Customer Churn Prediction Project Report

1. Introduction

Customer churn refers to the rate at which customers stop doing business with a company. In the competitive telecom industry, retaining customers is more cost-effective than acquiring new ones. This project aims to predict customer churn and identify key factors influencing attrition, helping businesses develop data-driven retention strategies.

Objective:

- Predict customer churn using machine learning models.
- Uncover key predictors of churn from customer data.
- Provide actionable insights for improving customer retention.

2. Dataset Overview

The dataset used in this project is the **Telco Customer Churn dataset**, sourced from Kaggle. It contains 7,032 rows and 21 columns with customer demographic, account, and service details.

- **Key Features**:
    - Demographics: Gender, Senior Citizen, Partner, Dependents.
    - Services: Phone Service, Internet Service, Online Security, Streaming TV/Movies, etc.
    - Account Details: Contract type, Paperless Billing, Payment Method.
    - Usage and Charges: Tenure, Monthly Charges, Total Charges.
    - Target Variable: Churn (Yes/No).

3. Data Preprocessing

1. **Handling Missing Values**:
    - The TotalCharges column had missing values, which were addressed by converting it to numeric and dropping rows with missing entries.
2. **Feature Encoding**:
    - Binary features (e.g., gender, Partner) were encoded as 0 and 1.

    - Categorical features with multiple categories (e.g., Internet Service, Contract) were one-hot encoded.

3. **Normalization**:
    - Features were normalized to ensure compatibility with the machine learning model.

Final Dataset:
After preprocessing, the dataset contained 7,032 rows and 31 features.

4. Exploratory Data Analysis (EDA)

Key findings from EDA:

- **Churn Rate**:
  The overall churn rate was **26.6%**, indicating that over one-fourth of the customers left the service.
- **Tenure**:
  Customers with shorter tenure (newer customers) had a significantly higher likelihood of churning.
- **Monthly Charges**:
  Higher monthly charges correlated with a higher churn rate.
- **Contract Type**:
  Month-to-month contract customers showed the highest churn rate compared to those on one-year or two-year contracts.

Visualization Highlights:

- Histograms of tenure and monthly charges revealed clear trends linked to churn.
- Correlation analysis identified significant relationships between key features and churn.

5. Machine Learning Model

1. **Chosen Model**:
   The **Random Forest Classifier** was selected for its ability to handle mixed data types, robustness against overfitting, and ease of interpretability (feature importance).
2. **Why Random Forest?**
    - It handles non-linear relationships effectively.
    - It is less sensitive to outliers and missing values.

- Provides insights into feature importance, aiding interpretability.

3. **Performance Metrics**:
   - **Accuracy**: 78.67%
   - **Recall for Churned Customers**: 48.8%
   - **Precision for Churned Customers**: 62.7%
   - **Key Predictors**: Monthly Charges, Tenure, and Contract Type.

6. Insights and Key Findings

1. **High-Risk Customers**:
   - Month-to-month contract customers with high monthly charges and low tenure are at the highest risk of churning.
2. **Retention Strategies**:
   - Focus on offering incentives or long-term contracts to customers with shorter tenures and high charges.
3. **Feature Importance**:
   - The top predictors of churn were:
     - Monthly Charges.
     - Tenure.
     - Contract Type.
4. **Model Strengths**:
   - The model balances accuracy and recall, making it effective for identifying at-risk customers while minimizing false positives.

7. Conclusion

This project demonstrates how predictive analytics can empower businesses to mitigate customer churn. The Random Forest Classifier provided robust predictions, identifying key factors like contract type, tenure, and monthly charges. By leveraging these insights, telecom companies can implement targeted retention strategies, such as offering discounts or promoting long-term contracts to at-risk customers. Future improvements could involve hyperparameter tuning, balancing the dataset, and exploring advanced models like XGBoost for enhanced performance.