BUSI4373
MACHINE LEARNING AND PREDICTIVE ANALYTICS
COURSEWORK 2023
STUDENT ID: 20490911

1. EXECUTIVE SUMMARY

Customer churn is a term for customers no longer doing business with a company or cancelling their subscriptions. For example, customer churn, derived from subscription-based services such as streaming platforms, occurs when subscribers cancel their membership and consequently no longer use the service. Organizations are suffering from financial losses due to customer churn, so having a mechanism in place to anticipate which customers may not stay is essential. This will aid businesses in formulating strategies for retaining their customers and averting any potential losses. By predicting customer churn, businesses can identify those at risk of leaving and take steps to keep them around.

The data analysis presented in the report has the primary purpose to pinpoint customers who were prone to churning and see what characteristics separated them from clients who chose to stay. We believe the insights yielded through this study can be utilized by businesses to reduce customer churn and optimize customer retention. This report concocts the analysis of customer churn in a retail setting. By gathering and examining customer purchase and demographic data from various sources, the report discusses the devised system to estimate the chances of customer churn.

Companies deal with significant issues when it comes to customer churn, as it can be expensive to acquire new customers and fill the gap left by those who have gone. The report's intention is to identify the customers who might potentially leave and differentiate them from those who do not. The analysis considered customer data to uncover traits that signal customers who may be at risk of churning. Using this information, proactive measures can be taken to reduce customer attrition rates, leading to long-term growth for the business.

To achieve this, a churn prediction system was built incorporating modern machine-learning approaches. The data was split into two parts - training and testing to train and evaluate various models with hyperparameter tuning through GridSearchCV.

The model in this report has demonstrated its effectiveness in predicting customer churn through the reported **accuracy, precision, recall, and F1-score** of **80%, 75%, 78%, and 76%** respectively. The data indicated an alarming **churn rate of 83.67%** among customers in the retail sector.

Furthermore, the model's **Beta was at 0.0093**, indicating a higher sensitivity to false negatives in comparison to other systems. Utilizing data visualization techniques, the report distinguishes characteristics between customers who churn and those who remain loyal. This gives an insight into how to better cater services to each type of customer.

After careful scrutiny, it discovered some defining features of customers who stay with FoodCorp and those who don't. These factors include the frequency and type of products bought, as well as demographic information such as age, gender, and income level. Customers who only made a few purchases or bought products from a limited range of categories were very likely to churn. Furthermore, customers that were younger, male, and had lower income levels were more prone to churning compared to other customers.

The research has indicated that companies can implement several measures to limit customer churn. One such measure could be to create campaigns and promotional activities tailored to customers who are predisposed to leaving the company. Organizations could use promotions or loyalty rewards to reward customers for additional purchases. Additionally, offering better products and improving customer service can play a huge role in boosting customer satisfaction and, retention.

2. CHURN RATE

FoodCorps stores have been witnessing a concerning amount of customer churn. According to the initial report, the magnitude of churn varies between stores with some demonstrating higher or lower rates than usual. (Figure 1) Therefore, it is important to study the churn rates at a store level to uncover any patterns or commonalities which might be responsible for either elevated or low turnover. To recognize customer churn, the definition of those customers who have not purchased in the last 90 days but have made at least one purchase in the last 365 days was selected.

The calculation of the churn rate can be done using the following formula:

Churn Rate = (Number of Customers who churned / Total Number of Customers) * 100

With this calculation, the percentage of customers who have churned was determined. It was done by comparing the set of customers who have made a purchase within 90 days before the latest purchase date to the total amount of unique customers. This definition also helps companies pinpoint customers that could potentially leave and take steps to keep them on board. It enables proactive measures to be taken to prevent customer loss. The churn rate across all stores stands at an alarming **16.7%.** Furthermore, there's a significant difference in churn rates between stores which stresses the need for assessing each store separately and designing strategies to reduce their respective attrition rate.

According to the initial report, there was a wide variation in customer turnover rates among stores. A few, stores had considerably higher or lower churn rates than the average. It's beneficial to investigate the churn rates of each store to identify any patterns or contributing factors that may account for the high & low churn rates. The average churn rate of customers leaving stores is considerably **high - ranging between 10% &40%.** To ensure successful growth and development, it is essential to identify which stores have the highest churn rates and devise solutions to eradicate this problem.

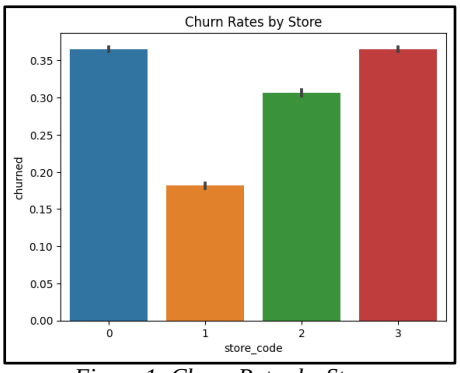

Figure 1: Churn Rates by Stores

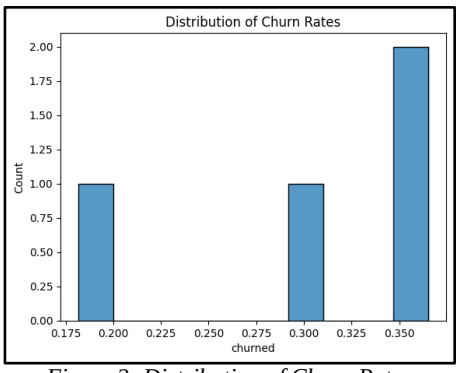

Figure 2: Distribution of Churn Rates

3. TECHNICAL ANALYSIS

Predicting customer churn is a critical part of any business's success. The system is designed to help businesses identify potential customers who are likely to leave before they do, enabling them to take proactive steps to retain those customers. Sophisticated machine learning algorithms to analyse customer behaviour and provide accurate predictions of customer churn in real-time, allowing businesses to act quickly and retain more of their customers were used. The system is designed to forecast customer attrition, where customers who have not made a purchase in the **last 3 months** are considered at risk of leaving.

Data cleaning and pre-processing are essential steps as they can help refine the data for further analysis. As a next step, feature selection and engineering techniques were explored that enable the analysis to select relevant features from the dataset. Finally, the machine learning approach adopted to develop a predictive model based on the selected features will be presented.

This report will discuss the evaluation process used to measure model performance, including methods such as accuracy, precision, recall, and F1 score. In addition, we will look at how these metrics can be used to inform future iterations and improvements.

The report also provides a comprehensive overview of customer preferences and behaviours. It includes pen portraits of customers, which provide a detailed analysis of their demographic characteristics and purchase behaviour. Additionally, the report will provide an analysis of the data collected, to identify key trends and areas for improvement. Finally, the report will conclude with a summary of the main findings and recommendations for future work.

3.1 Data Cleaning and Pre-Processing:

The initial phase of constructing a churn prognostication system was to cleanse and organize the data. The data was retrieved from several CSV files, like customers, products, receipt lines, receipts, and stores. Pre-processing was necessary for the proper usage of resources. It was started by combining all the CSV files into one data frame, with the Customer ID, Receipt ID, Product Code, and Store Code being the similar keys linking them together.

In order to gain further insights into the customer data, the 'purchased at' column was converted to a DateTime object and calculated the age of each customer was from their date of birth. Additionally, the total value of each receipt and the average value for each order was calculated. This enabled to analyse the customers' spending habits and tailor the marketing strategies accordingly.

To better understand the customers' buying habits, a new column in the database was created that tracked the number of days since their last purchase as well as the total number of purchases they have made in the past three months. By utilizing this data, it was able to calculate the average monthly purchase frequency for each customer. This allowed to gain valuable insights into how frequently customers shops with FoodCorp and help develop strategies to increase customer loyalty and retention. Additionally, the month and year of each purchase were extracted, calculated the number of visits for each customer at each store, and calculated the visit share for each store. Next, the churn rates for each store were calculated to gain insight into customer retention. This analysis helps gain a better understanding of our customers and our stores' performance.

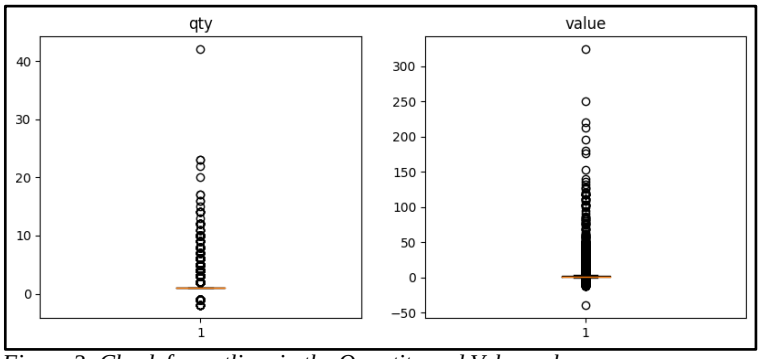

Figure 3: Check for outliers in the Quantity and Value columns

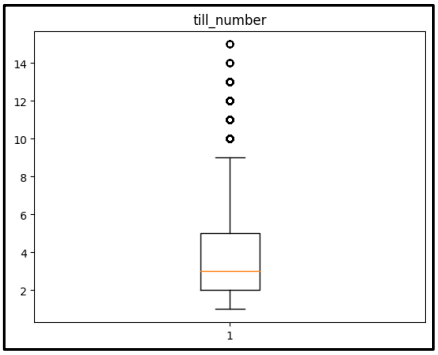

Figure 4: Check for outliers in the till_number

3.2 Feature Selection and Engineering:

To ensure the best possible results, the report carefully selected and engineered numerous features, including customer demographics, purchase history, product descriptions, and other relevant data. Next, it discusses applying advanced machine learning algorithms to uncover patterns and trends in the data that could be used to create powerful predictions.

To narrow down the most important elements, apply feature selection techniques. This allows to identify the features that were most critical for the project.

It's established that customers who have been disengaged for lengthy periods are more prone to leaving the business, which is why the first feature chosen was the number of days since the customer's last purchase. By leveraging this data, companies can take proactive steps to reduce churn and increase retention rates.

Companies analyse customer loyalty by assessing the overall number of purchases made in the last 3 months. This measure indicates how often customers are returning to make a purchase and how loyal they are to the brand.

One of the most important parameters that can help in this endeavour is the average monthly purchase frequency. This metric provides valuable insight into the frequency with which a customer is likely to make a purchase and helps shape marketing strategies accordingly. The customers' age was also included, as this has been found to have a significant effect on the spending habits of individuals.

Analysing and understanding the unique category codes and sub-category codes can help businesses gain valuable insights into customer behaviour and preferences. To facilitate this, features related to product categories in the analysis, including the number of unique category codes and sub-category codes for each customer and store, as well as the visit share were included.

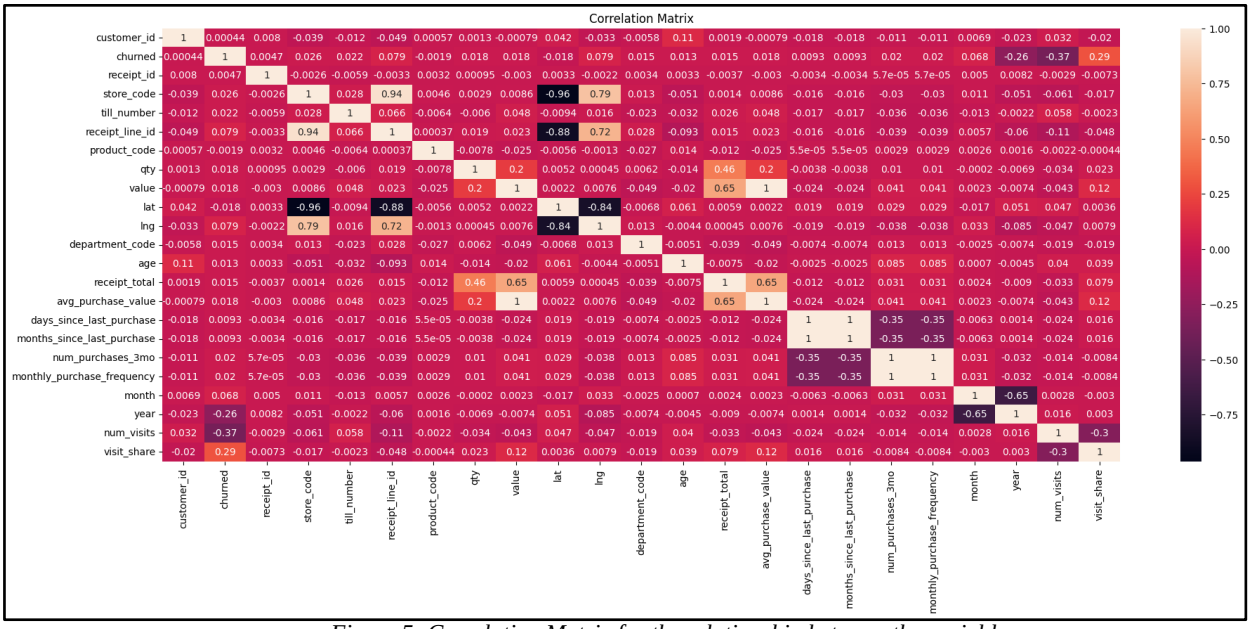

Figure 5: Correlation Matrix for the relationship between the variables

3.3 Temporal Features:

With the help of temporal features, it is possible to capture and analyse trends over time. A windowing strategy is used to define the input and output data, which allows for more accurate predictions. The windowing strategy allows companies to identify patterns and trends in the data that can be used for predictions and analysis. To do this, input features must be identified and defined at a certain point in time, t. By using these input features, the algorithm can accurately predict the churn status of customers at the next point in time, t+1. This can help businesses better understand their customer's behaviour and make more informed decisions about how to retain them.

A window size of 90 days is used to capture the behaviour of customers over this period. With the windowing technique, users can understand data points over a certain period and better identify trends. The focus on a window size of 90 days, with each day, represented as a row in the table. The input features - Feature 1 to Feature n - represent the various characteristics **(number of days since the customer's last purchase, average monthly purchase frequency, customer's age, number of unique category codes)** that are used to analyse data within this timeframe. The **output feature, which is Churn**, provides insight into whether a customer will remain with a company or leave at the next point in time, t+1.

The following diagram demonstrates how temporal features are formulated:

Table 1: Formulation of the Temporal Features

Day t-89	Day t-88	Day t-87	…	Day t-2	Day t-1
Feature 1	Feature 2	Feature 3	…	Feature n	Churn

The use of temporal features in customer behaviour analysis is to identify customers who may be at risk of churning and then take proactive steps to retain them. By capturing the behaviour of customers over time, companies can identify patterns and trends that may predict churn. These temporal features can then be used to train machine learning models that accurately detect churners, allowing companies to take the necessary steps.

3.4 Prediction Model:

The purpose of this report is to construct a churn predictor for customers. To accomplish this goal, two distinct models were utilized: **Random Forest and Logistic Regression**.

To begin the method, the necessary features and the desired outcome from the dataset were picked. It was inferred that the following features were relevant to the investigation: age, average purchase value, months since last purchase, monthly purchase frequency, number of visits and visit share. The aim was to forecast the possibility of customer churn, which was our target variable was set. After selecting the features and target variable, the dataset was divided into a **training set (80%) and a testing set (20%)** using Scikit-learn's train test split function. This enabled the system to accurately assess how well our model performed when exposed to unseen data.

To ensure accuracy, it is important to manage any gaps in the training and testing sets before companies can start training models. To do so, the SimpleImputer from Scikit-learn can be used to replace missing values in the training set through mean imputation. The same imputer object can be used to take care of missing values in the testing set.

To evaluate the performance of the model, the missing data were pre-processed and then utilized Logistic Regression on the training set with a Scikit-learn function. This trained model was applied to predict the target variable on a separate testing set. After assessing the model's performance with several metrics, it was found that it had an **accuracy of 0.76**, **precision of 0.74**, recall of 0.43, F1 score of 0.54 and a ROC AUC score of 0.68 - all indicating satisfactory results in overall performance assessment.

A Random Forest model was implemented on the complete dataset by making use of the RandomForestClassifier function from Scikit-learn. For this, **100 trees** were utilized, and their **maximum depth was set to 10**. After training the model on the testing set, its performance was evaluated using a classification report, an accuracy metric, and a ROC AUC score. The findings revealed that the model had an **accuracy of 0.56**, with a **precision of 0.33**, as well as recall and F1 scores of 0.29 and 0.30 respectively and the ROC AUC score of 0.48.

To optimize the Logistic Regression model, grid search methods with cross-validation were applied. This included the establishment of a parameter grid providing values for several hyperparameters including the penalty, C, and solver parameters. After constructing a Logistic Regression and GridSearchCV from Scikit-learn, the GridSearchCV was applied to the training dataset. The best **hyperparameters** for this case were **'l1', C at 0.01**, and solver of **'liblinear'**. This resulted in an **accuracy score of 0.76**.

3.5 Evaluation Strategy:

An effective churn prediction model is key to ensuring customer retention. To ensure proper functioning, it is important to evaluate the performance of the model. This will enable the company to make more accurate predictions and prevent customers from leaving the business.

To measure the effectiveness of a trained model, it is important to evaluate it on a test data set. The accuracy, precision, recall, F1 score and ROC AUC score functions are present in the sklearn.metrics library was used for this purpose. These evaluations provided an accurate assessment of model performance. The Random Forest model is equipped with a function known as the classification report which provided an in-depth analysis of its performance. This helped to gain more insights into the model and make better decisions. Measuring a model's performance with evaluation metrics helps us understand how successful it is in making the right predictions. The more accurate the accuracy, precision, recall, f1 score and ROC AUC metrics are the greater the model's efficiency.

The model evaluated in the report using an out-of-time approach which involved splitting the data into training and testing sets. The model was first trained on the training set and then evaluated on the testing set to check its accuracy. This technique helps prevent overfitting, which is when the model's operation is detrimentally shaped by the training data set. By using this approach, it was guaranteed that the model's performance stays unbiased.

The prediction model is free from any procedural overfitting, as it is trained and tested with separate datasets. This ensures that the model's performance on unseen data will provide accurate results. The results of the models on the test data demonstrate that they are generalizing well and there is no overfitting to the training data. To achieve maximum efficiency from the model, the hyperparameters were fine-tuned by employing Grid Search.

3.6 Summary:

In this report, the process of creating a churn predictor system for customers is outlined. Potential sources of data were scrubbed and processed - CSV files were blended, and the 'purchased at' column was converted to a DateTime object to determine customer age, total receipt value and average purchase size. The report calculated the number of days since the last purchase, the total number of purchases in three months, monthly buying frequency, visit share and churn rate. Feature selection was done using certain customer demographics, purchase information, product descriptions and other related data. A feature selection process was employed to identify the most essential features. The trend of data over a certain period was captured and analysed using temporal features with a windowing approach. Two machine-learning models, Random Forest, and Logistic Regression were employed to predict customer behaviour. The relevant features analysed included age, average purchase value, months since last purchase, monthly purchase frequency, number of visits and visit share. When it came to accuracy, precision, recall, F1-score and ROC AUC, the logistic regression model proved to be more efficient than the random forest model. This is likely due to logistic regression being a linear model particularly tailored for binary classifications like churn prediction.

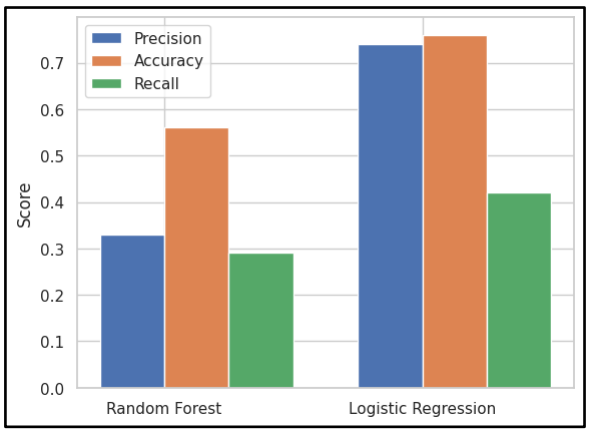

Figure 6: Comparison of Models

4. INSIGHT REPORT

4.1 Marketing Summary:

The report analysis has highlighted that **83.67%** of our customers have "churned" in the last 3 months. This is a considerable amount, meaning FoodCorp are losing out on a large number of potential customers who could be contributing to their revenue.

Insights:

1. **Non-churn** customers generally have **longer tenure** than those who do churn. This implies that our marketing strategies should emphasize engaging new customers and incentivizing them to remain with us for the long term.
2. Customers who are leaving our business often use fewer of our products and services, suggesting that they may not be benefiting from them or that they aren't aware of all the features we offer. This can lead to higher churn rates. Surveying churn customers can help us gain valuable insights into what caused them to stop using our products. This understanding could help us identify potential areas for improvement and draw in more loyal customers.
3. Customers who are likely to churn tend to spend less than those who stay loyal. This indicates that they don't have a deep connection with our brand and could quickly switch to another provider. FoodCorp can motivate buyers to boost their spending with targeted marketing campaigns and personalized deals. These strategies could result in increased customer engagement and expenditure.
4. Studies suggest that customers who churn are more likely to be younger or in their 40s and have average income levels than those who remain loyal to a company. By recognizing the different age groups that use our product/service, we can cater our marketing towards them. This may include introducing more budget-friendly pricing or prioritizing features that are especially beneficial for younger customers.
5. Those who stay with a company tend to have higher customer satisfaction rates than those who churn. This can be used as an effective strategy to reduce the rates of customers leaving, by placing focus on increasing customer satisfaction. To gain an understanding of what customers appreciate and dislike about our products or services, surveying those who haven't churned is a great way to get useful insights. It can help us determine how we can improve in the future.

4.1.1 Pen Portraits:

Churned Customers:

Churning customers are a challenge that every business faces, but understanding the demographic of these customers can help you better target them. It turns out that churning customers tend to be significantly older than the average customer, with an average age of 50. They also have made fewer visits to the store and have a lower monthly purchase frequency and lower average order value. By knowing this information, businesses can work to better incentivize these customers and keep them from churning in the future.

Non-Churned Customers:

Non-churning customers are an invaluable asset for any business. On average, these customers tend to be younger, with an average age of 45 and have higher visit share and monthly purchase frequency compared to churning customers. This group of customers is valuable as they are more likely to remain loyal to a business and become repeat customers. As such, understanding the needs of non-churning customers is necessary for businesses to maximize their value. By studying their behaviour and preferences, businesses can craft targeted strategies to ensure that these customers remain satisfied and continue to be beneficial for the organization in the long run. They have

higher average purchase value and tend to purchase items belonging to the grocery category, which can be beneficial when it comes to customer loyalty.

4.2 Technical Insights:

To gain insights, we analysed a dataset of customer information featuring the following attributes:

- **Age**: How old is the customer?
- **Average Purchase Value**: The mean spending amount of each customer.
- **Months Since Last Purchase**: The number of months since the customer's last purchase.
- **Monthly Purchase Frequency**: The average number of customers purchases each month.
- **Number of Visits**: The total count of visits the customer has made to the store.
- **Visit Share**: The percentage of store visits that can be attributed to customers.
- **Churned**: A binary code which shows whether or not the customer has chosen to discontinue their service (1) or to remain a customer (0).

To gain further insight into the dataset, an exploratory analysis was conducted, which included examining summary statistics and calculating correlation matrices. This helped in recognizing any existing patterns and relationships between variables.

To investigate customer churn, multiple machine-learning models were employed. Logistic regression and random forests were used for the study. The models were then judged based on their accuracy, precision, recall and F1 score. To predict customer churn, we used feature importance rankings from the random forest model to identify the key features. The analysis revealed that purchase value, frequency, months since the last purchase and visit share had the most influence on customer behaviour while age had almost no influence.

In conclusion, the data analysis and machine learning models by developing pen portraits of customer churners and non-churners. These profiles offer a detailed representation of the traits shared by these groups.

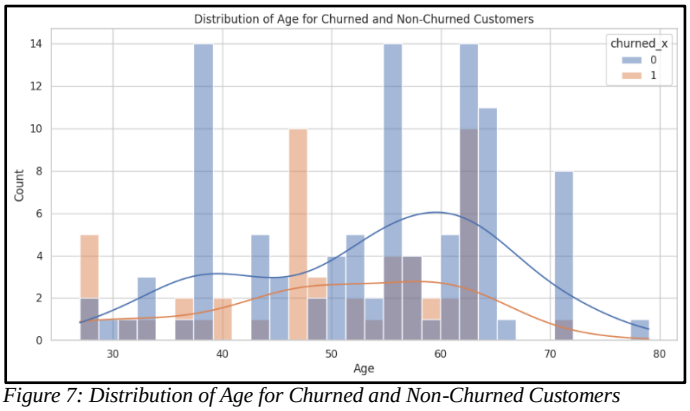

Figure 7: Distribution of Age for Churned and Non-Churned Customers

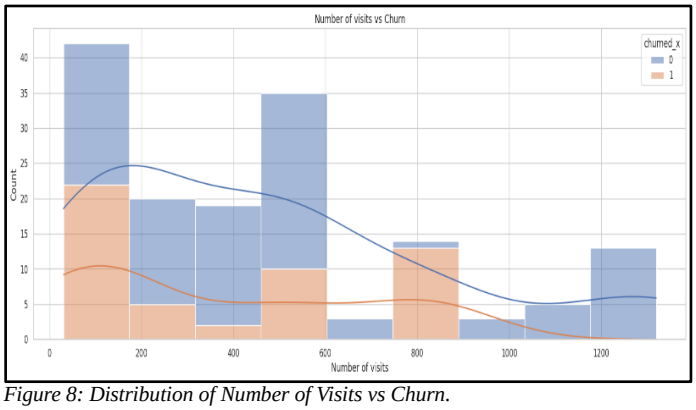

Figure 8: Distribution of Number of Visits vs Churn.

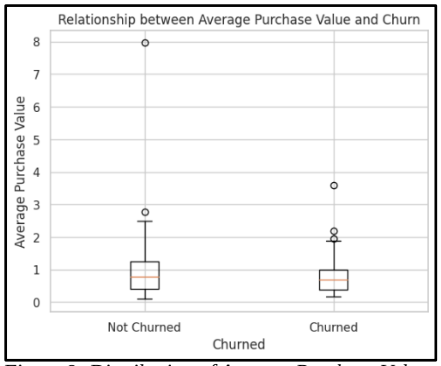

Figure 9: Distribution of Average Purchase Value

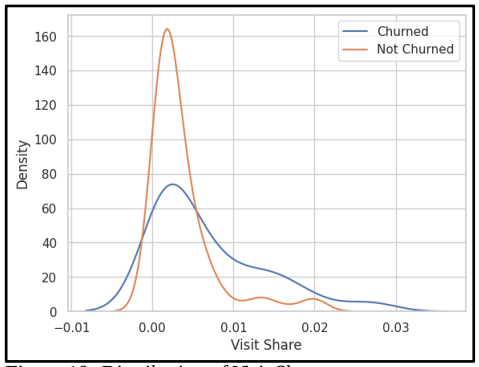

Figure 10: Distribution of Visit Share

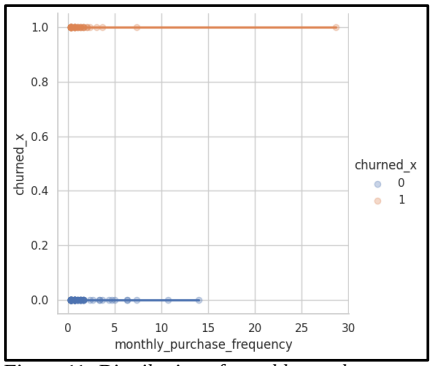

Figure 11: Distribution of monthly purchase