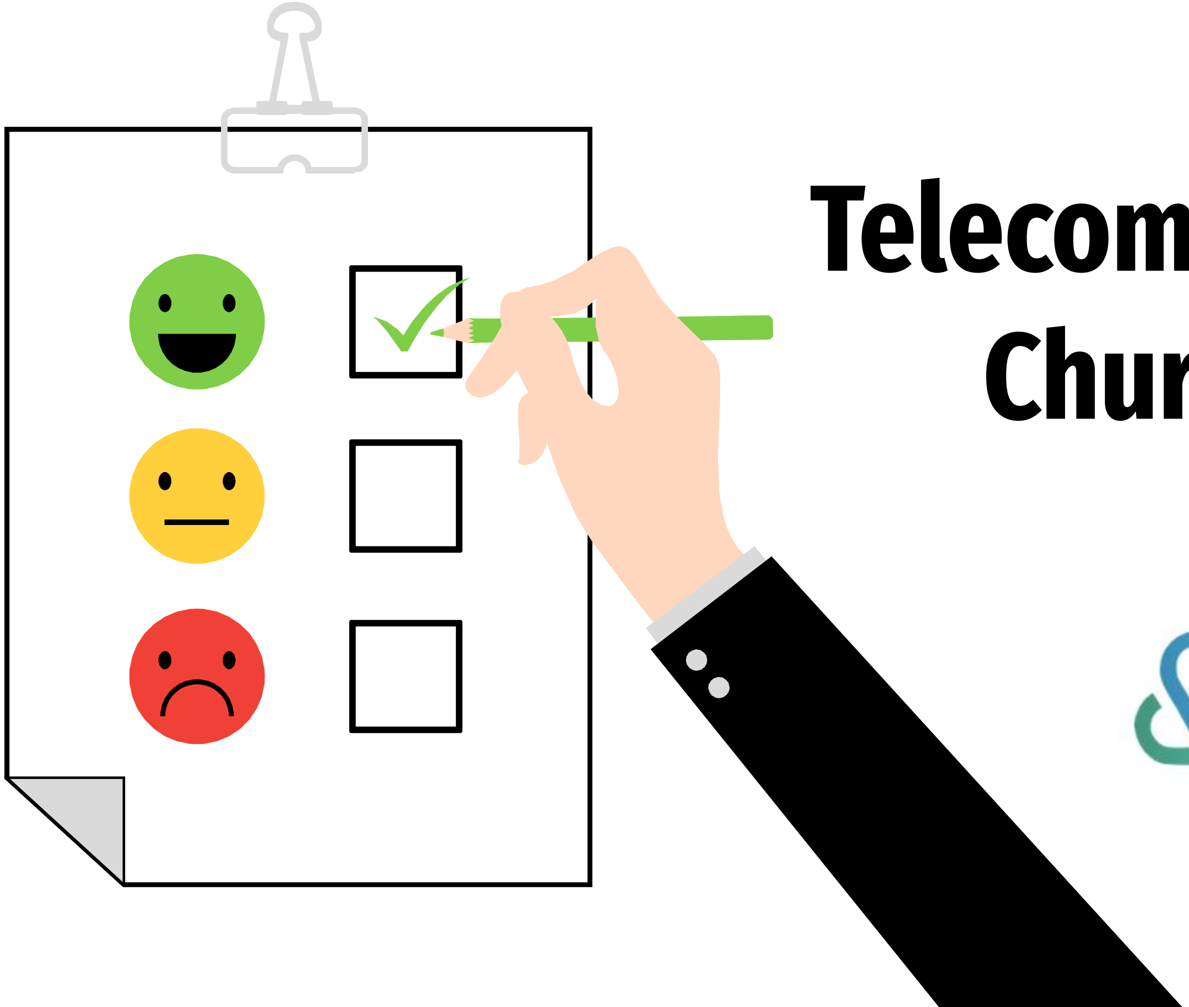


Telecom Customer Churn Analysis

Yoga Aditia Nugroho

ibimbing

Outline

Definition of and Main Cause of Customer Churn 1

Objective, Solution, and Data Introduction 2

Flow of Data Processing 3

Result of Processing and Machine Learning installation 4

Business Insight and and Summary 5

Definition and Main Cause of Customer Churn

Definition

Customer churn is the percentage of customers that stopped using your company's product or service during a certain time frame.

For example, if you start your quarter with 100 customers and end with 95, your churn rate is 5% because you lost 5% of your customers.

The Main Cause of Customer Churn

There are several condition the customer categorical being churn or not churn the motives are

- Poor customer service
- Nonexistent or failed onboarding
- Lack of perceived value
- Poor market fit
- Involuntary churn
- Switch to competitor

Objective and Solution

Using Machine Learning for predict how the user churn and what is the most influence from it based on their variable and analysis for business

Objective and Solution

The customer churn analysis are good for the telecommunication company to know the motive and what is the most variable affect the churn and how to maintain it.

Business Problem

Data Intoduction

"Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs." [IBM Sample Data Sets]

Content

Each row represents a customer, each column contains customer's attributes described on the column Metadata.

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

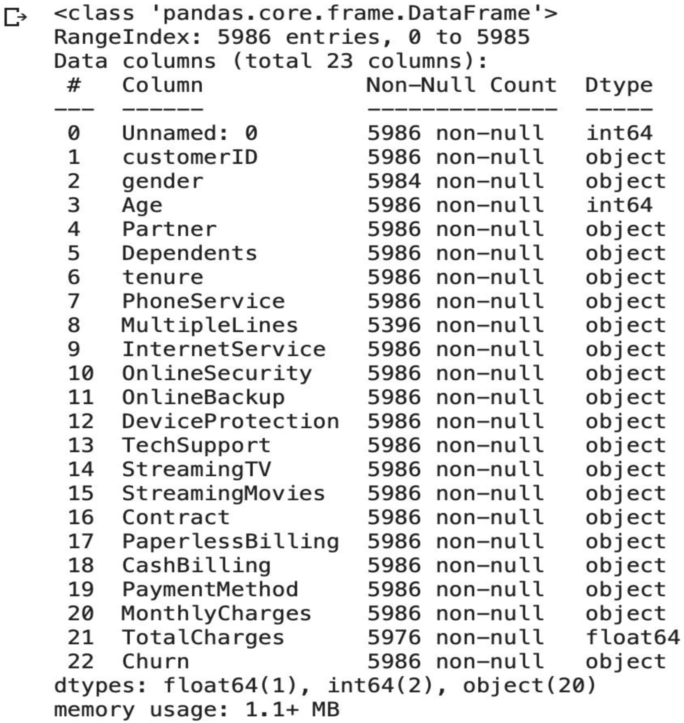
Flow of Data Processing

1 ★ Cleaning, Data Manipulation, Scaling the Data, and Check Correlation

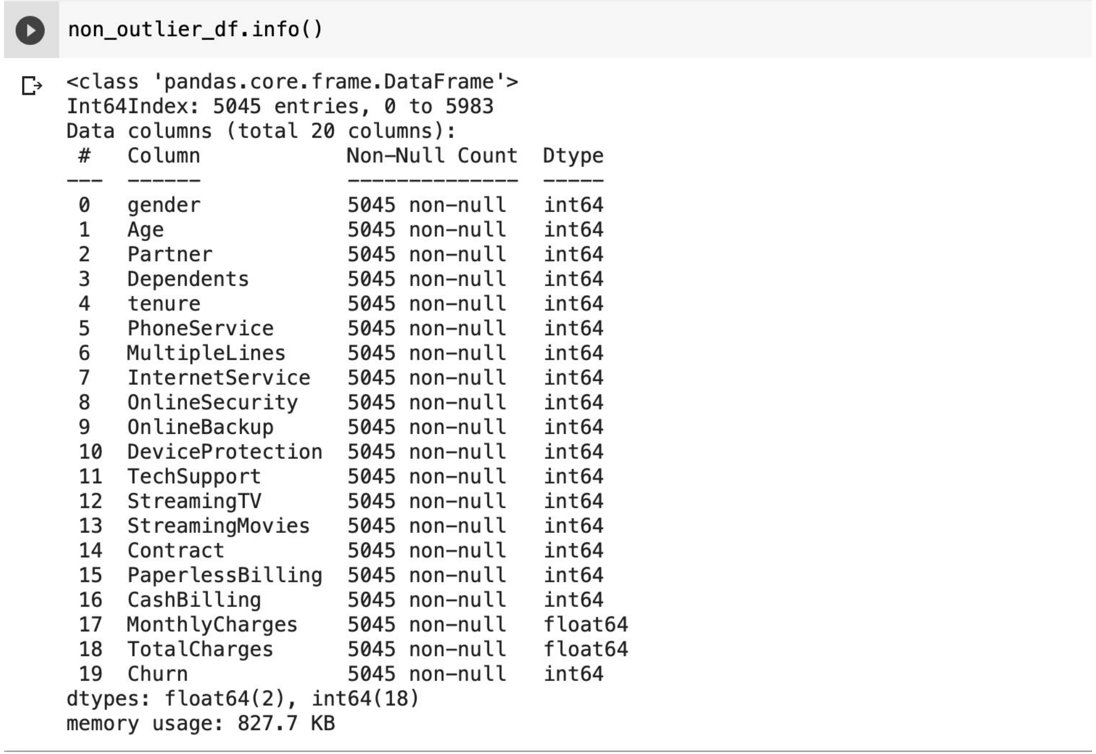
2 ★ Creating Logistic Regression Model

3 ★ Result, Business Insight, and Summary

Cleaning Data and Transform the Data



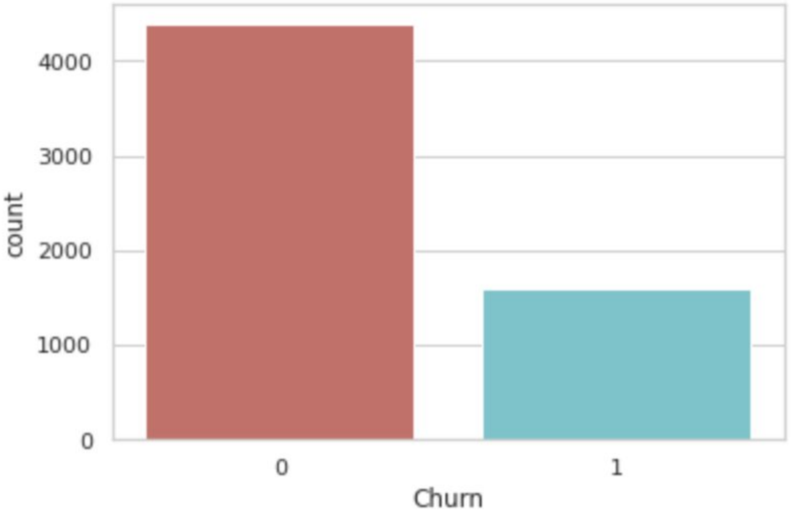
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5986 entries, 0 to 5985
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   Unnamed: 0        5986 non-null   int64  
 1   customerID        5986 non-null   object 
 2   gender            5984 non-null   object 
 3   Age               5986 non-null   int64  
 4   Partner           5986 non-null   object 
 5   Dependents        5986 non-null   object 
 6   tenure            5986 non-null   object 
 7   PhoneService      5986 non-null   object 
 8   MultipleLines     5396 non-null   object 
 9   InternetService   5986 non-null   object 
 10  OnlineSecurity    5986 non-null   object 
 11  OnlineBackup      5986 non-null   object 
 12  DeviceProtection  5986 non-null   object 
 13  TechSupport       5986 non-null   object 
 14  StreamingTV       5986 non-null   object 
 15  StreamingMovies   5986 non-null   object 
 16  Contract          5986 non-null   object 
 17  PaperlessBilling  5986 non-null   object 
 18  CashBilling       5986 non-null   object 
 19  PaymentMethod     5986 non-null   object 
 20  MonthlyCharges    5986 non-null   object 
 21  TotalCharges      5976 non-null   float64
 22  Churn             5986 non-null   object 
dtypes: float64(1), int64(2), object(20)
memory usage: 1.1+ MB
```



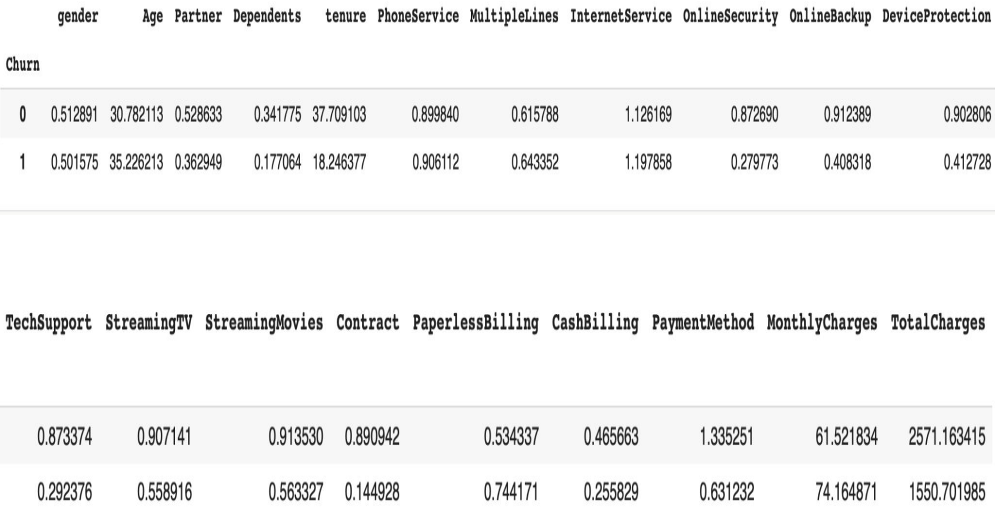
```
non_outlier_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5045 entries, 0 to 5983
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   gender            5045 non-null   int64  
 1   Age               5045 non-null   int64  
 2   Partner           5045 non-null   int64  
 3   Dependents        5045 non-null   int64  
 4   tenure            5045 non-null   int64  
 5   PhoneService      5045 non-null   int64  
 6   MultipleLines     5045 non-null   int64  
 7   InternetService   5045 non-null   int64  
 8   OnlineSecurity    5045 non-null   int64  
 9   OnlineBackup      5045 non-null   int64  
 10  DeviceProtection  5045 non-null   int64  
 11  TechSupport       5045 non-null   int64  
 12  StreamingTV       5045 non-null   int64  
 13  StreamingMovies   5045 non-null   int64  
 14  Contract          5045 non-null   int64  
 15  PaperlessBilling  5045 non-null   int64  
 16  CashBilling       5045 non-null   int64  
 17  MonthlyCharges    5045 non-null   float64
 18  TotalCharges      5045 non-null   float64
 19  Churn             5045 non-null   int64  
dtypes: float64(2), int64(18)
memory usage: 827.7 KB
```

Result of Data

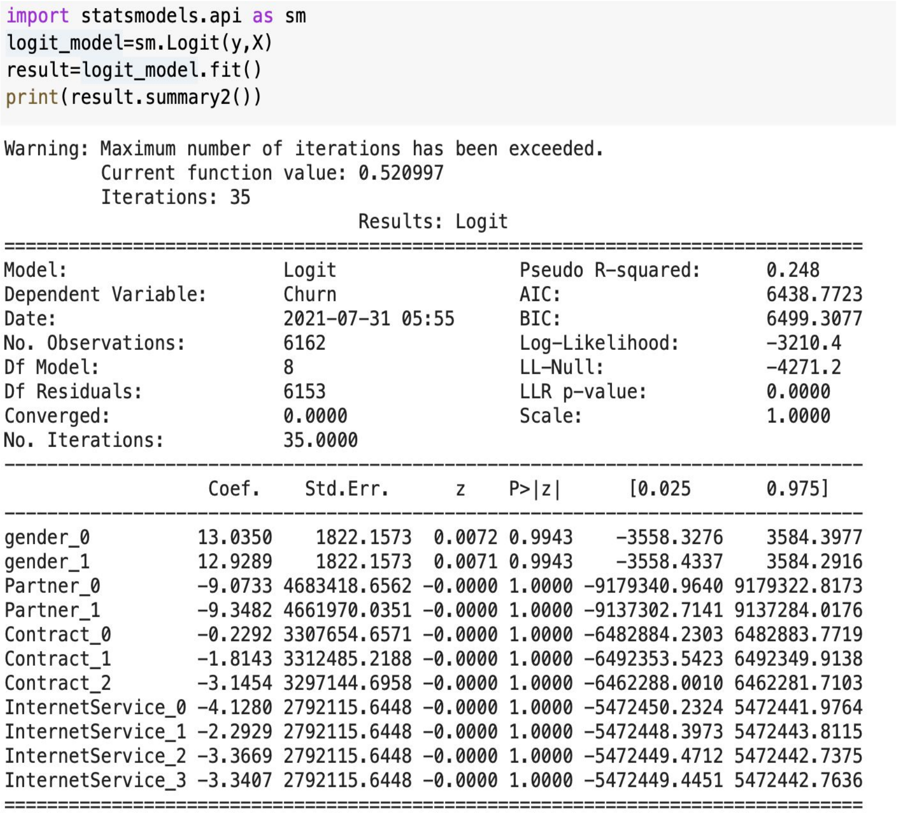


```
percentage of no churn is 73.41708542713567
percentage of churn 26.58291457286432
<Figure size 432x288 with 0 Axes>
```



Churn	gender	Age	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	CashBilling	PaymentMethod	MonthlyCharges	TotalCharges
0	0.512891	30.782113	0.528633	0.341775	37.709103	0.899840	0.615788	1.126169	0.872690	0.912389	0.902806	0.873374	0.907141	0.913530	0.890942	0.534337	0.465663	1.335251	61.521834	2571.163415
1	0.501575	35.226213	0.362949	0.177064	18.246377	0.906112	0.643352	1.197858	0.279773	0.408318	0.412728	0.292376	0.558916	0.563327	0.144928	0.744171	0.255829	0.631232	74.164871	1550.701985

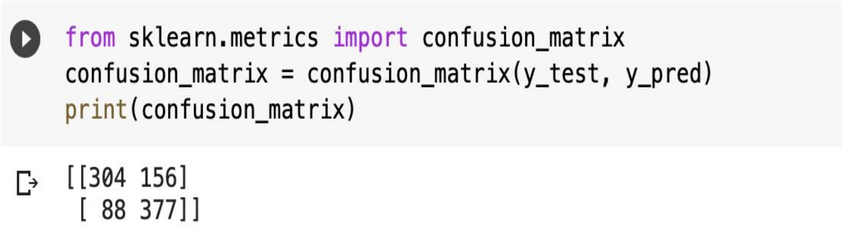
Logistic Regression Model



```
import statsmodels.api as sm
logit_model=sm.Logit(y,X)
result=logit_model.fit()
print(result.summary2())
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.520997
         Iterations: 35
                                 Results: Logit
================================================================================
Model:                    Logit                  Pseudo R-squared:       0.248
Dependent Variable:       Churn                  AIC:                    6438.7723
Date:                     2021-07-31 05:55       BIC:                    6499.3077
No. Observations:         6162                   Log-Likelihood:         -3210.4
Df Model:                 8                      LL-Null:                -4271.2
Df Residuals:             6153                   LLR p-value:            0.0000
Converged:                0.0000                 Scale:                  1.0000
No. Iterations:           35.0000
--------------------------------------------------------------------------------
                    Coef.    Std.Err.      z     P>|z|      [0.025       0.975]
--------------------------------------------------------------------------------
gender_0           13.0350    1822.1573  0.0072 0.9943    -3558.3276    3584.3977
gender_1           12.9289    1822.1573  0.0071 0.9943    -3558.4337    3584.2916
Partner_0          -9.0733 4683418.6562 -0.0000 1.0000 -9179340.9640 9179322.8173
Partner_1          -9.3482 4661970.0351 -0.0000 1.0000 -9137302.7141 9137284.0176
Contract_0         -0.2292 3307654.6571 -0.0000 1.0000 -6482884.2303 6482883.7719
Contract_1         -1.8143 3312485.2188 -0.0000 1.0000 -6492353.5423 6492349.9138
Contract_2         -3.1454 3297144.6958 -0.0000 1.0000 -6462288.0010 6462281.7103
InternetService_0 -4.1280 2792115.6448 -0.0000 1.0000 -5472450.2324 5472441.9764
InternetService_1 -2.2929 2792115.6448 -0.0000 1.0000 -5472448.3973 5472443.8115
InternetService_2 -3.3669 2792115.6448 -0.0000 1.0000 -5472449.4712 5472442.7375
InternetService_3 -3.3407 2792115.6448 -0.0000 1.0000 -5472449.4451 5472442.7636
================================================================================
```

Check confusion matrix

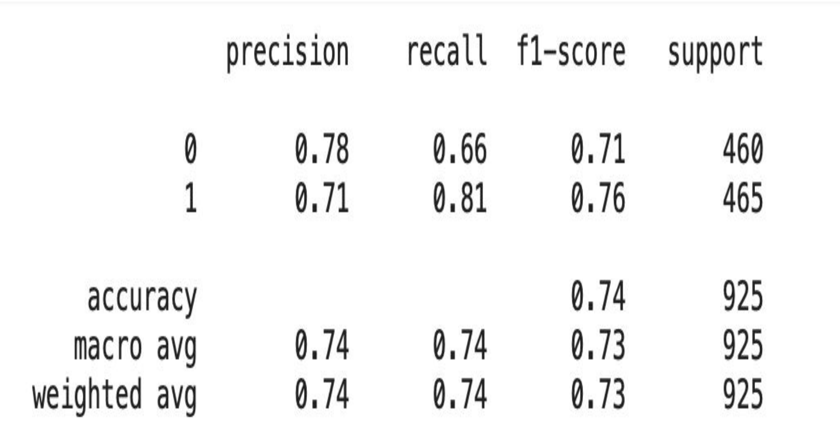


```
from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(y_test, y_pred)
print(confusion_matrix)
```

```
[[304 156]
 [ 88 377]]
```

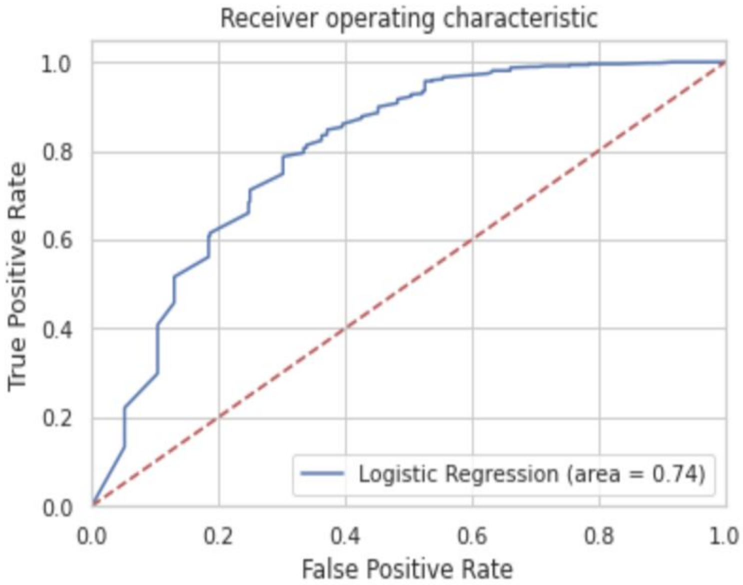
From this data I have 681 data correct and 244 incorrect data

The Result of Logistic Regression

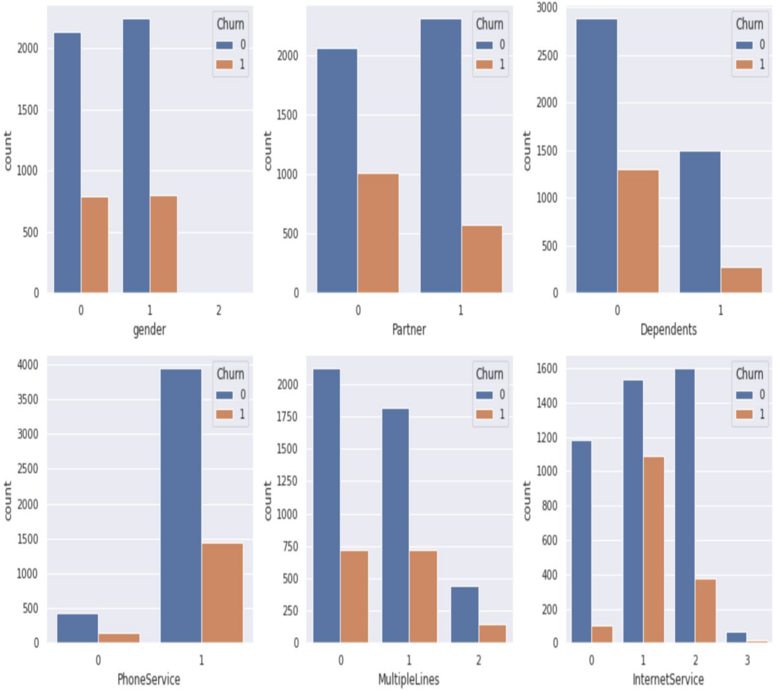


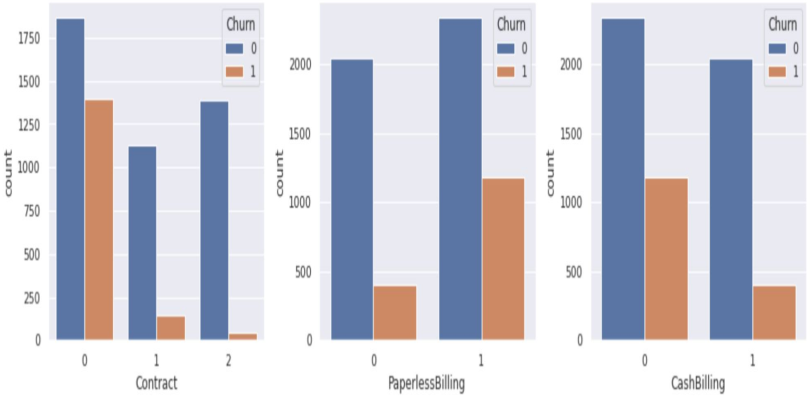
```
              precision    recall  f1-score   support

           0       0.78      0.66      0.71       460
           1       0.71      0.81      0.76       465

    accuracy                           0.74       925
   macro avg       0.74      0.74      0.73       925
weighted avg       0.74      0.74      0.73       925
```

From this data, the accuracy of mode around 74% with using Logistic Regression and precision 78% and 71%



Graph of Customer





Summary

- Gender variable is not impact the most of telecom customer churn
- Phone Service impact the customer churn because the customer want easier to use and easy to access for them
- Internet Service with Fiber Optic has more customer churn than before because at now mos of the company have Fiber Optic installation and product, so need extra effort for this sector.
- Contract month to month is not recommended for the company to give some contract better use 1 year or 2 years for subscription fee with certain discount or packages to customer.
- The model is 74% accuracy and 78% of precision, so need other machine learning model to get a new insight and more accuracy.

Customer Churn

T

H

A

N

K

YOU