Customer Churn Prediction Project Report

Project Overview

This project aims to predict customer churn using the Telco Customer Churn dataset. Churn prediction helps businesses identify customers likely to leave and take proactive measures to retain them.

Dataset

- Source: IBM Telco Customer Churn
- Rows: 7043
- Columns: 21
- Target: Churn (Yes/No)

Methodology

1. **Data Acquisition:**
    - Dataset downloaded automatically from a public source.
2. **Preprocessing:**
    - Handled missing values.
    - Encoded categorical features.
    - Scaled numerical features.
3. **Modeling:**
    - Used Logistic Regression for baseline prediction.
    - Split data into training and test sets (80/20).
4. **Evaluation:**
    - Accuracy, confusion matrix, classification report.
    - Visualizations for churn distribution, confusion matrix, and feature importance.

Results

- **Test Accuracy:** ~81.7%

- **Confusion Matrix:**
  - True Negatives: 940
  - False Positives: 96
  - False Negatives: 162
  - True Positives: 211
- **Top Features Contributing to Churn:**
  - MonthlyCharges
  - InternetService
  - PaperlessBilling
  - TotalCharges
  - MultipleLines
- **Classification Report:**
  - Precision, recall, and F1-score provided for both churn and non-churn classes.

Visualizations

- Confusion matrix heatmap
- Churn distribution bar chart
- Top 10 feature importances

Conclusion

The logistic regression model provides a solid baseline with 81.7% accuracy. The most influential features for predicting churn are MonthlyCharges, InternetService, and PaperlessBilling. Further improvements can be made by using advanced models and feature engineering.