Churn Analysis, Prediction & Business Suggestions

—— Focusing on live streaming customers

Group members:

A0296697Y HE Wei
A0297466J HUANG Wenjie
A0296660U JI Xiangtian
A0296622X SU Hao
A0221781J YANG Zikun

Project Report: Churn Analysis, Prediction, and Business Suggestions

Group member:

- A0296697Y HE Wei
- A0297466J HUANG Wenjie
- A0296660U JI Xiangtian
- A0296622X SU Hao
- A0221781J YANG Zikun

1. Introduction

This project investigates churn prediction by analyzing various customer behaviors, demographics, and transaction characteristics. By understanding the factors that influence churn, we aim to provide actionable insights to improve customer retention strategies.

2. Data Collection and Preprocessing

The dataset, comprising 5,630 rows, was sourced from TIANCHI, Alibaba Cloud's developer competition platform, capturing live-stream customer interactions with the its platform.

In the data preprocessing stage, to handle missing values and outliers in our e-commerce dataset, we applied specific imputation strategies based on each field's characteristics.

For fields like HourSpendOnApp and WarehouseToHome, where the data showed clear outliers and a skewed distribution, we employed median imputation to reduce the impact of extreme values. Similarly, OrderAmountHikeFromLastYear, CouponUsed, and DaySinceLastOrder also exhibited skewed distributions, so median imputation was used to ensure stability in the dataset.

For the HourSpendOnApp field, where values were relatively centered around the mean, mean imputation was chosen to preserve this central tendency. In the case of OrderCount, missing values were considered to represent no orders placed in the previous month; therefore, we utilized special value imputation by filling these values with 0 to retain the semantic accuracy of the field.

Overall, by employing diverse imputation methods (median imputation, mean imputation, and special value imputation), we minimized the influence of missing values and outliers while maintaining the data distribution's inherent characteristics, thereby providing a solid foundation for further analysis and modeling.

In the subsequent EDA process, we will handle outliers using the IQR method to identify outlier data points. These outliers will then be removed prior to model training.

3. Exploratory Data Analysis

Hypothesis

The project explores ten hypotheses regarding churn behavior, including device preference, age group, and discount amounts, each hypothesized to impact customer churn differently. Details are listed below.

- **Preferred Login Device vs. Churn Rate:**

  Users who access the platform via PC have a higher churn rate compared to those using mobile phones, as PC users may engage less deeply with the platform.

- **Age Group vs. Churn Rate:**

  Users in younger age groups (e.g., 18-25) may exhibit higher churn rates, as they might be more inclined to explore other platforms compared to older age groups.

- **Tenure vs. Churn Rate:**

  Users with shorter tenure periods (e.g., less than 6 months) exhibit higher churn rates, as new users may not have yet formed loyalty or attachment to the platform.

- **Time Since Last Order vs. Churn Rate:**

  Users who made their last order more recently are more likely to churn, as this could indicate that they have obtained what they need and may not return to the platform until they have another need, if at all.

- **Marital Status vs. Churn Rate:**

  Single users have a higher churn rate compared to married users, possibly due to differing priorities and engagement levels with the app.

- **Discount Amount vs. Churn Rate:**

Users who receive higher discounts are less likely to churn, as discounts may encourage customer loyalty and repeat purchases.

- **Order Category Preference vs. Churn Rate:**

  Users who prefer certain order categories, like groceries, may have a lower churn rate due to the recurring need for essentials, whereas users with preferences for one-time purchases may churn more readily.

- **Complaints vs. Churn Rate:**

  Users who lodge complaints are more likely to churn, suggesting that dissatisfaction or unresolved issues significantly impact customer retention.

- **Hours Spent on App vs. Churn Rate:**

  Users who spend fewer hours on the app per session are more likely to churn, indicating that a lack of engagement is correlated with higher churn rates.

- **City Tier vs. Churn Rate:**

  Users in lower-tier cities might have lower churn rates due to fewer competitive delivery or service options, whereas users in higher-tier cities may face more alternatives, potentially affecting churn behavior.

3.1 The correlation matrix analysis

The correlation matrix provides critical insights into the relationships between various numerical features within the dataset, especially with respect to churn, customer behavior, and promotional strategies. This section outlines these relationships and discusses their potential implications.

3.1.1 Churn and Related Features

(1) Churn and Complain: There exists a notable positive correlation between churn rate and the number of complaints (correlation coefficient approximately 0.25), suggesting that users who have registered more complaints are more likely to discontinue their engagement with the platform. This relationship highlights the potential impact of service quality on customer retention. Reducing the number of complaints through improved service quality may serve as an effective strategy to decrease churn rates.

(2) Churn and Tenure: The churn rate exhibits a negative correlation with customer tenure (correlation coefficient around -0.31), indicating that users with a longer tenure on the platform are less likely to churn. This negative correlation suggests that longer-term customers tend to develop a stronger sense of loyalty or attachment to the platform. Consequently, strategies focused on extending customer tenure, particularly for new users, could contribute to enhanced retention rates.

(3) Churn and DiscountAmount: The correlation between churn rate and discount amount is slightly negative (correlation coefficient approximately -0.15), implying that discounts may play a role in reducing churn. This correlation could indicate that discounts increase users' willingness to engage with the platform, thereby lowering the likelihood of churn. As such, offering discounts or promotional activities may be a viable retention strategy.

3.1.2 Significant Correlations Among Other Features

(1) DiscountAmount and Tenure: A relatively high positive correlation (approximately 0.45) exists between discount amount and tenure. This correlation suggests that long-term users are more likely to receive discounts, possibly as part of a retention strategy. Such a strategy could indicate that discount policies are effective in reinforcing customer loyalty, as longer-tenured users might receive more incentives to continue their engagement with the platform.

(2) CouponUsed and OrderAmountHikeFromLastYear: Coupon usage exhibits a moderate positive correlation with the increase in order amount from the previous year (correlation coefficient around 0.62), suggesting that coupons may stimulate users to increase their spending. This correlation indicates that providing coupons can be an effective method for encouraging users to increase their order amounts. As a result, the platform might consider regularly issuing coupons to promote higher spending among users.

(3) OrderCount and DaySinceLastOrder: There is a moderate positive correlation between the total number of orders (OrderCount) and the days since the last order (DaySinceLastOrder), with a correlation coefficient of approximately 0.40. This suggests that users who place orders more frequently tend to have shorter intervals between their orders. Such frequent engagement may characterize a subset of highly active users, who could benefit from targeted recommendations or promotional offers to further enhance their loyalty and spending behavior.

3.1.3 Analysis of Low-Correlation Features

Most feature pairs exhibit low correlation values, with correlation coefficients close to zero, indicating weak or insignificant linear relationships between them. This observation implies that many user behaviors and preferences may be influenced by complex, non-linear factors that are not captured through simple linear correlations among these features. Therefore, future analyses might benefit from incorporating additional variables or employing more complex modeling techniques, such as decision trees or random forests, to explore potential non-linear relationships in user behavior patterns

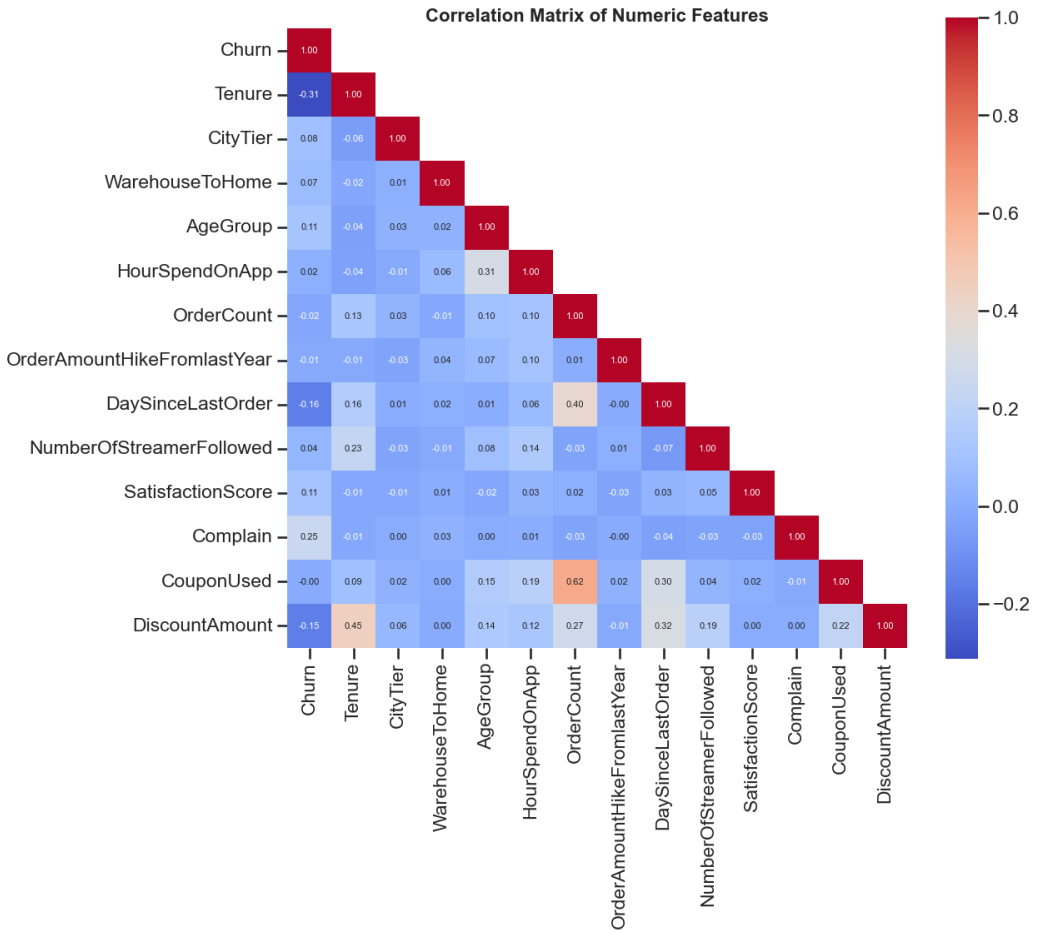


Figure 3.1.1 Correlation Matrix

3.2 Churn Distribution Analysis

The bar chart displays the distribution of churn within the dataset, with churn represented as a binary variable where 0 indicates customers who did not churn, and 1 indicates customers who did. The data reveals a significant imbalance between the two classes:

- **Non-Churned Customers (0):** There are 4,682 instances where customers did not churn, comprising the majority of the dataset.
- **Churned Customers (1):** Only 948 instances represent churned customers, indicating that the churned group is considerably smaller.

This imbalance has important implications for any predictive modeling aimed at churn analysis. When one class is disproportionately represented, models may be biased towards predicting the majority class, potentially resulting in a high accuracy that does not accurately capture the churn dynamics.

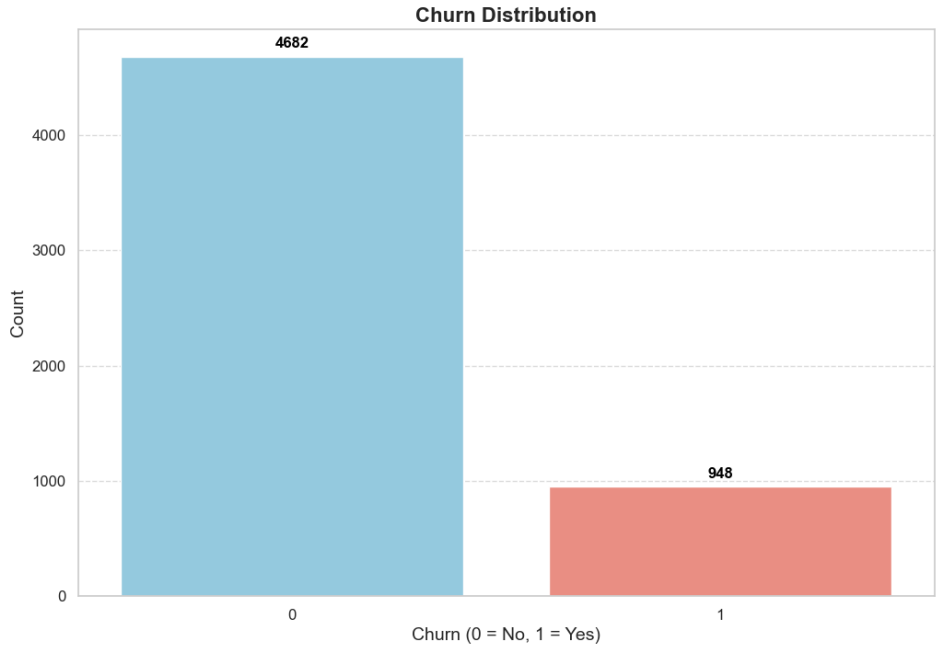


Figure 3.2.1 Churn Distribution

3.3 Preferred Login Device Analysis

The following section examines the distribution of preferred login devices among users and its relationship with churn rates. Two visualizations—one showing the general distribution of preferred devices and another analyzing churn rates by device type—provide insights into user engagement patterns and potential device-specific retention strategies.

3.3.1 Distribution of Preferred Login Device

The pie chart illustrates the breakdown of users' preferred login devices:

- **Mobile Phone:** Accounts for the majority, with 49.1% of users preferring to access the platform via mobile.
- **Pad:** Represents 29.0% of the user base.

- **PC:** Comprises 21.9% of the users.

This distribution suggests that nearly half of the users prefer mobile access, highlighting the importance of optimizing the platform's mobile experience to meet user expectations.

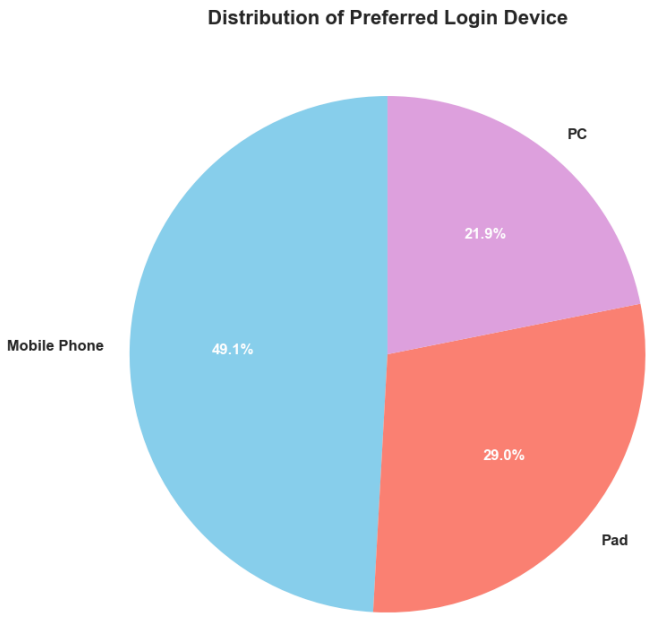


Figure 3.3.1 Preferred Login Device Distribution

3.3.2 Churn Analysis by Preferred Login Device

The bar chart compares churn rates across the three device types:

- **Pad Users:** Show a churn rate of 19.8%, indicating a higher risk of churn compared to mobile users but slightly lower than PC users.
- **PC Users:** Have the highest churn rate, with 22.4% of PC users discontinuing their engagement. This suggests potential issues with the desktop experience that may be contributing to customer attrition.

- **Mobile Phone Users:** Exhibit the lowest churn rate at 12.6%, which is significantly lower than the other two groups. This lower churn rate may reflect a stronger engagement or a more satisfactory user experience on mobile devices.

These findings have several practical implications:

- **Device-Specific Retention Strategies:** Given the higher churn rates for PC and Pad users, targeted improvements to the desktop and tablet experiences could enhance user retention. For example, optimizing the user interface or offering device-specific promotions could potentially lower churn rates for these groups.
- **Mobile-Centric Focus:** With nearly half of the users preferring mobile and a relatively low churn rate among mobile users, investing in mobile platform enhancements appears critical. Ensuring a seamless mobile experience could help maintain low churn and further strengthen user engagement.

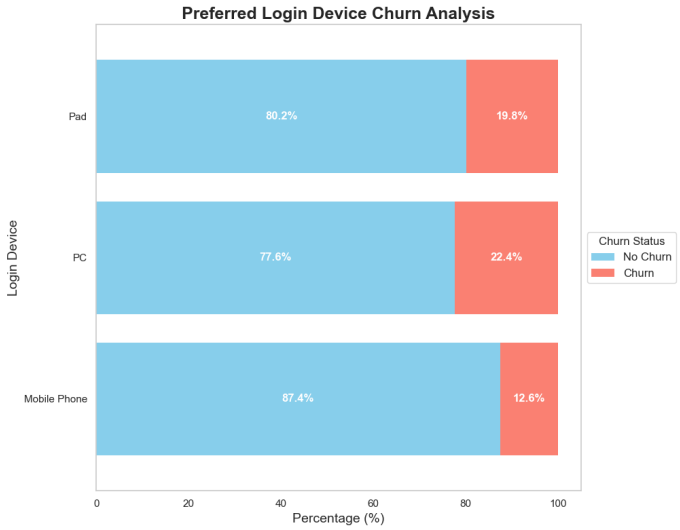


Figure 3.2.2 Preferred Login Device Churn Distribution

3.4 Age Group Churn Analysis

This bar chart illustrates churn rates across different age groups, providing insights into age-related patterns in customer retention:

- **10-19 Age Group:** This group has a very low churn rate, with only 22 out of 235 users (about 9.4%) churning. The small proportion of churn suggests that young users in this age bracket are relatively engaged and tend to retain their memberships.
- **20-29 Age Group:** Similar to the youngest age group, the 20-29 bracket also shows low churn, with only 26 out of 276 users (approximately 9.4%) churning. This consistency indicates that younger users are generally less likely to churn.
- **30-39 Age Group:** In this age group, churn becomes more noticeable, with 254 out of 1,699 users (about 15%) churning. Although this is higher than the previous groups, the majority of users in this age range remain active, suggesting moderate engagement.
- **40-49 Age Group:** This group has the highest churn count, with 392 out of 2,377 users (around 16.5%) churning. This elevated churn rate may indicate that users in their 40s are more likely to discontinue their membership, perhaps due to lifestyle changes or differing service expectations.
- **50-59 Age Group:** Churn remains relatively high in this age bracket, with 198 out of 881 users (about 22.5%) churning. This higher churn rate suggests a decreasing level of engagement as users age, possibly reflecting changing preferences or needs.
- **60-69 Age Group:** This age group exhibits a churn rate of approximately 34.6%

(56 out of 162 users), the highest across all groups. This significant churn rate could imply that older users face barriers to continued engagement or have specific service needs that are not being met.

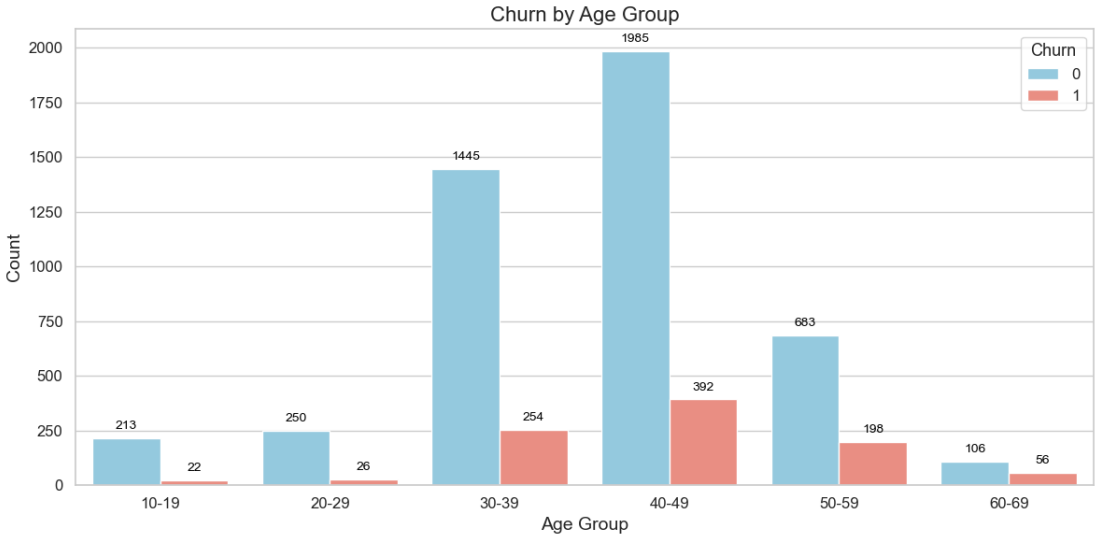


Figure 3.4.1 Churn by various Age Group

3.5 Analysis of Average Tenure by Churn Status

The bar chart illustrates the difference in average tenure between customers who churned (1) and those who did not churn (0). The following insights are evident:

- **Non-Churned Customers (Churn = 0):** The average tenure for customers who remained active on the platform is approximately 11.9 months. This suggests that users with longer tenures are more likely to continue their engagement with the platform.
- **Churned Customers (Churn = 1):** For customers who churned, the average tenure is significantly lower at around 5.0 months. This shorter tenure indicates that users who decide to leave the platform typically do so within the first few

months of their membership.

The substantial difference in average tenure between churned and non-churned users suggests that tenure is a critical factor in customer retention. Customers with longer engagement periods appear less likely to churn, possibly due to increased familiarity with the platform, accumulated benefits, or a greater sense of loyalty.

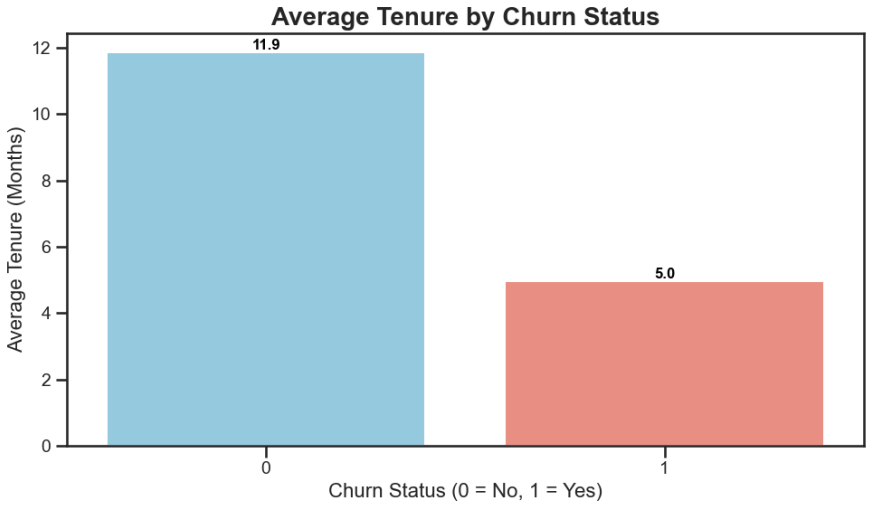


Figure 3.5.1 Average Tenure by Churn Status

3.6 Analysis of Days Since Last Order by Churn Status

The box plot compares the distribution of the number of days since the last order for churned and non-churned customers. This visualization provides insights into the recency of customer activity as a potential predictor of churn.

- **Non-Churned Customers**

1) The median value for non-churned customers is 3.0 days since their last order, indicating that they tend to have more recent activity on the platform.
2) The interquartile range (IQR) for non-churned customers is from 2.0 (Q1) to 8.0

(Q3), showing a relatively broad range of activity frequency within this group.

3) There are a few outliers in this group with values exceeding 20 days, suggesting that while most active customers engage frequently, some remain loyal despite longer gaps between orders.

- **Churned Customers**

1) The median value for churned customers is slightly lower, at 2.5 days since their last order, which implies that churned users may have slightly more recent last orders than expected. However, this value still reflects a smaller range of recent activity compared to non-churned customers.
2) The IQR for churned customers is narrower, from 1.0 (Q1) to 4.0 (Q3), indicating less variability in the recency of activity among churned customers.
3) There are fewer outliers compared to non-churned customers, with the majority of churned users having last-order times clustered closer to the median.

This analysis reveals a few important considerations：

1) **Recency as a Predictor:** The tendency for non-churned customers to have more recent orders may suggest that frequent engagement could be an indicator of customer retention. Customers who engage with the platform more regularly might be more likely to stay.
2) **Churn Dynamics:** The slightly lower median in days since last order for churned customers could imply that while they may have recent interactions, they are possibly dissatisfied or disengaged in other ways, leading to churn. This could indicate that recency alone may not fully capture churn risk, as customers who

recently placed orders may still churn if their needs or expectations are not met.

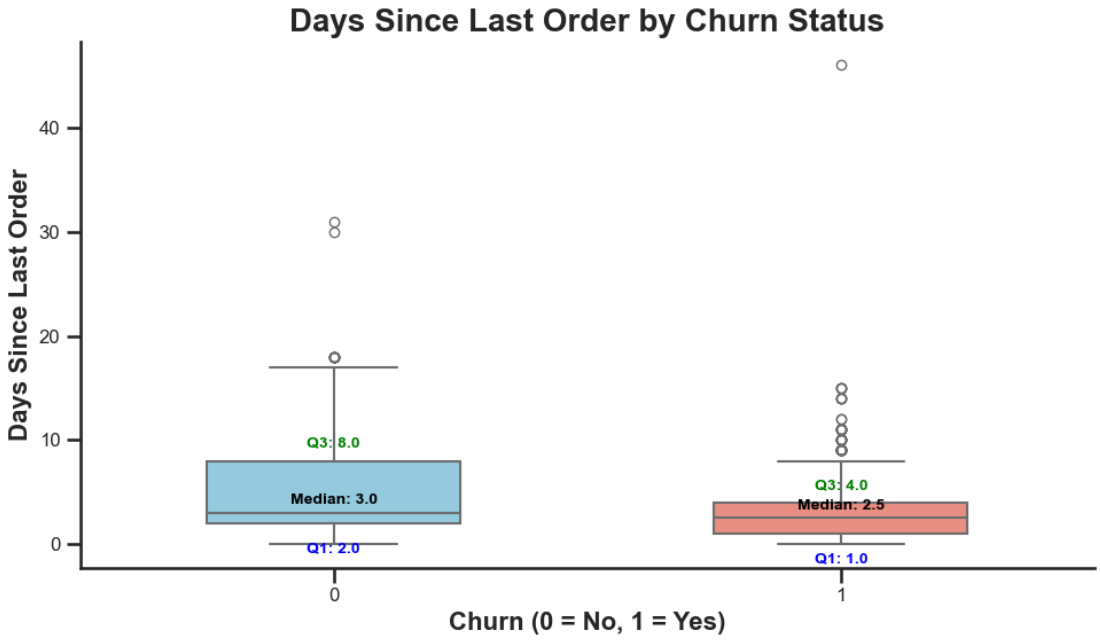


Figure 3.6.1 Days Since Last Order by Churn Status

3.7 Analysis of Churn Rate by Marital Status

The bar chart illustrates the churn rate segmented by marital status, revealing significant differences across the categories:

- **Single:** The churn rate among single customers is the highest at 26.73%. This suggests that single individuals may have different engagement patterns or loyalty levels compared to other groups, leading to a higher likelihood of discontinuing their relationship with the platform.
- **Divorced:** The churn rate for divorced customers is 14.62%, which is lower than for single individuals but still notably higher than for married customers. This could indicate that divorced individuals exhibit moderate levels of platform engagement and retention.

- **Married:** Married customers have the lowest churn rate at 11.52%. This lower churn rate might reflect higher stability or loyalty among married individuals, who may have a greater tendency to maintain long-term engagements with the platform.

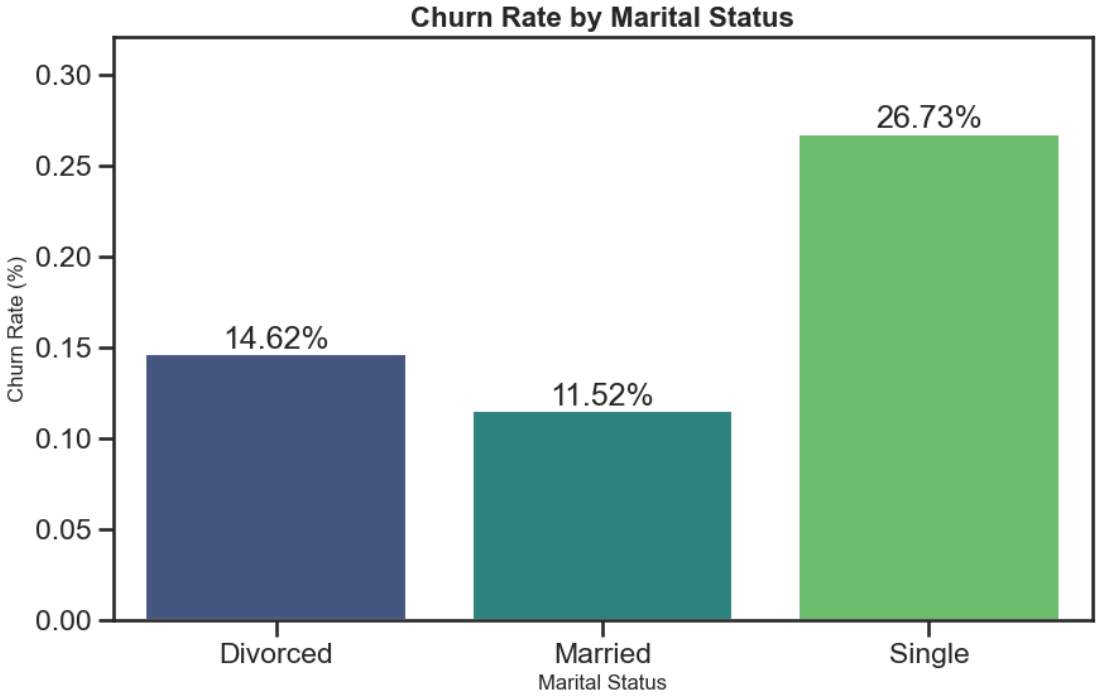


Figure 3.7.1 Churn Rate by Martial Status

The disparity in churn rates across marital status groups highlights potential differences in customer needs or satisfaction levels that could influence retention. Single customers, with the highest churn rate, may benefit from targeted engagement strategies, such as personalized offers or loyalty programs, to address their unique preferences and reduce churn. For married customers, the low churn rate suggests strong existing engagement, but retention efforts should continue to reinforce loyalty, perhaps by recognizing long-term commitment through special incentives.

3.8 Discount Amount Analysis by Churn Status

The set of visualizations provides insights into the distribution and average amount of discounts received by churned and non-churned customers. The analysis focuses on three aspects: density distribution, box plot summary, and average discount amount comparison.

3.8.1 Density Distribution of Discount Amounts

The density plot reveals differences in discount distribution between churned and non-churned customers:

- **Non-Churned Customers:** The distribution is slightly right-skewed, with a peak around 160, indicating that most non-churned customers received higher discount amounts.
- **Churned Customers:** The distribution for churned customers is more concentrated around a similar peak of 150 but exhibits a sharper drop-off. This suggests that churned customers generally received lower discounts compared to their non-churned counterparts.

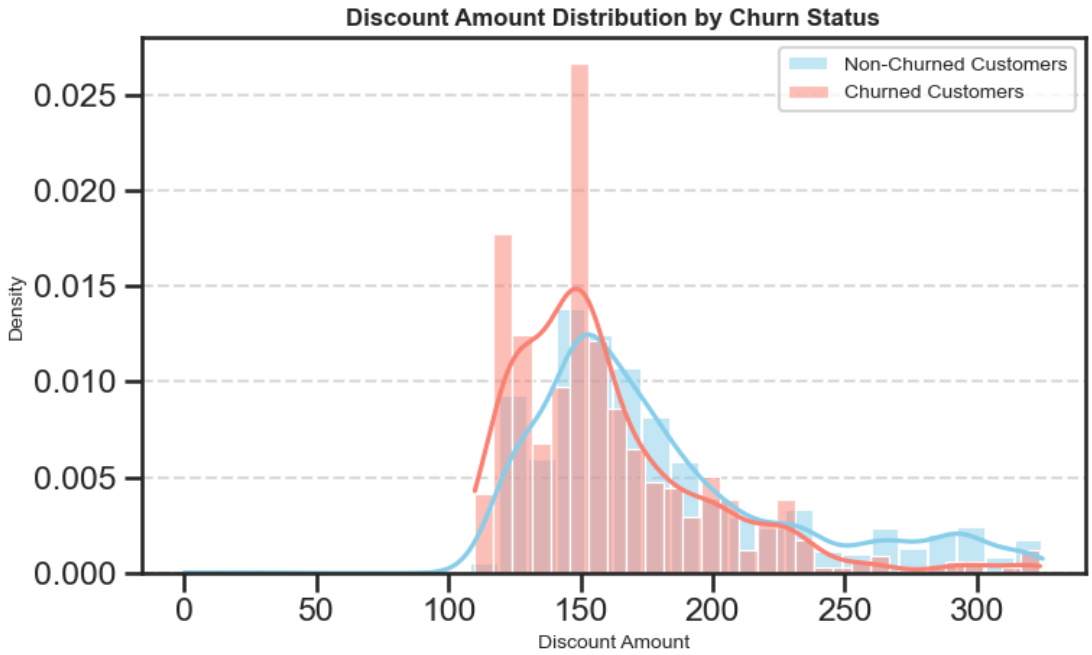


Figure 3.8.1 Density Distribution of Discount Amounts

3.8.2 Box Plot Summary of Discount Amounts

The box plot provides a summary of the distribution of discount amounts:

- **Non-Churned Customers:** The median discount amount for non-churned customers is 166, with an interquartile range (IQR) from 147 to 201. This group also has a few high-value outliers above 300, indicating some customers received very substantial discounts.
- **Churned Customers:** The median discount amount for churned customers is lower, at 150, with an IQR from 132 to 175. The range is narrower, and fewer high-value outliers are observed compared to non-churned customers.

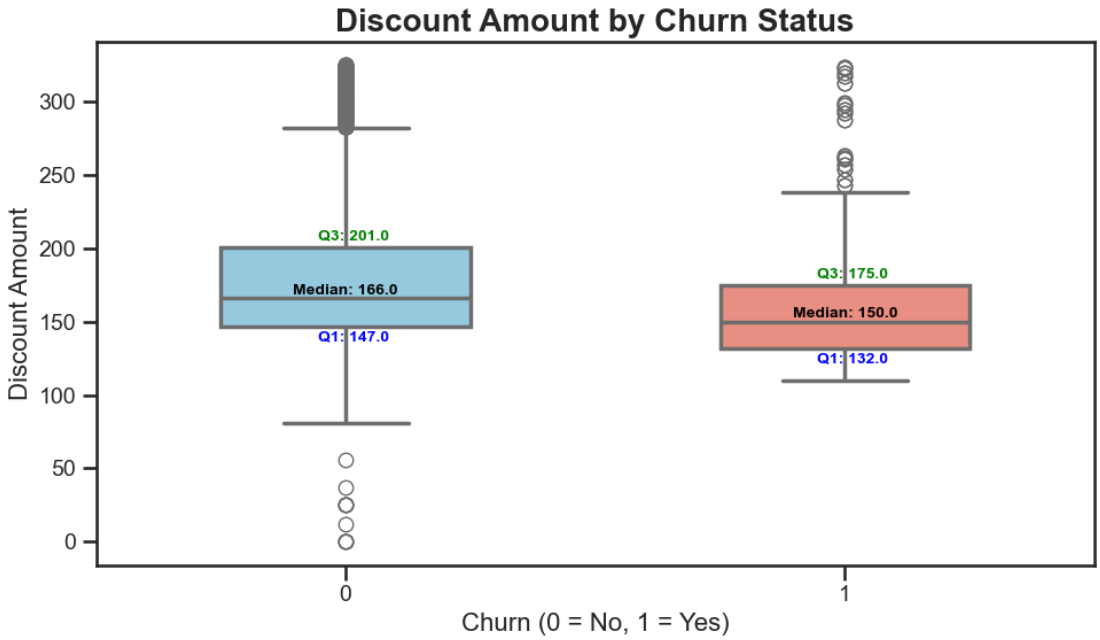


Figure 3.8.2 Discount Amount by Churn Status

3.8.3 Average Discount Amount Comparison

The bar chart comparing average discount amounts shows that:

1) **Non-Churned Customers** received an average discount of 180.6.
2) **Churned Customers** received an average discount of 160.4, which is approximately 11% lower than that of non-churned customers.

3.8.4 Implications

This analysis suggests that discount amount may have an impact on customer retention:

- Non-churned customers generally receive higher discounts, both on average and in terms of distribution, which could contribute to their continued engagement with the platform.

- Lower discount amounts among churned customers might indicate that they are not receiving sufficient incentives to stay, potentially leading to dissatisfaction or disengagement.

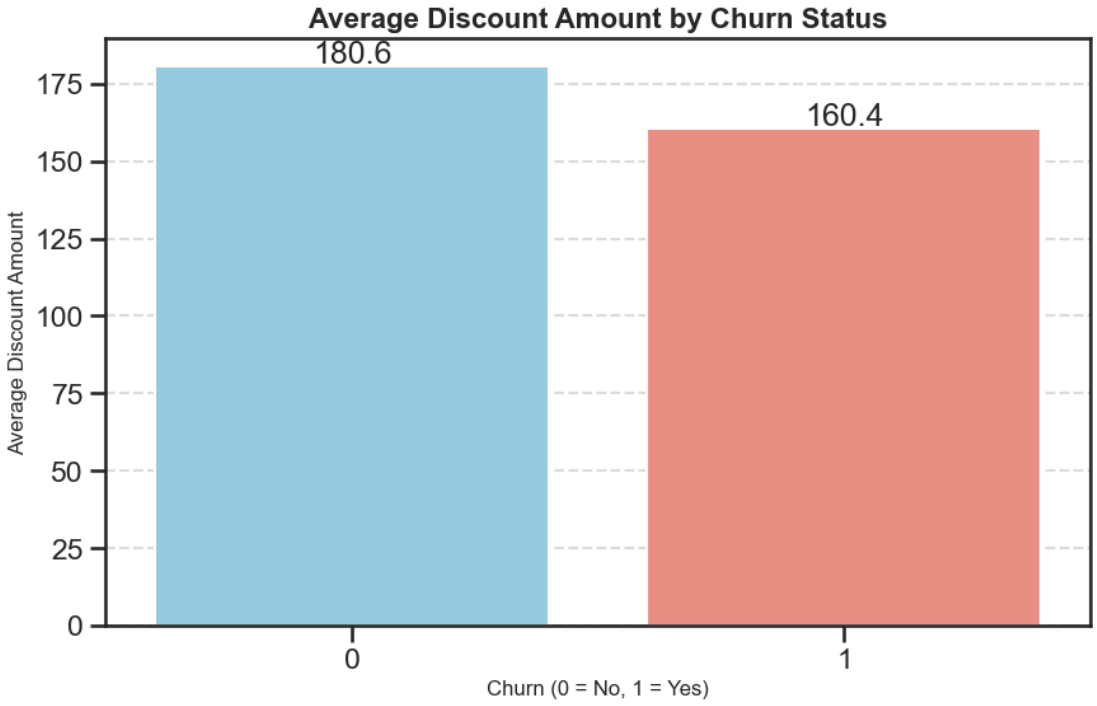


Figure 3.8.3 Average Discount Amount by Churn Status

3.9 Churn Rate Analysis by Preferred Order Category

The bar chart depicts the churn rate across various preferred order categories, highlighting significant differences based on customers' primary shopping preferences:

- **Mobile Phone:** This category exhibits the highest churn rate at 27.5%, indicating that customers whose preferred orders are in the mobile phone category are more likely to churn. This high churn rate could be due to factors such as market competition, high expectations, or the typically higher cost and longer purchase

cycles associated with electronics.

- **Household:** Following closely, the household category has a churn rate of 27.2%. This suggests that customers focused on household items may also have specific needs or expectations that are not being fully met by the platform, potentially leading to their disengagement.
- **Fashion:** The fashion category shows a moderate churn rate of 15.5%. While not as high as mobile phones or household items, it is still notable, suggesting that customers interested in fashion products may have moderate levels of retention, possibly due to varied product preferences and seasonal purchasing behaviors.
- **Laptop & Accessory:** The churn rate in this category is 10.2%, lower than that of fashion and household items. This may reflect a relatively stable customer base for laptop and accessory purchases, which could be due to their less frequent, but higher value, purchase patterns.
- **Others:** The churn rate for other unspecified categories stands at 7.6%, which indicates moderate customer retention and lower churn than the major product categories.
- **Grocery:** Grocery has the lowest churn rate at 4.9%, suggesting that customers with a preference for grocery items are highly engaged and exhibit strong retention. This lower churn rate could be attributed to the frequent, essential nature of grocery shopping, which encourages regular engagement with the platform.

The churn rate differences by order category suggest that customer retention varies significantly depending on the type of products customers are primarily interested in:

1) **High-Churn Categories (Mobile Phone and Household):** These categories may benefit from targeted retention strategies, such as tailored promotions, enhanced product variety, or improved customer support, to better meet customer expectations and reduce churn.
2) **Low-Churn Categories (Grocery):** The high retention among grocery-focused customers highlights the importance of consistent and reliable service in this area. Continuing to offer a high-quality experience for grocery shoppers could maintain this strong retention rate.

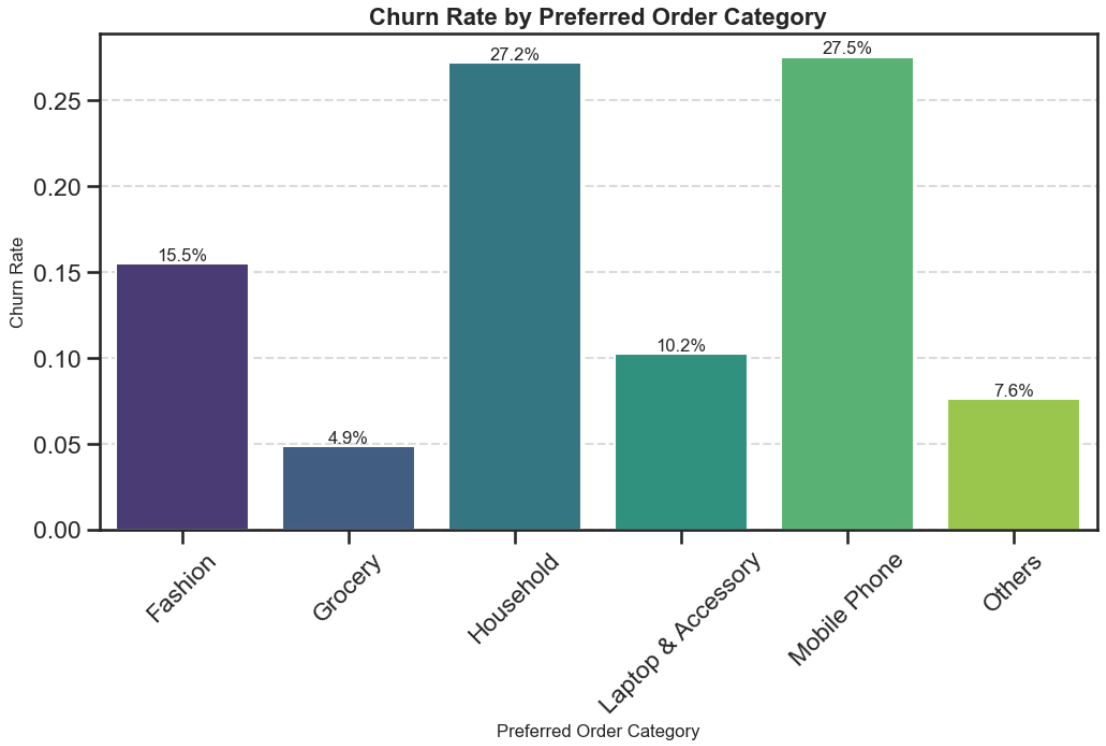


Figure 3.9.1 Churn rate by Preferred Order Category

3.10 Analysis of User Complaint by Churn Status

The bar chart displays the distribution of users with and without complaints segmented by churn status, offering insights into the potential relationship between customer complaints and churn.

- **Non-Churned Customers**

1) Out of 4,682 non-churned customers, 1,096 (approximately 23.4%) have lodged complaints, while the remaining 3,586 (76.6%) have not.
2) The majority of non-churned users did not have any complaints, suggesting that satisfaction or lack of issues is associated with retention.

- **Churned Customers:**

1) Among 948 churned users, 508 (about 53.6%) have made complaints, while 440 (46.4%) did not register any complaints.
2) A higher proportion of churned users have lodged complaints compared to non-churned users, indicating a strong correlation between customer complaints and the likelihood of churn.

This analysis highlights the importance of addressing customer complaints to improve retention:

1) **Complaint Impact:** The significant percentage of churned users with complaints (53.6%) suggests that unresolved issues may drive customer attrition. Addressing complaints promptly and effectively could reduce churn rates.
2) **Retention Among Non-Complainers:** The majority of non-churned users are

those who did not submit complaints, suggesting that a smooth, issue-free experience contributes to customer loyalty and retention.

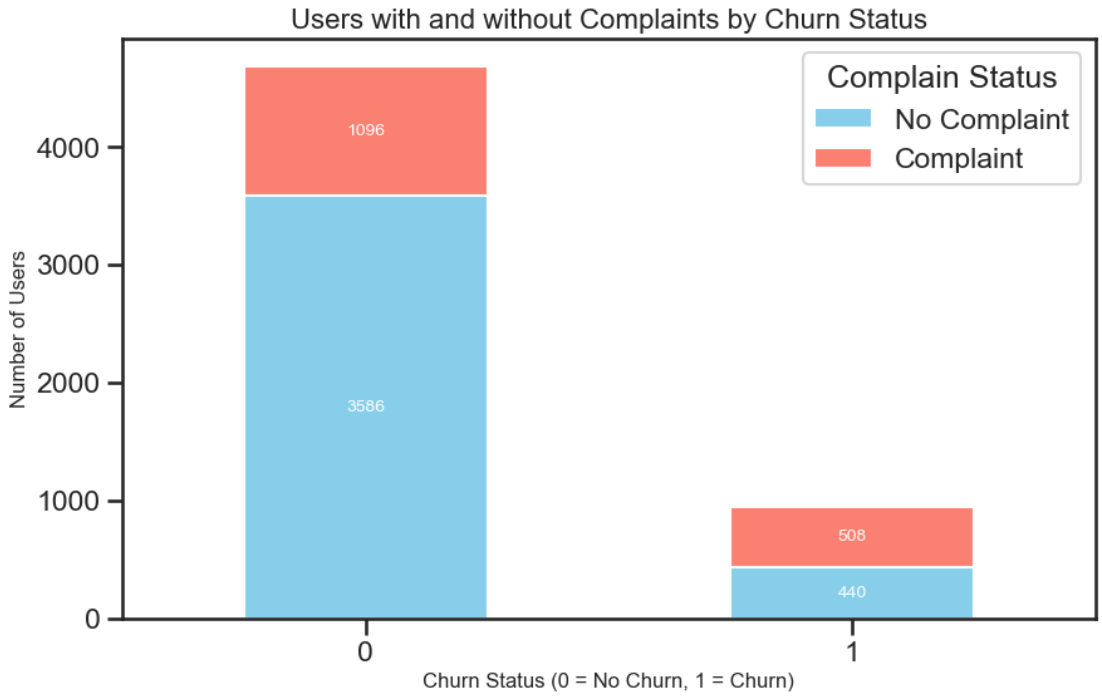


Figure 3.10.1 User Complaints by Churn Status

3.11 Analysis of Hour Spend on App by Churn Status

The set of visualizations examines the relationship between time spent on the app and customer churn status, including a density plot, box plot, and bar chart of average hours spent on the app.

3.11.1 Density Distribution of Hours Spent on App

The density plot shows the distribution of hours spent on the app for both churned and non-churned users:

- **Non-Churned Users:** The distribution for non-churned users exhibits multiple peaks, with the most significant around 3 hours, suggesting that users who spend

around this amount of time on the app are more likely to stay engaged.

- **Churned Users:** The distribution for churned users follows a similar pattern to non-churned users, with a peak also around 3 hours, indicating a comparable level of app usage between both groups.

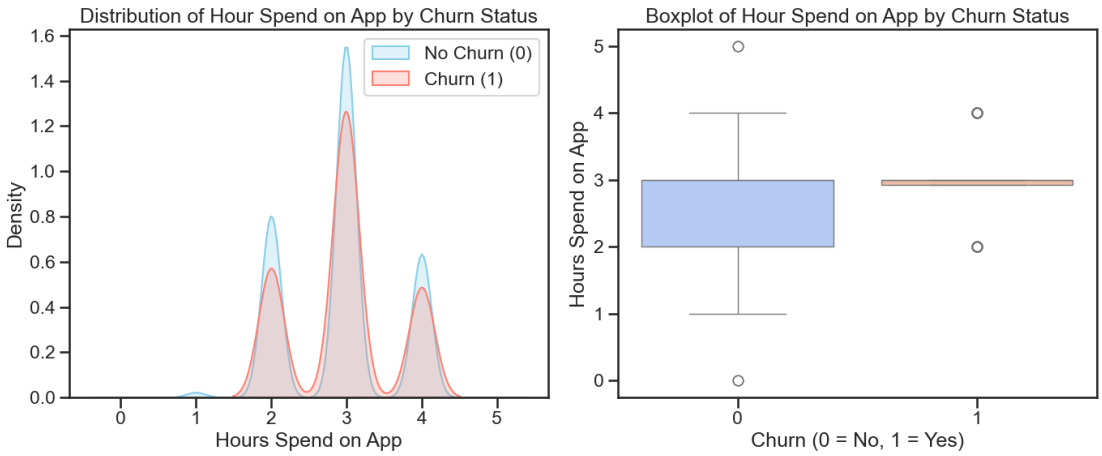


Figure 3.11.1 Density Distribution and Box Plot of Hour spend on App

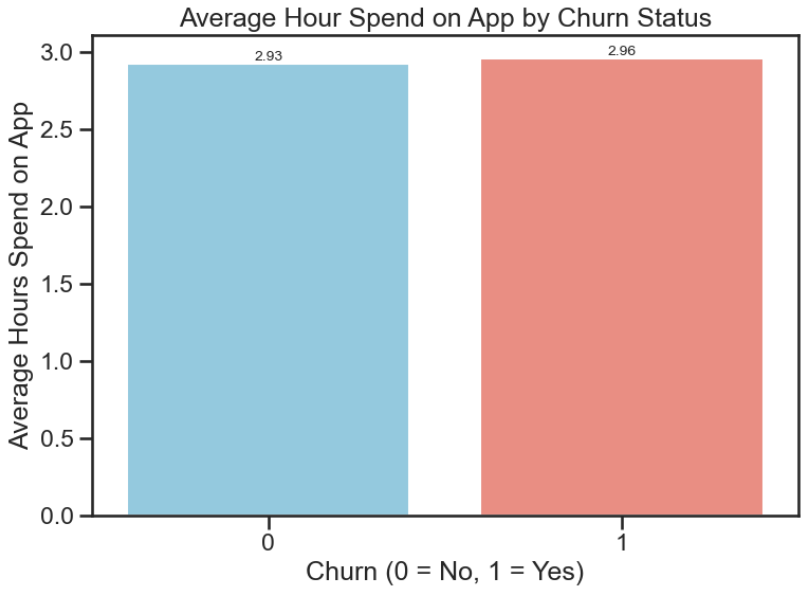


Figure 3.11.2 Average Hour Spend by Churn Status

3.11.2 Box Plot of Hours Spent on App

The box plot provides a summary of hours spent on the app:

- **Non-Churned Users:** The interquartile range (IQR) for non-churned users is broader, suggesting more variability in app engagement among users who continue to stay. Outliers indicate some non-churned users spend up to 5 hours on the app, showing high engagement levels.
- **Churned Users:** Churned users have a narrower IQR, indicating less variability in their app usage, with fewer outliers, suggesting a more consistent engagement pattern among those who eventually churn.

3.11.3 Average Hours Spent on App

The bar chart shows the average hours spent on the app by churn status:

- **Non-Churned Users** have an average of 2.93 hours.
- **Churned Users** have an average of 2.96 hours, which is nearly identical to that of non-churned users.

3.11.4 Implications

The analysis suggests that hours spent on the app does not significantly differentiate churned and non-churned users, as both groups exhibit similar engagement levels in terms of time. This finding implies that the amount of time spent on the app alone may not be a strong predictor of churn, as both retained and churned users display comparable usage patterns.

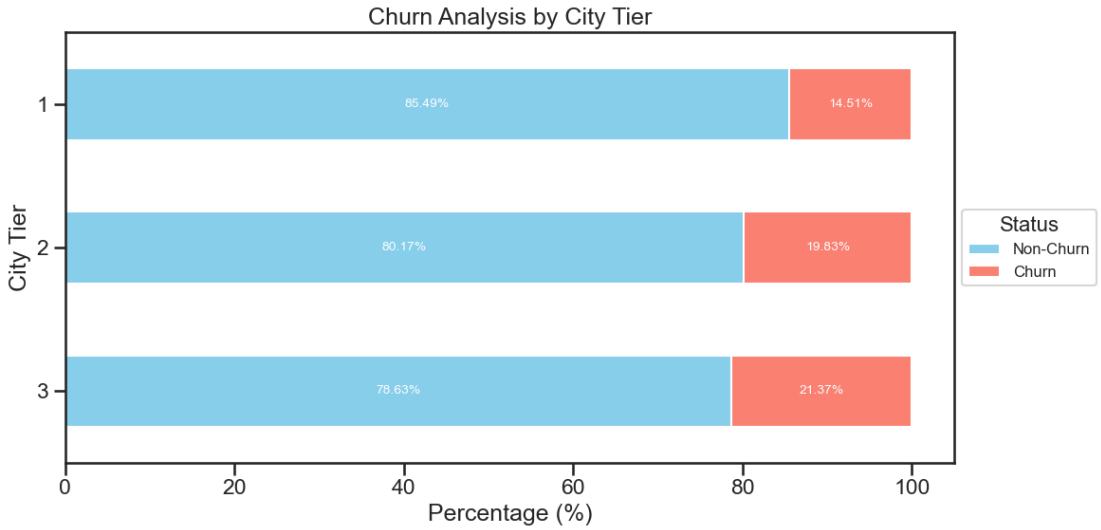


Figure 3.12.1 Churn rate by City Tier

3.12 Churn Analysis by City Tier

The bar chart shows the churn rate across different city tiers, revealing how geographic segmentation correlates with customer retention:

- **City Tier 1:** This tier has the lowest churn rate at 14.51%, with a high retention rate of 85.49%. Customers from Tier 1 cities are more likely to stay engaged with the platform, possibly due to better access to services, higher purchasing power, or stronger brand loyalty.
- **City Tier 2:** The churn rate in Tier 2 cities is higher at 19.83%, with a retention rate of 80.17%. This suggests a moderate level of engagement among customers in this segment, with churn higher than in Tier 1 cities, indicating that some users may have access or satisfaction issues.
- **City Tier 3:** Tier 3 cities exhibit the highest churn rate at 21.37%, with a retention rate of 78.63%. Customers in this tier are the most likely to churn, potentially due

to factors like limited access to the full range of services, lower purchasing power, or less brand affinity.

The analysis indicates that customers from lower-tier cities (Tiers 2 and 3) are more prone to churn, possibly due to differences in service access, socioeconomic factors, or varying levels of brand engagement. In contrast, Tier 1 cities have a higher retention rate, suggesting that customers in these areas are generally more satisfied or have better access to the platform's offerings.

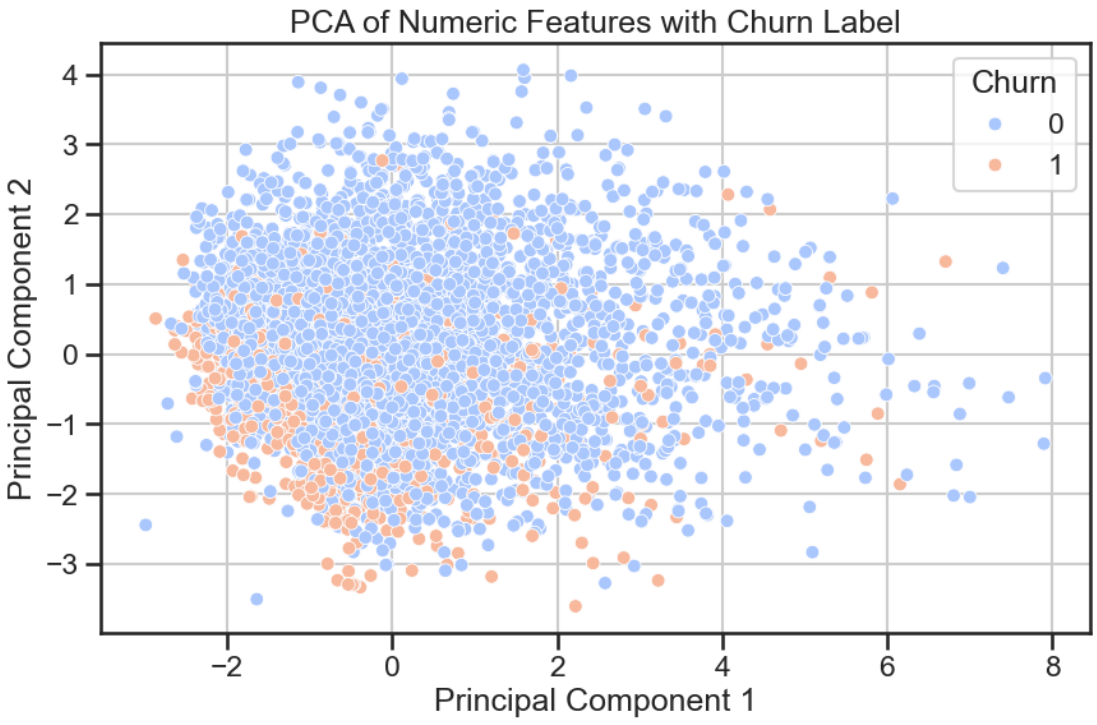


Figure 3.13.1 PCA Results

3.13 PCA of Numeric Features with Churn Label

The scatter plot visualizes the distribution of customers across two principal components, derived from Principal Component Analysis (PCA) on the numeric features in the dataset. Each point represents a customer, with colors indicating their churn status (blue for non-churned, orange for churned).

Insights

1) **Explained Variance:** The first two principal components explain approximately 28.4% of the total variance in the data, with Principal Component 1 accounting for 17.8% and Principal Component 2 accounting for 10.6%. This indicates that while PCA provides some dimensionality reduction, these two components do not capture the majority of variance, suggesting that other components or non-linear relationships might be significant in the data.
2) **Distribution and Overlap:** The plot shows considerable overlap between churned and non-churned customers, indicating that these two principal components do not effectively separate churned customers from non-churned ones. This overlap implies that linear combinations of features in these components may not provide clear differentiation between churned and non-churned groups.
3) **Potential Clusters:** While some loose clustering is visible, especially along the edges of the plot, there is no distinct separation by churn status. This lack of clustering suggests that churn might be influenced by complex interactions among features not captured in the linear PCA components.

Implications

The PCA plot suggests that churn behavior may not be easily captured by linear relationships in the primary features, as evidenced by the significant overlap in the two principal components. This implies that:

1) **Non-Linear Methods:** Machine learning models that capture non-linear relationships, such as decision trees, random forests, or gradient boosting, may perform better in distinguishing churned customers.
2) **Additional Feature Engineering:** Further feature engineering, such as interaction terms or non-linear transformations, might help in capturing the underlying patterns in the data.

4. Methodology

4.1 Feature Engineering

According to the heat map analysis, the correlation between each feature is relatively low. Therefore, we have decided to use all features as inputs to the model. This method ensures that the model fully utilizes available information during the training process to improve prediction performance. By retaining all features, we hope to capture potential complex patterns and interactions to better predict user churn behavior.

4.2 Model Selection

A range of machine learning classification models were applied to predict churn, with methods optimized using a rigorous hyperparameter tuning process. Specifically, we select 8 models for imbalanced classification tasks to leverage their unique strengths. **Logistic Regression** serves as our **benchmark model** due to its simplicity and low risk of overfitting. Among **ensemble methods**, **Random Forest** is chosen for its robustness against overfitting and ability to handle class weights. **AdaBoost** and **LightGBM** are incorporated for their adaptability to class imbalances, ensuring a well-rounded approach to model evaluation and selection. We opt for **Gradient Boosting** and its optimized version, **XGBoost**, for their superior performance in correcting errors iteratively. Lastly, **Support Vector Machine (SVM)** is included for its effectiveness in high-dimensional spaces, while **K-Nearest Neighbors (KNN)** provides insights into local class distributions.

The introduction of models are as follows:

- **Logistic Regression:** A supervised machine learning algorithm widely used for binary classification tasks.
- **Random Forest:** An ensemble method that uses multiple decision trees to improve generalization and avoid over-fitting.
- **AdaBoost:** A boosting algorithm which combines multiple weak classifiers together to build a strong classifier.
- **LightGBM:** A gradient boosting framework that uses decision trees for classification, ranking, and other tasks.

- **Gradient Boosting:** An ensemble method that builds models sequentially, each focusing on the errors of its predecessor.
- **XGBoost:** An optimized gradient boosting framework which is known for its speed and performance.
- **Support Vector Machine (SVM):** A model uses a hyperplane to separate classes, which is effective for high-dimensional spaces.
- **K-Nearest Neighbors (KNN):** A non-parametric method machine learning algorithm used for classification by calculating and comparing distances metrics.

We use a unified **pipeline** for the entire training process including column transformer, which streamlines model training and ensures consistency. **Grid search** is employed to finetune the hyperparameters of each model to optimize performance. To address class imbalance issues, we implement **SMOTE** (Synthetic Minority Over-sampling Technique) to generate synthetic samples, allowing us to evaluate the model with and without data balancing.

4.3 Evaluation Metrics

Performance was assessed using F1 Score and ROC-AUC, focusing on the models' ability to predict churn accurately, especially for the minority class.

The performance of the models is assessed using the following metrics:

- **F1 Score:** The harmonic means of precision and recall, providing a balance between the two, especially useful for imbalanced datasets.

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

- **ROC-AUC Score:** The ROC-AUC score is the area under the ROC curve. It sums up how well a model can produce relative scores to discriminate between positive or negative instances across all classification thresholds. The ROC-AUC score ranges from 0 to 1, where 0.5 indicates random guessing, and 1 indicates perfect performance.

5. Results and Interpretation

5.1 Model Performance

These models were evaluated based on their ability to predict customer churn, regardless of whether SMOTE was applied for data balancing. The main observations include:

Without SMOTE:

Based on the chart, it's evident that Logistic Regression (LR) and AdaBoost (ADA) have the lowest performance.

In contrast, Gradient Boosting (GB) and LightGBM (LGBM) demonstrate superior performance with F1 scores of 0.86 and 0.88, and ROC-AUC scores of 0.94 for both, showcasing their effectiveness in handling the dataset. XGBoost (XGB) also performs

well, with an F1 score of 0.85 and a ROC-AUC score of 0.92, making it another strong contender.

The Random Forest (RF), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) models show moderate performance, achieving decent scores but not quite matching the top-performing models.

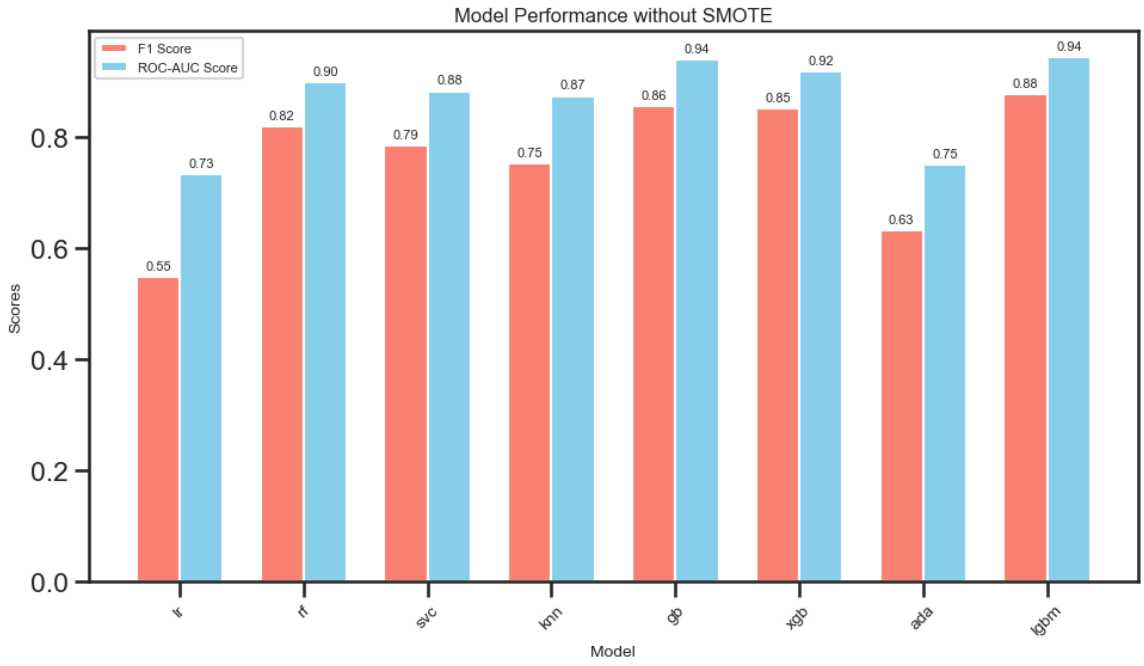


Figure 5.1.1 Model Performance without SMOTE

With SMOTE:

The chart shows that after applying SMOTE, all models have significantly improved performance in handling ethnic minority classes. LightGBM (LGBM) stood out with outstanding performance, scoring 0.94 in both ROC-AUC and F1, demonstrating its strong ability in classification tasks.

Random Forest (RF) and Support Vector Machine (SVC) both performed well in F1 score, reaching 0.97, indicating a good balance between accuracy and recall. Gradient enhancement (GB) also performed well, with a ROC-AUC score of 0.93.

Other models, such as K-Nearest Neighbors (knn) and AdaBoost (ada), perform stably with ROC-AUC scores remaining above 0.91. This indicates that the use of SMOTE effectively improves the generalization ability of these models.

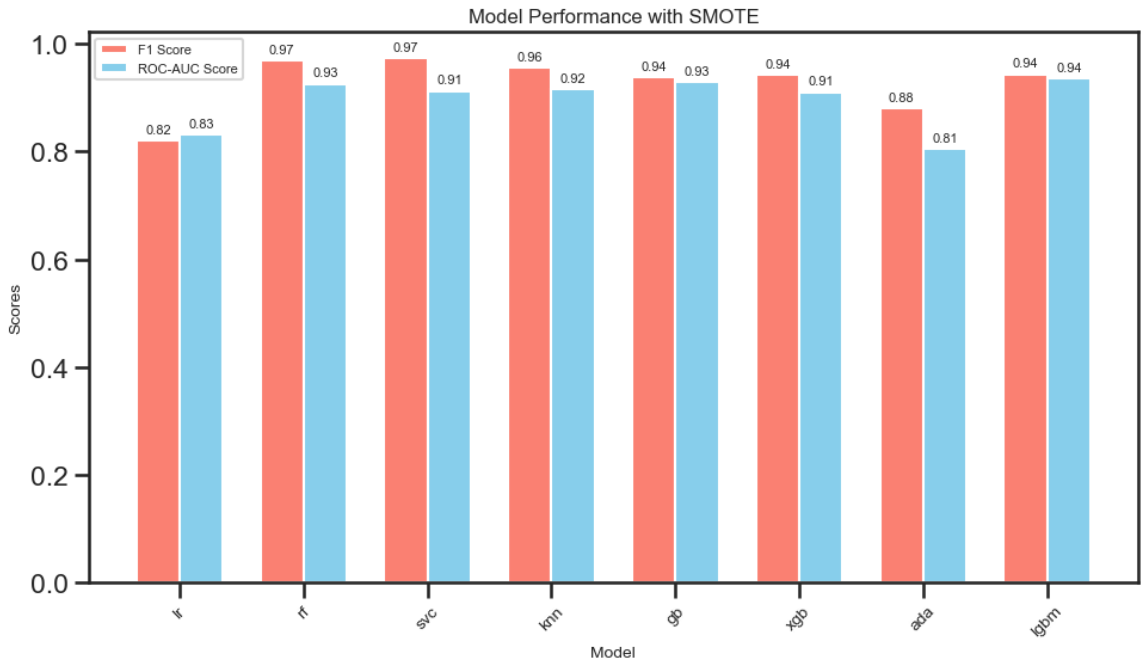


Figure 5.1.2 Model Performance with SMOTE

The chart below shows the comparison between model performance with and without SMOTE. SMOTE greatly improved recall for the minority class but sometimes led to a slight drop in overall accuracy and precision, particularly in models more sensitive to data distribution. This trade-off is common because SMOTE generates synthetic samples for the minority class, enhancing class balance but potentially

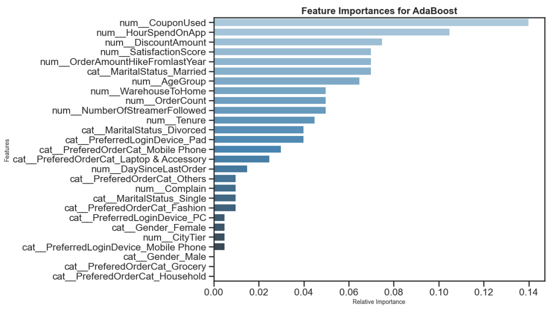
introducing noise. In imbalanced datasets, this technique is crucial to prevent models from heavily favoring the majority class, though it may not always benefit performance metrics that depend on exact distribution fidelity, such as precision.

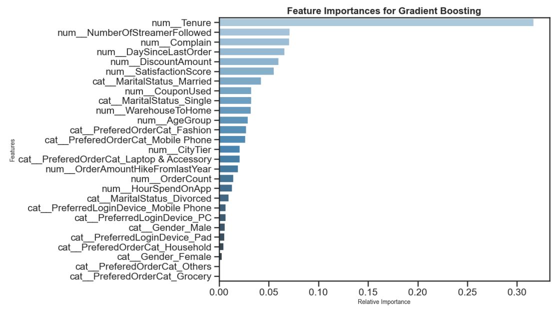
Figure 5.1.3 Model Performance Comparison

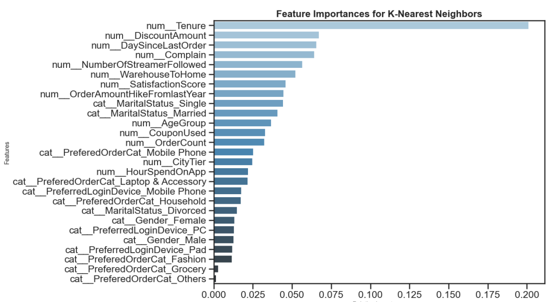
In conclusion, LGBM emerged as the top-performing model, achieving high accuracy, recall, and ROC-AUC scores both with and without SMOTE. This suggests that boosting techniques like LightGBM are highly effective in handling imbalanced datasets and can provide reliable predictions for both classes.

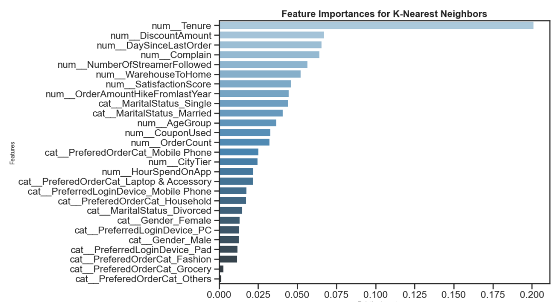
5.2 Feature Importance

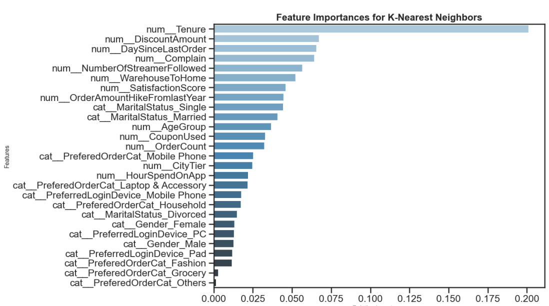
Through trained models, we can identify the key factors affecting customer churn. Different models highlight different factors of importance which is demonstrated as below:

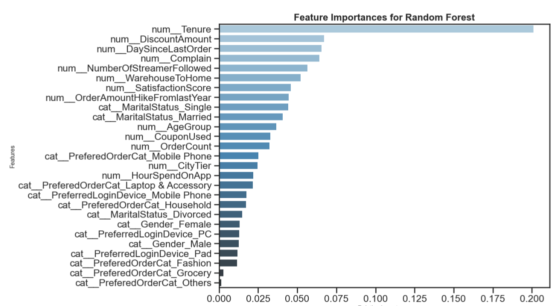


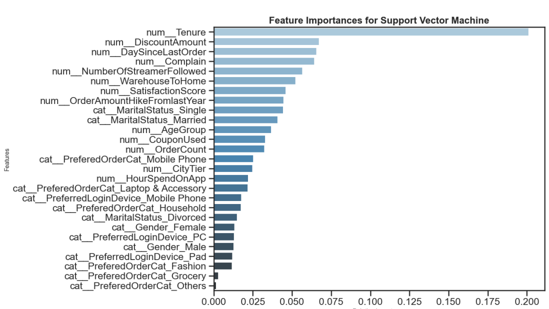


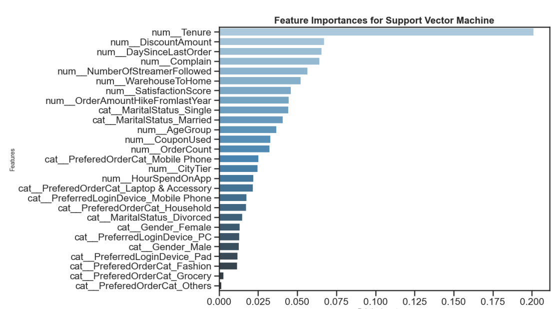












Based on the top 10 average relative importance scores, the following features have been identified as key factors in predicting customer behavior:

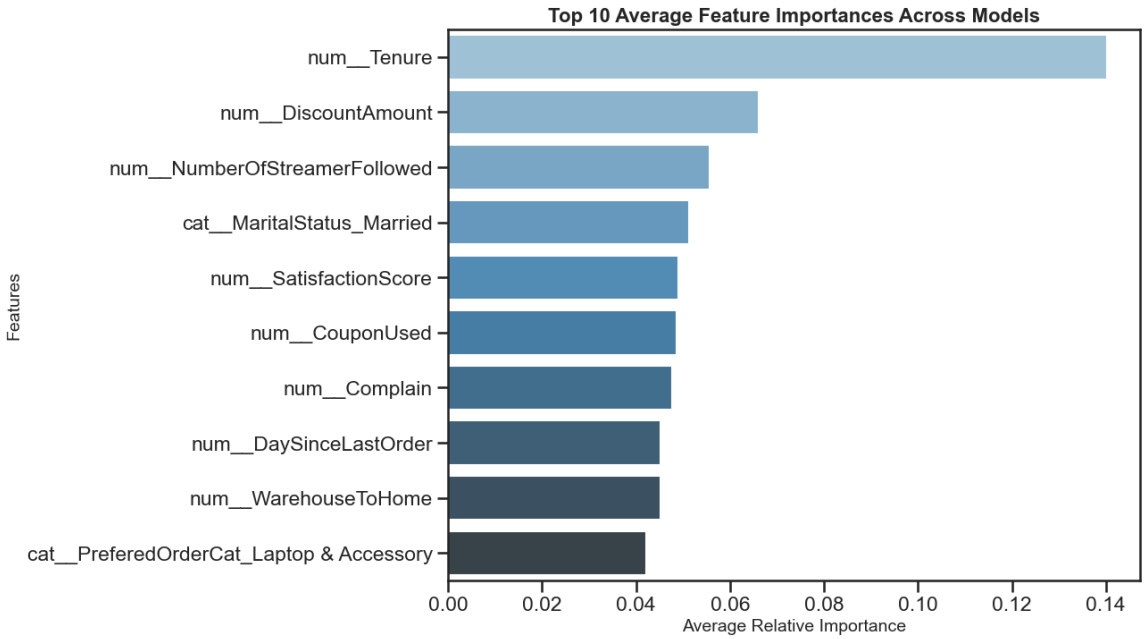


Figure 5.1.4 Top 10 Average Feature Importance

1. **Customer Characteristics**

- **Tenure:** The length of time a customer has been with the company is the most influential feature, reflecting customer loyalty and satisfaction.
- **Marital Status:** Being married may indicate differences in purchasing power and needs. Divorced customers exhibit purchasing behaviors influenced by changes in lifestyle and financial status.
- **Distance From Warehouse to Home:** The distance can impact delivery times and costs, influencing customer satisfaction and repeat purchases.

2. **Purchasing Behavior**

- **Preferred Order Category:** Preference for laptop and accessory orders indicates that customers who frequently purchase these items may have unique behavioral patterns.
- **Day Since Last Order:** The number of days since the last order serves as a key indicator of customer engagement.

3. **Economic Factors**

- **Discount Amount:** The amount of discount received plays a role in customer engagement and retention.
- **Coupon Used:** The number of coupons used is influential, suggesting that promotions can engage customers.

4. **Engagement**

- **Number of Streamer Followed:** a higher number of streamers followed indicates greater platform dependence, increasing retention likelihood.

5. **Customer Feedback**

- **Complaint:** The number of complaints can indicate customer dissatisfaction and potential churn risk.
- **Satisfaction Score:** Satisfaction scores provide a direct measure of customer happiness and loyalty.

5.3 Actionable Insights

Taking all the above factors into account, **implementing a loyalty program** can significantly increase the customer retention rate by rewarding long-term customers and offering them exclusive benefits. Additionally, **personalized offers** tailored to

individual customer characteristics and behaviors are essential. For example, bundles of family-oriented products may appeal to married customers.

Another important aspect is **customer feedback**, as complaints and satisfaction scores count. The platform should recognize and reward customers who provide valuable feedback, addressing issues and making improvements effectively. Beyond improving the online user experience, **offline infrastructure** must also keep pace. Consider opening additional warehouses closer to high-density customer areas to enhance delivery efficiency, especially for tier 2/3 cities.

6. AI-generated Intervention Advice with Personalized Massage

The "Intelligent Suggestions" advice generator aims to address customer retention challenges by leveraging AI-based personalization techniques (LLMs). This approach combines advanced data processing with natural language generation to create tailored messages and actionable intervention plans for at-risk customers. This report examines the mechanics, benefits, and business advantages of the "Intelligent Suggestions" generator.

6.1 Operational Flow of Intelligent Suggestions

The heart of the "Intelligent Suggestions" model is its robust personalization logic, which analyzes both customer-specific features and the relative importance of each feature. This process tailors recommendations to each individual, providing meaningful, context-specific advice.

- **Identify At-Risk Customer**: The model collects a variety of customer attributes—such as demographics, purchase history, engagement frequency,

and behavioral patterns. These attributes are processed to detect signs that a customer may be at risk of disengagement or churn.

- **Feature Importance Assessment**: The model weighs each customer attribute based on its relevance to retention. Machine learning models often present challenges in interpretation due to their complexity and the potential inclusion of biased or irrelevant variables. To prioritize clarity and reliability in our causal inference, we used **logistic regression results.** We also adjusted to handle Missing Not At Random (MNAR) data , addressing mitigate multicollinearity issues (see our code file for detailed procedures). Through this approach, we identified key indicators such as Tenure, City Tier, Warehouse-to-Home distance, among others. This nuanced weighting ensures that recommendations are relevant and aligned with each customer's unique relationship with the business.
- **Detailed Prompt Creation**: Once the customer attributes are gathered and assessed, they are used to create a tailored prompt for the GPT-3.5 API(openAI 2023). The API then generates two essential outputs:
    - **Personalized Intervention Plan**: A strategic plan that outlines specific actions or engagement tactics.
    - **Customized Message**: A customer-facing message designed to resonate with the individual's needs, preferences, and past behaviors.

Together, these outputs allow for targeted, impactful outreach aimed at strengthening customer loyalty and satisfaction.

6.2 Example of Intelligent Suggestions

Inputs:

Feature	Value
CustomerID	50179
Churn	-0.557086
Tenure	-0.454791
PreferredLoginDevice	Mobile Phone
CityTier	3
WarehouseToHome	-0.193475
MaritalStatus	Divorced
AgeGroup	-1.625078
Gender	Male
HourSpendOnApp	-0.112351
OrderCount	1.978651
OrderAmountHikeFromLastYear	-1.065143
DaySinceLastOrder	1.083724
PreferredOrderCat	Fashion
NumberOfStreamerFollowed	0.479049
SatisfactionScore	-0.839537
Complain	-0.870573
CouponUsed	0.587754
DiscountAmount	0.285067

Table 1: Customer Data Example Row (Note: Numerical values are in z-score format)

Outputs:

"Intervention Plan:
Offer Personalized Discount: Given the customer's high DiscountAmount and CouponUsed, we will provide a tailored discount on products from preferred category, Fashion.
Enhance Streamer Engagement: As the customer follows several streamers, we will collaborate with these influencers to create exclusive content or promotions that align with the customer's interests.
Address Complaints: Since the customer has expressed dissatisfaction (Complain), we will ensure that any outstanding issues are resolved promptly and offer additional support to enhance their experience.
Improve Retention through Engagement: Considering the low SatisfactionScore, we will engage the customer with personalized recommendations and invite them to participate in surveys or feedback sessions to better understand their needs.

Personalized Message:

Subject: 🌟 Exclusive Offer Inside! We Miss You at [Company Name] 📩

Dear [Customer Name],

We hope this message finds you well. It has come to our attention that we haven't seen you shopping with us as frequently, and we wanted to reach out personally.

As a valued member of our e-commerce family for [Tenure] months, your satisfaction and loyalty mean the world to us. To show our gratitude for your continuous support, we are thrilled to offer you an exclusive discount on our wide range of premium Fashion products—because we know these are your favorites.

Use code [DiscountCode] at checkout to avail of this special offer and add a touch of joy to your shopping experience with us. Should there be any queries, feedback, or assistance you require, our dedicated team is always here to lend a helping hand.
Your satisfaction matters to us deeply, and we look forward to continuing this wonderful journey together.

Warm regards,
[Your Name]
[Company Name] Customer Retention Team 🌟"

6.3 Business Advantages of Personalized Outreach

Implementing Intelligent Suggestions offers several competitive and financial advantages:

- **Competitive Edge**: In a marketplace where customers expect tailored experiences, Intelligent Suggestions enables businesses to stand out by providing hyper-personalized communications. This level of attentiveness shows customers that the business understands and values their needs, which can lead to stronger customer bonds and increased loyalty.
- **Customer Testimonial:** "It felt like they knew exactly what I needed, which made me more eager to stay." – This type of feedback highlights the positive impact of personalized outreach on customer sentiment and loyalty.

- **Revenue Growth Through Increased Retention**: Retaining existing customers is typically more cost-effective than acquiring new ones. By proactively engaging at-risk customers, Intelligent Suggestions helps businesses maximize the lifetime value of each customer. This increases profitability and strengthens the overall customer base.
- **Low-Cost Strategy for Customer Base Expansion:** Unlike many loyalty programs or discounts that require substantial investment, Intelligent Suggestions relies on AI to deliver meaningful interventions at minimal cost. This scalability makes it ideal for businesses aiming to rapidly grow their customer base while keeping overhead low.

7. Conclusions

This analysis highlights key churn predictors, demonstrating that customer engagement, purchasing behavior, and discounts significantly influence churn. Tree-based models, especially LGBM, proved most effective for imbalanced churn prediction, benefiting from SMOTE's balancing effect. The insights from this study provide targeted areas for improving retention through focused marketing and discount strategies.

A. References

[1] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

[2] Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., & Zeng, W. (2019). Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural water management*, *225*, 105758.

[3] OpenAI. (2023). ChatGPT API (Version GPT-4) [Large language model]. Retrieved from https://platform.openai.com/