Credit Card Fraud Detection

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
```

In [2]:
```
# loading the dataset to a pandas DataFrame
df = pd.read_csv("C:\\Users\\Ramees\\Project\\creditcard.csv")
df
```

Out[2]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
284802	172786.0	-11.881118	10.071785	-9.834783	-2.066656	-5.364473	-2.606837	-4.918215	7.305334	1.914428	...	0.213454	0.111864	1.014480	-0.509348	1.436807	0.250034	0.943651	0.823731	0.77	0
284803	172787.0	-0.732789	-0.055080	2.035030	-0.738589	0.868229	1.058415	0.024330	0.294869	0.584800	...	0.214205	0.924384	0.012463	-1.016226	-0.606624	-0.395255	0.068472	-0.053527	24.79	0
284804	172788.0	1.919565	-0.301254	-3.249640	-0.557828	2.630515	3.031260	-0.296827	0.708417	0.432454	...	0.232045	0.578229	-0.037501	0.640134	0.265745	-0.087371	0.004455	-0.026561	67.88	0
284805	172788.0	-0.240440	0.530483	0.702510	0.689799	-0.377961	0.623708	-0.686180	0.679145	0.392087	...	0.265245	0.800049	-0.163298	0.123205	-0.569159	0.546668	0.108821	0.104533	10.00	0
284806	172792.0	-0.533413	-0.189733	0.703337	-0.506271	-0.012546	-0.649617	1.577006	-0.414650	0.486180	...	0.261057	0.643078	0.376777	0.008797	-0.473649	-0.818267	-0.002415	0.013649	217.00	0

284807 rows × 31 columns

In [3]:
```
# 5 Rows of the Dataset
df.tail()
```

Out[3]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
284802	172786.0	-11.881118	10.071785	-9.834783	-2.066656	-5.364473	-2.606837	-4.918215	7.305334	1.914428	...	0.213454	0.111864	1.014480	-0.509348	1.436807	0.250034	0.943651	0.823731	0.77	0
284803	172787.0	-0.732789	-0.055080	2.035030	-0.738589	0.868229	1.058415	0.024330	0.294869	0.584800	...	0.214205	0.924384	0.012463	-1.016226	-0.606624	-0.395255	0.068472	-0.053527	24.79	0
284804	172788.0	1.919565	-0.301254	-3.249640	-0.557828	2.630515	3.031260	-0.296827	0.708417	0.432454	...	0.232045	0.578229	-0.037501	0.640134	0.265745	-0.087371	0.004455	-0.026561	67.88	0
284805	172788.0	-0.240440	0.530483	0.702510	0.689799	-0.377961	0.623708	-0.686180	0.679145	0.392087	...	0.265245	0.800049	-0.163298	0.123205	-0.569159	0.546668	0.108821	0.104533	10.00	0
284806	172792.0	-0.533413	-0.189733	0.703337	-0.506271	-0.012546	-0.649617	1.577006	-0.414650	0.486180	...	0.261057	0.643078	0.376777	0.008797	-0.473649	-0.818267	-0.002415	0.013649	217.00	0

5 rows × 31 columns

In [4]:
```
# dataset information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
 15  V15     284807 non-null  float64
 16  V16     284807 non-null  float64
 17  V17     284807 non-null  float64
 18  V18     284807 non-null  float64
 19  V19     284807 non-null  float64
 20  V20     284807 non-null  float64
 21  V21     284807 non-null  float64
 22  V22     284807 non-null  float64
 23  V23     284807 non-null  float64
 24  V24     284807 non-null  float64
 25  V25     284807 non-null  float64
 26  V26     284807 non-null  float64
 27  V27     284807 non-null  float64
 28  V28     284807 non-null  float64
 29  Amount  284807 non-null  float64
 30  Class   284807 non-null  int64  
dtypes: float64(30), int64(1)
memory usage: 67.4 MB
```

In [5]:
```
# checking the number of missing values in each columns
df.isnull().sum()
```

Out[5]:
```
Time      0
V1        0
V2        0
V3        0
V4        0
V5        0
V6        0
V7        0
V8        0
V9        0
V10       0
V11       0
V12       0
V13       0
V14       0
V15       0
V16       0
V17       0
V18       0
V19       0
V20       0
V21       0
V22       0
V23       0
V24       0
V25       0
V26       0
V27       0
V28       0
Amount    0
Class     0
dtype: int64
```

In [6]:
```
# distribution of legit transactions and fraudulent transactions
class_value = df["Class"].value_counts()
class_value
```

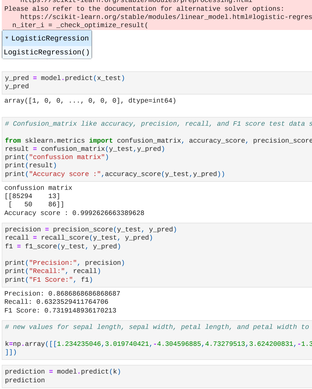
Out[6]:
```
0    284315
1       492
Name: Class, dtype: int64
```

0 ---> Normal Transaction 1 ---> Fraudulent Transaction

In [7]:
```
#visualizing the data using histograms

LABELS = ["Normal Transaction", "Fraud Transaction"]
class_value.plot(kind = "bar",rot=0)
plt.title("Class Distribution of Transaction")
plt.xticks(range(2), LABELS)
plt.xlabel("Classes")
plt.ylabel("No of Occurences")
```

Out[7]:
```
Text(0, 0.5, 'No of Occurences')
```



In [8]:
```
# separating the data for analysis

normal = df[df.Class == 0]
Fraud = df[df.Class == 1]
```

In [9]:
```
print(normal.shape)
print(Fraud.shape)
```

```
(284315, 31)
(492, 31)
```

In [10]:
```
normal.Amount.describe()
```

Out[10]:
```
count    284315.000000
mean         88.291022
std         250.105092
min           0.000000
25%           5.650000
50%          22.000000
75%          77.050000
max       25691.160000
Name: Amount, dtype: float64
```

In [11]:
```
Fraud.Amount.describe()
```

Out[11]:
```
count     492.000000
mean      122.211321
std       256.683288
min         0.000000
25%         1.000000
50%         9.250000
75%       105.890000
max      2125.870000
Name: Amount, dtype: float64
```

In [12]:
```
# select column
x = df.iloc[:,1:30].values
y = df.iloc[:,30].values
```

In [13]:
```
print("Input Range : ", x.shape)
print("output Range : ", y.shape)
```

```
Input Range :  (284807, 29)
output Range :  (284807,)
```

In [14]:
```
# Split the dataset into training and testing sets
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.3,random_state = 42)
```

In [15]:
```
x_train.shape
```

Out[15]:
```
(199364, 29)
```

In [16]:
```
y_train.shape
```

Out[16]:
```
(199364,)
```

In [17]:
```
# Algorithm like LogisticRegression
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
```

In [18]:
```
# training the Logistic Tegression model with Training data
model.fit(x_train,y_train)
```

```
I:\Install Software\Anaconda\Lib\site-packages\sklearn\linear_model\_logistic.py:460: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

Out[18]:
```
LogisticRegression()
```

In [19]:
```
y_pred = model.predict(x_test)
y_pred
```

Out[19]:
```
array([1, 0, 0, ..., 0, 0, 0], dtype=int64)
```

In [21]:
```
# Confusion_matrix like accuracy, precision, recall, and F1 score test data set

from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, f1_score
result = confusion_matrix(y_test,y_pred)
print("Confussion matrix")
print(result)
print("Accuracy score :",accuracy_score(y_test,y_pred))
```

```
confussion matrix
[[85294    13]
 [   50    86]]
Accuracy score : 0.9992626663389628
```

In [22]:
```
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
```

```
Precision: 0.8686868686868687
Recall: 0.6323529411764706
F1 Score: 0.7319148936170213
```

In [23]:
```
# new values for sepal length, sepal width, petal length, and petal width to see the predicted species

k=np.array([[1.234235046,3.019740421,-4.304596885,4.73279513,3.624200831,-1.357745663,1.713444988,+0.496358487,-1.28285782,-2.447469255,2.101343865,-4.609628391,1.464377625,-6.079337
]])
```

In [25]:
```
prediction = model.predict(k)
prediction
```

Out[25]:
```
array([0], dtype=int64)
```

In [26]:
```
if prediction[0] == 1:
    print("The transaction is fraudulent.")
else:
    print("The transaction is not fraudulent.")
```

```
The transaction is not fraudulent.
```

In [ ]:
```

```