Customer Churn Prediction Report

1. Introduction

The project predicts customer churn. The dataset used for this analysis includes customer information such as call details, account information, and churn status. The goal is to identify patterns and build a predictive model to determine whether a customer is likely to churn. Did you know that attracting a new customer costs five times as much as keeping an existing one? We want to minimize the churn rate by making a good predictor for this case.

2. Approach

2.1 Data Preprocessing

- **Data Loading**: The dataset was split into training/validation **80%** and testing **20%** sets.
- **Encoding Categorical Variables**: Binary categorical variables like `International plan`, `Voice mail plan`, and `Churn` were encoded as 1/0. The `State` and `Area code` columns were label-encoded for numerical representation.
- **Handling Missing Values**: No missing values were found in the dataset.
- **Outlier Treatment**: Outliers in numerical features like `Total day minutes`, `Total eve minutes`, and `Total night minutes` were handled by dropping them using the Interquartile Range (IQR) method, that will help in final value of recall and accuracy.
- **Feature Scaling**: Numerical features were standardized using `StandardScaler` to ensure consistent scaling across the dataset.

2.2 Exploratory Data Analysis (EDA)

- **Class Imbalance**: The dataset was highly imbalanced, with most customers not churning. SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the classes.
- **Correlation Analysis**: A correlation heatmap was generated to identify relationships between features and the target variable (`Churn`). Features like `Total day charge` and `Customer service calls` showed moderate correlation with churn.
- **Visualizations**: Histograms and boxplots were used to understand the distribution of numerical features and detect outliers.

2.3 Model Selection

- **Algorithm**: A Random Forest Classifier was chosen due to its ability to handle non-linear relationships and feature importance interpretation.
- **Training**: The model was trained on the resampled training data (after applying SMOTE for oversampling the minority class).
- **Validation**: The model was evaluated on a validation select the optimal threshold for churn prediction to increase the recall value of 0 class that will help in predict correctly.

2.4 Evaluation Metrics

- **Accuracy**: Measures the overall correctness of the model.
- **Precision-Recall Curve**: Used to evaluate the trade-off between precision and recall, especially for imbalanced datasets.
- **Confusion Matrix**: Provides insights into true positives, false positives, true negatives, and false negatives.

3. Findings

3.1 Model Performance

- **Validation Accuracy**: The model achieved an accuracy of **94.5%** on the validation set.
- **Test Accuracy**: The model achieved an accuracy of **95.6%** on the test set.
- **Classification Report**:
  - **Non-Churn (Class 0):**
    - **Precision: 0.954**
    - **Recall: 0.996**
    - **F1-Score: 0.975**
  - **Churn (Class 1):**
    - **Precision: 0.971**
    - **Recall: 0.715**
    - **F1-Score: 0.824**
- **Test Confusion Matrix:**
  - **True Positives (TP): 68** (correctly predicted churn)
  - **False Positives (FP): 2** (incorrectly predicted churn)
  - **True Negatives (TN): 570** (correctly predicted non-churn)
  - **False Negatives (FN): 27** (incorrectly predicted non-churn)

3.2 Key Insights

- **Feature Importance**: Features like `Total day charge`, `Customer service calls`, and `International plan` were found to be significant predictors of churn.
- **Class Imbalance**: The original dataset was highly imbalanced, with only **14%** of customers churning. SMOTE helped improve the model's ability to predict the minority class.
- **Threshold Optimization**: A threshold of **0.4** was used to increase recall, ensuring that more churn cases were correctly identified.

4. Conclusion

The Random Forest model performed well in predicting customer churn, achieving high accuracy and balanced precision-recall metrics. The use of SMOTE effectively addressed the class imbalance issue, improving the model's ability to identify churn cases. Key features like `Total day charge` and `Customer service calls` were identified as strong predictors of churn.

5. Libraries Used

- **Pandas**: Data manipulation and analysis.
- **NumPy**: Numerical computations.
- **Matplotlib/ Seaborn**: Data visualization.
- **Scikit-learn**: Model training, evaluation, and preprocessing.
- **Imbalanced-learn**: SMOTE for handling class imbalance.