End-to-End ML Pipeline with Scikit-learn Pipeline API

1. Problem Statement

Customer churn is when a customer stops using a company's services. Predicting churn is critical for businesses to take preventive actions and retain customers. The Telco Churn Dataset provides information about customer demographics, services, account details, and churn labels.

2. Objective

Build a **reusable and production-ready machine learning pipeline** to predict customer churn using Logistic Regression and Random Forest, with proper preprocessing, hyperparameter tuning, evaluation, visualization, and model export.

3. Dataset Overview

- **Source:** Telco Customer Churn Dataset (CSV format)
- **Number of Rows**: ~7,000
- **Columns (key ones):**(customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn)
- **Target Column:** Churn (Yes = 1, No = 0)
- **Dataset Challenges:**Missing values in **TotalCharges**.Mix of numerical and categorical features

4. Data Preprocessing

Steps Taken

1. Removed customerID column (not useful for prediction).
2. Converted TotalCharges to numeric and filled missing values with median.
3. Split columns into:
    - **Numerical features:** tenure, MonthlyCharges, TotalCharges
    - **Categorical features:** All other object-type columns
4. Preprocessing pipeline:
    - Numerical: Median imputation + StandardScaler
    - Categorical: Most frequent imputation + OneHotEncoder
5. Split data into train and test sets (80%-20%, stratified by Churn).

Pipeline Benefits

- Automatic handling of missing values
- Scaling of numerical features
- Encoding categorical features
- Fully reusable in production

5. Model Development & Training

Models Used

1. Logistic Regression
2. Random Forest Classifier

Hyperparameter Tuning

- Performed using GridSearchCV with 5-fold cross-validation.
- **Logistic Regression:** C and penalty tuned.
- **Random Forest:** n_estimators, max_depth, and min_samples_split tuned.
- **Scoring metric:** F1-score (balances precision & recall for churn prediction).

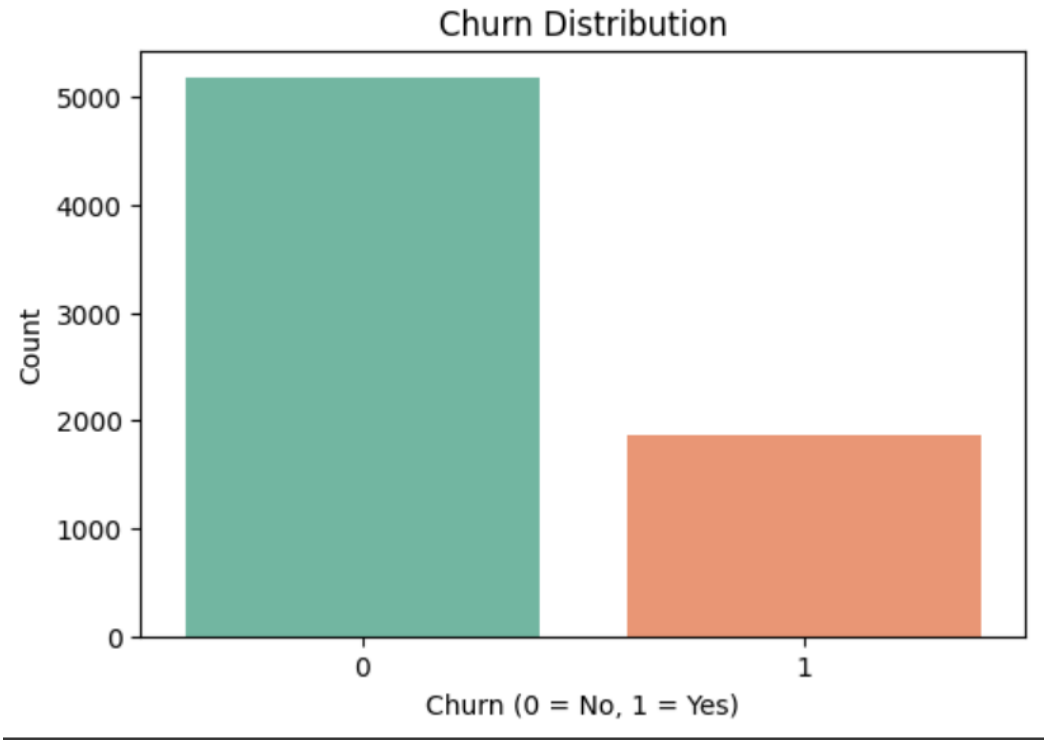
```
Logistic Regression Accuracy: 0.8240
Random Forest Accuracy: 0.7857
```

Best Model Selection

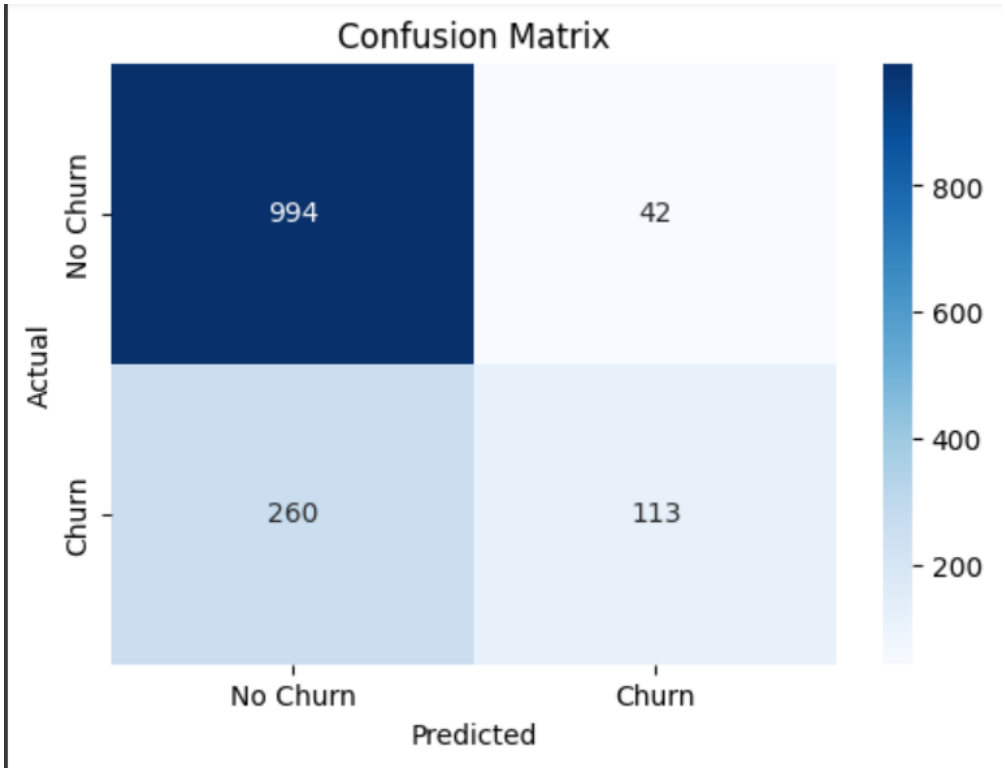
- GridSearchCV automatically selects the model and parameters with the highest F1-score on the training set.

6. Visualize Accuracy & Loss

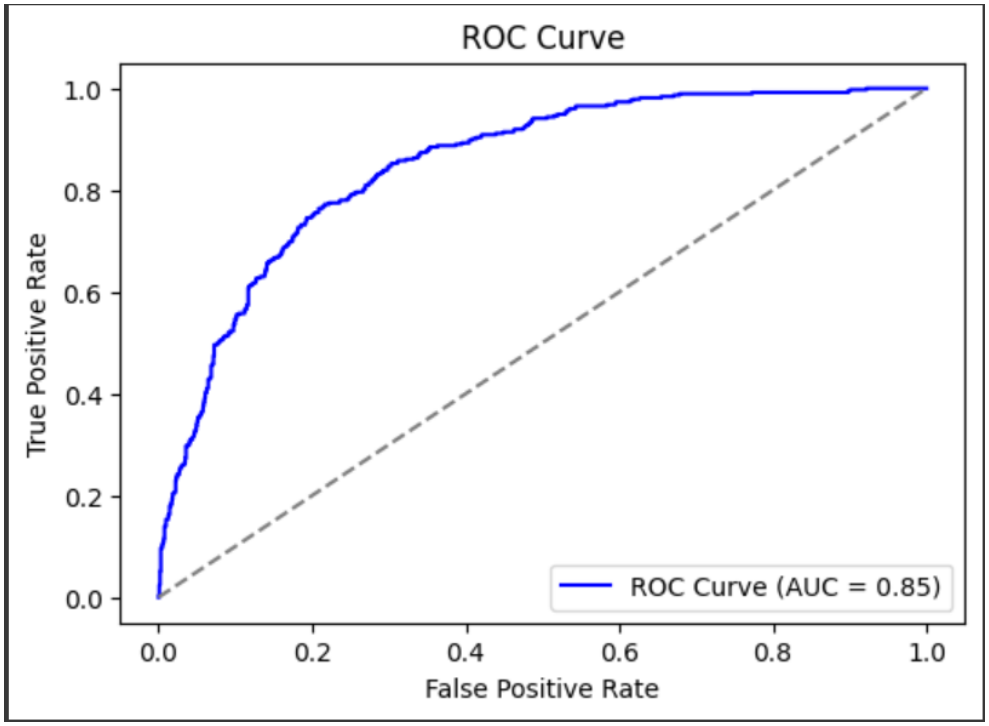
- **Class Distribution**: Shows imbalance between churn and non-churn customers.



- **Confusion Matrix**: Displays true positives/negatives and false positives/negatives.



- **ROC Curve**: Shows tradeoff between true positive rate and false positive rate.



- **Feature Importance**: Top 10 features impacting churn prediction (from Random Forest).