NANYANG TECHNOLOGICAL UNIVERSITY

SINGAPORE

BC2407 Analytics II: Adv Pred Tech
2023/24 Semester 2

Project Title: From Enigma to Engagement - Leveraging Predictive Analytics and Machine Learning to Optimise Customer Retention Strategies for Shopee Thailand

Prepared For: Prof Eric

Seminar Group: 4
Team: 6

Group Member	Matriculation No
Andy Chan Yan Meng	U2221216B
Dhiya'Diyana Binte Irwan	U2110746L
Gregory Goh Zong Jun	U2210004K
Lee Fang Hui, Jesslyn	U2211070B
Tan Yu Xiang (Javier)	U2210390H

Table of Contents

Executive Summary

In the dynamic landscape of Southeast Asia's e-commerce market, Shopee Thailand faces a critical challenge in retaining its user base amidst the region's rapid growth. Despite high internet penetration and smartphone adoption, Shopee user retention in Thailand trails behind regional counterparts, with a stagnant rate of approximately 37%. This poses significant revenue implications, compounded by Shopee Thailand’s current broad, one-size-fits-all approach to customer incentives.

To address this issue, we propose a solution leveraging predictive analytics and machine learning through a comprehensive three-step framework aimed at improving customer retention and revenue for Shopee Thailand. This framework involves the integration of advanced analytical models tailored to Shopee Thailand’s needs: (1) Gather real-time customer data to push into the following step, (2) Random Forest using customer Transactional and Session data to assess the likelihood of churn, (3) K-Prototypes Clustering to identify unique customer segment to offer tailored marketing and sales strategies. (4) Continuous monitoring of tailored strategies implemented in Step 3. Each step in the framework aims to assist Shopee Thailand in their efficiency and effectiveness of retaining its customers.

In the development of this framework, we combine two datasets comprising approximately 5000 entries, encompassing a wide range of attributes including user behaviour variables and customer demographics. These diverse dimensions collectively inform an important aspect: the churn status of the customer, providing invaluable insights for strategic decision-making and customer retention efforts.

While acknowledging certain limitations of the framework, such as scalability and data quality issues, it nevertheless serves as a guide for Shopee Thailand to leverage its existing systems and real-time customer data effectively. By implementing this framework, Shopee Thailand can maximise customer lifetime value, enhance revenue, and sustain a competitive edge in the unique Thailand e-commerce landscape.

Through continuous iteration and refinement, Shopee Thailand can adapt to evolving market dynamics and consumer preferences, further strengthening its position as a market leader in the Southeast Asian e-commerce space.

1. Business Understanding

1.1 Introduction

Today, Southeast Asia’s e-commerce market presents a lucrative opportunity for online retailers, driven by a population exceeding 600 million and a combined GDP of 3 trillion USD (Jaouadi & Chuidian, 2023). Among the prominent players in this vibrant market, Shopee stands out as a dominant e-commerce platform, leveraging the region's robust growth to capture a substantial market share. However, within this thriving ecosystem, a puzzling phenomenon emerges in Thailand.

1.2 Business Problem

Amidst Thailand's flourishing e-commerce scene characterised by high internet penetration, widespread smartphone adoption, and increasing online consumer confidence, Shopee Thailand confronts significant hurdles in user retention compared to its regional counterparts. While neighbouring countries boast retention rates exceeding 45%, Thailand's rate remains stagnant at approximately 37% even after a two-year period (Ai, 2023). This discrepancy translates to a potential annual loss of over 8% of Shopee's Thai user base, signifying substantial revenue implications.

One plausible explanation for this stagnation could be the emergence of local e-commerce platforms and department stores offering niche products tailored specifically to the Thai market. This diversification dilutes Shopee Thailand's market share and undermines its ability to secure returning customers. Notably, the marked difference in retention rates between Thailand and more developed e-commerce markets like Singapore underscores the urgency for a comprehensive analysis to identify underlying issues and devise effective growth strategies tailored to Thailand's unique dynamics.

Research indicates that a mere 5% increase in customer retention can yield an impressive 95% boost in profits (Saleh & Saleh, 2015), underscoring the critical need to address the churn issue in Thailand's e-commerce landscape. Moreover, the financial implications are compounded by the significantly higher cost – estimated to be 5-7 times more – of acquiring new customers compared to retaining existing ones. Shopee's struggle with elevated churn rates in Thailand not only escalates marketing expenditure but also poses a direct threat to overall profitability.

In the competitive e-commerce sector, the importance of retaining customers cannot be overstated. Shopee Thailand’s current retention challenges manifest in decreased sales, reduced customer lifetime value, and a consequent negative impact on revenue and market competitiveness.

The case of Shopee facing lower customer retention rates in Thailand, despite the industry's overall growth, underscores a unique set of challenges in the e-commerce landscape. Unlike traditional management approaches that might prioritise high-value customers, in the digital realm of e-commerce, every customer holds value. Thus, addressing the loss of even low-value customers becomes a critical imperative amidst the rapidly evolving consumer landscape, technological advancements, and competitive dynamics specific to the Thai market.

1.3 Current Strategies & Opportunities

Shopee's current approach to customer retention adopts a broad, one-size-fits-all strategy, extending free shipping, discounts, and other benefits universally to all users on the platform. While this strategy aims to foster general customer satisfaction and engagement, it inadvertently leads to a significant loss of revenue. This arises because the benefits are also extended to customers who are not at risk of churning and would likely have made purchases without these incentives. Consequently, this indiscriminate allocation of benefits dilutes the impact of Shopee's marketing efforts and financial resources, offering rewards to those who do not require them to maintain their loyalty (Dang, 2021). This lack of targeted incentives not only represents an inefficient use of resources but also misses the opportunity to effectively address and retain the segment of the customer base that is truly at risk of disengagement and churn (Lupu, 2024).

1.4 Opportunity Statement

As such, our team has identified the following opportunity statement: How can we leverage predictive analytics and machine learning to optimise customer retention strategies for Shopee Thailand to **improve customer retention and revenue**.

2. Proposed Solution

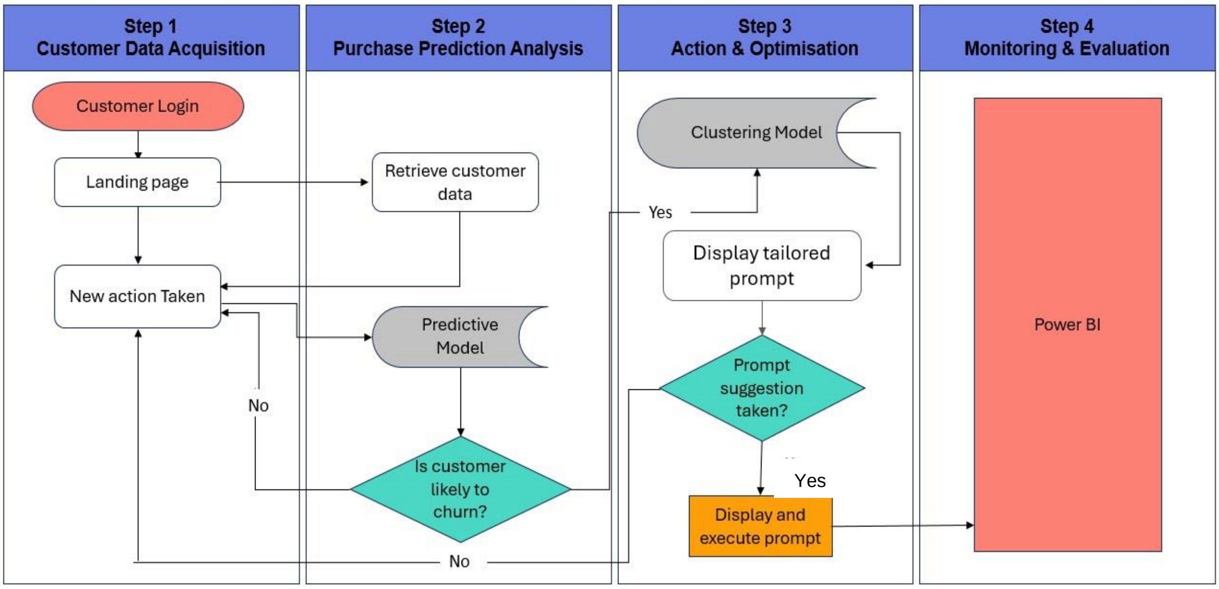


Figure 1: Maximising Customer Retention Framework

Our strategy centres on addressing churn within Shopee Thailand's customer base through a comprehensive three-step framework, emphasising proactive, real-time action to retain valuable customers. We define churned customers as those who previously placed orders but won't do so within the next 3 months, aligning with typical early-stage churn (Hill, 2020). This definition enables timely proactive measures to prevent attrition, allowing tailored solutions such as personalised offers or enhanced support to retain valuable customers.

(1) Step 1 of the framework involves **collecting data on customers** to Shopee's Thailand ecommerce platform. This data includes their product views, site duration, etc. **(2)** In Step 2, we utilise the collected data from Step 1 to **predict which customers are likely to churn** using a predictive algorithm. **(3)** Next, Step 3 focuses on taking action based on the predictions from the purchase prediction analysis. This entails leveraging a clustering model by **identifying the customer demographics and delivering targeted sales and marketing prompts** to customers who are likely to churn and improving their chances of making a purchase. (4) Finally, post-implementation of our clustering strategy, where stakeholders at Shopee Thailand can utilise Power BI to **continuously monitor and evaluate the effectiveness of tailored solutions**.

Overall, this framework aims to boost purchase rates and strengthen customer relationships in Thailand in real-time, through personalised engagement strategies driven by data insights. By prioritising such approaches, Shopee Thailand can reinforce long-term customer loyalty, retention, and improve revenue generation.

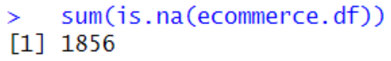
3. Data Cleaning and Preprocessing

3.1 Dataset Description

Our team selected 2 datasets to aid in our analysis. The *Ecommerce Customer Churn* Dataset features various customer attributes such as demographics, order details, and customer behaviour metrics. The second dataset, *Online Shoppers Purchasing Intention* Dataset, contains detailed attributes derived from user session data and is designed to study the purchasing intention of customers.

The two datasets comprise 20 variables and 17 variables respectively, as detailed in Appendix A-1 and A-2.

3.2 Handling Missing Values through rfImpute



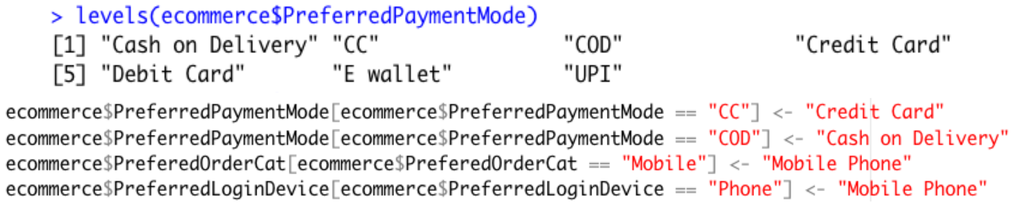
```
>   sum(is.na(ecommerce.df))
[1] 1856
```

Figure 2: Missing Values in Ecommerce Customer Churn Dataset

Notably in the Ecommerce Customer Churn Dataset, there are approximately 1800 missing values. However, it's important to note that such data gaps are inherent in customer demographic fields due to its reliance on user submission (Gomes & Meisen, 2023). Nevertheless, with appropriate data handling and imputation techniques, the integrity of analyses may still be maintained (Alam, 2023).

Adhering to the principle that these values are missing at random and have many outliers, our team has chosen to employ the `rfImpute` function for imputation, utilising rough fix median imputation and the Random Forest algorithm. This offers a more accurate estimation of missing values, particularly suitable for datasets with outliers or non-linear relationships. Median imputation is preferred due to the skewed distributions and outliers present in the continuous variables, as identified in Section 4.1.2 (Shiksha, 2023). This combined approach of RandomForest and Median rough fix Imputation aims to address missing values effectively while ensuring the quality, integrity and reliability of our modelling and analysis.

3.3 Data Standardisation



```
> levels(ecommerce$PreferredPaymentMode)
[1] "Cash on Delivery" "CC"               "COD"              "Credit Card"
[5] "Debit Card"       "E wallet"         "UPI"
```

```
ecommerce$PreferredPaymentMode[ecommerce$PreferredPaymentMode == "CC"] <- "Credit Card"
ecommerce$PreferredPaymentMode[ecommerce$PreferredPaymentMode == "COD"] <- "Cash on Delivery"
ecommerce$PreferedOrderCat[ecommerce$PreferedOrderCat == "Mobile"] <- "Mobile Phone"
ecommerce$PreferredLoginDevice[ecommerce$PreferredLoginDevice == "Phone"] <- "Mobile Phone"
```

Figure 3: Data Standardisation

Standardised data serves as the cornerstone for effective data modelling. Therefore, we have replaced abbreviations such as "CC" with their full counterparts like "Credit Card," ensuring uniformity across the dataset. This ensures consistency and comparability, thereby bolstering the accuracy and interpretation of our model training outlined below.

3.4 Final Dataset

To enhance our analysis, we combined two datasets by aligning the 'Revenue' column from the Online Shoppers Purchasing Intention Dataset with the 'Churn' column in the Ecommerce Customer Churn Dataset on the basis where 'Revenue' = 0 corresponds to 'Churn' = 1. In this process, we excluded 'CustomerID' for its lack of predictive value and 'Visitor_Type' to concentrate on returning visitors, given their direct relevance to churn analysis.

This integration allows us to utilise both user behaviour and demographic variables simultaneously for predictive modelling and tailored solutions.

4. Data Exploration and Analysis

The final dataset we are using comprises approximately 5000 entries. It encompasses key attributes ranging from user behaviour variables to customer demographics. These diverse dimensions collectively inform an important aspect: the churn status of the customer, providing invaluable insights for strategic decision-making and customer retention efforts.

To aid in our analysis and development of our models, we have categorised the features into different types based on their nature and potential impact on the variable 'Churn'.

Category	Features	Description
Demographic	Gender, CityTier, MaritalStatus, NumberOfAddress, WarehouseToHome	These features represent basic characteristics of customers that might influence their loyalty and satisfaction.
Transactional	PreferredLoginDevice, PreferredPaymentMode, OrderAmountHikeFromlastYear, CouponUsed, CashbackAmount, PreferredOrderCat, SpecialDay	Transactional features relate to the customer's purchasing behaviour and preferences, which can signal their satisfaction and likelihood to continue using the service.
Engagement	Tenure, HourSpendOnApp, NumberOfRegisteredDevice, OrderCount, DaySinceLastOrder, Complain, SatisfactionScore	These features are indicative of how engaged the customers are with the ecommerce platform, which can be critical for understanding churn.
Session	BounceRates, ExitRates, PageValues, Administrative, TrafficType, Administrative_Duration, Informational, Informational_Duration, ProductRelated, Month, Browser, Region, ProductRelated_Duration, OperatingSystems, Weekend	These features offer context about the session that might correlate with customer behaviour and preferences

Table 1: Categorisation of Independent Variables

4.1 Univariate Analysis

4.1.1 Categorical Variables

From Appendix B-2, we can see that the highest activity is recorded in the month of May, followed by December. There also seems to be a prevalence for OperatingSystems and Browser 2 which could indicate technological preferences. Region 1 also has higher activity and TrafficType 2 is how most visitors are getting into the site.

Furthermore, it is generally observed that **Cash on Delivery** reigns supreme as the preferred payment method, followed by digital options like Debit Card and E-wallet. **Mobile phones** are the go-to login device, with most users having 3 devices registered. Interestingly, Mobile Phone and Laptop & Accessory tops the preferred order category, while satisfaction scores appear rather positive, with the majority giving 3-5 scores. The majority of customers did not complain or churn, suggesting a positive overall sentiment.

4.1.2 Continuous Variables

Referring to Appendix B-4, the median time spent on product-related pages is also low compared to the maximum, suggesting that most users only viewed a few pages. There are also notable outliers in most variables which indicates differing preferences amongst users. This observation suggests the presence of distinct customer segments characterised by significantly varied duration patterns.

4.2 Multivariate Analysis

4.2.1 Analysis of Correlation Heatmap

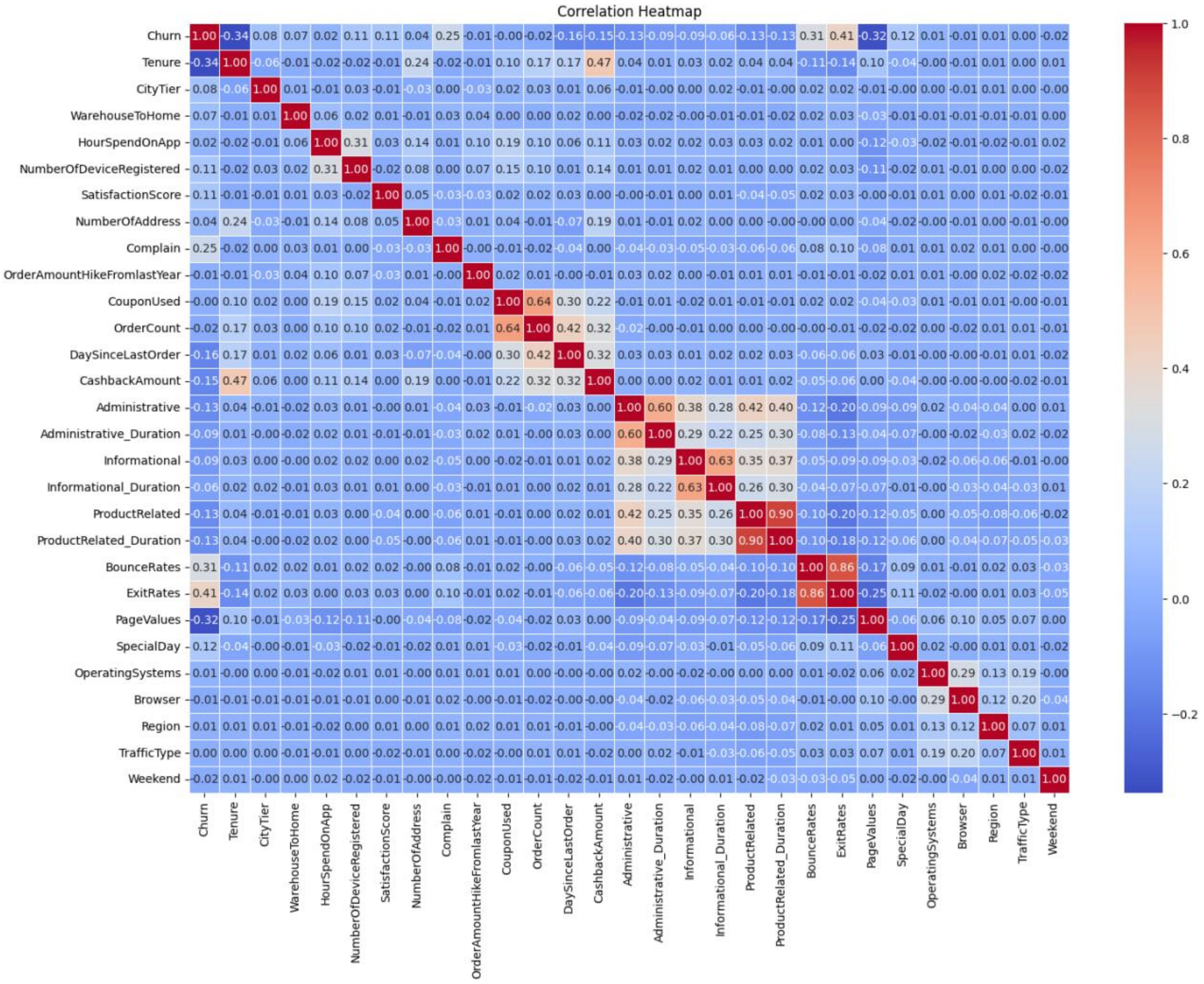


Figure 4: Correlation Heatmap of Variables

Figure 4 reveals a strong positive correlation between 'Administrative', 'Informational', 'ProductRelated' and their associated durations: 'Administrative_Duration', 'Informational_Duration', and 'ProductRelated_Duration'.

The features with the strongest correlation to 'Churn' are 'ExitRates', 'Tenure', 'PageValues' and 'BounceRates'. This highlights the potential of refining business strategies, such as reducing exit and bounce rates, and optimising administrative processes to reduce churn rates.

4.2.2 Analysis of Demographic Variables

With reference to Appendix B-5, when comparing the distribution of 'Gender', males emerge as a demographic with a higher propensity to churn, and 'MaritalStatus' with its high variance, forms distinct customer clusters. Yet variables like 'NumberOfAddress' lack predictive power for churn likelihood because its distribution remains **consistent**, irrespective of their churn status.

In geographical terms, customers residing in CityTier 3 exhibit a higher churn likelihood. Surprisingly, distance of home from the warehouse does not have a visible significant relationship with churn, prompting the exploration of other factors like delivery charges, which could influence purchasing behaviour, especially for remote customers.

4.2.3 Analysis of Transactional Variables

From Appendix B-6, we observe notable behaviour patterns with 'PreferredOrderCat', 'PreferredLoginDevice' and 'PreferredPaymentMode'. Customers favouring mobile phone logins and those opting for cash on delivery are more likely to churn. These preferences might reflect an underlying trend in convenience or security perception, which requires further investigation to tailor customer retention strategies effectively.

The 'CouponUsed' variable displays significant variance and possesses a direct correlation with churn likelihood. Contrastingly, the distribution of 'OrderAmountHikeFromLastYear' does not significantly differ between churned and retained customers.

The bar graphs of 'SpecialDay' reveals an unexpected trend where the lowest churn rates are observed during periods not closely aligned with special days. The boxplot comparing 'CashbackAmount' distributions show a wider interquartile range for non-churned customers, highlighting greater variability in the cashback amounts within this group. This may suggest that an unconventional timing in engagement efforts, including cashback rewards, may influence customer retention strategies.

4.2.4 Analysis of Engagement Variables

As seen in Appendix B-7, we observe a negative correlation between 'Tenure' and churn rate, illustrating that customers with longer associations with the service exhibit a lower propensity to churn. This trend underscores the value of fostering long-term customer relationships.

The median of 'HourSpendOnApp', which stabilises around 3 hours for both churned and retained customer groups, suggests that the duration of app engagement alone does not significantly influence churn risk. This finding challenges our assumption that higher app usage inherently leads to greater customer retention. In contrast, the churn rate is observed to increase with 'NumberOfRegisteredDevice'.

A weak inverse relationship is identified between 'SatisfactionScore' and 'Complain'. However, we noticed that high satisfaction scores also show a correlation with increased churn rates, indicating a complex relationship influenced by interaction complexities and segmentation differences.

On the other hand, a positive relationship can be observed between 'DaySinceLastOrder' and 'OrderCount' (refer to Figure 4), suggesting longer intervals between orders often accompany a higher total order count, possibly reflecting buying trends like bulk purchases or engagement in loyalty programs.

4.2.5 Analysis of Session Variables

Appendix B-8 shows that for sessions where 'PageValues' is 0, the majority of customers churned. In contrast, when 'PageValues' exceeds 0, a significantly larger portion of customers were retained. These could suggest a correlation between higher page values and improved customer retention rates, indicating that 'PageValues' may be an influential predictor of churn.

A time series plot of 'Month' against session count shows a spike in retained customers in May and November (refer to Appendix B-9), suggesting a potential seasonal effect or a successful campaign that might be linked to customer retention.

From Appendix B-10, there is a clear positive correlation between 'BounceRates' and 'ExitRates'. Additionally, it is noted that a higher number of customers less prone to churn predominantly cluster in areas exhibiting lower bounce and exit rates. This pattern suggests that sessions with increased exit and bounce rates are more inclined towards churning.

Among the 'TrafficType' levels, Traffic Type 13 stands out with an exceptionally high churn rate, suggesting it as a significant risk factor and a potential predictor for customer attrition (refer to Appendix B-11). Conversely, Traffic Type 2 demonstrates a notably low churn rate, indicating it may be associated with higher customer retention. These contrasting patterns signify that while some traffic types are prone to higher churn, others may possess protective qualities that could inform retention strategies.

From Appendix B-12, the purchase rates appear to be relatively uniform across 'OperatingSystems', 'Browser', 'Region', and 'Weekend', without any significant variations. This suggests that these variables hold minimal predictive value for identifying potential churn among customers.

5. Data Modelling & Evaluation

5.1 Churn Prediction Model

5.1.1 Methodology

We employ various supervised machine learning algorithms to enhance accuracy and robustness in predicting customer churn. The training process involved 3 models for classification: Random Forest, Classification and Regression Trees (CART), and Logistic Regression. Our target variable, 'Churn', is binary, representing whether a customer has churned (1) or retained (0).

5.1.2 Selection of Accuracy Metrics

Misclassifying churning customers as retained customers can result in missed opportunities, while misclassifying retained customers as churning customers can waste resources. To prevent these errors, we used 4 metrics to measure the accuracy of our model (Sharma, 2022).

Metric	Formula	Description
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Gauges the overall correctness of the model by calculating the proportion of true results (true positives and true negatives)
Precision	$\frac{TP}{TP + FP}$	Assesses the model's exactness by measuring the ratio of true positives to the total number of predicted positives
Recall	$\frac{TP}{TP + FN}$	Evaluates completeness by determining the ratio of true positives to the actual number of positives
F1 Score	$\frac{2 * Recall * Precision}{Recall + Precision}$	Provides a harmonic mean of precision and recall, offering a balance between the two and serving as a single measure of a model’s precision.

*TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

Table 2: Formula and Interpretation behind Accuracy Metrics

Using a combination of metrics to determine the accuracy of our models allows us to avoid being misled by the accuracy paradox, where a model has high accuracy but is poor in predicting the minority class (Afonja, 2017).

5.1.3 Addressing Outliers

In analysing our dataset, we've observed that certain variables, such as 'Informational', and 'Informational_Duration', exhibit notable outliers (refer to Appendix B-4). For instance, the 'Informational' variable is significantly right-skewed, displaying numerous data points that deviate from the central cluster.

Despite their statistical outlier status, we've decided to retain these data points. This decision is rooted in the specific domain of predictive churn analysis, where such outliers may not merely be anomalies, but could in fact signify early warnings of customer churn. In our context, these outliers could represent critical behavioural patterns that are precursors to churn and omitting these outliers could remove valuable predictive signals and reduce the model's ability to identify at-risk customers effectively.

5.1.4 Addressing Class Imbalance via Oversampling

Since the dataset is imbalanced (refer to Appendix B-1), this could result in suboptimal generalisation when it comes to accurately identifying instances of Churn, which are the minority class in the dataset. Thus, the train dataset was resampled using Synthetic Minority Oversampling Technique (SMOTE) to ensure an even proportion of cases in the train dataset (refer to Appendix C-2). This is to prevent the models from developing a bias toward predicting the more prevalent class — Churn = 0.

5.1.5 Feature Selection

As our dataset contains a large number of features, including all of them in our model can lead to overfitting, especially if some have only a marginal effect on the outcome (AWS, 2024). Therefore, we will streamline the feature set via feature selection.

Our predictive model's goal is to effectively predict churn in real time to offer retention strategies. Hence, we focused on selecting features from the **Engagement** and **Session** categories (refer to Section 4) due to their direct relevance to understanding customer behaviour and interaction patterns. Unlike Demographic or Transactional features, which may remain static or change infrequently, Engagement and Session features can provide a timely and sensitive indicator of shifting customer sentiments and potential churn risk, which is the keystone of our real-time predictive algorithm (Makad, 2023).

From Section 4.2.3 and 4.2.4, variables such as 'HourSpendOnApp', 'OperatingSystems', 'Browser', 'Region' and 'Weekend' are deemed less informative and will be excluded. Other variables such as 'NumberOfRegisteredDevice', 'SatisfactionScore' will also be excluded due to their marginal predictive value for a real time context.

Although indicative of transaction frequency, 'OrderCount' lacks the immediacy required for real time prediction. Since 'DaySinceLastOrder' provides a more current reflection of customer engagement, 'OrderCount' will be omitted to streamline the model for real time applicability.

From Section 4.2.5, we identified that 'Administrative', 'Informational', and 'ProductRelated' variables exhibit a high degree of correlation with their respective duration-based counterparts. Given their high correlation, including both would introduce redundancy, potentially diluting the model's predictive power and computational efficiency. However, as duration-based metrics not only signify that an interaction occurred but also how invested a user was during these interactions, they offer a richer view of user engagement. Therefore, we will choose to include 'Administrative_Duration', 'Informational_Duration', and 'ProductRelated_Duration' in our model.

We also employed the 3 models and obtained the top 10 important features (refer to Appendix C-3). Upon analysis, we observed that the features 'Tenure', 'Month', 'ExitRates', and 'PageValues', consistently ranked high across all models. These dynamic features change with customer behaviour over time, providing real-time insight into customer engagement levels, aligning with our decision to keep them in our model.

Final features selected: Tenure, Complain, DaySinceLastOrder, Administrative_Duration, Informational_Duration, ProductRelated_Duration, BounceRates, ExitRates, PageValues, Month, TrafficType

5.1.6 Model Development

We ran the 3 models against the 11 selected features and validated each model's performance using a confusion matrix, where the rows represent the model's predictions and the columns represent the actual instances of churn. From our confusion matrices (refer to Appendix C-4), we were able to obtain the necessary accuracy metrics, as shown in Table 3.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest Model	94.67%	82.27%	86.93%	84.54%
CART Model	92.36%	73.48%	85.16%	78.89%
Logistic Regression Model	89.69%	63.59%	87.28%	73.58%

Table 3: Results of Predictive Models

From our results, the Random Forest Model demonstrates the highest accuracy among the three. With an accuracy of 94.67%, it outperforms both the CART Model and the Logistic Regression Model. Moreover, it exhibits a robust recall rate of 86.93%, indicating its ability to correctly predict churned customers.

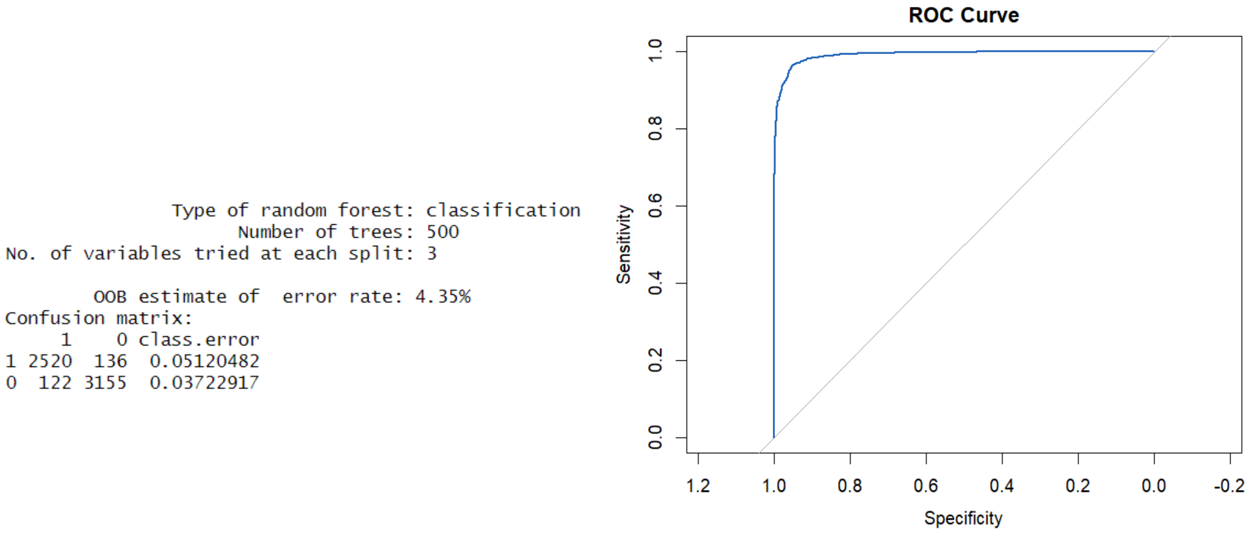
Additionally, due to the Random Forest’s ensemble nature, where multiple decision trees vote on the outcome, the effect of outliers is often diluted (Sellahewa, 2023). In contrast, Logistic Regression is more sensitive to outliers, which may distort the decision boundary and lead to biassed estimates.

The Random Forest Model's capability to maintain performance integrity in the presence of outliers, combined with its superior accuracy metrics, makes it an optimal choice for deployment in predicting customer churn.

5.1.7 Selected Model Evaluation

Using an ensemble of 500 decision trees, the Random Forest model has generalised the balanced train set into the following confusion matrix with a low Out-of-bag (OOB) error at 4.35%. With error rates of 5.12% and 3.72% for predicting churn and retention respectively, this demonstrates the model's reliability in distinguishing between the churn and retention classes.

The Receiver Operating Characteristic (ROC) curve serves as a tool for assessing the effectiveness of a binary classifier, analysing the true positive rate and false positive rate across various threshold levels (Narkhede, 2018). As seen in Figure 5, our model's ROC curve indicates a highly effective classifier. The curve traces a path close to the ideal top-left corner of the graph, indicative of a high true positive rate and a low false positive rate.



```
             Type of random forest: classification
                   Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 4.35%
Confusion matrix:
     1    0 class.error
1 2520  136  0.05120482
0  122 3155  0.03722917
```

Figure 5: Performance Metrics and ROC Curve for Random Forest Model

This performance instils confidence in the model's applicability, making it an asset in the strategic toolkit for mitigating customer churn.

5.2 Clustering Algorithm

5.2.1 Methodology

To segment the at-risk customers into meaningful groups for tailored marketing interventions, we employed the K-prototypes algorithm, an unsupervised machine learning technique. K-prototypes is chosen for its proficiency in handling the mixed data types, characterised by our dataset. By processing both data types within a unified framework, the algorithm effectively groups customers into 'k' clusters where 'k' is pre-determined through an iterative optimisation approach.

5.2.2 Data Preparation

We begin by isolating a subset of data representing customers who have churned, which will be used in the K-Prototypes algorithm. A targeted clustering approach ensures that the resulting customer profiles are distinctly relevant to the context of churn. By homing in on churned customers, this specificity is directly aligned with our strategic goals, ensuring that all subsequent actions are tailored to re-engaging lost customers and nurturing loyalty.

5.2.3 Feature Selection

Our clustering model's goal is to generate targeted sales and marketing prompts. Hence, we focus on selecting variables from the **Demographic** and **Transactional** categories (refer to Section 4). This ensures that the features used for clustering are actionable and can be easily translated into interventions.

As Demographic features are static and do not change frequently, they are reliable for segmenting customers into stable clusters that are important for crafting marketing strategies (Heath, 2023). Transactional features provide direct insights into a customer's purchasing habits and preferences, which is crucial for personalising marketing efforts and promotions (Heath, 2023). Moreover, this category includes direct indicators of customers' value to the company and their response to past marketing stimuli, which are actionable for driving sales.

From Section 4.2.2 and 4.2.3, variables such as 'OrderAmountHikeFromLastYear' and 'NumberOfAddress' are deemed less informative and will be excluded.

Although 'Gender' shows variability in churn propensity, it is not a consistent predictor within clustering models due to diverse influencing factors, and its impact varies across datasets (Kaya et al., 2018). The periodic nature of 'SpecialDay' makes it challenging to directly leverage within our current clustering model for implementing targeted retention strategies. Hence we will also exclude these two variables.

Engagement and Session features, while valuable for understanding overall customer behaviour, do not offer the same level of direct actionability for the purpose of creating sales and marketing prompts. Including these features would add complexity to the model without contributing to the primary goal. However, in a strategic extension of our model, we have decided to include 'OrderCount' and 'DaySinceLastOrder' from the Engagement category. When analysed alongside 'CouponUsed', which signals responsiveness to promotions, these three variables offer a holistic view of customer behaviour capturing the nuances of purchase frequency, the timing of engagements, and promotional influence.

Final features selected: PreferredOrderCat, PreferredLoginDevice, PreferredPaymentMode, CityTier, WarehouseToHome, MaritalStatus, CashbackAmount, CouponUsed, OrderCount, DaySinceLastOrder

5.2.4 Determining the Optimal Number of Clusters

To determine the optimal 'k', we measure the within-cluster sum of squares (WSS), which quantifies the compactness of the clusters. WSS calculates the sum of the squared distances between each customer and the centroid of their respective cluster.

We employed the elbow method to find the most appropriate value for 'k'. This technique involves running the K-prototypes algorithm across a range of 'k' values and plotting the corresponding WSS. As 'k' increases, WSS tends to decrease as the clusters become tighter. We look for the 'elbow' point in the plot where the rate of decrease sharply changes, suggesting that adding more clusters beyond this point does not lead to a significant reduction in WSS.

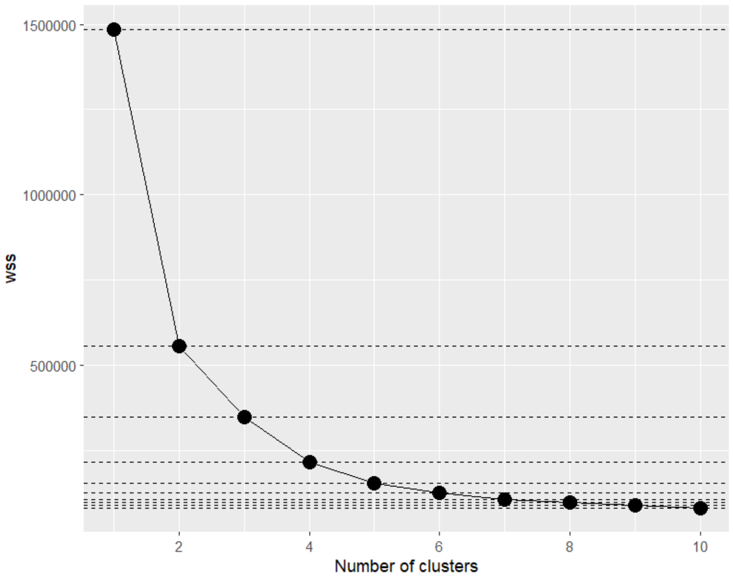
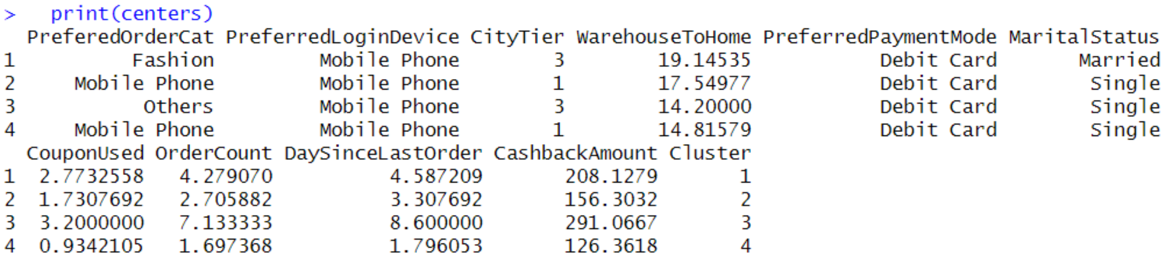


Figure 6: Elbow Method for Optimal K-Prototypes Clusters

Figure 6 illustrates the change in WSS values as the number of clusters increases. The point where the plot bends like an elbow, at 'k' = 4, can be identified from this plot. After 4 clusters, the WSS curve starts to flatten, and the rate of decrease in WSS from increasing the number of clusters is marginal.

5.2.5 Evaluation of Cluster Quality



```
>   print(centers)
  PreferedOrderCat PreferredLoginDevice CityTier WarehouseToHome PreferredPaymentMode MaritalStatus
1          Fashion         Mobile Phone        3        19.14535           Debit Card       Married
2     Mobile Phone         Mobile Phone        1        17.54977           Debit Card        Single
3           Others         Mobile Phone        3        14.20000           Debit Card        Single
4     Mobile Phone         Mobile Phone        1        14.81579           Debit Card        Single
  CouponUsed OrderCount DaySinceLastOrder CashbackAmount Cluster
1  2.7732558   4.279070          4.587209       208.1279       1
2  1.7307692   2.705882          3.307692       156.3032       2
3  3.2000000   7.133333          8.600000       291.0667       3
4  0.9342105   1.697368          1.796053       126.3618       4
```

Figure 7: K-Prototypes Cluster Centers

With the optimal number of clusters established, we proceeded to examine the cluster centroids to evaluate the quality and distinctiveness of each group. The centroids show clear differentiation between clusters, especially in terms of 'PreferredOrderCat', 'CityTier', 'WarehouseToHome', 'DaySinceLastOrder' and 'CashbackAmount'. This suggests that the model has successfully separated customers into homogenous groups based on these key features. The distinct patterns observed in these clusters can then direct us to tailor specific marketing interventions.

6. Deployment of Solution

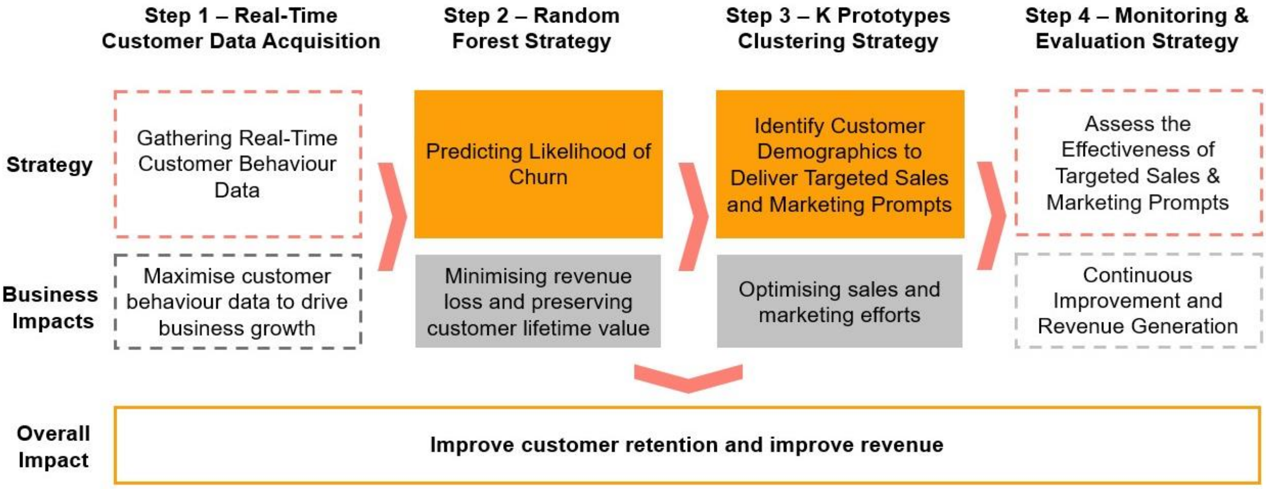

Step 1 – Real-Time Customer Data Acquisition
Step 2 – Random Forest Strategy
Step 3 – K Prototypes Clustering Strategy
Step 4 – Monitoring & Evaluation Strategy
Strategy
Gathering Real-Time Customer Behaviour Data
Predicting Likelihood of Churn
Identify Customer Demographics to Deliver Targeted Sales and Marketing Prompts
Assess the Effectiveness of Targeted Sales & Marketing Prompts
Business Impacts
Maximise customer behaviour data to drive business growth
Minimising revenue loss and preserving customer lifetime value
Optimising sales and marketing efforts
Continuous Improvement and Revenue Generation
Overall Impact
Improve customer retention and improve revenue

Figure 8: Overview of our Maximising Customer Retention Framework

Figure 8 provides a comprehensive overview of how each component of our holistic solution integrates to tackle the challenge of enhancing Shopee Thailand’s customer retention rate in Thailand's ecommerce market. In subsequent sections, we'll delve into the deployment of this framework within Shopee Thailand, aiming to elevate its customer retention rate to match that of other Southeast Asian countries. Our three-pronged solution guarantees an increased likelihood of retaining these valuable Thai customers.

6.1 Monitoring & Gathering Real-Time Customer Behaviour

In the digital landscape of e-commerce, it is imperative that Shopee Thailand leverages on the customer data to monitor and analyse customer behaviour data in real time. As such, the first step is to collect and analyse customer behaviour to gain insights to current consumer sentiments. To initiate this process effectively, Shopee Thailand can leverage its sophisticated, existing real-time customer data acquisition system by harnessing its robust web scraping and data extraction capabilities (Shukla, 2023).

6.2 Predicting Likelihood of Churn

After gathering customer’s real-time data, Shopee Thailand can deploy the Random Forest model to accurately forecast the likelihood of customer churn. By analysing on-site behaviour patterns in-depth, Shopee Thailand can proactively identify potential churn events in real-time.

The model, with a relatively low error rate of 4.35% (refer to Section 5.1.7), implies that it is not only feasible but also highly promising. To facilitate this integration, Shopee Thailand’s engineering team will develop a user-friendly API to interface with the Random Forest model. This API will serve as a bridge, allowing real-time customer data to be fed into the model effortlessly. Additionally, Shopee Thailand can leverage its streamlined data pipeline to ensure smooth data flow, transforming raw customer interactions into actionable insights.

The intention is to integrate this predictive model seamlessly into Shopee Thailand’s existing systems, including the in-house platform and Google Cloud Platform, which offers the reliability and scalability needed for this task, ensuring optimal performance for Shopee Thailand's operations (Netify, 2024). This integration will enable Shopee Thailand to leverage its robust data architecture effectively.

With the model in place, Shopee Thailand will automate its operational workflow to react promptly to potential churn risks. When the model identifies a customer at risk of churning, it will trigger Step 3 for tailored marketing strategies and prompts designed to retain their loyalty. This reactive strategy not only mitigates sales loss but also improves the preservation of customer lifetime value within the company. Furthermore, this strategy helps to address Shopee Thailand’s challenge of inefficient resource allocation and lack of targeted incentives by providing a data-driven solution that enables personalised engagement with at-risk customers, thereby optimising retention efforts and maximising the effectiveness of marketing strategies.

6.3 Identifying Customer Demographics to Deliver Targeted Prompts

Customers who are flagged by our predictive model as having a high risk of churn are then transitioned into the third step of our framework, which involves the deployment of the K-Prototypes model to cluster churning customers.

Established cluster centroids will be predetermined on Shopee Thailand’s vast historical customer datasets using the K-Prototypes model. This serves as a baseline that facilitates a deeper insight into the characteristic features and behaviours of each segment. For every flagged customer, the algorithm will measure the proximity between this newly identified potential churn customer and the established cluster centroids based on the customer’s data and assign them to the nearest centroid. This strategic alignment enables the deployment of cluster-specific retention tactics aimed at encouraging the customer to engage in further transactions with Shopee Thailand.

To further illustrate the framework, we have included several solutions designed to cater to the unique characteristics identified within specific clusters:

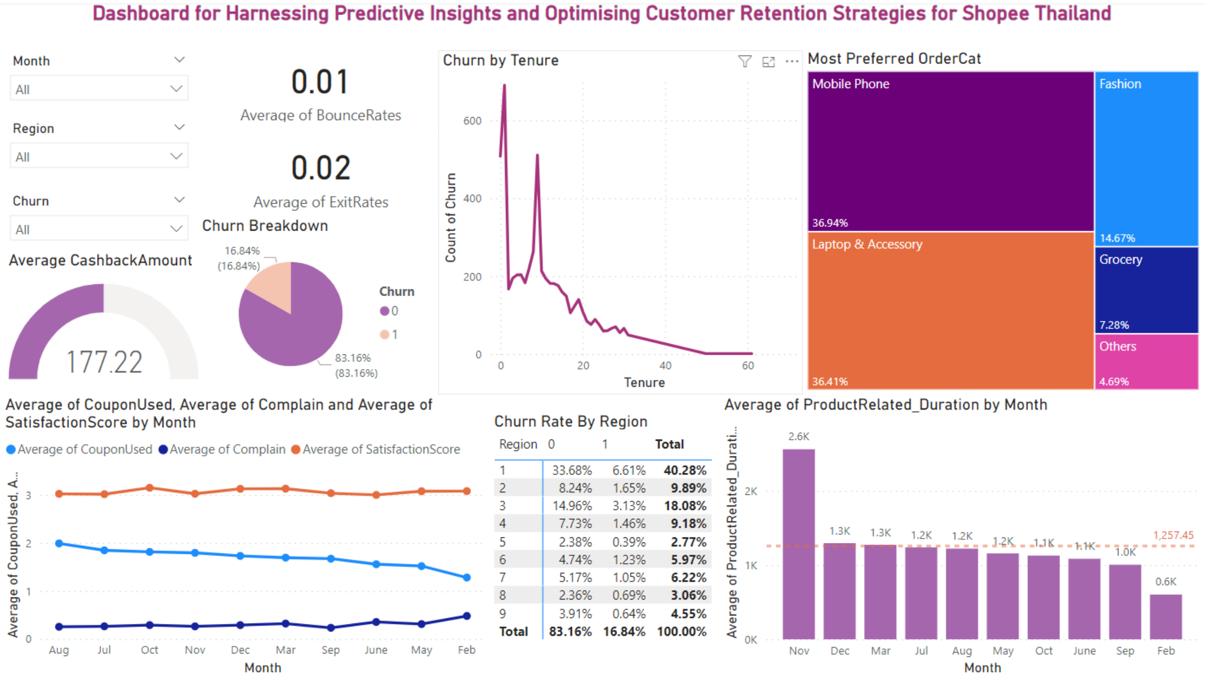
1. For clusters where customers are identified to make bulk purchases more frequently, we can introduce limited-time offers tailored specifically for this purchasing pattern. By providing exclusive discounts or free shipping for orders made within a set timeframe or for a limited quantity of products, this approach can motivate more regular buying patterns, allowing them to enjoy the benefits of bulk discounts without necessarily stockpiling.

2. For clusters which indicate strong preferences in cash as a preferred payment method, Shopee Thailand can consider a Cash Conversion strategy. In Thailand, there is still a preference for cash transactions over digital payments, particularly in smaller businesses, street markets, and local establishments (Suruga, 2023). Hence, promoting the option to convert their cashback vouchers into physical cash at selected partner outlets or through designated redemption points can provide flexibility for customers who prefer cash transactions while still benefiting from the rewards earned through their purchases.

3. Additionally, in Bangkok, the purchasing power is much higher compared to rural areas (Wyatt, 2024), where poverty is more imminent (World Bank, 2022). Making use of the 'CityTier' variable in the clustering model, Shopee Thailand can offer enhanced discounts and vouchers targeted at customers in less affluent regions. This initiative can bridge the spending power gap, making online purchases more accessible and appealing to a wider demographic.

Aligning the strategies with the identified clusters, Shopee Thailand can offer more personalised, effective retention tactics. This not only enhances customer satisfaction but also fosters a more loyal customer base, reducing churn and driving sustainable growth, addressing the lack of targeted incentives mentioned in Section 1.3.

6.4 Monitoring & Evaluation Strategy



Churn Rate By Region

Region	0	1	Total
1	33.68%	6.61%	**40.28%**
2	8.24%	1.65%	**9.89%**
3	14.96%	3.13%	**18.08%**
4	7.73%	1.46%	**9.18%**
5	2.38%	0.39%	**2.77%**
6	4.74%	1.23%	**5.97%**
7	5.17%	1.05%	**6.22%**
8	2.36%	0.69%	**3.06%**
9	3.91%	0.64%	**4.55%**
Total	**83.16%**	**16.84%**	**100.00%**

Figure 9: Dashboard for Stakeholders

To ascertain the impact and effectiveness of our Phase 2 and 3 strategies, Shopee Thailand can implement a robust Monitoring & Evaluation Strategy, focused on assessing the outcomes of our predictive analytics and tailored engagement efforts. This approach is essential for quantifying the influence of these strategies on churn rates, ensuring our actions directly contribute to improved customer retention.

To facilitate this, we propose the implementation of a real-time dashboard (Fig 9), designed to offer stakeholders immediate visibility into crucial customer interaction metrics. This dashboard will track key indicators such as exit and bounce rates, the volume of customer complaints, and preferred orders by customers. By closely monitoring these metrics, stakeholders can gain deep insights into customer satisfaction and overall experience.

For instance, if a decline in satisfaction scores is observed even after the deployment of targeted engagement strategies, this may signal to stakeholders that these strategies are not effective. Potential underlying issues, such as lack of content relevance or lapses in product quality, could be contributing factors to this disconnect, thereby influencing churn likelihood. Armed with this real-time data, stakeholders can identify and address these gaps promptly, ensuring that no customer dissatisfaction goes unattended.

Moreover, the agility afforded by this real-time dashboard enables Shopee Thailand to make swift, data-informed decisions to refine customer engagement strategies, significantly enhancing our capacity to reduce churn. This proactive monitoring and evaluation approach ensures that our targeted efforts are continuously optimised, responding dynamically to customer feedback and behaviours.

7. Evaluation of Proposed Solution

7.1 Benefits

7.1.1 Reducing Churn & Fostering Better Customer Relations

The use of machine learning algorithms for Shopee Thailand's churn prediction model allows continuous analysing of customer behaviour. Through the use of Random Forest, Shopee Thailand will be able to ensure accurate prediction of user's churn risk. This gives them the opportunity to intervene, reduce customer churn risk and take action to retain them before they switch to competitors.

By collecting and analysing customer data points, the framework provides valuable insights into customer behaviour and preference. This empowers Shopee Thailand to deploy K-Prototypes Clustering for customer segmentation. Through identification of distinct customer segments, Shopee Thailand can tailor engagement strategies based on demographics and preferences. This tailored approach enhances the relevance and effectiveness of personalised sales and marketing initiatives, ultimately resulting in improved customer satisfaction and retention. By leveraging personalised promotions based on customer data and preferences, Shopee Thailand can foster stronger customer relationships.

7.1.2 Increase Revenue Generation

Satisfied customers are more likely to increase purchase frequency and overall engagement (Oman, 2023). Loyal customers spend more and are more receptive to upsells and cross-sells, boosting Shopee Thailand's profitability and competitive edge within the e-commerce landscape. This will allow them to decrease customer churn rates and improve revenue growth.

7.2 Limitations

One significant constraint is the reliance on real-time customer data for predicting churn and targeting marketing efforts effectively. The initial step of data collection from Shopee Thailand’s ecommerce platform, encompassing customer interactions such as product views and site duration, appears readily achievable given the nature of the business. However, any slight inaccuracies in these data can lead to misguided predictions (classifying churning customers as retained customers), potentially resulting in customer dissatisfaction due to absent promotions. This limitation is compounded by the fact that our framework only involves one layer of algorithm to predict churn. Thus, any slight inaccuracy (classifying retained customers as churning customers) can cause wasted resources. More layers of prediction analysis may be required to enhance the accuracy and effectiveness of our predictive model.

The subsequent step of taking action based on predictions, leveraging K-Prototypes Clustering to personalise engagement strategies, is promising. However, basic demographic segmentation may fall short in capturing the complexities of customer behaviour. Other factors such as psychographic traits are extremely critical to consider since they provide deeper insights into customer preferences, motivations, and behaviours.

While our proposed framework for customer retention holds promise, it is important to recognise and address the limitations discussed. Ensuring the accuracy of customer data and considering additional layers of prediction analysis are crucial steps to mitigate the risk of misguided predictions and wasted resources. Moreover, incorporating factors beyond basic demographic segmentation, such as psychographic traits, is essential for a comprehensive understanding of customer behaviour. By acknowledging these limitations and refining our proposed solution accordingly, Shopee Thailand can optimise the effectiveness of our retention strategies and better serve its customers' needs.

7.3 Future Considerations

Shopee Thailand will need to adapt to evolving customer behaviour and market dynamics especially in Thailand, with intense competition in the ecommerce landscape. By incorporating new data and adjusting model parameters over time, businesses can enhance the accuracy and effectiveness of churn prediction, further minimising revenue loss and driving sustainable growth. Additionally, Shopee Thailand should consider exploring feature engineering involving creating new features from existing data that might be more predictive of churn. For example, combining purchase frequency and average order value might create a better indicator of customer loyalty than either metric alone.

8. Conclusion

In conclusion, this report highlighted the potential of integrating Random Forest and K-Prototypes clustering algorithms to meaningfully decrease churn rates and bolster customer retention for Shopee Thailand. It also elucidates the strategic advantage of real-time monitoring of customer behaviour, empowering stakeholders with deeper and more actionable customer insights. Our proposed solutions have demonstrated the capacity to not only to personalise the customer experience but also to refine marketing efforts, thereby streamlining costs and increased revenue for Shopee Thailand.

While acknowledging the inherent limitations, the report outlines clear strategies to overcome these barriers, ensuring the efficacy of our proposed solutions. In leveraging sophisticated machine learning models and advanced analytics, we are confident that our approach aligns with the business objective of minimising churn and enhancing customer loyalty. This would revolutionise the Shopee Thailand's business operation and sustain its revenue growth in the dynamic e-commerce sector.

9. References

Afonja, T. (2017, December 8). *Accuracy Paradox. "If you don't know anything about… | by Tejumade Afonja*. Towards Data Science. Retrieved March 31, 2024, from https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b

Alam, S. (2023, January 6). *An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity*. ScienceDirect. Retrieved March 31, 2024, from https://www.sciencedirect.com/science/article/pii/S2772662223001819

AWS. (2024). *What is Overfitting? - Overfitting in Machine Learning Explained - AWS*. Amazon AWS. Retrieved March 31, 2024, from https://aws.amazon.com/what-is/overfitting/

Dang, T. (2021). IMPROVING E-COMMERCE SERVICE FOR BETTER CUSTOMERS' SATISFACTION. *Shopee Viet Nam*, *51*. https://www.theseus.fi/bitstream/handle/10024/493585/Dang_Thao.pdf?isAllowed=y&sequence=3

Gomes, M. A., & Meisen, T. (2023, January 6). *A review on customer segmentation methods for personalized customer targeting in e-commerce use cases*. SpringerLink. Retrieved March 31, 2024, from https://link.springer.com/article/10.1007/s10257-023-00640-4

Heath, C. (2023, April 27). *Demographic Segmentation: Guide, Types & Examples*. Dovetail. Retrieved March 31, 2024, from https://dovetail.com/market-research/demographic-segmentation/

Hill, C. (2020). *How to beat customer churn and improve retention*. Yaagneshwaran Ganesh (Yaag). Retrieved March 29, 2024, from https://www.yaagneshwaran.com/blog/customer-churn/

Lupu, R. (2024, March 17). *The Market for Reviews: Strategic Behavior of Online Product Reviewers with Monetary Incentives*. SpringerLink. Retrieved March 31, 2024, from https://link.springer.com/article/10.1007/s41464-020-00094-y

Makad, S. (2023, September 21). *4 Powerful Strategies for Reducing Customer Churn*. MoEngage. Retrieved March 31, 2024, from https://www.moengage.com/blog/powerful-strategies-tackling-customer-churn/

MartechAsia. (2021). Wikipedia. Retrieved March 30, 2024, from http://martechasia.net/case-study/shopee-and-partners-developed-future-ready-data-mart-to-help-merchants-optimise-sales-performance/

Netify. (2024). *Shopee Taiwan: Providing fun and seamless online shopping experiences with cloud infrastructure*. Wikipedia. Retrieved March 30, 2024, from https://www.netify.ai/resources/ips/34.96.222.199

Oman. (2023). Wikipedia. Retrieved March 31, 2024, from https://www.linkedin.com/pulse/power-customer-satisfaction-building-loyalty-retention/

Sellahewa, K. (2023, May 15). *Random Forest Explained! (Regression and Classification Tasks)*. Wikipedia. Retrieved March 31, 2024, from https://www.linkedin.com/pulse/random-forest-explained-regression-classification-tasks-sellahewa/

Sharma, P. (2022, January 16). *Decoding the Confusion Matrix*. KeyToDataScience. Retrieved March 31, 2024, from https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb

Shiksha. (2023, January 27). *Handling missing data: Mean, Median, Mode*. Shiksha Online. Retrieved March 21, 2024, from https://www.shiksha.com/online-courses/articles/handling-missing-data-mean-median-mode/

Shukla, A. (2023, August 4). *Understanding the Shopee App: Functionality and Insights for Building a Similar E-commerce App*. Next Big Technology. Retrieved March 30, 2024, from https://nextbigtechnology.com/understanding-the-shopee-app-functionality-and-insights-for-building-a-similar-e-commerce-app/

Suruga, T. (2023, April 12). *Thailand, Japan and Vietnam lag in Asia's digital payments rush*. Nikkei Asia. Retrieved March 31, 2024, from https://asia.nikkei.com/Business/Finance/Thailand-Japan-and-Vietnam-lag-in-Asia-s-digital-payments-rush

Verma, A. (2021). *Ecommerce Customer Churn Analysis and Prediction*. Kaggle. Retrieved March 21, 2024, from https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/data

World Bank. (2022, October 21). *Rural Thailand Faces the Largest Poverty Challenges with High Income Inequality*. World Bank. Retrieved March 31, 2024, from https://www.worldbank.org/en/news/press-release/2022/10/21/rural-thailand-faces-the-largest-poverty-challenges-with-high-income-inequality

Wyatt, C. (2024). *The Average Cost of Living in Thailand in 2024*. Silk Estate. Retrieved March 31, 2024, from https://silkestate.io/cost-of-living-in-thailand/

10. Appendix

Appendix A: Dataset Description

Appendix A-1: Data Dictionary (Dataset 1)

Feature	Description	Data Type	Range/Values
CustomerID	Unique customer identifier	Continuous	50001 to 55630
Churn	Flag indicating if the customer churned or retained	Categorical	0 (Retained) 1 (Churned)
Tenure	Duration of customer relationship with ecommerce company	Continuous	0 to 61
PreferredLoginDevice	Preferred login device of customer	Categorical	Computer, Mobile Phone
CityTier	Level of development of the customer's city	Categorical	1 (Highly Developed) 2 (Developed) 3 (Developing)
WarehouseToHome	Distance in between warehouse to home of customer	Continuous	5 to 127 km
PreferredPaymentMode	Preferred payment method of customer	Categorical	Cash on Delivery Credit Card Debit Card E-wallet UPI (Refers to online payment applications eg. PromptPay)
Gender	Gender of customer	Categorical	Male Female
HourSpendOnApp	Number of hours spent on the mobile application or website	Continuous	0 to 5 hours
PreferredOrderCat	Customer's most ordered category in the last month	Categorical	Fashion Grocery Laptop and Accessory Mobile Phone Others
SatisfactionScore	Customer’s satisfaction score with the service	Categorical	1 (Very Dissatisfied) to 5 (Very Satisfied)
MaritalStatus	Customer's marital status	Categorical	Divorced MarriedSingle

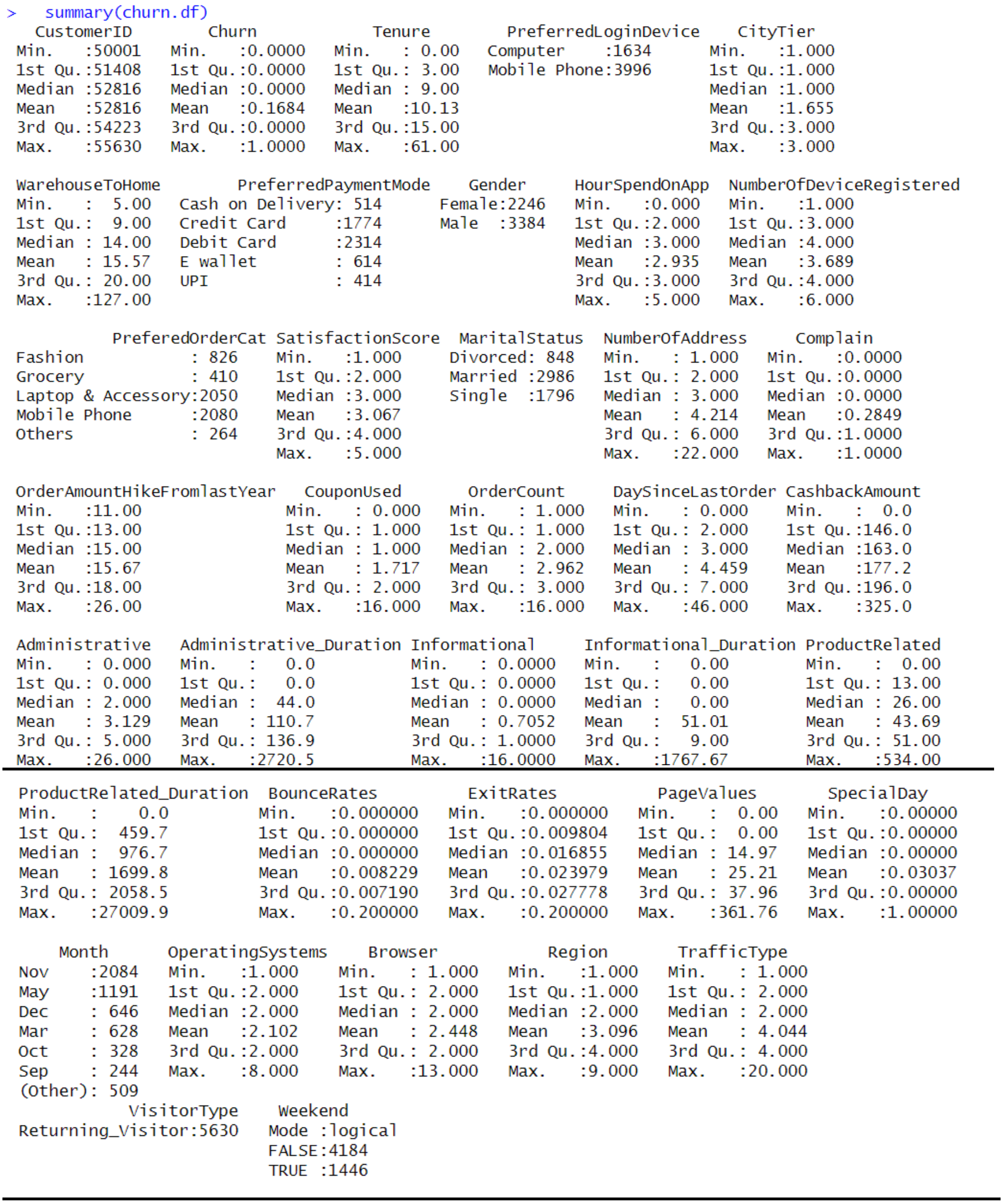
Feature	Description	Data Type	Range/Values
NumberOfAddress	Total number of addresses linked to the customer	Continuous	1 to 22
NumberOfRegisteredDevice	Total number of devices registered	Continuous	1 to 6
Complain	Indicates if the customer complained last month	Categorical	0 (No Complain) 1 (Complained)
OrderAmountHikeFromlastYear	Percentage increase in order value compared to last year	Continuous	11% to 26%
CouponUsed	Number of coupons used by the customer last month	Continuous	0 to 16
OrderCount	Total number of orders placed by the customer last month	Continuous	1 to 16
DaySinceLastOrder	Days since the customer's last order	Continuous	0 to 46
CashbackAmount	Average cashback received by the customer last month ($)	Continuous	0 to 325

Appendix A-2: Data Dictionary (Dataset 2)

Feature	Description	Data Type	Range/Values
Administrative	Number of pages of this type (administrative) visited by the user in that session.	Continuous	0-27 No. of Pages
Administrative_ Duration	Total amount of time (in seconds) spent by the user on administrative pages during the session.	Continuous	0-4000s
Informational	Number of informational pages visited by the user in that session.	Continuous	0-24 No. of Pages
Informational_ Duration	Total time spent by the user on informational pages.	Continuous	0-2600s
ProductRelated	Number of product-related pages visited by the user.	Continuous	0-705 No. of Pages
ProductRelated_ Duration	Total time spent by the user on product-related pages.	Continuous	0-64,000s
BounceRates	Average bounce rate of the pages visited by the user. The bounce rate is the percentage of visitors who navigate away from the site after viewing only one page.	Continuous	0-0.2
ExitRates	Average exit rate of the pages visited by the user. The exit rate is a metric that shows the percentage of exits from a page.	Continuous	0-0.2
PageValues	Average value of the pages visited by the user. This metric is often used as an indicator of how valuable a page is in terms of generating revenue.	Continuous	0-362
SpecialDay	Closeness of the site visiting time to a specific special day	Categorical	0 to 1, Intervals of 0.2

Feature	Description	Data Type	Range/Values
	(e.g., Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalised with a transaction.		
Month	Month of the year in which the session occurred.	Categorical	Jan-Dec
OperatingSystems	Operating system used by the user.	Categorical	1 to 8
Browser	Browser used by the user.	Categorical	1 to 13
Region	Region from which the user is accessing the website.	Categorical	1 to 9
TrafficType	Type of traffic (e.g., direct, paid search, organic search, referral).	Categorical	1 to 20
VisitorType	Type of visitor	Categorical	New_Visitor Returning_Visitor Other
Weekend	Whether the session occurred on a weekend.	Categorical	0 (True) 1 (False)

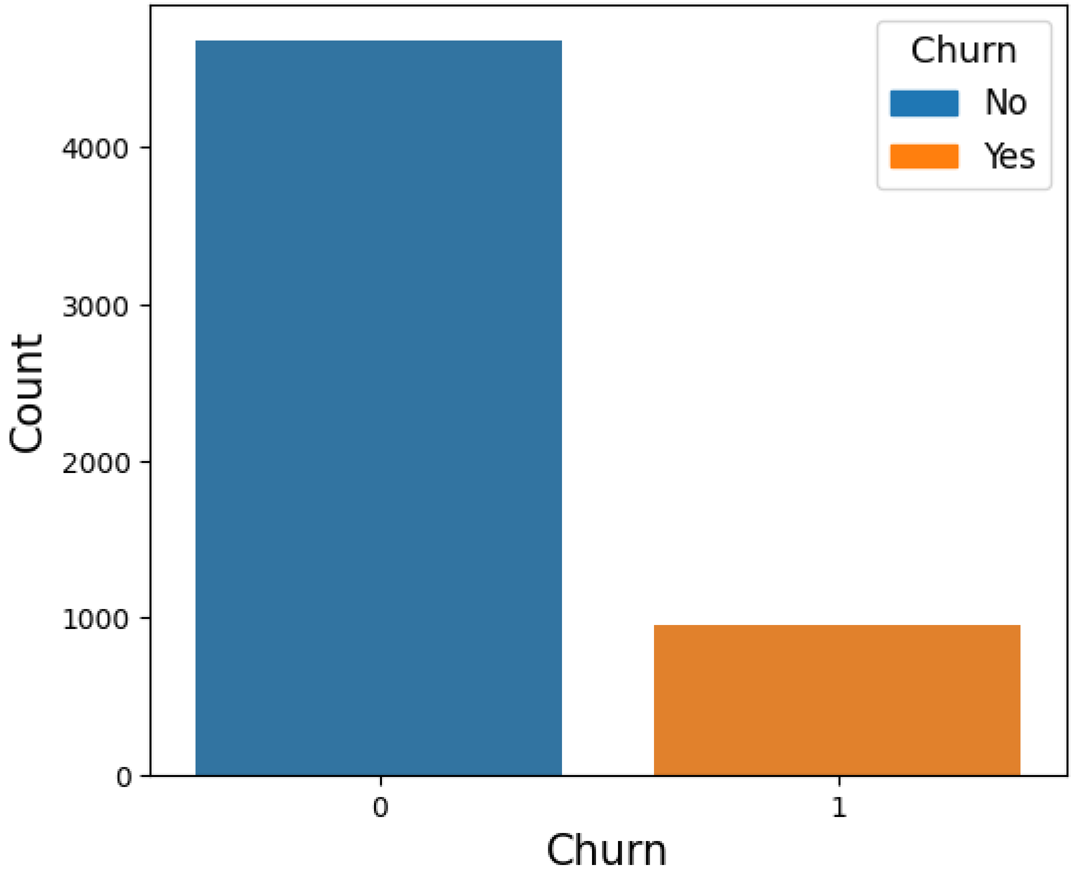
Appendix A-3: Summary of Final Dataset



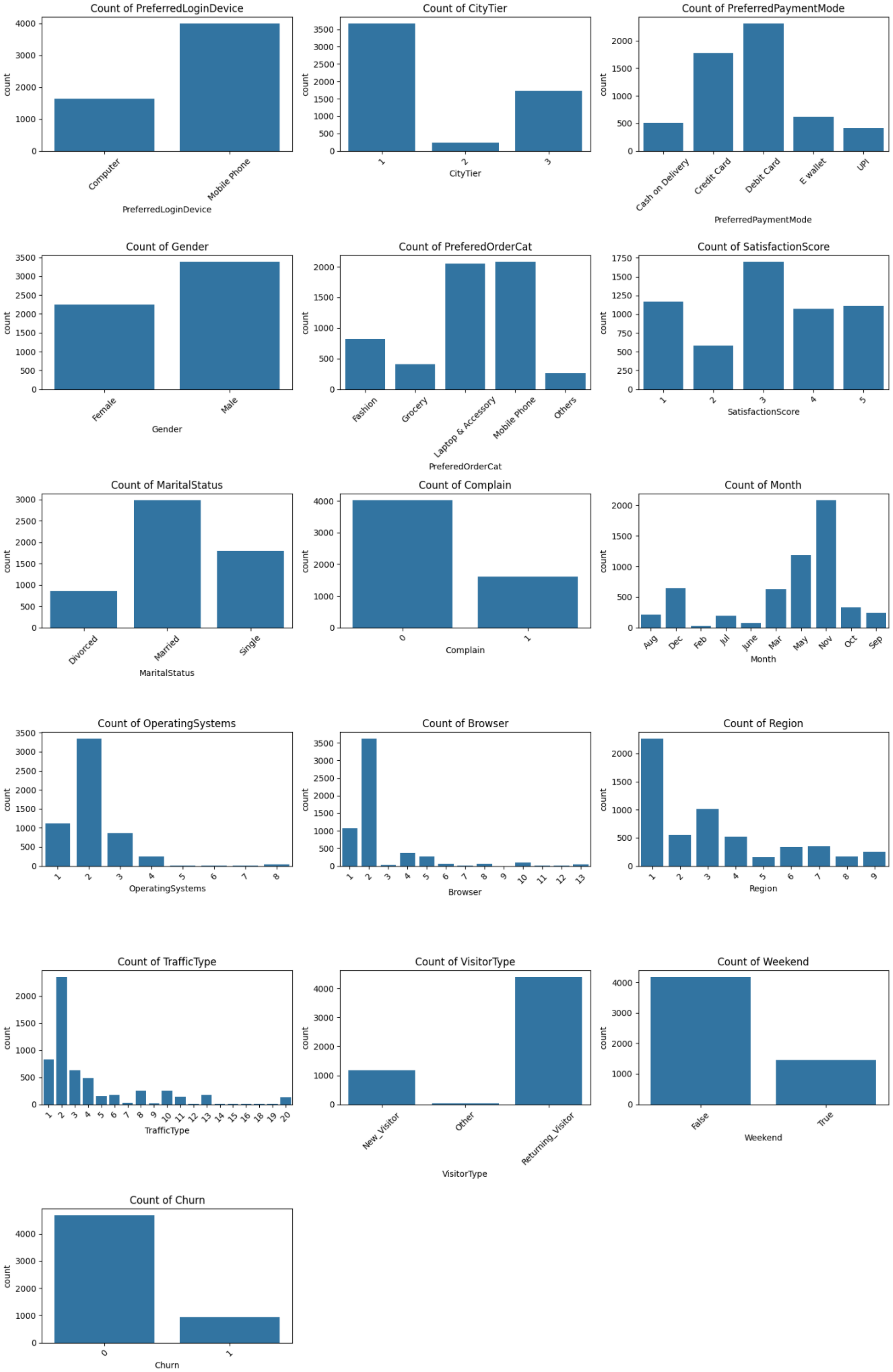
```
>   summary(churn.df)
   CustomerID         Churn              Tenure          PreferredLoginDevice     CityTier    
 Min.    :50001   Min.    :0.0000   Min.    : 0.00   Computer     :1634        Min.    :1.000  
 1st Qu.:51408   1st Qu.:0.0000   1st Qu.: 3.00   Mobile Phone:3996        1st Qu.:1.000  
 Median :52816   Median :0.0000   Median : 9.00                              Median :1.000  
 Mean    :52816   Mean    :0.1684   Mean    :10.13                              Mean    :1.655  
 3rd Qu.:54223   3rd Qu.:0.0000   3rd Qu.:15.00                              3rd Qu.:3.000  
 Max.    :55630   Max.    :1.0000   Max.    :61.00                              Max.    :3.000  

 WarehouseToHome        PreferredPaymentMode     Gender        HourSpendOnApp   NumberOfDeviceRegistered
 Min.    :  5.00   Cash on Delivery: 514       Female:2246   Min.    :0.000   Min.    :1.000          
 1st Qu.:  9.00   Credit Card      :1774       Male   :3384   1st Qu.:2.000   1st Qu.:3.000          
 Median : 14.00   Debit Card       :2314                      Median :3.000   Median :4.000          
 Mean    : 15.57   E wallet         : 614                      Mean    :2.935   Mean    :3.689          
 3rd Qu.: 20.00   UPI              : 414                      3rd Qu.:3.000   3rd Qu.:4.000          
 Max.    :127.00                                               Max.    :5.000   Max.    :6.000          

            PreferedOrderCat SatisfactionScore  MaritalStatus  NumberOfAddress     Complain     
 Fashion             : 826    Min.    :1.000     Divorced: 848   Min.    : 1.000   Min.    :0.0000  
 Grocery             : 410    1st Qu.:2.000     Married :2986   1st Qu.: 2.000   1st Qu.:0.0000  
 Laptop & Accessory:2050    Median :3.000     Single  :1796   Median : 3.000   Median :0.0000  
 Mobile Phone        :2080    Mean    :3.067                      Mean    : 4.214   Mean    :0.2849  
 Others              : 264    3rd Qu.:4.000                      3rd Qu.: 6.000   3rd Qu.:1.0000  
                              Max.    :5.000                      Max.    :22.000   Max.    :1.0000  

 OrderAmountHikeFromlastYear   CouponUsed        OrderCount      DaySinceLastOrder  CashbackAmount 
 Min.    :11.00              Min.    : 0.000   Min.    : 1.000   Min.    : 0.000    Min.    :  0.0  
 1st Qu.:13.00              1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 2.000    1st Qu.:146.0  
 Median :15.00              Median : 1.000   Median : 2.000   Median : 3.000    Median :163.0  
 Mean    :15.67              Mean    : 1.717   Mean    : 2.962   Mean    : 4.459    Mean    :177.2  
 3rd Qu.:18.00              3rd Qu.: 2.000   3rd Qu.: 3.000   3rd Qu.: 7.000    3rd Qu.:196.0  
 Max.    :26.00              Max.    :16.000   Max.    :16.000   Max.    :46.000    Max.    :325.0  

 Administrative    Administrative_Duration Informational     Informational_Duration ProductRelated  
 Min.    : 0.000   Min.    :    0.0        Min.    : 0.0000   Min.    :    0.00       Min.    :  0.00  
 1st Qu.: 0.000   1st Qu.:    0.0        1st Qu.: 0.0000   1st Qu.:    0.00       1st Qu.: 13.00  
 Median : 2.000   Median :   44.0        Median : 0.0000   Median :    0.00       Median : 26.00  
 Mean    : 3.129   Mean    : 110.7        Mean    : 0.7052   Mean    :   51.01       Mean    : 43.69  
 3rd Qu.: 5.000   3rd Qu.: 136.9        3rd Qu.: 1.0000   3rd Qu.:    9.00       3rd Qu.: 51.00  
 Max.    :26.000   Max.    :2720.5        Max.    :16.0000   Max.    :1767.67       Max.    :534.00  

 ProductRelated_Duration  BounceRates          ExitRates           PageValues        SpecialDay     
 Min.    :     0.0       Min.    :0.000000   Min.    :0.000000   Min.    :  0.00   Min.    :0.00000  
 1st Qu.:   459.7       1st Qu.:0.000000   1st Qu.:0.009804   1st Qu.:  0.00   1st Qu.:0.00000  
 Median :   976.7       Median :0.000000   Median :0.016855   Median : 14.97   Median :0.00000  
 Mean    :  1699.8       Mean    :0.008229   Mean    :0.023979   Mean    : 25.21   Mean    :0.03037  
 3rd Qu.:  2058.5       3rd Qu.:0.007190   3rd Qu.:0.027778   3rd Qu.: 37.96   3rd Qu.:0.00000  
 Max.    :27009.9       Max.    :0.200000   Max.    :0.200000   Max.    :361.76   Max.    :1.00000  

     Month       OperatingSystems    Browser           Region        TrafficType    
 Nov     :2084   Min.    :1.000   Min.    : 1.000   Min.    :1.000   Min.    : 1.000  
 May     :1191   1st Qu.:2.000   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 2.000  
 Dec     : 646   Median :2.000   Median : 2.000   Median :2.000   Median : 2.000  
 Mar     : 628   Mean    :2.102   Mean    : 2.448   Mean    :3.096   Mean    : 4.044  
 Oct     : 328   3rd Qu.:2.000   3rd Qu.: 2.000   3rd Qu.:4.000   3rd Qu.: 4.000  
 Sep     : 244   Max.    :8.000   Max.    :13.000   Max.    :9.000   Max.    :20.000  
 (Other): 509                                                                          
             VisitorType      Weekend       
 Returning_Visitor:5630    Mode  :logical  
                           FALSE:4184      
                           TRUE  :1446      
```

Appendix B: Data Exploration

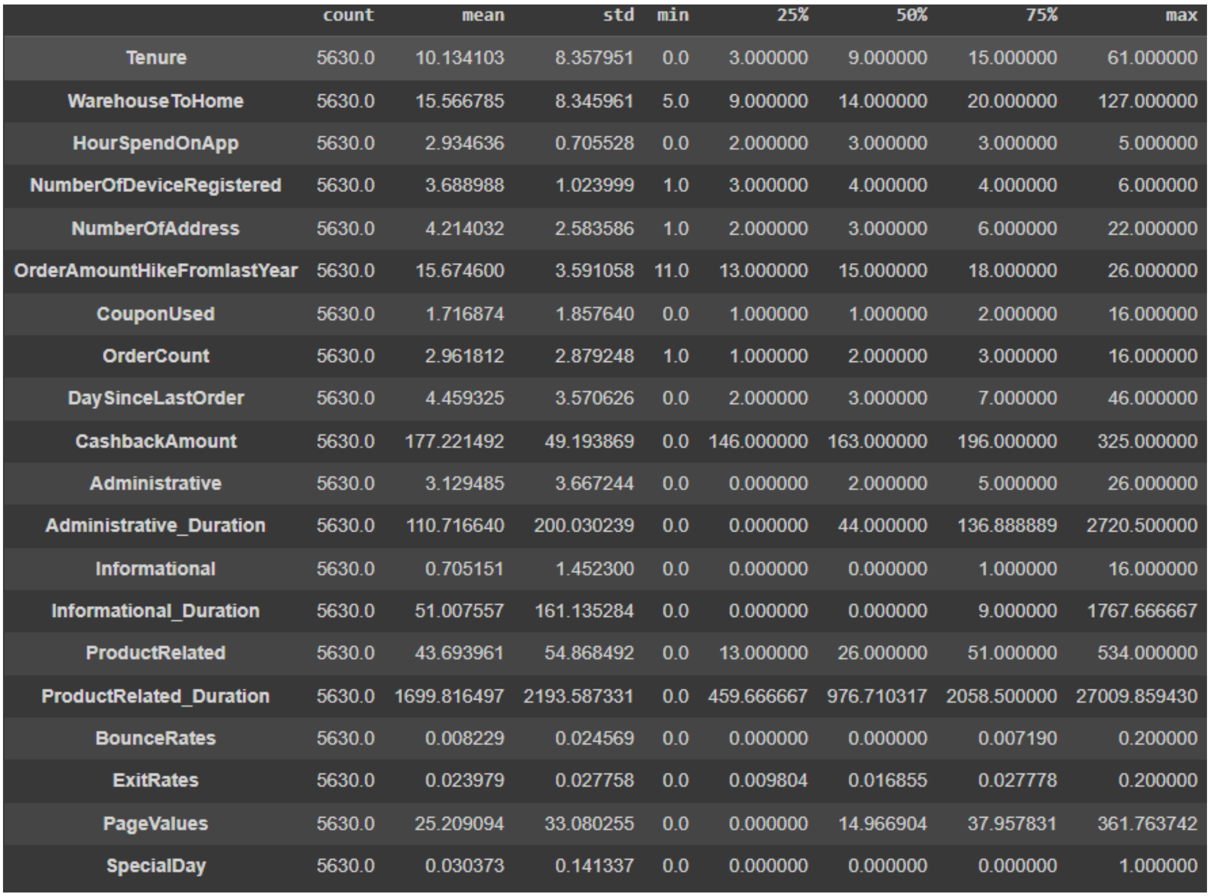
Appendix B-1: Distribution of Churn in the Dataset



Appendix B-2: Distribution of Categorical Variables in the Dataset

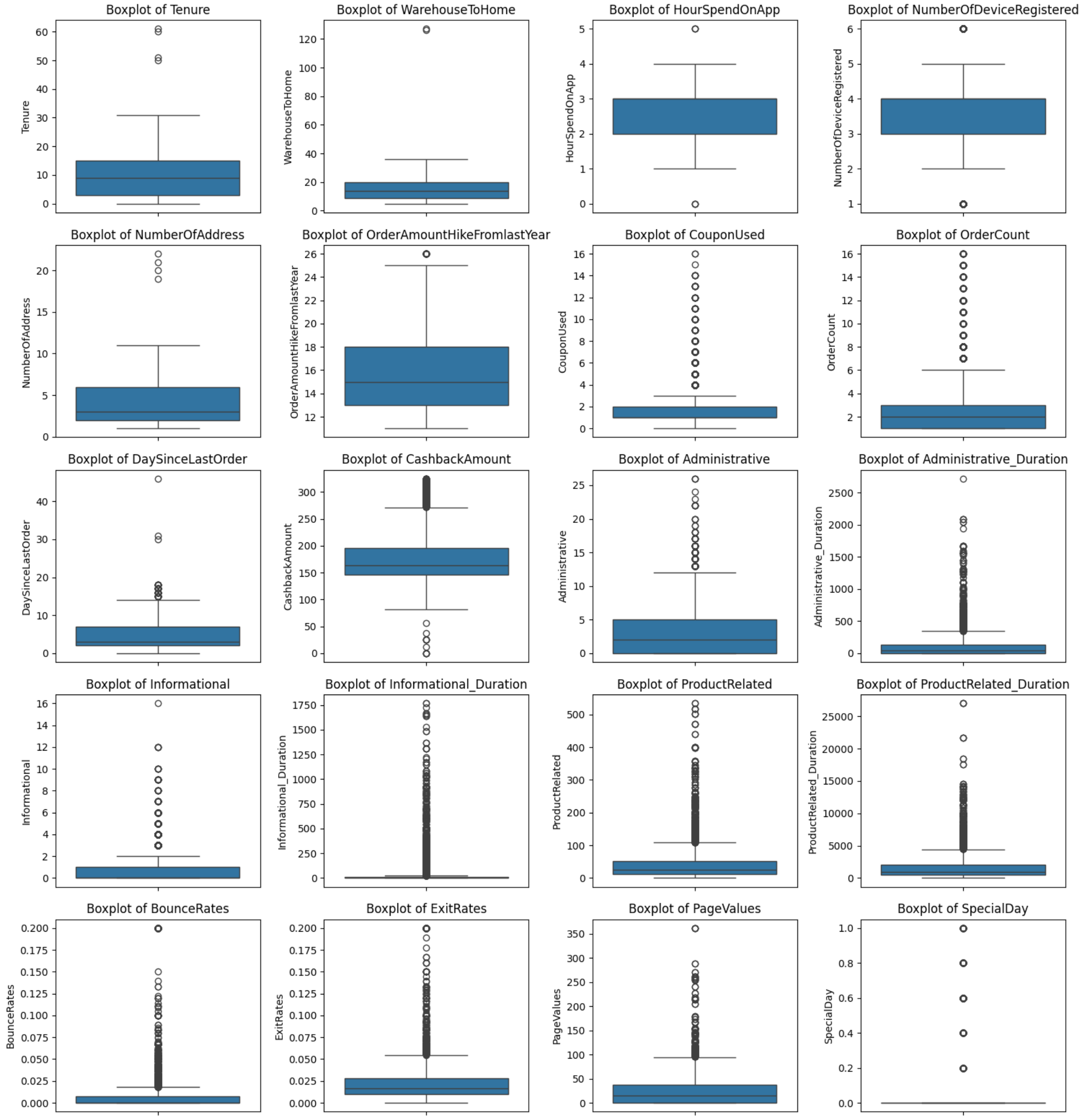


Appendix B-3: Distribution of Continuous Variables in the Dataset

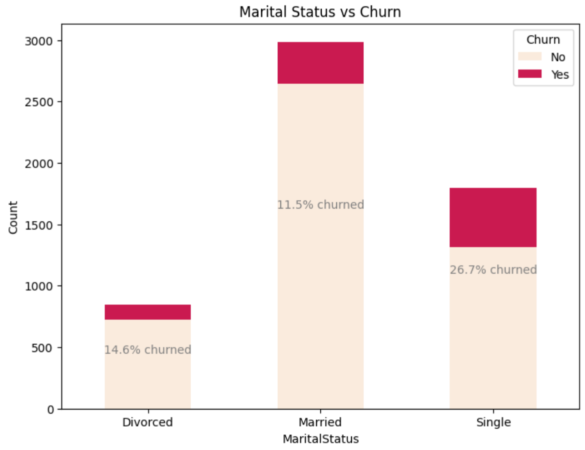


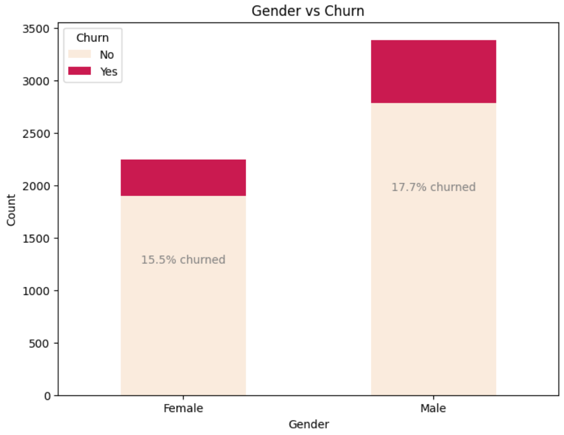
	count	mean	std	min	25%	50%	75%	max
Tenure	5630.0	10.134103	8.357951	0.0	3.000000	9.000000	15.000000	61.000000
WarehouseToHome	5630.0	15.566785	8.345961	5.0	9.000000	14.000000	20.000000	127.000000
HourSpendOnApp	5630.0	2.934636	0.705528	0.0	2.000000	3.000000	3.000000	5.000000
NumberOfDeviceRegistered	5630.0	3.688988	1.023999	1.0	3.000000	4.000000	4.000000	6.000000
NumberOfAddress	5630.0	4.214032	2.583586	1.0	2.000000	3.000000	6.000000	22.000000
OrderAmountHikeFromlastYear	5630.0	15.674600	3.591058	11.0	13.000000	15.000000	18.000000	26.000000
CouponUsed	5630.0	1.716874	1.857640	0.0	1.000000	1.000000	2.000000	16.000000
OrderCount	5630.0	2.961812	2.879248	1.0	1.000000	2.000000	3.000000	16.000000
DaySinceLastOrder	5630.0	4.459325	3.570626	0.0	2.000000	3.000000	7.000000	46.000000
CashbackAmount	5630.0	177.221492	49.193869	0.0	146.000000	163.000000	196.000000	325.000000
Administrative	5630.0	3.129485	3.667244	0.0	0.000000	2.000000	5.000000	26.000000
Administrative_Duration	5630.0	110.716640	200.030239	0.0	0.000000	44.000000	136.888889	2720.500000
Informational	5630.0	0.705151	1.452300	0.0	0.000000	0.000000	1.000000	16.000000
Informational_Duration	5630.0	51.007557	161.135284	0.0	0.000000	0.000000	9.000000	1767.666667
ProductRelated	5630.0	43.693961	54.868492	0.0	13.000000	26.000000	51.000000	534.000000
ProductRelated_Duration	5630.0	1699.816497	2193.587331	0.0	459.666667	976.710317	2058.500000	27009.859430
BounceRates	5630.0	0.008229	0.024569	0.0	0.000000	0.000000	0.007190	0.200000
ExitRates	5630.0	0.023979	0.027758	0.0	0.009804	0.016855	0.027778	0.200000
PageValues	5630.0	25.209094	33.080255	0.0	0.000000	14.966904	37.957831	361.763742
SpecialDay	5630.0	0.030373	0.141337	0.0	0.000000	0.000000	0.000000	1.000000

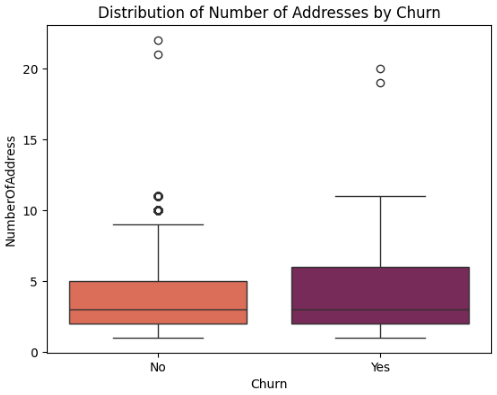
Appendix B-4: Boxplots of Continuous Variables

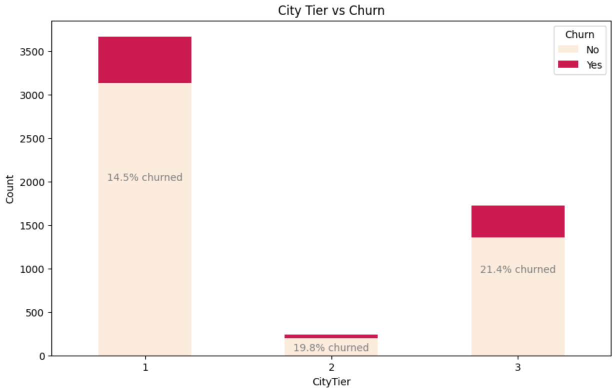


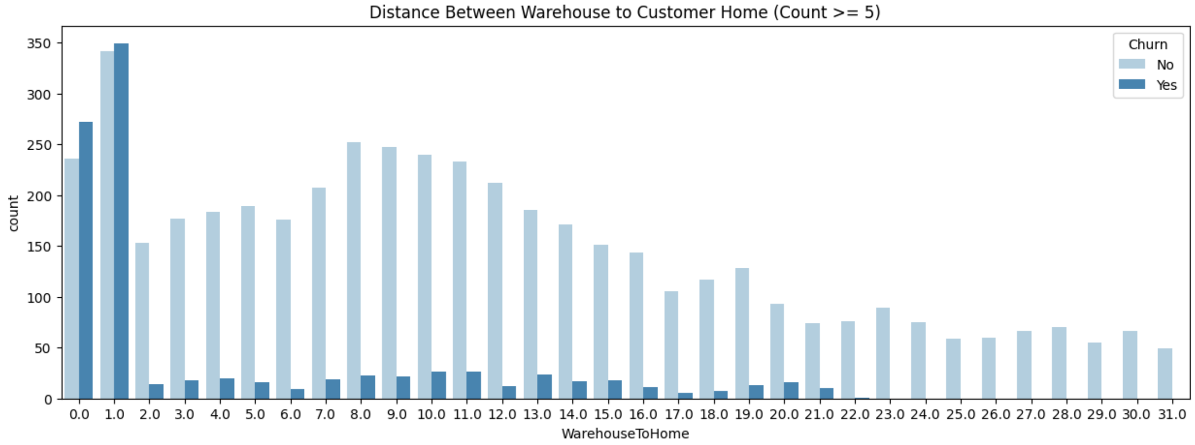
Appendix B-5: Customer Churn Analysis by Demographic Variables



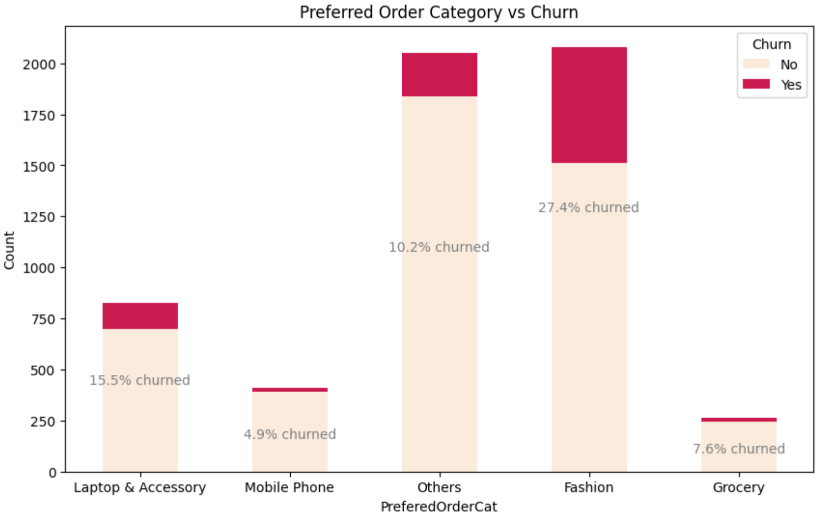


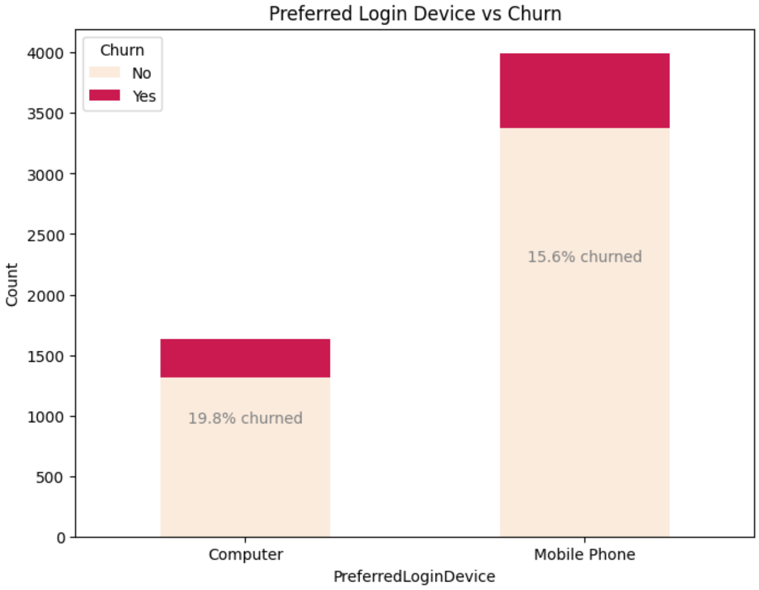


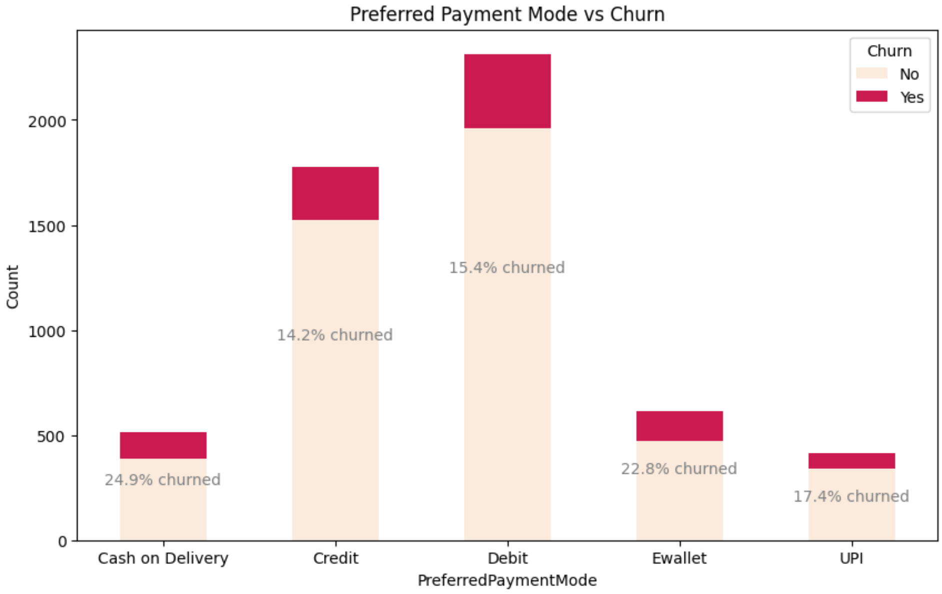


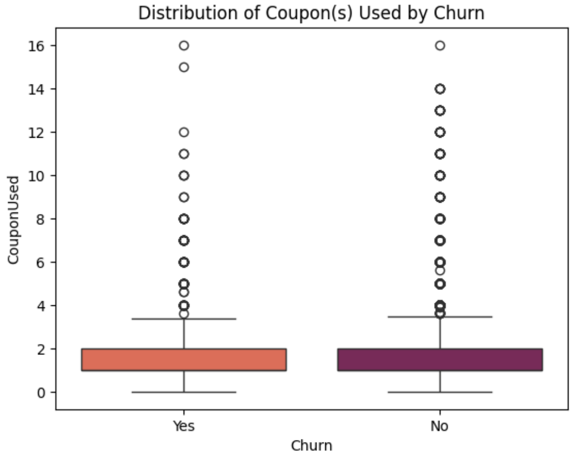


Appendix B-6: Customer Churn Analysis by Transactional Variables

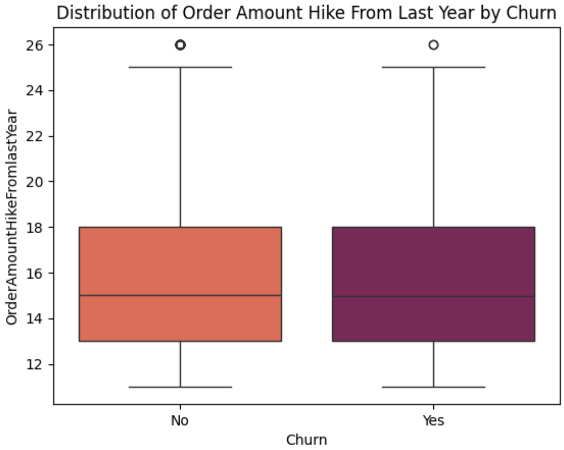




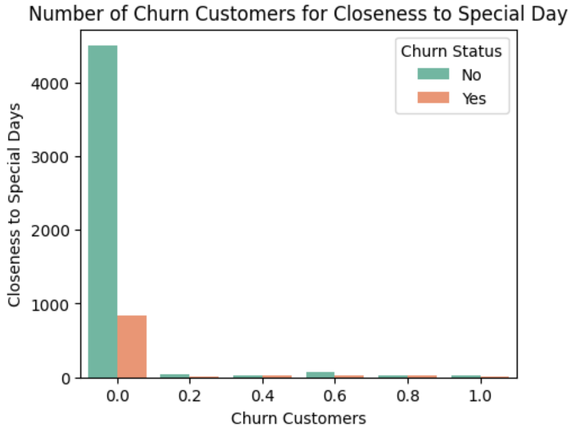




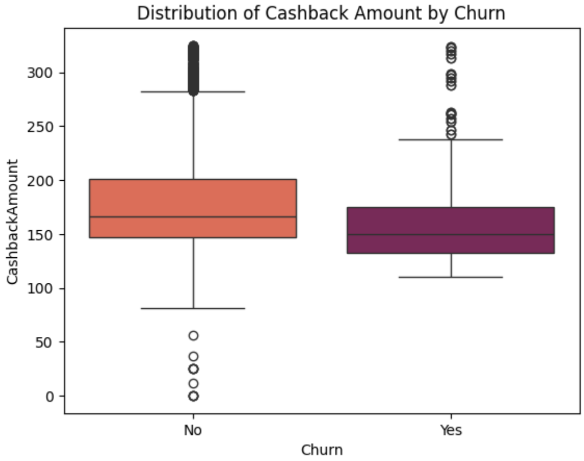
Distribution of Coupon(s) Used by Churn



Distribution of Order Amount Hike From Last Year by Churn

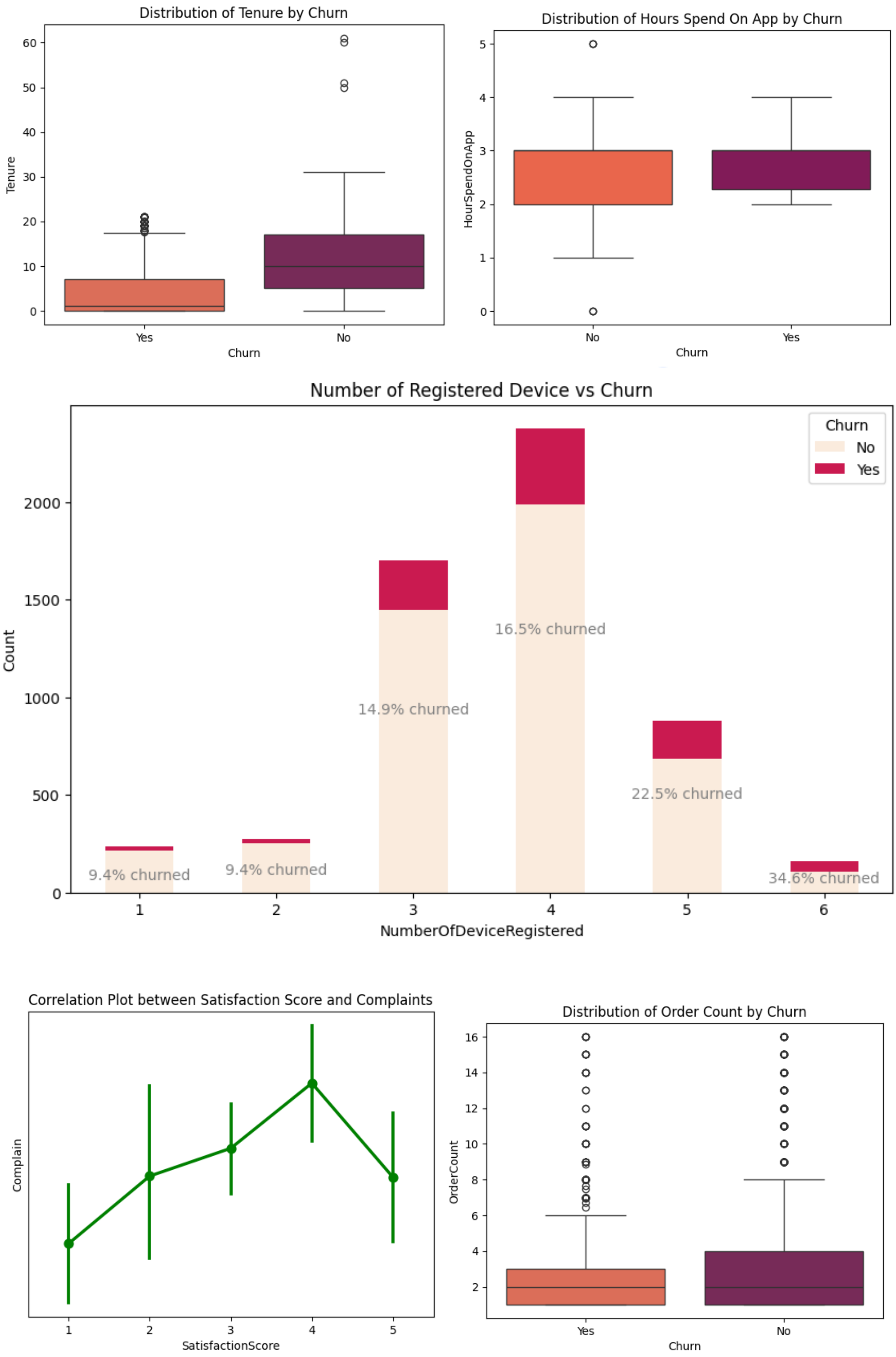


Number of Churn Customers for Closeness to Special Day



Distribution of Cashback Amount by Churn

Appendix B-7: Customer Churn Analysis by Engagement Variables



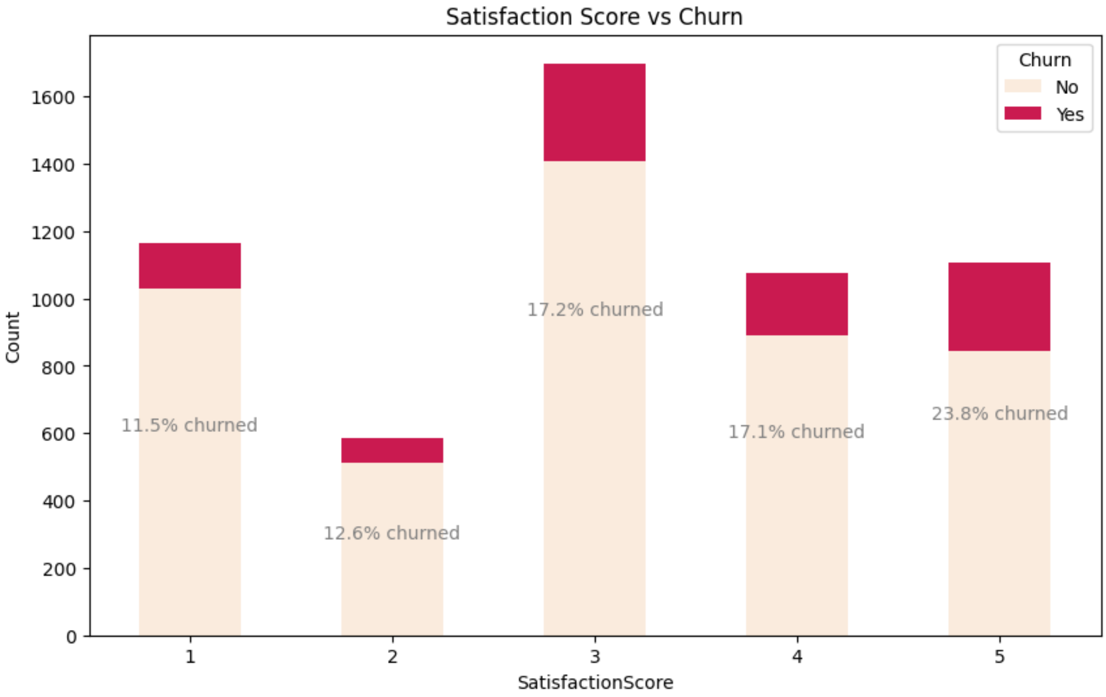
Distribution of Tenure by Churn

Distribution of Hours Spend On App by Churn

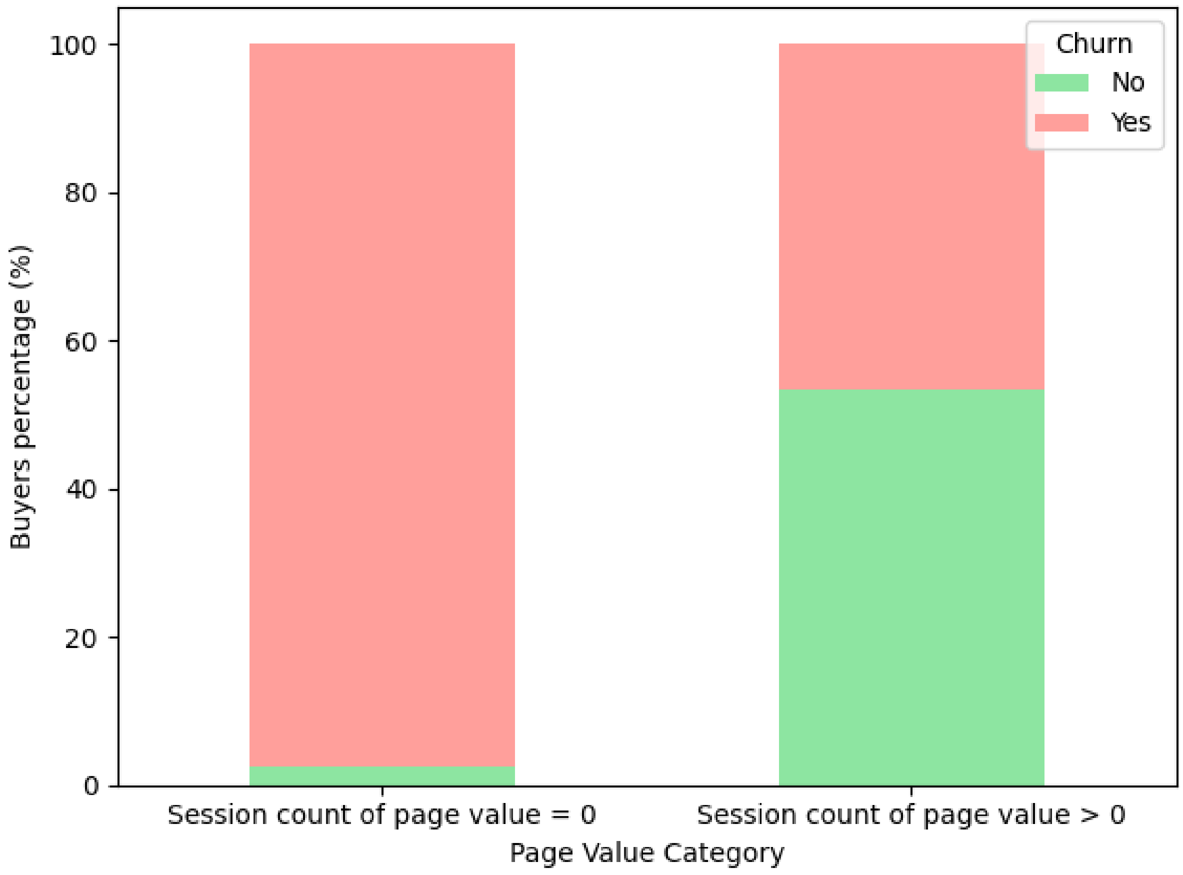
Number of Registered Device vs Churn

Correlation Plot between Satisfaction Score and Complaints

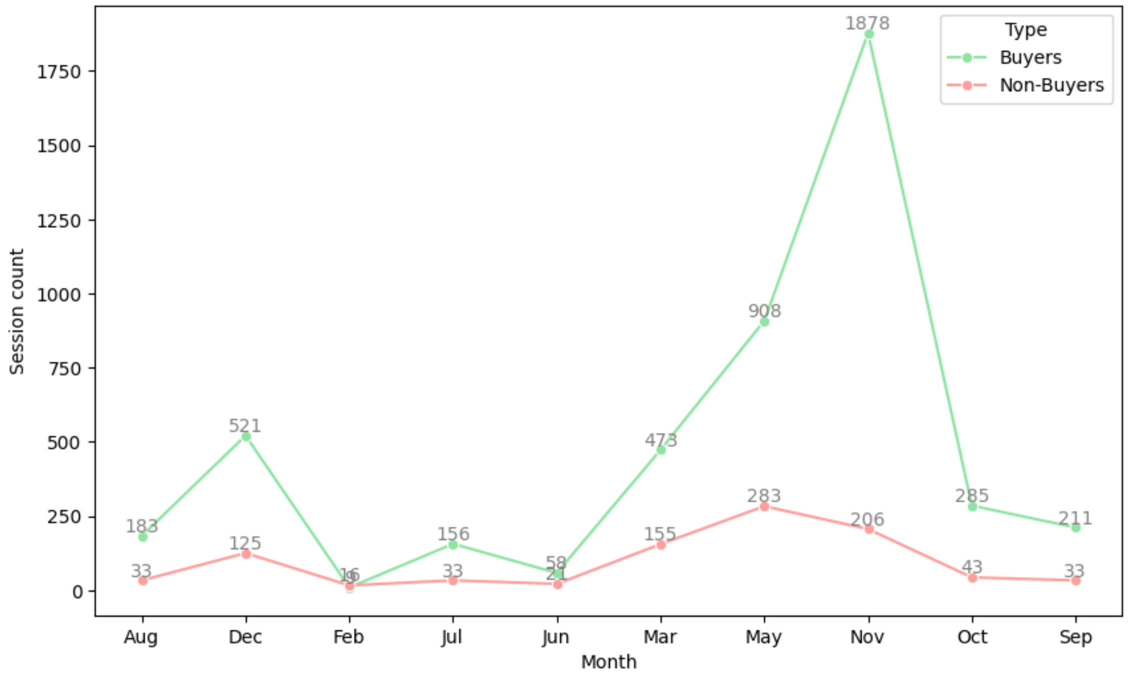
Distribution of Order Count by Churn


Satisfaction Score vs Churn
Churn
No
Yes
Count
1600
1400
1200
1000
800
600
400
200
0
11.5% churned
12.6% churned
17.2% churned
17.1% churned
23.8% churned
1
2
3
4
5
SatisfactionScore

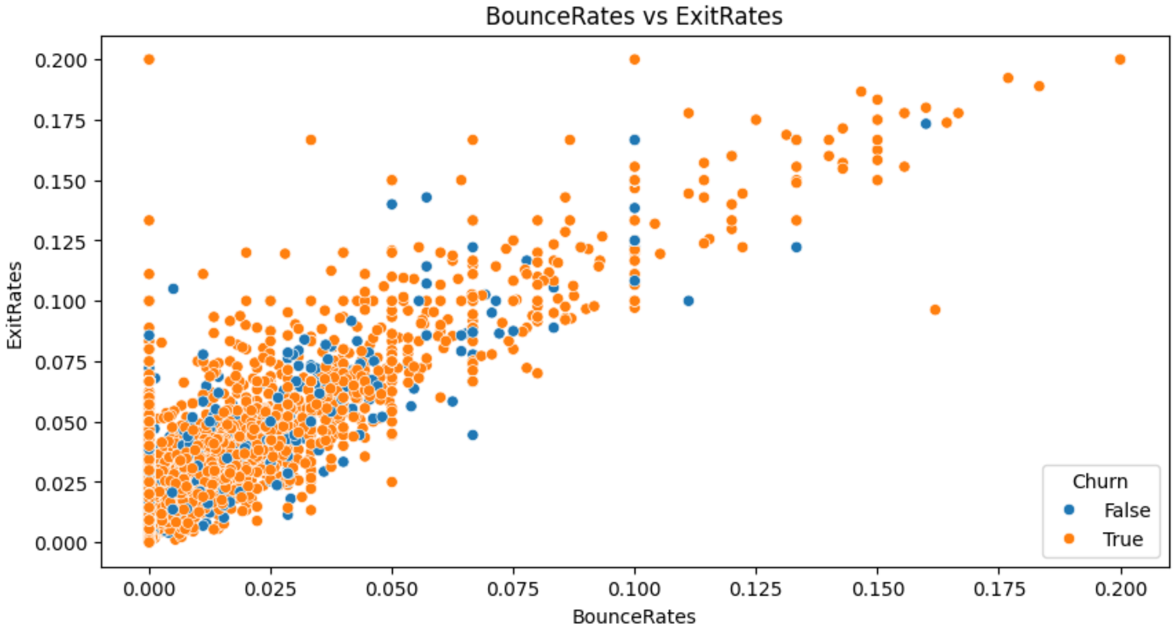
Appendix B-8: Churn Rate by Page Value Among Customers



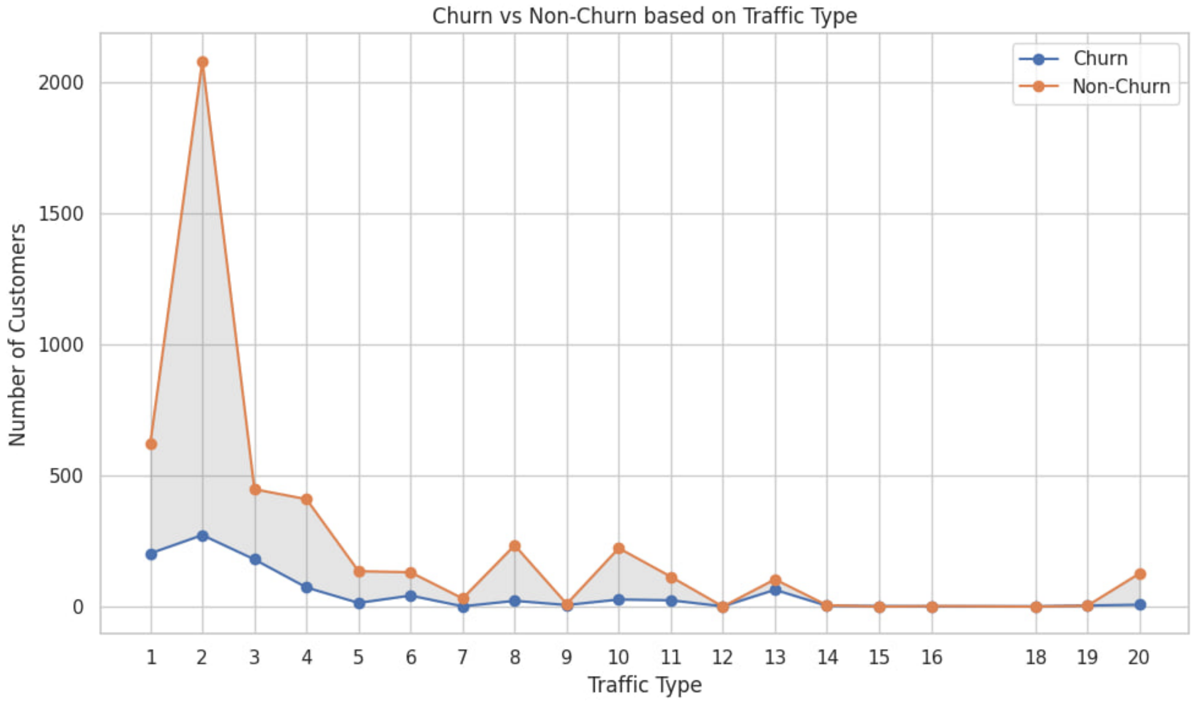
Appendix B-9: Monthly Session Count of Churned vs Retained Customers



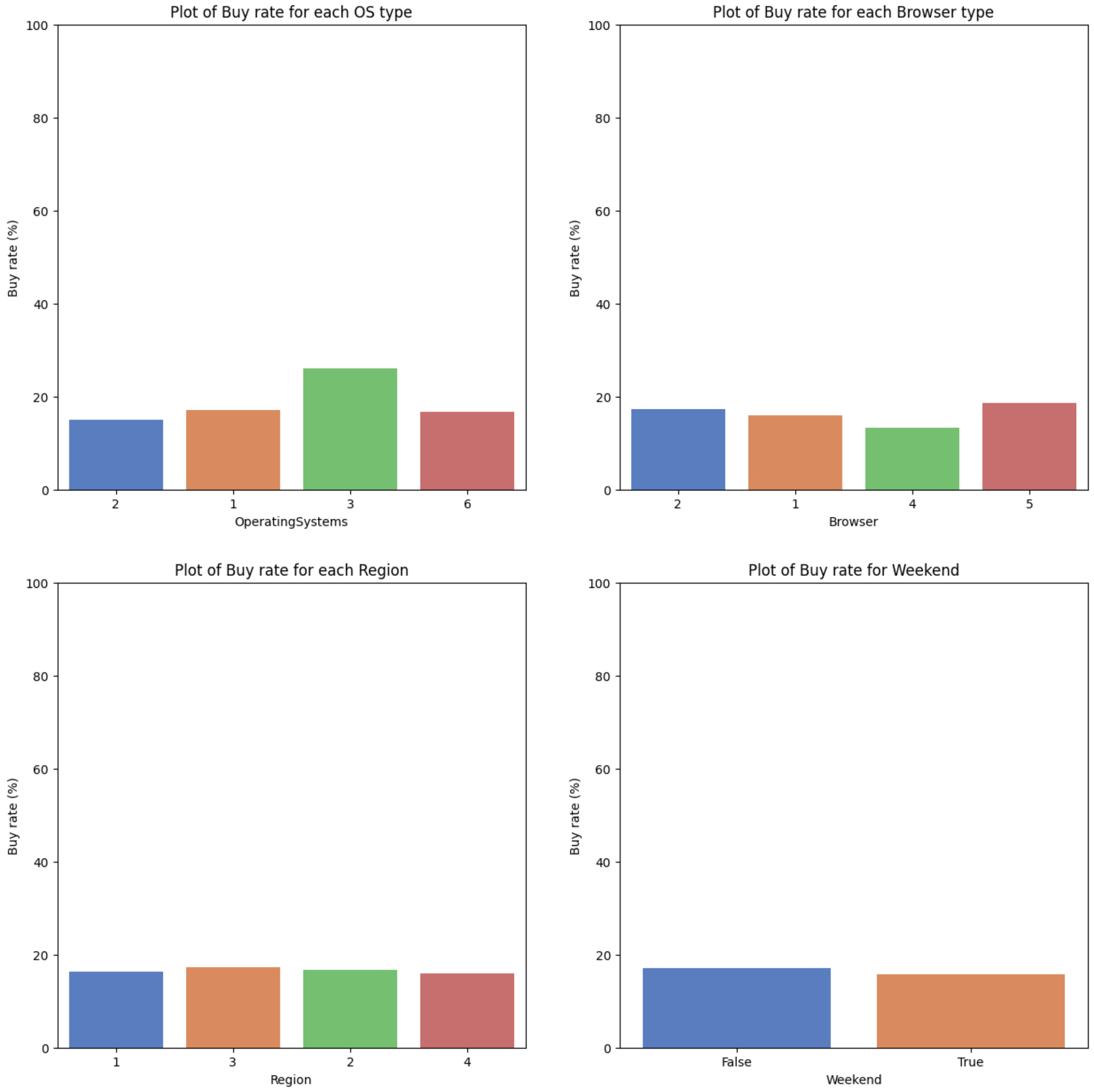
Appendix B-10: Scatter Plot of Bounce Rates vs Exit Rates with Churn Indication



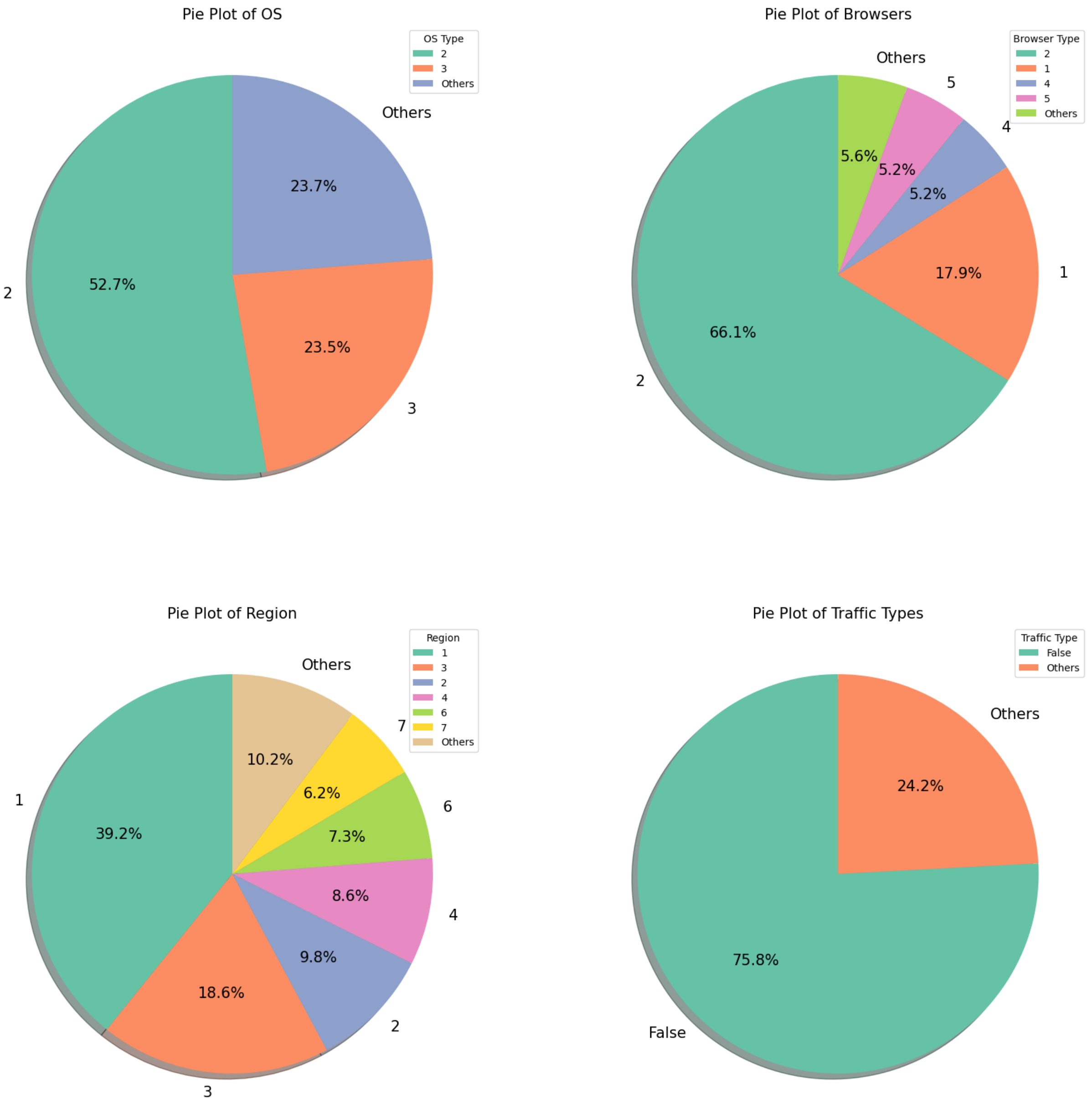
Appendix B-11: Customer Churn Comparison by Traffic Type



Appendix B-12: Bar Graphs of Buy Rates by OperatingSystem, Browser, Region, and Weekend

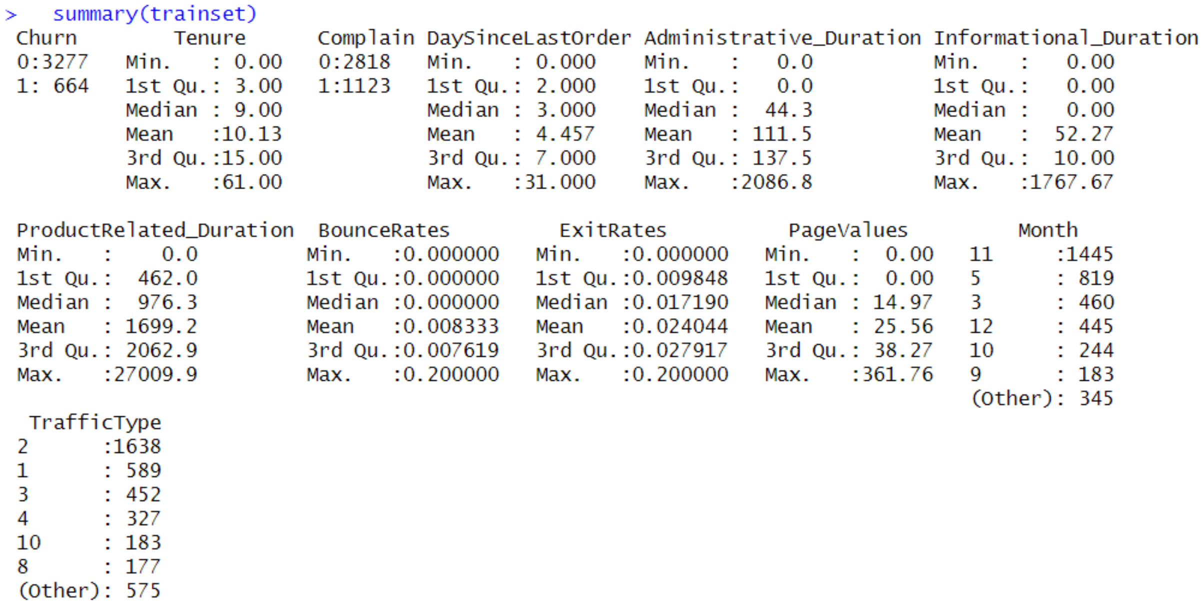


Appendix B-13: Pie Charts of Customer Distribution by OperatingSystem, Browser, Region, and TrafficType



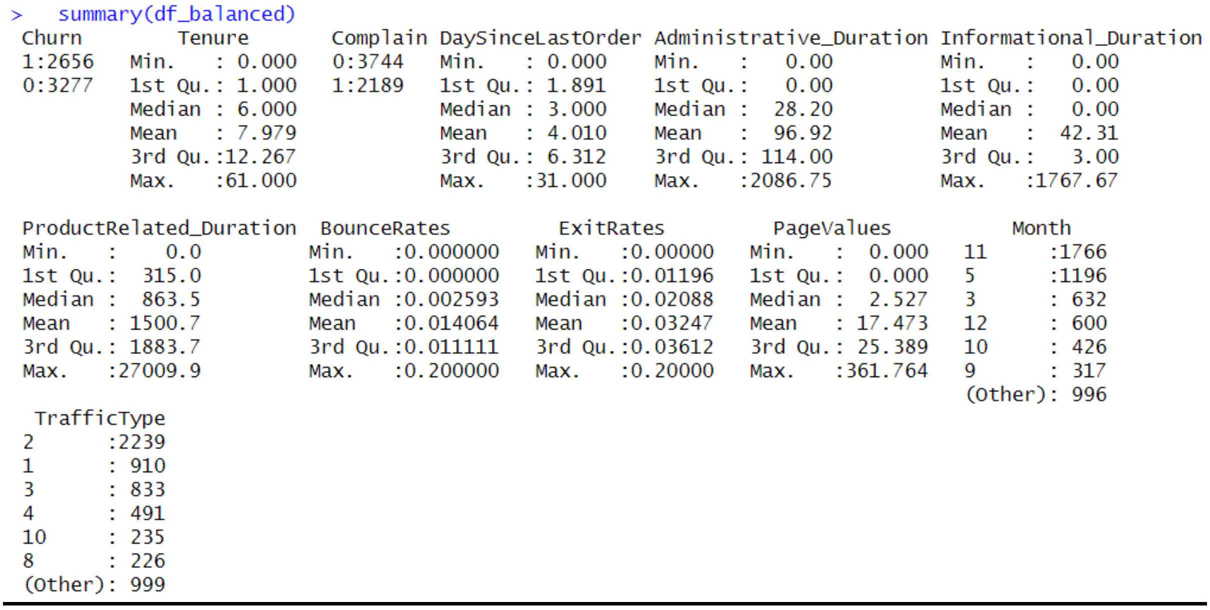
Appendix C: Data Modelling

Appendix C-1: Summary of Trainset before Oversampling



```
>   summary(trainset)
 Churn          Tenure        Complain DaySinceLastOrder Administrative_Duration Informational_Duration
 0:3277    Min.    : 0.00    0:2818   Min.    : 0.000    Min.    :    0.0          Min.    :    0.00
 1: 664    1st Qu.: 3.00    1:1123   1st Qu.: 2.000    1st Qu.:    0.0          1st Qu.:    0.00
           Median : 9.00             Median : 3.000    Median :   44.3          Median :    0.00
           Mean   :10.13             Mean   : 4.457    Mean   :  111.5          Mean   :   52.27
           3rd Qu.:15.00             3rd Qu.: 7.000    3rd Qu.:  137.5          3rd Qu.:   10.00
           Max.   :61.00             Max.   :31.000    Max.   :2086.8          Max.   :1767.67

 ProductRelated_Duration  BounceRates         ExitRates           PageValues          Month
 Min.    :     0.0       Min.    :0.000000   Min.    :0.000000   Min.    :  0.00    11      :1445
 1st Qu.:   462.0       1st Qu.:0.000000   1st Qu.:0.009848   1st Qu.:  0.00    5       : 819
 Median :   976.3       Median :0.000000   Median :0.017190   Median : 14.97    3       : 460
 Mean   :  1699.2       Mean   :0.008333   Mean   :0.024044   Mean   : 25.56    12      : 445
 3rd Qu.:  2062.9       3rd Qu.:0.007619   3rd Qu.:0.027917   3rd Qu.: 38.27    10      : 244
 Max.   :27009.9       Max.   :0.200000   Max.   :0.200000   Max.   :361.76    9       : 183
                                                                               (Other): 345
  TrafficType
 2       :1638
 1       : 589
 3       : 452
 4       : 327
 10      : 183
 8       : 177
 (Other): 575
```

Appendix C-2: Summary of Trainset after Oversampling

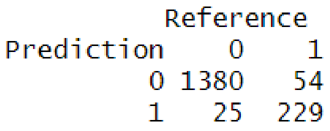
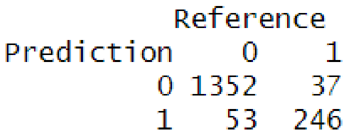
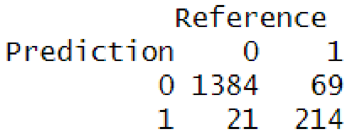
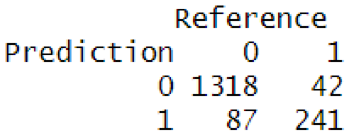
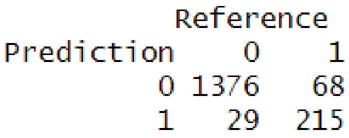
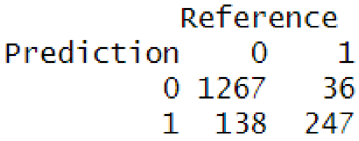


```
>   summary(df_balanced)
 Churn          Tenure         Complain DaySinceLastOrder Administrative_Duration Informational_Duration
 1:2656    Min.    : 0.000    0:3744   Min.    : 0.000    Min.    :    0.00         Min.    :    0.00
 0:3277    1st Qu.: 1.000    1:2189   1st Qu.: 1.891    1st Qu.:    0.00         1st Qu.:    0.00
           Median : 6.000             Median : 3.000    Median :   28.20         Median :    0.00
           Mean   : 7.979             Mean   : 4.010    Mean   :   96.92         Mean   :   42.31
           3rd Qu.:12.267             3rd Qu.: 6.312    3rd Qu.:  114.00         3rd Qu.:    3.00
           Max.   :61.000             Max.   :31.000    Max.   :2086.75         Max.   :1767.67

 ProductRelated_Duration  BounceRates         ExitRates          PageValues          Month
 Min.    :     0.0       Min.    :0.000000   Min.    :0.00000   Min.    :  0.000    11      :1766
 1st Qu.:   315.0       1st Qu.:0.000000   1st Qu.:0.01196   1st Qu.:  0.000    5       :1196
 Median :   863.5       Median :0.002593   Median :0.02088   Median :  2.527    3       : 632
 Mean   :  1500.7       Mean   :0.014064   Mean   :0.03247   Mean   : 17.473    12      : 600
 3rd Qu.:  1883.7       3rd Qu.:0.011111   3rd Qu.:0.03612   3rd Qu.: 25.389    10      : 426
 Max.   :27009.9       Max.   :0.200000   Max.   :0.20000   Max.   :361.764    9       : 317
                                                                               (Other): 996
  TrafficType
 2       :2239
 1       : 910
 3       : 833
 4       : 491
 10      : 235
 8       : 226
 (Other): 999
```

Appendix C-3: Top 10 Important Features Identified in each Model

Ranking	Random Forest	CART	Logistic Regression
1	PageValues	PageValues	Tenure
2	Tenure	Tenure	PageValues
3	Month	Month	Complain
4	ExitRates	ExitRates	NumberOfAddress
5	ProductRelated_Duration	ProductRelated_Duration	ExitRates
6	Complain	ProductRelated	MaritalStatus
7	TrafficType	CashbackAmount	DaySinceLastOrder
8	MaritalStatus	BounceRates	OrderCount
9	BounceRates	SpecialDay	NumberOfDeviceRegistered
10	ProductRelated	CouponUsed	SatisfactionScore

Appendix C-4: Confusion Matrix of Classification Models

Model	Before Oversampling	After Oversampling
Random Forest	 Reference Prediction 0 1 0 1380 54 1 25 229	 Reference Prediction 0 1 0 1352 37 1 53 246
CART	 Reference Prediction 0 1 0 1384 69 1 21 214	 Reference Prediction 0 1 0 1318 42 1 87 241
Logistic Regression	 Reference Prediction 0 1 0 1376 68 1 29 215	 Reference Prediction 0 1 0 1267 36 1 138 247