Logistic Regression

Use Case : Telecom Dataset

Fellow : Jithin J Kumar

Importing Libraries and Dataset

GRAD VALLEY DATA SCIENCE

Telecom Dataset

Understand the Telecom data provided by analysing and visualising the data.Build the model using Logistic Regression with the train data. Predict the customers churning for the test data provided based on the built and validate

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import linear_model
import seaborn as sns
from sklearn import preprocessing
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
```

```
df=pd.read_excel('train_telecom.xlsx')
```

Data Exploration

```
df.shape
```

```
(3333, 20)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3333 entries, 1 to 3333
Data columns (total 20 columns):
state                           3333 non-null object
account_length                  3333 non-null int64
area_code                       3333 non-null object
international_plan              3333 non-null object
voice_mail_plan                 3333 non-null object
number_vmail_messages           3333 non-null int64
total_day_minutes               3333 non-null float64
total_day_calls                 3333 non-null int64
total_day_charge                3333 non-null float64
total_eve_minutes               3333 non-null float64
total_eve_calls                 3333 non-null int64
total_eve_charge                3333 non-null float64
total_night_minutes             3333 non-null float64
total_night_calls               3333 non-null int64
total_night_charge              3333 non-null float64
total_intl_minutes              3333 non-null float64
total_intl_calls                3333 non-null int64
total_intl_charge               3333 non-null float64
number_customer_service_calls   3333 non-null int64
churn                           3333 non-null object
dtypes: float64(8), int64(7), object(5)
memory usage: 546.8+ KB
```

As seen : 20 Columns

Dependent Variable is Y
Independent Variables are as shown

3333 observations in the training data.

Categorical Variables are
State, Area_code, International_plan, Voice_mail_plan and Churn

Rest all are Continuous Variables.

Data Exploration (Some Intresting Insights)

churn state	no	yes	All	Percent
NJ	50	18	68	26.0
CA	25	9	34	26.0
TX	54	18	72	25.0
MD	53	17	70	24.0
SC	46	14	60	23.0
MI	57	16	73	22.0
MS	51	14	65	22.0
WA	52	14	66	21.0
ME	49	13	62	21.0
NV	52	14	66	21.0

churn area_code	no	yes	All	Percent
area_code_408	716	122	838	15.0
area_code_510	715	125	840	15.0
area_code_415	1419	236	1655	14.0
All	2850	483	3333	14.0

churn international_plan	no	yes	All	Percent
no	2664	346	3010	11.0
yes	186	137	323	42.0
All	2850	483	3333	14.0

churn voice_mail_plan	no	yes	All	Percent
no	2008	403	2411	17.0
yes	842	80	922	9.0
All	2850	483	3333	14.0

churn number_customer_service_calls	no	yes	All	Percent
0	605.0	92.0	697	13.0
1	1059.0	122.0	1181	10.0
2	672.0	87.0	759	11.0
3	385.0	44.0	429	10.0
4	90.0	76.0	166	46.0
5	26.0	40.0	66	61.0
6	8.0	14.0	22	64.0
7	4.0	5.0	9	56.0
8	1.0	1.0	2	50.0
9	NaN	2.0	2	100.0
All	2850.0	483.0	3333	14.0

Testing Our Model

With all X Variables

```
model=testmodel(x,y)
```

```
***Training Data***
accuracy : 0.8714
precision : 0.6429
recall : 0.1837


***Validation Data***
accuracy : 0.8683
precision : 0.5385
recall : 0.1045
```

With 6 main features

```
***Training Data***
accuracy : 0.8686
precision : 0.6154
recall : 0.1633


***Validation Data***
accuracy : 0.8663
precision : 0.5
recall : 0.1045
```

with 10 major features

```
***Training Data***
accuracy : 0.8714
precision : 0.6429
recall : 0.1837


***Validation Data***
accuracy : 0.8683
precision : 0.5385
recall : 0.1045
```

Final Test Score :

```
***Test Data***
accuracy : 0.8758
precision : 0.6
recall : 0.2277
```

Thank You….!

Open for Questions