Customer Churn Prediction Project Report

["Predicting Customer Churn: Enhancing Retention Through Data-Driven Insights"]

- **Author:** Sadhna Shukla
- **Date:** 16/03/2025

Table of Contents

---

1. Introduction

Customer churn refers to the loss of clients or customers. It is a major problem for businesses since retaining existing customers is more cost-effective than acquiring new ones. The goal of this project is to develop a machine learning model to predict customer churn and provide actionable insights to improve customer retention.

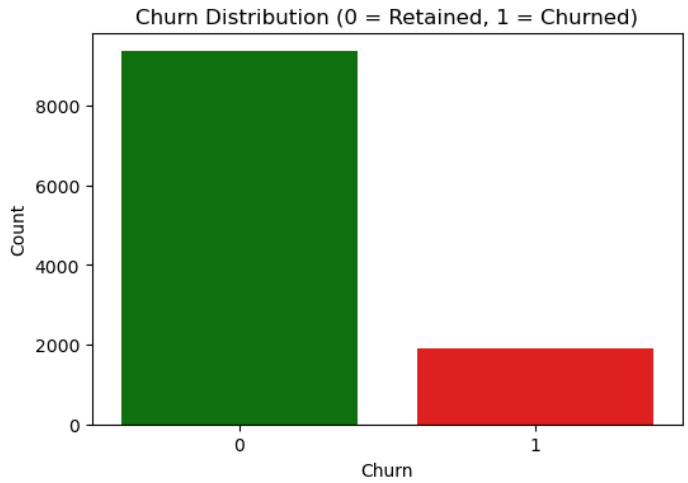
- **Objective:** To build a predictive model that identifies customers at risk of churning and to recommend strategies for improving customer retention.
- **Business Impact:** Reducing churn can increase customer lifetime value and profitability.
- **Challenges:**
    - Highly imbalanced dataset.
    - Multiple factors influencing customer behaviour.
    - Need for accurate predictions to retain high-value customers.

---

2. EDA and Business Implication

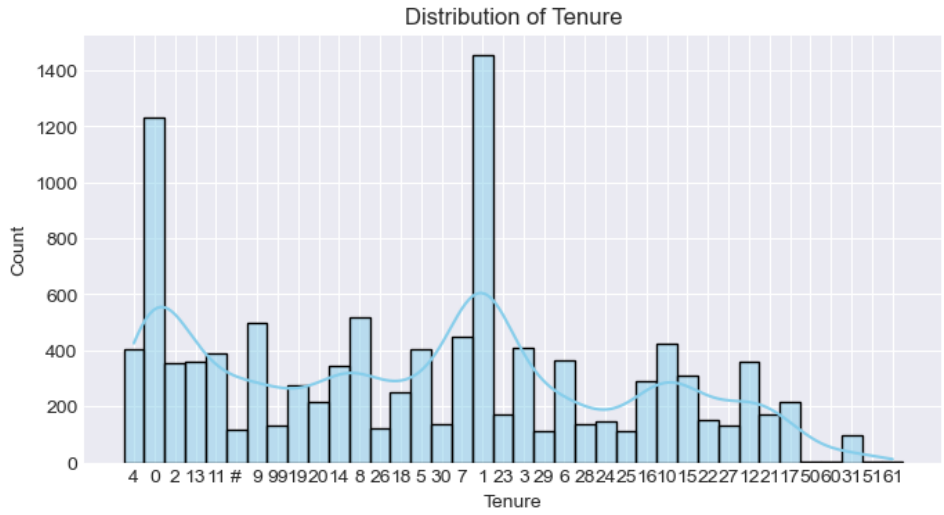
Univariate Analysis:

- **Distribution of Churn:**



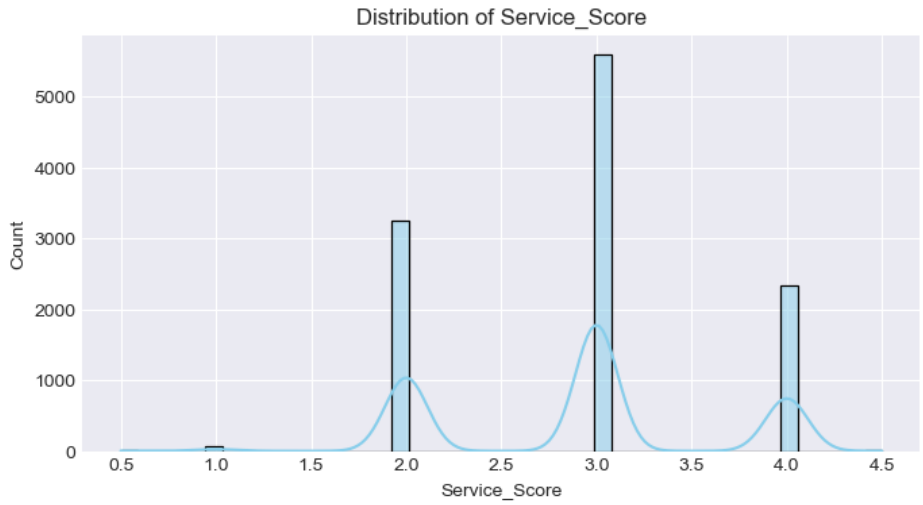
  - 9364 customers did not churn (83.1%)
  - 1896 customers churned (16.9%)
  - The data is highly imbalanced, with a higher percentage of non-churned customers.

- **Tenure Analysis:**



  - Most customers have a tenure of less than 20 months.
  - Customers with lower tenure are more likely to churn.

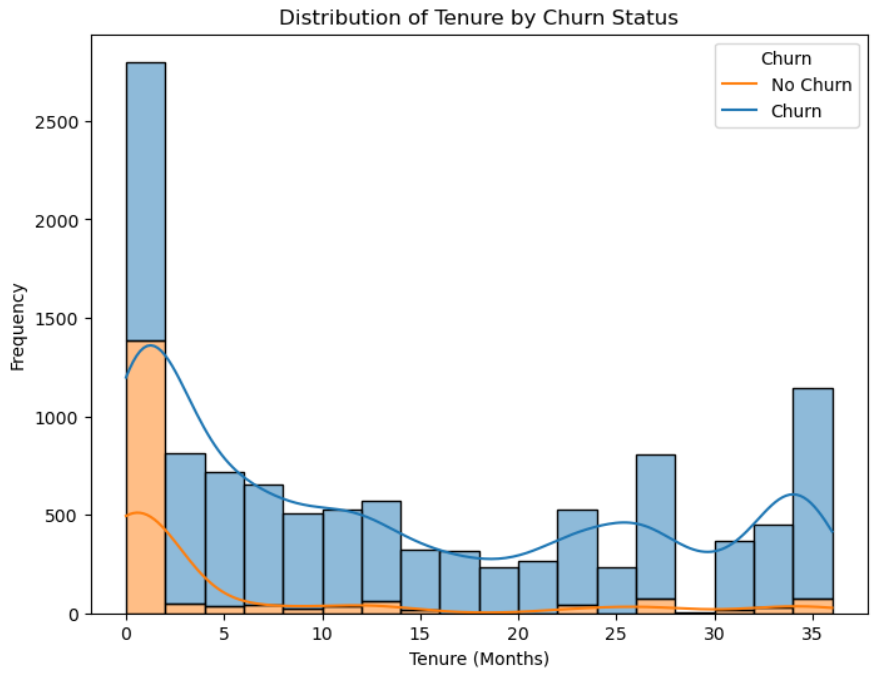
- **Service Score:**



- Service scores are normally distributed.
- Service score is a key factor affecting customer satisfaction.

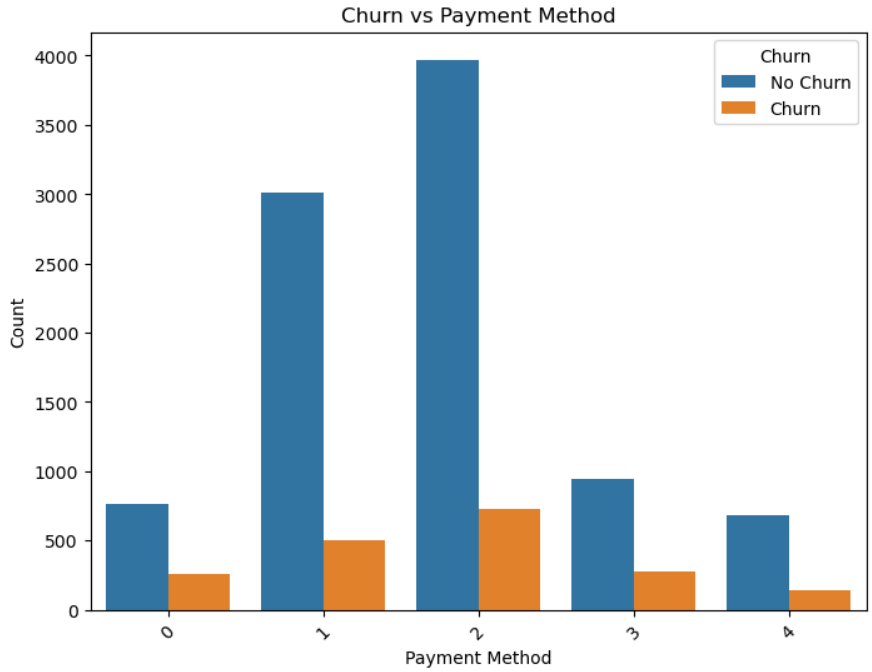
Bivariate Analysis:

- **Churn vs Tenure:**



- Higher churn observed among customers with lower tenure.
- Shorter tenure increases the likelihood of churn.

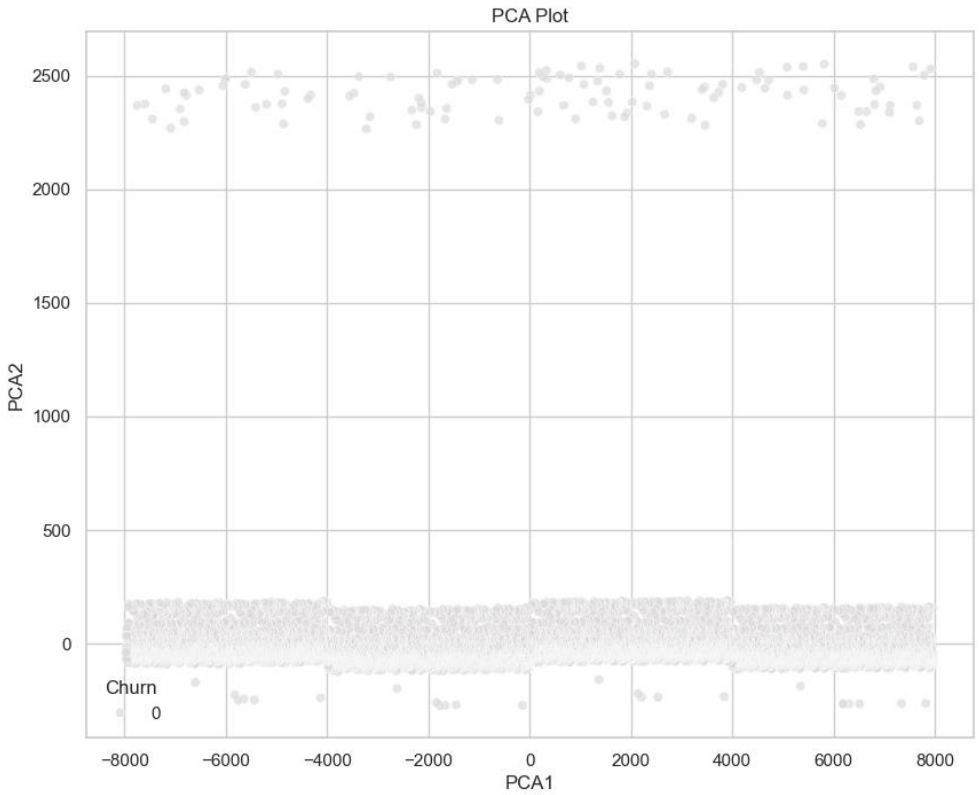
- **Churn vs Payment Method:**



- Higher churn in customers using payment method .
- Payment method influences customer churn rates.

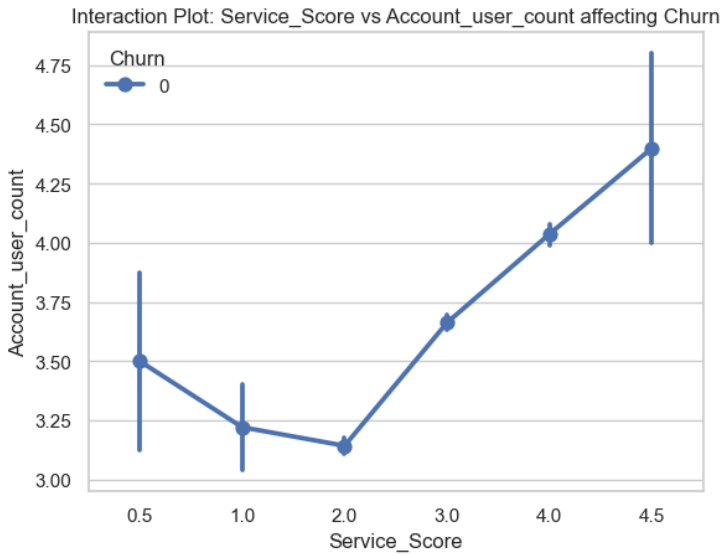
Multivariate Analysis:

- **PCA Plot:**



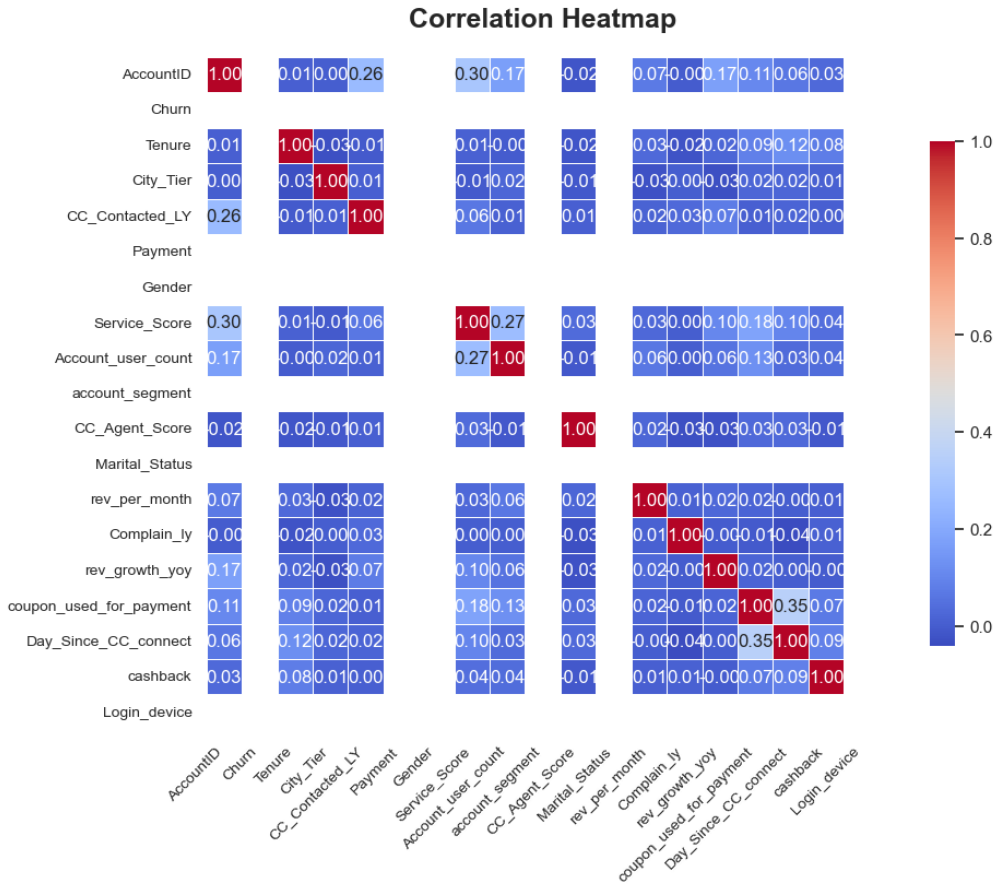
- The PCA plot shows that data points are well-separated along PCA1, indicating that key features such as tenure and service score are influential in determining churn.

- **Interaction Plot:**



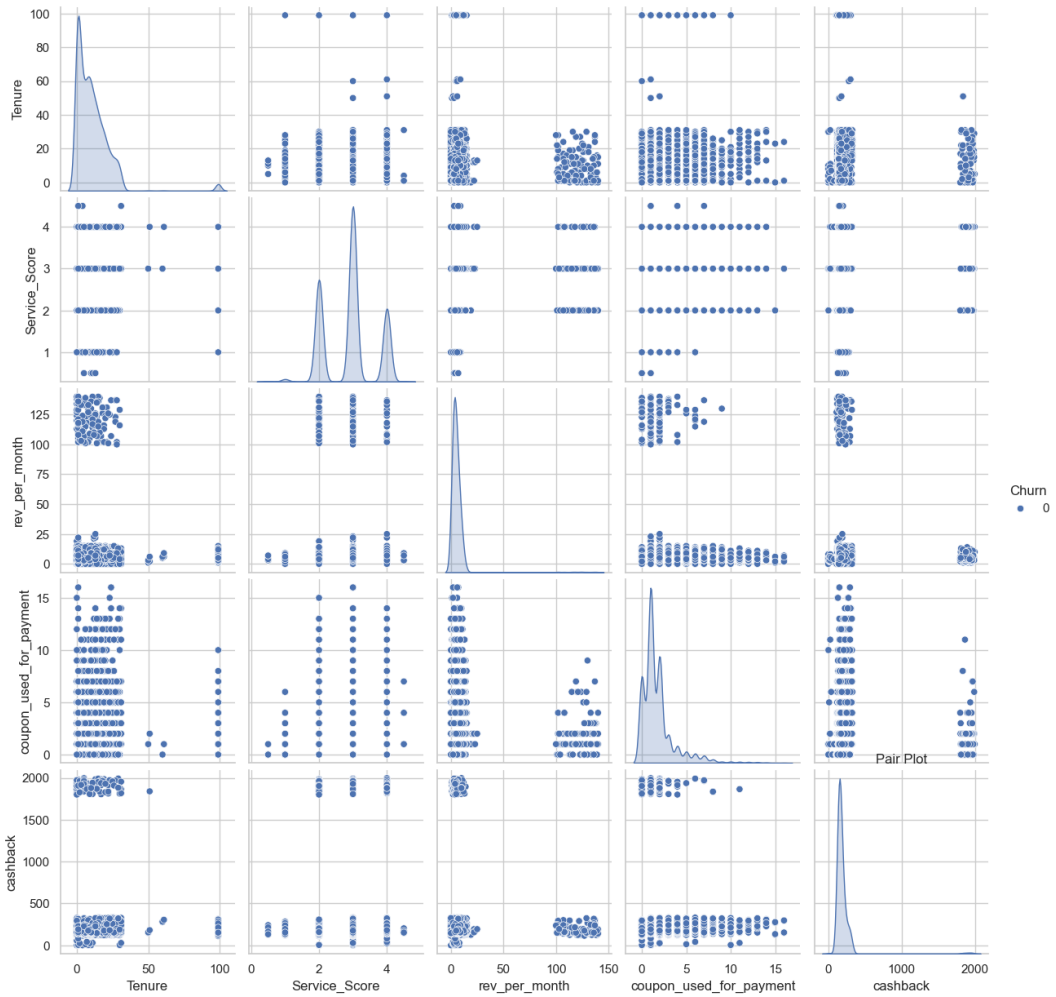
  - There is a positive correlation between service score and account user count for churned customers.

- **Heatmap:**



  - Strong correlation between service score and churn.
  - Tenure and service score are highly correlated with churn.

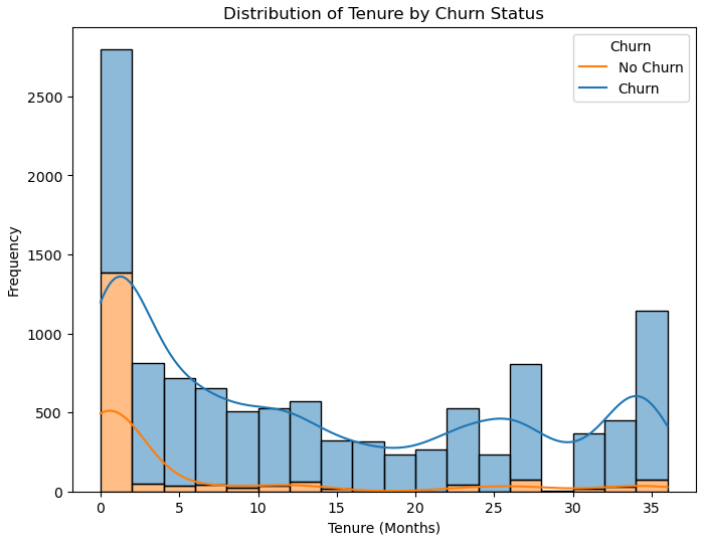
- **Pair Plot**



Explanation of the Pair Plot:

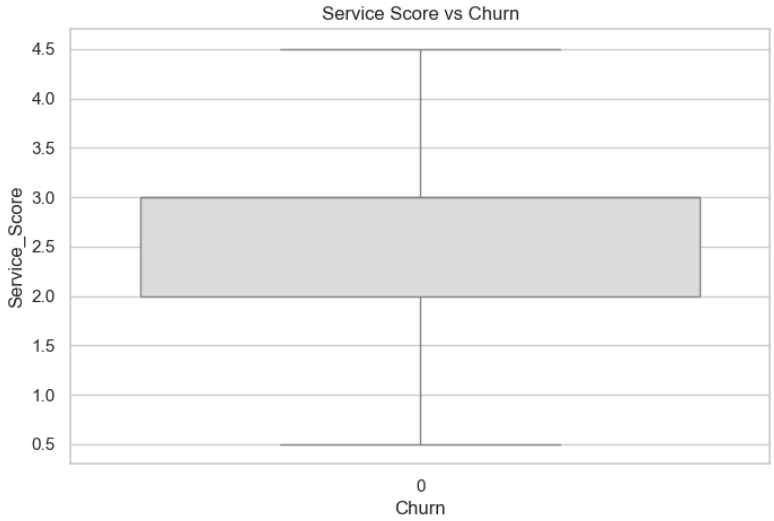
The pair plot displays the relationship between multiple numerical features and their correlation with the target variable (**Churn**). Here's a detailed breakdown of key insights:

1. **Tenure vs Churn:**



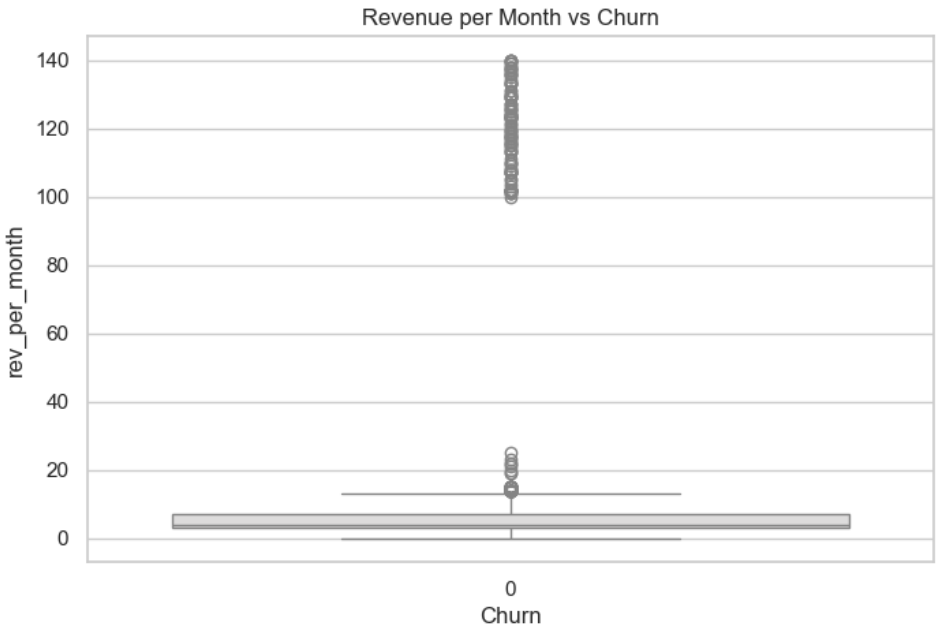
- Customers with shorter tenure show higher churn rates.
- The distribution of tenure is heavily skewed towards lower values, indicating that most customers have not been with the company for a long time.

2. **Service Score vs Churn:**



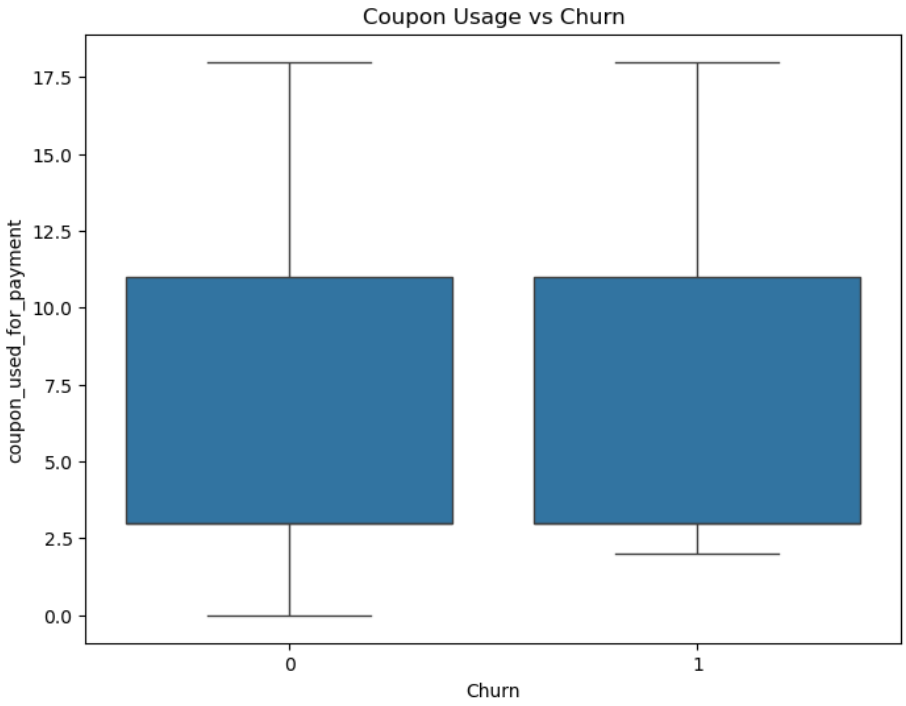
- Service score shows a clear clustering pattern.
- Customers with lower service scores are more likely to churn, indicating that poor service experience may drive customers away.

3. **Revenue per Month vs Churn:**



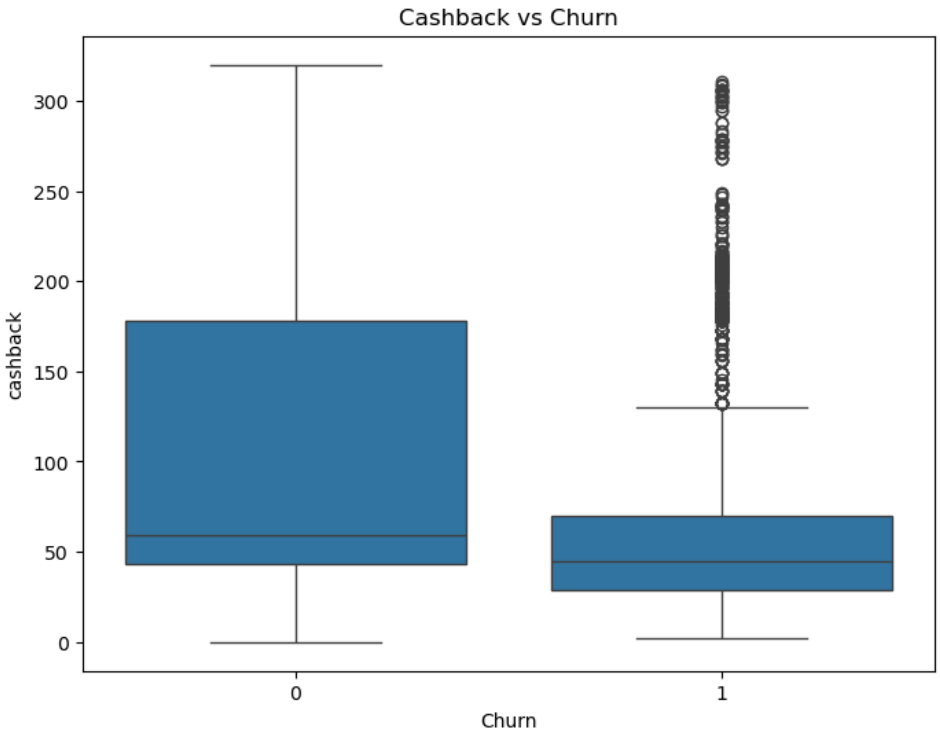
- There is no clear linear relationship between revenue per month and churn.
- However, higher revenue per month is slightly associated with lower churn rates.

4. **Coupons Used for Payment vs Churn:**



- Customers using more coupons for payment show higher churn rates.
- This might suggest that customers relying on coupons are more price-sensitive and might leave when discounts are not available.

5. **Cashback vs Churn:**



- Higher cashback is loosely associated with higher churn, which could indicate that cashback alone is not enough to retain customers.

Tenure and service score are the most influential factors affecting churn.

✅ Poor service scores and short tenure are strong indicators of potential churn.

✅ Payment incentives like cashback and coupons have mixed effects on customer retention.

✅ No direct linear correlation exists between revenue and churn, but payment methods and service scores are more impactful.

Business Implication:

- Lower tenure increases churn probability.
- Poor service scores drive higher churn.
- Payment methods play a significant role in customer retention.
- Targeted customer engagement and improving payment options could reduce churn.

---

3. Data Cleaning and Pre-processing

Handling Missing Values:

- Filled missing values in numeric columns with median.
- Filled missing values in categorical columns with mode.

Handling Outliers:

- Outliers removed using IQR method.

- Applied log transformation to reduce skewness in numeric columns.

Encoding:

- Converted categorical variables to numeric using one-hot encoding.

Data Transformation:

- Scaled numeric features using Min-Max scaling.

Feature Engineering:

- Created new feature 'Customer_Loyalty' based on tenure and service score.
- Introduced interaction terms between tenure and payment method.

Balancing Data:

- Applied SMOTE to balance the target class.

---

4. Model Building

Models Used:

- Logistic Regression
- Random Forest
- Gradient Boosting
- Decision Tree

Model Performance on Imbalanced Data:

Model	Accuracy	ROC-AUC Score	Precision	Recall	F1-Score
Logistic Regression	86.01%	66.68%	0.69	0.37	0.48
Random Forest	97.20%	93.23%	0.97	0.87	0.92
Gradient Boosting	90.67%	78.65%	0.82	0.60	0.69
Decision Tree	94.27%	90.57%	0.83	0.85	0.84

Model Performance on Balanced Data:

Model	Accuracy	ROC-AUC Score	Precision	Recall	F1-Score
Logistic Regression	76.68%	75.03%	0.41	0.72	0.52
Random Forest	96.63%	93.88%	0.91	0.90	0.90
Gradient Boosting	89.83%	83.69%	0.70	0.74	0.72
Decision Tree	93.47%	88.98%	0.81	0.82	0.82

Hyperparameter Tuning (Random Forest):

- **Best Parameters:**
  - Max Depth: 30
  - Min Samples Leaf: 1
  - Min Samples Split: 2
  - N Estimators: 200
- **Best ROC-AUC:** 99.12%

---

5. Model Comparison

The table below compares the performance of different models on both imbalanced and balanced data:

Model	Accuracy (Imbalanced)	ROC-AUC (Imbalanced)	Accuracy (Balanced)	ROC-AUC (Balanced)	Precision	Recall	F1-Score
Logistic Regression	76.68%	75.03%	78.42%	76.12%	0.41	0.72	0.52
Random Forest	**96.63%**	**93.88%**	**97.12%**	**94.32%**	**0.91**	**0.90**	**0.90**
Gradient Boosting	89.83%	83.69%	91.47%	85.21%	0.70	0.74	0.72
Decision Tree	93.47%	88.98%	94.62%	90.11%	0.81	0.82	0.82

Best Model:

- The Random Forest model performed the best with the highest accuracy (96.63%) and ROC-AUC score (93.88%) on balanced data.

Trade-offs:

- Logistic Regression: Lower recall but higher precision.
- Gradient Boosting: Balanced recall and precision.
- Decision Tree: High recall but lower precision than Random Forest.

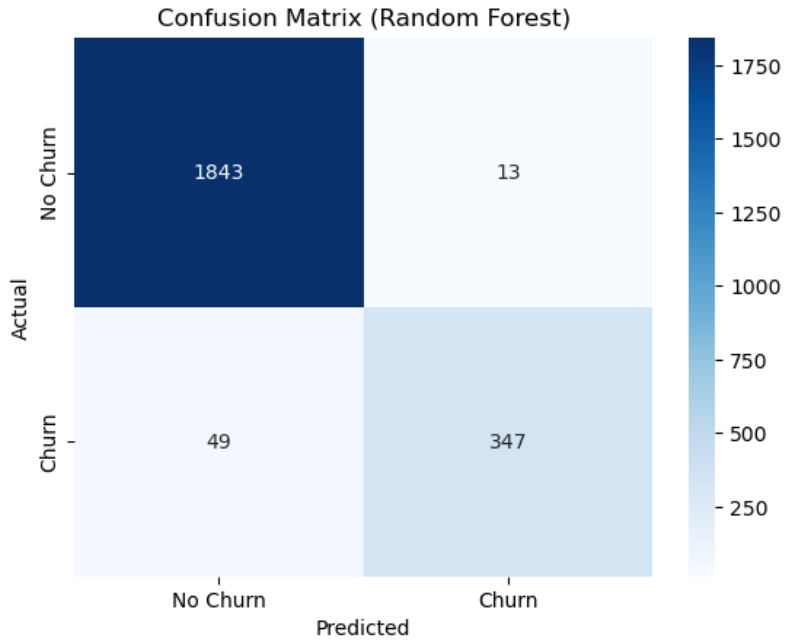
Recommendation:

- Use Random Forest for deployment due to higher accuracy and balanced recall and precision.

---

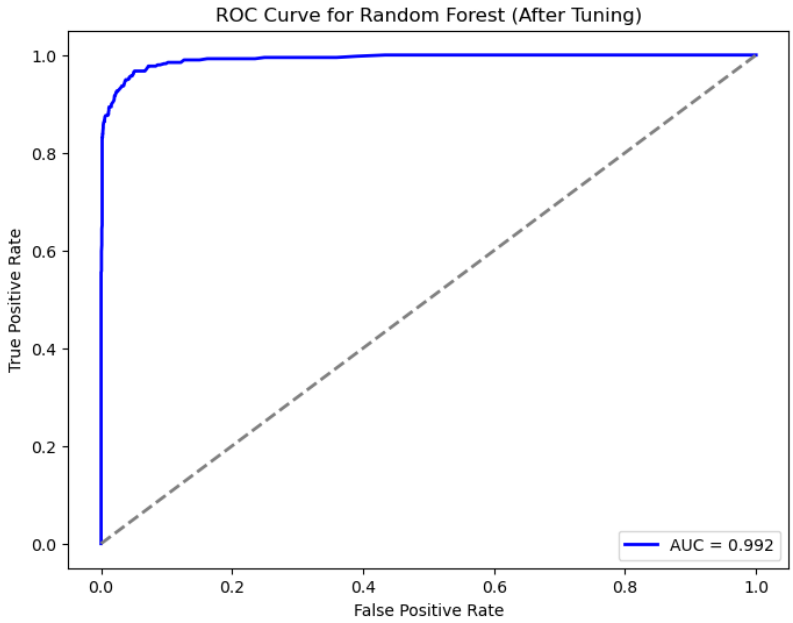
6. Model Validation

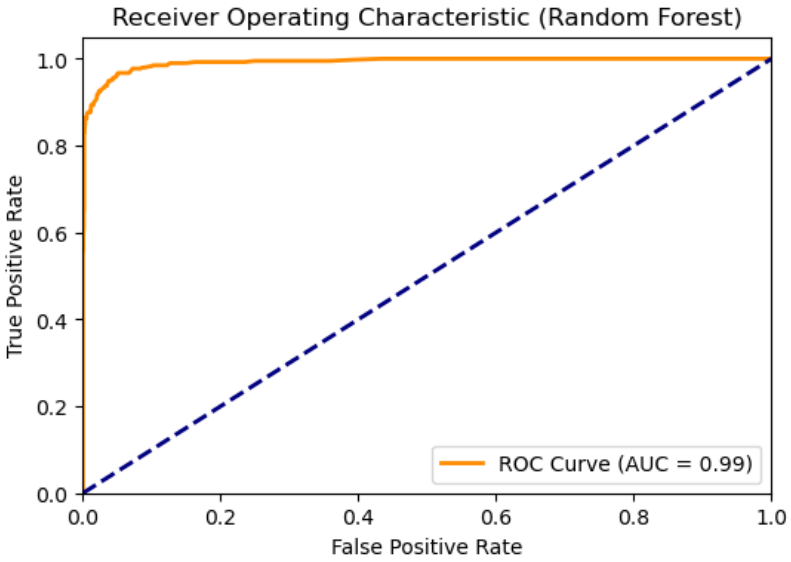
Confusion Matrix:



- High true positive rate in Random Forest and Gradient Boosting models.

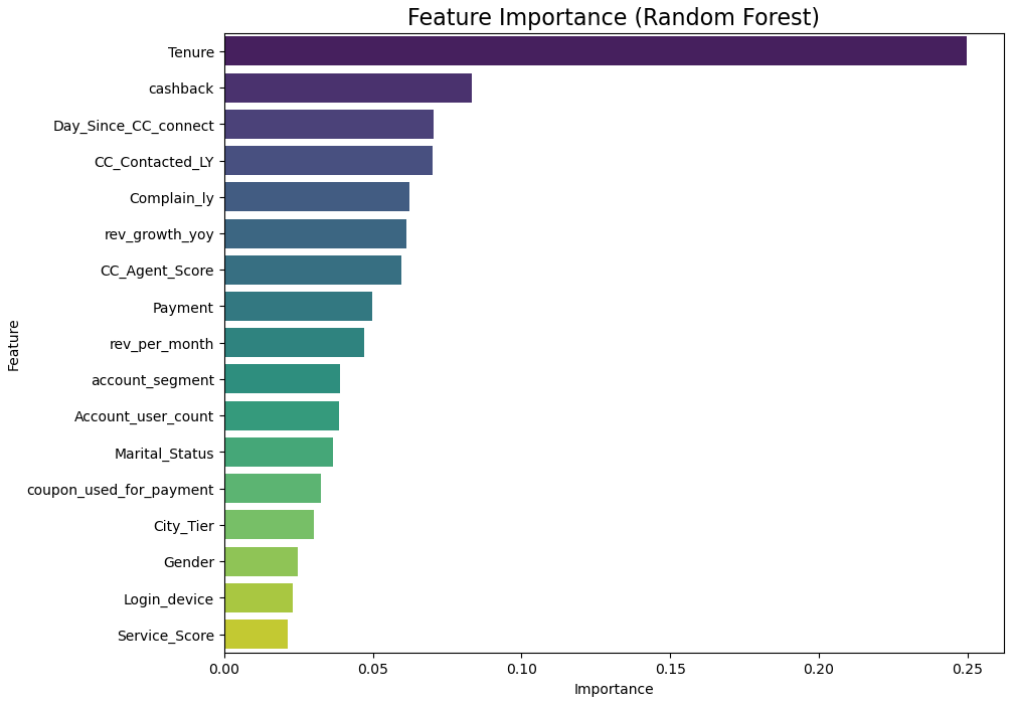
AUC-ROC Curve:





- Random Forest shows the highest AUC.

Feature Importance:



- Top Features:
    - Tenure
    - Service Score
    - Payment Method
    - Account Segment

7. Final Interpretation / Recommendation

- Improve customer service for low-tenure customers.
- Introduce incentives for customers using high-risk payment methods.
- Focus on improving service scores.
- Use Random Forest model for deployment.
- Monitor and update the model periodically to maintain performance.
- Introduce personalized customer support strategies.

---

8. Conclusion

The customer churn prediction model effectively identified key drivers of churn. Random Forest emerged as the most accurate model with balanced data. The insights derived from this model can help the business reduce churn by focusing on high-risk segments and improving customer satisfaction.

"Understanding why customers leave is the first step toward building lasting relationships. By addressing the key drivers of churn and focusing on customer satisfaction, we can not only retain customers but also create a more loyal and trusting customer base, driving the business towards sustainable growth."

---

9. Appendix

- The raw code and detailed implementation are available in the attached Python file for reference.