In [2]:
```
import warnings
warnings.filterwarnings("ignore")

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
from imblearn.over_sampling import SMOTE
```

Train a base line model

In [3]:
```
df = pd.read_csv('creditcard.csv')
y = df['Class']
X = df.drop('Class',axis=1)
```

In [4]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42
```

In [5]:
```
clf = RandomForestClassifier(max_depth=5, random_state=0)
clf = clf.fit(X_train,y_train)
```

In [6]:
```
res = clf.predict(X_test)
confusion_matrix(y_test, res)
```

Out[6]:
```
array([[7972,    1],
       [   7,   20]])
```

In [7]:
```
from sklearn.metrics import classification_report
print(classification_report(y_test, res))
```

```
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      7973
          1       0.95      0.74      0.83        27

avg / total       1.00      1.00      1.00      8000
```

Train a better model with data imbalanced solved

In [8]:
```
sm = SMOTE(random_state=42)
X_res, y_res = sm.fit_sample(X_train, y_train)
```

In [9]:
```
clf = RandomForestClassifier(max_depth=5, random_state=0)
clf = clf.fit(X_res,y_res)
```

In [10]:
```
res = clf.predict(X_test)
confusion_matrix(y_test, res)
```

Out[10]:
```
array([[7967,    6],
       [   4,   23]])
```

In [11]:
```
print(classification_report(y_test, res))
```

```
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      7973
          1       0.79      0.85      0.82        27
```

```
avg / total       1.00      1.00      1.00      8000
```

Report and Evaluation

In the base model without SMOTE method, we are seeing that we are getting a good result on class 0 with only 1 wrong prediction, but the class 1 gets a lot worse since there is not enough data of class 1 being fed to the model.

After applying the SMOTE imbalanced solutions, the accuracy of the Class 1 (monority) has gone up quite a bit, reduced almost a half. However, the sacrifices is that there is a slight increase of error in class 0 (majority) predictions, which I think is the trade off.