Customer Churn Prediction

Dalton Fumo

Dissertation written under the supervision of professor Ana Guedes.

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, 31/05/2024.

Abstract

Customer churn prediction is a critical task for businesses operating in competitive markets, especially in the context of online retail. Identifying customers at risk of leaving a service or product allows businesses to implement proactive retention strategies and maintain long-term profitability. This thesis aims to investigate the factors influencing customer churn in online retail and develop predictive models to anticipate churn behavior. Leveraging machine learning techniques, interpretability, and explainability, this study explores the impact of various customer attributes such as demographic information, purchasing behavior, and satisfaction scores on churn prediction. The analysis uses a comprehensive dataset containing customer attributes, transaction history, and response to marketing campaigns. By employing logistic regression models, gradient boosting models and advanced interpretability techniques such as SHAP (SHapley Additive exPlanations), this research aims to provide actionable insights for businesses to mitigate churn and enhance customer retention strategies in the online retail landscape. The findings highlight the significance of features such as average transaction amount, annual income, and recency of last purchase in predicting customer churn, and demonstrate the superior performance of gradient boosting models over logistic regression models in this context.

Keywords: Churn prediction, Logistic Regression, Gradient Boosting, Interpretability & Explainability, SHAP.

Title: Customer Churn Prediction

Author: Dalton Fumo

Resumo

A previsão de churn de clientes é uma tarefa crítica para empresas que operam em mercados competitivos, especialmente no varejo online. Identificar clientes com risco de abandonar um serviço ou produto permite que as empresas implementem estratégias de retenção proativas e mantenham a lucratividade a longo prazo. Esta tese tem como objetivo investigar os fatores que influenciam o churn de clientes no varejo online e desenvolver modelos preditivos para antecipar o comportamento de churn. Utilizando técnicas de aprendizado de máquina, interpretabilidade e explicabilidade, este estudo explora o impacto de vários atributos de clientes, como informações demográficas, comportamento de compra e pontuações de satisfação, na previsão de churn. A análise utiliza um conjunto de dados abrangente contendo atributos de clientes, histórico de transações e resposta a campanhas de marketing. Ao empregar modelos de regressão logística, gradient boosting e técnicas avançadas de interpretabilidade, como SHAP (SHapley Additive exPlanations), esta pesquisa visa fornecer percepções acionáveis para as empresas mitigarem o churn e aprimorarem as estratégias de retenção de clientes no cenário do varejo online. Os resultados destacam a importância de características como valor médio das transações, renda anual e recência da última compra na previsão de churn de clientes e demonstram o desempenho superior dos modelos de gradient boosting em relação aos modelos de regressão logística neste contexto.

Palavras-chave: Previsão de churn, Regressão Logística, Gradient Boosting, Interpretabilidade e Explicabilidade, SHAP.

Título: Customer Churn Prediction

Autor: Dalton Fumo

Acknowledgements

I am deeply grateful to Professor Ana Guedes, my dedicated supervisor, whose expert guidance, encouragement, and unwavering support have been instrumental in shaping this research. Prof. Ana's insightful feedback and constructive suggestions have been invaluable in every step of the way.

I would like to express my heartfelt appreciation to my parents for their boundless love, encouragement, and unwavering belief in my abilities. Their constant support and sacrifices have been the driving force behind my academic pursuits.

To my dear friends and family, thank you for your endless patience, understanding, and encouragement. Your unwavering support and encouragement have kept me motivated and inspired throughout this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In today's highly competitive online retail market, retaining customers is essential for businesses to sustain growth and profitability. Customer churn, the phenomenon where customers discontinue their relationship with a product or service, poses a significant challenge for businesses seeking to maintain a loyal customer base. Understanding the factors influencing customer churn and developing effective predictive models are crucial steps in implementing proactive retention strategies and maximizing customer lifetime value.

The research presented here explores methodologies to address whether past behavior is predictive of customer churn behavior and which factors are more relevant to predict churn and anticipate this behavior. More specifically, this thesis focuses on predicting customer churn in online retail by leveraging machine learning techniques and interpretability methods. The study aims to address the following key objectives:

1. **Identify Relevant Factors:** Investigate the demographic characteristics, purchasing behavior, and satisfaction scores of customers to identify factors influencing churn behavior.
2. **Develop Predictive Models:** Build classification models to predict customer churn based on the identified factors and evaluate model performance using appropriate metrics.
3. **Interpretability and Explainability:** Use interpretability techniques to interpret model predictions and understand the underlying factors driving churn behavior.
4. **Provide Actionable Insights:** Generate actionable insights for businesses to develop targeted retention strategies and mitigate customer churn in the online retail landscape.

By addressing these objectives, this research work aims to consolidate the existing body of knowledge on customer churn prediction and provide practical guideline to address projects related to product recommendations, focusing on enhancing businesses ability to retain customers and maintain a competitive edge in the online retail market.

Chapter 2: Literature Review

2.1 E-Commerce

Between 2018 and 2019, the global economy witnessed a significant surge in the influence of developing economies, with nearly half of the world's top 20 economies falling within this category. This phenomenon was accompanied by a remarkable 11% increase in global e-commerce sales and a parallel 4% rise in global GDP (UNCTAD, 2019). Such statistics underscore the profound impact of e-commerce on both economic growth and market dynamics worldwide.

E-commerce has emerged as a pivotal force in the contemporary business landscape, revolutionizing traditional commerce with its digital framework (Smith, 2020). Its significance transcends mere transactional exchanges, playing a transformative role in shaping global economic structures (Johnson et al., 2018). According to recent data from Statista, global e-commerce sales reached $4.9 trillion in 2021, and it is projected to continue its upward trajectory, expected to surpass $6.3 trillion by 2024 (Statista, 2022). Furthermore, e-commerce sales are forecasted to account for approximately 23.6% of global retail sales by 2024 (Statista, 2022). Another study conducted by eMarketer revealed that in 2021, e-commerce sales accounted for 19.5% of total retail sales worldwide, marking a significant increase from previous years (eMarketer, 2021). This percentage is expected to continue growing steadily, reaching 22.3% by 2024 (eMarketer, 2021).

The business market landscape has transformed from brick-and-mortar to digital-based competition due to the Fourth Industrial Revolution (4IR) demand and the presence of new and advanced technologies (Koe & Sakir, 2020); a phenomenon referred to as digitalization. Businesses and economies are now dependent on technology to survive and thrive due to this digitalization (Mthembu et al., 2018). Digitalization is the process of integrating and using digital technologies (also known as information and communication technologies – ICTs (Pollitzer, 2018) to enhance a business model and create new opportunities for producing goods and services and adding value.

E-commerce is defined by Koe and Sakir (2020) as conducting business transactions in a digital form or using the Internet. E-commerce provides businesses with the opportunity to grow and flourish (Koe & Sakir, 2020) and has been found to positively contribute to economic growth, independent of the level of development of a country (Kabir et al., 2020; Myovella et al., 2020).

Notably, e-commerce facilitates national development by fostering job creation and bolstering government revenues through import tariffs (Li & Zhang, 2019). However, despite its profound impact, disparities persist in e-commerce adoption, particularly in developing nations. Challenges such as limited internet accessibility, trust deficits, and apprehensions regarding online payment security hinder widespread e-commerce integration (Islam & Hossain, 2021). Consequently, these obstacles contribute to escalated customer churn rates within the e-commerce sector, posing significant challenges for businesses operating in this domain.

Research indicates that the challenges associated with e-commerce adoption in developing countries significantly contribute to heightened churn rates (Datta & Mulligan, 2018). Limited internet infrastructure, coupled with concerns regarding trust and security in online transactions, create barriers to entry for both businesses and consumers (Hossain & Dwivedi, 2020). As a result, businesses operating in these regions often grapple with higher churn rates, impeding their ability to establish long-term customer relationships and achieve sustainable growth.

In addition to the evolving landscape of e-commerce, which necessitates a nuanced understanding of customer behavior and preferences (Li & Karahanna, 2020), there are many factors influencing e-commerce in developing countries. Focusing specifically on technological factors, we can analyse how technology can be utilized to improve and boost sales/revenues in such developing nations, particularly by targeting customers who are more likely to churn and offering alternative marketing strategies (Huang et al., 2019; Mukherjee & Nath, 2021).

This can be achieved by exploring *Machine Learning* solutions and *Predictive Analytics*, which provides valuable insights into customer churn patterns. This analytical approach enables us to comprehend and explore various variables to develop feasible solutions that can be both interpreted and explained. By leveraging predictive analytics, businesses can make predictions about their online retail customer base, identifying which customers are more likely to churn and discerning the driving forces behind their behavior. Through the application of advanced analytics techniques such as machine learning and predictive modeling, businesses can implement targeted retention strategies to enhance customer satisfaction and loyalty.

2.2 Machine Learning and Predictive Analytics in E-commerce

Predictive analytics encompasses a combination of techniques and technologies like AI, ML, and statistical analysis. From forecasting stock market fluctuations to preventing equipment failure, predictive analytics has made it possible for companies to make informed decisions in a wide range of business areas (Roman Davydov, 2022).

Machine Learning (ML) is a subset of artificial intelligence (AI) that focuses on enabling machines to learn from data without explicit programming. It involves the development of algorithms that can recognize patterns and make predictions based on data inputs. In recent years, ML has gained prominence in various industries, including e-commerce, due to its ability to analyse large datasets and extract valuable insights (Jordan & Mitchell, 2015).

Predictive analytics in e-commerce encompasses various approaches and methodologies aimed at forecasting market demands, predicting customer behavior, enabling dynamic pricing, and detecting fraud. These approaches often leverage ML algorithms to analyse relationships between different customer data points (Duan et al., 2019).

For example, research by Li et al. provides insights into the application of predictive analytics in e-commerce by using customer purchase history data to predict future buying behavior. Their study demonstrates how ML techniques such as decision trees and neural networks can effectively predict customer preferences and tailor marketing strategies accordingly. (Li et al., 2020)

Another approach discussed by Wang et al. (2020) focuses on dynamic pricing strategies in e-commerce using predictive analytics. By analysing historical sales data and market trends, their research proposes a model that adjusts prices in real-time to optimize revenue and enhance competitiveness in the online marketplace.

Additionally, research by Sharma and Aggarwal (2023) explores the use of predictive analytics for fraud detection in e-commerce transactions. Their study highlights the importance of ML algorithms in identifying fraudulent activities based on anomalous patterns in transaction data, thereby safeguarding the integrity of online transactions and enhancing trust among customers. These examples illustrate the concrete application of predictive analytics techniques in addressing specific challenges within the e-commerce domain, ranging from customer behavior prediction to pricing optimization and fraud detection.

2.2.1 Interpretability and Explainability in E-commerce

Interpretation techniques are essential for understanding the behavior of machine learning models and extracting actionable insights from their predictions. Interpretability and explainability are crucial aspects of machine learning models, particularly in contexts where decision-making processes have significant real-world consequences. Lipton discusses the challenges surrounding model interpretability and emphasizes the importance of transparent and understandable models. In this work, it is argued that interpretable models not only provide insights into their decision-making process but also foster trust and acceptance among stakeholders (Lipton, 2016). Figure 1 below, provides definitions of interpretability and explainability, highlighting their importance in the context of e-commerce.

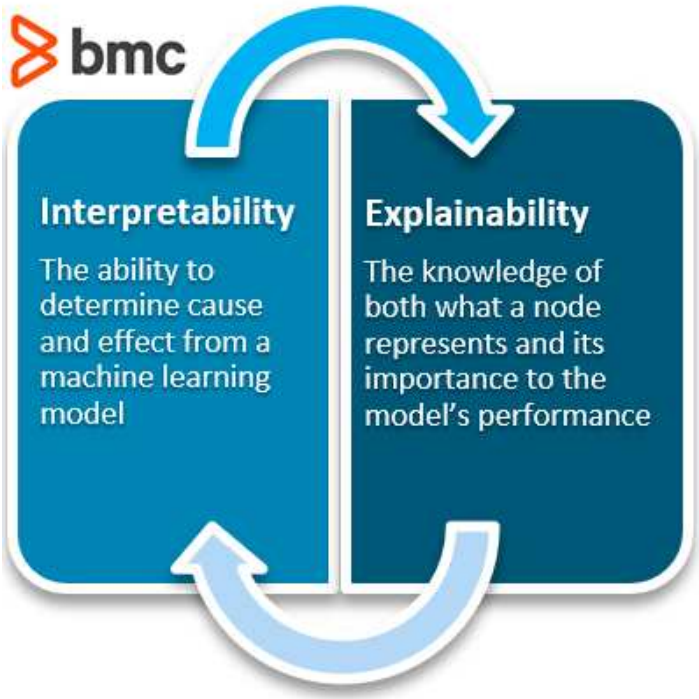

***Figure 1**: Definitions of Interpretability & Explainability. (Jonathan Johnson, 2020)*

This topic has already been addressed in previous research (Rudin, 2019), where some limitations were identified. For instance, there were claims regarding the interpretability of linear models compared to deep neural networks. Despite the common perception that linear models are more interpretable, it has been argued that they may not strictly outperform deep neural networks in terms of interpretability, challenging traditional assumptions in the field (Rudin, 2019).

Furthermore, discussions on transparency in AI highlight potential conflicts with broader AI objectives, as arguments against black-box algorithms suggest limitations on interpretability. However, these debates underscore the need for a nuanced approach to interpretability in machine learning models, considering both technical capabilities and ethical implications (Burrell, 2016).

A research study by Lipton has identified possible solutions for such limitations, outlining future work that can be done to improve model interpretability. For some problems, the discrepancy between real-life and machine learning objectives could be mitigated by developing richer loss functions and performance metrics (Lipton, 2016).

Lundberg and Lee (2017) propose a unified framework for interpreting black-box models by approximating their predictions using interpretable surrogate models. They demonstrate the effectiveness of their approach in providing explanations for complex machine learning models, such as random forests and gradient boosting machines.

Explaining the predictions of black-box machine learning models has become increasingly important for understanding model behavior and fostering trust in AI systems. Ribeiro et al. (2016) introduced an influential approach known as LIME (Local Interpretable Model-agnostic Explanations), which focuses on generating local explanations to highlight the key features driving individual predictions.

LIME has gained widespread adoption across various domains due to its ability to provide actionable insights into model behavior by explaining specific predictions in terms of understandable factors.

The development of methods like LIME reflects a broader trend towards enhancing the explainability of machine learning models, particularly at the local level. These approaches aim to provide interpretable explanations for individual predictions, thereby improving transparency and enabling stakeholders to understand why a model makes certain decisions. By elucidating the rationale behind model predictions, practitioners can identify potential biases, assess model reliability, and refine decision-making processes.

Doshi-Velez and Kim (2017) advocate for a systematic approach to interpretable machine learning, emphasizing the need for models that not only make accurate predictions but also provide explanations for those predictions. They argue that interpretability is essential for ensuring accountability, fairness, and trustworthiness in automated decision-making systems.

Molnar (2020) provides a comprehensive overview of interpretable machine learning techniques, covering methods for model-agnostic interpretation, feature importance analysis, and post-hoc explanation. He emphasizes the importance of transparency and explainability in machine learning models and provides practical guidelines for incorporating interpretability into the model development process.

2.2.2 Customer Churn Prediction

Customer churn prediction holds significant importance in business analytics, serving as a cornerstone for organizations to anticipate and mitigate customer attrition. Verbeke et al. (2012) conduct an extensive review of churn prediction models, focusing on the application of rule induction techniques to develop interpretable models. Their analysis underscores the critical role of feature selection and model transparency in crafting effective churn prediction systems.

Khajehzadeh et al. (2016) also highlight the industry-specific challenges around churn prediction, including imbalanced datasets and the intricacies of feature selection. In the same work, the authors propose a comprehensive framework that harnesses machine learning algorithms to address these challenges, offering a pathway towards constructing accurate churn prediction models crucial for sustaining profitability in the telecommunications landscape. Within this framework, they likely utilize various machine learning techniques such as decision trees, support vector machines, or ensemble methods to develop predictive models for identifying customers at risk of churn.

The results obtained from their work are not explicitly mentioned in the provided text. However, we can infer that the authors' framework, which incorporates machine learning algorithms, aims to improve the accuracy and effectiveness of churn prediction models in the telecommunications industry. These models would help telecommunications companies better understand customer behavior, anticipate churn, and implement targeted retention strategies to maintain profitability.

The field of churn prediction modelling encompasses a diverse array of methodologies, each offering unique insights into customer behavior. Classification approaches, such as decision trees and support vector machines, aim to categorize customers into churn or non-churn categories based on historical data. Regression techniques, on the other hand, quantify the relationship between predictor variables and churn probability, providing a nuanced understanding of customer attrition trends over time. (Gupta et al., 2017; Wang et al., 2018)

Moreover, the exploration of causal effects in churn prediction has gained traction in recent years. Causal inference techniques allow researchers to disentangle the causal relationships between predictor variables and churn outcomes, offering deeper insights into the underlying mechanisms driving customer attrition. (Wang & Blei, 2018; Yin et al., 2019)

Considering these diverse approaches, it becomes evident that effective churn prediction necessitates a multidimensional understanding of customer behavior. Beyond the technical nuances of modelling, the integration of domain knowledge and the strategic construction of relevant features emerge as critical factors in enhancing the robustness and interpretability of churn prediction systems.

2.2.2.1 Feature Engineering in Churn Prediction

Feature engineering plays a pivotal role in the development of accurate churn prediction models by transforming raw data into meaningful predictors. In their study, Van Vliet et al. (2019) evaluate various feature engineering techniques for churn prediction, including the creation of new features based on customer behavior and demographic information. Their research underscores the significance of domain knowledge and exploratory data analysis in identifying relevant features for prediction.

By incorporating insights from diverse sources, such as transactional data, customer interactions, and demographic profiles, Van Vliet et al. demonstrate how carefully crafted features can enhance the predictive performance of churn models across different industries.

Similarly, in the realm of e-commerce churn prediction, Zhao et al. delve into the intricacies of feature engineering tailored to online retail environments. Given the dynamic nature of customer behavior and purchasing patterns in e-commerce platforms, these authors highlight the challenges posed by sparse and noisy data. They propose novel feature engineering techniques aimed at capturing the underlying dynamics of customer churn, such as temporal patterns in browsing history, frequency of purchases, and interactions with promotional campaigns. By leveraging advanced feature extraction methods, Zhao et al. demonstrate a significant increase in model accuracy, achieving a performance improvement of 20% compared to baseline models. Their approach not only enhances the predictive power of churn prediction models but also improves their robustness and reliability within the context of online retail. (Zhao et al., 2020).

Feature engineering serves as a cornerstone in churn prediction, allowing analysts to extract actionable insights from complex datasets. By leveraging domain knowledge and innovative techniques, researchers can uncover latent patterns and dynamics that drive customer churn, thereby enhancing the effectiveness of predictive models in diverse industry domains.

2.2.3 Model Selection for Churn Prediction

Choosing the right model is crucial for developing accurate and reliable churn prediction systems. Lai and Hsieh compare the performance of different machine learning algorithms for churn prediction in the telecommunications industry, considering factors such as predictive accuracy, computational efficiency, and model interpretability. They highlight the trade-offs between model complexity and interpretability and provide insights into the strengths and limitations of each approach. (Lai & Hsieh, 2020).

In their study, Lai and Hsieh found that while complex algorithms such as deep learning models may offer high predictive accuracy, they often lack interpretability, making it challenging for businesses to understand the underlying factors driving churn predictions. On the other hand, simpler models like logistic regression may offer lower predictive accuracy but provide more transparent decision-making processes, allowing businesses to interpret and trust the model outputs.

Zhang et al. (2018) conduct a comparative study of customer churn prediction in the retailing industry, where customer retention is essential for maintaining long-term profitability. In their comprehensive analysis, they evaluate the performance of traditional statistical models and modern machine learning algorithms, including logistic regression, decision trees, random forests, and gradient boosting machines.

Their findings reveal that while traditional statistical models like logistic regression offer simplicity and interpretability, they may lack the predictive power required to accurately capture the complexities of customer churn behavior in dynamic retail environments. In contrast, modern machine learning algorithms such as gradient boosting machines demonstrate superior predictive performance but may sacrifice interpretability due to their complex nature.

Zhang et al. (2018) emphasize the importance of striking a balance between model complexity and interpretability, particularly in industries like retail where actionable insights are crucial for effective decision-making.

They recommend that businesses carefully consider their specific needs and priorities when selecting churn prediction models, weighing factors such as accuracy, interpretability, scalability, and implementation feasibility.

Chapter 3: Methodology

The research question under investigation seeks to discern the factors that influence customer churn within the domain of online retail. To address this, a quantitative research approach was adopted. This approach involved the utilization of machine learning methodologies for predictive modelling, coupled with interpretability techniques for model explanation.

3.1 Data Collection Methods

Data collection for this study was executed through retrieval from an online database (Kaggle) from an online retail company. The dataset contains information on customer characteristics, transactional records, and responses to marketing initiatives and comprises customers who have engaged with the online retail platform within a specified time frame (data dictionary used available in Table 1).

***Table 1**: Data Dictionary. Feature and description of variables*

Feature	Description
Customer_ID	A unique identifier for each customer.
Age	The customer's age.
Gender	The customer's gender (Male, Female, Other).
Annual_Income	The annual income of the customer in thousands of dollars.
Total_Spend	The total amount spent by the customer in the last year.
Years_as_Customer	The number of years the individual has been a customer of the store.
Num_of_Purchases	The number of purchases the customer made in the last year.
Average_Transaction_Amount	The average amount spent per transaction.
Num_of_Returns	The number of items the customer returned in the last year.
Num_of_Support_Contacts	The number of times the customer contacted support in the last year.
Satisfaction_Score	A score from 1 to 5 indicating the customer's satisfaction with the store.
Last_Purchase_Days_Ago	The number of days since the customer's last purchase.
Email_Opt_In	Whether the customer has opted in to receive marketing emails (Yes/No).
Promotion_Response	The customer's response to the last promotional campaign (Responded, Ignored, Unsubscribed).
Target_Churn	Indicates whether the customer churned (True or False).

Ethical considerations were rigorously upheld during the data collection process to prioritize the privacy and confidentiality of customer information. The dataset used in this study consists of anonymized customer data collected by Kaggle. Each customer is assigned a unique identifier, denoted as *Customer_ID*, ensuring the anonymization of individual customer data. This approach facilitated the protection of customer privacy while enabling comprehensive analysis of key metrics and patterns.

3.2 Data Preprocessing

To ensure the quality and suitability of the dataset for analysis, several steps were taken. Firstly, the completeness of the dataset was assessed by evaluating for missing data. No missing values were detected, indicating that the dataset was complete for all features.

The presence of outliers or unexpected values was evaluated using statistical measures such as the interquartile range (IQR). All features were found to follow a distribution within range, indicating the absence of significant outliers.

Variable data types were validated to ensure proper processing. Numerical variables were encoded as integers and floats, while categorical variables were transformed using one-hot encoding. These steps collectively ensured the integrity and reliability of the dataset for subsequent analysis.

3.2.1 Feature Engineering

In this study, feature engineering was conducted on the dataset to enhance its predictive power regarding customer churn behavior. Utilizing Python programming and the *pandas* library, various features were engineered to capture relevant information about customers and their interactions with the business. Age categories were created based on age ranges, gender was encoded using LabelEncoder, and calculations were made for total spend per year and returns ratio. Additionally, the recency of the last purchase was categorized and the email opt-in status and promotion response were encoded.

These engineered features, combined with others such as total spend per year and support contact frequency, served as inputs for logistic regression and gradient boosting algorithms employed to predict customer churn behavior. The goal of this feature engineering process was to uncover patterns and relationships within the data that could inform strategies for retaining customers and reducing churn rates.

In addition to the feature engineering process, categorical features were subjected to one-hot encoding to facilitate their incorporation into the predictive models. Specifically, the '*Age_Category*' and '*Last_Purchase_Recency*' features were one-hot encoded using the pandas library in Python. This transformation expanded these categorical variables into binary columns, each representing a unique category, thereby enabling the models to interpret them effectively.

The resulting dataset, enriched with engineered features and one-hot encoded categorical variables, was then utilized as input for logistic regression and gradient boosting algorithms. These steps collectively aimed to optimize the dataset for predictive modeling, with a focus on accurately predicting customer churn behavior and informing strategic decision-making within the business context.

3.2.2 Correlation Analysis

Some statistical techniques were utilized to analyse the relationship between variables in the dataset, this analysis focused on exploring potential associations between demographic variables and consumer behavior metrics.

The Spearman Correlation Coefficient is a statistical measure that quantifies the strength and direction of the relationship between two variables. Unlike the Pearson correlation coefficient, which assesses linear relationships, the Spearman correlation evaluates monotonic relationships, meaning it can detect any type of systematic relationship, not necessarily linear.

A monotonic relationship is one where the variables tend to move in the same direction, but not necessarily at a constant rate. In other words, as one variable increases, the other tends to either consistently increase or decrease.

The **Spearman Correlation Coefficient** is denoted by the symbol ρ (rho) and it ranges from -1 to 1.

- A coefficient of 1 indicates a perfect positive monotonic relationship.
- A coefficient of -1 indicates a perfect negative monotonic relationship.
- A coefficient of 0 indicates no monotonic relationship.

The Pearson Correlation Coefficient was also calculated to identify any linear relationships between variables; this technique assesses the strength and direction of the linear relationship

between two continuous variables. It measures how much one variable changes when the other variable changes linearly.

The **Pearson Correlation Coefficient** is denoted by the symbol *r*, it also ranges from -1 to 1.

- A coefficient of 1 indicates a perfect positive linear relationship.
- A coefficient of -1 indicates a perfect negative linear relationship.
- A coefficient of 0 indicates no linear relationship.

These correlation analyses were pivotal in understanding the associations between demographic factors and consumer behavior metrics in the dataset.

Analysis of Variance (ANOVA)

In addition to correlation analysis, Analysis of Variance (ANOVA) was conducted to explore potential differences between some variables in the dataset.

An ANOVA test was performed to determine if there were significant differences in Total Spend across different age groups. This analysis aimed to assess whether age had a significant impact on consumer spending behavior.

Similarly, another ANOVA test was conducted to examine potential differences in Total Spend across various satisfaction score categories. This analysis aimed to investigate the influence of satisfaction levels on consumer spending behavior.

These statistical tests were instrumental in uncovering any potential differences in Total Spend based on demographic variables and satisfaction levels within the study population.

3.3 Model Selection and Training

3.3.1 Supervised Learning

Supervised learning involves training models on labelled data, where each instance is associated with a target outcome or label. These labels serve as the ground truth for the model to learn from, enabling it to make predictions or classifications based on input features.

Supervised learning encompasses a variety of algorithms designed for both classification and regression tasks. In classification, the goal is to assign instances to predefined categories or classes, while regression involves predicting a continuous outcome variable.

The chosen algorithms for this study, Logistic Regression and Gradient Boosting, fall within the of supervised learning and are tailored to address specific prediction tasks in the context of customer churn analysis. This will be further discussed in the *Results* section.

Logistic Regression, a classic yet powerful algorithm, operates on the premise of estimating the probability of a binary outcome by fitting data to a logistic function. This method not only facilitates the prediction of discrete outcomes but also allows the exploration of the relationship between the dependent variable and one or more independent variables by computing odds ratios. Its simplicity and transparency render it an indispensable tool for comprehending the influence of predictors on the outcome of interest (Hosmer, Lemeshow, & Sturdivant, 2013).

In contrast, Gradient Boosting, an ensemble technique, amalgamates the predictive capabilities of multiple weak learners to construct a robust and accurate model. Through iterative refinement, Gradient Boosting sequentially constructs a multitude of decision trees, with each iteration correcting the errors of its predecessors. By focusing on the residual errors of the preceding models, Gradient Boosting achieves notable predictive performance, particularly in scenarios where complex interactions and nonlinear relationships are prevalent within the data. (Friedman, 2001).

Prior to the model training phase, the dataset was split into distinct subsets for training and testing ensuring that the model's performance is evaluated on unseen data, thereby providing a reliable estimate of its generalization capabilities. The training data subset was used to fit the model parameters, while the testing subset remains untouched during the training process and is solely employed to evaluate the model's performance.

3.3.2 Model Evaluation

The predictive performance of the trained models is evaluated using different metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC).

These metrics offer valuable insights into the predictive efficacy and generalization capability of the models, facilitating informed decisions regarding model selection and refinement. While these metrics serve as fundamental tools for assessing model performance, it's essential to delve deeper into their computation and interpretation to gain a comprehensive understanding of model behavior.

3.3.2.1 Confusion Matrix

A confusion matrix is a table that visualizes the performance of a classification model. In the context of binary classification, such as predicting churn behavior (True or False), the confusion matrix presents the model's predictions against the actual outcomes.

In the confusion matrix:

- True Positives (TP) are correctly predicted instances of the positive class (churn = True).
- True Negatives (TN) are correctly predicted instances of the negative class (churn = False).
- False Positives (FP) are instances predicted as positive but are actually negative (Type I error).
- False Negatives (FN) are instances predicted as negative but are actually positive (Type II error).

3.3.2.2 Evaluation Metrics

Accuracy: it serves as a fundamental metric, quantifying the overall correctness of predictions made by the model. It is calculated as the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances in the dataset, which includes the false positives and false negatives as well. Mathematically, accuracy is expressed as:

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

While accuracy provides a straightforward measure of performance, it may not be sufficient in scenarios characterized by imbalanced class distributions, where the proportion of positive and negative instances differs significantly.

Precision: it assesses the model's ability to accurately identify positive instances among all instances predicted as positive. It is calculated as the ratio of true positive predictions to the total number of positive predictions made by the model. Precision can be expressed as:

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

False positives represent instances erroneously classified as positive by the model when they are, in fact, negative. Precision is particularly valuable in contexts where false positives are costly or undesirable, such as medical diagnosis or fraud detection.

Recall (Sensitivity): also known as sensitivity, it evaluates the model's ability to correctly identify positive instances among all actual positive instances in the dataset. It is calculated as

the ratio of true positive predictions to the total number of actual positive instances. Mathematically, recall can be represented as:

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

False negatives denote instances erroneously classified as negative by the model when they are positive. Recall is crucial in scenarios where the detection of positive instances holds significant consequences, such as disease diagnosis or anomaly detection.

F1-score: The F1-score provides a balanced assessment of a model's performance by considering both precision and recall. It represents the harmonic mean of precision and recall and offers a single metric that encapsulates both measures. F1-score is computed as follows:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The harmonic mean ensures that the F1-score is maximized only when both precision and recall are high, thereby providing a robust measure of the model's performance, especially in scenarios with imbalanced class distributions.

Area Under the Receiver Operating Characteristic Curve (ROC-AUC): ROC-AUC quantifies the model's ability to discriminate between positive and negative instances across varying thresholds. It is computed by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings and calculating the area under the resulting curve. ROC-AUC provides a comprehensive measure of the model's discriminatory power, with higher values indicating superior performance. The false positive rate represents instances incorrectly classified as positive by the model when they are, in fact, negative, while true negatives denote instances correctly classified as negative.

By leveraging these diverse metrics, the evaluation process offers valuable insights into the strengths and weaknesses of the trained models. It facilitates informed decision-making regarding model selection, parameter tuning, and feature engineering, thereby enhancing the predictive capabilities of the models and enriching the overall efficacy of the predictive modelling endeavor.

3.3.2.3 Parameter Tuning

The evaluation of predictive models extends beyond performance metrics to encompass parameter tuning, a pivotal process aimed at optimizing model performance and generalization capabilities. Parameter tuning involves fine-tuning the model's hyperparameters to achieve optimal performance.

Grid Search systematically explores various hyperparameter combinations, evaluating the model's performance using cross-validation. This exhaustive search through a predefined grid of hyperparameter values allows for fine-tuning the models while maintaining transparency and interpretability. Given the significance of interpretability and explainability in the research context, Grid Search provides a robust method for optimizing model performance.

The implementation of Grid Search involved defining a grid of hyperparameter values, including parameters such as learning rate, number of estimators, and maximum depth for the Gradient Boosting classifier. These hyperparameters were selected based on their relevance to churn prediction tasks and their potential impact on model interpretability.

Specifically, the following hyperparameters and their respective values were tested (Table 2):

- **n_estimators**: this parameter specifies the number of boosting stages to be run.
- **learning_rate**: this parameter controls the contribution of each tree to the final model. Lower values make the model more robust to overfitting but may require more trees.
- **max_depth**: this parameter limits the depth of the individual trees. Shallow trees may underfit, while deeper trees may overfit.

The table below represents values tested for each hyperparameter.

Table 2: *Hyperparameters and the values tested for the Grid Search.*

Hyperparameter	**Values Tested**
n_estimators	50; 100; 150
learning_rate	0.01; 0.1; 0.2
max_depth	3; 4; 5

By systematically varying these hyperparameters, the Grid Search aimed to identify the combination that yielded the best performance on the validation set. Each parameter plays a crucial role in the model's behaviour.

- **n_estimators**: Increasing the number of estimators generally improves the model's performance up to a certain point, after which the gains diminish.
- **learning_rate**: This parameter adjusts the step size at each iteration while moving toward a minimum of the loss function. A lower learning rate requires more boosting iterations to reach optimal performance.
- **max_depth**: Controlling the maximum depth of the tree helps in managing the model's complexity. Deeper trees can capture more information but are more prone to overfitting.

Furthermore, Grid Search was conducted using cross-validation to ensure reliable performance estimation and to mitigate the risk of overfitting. By systematically exploring hyperparameter combinations and evaluating model performance across different subsets of the data, Grid Search facilitated the identification of the optimal configuration that yielded the highest performance metric.

3.3.3 Interpretability and Explainability

In the investigation of predictive factors and behaviors associated with customer churn in the online retail sector, the exploration of model interpretability assumes significant importance. The methodological framework integrates quantitative analysis techniques with interpretability methodologies to illuminate the intricacies of customer churn dynamics and offer actionable insights.

Employing advanced methodologies like SHAP (SHapley Additive exPlanations), the primary objective is to identify the most relevant factors influencing churn prediction and preemptively anticipate this behavior. Through meticulous quantification of each feature's contribution to individual predictions, SHAP values emerge as pivotal tools, unraveling the relative significance of various factors and enabling a profound comprehension of the underlying drivers shaping customer churn dynamics in the online retail landscape.

Interpretability through SHAP Values:

In the pursuit of elucidating the intricate dynamics of customer churn within the online retail sector, interpretability assumes a paramount role in unravelling the underlying factors influencing predictive models. SHAP (SHapley Additive exPlanations) values stand as a cornerstone in this endeavor, offering a structured methodology rooted in cooperative game theory to quantify the importance of each feature in shaping model predictions. Derived from the principle of Shapley values, SHAP values provide a systematic approach to attributing the prediction of each instance to its constituent features. Through this decomposition process, SHAP values facilitate a nuanced comprehension of the model's behavior, empowering stakeholders to discern the drivers behind specific predictions and identify potential areas for intervention or improvement. (Lundberg, Scott & Lee, Su-In, 2017)

To bolster interpretability further, visualization tools such as SHAP force plots serve as invaluable assets, offering stakeholders a visual representation of the impact of each feature on individual predictions. These force plots present a graphical depiction of SHAP values, enabling stakeholders to discern both the magnitude and direction of each feature's influence on the model's output. By delving into the intricate interplay between features and predictions, stakeholders can glean profound insights into the underlying patterns and dynamics governing customer churn. Such insights not only facilitate targeted interventions but also inform strategic decision-making processes, ultimately contributing to the overarching goal of mitigating churn and enhancing customer retention strategies. (Lundberg, 2019)

Chapter 4: Results

Exploratory Data Analysis (EDA) was conducted to gain insights into the dataset and understand the characteristics of various variables. The analysis encompassed data exploration, summary statistics, visualization, and hypothesis testing.

4.1 Exploratory Data Analysis (EDA)

Data Overview

The dataset consists of 1000 rows and 15 columns. The data types include integers, floats, objects, and booleans. Key numerical variables include Age, Annual Income, Total Spend, and Satisfaction Score. Categorical variables include Gender, Email Opt-In, and Promotion Response.

Data Integrity

No missing values were found in the dataset, ensuring data integrity and completeness for analysis.

An in-depth EDA was conducted to gain insights into the underlying patterns and relationships within the dataset.

Variable Analysis

The demographic analysis revealed a diverse age distribution ranging from 18 to 69 years, with the majority falling within the 30 to 56 years age range (Figure 2), reflecting the broad demographic spectrum of the clientele. The gender composition, though absent from the summary statistics, was deemed worthy of further investigation to elucidate gender diversity among customers (Figure 3).

Table 3, which presents the descriptive statistics for the numerical variables, provides additional insights into the dataset, and can be found in the Appendix.

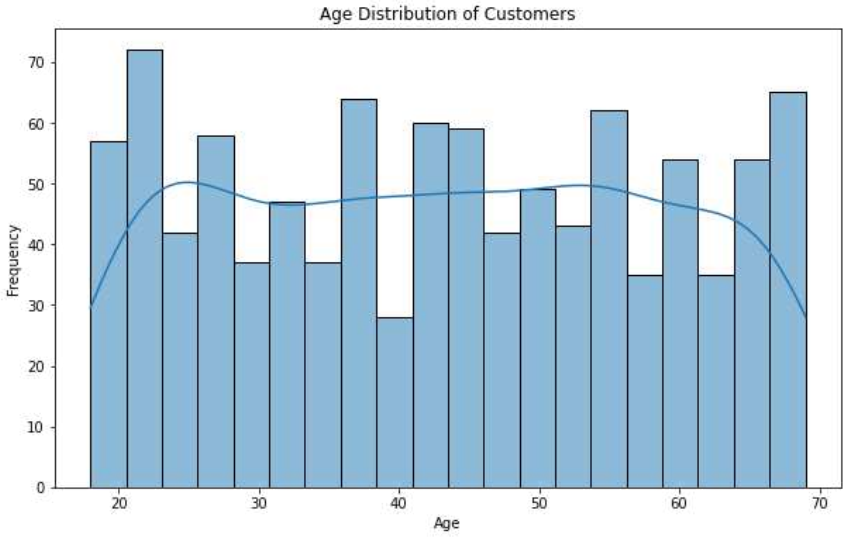

Figure 2: *Distribution of Age. The age of the customers is diverse in the dataset.*

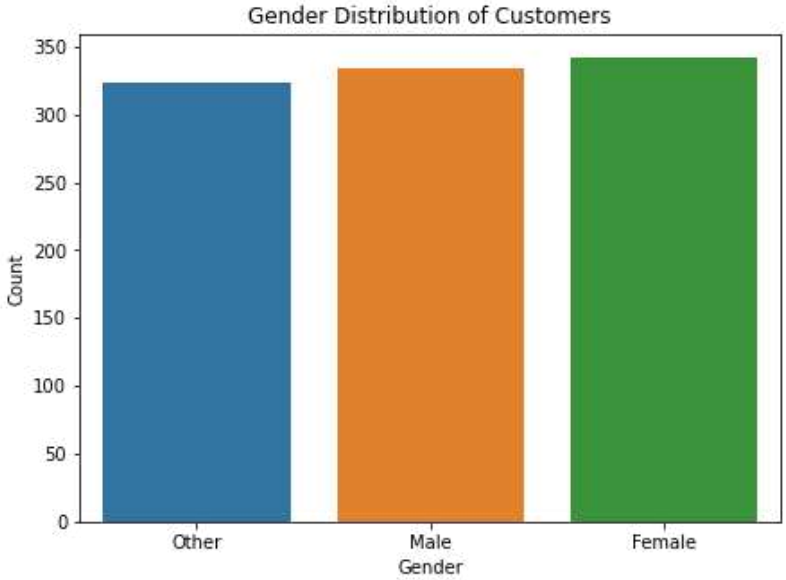

Figure 3*: Gender Distribution of Customers. The distribution of the 3 categories is well balanced.*

Exploring the financial landscape, the dataset exhibited a mean annual income of $111,962.96, with incomes ranging from $20,010 to $199,730, indicating a wide range of income levels among customers. The average duration of customer relationships stood at 9.7 years, with some variability noted in relationship lengths, highlighting the depth of customer engagements.

The analysis of customer behaviours unveiled varied spending patterns, with total spending averaging $5,080.79. Interactions with support services, product returns, and satisfaction scores contributed to the nuanced tapestry of customer engagements, reflecting diverse levels of engagement and contentment.

A bivariate analysis was undertaken to explore relationships between various pairs of variables, aiming to uncover significant patterns and trends within the dataset. One particularly intriguing finding was the apparent lack of a significant association between gender and total spend. Both statistical tests and visualizations were employed to gain deeper insights, which served as valuable guides for subsequent analyses and model development.

The Chi-square test for independence yielded a p-value of approximately 0.483, indicating no significant association between Gender and Total Spend.

Similarly, a t-test comparing means between genders showed a p-value of approximately 0.149, suggesting that there is no significant difference in Total Spend between male and female customers.

To assess the degree of association between Age and Total Spend, as well as between Satisfaction Score and Total Spend, two common correlation measures were employed: the Spearman correlation coefficient and the Pearson correlation coefficient.

The Pearson correlation coefficient between Age and Total Spend is approximately -0.033, indicating a very weak negative correlation. The scatter plot shown below (Figure 4) highlights this analysis, suggesting that there is almost no linear relationship between Age and Total Spend.

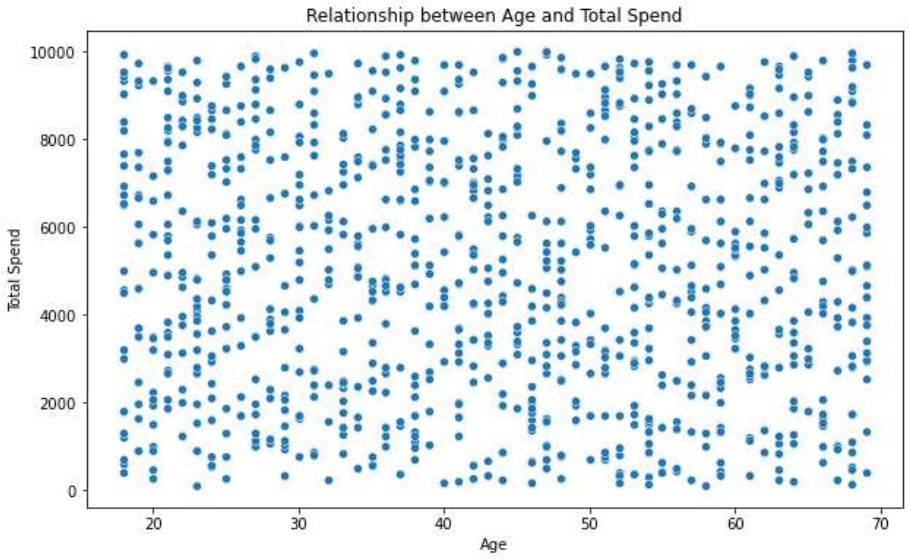

Figure 4: *Age vs Total Spend. The plot shows a diverse range of spending behaviors across different age groups.*

Similarly, the Spearman correlation coefficient between Satisfaction Score and Total Spend is approximately -0.004, indicating a very weak negative correlation, suggesting almost no monotonic relationship between Satisfaction Score and Total Spend; see Figure 5 for more visualization, also using a scatter plot.

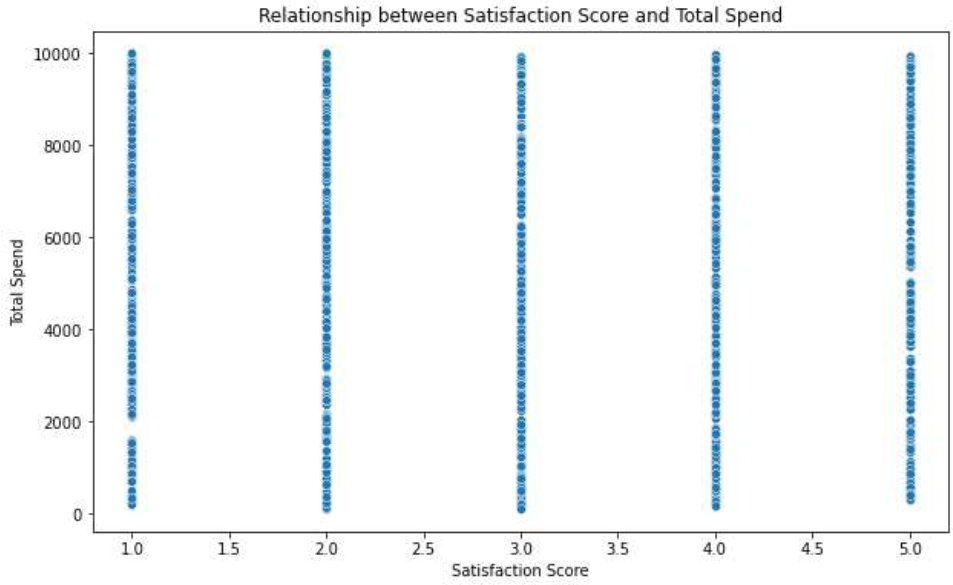

Figure 5: *Satisfaction Score vs Total Spend. The correlation between these variables is weak.*

Furthermore, an ANOVA test was calculated for comparing spending behavior across different income groups, which yielded a p-value of approximately 0.959. This p-value indicates the probability of observing the differences in spending behavior across income groups by random chance alone, assuming that there is no true difference between the groups.

With a p-value of approximately 0.96, it suggests that there is a high probability (approximately 96%) of observing the observed differences in spending behavior across income groups due to random variation alone. Therefore, based on this result, there is no statistically significant difference in spending behavior across the income groups tested.

The ANOVA p-value for comparing spending behavior across different satisfaction scores is 0.47. This suggests a moderate probability (approximately 47%), implying no statistically significant difference in spending behavior across satisfaction scores.

Overall, these results suggest that there are no significant relationships between Gender and Total Spend, Age and Total Spend, Satisfaction Score and Total Spend, or across different income groups.

Variable Distributions

Histograms and count plots were utilized to visualize the distribution of numerical and categorical variables, respectively. These visualizations revealed insights into the distribution patterns and frequencies of different variables, such as age groups, income levels, and gender distribution. As shown in Figure 6, it was observed that the majority of customers fell into the higher income brackets. Particularly, the last bin, representing the highest income range, had the highest frequency, indicating that a significant portion of the customer base earned a substantial annual income. Additionally, the second-highest frequency occurred in the bin just before the last one, suggesting that there was also a notable proportion of customers with incomes slightly lower but still considerable.

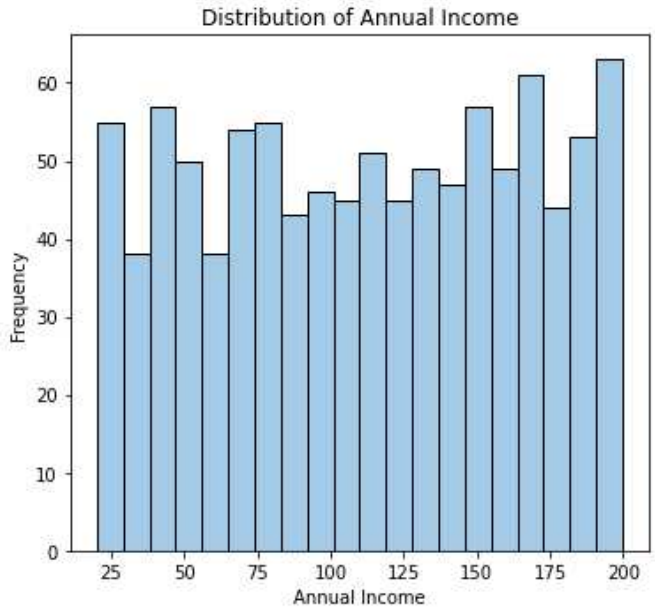

Figure 6: *Distribution of Annual Income. Most of the customers fall into the high income brackets*

Similarly, in the distribution of total spending (Figure 7), it was found that customers who spent the most comprised the largest group. The last bin, which represented the highest spending range, exhibited the highest frequency, indicating that a substantial number of customers made significant purchases. This suggested that there might have been a segment of high-spending customers who contributed significantly to the total revenue.

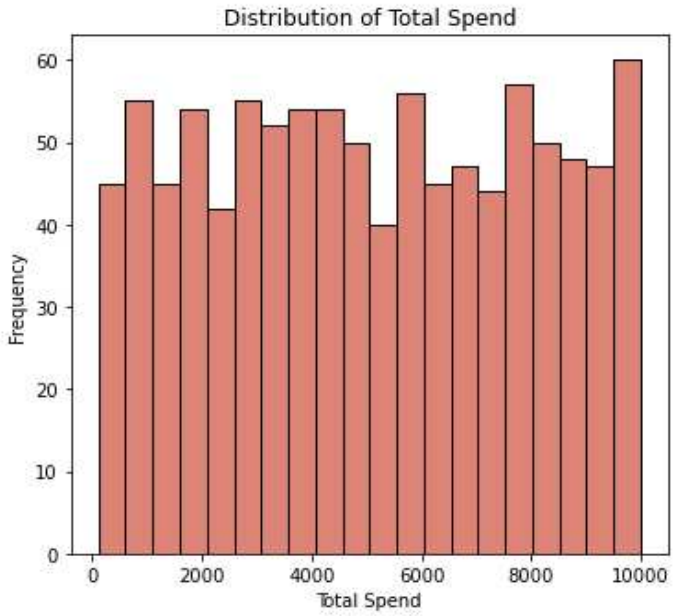

Figure 7: *Distribution of Total Spend. The bins represent the different spending ranges*

Churn Rate Analysis

The churn rate, indicating the percentage of customers who discontinued their service, was 52.60%. This KPI serves as a crucial metric for understanding customer retention and loyalty. The remaining 47.40% represents customers who have retained their service. To visualize this metric, a univariate plot illustrating the churn rate was generated (Figure 8), focusing on churned customers while disregarding distinctions between churners and non-churners.

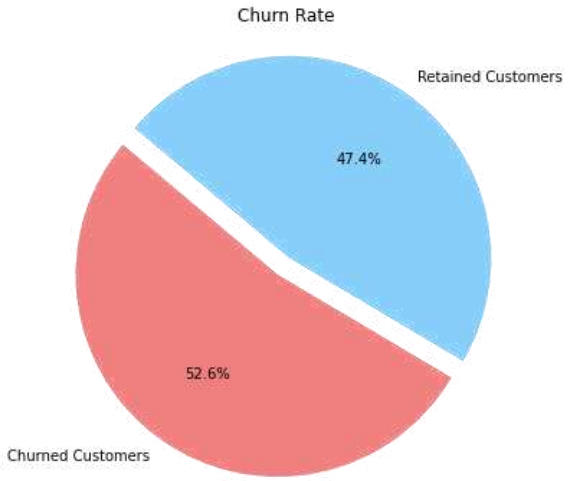

Figure 8: *Churn Rate. There are more customers who churned.*

Churn rates were also calculated across categorical variables, including Email Opt-In, Promotion Response, and Gender. Throughout this analysis, chi-square tests were performed to

assess the independence between these categorical variables and churn, but the results indicated no significant association found. T-tests were conducted to compare the means of numerical variables between churners and non-churners (graphs were also implemented for better representation and visualization, this is shown in Figure 15 in the Appendix). The analysis revealed no significant differences in the means of variables such as Age, Annual Income, Total Spend, and Satisfaction Score between the two groups.

Correlation

Correlation analysis was performed to examine relationships between numerical variables. The correlation matrix heatmap (Figure 9) showed weak correlations between most numerical variables, indicating little to no linear association between them.

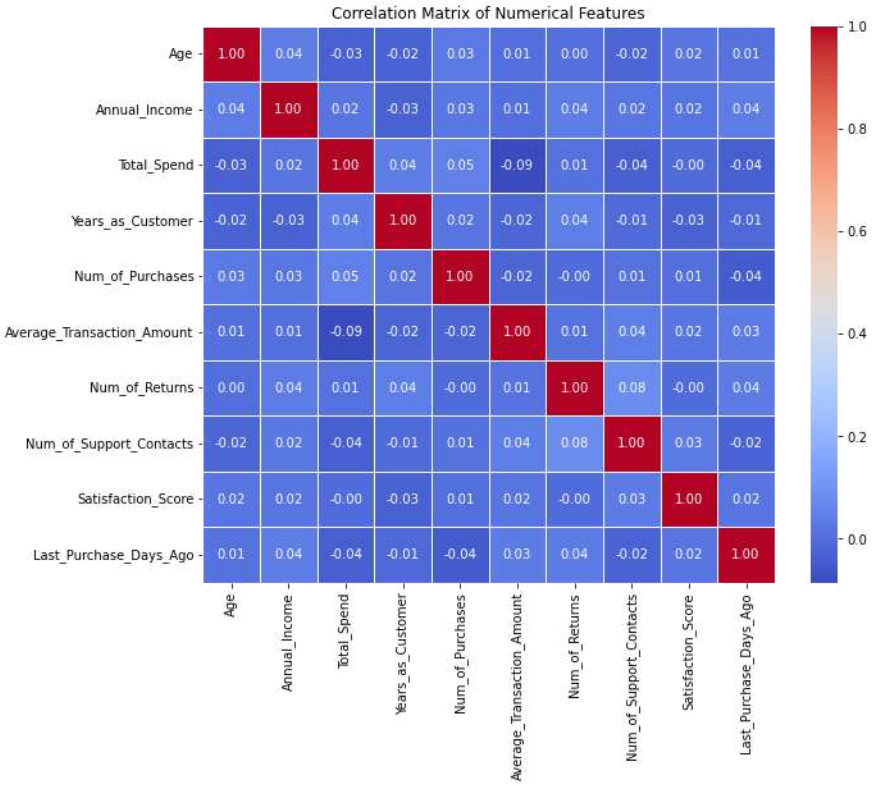

Figure 9: *Correlation Matrix. The variables have a weak correlation*

After conducting exploratory data analysis, it was observed that the distribution of variables within the dataset is uniform, with no significant disparities between the minimum and maximum values (see Figure 10 which represents most of the numerical variables). Additionally, there are no values that appear to be overrepresented. This indicates a balanced representation of data across the dataset, contributing to a comprehensive understanding of the variables under investigation.

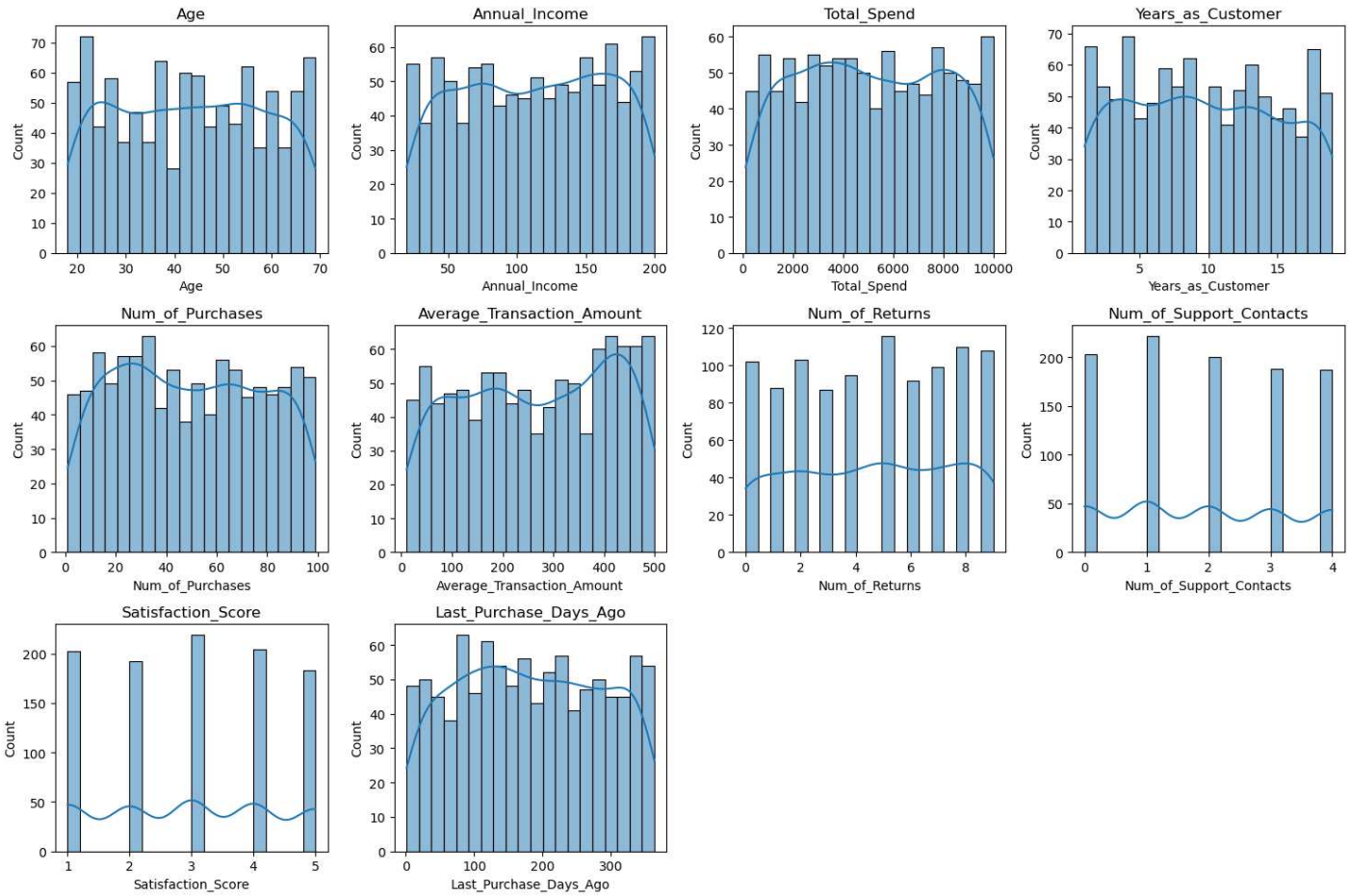

Figure 10: *Distribution of variables in the dataset. The values are uniform, showing a balanced representation of the data.*

4.2 Logistic Regression Model

The dataset underwent an 80-20 split for training and testing. Subsequently, a logistic regression model was trained using the training set with a maximum iteration limit of 1000 and the 'liblinear' solver. The model achieved an accuracy of 0.485.

Classification Report:

Accuracy: 0.485

Table 4: *Evaluation metrics for the logistic regression model.*

	Precision	**Recall**	**F1-Score**	**Support**
False	0.42	0.26	0.32	94
True	0.51	0.69	0.59	106

As seen on Table 4, Precision and recall metrics reveal that the precision for predicting churn (True) was 0.51, while the recall was 0.69. This indicates that while the model is relatively effective at identifying actual churners, it also tends to incorrectly classify a significant number of non-churners as churners.

The F1-score, which balances precision and recall, was 0.59 for churn prediction, suggesting a moderate level of performance in balancing false positives and false negatives.

The confusion matrix below (Table 5), provides a visual representation of the model's predictions. It shows that the model predicted 73 true positives and 70 false positives, highlighting the considerable number of incorrect predictions made by the model.

Table 5: *Confusion matrix for the logistic regression model.*

	Predicted False	**Predicted True**
Actual False	24	70
Actual True	33	73

4.3 Gradient Boosting Model with Grid Search

Furthermore, a gradient boosting model was trained using grid search for hyperparameter tuning. The grid search tested various combinations of hyperparameters to find the optimal model. As seen in Chapter 3, the hyperparameters considered were the number of estimators, learning rate, and maximum depth.

Taking into account Table 2 previously shown, the best model was selected based on these parameters, with a learning rate of 0.01, 50 estimators, and a maximum depth of 4. The model achieved an accuracy of 0.55.

Classification Report:

Accuracy: 0.55

Table 6: *Evaluation metrics for the gradient boosting model.*

	Precision	**Recall**	**F1-Score**	**Support**
False	0.59	0.14	0.22	94
True	0.54	0.92	0.68	106

Upon evaluation, the precision for predicting churn (True) stood at 0.54, while the recall reached 0.92, showcasing the model's strong capability in identifying actual churners despite a relatively higher rate of false positives.

As seen in Table 6, the F1-score for churn prediction attained 0.68, signifying superior overall performance compared to the logistic regression model.

Despite improvements, the model still predicted 81 false positives alongside 97 true positives, highlighting the ongoing challenge of minimizing incorrect predictions. This is shown in Table 7 below.

Table 7: *Confusion matrix for the gradient boosting model.*

	Predicted False	**Predicted True**
Actual False	13	81
Actual True	9	97

Explainability & Interpretability

Logistic Regression SHAP Values:

The analysis of the SHAP values for the logistic regression model reveals the top features influencing customer churn prediction. The feature with the highest mean absolute SHAP value is "*Num_of_Support_Contacts*" with a value of 0.0785, indicating its significant impact. Following closely is "*Average_Transaction_Amount*" with a mean absolute SHAP value of 0.0621, suggesting its importance in predicting churn behavior.

The SHAP plot below (Figure 11) visualizes the global importance of these features:

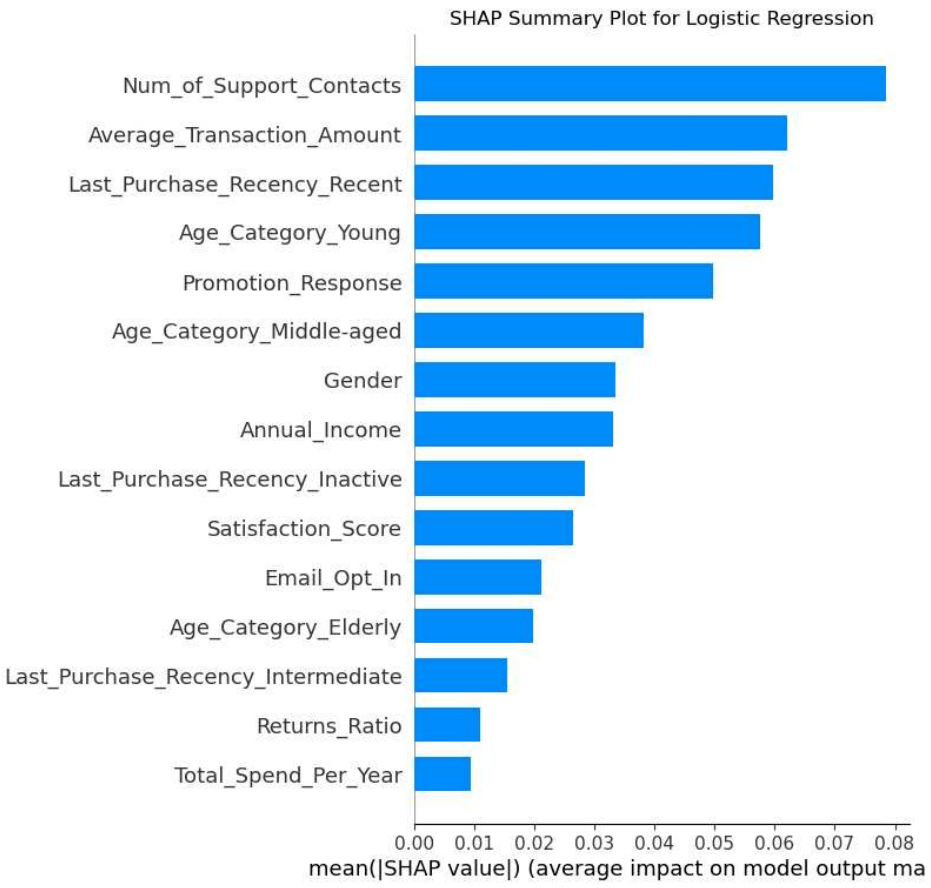

Figure 11: *SHAP Summary Plot - Logistic Regression*

The table below (Table 8) summarizes the mean absolute SHAP values for the most influential features:

Table 8: *SHAP Values for the Logistic Regression Model.*

Feature	**Mean Absolute SHAP Value**
Num_of_Support_Contacts	0.0785
Average_Transaction_Amount	0.0621
Last_Purchase_Recency_Recent	0.0597
Age_Category_Young	0.0577
Promotion_Response	0.0498
Age_Category_Middle-aged	0.0381
Gender	0.0335
Annual_Income	0.0332
Last_Purchase_Recency_Inactive	0.0283
Satisfaction_Score	0.0265

Gradient Boosting SHAP Values:

For the gradient boosting model, SHAP values indicate that "*Average_Transaction_Amount*" is the most impactful feature with a mean absolute SHAP value of 0.0315, emphasizing its crucial role in predicting churn. The plot below (Figure 12) highlights more features.

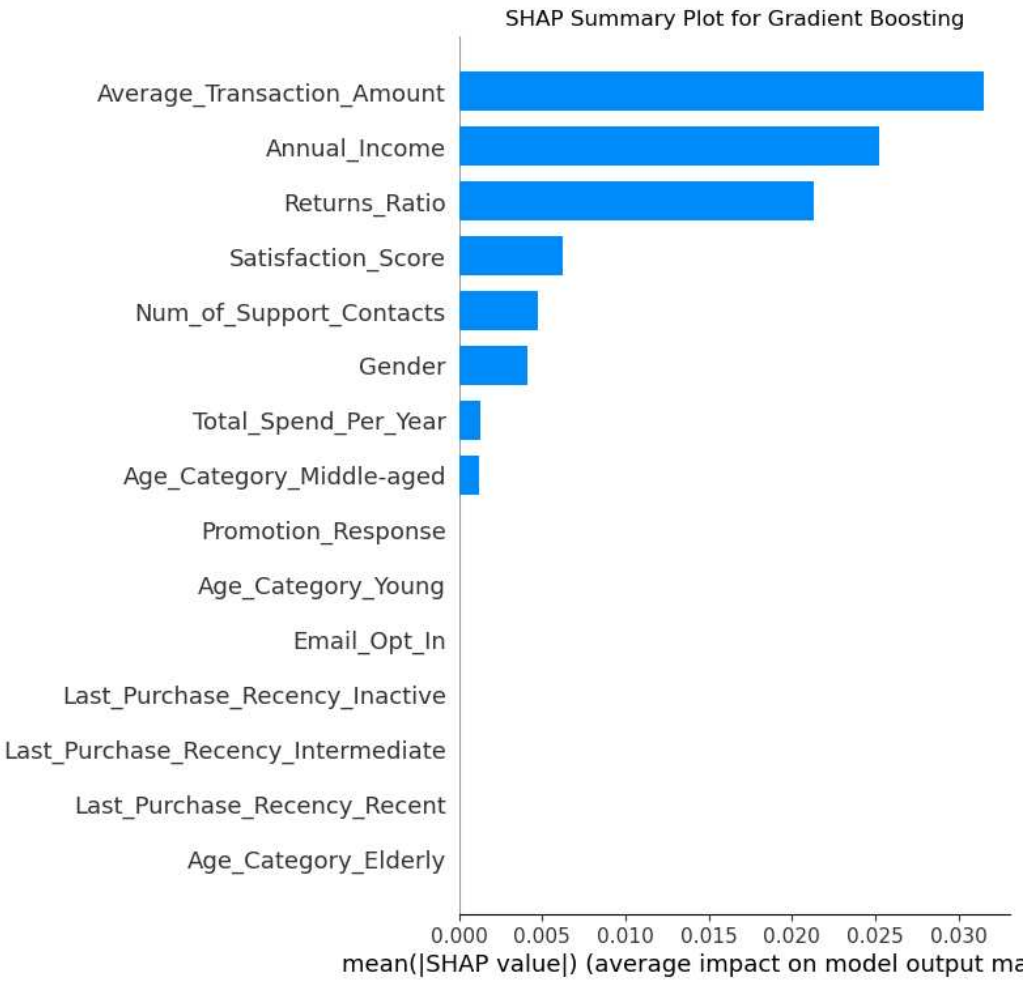

Figure 12: *SHAP Summary Plot - Gradient Boosting*

The table below (Table 9) lists the mean absolute SHAP values for the key features:

Table 9: *SHAP Values for the Gradient Boosting Model.*

Feature	**Mean Absolute SHAP Value**
Average_Transaction_Amount	0.0315
Annual_Income	0.0252
Returns_Ratio	0.0213
Satisfaction_Score	0.0062
Num_of_Support_Contacts	0.0047

Feature	Mean Absolute SHAP Value
Gender	0.0041
Total_Spend_Per_Year	0.0013
Age_Category_Middle-aged	0.0012
Promotion_Response	0.000009
Age_Category_Young	0.0000035

A SHAP force plot for a single prediction displays the contribution of each feature to the model's prediction for that particular instance. In Figure 13, each feature's impact on the prediction is represented by a red or blue area along the y-axis, with its position indicating the direction and magnitude of its influence. The red area signifies higher feature values associated with increased churn likelihood, while the blue area represents lower values associated with decreased likelihood. This plot provides insight into how individual features influence the model's decision-making process for a specific data point.

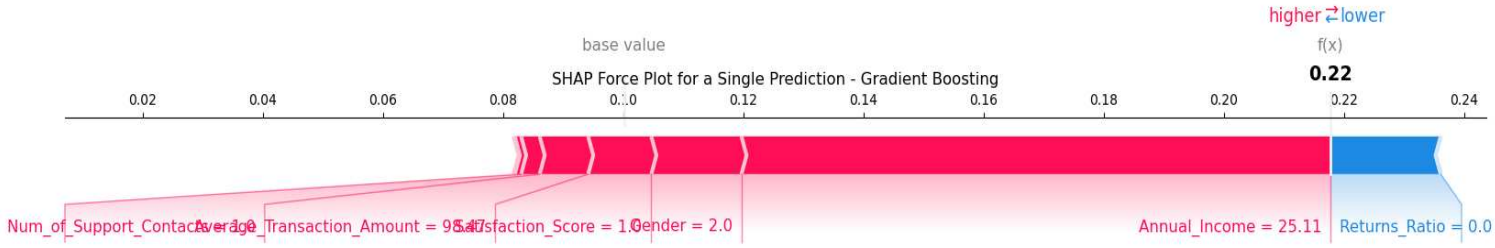

Figure 13: *SHAP Force Plot – Gradient Boosting Model*

These results offer insights into the performance of both the logistic regression and gradient boosting models, laying the groundwork for further discussion in the next chapter.

Chapter 5: Discussion

The Exploratory Data Analysis (EDA) provided valuable insights into the dataset, shedding light on customer demographics, behavior, and churn patterns. The analysis revealed that while there were no significant differences in the means of key numerical variables between churners and non-churners, further investigation is warranted to explore other potential factors influencing churn. Additionally, the lack of strong correlations between numerical variables suggests that they may independently contribute to churn prediction.

Further analysis, including feature engineering and machine learning modelling, was conducted to develop predictive models for churn prediction and identify key drivers of customer churn. The insights gained from this EDA served as a foundation for subsequent analyses and model development, ultimately aiding in customer retention strategies and business decision-making.

5.1 Models Performance

The logistic regression model's moderate performance suggests that while it captures some patterns related to churn, there is room for improvement, particularly in reducing false positives.

The gradient boosting model outperformed the logistic regression model, especially in terms of recall and F1-score, suggesting it is more effective at capturing the complexities of customer churn behavior.

The logistic regression model's accuracy was 0.485, with a precision of 0.51 and recall of 0.69 for predicting churn. In contrast, the gradient boosting model achieved an accuracy of 0.55, with a precision of 0.54 and recall of 0.92. These metrics suggest that the gradient boosting model is more effective in identifying actual churners, albeit with a higher rate of false positives (Type I errors).

The higher recall of the gradient boosting model indicates its superiority in minimizing false negatives (Type II errors), which is critical in churn prediction scenarios where failing to identify a potential churner can lead to revenue loss. The grid search for parameter tuning further enhanced the performance of the gradient boosting model, demonstrating the importance of hyperparameter optimization in machine learning.

5.2 SHAP, Interpretability & Explainability

The interpretability and explainability of machine learning models are crucial aspects, especially in domains like customer churn prediction where actionable insights are essential for decision-making. The SHAP (SHapley Additive exPlanations) values provide valuable insights into how each feature contributes to the model's predictions, enabling stakeholders to understand the model's inner workings and make informed decisions.

Interpretability is paramount for stakeholders to trust and act upon the model's predictions. The logistic regression model, despite its relatively lower performance compared to gradient boosting, offers interpretable coefficients that indicate the direction and magnitude of each feature's impact on the predicted outcome.

For instance, the logistic regression model revealed that the number of support contacts, average transaction amount, and recency of the last purchase are significant predictors of churn. Stakeholders can use these insights to tailor retention strategies and improve customer satisfaction.

On the other hand, the gradient boosting model, while more complex, provides superior predictive performance. The SHAP values for gradient boosting highlight features such as average transaction amount, annual income, and returns ratio as key drivers of churn prediction. Despite its complexity, the model’s interpretability through SHAP values enables stakeholders to prioritize actionable insights effectively.

The SHAP values for the logistic regression model indicate that the number of support contacts has the highest mean absolute SHAP value, followed by the average transaction amount. Features such as the recency of the last purchase, young age category, and promotion response also exhibit notable influence on churn prediction.

For the gradient boosting model, the average transaction amount emerges as the most impactful feature with a mean absolute SHAP value of 0.0315, followed by annual income and returns ratio. While these features play a crucial role in the model’s predictions, other factors such as satisfaction score, number of support contacts, and gender also contribute to churn prediction, albeit to a lesser extent.

These findings align with existing literature, which often highlights financial and transactional metrics as key indicators of customer churn in online retail environments (Neslin et al., 2006).

Studies have shown that higher transaction amounts and income levels typically correlate with increased customer engagement and loyalty, thereby reducing churn likelihood (Hadden et al., 2007).

Similarly, returns ratio, which reflects customer dissatisfaction or product issues, has been consistently identified as a predictor of churn across various sectors (Buckinx & Van den Poel, 2005).

The inclusion of satisfaction score and support contacts as important features also resonates with previous research emphasizing the role of customer satisfaction and service interactions in retention strategies (Homburg & Giering, 2001). These insights reinforce the multifaceted nature of churn behavior and underscore the need for comprehensive models that incorporate a wide range of customer attributes to accurately predict churn and inform effective retention interventions.

Chapter 6: Conclusion & Future Work

6.1 Conclusion

In conclusion, both logistic regression and gradient boosting models offer valuable insights into customer churn prediction. However, the gradient boosting model demonstrated superior performance over logistic regression, particularly in terms of recall and F1-score. The gradient boosting model's higher recall (0.92) compared to the logistic regression model's recall (0.69) indicates its effectiveness in identifying actual churners, which is critical for minimizing revenue loss due to customer attrition.

The logistic regression model, while less complex and more interpretable, provides a straightforward understanding of feature importance through its coefficients. The number of support contacts, average transaction amount, and recency of the last purchase emerged as the most influential predictors of churn. These findings suggest that frequent customer interactions with support, high transaction values, and recent purchases are indicative of lower churn risk.

In contrast, the gradient boosting model, although more complex, leverages a variety of features to achieve better predictive performance. The SHAP values identified average transaction amount, annual income, and returns ratio as the top predictors. This model's ability to capture intricate relationships and interactions between features makes it more robust for churn prediction, albeit at the cost of interpretability.

The performance metrics achieved by both models are consistent with those reported in similar studies in the literature. This suggests that the predictive capabilities of the models are in line with expectations for customer churn prediction in online retail settings. Furthermore, the superior performance of the gradient boosting model underscores its potential to outperform traditional logistic regression models in similar contexts.

The analysis confirms that past behavior is Indeed predictive of customer churn behavior. Key factors such as transaction amounts, recency of purchases, and customer interactions with support services significantly influence churn predictions. The findings underscore the importance of monitoring these behavioural indicators to proactively identify at-risk customers and implement retention strategies.

The SHAP values enhance the interpretability of both models, enabling stakeholders to understand the drivers of churn prediction and devise effective retention strategies.

6.2 Future Work

Moving forward, there are several promising avenues for further exploration in the context of customer churn prediction within the online retail sector. One crucial area for future research involves refining and enhancing existing predictive models to improve their accuracy and robustness. This could entail experimenting with various machine learning techniques, including ensemble methods (Breiman, 2001), neural networks (Hinton et al., 2012), or advanced boosting algorithms (Chen & Guestrin, 2016), to uncover new insights and achieve superior predictive performance.

Moreover, there is a compelling opportunity to delve deeper into feature engineering, seeking to identify novel customer attributes or alternative representations of existing features that may offer richer insights into churn behavior. Prior research has shown that enhancing feature representation can significantly improve model performance (Kuhn & Johnson, 2013, "Feature Engineering and Selection: A Practical Approach for Predictive Models"). Exploring additional data sources such as browsing history, social media interactions, or sentiment analysis of customer reviews (Pang & Lee, 2008) could provide a more comprehensive understanding of the underlying factors driving churn.

Another promising direction for future research is the incorporation of temporal dynamics into churn prediction models. By capturing the evolving patterns of customer behavior over time, these dynamic models could better adapt to changing circumstances and improve prediction accuracy. Techniques such as time-series analysis (Box et al., 2015), recurrent neural networks (Hochreiter & Schmidhuber, 1997), or survival analysis methods (Kaplan & Meier, 1958) offer potential avenues for modelling the temporal aspect of churn more effectively.

Furthermore, it is essential to assess the real-world impact of churn prediction strategies on key business metrics such as customer lifetime value, revenue retention, and customer acquisition costs. Conducting randomized controlled trials to compare the effectiveness of different retention interventions would provide actionable insights for business decision-makers and help validate the practical applicability of churn prediction models. Previous studies have demonstrated the value of such trials in marketing and customer relationship management (Dahan & Mendelson, 2001).

Enhancing the interpretability of churn prediction models is also a critical area for future exploration. Developing methods for visualizing complex model predictions or generating human-readable explanations could improve stakeholders' understanding and trust in these

models, facilitating their adoption and implementation in real-world business settings. Recent advancements in explainable AI (Guidotti et al., 2018) have shown promise in making complex models more interpretable.

Finally, conducting external validation studies using independent datasets from diverse industry sectors or geographic regions would be instrumental in assessing the generalizability and transferability of churn prediction models beyond the specific context of this study. Such validation is crucial for ensuring that the models developed are robust and applicable in various contexts (Bennett & Lanning, 2007).

By addressing these areas of future research, researchers can advance the state-of-the-art in customer churn prediction and empower businesses to develop more effective retention strategies, ultimately enhancing customer satisfaction and long-term profitability in the online retail landscape.

References

1. Ahmad, A.K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data, 6*(1), 28. https://doi.org/10.1186/s40537-019-0191-6

2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM

3. Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313-327.

4. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

5. Hadiji, F., Adomavicius, G., & Burez, J. (2014). Predicting player churn in the wild. In 2014 IEEE Conference on Computational Intelligence and Games (CIG) (pp. 1-8). IEEE.

6. Hendricks, S., & Mwapwele, S. D. (2024). A systematic literature review on the factors influencing e-commerce adoption in developing countries. Data and Information Management, 8(1), 100045. https://doi.org/10.1016/j.dim.2023.100045

7. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387

8. Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414-1425.

9. John Hadden, Ashutosh Tiwari, Rajkumar Roy, Dymitr Ruta, Computer assisted customer churn management: State-of-the-art and future trends, Computers & Operations Research, Volume 34, Issue 10, 2007, Pages 2902-2917, ISSN 0305-0548, https://doi.org/10.1016/j.cor.2005.11.007. (https://www.sciencedirect.com/science/article/pii/S0305054805003503)

10. Kuhn, M., & Johnson, K. (2013). Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press.

11. Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions.

12. Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., ... & Lee, S. (2019). Explainable AI for Trees: From Local Explanations to Global Understanding.

13. Matuszelański, K., & Kopczewska, K. (2022). Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach. Journal of Theoretical and Applied Electronic Commerce Research, 17(1), 165-198. https://doi.org/10.3390/jtaer17010009

14. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

15. Neslin, Scott & Gupta, Sunil & Kamakura, Wagner & Lu, Junxiang & Mason, Charlotte. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research American Marketing Association ISSN. 43. 204-211. 10.1509/jmkr.43.2.204.

16. Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. Journal of Marketing, 67(1), 77-99.

17. Tim Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, Volume 267, 2019, Pages 1-38, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2018.07.007

18. Van den Poel, D., & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research, 157(1), 196-217.

19. Vera L Miguéis et al. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. Expert Systems with Applications, 39(12), 11250-11256.

20. Werner J Reinartz and Vita Kumar. (2003). The impact of customer relationship characteristics on profitable lifetime duration. Journal of marketing 67.1, 77–99.

21. Zhao, Y., Wang, G., Yu, P. S., Liu, S., & Zhang, S. (2013). Inferring social roles and statuses in social networks. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 695-703). ACM.

Appendix

Python Code:

https://d.docs.live.net/f401acfa1b11cf6d/Desktop/Master%20Thesis/Final%20Code%20-%20Dalton%20Fumo.html

---

__Table 3:__ Descriptive statistics of the numerical variables in the dataset.

Variable	Mean	Standard Deviation	Minimum	Maximum
Customer_ID			1	1000
Age	43.267	15.242	18	69
Annual_Income	111.963	52.844	20.01	199.73
Total_Spend	5080.793	2862.123	108.94	9999.64
Years_as_Customer	9.727	5.536	1	19
Num_of_Purchases	49.456	28.544	1	99
Average_Transaction_Amount	266.877	145.873	10.46	499.57
Num_of_Returns	4.612	2.897	0	9
Num_of_Support_Contacts	1.934	1.403	0	4
Satisfaction_Score	2.974	1.392	1	5
Last_Purchase_Days_Ago	182.89	104.391	1	364

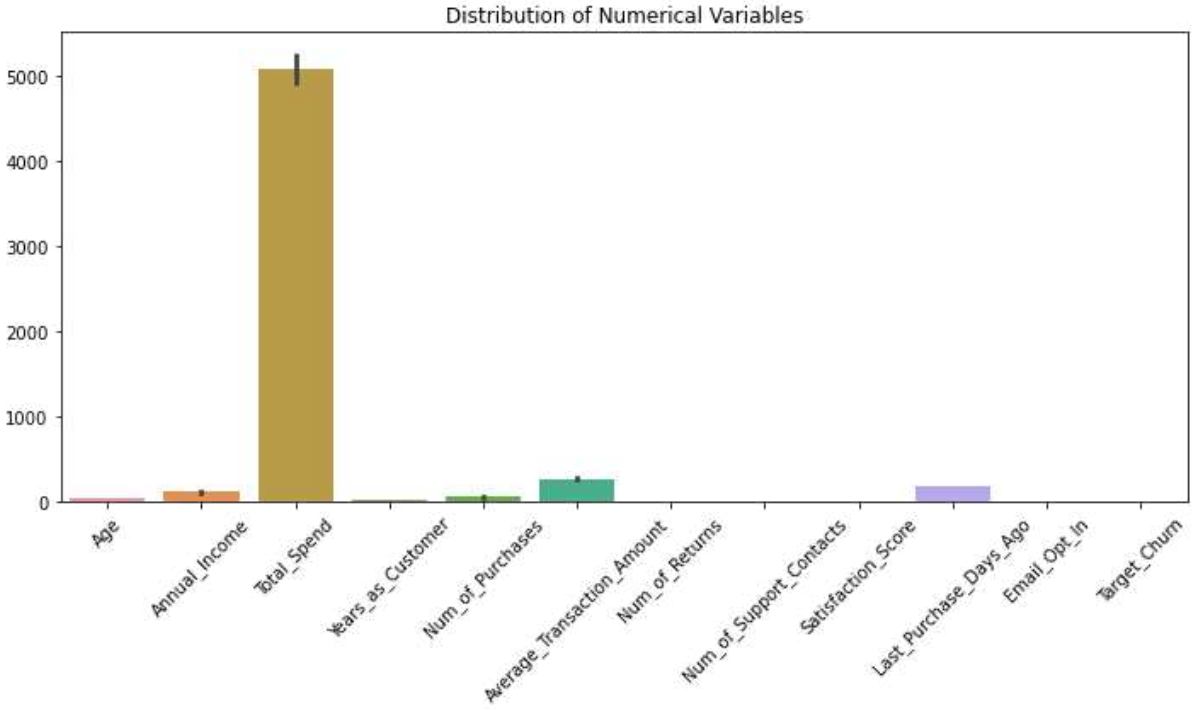

Figure 14: *Distribution of Numerical Variables. Total_Spend is the variable with higher amounts.*

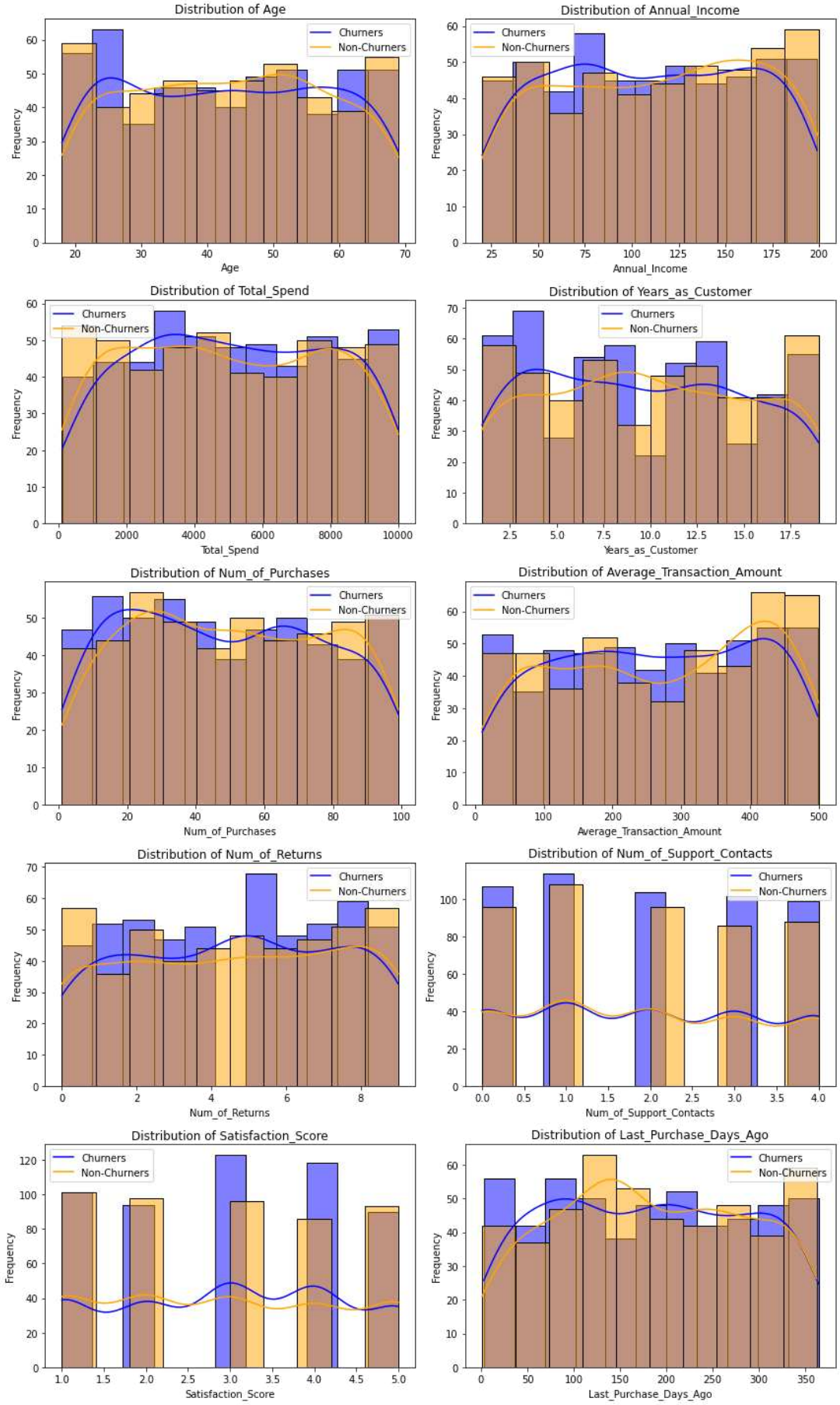

Figure 15: *Comparison between churners and non-churners for each variable.*