**Github** : <https://github.com/Arunking007/Customer-churn>

**Predicting Customer**

**Churn using Machine Learning**

Uncover Hidden Patterns to Improve Retention

# DETAILS

NAME : ARUNKUMAR.R

REG NO : 612823205008

DEGREE : B.TECH-INFORMATION TECHNOLOGY

INSTITUTION : VARUVAN VADIVELAN INSTITUTION OF TECHNOLOGY DATE OF SUBMISSION : 08.05.2025

# Problem Statement

* Companies face challenges in identifying customers at risk of leaving.
* Without accurate predictions, resources are wasted on retaining customers who may stay and neglecting those who may leave
* Customer churn impacts business revenue
* It's challenging to predict which customers will leave
* Goal: Use ML to detect churn patterns and reduce losses

# Project Objective

* Build a predictive model using customer data
* Help businesses identify customers likely to churn
* Provide insights to improve customer retention strategies

## Flowchart of Project

**Data**

**Collection**

**Preprocessin**

**g**

**EDA**

**Feature**

**Engineering**

**Model**

**Building**

**Evaluation**

**Insights &**

**Deployment**

# Data Description

* Dataset Name: Telco Customer Churn • Dataset Source: Kaggle
* Dataset link:

[https://www.kaggle.com/datasets/shantanudhakadd/b ank-customer-churn-prediction](https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction)

* Type of Data: Structured tabular data
* Records and Features: 7,043 customer records and 21 features (categorical + numerical)
* Target Variable: Churn (Yes/No – indicates if the customer left)
* Static or Dynamic: Static dataset

# Data Preprocessing

* Verified dataset integrity: Confirmed that there were no missing or null values.
* Removed irrelevant features: Dropped columns with very low variance that don't contribute to predictions (e.g., features with only one unique value).
* Duplicate check: Ensured the dataset contained no duplicate rows.
* Categorical encoding: Converted all categorical features to numerical using one-hot encoding to make them compatible with machine learning algorithms.
* Normalization: Applied StandardScaler to numerical columns

(MonthlyCharges, TotalCharges, tenure) to bring them to a similar scale.

* Outlier detection: Used boxplots and z-score methods to detect and analyze extreme outliers; these were further investigated to decide on retention or removal

## Exploratory Data Analysis (EDA)

* Analysis :
  + Explored relationships between features and churn.
  + Visualized data using charts like bar graphs, histograms, and box plots.
* Key Findings :
  + High correlation between churn and factors like service tenure and monthly charges.
  + Certain account features like contract type and payment method significantly influence churn predictions.

# Feature Engineering

• New Features :

* Created binary features for churn (Yes/No).
* Derived tenure categories (short, medium, longterm).
* Transformed continuous features like monthly charges into categorical groups.

# Model Building

* Data Preprocessing:
  + Handle missing values and encode categorical features using One-Hot Encoding.
  + Scale numerical features using StandardScalar.
* Model Selection:
  + Random Forest Classifier: An ensemble method using multiple decision trees.
  + Logistic Regression: A linear model predicting churn probability.
* Train-Test Split:
  + Split data into 80% training and 20% testing sets.
* Model Training:
  + Train both models on the training set using respective algorithms.
* Model Evaluation:
  + Evaluate using Confusion Matrix, Classification Report, and Accuracy.
  + Compare Precision, Recall, F1-Score, and Accuracy to choose the best model.
* Prediction:
  + Predict customer churn for new data using the trained model.

# Visualization of Results

* Confusion Matrix
  + Shows True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).
  + Helps in understanding model accuracy visually.
* ROC Curve (Receiver Operating Characteristic)
  + Plots True Positive Rate vs False Positive Rate.
  + AUC (Area Under Curve) score closer to 1 = better performance.
* Precision-Recall Curve
  + Especially useful when dealing with imbalanced classes (like churn vs no churn).
  + Shows trade-off between precision and recall.
* Feature Importance Bar Chart
  + Highlights which features influenced churn prediction the most.
  + Helps in explaining the model's decisionmaking process.
* Classification Report Table
  + Metrics like Accuracy, Precision, Recall, F1Score.
  + Can be shown as a colored heatmap or plain table.

# Tools & Technologies Used

* Python (Pandas, Sklearn, Matplotlib, etc.)
* Google Colab
* Kaggle (Dataset)
* Streamlit (for App)

## TEAM MEMBERS AND CONTRIBUTION

* AJAY K
  + Data Cleaning & EDA
* ALLAUDDIN K
  + Feature Engineering & Model Development
* ARUN KUMAR R
  + Documentation & Reporting