TASK 2: Customer Churn Prediction Report

1. Introduction

This project utilizes the Telco Customer Churn dataset to analyze factors influencing churn and develop machine learning models to predict churn likelihood.

2. Dataset Overview

The dataset contains customer information, such as tenure, monthly charges, and subscription services. The target variable is **Churn**, which indicates whether a customer has left ("Yes") or stayed ("No").

Class Imbalance

- The dataset has a class imbalance (Non-Churn: 73.46%, Churn: 26.54%).
- SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the dataset.

3. Data Pre-processing

- **Handling Missing Values**: Any missing values were either filled or dropped.
- **Encoding Categorical Variables**:
  - Binary categorical values ("Yes"/"No") were encoded as 1 and 0.
  - "No internet service" responses were treated as "No".
- **Feature Scaling**: Continuous features were scaled to improve model performance.

4. Exploratory Data Analysis

Key Insights:

- Customers with **shorter tenure** have a higher likelihood of churn.
- High **monthly charges** correlate with increased churn probability.
- Customers subscribed to **Tech Support and Online Security** have a lower churn rate.
- **Senior citizens** are highly likely to churn.
- Customers **without a partner** have a **higher churn probability**.
- Customers **without dependents** also have a **higher churn probability**.
- **Electronic check payment** method has the **highest churn probability.**
- Internet Service Type Impact:
  - **Fibre optic** users have the **highest churn rate**.
- Service Subscription Impact on Churn:
  - **No Online Security → High churn**
  - **No Device Protection → High churn**
  - **No Tech Support → High churn**
  - **No Online Backup → High churn**

5. Model Selection & Training

Two models were trained and evaluated:

1. Logistic Regression (Baseline Model)

- **Accuracy:** 81.58%
- **Precision:** 86% (Non-Churn), 78% (Churn)
- **Recall:** 75% (Non-Churn), 88% (Churn)
- **F1-score:** 80% (Non-Churn), 83% (Churn)
- **Confusion Matrix:**
  - True Positives: 1370
  - False Positives: 390
  - False Negatives: 182
  - True Negatives: 1163

2. XGBoost (Optimized Model)

- **Accuracy:** 83.96%
- **Precision:** 84% (Non-Churn), 84% (Churn)
- **Recall:** 83% (Non-Churn), 84% (Churn)
- **F1-score:** 84% (Non-Churn), 84% (Churn)
- **Confusion Matrix:**
  - True Positives: 1311
  - False Positives: 257
  - False Negatives: 241
  - True Negatives: 1296

Model Comparison

Model	Accuracy	Precision (Churn)	Recall (Churn)	F1-Score (Churn)
Logistic Regression	81.58%	78%	88%	83%
XGBoost	**83.96%**	**84%**	**84%**	**84%**

6. Feature Importance Analysis

- **Key Features Influencing Churn:**
  - **Tenure**: Shorter tenure increases churn likelihood.
  - **Monthly Charges**: Higher charges lead to more churn.
  - **Online Security & Tech Support**: Customers using these services are less likely to churn.

7. Conclusion & Recommendations

Key Findings:

- XGBoost outperforms Logistic Regression with better accuracy and balanced precision/recall.
- Subscription to security and tech support services significantly reduces churn.
- Customers with high monthly charges tend to churn more.

Name: Divyanshu Kumar

CID: TI_JAD_#12369