# Customer Churn Prediction Project

## Project Objective

The goal of this project was to build a machine learning model to predict customer churn for a telecom company, enabling targeted retention strategies to reduce churn. This is critical in the telecom industry, where high acquisition costs and low loyalty make churn prediction essential for profitability. The model’s predictions were intended to inform financial outcomes, such as revenue saved and retention costs, using a planned business impact analysis.

## Dataset Description

The telecom\_churn.csv dataset contains 243,553 rows and 14 columns, including:

* Demographic features: age, gender, state, city, pincode, num\_dependents, estimated\_salary.
* Usage metrics: calls\_made, sms\_sent, data\_used.
* Service details: telecom\_partner, date\_of\_registration.
* Target: churn (0 = stayed, 1 = churned).

Negative values were observed in data\_used, indicating potential data quality issues requiring preprocessing.

## Data Cleaning and Preprocessing

* **Missing & Duplicate Values**: I confirmed no missing values and no duplicate rows.
* **Negative Values**: Negative values in data\_used were removed from the dataset
* **Outlier Detection**: No outliers were detected in estimated\_salary, data\_used, and calls\_made using the IQR method. I tested capping outliers at the 1st and 99th percentiles to assess impact.
* **Encoding and Scaling**: Categorical variables (gender, telecom\_partner, state, city) were encoded using LabelEncoder. Numerical features were scaled with StandardScaler for models like Logistic Regression.

## Exploratory Data Analysis (EDA)

I generated a pandas\_profiling report and used seaborn and matplotlib for visualizations. Key insights:

* Customers with lower calls\_made, sms\_sent, and data\_used had a higher churn rate (e.g., 15% churn for low-usage customers).
* Certain states had higher churn density.
* The churn column was imbalanced, necessitating balancing techniques.

I created some visualizations using histogram, barplot, boxplot for numerical variables distribution and for churn rate for categorical variables.

## Handling Class Imbalance

The dataset had a 20% churn rate. I applied SMOTE to create synthetic samples for the minority class (churn = 1).

## Models Trained

I trained five models: Random Forest, Decision Tree, Logistic Regression, XGBoost, and Gradient Boosting.

**Final Model**: Random Forest was the model that performed better with an accuracy 76%.

## Evaluation Metrics

I used Accuracy, Precision, Recall, F1-Score, and ROC-AUC, calculated with classification\_metrics and visualized with roc\_curve\_plot. Random Forest’s high ROC-AUC and balanced Precision/Recall made it ideal for churn prediction, minimizing missed churners and false positives.

## Results Interpretation

* **Random Forest**: Best performer due to its robustness, handling of categorical features, and high ROC-AUC. Feature importance plots (from plot\_feature\_importance) showed data\_used, calls\_made, and sms\_sent as top predictors.
* **XGBoost**: Slightly higher Recall but required more tuning. Feature importance aligned with Random Forest.
* **Gradient Boosting**: Competitive but slightly lower performance than Random Forest.
* **Decision Tree**: Overfit, with lower test metrics.
* **Logistic Regression**: Limited by linear assumptions, lower Recall.