Customer Churn Prediction and Analysis Project

Project Overview

This project focused on developing a machine learning solution to predict customer churn for a telecommunications company. By accurately identifying customers at risk of churning, the company can take proactive retention measures to maintain its customer base and revenue.

The project was completed by a team of five members, each responsible for a specific milestone:

- Zeyad Sami Tahoun: Milestone 1 - Data Collection, Exploration, and Preprocessing
- Hager Essa: Milestone 2 - Advanced Data Analysis and Feature Engineering
- Ahmed Reda: Milestone 3 - Machine Learning Model Development and Optimization
- Doha Sayed: Milestone 4 - MLOps, Deployment, and Monitoring

Milestone 1: Data Collection, Exploration, and Preprocessing

Data Collection:

The IBM Telco Customer Churn dataset was used. It contains data on ~7,000 customers.

Data Exploration:

- Analyzed feature distributions and relationships
- Checked for missing values, duplicates, and outliers
- Identified the churn rate (~26-27%)

Data Preprocessing:

- Standardized column names
- Handled missing values in TotalCharges
- Converted SeniorCitizen to categorical
- Created numeric Churn Value column

Key Visualizations:

- Tenure and Monthly Charges by churn
- Churn rate by contract type and internet service
- Churn across demographic factors

Customer Churn Prediction and Analysis Project

Initial Key Findings:

- Shorter tenure = higher churn
- Month-to-month contracts = highest churn (~40-45%)
- Fiber optic = higher churn
- Lack of online services = higher churn
- Electronic check = higher churn
- Higher monthly charges = higher churn

Milestone 2: Advanced Data Analysis and Feature Engineering

Feature Engineering:

- Tenure Groups, Service Count, Tech/Streaming Service Flags, New Customer and Family Flags
- Monthly Charges Category
- Customer Segments: New/Established, Low/High-Value

Statistical Analysis:

- Chi-square, T-tests, Correlation analysis

Advanced Feature Selection:

- ANOVA F-value, Mutual Information

Key Findings:

- Contract type strongly associated with churn
- Tenure and segments negatively correlated with churn
- Tech services reduce churn
- More services = lower churn

Milestone 3: Machine Learning Model Development and Optimization

Data Preparation:

- One-hot encoded categorical variables
- Train-test split (80/20)

- StandardScaler used on numerical features

Models Trained:

1. Logistic Regression
2. Random Forest
3. Gradient Boosting
4. XGBoost

Metrics: Accuracy, Precision, Recall, F1 Score, ROC-AUC

Feature Importance:

- Confirmed contract type, tenure, and services as key factors

Model Tuning:

- Grid Search for hyperparameter tuning based on F1 Score

Model Summary:

- Best model saved with scaler and feature names for deployment

Milestone 4: MLOps, Deployment, and Monitoring

Prediction Function:

- Applies preprocessing, scaling, and returns probability results

Flask API Design:

- Prediction and health check endpoints
- Error handling and API documentation