Customer Churn Prediction Model for a Subscription Service

1. Introduction

Customer churn prediction is a crucial task for businesses to understand which customers are likely to leave and to develop effective retention strategies. In this project, we built a machine learning model to identify customers likely to churn using a Telco Customer Churn dataset.

2. Objectives.

- Analyze customer behavior and identify factors that contribute to churn.
- Develop and evaluate machine learning models to predict customer churn.
- Provide actionable insights to help reduce churn and improve customer retention.

3. Problem Statement

High customer churn can significantly impact a company's revenue and growth. By predicting which customers are at risk of churning, companies can take proactive steps to retain them and mitigate the cost of acquiring new customers.

4. Scope

We used a publicly available dataset containing customer information such as tenure, monthly charges, contract type, payment method, and whether they churned or not.

Dataset: Telco Customer Churn dataset from GitHub (WA_Fn-UseC_-Telco-Customer-Churn.csv).

5. Methodology

Data Acquisition: Load the dataset directly from GitHub.

Data Preprocessing & EDA: Clean the data, convert data types, handle missing values, encode categorical variables, scale numerical features and perform exploratory data analysis.

Model Training: Split the data into training and testing sets and train machine learning models (e.g., Logistic Regression, Random Forest).

Model Evaluation: Evaluate models using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

6. Implementation in Python

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Load dataset
url =
"https://raw.githubusercontent.com/dsrscientist/IBM_HR_Analytics_HR_Employee_Chu
rn/master/WA_Fn-UseC_-Telco-Customer-Churn.csv"
df = pd.read_csv(url)

# Preprocessing
df = df.drop(columns=['customerID'])  # Drop irrelevant column
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')  # Convert to
numeric
df.fillna(df['TotalCharges'].median(), inplace=True)  # Fill missing values

# Encode categorical variables
label_encoders = {}
for col in df.select_dtypes(include=['object']).columns:
    if col != 'Churn':
        le = LabelEncoder()
        df[col] = le.fit_transform(df[col])
        label_encoders[col] = le

# Convert target variable
df['Churn'] = df['Churn'].apply(lambda x: 1 if x == 'Yes' else 0)

# Split data
X = df.drop(columns=['Churn'])
y = df['Churn']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# Standardize numerical features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Train logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# Predictions
y_pred = model.predict(X_test)

# Evaluation
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
report = classification_report(y_test, y_pred)

# Display results
print(f"Accuracy: {accuracy:.2f}")
print("Confusion Matrix:")
print(conf_matrix)
print("Classification Report:")
print(report)
```

7. Model Training

We trained a Logistic Regression model to predict customer churn. Other models such as Random Forest and XGBoost were also considered.

8. Model Evaluation

The model was evaluated using Accuracy, Precision, Recall, and F1-score. The confusion matrix helped assess how well the model classified churners vs non-churners.

9. Findings

The analysis revealed that factors such as contract type, tenure, monthly charges, and service usage are highly correlated with customer churn. The machine learning models achieved reasonable accuracy, with Logistic Regression and Random Forest performing competitively.

Key insights from the model:

• Customers on month-to-month contracts are more likely to churn.

• Higher monthly charges correlate with increased churn.

• Customers using paperless billing showed a higher churn rate.

• Longer contract periods reduced churn significantly.

10. Conclusion

The churn prediction model provides valuable insights into customer behavior. It enables proactive retention efforts, such as targeted incentives for high-risk customers. Further improvements using advanced models and feature engineering could enhance accuracy. Accurate churn prediction enables targeted customer retention strategies. Companies should focus on customers with month-to-month contracts and high charges by offering personalized incentives or service improvements. Regular model updates and incorporating additional customer data can further enhance prediction accuracy.