Telco Customer Churn Prediction
End-to-End ML Pipeline Project

Author: Tayyab Malik

Date: July 2025

OBJECTIVE

Build a production-ready machine learning pipeline using Scikit-learn to predict customer churn based on telecom usage data.

DATASET

- Telco Customer Churn Dataset
- Source: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

TOOLS & LIBRARIES

- Python, pandas, numpy, scikit-learn, joblib, matplotlib, seaborn

STEP-BY-STEP PIPELINE

1. DATA LOADING & CLEANING

- Loaded CSV dataset, removed unnecessary columns (customerID), handled missing TotalCharges.
- Converted 'Churn' column to binary labels.

2. TRAIN-TEST SPLIT

- Used train_test_split (80/20) with stratification for balanced classes.

3. PREPROCESSING

- Used ColumnTransformer to scale numerical features and one-hot encode categorical ones.
- Applied within a Pipeline for consistency.

4. MODEL TRAINING

- Built two pipelines: Logistic Regression and Random Forest.
- Combined preprocessing + model into one unit.

5. HYPERPARAMETER TUNING

- Used GridSearchCV with 5-fold CV:

- Logistic Regression (C values)

- Random Forest (n_estimators, max_depth)

6. MODEL EVALUATION

- Accuracy Scores:

- Logistic Regression ~80%

- Random Forest ~83%

- Used classification_report for precision, recall, F1-score.

7. MODEL EXPORT

- Exported best-performing model (Random Forest pipeline) using joblib.

8. MODEL REUSE

- Demonstrated loading and using pipeline for future predictions on new data.

FINAL INSIGHTS

- Customers with month-to-month contracts, high charges, and short tenure are more likely to churn.

- Random Forest captured patterns better than Logistic Regression.

- Exported pipeline is reusable and ready for deployment.

FUTURE IMPROVEMENTS

- Deploy with Flask or Streamlit

- Add dashboard and monitoring

- Experiment with more advanced models

PROJECT STRUCTURE

- telco_churn_pipeline.pkl (exported model)

- Notebook.ipynb (complete code)

- Dataset.csv (Kaggle)

- README.md

- Report.pdf

AUTHOR

Tayyab Malik