. METHODOLOGY

In this chapter, we explore the methodologies employed in the dissertation. The primary focus is on machine learning algorithms and their applications in predictive modeling. We will explore three main techniques: Binary Logistic Regression, Decision Trees with Random Forest and Boosted Random Forest. We will then investigate cross validation, feature importance and classification metrics.

4.1 MOTIVATION FOR MODEL SELECTION

Understanding customer churn in the credit card industry is a multifaceted endeavor, shaped by diverse variables ranging from individual behavior to macroeconomic factors. The inherent complexity of this domain necessitates the utilization of sophisticated predictive models that can adequately capture the nuances and relationships within the data. To determine the selection of models, their suitability for the task at hand is highlighted:

- Decision trees with random forest: Decision trees effectively handle categorical variables and missing values, providing clear decision rules. Random Forest, an ensemble method of decision trees, offers a more nuanced approach by reducing overfitting and improving generalization. This is achieved by constructing multiple decision trees during training and outputting the mode of the classes (classification) of the individual trees for prediction (Hastie, Tibshirani, & Friedman, 2008). Decision trees was also the second most widely estimation technique in churn prediction in the paper "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models" (Neslin, Gupta, Kamakura, Lu, & Mason, 2006).
- Boosted random forest (XGBoost): Boosting is a technique that adjusts for the weaknesses of individual models by combining them. When applied to Random Forests, it iteratively enhances the performance by focusing on misclassified data points from prior trees. In a domain like customer churn, where specific subtle patterns may determine churn behavior, the boosted random forest can uncover these intricate relationships (Friedman, 2001). The primary rationale for utilizing XGBoost in this thesis stems from its performance in terms of accuracy, as extensively detailed later in the analysis section of Chapter 5. The selection of XGBoost is also justified by its adoption in previous literature in the context of customer churn, for instance in the well-cited paper "Defection Detection" (Neslin, Gupta, Kamakura, Lu, & Mason, 2006).

- Logistic regression: As a foundational method in binary classification problems, logistic regression is ideal for providing a baseline model. Its transparency allows for the easy interpretation of relationships between predictors and the target variable, making it a natural starting point. Given the dichotomous nature of churn – a customer either churns or doesn't – logistic regression stands as an intuitive choice (Hosmer, Lemeshow, & Sturdivant, 2013). Logistic regression was also the most widely estimation technique in churn prediction in the paper "Defection Detection" (Neslin, Gupta, Kamakura, Lu, & Mason, 2006)**.**
- Cross validation: Given the models employed and the risk of overfitting, cross-validation is imperative. It provides a robust measure of a model's predictive power, ensuring that our conclusions are not merely an artifact of a particular train-test data split but are representative of broader trends (Wu & Vos, 2018)**.**

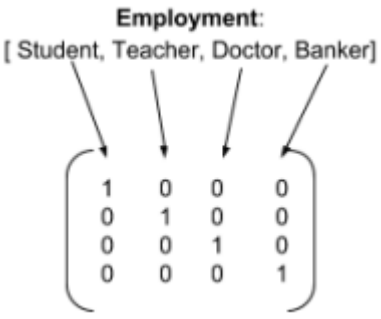
4.2 ONE HOT ENCODING

First, to avoid issues due to the categorical variables Gender, Income_Level, Marital_Status, Education_Level and Card_Category, we have applied the method of one-hot encoding. This method creates a new column for each unique value of the variable, signifying whether that value was active with a binary 1 or 0. This is a technique used to convert categorical variables into a form that could be provided to machine learning algorithms to do a better job in prediction. It works by creating a new binary column for each category of the variable (Rodríguez, 2018). For example, if a variable has three categories, like low, medium, and high, one-hot encoding will create three new columns, one for each category, with binary values:

Low: [1, 0, 0]

Medium: [0, 1, 0]

High: [0, 0, 1]



1

[1] Graphical representation of one-hot encoding, adapted from "An Investigation of Categorical Variable Encoding Techniques in Machine Learning: Binary Versus One-Hot and Feature Hashing" by Cedric Seger, KTH Royal Institute of Technology.

The utility of one-hot encoding becomes particularly evident in scenarios where categorical variables exhibit irregular intervals or lack ordinality. For instance, within our dataset, “Marital_Status” presents as a categorical variable without any inherent sequential order in its values. A similar observation applies to “Education_Level”. While it might be conceivable to utilize “Income_Level” without one-hot encoding, the inconsistent ranges within this variable present a more intriguing analytical challenge. One-hot encoding, by transforming these categorical variables into a binary format, simplifies their representation. This preprocessing technique is especially beneficial when deploying complex models such as binary logistic regression and boosted random forests, as it aids in effectively handling non-ordinal categorical data.

4.3 Binary Logistic Regression

Binary logistic regression is a statistical technique used to model the relationship between a binary dependent variable and one or more independent variables (Hosmer, Lemeshow, & Sturdivant, 2013). In this context, the binary outcome denotes two categories or classes, "1" or "0". In our thesis, this distinction indicates whether an individual has churned. In this chapter, we delve into the methodologies employed in this research. The primary focus is on machine learning algorithms and their applications in predictive modeling. We will explore five main techniques: Binary Logistic Regression, Decision Trees with Random Forest, Boosted Random Forest, Neural Networks, and Hybrid Models.

The logistic function is an S-shaped curve, known as the sigmoid function, which maps any real-valued number into a value between 0 and 1. This function is represented as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Where σ(z) is the output probability estimate, e is the base of natural logarithms, and z is the input to the function, derived from the linear combination of predictors. The primary equation for binary logistic regression is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- p denotes the probability of the dependent event occurring.
- β0 is the intercept.
- β1, β2 are the coefficients of the predictor variables X1, X2, ...

The coefficients in the model signify the change in the log odds of the dependent variable for a one-unit change in the predictor variable. It's crucial to note that the relationship between predictors and log odds is assumed to be linear. Binary logistic regression operates under several assumptions (Hosmer, Lemeshow, & Sturdivant, 2013):

- The response variable is binary.
- The observations are independent of each other.
- There's no multicollinearity among predictor variables, which means predictors are not highly correlated.
- A sufficiently large sample size is required for reliable estimates.

While binary logistic regression offers robust analysis under these conditions, it's essential to explore other methods that can complement or provide alternative perspectives in situations where these assumptions might not hold or where different model characteristics are beneficial.

4.4 Decision Trees with Random Forest

Decision trees are a foundational machine learning algorithm used for both classification and regression tasks. They segment the dataset into subsets based on the values of input features. This segmentation aims to achieve purity in the target variable within each subset. The tree comprises nodes, which represent tests on features, and branches, which indicate the test outcomes (Hastie, Tibshirani, & Friedman, 2008).

Random Forest enhances the decision tree algorithm by constructing multiple trees during training and aggregating their predictions during testing. This ensemble approach aims to improve accuracy and robustness by mitigating overfitting, a common issue with standalone decision trees. Random Forest introduces diversity among trees through techniques like bootstrap aggregating and feature randomness, ensuring a more generalized model (Hastie, Tibshirani, & Friedman, 2008).

As cited by scikit-learn on chapter 1.10.7 (scikit learn, 2023): “Decision trees segment the dataset into subsets based on the values of input features. Given training vectors $x_i \in R^n$, i=1,..., l and a label vector $y \in R^l$, a decision tree recursively partitions the feature space such that the samples with the same labels or similar target values are grouped together. Let the data at node m be represented by Qm with n_m samples. For each candidate

split $\theta = (j, t_m)$ consisting of a feature j and threshold t_m, partition the data into $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$ subsets":

$$Q_m^{left}(\theta) = \{(x, y)|x_j \leq t_m\}$$
$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta)$$

The quality of a candidate split of node m is then computed using an impurity function or loss function H(), the choice of which depends on the task being solved (classification or regression):

$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta))$$

Then, the parameters that minimizes the impurity are selected:

$$\theta^* = \text{argmin}_\theta \, G(Q_m, \theta)$$

Recurse for subsets $Q_m^{left}(\theta^*)$ and $Q_m^{right}(\theta^*)$ until the maximum allowed depth is reach, $n_m < \min_{samples}$ or $n_m = 1$.

In this case, the target is a classification outcome taking on values 0, 1 for node m of the proportion of class k observations:

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k)$$

Common measures of impurity are the following:

Gini:

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

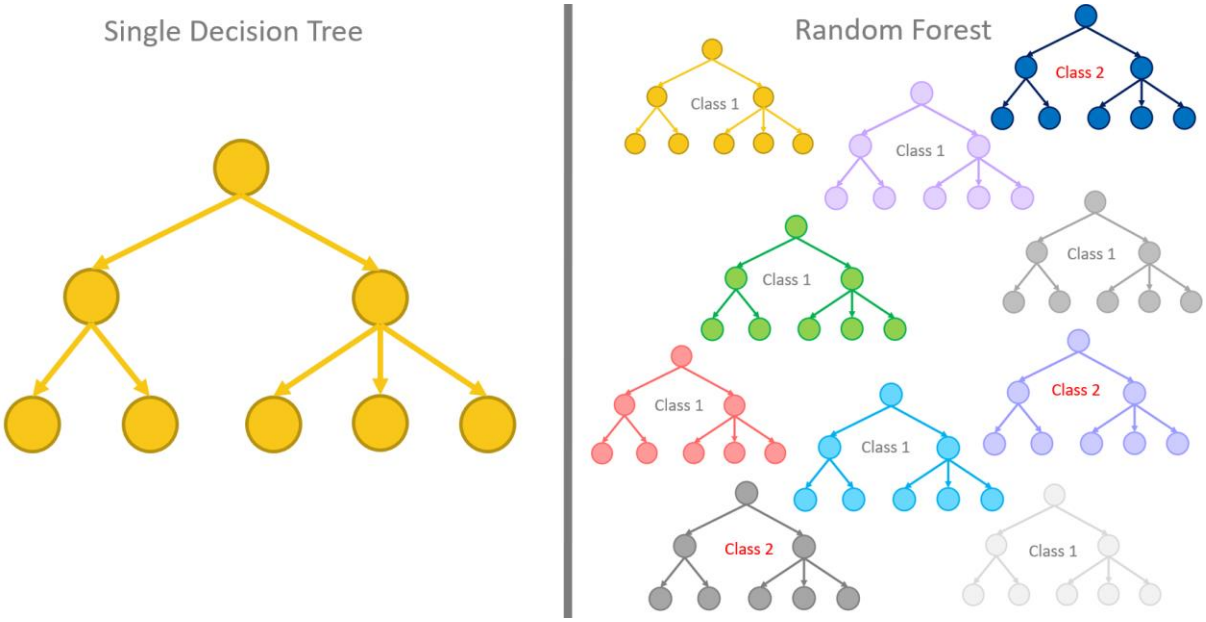
Log Loss or Entropy:

$$H(Q_m) = -\sum_k p_{mk} \log(p_{mk})$$

Random Forest construct many individual decision trees at training, where predictions from all trees are pooled to make the final prediction.

Random forest aggregates predictions from N trees. If Yi is the prediction from the i tree, the final prediction Y for classification is the mode of all Yi, and for regression, it's the average (Hastie, Tibshirani, & Friedman, 2008):

$$Y = \frac{1}{N}\sum_{i=1}^{N} Y_i$$



2

4.5 Boosted Random Forest

Building upon the concept of Random Forest, Boosted Random Forest introduces an additional layer of sophistication to this approach. While Random Forest focuses on reducing variance by averaging multiple deep decision trees, Boosted Random Forest shifts the focus towards reducing bias through a boosting technique. In this method, trees are built sequentially, with each new tree being trained to correct the errors made by the previous ones. This process creates a series of trees that are 'boosted' to improve model performance, particularly in cases where

[2] Figure adapted from Carrillo-Perez, F., et al., 'Applications of artificial intelligence in dentistry: A comprehensive review,' Journal of Esthetic and Restorative Dentistry, 2021. The figure illustrates examples of (A) a decision tree predicting one class and (B) a random forest method, where each tree predicts a class and the overall majority predicted class is taken.

Random Forest might still leave room for improvement in terms of bias. Boosted Random Forest combines the robustness of random forests with the precision of boosting, creating a powerful tool for tackling complex predictive modeling challenges, and a technique that focuses on training instances that are complicated to predict (Friedman, 2001).

The primary principle behind boosting is to assign weights to training instances. Instances that are misclassified by previous trees are given higher weights, ensuring that subsequent trees give them more attention. This iterative correction process enhances the model's accuracy. Boosted Random Forest combines the power of Random Forest with the adaptability of boosting, making it a potent tool for challenging datasets with intricate patterns (Friedman, 2001).

As cited by scikit learn chapter 1.11 (scikit learn, 2023): "Gradient boosting regression trees regressors are additive models whose prediction $\hat{y}_i$ for a given input x_i is of the following form:

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i)$$

Where the h_m are estimators called *weak learners* in the context of boosting. Gradient tree bosting uses decision tree regressors of fixed size as weak learners. The constant M corresponds to the n_estimators parameter. Similar to other boosting algorithms, gradient boosted regression trees are constructed as following:

$$F_m(x) = F_{m-1}(x) + h_m(x),$$

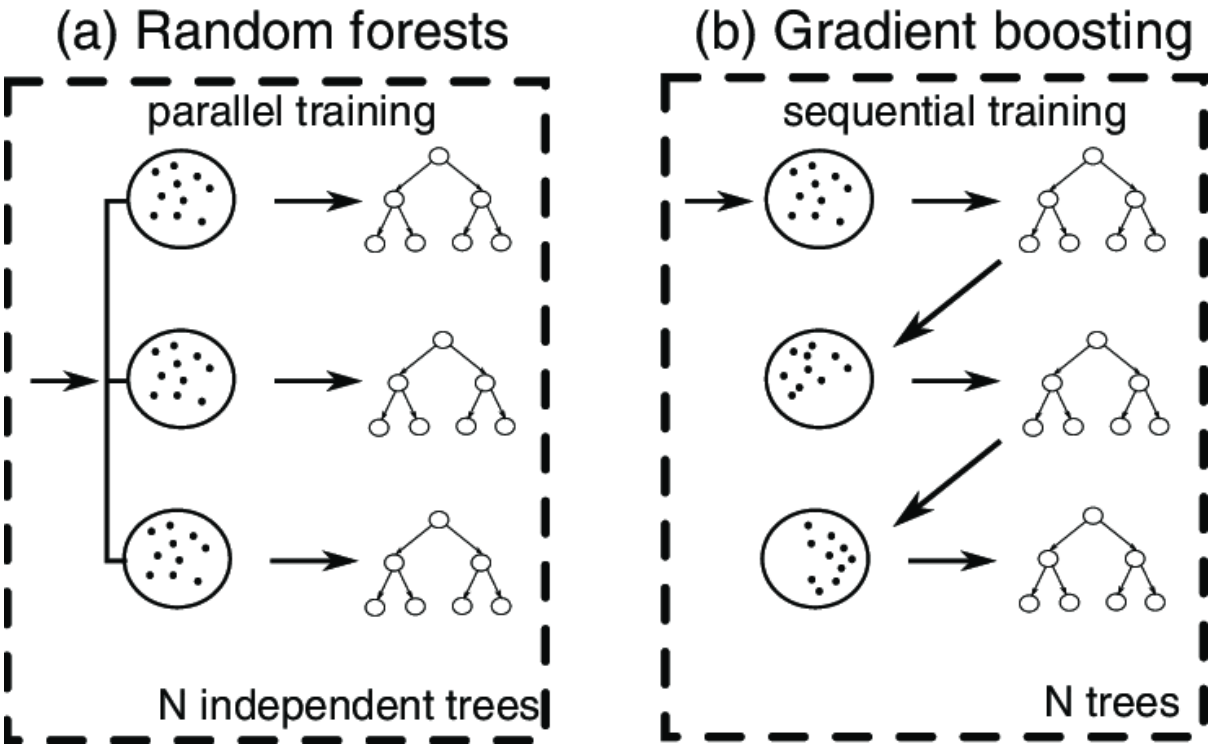
Where the newly added h_m is fitted in order to minimize a sum of losses L_m, given the previous ensemble F_{m-1}:

$$h_m = \arg\min_h L_m = \arg\min_h \sum_{i=1}^{n} l(y_i, F_{m-1}(x_i) + h(x_i)),$$

Where $l(y_i, F(x_i))$ is defined by the loss parameter. By default, the initial model F0 is chosen as the constant that minimizes the loss: for a least-squares loss, this is the empirical mean of

the target values. The initial model can also be specified via the "init"-argument. Using a first-order Taylor approximation, the value of the loss parameter can be approximated as follows:

$$l(y_i, F_{m-1}(x_i) + h_m(x_i)) \approx l(y_i, F_{m-1}(x_i)) + h_m(x_i) \left[\frac{\partial l(y_i, F(x_i))}{\partial F(x_i)} \right]_{F=F_{m-1}}.$$



[3]

According to Chen & Guestrin, XGBoost is an open source, widely known boosted tree package "widely recognized in a number of machine learning and data mining challenges." (Chen & Guestrin, 2016). It is also known for scalability and fast speeds over ten times that of other popular existing methods (Chen & Guestrin, 2016). This is the chosen model for this dissertation.

4.6 Feature Importance

In both the random forest and boosted random forest models, not all variables carry equal significance. It is crucial to identify and understand which variables exert the most influence in these models, as this is key to accurately interpreting their predictions and underlying factors.

[3] Figure 1. 'Comparison between (a) random forest and (b) gradient boosting methods,' adapted from Kowalek et al., Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach, Physical Review E, September 2019.

Feature Importance in Random Forest

In the random forest model, the importance of a variable is determined by calculating its contribution to reducing the squared error risk. This calculation is performed for each variable across every node within the random forest. Subsequently, the score for a particular feature is averaged relative to all other features. However, in a boosted model, such as the XGBoost model we employ, certain variables may be deemed redundant. In such instances, these variables might receive no feature importance score whatsoever (Huiting, Jiabin, & Long, 2017) (Hastie, Tibshirani, & Friedman, 2008).

Feature Importance in Extreme Gradient Boosting

In contrast to utilizing feature vectors for assessing similarities between forecasted and historical data, gradient boosting employs boosted trees to effectively determine feature scores, highlighting each feature's relevance in the training model. A feature's score escalates in proportion to its utilization in making pivotal decisions via boosted trees. The significance of a feature is measured using three metrics: “gain”, “frequency”, and “cover” (Zheng, Yuan, & Chen, 2017). "Gain" is the predominant metric assessing a feature's contribution in tree branches. "Frequency" refers to the count of a feature across all trees, a simpler metric than gain. "Cover" denotes a feature's relative observational value. This research focuses on "gain" for setting feature importance. Breiman proposed a formula involving the square of τˆ to quantify the significance of each predictor feature in a single decision tree, which comprises internal nodes that bifurcate into subregions based on the predictor feature (Breiman, 2001):

$$w_\ell^2(T) = \sum_{t=1}^{J-1} \hat{\tau}_t^2$$

The chosen feature is identified by its maximum potential to reduce squared error risk, compared to a constant fit across the entire region. This squared importance is calculated as the aggregate of such reductions across nodes where the feature was the splitting criterion. The formula for determining importance across multiple trees involves averaging these squared improvements over M trees.

$$w_\ell^2(T) = \frac{1}{M} \sum_{\mathrm{m}=1}^{M} \hat{\tau}_t^2(T_m)$$

The importance of a feature is also judged based on the extent to which prediction performance is impacted when the feature is substituted with random noise. Using data analysis from earlier sections, several features are fed into the XGBoost algorithm to ascertain their contribution to prediction accuracy in training (Zheng, Yuan, & Chen, 2017).

While these techniques are essential for building powerful predictive tools, they also highlight the importance of robust validation tools.

4.7 Cross-validation

Cross-validation is a widely employed technique to assess the predictive performance of models like Boosted Random Forest and logistic regression. The fundamental idea is to split the data into multiple subsets and to test the model's performance on one subset while training on the others. This provides a more robust measure of the model's generalization capability on unseen data as compared to a singular train-test split, which can exhibit biases depending on the specific data selection (Wu & Vos, 2018).

In this thesis, cross-validation is applied to measure the accuracy for each of the models with the overarching goal of affirming that the models can effectively generalize to new and unseen data. Single train-test splits, while simpler, might at times provide performance evaluations that are either too optimistic or too pessimistic based on the idiosyncrasies of the data split. In contrast, cross-validation minimizes this uncertainty by rotating through diverse train and test splits, giving a broader perspective on model performance. By utilizing different subsets for training and testing, it ensures that any bias potentially arising from a specific data split is minimized. Further, when aggregating results across various iterations, the variability in performance metrics is reduced, leading to a more consistent and dependable assessment. Unlike the conventional singular train-test splits, cross-validation guarantees that each data point plays a role in both training and validation, thus ensuring a more efficient use of the available data.

Cross-validation provides a method to directly estimate the test error of prediction, particularly when dealing with substantial sample sizes. This is distinct from criteria such as AIC and BIC, which focus on estimating in-sample error. In the K-fold cross-validation approach, the entire sample is divided into K equally sized segments. If K equals the sample size, termed "n", this method becomes what's known as leave-one-out cross-validation. During each iteration, one of the K subsamples is set aside for validation, while the remaining $K - 1$ subsamples serve the

purpose of model fitting. The resultant cross-validation (CV) estimate for test error is given by an equation, where the variable k(i) denotes the i-th observation in the k-th subsample as cited by (Wu & Vos, 2018)**:**

$$CV = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}^{-k(i)}(x_i)),$$

In practical applications, K often takes values like 5 or 10. However, leave-one-out cross-validation is also frequently used. Generally, as K increases, there's a reduction in model bias but a corresponding rise in model variance. With leave-one-out cross-validation, CV acts as an unbiased estimator for test error but might exhibit high variance. On the other hand, when K is small, say 5, there is less variance in CV, but bias might emerge. For larger original samples (e.g., 200 or more), a smaller stratified K might be adequate. Conversely, for smaller original samples, a bigger stratified K is preferable. Depending on computational capabilities, experimenting with various K values for the same problem can also be beneficial (Wu & Vos, 2018)**.**

4.8 Chi-Square Test

In this thesis, the Chi-Square test emerges as a pivotal tool in further understanding and validating our models. The Chi-Square test, denoted as X^2-test is widely used as a non-parametric statistical test to examine the association between categorical variables. It is particularly useful in determining whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. In the context of our research, the Chi-Square test serves as an essential tool for analyzing categorical data, providing insights into the relationships and dependencies among various variables (The University of Utah, 2022).

The test statistic is calculated using the formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where O_i denotes the observed frequency for the i^{th} category, and E_i denotes the expected frequency under the null hypothesis for the i^{th} category. The summation runs over all

categories. This formula quantifies the discrepancy between the observed and expected frequencies (The University of Utah, 2022).

In our research context, the Chia-square test serves a dual purpose. Firstly, it compares categorical features within a segment to the same categorical features in the entire dataset, analyzing one category from the segment at a time against the same category in the dataset. Secondly, it assesses the homogeneity of feature categories within the segment itself. This test examines whether the distribution of a particular feature, such as the churn rate, differs significantly across different categories within the segment. For instance, it allows us to question whether the distribution of churners in one category, like Gender_0, differs from that in another category, such as Gender_1, within a specific segment, say, “high inactive users”.

The application of the Chi-Square test begins with the establishment of a null hypothesis, which typically posits no association between the variables under study. The alternative hypothesis, conversely, suggests a significant association. The Chi-Square statistic is then computed based on the observed data and compared to a critical value from the Chi-Square distribution, considering the degrees of freedom (df) and the desired level of significance (usually 0.05). Degrees of freedom in this context are calculated as $df = (r - 1)(c - 1)$, where r is the number of rows and c is the number of columns in the contingency table (The University of Utah, 2022).

If the computed Chi-Square statistic exceeds the critical value from the Chi-Square distribution table, the null hypothesis is rejected, indicating a significant association between the categorical variables. Conversely, if the statistic is lower than the critical value, the null hypothesis cannot be rejected, suggesting no significant association.

4.9 Mann-Whitney U Test

The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is used to compare two independent groups on a continuous or ordinal outcome. It is particularly favored when the data do not follow a normal distribution, a common scenario in many practical research settings. This test assesses whether there is a statistically significant difference in the median values of the two groups, offering a robust alternative to the t-test when the assumption of normality is questionable (Wayne W. LaMorte, 2017).

The Mann-Whitney U test is based on ranking all observations from both groups together. The basic premise is to calculate the U statistic, which represents the number of times observations in one group precede observations in the other group in the ranking order. The U statistic for each group is given by:

$$U = n_1 n_2 + \frac{n(n+1)}{2} - R$$

Where n_1 and n_2 are the sample sizes of the two groups, n is the total number of observations, and R is the sum of the ranks for the group in question. The smaller of the two U values calculated for the two groups is used as the test statistic (Wayne W. LaMorte, 2017).

The Mann-Whitney U-test is employed to the dataset to assess whether the distribution of a dependent variable for continuous variables in one segment significantly differs from that I the remaining data. Essentially, it helps to determine whether the observed differences can be attributed to customers belonging to a specific segment or if they are merely a result of random chance.

To apply the Mann-Whitney U test, one must first rank all the observations from both groups together. The ranks are then summed for each group, and the U statistic is calculated. This statistic is compared against critical values from the Mann-Whitney U distribution, taking into account the sample sizes of both groups. The level of significance, alpha α, usually set at 0.05, determines whether the differences in ranks are large enough to be determined as statistically significant.

Interpreting the Mann-Whitney U test involves determining whether the U statistic falls within the region of rejection for the predetermined level of significance. If it does, the null hypothesis, which states that there is no difference in the median values of the two groups, is rejected. This implies that there is a statistically significant difference between the two groups. If the U statistic does not reach this critical value, the null hypothesis shout not be rejected, indicating no significant difference in the median values.

The Mann-Whitney U test assumes that the observations are independent and that the two groups are similar except for the treatment or condition being tested. It also assumes that the responses are at least ordinal. One limitation of the test is that it does not provide information about the magnitude of the difference between groups, only the significance of the difference.

Additionally, while it is more robust than the t-test in the presence of non-normal distributions, extreme outliers can still affect its performance (Wayne W. LaMorte, 2017).

(Kowalek, Loch-Olszewska, & Szwabinski, 2019)

4.10 Classification Metrics

In our thesis, the validation of the predictive models is of paramount importance to ensure the robustness and reliability of the results. To this end, we also employ a comprehensive set of evaluation metrics tailored to the nature of our binary classification problem: predicting customer churn. In developing a predictor model, the dataset was split into three parts: 80% for training and 20% for testing. The effectiveness of models predicting customer churn is measured using classification metrics such as recall, precision, accuracy, false omission rate, and the F1 score. Given the potential imbalance between churned and retained customers in the dataset, the F1 score becomes a crucial metric. It is the harmonic mean of precision and recall, offering a balance between the two (Sokolova & Lapalme, 2009).

To derive these performance metrics, a confusion matrix is created based on the classification results. This matrix consists of four components: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

The elements of the confusion matrix are defined as follows:

• True Positives (TP): The count of churn customers accurately identified as churn.

• True Negatives (TN): The count of non-churn customers accurately identified as non-churn.

• False Positives (FP): The count of non-churn customers incorrectly identified as churn.

• False Negatives (FN): The count of churn customers incorrectly identified as non-churn.

To compute the performance metrics, specific formulas utilizing these matrix elements are employed:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$False\ Omission\ rate\ (FOR) = \frac{TN}{TN + FN}$$

$$F1\ score = 2 * \frac{Precision\ *\ Recall}{Precision + Recall}$$

As an important note, the way the confusion matrix is set up as non-churn considered "positive". Hence, the confusion matrix reported shown in Table 2 will be used. In the classification report, the classification metrics for target value 1 will regard churn.

Table 2 – Confusion matrix report setup.

Actual / Predicted	Non-Churn (0)	Churn (1)
Non-Churn (0)	TN	FP
Churn (1)	FN	TP

ROC and AUC: The Receiver Operating Characteristic (ROC) curve is a graphical representation of the true positive rate against the false positive rate at various threshold settings. The Area Under the Curve (AUC) quantifies the overall ability of the model to discriminate between the positive and negative classes (Fawcett, 2006). An AUC of 1 indicates perfect discrimination, while an AUC of 0.5 suggests no discrimination.

$$ROC - AUC = \int_0^1 TPR(FPR)\,\mathrm{d}FPR$$

$$= \int_0^1 TPR(FPR^{-1}(x))\,\mathrm{d}x$$

By leveraging these metrics, the overall accuracy as well as the balance between precision and recall, the model's discriminatory power and its parsimony is considered.

5. ANALYSIS

This section presents the results of the predictive modeling analysis conducted to understand the factors contributing to customer churn. Utilizing a robust stratified K-fold cross-validation approach, the performance was assessed through binary logistic regression and Boosted Random Forest. First, we will present some general results from the dataset. We will then examine and analyze the proposed hypotheses. The primary objective is to validate or refute these hypotheses based on the data insights and patterns, as well as identify interesting connections to suggest strategies and managerial implications.

5.1 GENERAL RESULTS

Binary Logistic Regression

The first implementation is the binary logistic regression model. Figure 14 shows the results from running the regression on a stratified K-fold of five splits.

	1st fold	2nd fold	3rd fold	4th fold	5th fold
Stratified K-Fold Accuracy scores per fold:	0.88252715	0.89289240	0.88197531	0.87851852	0.88395062
Stratified K-fold Mean accuracy:	0.88397280				

Classification Report based on a threshold of 0.5:

	precision	recall	f1-score	support
0	0.91	0.95	0.93	1700
1	0.68	0.51	0.58	326
accuracy			0.88	2026
macro avg	0.80	0.73	0.76	2026
weighted avg	0.87	0.88	0.88	2026

Confusion Matrix based on a threshold of 0.5:

	Predicted existing	Predicted attrition
Actual existing	1623	77
Actual attrition	161	165

Figure 14 – Classification report from binary logistic regression with stratified K-fold using five splits.

The classification report in Figure 14 shows the accuracy of the model over five splits, as well as a classification report and a confusion matrix for the first fold. This model is created using the features determined through the VIF matrix shown on page 29. The dependent variable is Attrition_Flag, with 0 being existing customer and 1 being attrited customer. The independent variables are Customer_Age, Gender, Dependent_count, Months_on_book, Total_Relationship_Count, Months_Inactive_12_mon, Contacts_Count_12_mon,

Credit_Limit, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Trans_Ct, Total_Ct_Chng_Q4_Q1, Avg_Utilization_Ratio, Education_Level_1 through Education_Level_6, Marital_Status_1 through Marital_Status_3, Income_Category_1 through Income_Category_5, and Card_Category_1 through Card_Category_3.

The logistic regression model's classification report was applied with the default threshold of 0.5. The stratified K-fold approach ensured that the proportion of churn to non-churn customers was consistent in each fold of the model training process, providing a robust validation framework. The model's performance across the five folds yielded accuracy scores of 88.25 %, 89.29 %, 88.20 %, 87.85 %, and 88.40 %, respectively. These results suggest that the model's predictive capability is robust across different subsets of the data, with a mean K-fold accuracy of 88.40 %. This level of accuracy across folds indicates that the model is relatively stable and generalizes well to unseen data.

Classification report

In the classification report, a confusion matrix of the predicted values from the model is presented, shown in Figure 14. To illustrate an important consideration for business managers, we will present a theoretical profit analysis. Assume a retention campaign with 100 % success rate. Retained customers profit the business 500 and applying the retention campaign cost 100 per individual. Table 3 illustrates the possible outcomes based on the classification of customers.

Table 3 – Model profit scenario.

Theoretical profit analysis		
	Description	Business profit
True retention	TN (Retained customer correctly classified)	500
False retention	FN (Churner incorrectly classified)	0
True attrition	TP (Churner correctly classified)	400
False attrition	FP (Retained customer incorrectly classified)	400

This is a fundamental approach designed to illustrate how model configuration can impact business profits. Table 4 shows the profit results from the model compared to the same logistic regression model where the threshold value, the threshold deciding whether a customer is categorized as retained or churned, is set to the profit-maximizing threshold for this specific scenario. Table 5 illustrates the differing confusion matrixes.

Table 4 – Impact from the choice of threshold value, illustrating a threshold value of 0.5 versus the profit-optimizing threshold value of 0.30883088 for this specific scenario.

Threshold	0.5	0.30883088	Change
Accuracy	88,25 %	87,91 %	-0,35 %
Precision (1)	68,18 %	60,31 %	-7,88 %
Recall (1)	50,61 %	72,70 %	22,09 %
False Om. Rate (1)	90,98 %	94,55 %	3,57 %
F1 score (1)	58,10 %	65,92 %	7,83 %
Total profit	**908300**	**929200**	**20900**

Table 5 – Confusion matrix for the logistic regression model at a threshold of 0.5 compared to 0.30883088.

Confusion matrix of the first fold. Default and custom threshold.

		Predicted			
	Threshold value	0.5		0.30883088	
		Existing	Attrition	Existing	Attrition
Actual	Existing	1623	77	1544	156
	Attrition	161	165	89	237

This example is designed to illustrate how prioritizing specific classification metrics can play a pivotal role in profit maximization. It is crucial for the designer of a predictive model to incorporate key metrics such as the average customer lifetime value, likelihood of success for the retention campaign, and an acceptable range for margin of error. With improved models, the prioritization has diminishing returns as misclassifications are reduced. However, if the model is unable to capture discrepancies within the data, thereby providing inadequate results, such prioritizations can increase profits. The authors are not equipped with adequate information to pursue an optimized profit calculation nor an optimal classification priority. Therefore, the threshold value going forward is kept at the default value of 0.5.

While the binary logistic regression model demonstrates certain strengths, particularly in its overall high accuracy, it also reveals critical areas for improvement. Conversely, for the positive class (1), which denotes the attrition group, precision (1) is at 68.18 %. This indicates a considerable number of false positives, where non-attrition cases are incorrectly classified as attrition. Recall (1) at 50.61 % implies a suboptimal classification of true positives, the proportion of attrition cases correctly classified to the total number of attrition cases. The F1-score (1) standing at 58.10 %, though lower than that of the negative class (0), represents a moderate balance between precision and recall given the challenging nature of predicting the minority class.

While results from the logistic regression provides a robust foundation, other modeling techniques have the potential for advancing accuracy in prediction, thereby increasing business profitability, and improved managerial decision-making. Hence, a boosted random forest model was applied for further analysis.

Boosted Random Forest Results

In pursuit of more refined results, XGBoost was the chosen boosted random forest model for our prediction modelling. As previously stated in methodology, the selection of XGBoost is also justified by its adoption in previous literature in the context of customer churn, for instance in the well-cited paper "Defection Detection" (Neslin, Gupta, Kamakura, Lu, & Mason, 2006).

Figure 15 shows the classification report on the performance of the model on a specific test set, including an accuracy score averaged over the same stratified K-folds as the logistic regression was run on.

XGBoost K-folds accuracy:	1st fold	2nd fold	3rd fold	4th fold	5th fold
Stratified K-Fold Accuracy scores per fold:	0.96594274	0.97334650	0.96543210	0.96839506	0.97777778
Stratified K-fold Mean accuracy:	0.97017884				

XGBoost Classification Report for the 1st fold:

		precision	recall	f1-score	support
	0	0.98	0.98	0.98	1700
	1	0.91	0.87	0.89	326
accuracy				0.97	2026
macro avg		0.94	0.93	0.94	2026
weighted avg		0.97	0.97	0.97	2026

XGBoost Confusion Matrix for the 1st fold:

	Predicted existing	Predicted attrition
Actual existing	1659	40
Actual attrition	39	288

Figure 15 – XGBoost model performance

The results from the XGBoost model shown in Figure 15 display an average accuracy of 97.02 % across all folds. The precision, recall, and F1-score scores vary between 0.87 and 0.98. Compared to the logistic regression model, the XGBoost model illustrates advanced capabilities in classification accuracy.

The XGBoost model was re-run a excluding one-hot encoding and feature selection, testing the model's inherent ability to tackle multicollinearity. These results are shown in Figure 16. Due to the variation from stratification occurring when differing features are present, the k-folds are not identical. As such, we do not conduct a direct comparison between the confusion matrices. Nonetheless, the classification metrics and average accuracies are, for all practical purposes,

identical. This suggests a consistent performance of the model regardless of feature selection and one-hot encoding.

Stratified K-Fold Accuracy scores per fold:	0.96544916	0.9733465	0.96691358	0.96888889	0.9782716
Stratified K-fold Mean accuracy:	0.97057395				

XGBoost Classification Report:

		precision	recall	f1-score	support
	0	0.97	0.73	0.83	1699
	1	0.91	0.87	0.89	327
accuracy				0.97	2026
macro avg		0.94	0.93	0.94	2026
weighted avg		0.97	0.97	0.97	2026

XGBoost Confusion Matrix:

	Predicted existing	Predicted attrition
Actual existing	1671	28
Actual attrition	41	286

Figure 16 – XGBoost classification report and confusion matrix when model is trained on all features.

The comparative analysis between the logistic regression model and the XGBoost model reveals several key distinctions. Primarily, the XGBoost model demonstrates higher accuracy, and an inherent handling of multicollinearity. Additionally, while both models show similar performance on classifying existing customers (0), the improvement XGBoost shows in classification metrics regarding classification of attrited customers (1) is significant. This translates to an increase in correct classification of attrited customers.

XGBoost Feature Importance

Understanding the underlying reasons of the XGBoost model leads us to the analysis of feature importance. The importance of each variable applied in the XGBoost model is shown in Figure 17. In the XGBoost model, the "feature importance" attribute provides a calculated relative score depicting the feature's contribution in reducing squared errors, discussed in chapter 4.6 FEATURE IMPORTANCE. The results reveal that Total_Trans_Ct is characterized as the most important feature for lowering errors in the prediction. The two subsequent features are Total_Relationship_Count and Avg_Utilization_Ratio.

In section "3.3 CORRELATIONS BETWEEN VARIABLES", the linear correlation of Total_Trans_Ct and Total_Relationship_Count was assessed. Respectively, a value of -0.37 and -0.15 for the

features. Total_Revolving_Bal has a linear correlation of -0.26, suggesting an increase in revolving balance reduces the likelihood of churn.

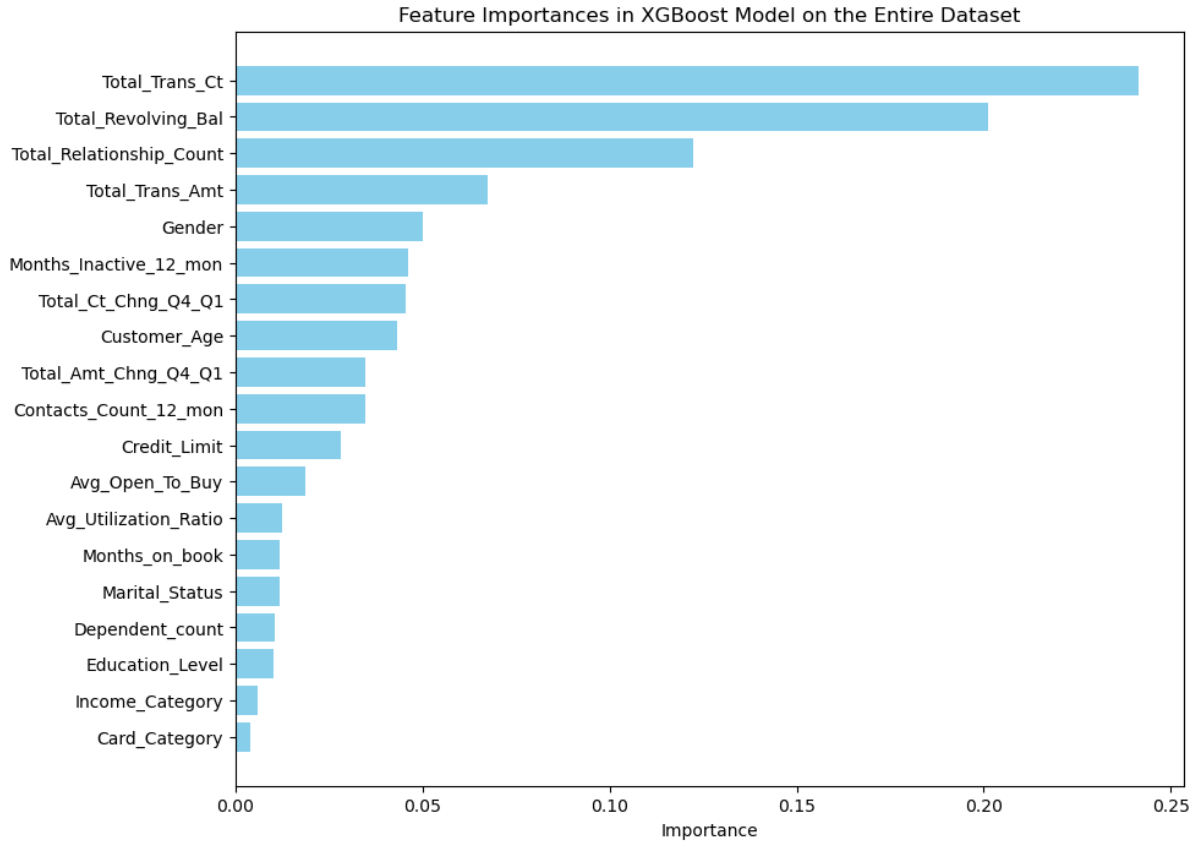


Figure 17 – Feature importances from the XGBoost model trained on the entire dataset.

The dataset used in the XGBoost model incorporated all variables, apart from the 'CLIENTNUM' feature as it provides no meaningful information except indexing. The XGBoost model, when trained on the dataset with preprocessing such as one-hot encoding and feature selecting, did not exhibit any notable improvement to the model trained with the preprocessing steps omitted. The independent variables included: age, gender, number of dependents, education level, marital status, income category, product variable, period of relationship with the bank, total number of products, number of months inactive in the last 12 months, no. of contacts in the last 12 months, credit limit, total revolving balance on the credit card, open to buy credit line (average of the last 12 months), change in transaction amount (Q4 over Q1), total transaction amount (last 12 months), total transaction count (last 12 months),

change in transaction count (Q4 over Q1), and average card utilization ratio. The dependent variable was the customer's status, classified binary as either churn or non-churn (1 or 0).

Figure 17 illustrates the significance of the variables that were incorporated as independent variables, merging the gathered data with the variables**.** The feature importance analysis identified the top three variables as the total transaction count, total revolving balance, and the total number of products held by the customer. In contrast, variables like dependent count, education level, and income categories demonstrated minimal relevance in prediction power in the XGBoost model.

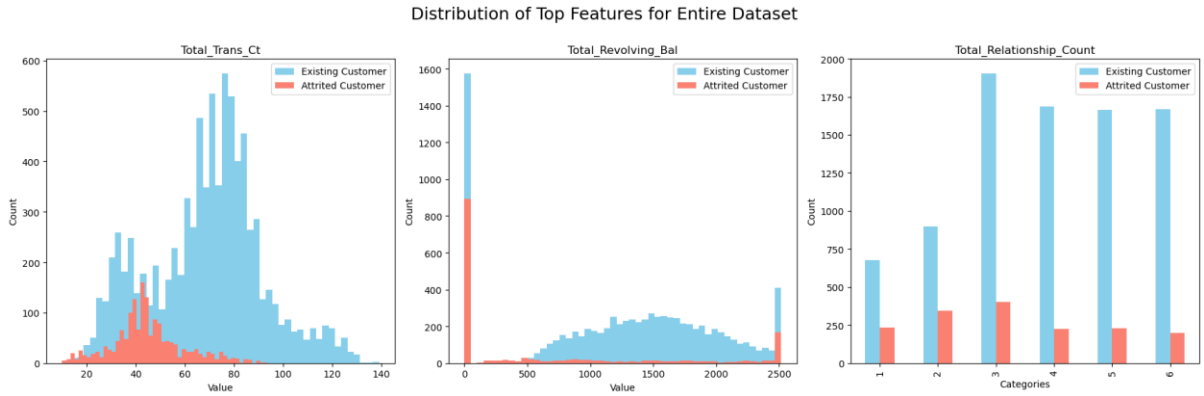


Figure 18 – Distribution of the top three features for the customers in the entire dataset for one of the five k-folds.

To understand the association between a feature and customer churn, we can examine the top three features. Firstly, Total_Trans_Ct appears to be a significant factor, particularly towards the low part of the distribution, with a notable peak for attrited customers around 42 transactions. This trend suggests a correlation between fewer transactions and a higher likelihood of churning. However, examining the Total_Trans_Ct plot in Figure 18, we see a similar concentration of existing customers in the lower transaction range. Attrited customers are predominantly below 70 transactions, suggesting that reviewing customers with lower transaction counts could yield interesting insights.

Secondly, the plot depicting Total_Revolving_Bal in Figure 18 reveal 508 customers with a revolving balance of 2 517, and 2 470 customers with a balance of 0. The frequent occurrence of 2 517 suggests an upper limit, as other balances fall within a continuous range. Both groups contain a significant number of churned customers. The zero-balance group represents

customers in two groups, those do not utilize their credit cards, or customers who consistently pay off their balances.

Thirdly, Total_Relationship_Count, reflecting the number of services or products a customer holds with the business, also shows a correlation. A higher proportion of churned customers is seen among those with fewer products or services, albeit among a slightly smaller sample.

A striking observation across these features is the trend of low values correlating with higher churn rates, suggesting that churning customers might gradually decrease their usage of the business's products or services. Differently from the linear correlation explored previously, the concentration of attrition in the lower values suggest a non-linear relationship between the features and attrition.

From the general results, we've observed intriguing trends, especially the association between low engagement in various features and increased churn rates. Appendix A and B provides extensive results for the general dataset in tables and figures. These insights prompt us to analyze deeper questions: What specific patterns exist within customer segments? Can we discern distinct behaviors or preferences that are predictive of churn? To systematically uncover some of these patterns, we analyze the targeted hypotheses about customer behavior.

5.2 Hypothesis 1

Low credit card usage rates are predictive of churn.

In addressing Hypothesis 1, our primary task is to define and contextualize what constitutes a 'low usage rate' of credit cards. We can only define a low rate of usage based on the dataset we have. Since most of the data is skewed and does not follow a normal distribution, using standard deviations as a measure up or down would be misleading. Instead, we establish a definition of “low” or “high” as the 20^{th} percentile or the 80^{th} percentile respectively, in addition to highlighting special cases.

In this hypothesis, we could define low credit card usage through the variables of utilization ratio, transaction count or transaction amount. In the context of this dataset, low utilization ratio is indicative of customers who seldom allow their revolving balance to reach significant amounts. Although this feature could provide insights into customer usage, it might not be as informative as transaction count or amount. Utilization ratio is

likely more descriptive of the customer's approach to their revolving balance, and not necessarily about their total use.

Considering both transaction count and amount, the 20th percentile for transaction count has a churn rate of 32.66 %, while the same percentile for transaction amount has a churn rate of 20.37 %. Although we will note that through reviewing other approaches to defining "low" and "high" for the feature, certain segments defined on transaction amount had a larger churn rate. The bottom 166 customers in transaction amount have 86.14 % churn. In comparison, a mirroring count of the bottom customers in transaction count has a churn rate between 55-60 %.

However, due to the larger churn rate in transaction count even at larger samples, in addition to transaction count as a refined descriptor of frequency of use, as well as the feature standing at a higher correlation to attrition flag, we chose to move on with the 20th percentile of transaction count in the hypothesis.

This segment has a count of 2 076, with a churn rate of 32.66 %. A chi-squared test comparing the segment to the dataset suggests the distribution of attrition is different statistically, with a chi-squared statistic of 308.52 and a P-value of 4.58e-69, far below the standard alpha of 0.05. A lower transaction count does correlate positively with churn.

The top three features calculated by XGBoost for the model trained exclusively on segment data, shown in Figure 19, shows Total_Trans_Ct now ranked as the lowest in feature importance. The feature is naturally lower as the variability is lower within the same feature as the segment is defined upon. As seen in the general model, total revolving balance is still an important feature to reduce misclassifications of for attrition. The two subsequential features are relationship count, and total count change between the fourth and first quarter.

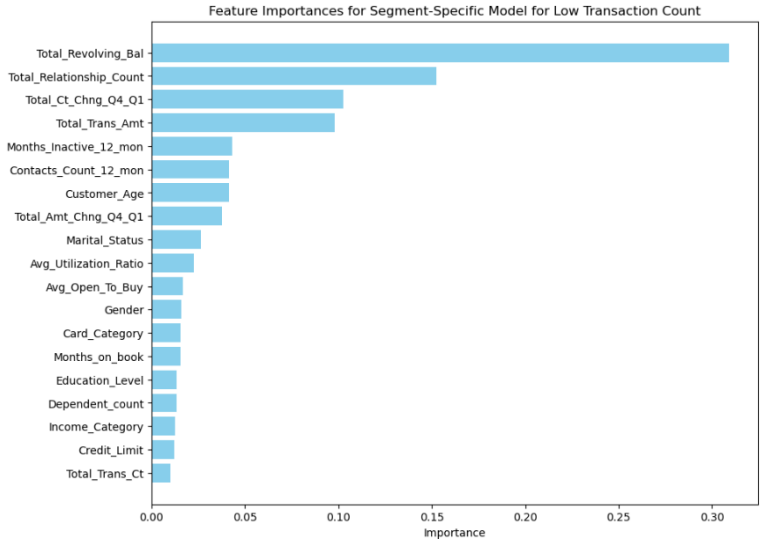


Figure 19 – Feature importance of the segment-specific model for the low transaction count segment.

Figure 20 and Figure 21 shows the distribution of existing and attrited customers for each of the top three features within the segment and the remaining data, respectively. A comparative analysis of these distributions reveals differences in the distributions. In the low transaction count segment, a higher proportion of attrited customers have 0 revolving balance. A similar trend is observed in total relationship count and total count change. However, the authors note that transaction count is not independent from total count change. All customers in the segment are expected to skew towards lower values for total transaction count change.

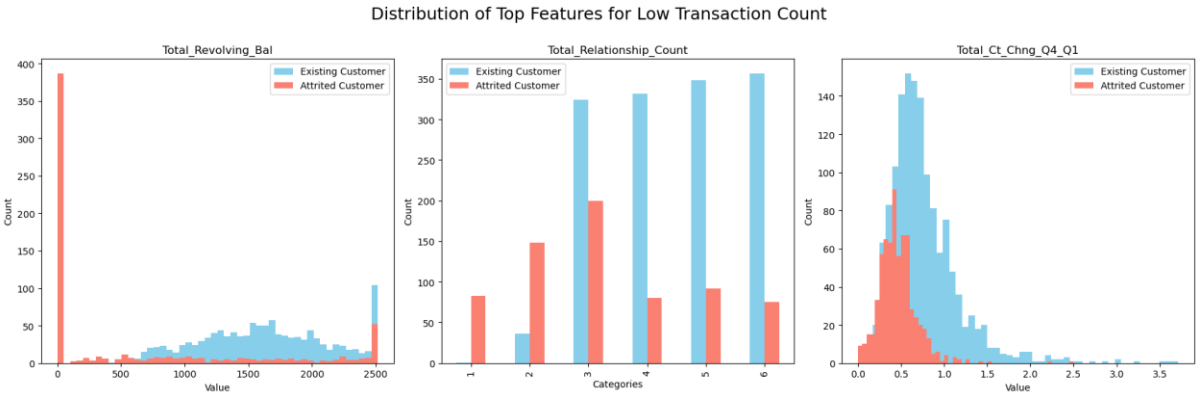


Figure 20 – Distribution of customers within the top three features of the segment decided by the segment-specific model for the Low Transaction Count segment. For the reader, the bar for Total_Revolving_Bal depicting value 0 does not show Existing Customers through the Attrited Customers because the count is lower. There are 170 Existing Customers and 387 Attrition Customers at 0.

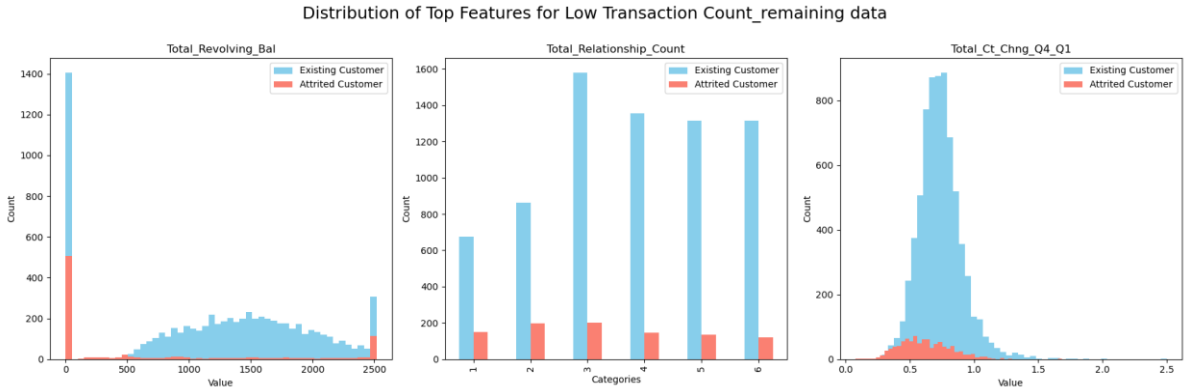


Figure 21 – Distribution of customers within the top three features of the remaining data decided by the segment-specific model for the Low Transaction Count segment.

The analysis of total relationship count reveals a noteworthy observation. Specifically, a higher relationship count correlates to reduced churn rates. However, from four counts, the reduction stabilizes.

Table 6 – Comparison of attrition rate in the low transaction count segment to the remaining data.

	Low Transaction Count	Remaining Data
Total_Relationship_Count_1	98.81 %	18.16 %
Total_Relationship_Count_2	80.43 %	18.70 %
Total_Relationship_Count_3	38.17 %	11.23 %
Total_Relationship_Count_4	19.42 %	9.67 %
Total_Relationship_Count_5	20.91 %	9.30 %
Total_Relationship_Count_6	17.36 %	8.44 %

Table 6 illustrates the attrition rate in the low transaction count segment compared to the corresponding relationship count in the remaining data. Both segments exhibit a reduction in churn rate the higher the relationship count, supporting the linear correlation between the attrition flag feature and the total relationship count feature of -0.15. Interestingly, four relationships and up appears to have diminishing effect.

5.3 Hypothesis 2

Reduction in credit card usage is highly indicative of customer churn.

Building upon the analysis of transaction counts in Hypothesis 1, Hypothesis 2 introduces a dynamic aspect of customer behavior, the reduction in credit card usage over time. In

hypothesis 1, we argued that transaction count was the most prominent descriptor of credit card use among the available features. Total_Ct_Chng_Q4_Q1 depicts the proportion of transaction counts in the fourth quarter compared to the first quarter. To assess reduced use, the value must be below 1. Although an arbitrary value could be set for defining reduction, the authors adapted the same approach as in hypothesis 1, defining the segment at the 20th percentile. This results in a segment with customers reducing their transaction count by a factor of 0.545 or more, a reduction of about 50 % or more. The segment has a count of 2032 customers with a churn rate of 42.13 %.

Performing a chi-squared test for statistical significance shows a statistic of 705.66 and a P-value of 1.79e-155. Not surprisingly, the test suggests that the difference in attrition ratio between the segment and the rest of the data is significant.

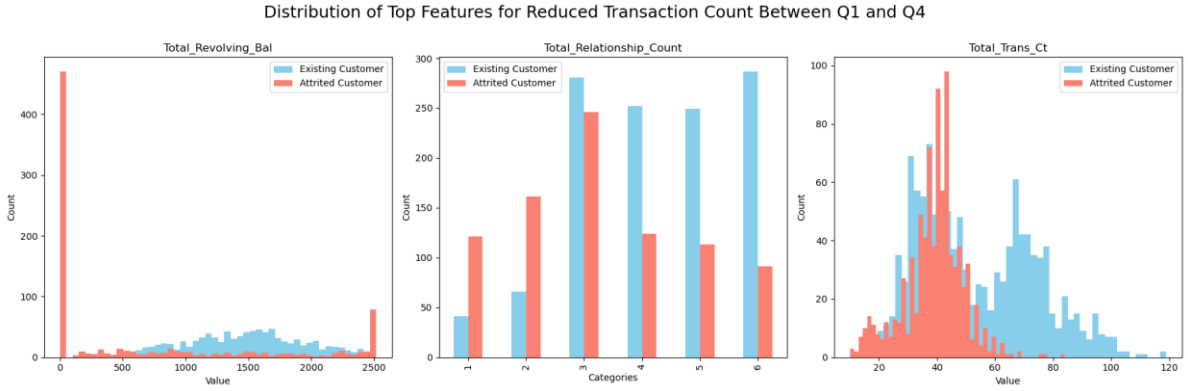


Figure 22 – Top three features from the segment-specific model on reduced transaction count. Note that the count of Existing Customers is not 0 for Total_Revolving_Bal 0.

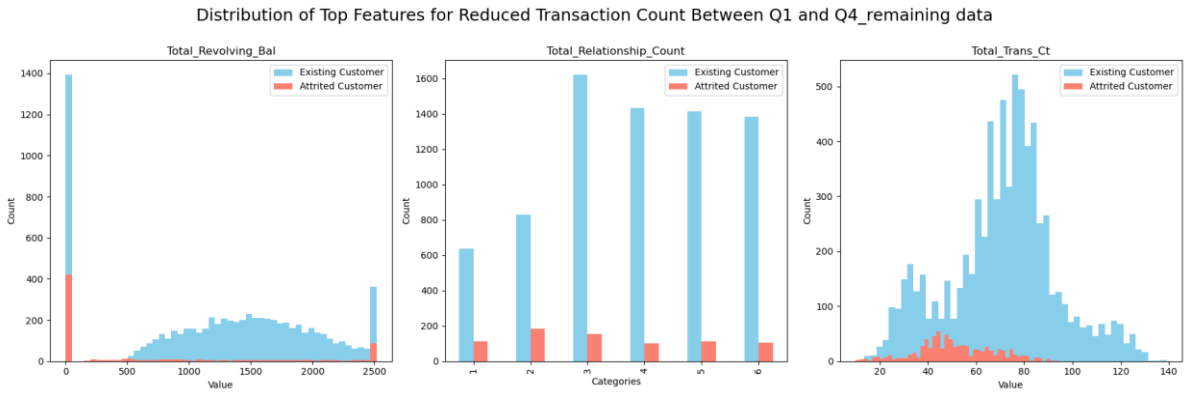


Figure 23 – Top three features from the segment-specific model, showing the distribution of customers within the remaining data.

Figure 22 and Figure 23 illustrate the distribution of attrited and existing customers within the segment and the remaining data, respectively. Total_Trans_Ct in the segment depicts a skew towards lower values, as expected due to the relationship between Total_Trans_Ct and Total_Ct_Chng_Q4_Q1.

Consistent with the findings of hypothesis 1, the top three predictive features for the reduced use segment contains the total relationship count as a prominent predictor. The inverse relationship between relationship count and attrition is observed in the segment, albeit more prominent. In contrast to hypothesis 1, in this segment the higher counts four through six do not stabilize, and instead appear to correlate to a further reduction churn rate.

Table 7 – Comparison of attrition rate in the reduced transaction count segment to the remaining data.

	Reduced Transaction Count	Dataset	Not in segment
Total_Relationship_Count_1	74.69 %	25.60 %	14.97 %
Total_Relationship_Count_2	70.93 %	27.84 %	18.21 %
Total_Relationship_Count_3	46.68 %	17.35 %	8.66 %
Total_Relationship_Count_4	32.98 %	11.77 %	6.58 %
Total_Relationship_Count_5	31.22 %	12.00 %	7.46 %
Total_Relationship_Count_6	24.07 %	10.50 %	7.06 %

Table 7 shows a comparison of the individual categories within Total_Relationship_Count in the reduced transaction count segment, with the corresponding category in the remaining data. The reduced count segment reveals a similar pattern to the low transaction count segment. The churn rate of the reduced transaction count segment appears to go down as the relationship count goes up. Supporting the proposition that a higher relationship count improves the likelihood of retaining a customer.

Although exploring the distribution within the top features can yield insights, another approach includes excluding the top-ranked predictors to assess the predictive abilities, and the distribution of customers within other features. The model was re-run excluding four top features, Total_Trans_Ct, Total_Revolving_Bal, Total_Relationship_Count and Avg_Utilization_Ratio. Accuracy and recall in the classification report went down from 93.8 % and 91 %, to 90.4 % and 85 % respectively. Note that the segment is unaltered.

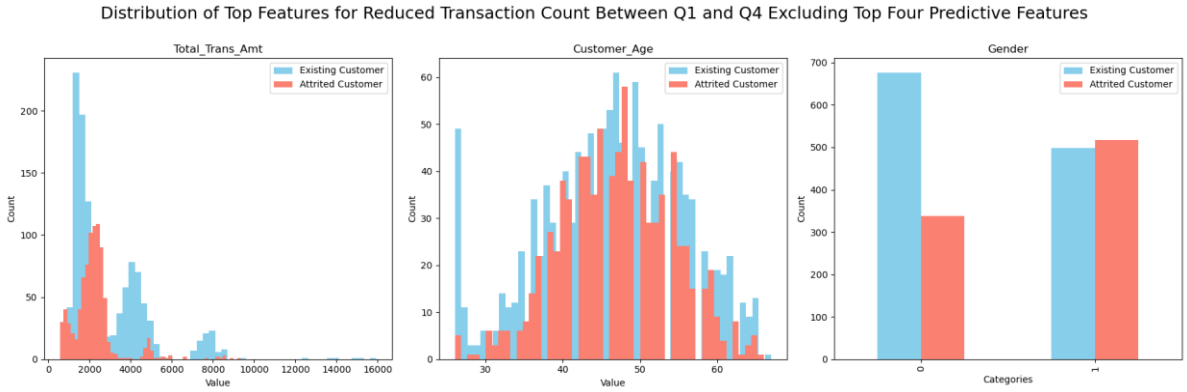


Figure 24 – Distribution of customers within the new top three features of the reduced usage segment.

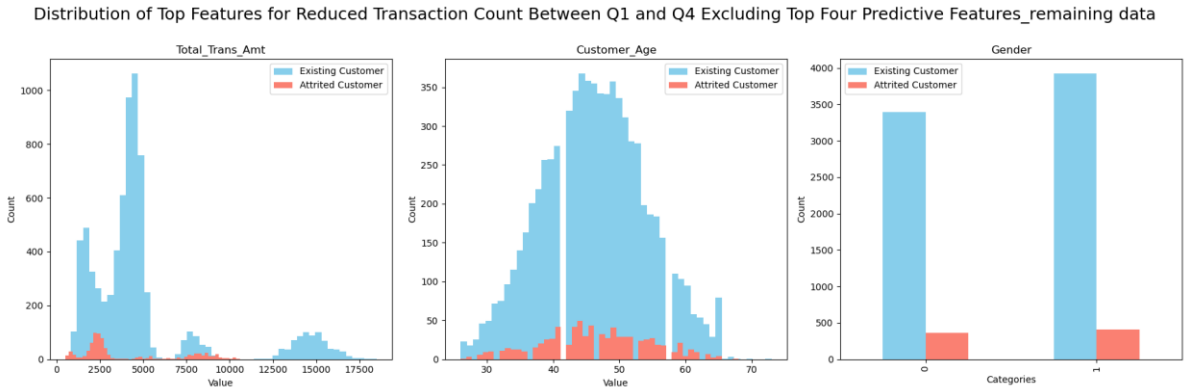


Figure 25 – Distribution of customers within the remaining customers for the new top three features.

Notably, the model now suggests gender as an effective predictor. Illustrated in Figure 24 and Figure 25, the segment with reduced count depicts a significantly higher churn rate for female customers in comparison to male customers.

Table 8 – Churn distribution comparison between the reduced transaction count segment and the remaining data for gender.

	Reduced Transaction Count	Dataset	Not in segment
Gender 0 (Male)	33.30 %	14.62 %	9.56 %
Gender 1 (Female)	50.93 %	17.36 %	9.49 %

Table 8 illustrates the variation in churn rates for males and females in the segment, in the dataset and individuals not in the segment. A linear regression analysis was performed on Attrition_Flag as the dependent variable with Gender as the independent variable. Revealing gender as a significant variable with a P-value of 0.000. The difference is relatively small with about a 3 % increase in churn likelihood from female customers. However, a linear regression

analysis was also performed to review the interaction of gender and being part of the segment, with results shown in Table 9. The report indicates that churn likelihood increases from being part of the segment, but a female customer is observed to have an increased churn likelihood by another 17.71 %, with a P-value of 0.000, indicating a significant relationship. We also see that for customers who are not part of this segment, gender is no longer a significant predictor, which the churn rates for the remaining data in Table 8 also indicates.

Table 9 – Regression report show relationships between gender and being part of the reduced count segment.

Regression report Attrition_Flag on reduced_count_seg##gender

Source	SS	df	MS	Number of obs	=	10127
				F(3, 10123)	=	540.15
Model	188.435357	3	62.8117858	Prob > F	=	0.0000
Residual	1177.17144	10,123	.116286816	R-squared	=	0.1380
				Adj R-squared	=	0.1377
Total	1365.60679	10126	.134861425	Root MSE	=	.34101

attrition_flag	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
1.reduced_count_seg	.2373736	.0120642	19.68	0.000	.2137254	.2610218
1.gender	.0007223	.0076003	-0.10	0.924	.0156204	.0141758
reduced_count_seg#gender 1 1	.1770586	.0169315	10.46	0.000	.1438695	.2102477
_cons	.0956313	.0055657	17.18	0.000	.0847215	.1065412

The result of this analysis suggests that the difference in churn rates between males and females in the dataset are more prominent for customers within the segment. As the remaining data describe a churn rate for both genders, it could suggest that the differing churn rates observed in the dataset between the genders could be attributed to customers who are part of this segment. However, this does not confirm a causal relationship between reduced count, gender or attrition as other prominent features within the segment could be explanatory. Nonetheless, the results suggest that females with reduced transaction counts churn at rate higher than their male counterparts.

5.4 Hypothesis 3

Customers with a greater number of banking services/products with the bank are less likely to exhibit attrition.

As we shift our focus to Hypothesis 3, we explore a different aspect of customer churn: the influence of relationship banking on customer churn. In this case, we first need to determine what constitutes “a greater number of services” with the bank. The approach

used with selecting a certain percentile does not work as well with distinct numerical variables such as relationship count, as the number of customers can vary greatly within each distinct value. Nor is standard deviation as a measure for selection particularly useful as the feature does not follow a normal distribution.

To create a segment encompassing 20 % of the customer base, mirroring the segments assessed in hypothesis 1 and 2, it was decided to address customers with six or more relationship counts. 1866 customers fit this description, 18.43 % of the customer base. The churn rate of this segment is 10.5 %, with a chi-squared test statistic of 37.39 and a P-value of 9.68e-10, the number of churners in this segment is statistically significantly different from the total customer base, with a lower churn rate.

A relationship count of five or more results in a segment with 11.26 % churn, and a P-value from the chi-square test of 1.59e-12, also statistically significant. This segment has almost 3800 customers, and we decided to go with the lower one. It will be noted that this finding also supports the hypothesis that a higher relationship count is correlated with a lower churn rate. This does, however, further support the findings from both hypothesis 1 and 2, suggesting that a higher relationship count is connected to lower churn rates.

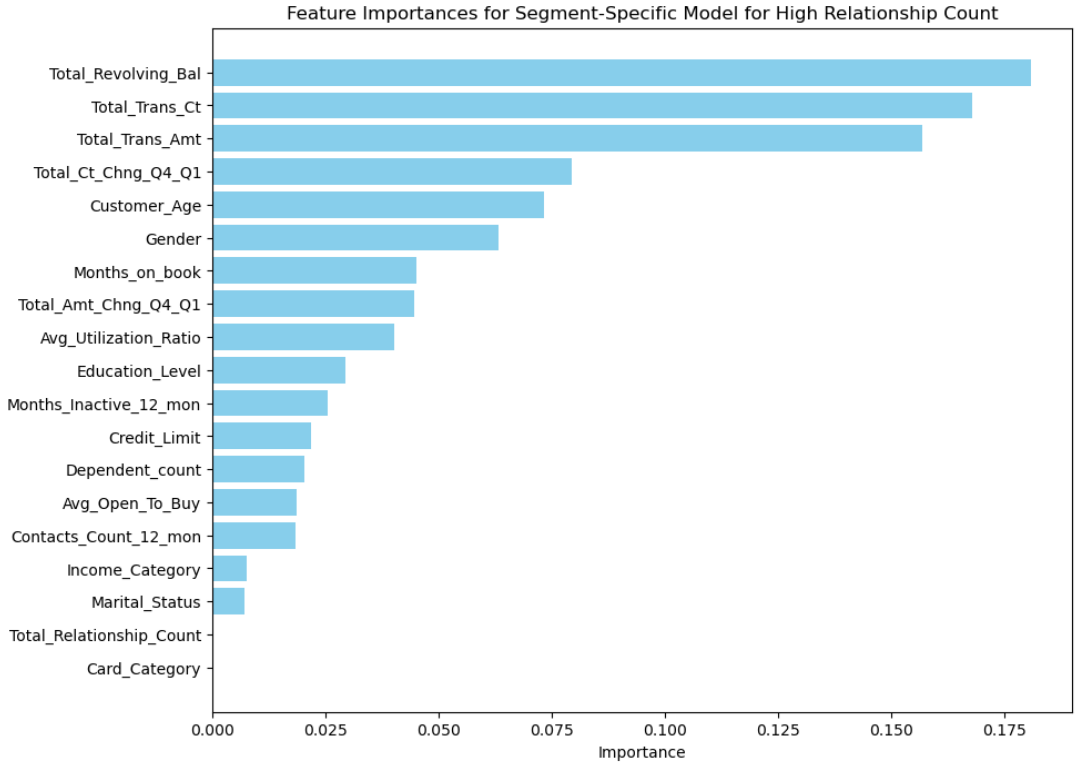


Figure 26 – Feature importance segment-specific model for high relationship segment.

Reviewing the top features presented in Figure 26 from the XGBoost model trained on the segment, total revolving balance, transaction count and -amount are the most prominent features for prediction. This is consistent with the findings from the general analysis, as well as findings from hypothesis 1 and 2.

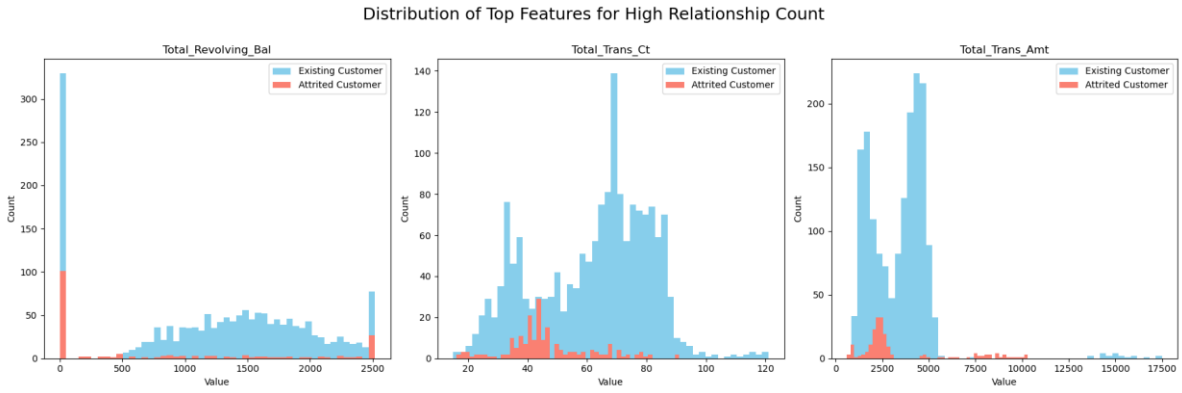


Figure 27 – Feature importance for high relationship count, customer distribution.

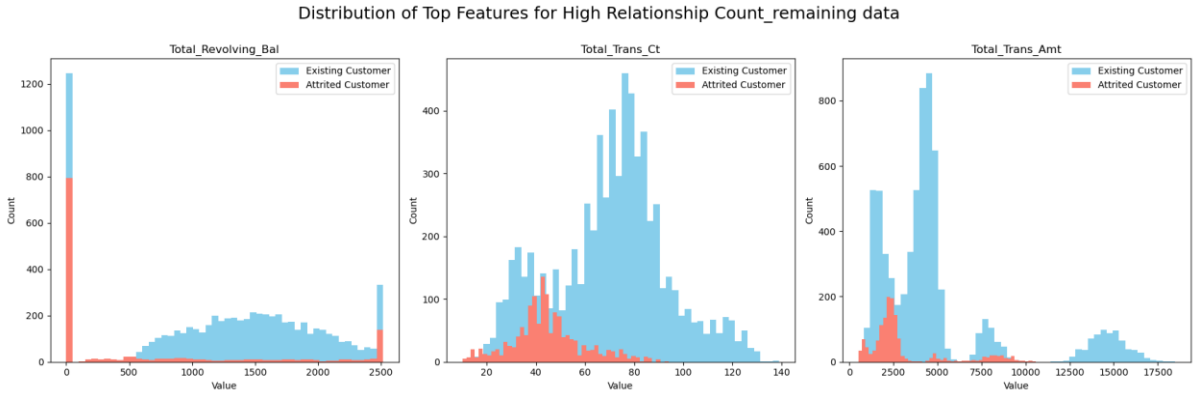


Figure 28 – Distribution of customers in the top three features of the segment specific model for high relationship count. Distribution shown for remaining data.

Reviewing the distribution of attrited and existing customers in the top features of the segment and the remaining data, shown in Figure 27 and Figure 28, all features show an expected reduction in attrited customers in the segment. However, a noteworthy observation is found in the Total_Trans_Amt feature. In the high relationship segment, both existing and attrited customers are concentrated at a total amount below 6 000, while the remaining customers have a notable count above these amounts. Suggesting a correlation between a high relationship count and a lower transaction amount.

Table 10 – Mann Whitney U test comparing distribution of churners in the high relationship segment to the remaining data.

	Variable	Test Used	Statistic	P-value
0	Total_Trans_Amt	Mann-Whitney U	6487489.0000	1.06e-26

A Mann-Whitney U test, as explained in the methodology chapter of 4.9 on page 39, was employed. This test was utilized to compare the distribution within the segment against the distribution in the remaining data. Results shown in Table 10 with a p-value of 1.06e-26 suggests that we should reject the assumption that the distribution is similar. The result supports the visual comparison of the distribution.

5.5 Hypothesis 4

Customers with higher months of inactivity, or a higher number of contacts made to the bank are more likely to churn.

In Hypothesis 4, we confront a different aspect of customer behavior – inactivity of credit card usage, and frequency of contact with the bank. To approach this hypothesis, we again need to define what “high” inactivity and “high” number of contacts means. The approach

of selecting a higher percentile is ineffective for the distribution of the customers within the feature.

Defining high inactivity poses a challenge. In this case, the categorical nature requires a decision between either three or four months of inactivity or more. The authors recognize the inconsistency with selecting segments including 20 % of the customer base. Defining high inactivity at three months and up results in a segment containing 4 583 customers. At four and up, 737 customers are observed with a churn percentage of 24.56 %. Since the feature is of categorical nature, the choice was made to consider the smaller segment.

A chi-squared test comparing the attrition rate to the dataset resulted in a statistic of 35.11 and a P-value of 3.11e-09, suggesting that the attrition rate is statistically speaking significantly higher compared to the dataset.

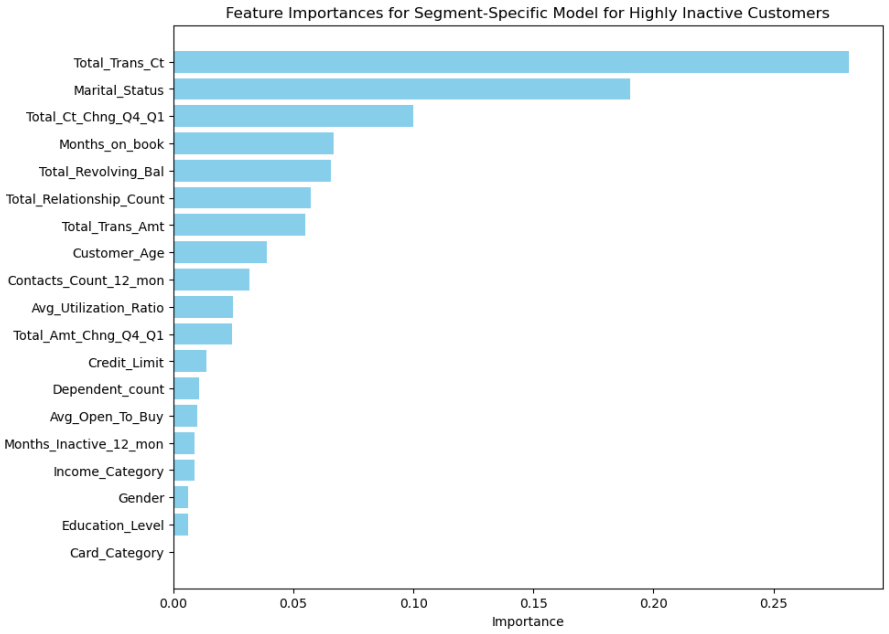


Figure 29 – Feature importance for high inactivity segment, segment specific model.

In Figure 29 the feature importance for the model trained on the high inactivity segment is illustrated, and interestingly, marital status is among the strong predictors. This is counterintuitive to literature on demographic data, which reports that demographics seemingly have little predictive power.

Figure 30 – Feature importance for high inactivity segment, customer distribution.

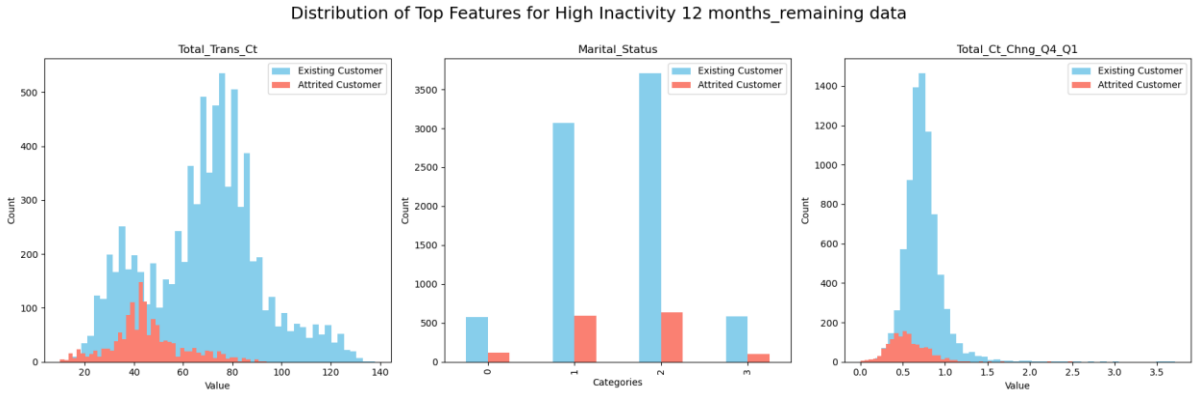


Figure 31 – Customer distribution for the top three features, shown for remaining data.

A comparative analysis of the distribution of customers within the top three features of the high inactivity segment and the remaining data in Figure 30 and Figure 31, the initial observation suggests that marital status, especially 1 and 2 tend toward a higher churn rate. However, by considering the inherently higher churn rate of the segment, as well as the number of observations for the martial status, the results are inconclusive. A goodness of fit chi-squared test for the category compared to expected numbers of observations, given the increased churn rate, gives us a p-value of 0.946.

Transitioning to the high contacts segment. This segment is defined with the same approach as the high inactivity segment. Accounting for the categorical nature, customers with four or more contact counts yields a segment containing 16 % of the customers. The segment has 1622 customers at a 26.39 % churn rate. Running a chi-squared test on the data to compare the attrition rate of segment to the dataset, we get a chi-squared statistic of 102.49, and a P-value of 4.34e-29. This is a low P-value, suggesting the churn rate is statistically significantly different.

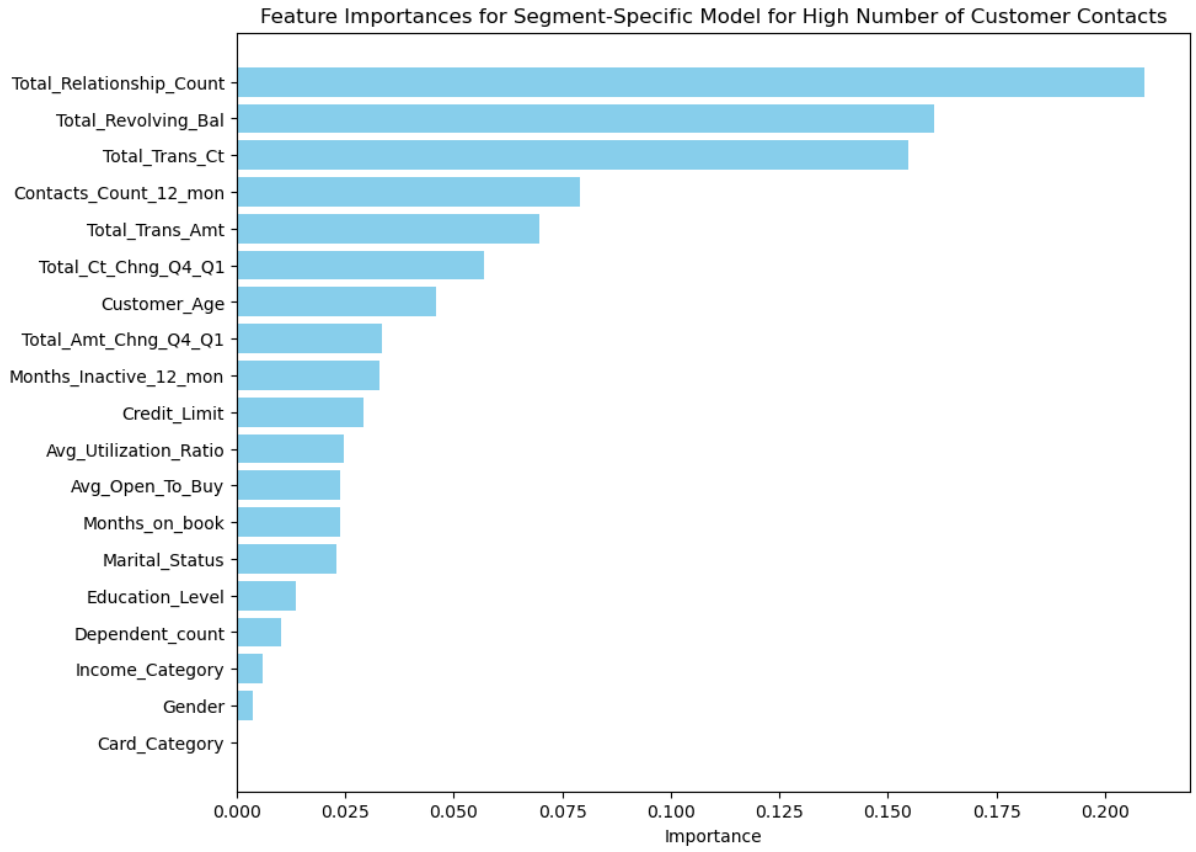


Figure 32 – Feature importance for segment specific model, high contact segment.

Total relationship count, revolving balance and transaction count are the most predictive features according to the segment specific model, as illustrated in Figure 32. Visualization of the distribution of attrited and existing customers within the top three features is shown in Figure 33.

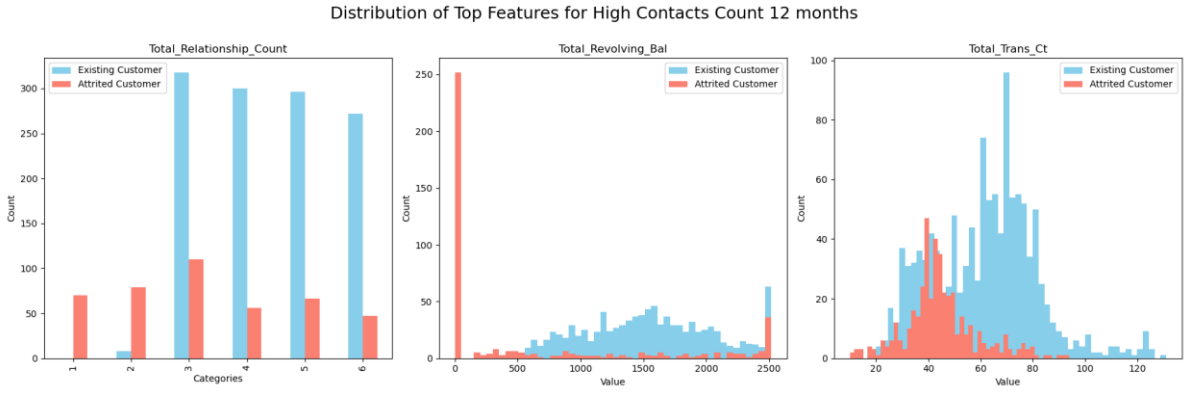


Figure 33 – Feature importance for high contact segment, customer distribution.

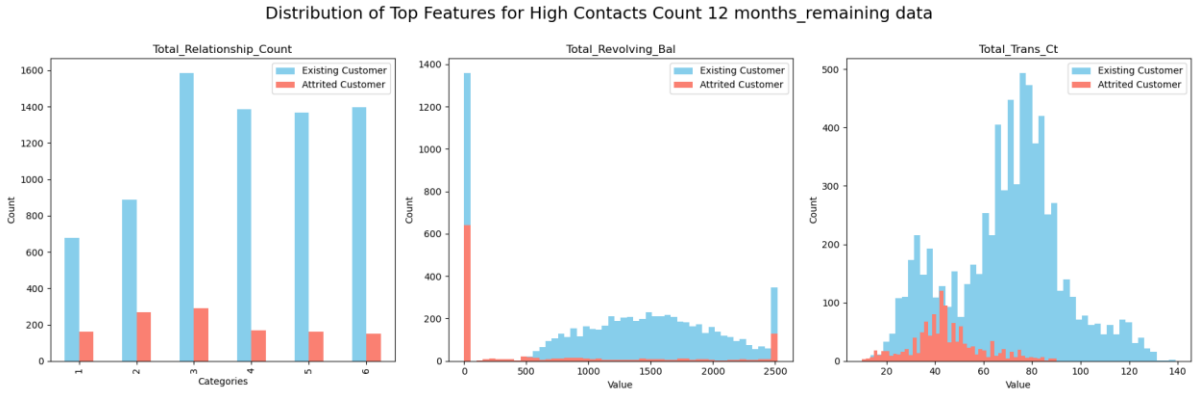


Figure 34 – Distribution of customers within the top three features of the segment specific model for the high contact segment. Distribution for the remaining customers.

Through observation of the customer distribution, comparing the distribution of customers within the segment and the remaining customers, depicted in Figure 33 and Figure 34 respectively, no distinct visual difference is observed apart from a higher concentration of attrited customers in the lower values of relationship count. This finding is consistent with the general analysis, hypothesis 1 and hypothesis 2. As relationship count is observed as a prominent feature in another high-churn segment, it suggests a stronger relationship between the feature and attrition.

In Table 11 the test results from a Mann-Whitney U test comparing the attrition distribution of Total_Trans_Ct feature of the segment to the remaining data results in a p-value of 7.21e-55, suggesting the difference in distribution we see is statistically significant. Most notably, the transaction count appears to have a higher concentration of churners towards the lower end, as well as there being fewer customers overall with higher transaction counts. It is also important to consider the increased churn rate of the segment when reviewing the differing distributions. A higher churn rate and as such a concentration of attrited customers is expected.

Table 11 – Mann-Whitney U test comparing distribution of attrition in transaction count in the high contacts segment to the remaining data.

	Variable	Test Used	Statistic	P-value
0	Total_Trans_Ct	Mann-Whitney U	5214360.5000	7.21e-55

Reviewing some of the demographics for the segment also revealed a slightly higher count of male customers in relation to female, even when accounting for the higher female count in the

overall data. The results of the chi-square test are shown in Table 12. What the results might suggest is that males might contact the bank more often.

Table 12 – Comparing the distribution of males and females in the high contacts segment.

Chi-squared test

	High Contacts Segment		
Gender	Not part of	Part of	Total
0	3927	842	4769
1	4578	780	5358
Total	8505	1622	10127
Pearson chi2(1) =		180.034	Pr = 0.000

Females in the segment did also churn at a higher rate in the high contacts segment as they did in the reduced count segment, although not as prominent. The attrition rate in this segment is also a little lower. Females do generally churn at a slightly higher rate in the dataset.

Table 13 – Churn rates in the high contacts segment for gender compared to remaining data.

Churn rate	Segment	Remaining data
Male	23.16 %	12.78 %
Female	29.87 %	15.22 %

The high contacts segment indicates that males might be more prone to having more contacts with the firm. Churn rates are higher in the segment, suggesting correlation with churning. The analysis of the high contacts segment supports the hypothesis that the relationship count is indicative of lower churn rates.

5.6 HYPOTHESIS 5

Different age groups exhibit distinct patterns in credit card usage, which in turn impacts customer churn.

Hypothesis 5 shifts our focus to another critical dimension of customer behavior – the influence of age on credit card usage patterns. A structured approach to this analysis would involve segmenting the customer base into distinct age groups. Commonly used age brackets include young adults (e.g., 25-36), middle-aged adults (36-45), adults (46-55) and older adults (56-65). This results in four distinct groups, shown in Table 14. The count

within each group varies between 919 and 4 135. Notice that group 66+ is excluded, due to having only seven customers.

Table 14 – Segments for Hypothesis 5

Segment age group	N	Churn rate
25-36	919	13.28 %
36-45	3 742	16.19 %
46-55	4 135	16.64 %
56-65	1 321	15.82 %

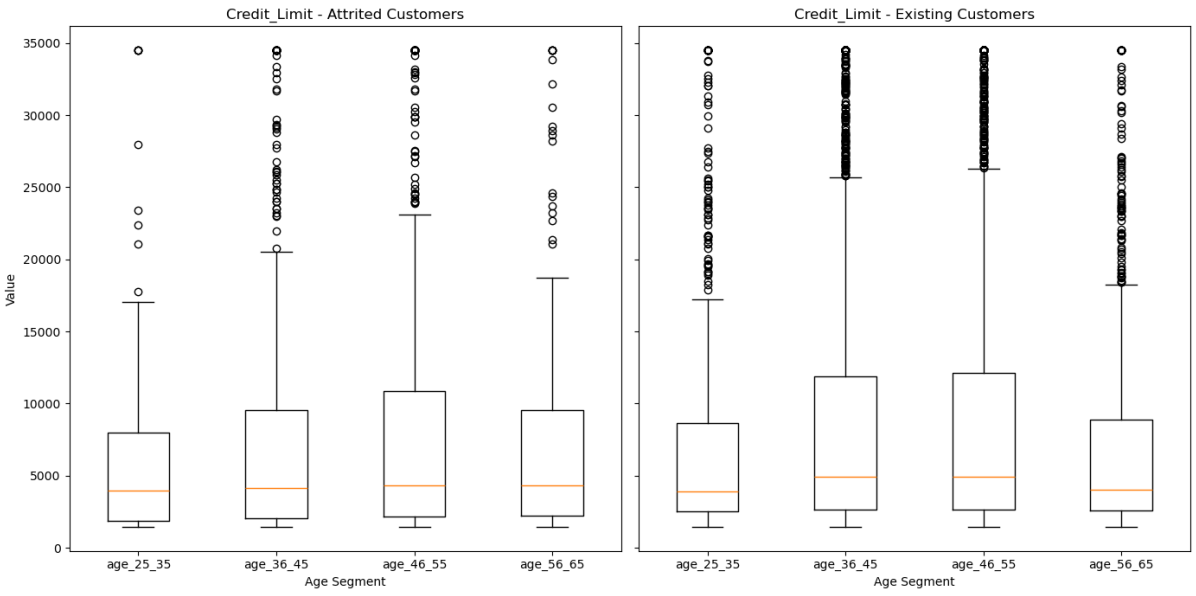


Figure 35 – Box plot comparison of age segments. The feature reviewed is Credit_Limit.

One of the features we might expect to see a difference in is the credit limit given in each group. Figure 35 illustrates the distribution of customers within the credit limit feature for the different age groups. Most notable in the credit limit comparison is the similar medians while the middle segments have more people with higher limits for the existing customers, while attrited customers seem to have a similar median. Although there seems to be more outliers in the existing customers, the number of observations is higher and could be arbitrary.

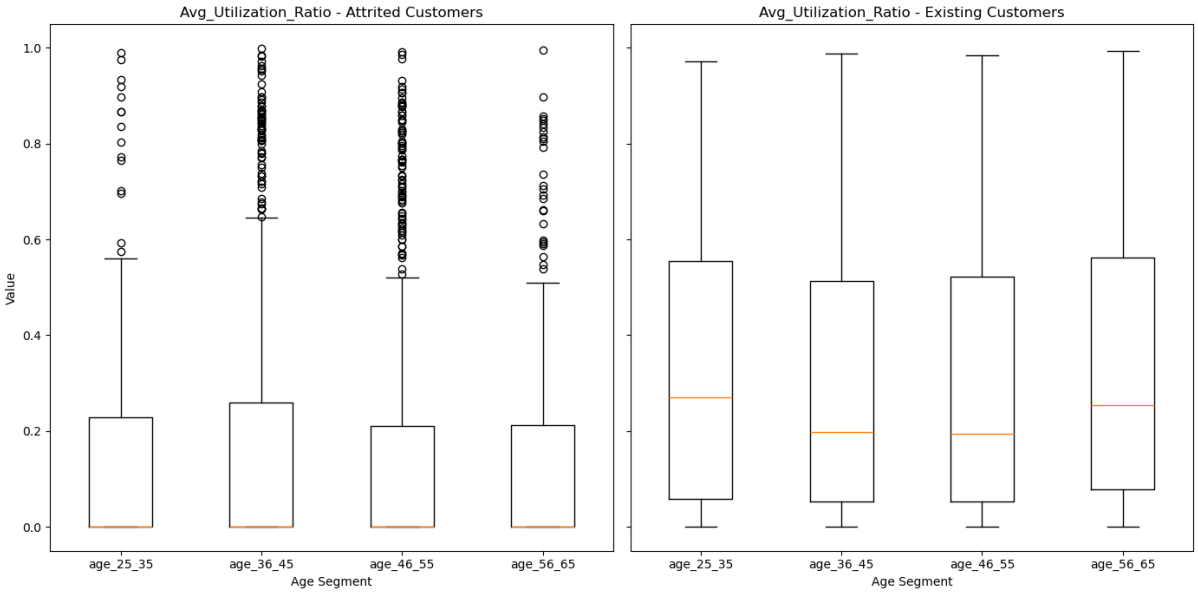


Figure 36 – Box plot comparison of age segments. The feature reviewed is Avg_Utilization_Ratio.

Regarding utilization ratio, a notable difference is observed between attrited customers and existing customers, where the median is 0 for all segments for attrited customers, while existing customers have a median around 20 %, illustrated in Figure 36. For further analysis, customers with zero utilization was excluded, resulting in a different distribution, shown in Figure 37.

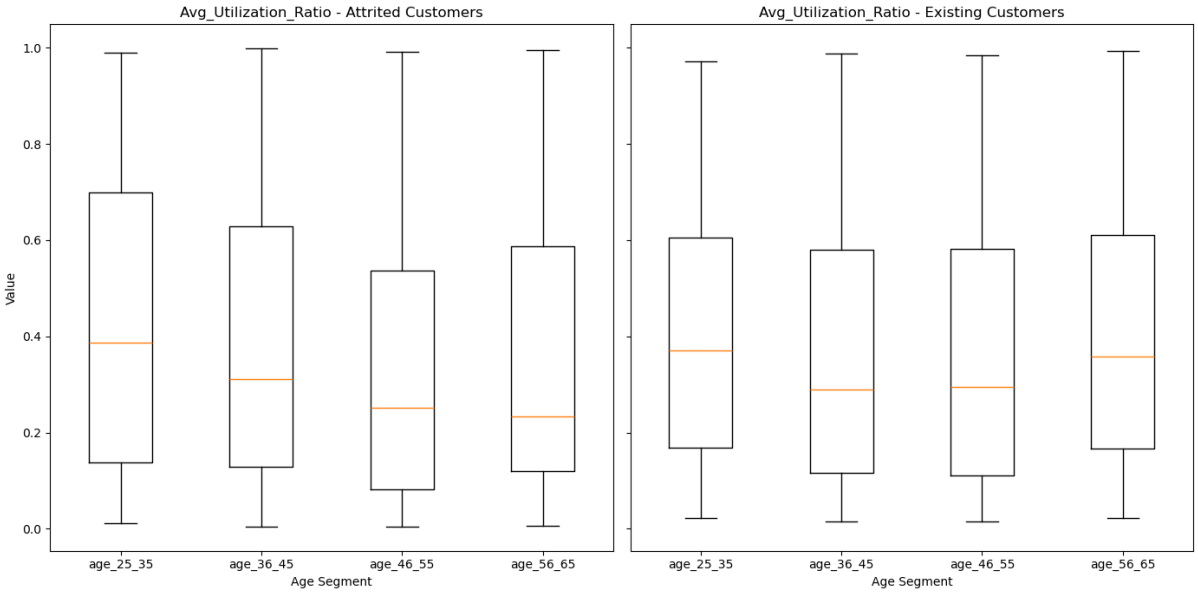


Figure 37 – Box plot comparison of age segments. The feature reviewed is Avg_Utilization_Ratio excluding customers with 0 average utilization ratio.

All age groups have an attrition rate of over 30 % for the group with zero utilization, and around 10 % for the groups with utilization above zero. Utilization ratio is slightly negatively correlated with attrition, meaning a higher utilization ratio tends to correlate with a lower churn rate. The correlation is weak, at -0.18, but the results from the age group analysis support the finding. Although, the linear negative correlation could be attributed to the higher churn rate within the zero-utilization segment, disregarding potentially important correlations existing in the continuous specter. Another notable observation is that females have a higher churn rate in the dataset, but female customers also tend to have a higher utilization ratio which is correlated with lower churn.

Reviewing the age groups in the context of usage reveals a differing distribution for the transaction count feature. Figure 38 illustrates the age groups, with existing customers to the left and attrited to the right. Specifically, attrited customers have a significantly lower median for transaction count. The youngest segment exhibits a slightly lower median count in parallel with the oldest group. Attrited customers in the youngest segment differ slightly by having a higher transaction count, suggesting transaction count is a weaker predictor among younger customers. However, this requires the model to account for outliers in the other age groups.

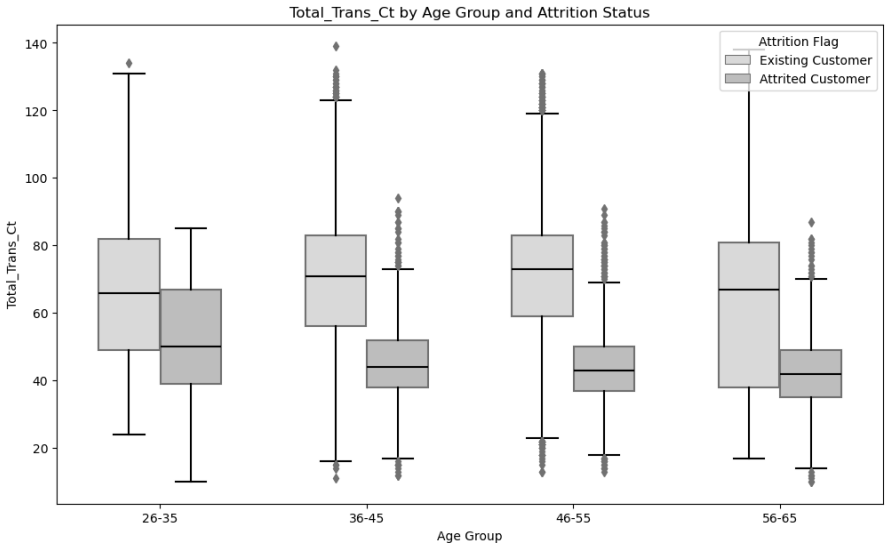


Figure 38 – Box plot showing transaction count between the age groups for existing and attrited customers.

Through the analysis of the different age groups, the most notable observation was the tendency of attrited, younger customers to exhibit higher transaction counts, suggesting that transaction count is a less effective feature for this specific group.

6. DISCUSSION

The purpose of this thesis was to contribute to research on customer churn by studying the predictive indicators and behavioral patterns that signal a customer's likelihood to discontinue service. This approach of analyzing customer relationships and behaviors in the credit card industry has yielded several intriguing insights. In what follows, we will discuss the findings of the hypotheses and then present potential strategies derived from these findings. The strategies are conceptualized based on the insights from the analysis part in Chapter 5. Each strategy aims to address specific customer behaviors and patterns, with the goal of enhancing customer retention and customer lifetime value.

Finally, we will detail the steps for the implementation of these strategies in churn intervention campaigns and propose directions for future studies. This will provide a clear roadmap for utilizing our insights to mitigate customer churn.

6.1 HYPOTHESIS 1

Hypothesis 1: Low usage rate of credit cards is predicative of churn.

On our analysis of hypothesis 1 in chapter 5.2 on page 50-52, we confirmed that a lower credit card usage rate is indeed correlated with a higher churn rate. However, it is still not clear whether low usage rates are a result of an already slipping customer, or that low usage rates cause people to cancel their card, or if other factors such as dissatisfaction could cause low usage. Engaging customers to increase use could be a good starting point to address this finding. Another interesting finding is that relationship count, especially from four counts and up seems to reduce churn drastically.

A reasonable assumption for most services is that a lower usage rate correlates with customer churn. Although it cannot be claimed as truth for all customers, lower usage could indicate less necessity of the service. Less frequent usage may also be associated with a diminished sense of loyalty and a reduced barrier to switching providers. This is where the significance of the relationship count becomes evident in supporting the claim. A higher number of relationships could indicate a stronger loyalty to the business, thereby increasing the barrier to switching. Moreover, the impact of relationship count in mitigating churn underscores the importance of customer engagement. A higher relationship count might reflect deeper integration of the service into the customer's financial habits, which could translate into higher perceived value and customer

satisfaction. This observation raises questions about the nature of these relationships: Are they indicative of genuine loyalty or simply a result of convenience? Do they reflect a customer's satisfaction with the service, or are they a consequence of inertia?

Feedback solicitation is crucial in understanding the barriers faced by low-usage customers. This feedback can then be applied to create tailored marketing strategies and product enhancements to better align with customer needs. This strategy can be tested by either measuring the churn rate before and after implementing the engagement strategy, or by measuring the usage rate of their credit card before and after.

While the connection between low usage rates and customer churn is evident, the complexity of this relationship calls for a deeper investigation into the various factors influencing customer behavior and decisions. Understanding these connections will be crucial in comprehensively analyzing customer churn patterns.

Potential Strategy from Findings of Hypothesis 1

Based on the findings from Hypothesis 1, a potential strategy is to engage customers who have a low number of transactions. This can be done by sending emails and directly contacting those customers who exhibit a low usage rate. The aim is to discover the reasons behind their infrequent use of their card. To encourage increased usage, it could be considered to implement bonuses such as waiving the card fee or increasing the reward rate, especially if they are customers of a competitor. This could incentivize them to use the card more often. Additionally, it's important to remind these customers about the benefits of the card. Previous results have indicated that customers with a low usage rate tend to have a higher churn rate, making it crucial to address this group specifically.

The findings from Hypothesis 1 reveal that within this segment, customers with a lower number of relationships with the business tend to have higher churn rates. Notably, the churn rate decreases progressively from one to three relationships, then stabilizes. This suggests that the business could potentially benefit in the long term by offering incentives to encourage customers to expand their engagement to four or more relationships, rather than stopping at three or fewer. It would be crucial to tailor these offers based on customer demographics to maximize the likelihood of uptake.

However, the precise reason why having four or more relationships appears to reduce the likelihood of churn is not entirely clear. It's uncertain whether this reduced likelihood of

churn from other sources leads to an increased chance of purchasing additional services, or if another causative factor is at play. While we can speculate that enhanced customer satisfaction might lead to reduced churn, further investigation is needed to understand this correlation fully.

One of the central elements of this strategy is direct communication with customers. By reaching out through emails and personalized contacts, the bank can gain valuable insights into why certain customers are not utilizing their cards as frequently. This approach is rooted in the principle of customer-centricity, prioritizing an understanding of individual customer needs and behaviors.

In a paper by Shini Renjith, it is stated that customers who are more prone to attrite need to be handled in a more personalized manner. She lists various approaches that can be adopted to generate personalized actions (Renjith, 2015):

- **Top-down Approach** – Based on managerial level experience from prior actions, actions are defined independent of customer profiles.
- **Bottom-up Approach** – Customer profiles are defined prior to defining actions, but requires considerable amount of supervision effort.
- **Customized Approach** - The offers have to be granular enough and the customers are free to choose from the wide set of alternatives.
- **Similarity-based Approach** - Actions are triggered based on customer preferences inferred from customer profiles.

Implementing bonuses or incentives serves as a crucial tactic. For example, offering to remove card fees or increasing reward rates, particularly for those customers who may be using a competitor's card, could effectively incentivize increased usage. Such incentives are not just about immediate gratification for the customer; they also communicate the bank’s commitment to providing value and acknowledging customer preferences. The implementation of bonuses, rewards, or incentives has a significant impact on customer churn, as detailed in the paper "Are You Back for Good or Still Shopping Around? Investigating Customers’ Repeat Churn Behavior." (Kumar, Leszkiewicz, & Herbst, 2017). The study indicates that such strategies can effectively mitigate repeat churn behavior, thereby extending the second lifetime tenure of customers. This approach not only helps retain customers but also increases profitability. Although in the paper written

by Eva Ascarza, she argues that the customer lifetime value must be estimated before determining if the incentive of a reacquisition is worth it. We also need to factor in customers who will pretend to churn to gain the implemented incentives (Ascarza, et al., 2017).

6.2 HYPOTHESIS 2

Hypothesis 2: Reduction in credit card usage is highly indicative of customer churn.

In the analysis of hypothesis 2 starting on page 53, it was confirmed that reduction in credit card usage is highly indicative of customer churn. With a churn rate exceeding 42%, it is crucial for the business to target this segment extensively. Reduced usage can act as a kind of "forewarning", which a low transaction count alone fails to provide, indicating that a customer is increasingly likely to churn as their usage diminishes. Although this segment represents just about 20% of the dataset, it accounts for a significant 52.6% of all churners within the entire dataset. However, it is crucial to remember that reduced usage could stem from various other factors that contribute to churn.

Additionally, it became evident that churn within this segment is more pronounced among females compared to males, even after accounting for both the higher churn rate of the segment and the overall higher churn rate among females in the dataset. This suggests that general usage might be more critical to female customers, making a reduction in usage an even stronger predictor of churn. Given that the segment with reduced transaction count is characterized by decreased usage, and females are approximately 17% more likely to churn within it, this appears to be a plausible assumption. However, the underlying reasons for the reduction in usage remain unexplained.

There are also additional factors contributing to this pattern observed in the business. This could be factors such as advertising or campaigns failing to effectively resonate with female customers, or some alteration in the customer experience that disproportionately affects females more than males. However, to empirically investigate these possibilities, the business would need time series data to examine changes from the date of implementation of these strategies or changes. The reason for this disparity in churn rates is not immediately clear from the current data, but it is imperative for the business to explore why their female customers are

churning at a higher rate than their male counterparts, particularly among those with reduced usage.

In the analysis of Hypothesis 2 in chapter 5.3 on page 52, the feature importance within the XGBoost model once again highlighted relationship count as one of the top predictors. Echoing the findings from Hypothesis 1, an increase in relationship count corresponded with a decrease in churn rates for each increment. Notably, from four relationships onwards, the reduction in churn rate appeared to plateau, indicating that beyond this point, churn rates were not significantly affected by further increases in relationship count. This plateau in churn rate reduction beyond four relationships suggests a threshold effect, where the benefits of additional relationships in terms of reducing churn become less pronounced. This finding is crucial for the business as it implies that while fostering customer relationships is beneficial, there is a point beyond which additional relationships may not contribute significantly to decreasing churn. This insight can inform customer relationship management strategies, indicating a potential shift of focus from merely increasing relationship counts to deepening the quality and engagement level of existing relationships.

Moreover, the significance of relationship count as a predictor in both Hypotheses 1 and 2 underscores its role in customer retention strategies. It highlights the importance of understanding not just the quantity but the nature of these relationships. Are they based on different products or services, or do they represent various types of interactions with the business?

The findings from the analysis of Hypothesis 2 in chapter 5.3 complement those from Hypothesis 1 in chapter 5.2, painting a more comprehensive picture of the factors influencing customer churn. They emphasize the need for a nuanced approach to customer retention that considers both the quantitative aspects, like relationship count and credit card usage, and the qualitative aspects, like the nature of customer relationships and individual customer experiences. The analysis of Hypothesis 2 reinforces the understanding that customer churn is a multifaceted issue. It requires a strategic approach that considers various predictive factors, including relationship count and usage patterns. By integrating these insights into their retention strategies, the business can more effectively address the challenges of customer churn, especially in segments identified as high risk.

Potential Strategy from Findings of Hypothesis 2

In Hypothesis 2, it was confirmed that customers with a reduction in their credit card usage have a high churn rate of 42%. This finding in the dataset opens a window for further proactive churn prevention, expanding upon the first strategy outlined on page 72. By identifying customers who have recently decreased their card usage, the business can intervene at an early stage in the churn process. Such early engagement is vital as it occurs while the customer is still in the preliminary phases of potentially leaving the service. By effectively targeting these individuals, understanding and addressing their specific concerns or evolving needs, there is a substantial opportunity to mitigate churn risks and retain these customers.

Although the exact causes of reduced credit card usage remain uncertain, we consider these types of customers to be an effective targeting audience in churn intervention campaigns. Combining this targeting with implementing extra bonuses or incentives, as discussed in strategy 1, it could serve as an effective way to increase card usage. In these communications, it is crucial to highlight the benefits of continuing to use the credit card, perhaps by reminding them of the rewards or advantages they are missing out on. This is where combining the targeting with extra bonuses or incentives, becomes powerful. For instance, offering a limited-time cashback bonus or extra reward points for certain types of purchases can motivate customers to increase their credit card usage.

Furthermore, feedback should be actively sought from these customers to understand the reasons for their reduced usage. This feedback can be invaluable in not only retaining the customer but also in improving overall service and product offerings.

6.3 Hypothesis 3

Hypothesis 3: Customers with a greater number of banking services/products with the bank are less likely to exhibit attrition.

In Hypothesis 3, examined in Chapter 5.4 on page 56, our analysis provided confirmation of the trends initially suggested in Hypotheses 1 and 2. We discovered a clear connection: a higher number of customer relationships with the business is indeed associated with lower churn rates.

Building on this key finding, it becomes evident that fostering and maintaining multiple touchpoints with customers is a critical factor in reducing the likelihood of churn. The presence of multiple relationships suggests a deeper engagement and reliance on the business's services, which in turn creates a stronger bond and a higher perceived value from the customer's perspective.

The XGBoost model identified transaction amount as one of the predictors in this segment, indicating that customers with higher transaction amounts are less common in the high relationship count segment. This is somewhat unexpected, given that increased transaction amounts are slightly negatively correlated with churn, and we are analyzing a segment with low churn rates. The relationship between attrition and transaction amount, however, may be more intricate. Customers with higher transaction amounts could primarily be using the credit card service, without much need for other services offered by the business. Conversely, customers with a high relationship count might view the credit card as an ancillary service and therefore use it less frequently.

Understanding whether a customer initiated their relationship through the credit card service or another offering could be crucial in deciphering their transaction behavior. Conducting surveys with these customers could yield valuable insights. Gaining a deeper understanding of a customer's journey, such as the initial service they used and the channel through which they learned about the business, could significantly enhance the refinement of targeted strategies.

The implications of this discovery are multifaceted. Firstly, it highlights the importance of cross-selling and upselling strategies in customer retention. By encouraging customers to engage with a variety of products and services, the business not only increases revenue opportunities but also enhances customer stickiness.

Potential Strategy from Findings of Hypothesis 3

A potential strategy would be to target and acquire the bank’s already existing customers. Targeting the bank’s current customer base who presently do not have a credit card can be effective, as the customers who use multiple banking services have a lower churn rate. The paper "Cross-Selling the Right Product to the Right Customer at the Right Time" by Li et al. (2011) examines the profitability of cross-selling in banking. It proposes a dynamic customer response model, considering customer demand evolution, cross-selling's

multifaceted roles (such as promotion, advertising, and education), and customer preferences for communication channels. The study demonstrates that tailored and dynamic cross-selling strategies significantly improve profitability. These strategies increased immediate response rates by 56%, long-term response rates by 149%, and long-term profit by 177%, indicating a substantial positive impact on profitability when cross-selling is executed effectively (Li, Sun, & Montgomery, 2011).

Results from Hypothesis 3 also suggests that deepening customer relationships can be an effective strategy in reducing customer attrition. This can for instance be done through introducing a diverse portfolio of products and services to the customer. The paper "Customer Retention through Customer Relationship Management: The Exploration of Two-way Communication and Conflict Handling" by Mornay Roberts-Lombard, investigates how customer relationship management (CRM) influences customer loyalty in South African short-term insurance organizations. It emphasizes two-way communication and conflict handling as significant factors in CRM. Overall, CRM, when implemented with proper conflict handling and balanced communication, leads to increased customer loyalty and retention (Roberts-Lombard, 2011).

The rationale behind this strategy is rooted in the principle of relationship banking, where the depth and breadth of a customer's relationship with a bank are directly correlated with their retention. Customers who use multiple products are more integrated into the bank’s ecosystem, often perceiving higher value and convenience, which can lead to increased loyalty and reduced churn. It is still not clear whether having more banking services make customers less prone to churning, or that higher customer satisfaction leads to acquiring more services. In economic terms, the axiom that customers behave rationally could imply that higher satisfaction leads to more services as the most logical path. Either way, providing tempting offers for more services could “tie down” the customer more. Keeping satisfaction high would still be imperative.

Another paper “Relationship Banking and Profitability - An Empirical Survey of German Banks" finds that relationship banking, characterized by the use of qualitative information and closer customer-bank relationships, is positively correlated with higher profitability for banks. The study, based on a survey of 129 German banks, indicates that banks more focused on relationship banking aspects tend to have higher return on equity (ROE). Banks

in the relationship group, which emphasize customer relationships and qualitative information in their credit processes, generally perform better than those with a transaction-focused or ambiguous process design. The paper underscores the positive impact of relationship banking on bank performance, especially in terms of profitability (Schwarze, 2007).

6.4 HYPOTHESIS 4

Hypothesis 4: Customers with higher months of inactivity and higher number of contacts made to the bank are more likely to churn. Results from the analysis of hypothesis 4 on page 63, customers with higher months of inactivity and customers with higher number of contacts have a statistically significant higher churn rate.

From the high inactivity segment, XGBoost did suggest marital status as an important predictive feature. The correlation between being part of the segment and having certain a marital status did not seem to be predictive of churn. Marital status 1 and 3 had higher rates of churn in the segment, but the sample size was too small to infer a relationship. The analysis did also suggest transaction count and transaction count change as predictors. From visual representation of the distribution within the segment and the remaining data, churning customers were again found towards the lower end.

From the high contacts count segment, the analysis revealed a statistically significant higher proportion of male customers within the segment. The churn rate of the segment was also found to be higher. Females, as within the reduced count segment discussed in hypothesis 2, churn at a higher rate within the segment, but the difference is not as large as in the high contacts segment.

Males being the more prominent gender in this segment in addition to the segment having a higher churn rate might suggest males are less withheld when it comes to contacting the business. Whether the nature of the contact is to complain, inquire about services or something else is unknown. That women are less represented in the segment could imply overall higher customer satisfaction. This would, however, lead us to question what other than customer satisfaction can explain the higher churn rate for women.

According to the analysis, higher contact is certainly correlated with churn. However, while the number of contacts is important, noting the nature of the contact, perhaps with a categorical keyword such as “Complaint”, “General question”, “Assistance”, and whether it was solved could give even better insights. Understanding how customer contacts are related to customer churn is important. Does the business have to improve customer service? Are there product- or service-related issues that customer service cannot provide answers to, or that the business is unable to improve on?

Potential Strategies from Findings of Hypothesis 4

Another potential strategy that utilizes the findings of hypothesis 4 is to tailor outreach and solutions to customers who have a high number of inactive months or to customers who have contacted the bank on several occasions. As analyzed in Chapter 5, inactivity is a strong predictor for churn, and customers that are defined as highly inactive have a churn rate of 24.56 %, and those defined as having a high number of contacts have a churn rate of 26.39 %.

A high contact count to customer service, on the other hand, may indicate unresolved problems for the customer, or a potential dissatisfaction about the card or the bank’s customer service. It therefore might be effective to allocate more customer service agents to reach out to these specific customers.

By understanding the specific reasons behind their dissatisfaction, high contact count or inactivity, whether it's pricing, service issues, or other factors, companies can tailor their responses, offerings and marketing campaigns. This targeted approach can be more effective in addressing the unique needs and concerns of each customer, potentially reducing churn and fostering better customer relationships.

The paper "Are You Back for Good or Still Shopping Around? Investigating Customers’ Repeat Churn Behavior" focuses on understanding the reasons behind customer churn and repeat churn in the context of a telecommunications provider. The study identifies various reasons for customer defection, noting that while some are beyond the firm's control, the most common critical incidents leading to customer churn involve pricing and service issues. These include direct service failures, unsatisfactory service encounters, and the firm's response to service failures. It also finds that offering tailored customer service and solutions can significantly mitigate repeat churn behavior. This approach extends the

second lifetime tenure of customers and increases profitability. Specifically, the study suggests that addressing repeat churn behavior can lead to an increase in profitability by approximately $150,000 over the lifetime of customers in the sample. This translates to gains of over $15 million for deferring second lifetime churn in a million returning customers, depending on the type of churn (Kumar, Leszkiewicz, & Herbst, 2017). Although further research is needed to analyze the triggers of inactivity.

The analysis indicates that a period of 3-4 months of inactivity serves as a critical threshold for increased customer churn risk. Developing a proactive retention strategy targeting specific customer segments around this time frame could also be highly effective. Tailoring the mode of communication based on the preferences of different segments could enhance the impact of these strategies. For instance, younger customers might respond more positively to text messages, appreciating their convenience and immediacy. On the other hand, older customers may appreciate the personal touch of a phone call, which can convey a sense of importance and personal attention. Additionally, understanding the reasons behind the inactivity in various segments could further refine these strategies. For example, if a segment shows inactivity due to dissatisfaction with services, personalized offers or service upgrades could be an effective approach. Meanwhile, for those who are inactive due to lack of awareness of available products, educational or informative messages might be more appropriate.

6.5 HYPOTHESIS 5

Hypothesis 5: Different age groups exhibit distinct patterns in credit card usage, which impacts customer churn.

This hypothesis stems from the premise that different age groups, shaped by their unique financial behaviors and life stages, might engage differently with credit card services. As such, this section of our thesis aimed to uncover whether younger, middle-aged, and older customers exhibit distinctly different patterns in their credit card utilization, credit limits, and transaction frequencies. While the analysis did not show any clear distinguishable patterns in usage between the groups, apart from marginally higher utilization ratio amongst younger individuals, and slightly higher usage and credit limits in the middle group, it did yet again confirm the correlation between transaction count and attrition.

A customer exhibiting a higher utilization ratio is not inherently indicative of more frequent card use compared to one with a lower ratio. The utilization ratio essentially measures the revolving balance in relation to the available credit limit. The tendency of younger customers to have a higher utilization ratio might imply that they are less prompt in settling their debts compared to other age groups. This could be interpreted as a sign of reduced financial responsibility, although such a conclusion is not definitively supported by our analysis. For instance, younger individuals often have lower credit limits due to shorter credit histories and potentially lower incomes at the start of their careers. This situation naturally leads to higher utilization ratios, even with similar or lower spending patterns compared to other age groups. Additionally, middle-aged customers typically exhibit greater financial stability and creditworthiness, which is reflected in their higher credit limits and increased transaction frequencies.

The age brackets of 36-45 and 46-55 appear to have, on average, a marginally higher credit limit. This increased limit could be contingent upon the business's policies allowing customers to modify their credit limits, potentially influenced by creditworthiness evaluations. Older customers typically have had more time to establish a robust credit history and may possess a more stable financial life characterized by well-defined expenses and incomes. Notably, the median credit limit for customers who have left the bank is slightly lower across all age groups. This raises a question: Are customers who do not receive an enhanced credit limit from the bank more inclined to switch to alternative service providers?

6.6 Designing an Optimal Churn Intervention Campaign

As we embark on Chapter 6.6, we will utilize previous insights and strategies to tailor an optimal churn intervention campaign. Drawing insights from "In Pursuit of Enhanced Customer Retention Management" by Eva Ascarza et al., this subchapter aims to design a churn intervention campaign that encompasses aspects such as campaign type, targeting strategies, incentive models, timing, and result analysis, in accordance with the recommended strategies.

Having identified which customers are at risk of churning, and why they are at risk, the first step to assess is the type of campaign. Eva Ascarza argues that there is a distinction between reactive campaigns and proactive campaigns. Reactive campaigns involve a wait-and-see tactic where the company acts only after a customer has churned, often using financial incentives to

'win back' the customer. In contrast, proactive campaigns are about preemptive action, addressing the root causes of churn before the customer decides to leave. Reactive campaigns are relatively straightforward because the at-risk customers are self-evident – those who contact to cancel. These campaigns allow for clear calculation of 'rescue rates' to assess their effectiveness and enable monitoring of customers' subsequent actions.

Gupta et al. introduced the concept of measuring company profits as a function of the total customer lifetime value (CLV). Their predictive model allows for targeting specific classification metrics by adjusting the threshold value. This becomes particularly relevant when a business considers the high cost of losing a customer and decides to balance the prediction model based on the importance of specific classifications. For instance, in churn prediction, the cost of false negatives could be substantial. This consideration may differ if the retention strategy of the business is relatively inexpensive.

The incentives to regain customers in these scenarios are usually substantial, due to the certainty of impending churn. However, reactive campaigns come with challenges. Not all customers can be retained, and there's a risk of customers learning to leverage their churn threat for high-value rewards, potentially undermining the long-term viability of such strategies. Conversely, proactive campaigns pose greater complexities, starting with the identification of at-risk customers. Advanced analytics are essential to minimize the risk of false positives (wrongly targeting customers not intending to leave) and false negatives (overlooking customers who are likely to churn)**.** These campaigns also have to weigh the likelihood of successfully retaining identified potential churners, requiring careful strategy and execution (Ascarza, et al., 2017).

The second step is to assess which customers to target. Targeting customers at the highest risk of not being retained might seem logical, but it's not always the most effective strategy. Highly at-risk customers may be so disillusioned with the company that they are beyond the reach of retention efforts. A more viable approach is to focus on customers who are considering leaving but could potentially be persuaded to stay (Ascarza, et al., 2017).

However, not all customers who are likely to respond to retention initiatives should be targeted. Some may have a lower lifetime value, making the cost of retention outweigh the benefits. It's crucial for companies to evaluate the profitability of retention actions, considering factors like

the likelihood of churn, response to the campaign, customer lifetime value (CLV), and the cost of incentives (Ascarza, et al., 2017).

In proactive campaigns, due to imperfect prediction models, there's a chance of targeting customers who would have remained loyal without intervention. This needs to be factored into the targeting decision (Ascarza, et al., 2017). For instance, (Berson, Smith, & Thearling, 2000) found that customers who declined a retention offer had a higher churn rate afterwards, possibly because the offer prompted them to reconsider their loyalty. This study opens up the debate on whether retention offers separate satisfied customers from dissatisfied ones. Ascarza, Iyengar, and Schleicher (2016) also showed that retention efforts might unintentionally prompt loyal customers to leave. Some customers continue their association with a firm out of habit or inertia. However, when these "habitual non-churners" are targeted with retention campaigns, it can disrupt their routine, lead them to question their satisfaction with the service, and ironically result in churn. On the flip side, such campaigns can also reinforce their loyalty and strengthen their habitual choice to stay with the company. Therefore, the impact of retention efforts on these customers can be double-edged and needs careful consideration (Ascarza, et al., 2017).

Additionally, challenges may arise in selecting the appropriate customer segment for targeting. Some retention campaigns might be ineffective with certain groups. An example of this would be using TikTok advertisements for customers over 60 years old. The success of a retention campaign can generally only be assessed post-implementation, although it can be guided by certain principles.

Thirdly, the type of incentives for the campaign must be selected to align with the goal of the campaign and the marketing strategy. One strategy is to understand the reasons behind customer churn and tailor retention efforts accordingly. An important decision in this process is choosing between price and non-price incentives. While price incentives might be effective initially, they can be easily replicated by competitors (as seen in the current telecom industry) and encourage customers to always look for the best deal. Therefore, non-price incentives, like product enhancements (for example, a gaming company introducing new levels in a game), may be more effective for long-term retention (Ascarza, et al., 2017).

Another method is to offer customers a choice of incentives, including the option to not choose any (do nothing). Research suggests that this increases customer persistence and retention**.** The

design of the retention effort should ideally appeal to both high-risk customers and the general customer base, such as through overall product or service enhancements (Ascarza, et al., 2017).

Incorporating elements of surprise can also be beneficial. This aligns with the concept of customer delight in services marketing, where unexpected positive experiences strengthen customer loyalty. This is particularly relevant as retention campaigns often target customers who might not have churned otherwise. If these customers are pleasantly surprised by offers like a new phone in a telecom context or exercise gear in a gym setting, it can bolster long-term loyalty even among those who weren't at risk of churning. This approach supports the idea of a service recovery paradox, where exceeding the expectations of dissatisfied customers can lead to increased engagement and reduced churn (Ascarza, et al., 2017).

There’s also a significant distinction to be made between one-to-one personalized retention efforts and offering a single incentive to all at-risk customers. Personalized retention strategies involve crafting unique offers or solutions based on the individual characteristics, preferences, and past interactions of each customer. This approach can be more effective in directly addressing the specific reasons a customer may consider leaving, thereby not only preventing churn but also potentially enhancing the customer's lifetime value (CLV) (Ascarza, et al., 2017). It demonstrates a high level of customer understanding and commitment, which can foster stronger loyalty and satisfaction. On the other hand, providing a uniform incentive to all at-risk customers can be more efficient and easier to manage. However, this approach might not resonate as deeply with each customer's unique needs and reasons for potential churn. While it may successfully retain a broad segment of customers, it might not significantly enhance the CLV in the same way that personalized efforts can.

Subsequently, the timing of the campaign is a critical factor to consider. There's a balance to be struck in the timing of retention campaigns: too late, and it becomes a reactive effort with the customer already on their way out, possibly leading to unsuccessful retention or high costs. Too early, and the campaign risks being irrelevant or even triggering churn considerations.

To manage this timing challenge, one effective method is to analyze data from past campaigns. This involves developing models to predict churn and the likelihood of successful retention over time from a specific start point, like the current period. Immediate targeting after acquiring a customer isn't always essential, but early consideration of retention is crucial. For instance, a telecom company should ensure new customers are on the most suitable data plan from the

outset. Pre-emptive timing targets customers before any signs of reduced loyalty appear. Proactive timing involves initiating campaigns for those identified as at-risk through predictive analytics. Reactive timing is about attempting to retain customers actively in the process of churning. Win-back strategies are employed once a customer has churned, with efforts to regain them. Post win-back actions are those taken after a customer has turned down a win-back proposal. Stauss and Friedge (1999) lay the groundwork for win-back strategies, advocating for in-depth conversations with churned customers to understand their reasons for leaving and to create tailored incentives (Ascarza, et al., 2017).

Finally, the results need to be evaluated. Evaluating campaigns is crucial, ideally through the use of a control group that is randomly chosen to not receive the campaign. This approach simplifies the assessment process, allowing for the straightforward compilation of top-line results without the need for complex causal modeling. Key metrics for evaluation should include overall profitability and various retention metrics. As suggested by Ascarza et al. (2017), calculating the "rescue-rate" of a campaign, which determines the percentage of potential churners who were retained, is essential. The company should consistently calculate this rate for all campaigns, enabling a meta-analysis across multiple campaigns to identify factors that influence the rescue rate. These factors can include the characteristics of the incentives, the profile of targeted customers, and how well the incentives align with customer needs.

Another critical aspect of evaluation is assessing the long-term impact of the campaign. While a customer may be retained in the short term, it's important to evaluate the subsequent effect on their future profitability. This involves analyzing changes in the customer's retention rate post-campaign, understanding whether an increase in retention is due to higher customer satisfaction, or if there's a decrease because the customer now expects or demands similar incentives in the future.

As we conclude our discussion on the design of an optimal churn intervention campaign, with a focus on assessing the long-term impact on customer profitability and retention rates, it is crucial to shift our focus to the following chapter, which delves into the inherent limitations and explores potential areas for further research.

6.7 Limitations and Further Research

Limitations

This chapter outlines the key limitations of the study, which include data constraints, lack of certain variables, potential biases, and issues related to generalizability and data quality. Understanding these limitations is crucial for interpreting the findings accurately and for guiding future research in this domain.

One of the primary limitations of this study is that the dataset consists of 1,600 churned customers. This limited sample size could restrict the robustness and the depth of the churn analysis. The smaller number of churned customers, compared to the total customer base, may not capture the full spectrum of reasons leading to customer attrition. Consequently, the predictive models developed might not fully encapsulate the diverse and complex factors influencing churn behavior.

The dataset lacks variables related to the types of credit card products used by customers. Credit card features such as reward programs, interest rates, and card fees vary widely and can significantly influence customer satisfaction and retention. The absence of these product-specific details hinders a more nuanced understanding of churn drivers. This limitation poses a challenge in developing targeted strategies to mitigate churn for specific credit card products.

Another critical limitation is the absence of contextual information such as economic factors or market changes. Economic downturns, regulatory changes, and competitive dynamics in the credit card industry can profoundly impact customer behavior and churn rates. Without considering these external factors, the study may overlook key elements that influence customer decisions to discontinue their credit card services.

The dataset's lack of time-series data is a notable limitation. Time-series data would allow for the analysis of customer behavior and churn patterns over specific periods. This data is essential for identifying trends, seasonal variations, and the evolution of customer behavior. The static nature of the dataset restricts the ability to observe dynamic changes in customer activity and churn, thereby limiting the predictive capability of the analysis.

The generalizability of the study findings is another area of concern. Given that the dataset may not be representative of the broader customer base in the credit card industry, the results of this study could be specific to the sample and not applicable to different customer segments or geographic regions. This limitation highlights the need for caution in extrapolating the findings to other contexts without additional validation.

Finally, the dataset includes columns with unknown values, which introduces an element of uncertainty in the analysis. These unknown values could represent a range of different customer attributes or behaviors, and their ambiguous nature makes it challenging to interpret their impact accurately. The presence of these values necessitates assumptions or imputations that could potentially bias the results.

Future research

A pivotal area of exploration is the use of time-series data and external factors. This involves leveraging time-series data to uncover long-term trends and patterns in customer behavior, which can be crucial for predicting future actions. The impact of external factors, such as economic changes, market trends, and societal shifts, on customer behavior is also a key area for investigation in future research. Understanding these dynamics can help businesses adapt to changing environments and maintain customer loyalty.

Future research could also investigate how the prediction of retention risk with the probability of continuous transactions for noncontractual businesses, known as P(Alive). This could aim to uncover the underlying factors that influence P(Alive). By identifying these factors, what keeps customers engaged over time could be better understood, enabling more targeted and effective retention strategies (Ascarza, et al., 2017).

The investigation of overlooked churn predictors stands as an area of inquiry. This involves exploring potential churn indicators that have not been adequately considered or understood in past research. By identifying these hidden predictors, businesses can gain a more comprehensive understanding of what drives customers away, allowing for more proactive and effective churn prevention strategies.

Understanding the sequences of customer experiences that lead to non-retention is also vital. Future research could focus on identifying specific customer journey touchpoints or sequences of events that increase the likelihood of churn. By pinpointing these critical moments or patterns, businesses can intervene more effectively to improve customer experiences and reduce churn rates.

Measuring churn and retention for low-frequency products could be another point to investigate in future research. Traditional retention metrics often focus on high-frequency transactions, leaving a gap in understanding for products purchased infrequently. Research in this area aims

to develop new metrics or methodologies that accurately reflect customer loyalty and retention in contexts where purchase frequency is low.

Exploring whether customers are actually retainable and under what circumstances they are retainable is another area of focus in future research. Future research could analyze to understand the factors that determine a customer's likelihood of remaining loyal to a brand or product. It involves analyzing customer profiles, past behaviors, and preferences to identify those who are most likely to be retained and under what conditions.

Incentives that increase both retention and customer lifetime value are crucial for business success. Future research can explore the types of incentives, both monetary and non-monetary, that effectively encourage customers to stay with a brand while also enhancing their lifetime value. The goal would be to identify incentives that not only prevent churn but also foster long-term customer relationships.

Separating predictors from causes in customer churn is essential for developing effective retention strategies. This would involve distinguishing between factors that indicate a higher risk of churn and those that directly cause customers to leave. By understanding these differences, retention efforts could be tailored more precisely as they would address the root causes of churn.

Different retention rates for various acquisition channels represent an important area of future study. This would examine how the method by which customers are acquired affects their likelihood of staying with a company. It involves analyzing retention rates across different channels to identify which are most effective for long-term customer retention and why.

Integrating retention risk, incentive response, and customer profitability to guide targeting decisions is a complex but critical area of research. This would involve developing models that consider a customer's likelihood of churning, their responsiveness to different types of incentives, and their overall profitability. Such models can help businesses make more informed decisions about whom to target with retention efforts and whom to let go.

The advisability of monetary versus non-monetary incentives is a key consideration in retention strategies. Future research in this area could focus on when and under what circumstances each type of incentive is most effective. It could also explore which retention efforts not only prevent customers from leaving but also enhance their long-term value to the company.

Integrating reasons why customers churn into the best action for retaining customers is crucial for effective retention strategies. Future research here would involve understanding the specific reasons behind customer churn and developing targeted interventions that address these issues directly. By tailoring retention efforts to the underlying causes of churn, businesses can more effectively prevent customers from leaving.

Optimally allocating acquisition and retention campaign budgets over time is a significant challenge for many businesses. Future research could explore how to best distribute resources between acquiring new customers and retaining existing ones. It involves developing strategies and models that balance short-term gains with long-term customer value, ensuring sustainable business growth.

In addition to these specific areas, the utilization of data encompassing behavioral and customer interaction information, along with marketing efforts, could be a potential area of future research. By leveraging this data, it is possible to gain deeper insights into customer behaviors and preferences, enabling more targeted and effective retention strategies. This comprehensive approach could significantly advance the field of customer relationship management and contribute to the development of more sophisticated and efficient business practices.

7. CONCLUSION

This dissertation aimed to contribute to the area of research on customer churn, specifically the features of customer churn in the credit card industry. With the increasing competition and evolving customer preferences in the modern marketplace, the significance of comprehending and mitigating customer churn cannot be overstated. Therefore, our empirical analysis of customer churn provides insights for future strategic planning and decision-making in certain aspects of customer retention.

In this thesis, we have focused on the credit card industry, analyzing customer behavior and churn patterns. Our approach involved developing robust models to accurately assess the factors influencing customer churn. Based on our extensive literature review and data analysis, we formulated and analyzed hypotheses addressing key aspects such as customer loyalty, engagement patterns, and the impact of various service attributes on churn rates.

Overall, our thesis presents a detailed examination of customer churn, offering some insights and practical recommendations for service providers seeking to enhance customer retention and satisfaction. This research not only contributes to the academic discourse on customer churn but also provides actionable strategies for businesses aiming to improve their customer relationship management and reduce churn rates. From the general results in chapter 5.1 on page 42, we found the following features to be the most impactful in customer churn: total transaction count, total revolving balance, and number of other products/services with the same bank. In the analysis part from chapter 5.2 to 5.6, we have successfully established that both low credit card usage rates and a reduction in credit card usage are strong predictors of customer churn. Additionally, we found that customers engaged with a greater variety of banking services or products from the same bank exhibit a lower tendency to churn. Furthermore, our thesis indicates a higher likelihood of churn among customers who display increased months of inactivity or who make a higher number of contacts with the bank. Our analysis in chapter 5.4 on page 62 also revealed differences in churn rates by gender. In this business, female customers had a 2.74 % increased churn rate over male customers overall, and a 17.63 % increase churn rate over males in the reduced count segment.

These findings provide a deeper understanding of factors influencing customer loyalty and can guide strategic decisions to enhance customer retention. Still, due to the dynamic and multifaceted aspects of customer churn, the exact causes of what leads to churn remains unknown, emphasizing the need for future studies of factors influencing customer churn.

REFERENCES

Ahmad, R., & Buttle, F. (2002). Customer retention management: a reflection of theory and practice. *Marketing Intelligence & Planning - June*, 149-161.

Ang, L., & Buttle, F. (2004). Customer retention management processes. *European Journal of Marketing*, 83-99.

Ascarza, E., Iyengar, R., & Schleicher, M. (2016). *The Perils of Proactive Churn Prevention Using Plan Recommendations: Evidence from a Field Experiment*.

Ascarza, E., Neslin, S. A., Netzer, O., Anderson, Z., Fader, P. S., Gupta, S., . . . Schrift, R. (2017). *In Pursuit of Enhanced Customer Retention Management: Review, Key Issues and Future Directions.* Alberta: University of Alberta.

Berson, A., Smith, S., & Thearling, K. (2000). Building Data Mining Applications for CRM. McGraw-Hill.

Breiman, L. (2001). *Random forests. Machine learning, 45(1), 5-32.*

Burton, B. K. (2008). Commonsense Morality. In R. W. Kolb, *Encyclopedia of Business Ethics and Society* (pp. 365-366). Chicago: SAGE Publications.

Carrillo-Perez, F., & al, e. (2021). Applications of artificial intelligence in dentistry. *Journal of Esthetic and Restorative Dentistry*.

Chen, T., & Guestrin, C. (2016, August 13-17). XGBoost: A Scalable Tree Boosting System. San Francisco, CA, USA.

Craney, T. A., & Surles, J. G. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering*, pp. 391-403.

Damghani, B. M., Welch, D., O'Malley, C., & Knights, S. (2013). The Misleading Value of Measured Correlation. *Wilmott*.

DeSouza, G. (1992). Designing a Customer Retention Plan. *The Journal of Business Strategy*, 24-28.

Engwall, L. (1976). Reponse Time of Organizations. *The Journal of Management Studies*, 1-15.

Fawcett, T. (2006). *An introduction to ROC analysis, 27(8), 861-874.* Pattern Recognition Letters.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics. *Annals of Statistics*.

Gamble, J. W. (1998). Client retention is key to profitable service marketing. *Marketing News*, 8-9.

Glady, N., Baesens, B., & Croux, C. (2006). Modeling churn using customer lifetime value. *European Journal of Operational Research*, 402-411.

Gupta, K., & Stewart, D. W. (1996). Customer Satisfaction and Customer Behavior: The Differential Role of Brand and Category Expectations. *Marketing Letters 7:3*, 249-263.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). The Elements of Statistical Learning. In *The Elements of Statistical Learning* (pp. 367, 593). Stanford, California: Springer.

Hennig-Thurau, T., & Klee, A. (1997). The Impact of Customer Satisfaction and Relationship Quality on Customer Retention: A Critical Reassessment and Model Development. *Psychology & Marketing*, 737-764.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression (3rd ed.).* John Wiley & Sons.

Huiting, Z., Jiabin, Y., & Long, C. (2017, August 08). *Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation.* Retrieved from MDPI: https://www.mdpi.com/1996-1073/10/8/1168

Jamal, Z., & Bucklin, R. E. (2006). Improving the Diagnosis and Prediction of Customer Churn: A Heterogeneous Hazard Modeling Approach. *Journal of Interactive Marketing*, 16-29.

Kowalek, P., Loch-Olszewska, H., & Szwabinski, J. (2019). Classification of diffusion modes in single-particle tracking data. *Physical Review*.

Kumar, V., Leszkiewicz, A., & Herbst, A. (2017). Are You Back for Good or Still Shopping Around? Investigating Customers' Repeat. *Forthcoming in the Journal of Marketing Research*.

Lambrecht, A., & Skiera, B. (2006). Paying Too Much and Being Happy About It: Existence, Causes, and Consequences of Tariff-Choice Biases. *Journal of Marketing Research Vol. XLIII*, 212-223.

Li, S., Sun, B., & Montgomery, A. L. (2011). Cross-Selling the Right Product to the Right Time. *Journal of Marketing Research*.

Meyers, J. (1996, July 22). GTE TSI gets ready to burn churn. *Wireless Networks, Telephony*, p. 9.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, 204-211.

Renjith, S. (2015). An Integrated Framework to Recommend Personalized Retention. *International Journal of Engineering Trends and Technology*.

Roberts-Lombard, M. (2011). Customer retention through customer relationship management. *African Journal of Business Management*.

Rodríguez, P. M. (2018). Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, pp. 75:21–31.

Schwarze, F. (2007). Relationship Banking and profitability - An empirical survey of German banks. *Goethe University Frankfurt*.

scikit learn. (2023). Retrieved from 1.11. Ensembles: Gradient boosting, random forests, bagging, voting, stacking: https://scikit-learn.org/stable/modules/ensemble.html

Sokolova, M., & Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks, 45(4), 427-437.* Information Processing & Management.

The University of Utah. (2022, November). Retrieved from The Chi-Square Test for Independence: https://soc.utah.edu/sociology3112/chi-square.php

Van den Poel, D., & Larivière, B. (2001). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research 157*, 196-217.

Wayne W. LaMorte, M. P. (2017, May). *Boston University School of Public Health*. Retrieved from https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/bs704_nonparametric4.html

Wu, Q., & Vos, P. (2018). Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications. Chapter 9.2. In *Handbook of Statistics.*

Zemke, R. (1990, November 26). Wooing customers is only half the battle. *AMA International News*, p. 22.

Zheng, H., Yuan, J., & Chen, L. (2017). hort-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. *Energies*.

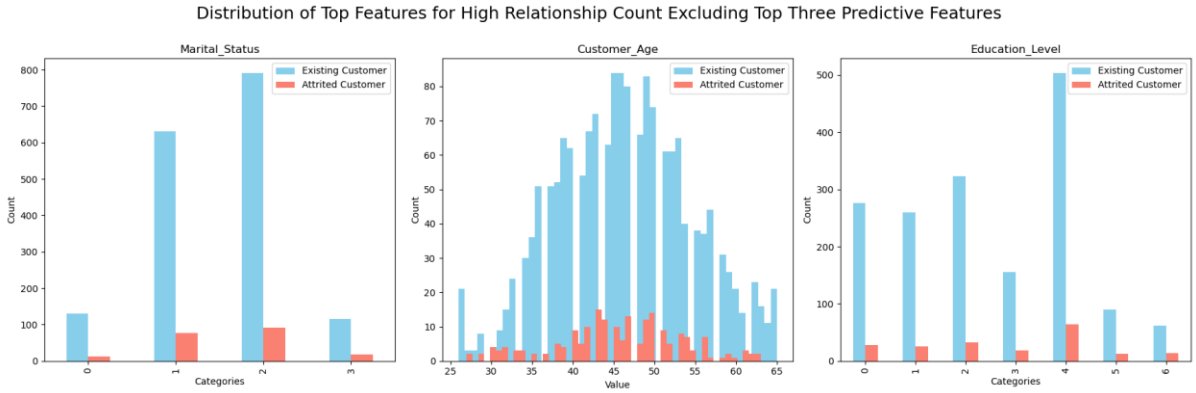
APPENDIX - SUPPLEMENTARY FIGURES

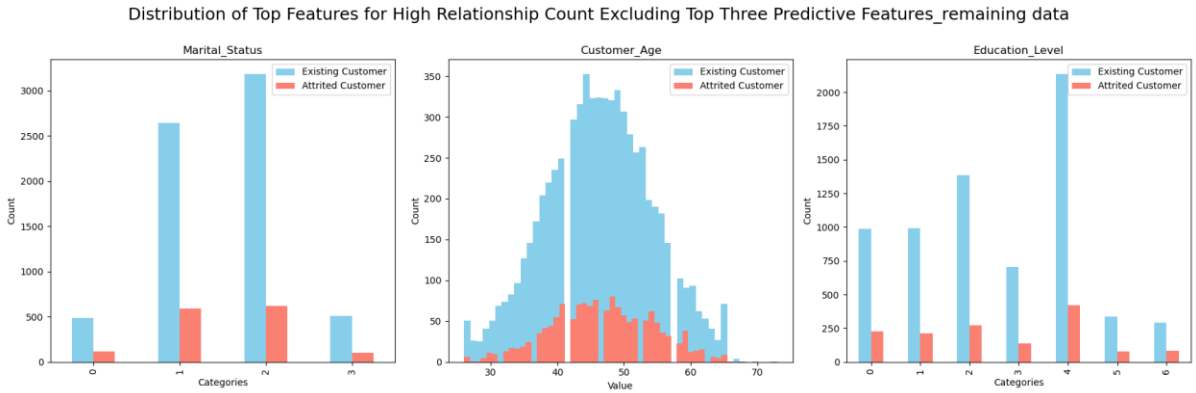
SEGMENT ANALYSIS:

A function was created to compare how the model trained on the entire dataset would handle the specific segments analyzed. Additionally, the function considers the fact that the segments might be a large part of the training data by comparing all the k-folds. Even so, the model might be too well trained on the segment. In fact, reviewing practically all segments we suggested, this was the case. Therefore, two more approaches are considered. First approach, the segment is removed from the dataset. Then, four individual splits are made, one with 0 % in the training set and 100 % in the test set, one with a 20/80 split, one with 50/50 and finally one with 80/20. This approach aims to lightly evaluate the model's ability to predict segments with which it has no, limited, medium, or high knowledge of. Second approach, another model is trained exclusively on the segment data itself with an 80/20 split through 5 k-folds. Feature importance from the highest accuracy combined model is compared to the segment-specific model. With this setup, we can easily remove top features to explore the performance of the model with less features to use for prediction, as well as exploring churn distribution within certain segments for given features.

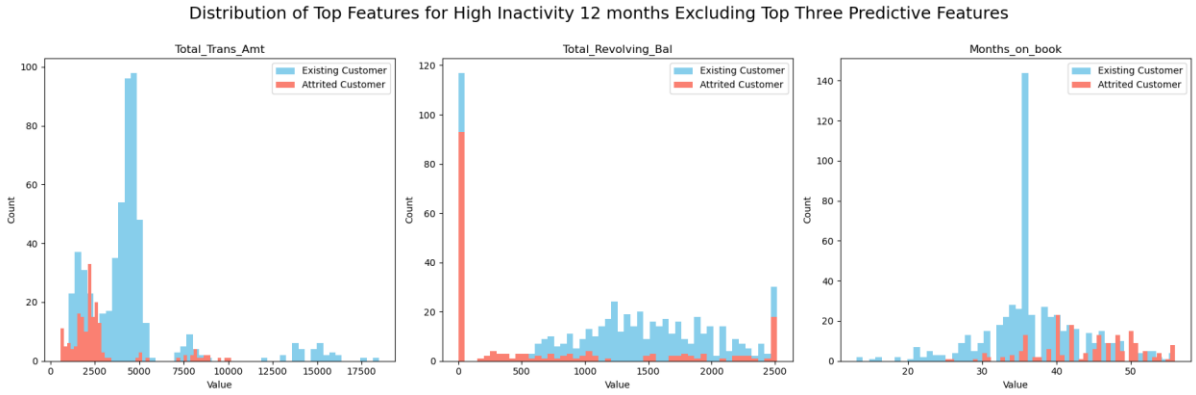
SEGMENT ANALYSIS RESULTS – HYPOTHESIS 1:

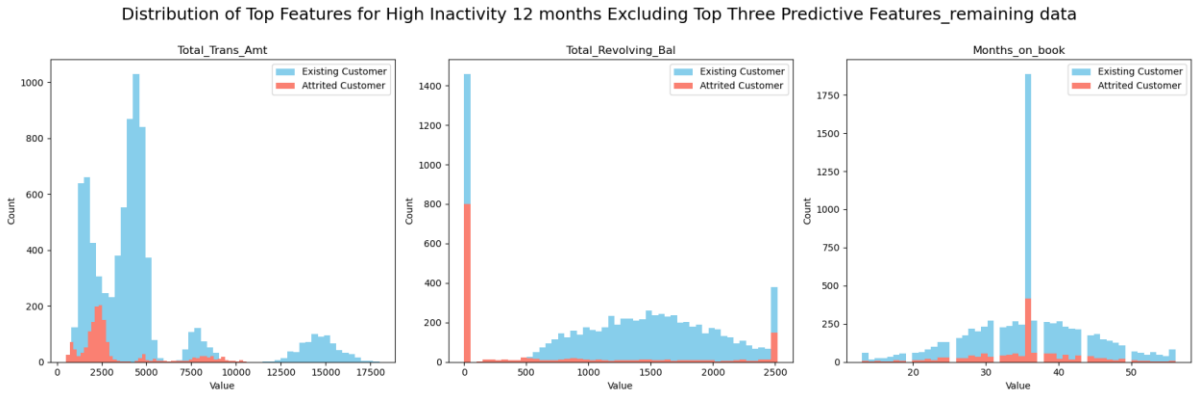
SEGMENT ANALYSIS RESULTS – HYPOTHESIS 3:



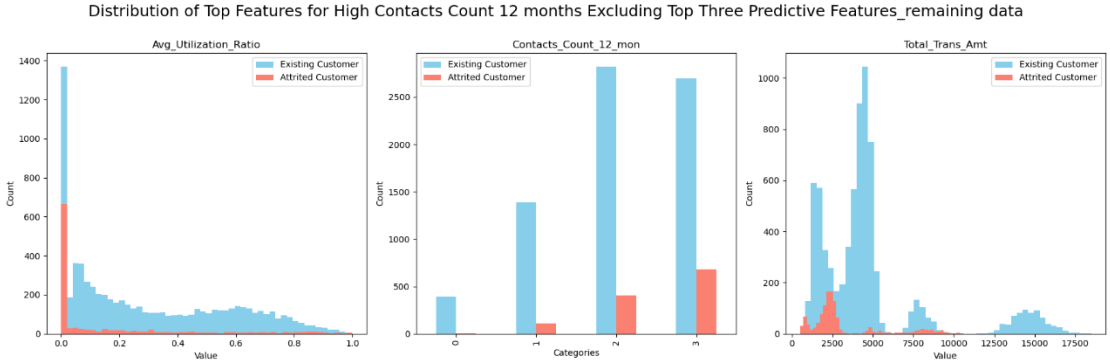


SEGMENT ANALYSIS RESULTS – HYPOTHESIS 4 – HIGH INACTIVITY:





SEGMENT ANALYSIS RESULTS – HYPOTHESIS 4 – HIGH CONTACTS COUNT:



CORRELATION MATRIX

Correlation Matrix	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio	Education_Level_1	Education_Level_2	Education_Level_3	Education_Level_4	Education_Level_5	Education_Level_6	Marital_Status_1	Marital_Status_2	Marital_Status_3	Income_Category_1	Income_Category_2	Income_Category_3	Income_Category_4	Income_Category_5	Card_Category_1	Card_Category_2	Card_Category_3
CLIENTNUM	1	-0,05	0,01	-0,02	0,01	0,13	0,01	0,01	0,01	0,01	0	0,01	0,02	-0,02	0	0,01	0	-0,01	0	0,02	0	0	-0,01	-0,01	0	0,01	0,02	0	0,02	0,01	-0,02	0,01	0	0
Attrition_Flag	-0,05	1	0,02	0,04	0,02	0,01	-0,15	0,15	0,2	-0,02	-0,26	0	-0,13	-0,17	-0,37	-0,29	-0,18	0	-0,01	-0,01	-0,01	0,01	0,03	0,02	-0,02	0	-0,01	-0,03	0	0,01	0,01	-0,01	0,01	0,01
Customer_Age	0,01	0,02	1	0,02	-0,12	0,79	-0,01	0,05	-0,02	0	0,01	0	-0,06	-0,05	-0,07	-0,01	0,01	0,01	0	-0,01	0	-0,02	0,03	-0,01	0,05	-0,04	-0,01	-0,02	0,01	0,04	0	-0,02	-0,01	0,01
Gender	-0,02	0,04	0,02	1	0	0,01	0	0,01	-0,04	-0,42	-0,03	-0,42	-0,03	-0,02	0,07	0,01	0,26	0,01	-0,02	0	0,01	-0,01	0,02	0,02	-0,01	0	0,03	-0,42	-0,45	-0,29	0,3	-0,07	-0,04	-0,01
Dependent_count	0,01	0,02	-0,12	0	1	-0,1	-0,04	-0,01	-0,04	0,07	0	0,07	-0,04	0,03	0,05	0,01	-0,04	0	-0,01	0	0	0,01	0	-0,04	0,01	0,01	-0,01	0,03	0,05	0,02	-0,02	0,02	0,03	0
Months_on_book	0,13	0,01	0,79	0,01	-0,1	1	-0,01	0,07	-0,01	0,01	0,01	0,01	-0,05	-0,04	-0,05	-0,01	-0,01	0	0	-0,01	0	-0,02	0,02	-0,01	0,03	-0,03	0	-0,02	0,01	0,03	-0,01	-0,01	-0,01	0
Total_Relationship_Count	0,01	-0,15	-0,01	0	-0,04	-0,01	1	0	0,06	-0,07	0,01	-0,07	0,05	-0,35	-0,24	0,04	0,07	0,01	0	-0,01	0,01	0,01	-0,01	-0,02	0,02	0,01	-0,01	0,01	0	-0,01	0	-0,06	-0,06	-0,04
Months_Inactive_12_mon	0,01	0,15	0,05	0,01	-0,01	0,07	0	1	0,03	-0,02	-0,04	-0,02	-0,03	-0,04	-0,04	-0,04	-0,01	0,01	-0,01	0	0,01	-0,01	0	0,01	-0,01	0	-0,02	0	-0,01	0	0,01	-0,02	0	0
Contacts_Count_12_mon	0,01	0,2	-0,02	-0,04	-0,04	-0,01	0,06	0,03	1	0,02	-0,05	0,03	-0,02	-0,11	-0,15	-0,09	-0,06	0,01	0	-0,01	0	-0,01	0	0,01	0	-0,01	0	0	0,01	0,02	0	0	0	0
Credit_Limit	0,01	-0,02	0	-0,42	0,07	0,01	-0,07	-0,02	0,02	1	0,04	1	0,01	0,17	0,08	0	-0,48	0,01	0	0	0	0,01	-0,01	0,03	-0,06	0,02	-0,16	0,09	0,33	0,34	0,03	0,44	0,23	0,11
Total_Revolving_Bal	0	-0,26	0,01	-0,03	0	0,01	0,01	-0,04	-0,05	0,04	1	-0,05	0,06	0,06	0,06	0,09	0,62	0	0,02	-0,01	0	0,01	-0,02	-0,04	0,04	0	0,01	0	0,02	0,02	-0,02	0,01	0,02	0,01
Avg_Open_To_Buy	0,01	0	0	-0,42	0,07	0,01	-0,07	-0,02	0,03	1	-0,05	1	0,01	0,17	0,07	-0,01	-0,54	0,01	0	0	0	0,01	0	0,04	-0,06	0,02	-0,16	0,09	0,33	0,34	0,04	0,44	0,23	0,11
Total_Amt_Chng_Q4_Q1	0,02	-0,13	-0,06	-0,03	-0,04	-0,05	0,05	-0,03	-0,02	0,01	0,06	0,01	1	0,04	0,01	0,38	0,04	0	0,01	0,01	-0,01	0,01	-0,02	-0,04	0,05	-0,01	0	0,02	0,01	-0,01	0,01	0	0,01	0
Total_Trans_Amt	-0,02	-0,17	-0,05	-0,02	0,03	-0,04	-0,35	-0,04	-0,11	0,17	0,06	0,17	0,04	1	0,81	0,09	-0,08	0,01	0	-0,01	0	0,01	-0,01	0,04	-0,06	0,01	0	0,01	0,01	0,01	-0,01	0,15	0,1	0,06
Total_Trans_Ct	0	-0,37	-0,07	0,07	0,05	-0,05	-0,24	-0,04	-0,15	0,08	0,06	0,07	0,01	0,81	1	0,11	0	0,01	0	0	0	0	-0,01	0,1	-0,12	0,02	0	-0,03	-0,04	-0,01	0,02	0,1	0,08	0,04
Total_Ct_Chng_Q4_Q1	0,01	-0,29	-0,01	0,01	0,01	-0,01	0,04	-0,04	-0,09	0	0,09	-0,01	0,38	0,09	0,11	1	0,07	0,01	0	0	0	-0,01	-0,02	-0,01	0,01	-0,01	0	0,01	0	-0,03	0,01	0	0	-0,01
Avg_Utilization_Ratio	0	-0,18	0,01	0,26	-0,04	-0,01	0,07	-0,01	-0,06	-0,48	0,62	-0,54	0,04	-0,08	0	0,07	1	0	0,01	-0,01	0,01	0	-0,01	-0,03	0,05	-0,02	0,07	-0,1	-0,17	-0,15	-0,07	-0,19	-0,09	-0,04
Education_Level_1	-0,01	0	0,01	0,01	0	0	0,01	0,01	0,01	0,01	0	0,01	0	0,01	0,01	0,01	0	1	-0,21	-0,14	-0,28	-0,1	-0,09	0	-0,02	0,03	-0,01	-0,01	-0,01	0,01	0,02	0	-0,02	0
Education_Level_2	0	-0,01	0	-0,02	-0,01	0	0	-0,01	0	0	0,02	0	0,01	0	0	0	0,01	-0,21	1	-0,17	-0,33	-0,12	-0,11	0	0,01	-0,02	0	0,02	0	0	0	-0,01	0,01	-0,02
Education_Level_3	0,02	-0,01	-0,01	0	0	-0,01	-0,01	0	-0,01	0	-0,01	0	0,01	-0,01	0	0	-0,01	-0,14	-0,17	1	-0,22	-0,08	-0,07	-0,01	0	0,01	0	-0,01	0,02	0	0	0	0,01	-0,01
Education_Level_4	0	-0,01	0	0,01	0	0	0,01	0,01	0	0	0	0	-0,01	0	0	0	0,01	-0,28	-0,33	-0,22	1	-0,15	-0,14	-0,01	0,01	0	0	-0,01	0	-0,02	-0,01	0,01	0	0,01
Education_Level_5	0	0,01	-0,02	-0,01	0,01	-0,02	0,01	-0,01	-0,01	0,01	0,01	0,01	0,01	0,01	0	-0,01	0	-0,1	-0,12	-0,08	-0,15	1	-0,05	-0,01	0	0	0,02	0,01	0	-0,01	-0,01	0,01	0	0,02
Education_Level_6	-0,01	0,03	0,03	0,02	0	0,02	-0,01	0	0	-0,01	-0,02	0	-0,02	-0,01	-0,01	-0,02	-0,01	-0,09	-0,11	-0,07	-0,14	-0,05	1	0,01	0	0	-0,01	0	-0,02	0,01	0,03	-0,01	0	0,01
Marital_Status_1	-0,01	0,02	-0,01	0,02	-0,04	-0,01	-0,02	0,01	0,01	0,03	-0,04	0,04	0,04	0,1	-0,01	-0,01	-0,03	0	0	-0,01	-0,01	-0,01	0,01	1	-0,74	-0,23	0	-0,01	-0,02	-0,01	0,01	0,03	0,02	0,01
Marital_Status_2	0	-0,02	0,05	-0,01	0,01	0,03	0,02	-0,01	0	-0,06	0,04	-0,06	0,05	-0,06	-0,12	0,01	0,05	-0,02	0,01	0	0,01	0	0	-0,74	1	-0,26	-0,01	0,01	0,01	0,01	-0,01	-0,04	-0,02	-0,01
Marital_Status_3	0,01	0	-0,04	0	0,01	-0,03	0,01	0	-0,01	0,02	0	0,02	-0,01	0,01	0,02	-0,01	-0,02	0,03	-0,02	0,01	0	0	0	-0,23	-0,26	1	0,01	0	-0,01	0	0,01	0,01	-0,01	0
Income_Category_1	0,02	-0,01	-0,01	0,03	-0,01	0	-0,01	-0,02	0	-0,16	0,01	-0,16	0	0	0	0	0,07	-0,01	0	0	0	0,02	-0,01	0	-0,01	0,01	1	-0,19	-0,2	-0,13	-0,16	0	-0,01	-0,01
Income_Category_2	0	-0,03	-0,02	-0,42	0,03	-0,02	0,01	0	0	0,09	0	0,09	0,02	0,01	-0,03	0,01	-0,1	-0,01	0,02	-0,01	-0,01	0,01	0	-0,01	0,01	0	-0,19	1	-0,17	-0,11	-0,14	0,02	0,03	0,01
Income_Category_3	0,02	0	0,01	-0,45	0,05	0,01	0	-0,01	0,01	0,33	0,02	0,33	0,01	0,01	-0,04	0	-0,17	-0,01	0	0,02	0	0	-0,02	-0,02	0,01	-0,01	-0,2	-0,17	1	-0,12	-0,15	0,04	0,01	-0,01
Income_Category_4	0,01	0,01	0,04	-0,29	0,02	0,03	-0,01	0	0,02	0,34	0,02	0,34	-0,01	0,01	-0,01	-0,03	-0,15	0,01	0	0	-0,02	-0,01	0,01	-0,01	0,01	0	-0,13	-0,11	-0,12	1	-0,1	0,03	0,03	0,02
Income_Category_5	-0,02	0,01	0	0,3	-0,02	-0,01	0	0,01	0	0,03	-0,02	0,04	0,01	-0,01	0,02	0,01	-0,07	0,02	0	0	0	-0,01	0,03	0,01	-0,01	0,01	-0,16	-0,14	-0,15	-0,1	1	-0,01	-0,01	0,02
Card_Category_1	0,01	-0,01	-0,02	-0,07	0,02	-0,01	-0,06	-0,02	0	0,44	0,01	0,44	0	0,15	0,1	0	-0,19	0	-0,01	0	0,01	0,01	-0,01	0,03	-0,04	0,01	0	0,02	0,04	0,03	-0,01	1	-0,03	-0,01
Card_Category_2	0	0,01	-0,01	-0,04	0,03	-0,01	-0,06	0	0	0,23	0,02	0,23	0,01	0,1	0,08	0	-0,09	-0,02	0,01	0,01	0	0	0	0,02	-0,02	-0,01	-0,01	0,03	0,01	0,03	-0,01	-0,03	1	0
Card_Category_3	0	0,01	0,01	-0,01	0	0	-0,04	0	0	0,11	0,01	0,11	0	0,06	0,04	-0,01	-0,04	0	-0,02	-0,01	0,01	0,02	0,01	0,01	-0,01	0	-0,01	0,01	-0,01	0,02	0,02	-0,01	0	1