Customer Churn Analysis

Prepared By :
Shakirah Benson

Customer Churn Overview

Executive Summary

This analysis explores the key drivers behind customer churn on an e-commerce platform. Using Python for data cleaning, visualization, segmentation, and predictive modeling, we identified actionable insights that can help improve customer retention. Key findings include the impact of order frequency, satisfaction scores, and complaint status on churn likelihood. We also developed customer segments using clustering and evaluated churn risk using a Random Forest classifier.

Project Goals

1. Clean and prepare the dataset for analysis
2. Explore data to identify patterns and trends
3. Visualize key relationships
4. Predict churn using classification modeling
5. Deliver actionable insights for retention strategies

Key Business Questions

1. Do customers with fewer orders or low engagement tend to churn more often?
2. What is the relationship between satisfaction scores, tenure, and churn?
3. How do other factors like complaints, payment modes, and cashback amounts correlate with churn?

Tools & Methods

- Python (pandas, matplotlib, seaborn, scikit-learn)
- Data Cleaning with Imputation and Type Conversion
- Correlation Analysis and Visualizations
- K-Means Clustering for Segmentation
- Random Forest Classification for Churn Prediction
- Customer Lifetime Value Estimation

Data Cleaning & Preparation

The dataset used in this analysis is sourced from Kaggle, specifically from the Customer Churn Dataset created by Muhammad Shahid Azeem. It contains detailed information about customer behavior on an e-commerce platform, including tenure, order count, satisfaction scores, preferred payment methods, and churn status.

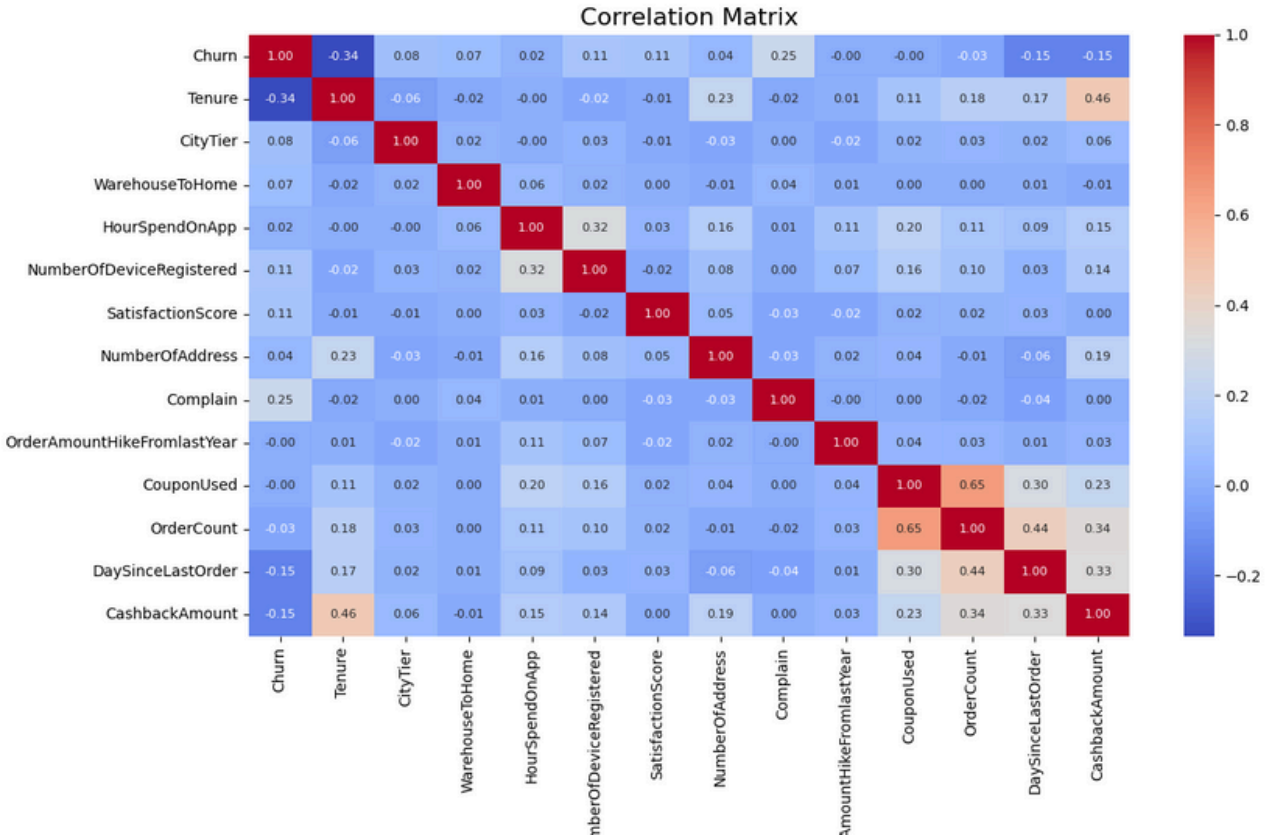
We first load the data, clean any missing or invalid entries, and ensure all numeric columns are correctly formatted. We used mean imputation for numeric columns with missing values. The dataset is loaded into a Pandas DataFrame, and missing or non-numeric values in important columns (e.g., `Tenure`, `SatisfactionScore`, `OrderCount`) are converted to numeric format, filling missing values with the column mean.

```
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   CustomerID                   5630 non-null   int64  
 1   Churn                        5630 non-null   int64  
 2   Tenure                       5630 non-null   float64
 3   PreferredLoginDevice         5630 non-null   object 
 4   CityTier                     5630 non-null   int64  
 5   WarehouseToHome              5630 non-null   float64
 6   PreferredPaymentMode         5630 non-null   object 
 7   Gender                       5630 non-null   object 
 8   HourSpendOnApp               5630 non-null   float64
 9   NumberOfDeviceRegistered     5630 non-null   int64  
 10  PreferedOrderCat             5630 non-null   object 
 11  SatisfactionScore            5630 non-null   int64  
 12  MaritalStatus                5630 non-null   object 
 13  NumberOfAddress              5630 non-null   int64  
 14  Complain                     5630 non-null   int64  
 15  OrderAmountHikeFromlastYear  5630 non-null   float64
 16  CouponUsed                   5630 non-null   float64
 17  OrderCount                   5630 non-null   float64
 18  DaySinceLastOrder            5630 non-null   float64
 19  CashbackAmount               5630 non-null   int64  
```

Visualizing the Data

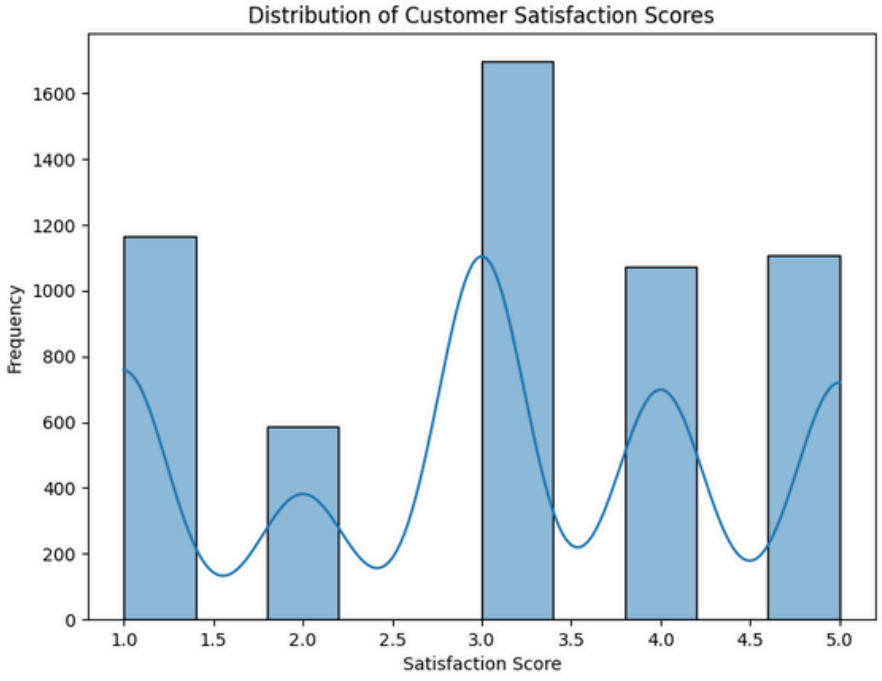
Correlation Matrix

To identify relationships between numeric variables, we generated a correlation matrix. This helps highlight key factors that could be contributing to customer churn. A correlation matrix was computed to identify key relationships among numerical variables. The matrix highlights significant correlations, such as between OrderCount and Churn (negative), and between SatisfactionScore and Churn (negative). This indicates that higher satisfaction and order frequency are associated with lower churn. These findings direct attention to customer engagement and satisfaction as key areas for retention.



Customer Satisfaction Scores

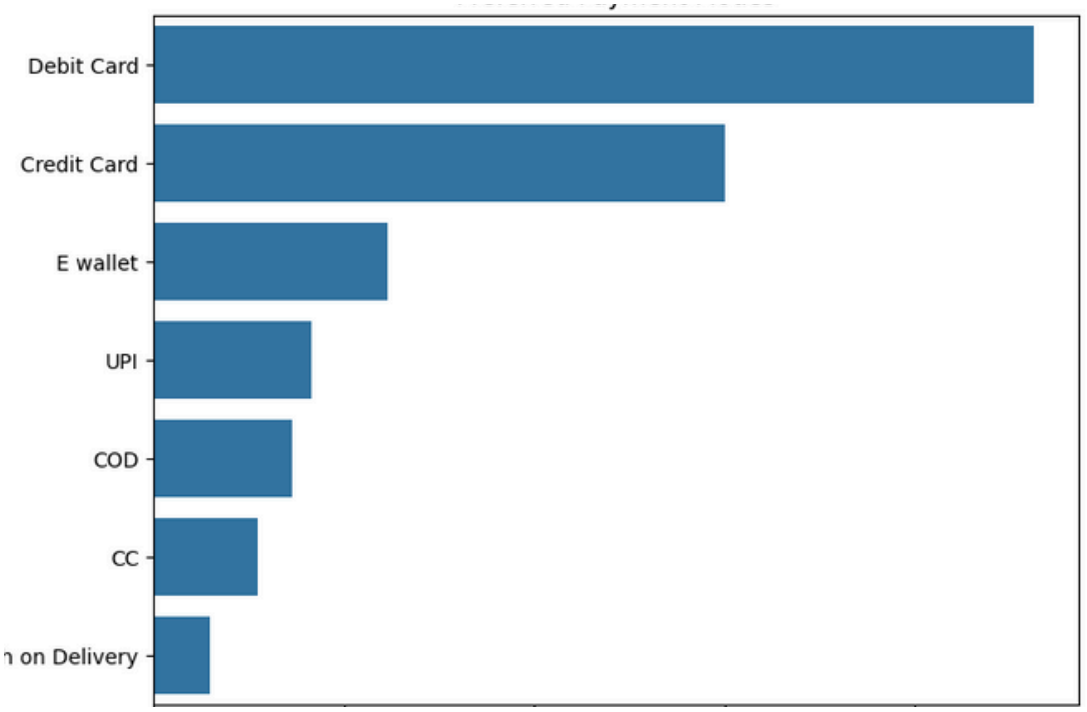
The histogram reveals that the majority of customers have moderate satisfaction scores. However, there is a notable group with low scores, potentially contributing to churn. Addressing these customers' concerns could improve retention.



Visualizing the Data

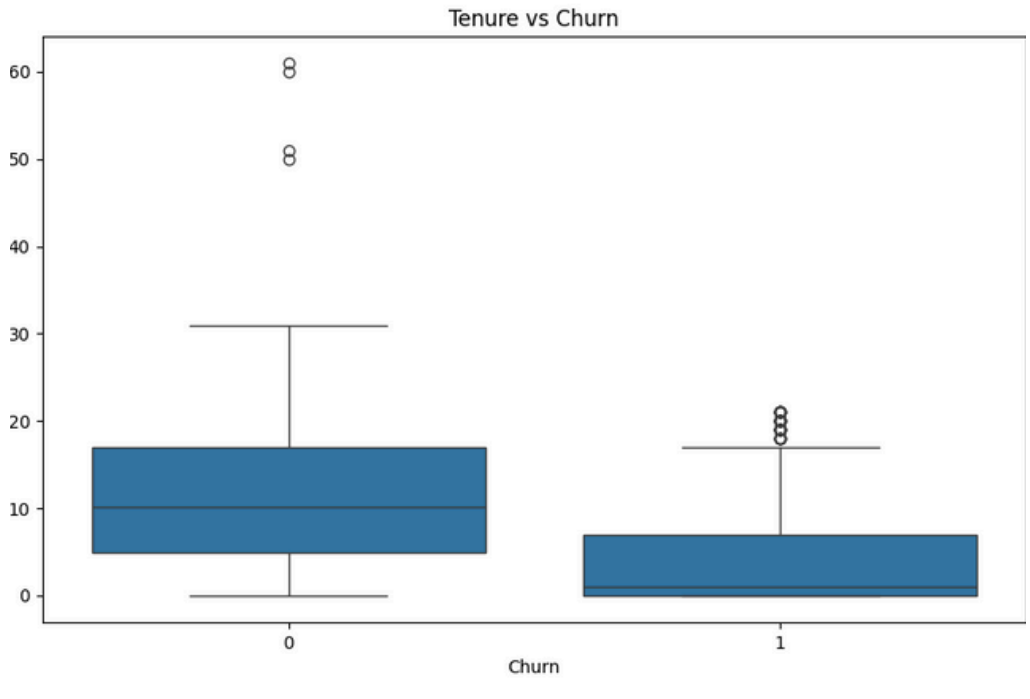
Payment Method Preferences

A count plot was used to visualize the distribution of preferred payment methods among customers. This helped identify the most and least popular options, providing insight into customer preferences that could inform product or UX decisions. Customers using less preferred payment methods may be less satisfied and more prone to churn, warranting exploration into whether certain payment modes affect user experience.



Tenure vs Churn

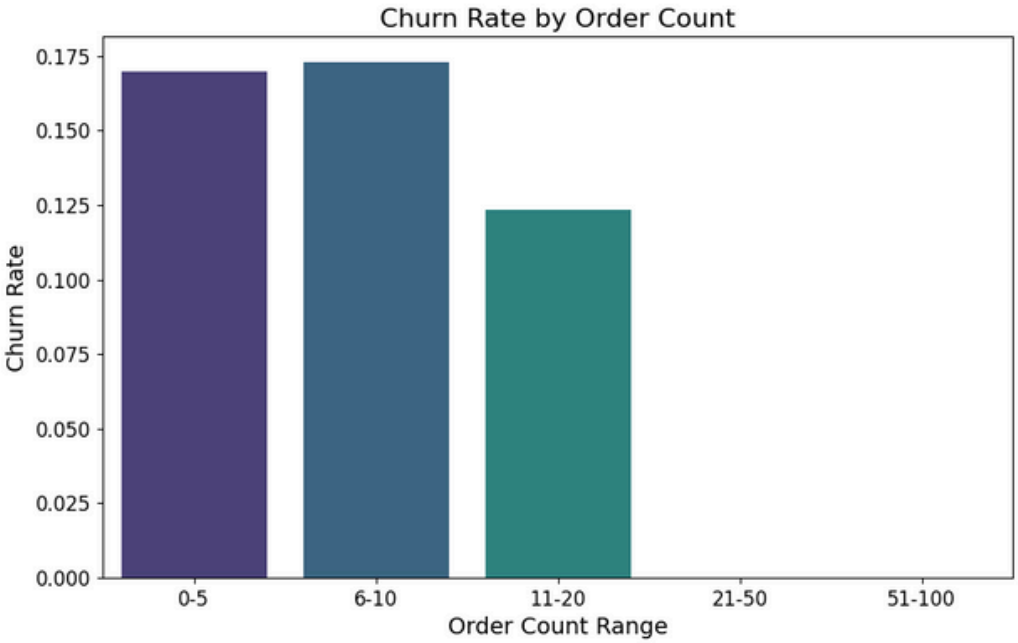
Customers with shorter tenures are significantly more likely to churn, highlighting the importance of strong onboarding and early engagement strategies. This insight was revealed through a boxplot comparing tenure across churned and retained users.



Visualizing the Data

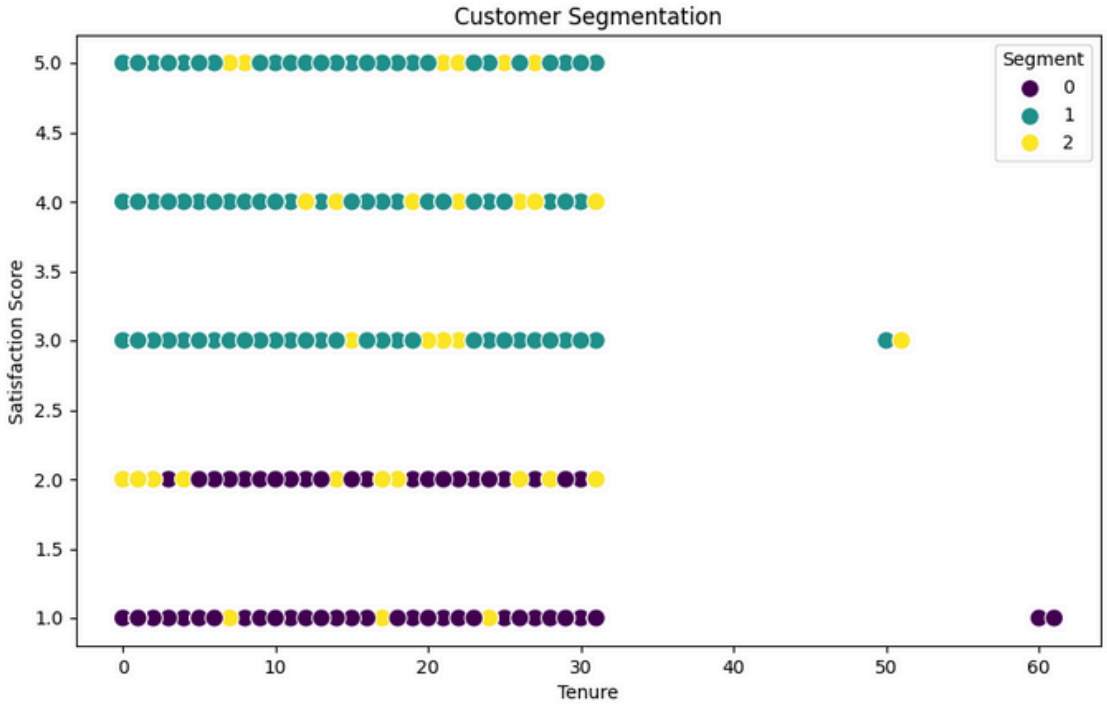
Churn Analysis by Order Count

A bar plot was used to examine how churn rates vary across different ranges of order counts. The results showed that customers with lower order counts (particularly in the 0–5 range) exhibited the highest churn rates. This insight highlights the importance of encouraging early repeat purchases to improve retention.



Customer Segmentation

K-Means clustering was applied to tenure, satisfaction scores, and order volume, resulting in three distinct customer segments. Segment 2 included high-tenure, high-frequency customers , our most valuable group for retention. Segment 0 showed low satisfaction and low order volume, signaling a churn-risk group while 1 showed satisfaction, but low activity. This segmentation supported data-driven targeting strategies for loyalty programs and re-engagement.



Visualizing the Data

Customer Lifetime Value (CLV)

Customer Lifetime Value is calculated by combining cashback earnings with churn risk estimates. This allowed us to identify high-value customers who were likely to stay, helping prioritize retention and upsell efforts to maximize ROI. Customers who receive higher cashback and have a lower likelihood of churning represent greater long-term value. By calculating Customer Lifetime Value (CLV), the top 10 most valuable customers were identified and prioritized for targeted retention efforts.

	CUSTOMERID	CLV
3541	53542	322.440136
5011	55012	322.231442
2880	52881	322.054895
4350	54351	321.447762
5169	55170	320.829107
3699	53700	320.829107
3646	53647	320.527812
5116	55117	320.480842
5276	55277	320.278600
4325	54326	320.272997

Customer Churn Predictor

A Random Forest Classifier was used to predict customer churn based on features such as tenure, satisfaction score, order frequency, and coupon usage. The model achieved an overall accuracy of 86%, performing especially well in identifying customers likely to remain active. Although recall for churned customers was lower, the model provides a strong foundation for an early warning system. With further tuning or class rebalancing, it could be deployed to trigger automated retention efforts for at-risk users.

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.89	0.94	0.92	941
1	0.58	0.42	0.49	185
ACCURACY			0.86	1126
MACRO AVG	0.74	0.68	0.70	1126
WEIGHTED AVG	0.84	0.86	0.85	1126

Insights & Recommendations

1. High-risk segments:
   a. Target customers with 0-5 orders with retention campaigns
   b. Implement immediate response protocols for customer complaints, as complaint-filing customers show higher churn rates
2. Customer Lifetime Value (CLV):
   a. Prioritize retention efforts on Segment 2 customers (high-tenure, high-satisfaction, frequent orders)
   b. Develop programs targeted towards customers with high CLV scores above 320
3. Measurement and Monitoring:
   a. Track customer satisfaction scores monthly
   b. Monitor complaint resolution time and impact
   c. Measure the effectiveness of retention campaigns using the developed metrics
   d. Develop an early warning system using the Random Forest model (86% accuracy) to identify potential churners

Next Steps

Future improvements could include tuning the Random Forest model for better churn recall, incorporating additional behavioral features (e.g., browsing activity or customer support interactions), and testing A/B retention strategies on high-risk segments. Deploying a real-time dashboard to monitor churn indicators and CLV could further support proactive decision-making.

Github Script Link

https://github.com/Skirah/customer-churn-analysis