Customer Churn Prediction Using Machine Learning

Prepared by: Mohamed Shahid

Abstract

This report presents a machine learning project focused on customer churn prediction. The goal of the project is to identify customers who are likely to discontinue using a company's services, enabling the business to take proactive retention measures. Two classification algorithms — Logistic Regression and Random Forest — were applied to a real-world customer dataset to predict churn effectively.

Introduction

Customer churn, or customer attrition, refers to the loss of clients or subscribers. It is a significant problem for businesses, particularly subscription-based services. Retaining existing customers is more cost-effective than acquiring new ones. Therefore, being able to predict which customers are at risk of churning allows businesses to take timely action and reduce churn rates.

In this project, machine learning techniques are applied to historical customer data to predict churn. The analysis involves data preprocessing, model building, evaluation, and performance comparison between two classification algorithms.

Dataset Description

The dataset used in this project contains customer information such as:

- Customer ID
- Tenure (length of service)
- Phone service status
- Internet service details
- Contract type
- Payment method
- Monthly charges
- Total charges
- Churn status (Yes/No)

The **target variable** is `Churn`, which indicates whether a customer has discontinued the service (Yes) or not (No).

Data Preprocessing

Before building the models, several data preprocessing steps were performed:

- **Missing Values Handling:**
  The `TotalCharges` column contained missing or invalid values which were handled by converting to numeric and dropping nulls.
- **Encoding Categorical Variables:**
  Many features were categorical. These were encoded using:
  - Label encoding for binary columns (e.g., `Churn` → 1/0).
  - One-hot encoding for columns with multiple categories.
- **Feature Scaling:**
  Numerical features were scaled using `StandardScaler` to ensure fair weightage during model training.
- **Train-Test Split:**
  The dataset was split into training and testing sets with 80% data for training and 20% for testing.

Modeling

1 Logistic Regression

A simple linear classification model used for binary outcomes. It was applied after scaling the features. Logistic Regression is interpretable and less prone to overfitting.

2 Random Forest Classifier

An ensemble technique that builds multiple decision trees. It handles non-linear relationships better and can rank feature importance.

Model Evaluation

The performance of both models was evaluated using:

- Confusion Matrix
- Classification Report (Precision, Recall, F1-Score)
- Accuracy Score
- ROC-AUC Score

Both models demonstrated good performance on the dataset. Random Forest slightly outperformed Logistic Regression due to its ability to capture complex feature interactions.

Feature Importance (Random Forest)

The Random Forest model identified the most influential features in predicting churn. Key features included:

- Tenure
- Contract type
- Monthly Charges
- Total Charges

- Internet Service type

This information can help businesses focus on important factors that contribute to customer churn.

Conclusion

Machine learning models such as Logistic Regression and Random Forest can accurately predict customer churn when provided with well-prepared data. Random Forest performed slightly better in this project. Predicting churn allows businesses to identify at-risk customers and take proactive steps to retain them, improving profitability and customer satisfaction.

References

- Dataset: Telco Customer Churn Dataset (or provided internal dataset)
- Scikit-learn Documentation
- Pandas for data preprocessing
- Seaborn and Matplotlib for data visualization