Machine Learning

(UE23CS352A)

Project Title : Customer Churn Prediction in Telecom Industry

Project Team:

Poojitha CV PES2UG23CS415

Podamala Pragna PES2UG23CS411

Problem statement

Customer churn is a critical issue in the telecom industry, where retaining existing customers is often more cost-effective than acquiring new ones. Churn occurs when a customer discontinues a company's services, leading to a direct loss in revenue and market share.
This project aims to develop a **machine learning-based predictive model** that identifies customers likely to churn based on their demographics, account details, and service usage behavior.

The primary goals are to:
• Predict whether a customer is likely to churn using historical telecom data.
• Identify key factors influencing customer retention and churn behavior.
• Provide actionable insights to help telecom companies design effective retention strategies.

Approach

The project follows a data-driven machine learning approach involving the following key steps:

1. Data Collection:
   - Used the **Telco Customer Churn** dataset containing 7043 records and 21 features.
   - Features included demographics, service details, contract type, payment method, and monthly/total charges.

2. Data Preprocessing:
   - Handled 11 missing values in the **TotalCharges** column by replacing them with the median.
   - Removed the customerID column (non-informative).
   - Ensured data consistency across categorical and numerical fields.

3. Exploratory Data Analysis (EDA):
   - Visualized churn trends by gender, senior citizen status, contract type, payment method, tenure, and monthly charges.
   - Found that month-to-month contracts, higher charges, and shorter tenures are strongly linked to higher churn.

4. Feature Engineering:
   - Created a tenure_group feature to categorize customers into groups (0–12, 13–24, 25–36, 37–48, 49–60, 60+ months).
   - Applied one-hot encoding to categorical variables.
   - Scaled numerical features using StandardScaler.
   - Split the dataset into 80% training and 20% testing sets.

5. Model Building:

- Implemented and compared the following models:
  - Logistic Regression
  - Random Forest Classifier
  - Gradient Boosting Classifier

6. Model Evaluation:

- Evaluated performance using Accuracy, Confusion Matrix, and Classification Report.
- Compared models to identify the best-performing one.

Implementation Overview

The project was implemented using **Python** as the primary programming language due to its robust ecosystem for data science and machine learning.

Libraries Used:

- **pandas, numpy** – for efficient data handling and preprocessing
- **matplotlib, seaborn** – for insightful data visualization and exploratory analysis
- **scikit-learn** – for building, training, and evaluating machine learning models

Workflow Summary:

1. Imported all the required libraries and configured the environment for clean, reproducible outputs.
2. Loaded the **Telco Customer Churn** dataset and carried out data cleaning to handle missing and inconsistent values.
3. Performed **exploratory data analysis (EDA)** to identify key trends and factors influencing customer churn.

4. Applied **feature engineering** techniques such as label encoding and one-hot encoding for categorical variables.
5. Standardized numerical features using **StandardScaler** to maintain consistent scale across attributes.
6. Split the dataset into **training (80%)** and **testing (20%)** subsets for model evaluation.
7. Trained and compared three machine learning models — **Logistic Regression**, **Random Forest Classifier**, and **Gradient Boosting Classifier**.
8. Assessed model performance using **Accuracy**, **Confusion Matrix**, and **Classification Report** metrics to identify the best-performing model

Results and Discussion

Model Name	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.8041	0.84	0.90	0.87
Random Forest Classifier	0.7949	0.83	0.90	0.87
Gradient Boosting Classifier	0.8041	0.84	0.90	0.87

From the results, **Logistic Regression** achieved the best accuracy of **80.41%**, making it both effective and interpretable for churn prediction.
Although ensemble models like Random Forest and Gradient Boosting performed closely, Logistic Regression provided a simpler yet equally accurate solution.

Challenges Faced

- **Data Imbalance:** The churn dataset showed class imbalance, which required attention to prevent biased predictions.
- **Feature Encoding:** Handling multiple categorical features (like contract type and payment method) required careful encoding to retain data meaning.
- **Model Tuning:** Optimizing ensemble models to balance accuracy and avoid overfitting involved several tuning iterations.
- **Computation Time:** Ensemble methods such as Random Forest and Gradient Boosting were computationally intensive compared to Logistic Regression.

Conclusion

The project successfully built a **machine learning-based churn prediction system** capable of identifying customers likely to discontinue telecom services.
Among all tested models, **Logistic Regression achieved the highest accuracy (≈80.4%)**, demonstrating that a well-preprocessed dataset can yield strong results even with simpler models.

This solution can assist telecom companies in:

- Predicting potential churners with high accuracy,
- Designing targeted retention campaigns, and
- Enhancing customer satisfaction while reducing revenue loss.