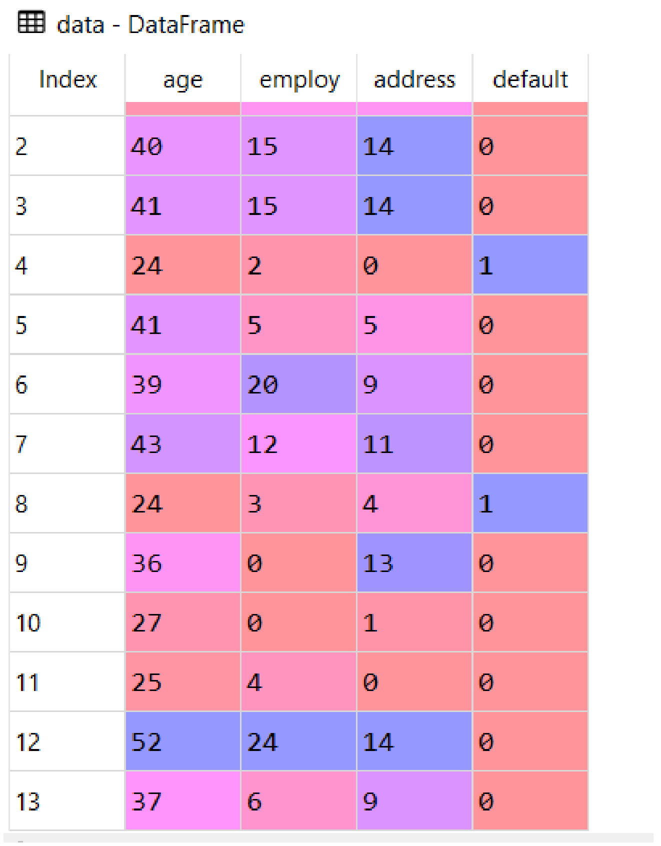
SMOTE

Data: (1) UpSamp_Demo.csv (2) data_final.csv

SMOTE concept

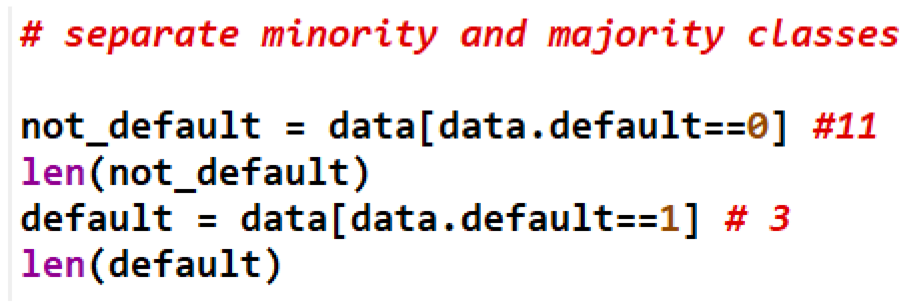


data - DataFrame

Index	age	employ	address	default
2	40	15	14	0
3	41	15	14	0
4	24	2	0	1
5	41	5	5	0
6	39	20	9	0
7	43	12	11	0
8	24	3	4	1
9	36	0	13	0
10	27	0	1	0
11	25	4	0	0
12	52	24	14	0
13	37	6	9	0

```python
# Jesus is my Saviour!
import pandas as pd
import sklearn
from sklearn.utils import resample

data = pd.read_csv("C:/Users/Dr Vinod/Desktop/DataSets1/UpSamp_ Demo.csv")
data = pd.DataFrame(data)
data.shape # 14 by 9
data.info()

data.default.value_counts()
'''
0    11
1     3
Name: default, dtype: int64
'''
3/11 # 27% are 1
```

Separate yes and no



```
# separate minority and majority classes

not_default = data[data.default==0] #11
len(not_default)
default = data[data.default==1] # 3
len(default)
```

Increase counts of 1 = counts of 0

Resampling code

Resampling will happen to this class (1)

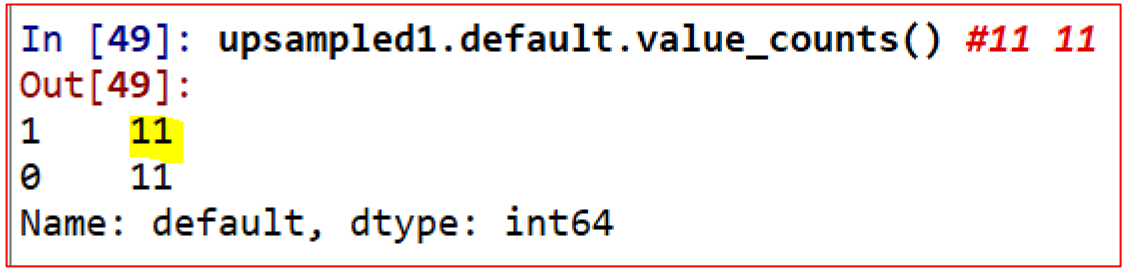
```
#___________________________________ upsample minority_with replacement
from sklearn.utils import resample
default_upsampled1 = resample(default,
                              replace=True, # sample with replacement
                              n_samples=len(not_default), # match number in majority class
                              random_state=27) # reproducible results
```

Without replacement it will not be possible!

Nos of sample (1) = class (0)

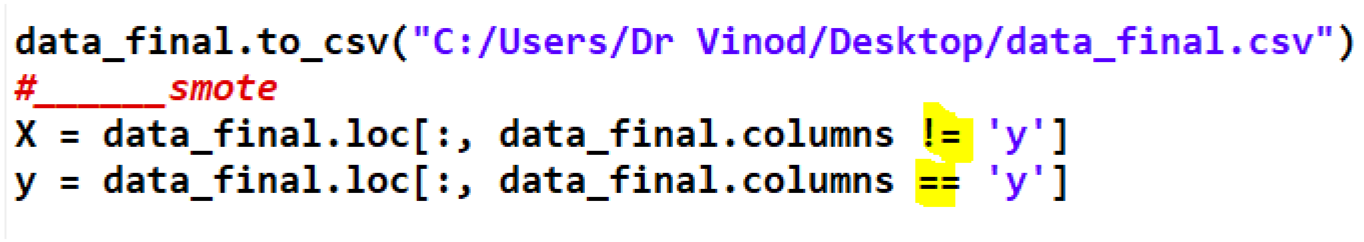
Combine vertically

```
# combine majority and upsampled minority

upsampled1 = pd.concat([not_default, default_upsampled1]) #22, 11, 11

# check new class counts
upsampled1.default.value_counts() #11 11

upsampled1.to_csv('C:/Users/Dr Vinod/Desktop/DataSets1/upsampled1.csv')
```



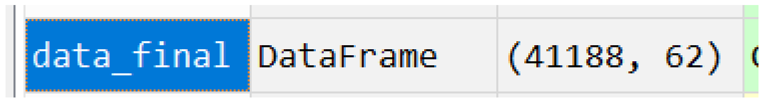
```
In [49]: upsampled1.default.value_counts() #11 11
Out[49]:
1     11
0     11
Name: default, dtype: int64
```

Create predictors and target variable

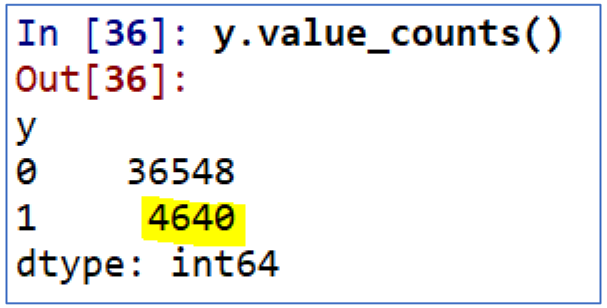


```
data_final.to_csv("C:/Users/Dr Vinod/Desktop/data_final.csv")
#______smote
X = data_final.loc[:, data_final.columns != 'y']
y = data_final.loc[:, data_final.columns == 'y']
```

How you should apply on a large data set?

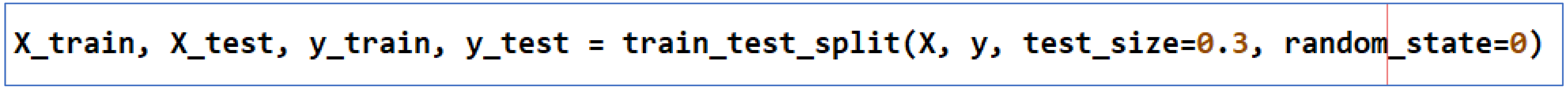


```
data_final  DataFrame  (41188, 62)
```

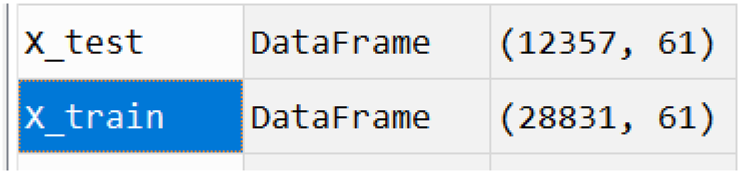


```
In [36]: y.value_counts()
Out[36]:
y
0    36548
1     4640
dtype: int64
```

30 % test data

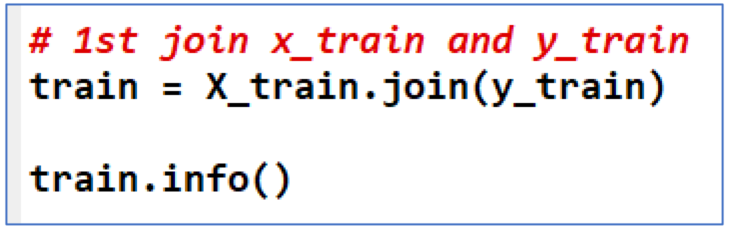


```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

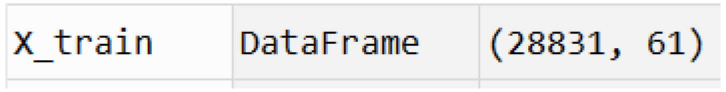


X_test	DataFrame	(12357, 61)
X_train	DataFrame	(28831, 61)

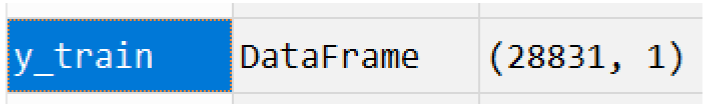
Join horizontally x_train and y_train



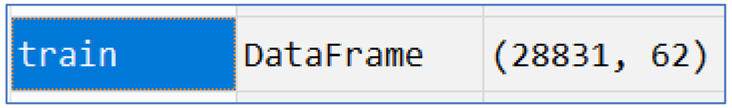
```
# 1st join x_train and y_train
train = X_train.join(y_train)

train.info()
```



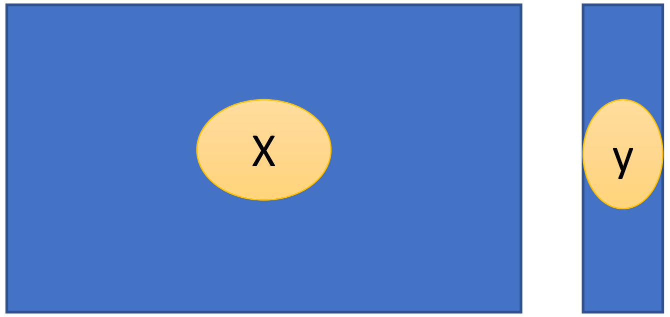
```
X_train     DataFrame     (28831, 61)
```



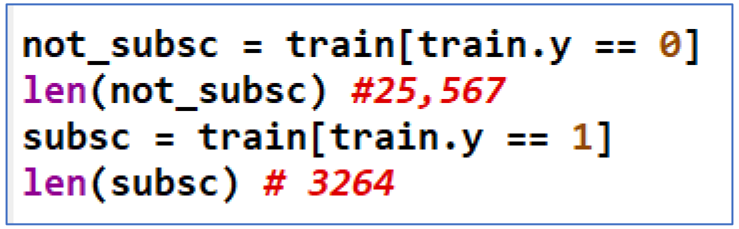
```
y_train     DataFrame     (28831, 1)
```



```
train       DataFrame     (28831, 62)
```



Count the imbalanced categories



```
not_subsc = train[train.y == 0]
len(not_subsc) #25,567
subsc = train[train.y == 1]
len(subsc) # 3264
```

Make minority class = majority class

```
# 2 upsample; minor catg 'subsc' to be incraesed to counts = not_subsc
from sklearn.utils import resample
subsc_os = resample(subsc,
                    replace=True, # sample with replacement
                    n_samples=len(not_subsc), # match number in majority class
                    random_state=27) # reproducible results

train_os = pd.concat([not_subsc, subsc_os])

train_os.y.value_counts()
'''
1    25567
0    25567
Name: y, dtype: int64
'''
```

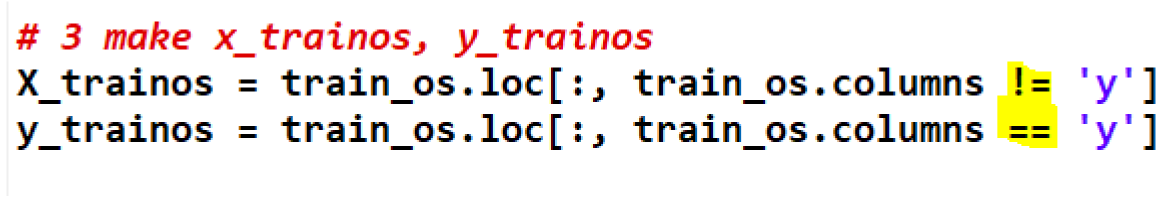
```
not_subsc = train[train.y == 0]
len(not_subsc) #25,567
subsc = train[train.y == 1]
len(subsc) # 3264
```

not_subsc

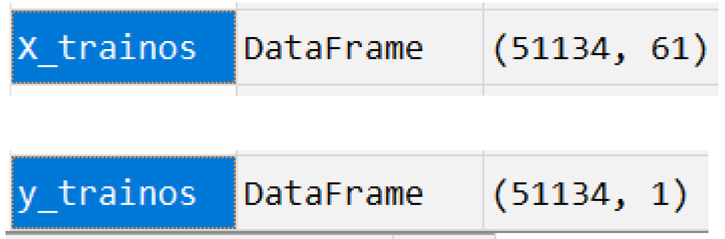
subsc_os

```
train_os    DataFrame    (51134, 62)
```

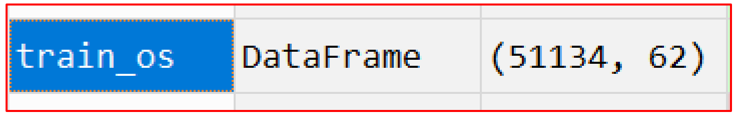
Now make oversampled x_train(os) & y_train(os)



```
# 3 make x_trainos, y_trainos
X_trainos = train_os.loc[:, train_os.columns != 'y']
y_trainos = train_os.loc[:, train_os.columns == 'y']
```



X_trainos	DataFrame	(51134, 61)
y_trainos	DataFrame	(51134, 1)



train_os	DataFrame	(51134, 62)

BECOME
THE HARDEST
WORKING
PERSON
YOU KNOW
shutterstock

"No matter how hard you work, someone else is working harder."
- by Elon Musk
pixelsquote.net