Part 2 : Feature Reduction (Extraction/Selection)¶

In this phase of the project, we will examine the features and remove or convert them

List of columns in the dataset are as follow:

In [21]:

```
data.columns
```

Out[21]:

```
Index(['RowNumber', 'CustomerId', 'Surname', 'CreditScore', 'Geography',
       'Gender', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCard',
       'IsActiveMember', 'EstimatedSalary', 'Exited'],
      dtype='object')
```

Step 11- remove columns **RowNumber**,**CustomerId**, and **Surname**¶

```
New columns:
```

Out[23]:

```
Index(['CreditScore', 'Geography', 'Gender', 'Age', 'Tenure', 'Balance',
       'NumOfProducts', 'HasCrCard', 'IsActiveMember', 'EstimatedSalary',
       'Exited'],
      dtype='object')
```

Step 12 - Onehot code Geography¶

In [24]:

```
The three countries are now coded.
```

Out[24]:

```
array([[1, 0, 0],
       [0, 0, 1],
       [1, 0, 0],
       ...,
       [1, 0, 0],
       [0, 1, 0],
       [1, 0, 0]], dtype=int32)
```

In [25]:

```
Coded countries
```

Out[25]:

```
array(['France', 'Germany', 'Spain'], dtype='<U7')
```

	France	Germany	Spain
0	1	0	0
1	0	0	1
2	1	0	0
3	1	0	0
4	0	0	1

In [27]:

Add Geography dummies to the dataset¶

In [28]:

Out[28]:

	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Fran
0	619	Female	42	2	0.00	1	1	1	101348.88	1	1
1	608	Female	41	1	83807.86	1	0	1	112542.58	0	0
2	502	Female	42	8	159660.80	3	1	0	113931.57	1	1
3	699	Female	39	1	0.00	2	0	0	93826.63	0	1
4	850	Female	43	2	125510.82	1	1	1	79084.10	0	0

Onehot code Gender¶

	Female	Male
0	1	0
1	1	0
2	1	0
3	1	0
4	1	0
...	...	...
9995	0	1
9996	0	1
9997	1	0
9998	0	1
9999	1	0

10000 rows × 2 columns

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	France	Germ
0	619	42	2	0.00	1	1	1	101348.88	1	1	0
1	608	41	1	83807.86	1	0	1	112542.58	0	0	0
2	502	42	8	159660.80	3	1	0	113931.57	1	1	0
3	699	39	1	0.00	2	0	0	93826.63	0	1	0
4	850	43	2	125510.82	1	1	1	79084.10	0	0	0

Dataset with geography and gender dummied and 1 dummy removed to avoid dummy trap.¶

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	France	Germ
0	619	42	2	0.00	1	1	1	101348.88	1	1	0
1	608	41	1	83807.86	1	0	1	112542.58	0	0	0
2	502	42	8	159660.80	3	1	0	113931.57	1	1	0
3	699	39	1	0.00	2	0	0	93826.63	0	1	0
4	850	43	2	125510.82	1	1	1	79084.10	0	0	0

Move dependent variable to last column¶

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	France	Germany	Fe
0	619	42	2	0.00	1	1	1	101348.88	1	0	1
1	608	41	1	83807.86	1	0	1	112542.58	0	0	1
2	502	42	8	159660.80	3	1	0	113931.57	1	0	1
3	699	39	1	0.00	2	0	0	93826.63	1	0	1
4	850	43	2	125510.82	1	1	1	79084.10	0	0	1

Step 13 - Set up independent variable and dependent variables and perform feature reduction¶

```
[[619.  42.   2. ...   0.   1.   1.]
 [608.  41.   1. ...   0.   1.   0.]
 [502.  42.   8. ...   0.   1.   1.]
 ...
 [516.  35.  10. ...   0.   0.   0.]
 [709.  36.   7. ...   0.   1.   1.]
 [772.  42.   3. ...   1.   0.   1.]]
[1 0 1 ... 1 1 0]
```

Attempt at feature reduction using PCA Before feature scaling¶

In [35]:

```
Original number of features: 12
Reduced number of features: 2
```

In [36]:

```
Scale Independent variables
```

```
[[-0.32609367  0.29341451 -1.041749   ... -0.57877454  1.09610816
   1.97704053]
 [-0.4399147   0.19806052 -1.38751174 ... -0.57877454  1.09610816
  -0.50580653]
 [-1.53673561  0.29341451  1.03282743 ... -0.57877454  1.09610816
   1.97704053]
 ...
 [-1.39187247 -0.37406345  1.72435291 ... -0.57877454 -0.91231872
  -0.50580653]
 [ 0.60516937 -0.27870946  0.68706469 ... -0.57877454  1.09610816
   1.97704053]
 [ 1.25705349  0.29341451 -0.69598626 ...  1.7277885  -0.91231872
   1.97704053]]
```

Attempt at feature reduction using PCA After feature scaling¶

In [37]:

```
Original number of features: 12
Reduced number of features: 12
```

In [ ]:

We removed irrelevant columns, onehot coded Gender and Geography columns. We performed feature scaling and performed a PCA feature reduction on pre and post scaling.

Our next step will be model selection and prediction.