1. Introduction

The objective of this project is to predict customer churn for a telecom company. Customer churn refers to the loss of clients or customers. Accurate prediction of churn can help the company take proactive measures to retain customers and reduce revenue loss.

2. Data Preprocessing

The dataset used for this project was obtained from the Telco Customer Churn dataset on Kaggle. It contains 7,032 rows and 21 columns, with features like customer demographics, account information, and services subscribed.

Steps taken:

Handling missing values: Replaced spaces with NaN and removed rows with missing values.

Encoding categorical variables: Used LabelEncoder to convert categorical features to numeric values.

3. Exploratory Data Analysis (EDA)

EDA was performed to understand the data better and identify key features that influence customer churn.

Findings:

Correlation Matrix: A heatmap was generated to visualize the correlation between numerical features. This helped in understanding relationships between different features.

Churn by Contract Type: A count plot was created to show churn distribution across different contract types. This revealed that customers with month-to-month contracts have a higher churn rate compared to those with one or two-year contracts.

4. Feature Engineering

To enhance the predictive power of the model, a new feature TotalServices was created, representing the sum of all service subscriptions a customer has.

5. Model Building

A Logistic Regression model was chosen for this project due to its simplicity and interpretability.

Steps:

Data Preparation: Split the data into training and testing sets with an 80-20 split.

Standardization: Standardized the features to have a mean of 0 and a standard deviation of 1.

Model Training: Trained a Logistic Regression model with a maximum iteration of 1000.

6. Model Evaluation

The model's performance was evaluated using various metrics:

Accuracy: 0.80

Precision: 0.68

Recall: 0.53

F1 Score: 0.59

Conclusion

The Logistic Regression model provides a good baseline for predicting customer churn with an AUC of 0.82. Further improvements could be made by addressing class imbalance and exploring more advanced models like Random Forest or Gradient Boosting.

This project demonstrated the importance of data preprocessing, feature engineering, and model evaluation in building effective predictive models. The insights gained from EDA, especially regarding contract types and service subscriptions, could help the telecom company devise better retention strategies.