📄 Customer Churn Prediction in Telecom Industry Project Report

1. Introduction

In the highly competitive telecom sector, customer retention is critical for profitability. Predicting customer churn allows businesses to proactively retain high-risk users through targeted interventions. This project involves building a machine learning model to predict churn likelihood and segment users based on their churn risk. The goal is to support data-driven retention strategies for improving customer loyalty and lifetime value.

2. Abstract

This project focuses on analyzing customer behavior to predict churn in a telecom company. Using a publicly available dataset, we performed exploratory data analysis (EDA), preprocessing, and applied three machine learning models — Logistic Regression, Random Forest, and XGBoost — to classify customers based on their likelihood of churn. SHAP explainability and segmentation logic were used to interpret the results and group customers into risk categories. Random Forest emerged as the best-performing model with 79.2% accuracy. Business strategies were recommended for each segment to improve retention rates.

3. Tools Used

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Seaborn, Matplotlib, Sklearn, SHAP, XGBoost
- **IDE:** Jupyter Notebook
- **Optional:** Power BI/Tableau (for dashboard, if included)

4. Steps Involved in Building the Project

1. **Data Collection & Cleaning:**

    - Dataset: 7043 customer records with 21 attributes.
    - Cleaned TotalCharges, handled missing values, encoded categorical features.

2. **Exploratory Data Analysis (EDA):**
    - Visualized churn distribution, feature correlations with churn (e.g., Contract, tenure, MonthlyCharges).
    - Observed that short tenure and high monthly charges are strong churn indicators.
3. **Model Building:**
    - Trained three models: Logistic Regression, Random Forest, XGBoost.
    - Evaluated using Accuracy, Precision, Recall, and ROC-AUC.
    - Random Forest gave the highest accuracy of 79.2%.
4. **Model Explainability:**
    - Used SHAP to interpret feature importance globally and individually.
    - Key churn drivers: Month-to-month contracts, high charges, low tenure.
5. **Customer Segmentation:**
    - Customers grouped into Loyal (0–30%), At Risk (30–70%), and High Risk (70–100%) based on churn probability.
    - Segment-wise strategy recommendations were made.

6. Conclusion

This project successfully demonstrated the application of machine learning for churn prediction in the telecom domain. By analyzing customer behavior patterns and predicting churn likelihood, the company can take proactive steps to reduce customer loss. In future iterations, advanced ensemble tuning, class balancing (SMOTE).