TELCO CUSTOMER CHURN PREDICTION

By: Manas Nayan Mukherjee

# ABSTRACT

This project analyzes a telecom customer dataset to predict customer churn. The dataset includes customer demographic information, account details, service usage, and billing information. Our goal is to identify patterns and build a predictive model to determine if a customer is likely to leave the service (churn). This enables telecom companies to design retention strategies and reduce attrition rates.

# INTRODUCTION

Telecommunication companies lose a significant portion of their customer base every year. Customer churn prediction is vital to proactively identify and retain at-risk customers. With the help of historical data and machine learning, we can develop a model that flags potential churners in advance.

# PROBLEM STATEMENT

We aim to explore a telco dataset and predict whether a customer is likely to churn. This will help the company to:  
- Identify customer segments at risk  
- Understand key drivers of churn  
- Improve customer retention efforts

# FEATURE DESCRIPTION

|  |  |
| --- | --- |
| Feature Name | Description |
| gender | Customer's gender |
| SeniorCitizen | Whether the customer is a senior |
| Partner | Whether the customer has a partner |
| tenure | Number of months the customer has stayed |
| PhoneService | Whether the customer has phone service |
| MultipleLines | Whether the customer has multiple lines |
| InternetService | Type of internet service |
| Contract | Type of contract (Month-to-month, etc.) |
| PaperlessBilling | Whether the customer uses paperless billing |
| MonthlyCharges | Monthly charges incurred by the customer |
| TotalCharges | Total charges incurred |
| Churn | Target variable – has the customer churned? |

# EXPLORATORY DATA ANALYSIS (EDA)

* DATA PREPARATION
* - Imported necessary libraries: pandas, numpy, matplotlib, seaborn, warnings
* - Validated dataset types, checked for duplicates
* - Converted TotalCharges to numeric type

# MISSING VALUES AND OUTLIERS

* - Used IQR and Boxplots to handle outliers
* - Applied capping for extreme values in MonthlyCharges and TotalCharges
* - Missing values in TotalCharges treated using median imputation

# DATA PREPROCESSING

* - Removed whitespace and formatting issues
* - Encoded categorical variables using LabelEncoding or OneHotEncoding
* - Scaled numerical columns (StandardScaler)

# UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

* UNIVARIATE: Barplots for categorical features, histograms for numerical features
* BIVARIATE: Churn % vs Contract Type, Correlation matrix
* MULTIVARIATE: Heatmaps, effects of Contract, tenure, and MonthlyCharges on Churn

# MACHINE LEARNING MODELS

* - Logistic Regression, Random Forest, XGBoost, SVM
* - Accuracy, ROC-AUC, Precision/Recall, Confusion Matrix

# CHALLENGES

* - Handling skewed churn classes (applied SMOTE)
* - Cleaning ambiguous TotalCharges entries
* - Hyperparameter tuning

# CONCLUSION

* - Month-to-month contracts significantly impact churn
* - Customers with fiber optic internet are more likely to churn
* - Customers with high monthly charges tend to churn more
* - Churn prediction accuracy ~85% using XGBoost model
* - Helps proactive customer retention