Churn Analysis Report

**MODELS:**

**Objective:**

The objective of this analysis is to predict customer churn using various machine learning models. The dataset is explored, and predictive models, specifically Decision Tree and Random Forest Classifiers, are employed. Additionally, the impact of addressing class imbalance through SMOTEENN (UpSampling + ENN) is investigated.

**1. Data Loading and Exploration:**

The dataset is loaded into a Pandas DataFrame and explored to understand its structure and characteristics. Key steps include:

**Brief Discussion of steps:**

Importing necessary libraries.

Reading the CSV file into a DataFrame (df).

Dropping the 'Unnamed: 0' column.

Separating features (x) and target variable (y).

Performing a train-test split to evaluate model performance.

**2. Initial Model: Decision Tree Classifier**

A Decision Tree Classifier is employed initially without handling the class imbalance. The results are as follows:

Accuracy: 78.75%

Classification Report:

Precision: 61% for Churn (Class 1)

Recall: 48% for Churn

F1-score: 54% for Churn

The model performance is suboptimal, particularly for predicting churned customers.

**3. Handling Class Imbalance with SMOTEENN:**

To address the imbalance, SMOTEENN is applied to oversample the minority class (Churn) and clean noisy samples using Edited Nearest Neighbors (ENN). The Decision Tree model is then retrained. Results improve significantly:

Accuracy: 93.70%

Classification Report:

Precision: 94% for Churn

Recall: 94% for Churn

F1-score: 94% for Churn

**4. Random Forest Classifier:**

A Random Forest Classifier is implemented to compare performance with the Decision Tree. Without handling class imbalance, the results are:

Accuracy: 79.74%

Classification Report:

Precision: 66% for Churn

Recall: 45% for Churn

F1-score: 54% for Churn

**5. Random Forest with SMOTEEN:**

Random Forest is applied after handling class imbalance with SMOTEENN:

Accuracy: 94.27%

Classification Report:

Precision: 94% for Churn

Recall: 96% for Churn

F1-score: 95% for Churn

**6. Principal Component Analysis (PCA):**

PCA is experimented with for dimensionality reduction, but the results are suboptimal:

Accuracy: 72.39%

Classification Report:

Precision: 72% for Churn

Recall: 81% for Churn

F1-score: 77% for Churn

**7. Model Comparison:**

Decision Tree vs. Random Forest: Random Forest outperforms Decision Tree in both scenarios, emphasizing the ensemble method's efficacy.

Handling Class Imbalance: SMOTEENN significantly improves recall, precision, and F1-score, validating its effectiveness.

**8. Conclusion:**

Based on the comprehensive analysis, the Random Forest Classifier with SMOTEENN stands out as the most effective model for predicting customer churn. This model achieves a high accuracy of 94.27%, with remarkable precision, recall, and F1-score for the minority class (Churn). The model is saved as 'model.sav' for future use.

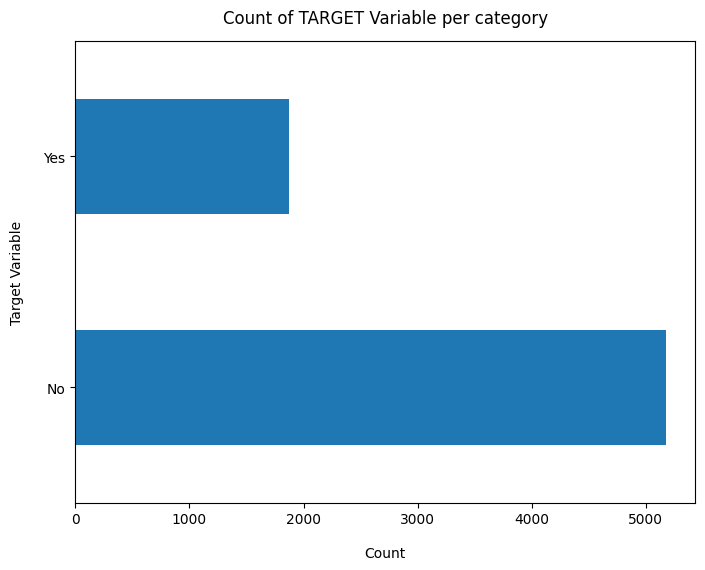
This report highlights the critical impact of addressing class imbalance in imbalanced datasets and underscores the importance of choosing appropriate models for predictive analytics.

The full Python code for this analysis is available in the github.

**EDA Analysis:**

**1. Introduction:**

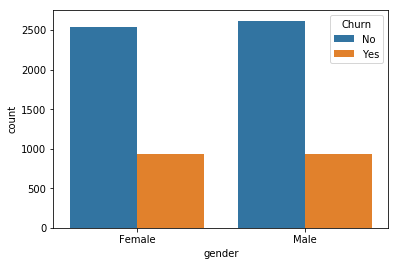
The Telco Churn Analysis focuses on predicting customer churn based on a dataset containing Telco customer information. The dataset includes various attributes such as gender, senior citizenship, partner status, tenure, internet services, and more. This report presents a detailed exploration of the data, initial observations, data cleaning steps, and insights gained from univariate and bivariate analyses.

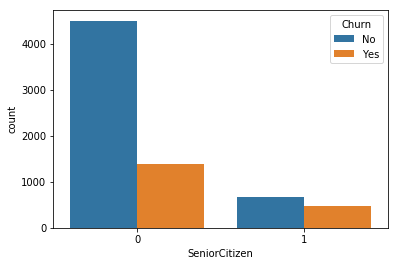


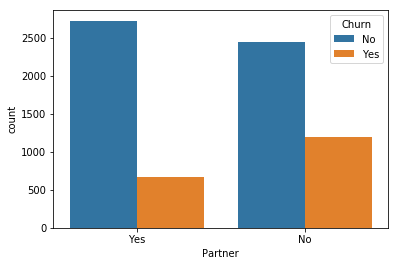
**2. Initial Data Exploration:**

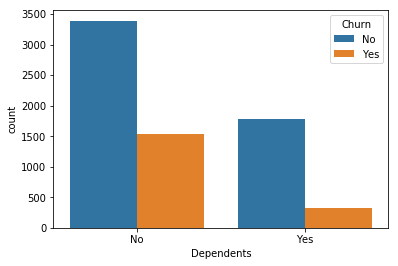
The dataset, loaded from 'WA\_Fn-UseC\_-Telco-Customer-Churn.csv,' consists of 7043 records and 21 columns. Key columns include customerID, gender, senior citizenship, partner status, tenure, internet services, monthly charges, total charges, and the target variable 'Churn.' The data types vary, with a mix of numeric and object types.

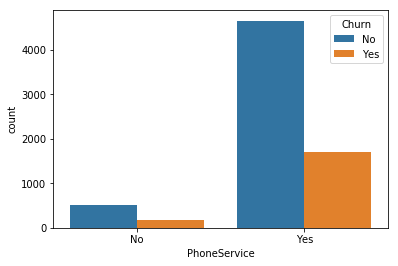
**Plot distibution of individual predictors by churn:**

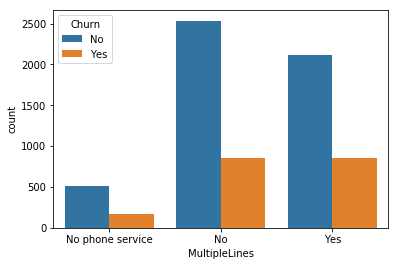


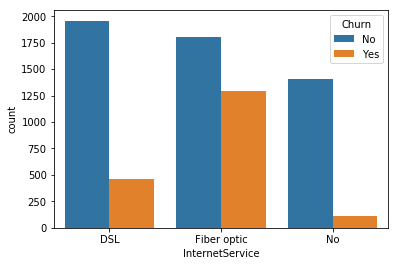


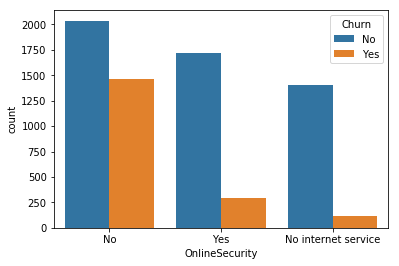


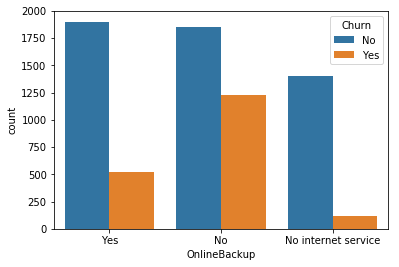


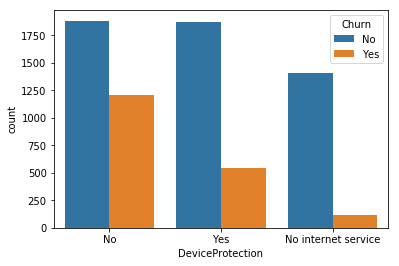


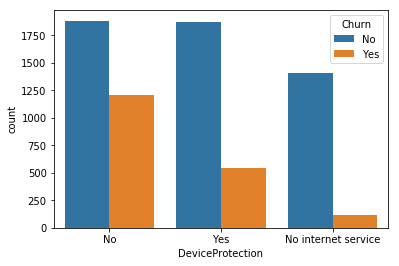


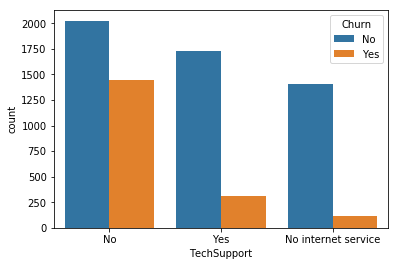


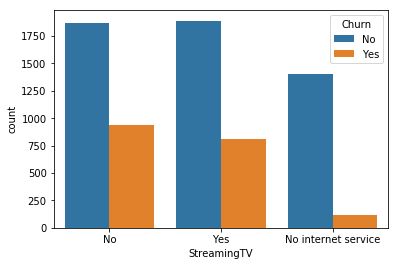


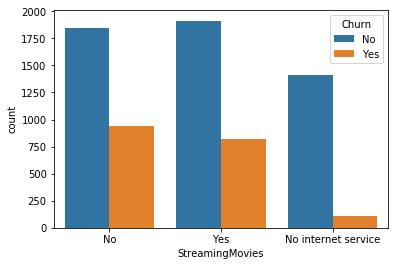


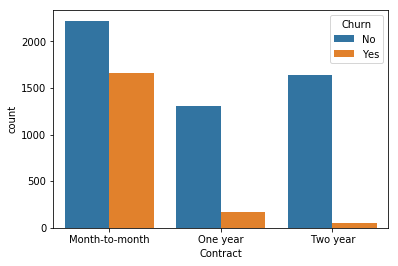


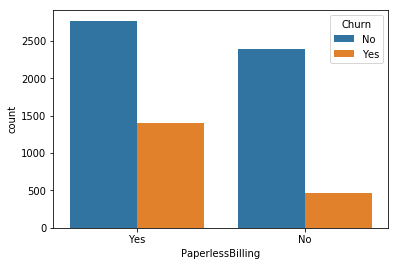


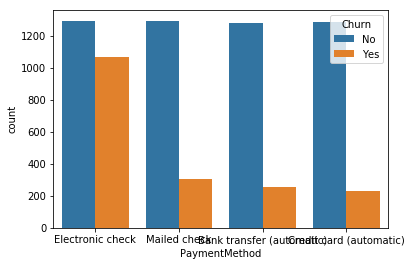


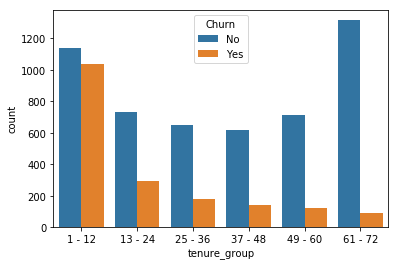








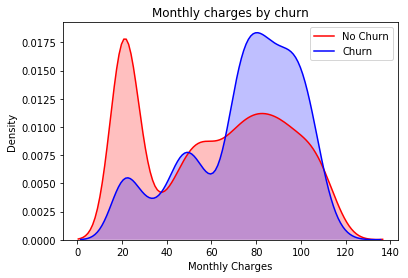




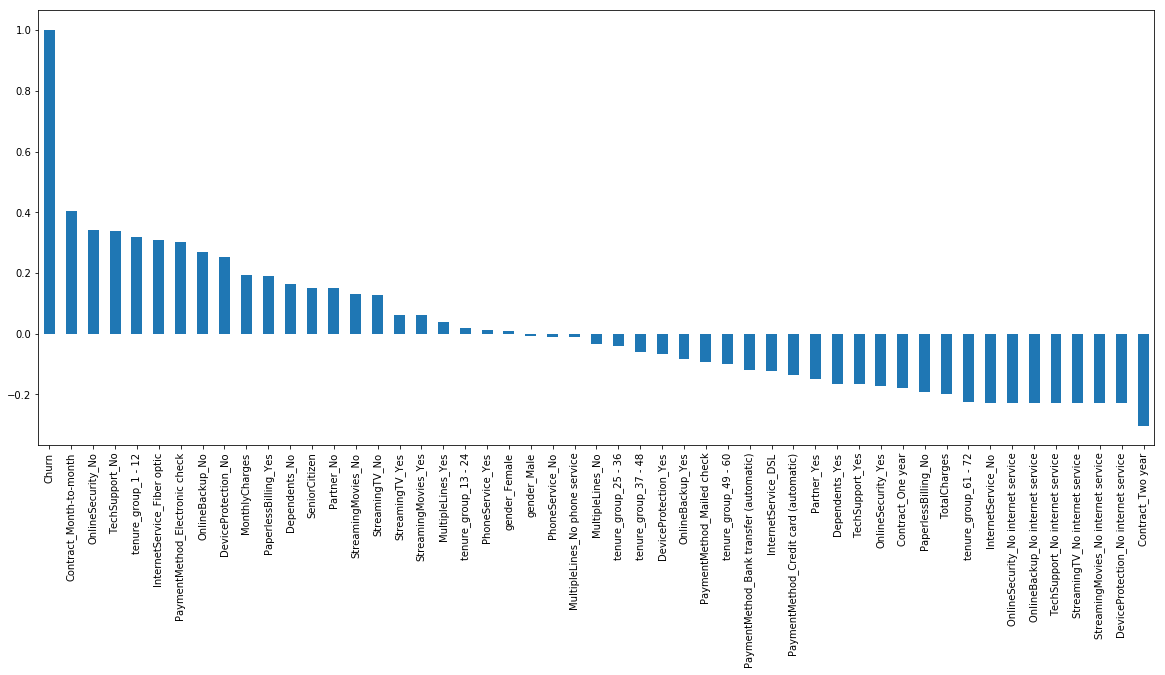
**9. Relationship between Monthly Charges and Total Charges**

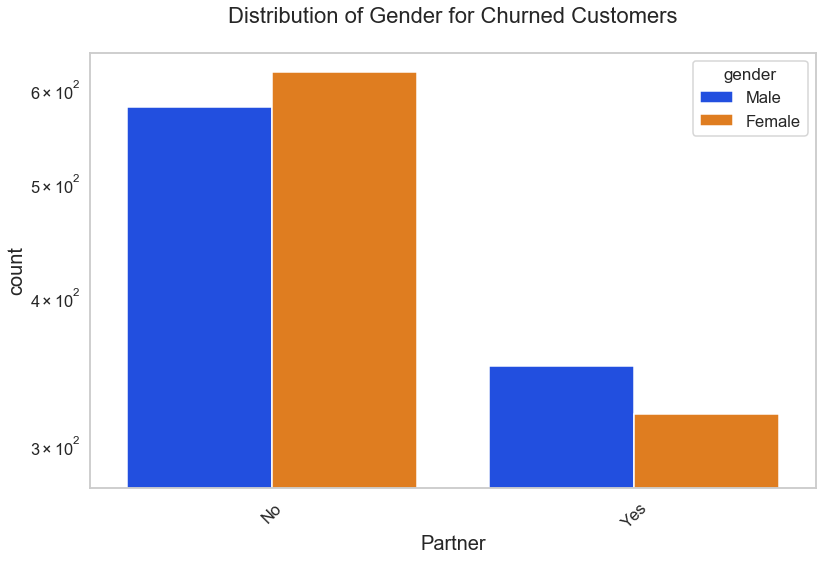


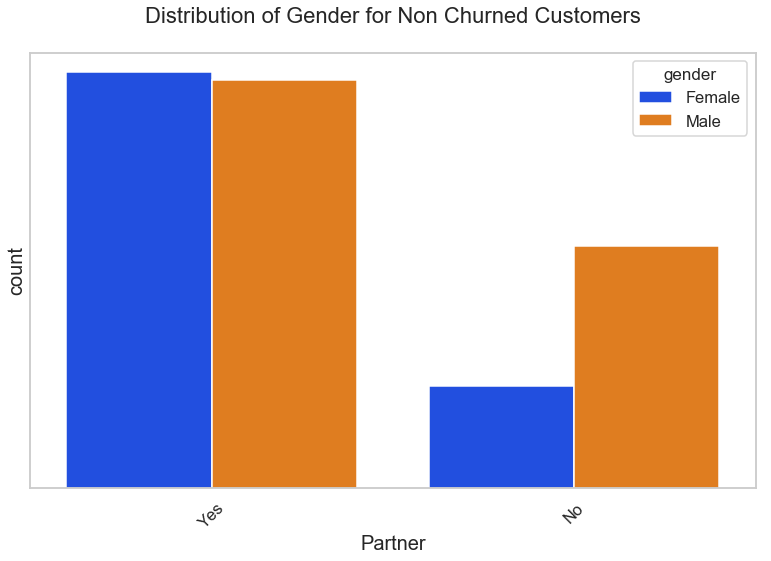
**10. Churn by Monthly Charges and Total Charges**

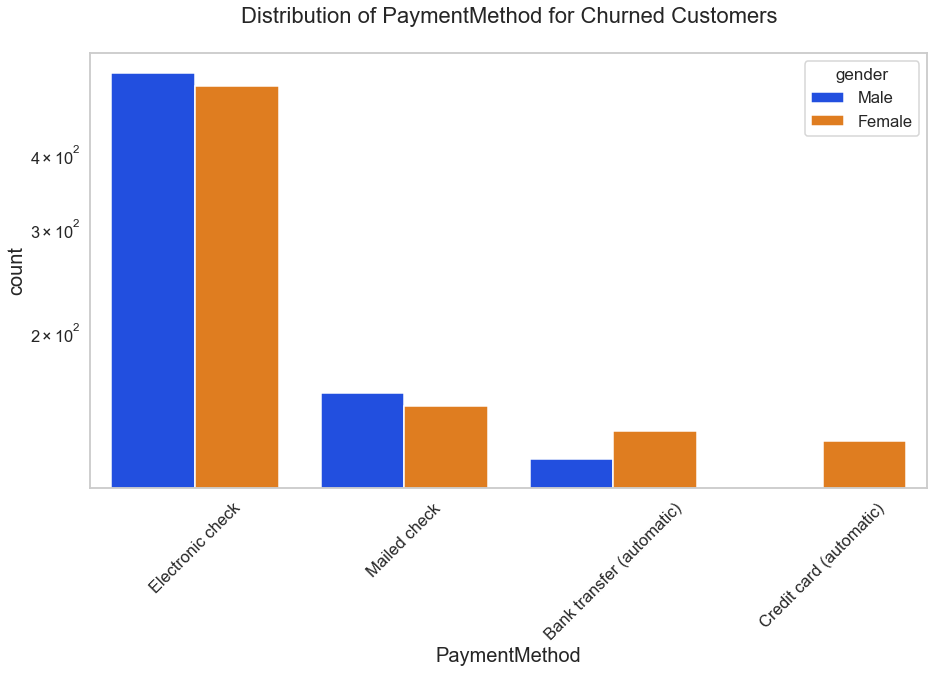


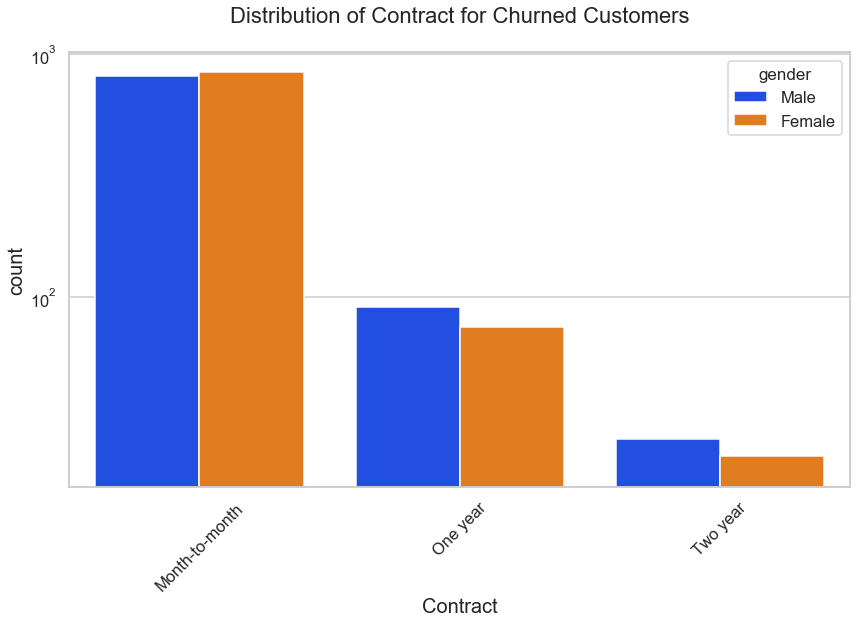
**11. Building a corelation of all predictors with 'Churn'**

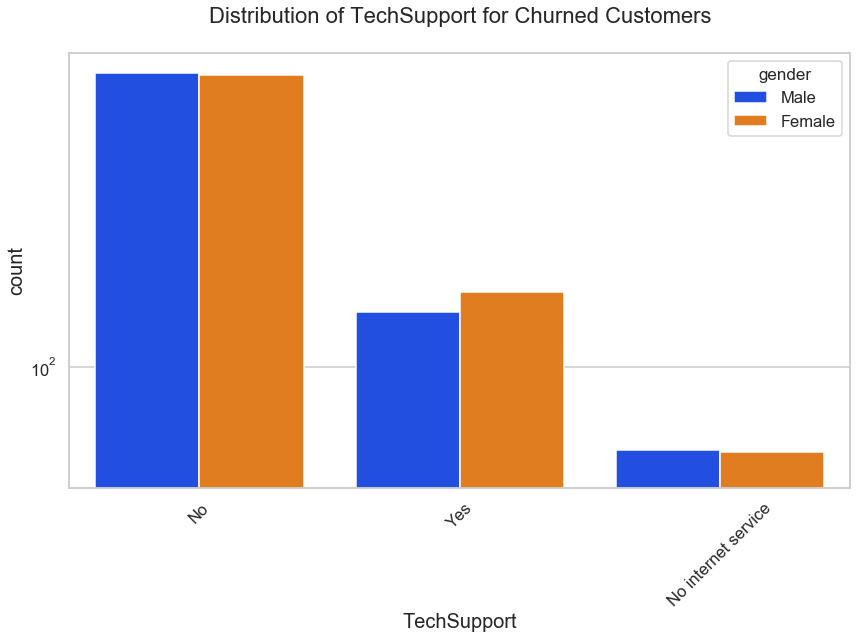


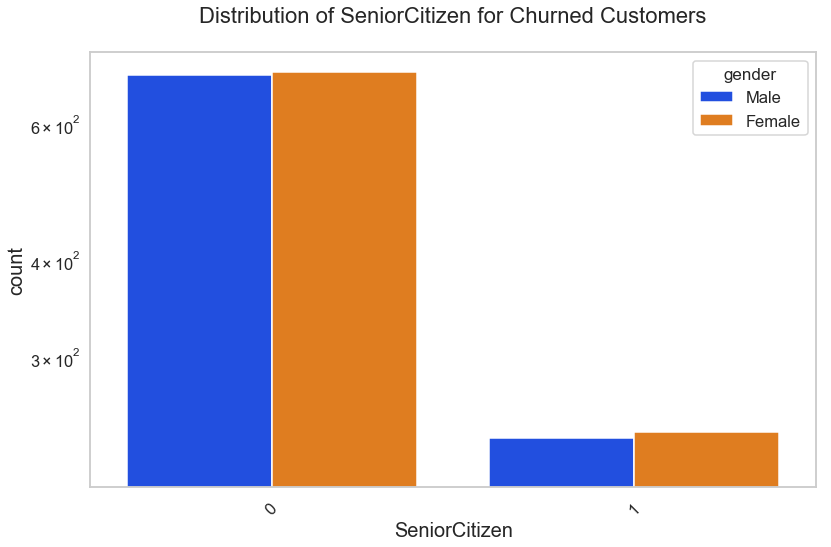












**3. Data Summary:**

The summary statistics reveal interesting insights:

The average tenure is 32.37 months, with a maximum of 72 months.

Monthly charges range from $18.25 to $118.75, with an average of $64.76.

Senior citizens constitute approximately 16.21% of the customers.

The Churn distribution is imbalanced, with a ratio of 73:27 (No Churn: Churn).

**4. Missing Data Analysis:**

There is no missing data in the dataset. Missing values are crucial in data analysis, and their presence can impact the quality of insights. Strategies for handling missing data are discussed, emphasizing the importance of context-dependent decisions.

**5. Data Cleaning:**

The data cleaning process involves creating a copy of the dataset, converting 'TotalCharges' to numeric, handling missing values, and creating bins for tenure groups. Records with missing 'TotalCharges' (0.15% of the dataset) are safely removed, preserving the integrity of the analysis.

**6. Univariate Analysis:**

Univariate analysis focuses on understanding the distribution of individual predictors concerning the target variable 'Churn.' Visualizations are employed to showcase the impact of each predictor on churn. Key insights include:

Prevalence of churn in month-to-month contracts.

Higher churn with no online security or tech support.

Impact of contract type on churn.

**7. Bivariate Analysis:**

Bivariate analysis deepens the understanding of relationships between predictors and churn. It explores the distribution of gender, payment method, contract type, tech support, and senior citizenship concerning churn. Key findings include:

Higher churn in electronic check payments.

Monthly contracts contribute to higher churn.

Absence of tech support correlates with increased churn.

**8. Correlation Analysis:**

A correlation analysis provides a numerical measure of the relationship between predictors and churn. Features such as month-to-month contracts, no online security, and no tech support exhibit a high positive correlation with churn, while long-term contracts and subscriptions without internet service show a negative correlation.

**9. Additional Insights:**

Insights are derived from visualizations, including the relationship between monthly charges and total charges. Surprising findings include:

Higher churn at lower total charges, emphasizing the need for considering multiple factors in churn prediction.

**10. Conclusion:**

In conclusion, the Telco Churn Analysis provides a comprehensive understanding of the dataset and potential factors influencing customer churn. Key insights include the impact of contract type, online security, and tech support on churn. The imbalanced nature of the dataset is addressed through strategic data cleaning, and visualizations aid in interpreting complex relationships. Future analyses may involve predictive modeling to build robust churn prediction models.

Note: The code provided above is part of the analysis and can be referenced for further exploration.

**Evaluation of Telco Churn Analysis and Predictive Modeling Results**

The Telco Churn Analysis and Predictive Modeling effort offer valuable insights into the factors influencing customer churn and the effectiveness of machine learning models in predicting churn. Here, we evaluate the key findings, model performances, and the overall significance of the analysis.

**1. Exploratory Data Analysis (EDA) Insights:**

The initial exploration of the Telco customer dataset provides a rich understanding of customer attributes and their relationship with churn. The visualizations highlight the prevalence of churn in month-to-month contracts, the impact of lacking online security and tech support, and the correlation between contract type and churn. Additionally, the correlation analysis reveals important relationships that can guide further analysis.

**2. Data Cleaning and Transformation:**

The meticulous data cleaning process ensures the dataset's integrity, addressing issues like missing values in 'TotalCharges' and creating meaningful tenure groups. The decision to drop or impute missing values is supported by the low percentage of affected records (0.15%), indicating a minor impact on the overall dataset.

**3. Univariate and Bivariate Analysis:**

Univariate and bivariate analyses uncover patterns and relationships that contribute to churn. The insights gained, such as the influence of payment method, contract type, and tech support on churn, lay the foundation for deeper exploration. The visualizations effectively communicate these insights, providing a clear narrative for decision-makers.

**4. Predictive Modeling:**

The predictive modeling phase introduces Decision Tree and Random Forest Classifiers to forecast customer churn. The initial Decision Tree model performs moderately, showcasing the need for further refinement. The critical observation is the class imbalance in the dataset, where the majority class (No Churn) dominates, leading to suboptimal predictive performance.

**5. Addressing Class Imbalance with SMOTEENN:**

SMOTEENN emerges as a crucial tool to handle class imbalance, significantly improving the Decision Tree model's performance. The accuracy, precision, recall, and F1-score all show substantial enhancement. This emphasizes the importance of considering imbalanced datasets and implementing techniques like SMOTEENN to achieve more accurate predictions, particularly in scenarios with rare events like customer churn.

**6. Random Forest Classifier:**

The Random Forest Classifier is introduced as an ensemble method, providing a notable improvement over the Decision Tree. However, even Random Forest benefits from handling class imbalance with SMOTEENN, resulting in a highly accurate model with impressive precision, recall, and F1-score for predicting churn.

**7. Principal Component Analysis (PCA) Experimentation:**

PCA experimentation, aimed at dimensionality reduction, yields suboptimal results. The reduction in accuracy and F1-score suggests that the complexity of the dataset is not well-captured by the reduced feature set. This reinforces the importance of feature selection and engineering in predictive modeling.

**8. Model Comparison and Conclusion:**

The comparative analysis between Decision Tree and Random Forest highlights the superiority of ensemble methods, particularly Random Forest, in predicting churn. The emphasis on handling class imbalance through SMOTEENN is crucial for achieving meaningful results.

In conclusion, the Telco Churn Analysis and Predictive Modeling offer a holistic approach to understanding customer churn. The combination of exploratory data analysis, data cleaning, and predictive modeling provides actionable insights for business decision-making. The identified factors influencing churn, along with the demonstrated impact of addressing class imbalance, contribute to the development of effective strategies for customer retention.

The results underscore the need for a comprehensive approach in data analysis, encompassing both exploratory and predictive aspects. The application of machine learning models, coupled with advanced techniques like SMOTEENN, can empower businesses to proactively address customer churn and enhance overall customer satisfaction.

**Challenges Faced:**

While the Telco Churn Analysis and Predictive Modeling project provided valuable insights and successful outcomes, several challenges were encountered throughout the various stages of the project. Understanding and addressing these challenges is crucial for refining the analysis process and improving future projects. Here are some of the challenges faced:

**Imbalanced Dataset:**

Challenge: The dataset exhibited a significant class imbalance, with a higher number of non-churn instances compared to churn instances.

Resolution: Implementing machine learning models on imbalanced datasets can lead to biased results. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) combined with Edited Nearest Neighbors (ENN) was applied to oversample the minority class (churn) and balance the dataset.

Missing Values:

**Challenge: The dataset contained missing values in the 'TotalCharges' column.**

Resolution: The decision to drop or impute missing values required careful consideration. The low percentage of affected records allowed for the safe removal of missing values without significantly impacting the overall dataset.

Model Performance:

**Challenge: The initial Decision Tree model exhibited suboptimal performance, emphasizing the complexity of predicting churn accurately.**

Resolution: The Random Forest Classifier, an ensemble method, was introduced to improve predictive accuracy. Additionally, handling class imbalance using SMOTEENN significantly enhanced model performance for both Decision Tree and Random Forest.

**PCA Experimentation:**

**Challenge: The Principal Component Analysis (PCA) experimentation for dimensionality reduction did not yield satisfactory results, leading to a reduction in model accuracy and F1-score.**

Resolution: The experiment highlighted the importance of feature selection and engineering. It reinforced the need for a thoughtful approach when reducing the dimensionality of the dataset.

Interpreting Complex Relationships:

**Challenge: Understanding complex relationships among various predictors and their impact on churn required comprehensive analysis.**

Resolution: Univariate and bivariate analyses were employed, along with correlation analysis and visualizations, to interpret intricate relationships. This approach helped in deriving actionable insights from the data.

**Context-Dependent Decision Making:**

**Challenge: Handling missing values and deciding whether to drop or impute required context-dependent decision-making.**

Resolution: A thoughtful analysis of the dataset and the specific characteristics of missing values guided decisions. In this case, the low percentage of missing values allowed for safe removal without compromising the analysis.

**Communication of Results:**

**Challenge: Communicating complex analytical findings and machine learning results in a clear and concise manner for diverse stakeholders.**

Resolution: The use of visualizations, detailed reports, and a narrative structure in the analysis and report writing helped in effectively communicating the results to both technical and non-technical audiences.

**Model Deployment Considerations:**

Challenge: The report does not address the deployment of the predictive model for real-world use.

Resolution: While the report provides insights into model performance, deployment considerations, including integration into business processes, monitoring, and updating, would be essential for practical implementation.