A PROJECT REPORT

ON

PREDICTING CUSTOMER CHURN IN A TELECOMMUNICATION COMPANY

BY

SRIRAMSETTY BHANU TEJA

Contents

OBJECTIVE 3
Introduction 3
Data Collection and Preprocessing 3
Dataset 3
Data Preprocessing 3
Loading and Overview 3
Handling Missing Values 3
Data Cleaning 3
Feature Engineering 4
Data Exploration 4
Data Export 4
Exploratory Data Analysis (EDA): 5
Summary Statistics 5
Univariate Analysis 5
One-Hot Encoding 6
Bivariate Analysis 11
Prediction Model 13
Conclusion 16

OBJECTIVE

The primary objective of this project is to develop a predictive model that can identify customers at risk of churning, enabling the company to take proactive measures to retain them.

Introduction

The telecommunications industry faces the challenge of customer churn, which can impact revenue and business sustainability. This project aims to develop a predictive model to identify customers at risk of churning, enabling proactive retention strategies.

Data Collection and Preprocessing

Dataset

The dataset used for this project is obtained from Kaggle: https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

Data Preprocessing

Loading and Overview

- The dataset was loaded using Pandas, and the first five records were displayed for an initial overview (head ()).
- The shape of the dataset was checked to understand the number of rows and columns (shape).

Handling Missing Values

- The dataset was examined for missing values, and a point plot was used to visualize the percentage of missing values for each column (info ()).
- No missing values were found, ensuring the dataset's integrity.

Data Cleaning

- 'TotalCharges' was converted to numeric format using pd.to_numeric() to facilitate numerical analysis.
- 11 null values in 'TotalCharges' were identified and subsequently dropped as they constituted a small percentage (0.15%).

Feature Engineering

- 'tenure_grp' was created by dividing customers into groups based on tenure using pd.cut() with predefined labels.
- Unnecessary columns ('customerID' and 'tenure') were dropped from the dataset.

```
data_teleco2['tenure_grp'].value_counts()
```

```
1 - 12     2175
61 - 72    1407
13 - 24    1024
25 - 36     832
49 - 60     832
37 - 48     762
Name: tenure_grp, dtype: int64
```

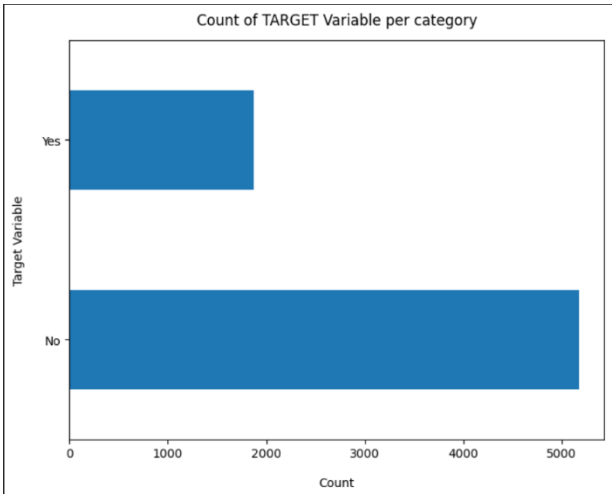
Data Exploration

- Exploratory Data Analysis (EDA) was conducted to understand the distribution of variables and identify patterns.

Data Export

- The cleaned and processed dataset, named 'Telecom_churn.csv,' was saved for future analysis.

These streamlined preprocessing steps ensured data cleanliness, handled missing values, engineered relevant features, and prepared the dataset for further analysis and modeling.



Exploratory Data Analysis (EDA):

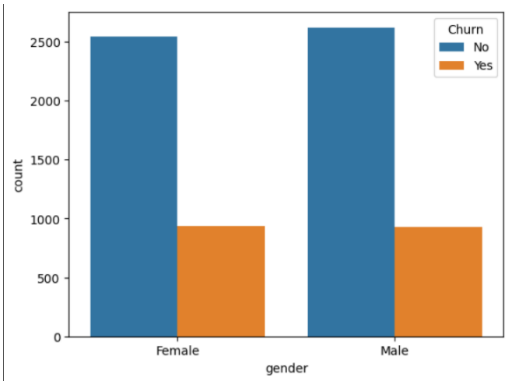
Exploratory Data Analysis is a crucial phase in understanding the characteristics of the dataset, identifying patterns, and gaining insights into the underlying structure of the data. In the context of the provided code, the EDA process includes the following steps:

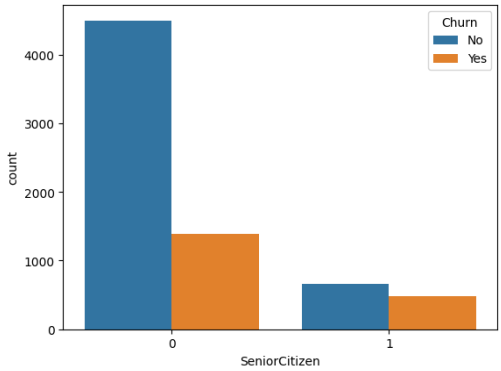
Summary Statistics

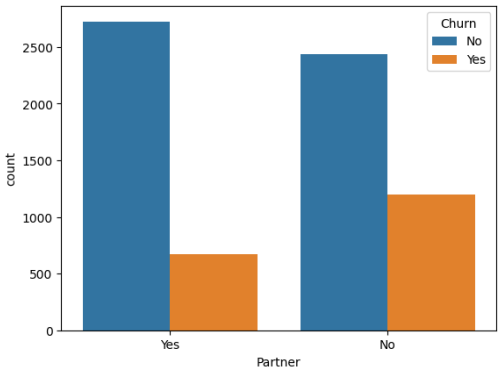
- Descriptive statistics of key variables were computed using the **describe()** function, providing insights into central tendency, dispersion, and shape of the distribution.
- The distribution of numerical features was visualized through histograms, offering a quick overview of data spread and skewness.

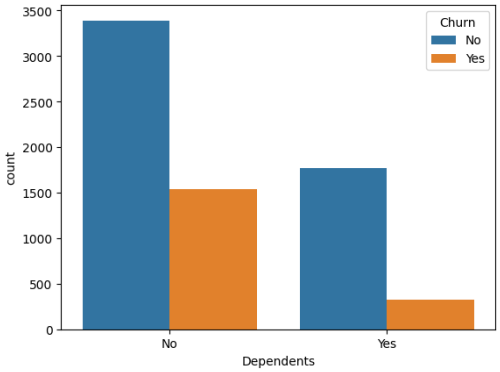
Univariate Analysis

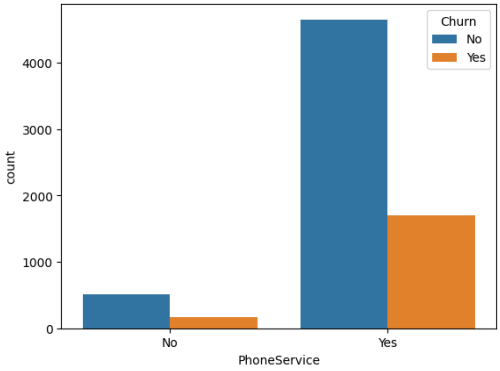
Individual predictors were analysed using count plots to visualize the distribution of churn and non-churn instances for each category. This provided a clear representation of the impact of different factors on customer churn.

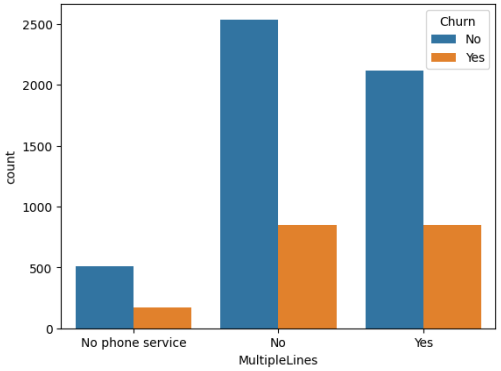


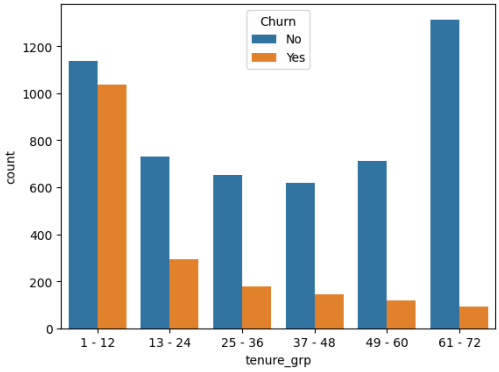


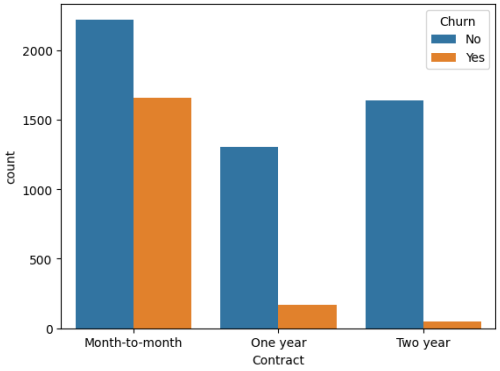












One-Hot Encoding

To handle categorical variables and ensure compatibility with machine learning algorithms, one-hot encoding was employed. This process involves converting categorical variables into binary vectors, creating new binary columns for each category.

Categorical Variables:

- Identified categorical variables in the dataset, including but not limited to 'Contract,' 'PaymentMethod,' 'TechSupport,' and 'tenure_grp.'

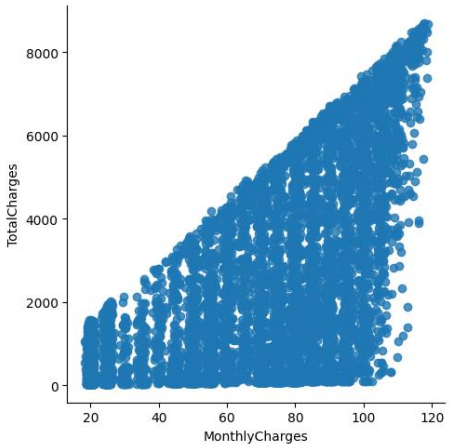
One-Hot Encoding Procedure

- Utilized the **get_dummies()** function from the Pandas library to perform one-hot encoding on the identified categorical variables.
- The function created binary columns for each category within the categorical variables, assigning binary values (0 or 1) based on the presence of each category.

Resulting Dataset

- The one-hot encoded dataset, named 'data_teleco_dummies,' was created, incorporating the newly generated binary columns for categorical variables.

Scatter plots were employed for visualizing relationships between numerical variables. Notably, a scatter plot was used to depict the relationship between Monthly Charges and Total Charges.



Comparing Monthly Charges by Churn and Total Charges by Churn

1. **Relationship Analysis:**

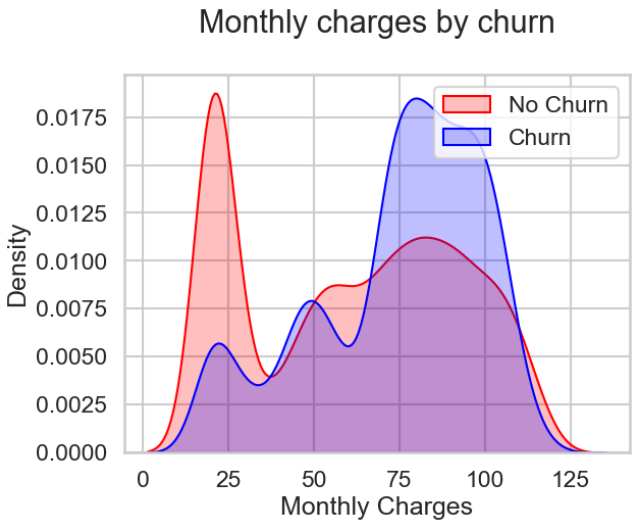
   A scatter plot indicated a positive linear relationship between Monthly Charges and Total Charges.

2. **Churn Analysis:**

   KDE plots were used to analyse Monthly Charges and Total Charges concerning customer churn.

3. **Monthly Charges and Churn:**

   Higher Monthly Charges correlated with increased churn rates, suggesting that customers with higher monthly expenses are more likely to churn.



Total Charges and Churn:

Unexpectedly, higher churn rates were observed at lower Total Charges, indicating that customers with lower overall spending are more prone to churn.

Combined Insights:

Customers with higher Monthly Charges, lower tenure, and lower Total Charges demonstrated higher churn rates.

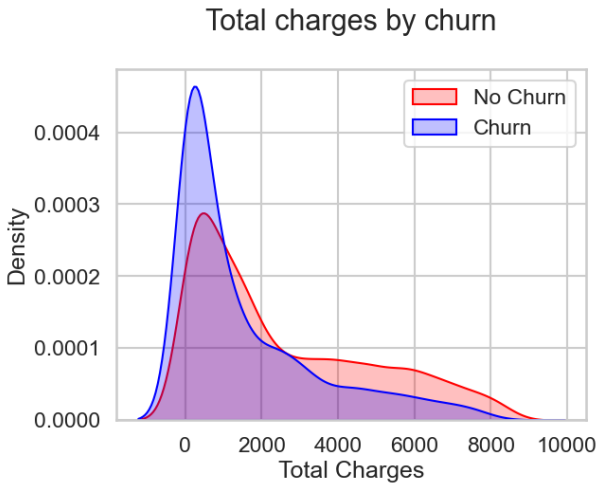
Implications:

- Pricing strategies or targeted offers may be considered to retain customers with higher monthly expenses.
- Further investigation into the specific services associated with higher Monthly Charges could provide valuable insights.

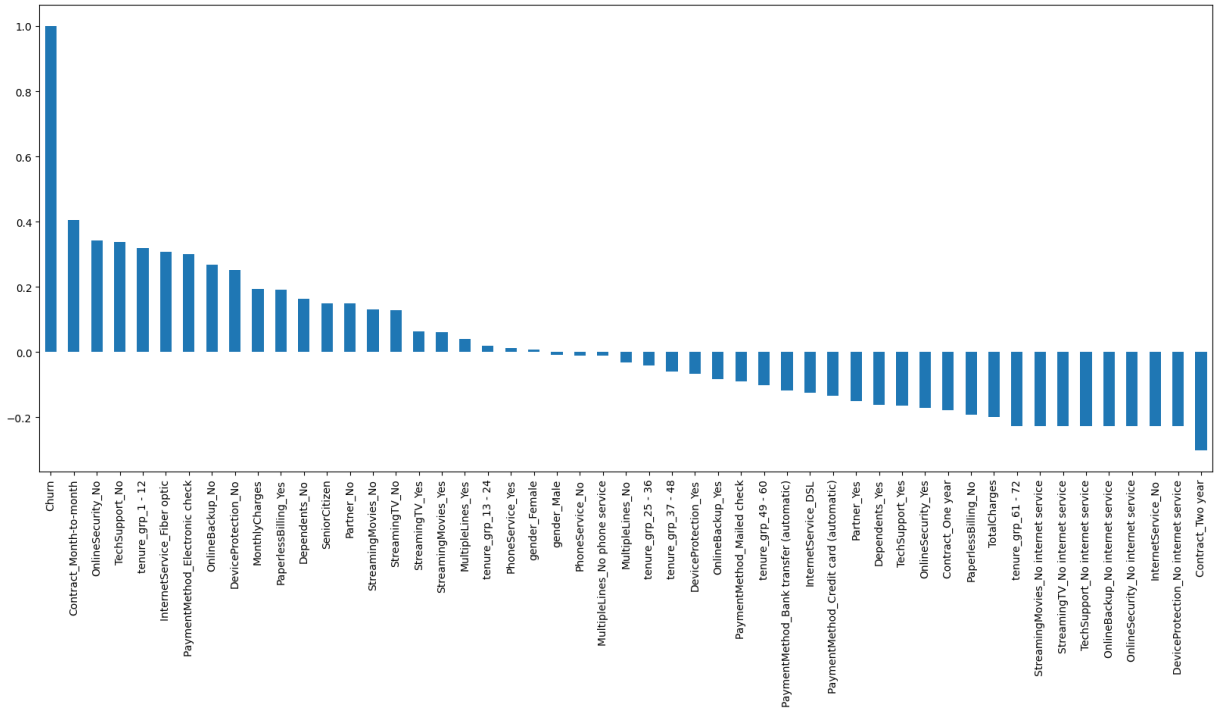
Future Considerations:

- Future analysis could explore additional factors such as tenure and service quality to refine churn prediction strategies.

These insights provide actionable information for the telecommunications company to tailor retention efforts based on customer billing patterns and potential indicators of dissatisfaction.



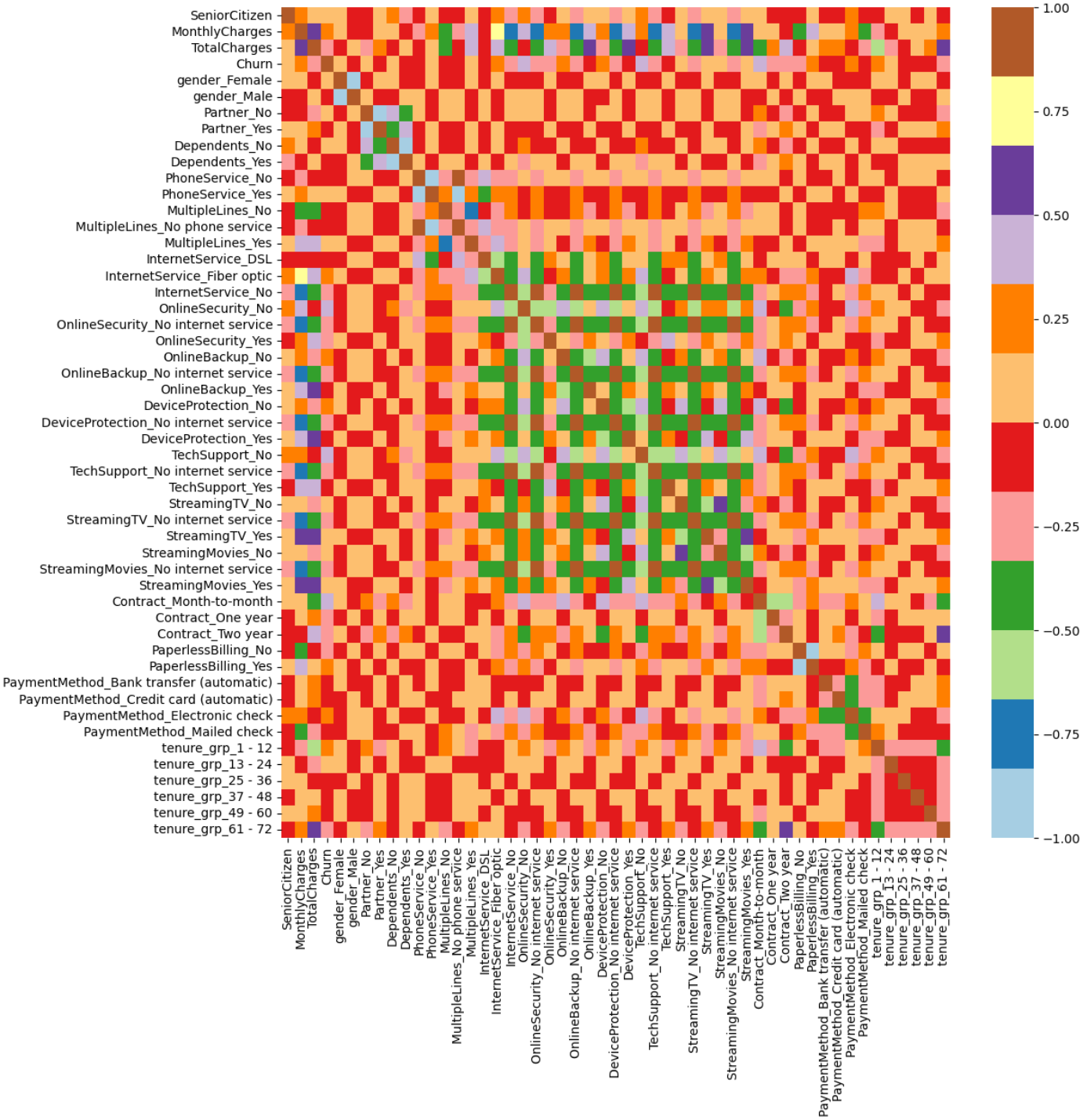
1. **Correlation Plot:**
    - Utilized a heatmap to visualize the correlation coefficients between all predictors and 'Churn.'
2. **Key Observations:**
    - Positive correlations indicate factors that increase the likelihood of churn, while negative correlations suggest protective effects.



3. **High Churn Factors:**
    - Identified predictors with high positive correlations, such as month-to-month contracts, no online security, and no tech support.
4. **Low Churn Factors:**
    - Recognized predictors with negative correlations, including long-term contracts, subscriptions without internet service, and extended customer engagement.
5. **Neutral Impact Variables:**
    - Acknowledged variables with minimal correlation to 'Churn,' like gender, phone service availability, and the number of multiple lines.
6. **Implications for Strategy:**
    - Informed potential strategies to address high-churn factors and reinforce customer engagement through long-term contracts.
7. **Heatmap Insights:**
    - The heatmap visually guides the strength and direction of correlations, aiding in pattern recognition.
8. **Limitations and Further Exploration:**
    - Recognized that correlation does not imply causation, suggesting the need for further exploration and validation.

9. **Strategic Decision-Making:**

    - Insights from the correlation analysis guide strategic decisions, focusing on mitigating high-churn factors and strengthening customer retention strategies.



Bivariate Analysis

Objective:

Explored relationships between pairs of variables to understand their joint impact on customer churn.

Methodology:

Utilized count plots and visualizations to analyse the interaction between two variables simultaneously.

Insights Drawn:

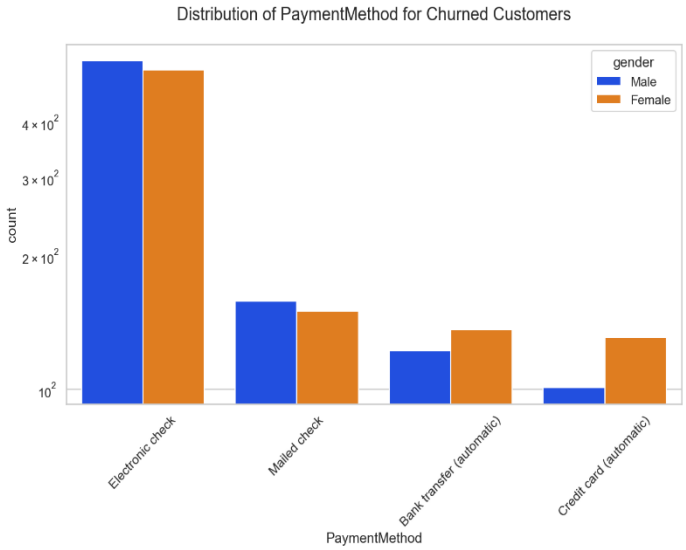
Derived insights into how specific variables influence the likelihood of churn when considered together.

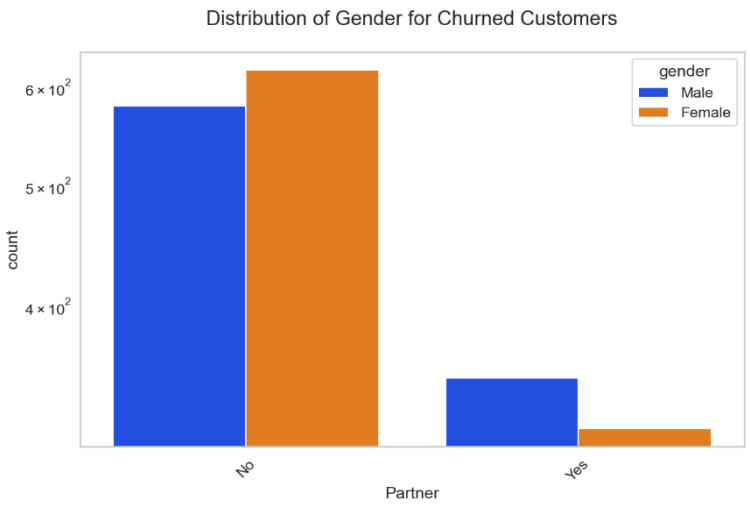
Visual Representations:

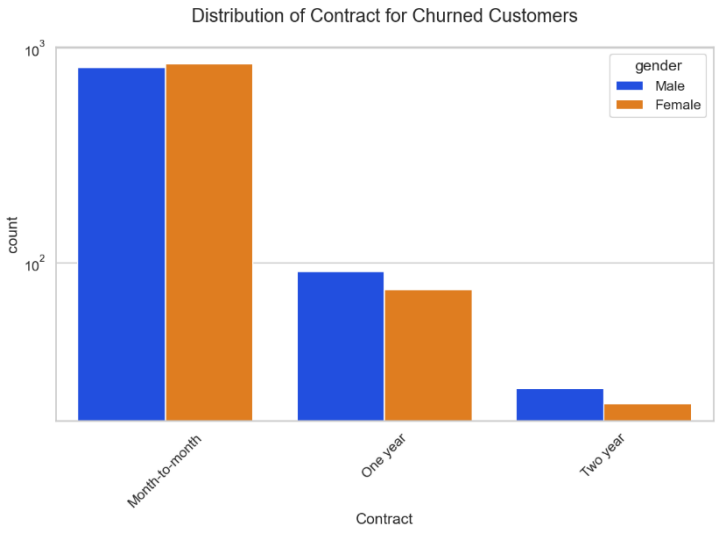
Generated visual representations, such as count plots, to illustrate the distribution of churn and non-churn instances for pairs of variables.

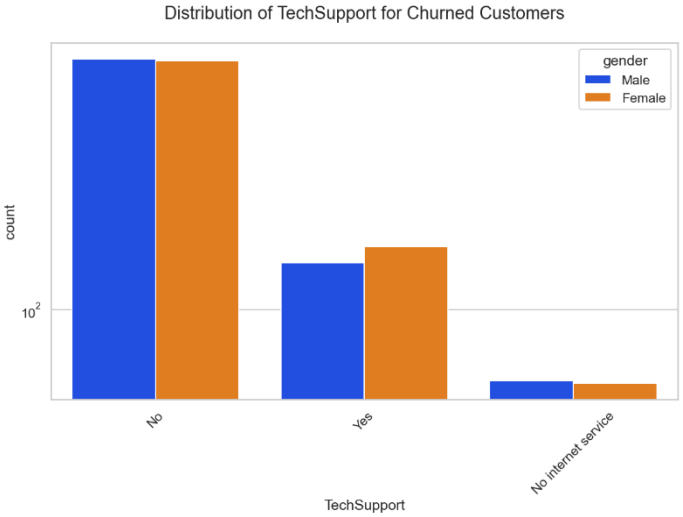
Concise Interpretation:

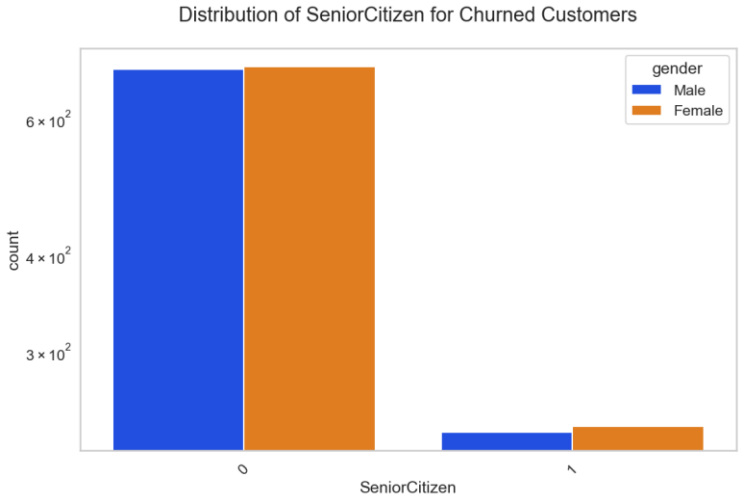
Bivariate analysis revealed nuanced relationships, aiding in the identification of specific factors that jointly influence customer churn. This insight guides targeted retention efforts.











Key Insights Summary

1. **Electronic Check Usage:**
    - Customers using electronic checks have the highest churn rates. Consider diversifying payment options.
2. **Monthly Contract Impact:**
    - Monthly contract subscribers are more likely to churn. Focus on converting them to longer-term contracts.

3. **Online Security and Tech Support:**
    - Lack of online security and tech support correlates with high churn. Enhance these services for improved customer satisfaction.
4. **Senior Citizen Segment:**
    - Non-senior citizens exhibit higher churn rates. Tailor retention efforts to address their specific needs.

Strategic Actions

These insights guide strategies for payment options, contract structures, service enhancements, and targeted retention efforts to mitigate churn effectively.

Prediction Model

Decision Tree Classifier

Implemented a Decision Tree Classifier with parameters tuned for optimal performance. Evaluated the model on the original dataset, providing classification metrics.

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=6, min_samples_leaf=8, random_state=100)
```

```
print(classification_report(y_test, y_pred, labels=[0,1]))
```

```
              precision    recall  f1-score   support

           0       0.84      0.90      0.87      1016
           1       0.69      0.55      0.61       391

    accuracy                           0.81      1407
   macro avg       0.77      0.73      0.74      1407
weighted avg       0.80      0.81      0.80      1407
```

Handling Imbalance with SMOTEENN:

Addressed class imbalance using the SMOTEENN (SMOTE + Edited Nearest Neighbours) technique. Resampled the dataset and retrained the Decision Tree Classifier.

```
print(classification_report(yr_test, y_pred_smote, labels=[0,1]))
```

	precision	recall	f1-score	support
0	0.93	0.94	0.94	532
1	0.95	0.94	0.95	647
accuracy			0.94	1179
macro avg	0.94	0.94	0.94	1179
weighted avg	0.94	0.94	0.94	1179

Random Forest Classifier:

Applied a Random Forest Classifier with tuned parameters on the original dataset. Assessed the model's performance using classification metrics.

```
print(classification_report(y_test, y_pred_rf, labels=[0,1]))
```

	precision	recall	f1-score	support
0	0.83	0.93	0.88	1016
1	0.74	0.49	0.59	391
accuracy			0.81	1407
macro avg	0.78	0.71	0.73	1407
weighted avg	0.80	0.81	0.80	1407

Handling Imbalance for Random Forest:

Employed SMOTEENN for balancing the classes in the Random Forest model. Trained the model on the resampled dataset.

```
print(classification_report(yr_test, y_pred_smote_rf, labels=[0,1]))
```

	precision	recall	f1-score	support
0	0.95	0.91	0.93	491
1	0.94	0.97	0.95	688
accuracy			0.94	1179
macro avg	0.94	0.94	0.94	1179
weighted avg	0.94	0.94	0.94	1179

Model Evaluation:

- Evaluated all models using metrics such as precision, recall, and F1-score.
- Confusion matrices provided insights into true positive, true negative, false positive, and false negative predictions.

Model Comparison:

- Compared the performance of Decision Tree and Random Forest models on both the original and resampled datasets.

Model Saving and Loading:

- Saved the best-performing model (Random Forest with SMOTEENN) using pickle for future use.
- Loaded the saved model to confirm its accuracy on the test set.

Conclusion

- The predictive models showcase promising performance in identifying potential customer churn.
- The Random Forest model with SMOTEENN stands out as an effective solution for handling class imbalance and improving model accuracy.

This report provides a comprehensive overview of the churn prediction model, its implementation, and evaluation. The models exhibit potential for assisting the telecommunications company in proactively retaining customers at risk of churning.