Project Report

# Customer Churn Prediction

## 1. Introduction

- **Objective:**
    - The primary objective of this project is to develop a machine learning model that predicts customer churn for a subscription-based service. Predicting churn helps businesses understand which customers are likely to stop using their service, allowing them to take proactive measures to retain those customers.
- **Importance:**
    - Reducing churn is crucial for maintaining revenue and ensuring the long-term sustainability of a subscription-based business. By identifying patterns that lead to churn, companies can develop targeted strategies to improve customer satisfaction and retention.

## 2. Data Understanding

- **Dataset Description:**
    - The dataset used in this project is the **Telco Customer Churn** dataset, which contains 7,043 rows and 21 columns. Each row represents a customer, and each column contains customer attributes such as demographic information, account details, and service usage data.
- **Key Attributes:**
    - **Churn:** The target variable indicating whether the customer has churned (1) or not (0).
    - **Tenure:** The number of months the customer has been with the company.
    - **Contract:** The type of contract the customer has (Month-to-month, One year, Two years).
    - **MonthlyCharges:** The amount charged to the customer monthly.
    - **TotalCharges:** The total amount charged to the customer.
- **Data Source:**
    - The dataset is publicly available on Kaggle.

## 3. Data Preprocessing

- **Dropping Irrelevant Columns:**
  - The customerID column was dropped as it does not contribute to predicting churn.
- **Handling Missing Values:**
  - The TotalCharges column had some missing values. These were replaced with the median value of the column.
- **Encoding Categorical Variables:**
  - Categorical variables were converted into numerical values using Label Encoding. This was necessary for the machine learning models to process the data.
- **Feature Scaling:**
  - Continuous features were scaled using StandardScaler to standardize the range of the features, which is essential for models like Logistic Regression.

## 4. Model Development

- **Data Splitting:**
  - The dataset was split into training (80%) and testing (20%) sets using train_test_split.
- **Model Selection:**
  - Three models were selected for training:
    1. **Logistic Regression:** A baseline model that is easy to implement and interpret.
    2. **Decision Tree:** A model that can capture non-linear relationships and interactions between features.
    3. **Random Forest:** An ensemble model that combines multiple decision trees to improve performance and reduce overfitting.

## 5. Model Evaluation

- **Performance Metrics:**
  - The models were evaluated using the following metrics:
    - **Accuracy:** The proportion of correctly predicted churn cases.
    - **Precision:** The proportion of positive identifications that were actually correct.
    - **Recall:** The proportion of actual positives that were correctly identified.

        - **F1-Score:** A balance between precision and recall.
        - **ROC-AUC Score:** A measure of the model's ability to distinguish between classes (churn vs. no churn).
- **Results:**
    - **Logistic Regression:**
        - **Accuracy:** 80.4%
        - **Precision:** 72.3%
        - **Recall:** 66.2%
        - **F1-Score:** 69.1%
        - **ROC-AUC:** 0.84
    - **Decision Tree:**
        - **Accuracy:** 78.5%
        - **Precision:** 70.1%
        - **Recall:** 68.7%
        - **F1-Score:** 69.4%
        - **ROC-AUC:** 0.81
    - **Random Forest:**
        - **Accuracy:** 82.1%
        - **Precision:** 74.6%
        - **Recall:** 69.3%
        - **F1-Score:** 71.9%
        - **ROC-AUC:** 0.87
- **ROC Curves:**
    - ROC curves were plotted for all models, indicating that the Random Forest model had the best performance in terms of ROC-AUC score.

## 6. Conclusion

Among the three models, the Random Forest classifier performed the best, achieving an accuracy of 82.1% and an ROC-AUC score of 0.87. The model successfully identified the most important factors contributing to customer churn, such as Contract, tenure, and Monthly Charges. By deploying this model, businesses can proactively identify customers who are at risk of churning. This allows the company to take targeted actions such as offering discounts, improving customer service, or providing incentives to retain these customers.