Table of Contents

ABSTRACT

The issue of customer retention is very important in any business setting, most especially a telecom industry with millions of customers.

Customer churn in the telecom industry usually describes a situation where a customer stops the service of one telecom company during the contract and switches to a competitor to obtain a better, cheaper and more satisfactory service for the customer's needs. (Liu et al., 2022)

This research aims to identify the key elements that lead to a customer leaving a telecom provider. Additionally, this initiative aims to offer business advice to the telecom company on how to reduce the rate of customer churn using data from statistical analysis.

In order to achieve the aims of the project, sample data from a telecom provider of phone and home services in California was gathered. This data included crucial facts and demographics about customers who used the company's services for a while before leaving. With this data, descriptive statistics were used to examine and develop understanding of the factors influencing customer churn.

Section A

Data collection

The data collected was gotten from https://www.kaggle.com/datasets/blastchar/telco-customer-churn,Here we will be using python, specifically Anaconda.

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

This dataset contains the information of 7043 customers with 21 features

```
<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      7032 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

Figure 1

Data cleaning

Some missing values were discovered when the descriptive summary of the data was reviewed using Panda's info () function. As seen in fig. 1 the feature 'TotalCharges' contains 11 missing values.

Looking at missing values the percentage of rows with missing values is less than 1%. With this we can safely remove the rows with missing values using the pandas drop () method.

On checking through each variable, the customerID variable has high variability (all records in this feature are unique). This would offer no insight to any further statistical analysis, so the column was safely dropped with pandas drop () function.

The ‘Churn’ feature is an object type with two unique values ‘Yes’ and ‘No’. The ‘Yes’ value was replaced with 1 while the ‘No’ value was replaced with 0. This converts the datatype to integer type and aids further statistical analysis.

Section B

Data Analysis

The first step taken was to divide the dataset into categorical and continuous variable. To make this segregation the unique values in each feature was observed. The features with less than 50 unique values were categorized as a categorical feature while the rest were categorized as continuous features.

This was done because the numerical analysis for continuous features is different from categorical function

```
df_category=pd.DataFrame()
df_float=pd.DataFrame()
for x in df.columns:
    if df[x].nunique()>50:
        df_float[x]=df[x]
    else:
        df_category[x]=df[x]

```

Figure 2 separating categorical features from continuous features

Analysis of continuous data

The first step to the analysis of continuous data was to obtain the statistical summary of the data

	tenure	MonthlyCharges	TotalCharges	Churn
count	7032.000000	7032.000000	7032.000000	7032.000000
mean	32.421786	64.798208	2283.300441	0.265785
std	24.545260	30.085974	2266.771362	0.441782
min	1.000000	18.250000	18.800000	0.000000
25%	9.000000	35.587500	401.450000	0.000000
50%	29.000000	70.350000	1397.475000	0.000000
75%	55.000000	89.862500	3794.737500	1.000000
max	72.000000	118.750000	8684.800000	1.000000

Figure 3 statistical summary of continuous data

From table 3 the mean, max, 25th percentile, 50th percentile and 75th percentile was easily gotten using the pandas describe () function.

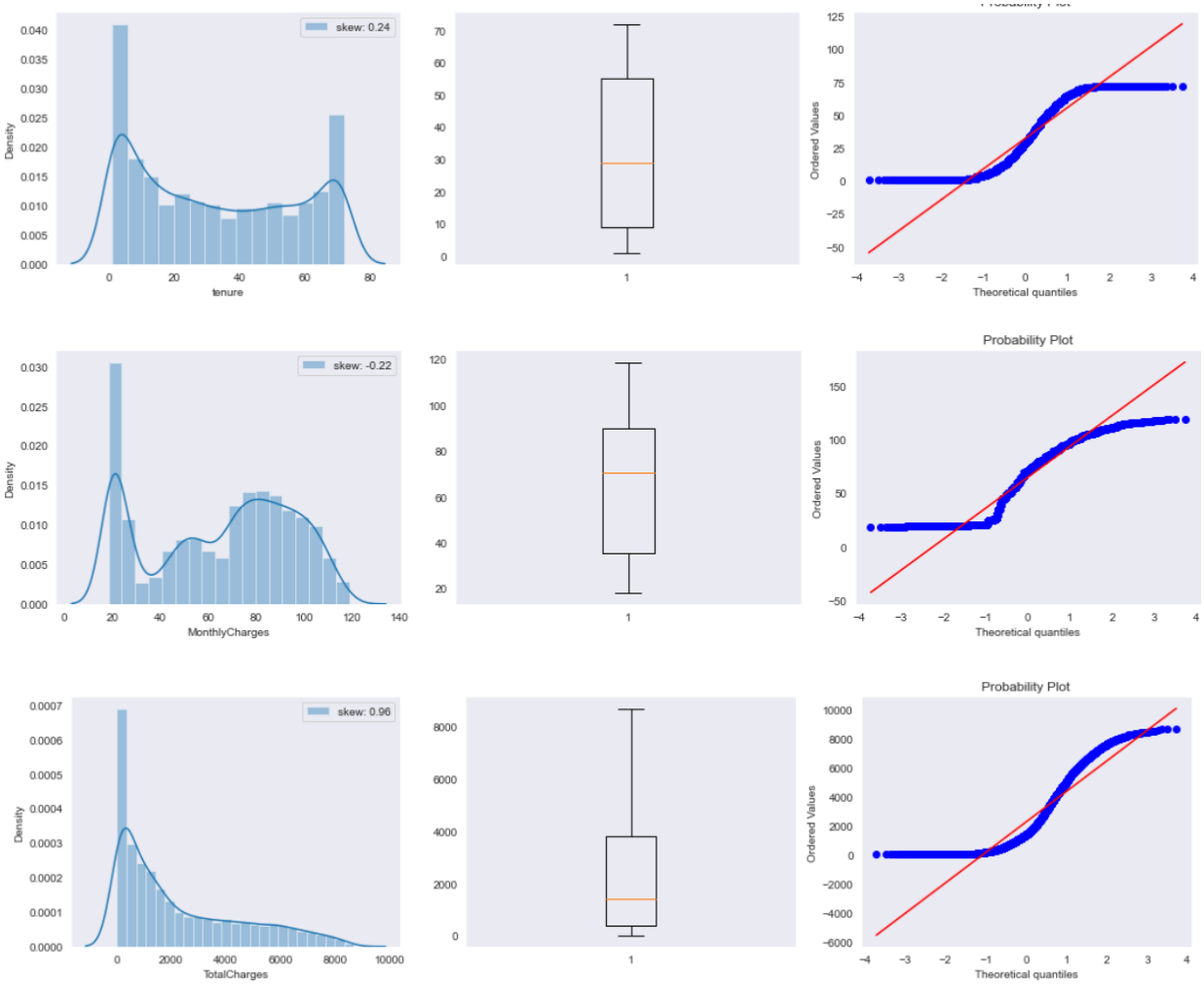
From statistical summary, the presence of outlier was hinted in the TotalCharges column because the difference between the mean and medium appears to be significant.

Outlier detection

To confirm the presence of outliers, the skewness, distribution, top percentile and lower percentiles were checked.

Data Distribution

The boxplots and distribution were plotted with the aid of the seaborn library



On observing the boxplot and distribution of the TotalCharges feature, the indication for an outlier was further indicated but not yet confirmed.

to confirm the presence of outliers, the top 5 percentile and bottom and 5 percentile was observed.

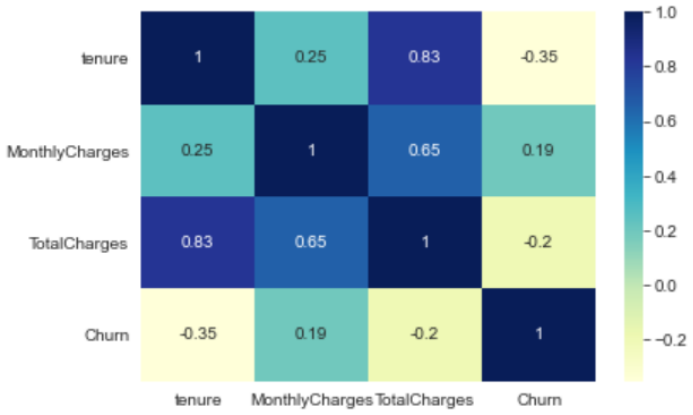
```
0.05      49.6050
0.10      84.6000
0.25     401.4500
0.50    1397.4750
0.75    3794.7375
0.90    5976.6400
0.95    6923.5900
0.96    7147.7060
0.97    7414.1485
0.98    7721.0960
0.99    8039.8830
1.00    8684.8000
Name: TotalCharges, dtype: float64
```

From the observation of the percentiles, it was concluded that there were no outliers in the TotalCharges columns.

Bivariate Analysis

Correlation

Correlation was used to check the strength of the linear relationship between each feature in the continuous dataset.



It was noticed that tenure and TotalCharges was highly positively correlated. We can infer that as the TotalCharges of a customer increases, the tenure of the customer increases.

No further inference could be made from the continuous data, so it was converted to categorical data by binning each feature into 2 groups (high and low).

Analysis of Categorical Data

The first step was to get the statistical summary of the data

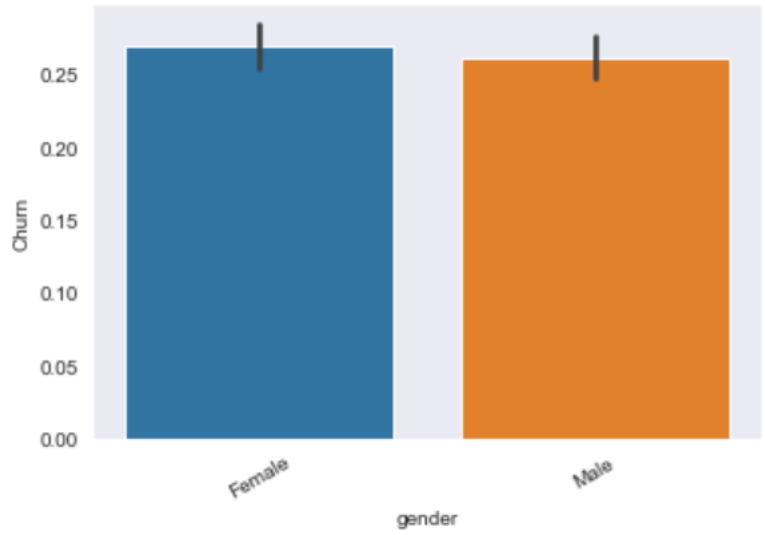
	count	unique	top	freq
gender	7032	2	Male	3549
Partner	7032	2	No	3639
Dependents	7032	2	No	4933
MultipleLines	7032	3	No	3385
InternetService	7032	3	Fiber optic	3096
OnlineSecurity	7032	3	No	3497
OnlineBackup	7032	3	No	3087
DeviceProtection	7032	3	No	3094
TechSupport	7032	3	No	3472
StreamingTV	7032	3	No	2809
StreamingMovies	7032	3	No	2781
Contract	7032	3	Month-to-month	3875
PaperlessBilling	7032	2	Yes	4168
PaymentMethod	7032	4	Electronic check	2365
tenure	7032	2	Low tenure	3558
MonthlyCharges	7032	2	Low MonthlyCharges	3519
TotalCharges	7032	2	Low TotalCharges	3516

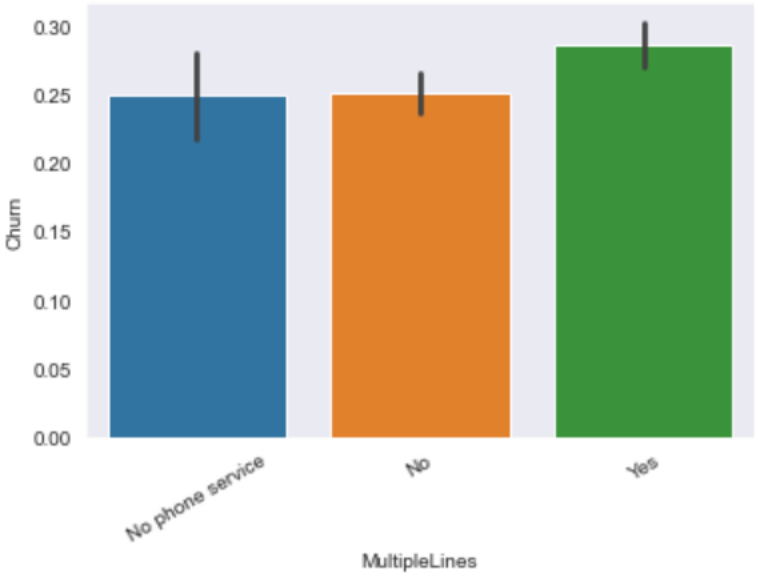
The statistical summary revealed the mode, its frequency and the number of unique features in each column.

Analysis

Using each feature as reference, the customers were divided into churned and non-churned.

This was to see how much each feature segregates the churned customers from the non-churned customers.





The figures above showed the churn segregation of customers based on gender and multiple lines. It could be inferred that these feature on their own does not create a clear distinction for analyzing customer behavior.

Feature engineering with these types of features could help gain better insights.

Top features influencing customer's churn

Using each variable, the data could be segregated into churn and non-churn. To get the top features influencing churn rate, the percentage of churn and non-churn for each unique values in a feature was gotten.

The features with the maximum difference between the churn percentage and non-churn percentage was gotten. The features gotten are the features that does a good job in segregating the churned customers from the non-churned customers.

Dashboard

The dashboard for this project was created with the use of python Dash visualization library.

For continuous data, the dashboard displays the scatterplots and the boxplots. For categorical data the dashboard displays the bar plots and grouped bar plots.

Running the dashboard application

To run the dashboard, the following steps should be taken

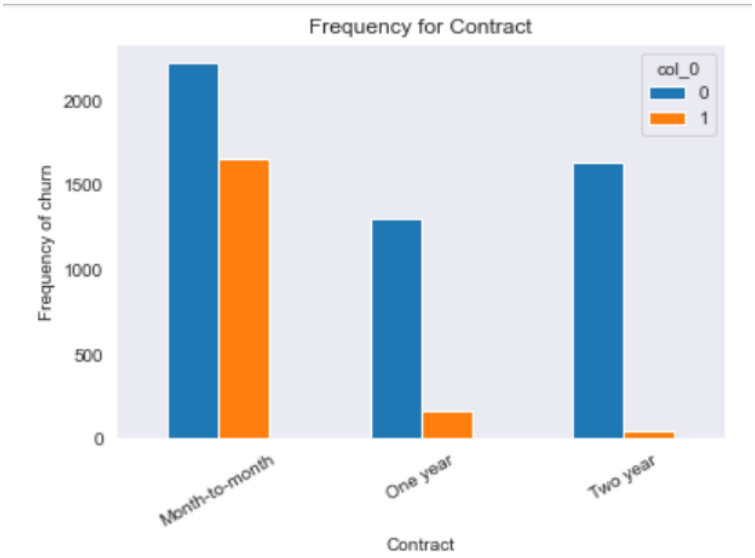
1. Run "pip install requirements.txt" : this installs the dash library
2. From the project folder run "python dashboard.py"
3. The local server link would be displayed on the terminal, copy it
4. Paste and run the link on a browser to view dashboard

SECTION C

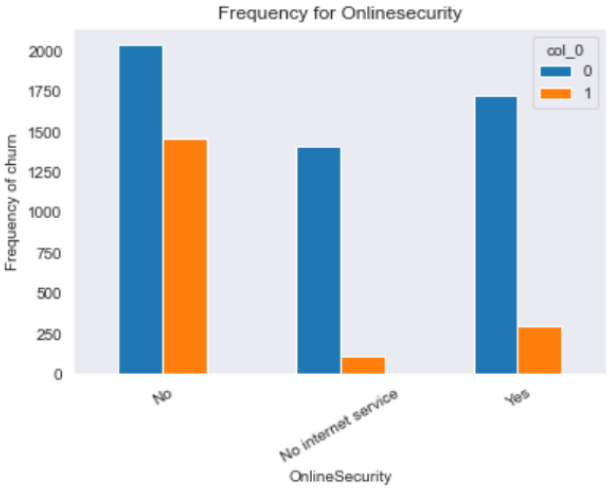
Results and conclusion

On segregating the data based on each feature in the dataset, the percentage of churn in each segregated data was then calculated. The variability in churn rate of sub dataset divided by each feature was used to analyze the top factors influencing churn rate.

The figure below shows the top 4 influencing factors with the churn percentage



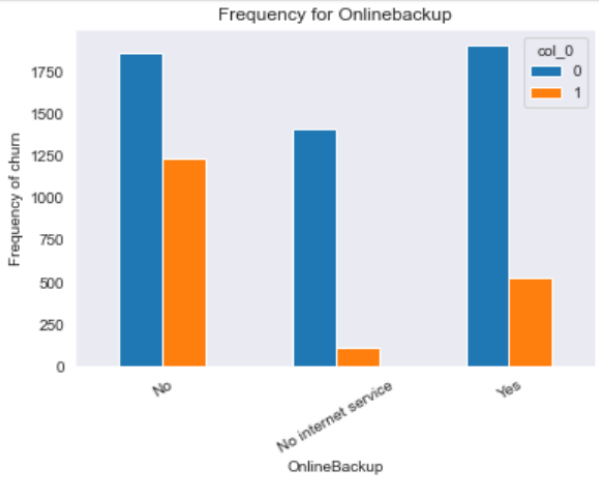

```
Month-to-month
of 3875 Month-to-month in Contract 42.7% churned


One year
of 1472 One year in Contract 11.3% churned


Two year
of 1685 Two year in Contract 2.8% churned
```



No
of 3497 No in OnlineSecurity 41.8% churned

Yes
of 2015 Yes in OnlineSecurity 14.6% churned

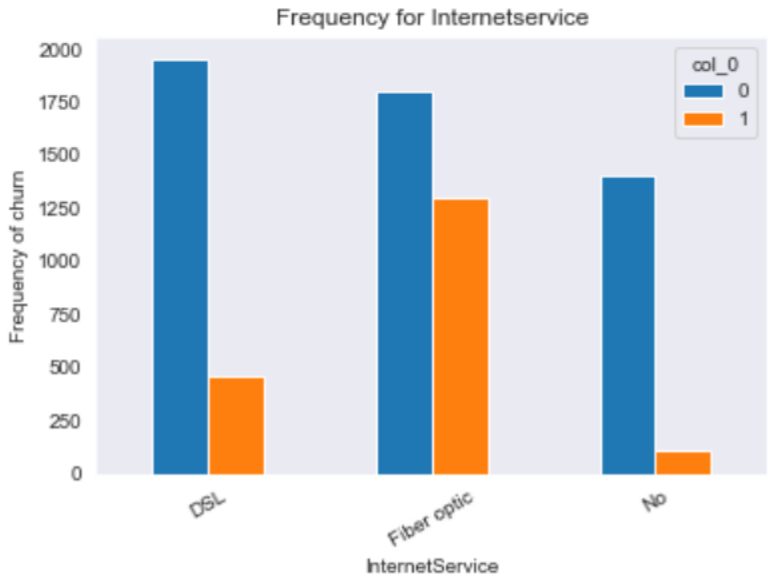
No internet service
of 1520 No internet service in OnlineSecurity 7.4% churned



Yes
of 2425 Yes in OnlineBackup 21.6% churned

No
of 3087 No in OnlineBackup 39.9% churned

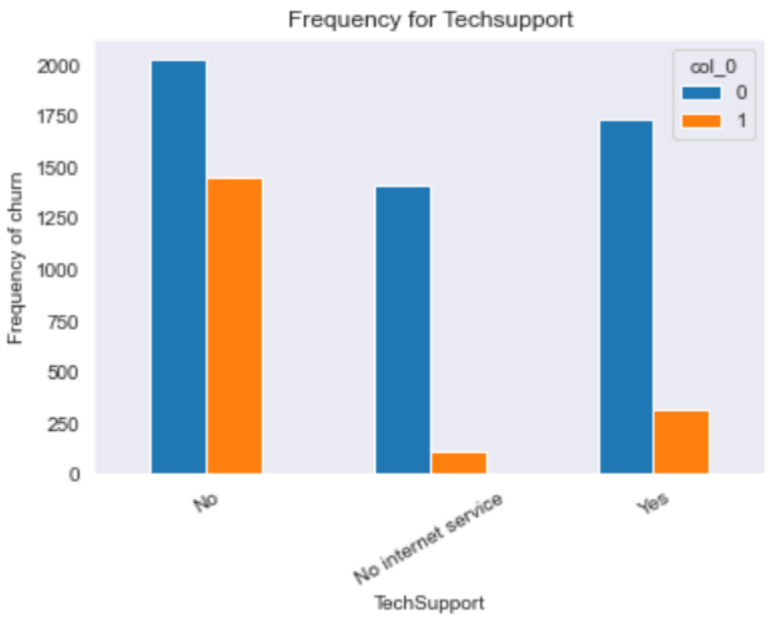
No internet service
of 1520 No internet service in OnlineBackup 7.4% churned



DSL
of 2416 DSL in InternetService 19.0% churned

Fiber optic
of 3096 Fiber optic in InternetService 41.9% churned

No
of 1520 No in InternetService 7.4% churned




```
No
of 3472 No in TechSupport 41.6% churned


Yes
of 2040 Yes in TechSupport 15.2% churned


No internet service
of 1520 No internet service in TechSupport 7.4% churned
```

The features internet service, online backup, online security, contract and tech support were determined to be the most influencing features for customer churn.

The following conclusion could be made from the results gotten:

1. Customers on Month-to-month contract are 39% more likely to churn when compared with customers on 2-year contract.

2. customers on Fiber optic internet service are 30% more likely to churn when compared with customers with no internet service or DSL.

3. customers with no Tech support are 34% more likely to churn when compared with customers with Tech support.

4. customers with no online backup are 32% more likely to churn when compared with customers with no online backup.

Business insights/ Recommendation

- Add more discount to the 1 and 2 years plans to enable more people subscribe to them

- Create plans and strategies to make customers depend more online backup

- Move customers from fiberoptics to DSL as there might be persistent issues with the fiber optics internet connection

- Create Tech support awareness to customers

- Provide customers with more assurance of online security

BIBLIOGRAPHY

Liu, Y., Fan, J., Zhang, J., Yin, X. and Song, Z. (2022). Research on telecom customer churn prediction based on ensemble learning. *Journal of Intelligent Information Systems*. doi:10.1007/s10844-022-00739-z.