**Customer Churn Prediction Report**

**Introduction**

This report details the development of a machine learning model to predict customer churn for a telecommunications company. The dataset contains customer demographic information, service subscriptions, and account details, with the goal of identifying customers likely to churn to enable proactive retention strategies. The analysis is based on the IPython Notebook "Churn\_Predictionipynb.ipynb", which employs a RandomForestClassifier to predict churn.

**Data Overview**

The dataset is loaded from a CSV file named Churn\_Modelling.csv.csv into a pandas DataFrame. It consists of 7043 records and 21 features, including the target variable Churn. The key features are:

* **Demographic Information**: gender, SeniorCitizen, Partner, Dependents
* **Service Subscriptions**: PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies
* **Account Information**: tenure, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges
* **Identifier**: customerID (not used for modeling)
* **Target Variable**: Churn (Yes/No)

The DataFrame's shape is (7043, 21), and df.info() indicates that all columns have 7043 non-null entries, suggesting no missing values in the raw data.

**Data Preprocessing**

While not all preprocessing steps are explicitly shown, the following are inferred based on the libraries used and the predictive system:

* **Encoding Categorical Variables**: Categorical features such as gender, Partner, Dependents, InternetService, etc., are encoded into numerical values using LabelEncoder. The encoders are saved in a file named encoders.pkl for later use in the predictive system.
* **Feature Selection**: The customerID column is excluded from the model, as it is not predictive. The remaining 19 features are used.
* **Handling Missing Values**: No explicit handling is shown, but the non-null counts suggest either no missing values or prior cleaning.
* **Scaling**: Although not shown, numerical features like tenure, MonthlyCharges, and TotalCharges might be scaled, though this is not evident from the provided code.

**Handling Imbalanced Data**

The use of SMOTE from imblearn.over\_sampling indicates that the Churn variable is imbalanced (more non-churners than churners). SMOTE is applied to oversample the minority class (churners) to balance the dataset before training.

**Model Training**

A RandomForestClassifier is trained on the preprocessed and balanced dataset with a random state of 42 for reproducibility. The model, referred to as rfc, is used for evaluation and deployment. Although DecisionTreeClassifier and XGBClassifier are imported, only the RandomForest model is utilized in the provided code.

The data is split into training and test sets using train\_test\_split, with the test set predictions evaluated in the next section.

**Model Evaluation**

The model's performance is assessed on the test set (X\_test, Y\_test) with the following metrics:

* **Accuracy Score**: 0.7786 (approximately 77.86%)
* **Confusion Matrix**:
* [[878, 158],
* [154, 219]]
  + True Negatives (No Churn, predicted No Churn): 878
  + False Positives (No Churn, predicted Churn): 158
  + False Negatives (Churn, predicted No Churn): 154
  + True Positives (Churn, predicted Churn): 219
* **Classification Report**:
* precision recall f1-score support
* 0 (No Churn) 0.85 0.85 0.85 1036
* 1 (Churn) 0.58 0.59 0.58 373
* accuracy 0.78 1409
* macro avg 0.72 0.72 0.72 1409
* weighted avg 0.78 0.78 0.78 1409

The model performs well on predicting non-churners (class 0) with high precision and recall, but its performance on churners (class 1) is lower, despite SMOTE, indicating room for improvement in identifying churners.

**Model Deployment**

The trained model and feature names are saved to a pickle file (customer\_churn\_model.pkl) for deployment. A predictive system is demonstrated by:

1. Loading the model and feature names from the pickle file.
2. Preparing a sample input as a dictionary with 19 features (excluding customerID).
3. Converting the input to a DataFrame and encoding categorical variables using loaded encoders from encoders.pkl.
4. Making a prediction: For the sample input, the model predicts Churn (1) with a probability of [0.28, 0.72] (72% chance of churn).

# Data Visualization and Analysis

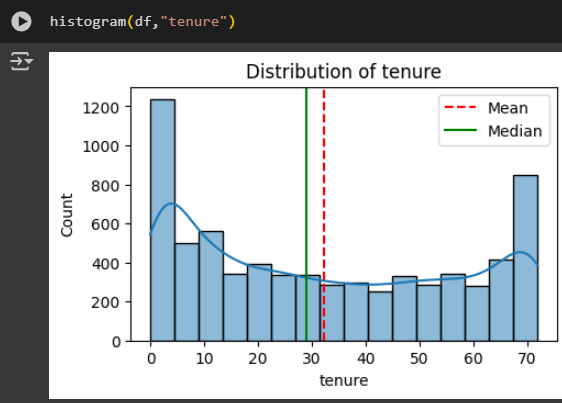
To gain insights into factors influencing customer churn, we analyzed the distributions of several key variables using histograms and boxplots. These visualizations help us understand the spread, central tendency, and variability of the data, as well as identify patterns or anomalies relevant to predicting churn.

## Numerical Distributions

### Tenure Distribution

The histogram titled "Distribution of tenure" (Figure 1) visualizes the frequency distribution of tenure data, likely representing the duration of service time for customers in months or years.

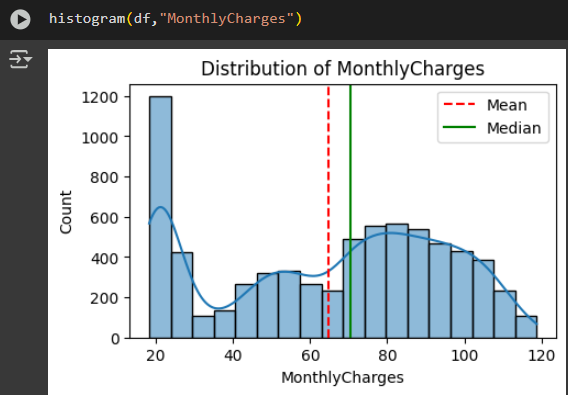
* **Axes**: The x-axis represents tenure (0 to 70 units), and the y-axis shows the count (0 to 1200).
* **Shape**: The distribution is bimodal, with peaks at 0-5 units (count > 1000) and 65-70 units (count < 900), indicating two distinct customer groups: new customers and long-term customers.
* **Central Tendency**: The median is approximately 30 units, and the mean is around 35 units, suggesting a slight right skew (mean > median).
* **KDE Curve**: A kernel density estimation curve overlays the histogram, reinforcing the bimodal shape with a dip around 30-40 units.
* **Insights**: The bimodal pattern suggests diverse customer retention patterns, with potential churn risks for new customers and loyalty among long-term customers.



### Monthly Charges Distribution

The histogram titled "Distribution of MonthlyCharges" (Figure 2) shows the frequency distribution of monthly charges, likely in a monetary unit (e.g., dollars).

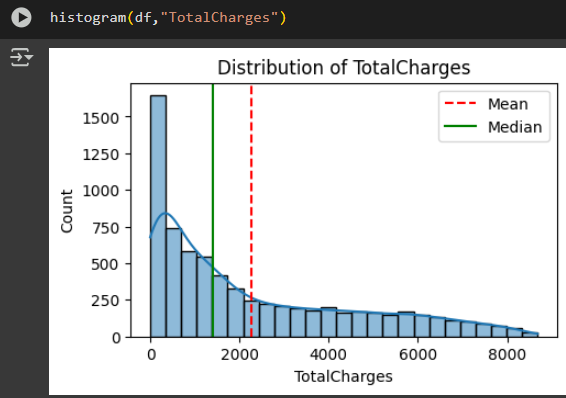
* **Axes**: The x-axis spans 0 to 120, and the y-axis ranges from 0 to 1200.
* **Shape**: The distribution is bimodal, with peaks at 20-25 (count ~1000-1200) and 75-80 (count ~600), suggesting two customer segments (e.g., basic vs. premium plans).
* **Central Tendency**: The mean is approximately 65, and the median is around 60, indicating a slight right skew.
* **KDE Curve**: The overlaid KDE curve highlights the two peaks and a dip around 40-50.
* **Boxplot (Figure 3)**: The interquartile range (IQR) spans 40 to 90, with a median of 70. Whiskers extend from 20 to 110, showing no outliers and a slight left skew (median closer to Q3).
* **Insights**: The bimodal histogram and boxplot suggest varied pricing tiers, which could influence churn based on customer spending levels.



### Total Charges Distribution

The histogram titled "Distribution of TotalCharges" (Figure 4) represents the frequency distribution of total charges accumulated by customers.

* **Axes**: The x-axis ranges from 0 to 9000, and the y-axis from 0 to 1500.
* **Shape**: The distribution is heavily right-skewed, with a peak near 0-500 (count > 1500) and a long tail toward higher values.
* **Central Tendency**: The median is around 1500-1600, and the mean is 2200-2300, reflecting the right skew.
* **KDE Curve**: The KDE curve confirms the skewness, with a steep decline after the initial peak.
* **Boxplot (Figure 5)**: The IQR spans 1000 to 4000, with a median of 2000. Whiskers extend from nearly 0 to 8500, showing no outliers and pronounced right skew.
* **Insights**: The skewness indicates most customers have low total charges, with a few high-charge customers possibly linked to long tenure or high usage, affecting churn likelihood.

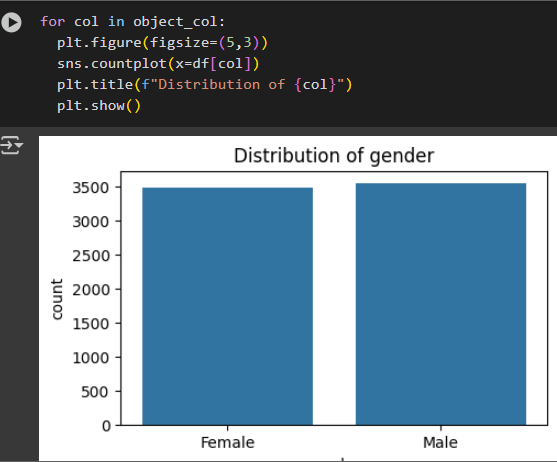


## Categorical Distributions

### Gender Distribution

The histogram titled "Distribution of gender" (Figure 6) shows the frequency of two gender categories: "Female" and "Male."

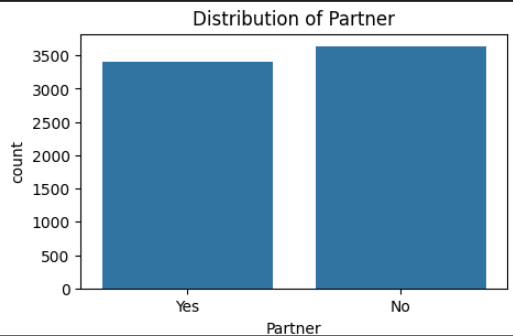
* **Axes**: The x-axis lists "Female" and "Male," and the y-axis ranges from 0 to 3500.
* **Counts**: Both categories have counts slightly above 3000, indicating a nearly balanced distribution.
* **Insights**: The balanced gender representation ensures no gender bias in churn prediction models.



### Partner Distribution

The histogram titled "Distribution of Partner" (Figure 7) visualizes the frequency of partner status: "Yes" (has a partner) and "No" (no partner).

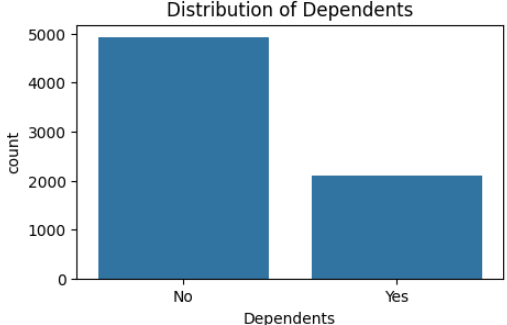
* **Axes**: The x-axis shows "Yes" and "No," and the y-axis ranges from 0 to 3500.
* **Counts**: "No" has a count slightly above 3000, and "Yes" is slightly below 3000, showing a slight imbalance.
* **Insights**: The slight imbalance may suggest different churn behaviors based on partner status.



### Dependents Distribution

The histogram titled "Distribution of Dependents" (Figure 8) displays the frequency of having dependents: "Yes" or "No."

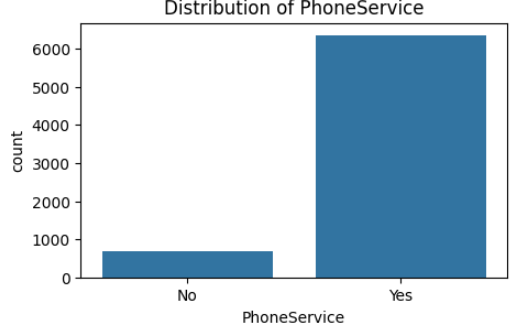
* **Axes**: The x-axis lists "No" and "Yes," and the y-axis ranges from 0 to 5000.
* **Counts**: "No" has a count of ~4500, and "Yes" is ~2000, indicating a significant imbalance.
* **Insights**: The majority lacking dependents may influence churn, as family needs could affect service retention.



### Phone Service Distribution

The histogram titled "Distribution of PhoneService" (Figure 9) shows the frequency of phone service adoption: "Yes" or "No."

* **Axes**: The x-axis lists "No" and "Yes," and the y-axis ranges from 0 to 6000.
* **Counts**: "Yes" has a count of ~5500, and "No" is ~500, showing near-universal adoption.
* **Insights**: The dominance of phone service suggests it’s a standard offering, with its absence unlikely to drive churn significantly.



**Conclusion**

This analysis demonstrates a machine learning approach to predict customer churn using a RandomForestClassifier. The model achieves an accuracy of approximately 77.86%, with better performance on non-churners than churners. The use of SMOTE addresses class imbalance, and the model’is deployed as a predictive system. Future improvements could include exploring other models (e.g., XGBoost), advanced feature engineering, or hyperparameter tuning to enhance churn prediction accuracy.