Customer Churn Prediction Project

Predicting churn and enabling proactive retention strategies

Overview

- This project predicts customer churn in a telecommunications dataset using machine learning models. The primary objective is to enable proactive retention strategies by accurately identifying customers likely to churn and uncovering key drivers influencing churn.

Business Understanding

- In subscription-based businesses, customer retention is critical to profitability. This project:
- - Identifies reasons behind churn.
- - Predicts churn to enable targeted retention strategies.
- - Builds long-term customer relationships through data-driven insights.

Data Understanding

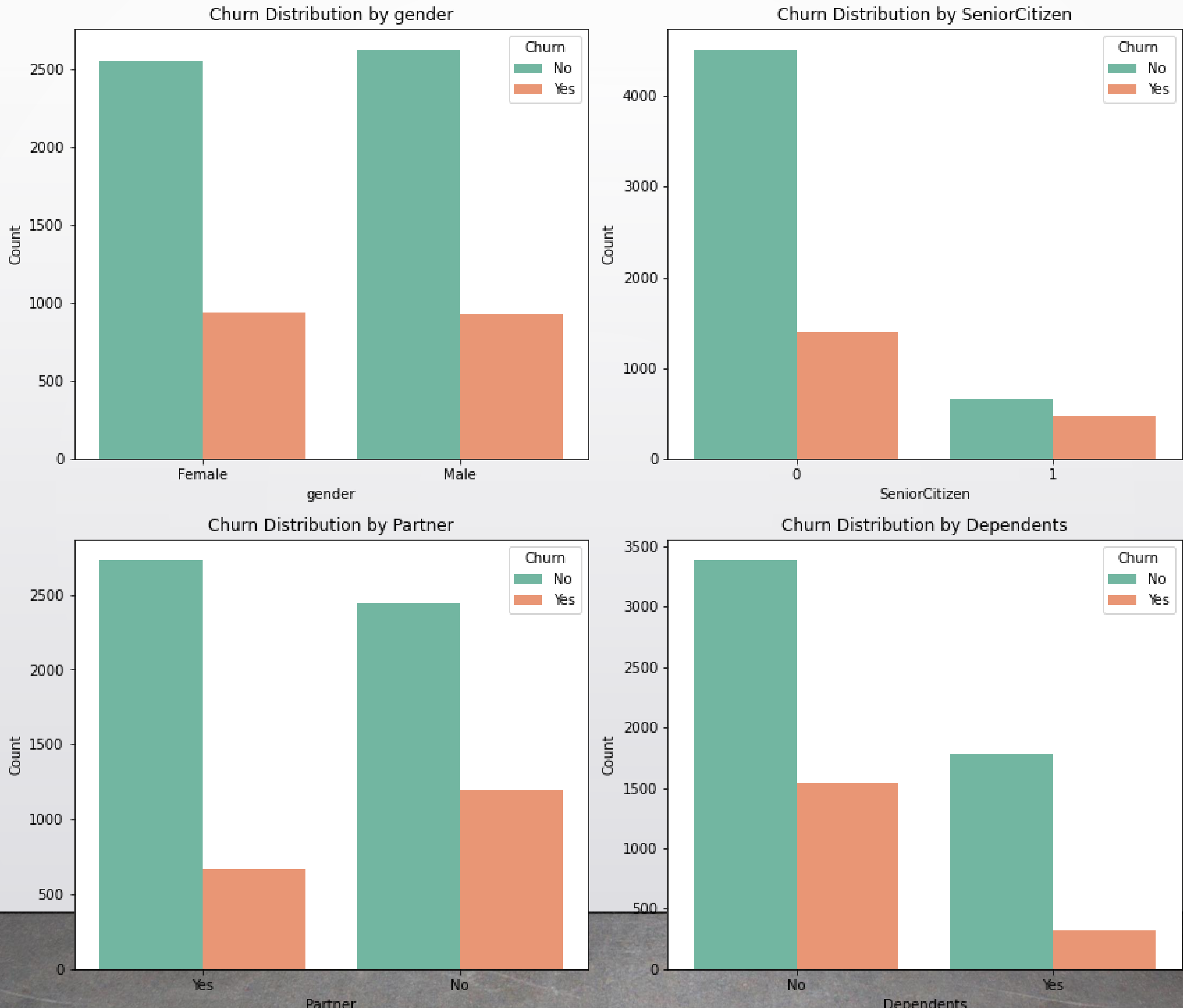
- • Dataset: Sourced from Kaggle, containing telecom customer data.
- • Features: Tenure, MonthlyCharges, TotalCharges, Contract Type, Payment Method, etc.
- • Target Variable: Churn (Yes/No).
- • Relevance: Provides a comprehensive view of customer behavior and service usage.

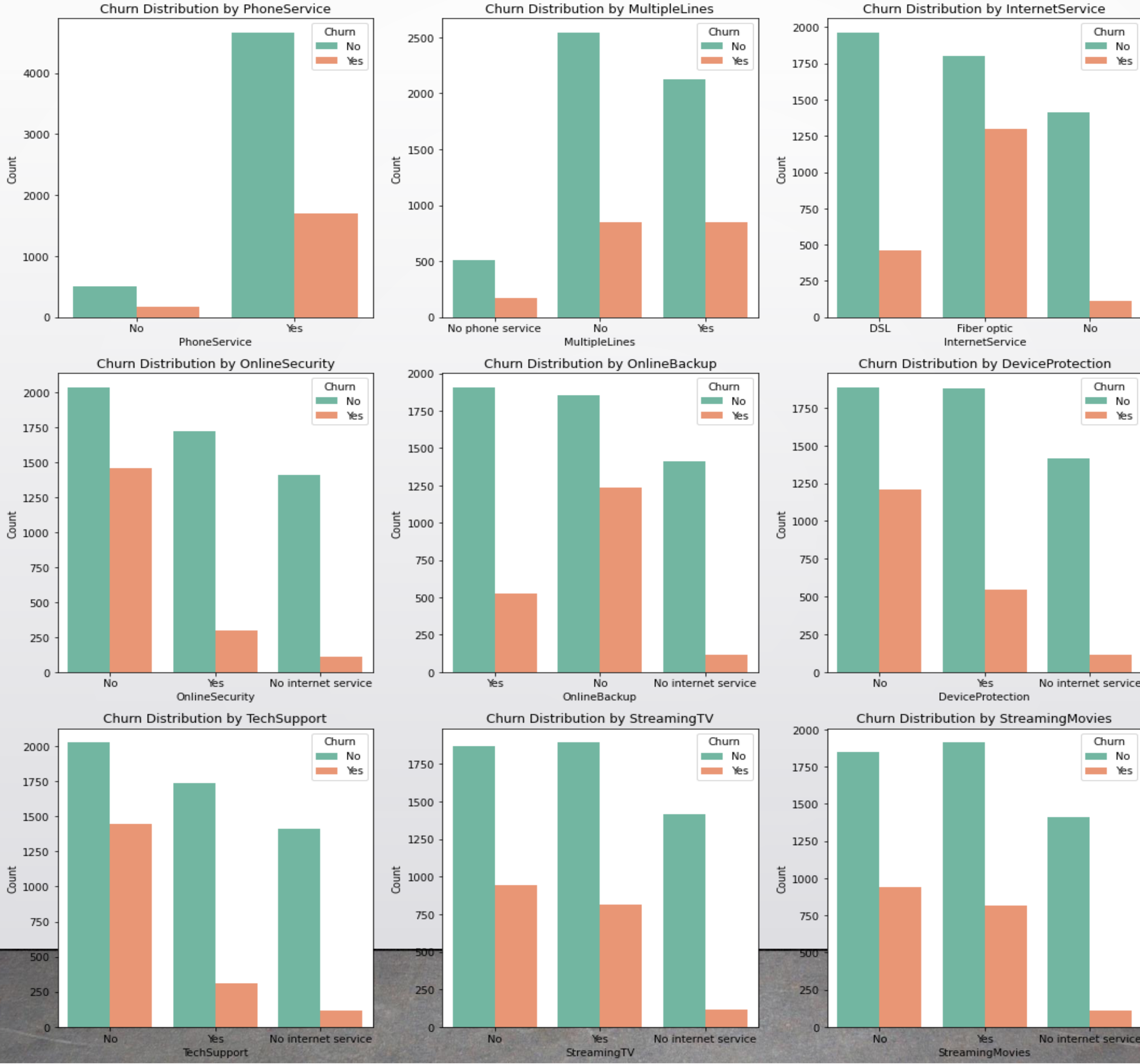
Data Preparation

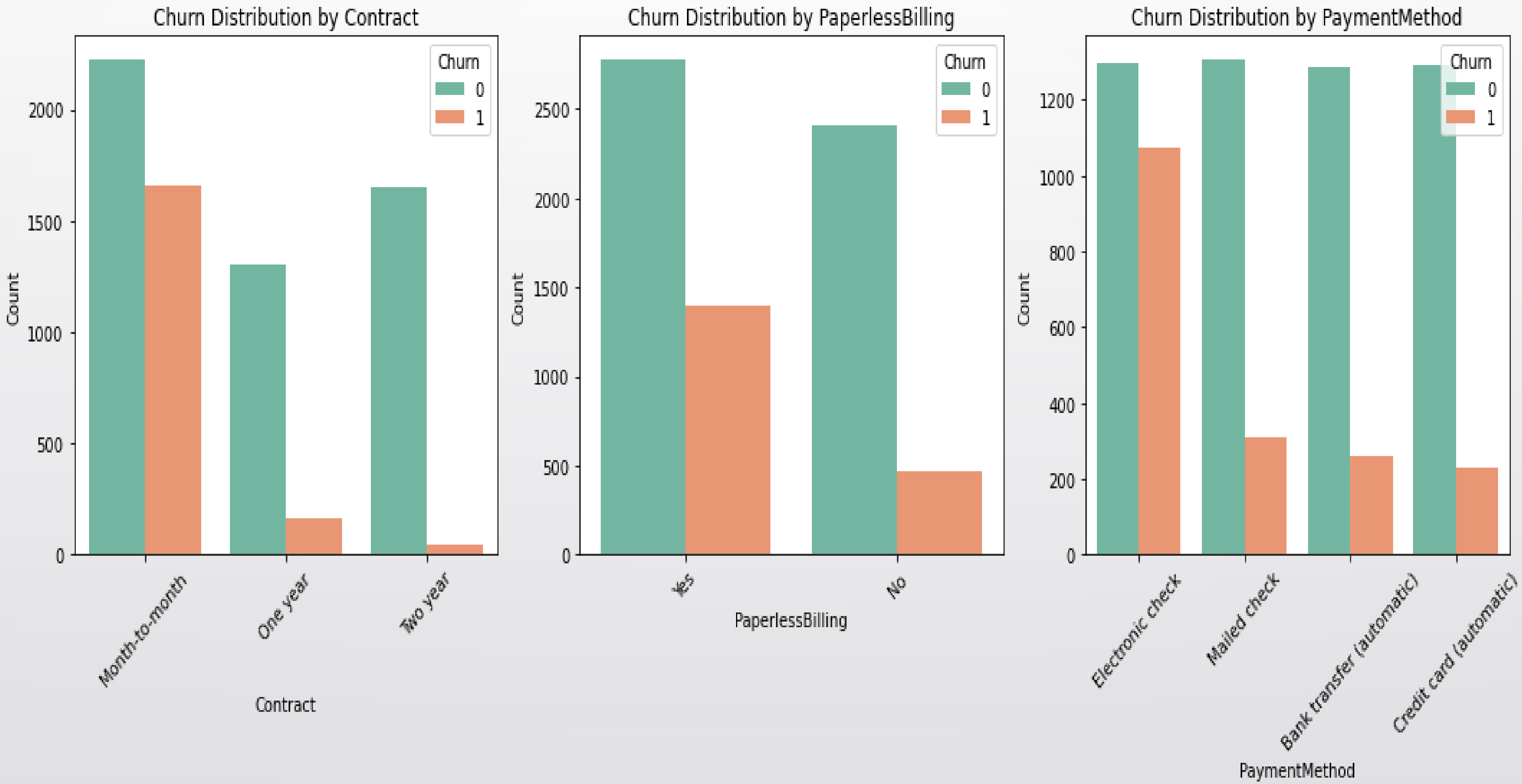
- • Handled missing values.
- • Converted TotalCharges from object to numeric type.
- • Scaled numerical variables for consistency.
- • One-hot encoded categorical variables.
- • Split the data into train and test sets.

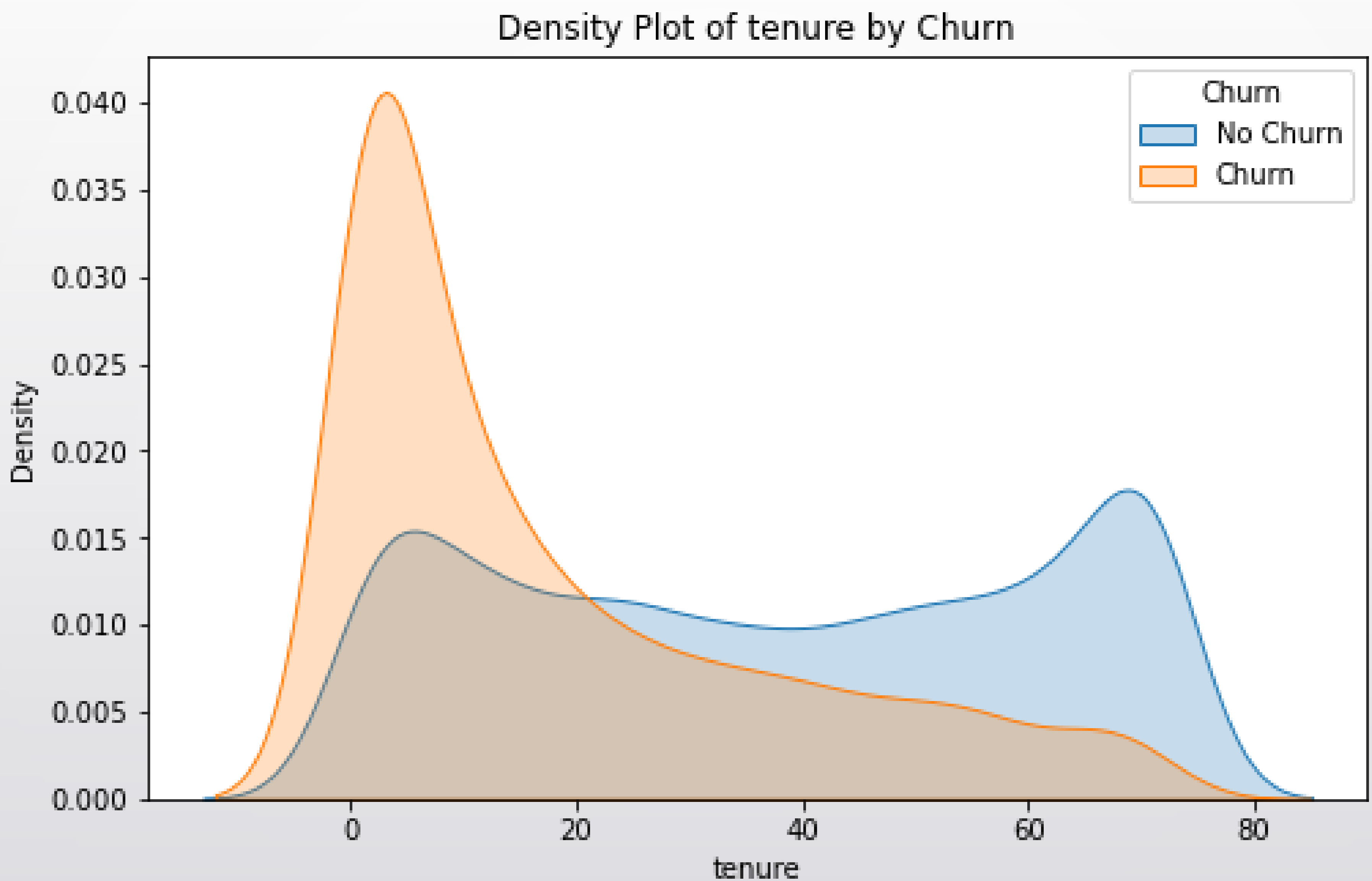
Exploratory Data Analysis (EDA)

- 1. Demographics: Certain groups show higher churn rates.
- 2. Services: Churn is higher among customers with fewer additional services.
- 3. Monthly Charges: Higher charges correlate with increased churn.
- 4. Tenure: Lower tenure customers are more likely to churn.

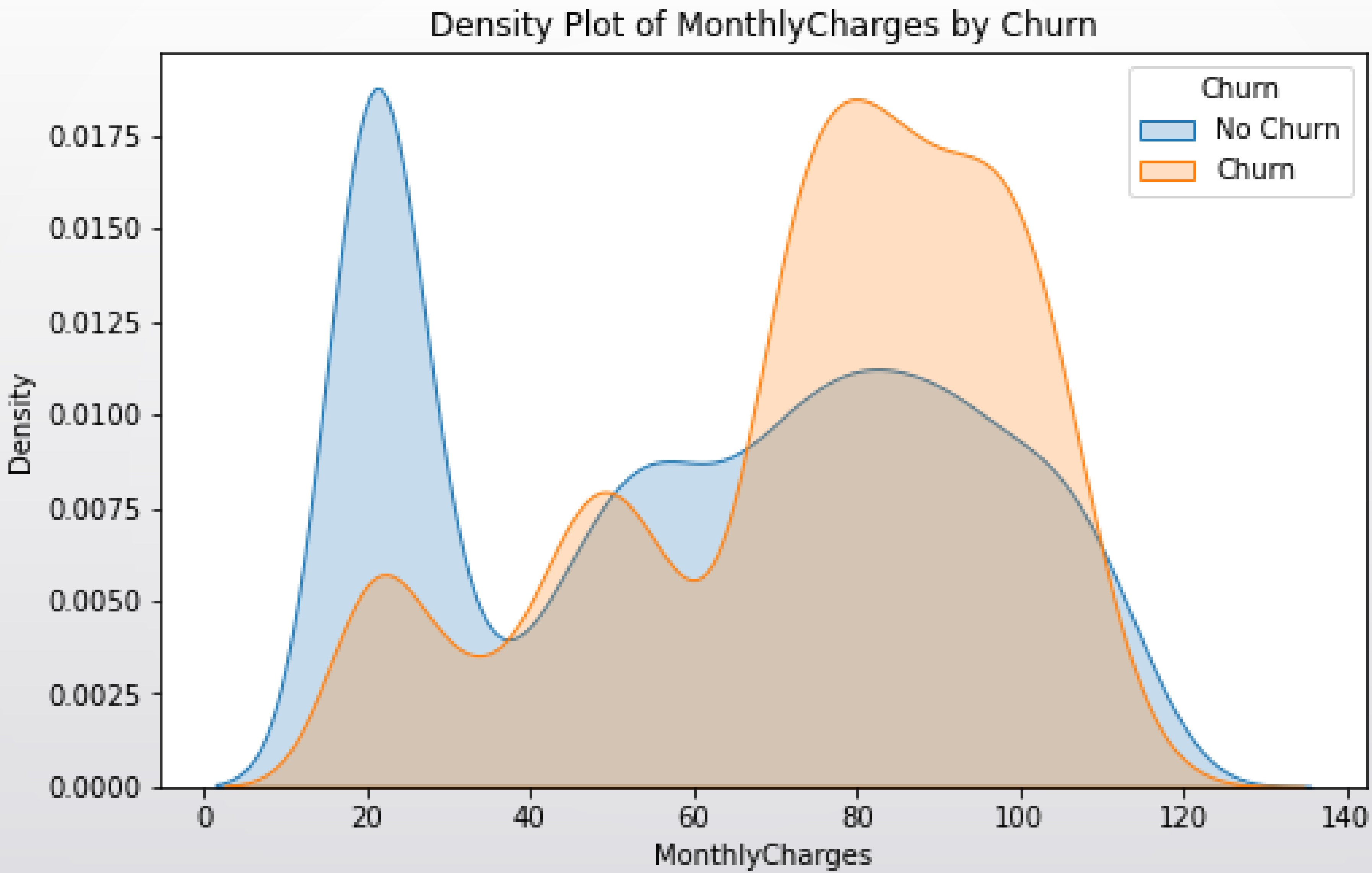

Churn Distribution by gender
Churn
No
Yes
Count
0
500
1000
1500
2000
2500
Female
Male
gender
Churn Distribution by SeniorCitizen
Churn
No
Yes
Count
0
1000
2000
3000
4000
0
1
SeniorCitizen
Churn Distribution by Partner
Churn
No
Yes
Count
0
500
1000
1500
2000
2500
Yes
No
Partner
Churn Distribution by Dependents
Churn
No
Yes
Count
0
500
1000
1500
2000
2500
3000
3500
No
Yes
Dependents

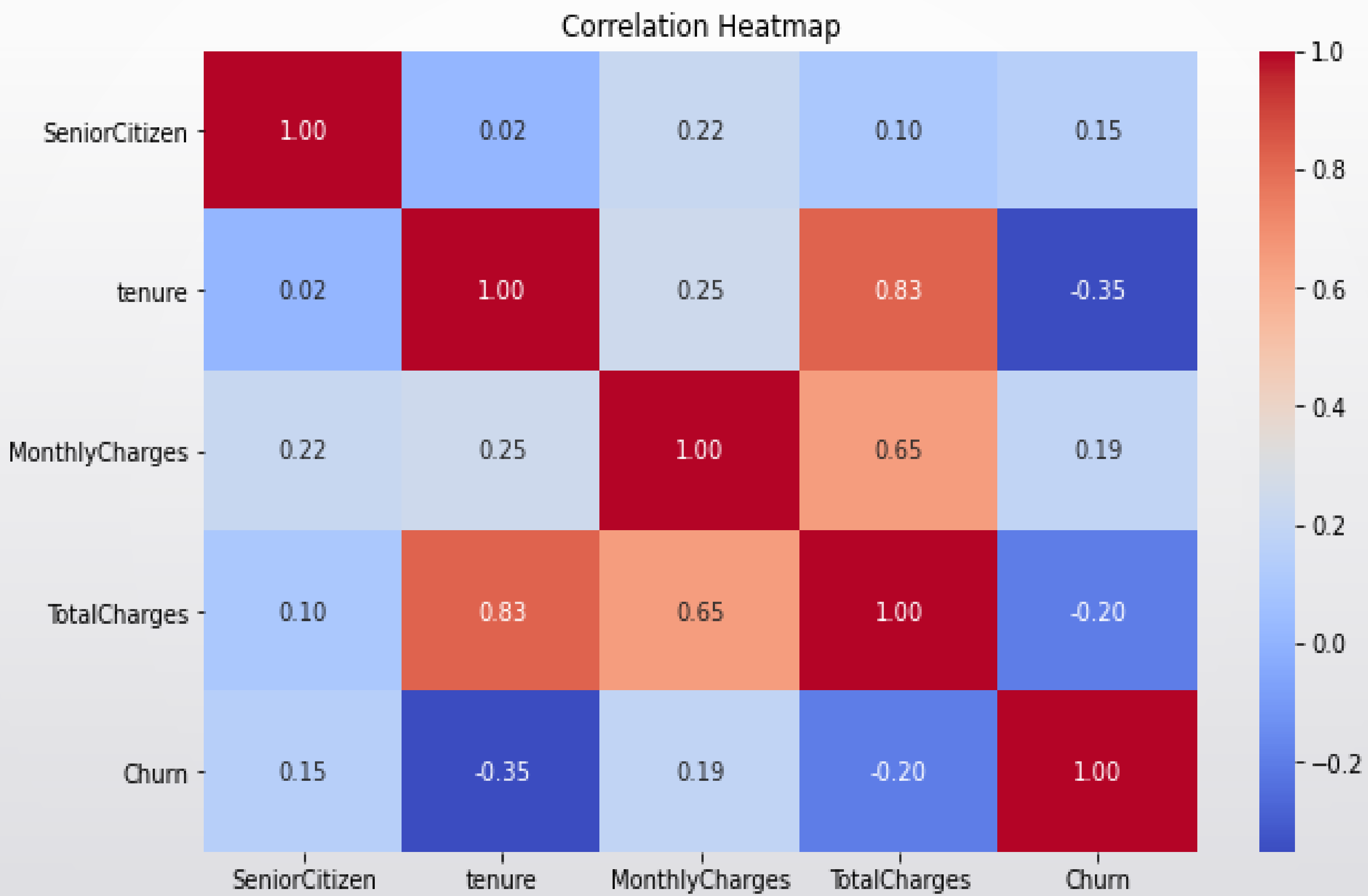

Churn Distribution by PhoneService
Churn
No
Yes
Count
0
1000
2000
3000
4000
No
Yes
PhoneService
Churn Distribution by MultipleLines
0
500
1000
1500
2000
2500
No phone service
No
Yes
MultipleLines
Churn Distribution by InternetService
0
250
500
750
1000
1250
1500
1750
2000
DSL
Fiber optic
No
InternetService
Churn Distribution by OnlineSecurity
No
Yes
No internet service
OnlineSecurity
Churn Distribution by OnlineBackup
Yes
No
No internet service
OnlineBackup
Churn Distribution by DeviceProtection
No
Yes
No internet service
DeviceProtection
Churn Distribution by TechSupport
No
Yes
No internet service
TechSupport
Churn Distribution by StreamingTV
No
Yes
No internet service
StreamingTV
Churn Distribution by StreamingMovies
No
Yes
No internet service
StreamingMovies


Churn Distribution by Contract
Churn
0
1
Count
0
500
1000
1500
2000
Month-to-month
One year
Two year
Contract
Churn Distribution by PaperlessBilling
Churn
0
1
Count
0
500
1000
1500
2000
2500
Yes
No
PaperlessBilling
Churn Distribution by PaymentMethod
Churn
0
1
Count
0
200
400
600
800
1000
1200
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
PaymentMethod



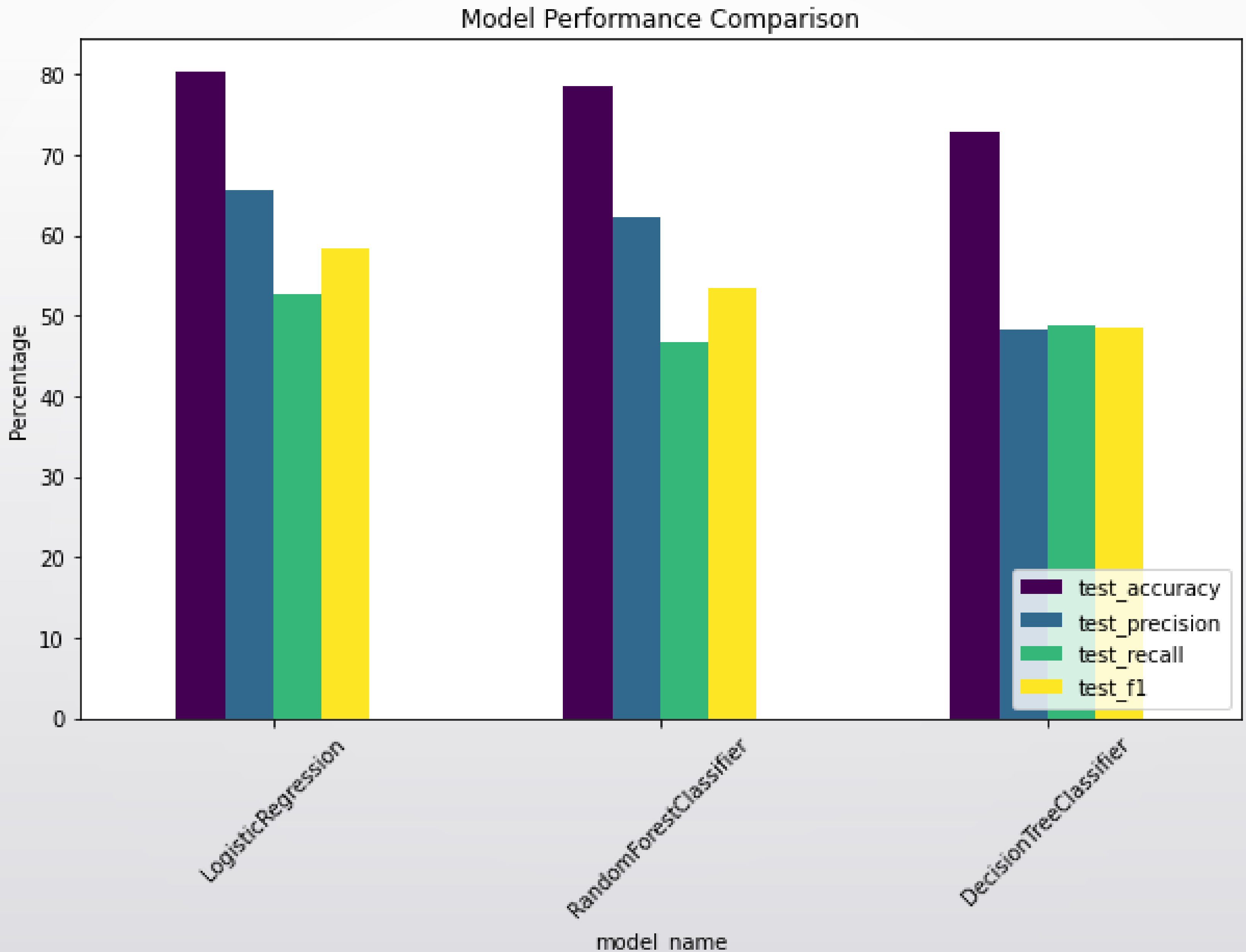
Density Plot of tenure by Churn
Churn
No Churn
Churn
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
0
20
40
60
80
tenure


Density Plot of MonthlyCharges by Churn
Churn
No Churn
Churn
Density
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0
20
40
60
80
100
120
140
MonthlyCharges


Correlation Heatmap
SeniorCitizen
tenure
MonthlyCharges
TotalCharges
Churn
1.00
0.02
0.22
0.10
0.15
0.02
1.00
0.25
0.83
-0.35
0.22
0.25
1.00
0.65
0.19
0.10
0.83
0.65
1.00
-0.20
0.15
-0.35
0.19
-0.20
1.00
SeniorCitizen
tenure
MonthlyCharges
TotalCharges
Churn
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

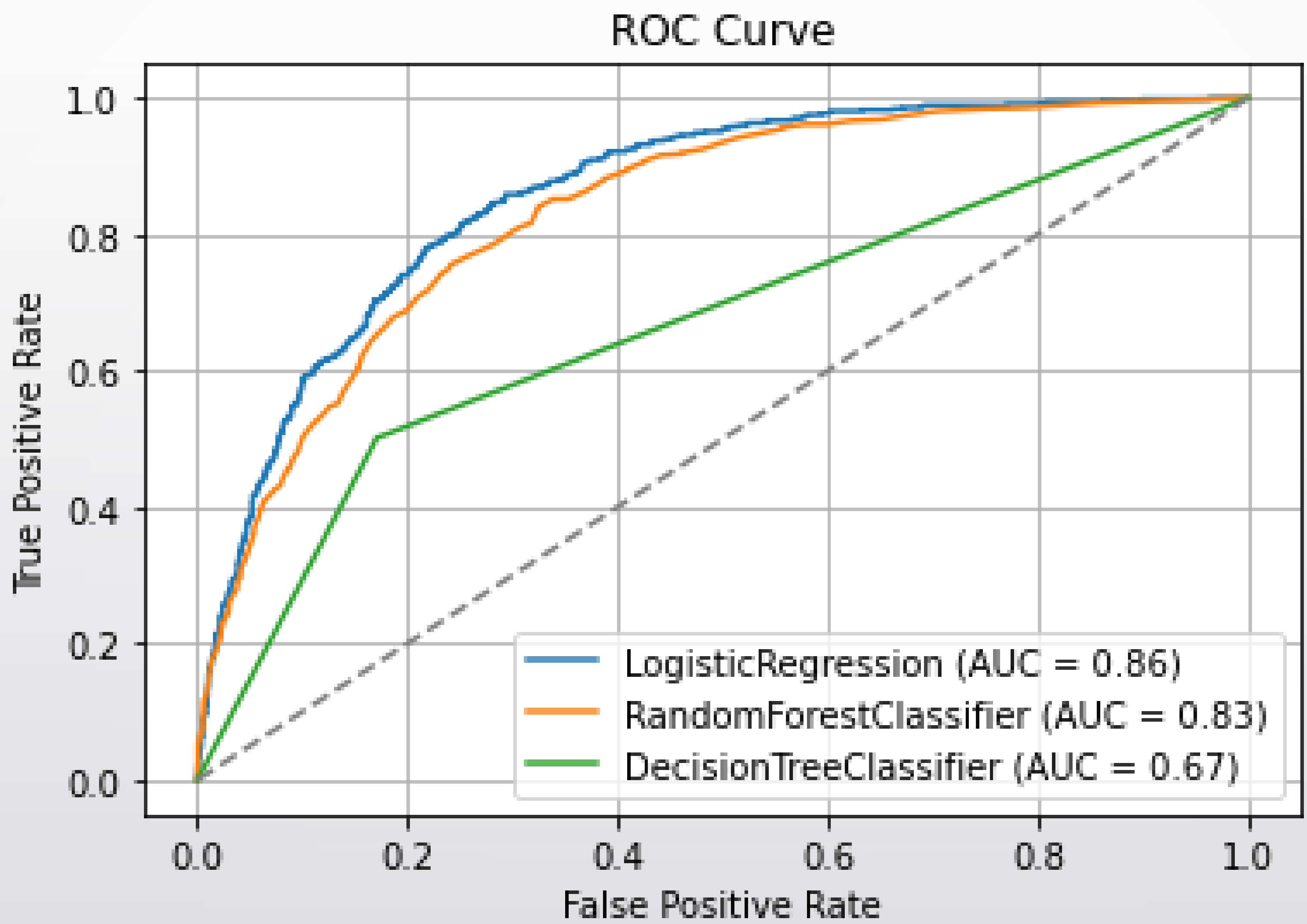
Modeling

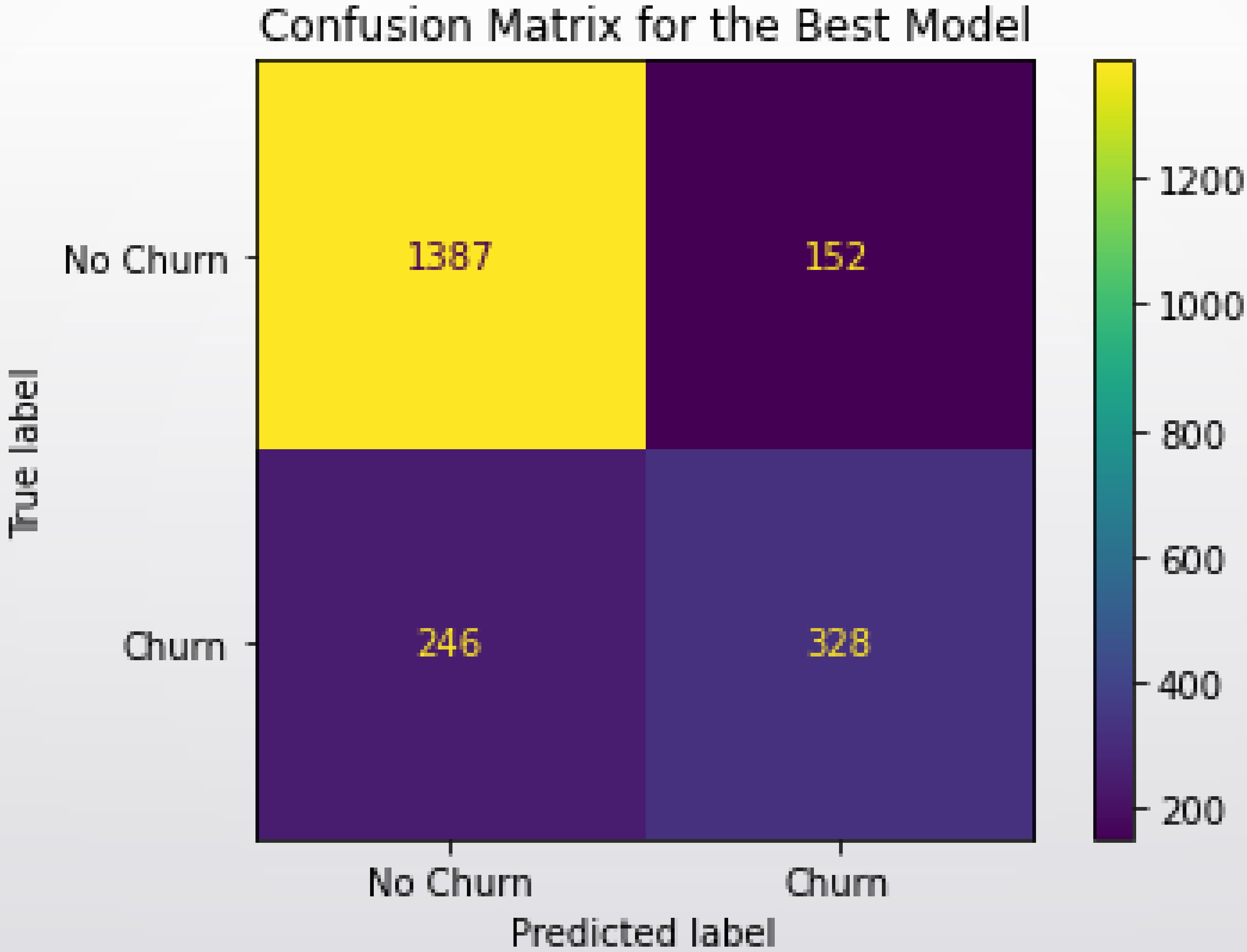
- • Models Built: Logistic Regression, Decision Tree, and Random Forest.
- • Baseline Model: Logistic Regression for interpretability.
- • Advanced Models: Hyperparameter-tuned Logistic Regression for improved accuracy.
- • Metrics: Evaluated using accuracy, precision, recall, F1-score, and ROC-AUC.


Model Performance Comparison
Percentage
0
10
20
30
40
50
60
70
80
test_accuracy
test_precision
test_recall
test_f1
LogisticRegression
RandomForestClassifier
DecisionTreeClassifier
model_name

Evaluation

- • Best Model: Logistic Regression achieved 81% accuracy and 0.86 ROC-AUC.
- • Key Metrics: Balanced precision and recall to address business needs.
- • Real-World Implications: Identified high-risk customers for proactive retention.


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LogisticRegression (AUC = 0.86)
RandomForestClassifier (AUC = 0.83)
DecisionTreeClassifier (AUC = 0.67)


Confusion Matrix for the Best Model
No Churn
Churn
True label
1387
152
246
328
No Churn
Churn
Predicted label
1200
1000
800
600
400
200

Model Selection

- • Selected Model: Logistic Regression
- • Performance: 81%
- - Accuracy: 0.81
- - ROC-AUC: 0.86
- • Reason: Balanced precision and recall, effectively identifying churners.

Limitations

- 1. Class Imbalance: Fewer churn cases affect model performance.
- 2. Limited Features: Missing customer feedback or competitor data.
- 3. Recall vs. Precision: High recall but moderate precision leads to false positives.

Recommendations and Areas of Improvement

- 1. Use Model Predictions: Implement targeted retention campaigns for high-risk customers.
- 2. Enhance Data Quality: Incorporate customer support interactions and market data.
- 3. Regular Monitoring: Retrain the model periodically to adapt to evolving patterns.

Conclusion

- The project successfully predicts churn and provides actionable insights for retention strategies. Future work will address class imbalance and expand feature diversity for better accuracy and impact.