In [3]:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from collections import Counter
from sklearn.preprocessing import StandardScaler

import imblearn as il
from imblearn.over_sampling import SMOTE
from imblearn.combine import SMOTEENN
from sklearn.model_selection import train_test_split
%matplotlib inline



from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import recall_score
from sklearn.metrics import average_precision_score
from sklearn.metrics import precision_recall_curve



pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 500)
pd.set_option('display.float_format', lambda x: '%.4f' % x)
```

In [2]:

```
data = pd.read_csv("/Users/sonia/Downloads/creditcard.csv")
```

In [4]:

```
count_classes = pd.value_counts(data['Class'], sort = True).sort_index()

count_classes
```

Out[4]:

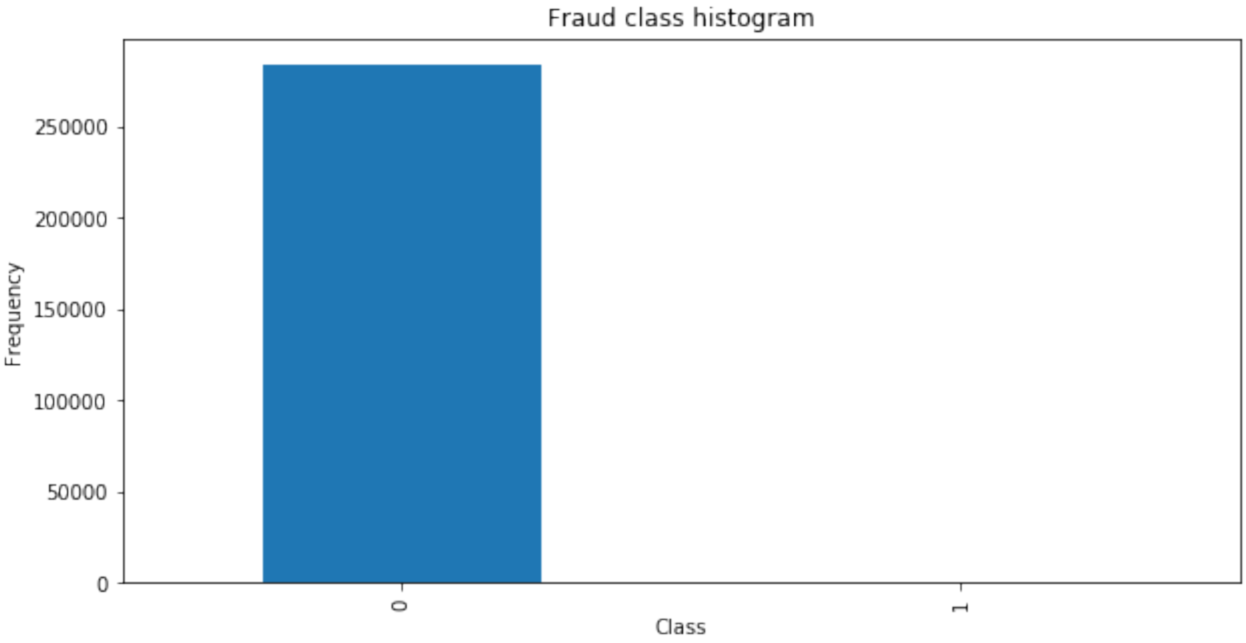
```
0    284315
1       492
Name: Class, dtype: int64
```

In [5]:

```
plt.figure(figsize=(10,5))
count_classes.plot(kind = 'bar')
plt.title("Fraud class histogram")
plt.xlabel("Class")
plt.ylabel("Frequency")
```

Out[5]:

```
Text(0,0.5,'Frequency')
```



In [6]:

```
data.isnull().sum()
```

Out[6]:

```
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 [7]: data.describe()
```

```
Out[7]:
```

	Time	V1	V2	V3	V4	V5	V
count	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.00
mean	94813.8596	0.0000	0.0000	-0.0000	0.0000	-0.0000	0.00
std	47488.1460	1.9587	1.6513	1.5163	1.4159	1.3802	1.33
min	0.0000	-56.4075	-72.7157	-48.3256	-5.6832	-113.7433	-26.16
25%	54201.5000	-0.9204	-0.5985	-0.8904	-0.8486	-0.6916	-0.76
50%	84692.0000	0.0181	0.0655	0.1798	-0.0198	-0.0543	-0.27
75%	139320.5000	1.3156	0.8037	1.0272	0.7433	0.6119	0.39
max	172792.0000	2.4549	22.0577	9.3826	16.8753	34.8017	73.30

```
In [9]: cols = list(data.columns.values)
```

In [10]:
```
cols
```

Out[10]:
```
['Time',
 'V1',
 'V2',
 'V3',
 'V4',
 'V5',
 'V6',
 'V7',
 'V8',
 'V9',
 'V10',
 'V11',
 'V12',
 'V13',
 'V14',
 'V15',
 'V16',
 'V17',
 'V18',
 'V19',
 'V20',
 'V21',
 'V22',
 'V23',
 'V24',
 'V25',
 'V26',
 'V27',
 'V28',
 'Amount',
 'Class']
```

In [11]:
```
X = data.iloc[:, data.columns != 'Class']
Y = data.iloc[:, data.columns == 'Class']
```

In [12]:
```
Y.shape
```

Out[12]:
```
(284807, 1)
```

In [13]:
```
X.shape
```

Out[13]:
```
(284807, 30)
```

In [14]:
```
print('Class labels:', np.unique(Y))
```

```
Class labels: [0 1]
```

In [15]:
```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.1,
```

In [16]:
```
train, x_val, y_train, y_val = train_test_split(X_train, Y_train,test_size
```

In [17]:
```
lr = LogisticRegression(random_state=0)
```

In [18]:
```
lr.fit(x_train, y_train)
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/linear_model/logistic.p
y:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:761
: DataConversionWarning: A column-vector y was passed when a 1d array
was expected. Please change the shape of y to (n_samples, ), for examp
le using ravel().
  y = column_or_1d(y, warn=True)
```

Out[18]:
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept
=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=0, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
```

In [19]:
```
print('Validation Results')
print('Accuracy: ', lr.score(x_val, y_val))
print('Recall Score: ', recall_score(y_val, lr.predict(x_val)))
```

```
Validation Results
Accuracy:  0.9990246947294503
Recall Score:  0.5121951219512195
```

In [20]:
```
print('Test Results')
print('Accuracy: ', lr.score(X_test, Y_test))
print('Recall Score: ', recall_score(Y_test, lr.predict(X_test)))
```

```
Test Results
Accuracy:  0.9990168884519505
Recall Score:  0.5636363636363636
```

In [21]:

```
sm = SMOTE(random_state=12, ratio = 'auto', kind = 'regular')
```

In [22]:

```
x_train_res, y_train_res = sm.fit_sample(x_train, y_train)

print('Resampled dataset shape {}'.format(Counter(y_train_res)))
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:761
: DataConversionWarning: A column-vector y was passed when a 1d array
was expected. Please change the shape of y to (n_samples, ), for examp
le using ravel().
  y = column_or_1d(y, warn=True)

Resampled dataset shape Counter({0: 230297, 1: 230297})
```

In [23]:

```
sm_logr = LogisticRegression(random_state=0)

sm_logr.fit(x_train_res, y_train_res)
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/linear_model/logistic.p
y:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
```

Out[23]:

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept
=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=0, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
```

In [24]:

```
print('Validation Results')
print('Accuracy: ', sm_logr.score(x_val, y_val))
print('Recall Score: ', recall_score(y_val, sm_logr.predict(x_val)))
```

```
Validation Results
Accuracy:  0.9855264697850428
Recall Score:  0.7804878048780488
```

In [25]:

```
print('Test Results')
print('Accuracy: ', sm_logr.score(X_test, Y_test))
print('Recall Score: ', recall_score(Y_test, sm_logr.predict(X_test)))
```

```
Test Results
Accuracy:  0.9853937712861206
Recall Score:  0.9090909090909091
```

In [ ]: