Original Analysis Case Study

Part 1 : Graphics Analysis

Part 2 : Feature Reduction (Extraction/Selection)

Part 3 : Filling in Missing Values

Part 1 : Graphics Analysis

In this case study, as part of phase I, we will perform exploratory data analysis by graphing the features in the dataset.

The dataset is composed of 10,000 customer's record at a bank. The dataset has a total of 14 features 13 of which can be considered as independent variables and 1 as the dependent variable. The goal is to build a model that can predict whether a customer is likely to stay or exit the bank. The model will predict the dependent variable 'Exited' using the approrpiate set of independent variables
'CreditScore','Geography','Gender','Age','Tenure','Balance','NumberOfProducts','HasCrCard', and 'IsActiveMember'.

We will perform model selection and model validation exercises and use the model the make the desired prediction. The accuracy and percision of the model will be analyzed in the next phases of the study.

```
The dimension of the table is:  (10000, 14)
```

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	\
0	1	15634602	Hargrave	619	France	Female	42	
1	2	15647311	Hill	608	Spain	Female	41	
2	3	15619304	Onio	502	France	Female	42	
3	4	15701354	Boni	699	France	Female	39	
4	5	15737888	Mitchell	850	Spain	Female	43	

	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	\
0	2	0.00	1	1	1	
1	1	83807.86	1	0	1	
2	8	159660.80	3	1	0	
3	1	0.00	2	0	0	
4	2	125510.82	1	1	1	

	EstimatedSalary	Exited
0	101348.88	1
1	112542.58	0
2	113931.57	1
3	93826.63	0
4	79084.10	0

Describe Data

	RowNumber	CustomerId	CreditScore	Age	Tenure
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000

	Balance	NumOfProducts	HasCrCard	IsActiveMember
count	10000.000000	10000.000000	10000.00000	10000.000000
mean	76485.889288	1.530200	0.70550	0.515100
std	62397.405202	0.581654	0.45584	0.499797
min	0.000000	1.000000	0.00000	0.000000
25%	0.000000	1.000000	0.00000	0.000000
50%	97198.540000	1.000000	1.00000	1.000000
75%	127644.240000	2.000000	1.00000	1.000000
max	250898.090000	4.000000	1.00000	1.000000

	EstimatedSalary	Exited
count	10000.000000	10000.000000
mean	100090.239881	0.203700
std	57510.492818	0.402769
min	11.580000	0.000000
25%	51002.110000	0.000000
50%	100193.915000	0.000000
75%	149388.247500	0.000000
max	199992.480000	1.000000

Summarized Data

	Surname	Geography	Gender
count	10000	10000	10000
unique	2932	3	2
top	Smith	France	Male
freq	32	5014	5457

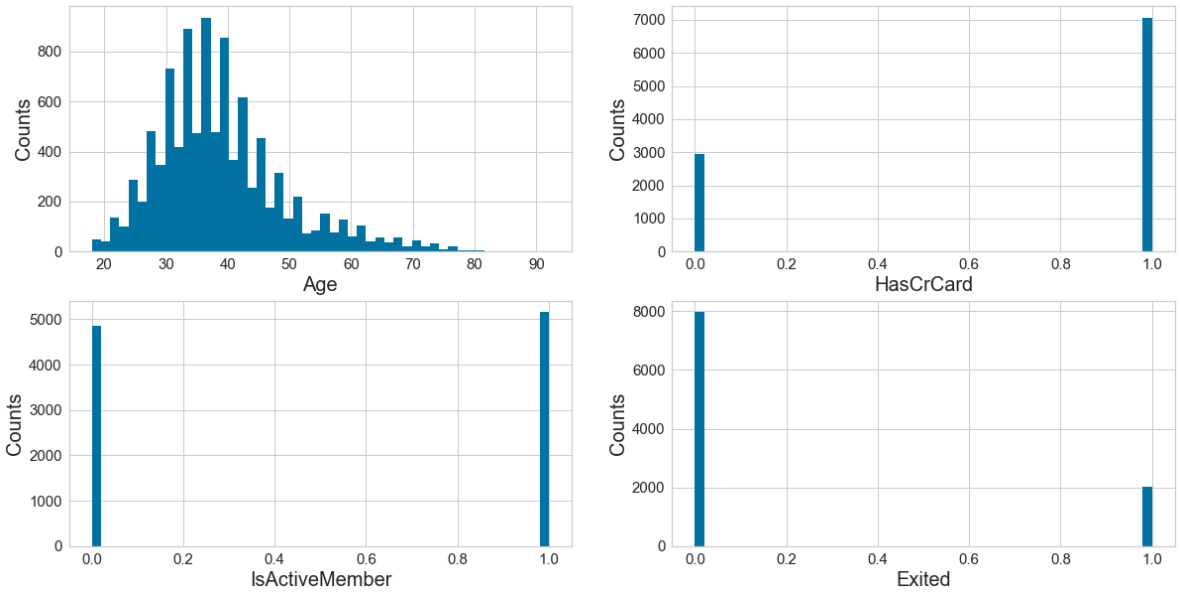
Summarized Data

	RowNumber	CustomerId	CreditScore	Age	Tenure
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000

	Balance	NumOfProducts	HasCrCard	IsActiveMember
count	10000.000000	10000.000000	10000.00000	10000.000000
mean	76485.889288	1.530200	0.70550	0.515100
std	62397.405202	0.581654	0.45584	0.499797
min	0.000000	1.000000	0.00000	0.000000
25%	0.000000	1.000000	0.00000	0.000000
50%	97198.540000	1.000000	1.00000	1.000000
75%	127644.240000	2.000000	1.00000	1.000000
max	250898.090000	4.000000	1.00000	1.000000

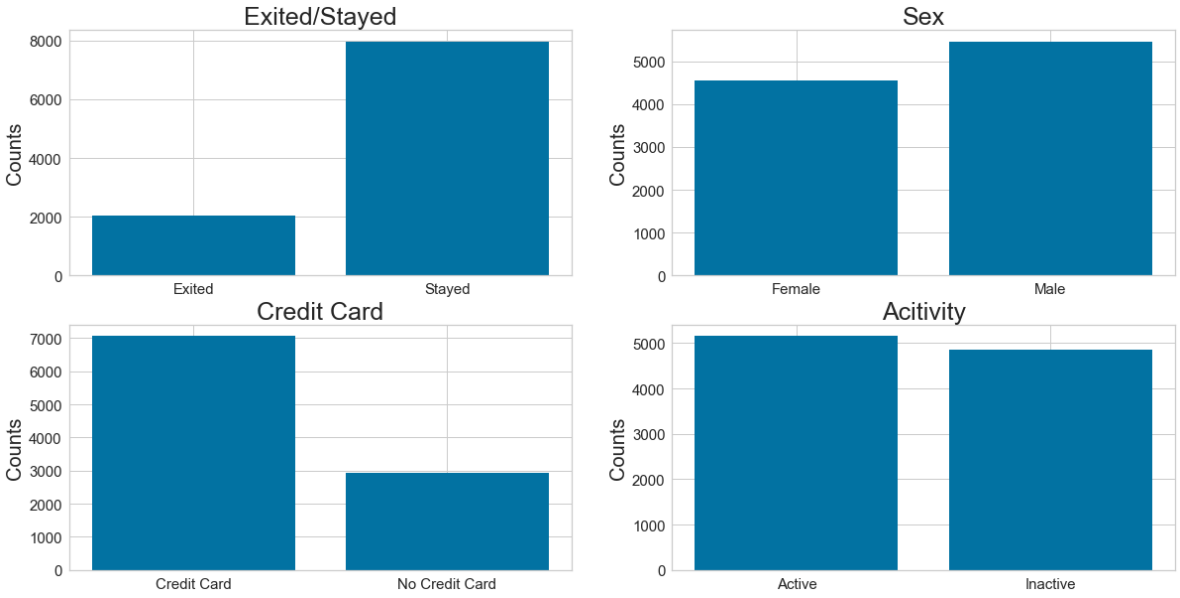
	EstimatedSalary	Exited
count	10000.000000	10000.000000
mean	100090.239881	0.203700
std	57510.492818	0.402769
min	11.580000	0.000000
25%	51002.110000	0.000000
50%	100193.915000	0.000000
75%	149388.247500	0.000000
max	199992.480000	1.000000

Histogram of ['Age', 'HasCrCard', 'IsActiveMember', 'Exited']

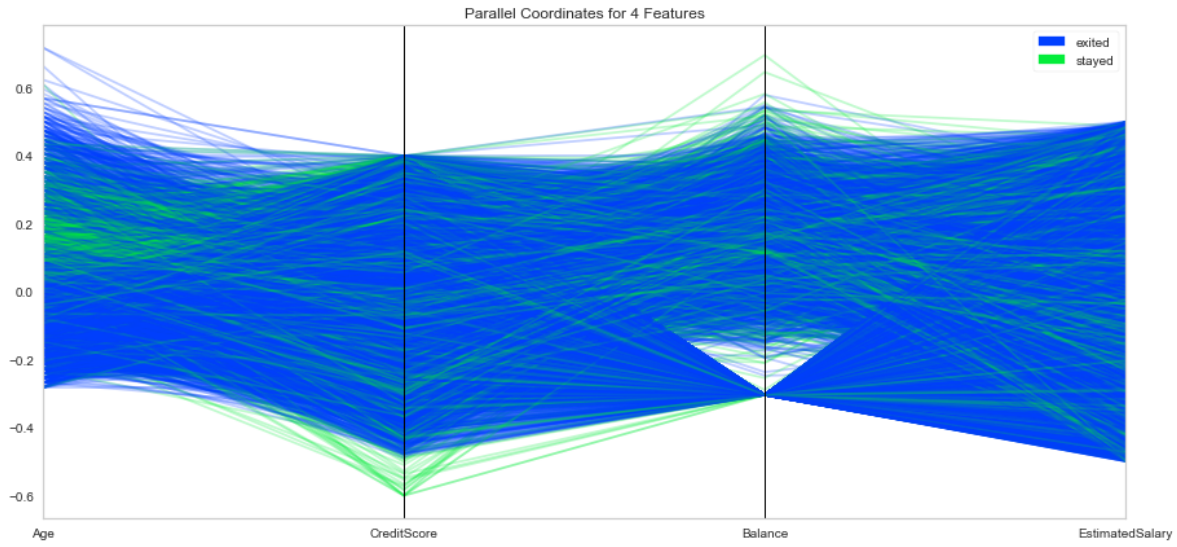


Barchart comparing the number of:

- **Exits vs stays**
- **Males vs. Female**
- **Has credit card vs does not have credit card**
- **active members vs inactive members**

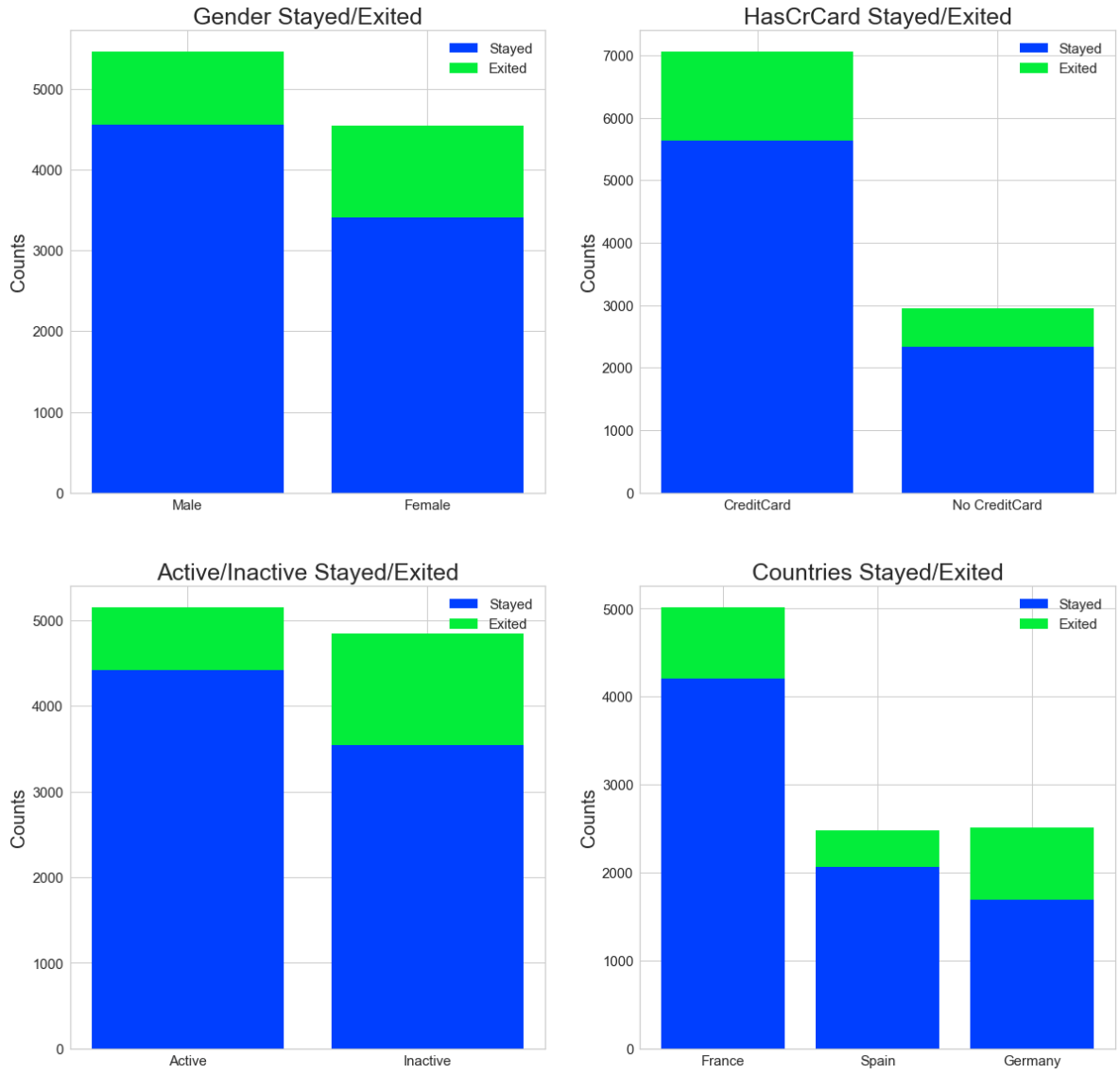


Parallel Coordinate graphe comparing ['Age', 'CreditScore', 'Balance', 'EstimatedSalary']



Stacked bar charts showing stays and exits based on:

- **Gender**
- **Has Credit card**
- **banking activity**
- **gegraphic location(Country)**



Part 2 : Feature Reduction (Extraction/Selection)

```
Out[16]: Index(['RowNumber', 'CustomerId', 'Surname', 'CreditScore', 'Geography',
                'Gender', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCar
         d',
                'IsActiveMember', 'EstimatedSalary', 'Exited'],
               dtype='object')
```

```
Out[18]: Index(['CreditScore', 'Geography', 'Gender', 'Age', 'Tenure', 'Balance',
                'NumOfProducts', 'HasCrCard', 'IsActiveMember', 'EstimatedSalary',
                'Exited'],
               dtype='object')
```

Out[19]:
```
array([[1, 0, 0],
       [0, 0, 1],
       [1, 0, 0],
       ...,
       [1, 0, 0],
       [0, 1, 0],
       [1, 0, 0]], dtype=int32)
```

Out[20]:
```
array(['France', 'Germany', 'Spain'], dtype='<U7')
```

Out[21]:

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

Out[23]:

	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActi
0	619	Female	42	2	0.00	1	1	
1	608	Female	41	1	83807.86	1	0	
2	502	Female	42	8	159660.80	3	1	
3	699	Female	39	1	0.00	2	0	
4	850	Female	43	2	125510.82	1	1	

Out[24]:

	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

Out[25]:

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMembe
0	619	42	2	0.00	1	1	
1	608	41	1	83807.86	1	0	
2	502	42	8	159660.80	3	1	
3	699	39	1	0.00	2	0	
4	850	43	2	125510.82	1	1	

Out[26]:

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMembe
0	619	42	2	0.00	1	1	
1	608	41	1	83807.86	1	0	
2	502	42	8	159660.80	3	1	
3	699	39	1	0.00	2	0	
4	850	43	2	125510.82	1	1	

Out[97]:

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMembe
0	619	42	2	0.00	1	1	
1	608	41	1	83807.86	1	0	
2	502	42	8	159660.80	3	1	
3	699	39	1	0.00	2	0	
4	850	43	2	125510.82	1	1	

```
<class 'numpy.ndarray'>
['Existed' 'Stayed' 'Existed' ... 'Existed' 'Existed' 'Stayed']
```

Out[100]: (10000, 12)

Out[101]: (10000, 11)

Out[102]: (10000,)

```
Original number of features: 11
Reduced number of features: 2
```

```
[[-0.32622142  0.29351742 -1.04175968 ...  0.99720391 -0.57873591
   1.09598752]
 [-0.44003595  0.19816383 -1.38753759 ... -1.00280393 -0.57873591
   1.09598752]
 [-1.53679418  0.29351742  1.03290776 ...  0.99720391 -0.57873591
   1.09598752]
 ...
 [ 0.60498839 -0.27860412  0.68712986 ...  0.99720391 -0.57873591
   1.09598752]
 [ 1.25683526  0.29351742 -0.69598177 ... -1.00280393  1.72790383
  -0.91241915]
 [ 1.46377078 -1.04143285 -0.35020386 ...  0.99720391 -0.57873591
   1.09598752]]


Original number of features: 11
Reduced number of features: 11
```

Part 3 : Filling in Missing Values

Summary of parts1 1 and 2: We have performed feature reduction and scaled the independent variables. The X and y variables are the independent variables dataset and the dependent variables respectively. The value of 0 or 1 for the depended variable has been converted to 'Stayed' and 'Exited" respectively in anticipation of using logistic regression classifier for modeling.

- Split_Train_Test

- Model Selection and Evaluation

```
Indpenden variables matrix:

[[-0.32622142  0.29351742 -1.04175968 ...  0.99720391 -0.57873591
   1.09598752]
 [-0.44003595  0.19816383 -1.38753759 ... -1.00280393 -0.57873591
   1.09598752]
 [-1.53679418  0.29351742  1.03290776 ...  0.99720391 -0.57873591
   1.09598752]
 ...
 [ 0.60498839 -0.27860412  0.68712986 ...  0.99720391 -0.57873591
   1.09598752]
 [ 1.25683526  0.29351742 -0.69598177 ... -1.00280393  1.72790383
  -0.91241915]
 [ 1.46377078 -1.04143285 -0.35020386 ...  0.99720391 -0.57873591
   1.09598752]]
```

```
Dependent variable array:

['Existed' 'Stayed' 'Existed' ... 'Existed' 'Existed' 'Stayed']
```

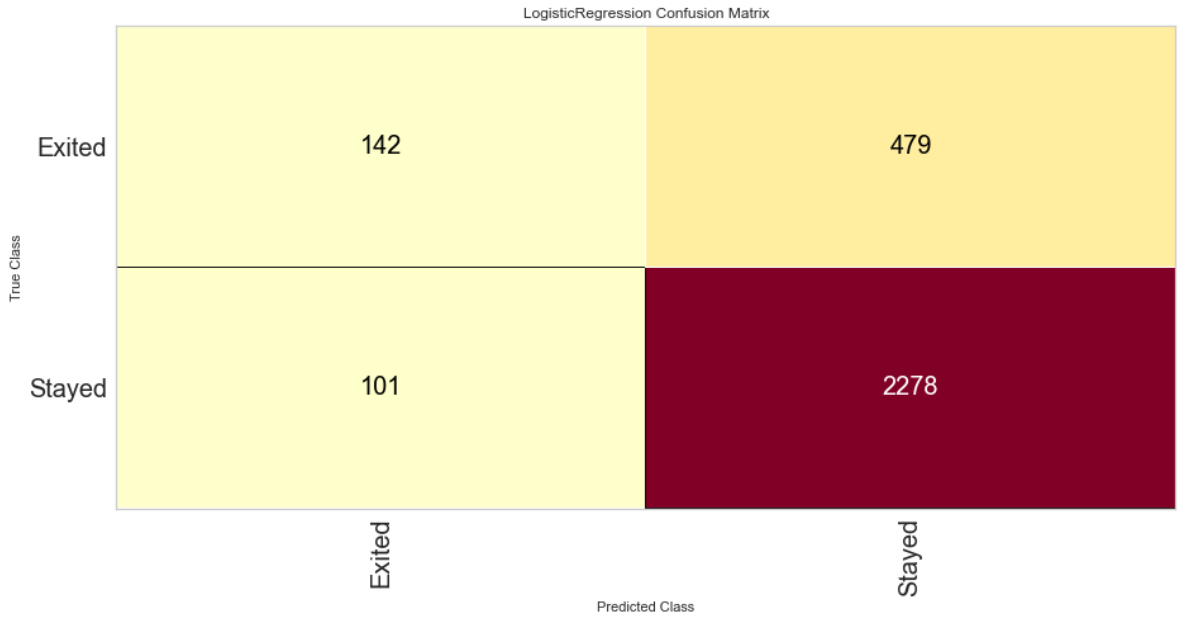
Step 14 - Split the dataset to 30% test set and 70% training dataset

```
Total sample in dataset:  10000
No. of samples in training set:  7000
No. of samples in validation set: 3000


(7000,)


(3000,)


No. of customer who stayed and exited in the training set:
Stayed     5584
Existed    1416
dtype: int64


No. of customer who stayed and exited  in the validation set:
Stayed     2379
Existed     621
dtype: int64
```

Step 15 - Model evaluation and metrics

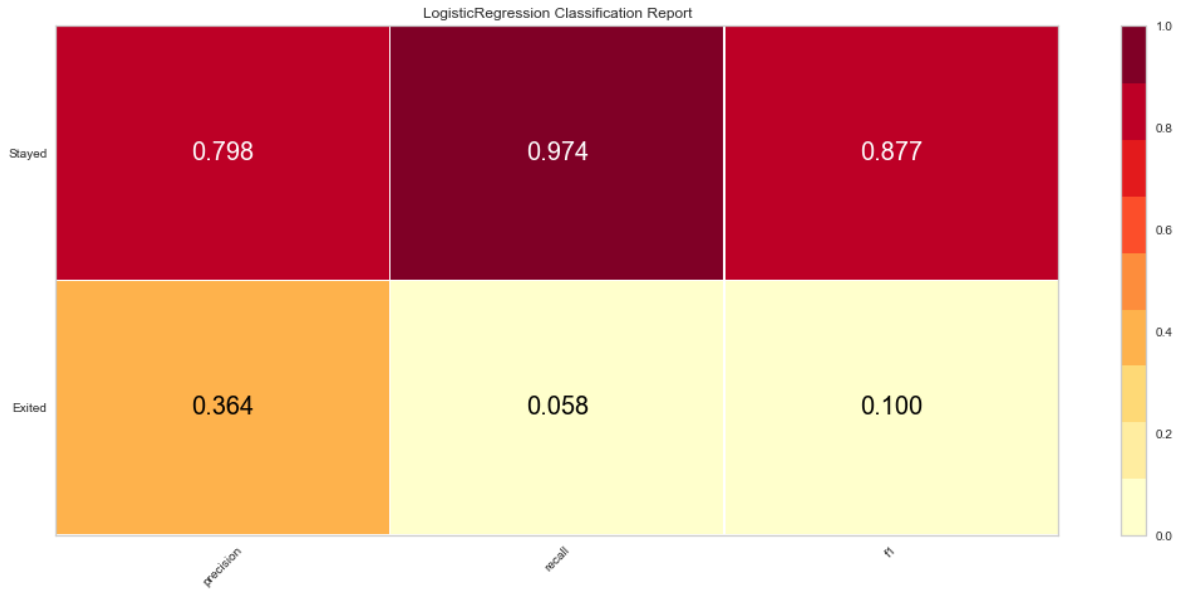
Create a logistics regression model

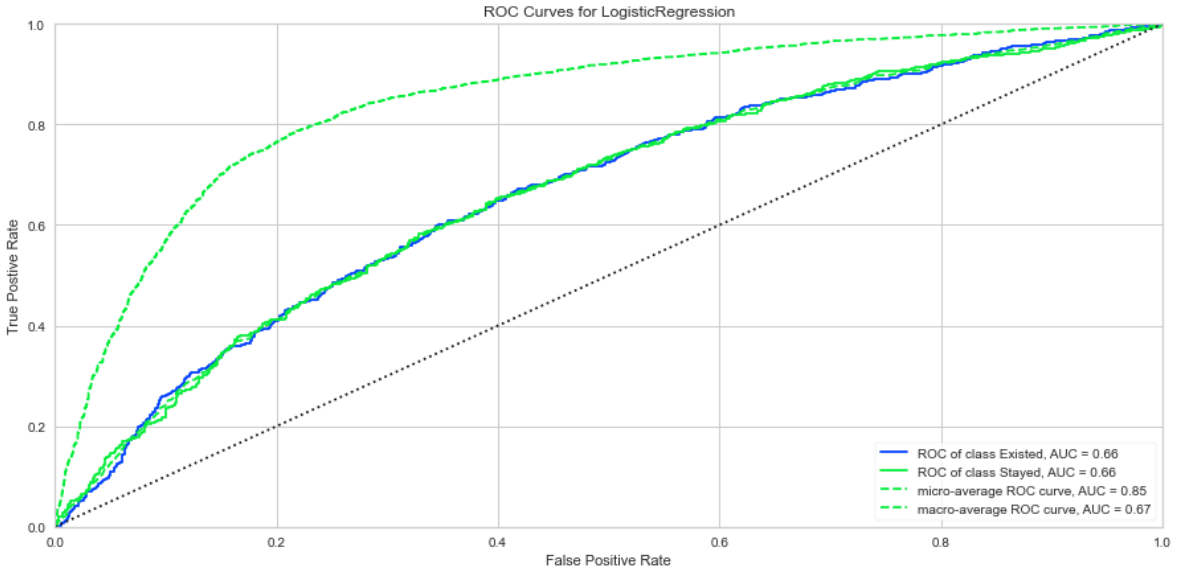
Define class for 'Exited' and 'stayed' to create confusion metrix and fit it into the trainign sets. Then display the confusion metric



Out[120]: <matplotlib.axes._subplots.AxesSubplot at 0x10be7bb0>

Precision, Recall, and F1 Score metrics:


LogisticRegression Classification Report
Stayed
Exited
0.798
0.974
0.877
0.364
0.058
0.100
precision
recall
f1
1.0
0.8
0.6
0.4
0.2
0.0


ROC Curves for LogisticRegression
True Postive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class Existed, AUC = 0.66
ROC of class Stayed, AUC = 0.66
micro-average ROC curve, AUC = 0.85
macro-average ROC curve, AUC = 0.67