Telecom Churn Prediction

Analyzing Customer Behavior and Predicting Churn

Introduction

- Identifying high-value customers who may churn
- Understanding key factors influencing churn
- Developing predictive models to prevent churn

Understanding Churn

- Customers switch over time, not instantly
- Three phases: Good, Action, and Churn
- Key to identifying early churn signs

Dataset Description

1) Customer-Level Data – The dataset contains telecom customer usage details across four months (June, July, August, September), labeled as 6, 7, 8, and 9.

2) Feature Types – It includes demographic details, recharge patterns, call & internet usage, and revenue-related attributes.

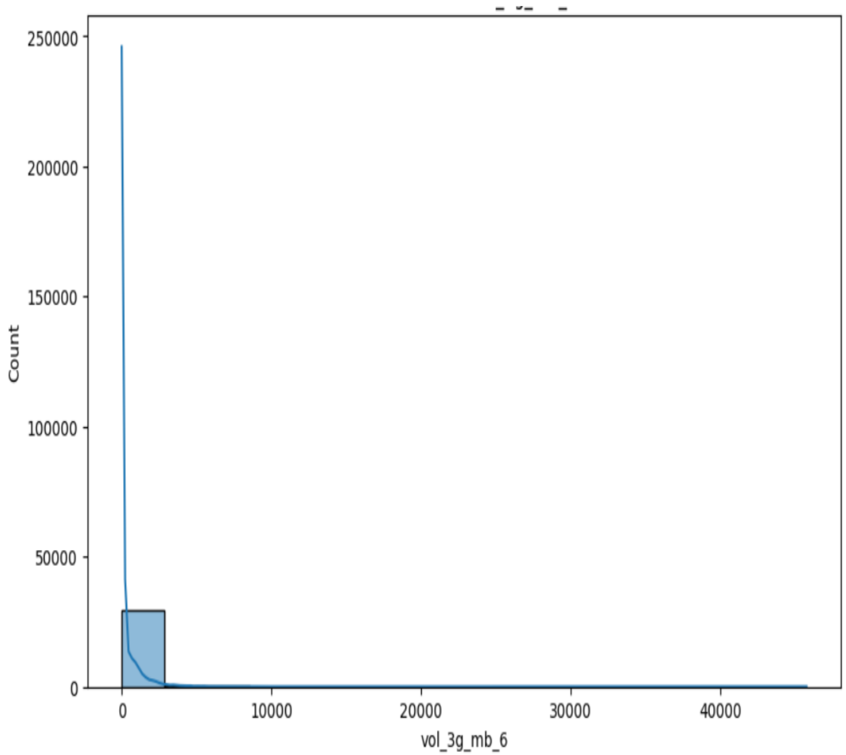
3) Churn Labeling – Churn is defined based on the September (9th month): Customers who made no calls (incoming/outgoing) and did not use mobile internet in September are marked as churners (1), else (0).

4) High-Value Customers – The analysis focuses on the top 30% revenue-generating customers, determined using the 70th percentile of the average recharge amount in the first two months (June & July).

Data Preprocessing

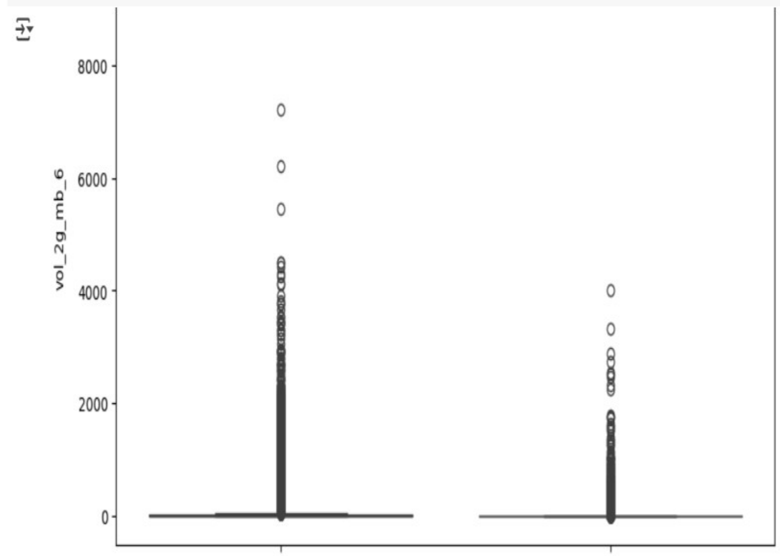
- Handling missing values
- Filtering high-value customers
- Addressing class imbalance (SMOTE/Undersampling)
- Feature scaling

EDA(Exploratory data analysis)



Bivariate Analysis (Churn vs Features)

```
# Plot churn vs features for key variables
for col in numerical_cols:
    plt.figure(figsize=(10, 6))
    sns.boxplot(x='churn', y=col, data=df_high_value)
    plt.title(f'Churn vs {col}')
    plt.show()
```



- The correlation matrix revealed that variables related to usage (such as total incoming and outgoing calls, as well as mobile internet usage) are strongly correlated with each other. However, these variables show a weak negative correlation with churn, highlighting that lower usage is a significant predictor of churn.

Model Building

- Logistic Regression to identify key churn predictors
- Handling multicollinearity
- Random Forest and Decision Tree for comparison
- Hyperparameter tuning

Evaluation Metrics

- Accuracy, Sensitivity (Recall), Specificity

Evaluate the Model

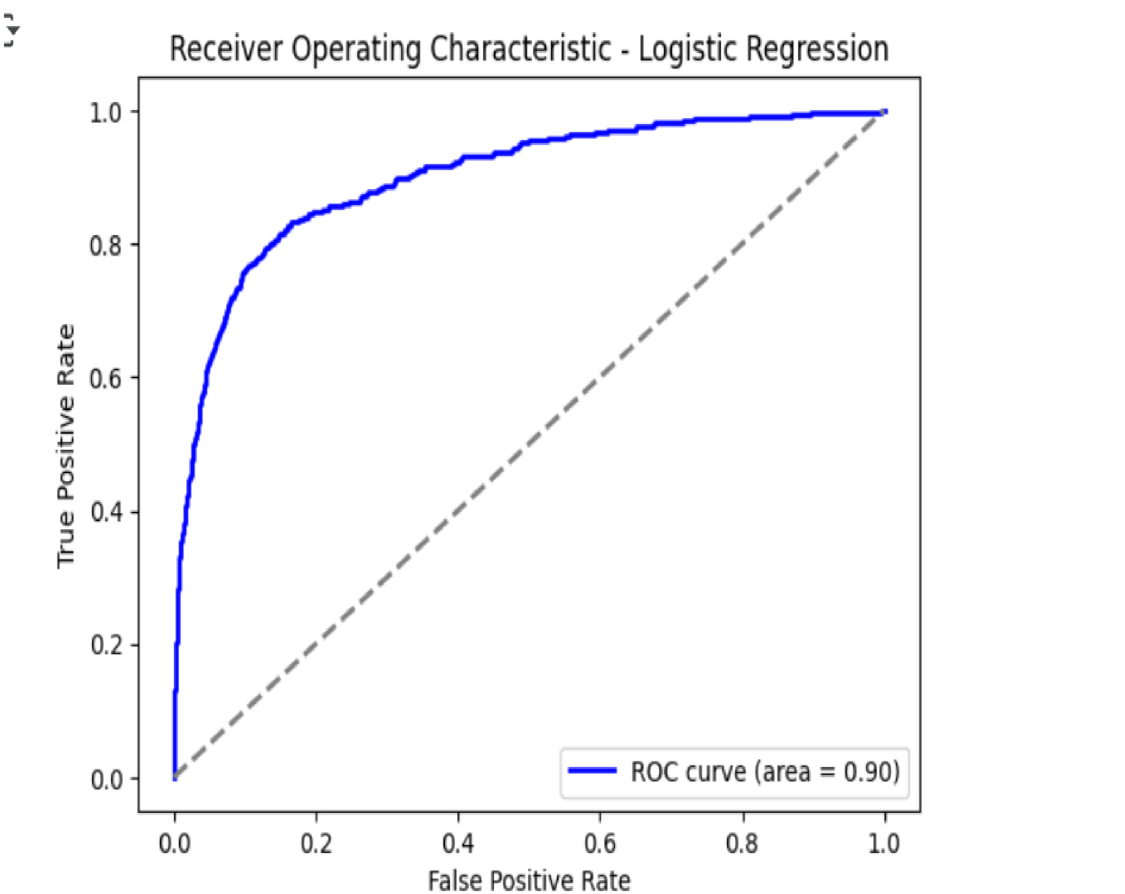
```
[ ]
# Print evaluation metrics for Logistic Regression
print(f"Logistic Regression - Accuracy: {accuracy_log_reg}")
print(f"Logistic Regression - Sensitivity: {sensitivity_log_reg}")
print(f"Logistic Regression - Specificity: {specificity_log_reg}")
```

```
Logistic Regression - Accuracy: 0.9345327336331835
Logistic Regression - Sensitivity: 0.35067437379576105
Logistic Regression - Specificity: 0.9897884755652808
```

- ROC curve for logistic regression :

The Receiver Operating Characteristic (ROC) curve demonstrates the model's performance, achieving an area under the curve (AUC) score of 0.90, indicating strong predictive capability.

```
27 cm_log_reg = confusion_matrix(y_test, y_pred_log_reg) # Changed 'confusion' to 'confusi
```



Feature Importance & Insights

- Top Features influencing churn:
  - Total recharge amount (last 3 months)
  - Declining call duration
  - Reduced data usage
- These insights help target at-risk customers effectively.
- Identified key churn indicators
- Used plots and summary tables for visualization
- Actionable insights for customer retention

Recommendations

- Offer personalized plans & discounts
- Improve customer service & network quality
- Monitor high-risk customers proactively
- Monitor Recharge Behavior
- Target Low-Usage Customers

Conclusion

The analysis identifies key factors like low service usage, particularly in the churn phase, and high recharge amounts as critical indicators of churn. By focusing on early intervention for high-risk customers, especially those with declining service usage and low recharges, telecom companies can significantly reduce churn rates, leading to better retention and sustained revenue. Implementing targeted, data-driven retention strategies based on these findings can provide telecom operators with a competitive edge in a highly competitive market.