Customer Churn Prediction - Detailed Project Report

Project Overview

The objective of this project is to **build a classification model that predicts customer churn** for a telecom company. By predicting which customers are likely to leave, the company can implement strategies to improve customer retention.

Dataset Information

- Dataset Name: Telecom Customer Churn Dataset
- **Description:** Contains customer demographic information, account details, services opted, and whether the customer churned.
- Target Variable: Churn (Yes / No)
- Features: Includes:
 - o Customer Info: Gender, SeniorCitizen, Partner, Dependents
 - o Account Info: Tenure, MonthlyCharges, TotalCharges
 - o Services: PhoneService, InternetService, Contract type, PaymentMethod, etc.

Tools & Technologies

- Language: Python
- Libraries:
 - o Pandas for data manipulation
 - o NumPy for numerical operations
 - o Matplotlib & Seaborn for data visualization
 - o Scikit-learn for machine learning models and evaluation
 - o XGBoost for advanced boosting-based classification

✓ Data Preprocessing

- 1. Loading the Data
 - o Loaded CSV data using Pandas.
- 2. Data Cleaning
 - o Handled missing values in TotalCharges.
 - o Converted data types where required.
- 3. Feature Encoding
 - o Applied Label Encoding for binary categories.
 - o Applied One-Hot Encoding for multiclass categorical features.
- 4. Feature Scaling
 - o Used **StandardScaler** to normalize numeric features for Logistic Regression.
- 5. Data Splitting
 - o Split data into Training (80%) and Testing (20%) using train test split.

Modeling

Logistic Regression

• Scaler: StandardScaler

• **Hyperparameter:** Increased max iter=2000 to ensure convergence.

• Accuracy: ~80%

• Evaluation:

Confusion Matrix

o Classification Report (Precision, Recall, F1-Score)

♦ XGBoost Classifier

- Model trained without label encoding (as it's deprecated in newer XGBoost versions).
- **Hyperparameters:** Default (with eval_metric='logloss')
- Accuracy: Higher than Logistic Regression (Typically ~82-85%)
- **Feature Importance:** Extracted to understand the most impactful features.

Results

Model Accuracy

Logistic Regression ~80% XGBoost ~82-85%

- **Key Metrics:** Precision, Recall, F1-Score evaluated on test data.
- **Feature Insights:** Contract type, tenure, and monthly charges are significant indicators of churn.

✓ Visualizations

- Correlation Heatmap for feature relationships.
- Churn Distribution Plot
- Bar plot for Feature Importance from XGBoost

Conclusions

- XGBoost outperforms Logistic Regression in churn prediction.
- The model can aid telecom companies to proactively target customers likely to churn.
- Business strategies can focus on improving contract offerings and reducing monthly charges.

X Tools & Environment

- Python 3.11
- Jupyter Notebook & Python Scripts
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost

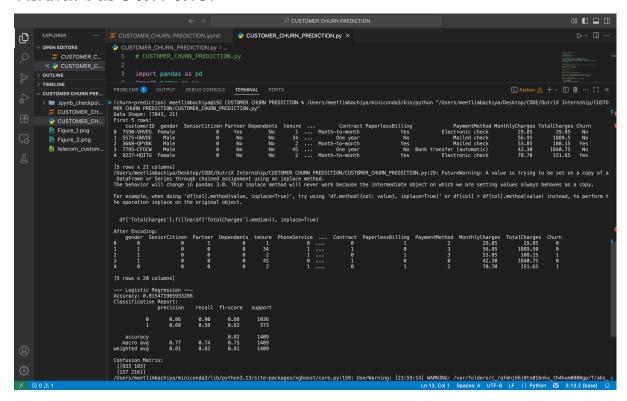
Future Enhancements

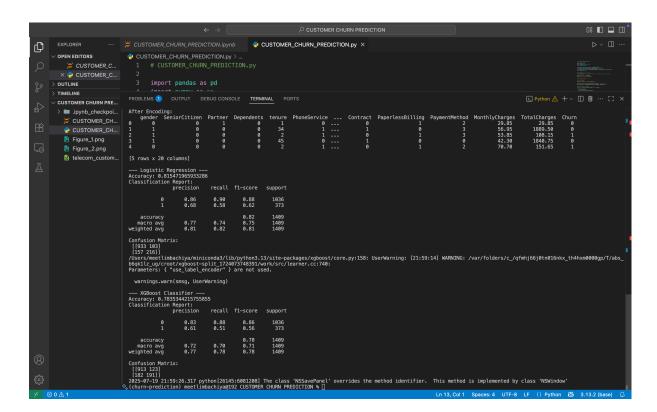
- Hyperparameter tuning (e.g., Grid Search, Randomized Search) for XGBoost.
- Explore more complex models like Random Forest or Neural Networks.
- Deploy model via a web application for real-time predictions.



Meet Limbachiya July 2025 | OutriX Internship Project

TERMINAL OUTPUTS:





GRAPHS:

