INTRODUCTION TO BUSINESS ANALYTICS

Group 3

Name	Roll No
Aishwarya Awchar	MBAA24073
Anup Kumar Behera	MBAA24078
Banoth Balaraju	MBAA24085
Deepshikha	MBAA24088
Vaishnavi Gupta	MBAA24141
Sarella Rahul Alex	MBAA24121
Shubham Keshwani	MBAA24126
Sritanu Debbarma	MBAA24132
Yash Patil	MBAA24146

Predicting Customer Churn in the Telecom Sector

Customer churn is a significant concern for telecom companies, as retaining customers is often more cost-effective than acquiring new ones. Churn refers to customers who have terminated their association with a service, and understanding the reasons behind churn is vital for minimizing customer loss and maximizing revenue.

By predicting churn, companies can proactively engage with at-risk customers, improving retention rates and revenue. This project utilizes the Telco Customer Churn Dataset from Kaggle to analyze factors influencing churn and apply machine learning techniques to predict it.

Data Processing Overview

1. Handling Missing Values

- Issue: Missing values in TotalCharges due to new customers with zero tenure.
- Action: Dropped rows with missing values (small fraction of data).
- Alternative: Impute with MonthlyCharges × Tenure (not applied here).

2. Encoding Categorical Variables

- Label Encoding: Binary variables (e.g., Gender, Churn) converted to 0 and 1.
- One-Hot Encoding: Multiclass variables (e.g., PaymentMethod) transformed into binary columns to preserve nominal relationships.

3. Standardizing Numerical Features

- Features like MonthlyCharges and TotalCharges scaled with StandardScaler (mean = 0, std dev = 1) to ensure uniform feature contribution.

4. Feature Selection

- Method: Correlation analysis to remove irrelevant/redundant features.
- Example: CustomerID excluded due to lack of predictive value.

5. Addressing Class Imbalance

- Observation: Only 26% of customers were classified as churners.
- Future Action: Techniques like SMOTE or cost-sensitive learning can address imbalance if needed.

6. Train-Test Split

- Split Ratio: 80% training, 20% testing.
  - Training Set: 5,634 records.
  - Testing Set: 1,409 records.
- Ensures evaluation on unseen data while maintaining a representative training sample.

Model Development

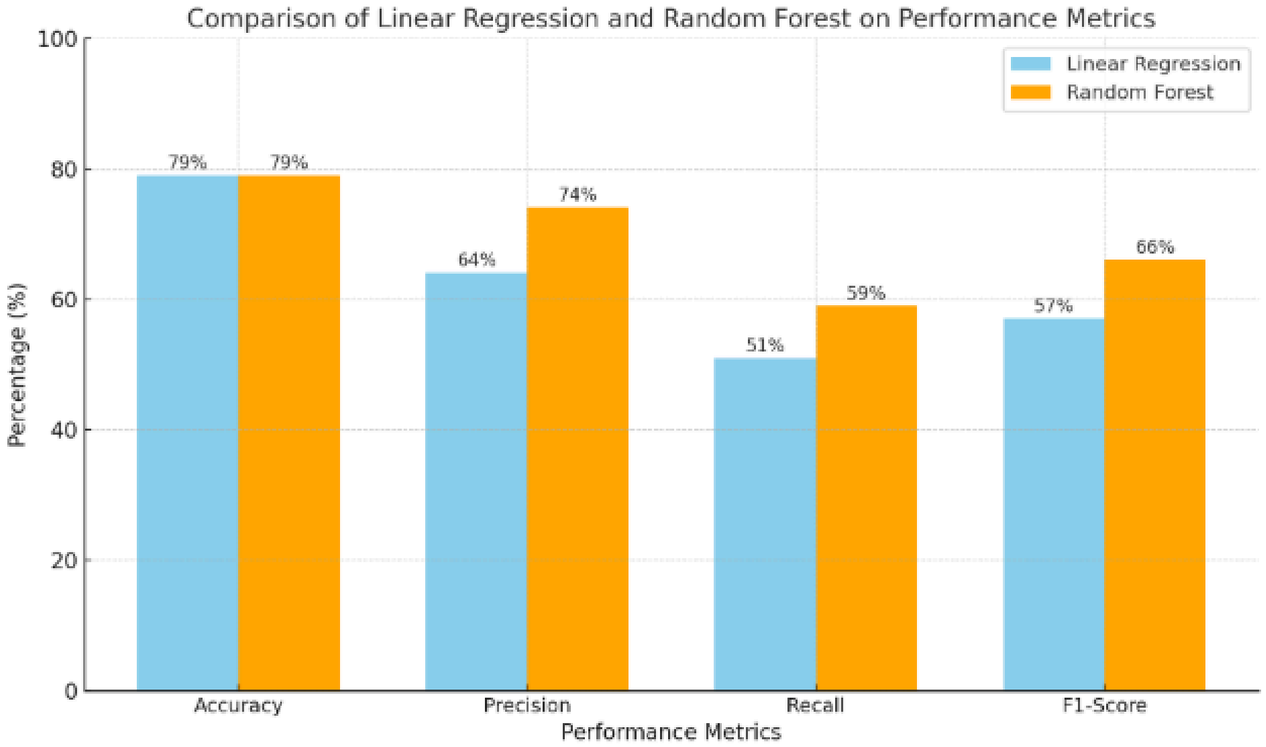
Chosen Models: Random Forest Classifier and Linear Regression

Aspect	Linear Regression	Random Forest
Model Type	Linear (global relationship)	Ensemble (non-linear, local patterns)
Interpretability	High (simple and explainable)	Moderate (less intuitive)
Complexity	Low (easy to compute)	High (computationally intensive)
Performance	Good for linear relationships	Good for complex, non-linear data
Overfitting Risk	Low (prone to underfitting)	Higher risk (mitigated by averaging)
Feature Handling	Requires scaled and numeric features	Handles categorical and numeric features

MODEL EVALUATION

Key Metrics for Churn Class:

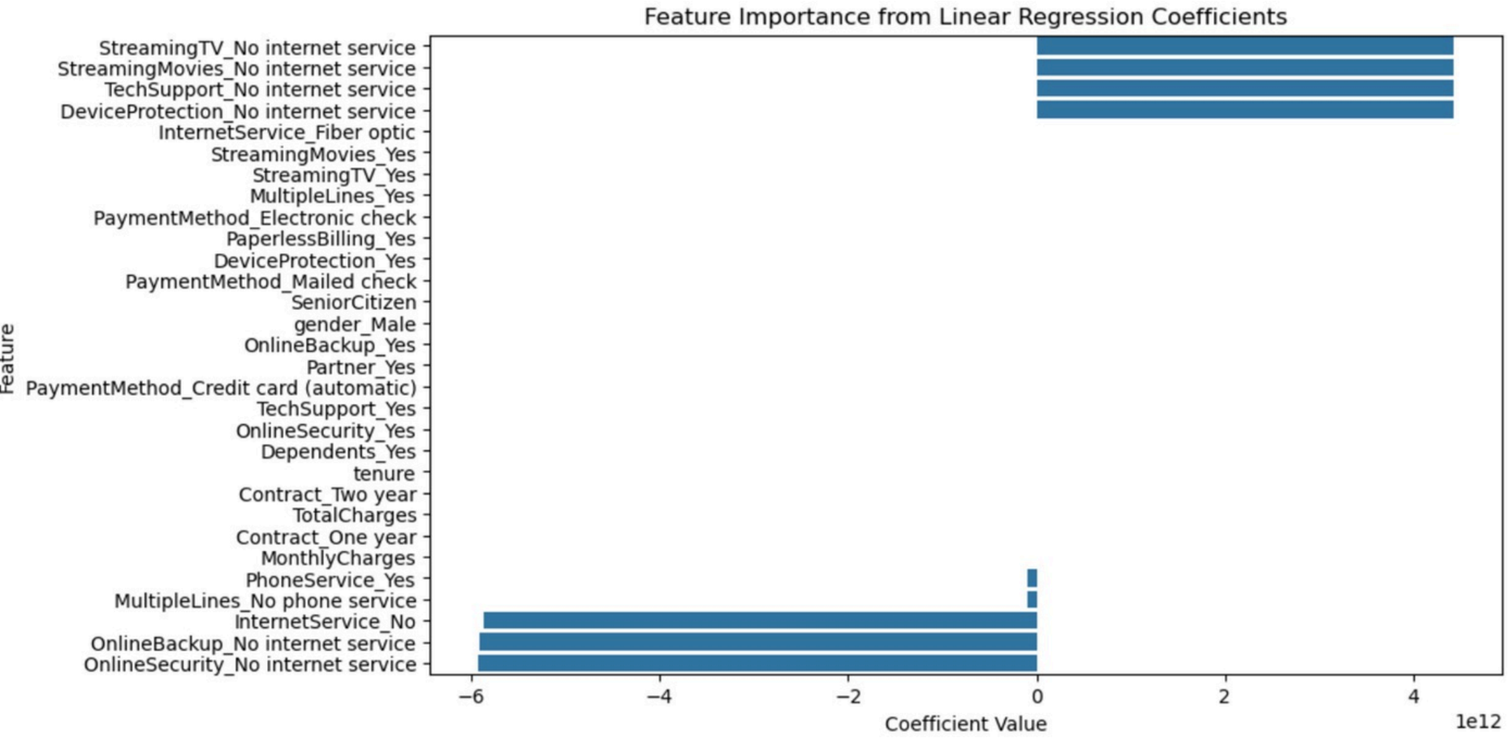
- Accuracy: Both models achieved an accuracy of 79%, indicating that it correctly classified 79% of the samples in the dataset.
- Precision: For the churn class, precision was 64% for linear regression & 74% for the random forest, meaning that out of all customers predicted to churn, 64% actually churned for linear regression & 74% for the random forest
- Recall: The recall for the churn class was 51% for linear regression & 59% for random forest, showing that the model identified that % of the actual churners.
- F1-Score: With a score of 57% for linear regression & 66% for random forest for the churn class, the F1-score balanced precision and recall, reflecting the models' overall effectiveness in managing the trade-off between false positives and false negatives.



LINEAR REGRESSION

Linear regression is a statistical method used for analyzing relationships between a dependent variable and one or more independent variables, assuming a linear relationship. It is particularly suited for continuous data where the dependent variable (response) is numeric, such as sales revenue, temperature, or weight. The independent variables (predictors) can be numeric or categorical (encoded appropriately). Linear regression is ideal for datasets with minimal multicollinearity, where predictors are not highly correlated.

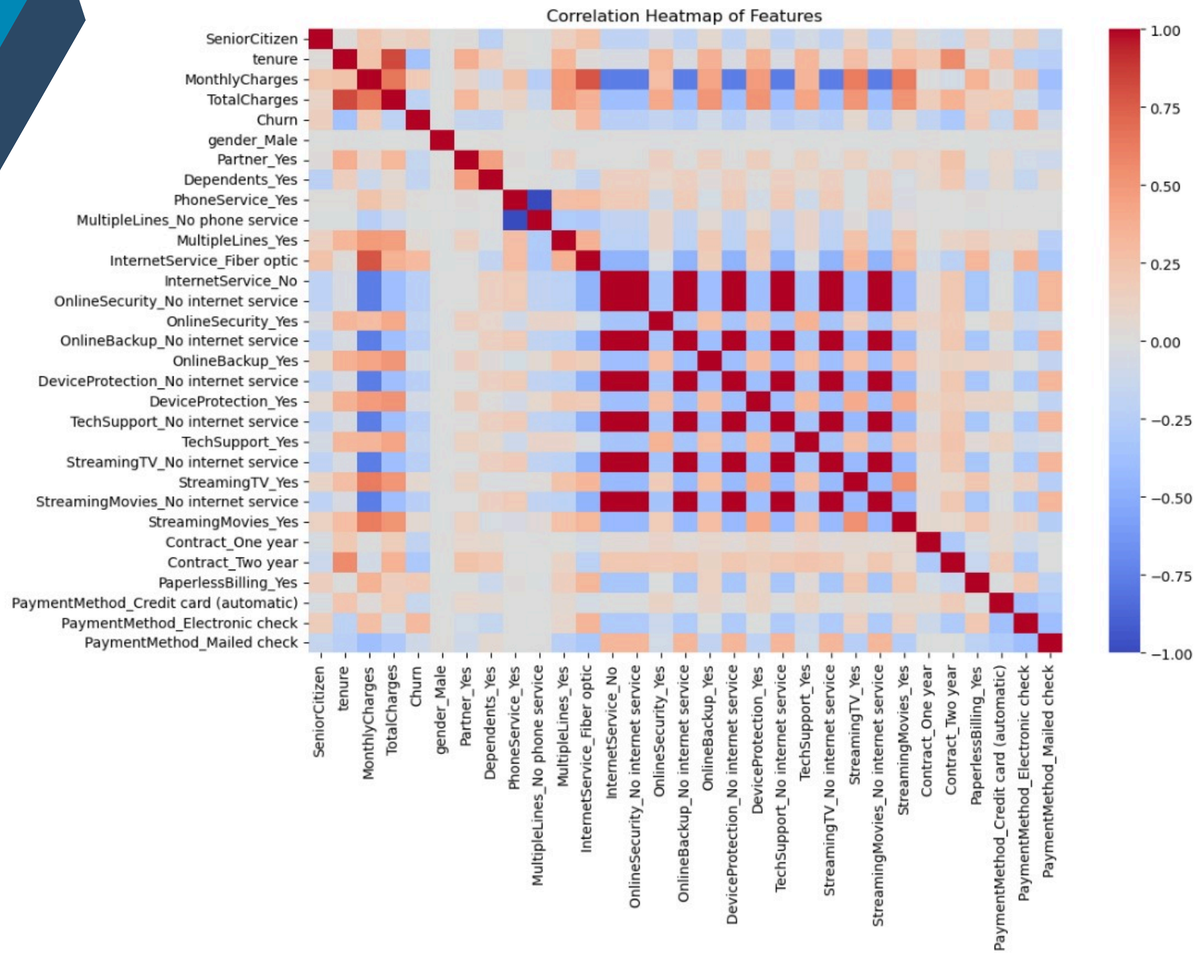
```
Mean Squared Error: 0.14162402244049385
R^2 Score: 0.2737580339695991
ROC AUC Score: 0.8362609570538618
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.90      0.87      1552
           1       0.64      0.51      0.57       561

    accuracy                           0.79      2113
   macro avg       0.74      0.70      0.72      2113
weighted avg       0.78      0.79      0.79      2113
```

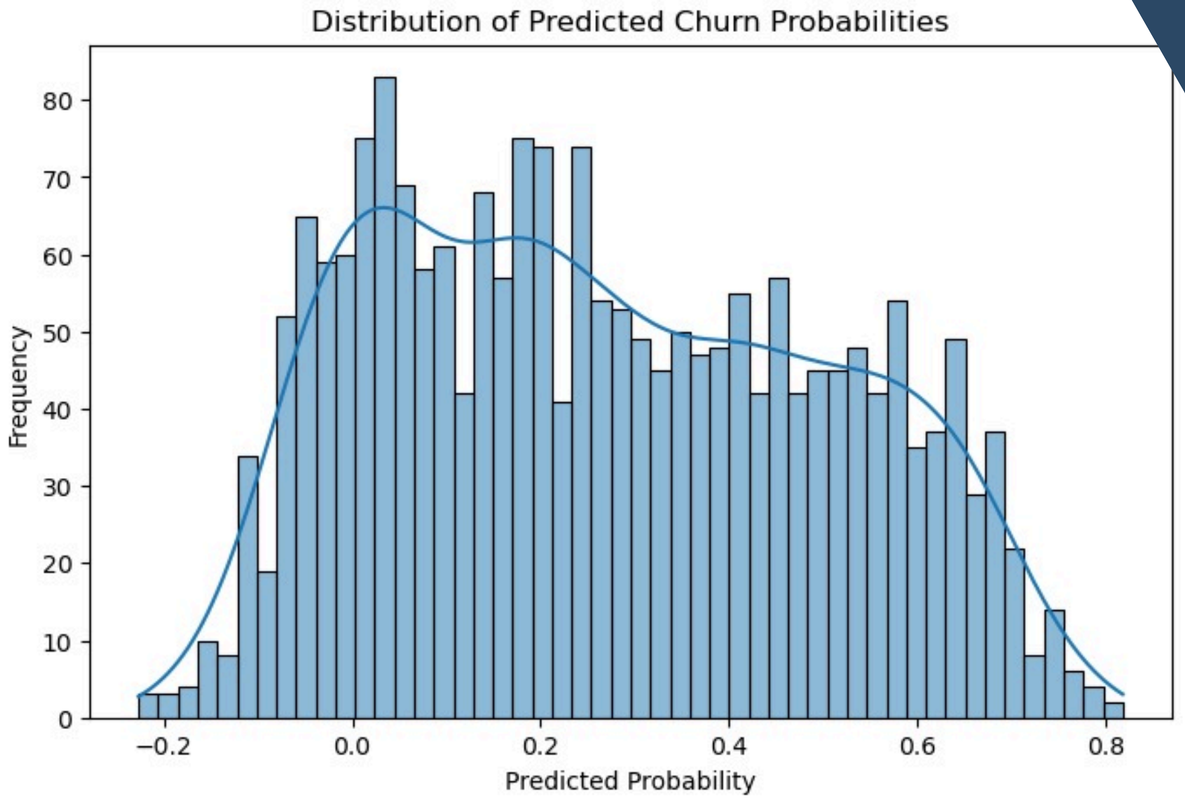


Feature Importance:

The feature importance derived from linear regression coefficients, highlights the variables that influence the target outcome (e.g., customer churn). Features such as "StreamingTV: No internet service" and "OnlineSecurity: No internet service" have the most significant negative coefficients, indicating their strong impact on the dependent variable. Similarly, contract type, monthly charges, and online service availability also show varying degrees of importance.

Linear Regression





The heatmap helps identify multicollinearity, where certain features may strongly correlate with each other (e.g., service-related variables). This might indicate redundancy and the need for dimensionality reduction or feature selection to improve model performance.

1. Most predicted probabilities cluster around 0, indicating the model leans toward low churn probabilities for many instances.
2. The distribution is slightly right-skewed, with some higher churn probabilities, suggesting the model does identify some cases with a high likelihood of churn.

CONFUSION MATRIX

Visual Representation:

Bar Graph or Heatmap:

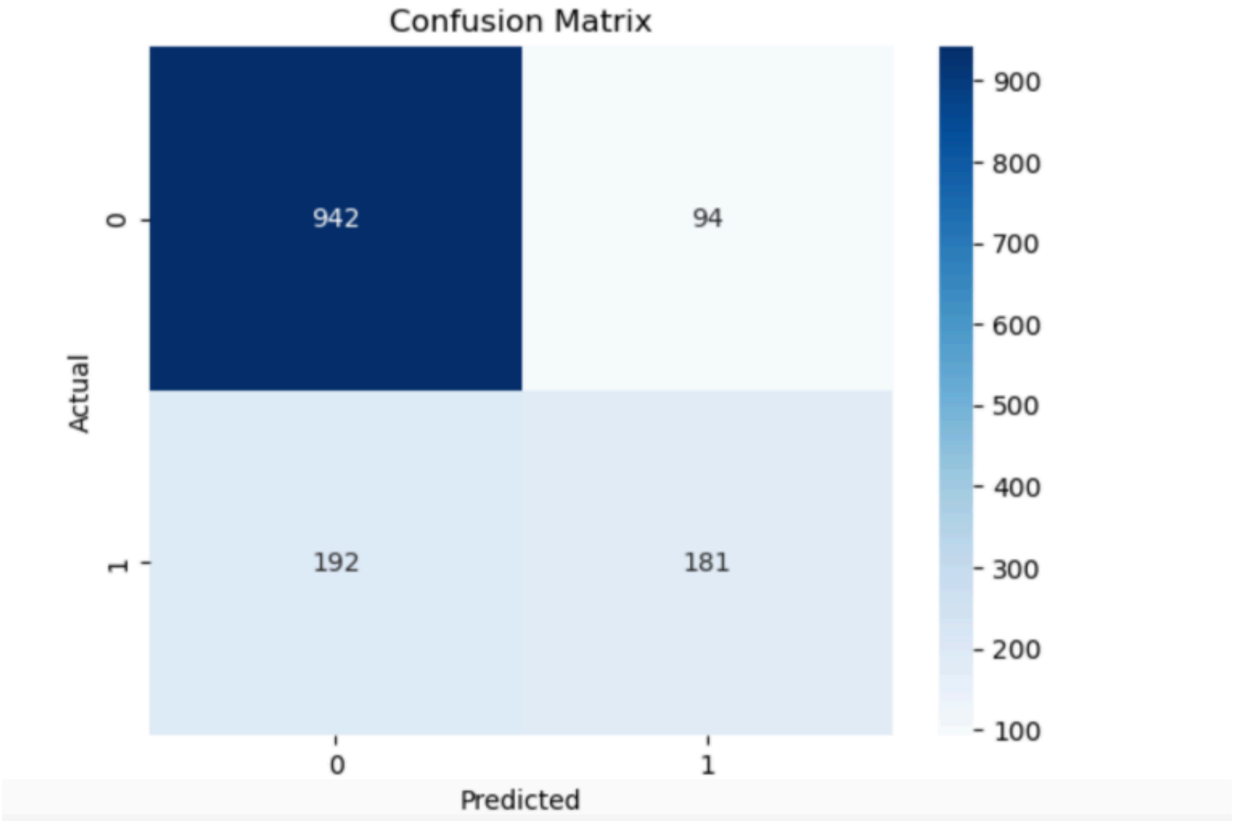
- Confusion matrix visualization with values (TP, TN, FP, FN).
- Bar Graph: Show precision, recall, F1-score comparison for the churn class.

Insights:

- Strengths: Correctly predicts most churners (TP = 942).
- Improvement Areas: Reduce false negatives (FN = 192) to better identify at-risk customers.
- True Positives (TP):942 customers were correctly identified as churners.
- True Negatives (TN):181 customers were accurately classified as not churners.
- False Positives (FP): 94 customers were incorrectly predicted to churn when they did not.
- False Negatives (FN): 192 actual churners were missed by the model, being classified as non-churners.

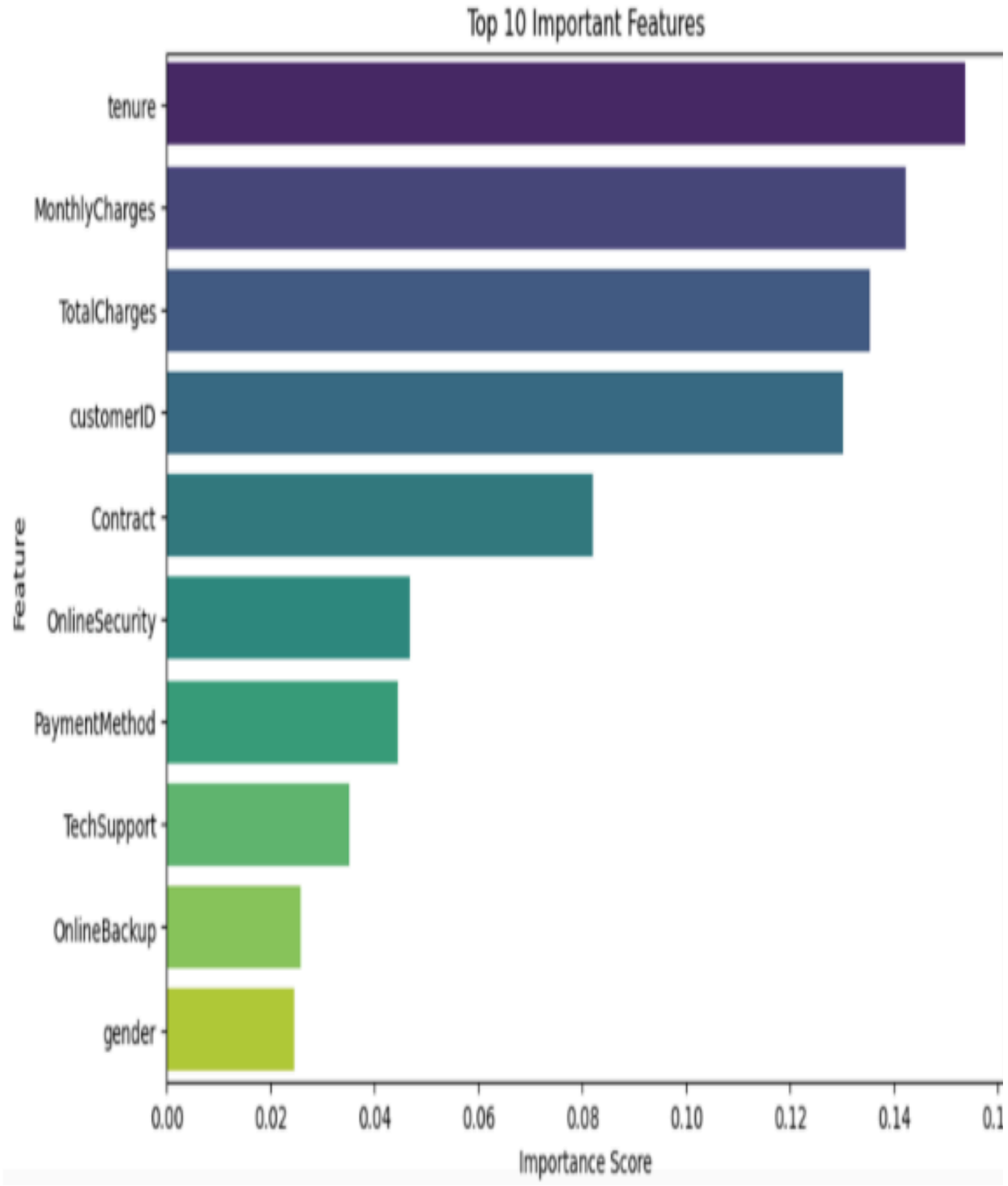
Accuracy: 0.7970191625266146
Classification Report:

	precision	recall	f1-score	support
0	0.83	0.91	0.87	1036
1	0.66	0.49	0.56	373
accuracy			0.80	1409
macro avg	0.74	0.70	0.71	1409
weighted avg	0.79	0.80	0.79	1409



Key Drivers of Customer Churn

1. **Contract Type (Most Influential):**
    a. Insight: Month-to-month contracts show higher churn.
    b. Action: Promote long-term contracts with incentives or discounts.
2. **Tenure:**
    a. Insight: Longer-tenured customers are less likely to churn.
    b. Action: Strengthen early-stage engagement strategies to build loyalty.
3. **Monthly Charges:**
    a. Insight: Higher charges lead to increased churn.
    b. Action: Offer personalized pricing plans or discounts to address cost concerns.
4. **Internet Service:**
    a. Insight: Fiber optic users show higher churn.
    b. Action: Investigate service quality or pricing issues to improve satisfaction.
5. **Payment Method:**
    a. Insight: Customers using electronic checks churn more.
    b. Action: Promote automated and secure payment methods like bank transfers or credit cards.



Recommendations

Based on the results, several actionable recommendations were proposed:

Offer Incentives for Long-Term Contracts:

Encourage month-to-month customers to switch to longer-term plans through attractive offers.

Address High Monthly Charges:

Provide personalized plans or discounts for customers with high bills to enhance value perception.

Investigate Fiber Optic Services:

Conduct surveys or feedback sessions to identify and resolve issues affecting fiber optic users.

Improve Payment Options:

Promote the use of credit cards or automatic bank transfers with rewards or discounts to reduce churn risk.

THANK YOU