Predicting Customer Churn For E-Commerce

Colin Bowers

Department of Continuing Education, McMaster University

DAT 203: Predictive Modelling & Data Mining

Dr. Haitham Amar

August 4, 2023

Abstract

This study presents a comprehensive analysis and predictive model development on customer churn for an anonymous leading online ecommerce business. By leveraging advanced data analytics and machine learning techniques, the analysis explores historical customer data and transactional patterns to proactively identify potential churners. Key insights reveal the top factors that significantly influence customer churn are: tenure, distance from the warehouse, time spent on the app, the number of devices registered, satisfaction score, and the number of addresses registered. A Random Forest model was developed and optimized to minimize false negatives yielding the ability to predict if a customer will churn with 97% accuracy. Armed with these actionable insights, the company can implement tailored strategies, nurture customer loyalty, and ensure a competitive edge in the dynamic ecommerce landscape.

Introduction

In the competitive world of ecommerce, customer churn is a critical factor of business success. Customer churn, also called customer attrition or turnover, refers to the phenomenon where customers discontinue their engagement with a company resulting in lost revenue and potential market share. Understanding and predicting customer churn is of paramount importance as it allows companies to proactively address any underlying issues and implement targeted retention strategies.

Problem Statement

This study explores customer churn in the context of an anonymous leading ecommerce company and its data on customer demographics and purchasing history. It has two main objectives:

(1) To develop a classification model that would be able to identify any existing customers that would be at risk of leaving.
(2) Provide some interpretations and recommendations that would be useful for account managers and marketing teams to optimize their customer retention strategy.

Data Overview

The data set used in this project belongs to an unnamed leading online ecommerce company. It consists of 5,630 customer records across 20 attributes, including a “Churn” field indicating if a customer has actually churned. This field is the target of our analysis and subsequent modelling. A full list of each variable and its description is given in the Appendix.

Methodology

The analysis and development of the predictive model for customer churn followed the following phases.

Data Preparation. The first phase includes preparing the data for analysis, handling missing values and removing outliers. Categorical variables were transformed using one-hot encoding.

Analysis. By comparing each variable against churn, it was observed which value ranges were associated with higher than average churn rates. Additionally, Pearson correlation coefficients allowed further

understanding of the interdependence between different attributes. These results are given in a separate section called “Supplementary Analysis”.

Modeling. For the development of a predictive model, four classifier algorithms were evaluated using the same training data: Logistic Regression, Decision Tree, Random Forest, and K Nearest Neighbors. The dataset was split into training and testing subsets (sized 80%/20%) to avoid overfitting and sufficiently evaluate performance using unseen data. Performance evaluation was carried out based on the following metrics: accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC). The importance of each feature in each of the models, where available, to identify potential optimization opportunities for feature selection was also assessed.

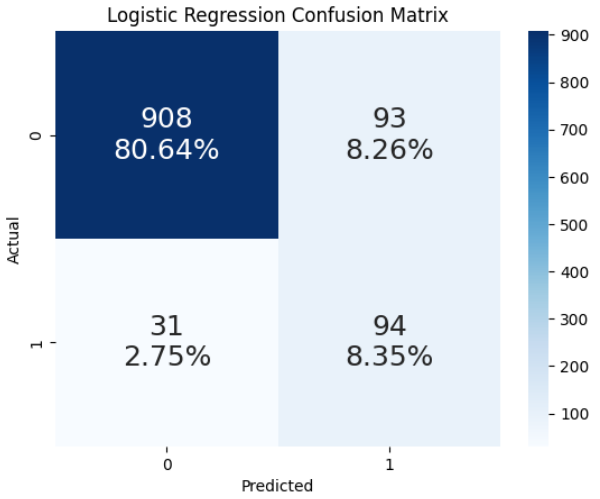
Optimization. The best performing model from the previous phase was then optimized. Recursive feature elimination with cross validation (RFECV) was employed as well as hyperparameter tuning (GridSearchCV) was conducted.

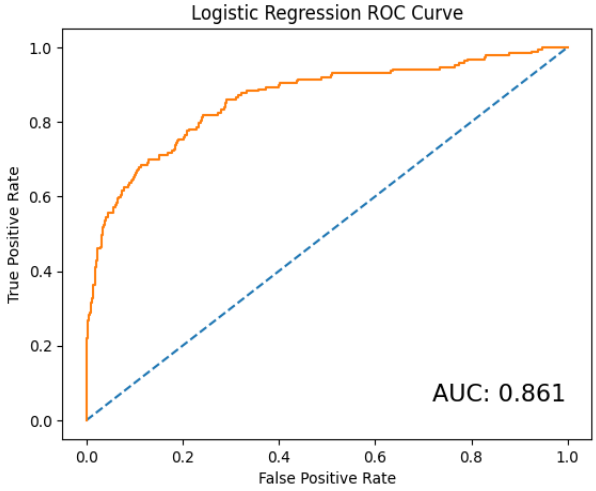
Results

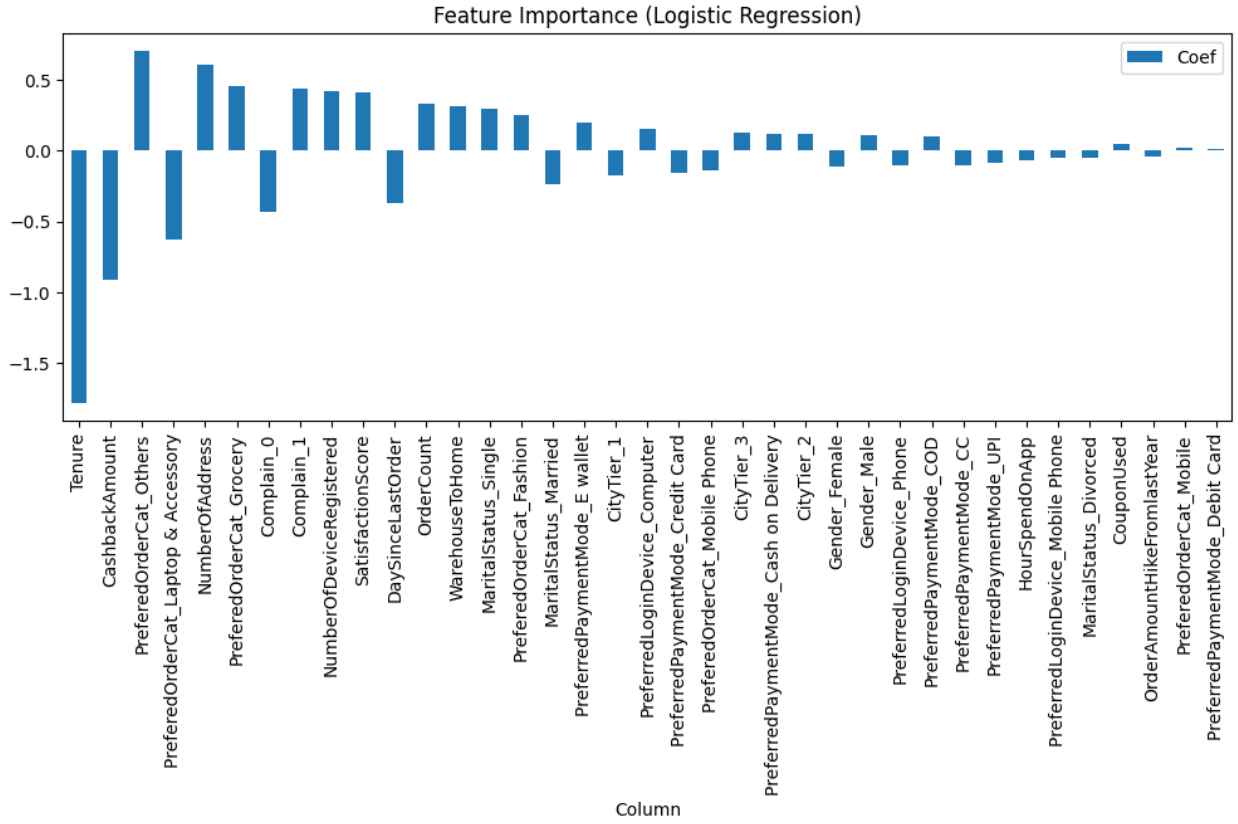
The Random Forest algorithm emerged as the top-performing model with the highest score in most of the performance metrics. The Decision Tree algorithm followed closely making it a viable alternative for prediction.

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.890	0.752	0.503	0.603	0.861
Decision Tree	0.954	0.869	**0.850**	0.859	0.912
Random Forest	**0.969**	**0.981**	0.829	**0.899**	**0.983**
K Nearest Neighbours	0.921	0.938	0.561	0.702	0.928

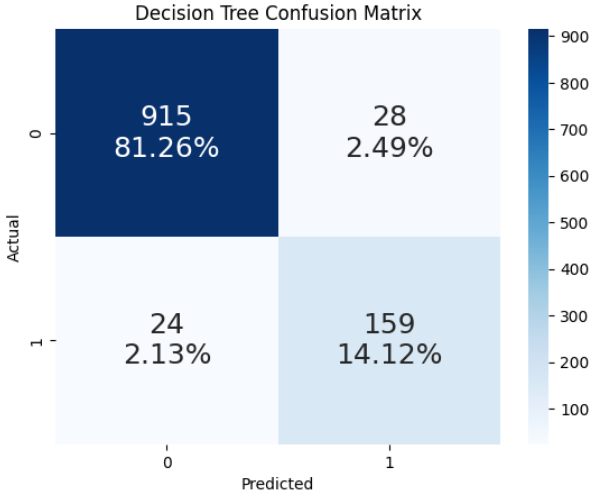
Logistic Regression

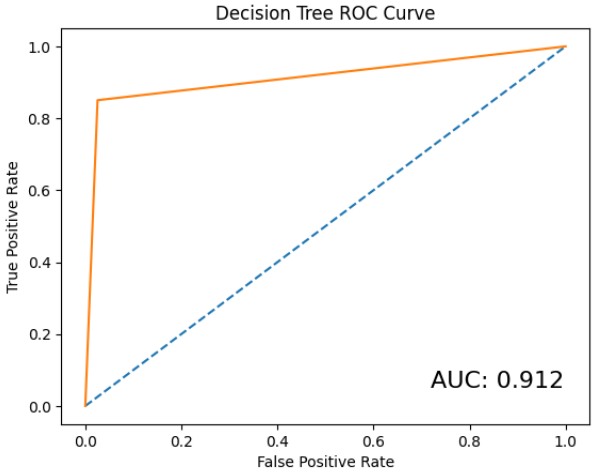


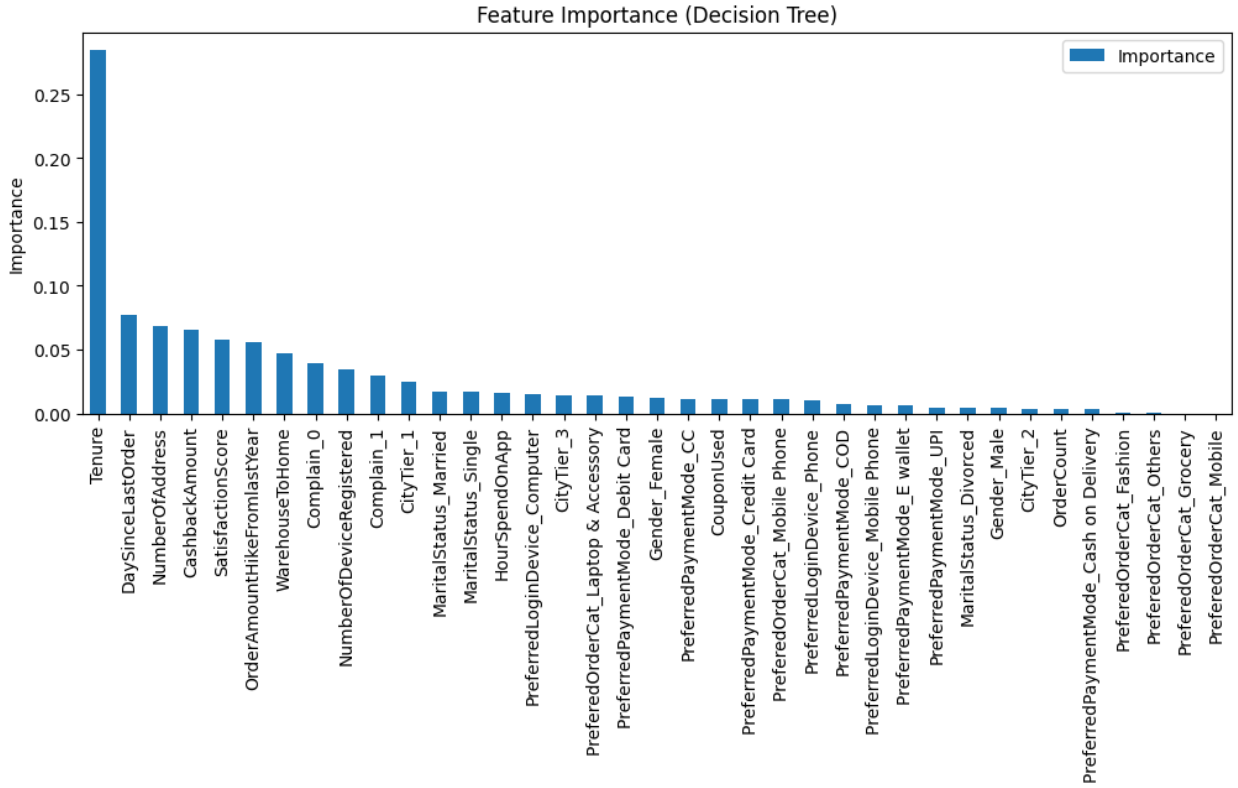




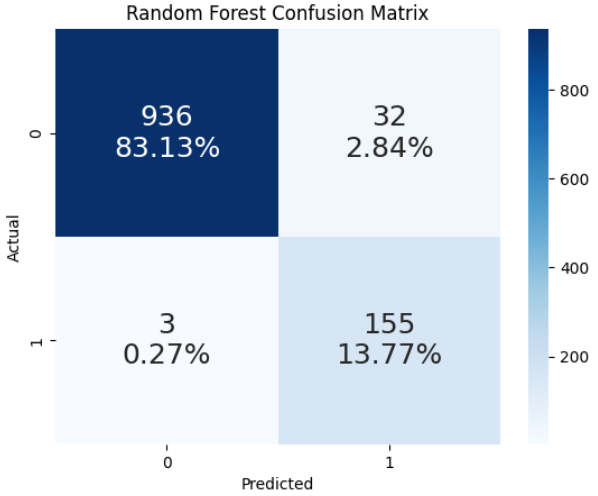
Decision Tree Classifier

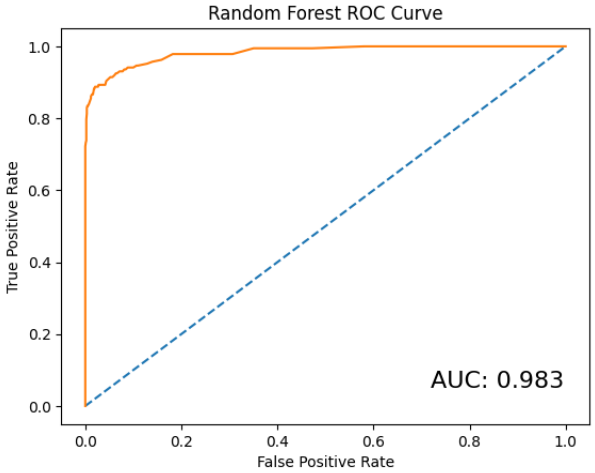


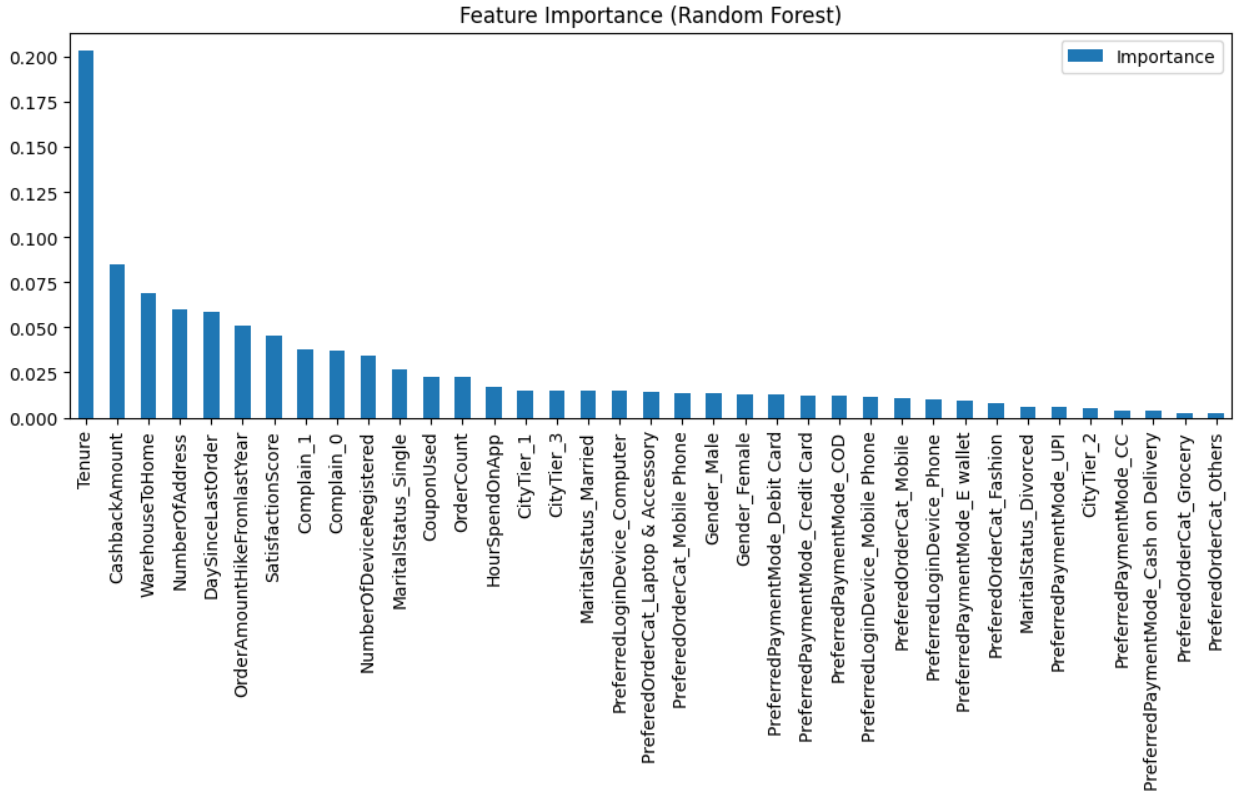




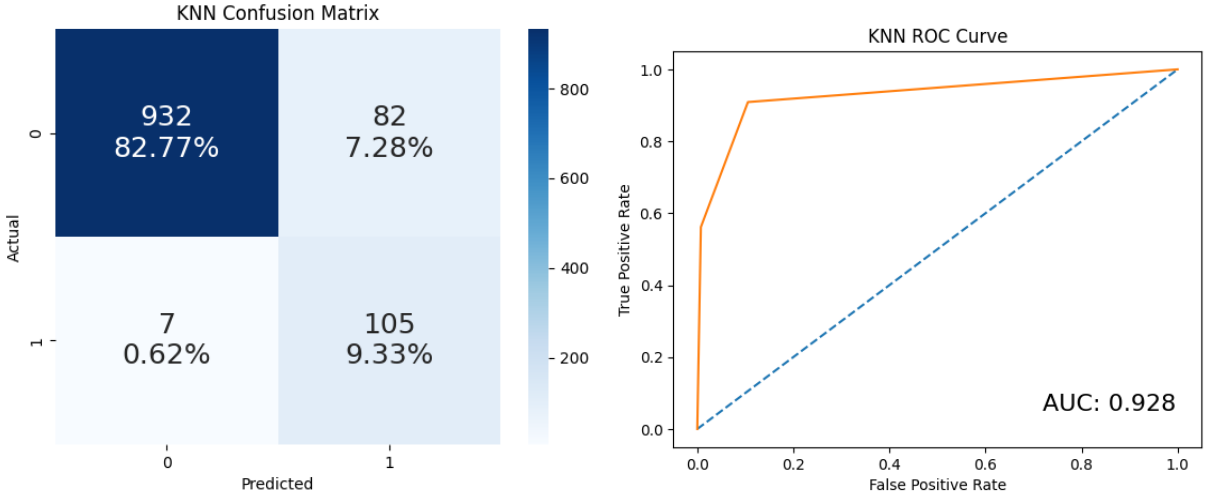
Random Forest Classifier







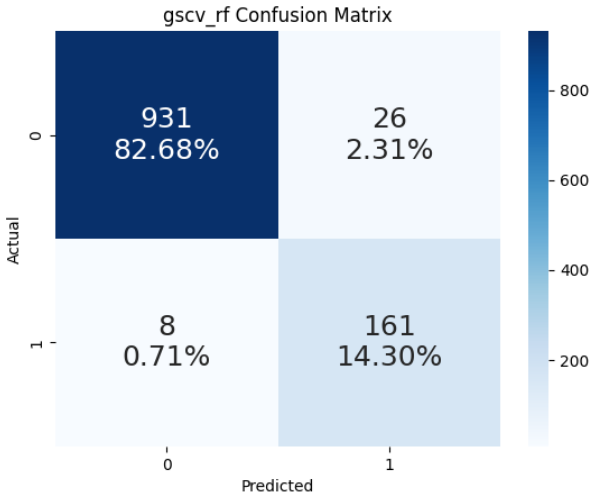
K Nearest Neighbours Classifier

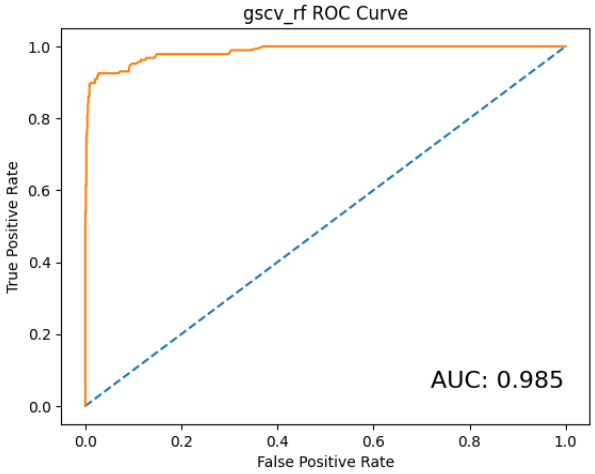


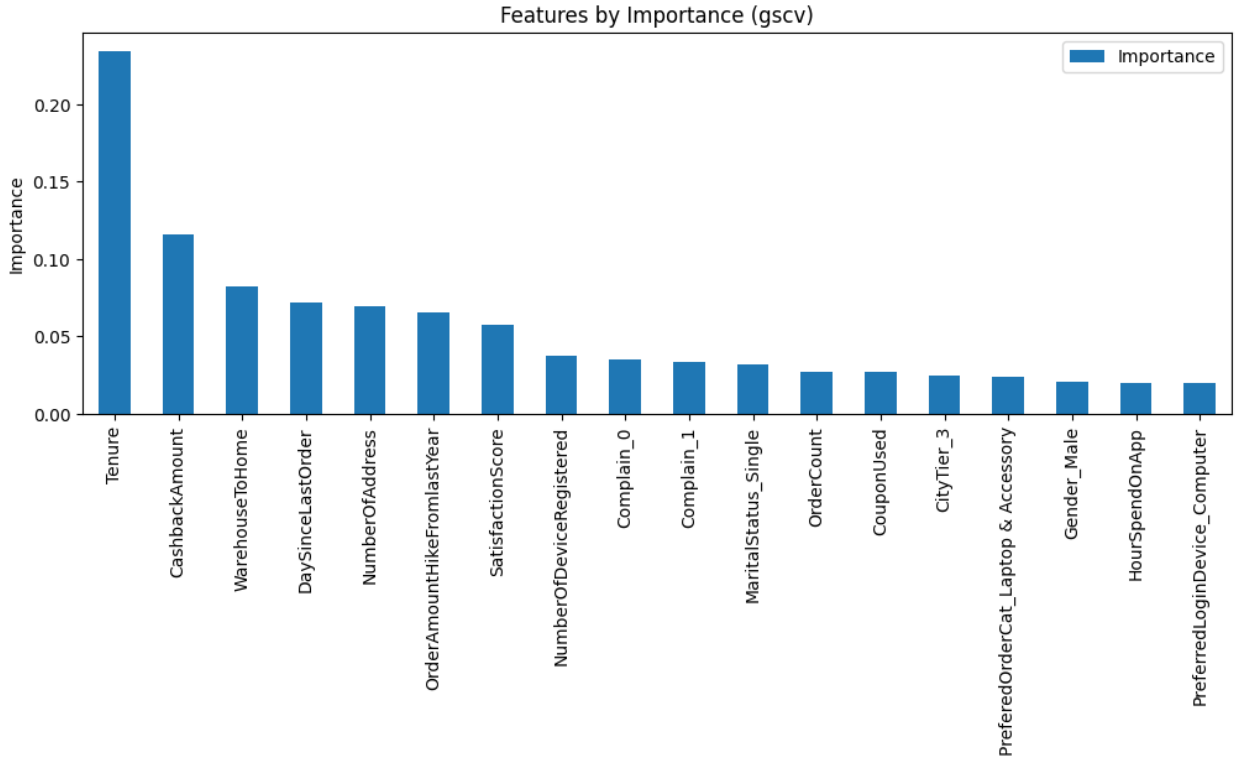
Optimization

The random forest model was selected for further optimization. Recursive feature elimination with cross validation (k=5) was used as a first step in optimizing the model. This determined that the optimal number of features is 18 (down from 38 after the one-hot encoding) while maintaining an accuracy score of 97.0%. For hyperparameter tuning, grid search with cross validation (k=5) was employed. The results of this optimization showed slight improvements overall. A small amount of precision was lost with a nominal amount gained in recall - the next section addresses this tradeoff.

	Accuracy	Precision	Recall	F1 Score	AUC
Before	0.969	**0.981**	0.829	0.899	0.983
After	**0.970**	0.953	**0.861**	**0.904**	**0.985**
Difference	+0.001	-0.028	+0.032	+0.006	+0.003







Discussion

Having analyzed the raw performance measurements of the models for customer churn and performing some optimization using machine learning best practices, a simplified yet high performing predictive model has been developed.

Measuring Performance

Accuracy should not be used as a primary measure since there are many more non-churned customers in our data than those that churned - the proportion is highly unbalanced.

Between precision and recall, In the context of identifying potential high risk customers for the internal sales, marketing and other retention teams to monitor, recall should be prioritized (Majumdar, 2016). By favoring recall, the aim is to minimize false negatives, ensuring that any customer with an increased probability of churning is flagged. That is, it is preferred to flag a customer incorrectly (a false positive) and waste some retention efforts than to not flag them (a false negative) and lose the customer.

This is assuming the company's costs to retain are less than the lifetime value of each customer. In general, acquiring new customers can cost up to five times more than retaining existing ones (Landis, 2022).

Model Generalization and Stability

After training on a large dataset (total of 5,630 samples), the tuned model demonstrated consistent and accurate predictions when evaluated on unseen data (20% of the data, or 1126 rows, were reserved for this purpose). Furthermore, the model maintained its performance across multiple cross-validation folds (k=5 fold in this case).

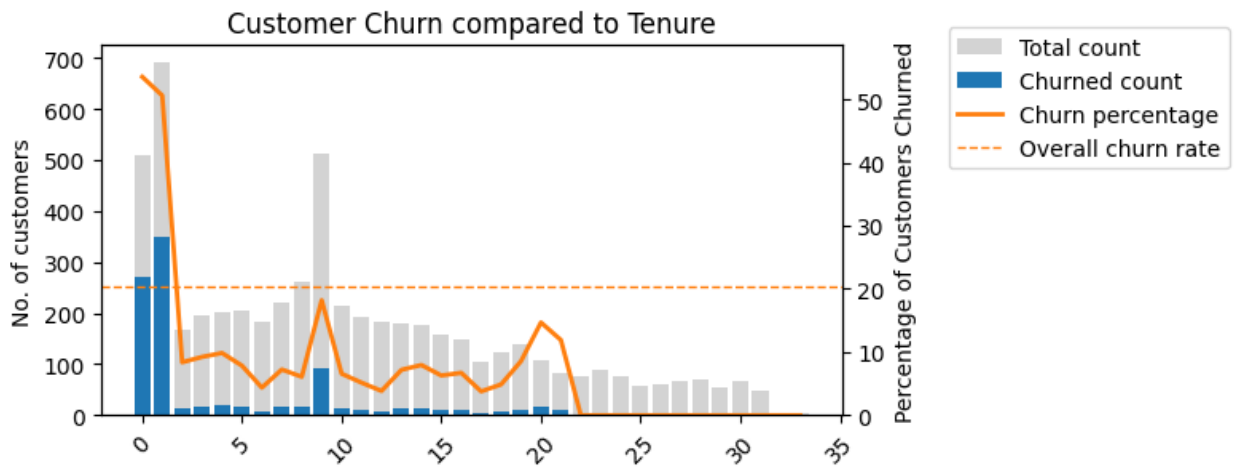
The model's consistent performance with its predictions across these folds, and on previously unseen data, highlights its stability and generalizability for guiding business decisions regarding customer churn.

Supplementary Analysis

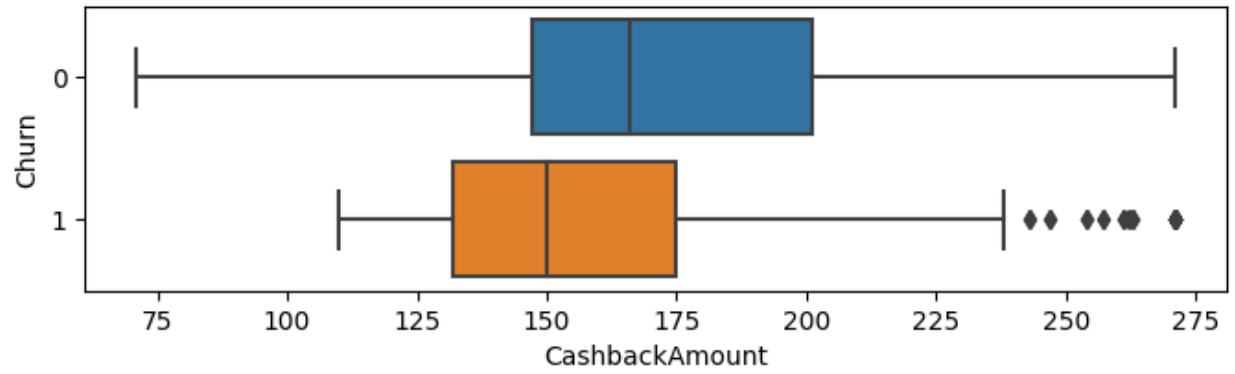
Some additional analysis was performed by comparing the overall churn rate to the rate across the full range of each other variable in the dataset. The following is a summary of results, along with a brief discussion, of the customer attributes where the probability of churning is higher than the average.

Overall Average Churn: 20.25%.

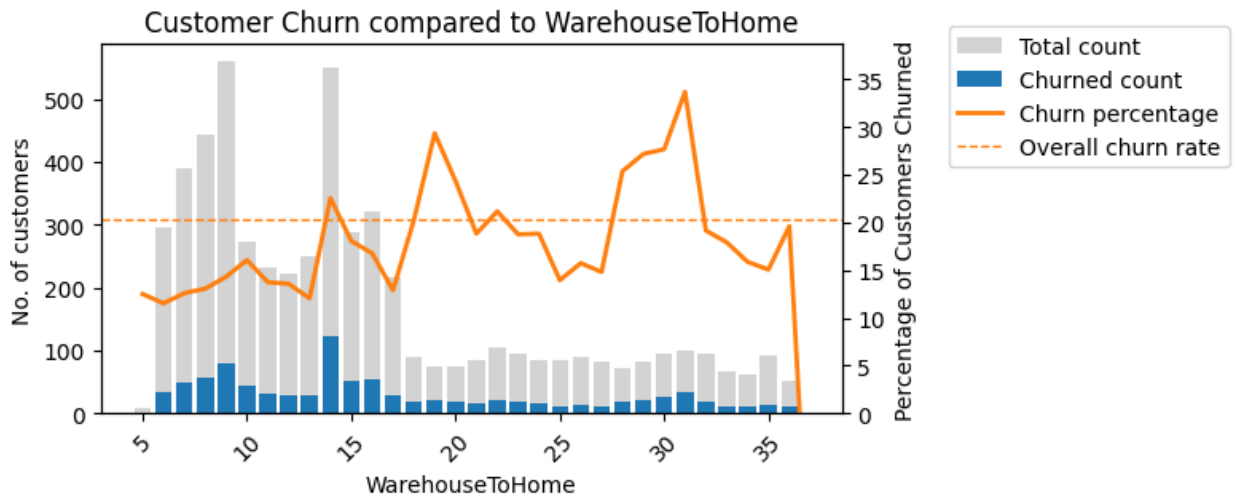
Less than 2 Years of Tenure. Customers with less than 2 years of tenure demonstrated a higher likelihood of churn (~50%). This indicates that newer customers may be more prone to disengagement compared to long-standing ones, warranting focused retention efforts to nurture their loyalty.



Cashback Amount. The mean cashback amount for churned customers was lower than for non-churned customers. Enhancing cash back incentives and other loyalty rewards might positively impact customer loyalty.

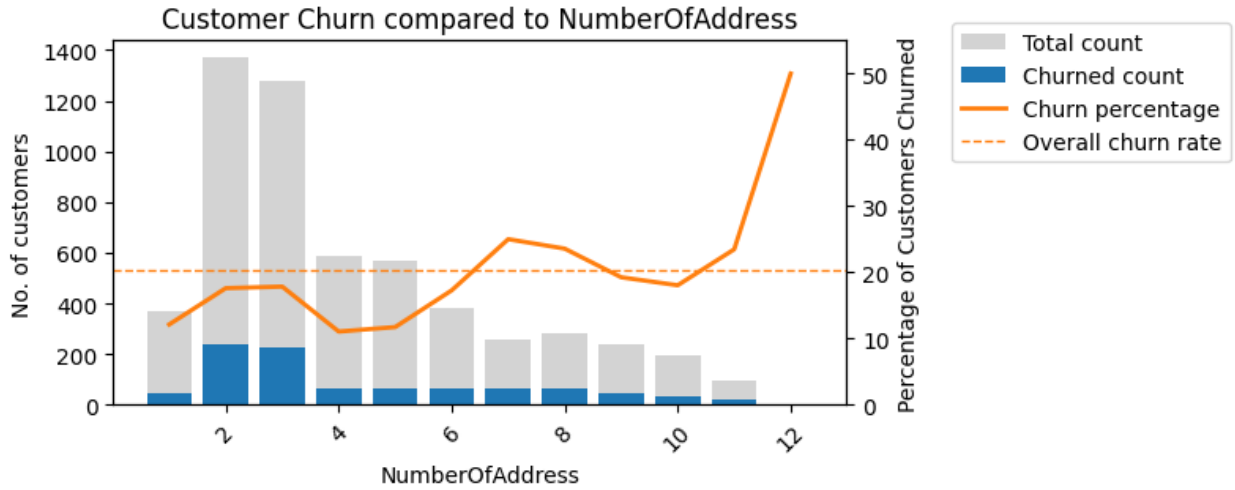


Distance from Warehouse. Customers residing at a further distance from the warehouse, especially beyond 17, demonstrated higher churn rates (~30-35%) with an observed dip back to average in the 23-28 range. Understanding the reasons behind this trend could unlock opportunities for logistics optimization and personalized customer support.

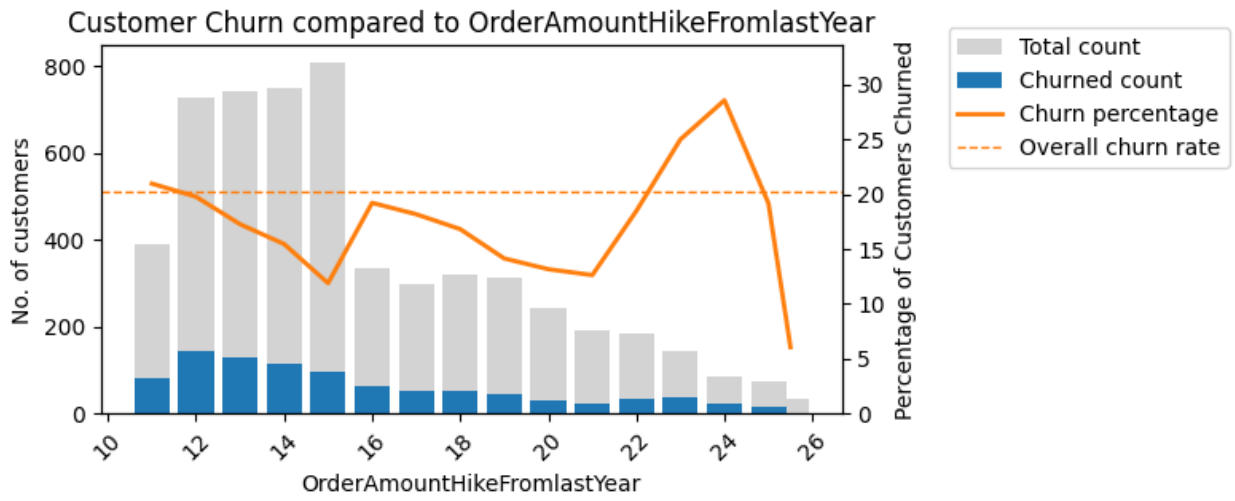


Time Since Last Order. Customers who churned typically had less than 2 days since their last order (35%). Focusing on maintaining consistent engagement and providing personalized incentives might prevent churn.

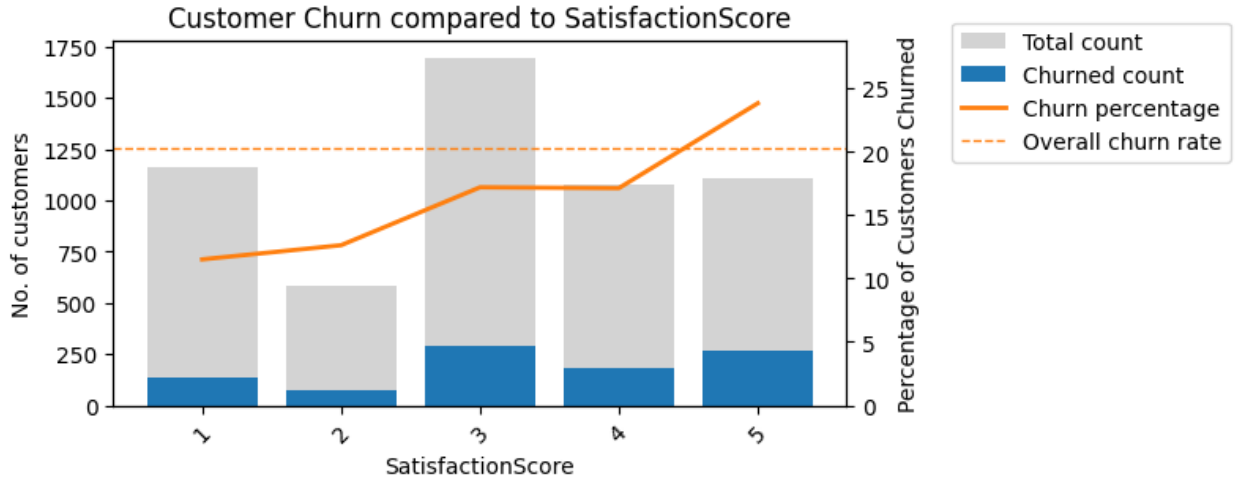
Number of Addresses. Customers with more than 6 addresses had a higher churn rate which increases as the number of addresses increases (22% - 50%).



Order Amount Increase. Customers experiencing a 23-25% increase in their order amount from the previous year demonstrated higher churn rates (29%). Understanding the reasons behind this correlation may inform pricing and discount strategies.

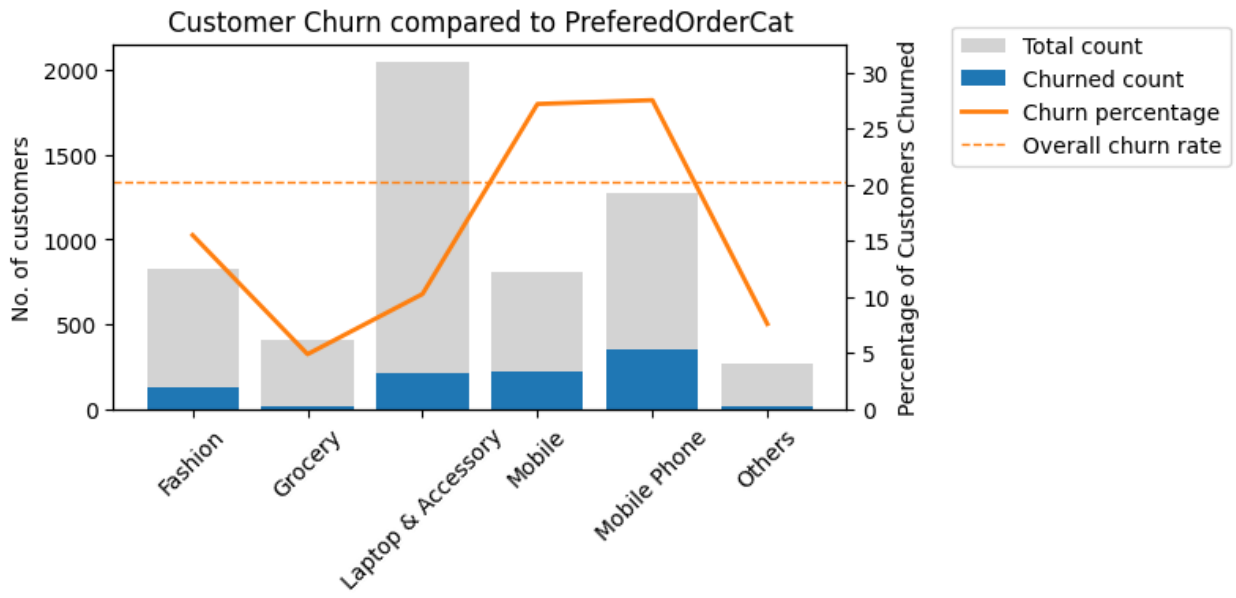


Satisfaction Score Anomaly. Surprisingly, customers with a satisfaction score of 5 showed a higher churn rate (~25%). This anomaly warrants further investigation to ascertain potential underlying factors affecting customer satisfaction and churn behavior.

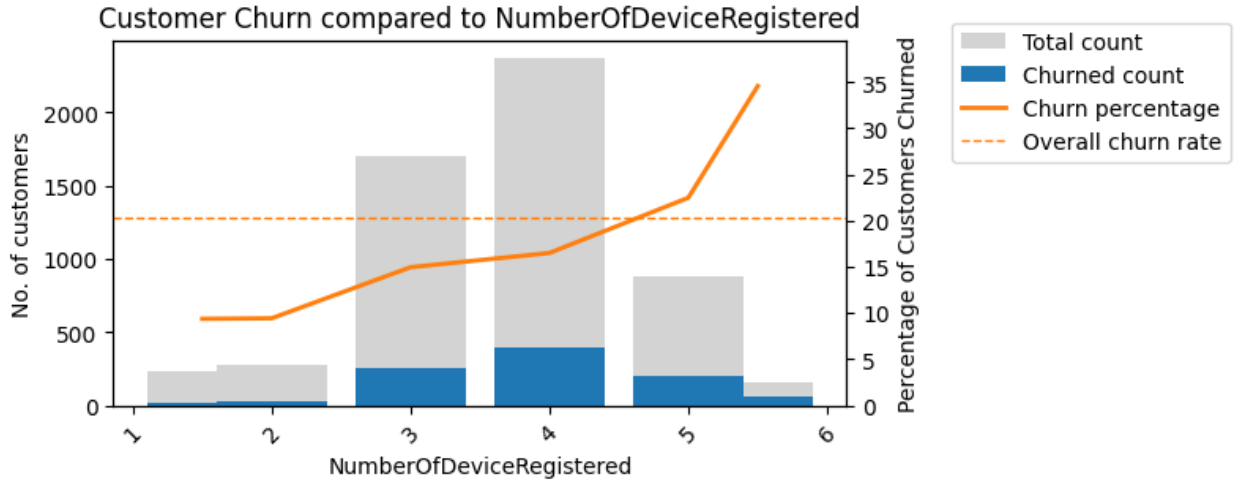


Customer Complaints. Customers who lodged complaints were more likely to churn (32%). The Customer Care Operations teams should continue tracking which customers complain in order to identify those who are at higher risk and report this to Account Managers.

Preferred Order Category. Customers whose preferred order category was Mobile or Mobile Phone exhibited higher churn rates (~28%). Offering tailored incentives and promotions in these categories might boost customer loyalty.



Number of Registered Devices. Customers with 5 or more registered devices were more likely to churn (35%). Implementing targeted cross-device experiences and support may enhance customer engagement and reduce churn risk.



Preferred Payment Methods. Customers who predominantly used Cash on Delivery (COD) and e-wallets had a higher likelihood of churning (28% and 22% respectively). Encouraging alternative payment methods and streamlining payment processes might contribute to better customer retention.

Marital Status - Single. Customers with a marital status of Single were more likely to churn (~28%). Understanding the needs and preferences of this segment could guide tailored retention strategies.

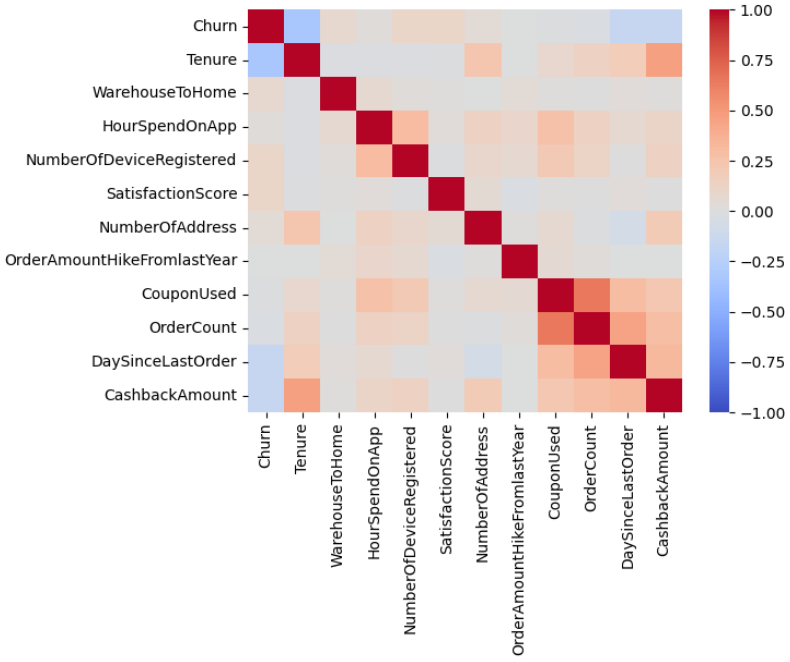
Phone and Preferred Login Device. Customers who have designated their phone - as opposed to a mobile phone or computer - as their preferred device exhibited a slightly higher churn rate (~22%). A potential opportunity to investigate the customer experience when making voice calls via telephone.

City Tier 3. Customers located in City Tier 3 areas were slightly more prone to churn (22%). There is a potential opportunity to create a region-specific marketing to increase customer engagement and mitigate churn.

Male. Males displayed a higher churn rate compared to females (18%), but this was still lower than the overall average indicating that gender is not a significant factor to identify churn.

Bivariate Correlation

The Pearson correlation was calculated for the pairwise combination of each variable and presented here in a visual heatmap plot. Key takeaways here include: tenure, days since last order and cash back amount all appear to reduce the likelihood of churn. On the other hand, the number of devices, satisfaction score and distance to warehouse causes an increase.



Conclusion

This study on customer churn prediction for the unnamed company yielded valuable insights and an accurate predictive model. Through data analysis and machine learning techniques, key attributes were identified that significantly influence customer churn and developed a Random Forest model with an impressive accuracy of 97%.

The results of this study have significant implications for the company's business operations and strategy. By proactively identifying potential churners with precision-focused retention efforts, the company can maximize customer loyalty and minimize revenue losses. Tailored marketing campaigns, optimized logistics, and improved mobile app user experience are some potential areas of focus to enhance customer engagement and satisfaction. Moreover, addressing customer complaints promptly and offering personalized incentives can further bolster customer retention. The insights provided by the model, combined with continuous monitoring and refinements, will enable the company to strengthen its market position, drive sustainable growth, and thrive in the competitive ecommerce landscape.

References

Landis, T. (April 12, 2022). *Customer Retention Marketing vs. Customer Acquisition Marketing.* Outbound Engine. https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing/

Majumdar, Abhik. (June 18, 2016). *Reducing User Churn with Machine Learning – Precision and Recall.* Vidora. https://www.vidora.com/ml-in-business/reducing-user-churn-with-machine-learning-precision-and-recall/

Appendix

Data Dictionary

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