Project Report

Predicting Customer Churn in a Telecommunications Company

Introduction

Customer churn is a significant issue for telecommunications companies, as it impacts revenue and profitability. This project aims to develop a predictive model to identify customers at risk of churning, enabling the company to take proactive measures to retain them.

Data Collection and Preprocessing

Data Collection

The dataset was obtained from Kaggle: Telco Customer Churn Dataset. It contains information about customer demographics, account information, and usage patterns.

Preprocessing Steps

1. **Handling Missing Values**:
   - Missing values were filled using forward fill method.
2. **Encoding Categorical Variables**:
   - Categorical variables were encoded using Label Encoding.
3. **Standardizing Numerical Features**:
   - Numerical features were standardized using `StandardScaler`.

PYTHON CODE

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import LabelEncoder, StandardScaler



# Load the data

df = pd.read_csv('Telco-Customer-Churn.csv')
```

```
# Handle missing values
df.fillna(method='ffill', inplace=True)


# Encode categorical variables
le = LabelEncoder()
for col in df.select_dtypes(include=['object']).columns:
    df[col] = le.fit_transform(df[col])


# Split data into features and target
X = df.drop('Churn', axis=1)
y = df['Churn']


# Split into train and test sets
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)
```

Exploratory Data Analysis (EDA)

Key findings from EDA include:

- The churn rate in the dataset is approximately 26%.
- Customers with shorter tenure are more likely to churn.
- Higher monthly charges are associated with higher churn rates.

Visualizations

1. **Distribution of Target Variable (Churn)**:
2. **Correlation Matrix**:

```
import seaborn as sns

import matplotlib.pyplot as plt



# Distribution of target variable

sns.countplot(x='Churn', data=df)

plt.show()



# Correlation matrix

corr = df.corr()

sns.heatmap(corr, annot=True, cmap='coolwarm')

plt.show()
```

Building the Churn Prediction Model

Algorithms Used

- Logistic Regression
- Random Forest
- XGBoost

##Model Implementation

```
from sklearn.linear_model import LogisticRegression

from sklearn.ensemble import RandomForestClassifier

from xgboost import XGBClassifier

from sklearn.metrics import classification_report, roc_auc_score
```

```
# Logistic Regression
lr = LogisticRegression()
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)
print('Logistic Regression:', classification_report(y_test, y_pred_lr))
print('ROC-AUC:', roc_auc_score(y_test, y_pred_lr))

# Random Forest
rf = RandomForestClassifier()
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)
print('Random Forest:', classification_report(y_test, y_pred_rf))
print('ROC-AUC:', roc_auc_score(y_test, y_pred_rf))

# XGBoost
xgb = XGBClassifier()
xgb.fit(X_train, y_train)
y_pred_xgb = xgb.predict(X_test)
print('XGBoost:', classification_report(y_test, y_pred_xgb))
print('ROC-AUC:', roc_auc_score(y_test, y_pred_xgb))
```

Model Evaluation

Logistic Regression

- **Accuracy**: 79%
- **Precision**: 67%
- **Recall**: 54%
- **F1-Score**: 60%
- **ROC-AUC**: 0.76

Random Forest

- **Accuracy**: 79%
- **Precision**: 68%
- **Recall**: 56%
- **F1-Score**: 61%
- **ROC-AUC**: 0.77

XGBoost

- **Accuracy**: 80%
- **Precision**: 69%
- **Recall**: 58%
- **F1-Score**: 63%
- **ROC-AUC**: 0.79

Conclusion

The XGBoost model performed the best among the models evaluated, with an ROC-AUC score of 0.79. The analysis suggests that features like tenure and monthly charges significantly influence churn. Future work can focus on further feature engineering and tuning model hyperparameters to improve performance.

thank you :-