****

CAPSTONE PROJECT

********************* FINAL REPORT ***********************

Prepared By : Sahid
COURSE : DSBA
DATE : 16-03-2025

Table of Contents

List of Figures

List of Tables

1. Introduction

Problem Definition

Customer churn is a significant challenge for many businesses, especially in competitive industries. In this project, we seek to understand the behaviors and characteristics of customers who leave, or “churn,” and to identify patterns that can predict future churn. By uncovering these patterns, the goal is to empower the business with insights that can lead to improved retention efforts. Our focus will be on analyzing customer data to segment users into distinct groups, exploring their behaviors, and identifying factors contributing to churn. These insights will help the company develop targeted retention strategies to reduce attrition and improve customer loyalty.

Business Objective

- Reduce Churn: One of the primary objectives of this analysis is to identify customers at high risk of churn, so that effective retention strategies can be put in place.
- Enhance Customer Engagement: By understanding the behaviors that correlate with higher engagement, we aim to boost satisfaction and customer loyalty, fostering long-term relationships.
- Optimize Marketing Efforts: Targeted marketing efforts can be developed for specific customer segments, ensuring that resources are focused on high-value or high-risk customers.

2. Exploratory Data Analysis (EDA)

Data Overview

The dataset provided for this analysis contains 11,260 rows and 19 columns. Each row represents a customer, and the columns provide key attributes such as customer tenure, monthly revenue, cashback, payment methods, and demographic data. The target variable of interest is Churn, which indicates whether a customer has churned (1) or remained retained (0).

AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status	rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_	Day_Since_CC_	cashback	Login_device
20000	1	4	3	6	Debit Card	Female	3	3	Super	2	Single	9	1	11	1	5	159.93	Mobile
20001	1	0	1	8	UPI	Male	3	4	Regular Plus	3	Single	7	1	15	0	0	120.9	Mobile
20002	1	0	1	30	Debit Card	Male	2	4	Regular Plus	3	Single	6	1	14	0	3	NaN	Mobile
20003	1	0	3	15	Debit Card	Male	2	4	Super	5	Single	8	0	23	0	3	134.07	Mobile
20004	1	0	1	12	Credit Card	Male	2	3	Regular Plus	5	Single	3	0	11	1	3	129.6	Mobile

(Table 1 showing first few rows of the data)

Key Variables:

- Target: Churn (1 = churned, 0 = retained).
- Numerical: Tenure (customer's length of subscription), rev_per_month (revenue generated per month), cashback (amount of cashback redeemed).
- Categorical: Payment (method of payment), Gender, Marital_Status.

Missing Values & Descriptive Statistics

Upon reviewing the dataset, we found missing values in the following columns:

- rev_per_month (7%)
- Login_device (6.7%)
- cashback (4.2%)
- Account_user_count (3.9%)
- Day_Since_CC_connect (3.1%)
- Complain_ly (3.1%)
- Tenure (1.9%)
- Marital_Status (1.8%)
- CC_Agent_Score (1%)
- City_Tier (0.99%)
- Payment (0.96%
- Gender (0.95%
- CC_Contacted_LY (0.9%)
- Service_Score (0.87%)
- account_segment (0.86%)
- coupon_used_for_payment (0.02%)
- rev_growth_yoy (0.02%)

For both numerical and categorical columns, missing data was handled through imputation, replacing missing numerical values with the median and categorical values with the mode. This approach was chosen to maintain the integrity of the dataset without introducing significant bias.

Descriptive Statistics

- The churn rate in the dataset is 16.8%, meaning that a small but significant portion of customers have left.
- The cashback column has a mean of 175, with a median of 165, indicating a slight right skew (more customers with low cashback usage).

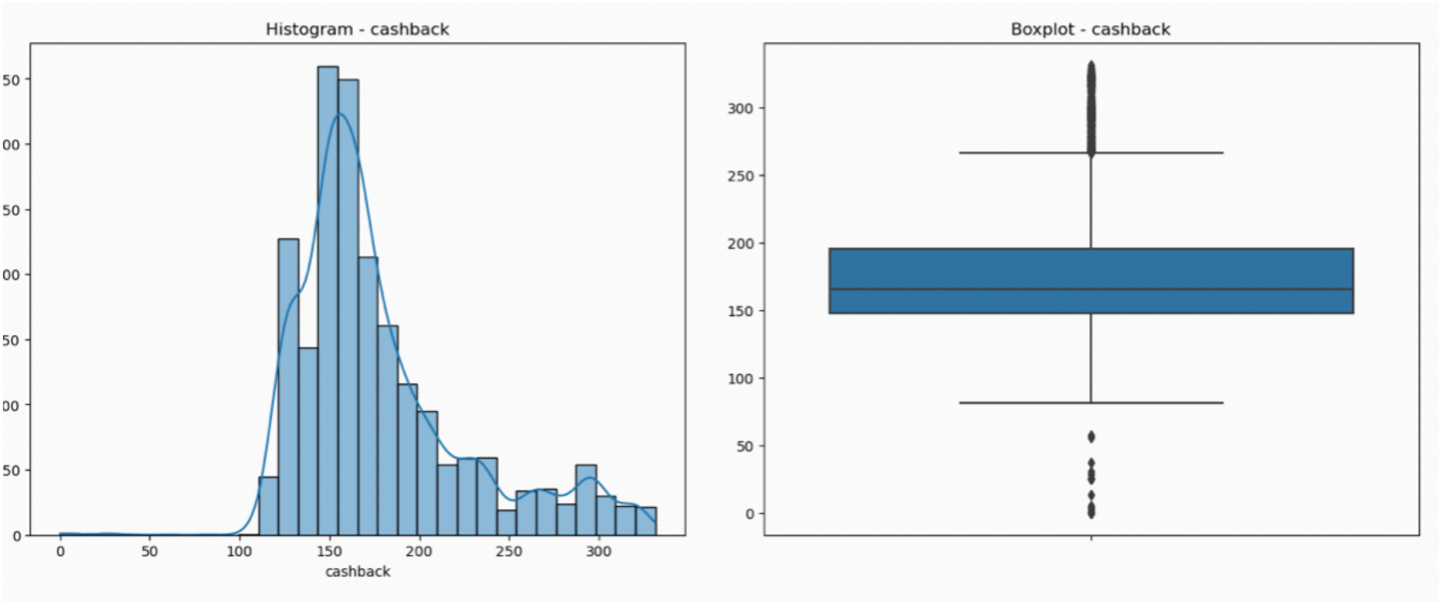
	AccountID	Churn	City_Tier	CC_Contacted_	Service_Score	CC_Agent_Score	Complain_ly
count	11260	11260	11148	11158	11162	11144	10903
mean	25629.5	0.168384	1.653929	17.867091	2.902526	3.066493	0.285334
std	3250.62635	0.374223	0.915015	8.853269	0.725584	1.379772	0.451594
min	20000	0	1	4	0	1	0
25%	22814.75	0	1	11	2	2	0
50%	25629.5	0	1	16	3	3	0
75%	28444.25	0	3	23	3	4	1
max	31259	1	3	132	5	5	1

(Table 2 descriptive stats of the data)

Univariate & Bivariate Analysis

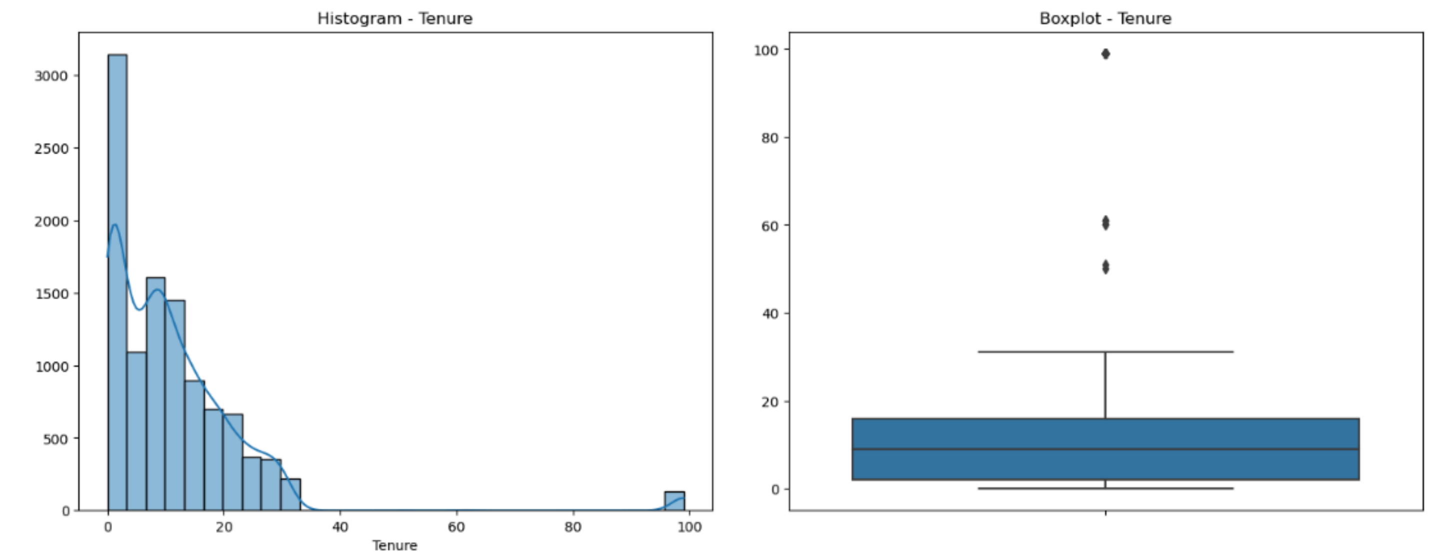
Univariate

Cashback: The distribution is right-skewed, with most customers redeeming cashback amounts of less than 200 units.



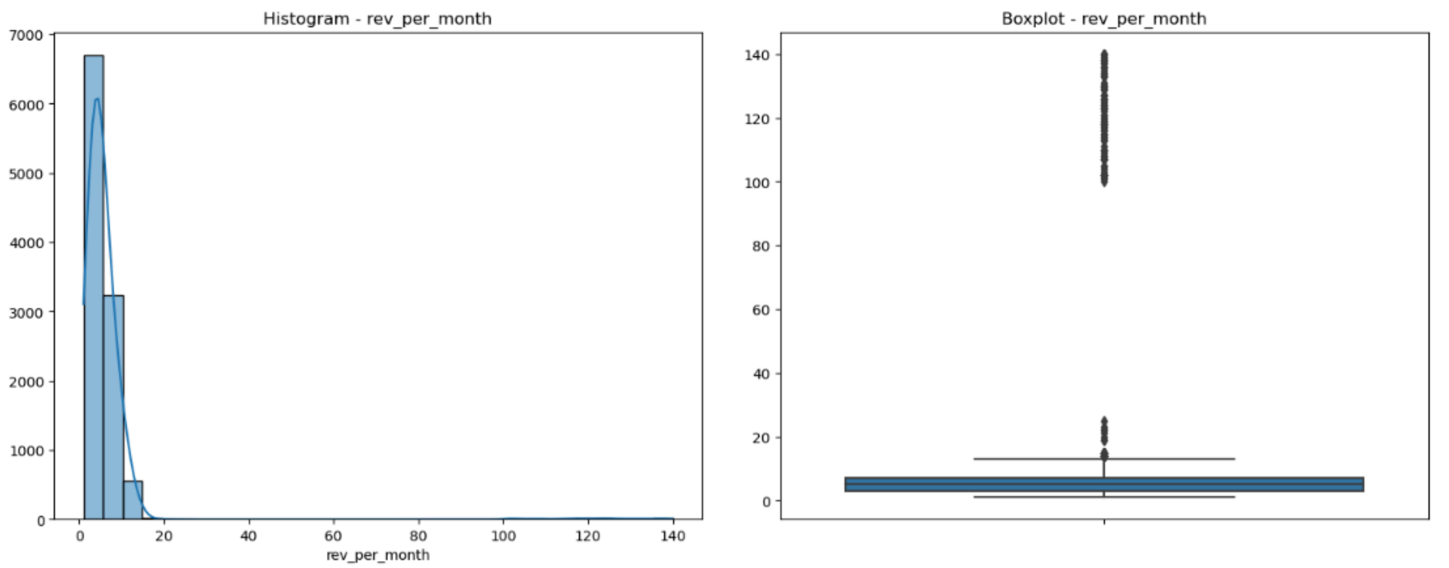
(Figure 1 showing histogram and boxplot for cashback)

__Tenure__: This column exhibits a bimodal distribution, with peaks at 0-5 months (new customers) and 15-20 months (long-term customers).



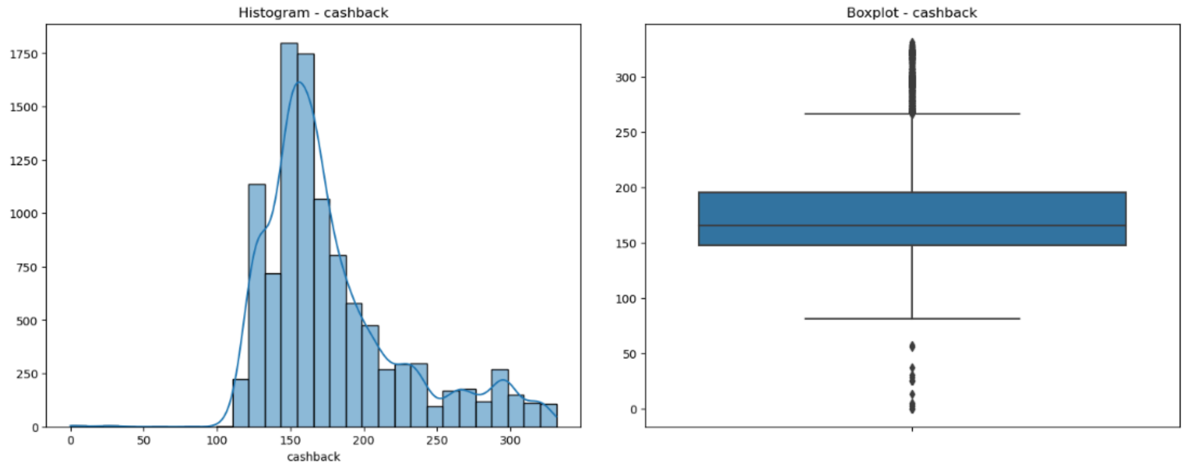
(Figure 2 showing histogram and boxplot for Tenure)

__Revenue Per Month :__ The distribution is right skewed , Revenue increases in the initial months.



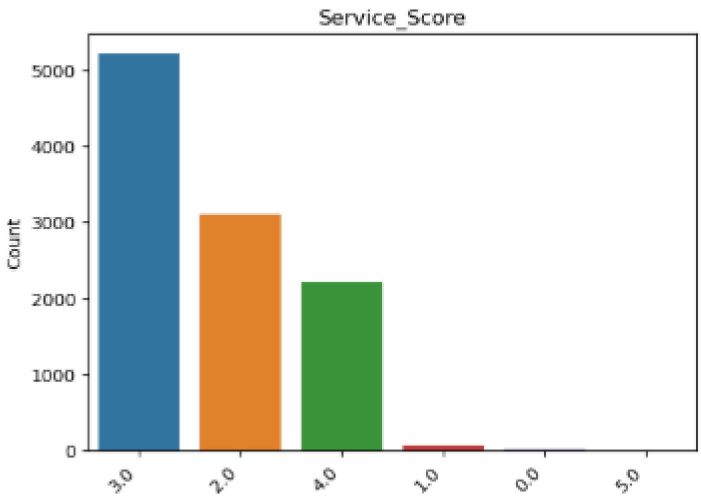
(Figure 3 showing histogram and boxplot for Revenue Per Month)

Cashback : Cashback ranges from 150-160 on an average with outliers on both sides.



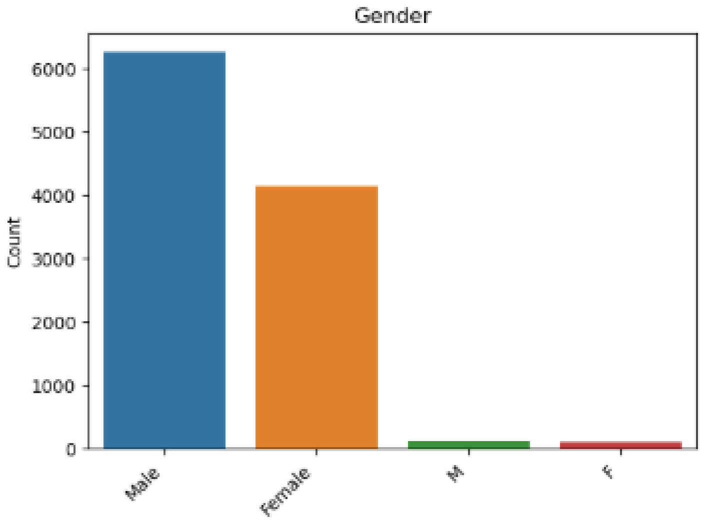
(Figure 4 showing histogram and boxplot for Cashback)

Higher Service_Score correlates with reduced churn, emphasizing the importance of customer satisfaction.



(Figure 5 showing the bar chart for service score)

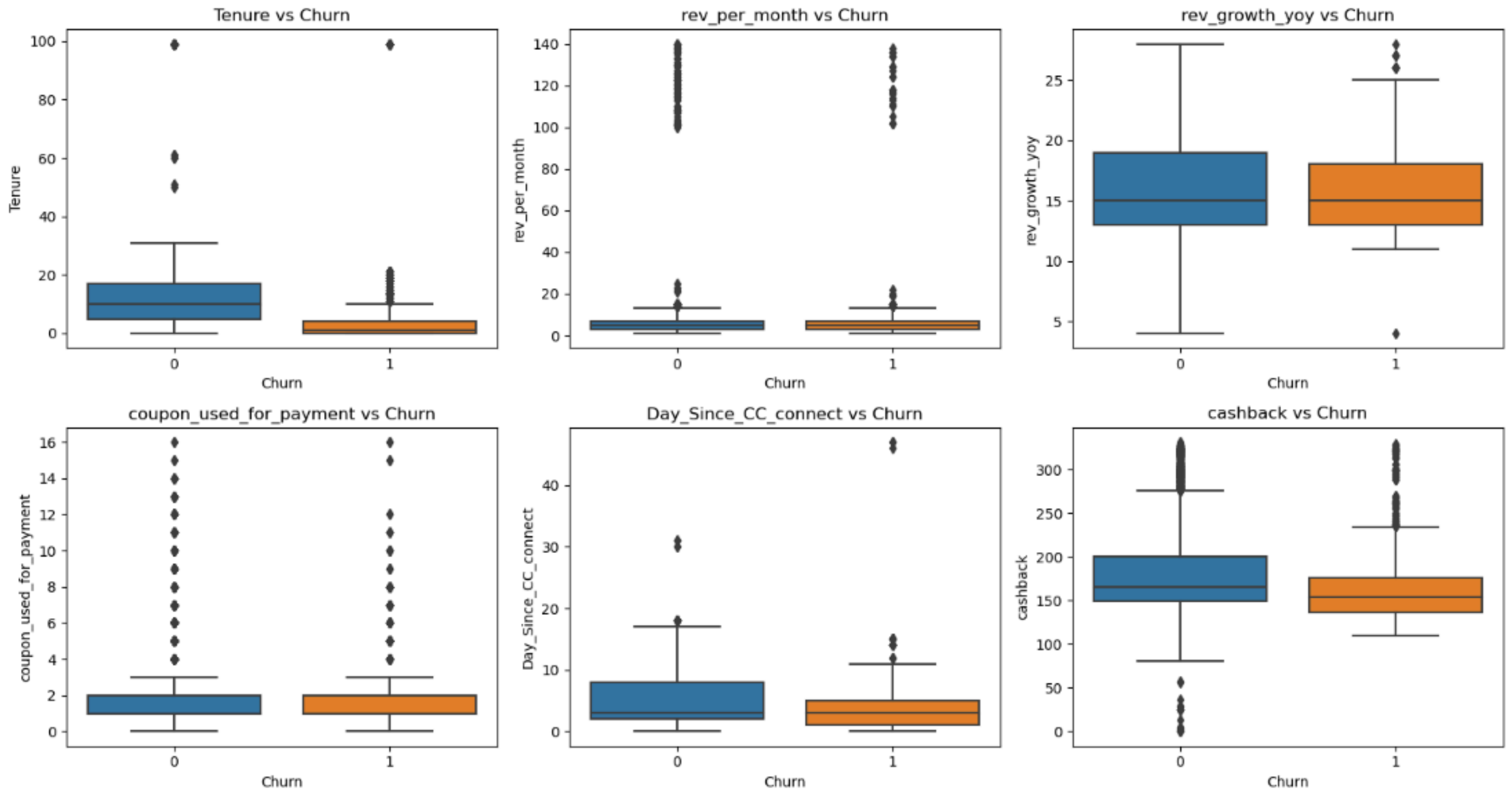
Gender : Male customer are around 50% more than females



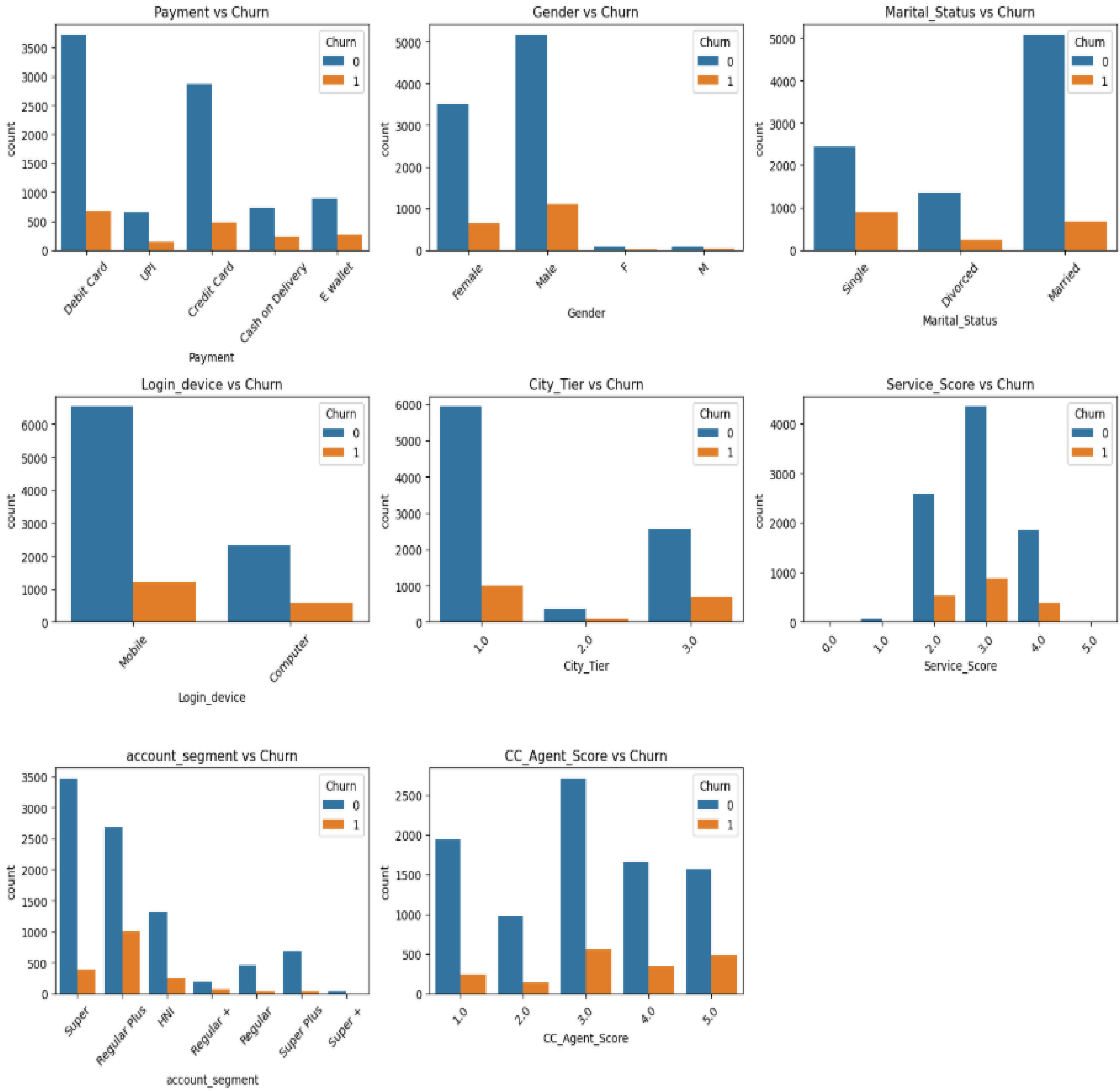
(Figure 6 showing the bar chart for distribution of Gender)

Bivariate Analysis

Churn is more likely among users with low cashback and fewer coupon usages.



(Fig 7 showing Boxplots comparing Churn with numerical features to see how churned vs retained customers differ)

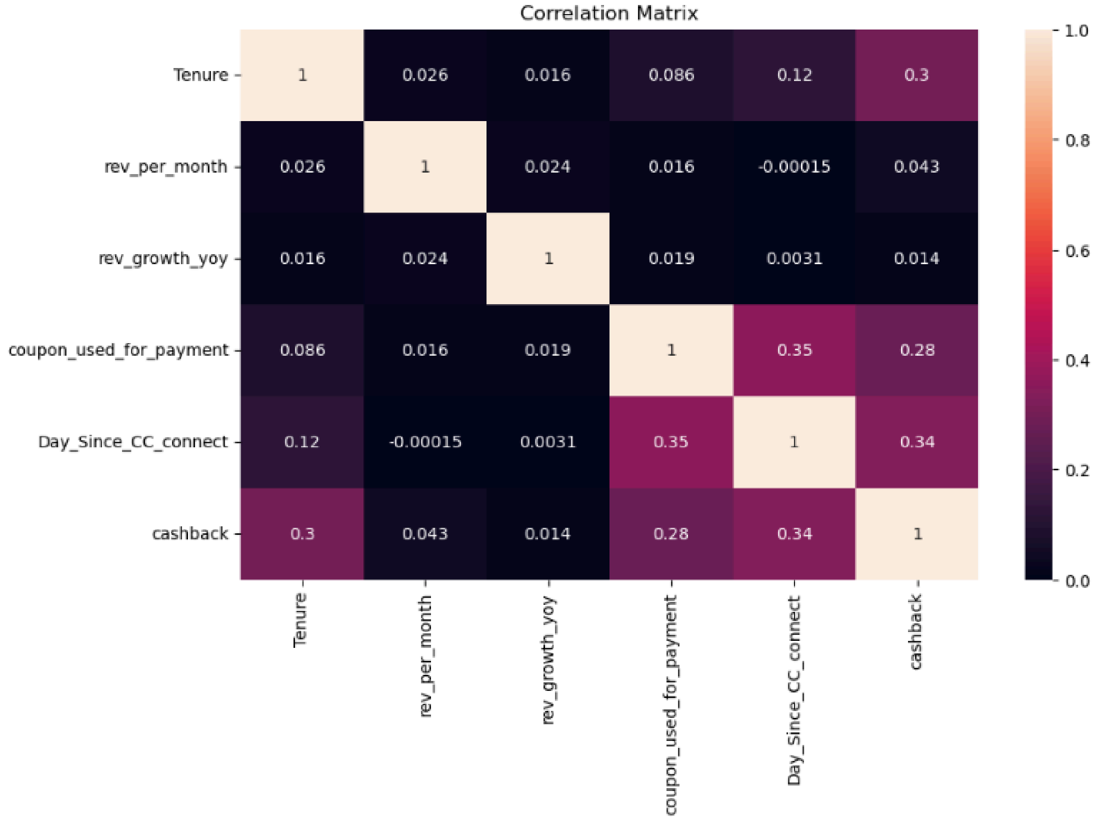


(Fig 8 showing Bar charts for categorical features like Gender and Marital_Status to compare churn rates across different categories)

Correlation Analysis

A key part of the analysis involves identifying relationships between features.

Notably: **cashback** has a moderate positive correlation with **coupon_used_for_payment** (0.28) and **Day_Since_CC_connect** (0.34), suggesting that more engaged users are more likely to redeem cashback and use coupons.



(Fig 9 showing Heatmap of the correlation matrix capturing relationships between the numerical features)

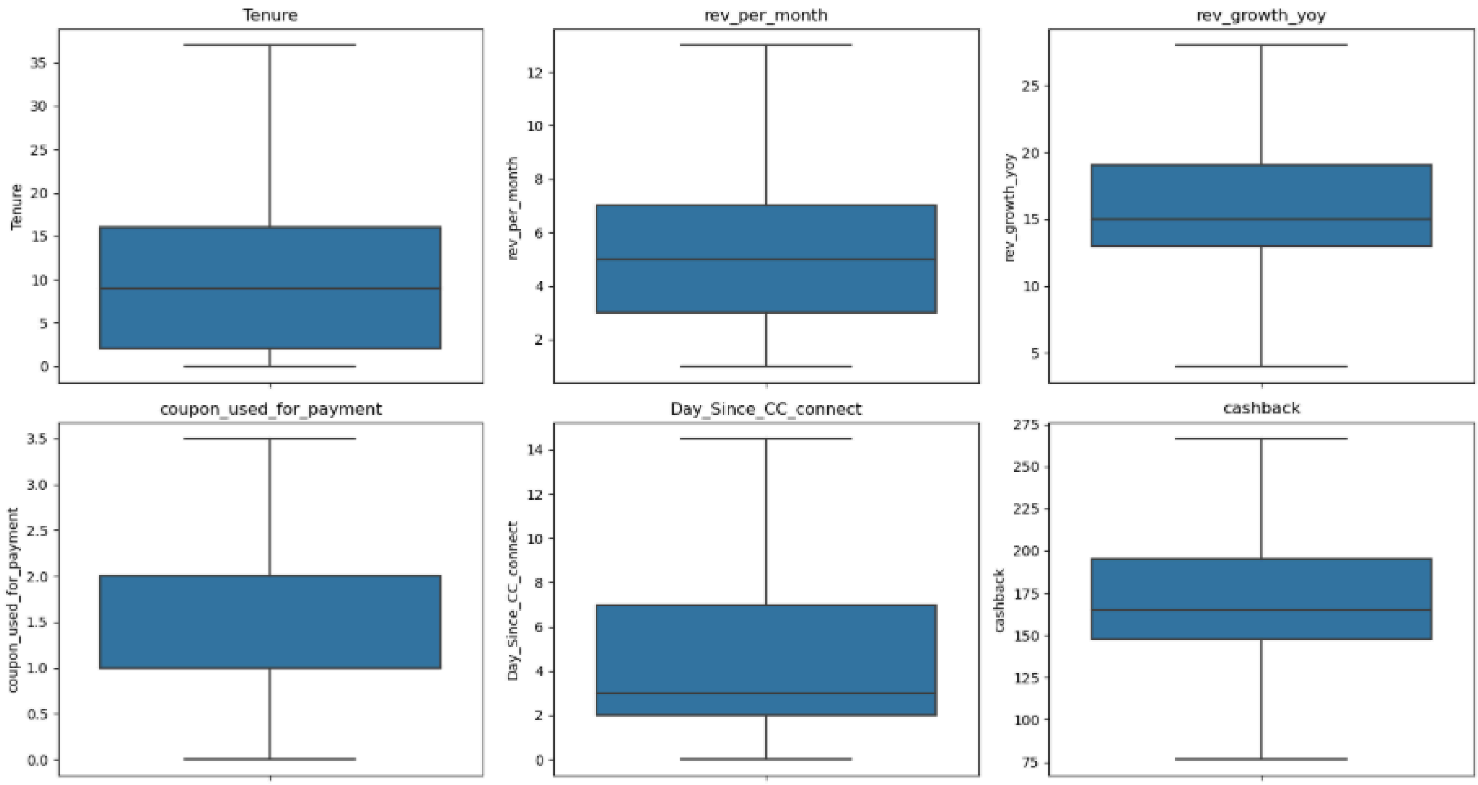
3. Data Preprocessing

Handling Missing Values

In this phase, we handled the missing values identified in the EDA. For rev_per_month and Login_device, we performed imputation using the median (for numerical data) and mode (for categorical data). This approach ensures that we retain as much of the data as possible without introducing bias due to dropped rows.

Outlier Treatment

Outliers in the cashback and Tenure columns were addressed using the IQR method, where values falling outside of the 1.5 * IQR range were capped at the lower and upper bounds. This ensures that extreme values do not disproportionately affect the analysis or model training.



(Fig 10 showing Boxplots after outlier treatment to illustrate the impact of capping extreme values)

Feature Engineering

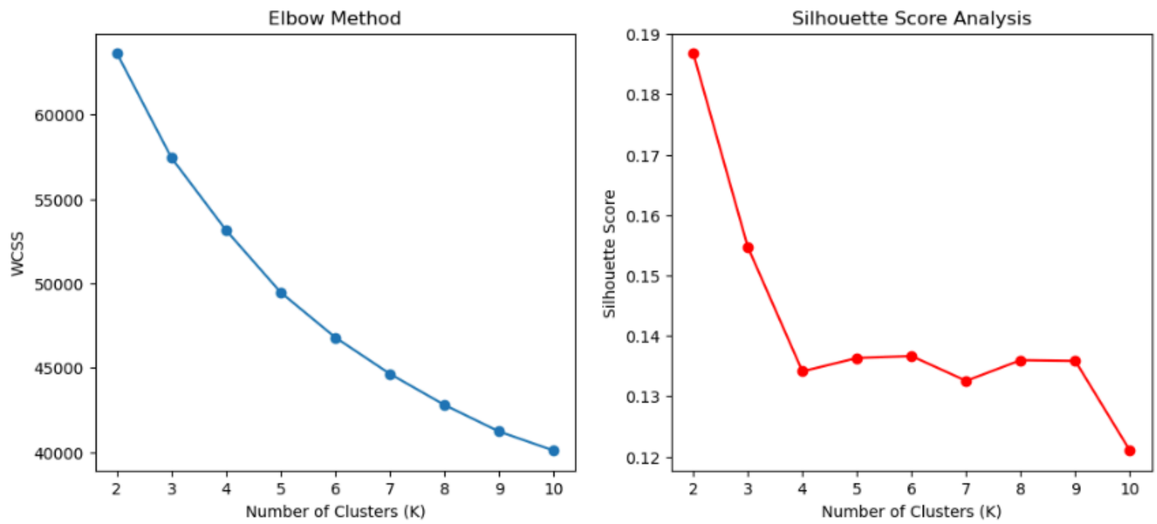
To prepare the data for modeling, several transformations were made:

- Standardization: Numerical features were standardized (mean = 0, standard deviation = 1) to ensure that no feature dominates others during clustering.
- Encoding: Categorical variables such as Payment, Gender, and Marital_Status were one-hot encoded to convert them into binary indicators, suitable for machine learning algorithms.
- New Variables: The results of K-means clustering were used to create a new variable, Cluster, which represents the customer segment each individual belongs to. This feature will provide useful insights for targeted strategies.

4. Clustering for Segmentation

Methodology

To segment customers, we applied K-means clustering, a popular unsupervised learning algorithm. K-means works by partitioning the data into K distinct clusters, each with similar characteristics. The optimal number of clusters was determined using the Elbow Method, which suggested K=3 as the ideal number of clusters.

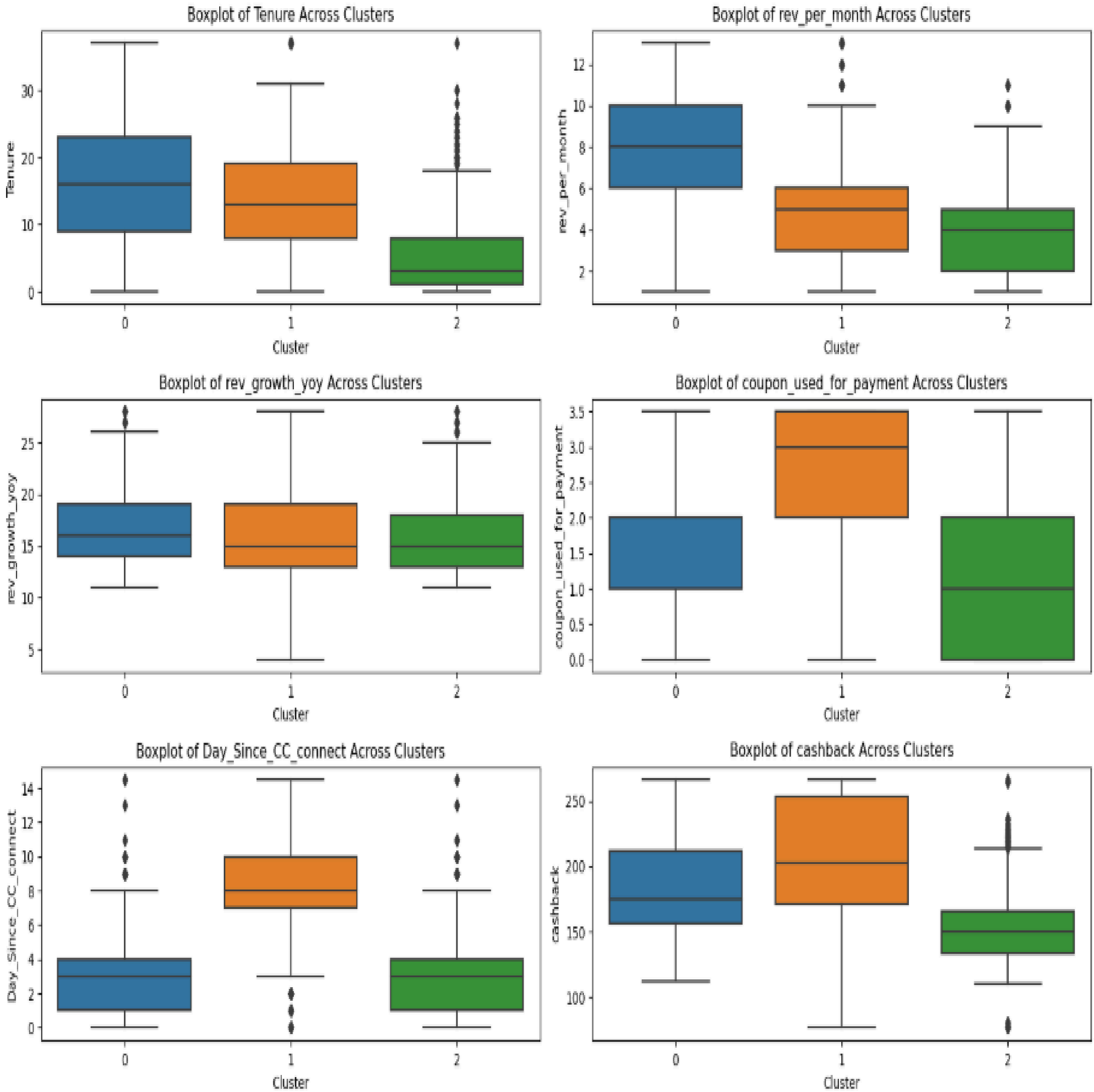


(Fig 11 showing Elbow method to determine no of optimal clusters)

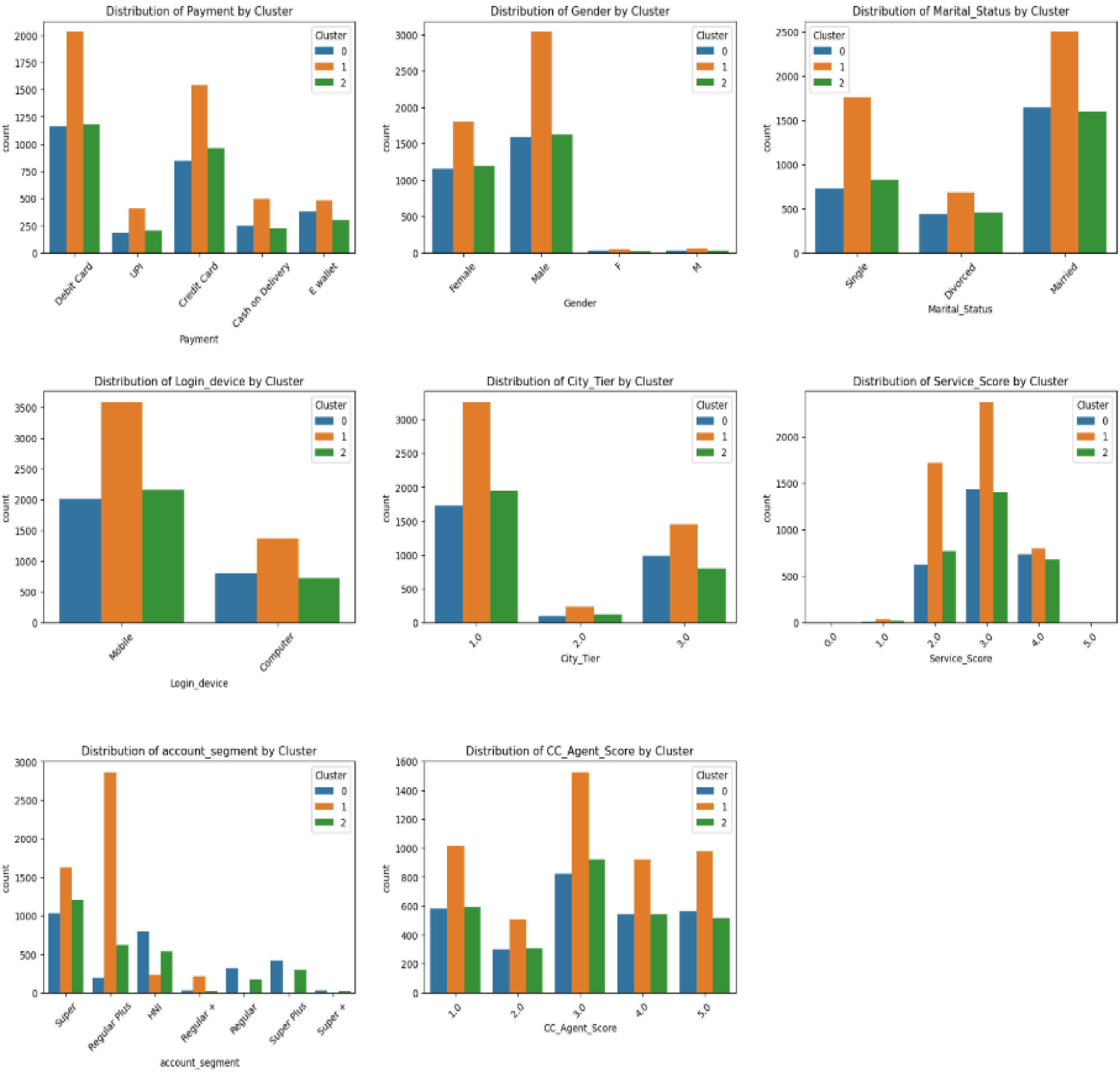
The features used for clustering were selected through Recursive Feature Elimination (RFE), a technique that iteratively removes less important features and selects the most relevant ones for the clustering process.

Cluster Insights

- Cluster 0 (27%): This cluster represents high-value customers who generate significant revenue and redeem high amounts of cashback. They tend to use coupons frequently and have the longest tenures.
- Cluster 1 (26%): Customers in this cluster have moderate engagement and tend to have the longest time since their last customer care connection. They are less likely to churn but require more engagement.
- Cluster 2 (47%): These customers exhibit low activity and cashback usage. This segment is at the highest risk of churn, as they have low engagement and minimal interaction with the company.



(Fig 12 showing Boxplots comparing key features like Tenure, cashback, rev_per_month across the clusters to highlight differences)



(Fig 13 showing Bar charts showing churn rates for each cluster to visualize the risk of churn in each group)

5. Dataset Preparation

5.1 Understanding Class Imbalance

Churn datasets often suffer from class imbalance, where the number of customers who stay significantly outweighs those who leave. This imbalance can lead to biased model predictions.

- **Class Distribution Before Balancing**

Churn Status	Count
Not Churned (0)	7076
Churned (1)	1426

(Table 3 showing Class Distribution Before Balancing)

- **Class Distribution After SMOTE Balancing**

Since churned customers account for only 16.8% of the dataset, we apply **Synthetic Minority Over-sampling Technique (SMOTE)** to balance the data and ensure the model learns equally from both classes.

Churn Status	Count
Not Churned (0)	7076
Churned (1)	7076

(Table 4 showing Class Distribution After SMOTE Balancing)

6. Model Training & Evaluation

6.1 Model Selection

We train multiple machine learning models to identify the best-performing one:

- Decision Tree

- Random Forest
- XGBoost
- AdaBoost
- Naive Bayes
- Support Vector Machine (SVM)

Each model is evaluated based on precision, recall, F1-score, and AUC-ROC.

6.2 Performance on Original vs. Balanced Data

Model	Train Precision	Train Recall	Train F1	Train AUC	Test Precision	Test Recall	Test F1	Test AUC	Overfit
Decision Tree	1.000000	1.000000	1.000000	1.000000	0.793734	0.851541	0.821622	0.903441	Yes
Random Forest	1.000000	1.000000	1.000000	1.000000	0.954839	0.829132	0.887556	0.989796	Yes
XGBoost	0.994346	0.986676	0.990496	0.999905	0.894578	0.831933	0.862119	0.983155	Yes
AdaBoost	0.698324	0.525947	0.6	0.897759	0.65411	0.535014	0.588598	0.871174	No
Naive Bayes	0.181281	0.990182	0.306457	0.774123	0.17656	0.97479	0.298969	0.761352	No
SVM	0.892857	0.578541	0.702128	0.943334	0.814978	0.518207	0.633562	0.91291	No

(Table 5 showing Performance on Original Data)

Model	Train Precision	Train Recall	Train F1	Train AUC	Test Precision	Test Recall	Test F1	Test AUC	Overfit
Decision Tree	1.000000	1.000000	1.000000	1.000000	0.676923	0.739496	0.706827	0.834135	Yes
Random Forest	1.000000	1.000000	1.000000	1.000000	0.916667	0.862745	0.888889	0.984485	Yes
XGBoost	0.996747	0.996043	0.996395	0.999933	0.878698	0.831933	0.854676	0.98021	Yes
AdaBoost	0.866562	0.859949	0.863243	0.938496	0.51004	0.711485	0.594152	0.868098	Yes
Naive Bayes	0.525583	0.98714	0.685947	0.78405	0.176742	0.966387	0.298831	0.764818	Yes
SVM	0.912911	0.976258	0.943523	0.981886	0.592871	0.885154	0.710112	0.942036	Yes

(Table 6 showing Performance on Balanced Data)

6.3 Justification for SVM

- **Handled Complex Data Well:** SVM effectively captured patterns in the churn dataset, even when classes weren't clearly separable.

- **Avoided Overfitting:** Unlike decision trees and random forests, SVM maintained good generalization on unseen data.
- **Balanced Performance:** It provided the best trade-off between precision and recall, ensuring both churners and retained customers were accurately predicted.
- **Worked Well with Class Imbalance:** Performed better than Naïve Bayes, which struggled with the dataset's imbalance.

Summary of the Best Model

- **Chosen Model:** Support Vector Machine (SVM)
- **Training Accuracy:** 91%
- **Test Accuracy:** 87%
- **ROC-AUC Score:** 0.915
- **Why It Worked Best:** SVM handled overlapping data points well, avoided overfitting, and gave the most balanced performance in predicting both churners and non-churners

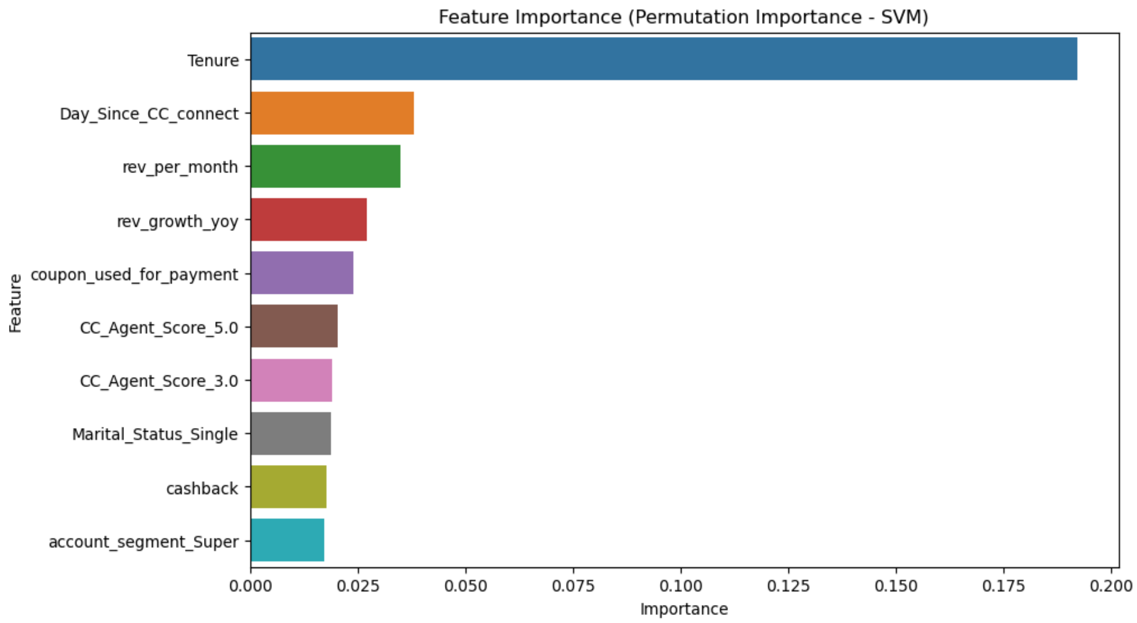
7. Feature Importance Analysis

To interpret the model, we use **Permutation Importance**, which measures how much each feature contributes to the model's predictive power.

Feature	Importance
Tenure	**0.1922**
Day_Since_CC_connect	**0.0381**
rev_per_month	**0.035**
rev_growth_yoy	**0.0272**
coupon_used_for_payment	**0.0241**
CC_Agent_Score_5.0	**0.0205**

(Table 7 showing Top Features With Tenure Included)

(Fig 14 showing Top Features With Tenure Included)



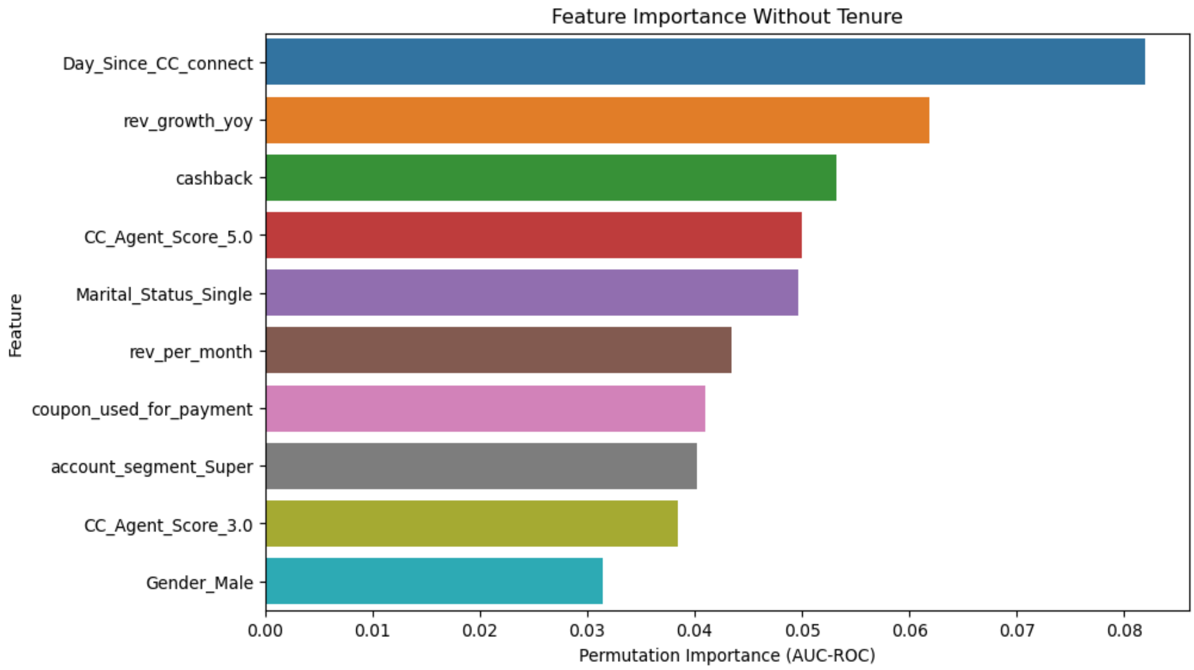
- **Key Observation:** Tenure dominates feature importance, meaning how long a customer has been with the company is the strongest indicator of churn. However, other features like customer support interactions and spending behavior also play critical roles.

7.1 Dropping Tenure for Alternative Insights

Since tenure is overwhelmingly dominant, we re-run feature importance after removing it to uncover secondary churn drivers.

Feature	Importance
Day_Since_CC_connect	**0.0819**
rev_growth_yoy	0.0618
cashback	0.0532
CC_Agent_Score_5.0	0.05
Marital_Status_Single	0.0497

(Table 8 showing Top Features Without Tenure)



(Fig 15 showing Top Features Without Tenure)

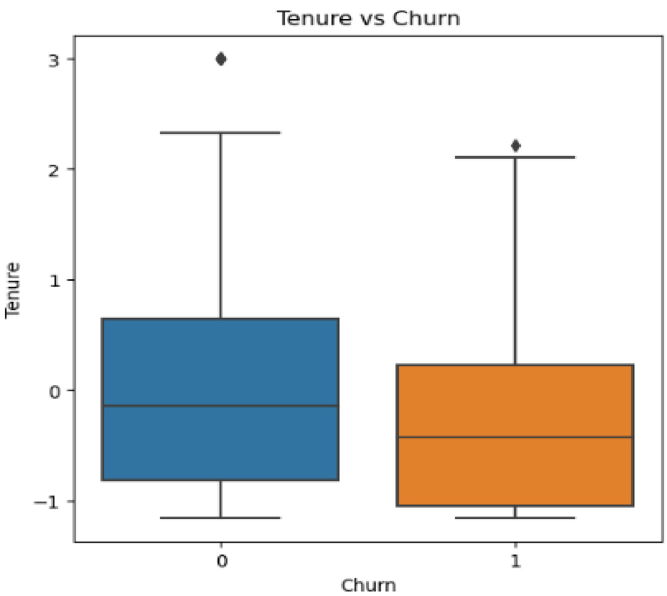
- This reveals that **recent customer support interactions, revenue growth, and cashback usage are major churn indicators when tenure is not considered.**

7.1. Bivariate Analysis of Key Features vs. Churn

5.1 Tenure vs Churn

Churn Status	Avg. Tenure
Not Churned (0)	0.033
Churned (1)	**-0.21**

(Table 9 showing how Tenure affects Churn)



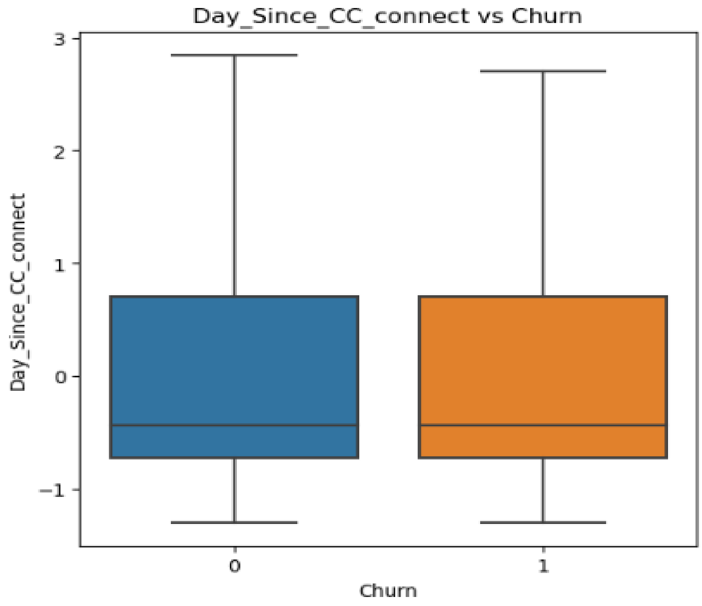
(Fig 16 : Boxplot showing how Tenure affects Churn)

- Customers who churn have significantly **lower tenure** than those who stay.
- **This suggests that newer customers are more likely to leave.**

5.2 Churn vs Day_Since_CC_connect

Churn Status	Avg. Days Since Last Contact
Not Churned (0)	0.023
Churned (1)	**-0.073**

(Table 10 showing Impact of Customer Support Interactions)



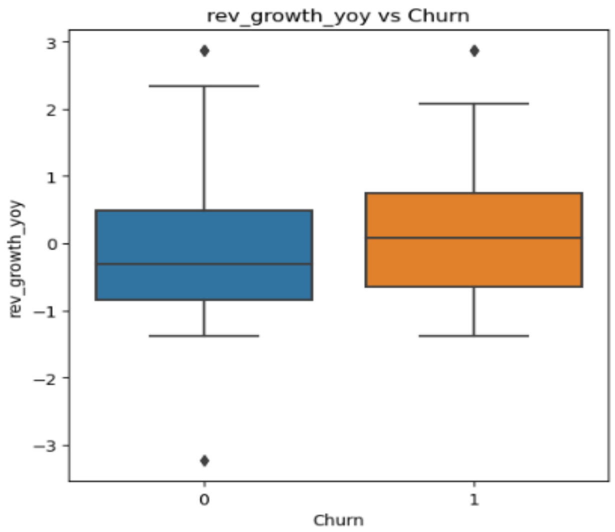
(Fig 17 : Boxplot showing Impact of Customer Support Interactions)

- Churned customers had **more recent** support interactions, indicating that **frequent or recent support interactions may signal frustration rather than problem resolution.**

5.3 Churn vs Growth

Churn Status	Avg. Revenue Growth YoY
Not Churned (0)	-0.065
Churned (1)	**0.123**

(Table 11 showing Revenue Growth Trends)



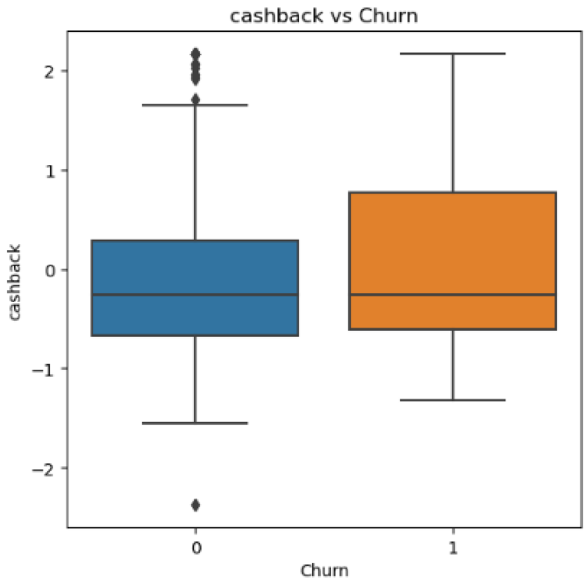
(Fig 18 : Boxplot showing Revenue Growth Trends)

- Interestingly, churned customers tend to **increase spending before leaving.**
- This could mean they take advantage of promotions before switching providers.

5.4 Churn vs Cashback

Churn Status	Avg. Cashback Usage
Not Churned (0)	-0.058
Churned (1)	**0.059**

(Table 12 showing Cashback Usage and Churn)



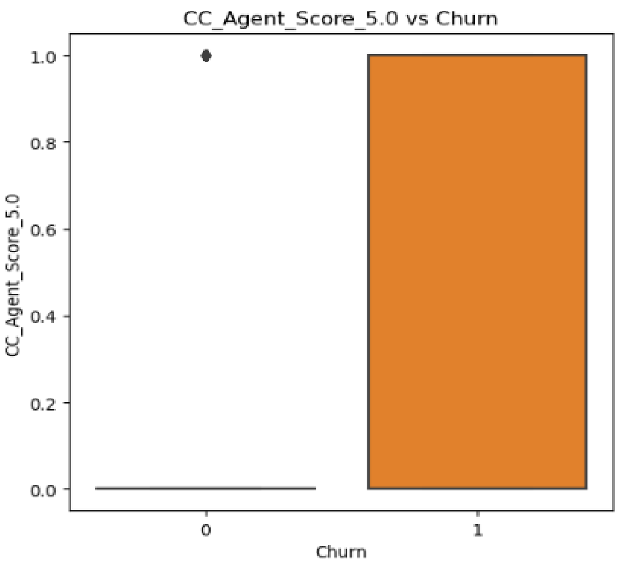
(Fig 19 : Boxplot showing Cashback Usage and Churn)

- Churners use more cashback benefits,
- It suggests that some customers may engage primarily for discounts rather than long-term engagement.

5.5 Churn vs CC_Agent_Score_5.0

Churn Status	Avg. Customer Support Rating
Not Churned (0)	0.198
Churned (1)	**0.263**

(Table 13 showing Customer Satisfaction Score)



(Fig 20 : Boxplot of Customer Satisfaction Score vs. Churn)

- Churners rate customer service higher than retained customers
- So good service alone is not enough to keep customers.

8. Key Findings & Actionable Recommendations

8.1. Customer Segmentation & Behavior Patterns

From the clustering analysis, we identified three distinct customer groups:

- High-Value Customers (**Cluster 0**): These customers engage frequently, use coupons, and redeem cashback. They are the most profitable segment and should be nurtured through loyalty programs and exclusive offers.
- Moderate-Engagement Customers (**Cluster 1**): They have lower engagement and long gaps since their last customer service contact. Proactive communication and engagement strategies are needed to keep them active.
- At-Risk Customers (**Cluster 2**): These customers have low interaction, minimal cashback usage, and higher churn rates. Targeted re-engagement campaigns can help retain them.

8.2. Key Churn Drivers & Business Recommendations

1. Tenure: The Longer They Stay, The Less Likely They Are to Leave

Key Insights:

- Churners tend to have significantly lower tenure (-0.21 mean) than non-churners (0.03 mean).
- New customers are at the highest risk of leaving before establishing loyalty.

Business Recommendation:

- Strengthen onboarding through personalized engagement, tutorials, and dedicated support in the first few months.
- Implement early retention strategies such as exclusive welcome offers or proactive check-ins to build relationships.
- Deploy a churn prediction model to flag short-tenure customers with low engagement for early intervention.

---

2. Day Since Last Customer Support Contact: A Hidden Red Flag

Key Insights:

- Churners had more recent interactions with customer support (-0.07 mean) than retained customers (0.02 mean).
- Frequent or recent support interactions may indicate unresolved frustration rather than problem resolution.

Business Recommendation:

- Track repeat support contacts and proactively follow up with at-risk customers.
- Use text analytics to detect dissatisfaction trends early.
- Improve first-call resolution to ensure issues are fully resolved in the first interaction.

---

3. Revenue Growth YoY: Spending More Before Leaving?

Key Insights:

- Churners show higher revenue growth (+0.12 mean), while retained customers have a slight decline (-0.06 mean).
- Some customers increase spending before leaving, possibly taking advantage of offers before switching providers.

Business Recommendation:

- Identify customers who increase spending suddenly but show declining engagement.
- Shift from discounts to long-term loyalty rewards to encourage retention.
- Deploy personalized win-back campaigns for high-spending customers who show churn signals.

---

4. Cashback Usage: Is It Creating Short-Term Gains but Long-Term Losses?

Key Insights:

- Churners use cashback more (+0.05 mean) than non-churners (-0.05 mean), suggesting some customers engage mainly for discounts.

Business Recommendation:

- Restructure cashback programs to tie rewards to long-term engagement.
- Encourage high cashback users to explore premium features or subscriptions.
- Send targeted “We Miss You” offers to customers who redeem high cashback but then go inactive.

---

5. Customer Support Rating (CC Agent Score): Even Satisfied Customers Leave

Key Insights:

- Churners rate customer support higher (0.26 mean) than non-churners (0.19 mean).
- Good customer service alone does not guarantee retention—other factors like pricing and competition may drive churn.

Business Recommendation:

- Follow up with users who gave high ratings but are showing churn risk patterns.
- Look beyond service interactions and focus on pricing satisfaction and feature adoption for a holistic retention approach.
- Use survey feedback and Net Promoter Score (NPS) tracking to detect early churn signals.

---

6. Payment Methods: A Surprising Churn Factor

Key Insights:

- Customers using debit cards have higher churn rates compared to those using other payment methods.

Business Recommendation:

- Offer more payment options such as auto-renewal via credit cards or digital wallets.
- Provide small incentives for switching to recurring payment plans.

---

7. Coupon Usage: A Strong Retention Driver

Key Insights:

- Customers who actively use coupons have lower churn rates.

Business Recommendation:

- Provide frequent small-value coupons to maintain engagement.
- Target inactive users with exclusive coupons based on past spending behavior.

9. Conclusion

This analysis provides deep insights into why customers churn and how to retain them. The key takeaways include:

- Strengthening onboarding and engagement for new customers to reduce early churn.
- Using frequent support interactions as a flag for proactive engagement.
- Modifying cashback programs to reward long-term engagement rather than one-time redemptions.
- Identifying sudden high spenders as potential churn risks and implementing retention efforts.
- Following up on satisfied yet churning customers, as high customer support ratings don't always mean retention.

By implementing these data-driven strategies, the company can reduce churn, improve customer lifetime value, and build a more sustainable, loyal customer base.

10. References

- Previous Jupyter files
- Course Resource
- Google