In [1]:
```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```python
data = pd.read_csv('santander-train.csv')
```

In [3]:
```python
data.head()
```

Out[3]:

	ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_comer_ult3	imp_op_var40_comer_ult1	imp_op_var40_comer_ult3
0	1	2	23	0.0	0.0	0.0	0.0	0.0
1	3	2	34	0.0	0.0	0.0	0.0	0.0
2	4	2	23	0.0	0.0	0.0	0.0	0.0
3	8	2	37	0.0	195.0	195.0	0.0	0.0
4	10	2	39	0.0	0.0	0.0	0.0	0.0

5 rows × 371 columns

In [4]:
```python
data.tail()
```

Out[4]:

	ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_comer_ult3	imp_op_var40_comer_ult1	imp_op_var40_co
76015	151829	2	48	0.0	0.0	0.0	0.0	
76016	151830	2	39	0.0	0.0	0.0	0.0	
76017	151835	2	23	0.0	0.0	0.0	0.0	
76018	151836	2	25	0.0	0.0	0.0	0.0	
76019	151838	2	46	0.0	0.0	0.0	0.0	

5 rows × 371 columns

In [5]:
```python
x = data.drop(columns='TARGET')
y = data['TARGET']
```

In [6]:
```python
x.shape,y.shape
```

Out[6]:
```
((76020, 370), (76020,))
```

In [52]:
```python
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
```

In [53]:
```python
x_train,x_test,y_train,y_test = train_test_split(x,y,train_size=0.80,random_state=42,stratify=y)
```

In [54]:
```python
 x_train.shape,x_test.shape
```

Out[54]:
```
((4319, 370), (15204, 370))
```

In [55]:
```python
x_train.duplicated().sum()
```

Out[55]:
```
0
```

In [56]:
```python
x_test.duplicated().sum()
```

Out[56]:
```
0
```

To remove the duplications.

In [57]:
```python
from sklearn.feature_selection import VarianceThreshold
```

In [58]:
```python
x_train_thresold = VarianceThreshold(threshold=0.01)
```

In [59]:
```python
x_train_unique = x_train_thresold.fit_transform(x_train)
x_test_unique =x_train_thresold.transform(x_test)
```

In [60]:
```python
x_train_unique.shape,x_test_unique.shape
```

Out[60]:
```
((4319, 222), (15204, 222))
```

In [69]:
```python
x_train_unique = pd.DataFrame(x_train_unique)
x_test_unique = pd.DataFrame(x_test_unique)
```

Use Feature Selection Techinique.

In [70]:
```python
#remove the co-reelated feature
```

In [71]:
```python
def corrlated_feature(data,thresh):
    cormat = data.corr()
    corr_col= set()
    for i in range(len(cormat.columns)):
        for j in range(i):
            if abs(cormat.iloc[i,j]>thresh):
                columns = cormat.columns[i]
                corr_col.add(columns)
    return corr_col
```

In [72]:
```python
correlated_feature = corrlated_feature(x_train_unique,0.80)
```

In [73]:
```python
len(correlated_feature)
```

Out[73]:
```
142
```

In [75]:
```python
x_train_uncorr = x_train_unique.drop(labels=correlated_feature,axis=1)
x_test_uncorr = x_test_unique.drop(labels=correlated_feature,axis=1)
```

In [76]:
```python
x_train_uncorr.shape,x_test_uncorr.shape
```

Out[76]:
```
((4319, 80), (15204, 80))
```

Feature Dimensions Reduction By LDA or Its A Classifier.

In [78]:
```python
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
```

Beacasue we having the classes 0 and 1.during the WE create one sapce so that we can seperate the two classes.

In [80]:
```python
lda = LDA(n_components=1)
```

In [81]:
```python
x_train_lda = lda.fit_transform(x_train_uncorr,y_train)
```

In [83]:
```python
x_train_lda.shape
```

Out[83]:
```
(4319, 1)
```

It reduces the dimensions from the 80 to 1.

In [84]:
```python
## TO avoid the overfitting
```

In [85]:
```python
x_test_lda = lda.transform(x_test_uncorr)
```

In [88]:
```python
from sklearn.ensemble import RandomForestClassifier
```

In [89]:
```python
def randomforest(x_train,x_test,y_train,y_test):
    clf = RandomForestClassifier(n_jobs=-1,n_estimators=1000,random_state=0)
    clf.fit(x_train,y_train)
    y_pred = clf.predict(x_test)
    print('Accuracy on Test Set :-')
    print('Accuracy :',accuracy_score(y_test,y_pred))
```

In [90]:
```python
%%time
randomforest(x_train_lda,x_test_lda,y_train,y_test)
```

```
Accuracy on Test Set :-
Accuracy : 0.5853064982899236
Wall time: 9.01 s
```

In [91]:
```python
%%time
randomforest(x_train,x_test,y_train,y_test)
```

```
Accuracy on Test Set :-
Accuracy : 0.703893712181005
Wall time: 11.4 s
```

Feature Reduction By PCA

In [92]:
```python
from sklearn.decomposition import PCA
```

In [93]:
```python
pca = PCA(n_components=2,random_state=42)
pca.fit(x_test_uncorr)
```

Out[93]:
```
PCA(copy=True, iterated_power='auto', n_components=2, random_state=42,
    svd_solver='auto', tol=0.0, whiten=False)
```

In [95]:
```python
x_train_pca = pca.transform(x_train_uncorr)
x_test_pca = pca.transform(x_test_uncorr)
```

In [96]:
```python
x_train_pca.shape,x_test_pca.shape
```

Out[96]:
```
((4319, 2), (15204, 2))
```

In [97]:
```python
%%time
randomforest(x_train_pca,x_test_pca,y_train,y_test)
```

```
Accuracy on Test Set :-
Accuracy : 0.09037095501183899
Wall time: 7.34 s
```

In [100…
```python
for component in range(1,79):
    pca = PCA(n_components=component,random_state=42)
    pca.fit(x_test_uncorr)
    x_train_pca=  pca.transform(x_train_uncorr)
    x_test_pca=   pca.transform(x_test_uncorr)
    print('Selected Comp',component)
    randomforest(x_train_pca,x_test_pca,y_train,y_test)
    print()
```

```
Selected Comp 1
Accuracy on Test Set :-
Accuracy : 0.08925282820310444

Selected Comp 2
Accuracy on Test Set :-
Accuracy : 0.09037095501183899

Selected Comp 3
Accuracy on Test Set :-
Accuracy : 0.2530255196001052

Selected Comp 4
Accuracy on Test Set :-
Accuracy : 0.4614575111812681

Selected Comp 5
Accuracy on Test Set :-
Accuracy : 0.4670481452249408

Selected Comp 6
Accuracy on Test Set :-
Accuracy : 0.4836885030255196

Selected Comp 7
Accuracy on Test Set :-
Accuracy : 0.4877663772691397

Selected Comp 8
Accuracy on Test Set :-
Accuracy : 0.48546435148645095

Selected Comp 9
Accuracy on Test Set :-
Accuracy : 0.5046698237305972

Selected Comp 10
Accuracy on Test Set :-
Accuracy : 0.5137463825309129

Selected Comp 11
Accuracy on Test Set :-
Accuracy : 0.528282031044462

Selected Comp 12
Accuracy on Test Set :-
Accuracy : 0.5316364114706656

Selected Comp 13
Accuracy on Test Set :-
Accuracy : 0.5305182846619311

Selected Comp 14
Accuracy on Test Set :-
Accuracy : 0.5274269928966061

Selected Comp 15
Accuracy on Test Set :-
Accuracy : 0.5255196001052355

Selected Comp 16
Accuracy on Test Set :-
Accuracy : 0.5347277032359905

Selected Comp 17
Accuracy on Test Set :-
Accuracy : 0.5376874506708761

Selected Comp 18
Accuracy on Test Set :-
Accuracy : 0.5357800578795054

Selected Comp 19
Accuracy on Test Set :-
Accuracy : 0.5378847671665351

Selected Comp 20
Accuracy on Test Set :-
Accuracy : 0.5372270455143383

Selected Comp 21
Accuracy on Test Set :-
Accuracy : 0.5395290712970271

Selected Comp 22
Accuracy on Test Set :-
Accuracy : 0.5403183372796633

Selected Comp 23
Accuracy on Test Set :-
Accuracy : 0.5440015785319653

Selected Comp 24
Accuracy on Test Set :-
Accuracy : 0.5450539331754801

Selected Comp 25
Accuracy on Test Set :-
Accuracy : 0.5434096290449881

Selected Comp 26
Accuracy on Test Set :-
Accuracy : 0.569060773480663

Selected Comp 27
Accuracy on Test Set :-
Accuracy : 0.5662325703762168

Selected Comp 28
Accuracy on Test Set :-
Accuracy : 0.5670876085240726

Selected Comp 29
Accuracy on Test Set :-
Accuracy : 0.5661667982109971

Selected Comp 30
Accuracy on Test Set :-
Accuracy : 0.5647198105761642

Selected Comp 31
Accuracy on Test Set :-
Accuracy : 0.5789265982636148

Selected Comp 32
Accuracy on Test Set :-
Accuracy : 0.57991318074191

Selected Comp 33
Accuracy on Test Set :-
Accuracy : 0.5784004209418574

Selected Comp 34
Accuracy on Test Set :-
Accuracy : 0.5855695869508024

Selected Comp 35
Accuracy on Test Set :-
Accuracy : 0.5952380952380952

Selected Comp 36
Accuracy on Test Set :-
Accuracy : 0.601223362273086

Selected Comp 37
Accuracy on Test Set :-
Accuracy : 0.6291765324914497

Selected Comp 38
Accuracy on Test Set :-
Accuracy : 0.6289134438305709

Selected Comp 39
Accuracy on Test Set :-
Accuracy : 0.648579321231255

Selected Comp 40
Accuracy on Test Set :-
Accuracy : 0.6816627203367535

Selected Comp 41
Accuracy on Test Set :-
Accuracy : 0.6784398842409892

Selected Comp 42
Accuracy on Test Set :-
Accuracy : 0.677058668771376

Selected Comp 43
Accuracy on Test Set :-
Accuracy : 0.6783083399105498

Selected Comp 44
Accuracy on Test Set :-
Accuracy : 0.6762694027887398

Selected Comp 45
Accuracy on Test Set :-
Accuracy : 0.6755459089713234

Selected Comp 46
Accuracy on Test Set :-
Accuracy : 0.675217048145225

Selected Comp 47
Accuracy on Test Set :-
Accuracy : 0.6760720862930808

Selected Comp 48
Accuracy on Test Set :-
Accuracy : 0.677453301762694

Selected Comp 49
Accuracy on Test Set :-
Accuracy : 0.677255985267035

Selected Comp 50
Accuracy on Test Set :-
Accuracy : 0.6792291502236254

Selected Comp 51
Accuracy on Test Set :-
Accuracy : 0.6804788213627992

Selected Comp 52
Accuracy on Test Set :-
Accuracy : 0.6806761378584583

Selected Comp 53
Accuracy on Test Set :-
Accuracy : 0.6816627203367535

Selected Comp 54
Accuracy on Test Set :-
Accuracy : 0.6823204419889503

Selected Comp 55
Accuracy on Test Set :-
Accuracy : 0.6869244935543278

Selected Comp 56
Accuracy on Test Set :-
Accuracy : 0.6860036832412523

Selected Comp 57
Accuracy on Test Set :-
Accuracy : 0.6838332017890029

Selected Comp 58
Accuracy on Test Set :-
Accuracy : 0.6859379110760326

Selected Comp 59
Accuracy on Test Set :-
Accuracy : 0.6858721389108129

Selected Comp 60
Accuracy on Test Set :-
Accuracy : 0.6837016574585635

Selected Comp 61
Accuracy on Test Set :-
Accuracy : 0.6834385687976848

Selected Comp 62
Accuracy on Test Set :-
Accuracy : 0.6837674296237832

Selected Comp 63
Accuracy on Test Set :-
Accuracy : 0.6824519863193896

Selected Comp 64
Accuracy on Test Set :-
Accuracy : 0.6837016574585635

Selected Comp 65
Accuracy on Test Set :-
Accuracy : 0.6817942646671928

Selected Comp 66
Accuracy on Test Set :-
Accuracy : 0.6842936069455406

Selected Comp 67
Accuracy on Test Set :-
Accuracy : 0.6814654038410944

Selected Comp 68
Accuracy on Test Set :-
Accuracy : 0.6830439358063668

Selected Comp 69
Accuracy on Test Set :-
Accuracy : 0.6843593791107603

Selected Comp 70
Accuracy on Test Set :-
Accuracy : 0.6831097079715864

Selected Comp 71
Accuracy on Test Set :-
Accuracy : 0.6831754801368061

Selected Comp 72
Accuracy on Test Set :-
Accuracy : 0.6840305182846619

Selected Comp 73
Accuracy on Test Set :-
Accuracy : 0.6827150749802684

Selected Comp 74
Accuracy on Test Set :-
Accuracy : 0.6843593791107603

Selected Comp 75
Accuracy on Test Set :-
Accuracy : 0.6833727966324652

Selected Comp 76
Accuracy on Test Set :-
Accuracy : 0.684227834780321

Selected Comp 77
Accuracy on Test Set :-
Accuracy : 0.6833070244672454

Selected Comp 78
Accuracy on Test Set :-
Accuracy : 0.6840305182846619
```