Customer Churn Prediction – Project Report

1. Introduction

Customer churn is a major challenge for subscription-based businesses, especially telecom companies.

This project aims to analyze customer behavior and develop predictive machine learning models to identify

customers likely to discontinue their service. Using the Telco Customer Churn dataset, the project applies

data analysis, preprocessing, visualization, and classification techniques to understand churn patterns

and create an accurate churn prediction system.

2. Dataset Overview

The dataset used contains 7,043 customer records with 21 attributes including demographic details,

service usage patterns, billing information, and churn status. Key features include gender, senior citizen

status, partner and dependent information, various service subscriptions, tenure, payment method, and monthly charges.

The target variable 'Churn' indicates whether the customer left the service (Yes/No).

3. Data Preprocessing

To ensure model accuracy, several preprocessing steps were applied:

- Handled missing values in TotalCharges by converting the column to numeric and imputing invalid values using the median.

- Removed irrelevant identifiers such as customerID.

- Standardized categorical values by replacing entries like 'No internet service' with 'No'.

- Encoded binary categorical variables using Label Encoding.

- Applied One-Hot Encoding to multi-class categorical features.

- Created a new feature 'tenure_group' by binning tenure into meaningful ranges.

4. Exploratory Data Analysis (EDA)

The dataset was explored using visualizations to identify churn-related trends. Key findings include:

- Customers with short tenure show significantly higher churn rates.

- High monthly charges correlate strongly with churn.

- Contract type has a major impact: month-to-month customers churn the most.

Count plots, histograms, and box plots were used to understand distributions and relationships within the data.

5. Feature Engineering

Feature engineering improved model performance by:

- Creating the 'tenure_group' categorical feature.

- Encoding categorical variables using label and one-hot encoding.

- Simplifying service-related columns by grouping similar values.

These steps helped models better capture complex relationships within the data.

6. Model Development

Two machine learning models were implemented:

1. Logistic Regression:

- Accuracy: ~80%

- ROC-AUC: ~0.84

- Strengths: interpretable, handles linear relationships well.

2. Random Forest Classifier:

- Accuracy: ~78%

- ROC-AUC: ~0.82

- Strengths: handles nonlinearity, identifies key predictive features.

The dataset was split into 80% training and 20% testing using stratified sampling to maintain churn distribution.

7. Model Evaluation

Models were evaluated using accuracy, ROC-AUC scores, classification reports, and confusion matrices.

Logistic Regression performed slightly better in AUC, while Random Forest offered useful insights through feature importance.

Important predictors included:

- Tenure

- Contract type

- Monthly charges

- Total charges

- Internet service type

8. Feature Importance Analysis

Random Forest feature importance analysis revealed that longer tenure and contract duration reduce churn likelihood.

On the other hand, customers with higher monthly charges or lacking tech support/security services showed increased churn risk.

9. Conclusion

This project successfully developed an end-to-end churn prediction pipeline. The analysis highlights that customers

with short tenure, high monthly charges, and flexible payment options are most likely to churn. The models provide

actionable insights that help businesses implement targeted retention strategies and enhance customer satisfaction.

10. Future Enhancements

Potential improvements include:

- Using advanced models like Gradient Boosting, XGBoost, or Neural Networks.

- Applying SMOTE to handle class imbalance.

- Deploying the model using a web app interface.

- Conducting real-time churn monitoring for business decision support.