**BA723- Business Analytics Capstone**

**Customer Churn Analysis for Jade**

**Data Analysis and Modeling**

**Submitted By: Nimish Shaji Abraham**

**301312186**

**Advisor: Bilal Hassanzadah**

Executive Summary

The primary objective of this project was to analyze and predict customer churn for Jade Inc. using various machine learning models. The goal was to identify which customers are most likely to churn and to develop strategic recommendations for customer retention. Through extensive data analysis and modeling, we identified key patterns and drivers of customer churn, while also suggesting actionable steps to enhance customer retention. A range of models, including Random Forest, Neural Networks, Gradient Boosting Machines, and Support Vector Machines, were explored to identify the most accurate and useful model for predicting churn.

To begin, we conducted an exploratory data analysis (EDA) to uncover key trends in customer behavior. We identified important features such as tenure, monthly charges, and the number of products used by the customer, which played a critical role in customer churn. Additionally, our analysis revealed that customers who had longer relationships with Jade were less likely to churn, whereas customers who were offered frequent discounts and promotions showed higher satisfaction rates. These insights laid the foundation for model building and informed the subsequent strategies for customer retention.

Following the EDA, we experimented with various models to predict customer churn. Among the models tested, Random Forest emerged as the most accurate, achieving a 91% accuracy rate. Its ability to handle non-linear relationships between features and capture the complexity of customer behavior made it an ideal choice for predicting churn. The Random Forest model allowed us to identify the most significant factors contributing to churn, including tenure and discounts, which provided valuable insights into customer behavior and preferences.

In addition to Random Forest, we experimented with Neural Networks and Gradient Boosting Machines. While Gradient Boosting Machines provided good accuracy, Neural Networks offered an opportunity to uncover more complex relationships in the data. Neural Networks were able to analyze large datasets and find patterns that could have been missed by other models. Although the accuracy of the Neural Network model was slightly lower than Random Forest, it provided a deeper understanding of the underlying relationships between customer features and churn.

Based on the model outputs, we provided strategic recommendations to Jade Inc. for improving customer retention. The data suggested that Jade should focus on strengthening long-term customer relationships through loyalty programs and discounts. In particular, we recommended that Jade offer more discounts to retain satisfied customers and use personalized rewards to keep customers engaged. Additionally, we advised Jade to simplify its user interface, as many customers reported frustration with the current layout. A more streamlined interface could lead to improved user experience, reducing customer frustration and further lowering the risk of churn.

In conclusion, this project successfully identified the key drivers of customer churn and provided actionable insights for Jade Inc. to enhance customer retention. By leveraging machine learning models, particularly Random Forest, Jade can focus its retention efforts on at-risk customers and implement targeted strategies to improve customer satisfaction. Moreover, by continuously retraining the models with updated data, Jade can stay ahead of changing customer behaviors and market conditions, ensuring the long-term success of its retention strategy.

Contents

[1.1 Background 4](#_Toc174650148)

[1.2 Business Problem 5](#_Toc174650149)

[1.3 Objectives and Measurement 5](#_Toc174650150)

[1.4 Assumptions and Limitations 6](#_Toc174650151)

[2.1 Data Set Introduction 8](#_Toc174650152)

[2.2 Data Dictionary 8](#_Toc174650153)

[2.3 Exclusions 13](#_Toc174650154)

[3.1 Data Exploration and Data Cleaning Techniques 14](#_Toc174650155)

[3.2 Data Visualization 21](#_Toc174650156)

[4.1 Feature Engineering 40](#_Toc174650157)

[5.1 Decision Tree 45](#_Toc174650158)

[5.2 Random Forest 46](#_Toc174650159)

[5.3 Linear Regression 47](#_Toc174650160)

[5.4 Logistic Regression 47](#_Toc174650161)

[5.5 Support Vector Machine 50](#_Toc174650162)

[5.6 Gradient Boosting Machine 51](#_Toc174650163)

[5.7 K-Nearest Neighbours 52](#_Toc174650164)

[5.8 Neural Network 53](#_Toc174650165)

[5.9 Model Comparison 54](#_Toc174650166)

[5.9.1 ROC-AUC Chart 56](#_Toc174650167)

[5.9.2 Feature Importance 58](#_Toc174650168)

[6.1 Recommendations for implementation 61](#_Toc174650169)

[6.2 Recommendations for future use 62](#_Toc174650170)

1. Introduction
   1. Background

Customer retention or managing a loyal relationship with a customer is one of the most typical and most difficult problem that a company has. Many a times, companies fail to change their business strategies with the changing tastes and preferences of the customers and also fail to capture a whole new set of audience. Surely, the customers a company has at a particular time may be happy with their products and services, but as the times change, the customers also change, and it is very important that a company reduces the customer churn as much as possible. Most of the previous studies have been conducted for industries in which customers are tied with contracts (such as telecom or banking), which limits the churn rate but it is very different for an ecommerce company to do the same (Matuszelański & Kopczewska, 2022).

The first issue faced by the ecommerce platforms is that a major portion of their customer base churn and there is no way to hold the customers in a contract like telecom or banking industries do. This increases the percentage of churn and the only way to mitigate this is to predict the type of customers which are going to churn and be proactive in actions. Another major issue which ecommerce companies face are increasing competition and a more difficult challenge to build personal relationships with customers (Rudälv, 2022).

Jade Inc. is a leading ecommerce company which in the North American region and caters to many customers. Their product categories include technology, fashion, grocery and much more. Some of the companies which are in constant competition with Jade are Amazon and Ebay. Over the years, Jade has become one of the most trusted ecommerce platforms because of their exceptional quality standards and pricing options.

* 1. Business Problem

The majority of firms, especially those in the e-commerce industry, face the difficult problem of customer retention. Wu et al. (2017) state that while attracting new consumers requires large financial investments to cultivate a loyal client base, retaining current customers is far more challenging than attracting new ones due to the significant value that current customers bring to businesses.

While customer acquisition is important, is also costs more than maintaining customers. Keeping and satisfying existing customers is generally regarded as more profitable than constantly renewing the customer base (Reinartz & Kumar, 2003). Therefore, to be a successful business, it is important to not only attract new customers, but also to gain their trust and retain them. In other words, the focus must not solely be on customer acquisition, but also and primarily on customer retention.

Despite its success in the recent years, Jade is facing challenges in maintaining low levels of customer churn, a critical factor for fostering stakeholder trust and sustaining business growth. Our current analysis aims to identify key factors influencing customer churn using a comprehensive dataset, which includes variables such as customer tenure, preferred login devices, satisfaction scores, and purchase behaviors. By examining these factors, we will propose actionable insights to enhance service quality and customer experience, ensuring Jade Inc. continues to thrive in the competitive e-commerce landscape.

* 1. Objectives and Measurement

The primary objective of this project is to take advantage of the historical data, customer survey forms and customer feedback forms and come up with models and strategies which would help Jade to reduce their churn levels in the future. The reason for taking the following types of data is to understand the needs of the customers in the present and to come up with solutions so that we can modify our business strategies to boost our customer retention level. Other major objectives of the project is to analyse why our business strategies are not working, what is the reason we are not able to captivate the audience and keep them in our company and also to see whether our modeling techniques are accurate and how useful will it be in the future. Some of the main objectives of the study are listed below:

1. Understanding and building predictive models which can accurately predict whether a customer will churn or not.
2. Comparing multiple models and selecting the one which would give us the best results.
3. Comparing the features which contribute the most toward churn and coming up with implementation techniques which would help us reduce churn.
4. Listing recommendations for future studies which would enhance the current modeling strategies and give recommendations for actions that should be taken.

The metrics that would be used to measure the model’s performance are as follows:

1. F1-Score
2. Accuracy Score
3. ROC-AUC chart
   1. Assumptions and Limitations

As is true with any and every study, this study also assumes some factors and have limitations in it. The **assumptions** that were made in the study are as follows:

1. The sample size of the data set is taken from the past year, 2023, and the dataset is not complete and has a lack of information.
2. The customer surveys and feedback are true and to the point.
3. The data is not biased.

The **limitations** of the study are as follows:

1. There is a lack of older studies to build upon as this might be the first project that focuses on customer churn.
2. The dataset is small, and the model needs much more data to give an accurate prediction
3. Time limitations restricted the scope of the study.

1. Data Sources
   1. Data Set Introduction

The current dataset includes information from historical data of Jade, customer surveys and feedback forms, all of which were obtained from Kaggle. The dataset had 5,630 observations in total and 20 unique variables. Few of the major variables include Tenure, Churn, PreferredLoginDevices, WarehouseToHome, SatisfactionScore, Complain and many more. These variables and observations helped the analysis to build models which would efficiently predict churn and also provide information about which of the variables have the strongest relationship with churn. Certain implementation techniques were also built based on these observations.

* 1. Data Dictionary

As mentioned above, this dataset consists of 5,500 observations and 20 unique variables. Though there were some cleaning methods used in order better the model, it is essential to note that the current data dictionary provided is the one before the data cleaning methods were used in order to show the full picture of what was done and be transparent about the data that was provided.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Description | Data Type | Values |
| CustomerID | Unique customer ID | Numerical | * Numerical Values of the customer id |
| Churn | Whether a customer churned or not | Categorical | * Churn: Yes * Churn: No |
| Tenure | Tenure of customer in organization | Numerical | Number of Years a customer has been in the company for. Ranging from 6 months to 5 years. |
| PreferredLoginDevice | The preferred login device of the customer | Categorical | * MobilePhone * Computer * Phone: Relates to customers placing phone call orders |
| CityTier | Ranking given to a city. The lower the number the more modern the city, | Categorical | * 3: Low Tier city * 2: Medium Tier City * 1: High tier city |
| WareHouseToHome | Distance between the warehouse and the customers’ home | Numerical | Containing distance in kilometers all the way from 5 to 127. |
| PreferredPaymentMode | Preferred payment method of customer | Categorical | * Debit Card * Credit Card * Cash on Delivery * CC (Cash Credit) * COD (Credit on Delivery) * E-Wallets (Paypal) * UPI (Apple Pay) |
| Gender | Gender of customer | Categorical | * Male * Female |
| HourSpendOnApp | Total Number of Hours spent on the app | Numerical | * 1: Used the app for 1 hour * 2: Used the app for 2 hours * 3: Used the app for 3 hours |
| NumberOfDeviceRegistered | Total Number of Devices registed on Jade | Numerical | 1 to 6. Each number lists the number of devices registered. |
| PreferredOrderCategory | The preferred category which customers like to buy from | Categorical | * Fashion * Laptop and Mobile Accessory * Mobilephones * Grocery * Others |
| Satisfaction Score | Is the customer satisfied with us or not | Numerical | Satisfaction values from 1 to 5. 1 being the least satisfied and 5 being the most satisfied. |
| Marital Status | Provides information on the customers’ marital status | Categorical | * ‘No’: Customers have not churned * ‘Yes’: Customers have churned |
| NumberOfAddress | Provides information on the number of addresses a customer has | Numerical | Has values from 1 to all the way to 22. |
| Complain | Provides information on whether a complain has been raised by the customer or not | Categorical | * 0: No * 1: Yes |
| OrderAmountHikeFromlastYear | Provides information on the order amount hike from last year in percentage | Numerical | Consists data from 11% to 26% |
| CouponUsed | Total number of coupon used by the customer in the last month | Numerical | Consists numbers from 0 to 16 |
| OrderCount | Total number of orders being placed in the last month | Numerical | Consists data from 1 all the way to 16 |
| DaySinceLastOrder | Day Since last order by customer | Numerical | Contains values from 0 days to 46 |
| CashbackAmount | Average cashback amount received in the last month | Numerical | Contains values from 0 all the way to 324.99. |

* 1. Exclusions

In order to help the model to run properly and to reduce the dimensions of the model so that the accuracy rate of the models improve, some irrelevant variables should be removed, so that the scope of the study is much clearer and can give the best possible accuracy rate.

For the current project, the researchers used a similar technique and removed one of the rows which was found to be irrelevant. The row which was removed from the original data set to improve visualisation and modeling was **CustomerID.**

1. Data Exploration  
   1. Data Exploration and Data Cleaning Techniques

In order to better understand the data, the researchers always use data exploration techniques so that they have a better view of what the data is about, how skewed it is, how many missing variables it has and a lot more. This enables the researchers to build predictive models which are best suited for the dataset they are using.

Similarly, there were multiple data exploration and data cleaning techniques which were used to improve the dataset.

**Conversion of Relevant Variables into Categorical Type**

Several variables in the dataset were represented as object and int data types where, the project would benefit if they were converted to categorical variables. This would help the researchers to improve the model’s predictive power and help to specify what type of prediction does this current study need. The variables that were converted to categories include:

1. Churn
2. Complain
3. Marital Status
4. PreferredLoginDevice

**Imputation of Missing Variables**

A screenshot of a computer

Description automatically generatedThe historical dataset and the dataset generated from the customer surveys had a lot of missing values which are shown as below:

To tackle the missing values, they were imputed with their respective rows’ median values. The researcher did not drop the missing values as firstly, it would not give the accurate information and secondly, it brought down the accuracy rate of the best model. After the imputation, the missing values in the dataset were as follows:

A screenshot of a computer

Description automatically generated

**Outlier Detection**

The dataset which was used for this project was sensitive to outliers and had some

values which were beyond the interquartile range. This was an important step to do as

if the outliers were not treated, the model would have been very skewed and would

have not let the model to give accurate predictions.

A group of blue boxes

Description automatically generated

In numerical terms:

A screenshot of a computer

Description automatically generated

**Removal of Outliers**

Now that we have detected the outliers present in all of the rows, it was time to remove

Them. In this current project, we used cap and floor to set an upper and lower limit

based on the interquartile range. The range that was built was as follows:

A screenshot of a computer

Description automatically generated

This helped in removing the outliers from the dataset and helped the researcher

to improve the model’s accuracy.

**Data Description**

After removing the outliers from the dataset, the descriptive statistics of it were as

follows:

A screenshot of a computer

Description automatically generated

**Skewness Detection and Removal**

The dataset which we received for our research had skews in it. Some variables were positively skewed and some were negatively skewed:

A screenshot of a computer program

Description automatically generated

**Log Transformation**

The skewness chart shows that there are some columns which are skewed. Some are slightly skewed but others are highly skewed. I'll be focusing on removing the skews from the moderate to highly skewed variables.

Also, since most of the the variables are positively skewed, I'll be applying log transformation to address the right skewness.

**Results of the Log Transformation**

**A screenshot of a computer

Description automatically generated**

Some of the variance has been removed from the variables and some positive skewness have been reduced. However, some of the variables which had lower positive skews went over to the negative side and because of that, a root transformation has to be applied.

**Results of the root transformation**

**A screenshot of a computer program

Description automatically generated**

Since the data is not severely skewed, I will be keeping these values as the final ones.

The reason for including log transformations in the data cleaning process is to have a perfect visualisation of the dataset available to us.

Explanation for removing uncapped columns and skewed columns- The researcher removed the initial columns which were raw as a part of the data cleaning process. The researcher did this because if they were to keep the unclean variables, they were getting a worse accuracy score for their best model. They were getting an F1-Score of 97% for their Random Forest when keeping the unclean variables whereas, if they removed them, it improved their score to 99%.

For visualisations, they used the clean data as well as it was free of errors and anomalies and gave them a better analysis of the data.

**Removal Of Duplicates**

The dataset used had a lot of duplicates in them. They were removed as they did not provide any additional information which would improve the models’ accuracy rate.

**A screen shot of a computer

Description automatically generated**

This was done in order to remove the duplicates that were present in the dataset.

* 1. Data Visualization

With the cleaned data now available to the researcher, it is now essential to view how the data is distributed. This would require the researcher to make use of data visualization techniques so that they are able to see the distribution and make assumptions and come up with modeling techniques which would suit this type of dataset.

A lot of data visualization techniques were used in the project and every visualization was made in Python. For this project, both univariate and multivariate analysis were done because the researcher thought it was relevant. This gave the researcher an idea about the relationship of different variables with churn and how they affect the rate of churn of a customer. The results of all the analysis are given below:

**Univariate Analysis**

A univariate analysis technique was used in this project to get to know about the customers of Jade and see the distribution of data. This was helpful for the researcher as it gave an idea to them about who the customers are, their tastes and preferences and helped with connecting the dots during the modeling process.

**3.2.1 Gender**

A blue and red circle with a number of percentages

Description automatically generated

Our dataset has around 40% of females and 60% of males.

* + 1. **Distribution of Churn**

A blue and red rectangular bar graph

Description automatically generated

After removing all the outliers and duplicates, our total churn population out of 5072 observations are 841 people.

**Bivariate Analysis**

A bivariate analysis was important in this project to know about the relationship between various variables and to know how they affect churn in the end. The researcher conducted lots of bivariate analyses which are as follows:

* + 1. **Correlation Matrix**

A screenshot of a graph

Description automatically generated

Interpretations:

1. **Tenure and Churn**: I've identified a moderate negative correlation (-0.33) between Tenure and Churn. This is our most significant lever for reducing churn. Customers who stay with us longer are less likely to leave, so I recommend we prioritize strategies to increase customer tenure.
2. **Cashback's Impact**: I've noticed that CashbackAmount has a moderate positive correlation with Tenure (0.46) and a slight negative correlation with Churn (-0.14). This suggests our cashback program could be an effective tool in reducing churn by encouraging longer tenure.
3. **Order Frequency**: I see that OrderCount has a slight negative correlation with Churn (-0.02). While small, this indicates that customers who order more frequently are somewhat less likely to churn. Encouraging repeat purchases could help reduce churn.
4. **Coupon Usage**: Interestingly, CouponUsed shows almost no correlation with Churn (-0.00). However, it's strongly correlated with OrderCount (0.67), which in turn slightly reduces churn. This suggests that while coupons don't directly impact churn, they could indirectly help by increasing order frequency.
5. **Satisfaction and Churn**: I've observed a weak positive correlation (0.10) between SatisfactionScore and Churn. This counterintuitive result suggests we may need to reassess how we measure customer satisfaction to ensure it aligns with reducing churn.
6. **Complaints and Churn**: There's a weak positive correlation (0.25) between Complain and Churn. This indicates that addressing customer complaints promptly and effectively could be crucial in our churn reduction efforts.
7. **App Engagement**: HourSpendOnApp shows a very weak positive correlation (0.01) with Churn. This suggests that while app engagement isn't strongly linked to churn, there might be an opportunity to use the app more effectively for retention.
   * 1. A group of graphs with numbers and text

        Description automatically generated with medium confidence**Bar Charts for Categorical Variables**

Interpretations:

1. When analyzing the **Preferred Login Device vs. Churn** graph, I notice that a significant number of users who prefer to log in using a mobile phone tend not to churn. Specifically, there are 2,190 users who continue to use the service without churning, compared to 324 who churn. In contrast, for users who prefer logging in via computer, 1,172 do not churn, while 285 do. This indicates a higher retention rate among mobile phone users. Interestingly, no users who prefer logging in via a computer churn, which could suggest high satisfaction within this group.
2. Looking at the **Preferred Payment Mode vs. Churn** graph, I observe that users preferring to pay with a debit card show the highest number of non-churners, with 1,783 not churning and 315 churning. Credit card users follow, with 1,174 non-churners and 174 churners. On the other hand, users who prefer cash on delivery exhibit the lowest retention rate, with only 113 not churning and 22 churning. This suggests that electronic payment methods, particularly debit cards, are associated with higher retention rates among users.
3. From the **Gender vs. Churn** graph, it is evident that male users are more likely to churn compared to female users. Specifically, 2,515 males do not churn, whereas 532 males churn. In comparison, 1,717 females do not churn, and only 309 females churn. This indicates a higher churn rate among male users, suggesting that females are generally more satisfied or engaged with the service.
4. I notice that the "**Laptop & Accessory**" category has the highest number of customers, with 1653 non-churned and 194 churned users. The "Mobile phone" category follows as the second most popular, showing 831 non-churned and 299 churned customers. Interestingly, the "Fashion" category has a notable difference between non-churned 639 and churned 124 customers, suggesting it might be an area where customer retention is relatively strong.
5. Looking at the "**MaritalStatus vs Churn**" graph, I observe that married customers form the largest group, with 2370 non-churned and 302 churned individuals. Single customers come next, showing 1138 non-churned and 415 churned customers. The divorced category is the smallest, with 724 non-churned customers and 124 churned ones. It's worth noting that the churn rate appears to be proportionally higher among single customers compared to married ones, which could be an interesting point for further analysis.

Therefore, customers who prefer logging in via mobile phones and those using debit cards for payments show higher retention rates, indicating their satisfaction with the service. Gender-wise, female users tend to stay with the service longer than male users. When looking at preferred order categories, "Laptop & Accessory" and "Mobile phone" have the highest customer base, with the "Fashion" category showing strong retention. Marital status also plays a role in churn, with married customers having a lower churn rate compared to single ones. These patterns suggest that targeted strategies to enhance user experience for mobile login users, electronic payment users, and single customers could effectively reduce churn rates.

* + 1. A group of blue and green rectangular shapes

       Description automatically generated**Bar Charts for Numerical Variables**

**Interpretations:**

1. **Mean Tenure by Churn:** The graph indicates that customers who have not churned exhibit a substantially higher mean tenure compared to those who have churned. This suggests that customers with longer tenures are more likely to remain with the service, highlighting the importance of customer retention efforts over time.
2. **Mean CityTier by Churn:** The mean CityTier for both churned and non-churned customers is nearly identical, indicating that the geographical tier of a city does not significantly influence customer churn. This suggests that other factors beyond geographic segmentation might be more critical in understanding customer retention.
3. **Mean WarehouseToHome by Churn:** The average distance from the warehouse to the customer's home is very similar for both churned and non-churned customers. This indicates that logistical factors, such as the proximity of the warehouse to the customer's residence, do not play a significant role in customer churn.
4. **Mean HourSpendOnApp by Churn:** The mean number of hours spent on the app is almost the same for both churned and non-churned customers, suggesting that the duration of app usage does not have a strong correlation with customer churn. This implies that app engagement metrics may need to be examined in conjunction with other variables to understand churn better.
5. **Mean NumberOfDeviceRegistered by Churn:** There is no substantial difference in the mean number of devices registered between churned and non-churned customers, indicating that the number of devices a customer registers does not significantly impact their likelihood of churning.
6. **Mean SatisfactionScore by Churn:** The satisfaction score is fairly consistent between both groups, implying that overall satisfaction levels, as measured, do not significantly differentiate between customers who churn and those who remain. This suggests that other aspects of customer experience might be influencing churn more strongly.
7. **Mean NumberOfAddress by Churn:** The mean number of addresses maintained by customers is similar regardless of their churn status, indicating that the number of addresses associated with a customer's account is not a significant factor in predicting churn.
8. **Mean Complain by Churn:** The data reveals that churned customers have a higher mean number of complaints compared to non-churned customers. This suggests a strong correlation between customer complaints and churn, indicating that addressing customer grievances promptly and effectively could be critical in reducing churn rates.
9. **Mean OrderAmountHikeFromlastYear by Churn:** The average increase in order amount from the previous year is similar for both groups. This indicates that changes in order value over time do not significantly influence whether customers decide to churn.
10. **Mean CouponUsed by Churn:** The mean number of coupons used is nearly the same for both churned and non-churned customers. This suggests that the usage of promotional coupons does not have a strong impact on customer retention or churn.
11. **Mean OrderCount by Churn**: The mean order count is comparable between both groups, indicating that the total number of orders placed does not significantly influence customer churn. This implies that frequency of orders alone may not be sufficient to predict churn behavior.
12. **Mean DaySinceLastOrder by Churn:** Customers who have churned have a lower mean number of days since their last order compared to those who have not churned. This might indicate that recent purchasing activity is not necessarily a deterrent to churn, and other factors such as customer satisfaction or competitive offerings might be at play.
13. **Mean CashbackAmount by Churn:** The mean cashback amount received by churned customers is slightly lower than that of non-churned customers, suggesting that cashback incentives may have a minor but notable impact on customer retention. This indicates that financial incentives could play a role in customer loyalty strategies.
    * 1. A group of graphs showing different sizes and shapes

         Description automatically generated with medium confidence**Histogram for Continuous Variables**

**Interpretations:**

1. The **distribution of tenure** indicates that the majority of customers have a tenure of fewer than 10 years, with a steep decline in frequency as tenure increases. This suggests that the company has a relatively young customer base with a high turnover rate in the initial years, while a smaller segment of loyal customers maintains a longer relationship with the company.
2. The **distribution of the WarehouseToHome** variable shows a concentration of customers within a short distance from the warehouse, primarily within 20 units of distance. There are a few outliers at much greater distances, but the majority of customers reside close to the warehouse, which could imply a regional business focus or an efficient local delivery system.
3. The **distribution of HourSpendOnApp** is characterized by several distinct peaks, indicating that customers tend to spend their time on the app in specific durations, notably around 2, 3, and 4 hours. These peaks suggest common usage patterns or behaviors among customers, which might be tied to specific app features or services that engage users for set periods.
4. The **distribution of OrderAmountHikeFromLastYear** reveals that most customers have experienced a moderate increase in order amounts, predominantly between 12 and 18 units. The frequency decreases steadily as the order amount hike increases, indicating that significant hikes are less common and the majority of customers see incremental increases in their order amounts year over year.
5. The **distribution of CouponUsed** shows that the majority of customers use between 0 to 2 coupons, with a sharp decline in frequency as the number of coupons used increases. This suggests that while coupon usage is prevalent, extensive use of coupons is rare, indicating either limited availability or lower perceived value of additional coupons.
6. The **distribution of OrderCount** is heavily skewed towards lower counts, with the majority of customers placing between 1 to 4 orders. The frequency sharply declines as the order count increases, highlighting that while occasional orders are common, frequent ordering is much less typical among customers.
7. The **distribution of DaySinceLastOrder** suggests that most customers have placed an order within the last 10 days, with frequency sharply decreasing as the days since the last order increase. This pattern indicates a high level of recent engagement, although there is a long tail of customers who have not ordered in a significant amount of time.
8. The **distribution of CashbackAmount** shows a concentration around 100 to 150 units, with the frequency tapering off as the cashback amount increases. This indicates that while a moderate amount of cashback is common, higher cashback amounts are less frequent, possibly reflecting company policy on cashback rewards or customer spending patterns.
   * 1. A group of blue squares

        Description automatically generated**Distribution Bar Charts**

**Interpretations:**

1. In analyzing the **distribution of churn**, I observe that the majority of customers have not churned, indicating a relatively stable customer base. However, a significant minority has churned, which suggests there are areas for improvement in customer retention strategies.
2. The **distribution of internet service** types reveals that the majority of customers use fiber optic services, followed by DSL, with a smaller portion using no internet service. This highlights the preference for high-speed internet among customers.
3. Examining the **payment methods**, I see that electronic check is the most common, followed by mailed check, bank transfer, and credit card. This suggests a diverse range of payment preferences among customers, with a notable inclination towards electronic transactions.
4. The **contract distribution** shows that month-to-month contracts are the most prevalent, followed by one-year and two-year contracts. This indicates a preference for flexibility among customers, with fewer committing to long-term contracts.
5. The **gender distribution** is fairly balanced, with a slight majority of male customers. This balance suggests that the services offered appeal equally to both genders, without significant bias.
6. In **the partner distribution**, a larger portion of customers do not have partners compared to those who do. This could imply that the services are particularly appealing to single individuals or those without partners.
7. The **distribution of dependents** shows that the majority of customers do not have dependents. This might indicate that the services are more attractive to individuals or couples without children.
8. The **marital status distribution** reveals that a significant portion of customers are married, followed by those who are single, divorced, or widowed. This suggests that the services cater well to a diverse range of marital statuses, with a strong appeal to married individuals.
   * 1. **Correlation between Numeric and Categorical Variables using ANOVA**

**A screenshot of a computer

Description automatically generated**

1. **Tenure**: The ANOVA results indicate a highly significant effect of Tenure on Churn, with an F-Statistic of 618.95 and an extremely low P-Value (5.22e-129). This suggests that the duration of a customer's engagement with the service strongly influences their likelihood to churn.
2. **CityTier**: The analysis shows that CityTier has a significant impact on Churn, evidenced by an F-Statistic of 49.28 and a P-Value of (2.51e-12). This suggests that customers from different city tiers exhibit varying propensities to churn.
3. **WarehouseToHome**: The significant F-Statistic of 24.17 and a low P-Value (9.11e-07) for WarehouseToHome indicate that the distance between the warehouse and the customer's home significantly affects their likelihood of churning.
4. **HourSpendOnApp**: The results for HourSpendOnApp show an F-Statistic of 0.91 and a P-Value of 0.34, suggesting that the amount of time a customer spends on the app does not have a significant impact on their likelihood to churn.
5. **NumberOfDeviceRegistered**: With an F-Statistic of 69.94 and a very low P-Value (7.81e-17), NumberOfDeviceRegistered is found to have a strong association with Churn, indicating that the number of devices registered by a customer is a significant predictor of churn.
6. **SatisfactionScore**: The analysis reveals a significant relationship between SatisfactionScore and Churn, with an F-Statistic of 52.28 and a P-Value of (5.54e-13). This highlights that customer satisfaction scores are a strong determinant of churn likelihood.
7. **NumberOfAddress**: The F-Statistic of 13.76 and a P-Value of (2.10e-04) indicate a significant effect of NumberOfAddress on Churn, suggesting that the number of addresses associated with a customer influences their propensity to churn.
8. **Complain**: The extremely high F-Statistic of 334.32 and a very low P-Value (2.23e-72) suggest that the number of complaints is a highly significant predictor of Churn, indicating a strong relationship between the frequency of complaints and the likelihood of churning.
9. **OrderAmountHikeFromlastYear**: The F-Statistic of 2.86 and a P-Value of 0.09 suggest that OrderAmountHikeFromlastYear does not have a significant effect on Churn, indicating that changes in order amounts over the past year are not a strong predictor of churn.
10. **CouponUsed**: The analysis shows an F-Statistic of 0.09 and a P-Value of 0.76, indicating that the use of coupons does not significantly affect Churn. This suggests that coupon usage is not a major factor influencing customer churn.
11. **OrderCount**: With an F-Statistic of 2.26 and a P-Value of 0.13, OrderCount does not have a significant effect on Churn, suggesting that the number of orders placed by a customer is not a strong predictor of their likelihood to churn.
12. **DaySinceLastOrder**: The high F-Statistic of 111.46 and a very low P-Value (8.69e-26) indicate a significant effect of DaySinceLastOrder on Churn, showing that the time elapsed since a customer's last order is a strong predictor of their likelihood to churn.
13. **CashbackAmount**: The significant F-Statistic of 105.95 and a low P-Value (1.32e-24) reveal that CashbackAmount has a strong impact on Churn, indicating that the amount of cashback received by customers is a significant factor in predicting churn.
    * 1. A screenshot of a computer

         Description automatically generated**PCA Analysis**

A graph with a red line and blue line

Description automatically generated

The PCA analysis has provided us with the variance explained by each principal component, which helps in understanding the dimensionality reduction of our dataset. The first principal component (PC1) explains approximately 18.96% of the variance, while the second principal component (PC2) accounts for about 10.36%. The cumulative explained variance ratio reveals that PC1 and PC2 together capture around 29.32% of the total variance. As we include more components, the cumulative variance increases, with all 13 components collectively accounting for 100% of the variance.

1. Data Preparation and Feature Engineering  
   1. Feature Engineering

Now that the dataset has been cleaned and visualized in a way which would help the researcher to make relevant models, it is time to do feature engineering. Just like data cleaning where the researcher changed data and made changes wherever they felt necessary, feature engineering does the somewhat the same thing but in a more complex manner.

Similar to the data pre-processing phase, several steps in the feature engineering process have been carried out on Python. By doing this, Jade’s customer data would be prepared for the crucial step of applying predictive analytics to the same data.

NOTE: The transformation of the data on the basis of log transformation and root transformation has been done in the data cleaning process which further enhanced our findings for the data visualisation. Please refer to the data cleaning process in order to see more about that feature engineering process.

* + 1. **Specifying the target and predictor variables**

To prepare our data for modeling, it is essential that we split the data into two parts: Target and Predictor. The reason we do this is to let the model know what we are predicting and how are we going to predict it. In this project, our goal is to predict the customer churn. So, our target variable was **Churn** and all the other variables that were present after the data cleaning process are the predictor variables.

* + 1. **Dummy Variable Creation**

After we are done with transforming our variables and specifying our target and predictor variables, we create the dummy variables. Creation of dummy variables I important to reduce the dimension of the dataset. This makes it easier for the model to predict our target and also protects from the curse of dimensionality. The Curse of Dimensionality refers to the various challenges and complications that arise when analyzing and organizing data in high-dimensional spaces (often hundreds or thousands of dimensions) ((Awan, 2023).

In this project dummy variables were created for Gender which was converted to Gender\_Male and the values in it were 0 and 1 where 0 means Yes and 1 means no. Since there were no other variables which could be binary classified, except for Churn which was already in 0’s and 1’s, we only created the dummy variables for Gender.

* + 1. **Standardisation and Normalisation**

Standardization in the context of Jade’s feature engineering component for customer churn prediction is the process of modifying numerical features to have a mean of 0 and a standard deviation of 1, making comparisons simpler. Standardizing numerical features improves predictive models' performance and accuracy, producing more trustworthy business insights.

* + 1. **Train-Test Split**

Partitioning the customer dataset into two subsets—one for training and the other for testing the predictive models—is the final stage in the feature engineering process of predicting customer attrition at Jade. The testing data assesses the prediction models' performance on fresh, unknown data, while the training data is used to build and optimize them. Each prediction model thus performs admirably in practical applications, resulting in wise business judgments. Jade’s customer dataset was split into two parts for this project: 70% was utilized for training and 30% was used for testing. To guarantee that the prediction models are trained with enough data while reserving enough unseen data, the 70-30 percent ratio is employed.

1. Model Exploration

The researcher, in order to predict churn, has created a lot of models which would enable them to come up with solutions that would help Jade reduce their churn levels. The modeling techniques used in this project are Decision Trees, Random Forest, Linear Regression, Logistic Regression (Full, Forward, Backward and Stepwise), Support Vector Machine, Gradient Boosting Machine, K-Nearest Neighbours and Neural Networks. Using many kinds of models helped the researcher to come up with solutions that would give them a clear understanding of which modeling process they should use.

In order to compare the models, there were a couple of metrics that were used namely, ROC-AUC, F1 Score and Accuracy. The detailed explanation of what each of these metrics are given below:

1. **ROC-AUC** **Curve-** The ROC-AUC curve, or Receiver Operating Characteristic - Area Under the Curve, is a graphical representation that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The ROC curve plots two parameters:

**True Positive Rate (TPR):** Also known as sensitivity or recall, it represents the proportion of actual positives that are correctly identified.

**False Positive Rate (FPR):** It represents the proportion of incorrect negatives identified as positives.

The AUC (Area Under the Curve) value measures the classifier's ability to distinguish between positive and negative classes. A model with an AUC score of 1 indicates perfect classification, while a score of 0.5 suggests random guessing.

**How It Helps:** The ROC-AUC score provides a single value summarizing the model's overall performance across all classification thresholds. It is beneficial when the class distribution is imbalanced because it evaluates how well the model distinguishes between the two classes, regardless of the decision threshold.

1. **F1 Score-** The F1 Score is the harmonic mean of precision and recall. The harmonic mean is a type of average that is calculated by dividing the number of observations by the reciprocal of each value. In the context of F1 Score, it is used to balance precision and recall, giving more weight to lower values. It is calculated as:

A black background with white text

Description automatically generated

Where:

* Precision is the ratio of correctly predicted positive observations to the total predicted positives.
* Recall (Sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class.

The F1 score balances the trade-off between precision and recall. It is beneficial when dealing with imbalanced datasets where focusing on accuracy alone might be misleading.

**How It Helps:** The F1 score is valuable when there is a need to balance false positives and false negatives. It is a more suitable metric than accuracy when the positive class is rare or when the cost of false positives and false negatives significantly differs. For instance, in a customer churn prediction model, the F1 score helps ensure that we do not just correctly predict churn but also minimize false predictions.

1. **Accuracy-** Accuracy is the ratio of correctly predicted observations to the total observations. It is calculated as:

A black background with white text

Description automatically generated

Where:

* **TP (True Positives):** The number of correctly predicted positive cases.
* **TN (True Negatives):** The number of correctly predicted negative cases.
* **FP (False Positives):** The number of incorrect positive predictions.
* **FN (False Negatives):** The number of incorrect negative predictions.

**How It Helps:** Accuracy is the most intuitive performance measure and is useful when the classes in the dataset are approximately balanced. However, in cases of class imbalance (e.g., where the churn rate is low), accuracy can be misleading. A model that always predicts the majority class can still achieve high accuracy but fail to identify the minority class correctly. This is why accuracy should be complemented with other metrics like ROC-AUC and F1 score to get a full picture of model performance.

* 1. Decision Tree

For both classification and regression tasks, a supervised learning approach called a decision tree is employed. It is a decision tree model that shows potential outcomes, resource costs, and utility of various choices. The decision rules are represented by the branches formed by the tree's division of the data according to feature values at different decision nodes (Seum, 2023). The last nodes, referred to as leaf nodes, stand for the continuous value or projected class in regression. Although decision trees are simple to use and intuitive, they are susceptible to overfitting, particularly when dealing with tiny datasets.

Recursively dividing the dataset into smaller subsets according to feature values is how Decision Trees operate. The splits that result in the highest information gain or Gini impurity reduction are then chosen.

We ran different decision tree models with different features. Out of those them, the full decision tree which was found to be the best was the decision tree 1 with the following parameters:

**Best parameters found: {'criterion': 'entropy', 'max\_depth': 30, 'max\_features': None, 'min\_samples\_leaf': 2, 'min\_samples\_split': 5}**

The Accuracy score and F1-Score was:

A screenshot of a computer

Description automatically generated

The accuracy score which the decision tree received was 0.94 or 94%. This means that the decision tree accurately predicts the correct target variable 94% of the time. This does not mean that the model is really good as the **94%** also includes it predicting No churn which is not our target for this project.

The F1-Score for predicting churn is 0.79 or **79%.** This says that the model is predicting churn correctly 79% time, which is good.

* 1. Random Forest

To create a forecast that is more reliable and accurate, a Random Forest ensemble learning technique generates several decision trees during training. Predictions are made by averaging the findings (in regression) or selecting the majority vote (in classification), with each tree in the forest being trained on a random subset of the data. This methodology improves robustness and lessens the overfitting issue that is frequently linked to individual decision trees.

The model becomes less sensitive to noise and more generalizable to data that hasn't been seen before thanks to Random Forest's introduction of randomness in feature selection and data sampling. It is highly accurate, adaptable, and resistant to overfitting, which makes it commonly utilized.

The best parameters found for the random forest are:

**Best parameters found: {'bootstrap': False, 'max\_depth': 30, 'max\_features': 'sqrt', 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300}**

The results of the Random Forest are as follows:

A screenshot of a computer screen

Description automatically generated

The accuracy score that random forest received was or 0.97 or **97%.** This means that the model was able to predict the target variable correctly 97% of the time.

The F1 Score for predicting churn was 91% which says that the model correctly predicted churned customers 91% of the time.

* 1. Linear Regression

A basic statistical technique called linear regression is used to forecast a continuous target variable based on one or more input factors. The connection between the independent variables (features) and the dependent variable (target) is assumed to be linear. The model minimizes the total of the squared discrepancies between the actual and predicted values (referred to as the residuals) in an attempt to determine the best-fitting line.

The accuracy for a linear regression model is measured by Root-Mean Squared Error and Mean Squared Error. The results of both measurement techniques are as follows:



The linear regression model's performance is quantified by a Mean Squared Error (MSE) of 0.0845 and an R² score of 0.3287. The MSE indicates that, on average, the squared differences between the predicted and actual values are 0.0845, suggesting moderate prediction accuracy. The R² score reveals that the model explains approximately 32.87% of the variance in the target variable, leaving 67.13% of the variance unexplained by the model. This indicates that while the model captures some of the underlying patterns in the data, it is not particularly strong and may benefit from additional features or more complex modeling techniques to improve its predictive power.

* 1. Logistic Regression

A classification method called logistic regression is used to forecast binary results depending on one or more predictor factors. Using the logistic function, which yields values between 0 and 1, it calculates the likelihood that a given input belongs to a particular class. By using a linear combination of the predictor variables, Logistic Regression models the log-odds of the outcome.

**Full Logistic Regression** models the association with the target variable by utilizing all of the features that are available. The classification report is as follows:

A screenshot of a computer screen

Description automatically generated

With a high accuracy of 91%, the full classification report demonstrates the model's overall performance. The model achieves good accuracy (0.93), recall (0.96), and F1-score (0.95) for the majority class (Class 0), meaning that it correctly detects most true positives with a low number of false positives. The model performs worse for the minority class (Class 1), with an F1-score of 0.67, a recall of 0.61, and a precision of 0.75. This suggests that although the model performs well overall, it has more difficulty accurately recognizing every instance of the minority class (churn).

**Forward Logistic Regression** approach begins with the model empty and gradually adds variables, choosing the most important variable at each stage until no additional meaningful improvement is seen. The classification report is as follows:

**A screenshot of a computer

Description automatically generated**

After using forward selection, the classification report reveals that while the model's performance on the minority class (Class 1) marginally decreases, its total accuracy stays at 91%. The F1-score falls to 0.63 as precision for Class 1 lowers to 0.74 and recall declines to 0.55. The reduced recall for Class 1 indicates that the forward selection process may have resulted in the exclusion of some features that are crucial for identifying the minority class (churn), making the model less sensitive to those instances. However, the majority class (Class 0) continues to perform very well, with a high recall of 0.97.

**Backward Logistic Regression** method starts with all of the model's candidate variables and gradually eliminates the least important ones until only the important ones are left. The classification report for backward logistic regression is as follows:

**A screenshot of a computer

Description automatically generated**

Similar to the full and forward selection models, the backward selection classification report demonstrates a model accuracy of 91%. With a high recall of 0.97 and an F1-score of 0.95, the performance of the majority class (Class 0) (No Churn) is still strong. With a recall of 0.58 and an F1-score of 0.66 for the minority class (Class 1), the backward selection method slightly outperforms forward selection. This implies that although backward selection may be able to identify the minority class (Churn) more correctly than forward selection by retaining some crucial properties, it is still not able to match the accuracy of the original model.

**Stepwise Logistic Regression** method combines forward and backward selection, allowing for the addition and removal of variables from the model at each step. Variables are added or deleted iteratively based on their statistical significance.

**A screenshot of a computer screen

Description automatically generated**

The model has an approximate overall accuracy of 91.5%. Class 0 (No Churn) has better precision (0.94), recall (0.97), and F1-score (0.95) than class 1 (Churn), indicating that the model works exceptionally well for this class. Class 1 performance, on the other hand, is poorer, with a precision of 0.77, recall of 0.61, and F1-score of 0.68, indicating that the model has more difficulty accurately detecting instances of this class. The weighted average F1-score is 0.91 and the macro average F1-score is 0.82, respectively, indicating that the majority class (0) has a greater overall influence on performance.

* 1. Support Vector Machine

One popular supervised learning approach for classification tasks is the Support Vector Machine (SVM), which can also be used for regression. The main goal of SVM is to locate the hyperplane in a high-dimensional space that best divides the data points of various classes. The objective is to produce the strongest possible classification borders by maximizing the margin between the closest points of the two classes, or support vectors.  
  
When applying kernel methods, SVMs may effectively process data that is both linearly separable and non-linearly separable in high-dimensional domains. Nonetheless, adjusting is crucial since they are highly dependent on the regularization and kernel A screenshot of a computer screen

Description automatically generatedparameters selected. The classification report for the SVM model is as follows:

An overall accuracy of 0.91 (91%), is achieved using the SVM model. With F1-score of 0.95, recall of 0.97, and precision of 0.93, it performs better on class 0 (No churn). With precision 0.76, recall 0.61, and F1-score 0.67, class 1 (Churn) performance is inferior. This indicates that the class imbalance (865 versus 150 support) may be the reason why the model performs better when it comes to class 0 instances but struggles a little bit with class 1. The average F1-score for the classes is 0.81, which indicates respectable performance.

* 1. Gradient Boosting Machine

Gradient Boosting is an ensemble technique wherein models are built one after the other, with each model fixing the mistakes of its predecessor. Both regression and classification problems frequently make use of it. The model is constructed iteratively, with decision trees usually serving as weak learners, and each new model is trained to forecast the residuals (errors) of the preceding models. The method is iterated until the model achieves the required degree of accuracy, with the aim of minimizing a certain loss function.

While gradient boosting can provide extremely accurate models, improper regularization can make it more prone to overfitting. Because of its capacity to capture complicated patterns and work well with complex datasets, this approach has gained popularity. The classification report for this model is as follows:

A screenshot of a computer

Description automatically generated

With an overall accuracy of 0.93 (93%), the GBM model demonstrates a minor improvement. In class 0 (No Churn), it continues to perform well (precision 0.94, recall 0.97, F1-score 0.96), while in class 1 (Churn), it outperforms SVM (precision 0.82, recall 0.66, F1-score 0.73). Class imbalance effects are still evident, but the macro average F1-score of 0.84 indicates more balanced performance across classes.

* 1. K-Nearest Neighbours

For applications involving regression and classification, the K-Nearest Neighbors (KNN) algorithm is an easy-to-understand technique. The majority class of a data point's closest neighbors is used to classify it. Within KNN, the "K" stands for the number of nearest neighbors taken into account. The majority class label of the K nearest points is assigned after the algorithm calculates the distance between the test data point and all training data points.

Being non-parametric, KNN does not assume anything about the distribution of the underlying data. Although KNN is easy to build, it can be computationally expensive for large datasets and depends on the distance metric and K values chosen. The classification report for the model is as follows:

A screenshot of a computer

Description automatically generated

With a 92% accuracy rate, the KNN model performs well. Class 0 (No Churn) performance (precision 0.93, recall 0.98, F1-score 0.95) is comparable to that of the other models. It has lesser precision (0.81) but better recall (0.55) for class 1 (Churn) when compared to GBM, yielding an F1-score of 0.66. Although there is still opportunity for improvement in class 1, the macro average F1-score of 0.81 is similar to the SVM model and indicates good overall performance.

* 1. Neural Network

An algorithmic system called a Neural Network is made up of layers of connected "neurons" that resemble the structure of the human brain in order to identify patterns. Complex tasks like natural language processing, image and audio recognition, and other high-dimensional data challenges are where it excels. Every neuron takes in information, processes it, and then sends the result to the layer above. This process gradually changes the data into a format that may be used to anticipate the desired variable.

Although they need a lot of data and substantial processing capacity, neural networks are an effective tool for simulating complex and highly non-linear connections. When the model is deep (has many layers) and there is a lack of training data, they may be prone to overfitting. Overfitting is frequently avoided by using regularization strategies like dropout. The classification report for the model is as follows:

A screenshot of a computer

Description automatically generated

The Neural Network model has an impressive accuracy of 0.97 (97%). It shows excellent and balanced performance across both classes, with high precision, recall, and F1-scores (0.98 for class 0 (No Churn), 0.89 for class 1 (Churn)). The macro average F1-score of 0.94 indicates superior and more balanced classification ability. This model appears to handle the class imbalance most effectively, providing the best overall performance among the four models.

* 1. Model Comparison

Model comparison involves assessing various machine learning algorithms to identify which one is the most effective for a particular job. Training multiple models on the same dataset, evaluating their performance with metrics such as accuracy, precision, recall, and F1-score, and choosing the model that achieves the best balance among these factors is the process involved. This comparison aids in determining the best model for a particular issue, guaranteeing precise and trustworthy predictions. Different algorithms such as decision trees, random forests, neural networks, and gradient boosting are frequently evaluated based on their performance in areas like generalizing to new data, managing imbalanced classes, and identifying outliers.  
  
**Our actions or activities were as follows:**  
  
We utilized various machine learning algorithms such as Random Forest, Neural Network, Decision Tree, Gradient Boosting, and Logistic Regression in our project for Jade Inc. to forecast customer churn. The assessment of these models was based on important performance indicators like accuracy, precision, recall, and F1-scores. Confusion matrices and ROC-AUC curves were additionally employed to evaluate the effectiveness of every model. We evaluated the predictive power of different models, comparing them to choose the best one that balanced reducing false negatives with maintaining high accuracy and precision, enabling us to gain insights into customer churn patterns.

A graph with blue bars

Description automatically generated with medium confidence

The Random Forest and Gradient Boosting models exhibit the highest accuracy, outperforming the other models, including the Decision Tree. This suggests that these ensemble methods are better suited for the classification task at hand, as they can handle complex patterns and interactions within the data more effectively. Random Forest and Gradient Boosting also demonstrate higher precision, recall, and F1-scores, especially for the minority class (Class 1). But, while gradient boosting may excel with tabular data in the short term, neural networks provide more versatility and potential in the long term for capturing complex data patterns, making them a compelling choice despite any initial score differences Given these results, focusing on Random Forest and Neural Network for predictive modeling will likely yield more accurate and reliable outcomes, making them the preferred choices for decision-making processes.

* + 1. ROC-AUC Chart

A graph of a number of different colored lines

Description automatically generatedThe ROC curve visually displays how well a binary classification model performs at various threshold levels. The graph shows the relationship between True Positive Rate (TPR) - also referred to as sensitivity or recall - and False Positive Rate (FPR), which measures the proportion of false positives among the actual negatives. By changing the decision threshold, the ROC curve demonstrates the balance between accurately recognizing positive examples and incorrectly labeling negatives as positives. An optimal model would move the curve towards the upper-left corner, where both True Positive Rate is high and False Positive Rate is low, indicating strong classification ability.  
  
The AUC of the ROC curve measures the model's overall effectiveness. The AUC varies between 0 and 1, with 0.5 indicating a random classifier, and 1.0 denoting an ideal classifier. A greater AUC signifies improved model performance, showing how well the model can differentiate between positive and negative classes at various thresholds. The ROC-AUC metric is beneficial for comparing numerous models, as it offers a singular value to evaluate the model's effectiveness regardless of the specific classification threshold used.

Based on the ROC curves and AUC values displayed in the chart, we can draw several conclusions about the performance of each model:

**Random Forest**:

AUC: 0.99 This model has the highest AUC, indicating excellent performance in distinguishing between classes. It is likely the best model in this comparison.

**Neural Network**:

AUC: 0.94 This model also performs very well, with a high AUC, making it a strong candidate for accurate classification.

**Gradient Boosting**:

AUC: 0.93 Another strong performer, with a slightly lower AUC than the neural network but still indicating high classification accuracy.

**K-Nearest Neighbors (KNN)**:

AUC: 0.92 KNN shows good performance, although it is not as strong as the top three models.

**Decision Tree**:

AUC: 0.89 The decision tree model has a good AUC but is outperformed by random forest, gradient boosting, and neural network models.

**Support Vector Machine (SVM)**:

AUC: 0.89 SVM has the same AUC as the decision tree, indicating similar performance.

**Interpretation and Recommendations**: Random Forest is the top-performing model based on the AUC value, suggesting it has the best overall ability to discriminate between the classes. Neural Network and Gradient Boosting also show excellent performance and could be considered as strong alternatives.

Models like KNN, Decision Tree, and SVM perform well but are not as effective as the top three models.

* + 1. Feature Importance

Feature importance is a method that ranks input features by their impact on a model's ability to predict outcomes. In the field of machine learning, knowing feature importance is crucial for determining the features that hold the most sway in predicting the output. Different models have varying methods for determining feature importance; for example, decision trees and ensemble techniques such as Random Forest or Gradient Boosting assess importance by evaluating how each feature contributes to enhancing the model's performance, utilizing metrics like Gini impurity and information gain.  
  
Assessing the importance of features is essential for understanding and enhancing models, as it offers understanding on the connections between input variables and the desired result. By emphasizing the most influential characteristics, it allows analysts to concentrate on the main factors affecting a model's accuracy, diminish interference from unrelated data, and uncover possible prejudices. Comprehending the significance of features can aid in selecting features, streamlining the model, enhancing performance, and providing stakeholders with more interpretable outcomes.

Since our model comparison says our two best models are Random forest and Neural Networks, it is better to concentrate our resources and time on the top 3 features of both of the models.

**Feature Importance of Random Forest:**

A graph with a bar graph

Description automatically generated**Random Forest**

**Neural Network**

A graph with blue and white bars

Description automatically generated

For the Random Forest model, the top 3 features are:

1. **Tenure**
2. **CashbackAmount**
3. **WarehouseToHome**

For the Neural Network model, the top 3 features are:

1. **Tenure**
2. **Complain\_1**
3. **NumberOfAddress**

Both models concur that Tenure is the most significant attribute, suggesting it plays a crucial role in predicting the target variable, Churn. This indicates that how long a customer has been with the company is a strong indicator of their behaviour or worth.

Nevertheless, the models vary in their critical characteristics. The Random Forest prioritizes operational factors such as CashbackAmount and logistics (WarehouseToHome), whereas the Neural Network places greater emphasis on customer satisfaction (Complain\_1) and customer details (NumberOfAddress). The variance may arise from variations in the structures of the models and their approach to information processing. The Random Forest captures nonlinear patterns in numerical data more effectively, whereas the Neural Network is better at detecting categorical variables and intricate relationships. Using various models is crucial to fully comprehend the features' significance in predictive modelling.

1. Recommendations

Based on the models and the information that we received from them, the researcher has come up with certain recommendations. These recommendations are for both implementation and for future studies. This will allow the company to focus on the problem and give suggestions on how to work on them and provide clear limitations and what to do for future studies.

6.1 Recommendations for implementation

**Simplifying the user interface for new casual users**

Apparently, the current interface of the software is creating problems and hitches to the number of users, especially the first-time and intermittent users. To this end Jade should consider a total overhaul of the interface of the web based software especially in regard to usability. This redesign should involve the simplification of the layout structure of the site so as to provide only vital data and services in a sequential manner. The navigation should be clear, it also can be addressed with the clearer menu or, exactly guided user flow which leads the user to the desired location.

Furthermore, Jade should be equally careful about labeling functions. Any button, link, as well as any other feature, has to have a name that will inform a user what they are dealing with. Contemplate the use of hint or brief description for functionalities that are complicated to understand. User testing should also be done from time to time during redesign process in order to see the changes have enhanced the user experience. It may involve testing different layouts of the application by a selected pool of users and collecting feedback from new and occasional users who constitute the target of the improvement.

**Amplifying Savings for Greater Customer Value**

The current mode of operation of discounts is not fulfilling the expectations of the users of the services hence causing inconvenience and loss of potential cost savings. Jade should review its discounts and promotions policy in a way that makes them more easily comprehensible to the public. This could entail having a special tab such as ‘Deals’ or ‘Savings’ whereby all the existing offers are shown at the website’s interface prominently. A more transparent method of pricing as it will be possible to display the original price, the amount of the discount and the final price besides the item where necessary.

In addition, Jade should enhance the communication on the discounts available to the customers. This could be done by use of targeted message that inform the users about offers relevant to their shopping history or preferences. The company may need to come up with a loyalty programme that offers customers with a wholesome package of benefits which can be easily seen to scale up progressively with increased patronage. It should also be easy to apply for discounts – for instance, by making them happen in the background, which eliminates having to enter any code at check-out. Last but not least, Jade should offer information about the ways to save money on the platform, for example, in the form of short videos or the special list of Frequently Asked Questions related to the hot offers and deals. Such changes should offer more customer benefits and are likely to raise total customer benefit, customer satisfaction and customer loyalty.

6.2 Recommendations for future use

For future studies related to this type of problem, the researcher outlines several areas to explore:   
  
**Sample Size**: Future research should try to use large sample sizes to come up with more accurate results that will represent a large population. This could mean collecting information from a larger an more varied patronage than perhaps had been initially targeted.  
  
**Non-Linearity**: Exploring other polynomial trends could also be conceivable to achieve greater inquiries on the collected data. This might include such issues as studying more refined methods of statistics or machine learning that can accommodate more intricate structures.  
  
**Modeling Techniques**: Perhaps more work could be done to optimize the modeling process with respect to the peculiarities of the dataset, and if different methods could be used, maybe one of them would work better or at least be easier to explain. This may involve using different learning algorithms, or even the use of an ensemble method.  
  
**Missing Values**: Finding better ways to deal with the problems of missing values might help revolutionize the outcome of the analysis that is conducted. This might involve the use of sophisticated imputation methods or procedures that can operate with missing data.  
  
**No Outlier Reduction**: It was also interesting to think about how outliers affect the analysis and to contemplate ways of dealing with the problem which did not involve the exclusion of the outliers.  
  
Such future researches would contribute into the enhancement of the analytical framework used in this study, and, thus, into more precise forecasting and better understanding of users’ behavior and preferences. If these areas could be tackled, Jade could possibly increase its capabilities to make decisions based on data analysis and even better its products and services.

1. Conclusion

The project was really insightful and gave very important recommendations and conclusion for Jade. This part highlights the most important points for easier understanding. The important points are as follows:

**Random Forest emerged as the most accurate model with a 91% accuracy rate.**

The Random Forest algorithm has demonstrated superior performance in predicting customer churn for Jade, achieving an impressive 91% accuracy rate. This high level of accuracy suggests that the model has successfully captured complex patterns and relationships within the customer data, making it a reliable tool for identifying potential churners.

The success of the Random Forest model can be attributed to its ensemble learning approach, which combines multiple decision trees to make predictions. This method is particularly effective in handling non-linear relationships and interactions between features, which are common in customer behavior data. The model's strong performance indicates that it has effectively learned from the various customer attributes and behaviors to differentiate between likely churners and loyal customers.

**The model also demonstrated robust performance with a 99% accuracy in predicting no churn.**

The Random Forest model's ability to predict non-churners with 99% accuracy is particularly noteworthy. This high level of precision in identifying satisfied customers who are likely to remain loyal is crucial for Jade's customer retention strategies. It allows the company to focus its resources more efficiently by targeting retention efforts towards those customers who are genuinely at risk of churning.

This exceptional accuracy in predicting no churn suggests that the model has successfully identified key indicators of customer satisfaction and loyalty. It provides Jade with a powerful tool to understand the characteristics and behaviors of its most stable customer base. This insight can be leveraged not only for retention purposes but also to inform strategies for acquiring new customers with similar profiles to those who tend to remain loyal.

**Tenure and Cashback Amount were identified as the most influential features contributing to customer churn.**

The analysis has revealed that Tenure and Cashback Amount are the two most significant factors in predicting customer churn. This finding provides valuable insight into the drivers of customer loyalty and dissatisfaction. Tenure, representing the length of a customer's relationship with Jade, suggests that longer-standing customers are less likely to churn. This underscores the importance of nurturing customer relationships over time and potentially implementing loyalty programs that reward long-term customers.

The significance of Cashback Amount indicates that financial incentives play a crucial role in customer retention. Customers who receive higher cashback amounts are likely more satisfied and therefore less prone to churning. This insight can guide Jade's pricing and reward strategies, suggesting that enhancing cashback offers could be an effective tool for reducing churn. However, it's important to balance these incentives with profitability considerations.

**Improve Customer Experience: Simplify the user interface and enhance navigation based on customer feedback.**

Based on customer feedback, it's clear that the current user interface is a pain point for many users, particularly new and casual ones. Simplifying the interface should be a top priority for Jade. This could involve redesigning the layout to be more intuitive, reducing clutter, and ensuring that key features are easily accessible. The goal should be to create a streamlined experience that allows users to accomplish their tasks with minimal confusion or frustration.

Enhancing navigation is another crucial aspect of improving the customer experience. This could include implementing a more logical menu structure, adding clear signposts throughout the user journey, and possibly introducing a search function if one doesn't already exist. Additionally, Jade should consider implementing user onboarding processes or tutorials to help new users familiarize themselves with the platform quickly. Regular usability testing and ongoing collection of user feedback will be essential to ensure that these improvements are effective and continue to meet user needs over time.

**Amplify Promotions: Increase transparency and accessibility of discounts to improve customer satisfaction and retention.**

To address the issues with the current discount system, Jade should focus on increasing the transparency and accessibility of its promotions. This could involve creating a dedicated section within the platform where all current discounts and promotions are clearly displayed. The terms and conditions of each promotion should be presented in simple, easy-to-understand language to avoid confusion.

Improving the accessibility of discounts is equally important. This could be achieved by implementing a system that automatically applies eligible discounts at checkout, rather than requiring users to manually enter codes. Jade could also consider personalized discount notifications based on user behavior and preferences. Additionally, implementing a clear visual indicator of savings (e.g., showing original price, discount amount, and final price) can help users quickly understand the value they're receiving. These changes should lead to increased customer satisfaction as users feel they're getting better value, potentially improving retention rates.

**Implement Random Forest and Neural Network models for ongoing churn prediction.**

Given the strong performance of the Random Forest model and the potential complementary insights from Neural Networks, Jade should implement both these models for ongoing churn prediction. The Random Forest model can serve as the primary prediction tool, leveraging its high accuracy in identifying both churners and non-churners. Its ability to handle non-linear relationships and provide feature importance makes it valuable for understanding the factors driving churn.

The Neural Network model can be used in parallel to capture complex patterns that might not be evident in the Random Forest model. Neural Networks excel at identifying intricate relationships in large datasets and could potentially uncover additional insights. By using both models, Jade can cross-validate predictions and gain a more comprehensive understanding of churn risk. Regular retraining of these models with new data will be crucial to maintain their accuracy over time as customer behaviors and market conditions evolve.

**Focus on refining the user interface and promoting discounts more effectively.**

Building on the insights from the customer experience and promotion recommendations, Jade should prioritize refining its user interface. This refinement should focus on creating a clean, intuitive design that allows users to navigate the platform effortlessly. Special attention should be given to how discounts and promotions are presented within the interface. Consider implementing visual cues or a dedicated "Deals" section that's prominently displayed to ensure users don't miss out on potential savings.

In terms of promoting discounts more effectively, Jade should develop a multi-channel approach. This could include in-app notifications for relevant deals, email marketing campaigns highlighting current promotions, and possibly even push notifications for mobile users (if applicable). The key is to make discounts more visible without being intrusive. Additionally, Jade could consider implementing a personalized recommendation system that suggests relevant discounts based on a user's browsing and purchase history. This targeted approach can increase the likelihood of users taking advantage of promotions, thereby improving their perceived value and potentially boosting retention.

**Consider expanding the sample size and exploring non-linear relationships for future studies to gain deeper insights.**

For future studies, Jade should consider significantly expanding its sample size. A larger dataset would provide more robust and representative results, potentially uncovering patterns or segments that might be overlooked in a smaller sample. This expanded dataset could include a wider range of customer types, behaviors, and time periods, allowing for more comprehensive analysis. Additionally, a larger sample size would increase the statistical power of the models, potentially leading to even more accurate predictions and insights.

Exploring non-linear relationships within the data could uncover complex patterns that linear models might miss. This could involve using advanced techniques such as polynomial regression, decision trees, or more sophisticated machine learning algorithms that can capture non-linear interactions between variables. By investigating these non-linear relationships, Jade might discover unexpected factors influencing customer churn or retention. This deeper understanding could lead to more nuanced and effective strategies for improving customer satisfaction and reducing churn.

References

Awan, A. (2023, September). *The Curse of Dimensionality in Machine Learning: Challenges, Impacts, and Solutions*. Datacamp; Datacamp. https://www.datacamp.com/blog/curse-of-dimensionality-machine-learning

DHL. (2024, January 25). *Customer Satisfaction In E-Commerce - DHL Express LK*. Dhl.com. https://www.dhl.com/discover/en-lk/e-commerce-advice/e-commerce-best-practice/customer-satisfaction-in-e-commerce

Lin, G. T. R., & Sun, C. (2009). Factors influencing satisfaction and loyalty in online shopping: an integrated model. *Online Information Review*, *33*(3), 458–475. https://doi.org/10.1108/14684520910969907

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

Rudälv, A. (2022). *Predicting Customer Churn in E-Commerce Using Statistical Modeling and Feature Importance Analysis A Comparison of Random Forest and Logistic Regression Approaches*.