Lead Scoring Case Study

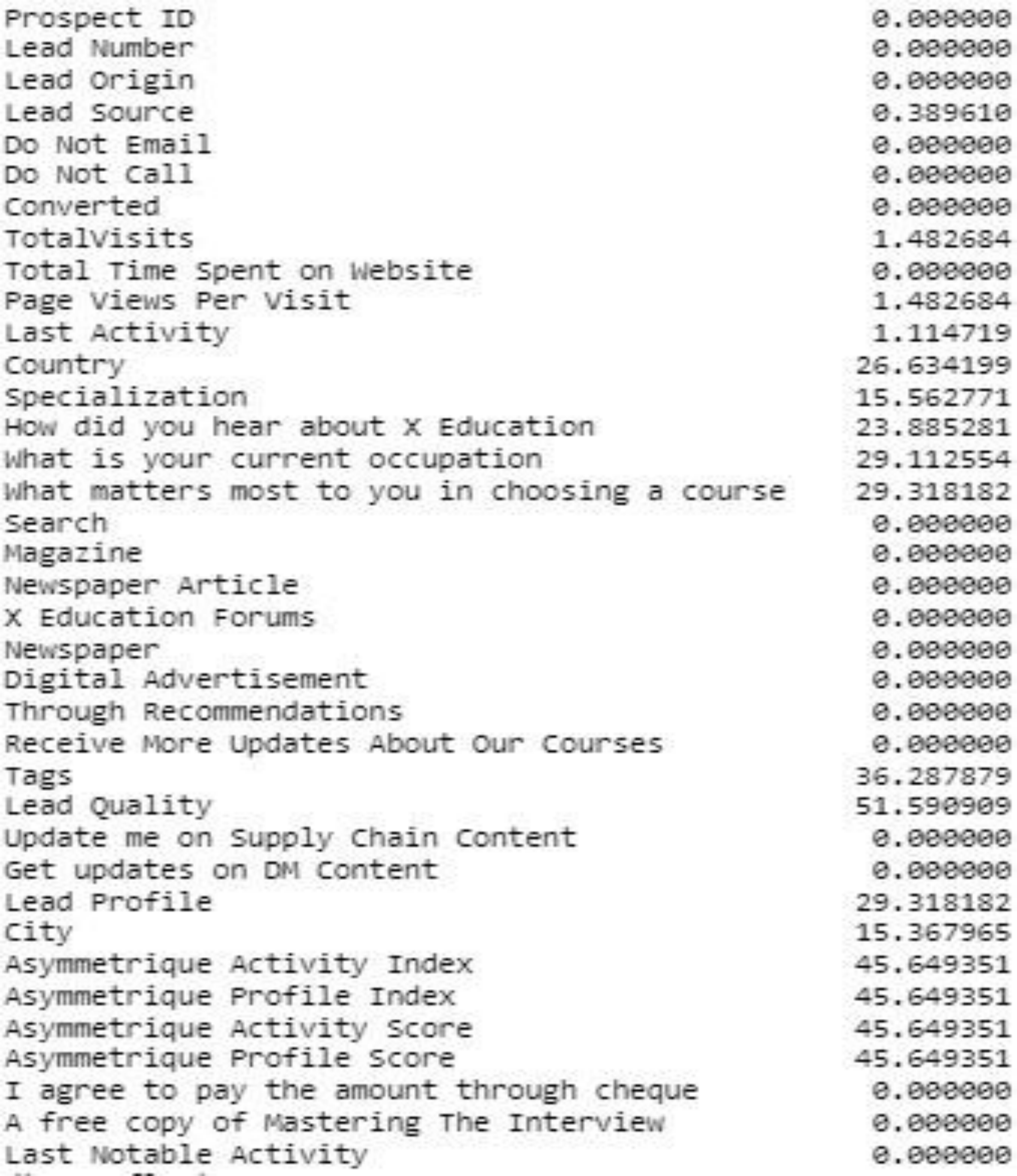
BY:
MANISH PANT
PALLAVI PATIL

Steps to be followed

- Data Cleaning
  - Dropping the columns
  - Dropping the rows with null values
- Univariate Analysis
- Bivariate Analysis
- Feature Selection
  - Manually dropping the columns
  - Calculating the VIF
- Predictions on train set
  - Model evaluation on the train dataset
  - Metrics on train set
  - ROC Curve
  - Finding an optimum cut-off value
- Predictions on test set
  - Performance metrics on the final test set.
  - Precision and recall
- Inferences for business decision- making

Data Cleaning

- We have these percentage of null valued present in our columns as shown in the adjacent picture.
- We dropped all those columns in which the percentage of null values is more then 32%.



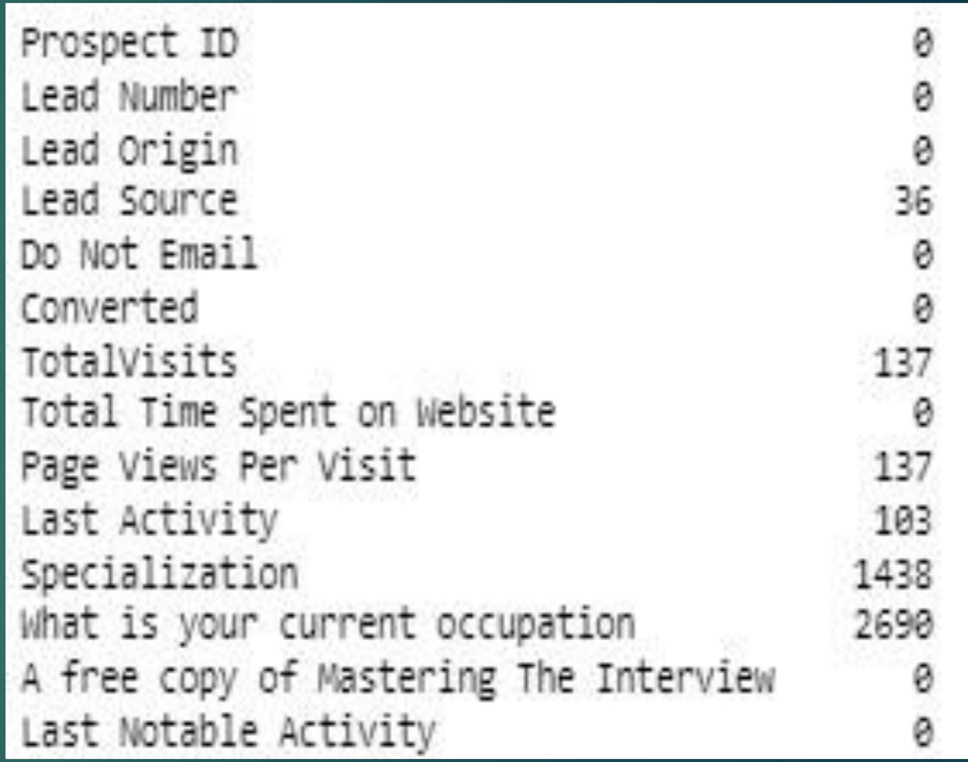
```
Prospect ID                                       0.000000
Lead Number                                       0.000000
Lead Origin                                       0.000000
Lead Source                                       0.389610
Do Not Email                                      0.000000
Do Not Call                                       0.000000
Converted                                         0.000000
TotalVisits                                       1.482684
Total Time Spent on Website                       0.000000
Page Views Per Visit                              1.482684
Last Activity                                     1.114719
Country                                          26.634199
Specialization                                   15.562771
How did you hear about X Education               23.885281
What is your current occupation                  29.112554
What matters most to you in choosing a course    29.318182
Search                                            0.000000
Magazine                                          0.000000
Newspaper Article                                 0.000000
X Education Forums                                0.000000
Newspaper                                         0.000000
Digital Advertisement                             0.000000
Through Recommendations                           0.000000
Receive More Updates About Our Courses            0.000000
Tags                                             36.287879
Lead Quality                                     51.590909
Update me on Supply Chain Content                 0.000000
Get updates on DM Content                         0.000000
Lead Profile                                     29.318182
City                                             15.367965
Asymmetrique Activity Index                      45.649351
Asymmetrique Profile Index                       45.649351
Asymmetrique Activity Score                      45.649351
Asymmetrique Profile Score                       45.649351
I agree to pay the amount through cheque          0.000000
A free copy of Mastering The Interview            0.000000
Last Notable Activity                             0.000000
```

Dropping the columns

- We have dropped columns such as Country, City, Lead Profile, Prospect Id and lead number as they will not affect our analysis.
- The columns which have very less or no variance are also dropped as they will not affect our analysis because there is only one value majorly present.

Dropping the rows with null values

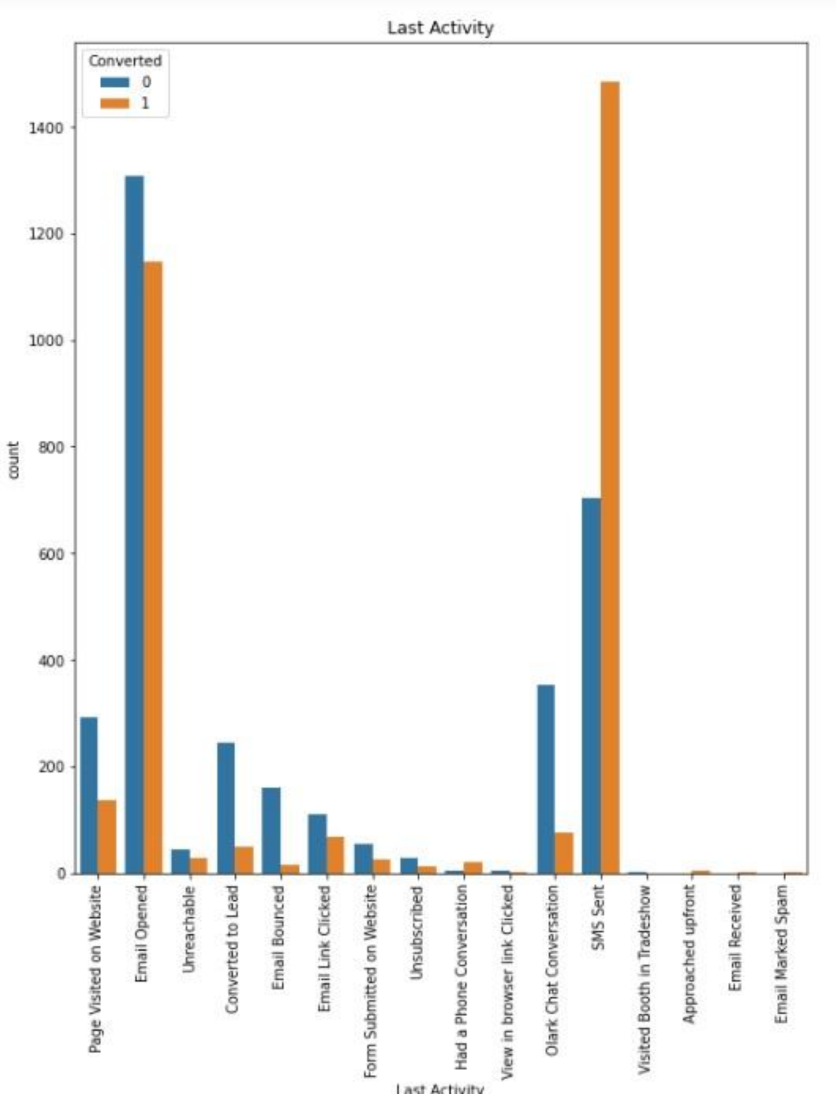
- After dropping the columns with more then 32% null values we will drop all the rows which contains null values.
- The number of null values in each column can be seen in the adjacent image.

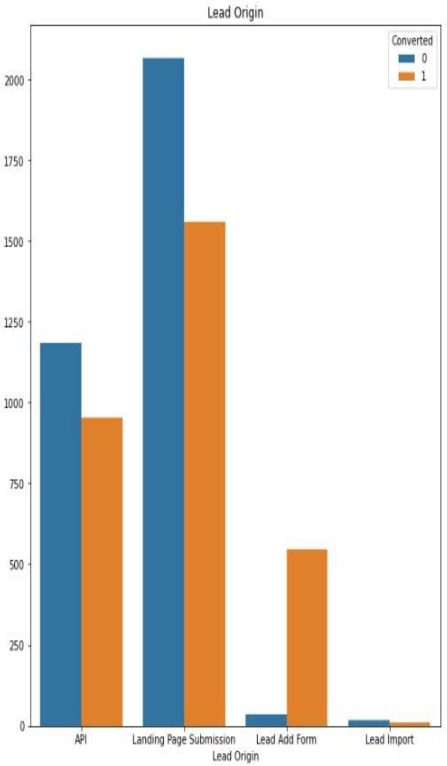


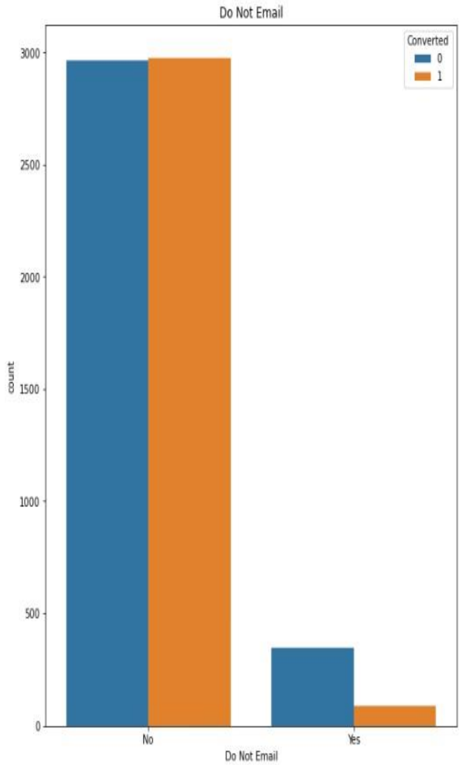
```
Prospect ID                                  0
Lead Number                                  0
Lead Origin                                  0
Lead Source                                 36
Do Not Email                                 0
Converted                                    0
TotalVisits                                137
Total Time Spent on Website                  0
Page Views Per Visit                       137
Last Activity                              103
Specialization                            1438
What is your current occupation           2690
A free copy of Mastering The Interview       0
Last Notable Activity                        0
```

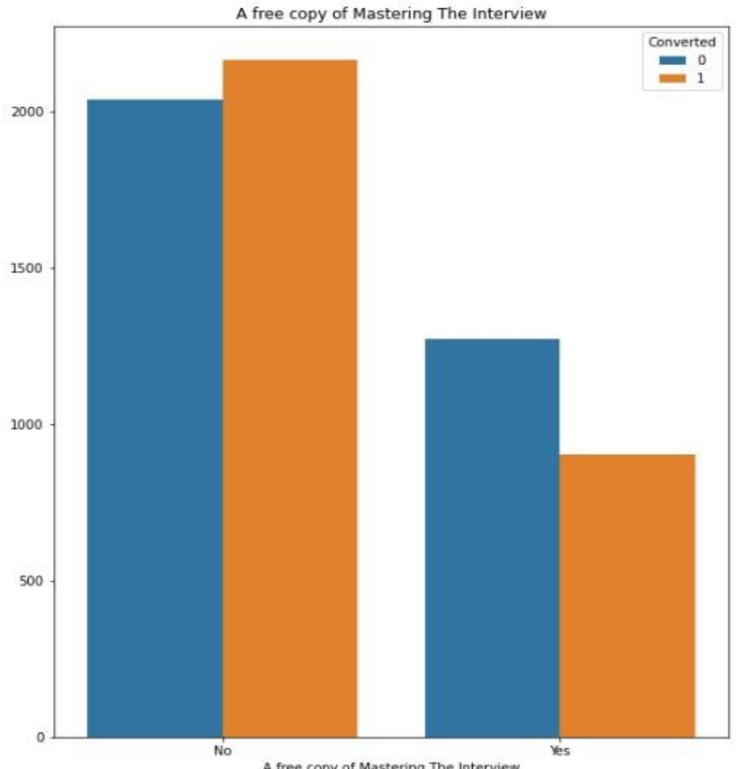
Univariate analysis

- Plotting various categorical variables by dividing the data according to the Converted column.
- The plots can be seen below.



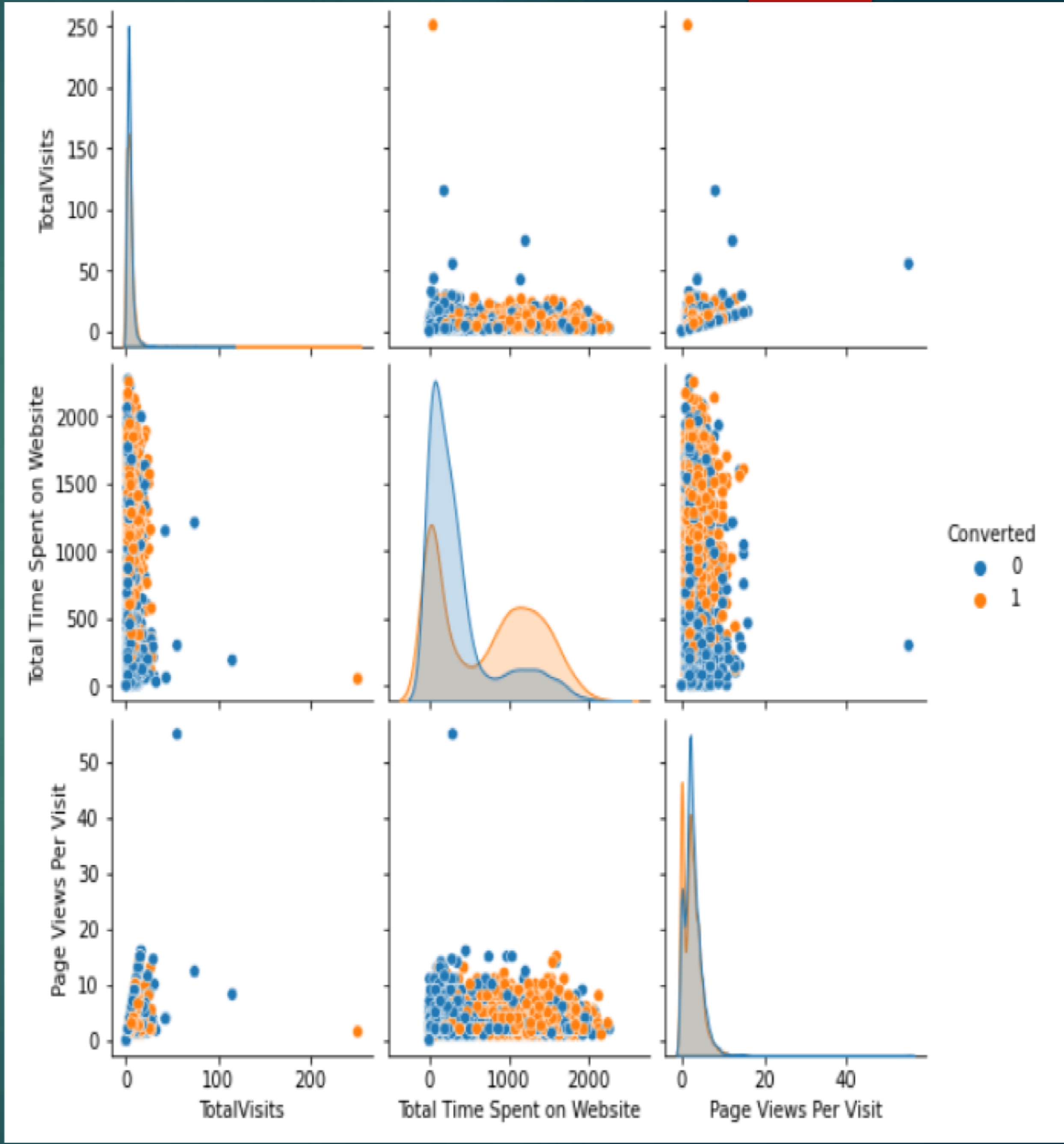






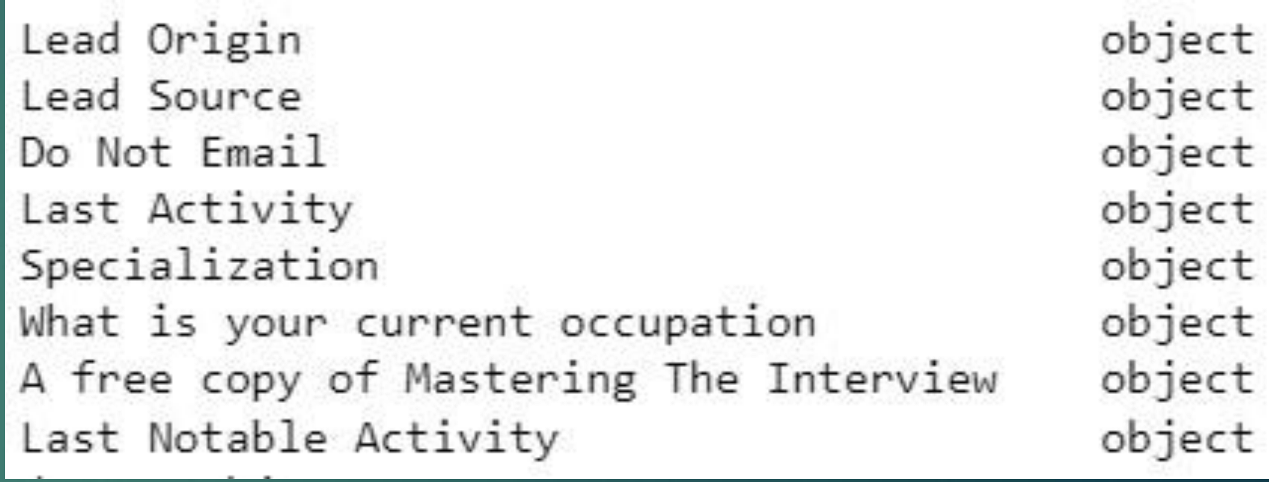
Bivariate Analysis

- Plotting numerical variables present in the data according the column Converted.



Dummy creation

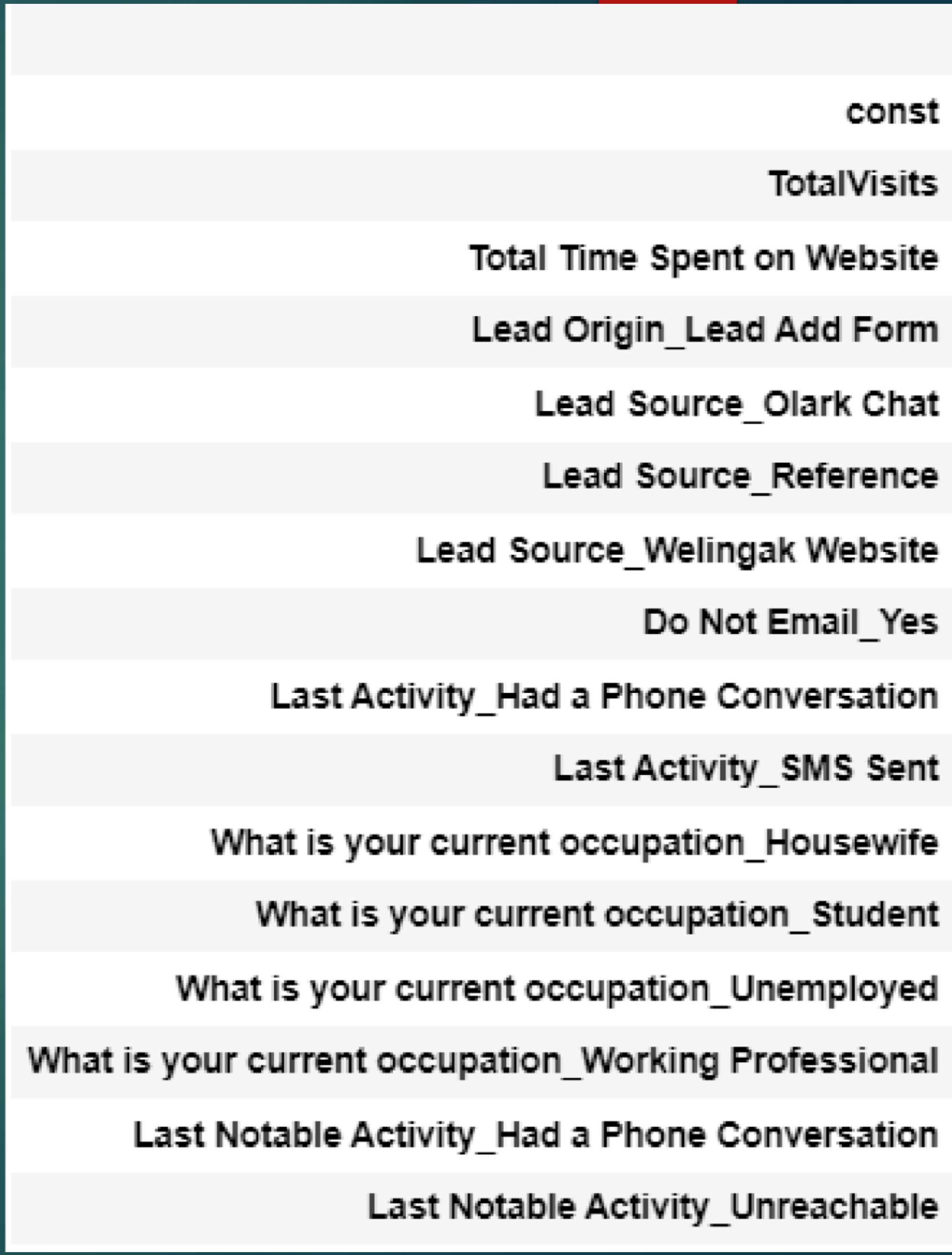
- We will make dummy variables for all the categorical columns present in the data set, as it will help us in building an optimal model.
- All the categorical columns can be seen in the adjacent image.



```
Lead Origin                                object
Lead Source                                object
Do Not Email                               object
Last Activity                              object
Specialization                             object
What is your current occupation            object
A free copy of Mastering The Interview     object
Last Notable Activity                      object
```

Feature Selection

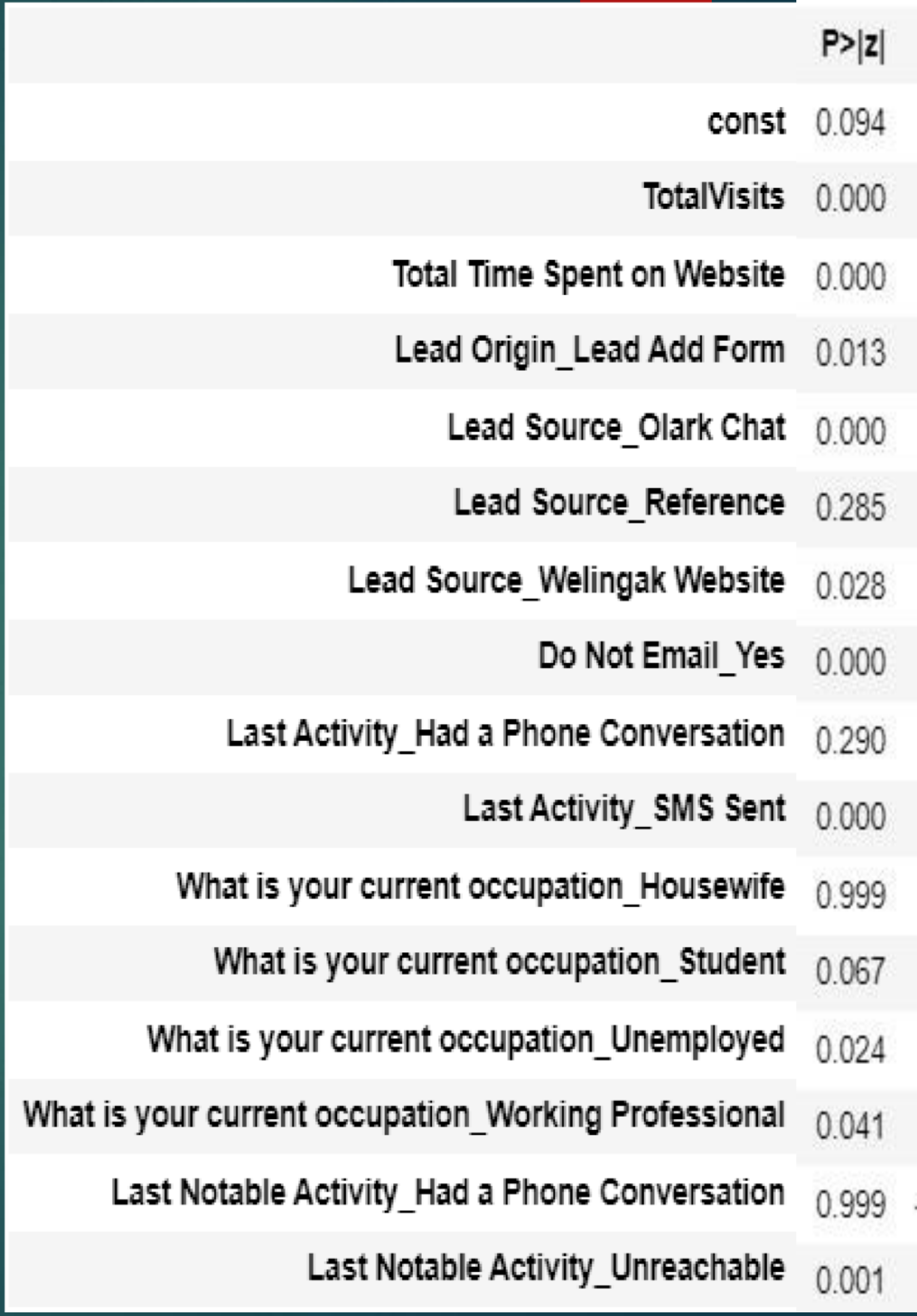
- We imported the RFE (recursive feature selection) method to select the 15 most important features from the data set.
- The most important 15 features can be seen in the adjacent image.



const
TotalVisits
Total Time Spent on Website
Lead Origin_Lead Add Form
Lead Source_Olark Chat
Lead Source_Reference
Lead Source_Welingak Website
Do Not Email_Yes
Last Activity_Had a Phone Conversation
Last Activity_SMS Sent
What is your current occupation_Housewife
What is your current occupation_Student
What is your current occupation_Unemployed
What is your current occupation_Working Professional
Last Notable Activity_Had a Phone Conversation
Last Notable Activity_Unreachable

Manually Dropping the features

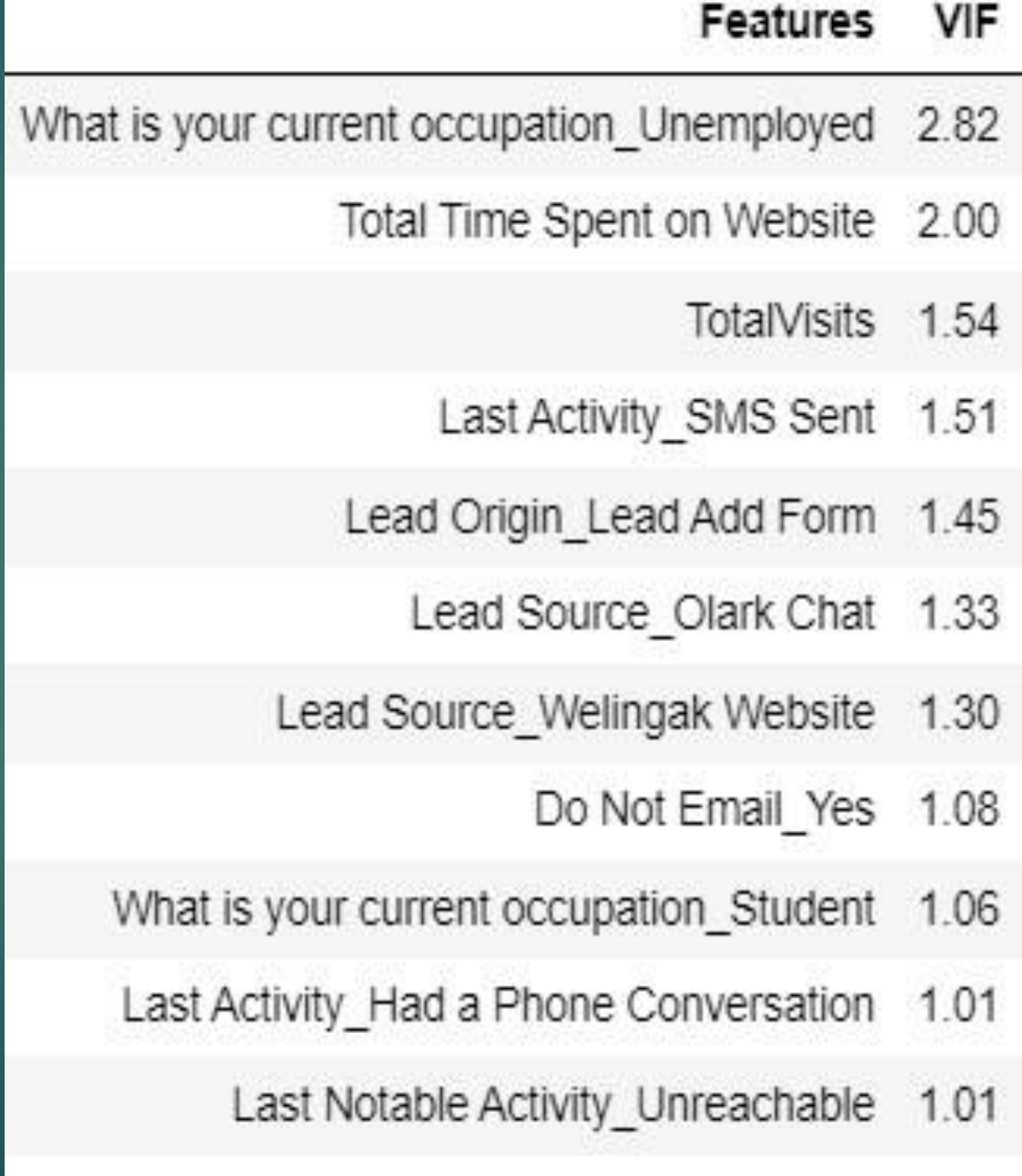
- After rfe method we will drop the features with more then 0.05 p-value in our logictic regression model.
- It can be seen in the image that the feature What is your current occupation_Housewife has the highest p-value in the model so we can drop it.
- Similarly, in the next models we will drop different columns like:
  - Last Notable Activity_Had a Phone Conversation
  - Lead Source_Reference
  - What is your current   occupation_Working Professional



	P>|z|
const	0.094
TotalVisits	0.000
Total Time Spent on Website	0.000
Lead Origin_Lead Add Form	0.013
Lead Source_Olark Chat	0.000
Lead Source_Reference	0.285
Lead Source_Welingak Website	0.028
Do Not Email_Yes	0.000
Last Activity_Had a Phone Conversation	0.290
Last Activity_SMS Sent	0.000
What is your current occupation_Housewife	0.999
What is your current occupation_Student	0.067
What is your current occupation_Unemployed	0.024
What is your current occupation_Working Professional	0.041
Last Notable Activity_Had a Phone Conversation	0.999
Last Notable Activity_Unreachable	0.001

Calculating the VIF

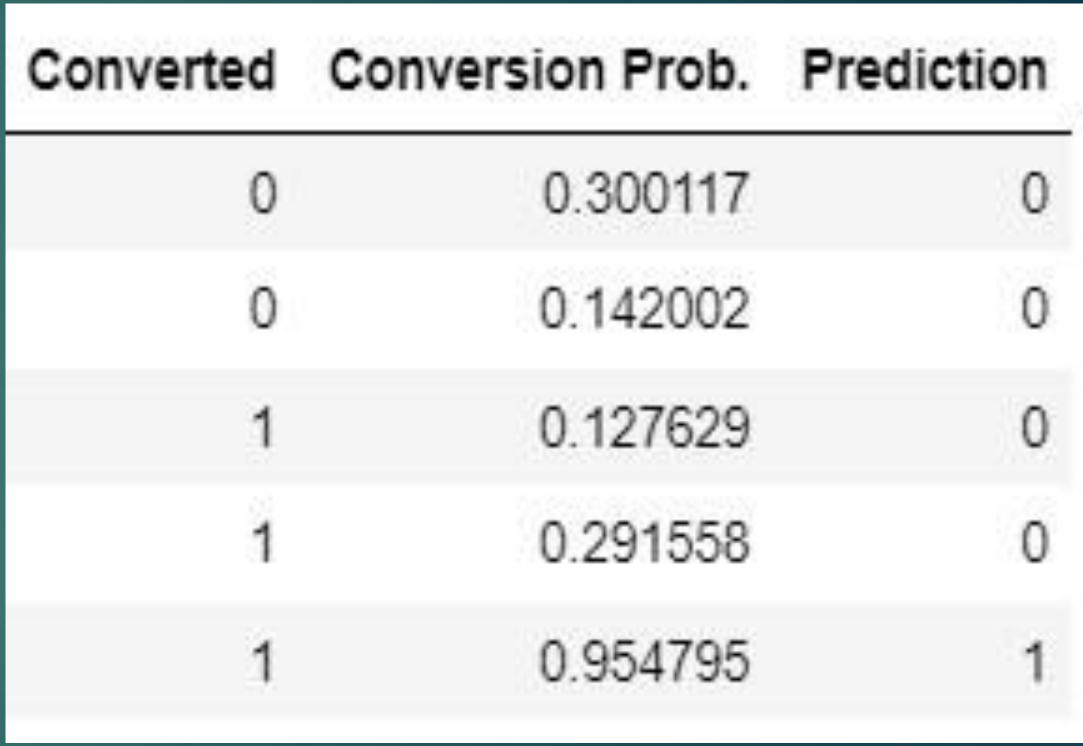
- Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.
- We will be calculating the VIF of the left over features in our model.
- The VIF values can be seen in the adjacent image.
- A VIF smaller then 5 is good to go.



Features	VIF
What is your current occupation_Unemployed	2.82
Total Time Spent on Website	2.00
TotalVisits	1.54
Last Activity_SMS Sent	1.51
Lead Origin_Lead Add Form	1.45
Lead Source_Olark Chat	1.33
Lead Source_Welingak Website	1.30
Do Not Email_Yes	1.08
What is your current occupation_Student	1.06
Last Activity_Had a Phone Conversation	1.01
Last Notable Activity_Unreachable	1.01

Predictions

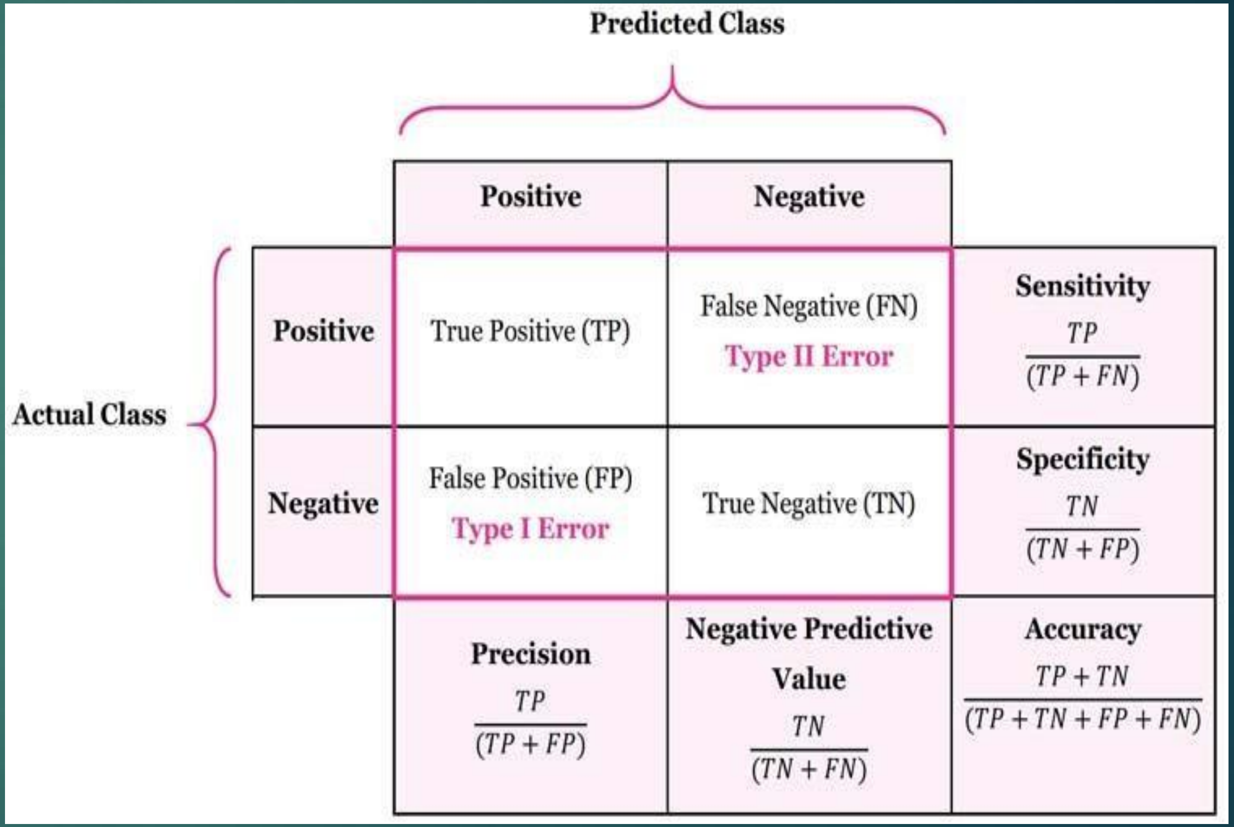
- We are left with 11 features in our final model. These 11 features will predict whether the lead has churned or not.
- The conversion probability can be predicted from our final model which is X_test_sm.
- We have chosen an arbitrary cut-off of 0.5 i.e. if the conversion probability of a lead is more then 0.5 then it will be predicted as 1 or the lead hasn't churned and vice versa with leads having probability less then 0.5.



Converted	Conversion Prob.	Prediction
0	0.300117	0
0	0.142002	0
1	0.127629	0
1	0.291558	0
1	0.954795	1

Model evaluation on the train dataset

- We will import the confusion matrix which will help us to calculate various performance matrix
- We will evaluate our model with the help of certain metrics like accuracy score, sensitivity, specificity.



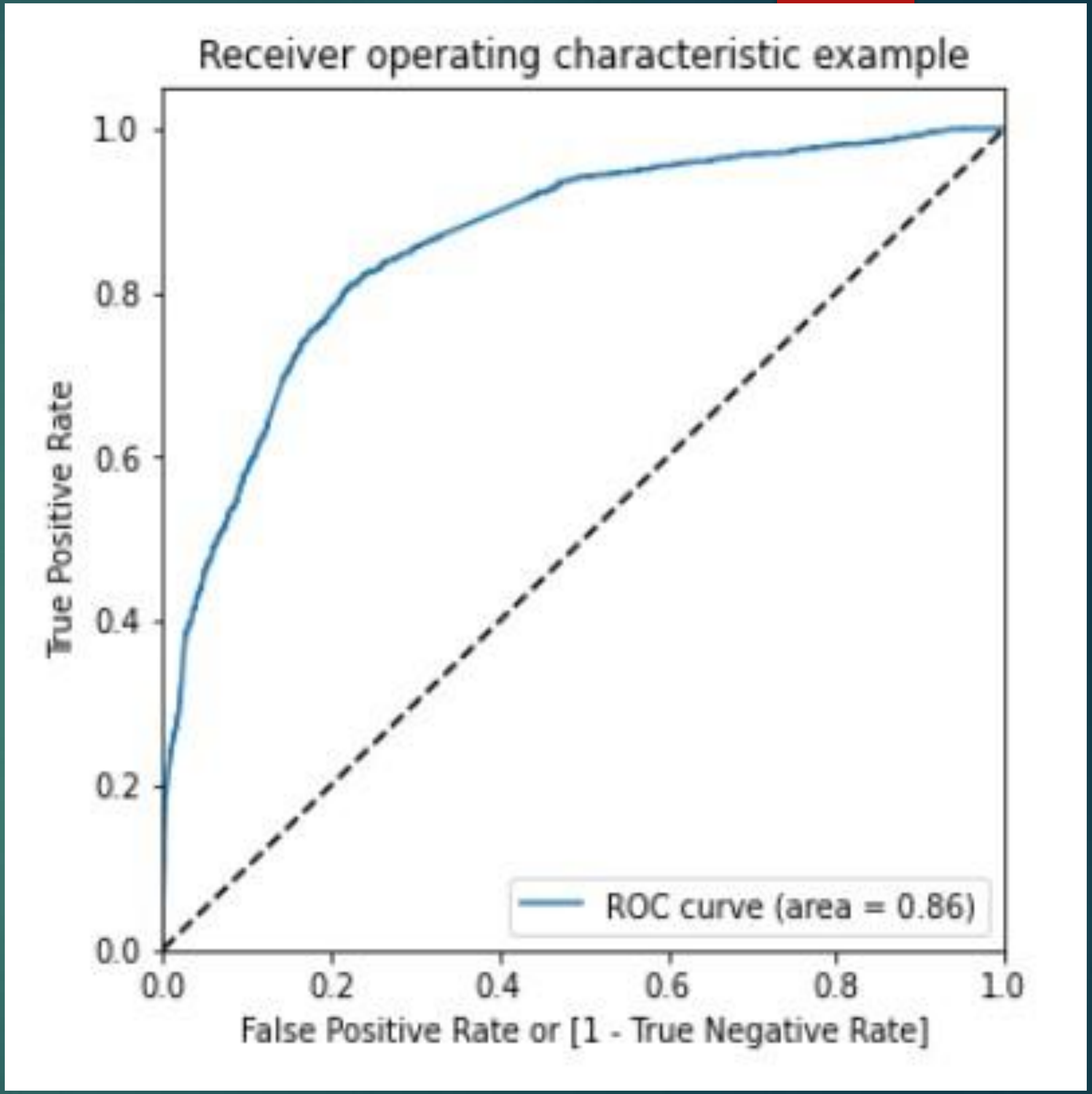
		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Metrics on train set

- Accuracy = 78.8%
- Sensitivity = 74%
- Specificity = 83.4%
- Precision = 77.7%
- Recall = 79.3%

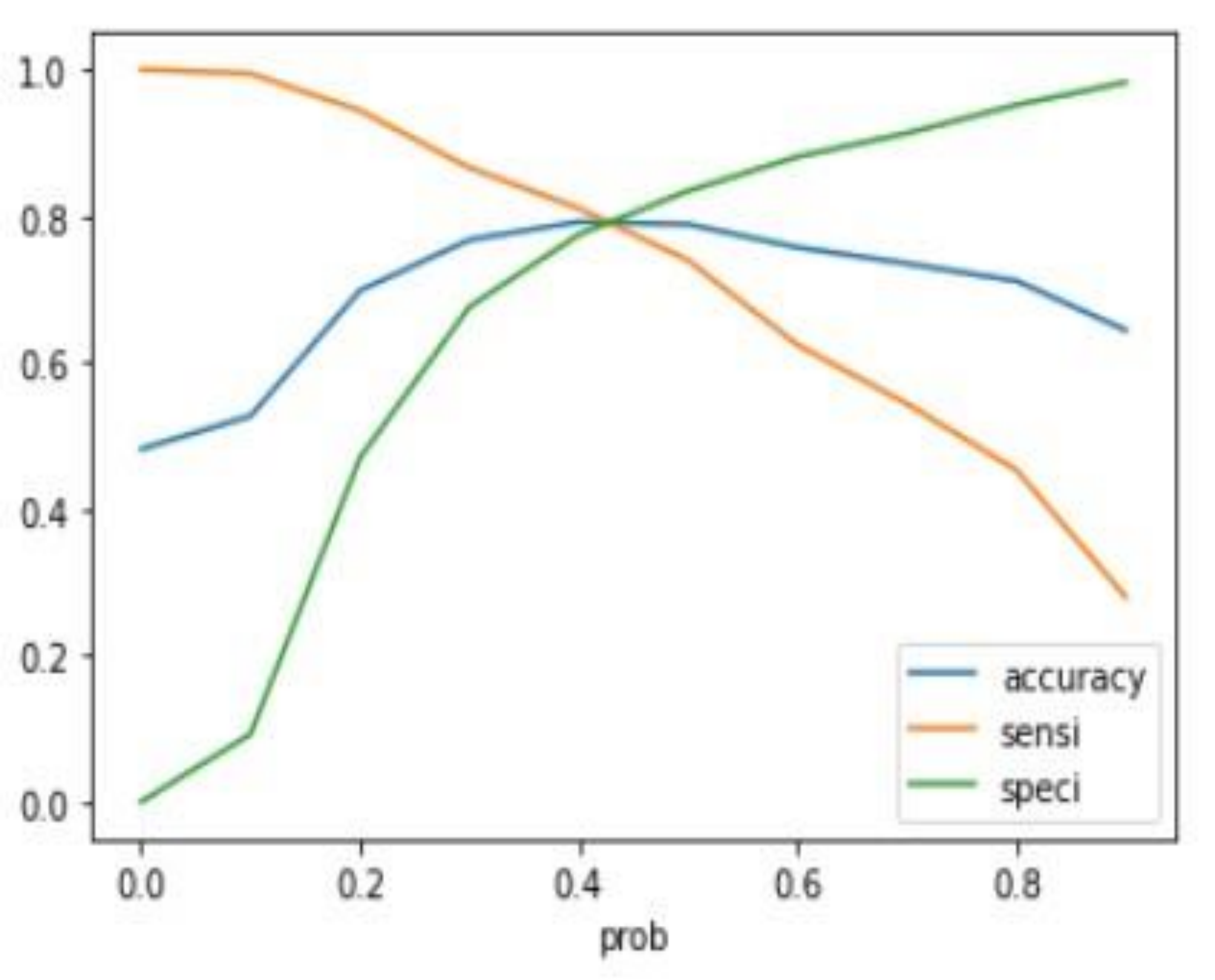
ROC Curve

- We will plot the roc curve as to check the auc (area under curve), more the area under the curve that means we have a good model which will cover most of the data points.
- As we can see in the adjacent image, auc is 0.86 which is quite decent amount of area.



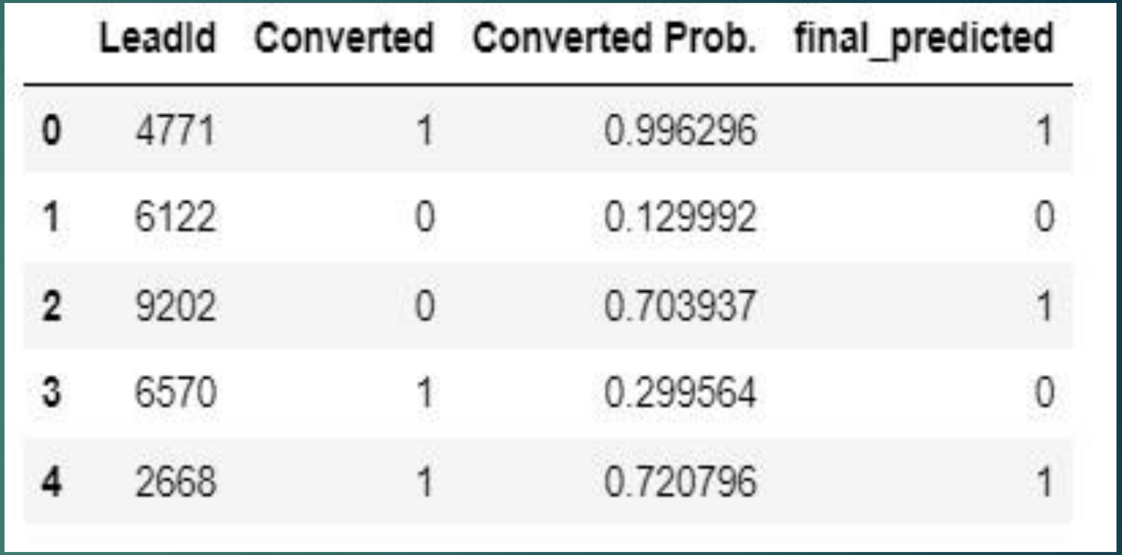
Finding an optimum cut-off value

- We have chosen an arbitrary cut-off of 0.5 for our model which is not accurate. Therefore, for finding an optimal cut-off we have to check the predictions of our model at every cut-off from 0.1 to 0.9 and then plotting the relationship between accuracy, sensitivity and specificity with the help of a graph.
- We can see that the three lines are intersecting at X=0.42, so we can take it as our optimal cut-off.
- The accuracy at this cut-off is 79% which is a good accuracy score.



Predictions on test set

- We have to retain only those columns which were there in our final train set model (X_test_sm), so we assign these columns to our test set (X_test_sm).
- We have to find the predictions i.e. conversion probability and merge it with the y_test series by converting both of them into y_test_sm dataframes and then merging them.
- Our final dataframe on test set is named as y_pred_final.



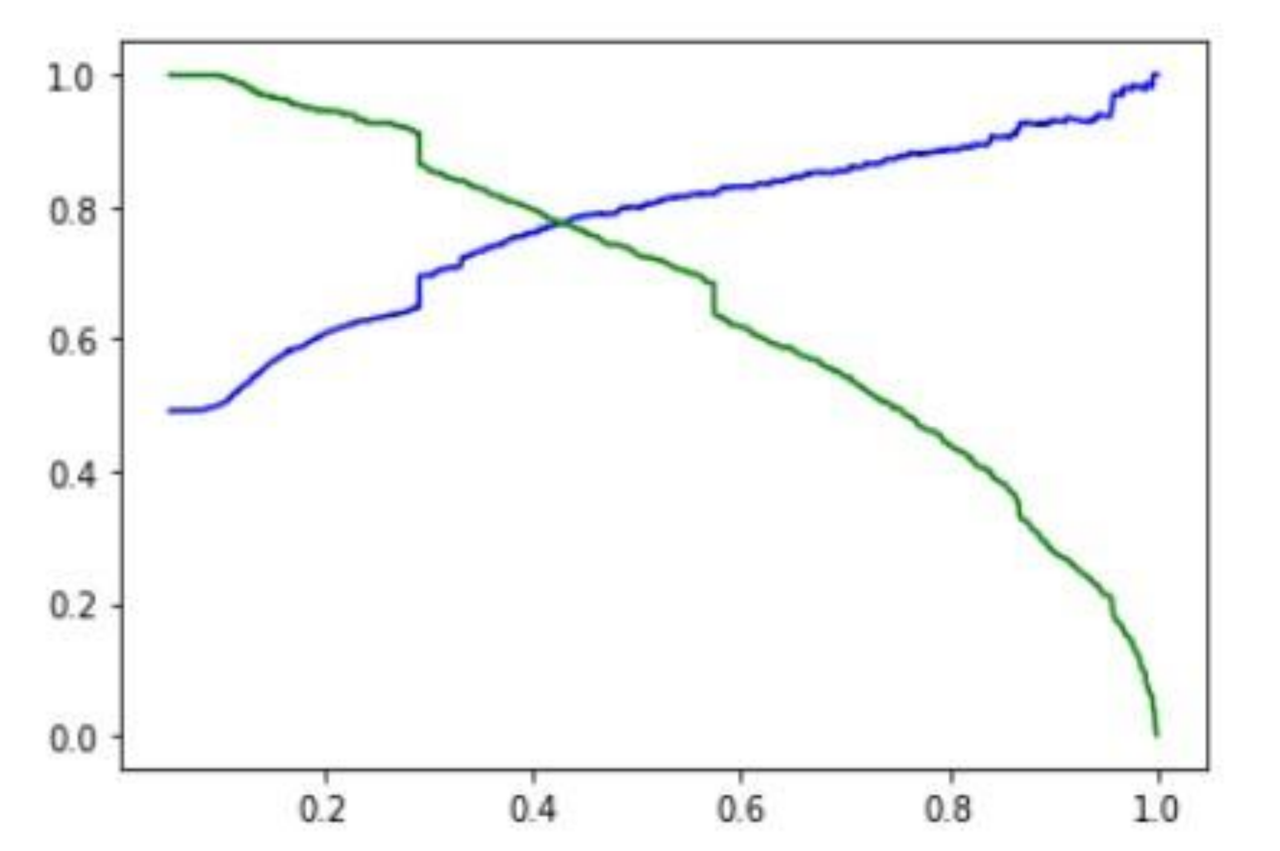
	LeadId	Converted	Converted Prob.	final_predicted
0	4771	1	0.996296	1
1	6122	0	0.129992	0
2	9202	0	0.703937	1
3	6570	1	0.299564	0
4	2668	1	0.720796	1

Performance metrics on the final test set.

- We plotted a confusion matrix for finding out the accuracy, sensitivity and specificity of out final model on test set.
- Metrics:
  - Accuracy = 77.6%
  - Sensitivity = 81.4%
  - Specificity = 74%

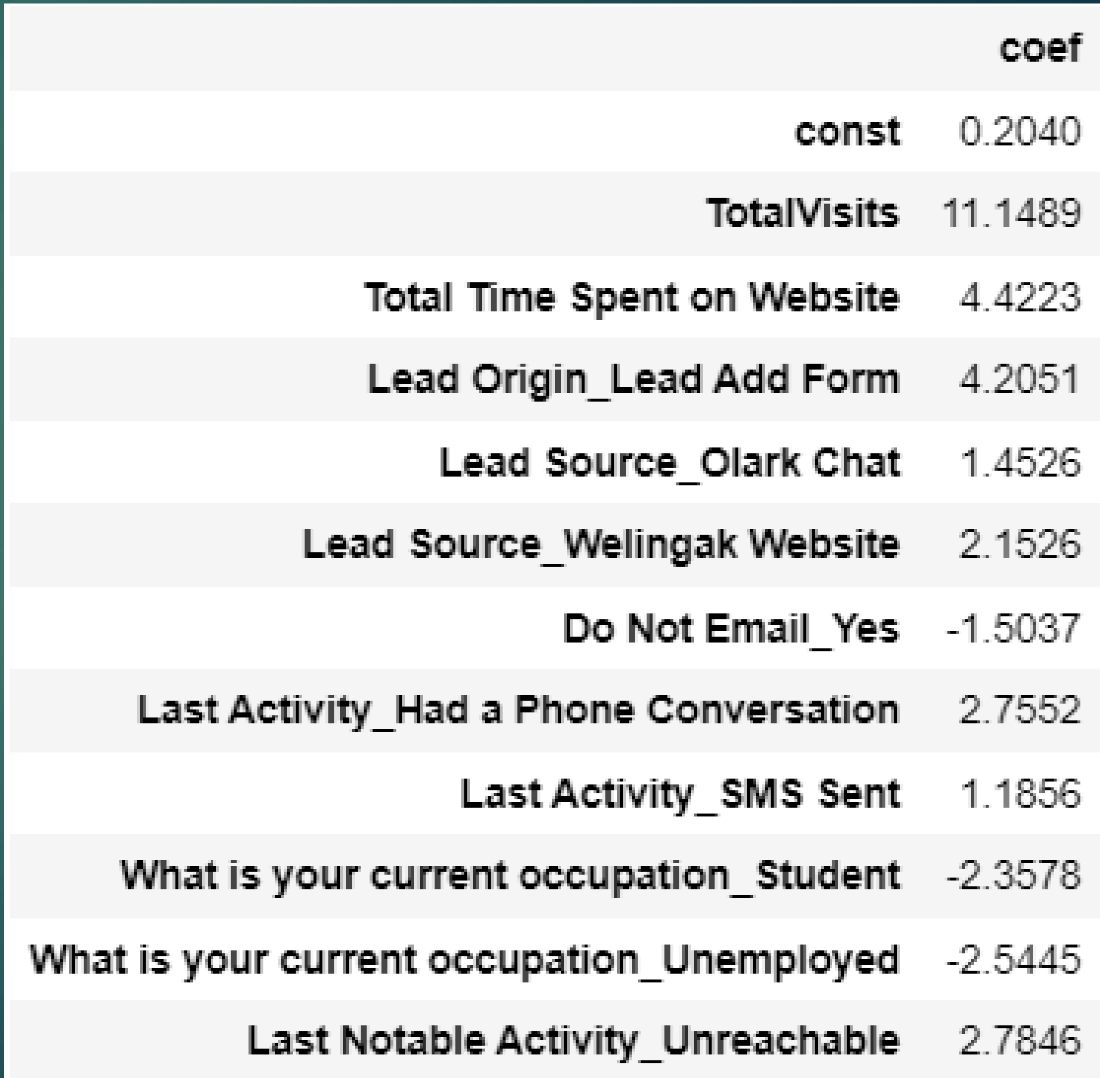
Precision and recall

- Precision = 74.37%
- Recall = 81.4%



Inferences for business decision-making

- The columns shown in the adjacent image are the features which will affect the conversion probability of a lead.
- The columns TotalVisits, Total Time Spent on Website and Lead Origin_Lead Add Form affects the conversion probability in a positive manner or we can say that:
- Conversion probability of a lead will be high if:
  - Total number of visits is high
  - Total time spent by that lead on the website is high .
  - And if the lead origin is lead add form.
- Conversion probability of a lead will be low if:
  - That lead is unemployed.
  - That lead is a student.
  - If he has chosen do not email as yes.



	coef
const	0.2040
TotalVisits	11.1489
Total Time Spent on Website	4.4223
Lead Origin_Lead Add Form	4.2051
Lead Source_Olark Chat	1.4526
Lead Source_Welingak Website	2.1526
Do Not Email_Yes	-1.5037
Last Activity_Had a Phone Conversation	2.7552
Last Activity_SMS Sent	1.1856
What is your current occupation_Student	-2.3578
What is your current occupation_Unemployed	-2.5445
Last Notable Activity_Unreachable	2.7846