Predicting Customer Churn for a Telecom Company

Team Insight

Nisarg Patel, Parth Patel

Project Overview

- Objective: Develop a model to predict customer churn using the Telco Customer Churn dataset.
- Importance: Churn prediction is crucial for improving customer retention and enhancing profitability in telecom.

Steps To Follow

- **Set Up and Initial Data Load**
- **Exploratory Data Analysis (EDA)**
- **Data Cleaning and Preprocessing**
- **Data Summary**
- **Data Visualization**
- **Model Selection and Initial Training**
- **Performance Evaluation and Metrics Selection**
- **Reflection and Next Steps**

Achievements of the Past Two Weeks

- Key Accomplishments:
- Completed initial data load and inspection.
- Performed Exploratory Data Analysis (EDA) with key insights.
- Cleaned and preprocessed the dataset.
- Started baseline model training.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

Data Load and Analysis

- Numerical Features: SeniorCitizen, tenure, MonthlyCharges, TotalCharges
- Categorical Features: Gender, Partner, Dependents, Contract, PaymentMethod, etc.

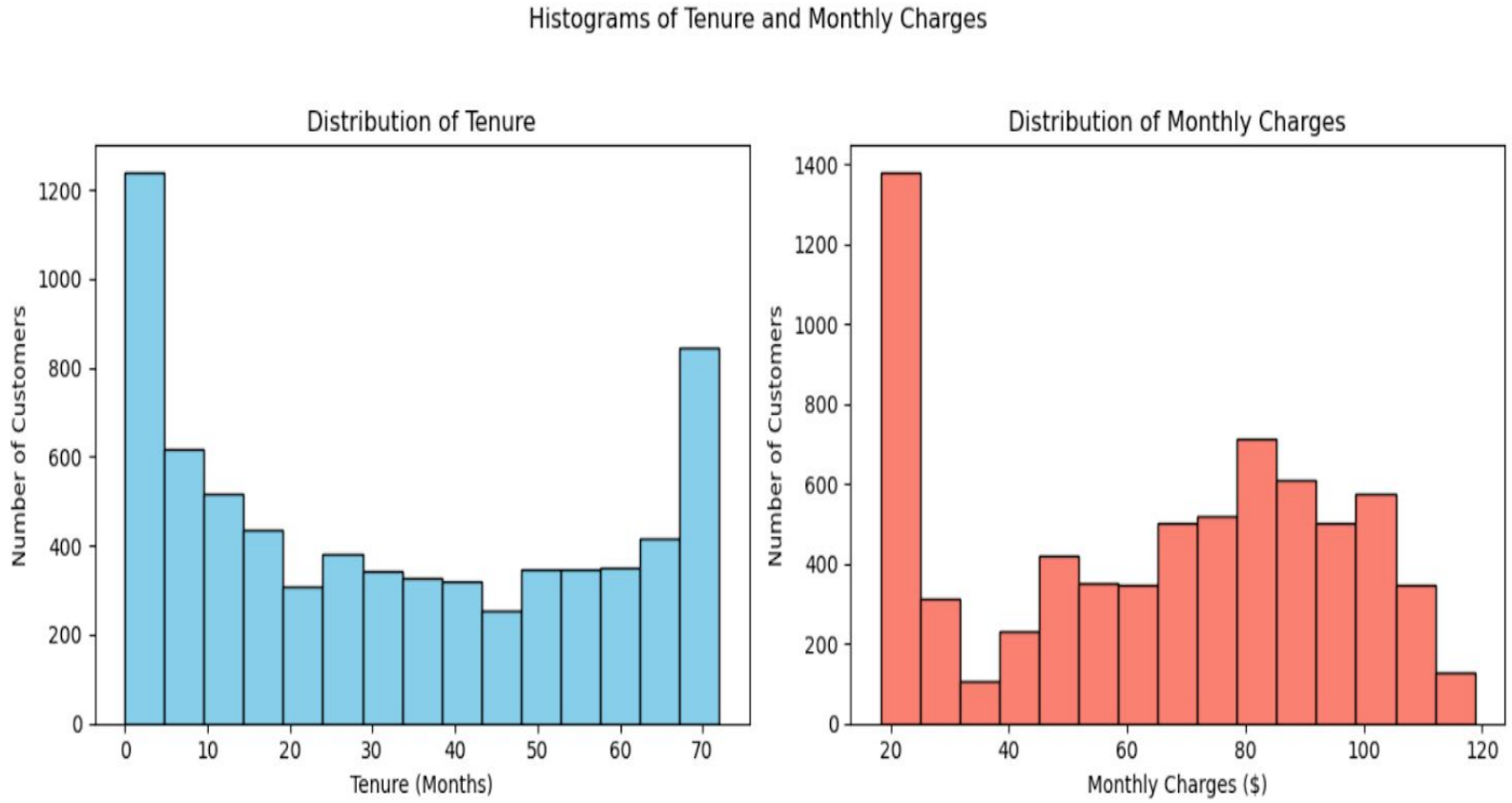
The TotalCharges column was incorrectly formatted as an object type due to missing or non-numeric values.

Basic Statistics

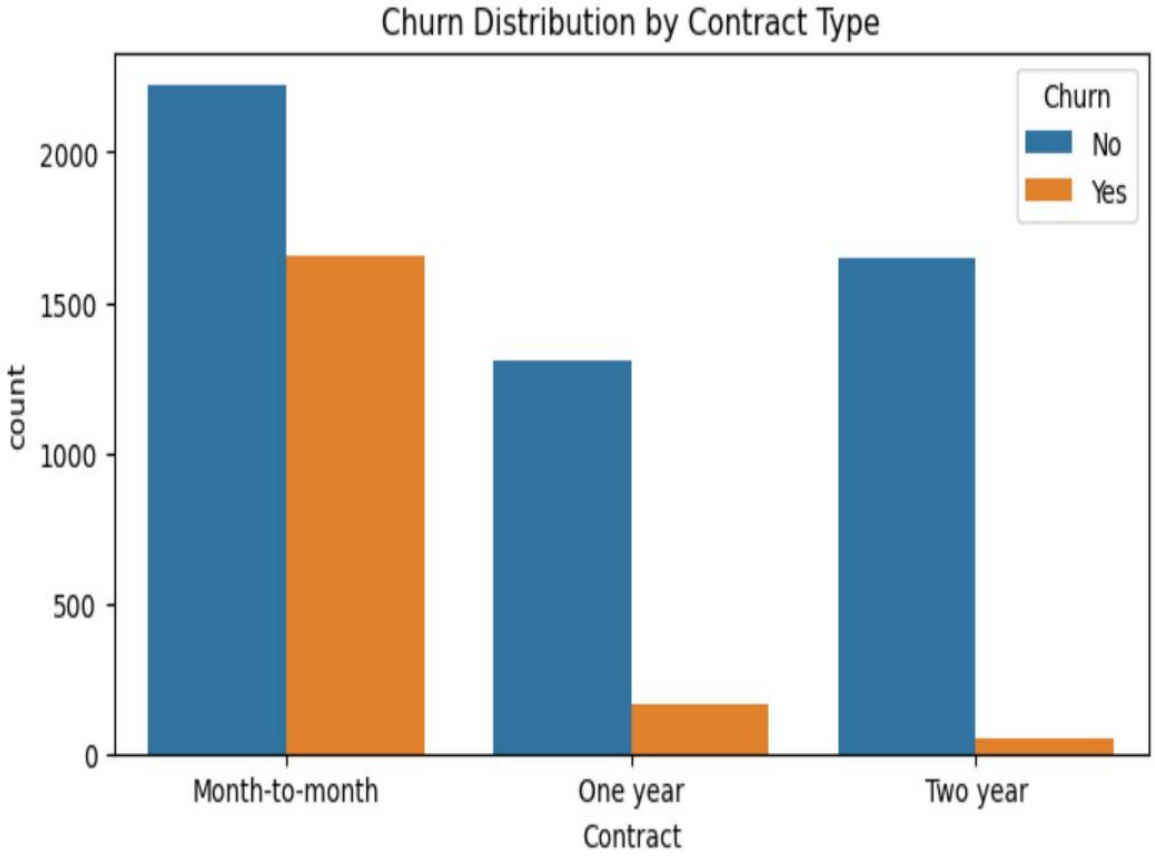
- Approximately 16.2% of customers are seniors
- Ranges from 0 to 72 months, with a mean of 32.4 months, indicating a mix of new and long-term customers
- MonthlyCharges: Values range from $18.25 to $118.75, suggesting diverse service plans

Next Step

- Clean the dataset by handling missing values and converting TotalCharges to numeric.
- Prepare the data for exploratory data analysis to uncover trends and relationships.
- Build a robust predictive model using the cleaned data.
- Focus on features such as tenure, MonthlyCharges, and contract type for predicting churn.



- Customers with low tenure might have a higher churn rate, while those with high tenure are probably loyal
- Higher monthly charges could correlate with churn if customers are dissatisfied with the cost-to-value ratio



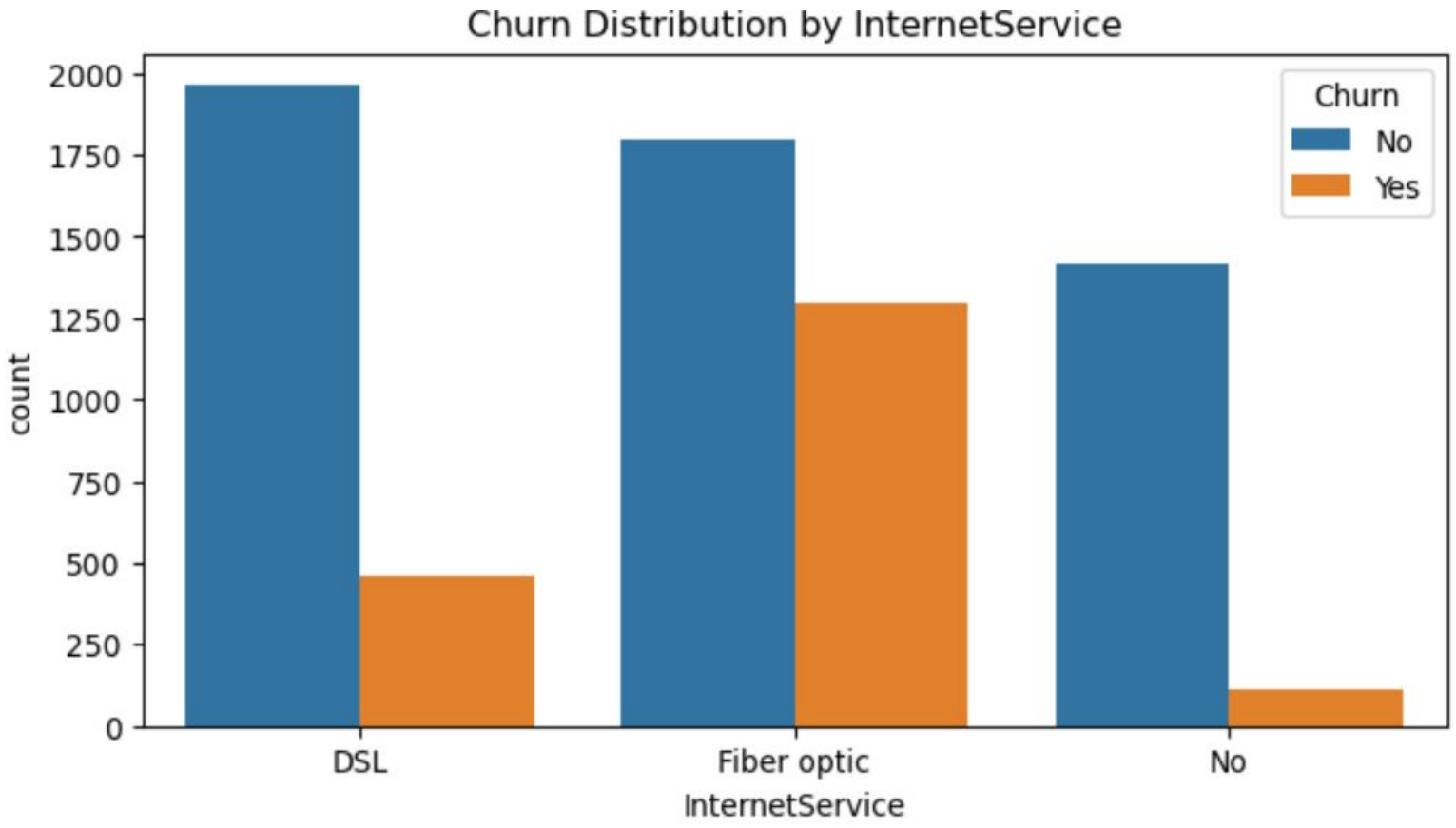
- Customers on month-to-month contracts have a significantly higher churn rate compared to those on one-year or two-year contracts
- Contract type appears to be a strong predictor of churn, as customers on longer-term contracts (one-year or two-year) are less likely to churn, possibly due to early termination fees or a higher commitment level

Based on insights

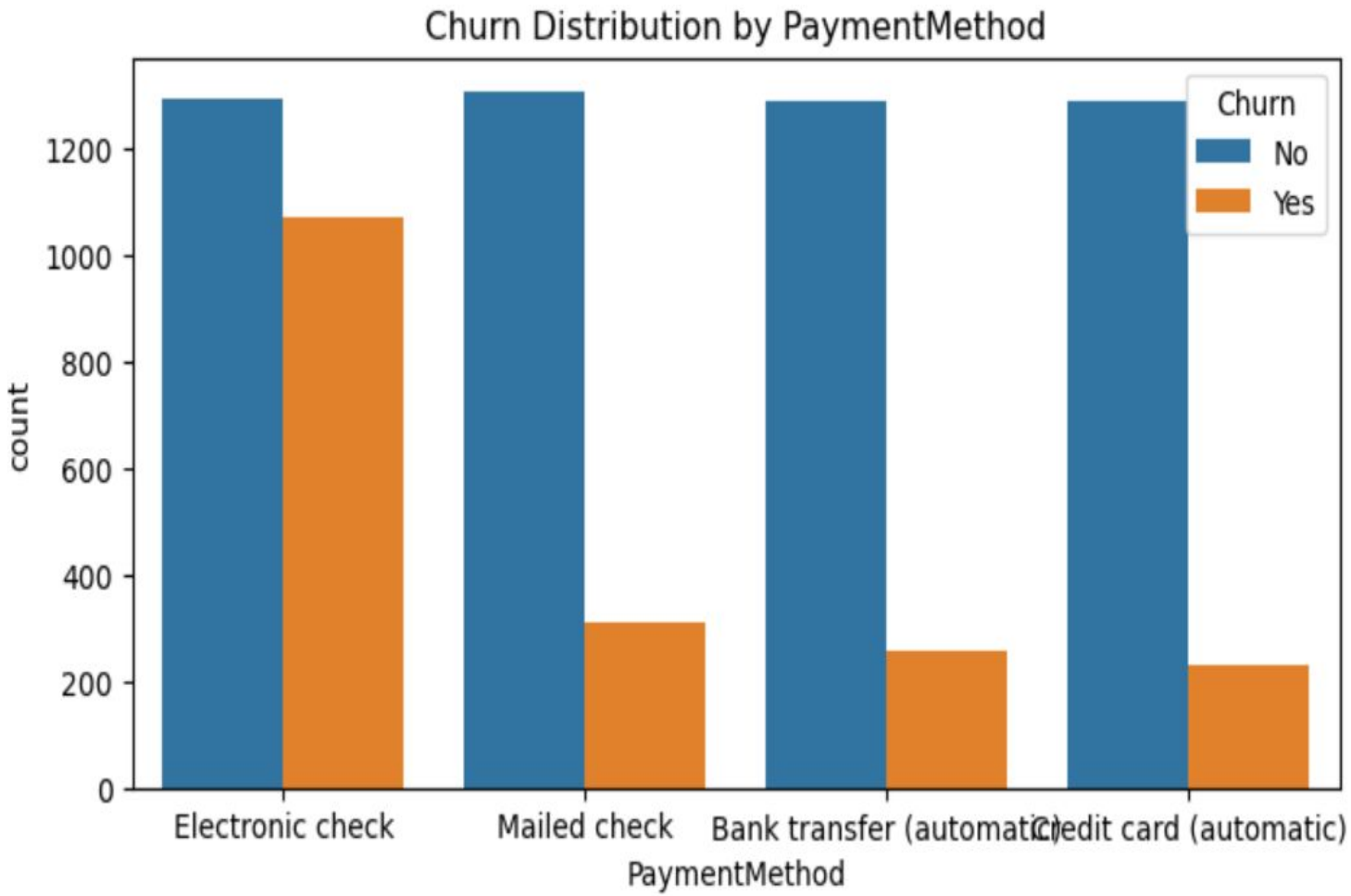
- **Tenure** and **Contract Type** appear to be strong indicators of churn
- **Monthly Charges** might influence churn but require further examination
- We may consider using either **TotalCharges** or **tenure** to avoid high correlation redundancy in the model

Extended EDA Plan

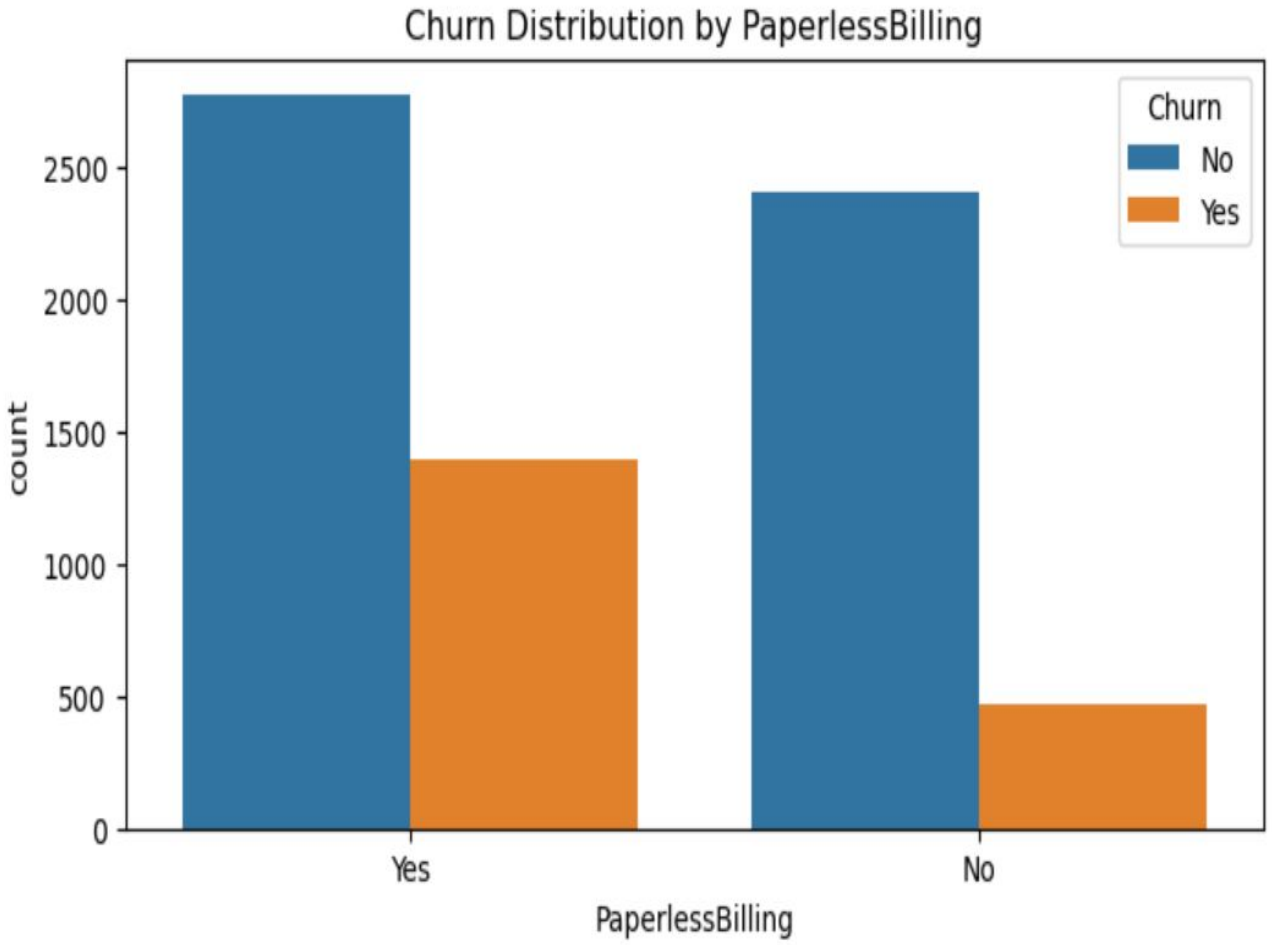
- **Churn Rate by Categorical Variables**
- **Numerical Feature Analysis by Churn**
- **Bivariate Analysis**
- **Outlier Detection**
- **Further Correlation Analysis**



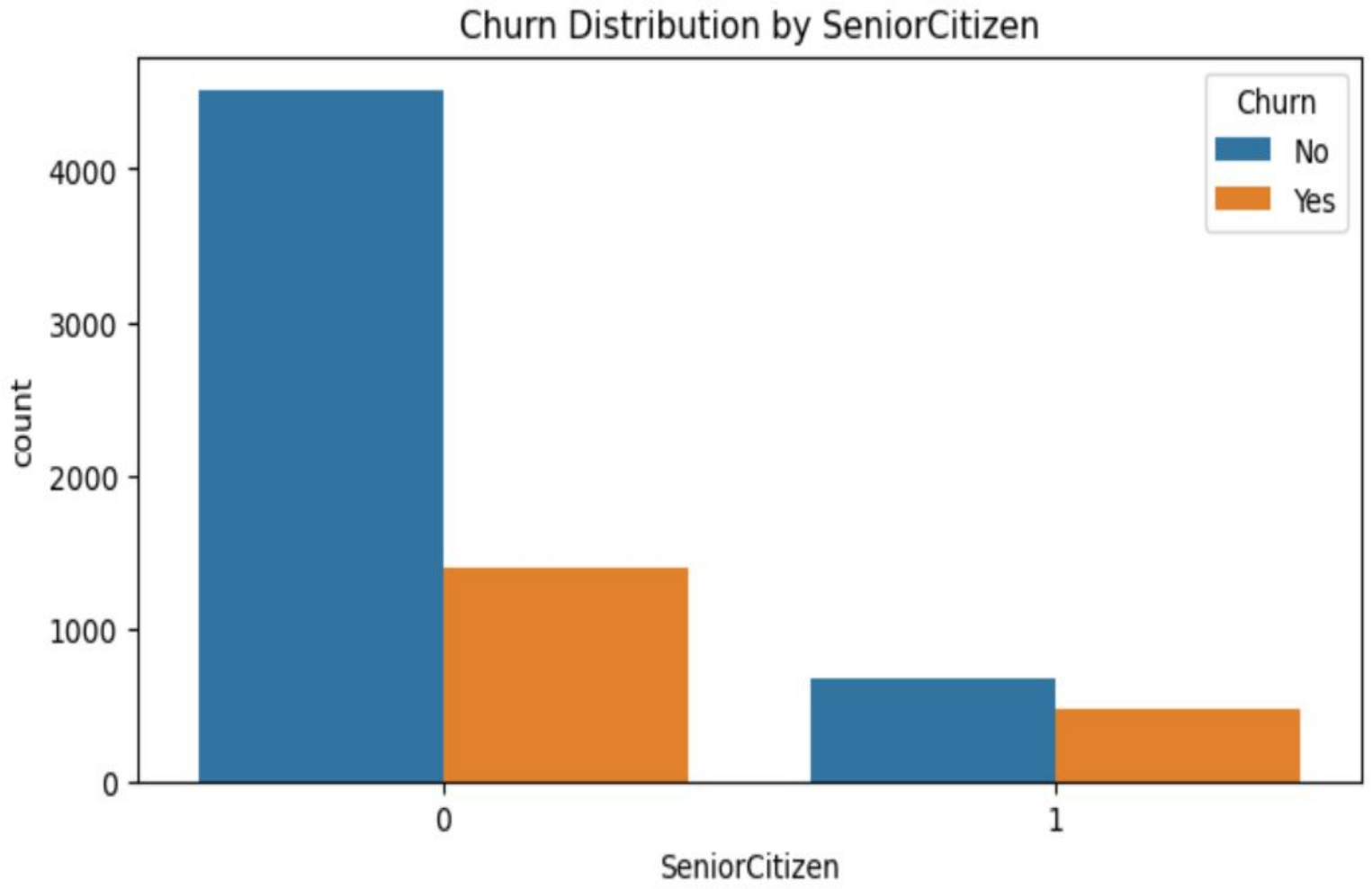
- Customers with **Fiber Optic** services show a higher churn rate compared to those with **DSL** or **No Internet Service**
- Fiber optic services might have a higher cost or other factors that contribute to dissatisfaction, making this a significant factor to consider in churn prediction



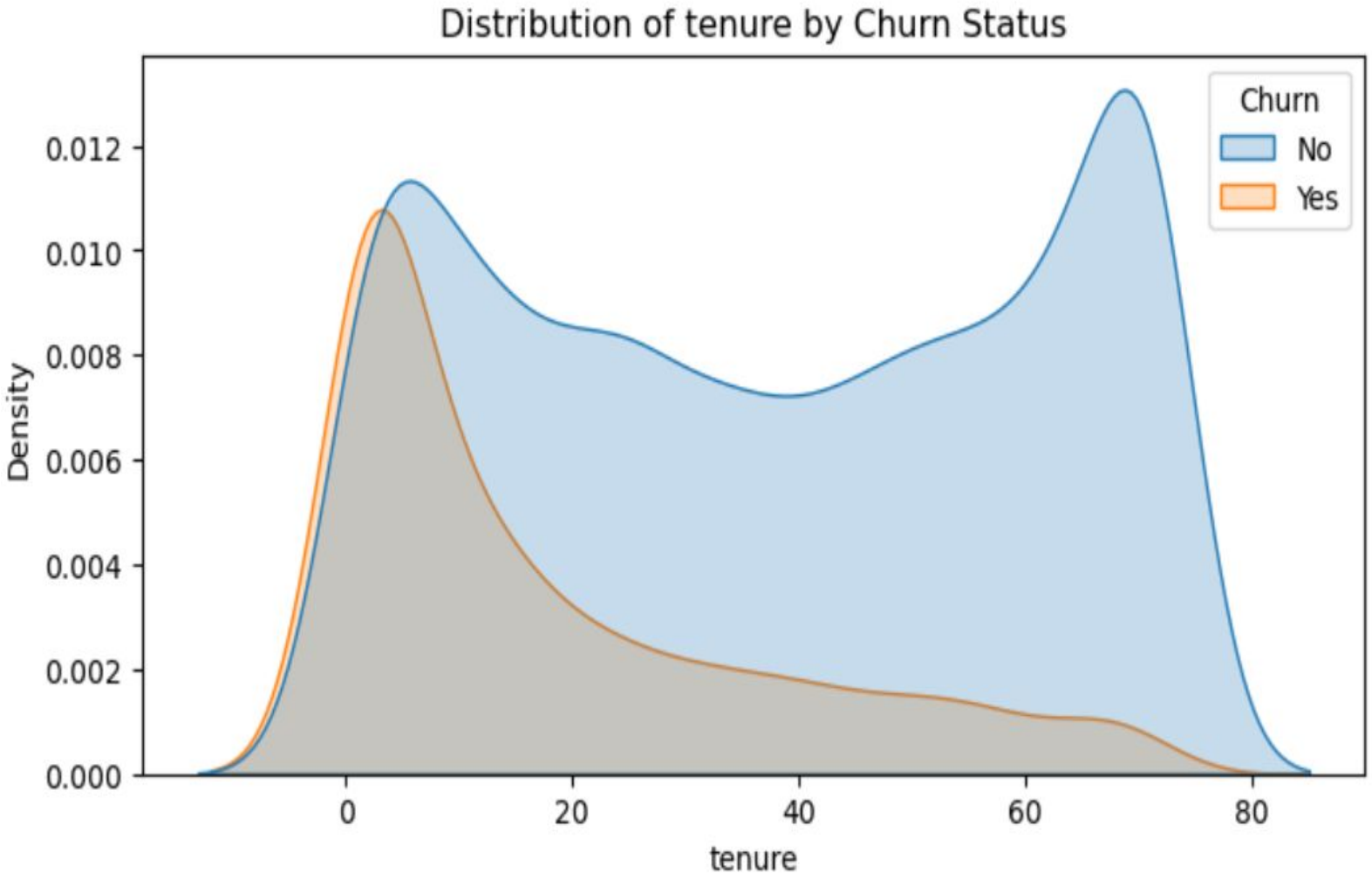
Customers using **Electronic Check** have a notably higher churn rate than those using other payment methods like **Mailed Check**, **Bank Transfer**, or **Credit Card** (automatic)



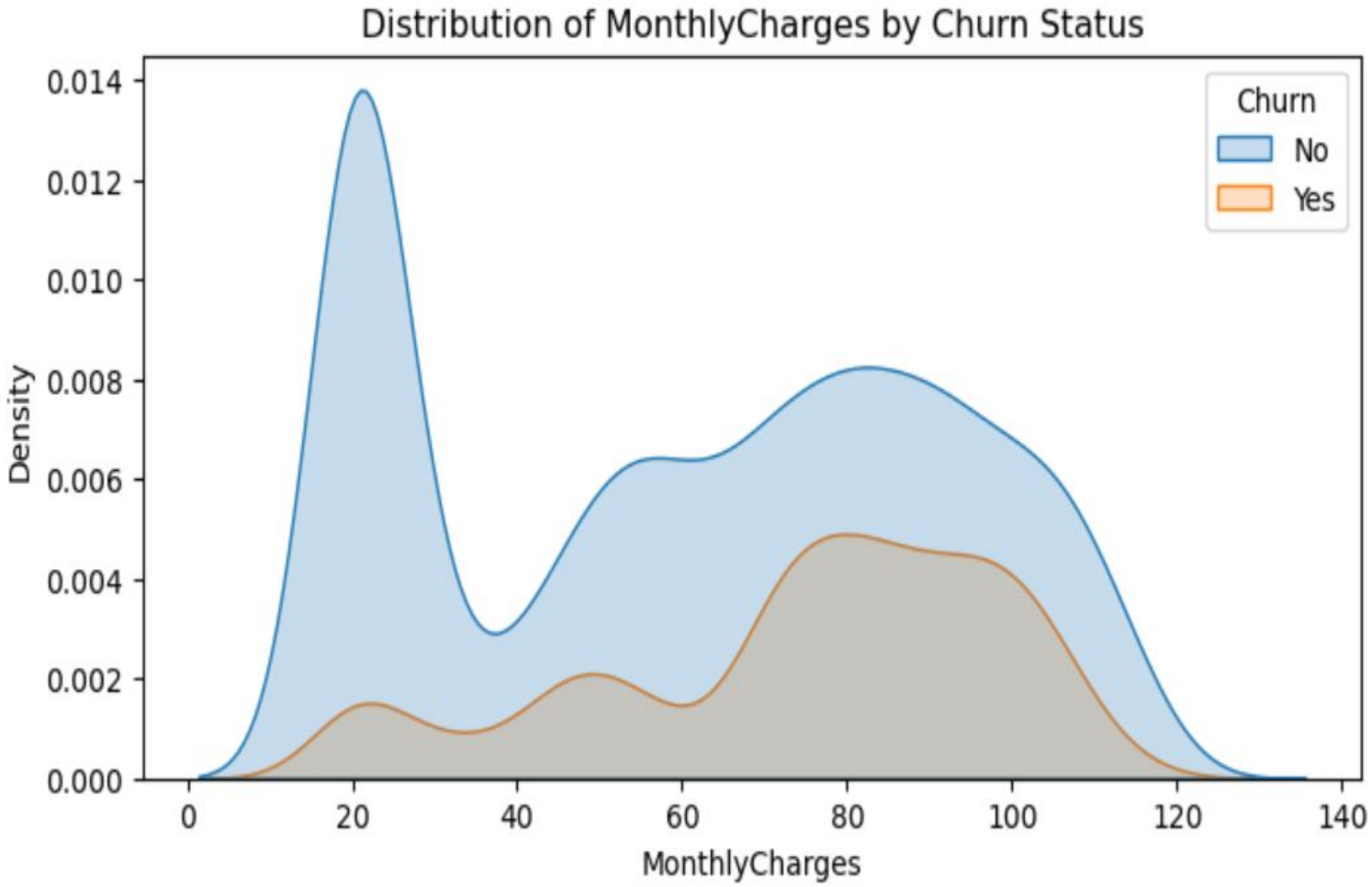
Customers with **Paperless Billing** show a higher churn rate than those who receive paper bills.



Senior citizens have a slightly higher churn rate compared to younger customers.



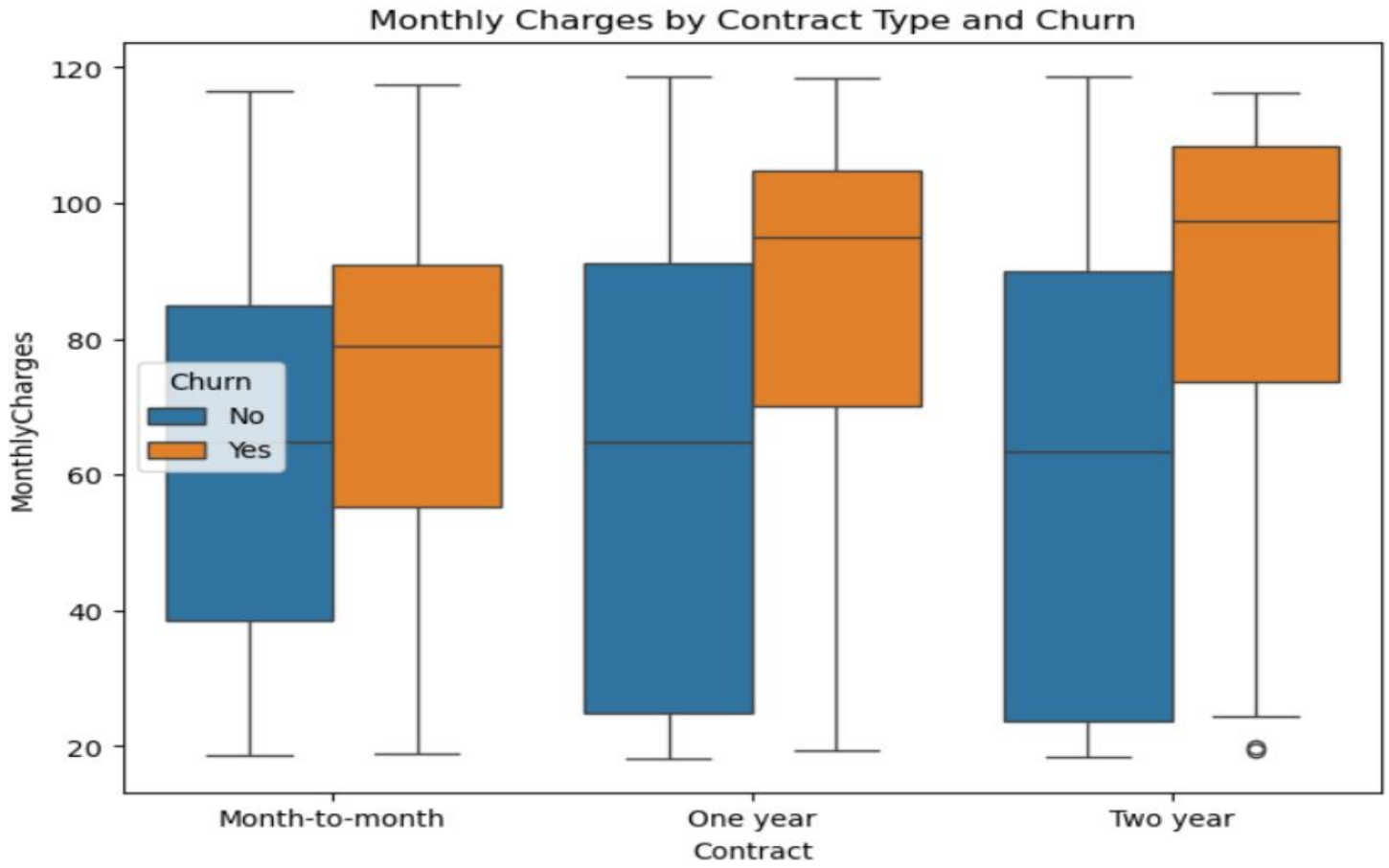
Customers with lower tenure (newer customers) are more likely to churn, while those with longer tenure show lower churn rates.



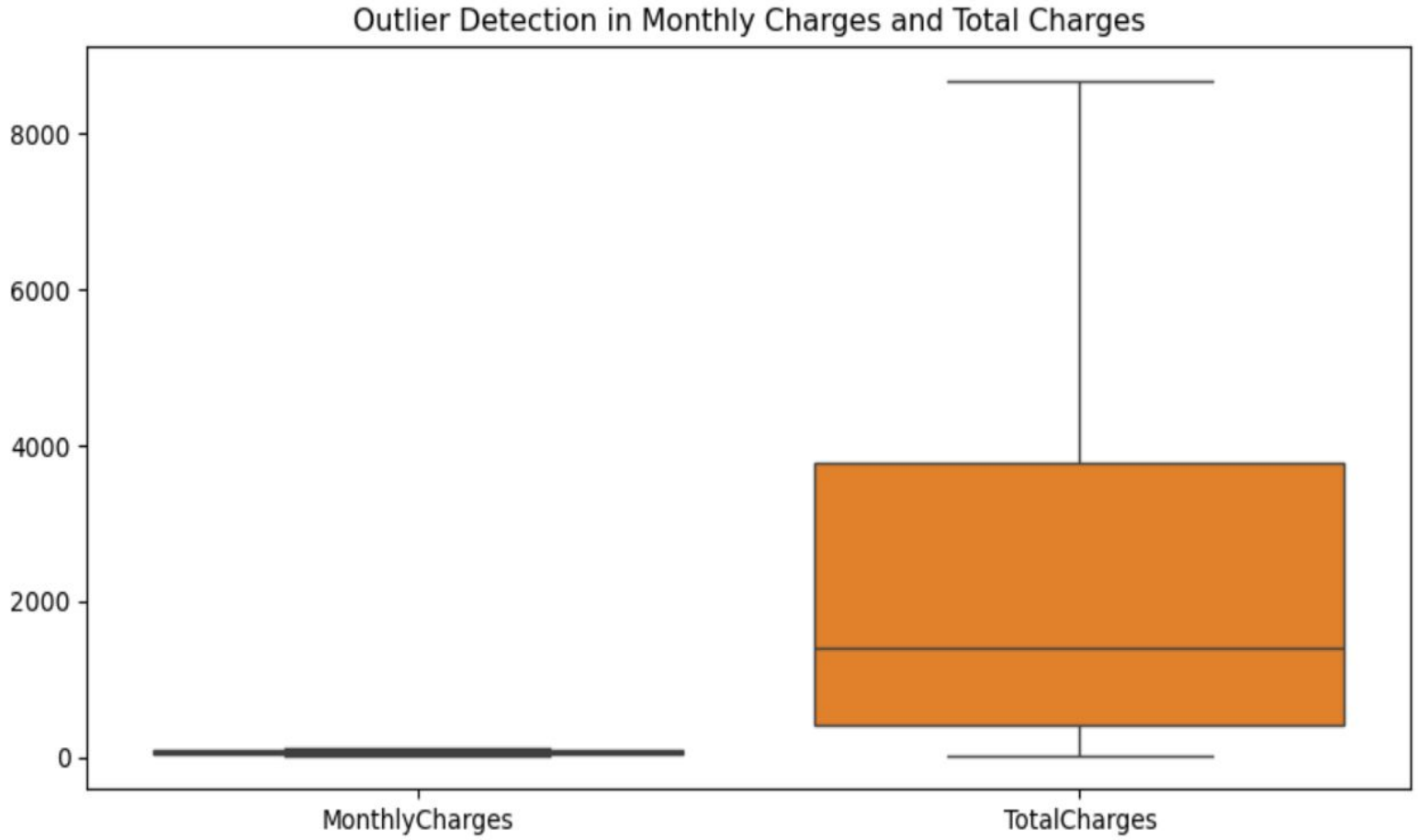
Higher monthly charges are associated with higher churn rates.

Summary of EDA Insights

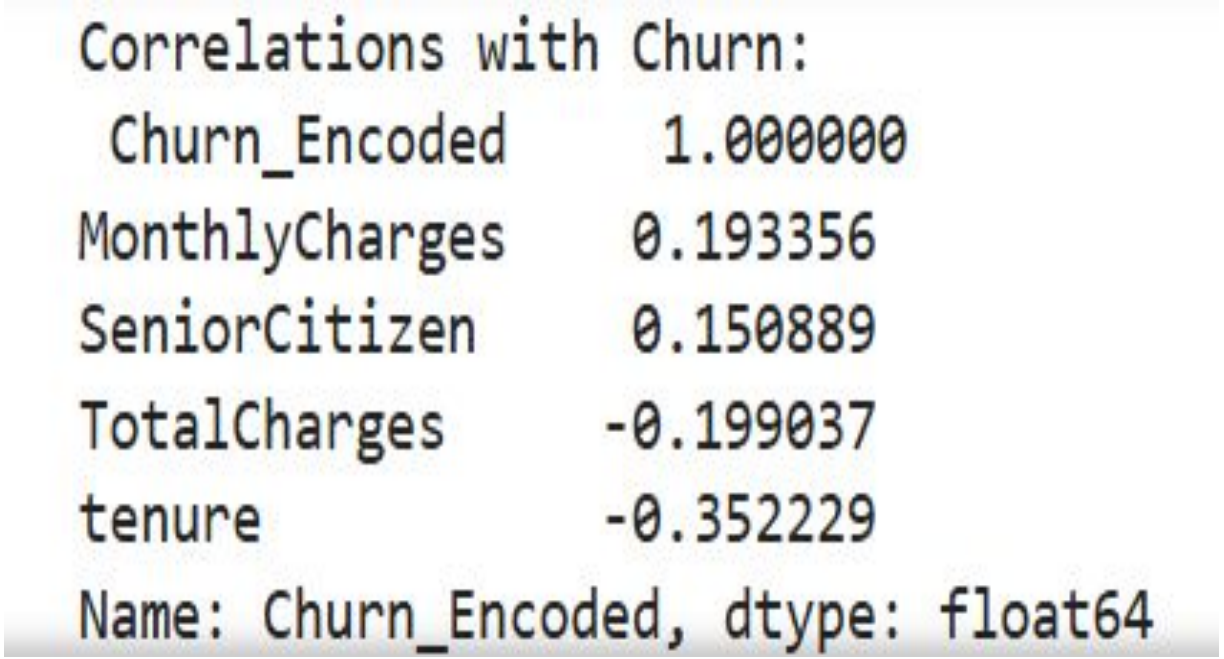
- **Internet Service Type**: Fiber optic users are more likely to churn
- **Payment Method**: Electronic check users show a higher churn tendency.
- **Paperless Billing**: Customers with paperless billing have a higher churn rate.
- **Tenure**: Shorter-tenured customers are more likely to churn.
- **Monthly Charges**: Higher monthly charges correlate with churn.



For month-to-month contracts, churned customers tend to have higher monthly charges than non-churned customers.This trend is less pronounced for one-year and two-year contracts.



- The box plot highlights that **TotalCharges** has several outliers, especially for customers with high tenure and extensive service use. Monthly charges appear to have fewer extreme outliers.
- Outliers in **TotalCharges** reflect customers with high cumulative costs, which could relate to long tenure or high service usage. These outliers should be carefully managed during preprocessing to avoid skewing the model.



- **MonthlyCharges** (0.19): A positive correlation with churn, suggesting that customers with higher monthly charges are slightly more likely to churn.
- **SeniorCitizen** (0.15): A weak positive correlation with churn, indicating that older customers might have a marginally higher tendency to leave.
- **TotalCharges** (-0.20): A negative correlation, which aligns with our observation that customers with higher total charges (usually long-term customers) tend to stay.
- **Tenure** (-0.35): The strongest negative correlation with churn, reinforcing that longer-tenured customers are less likely to churn.

Data Preprocessing and Feature Engineering Steps

- Handle Missing Values
- Encode Categorical Variables
- Scale Numerical Features
- Feature Engineering Based on EDA Insights
- Train-Test Split

Encode Categorical Variables

Categorical variables include:

gender, Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, and PaymentMethod.

- Apply one-hot encoding to these features to convert them into numerical form
- By using **drop_first=True**, we avoid multicollinearity by dropping one category per variable

Derived Features

- **Average Monthly Charges :-** dividing TotalCharges by tenure
- **Binned Tenure :-** tenure into categories such as New, Mid, and Loyal
- **High Monthly Charges Indicator :-** sets 1 for customers whose MonthlyCharges exceed a chosen threshold and 0 otherwise

Data Preprocessing for Modeling

- Feature Selection and Engineering
- Data Splitting
- Preprocessing Pipelines
- Model Training
- Evaluation and Next Steps

Logistic Regression Model

- Logistic Regression is easy to understand and provides coefficients that can help us see which features contribute most to predictions.
- Model Training Steps:

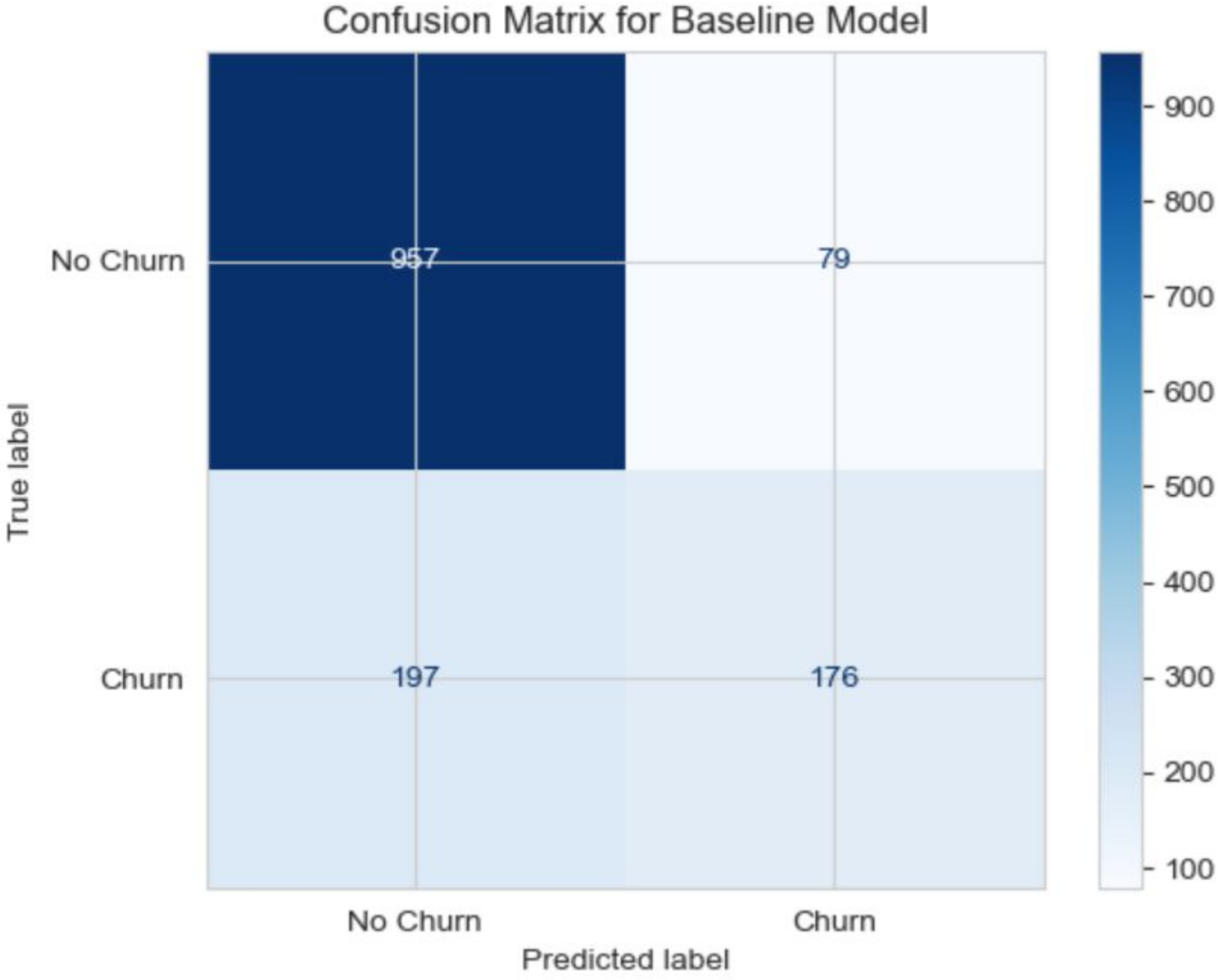
1. Train a Logistic Regression model on the training data.
2. Evaluate the model using accuracy, precision, recall, and F1-score.
3. Analyze the initial results to understand model performance.

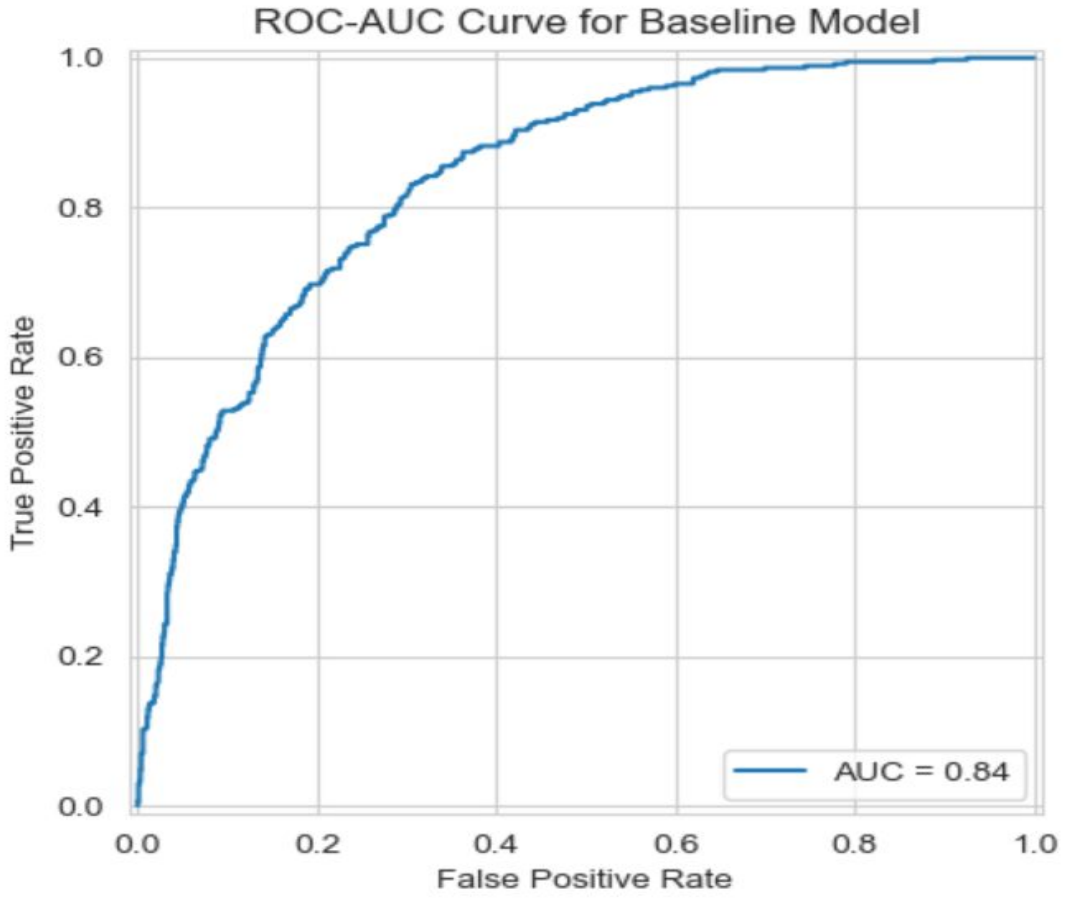
Accuracy: 0.8041163946061036

AUC-ROC: 0.8389570631527737

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.92	0.87	1036
1	0.69	0.47	0.56	373
accuracy			0.80	1409
macro avg	0.76	0.70	0.72	1409
weighted avg	0.79	0.80	0.79	1409


Confusion Matrix for Baseline Model
No Churn
Churn
True label
957
79
197
176
No Churn
Churn
Predicted label
900
800
700
600
500
400
300
200
100



Accuracy: 0.8041163946061036

AUC-ROC: 0.8389570631527737

Analysis of Results

- **Accuracy (80%)**: The model correctly predicts about 80% of the cases, which is a good starting point.
- **AUC-ROC (0.83)**: This value indicates that the model has a good ability to distinguish between churn and non-churn cases.
- **Recall for Churn (47%)**: The model only identifies 47% of actual churn cases, which shows room for improvement in detecting churn customers.
- **Cross-Validation Mean Score (0.79)**: Consistent with the test accuracy, suggesting the model generalizes reasonably well but can be optimized further.

Next Steps for Improvement

- **Address Class Imbalance**: Implement techniques such as SMOTE (Synthetic Minority Oversampling Technique) to balance the training set.
- **Feature Engineering**: Create new features that may better capture patterns in customer behavior (e.g., interaction between Contract type and MonthlyCharges).
- **Model Complexity**: Move to more complex models like Random Forests, Gradient Boosting.
- **Evaluate with Precision-Recall Curve**: Analyze the precision-recall trade-off, especially since recall for churn is low.