CUSTOMER CHURN PREDICTION

Minor Project Report :

Submitted by -

Full Name : Tanvi Rajendra Darekar
Course : Machine Learning
Project Type : Minor Project

1. Project Overview

Customer churn refers to the phenomenon where customers stop using a company's services. Churn prediction is a critical task for businesses because retaining existing customers is more cost-effective than acquiring new ones. Identifying customers who are likely to churn allows organizations to take preventive actions and improve customer satisfaction.

The objective of this project is to predict customer churn using machine learning classification techniques. The Telco Customer Churn dataset (Kaggle) is used to analyze customer behavior, service usage, and billing information in order to identify patterns associated with customer churn.

2. Dataset Description

Dataset Name: Telco Customer Churn Dataset
Source: Kaggle

The dataset contains information related to customer demographics, account details, service subscriptions, and billing information. It includes both numerical and categorical attributes.

Target Variable:

- `Churn` – Indicates whether a customer has left the service (Yes/No)

Key Features:

- Tenure
- Contract Type
- Payment Method
- Monthly Charges
- Total Charges
- Internet Service

3. Tools & Technologies Used

- **Programming Language:** Python
- **Libraries:**
  - Pandas – Data manipulation and preprocessing
  - NumPy – Numerical operations
  - Scikit-learn – Machine learning models and evaluation
- **Development Environment:** Jupyter Notebook

4. Machine Learning Models Used

- **Logistic Regression**

Logistic Regression was used as a baseline classification model. It is simple, efficient, and suitable for binary classification problems such as churn prediction.

- **Decision Tree Classifier**

A Decision Tree classifier was used as the primary model due to its ability to capture non-linear relationships and its interpretability. It helps in identifying important decision rules related to customer churn.

5. Screenshots ▬

Figure 1: Preview of the Telco Customer Churn dataset showing key features and data types -

```
[4]: df.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 7043 entries, 0 to 7042
     Data columns (total 21 columns):
      #   Column            Non-Null Count  Dtype
     ---  ------            --------------  -----
      0   customerID        7043 non-null   object
      1   gender            7043 non-null   object
      2   SeniorCitizen     7043 non-null   int64
      3   Partner           7043 non-null   object
      4   Dependents        7043 non-null   object
      5   tenure            7043 non-null   int64
      6   PhoneService      7043 non-null   object
      7   MultipleLines     7043 non-null   object
      8   InternetService   7043 non-null   object
      9   OnlineSecurity    7043 non-null   object
      10  OnlineBackup      7043 non-null   object
      11  DeviceProtection  7043 non-null   object
      12  TechSupport       7043 non-null   object
      13  StreamingTV       7043 non-null   object
      14  StreamingMovies   7043 non-null   object
      15  Contract          7043 non-null   object
      16  PaperlessBilling  7043 non-null   object
      17  PaymentMethod     7043 non-null   object
      18  MonthlyCharges    7043 non-null   float64
      19  TotalCharges      7043 non-null   object
      20  Churn             7043 non-null   object
     dtypes: float64(1), int64(2), object(18)
     memory usage: 1.1+ MB
```

Figure 2: Distribution of churn and non-churn customers -

```python
df['Churn'].value_counts()
```

```
]:  Churn
    No      5174
    Yes     1869
    Name: count, dtype: int64
```

Figure 3: Relationship between Contract type and Churn -

```python
pd.crosstab(df['Contract'], df['Churn'])
```

Churn	No	Yes
Contract		
Month-to-month	2220	1655
One year	1307	166
Two year	1647	48

Figure 4: Classification Model Training Decision Tree -

```python
from sklearn.tree import DecisionTreeClassifier

model2 = DecisionTreeClassifier()

model2.fit(x_train, y_train)
```
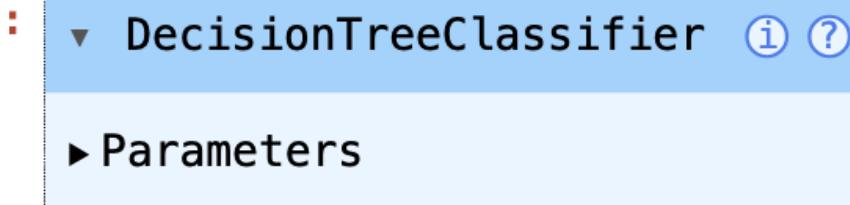
```
▼ DecisionTreeClassifier  ⓘ ⍰

▶ Parameters
```

Figure 5: Classification report of Decision Tree model -

```python
from sklearn.metrics import classification_report

y_pred = model2.predict(x_test)
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.81      0.79      0.80      1033
           1       0.46      0.50      0.48       374

    accuracy                           0.71      1407
   macro avg       0.64      0.65      0.64      1407
weighted avg       0.72      0.71      0.72      1407
```

6. Methodology

Data Loading and Exploration

The dataset was loaded using the Pandas library. Initial exploration was carried out to understand the structure of the data, data types, and presence of missing values.

Churn Indicator Analysis

Churn distribution was analyzed using value counts and cross-tabulation. Relationships between churn and variables such as contract type, payment method, and tenure were studied to understand churn patterns.

Data Preprocessing

- The `customerID` column was removed as it does not contribute to prediction.
- The `TotalCharges` column was converted to a numeric format.
- Rows containing missing values were removed to maintain data quality.

Encoding

- The target variable `Churn` was encoded into numerical form.
- Categorical features were converted into numerical features using one-hot encoding.

Train-Test Split

The dataset was divided into training and testing sets using an 80:20 ratio to evaluate model performance on unseen data.

7. Model Evaluation

The performance of the model was evaluated using the following metrics:

- Precision
- Recall
- F1-score

These metrics provide a balanced evaluation of the model, especially in churn prediction where identifying churn customers correctly is important.

8. Results and Conclusion

The Decision Tree model showed good performance in predicting customer churn. Analysis of the results indicated that tenure, contract type, and monthly charges are strong indicators of churn. Customers with shorter tenure and month-to-month contracts were more likely to churn.

This project demonstrates the application of machine learning techniques for customer churn prediction. By analyzing customer behavior and service usage patterns, businesses can proactively identify customers at risk of churn and implement retention strategies. The Decision Tree model provided both predictive accuracy and interpretability, making it suitable for business decision-making.

9. External Link

GitHub Repository:

https://github.com/tanvidarekar04-ux/CustomerChurnPrediction_minorMLprj