Machine Learning Project

Telco Customer Churn

Malak Magdy Mohammed 2300512

Goal: Predict whether a customer will churn (i.e., leave the company).

Type: Supervised binary classification.

A Quick Look at our Data Frame

The dataset used in this project is based on customer data from a telecommunications company. The goal was to analyze this data and predict whether a customer is likely to churn (i.e., leave the service).

Some key features in the dataset include:

gender, SeniorCitizen, Partner, Dependents: Demographic attributes
tenure, MonthlyCharges, TotalCharges: Numeric indicators of service length and billing
InternetService, OnlineSecurity, TechSupport, etc.: Service-specific usage details
Churn: The target variable indicating whether the customer left or stayed

METHODOLOGY

This project followed a structured machine learning workflow to predict customer churn using a telco dataset. The major steps were as follows:

Data Cleaning:

We began by preprocessing the dataset to handle missing or inconsistent values. In particular, we ensured the TotalCharges column was converted to numeric, removed rows with missing values, standardized the format for further analysis , and Categorical variables were encoded using one-hot encoding (dummy variables).

before

n	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	Monthly
o	No	No	No	Month-to-month	Yes	Electronic check	
es	No	No	No	One year	No	Mailed check	
o	No	No	No	Month-to-month	Yes	Mailed check	
es	Yes	No	No	One year	No	Bank transfer (automatic)	
o	No	No	No	Month-to-	Yes	Electronic check	

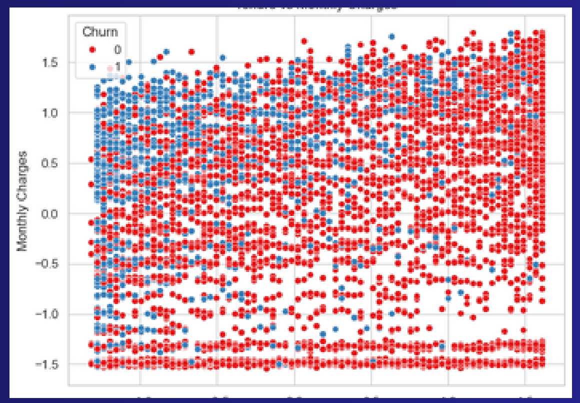
after

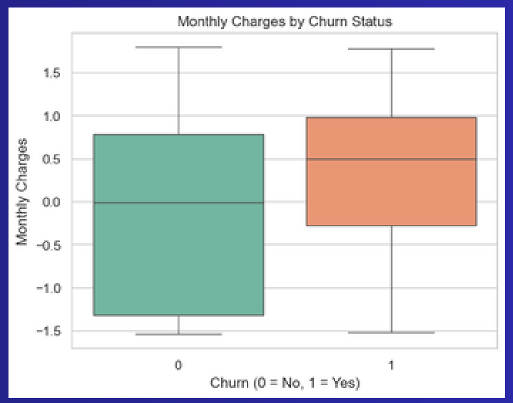
es	Churn	MultipleLines_No phone service	MultipleLines_Yes	InternetService_Fiber optic	InternetService_No	OnlineSecu internet
71	0	1	0	0	0	
76	0	0	0	0	0	
99	1	0	0	0	0	
00	0	1	0	0	0	
93	1	0	0	1	0	
74	1	0	1	1	0	
28	0	0	1	1	0	
52	0	1	0	0	0	

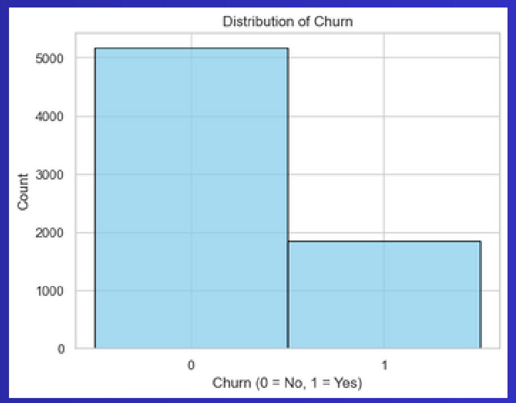
Exploratory Data Analysis (EDA):

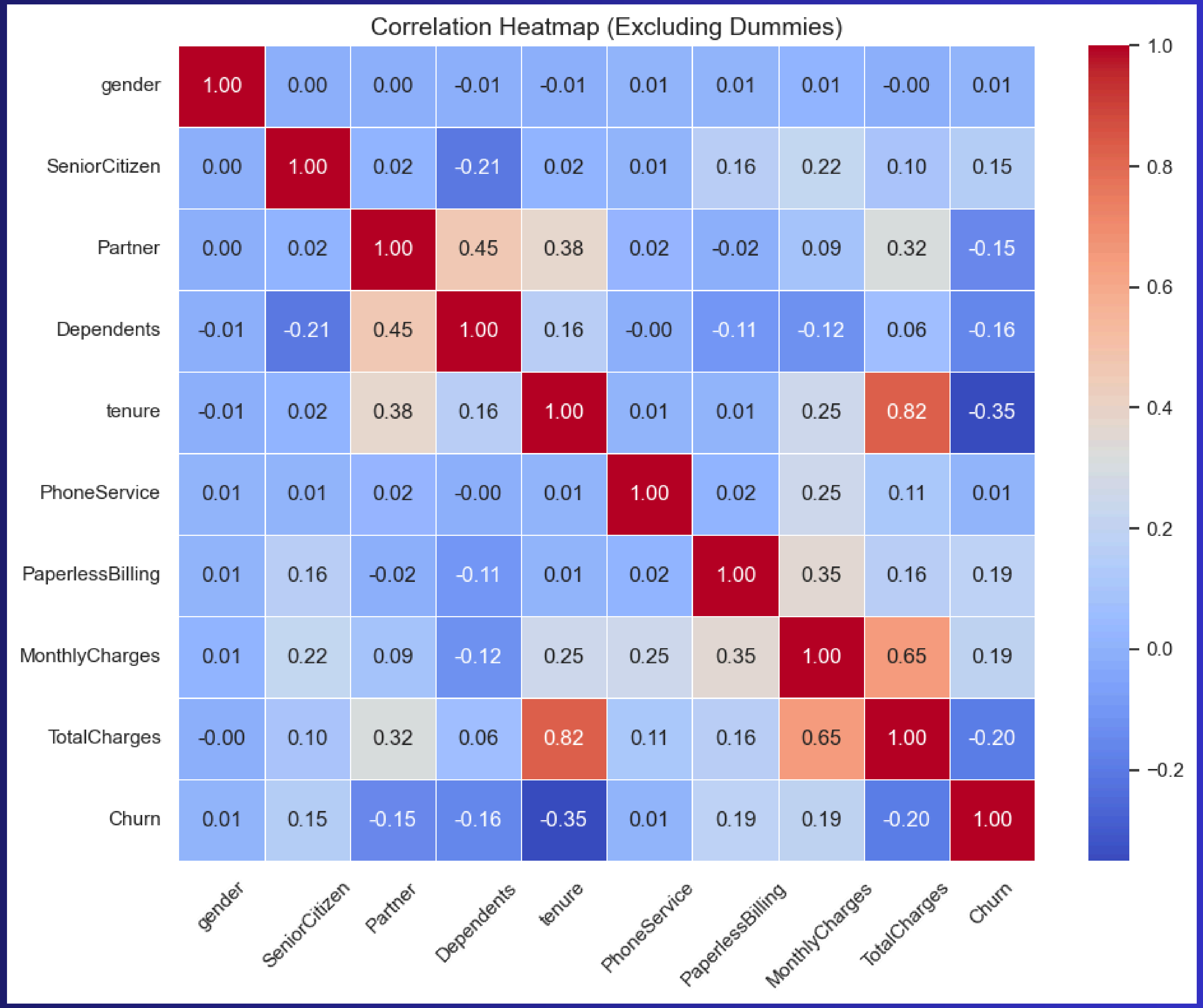
Visual insights were used to understand patterns and relationships in the data:

- A Churn Bar Plot The dataset is slightly imbalanced, with more customers not churning (0) than churning (1).
- A **box plot** of MonthlyCharges vs Churn revealed that pricing affect churn.
- A tenure with monthelyCharges scatterplot: High charges + low tenure might relate to higher churn
- A **correlation heatmap** (excluding dummy variables):TotalCharges and tenure likely show a strong positive correlation ,means customers who stay longer naturally accumulate more charges and MonthlyCharges and tenure might have low or even negative correlate, means that higher monthly = shorter stays.

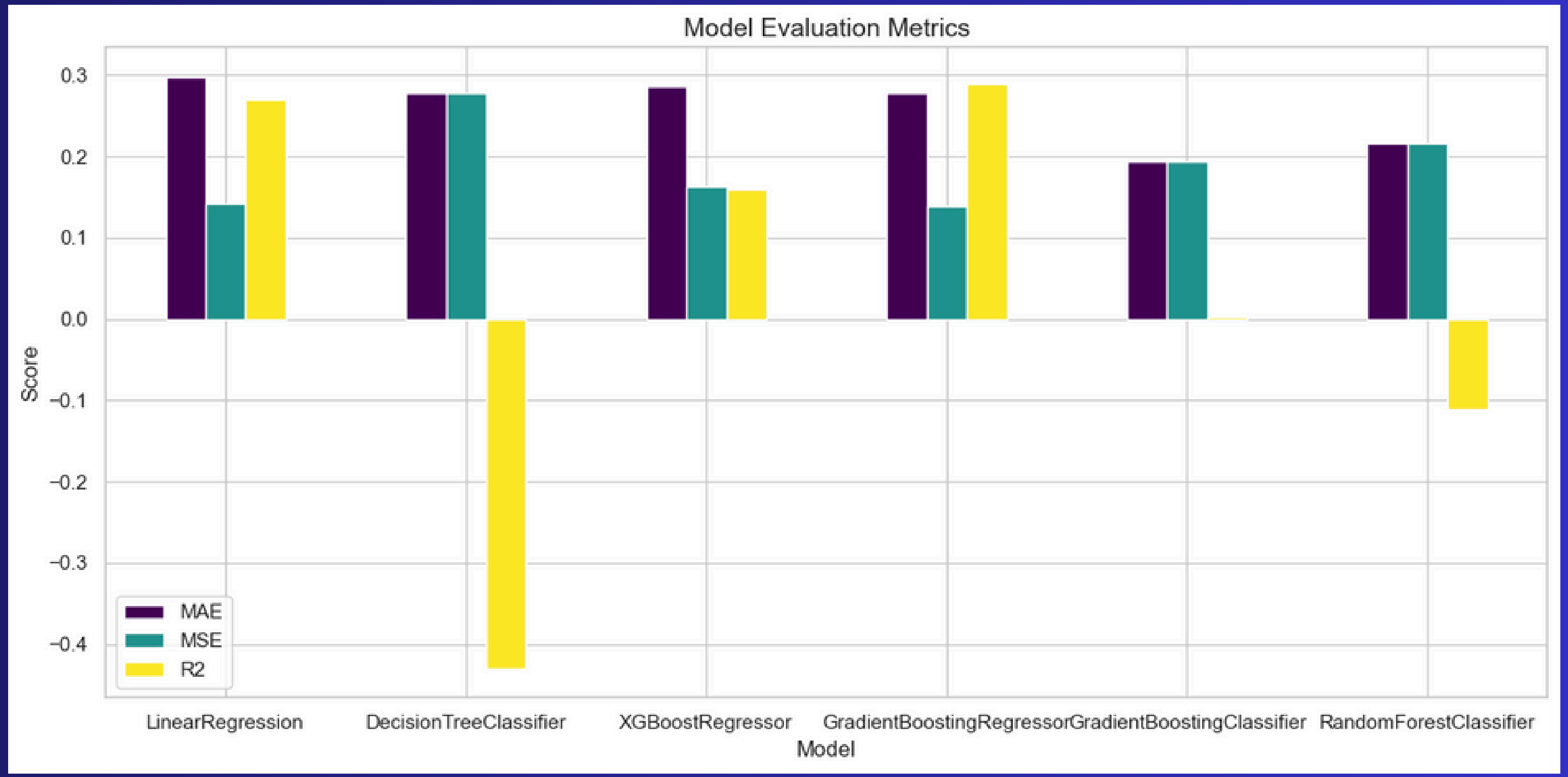








Preparation and Model Selection:



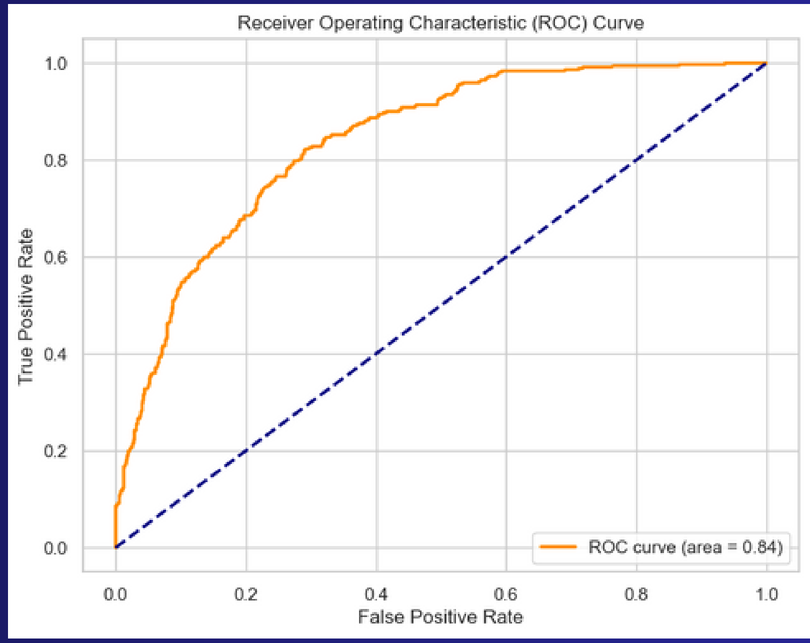
the dataset was split into training and testing sets then six models were tried:

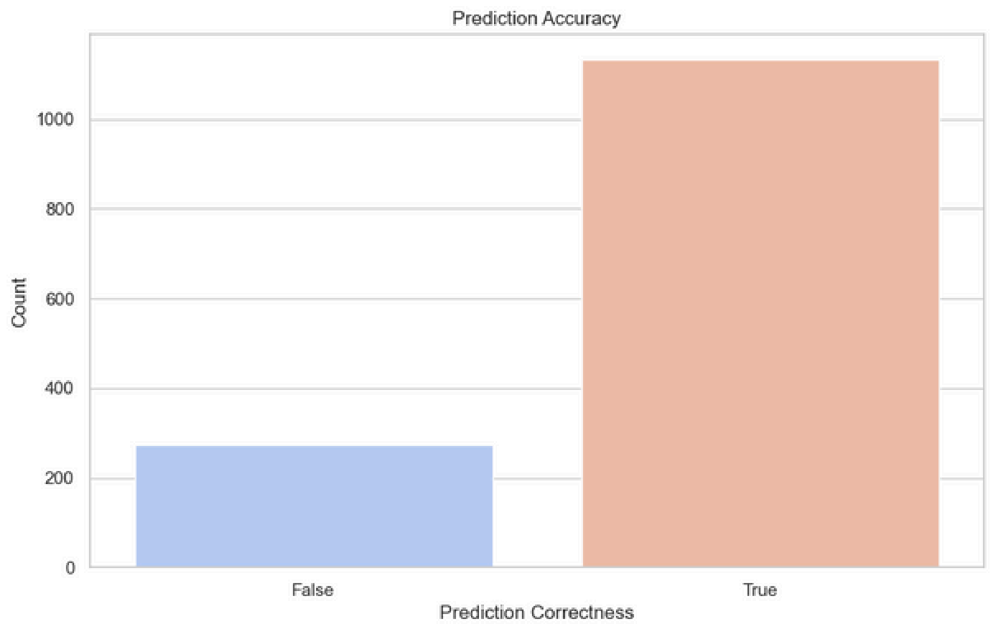
- LinearRegression
- DecisionTreeClassifier
- XGBoostRegressor
- RandomForestClassifier
- GradientBoostingRegressor: Gave good performance numerically but wasn't suited for classification metrics.
- GradientBoostingClassifier: Was ultimately selected as it supported probability predictions and allowed for classification evaluation using thresholding and ROC curves.

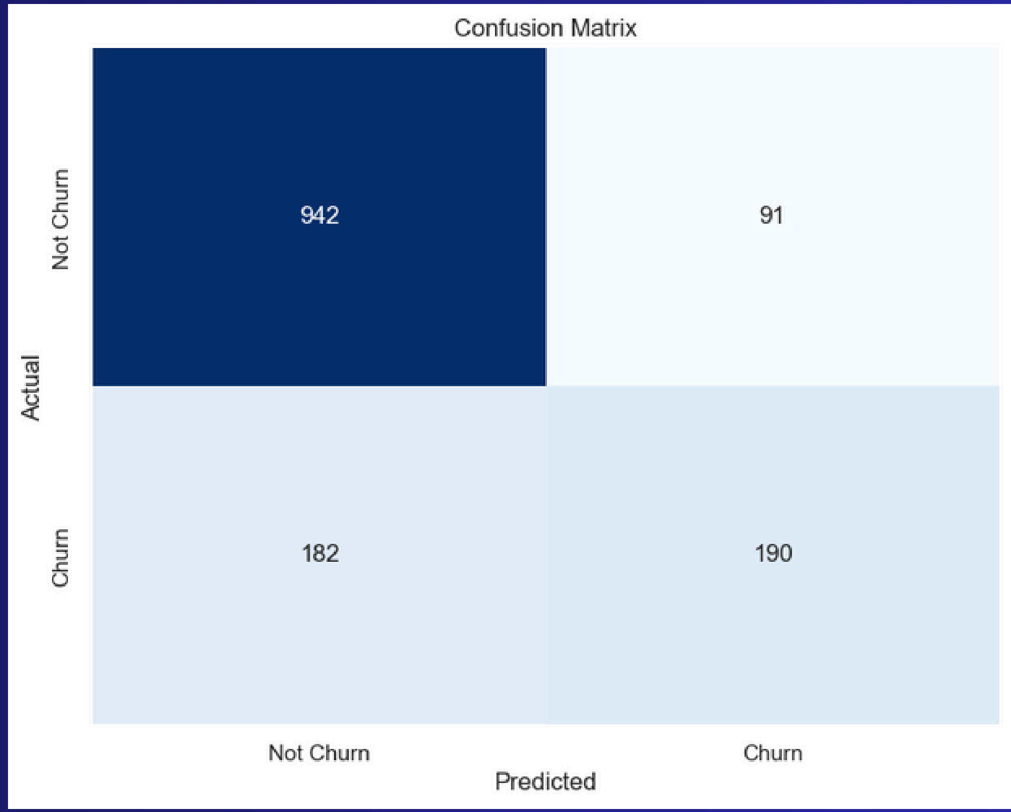
Model Evaluation:

The classifier was evaluated using:

- Accuracy score, confusion matrix, and classification report (precision, recall, F1).
- A ROC curve was plotted to assess the model's ability to separate classes.
- A threshold of 0.5 was used to convert probabilities into binary predictions.
- The model achieved a reasonable balance between false positives and false negatives.
- Prediction Visualization: A side-by-side DataFrame of actual vs predicted values was displayed to validate model outputs, along with a scatter plot styled for better interpretability.








```
Accuracy: 0.81
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.91      0.87      1033
           1       0.68      0.51      0.58       372

    accuracy                           0.81      1405
   macro avg       0.76      0.71      0.73      1405
weighted avg       0.80      0.81      0.80      1405
```

	Actual	Predicted	Correct
5627	0	1	False
6126	1	0	False
2361	1	1	True
2201	0	0	True
832	0	0	True
683	0	0	True
3403	0	0	True
6240	1	1	True
3996	0	0	True
4285	0	0	True
2286	0	0	True
2145	0	0	True
5317	1	0	False
4767	0	0	True
5469	0	0	True
3289	0	0	True
5915	0	0	True
6898	0	0	True
1785	0	0	True
4172	0	0	True
4711	0	0	True
6323	0	0	True
53	1	1	True
4940	0	1	False
5940	0	0	True
2482	0	1	False
2975	0	0	True
3730	1	0	False
3050	1	1	True
3371	1	0	False

Findings / Results:

After preprocessing, analyzing, and modeling the data, we arrived at several important insights and outcomes:

1. **Key Findings from Exploratory Data Analysis:**

- Churned customers often had higher MonthlyCharges and lower TotalCharges, indicating shorter tenure but higher-cost services.
- Features like Contract Type, Internet Service, and Tech Support showed strong correlations with churn behavior.
- The boxplot of TotalCharges vs Churn highlighted that customers who did not churn tended to have higher total charges, suggesting longer and more stable customer relationships.
- Heatmap Observations: The heatmap of numerical features (excluding dummy variables) showed that:
  1. TotalCharges and tenure are strongly positively correlated, as expected.
  2. MonthlyCharges had a weaker correlation with churn-related features, indicating it alone isn’t a strong predictor.

2.**Modeling Results:**

We tested both regression and classification models but focused on GradientBoostingClassifier for churn prediction.

- Accuracy: ~78.8%
- Confusion Matrix:
  - True Negatives: 934
  - False Positives: 99
  - False Negatives: 199
  - True Positives: 173
- Classification Report Highlights:
  - Precision (Class 1 - Churn): 0.64
  - Recall (Class 1 - Churn): 0.47
  - F1-Score (Class 1 - Churn): 0.54
- ROC AUC Score: AUC curve showed the model performs better than random guessing, indicating it has learned meaningful patterns.

Conclusions

What Did We Learn from the Data?

- Customer churn is significantly influenced by contract type, tenure, and monthly charges.
- Short-term contracts (especially month-to-month) are strongly associated with higher churn rates.
- Customers with lower total charges often churn, suggesting early-stage dissatisfaction.
- Services like Tech Support, Online Security, and Internet Type play a crucial role in customer retention.

Recommendations:

- Encourage longer-term contracts through incentives or discounts to reduce churn.
- Provide targeted support and offers to high-risk customers based on their tenure and service usage.
- Improve services related to tech support and online security, as their absence correlates with higher churn.
- Monitor customers with high MonthlyCharges and short tenure, as they are more likely to leave early.

What Can Be Improved in the Future?

- Enhance the model's recall by addressing class imbalance or exploring ensemble techniques like XGBoost or LightGBM.
- Incorporate customer feedback or sentiment data to better understand dissatisfaction reasons.
- Develop a real-time churn prediction system integrated with a CRM to act on predictions instantly.
- Perform A/B testing with retention strategies and track the impact on churn rates.

Thank you