

TITLE: CUSTOMER CHURN PREDICTION

Step-1: Problem Definition

- The aim of this project is to perform a thorough analysis of customer churn prediction data, with a primary emphasis on understanding and forecasting customer attrition in a business context.
- The ultimate goal is to offer actionable insights that can empower businesses to develop effective retention strategies, enhance customer satisfaction, and reduce churn rates.
- This comprehensive project encompasses stages like data collection, data preprocessing, exploratory data analysis (EDA), statistical analysis, visualization, and the development of practical recommendations for minimizing customer churn and boosting overall business performance.

Step 2: Data Collection

- We will collect customer churn prediction data from reputable sources pertinent to our industry, such as internal databases, customer records, and transaction histories.
- The primary data source for this project will be our company's customer database, encompassing historical customer interactions, subscription details, and churn-related information.
- Daily data updates will be drawn from our internal customer records, allowing us to maintain a real-time understanding of churn.
- Subsequently, this data will be merged with additional information, such as demographic data, to enrich our analysis and develop a more comprehensive customer churn prediction model.

```
#import all relevant libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msno
import warnings
warnings.filterwarnings("ignore")
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
from sklearn.preprocessing import LabelEncoder
```

```
#Loading the dataset
data=pd.read_csv("C:\Users\HP\Downloads\Telco-Customer-Churn.csv")
```

```
data.head()
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	Device
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No

```
data.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
```

```
data.describe()
```

	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

```
data[[col for col in data.columns.difference(num_cols) if col != 'seniorcitizen']]
.describe().T
```

	count	unique	top	freq
churn	7043	2	No	5174
contract	7043	3	Month-to-month	3875
dependents	7043	2	No	4933
deviceprotection	7043	3	No	3095
gender	7043	2	Male	3555
internetservice	7043	3	Fiber optic	3096
multiplelines	7043	3	No	3390
onlinebackup	7043	3	No	3088
onlinesecurity	7043	3	No	3498
paperlessbilling	7043	2	Yes	4171
partner	7043	2	No	3641
paymentmethod	7043	4	Electronic check	2365
phoneservice	7043	2	Yes	6361
streamingmovies	7043	3	No	2785
streamingtv	7043	3	No	2810

Step 3: Data Preprocessing

- Data cleaning and preprocessing constitute crucial phases in the preparation of data for analysis.
- During this stage, we will address data issues including duplicate entries, inconsistent formats, the treatment of missing values, and the conversion of categorical variables into numerical formats.

```
data.dtypes
```

```
customerID      object
gender          object
SeniorCitizen   int64
Partner         object
Dependents      object
tenure          int64
PhoneService    object
MultipleLines   object
InternetService object
OnlineSecurity  object
OnlineBackup    object
DeviceProtection object
TechSupport     object
StreamingTV     object
StreamingMovies object
Contract        object
PaperlessBilling object
PaymentMethod   object
MonthlyCharges  float64
TotalCharges    object
Churn           object
dtype: object
```

```
data.isnull().sum()
```

```
customerID      0
gender          0
SeniorCitizen   0
Partner         0
Dependents      0
tenure          0
PhoneService    0
MultipleLines   0
InternetService 0
OnlineSecurity  0
OnlineBackup    0
DeviceProtection 0
TechSupport     0
StreamingTV     0
StreamingMovies 0
Contract        0
PaperlessBilling 0
PaymentMethod   0
MonthlyCharges  0
TotalCharges    0
Churn           0
dtype: int64
```

Step 4: Data Exploration

- Engage in exploratory data analysis (EDA) to grasp the dataset's characteristics, patterns, and relationships.
- In this phase, the objective is to delve into the data to comprehend its nuances. EDA entails computing statistical summaries, creating data visualizations, and recognizing patterns and anomalies.
- Key areas of exploration encompass customer demographics, historical usage patterns, and the impact of various features on churn predictions.
- Utilize visualizations to gain insights into the distribution of key features and the identification of influential factors affecting customer churn.

```
#data cleaning  data transformation  data reduction
#drop irrelevant variables
data=data.drop(['CustomerId'],axis=1)
#identifying and treating missing values
data.isnull().sum()
data=data.fillna(0)

data.head()
```

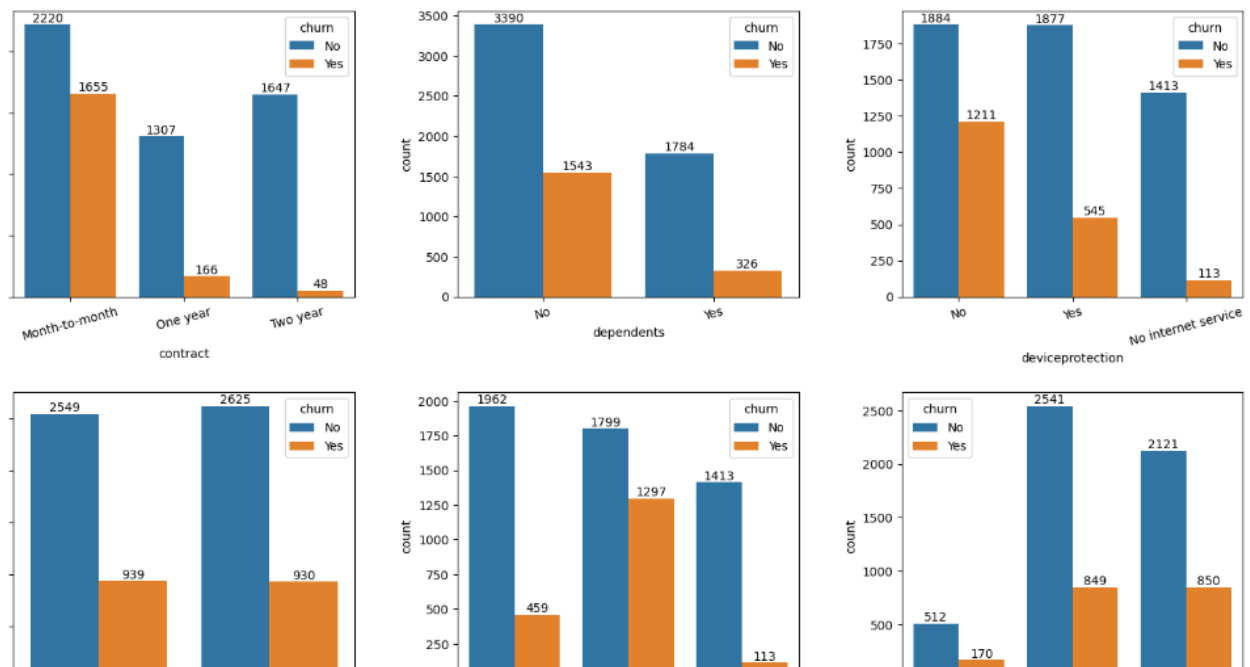
	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport
0	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No
1	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No
2	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No
3	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes
4	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No



```

num_cols = ['tenure','monthlycharges', 'totalcharges']
label="Churn"
plt.figure(figsize = (15, 26))
for i, col in enumerate(data.columns.difference(num_cols)[1:]):
    plt.subplot(6, 3, i+1)
    ax = sns.countplot(data, x = col, hue = label)
    ax.bar_label(ax.containers[0])
    ax.bar_label(ax.containers[1])
    plt.xticks(rotation = 15)
plt.tight_layout()
plt.show()

```



```
plt.figure(figsize = (4,4))  
ax = sns.countplot(data, x = label)  
ax.bar_label(ax.containers[0])  
plt.show()
```

