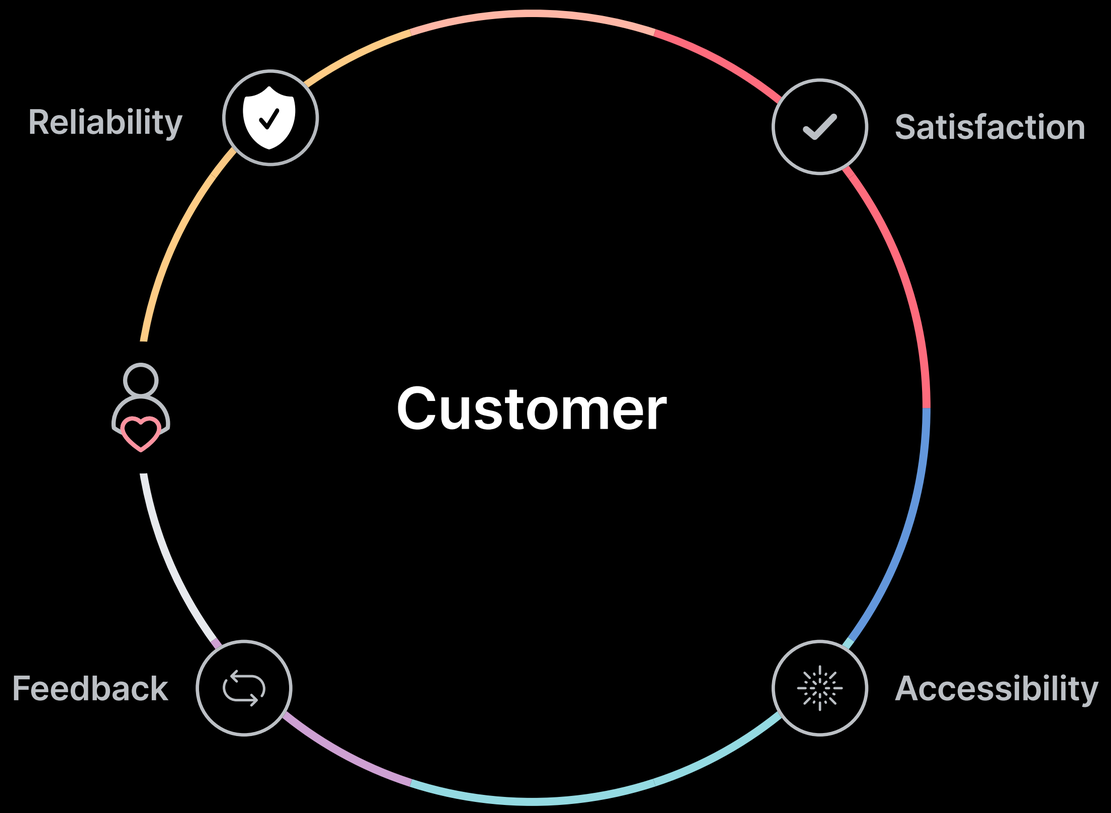
JADE

JADE INC.

Customer Churn Analysis

By: Nimish Abraham
Date: 2024/07/12

JADE



JADE INC.

Crafting Unmatched Customer Journeys

JADE

Business Problem

Jade Inc. is a leading e-commerce company that prides itself on delivering a seamless shopping experience to its diverse customer base. Despite its success, the company is facing challenges in maintaining low levels of customer churn, a critical factor for fostering stakeholder trust and sustaining business growth.

20%
Of Users
Churned

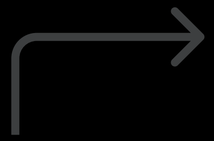
36%
Of Users Churn
Due to lower tenure with Jade

29%
Of Users reported
Lower levels of Satisfaction

15%
Of Users Are
Unhappy with the Customer Service

JADE

Executive Summary



1

Exploration

- Data Description
- Exploratory Data Analysis
- Data Cleaning
- Data Pre-Processing

2

Modeling

- Predictive Analysis
- Model Comparison
- Accuracy Measurement

3

Actions

- Implementations
- Important variables to tackle

4

Recommendations

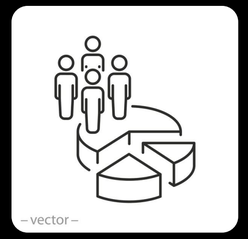
- Drawbacks in the study
- Recommendations for future studies

JADE

Data Sources

Historical Data

Analyzed the company’s historical data to view customer background and history.



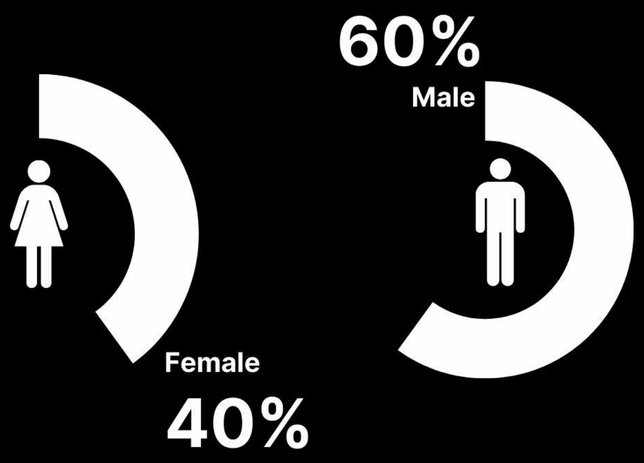
Customer Surveys

Surveyed customers of Jade to find out about their satisfaction levels and preferences.

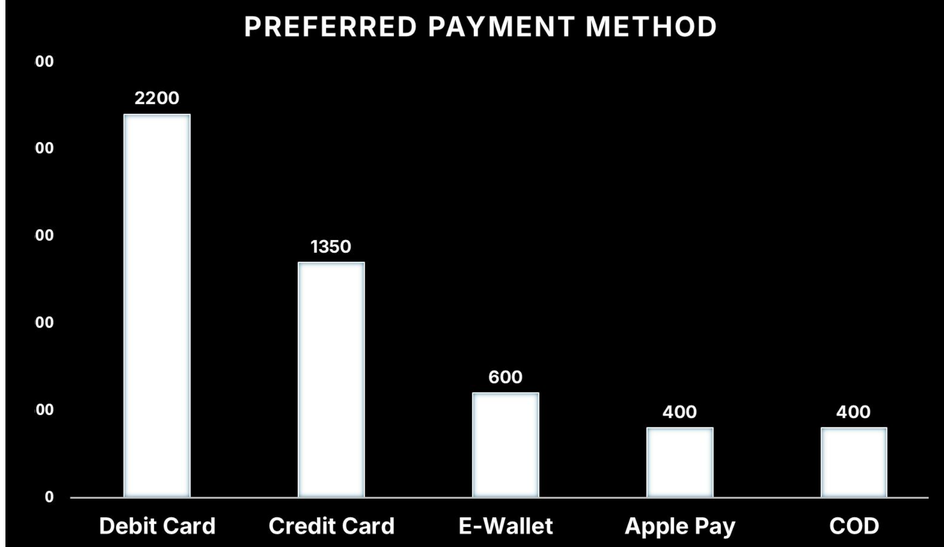
Community Insights

We collected feedback from social media and our community forum through polls.

Data Analysis

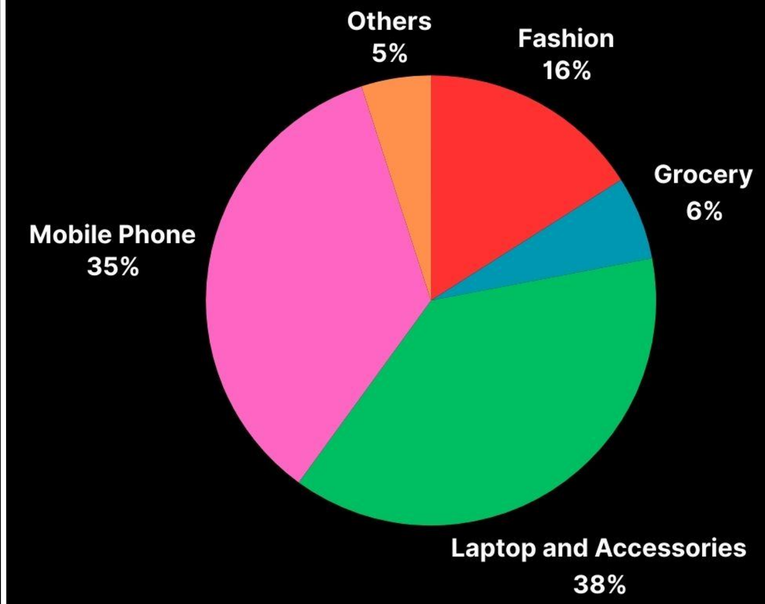


Majority of our customer Base is Male



Most Preferred Method of Payment- Debit Card

Least Preferred Method of Payment- Apple Pay

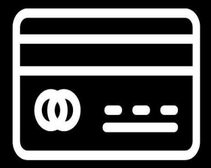


Top Selling Category- Laptop & Accessories

Least Selling Category- Grocery

Churn Analysis

20%
Of Users



who use debit card as the payment method churned

In contrast, only 5% of those who used credit cards churned.

25%
Of Male Users

Churned

In contrast, the churn rate among our female audience was significantly lower at 10%.

40%
Of Users

with less than 2years of tenure churned

In comparison, only 15% of the customers who have more than 3 years of tenure churned.

JADE

Data Preparation

After cleaning the data, we divided it into two distinct sets: the Training set and the Testing set. This separation allows us to evaluate the model's performance accurately.

Data Cleaning Methods Used

Cap and Floor

Cap and Floor was used to remove outliers in the dataset

Imputation with Median

The missing values in the dataset were imputed with median values

Skewness Control

Reduced skewness of the data using log transformation and root-square transformation

Total Data

The final count of observations in our dataset post-cleaning was 5072.

Training Dataset

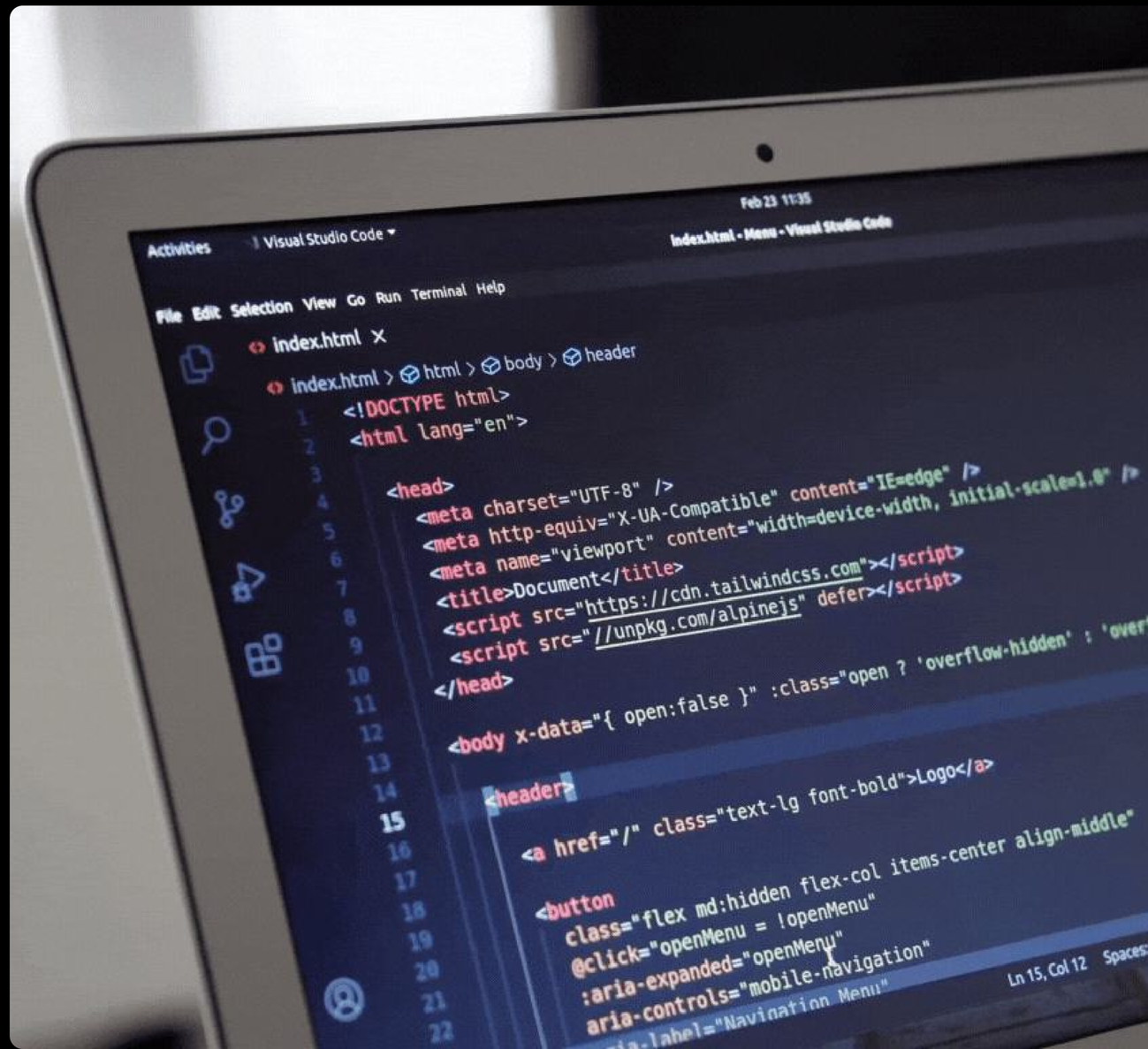
Following the split, the training dataset contained a total of 4057 observations.

Testing Dataset

Finally, our testing dataset, post-separation, comprised a total of 1015 observations.

Modeling- Types of models used

- Decision Trees
- Random Forest
- Linear Regression
- Support Vector Machine
- Gradient Boosting Machine
- Logistic Regression
- Neural Network
- K-Nearest Neighbors



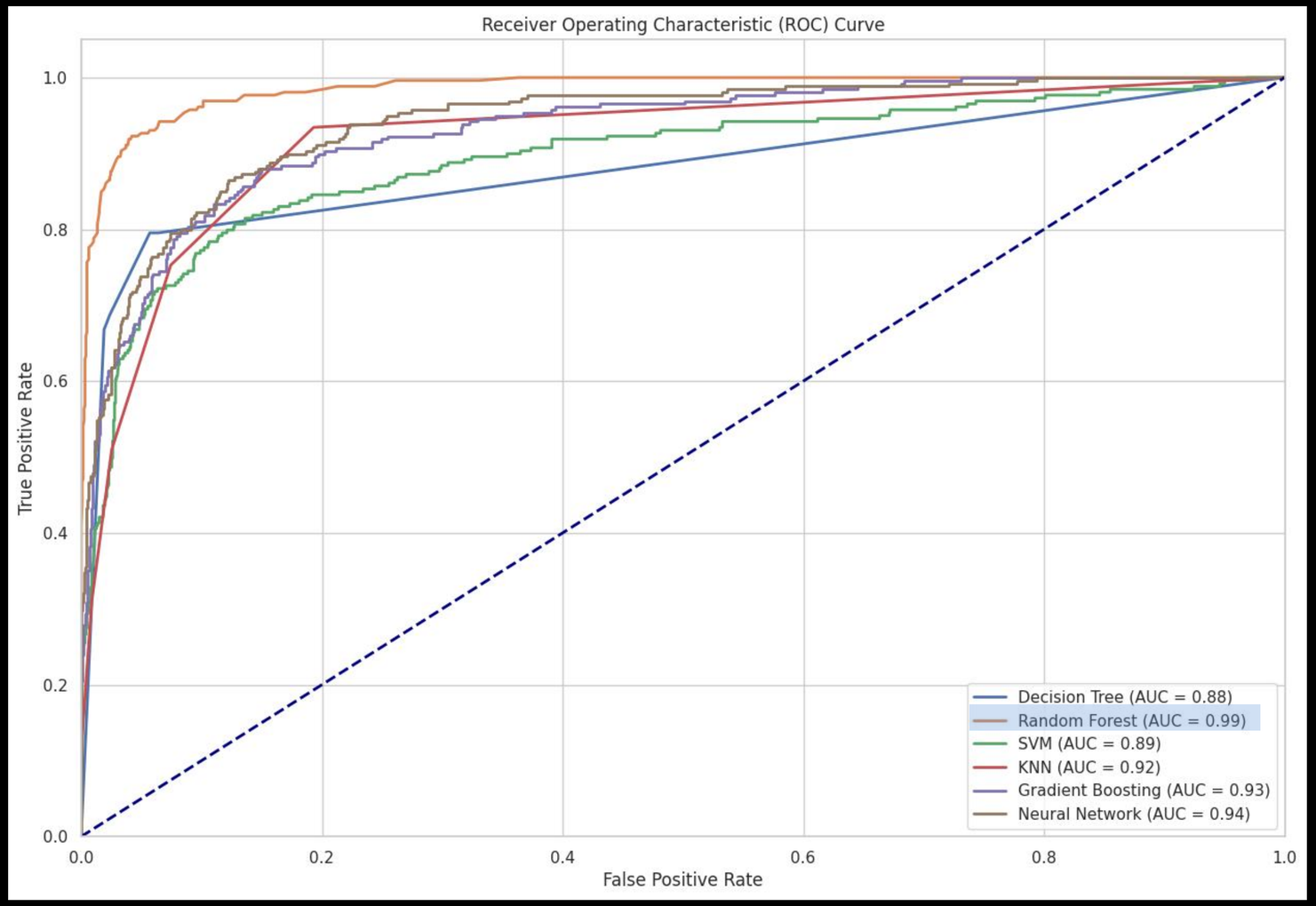
Model Comparison

When we conduct a comprehensive comparison between our top-performing Model and the rest of the models available, the outcomes can be summarized as below:

Models	Best Features	Accuracy Metrics- F1 Score	
		No Churn	Churn
Decision Tree	Tenure and Cashback Amount	96	79
Random Forest	Tenure and Cashback Amount	99	91
Logistic Regression	Tenure and Complain	95	68
Support Vector Machine	Tenure and Complain	95	67
Gradient Boosting Machine	Tenure and Complain	96	73
K-Nearest Neighbors	Tenure and Satisfaction Score	95	66
Neural Network	Number of Addresses and Order Amount Hike	98	89
Linear Regression	Tenure and Complain		

Model Comparison Contd.

ROC Chart



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.

Best Model

After conducting an exhaustive analysis, where we ran all the models and compared their accuracy rate and F1-scores, we identified the best performing model as:

Random Forest

Accuracy Rate Achieved:

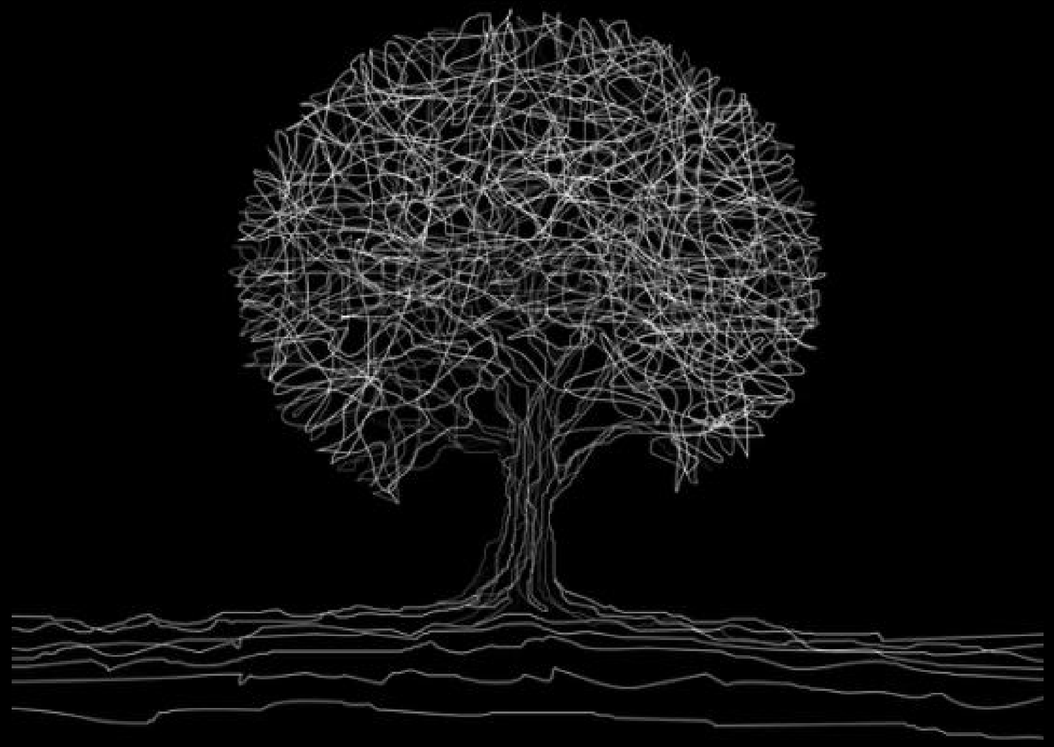
Best accuracy rate achieved for random forest for predicting churn was **91%**

F1 Score

The F1 Score of Random Forest showed that the model predicted No-Churn with an accuracy of **99%** .

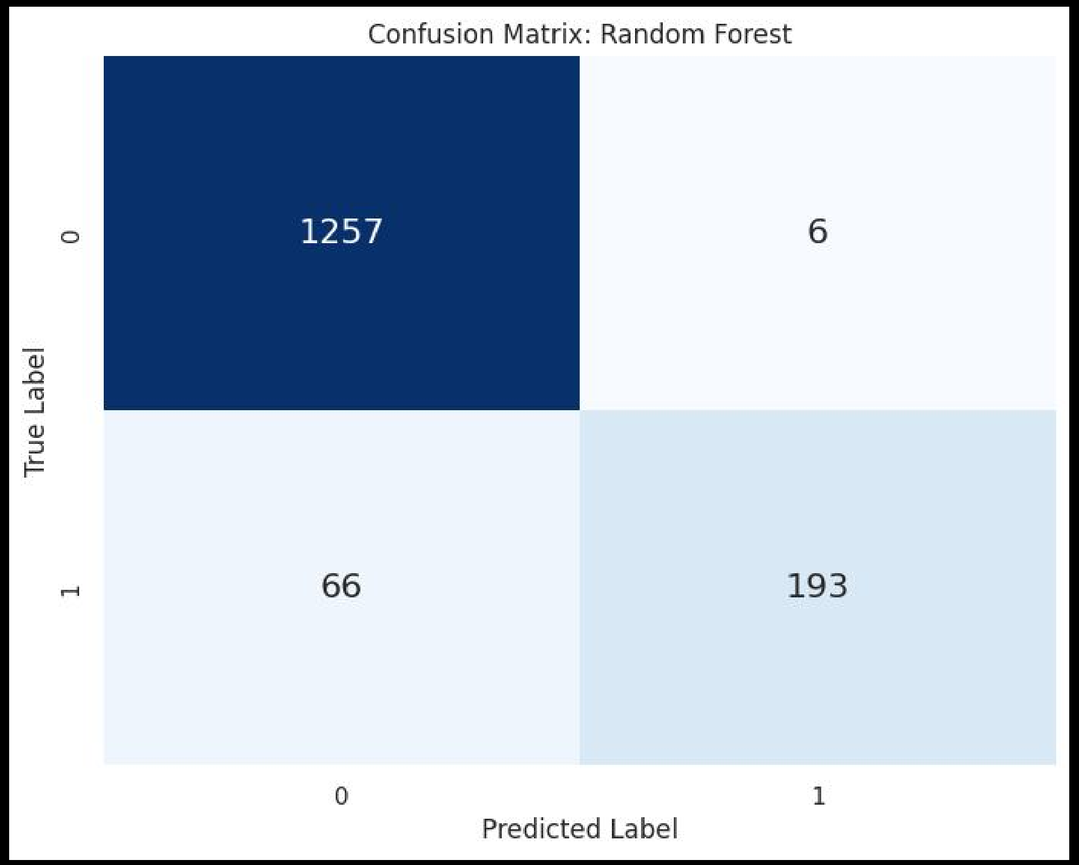
Reasons to Use

- Handles Overfitting
- Robustness to Outliers
- Handles Missing Data
- Low Bias



Confusion Matrix- Random Forest

Following the determination that our optimal model is Random Forest, we proceeded to conduct a confusion matrix analysis to assess the model's predictive performance. The outcomes are detailed below:



99.5%
correctly predicted

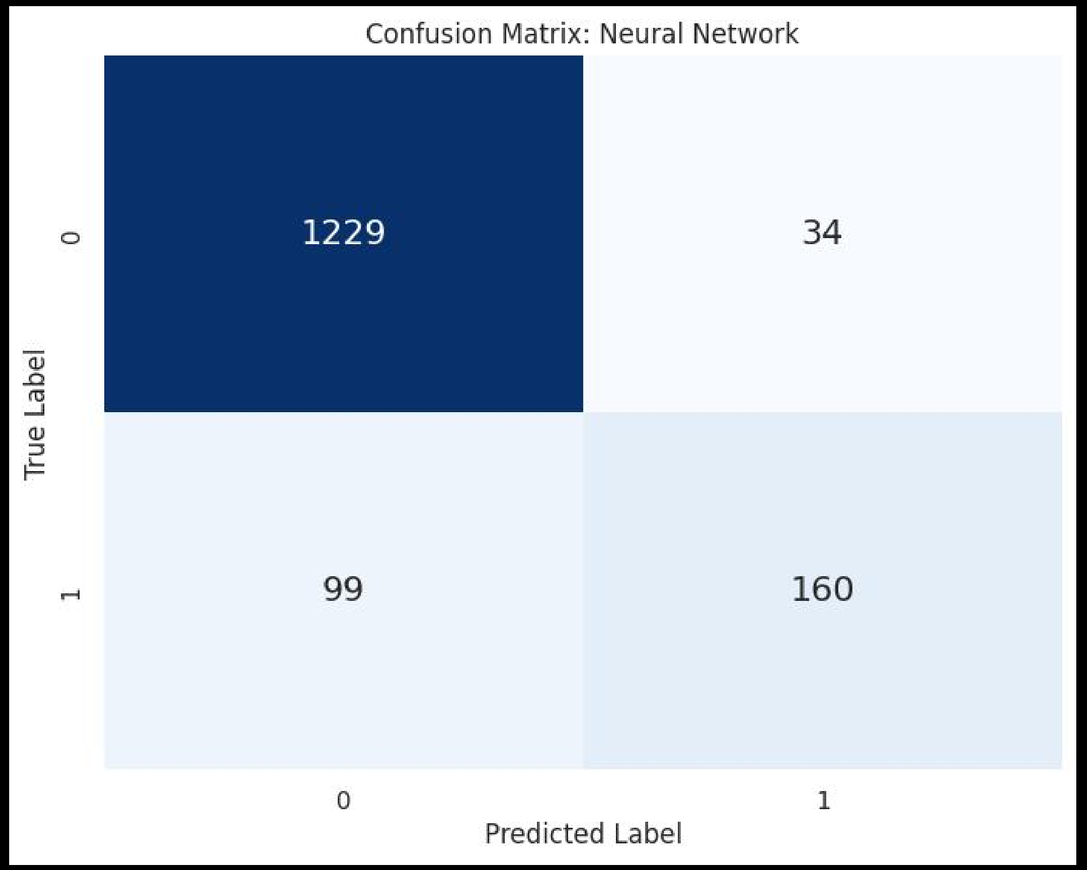
No Churn

75%
correctly predicted

Churn

Confusion Matrix- Neural Network

When we compare our top-performing model with the second-best model, which is the Neural Network, the outcomes are outlined below. The analysis reveals significant insights into the performance discrepancies between the two models.



97%
correctly predicted

No Churn

61%
correctly predicted

Churn

JADE

Feature Importance

Random Forest

During the analysis of our top model, we identified the key features that significantly impact churn. This process was pivotal in understanding the most critical variables influencing churn and establishing a foundation for our churn reduction strategies within the company.

NOTABLE FEATURES IDENTIFIED

- Tenure
- CashbackAmount
- WarehousetoHome
- Complain

68%

total contribution toward churn

JADE

Implementations

USER INTERFACE IMPROVEMENTS

Simplifying the user interface for new casual users

Users have expressed that the current interface is not intuitive, leading to confusion and inefficiencies. Common feature requests include a more streamlined layout, easier navigation, and clearer labeling of functions.

“I often find myself lost in the current interface. A more straightforward layout would save a lot of time.”

Jim, Sales Manager
Dunder Mifflin

“The navigation is confusing and not user-friendly. Simplifying it would greatly enhance our experience.”

Gavin
Hooli

RATING BY USER INTERFACE

ISSUE DATE

2024/07/15

ISSUED BY

FIGMA

JADE

Implementations

AMPLIFYING CURRENT PROMOTIONS

Amplifying Savings for Greater Customer Value

Users have expressed that the current discount system is not intuitive, leading to confusion and missed opportunities for savings. Common requests include more transparent pricing, easier access to promotions, and clearer communication of discount eligibility

" I often miss out on discounts because the system is confusing. A more transparent and straightforward approach would help me take full advantage of the savings. "

Phil, Real Estate Agent
Self Employed

"The way discounts are applied is unclear and frustrating. Simplifying the process would greatly improve our shopping experience.."

Michelle
Reddit

RATING BY DISCOUNTS PROVIDED

ISSUE DATE

2024/07/15

ISSUED BY

NHA

JADE

Recommendations

The predictive analytic methods gave us a lot of insights which would help us solve the business problem. However, modeling alone cannot solve the problem and as always, there is scope for future studies.

Future Studies

- Sample Size
- Non-Linearity
- Modeling techniques
- Missing Values
- No outlier reduction

JADE

Conclusion

- Random Forest emerged as the most accurate model with a 91% accuracy rate.
- The model also demonstrated robust performance with a 99% accuracy in predicting no churn.

- Tenure and Cashback Amount were identified as the most influential features contributing to customer churn.
- Improve Customer Experience: Simplify the user interface and enhance navigation based on customer feedback.
- Amplify Promotions: Increase transparency and accessibility of discounts to improve customer satisfaction and retention.

- Implement Random Forest and Neural Network models for ongoing churn prediction.
- Focus on refining the user interface and promoting discounts more effectively.
- Consider expanding the sample size and exploring non-linear relationships for future studies to gain deeper insights.

THANK YOU

Q&A?

JADE

Appendix

Data Dictionary

Data	Variable	Description
E Comm	CustomerID	Unique customer ID
E Comm	Churn	Churn Flag
E Comm	Tenure	Tenure of customer in organization
E Comm	PreferredLoginDevice	Preferred login device of customer
E Comm	CityTier	City tier
E Comm	WarehouseToHome	Distance in between warehouse to home of customer
E Comm	PreferredPaymentMode	Preferred payment method of customer
E Comm	Gender	Gender of customer
E Comm	HourSpendOnApp	Number of hours spend on mobile application or website
E Comm	NumberOfDeviceRegistered	Total number of deceives is registered on particular customer
E Comm	PreferedOrderCat	Preferred order category of customer in last month
E Comm	SatisfactionScore	Satisfactory score of customer on service
E Comm	MaritalStatus	Marital status of customer
E Comm	NumberOfAddress	Total number of added added on particular customer
E Comm	Complain	Any complaint has been raised in last month
E Comm	OrderAmountHikeFromlastYear	Percentage increases in order from last year
E Comm	CouponUsed	Total number of coupon has been used in last month
E Comm	OrderCount	Total number of orders has been places in last month
E Comm	DaySinceLastOrder	Day Since last order by customer
E Comm	CashbackAmount	Average cashback in last month

JADE

Appendix

Pre-Impute

Churn	0
Tenure	264
PreferredLoginDevice	0
CityTier	0
WarehouseToHome	251
PreferredPaymentMode	0
Gender	0
HourSpendOnApp	255
NumberOfDeviceRegistered	0
PreferedOrderCat	0
SatisfactionScore	0
MaritalStatus	0
NumberOfAddress	0
Complain	0
OrderAmountHikeFromlastYear	265
CouponUsed	256
OrderCount	258
DaySinceLastOrder	307
CashbackAmount	0

Post-Impute

Churn	0
Tenure	0
PreferredLoginDevice	0
CityTier	0
WarehouseToHome	0
PreferredPaymentMode	0
Gender	0
HourSpendOnApp	0
NumberOfDeviceRegistered	0
PreferedOrderCat	0
SatisfactionScore	0
MaritalStatus	0
NumberOfAddress	0
Complain	0
OrderAmountHikeFromlastYear	0
CouponUsed	0
OrderCount	0
DaySinceLastOrder	0
CashbackAmount	0

JADE

Appendix- Outlier Reduction

Pre-Cleaning

Tenure	4
CityTier	0
WarehouseToHome	2
HourSpendOnApp	3
NumberOfDeviceRegistered	0
SatisfactionScore	0
NumberOfAddress	4
OrderAmountHikeFromlastYear	0
CouponUsed	105
OrderCount	165
DaySinceLastOrder	40

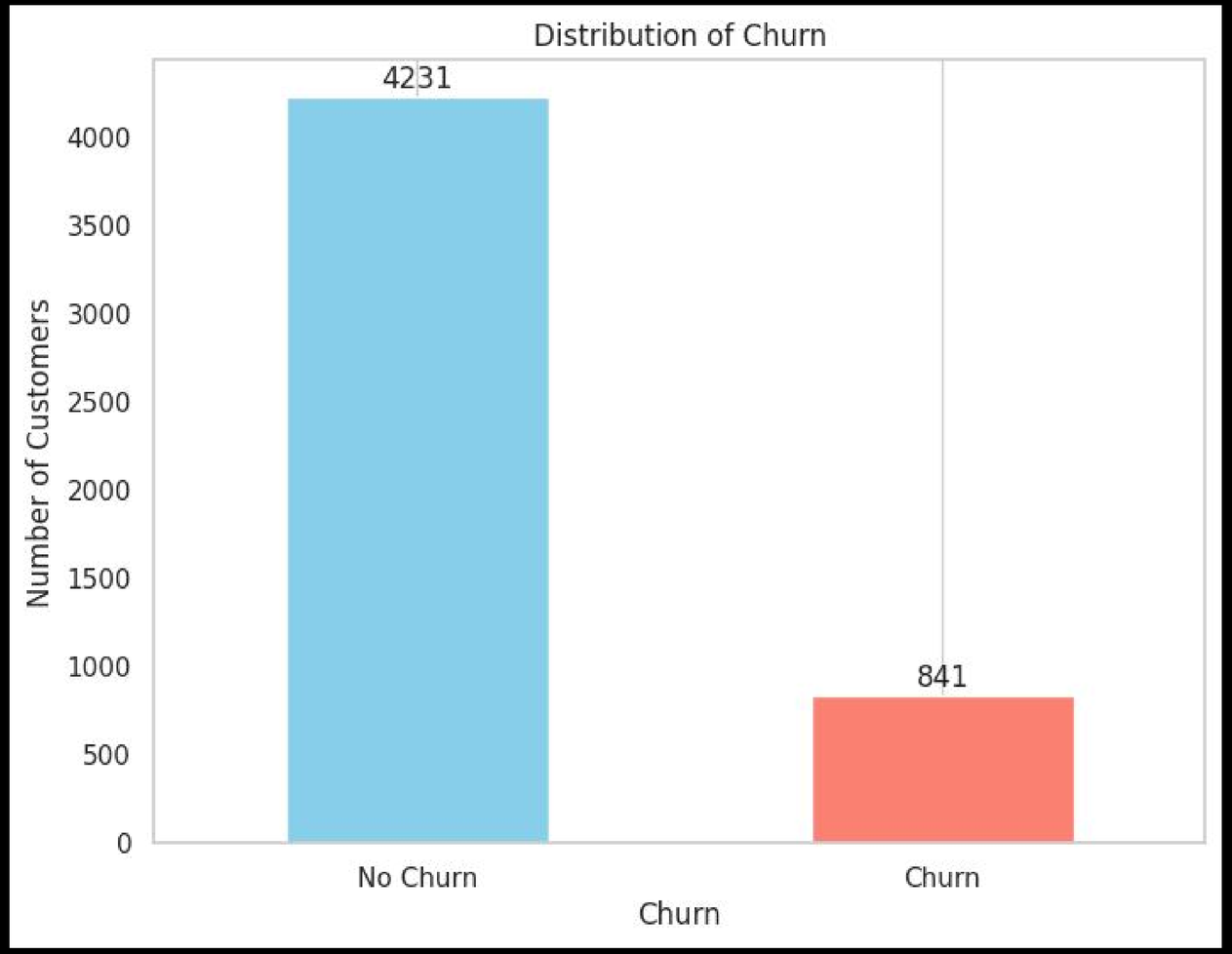
Post-Cleaning

Tenure_Capped	0
CityTier_Capped	0
WarehouseToHome_Capped	0
HourSpendOnApp_Capped	3
NumberOfDeviceRegistered_Capped	0
SatisfactionScore_Capped	0
NumberOfAddress_Capped	4
OrderAmountHikeFromlastYear_Capped	0
CouponUsed_Capped	0
OrderCount_Capped	0
DaySinceLastOrder_Capped	0
CashbackAmount_Capped	0

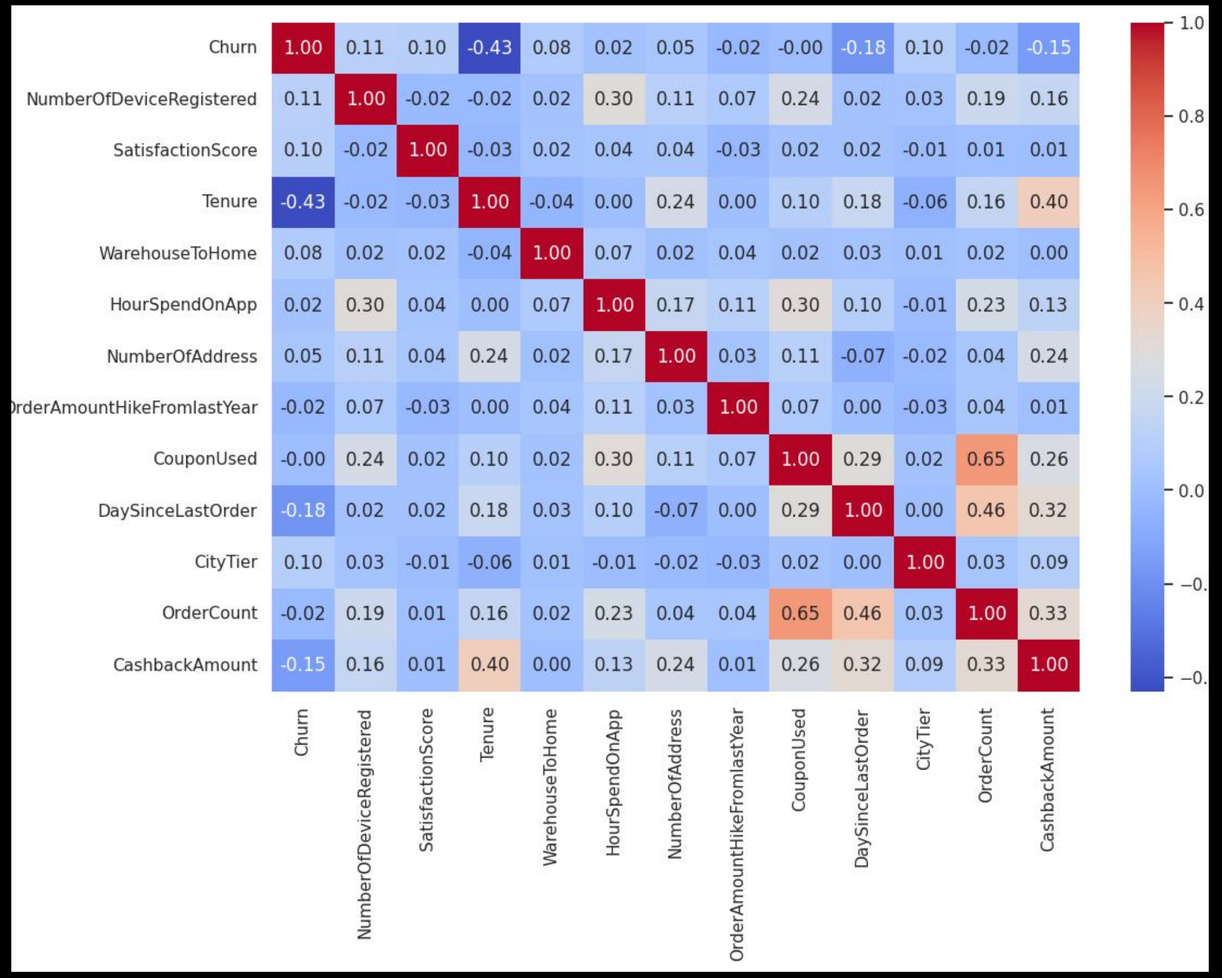
JADE

Appendix

Churn Distribution



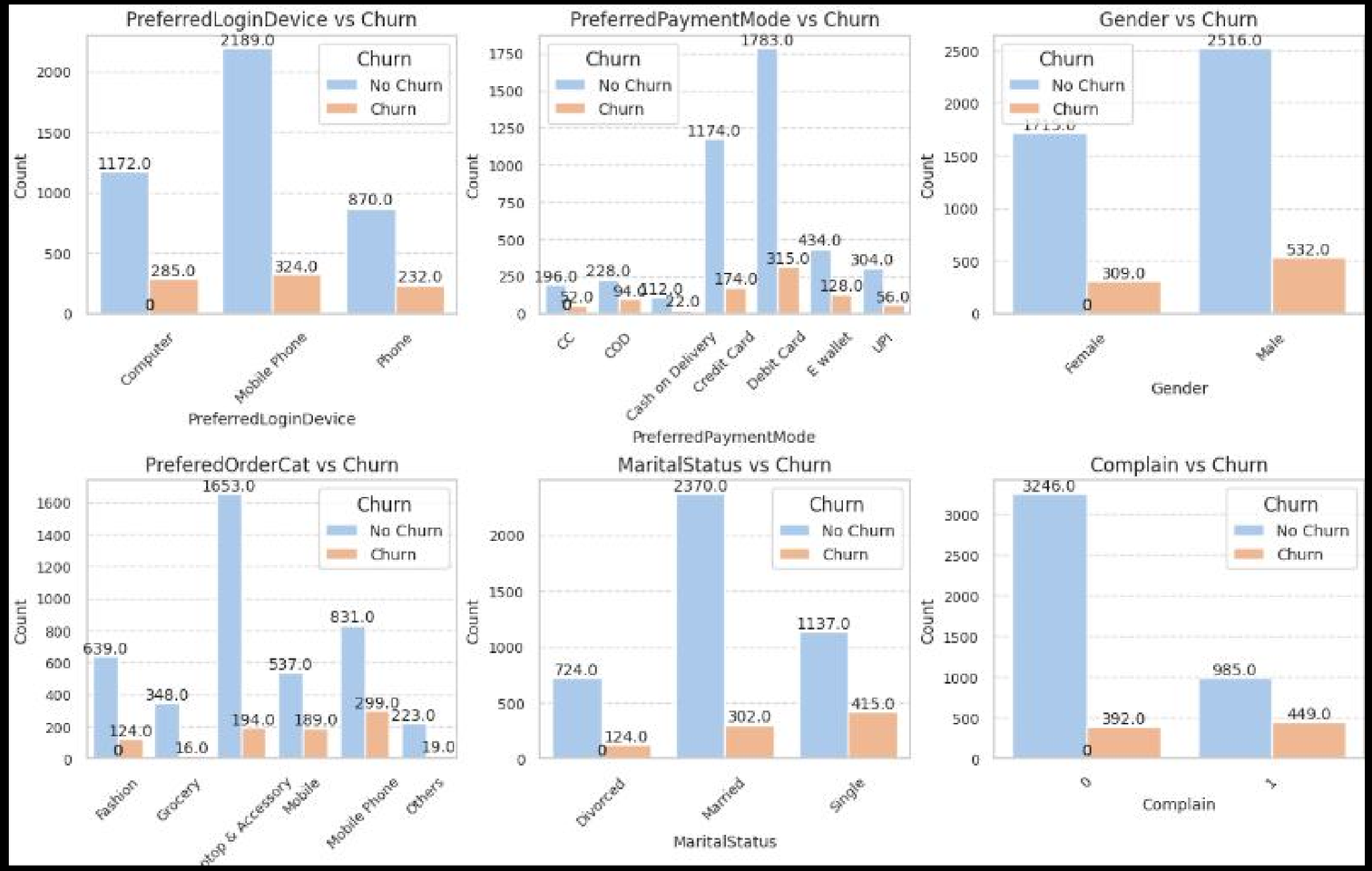
Correlation Matrix



JADE

Appendix

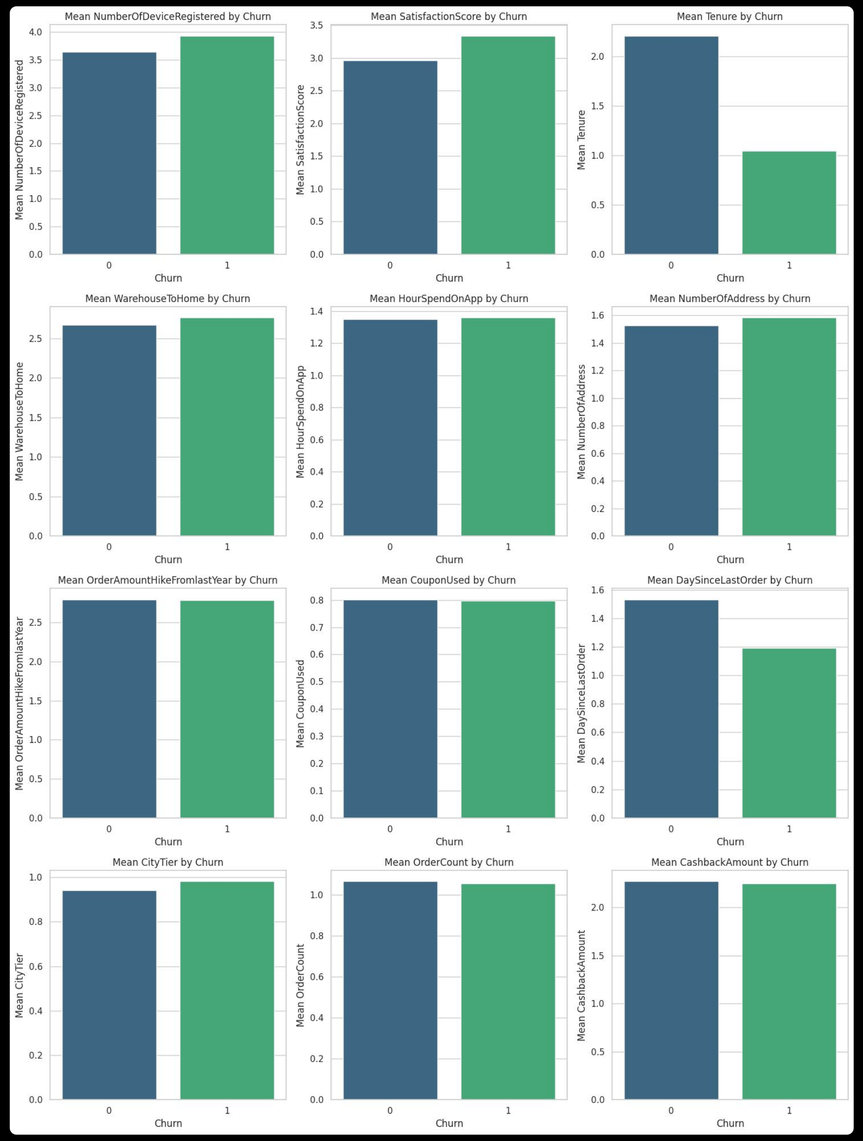
Bar Chart- Categorical



JADE

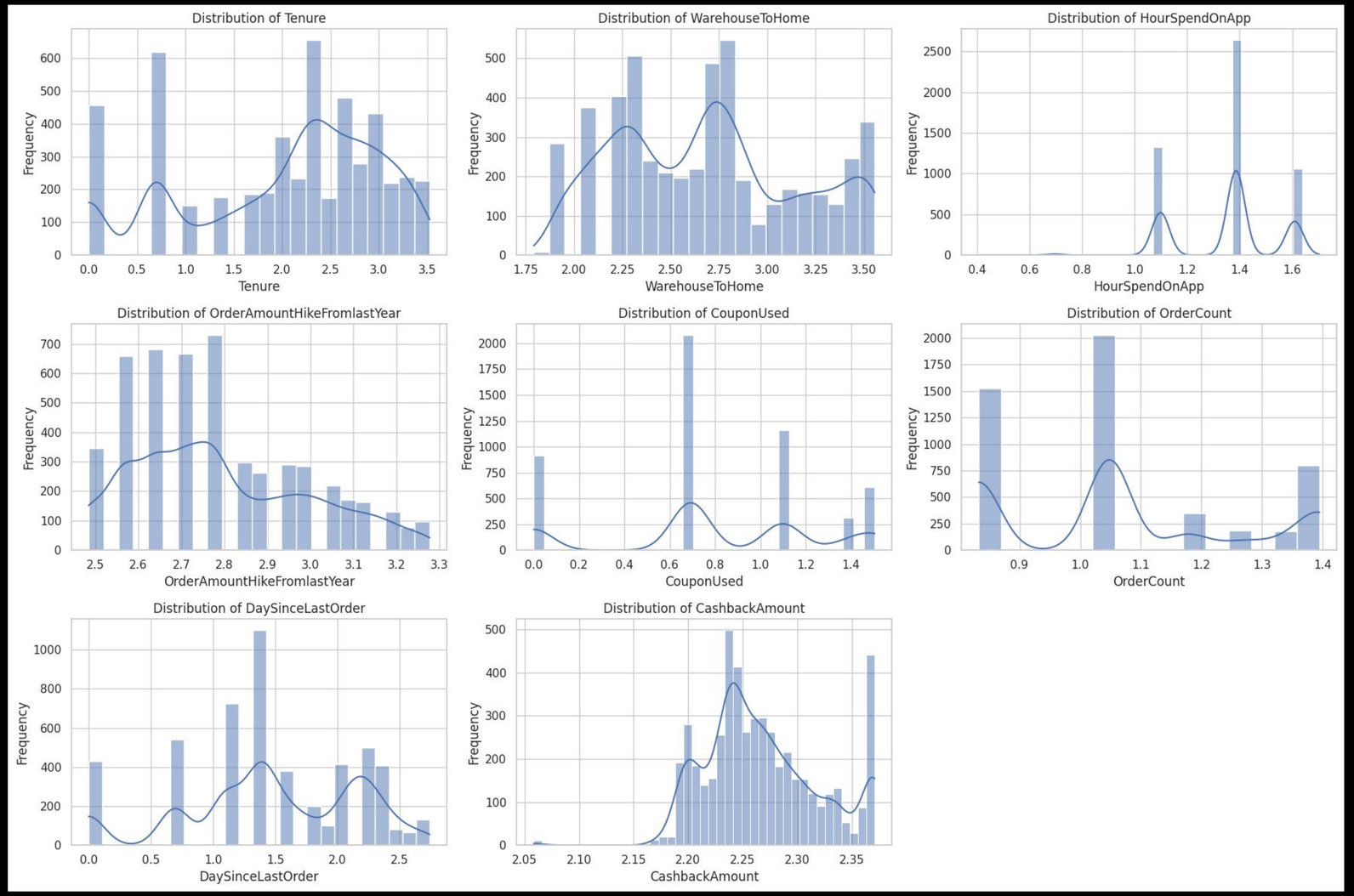
Appendix

Bar Chart- Numerical



Appendix

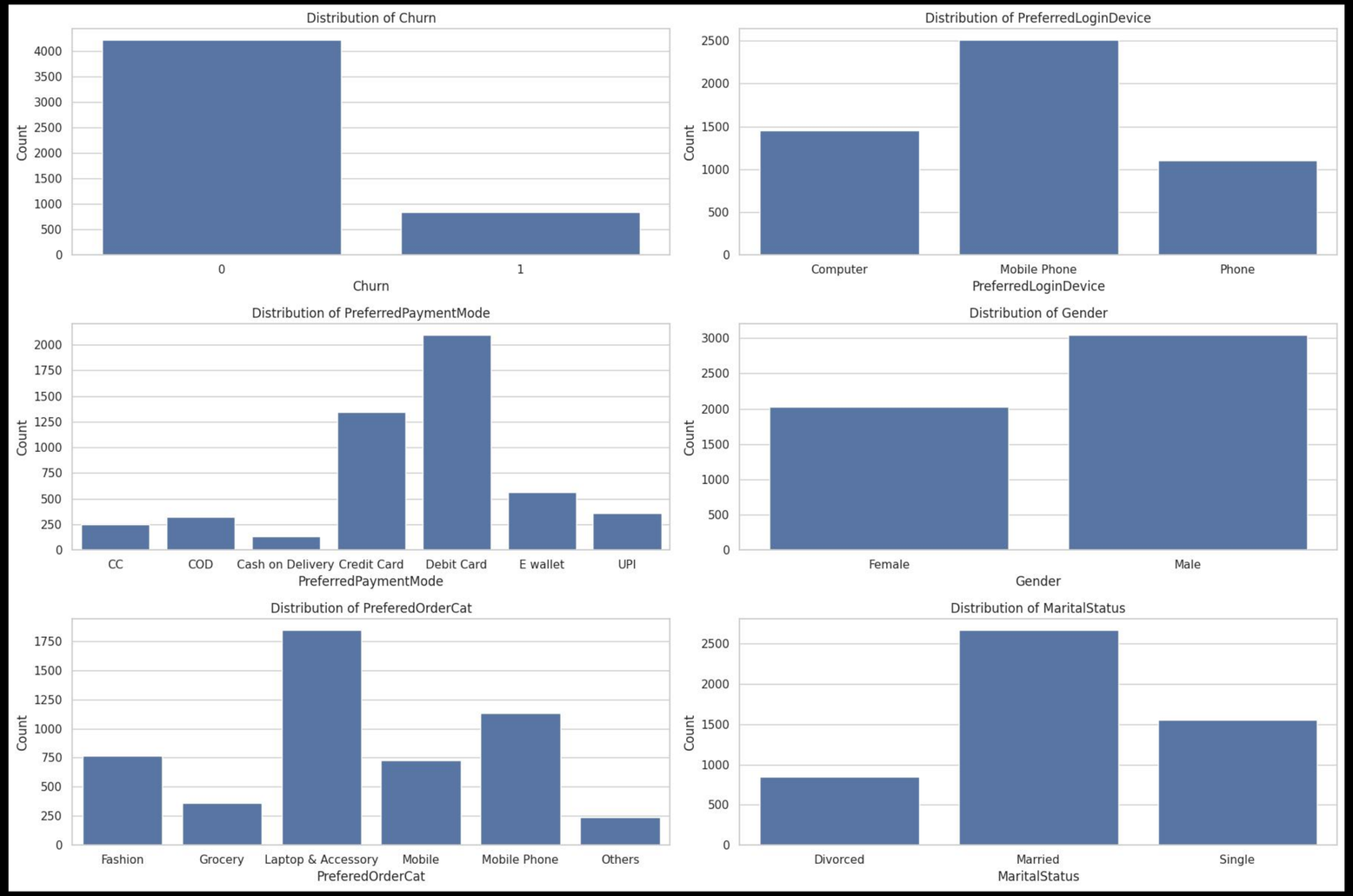
Histogram- Continuous



JADE

Appendix

Distribution Bar Chart



JADE

Appendix

ANOVA Statistic

```
ANOVA Summary:
                                F-Statistic        P-Value
NumberOfDeviceRegistered          67.527174   2.613954e-16
SatisfactionScore                 52.256980   5.594375e-13
Tenure                          1147.394626  6.513029e-227
WarehouseToHome                   30.048312   4.418105e-08
HourSpendOnApp                     1.551101   2.130305e-01
NumberOfAddress                   11.353046   7.588654e-04
OrderAmountHikeFromlastYear        3.105993   7.806490e-02
CouponUsed                         0.109570   7.406476e-01
DaySinceLastOrder                170.116715   2.864433e-38
CityTier                          48.936806   2.986710e-12
OrderCount                         2.348716   1.254487e-01
CashbackAmount                   119.362349   1.762665e-27
```

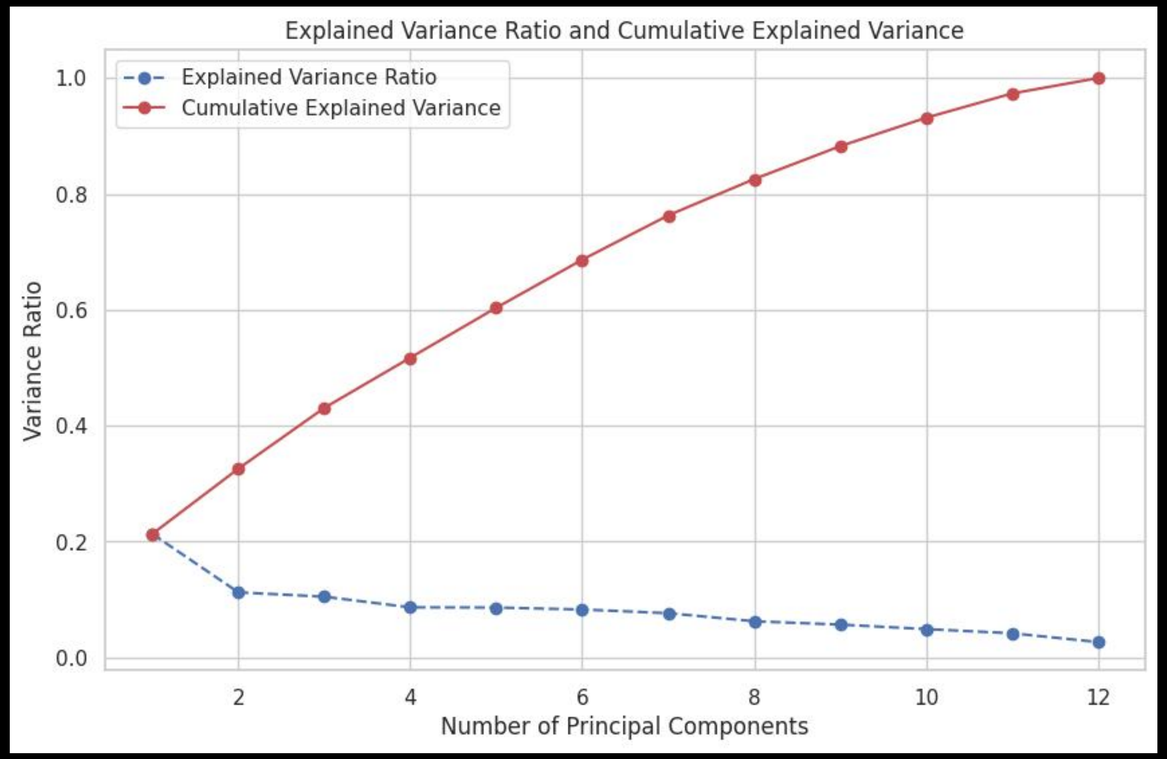
Interpretation

1. NumberOfDeviceRegistered: The ANOVA results indicate a highly significant effect of NumberOfDeviceRegistered on Churn, with an F-Statistic of 67.53 and an extremely low P-Value (2.61e-16). This suggests that the number of devices registered by a customer is a significant predictor of churn.
2. SatisfactionScore: The analysis reveals a significant relationship between SatisfactionScore and Churn, with an F-Statistic of 52.26 and a P-Value of (5.59e-13). This highlights that customer satisfaction scores are a strong determinant of churn likelihood.
3. Tenure: The ANOVA results indicate a highly significant effect of Tenure on Churn, with an F-Statistic of 1147.39 and an extremely low P-Value (6.51e-227). This suggests that the duration of a customer's engagement with the service strongly influences their likelihood to churn.
4. WarehouseToHome: The significant F-Statistic of 30.05 and a low P-Value (4.42e-08) for WarehouseToHome indicate that the distance between the warehouse and the customer's home significantly affects their likelihood of churning.
5. HourSpendOnApp: The results for HourSpendOnApp show an F-Statistic of 1.55 and a P-Value of 0.21, suggesting that the amount of time a customer spends on the app does not have a significant impact on their likelihood to churn.
6. NumberOfAddress: The F-Statistic of 11.35 and a P-Value of (7.59e-04) indicate a significant effect of NumberOfAddress on Churn, suggesting that the number of addresses associated with a customer influences their propensity to churn.
7. OrderAmountHikeFromlastYear: The F-Statistic of 3.11 and a P-Value of 0.08 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.
8. CouponUsed: The analysis shows an F-Statistic of 0.11 and a P-Value of 0.74, indicating that the use of coupons does not significantly affect Churn. This suggests that coupon usage is not a major factor influencing customer churn.
9. DaySinceLastOrder: The high F-Statistic of 170.12 and a very low P-Value (2.86e-38) 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.
10. CityTier: The analysis shows that CityTier has a significant impact on Churn, evidenced by an F-Statistic of 48.94 and a P-Value of (2.99e-12). This suggests that customers from different city tiers exhibit varying propensities to churn.
11. OrderCount: With an F-Statistic of 2.35 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. CashbackAmount: The significant F-Statistic of 119.36 and a low P-Value (1.76e-27) 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.

JADE

Appendix

PCA Analysis

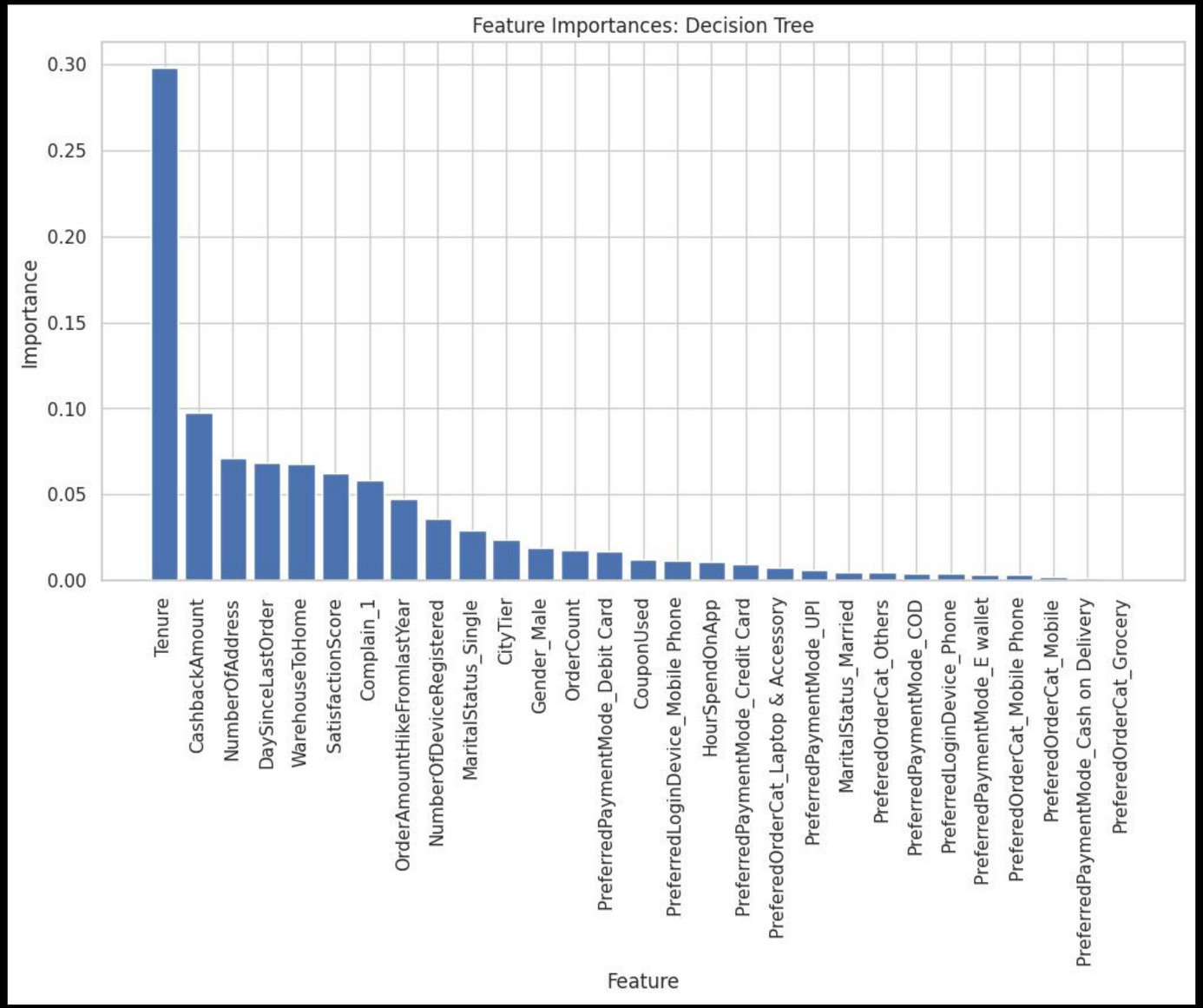


The first principal component (PC1) values range from 0.70 to 2.37, while the second principal component (PC2) values vary between -1.05 and 1.02. This transformation indicates that the original features have been reduced to two key components that capture the most variance in the data. For instance, the first row has a relatively moderate value on PC1 (0.70) but a negative value on PC2 (-1.05), suggesting it contributes differently to the variance captured by these components. In contrast, the second row shows a high positive value on PC1 (2.37) and is almost neutral on PC2 (0.01), indicating a strong influence on the first principal component.

JADE

Appendix

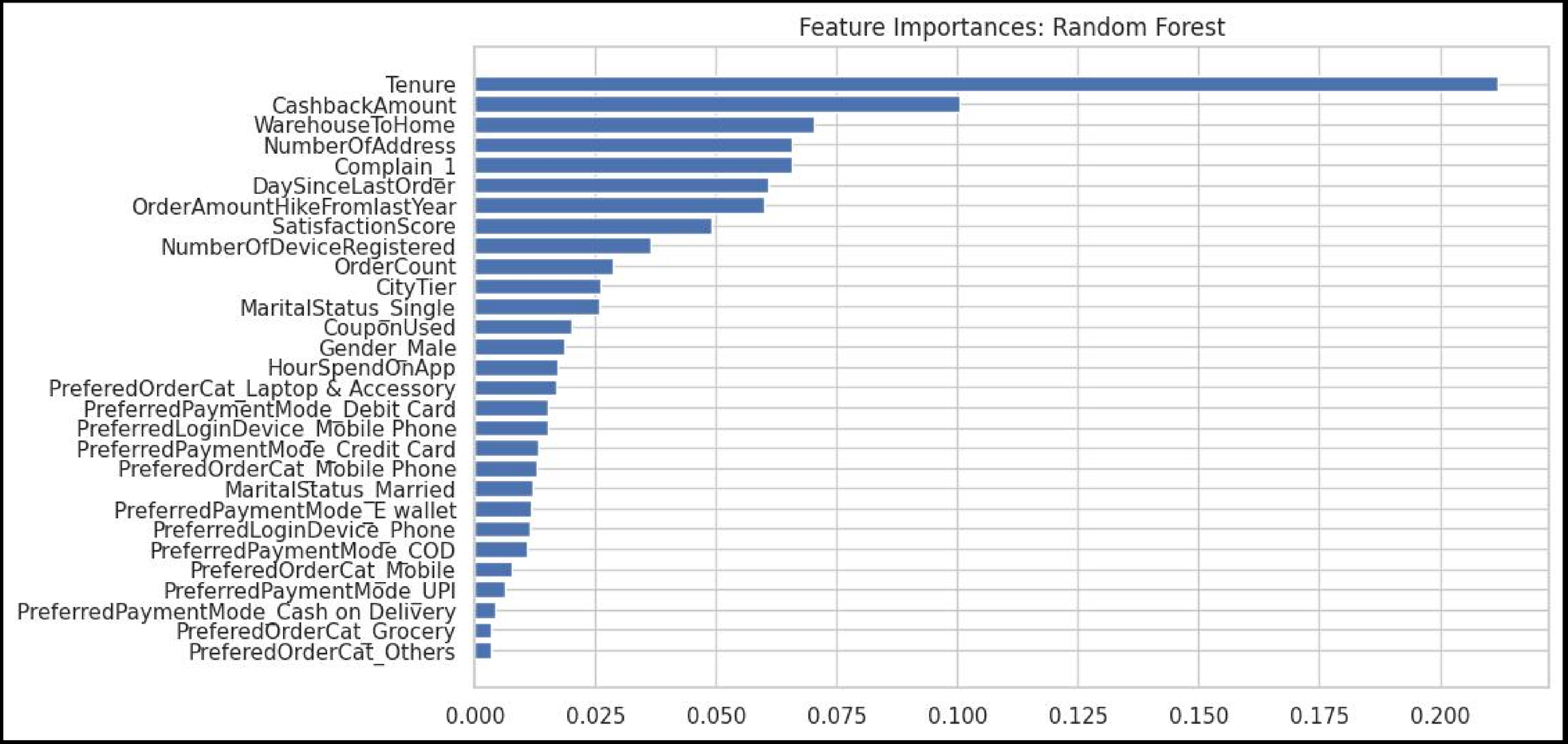
Feature Importances



JADE

Appendix

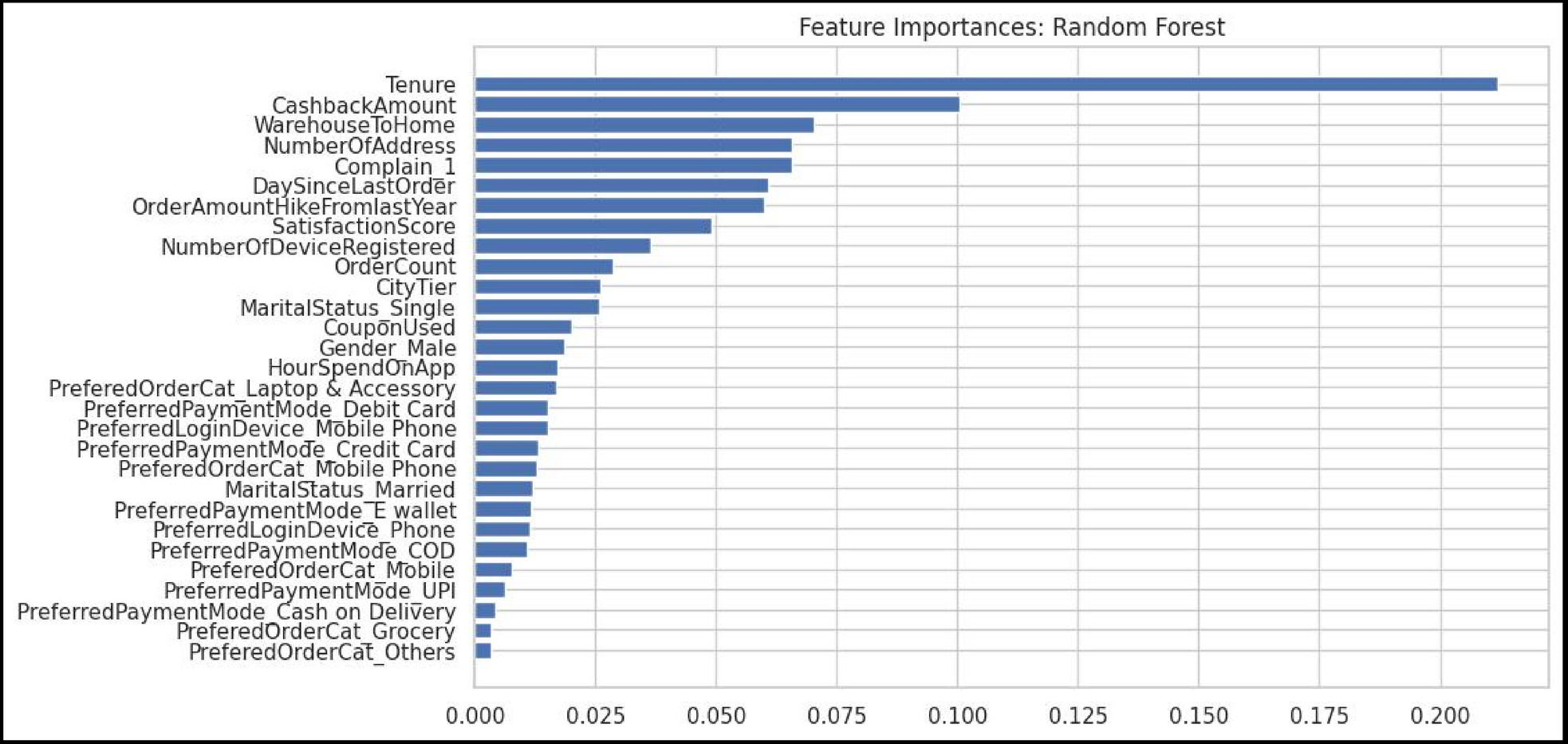
Feature Importances



JADE

Appendix

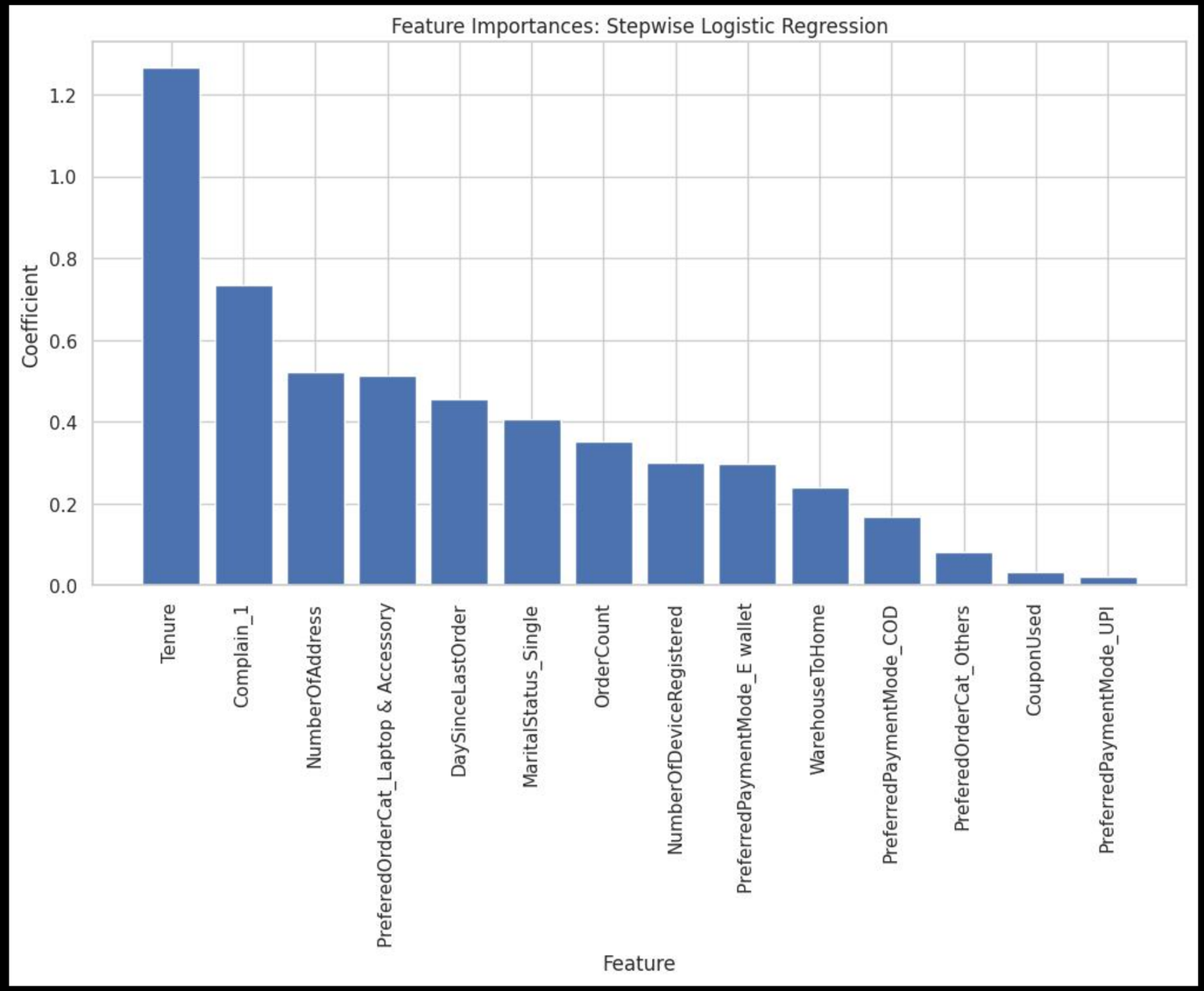
Feature Importances



JADE

Appendix

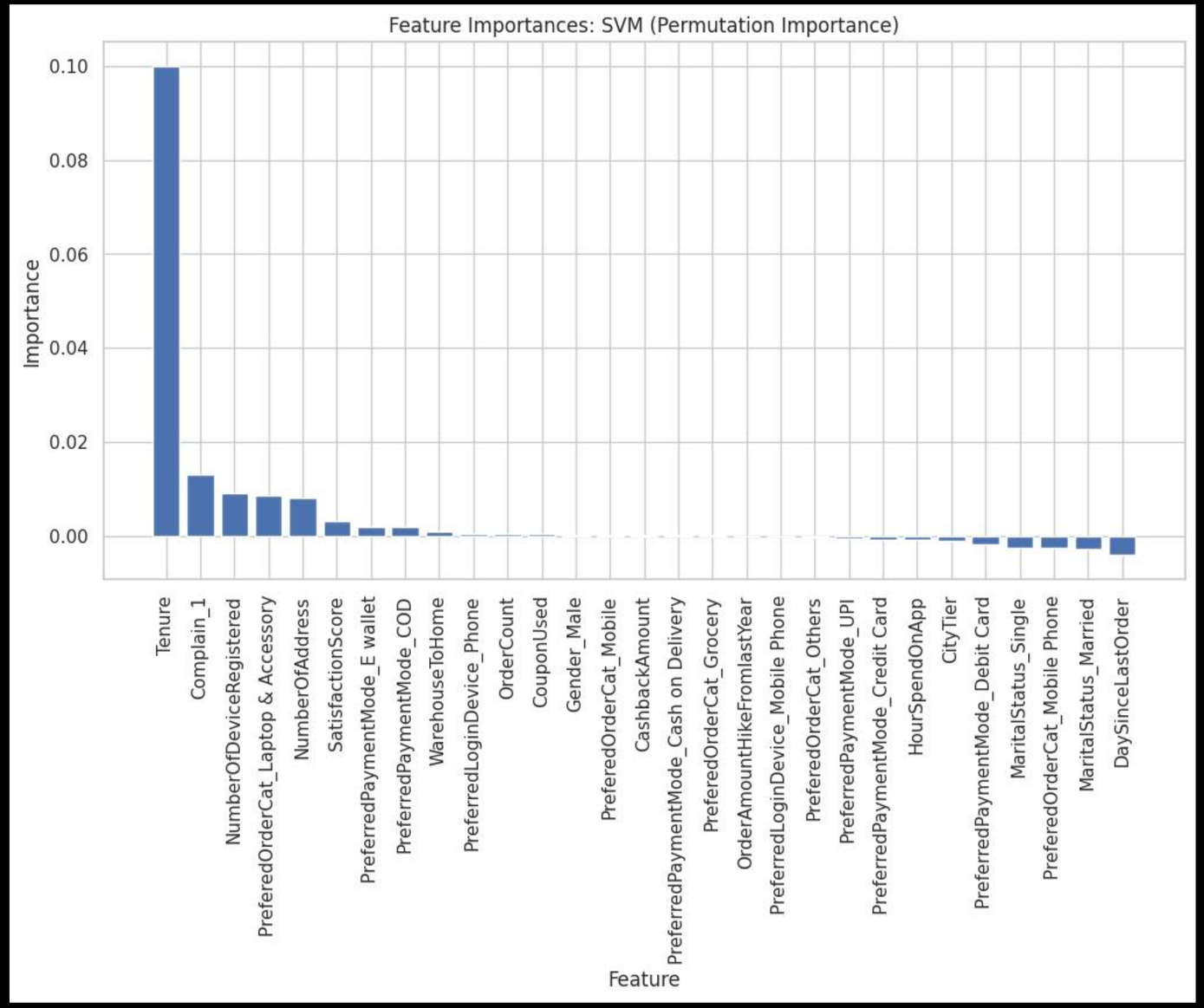
Feature Importances



JADE

Appendix

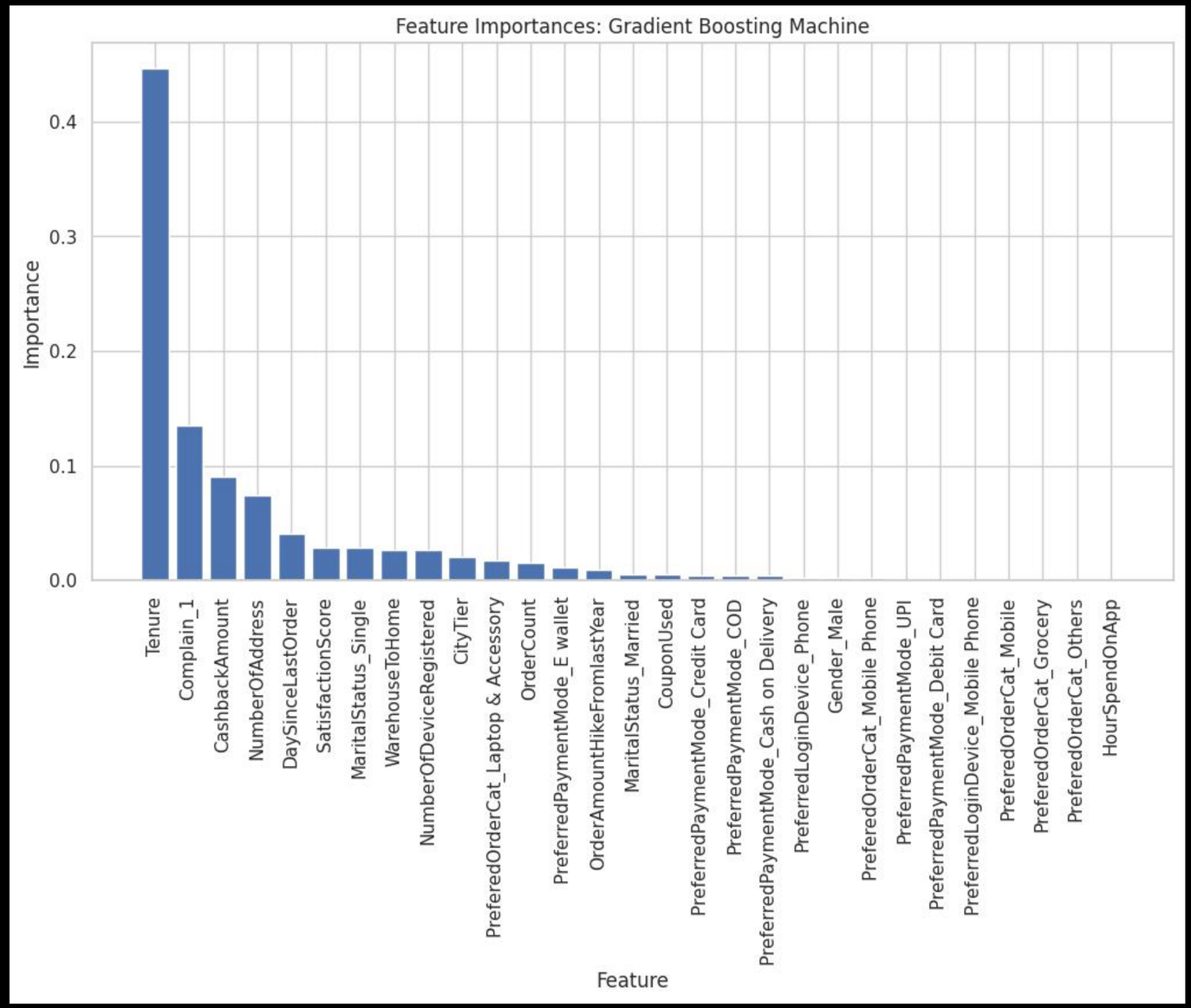
Feature Importances



JADE

Appendix

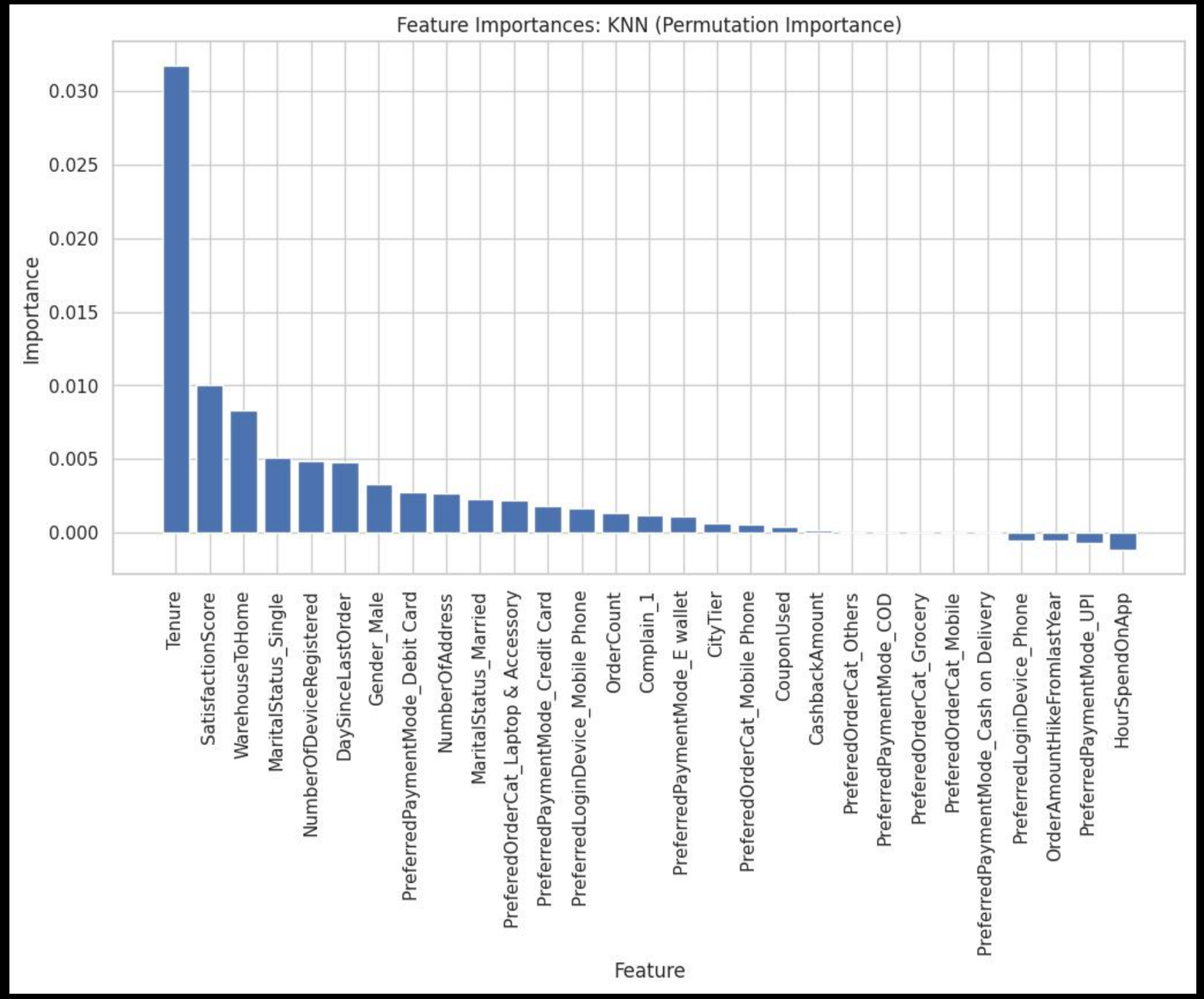
Feature Importances



JADE

Appendix

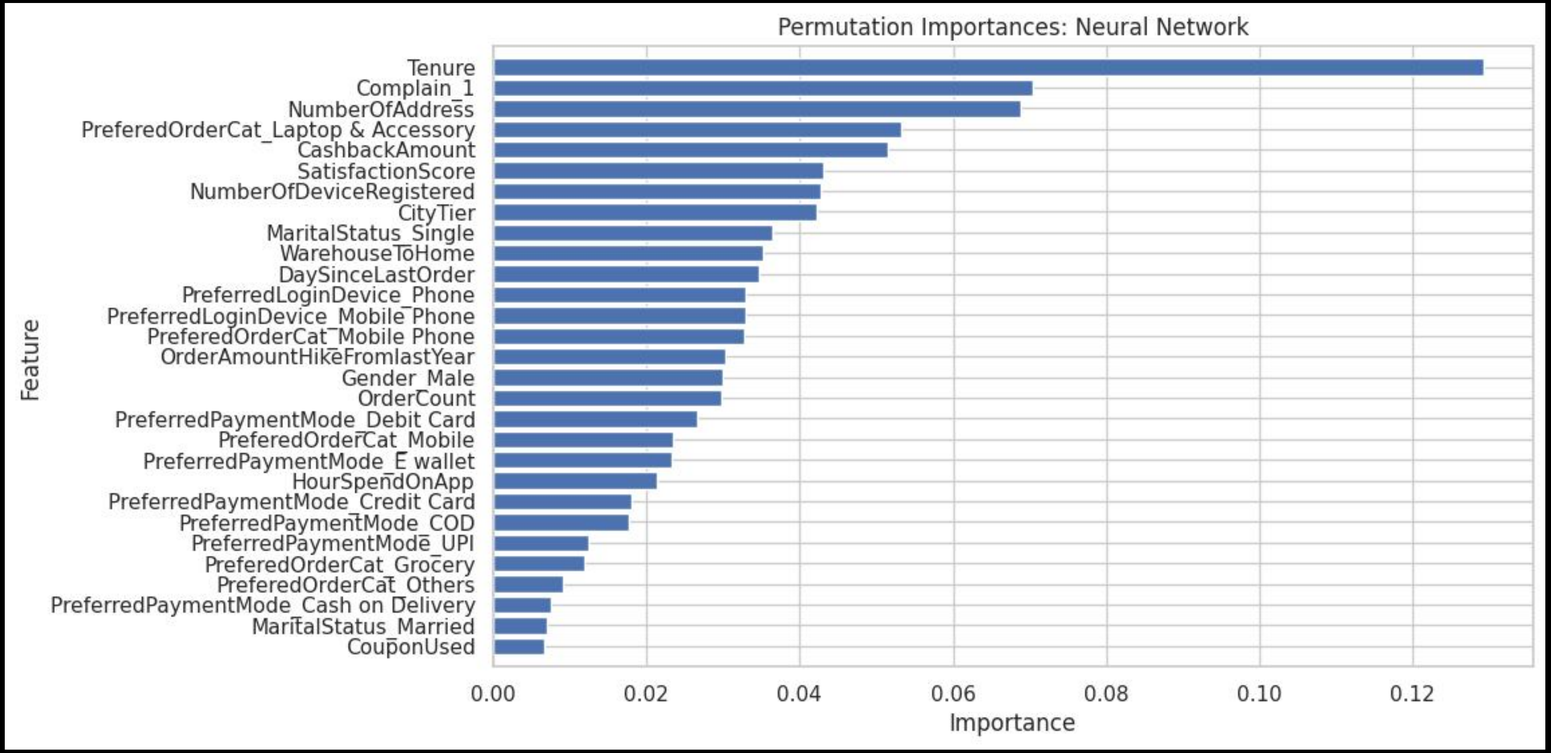
Feature Importances



JADE

Appendix

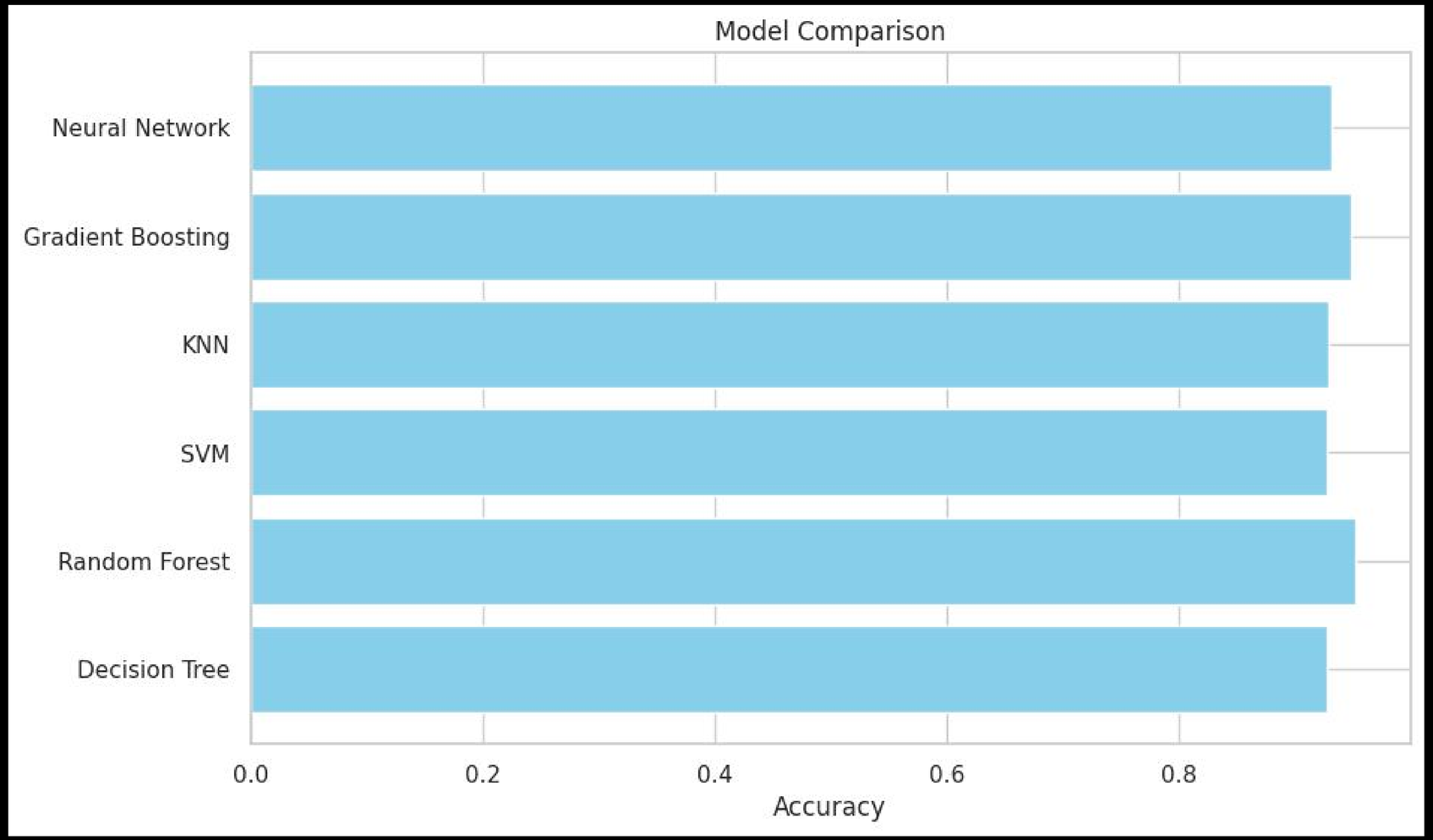
Feature Importances



JADE

Appendix

Model Comparison



JADE

Report Contents