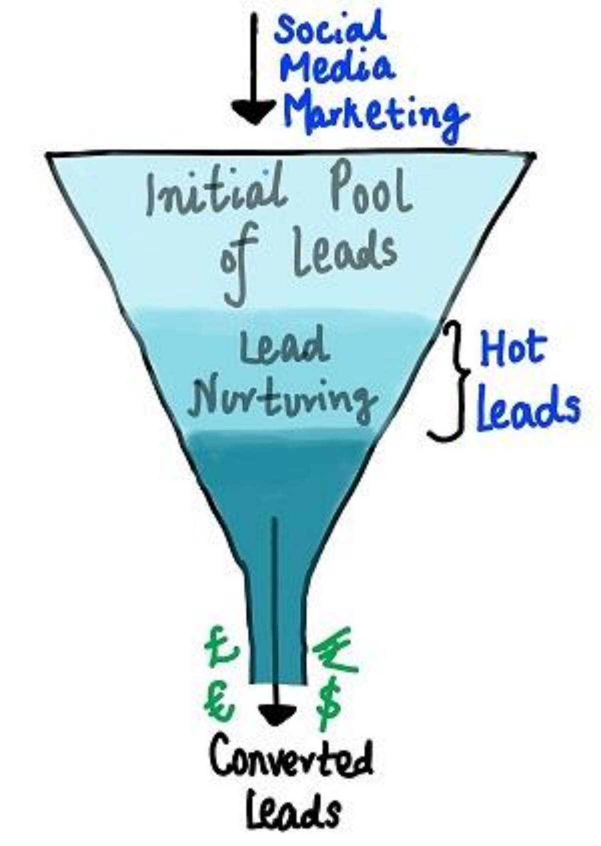
LEAD SCORING CASE STUDY

LOGISTIC REGRESSION

PRESENTED BY: DISHARI DEBNATH

PROBLEM STATEMENT

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. When people fill up a form providing their email address or phone number, they are classified to be a lead. Although X Education gets a lot of leads, its lead conversion rate is very poor.



BUSINESS GOAL

- The company wishes to identify the most potential leads, also known as ‘Hot Leads’. The goal is to build a model which assigns a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The target is to achieve a lead conversion rate of 80%

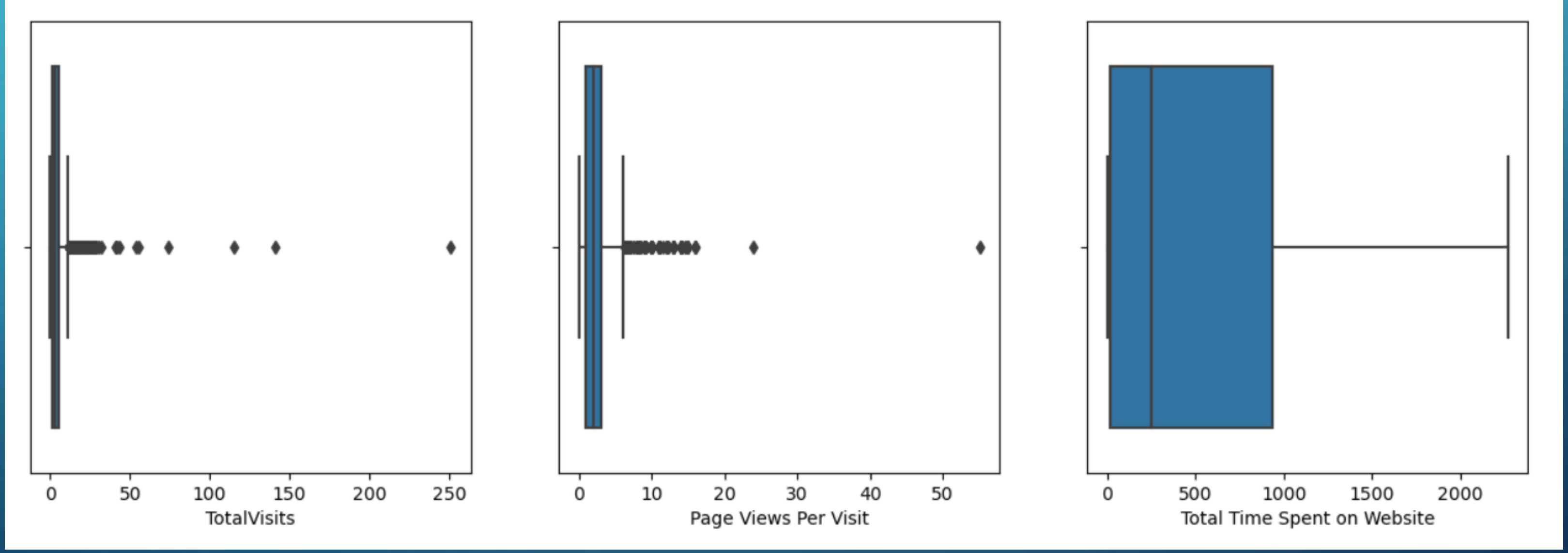
APPROACH

- Data importing/loading
- Data preparation
  - encoding categorical variable
  - handling null values
- EDA
  - Univariate Analysis
  - Outlier detection
  - Checking data imbalance
- Dummy Variable Creation
- Test Train Split
- Feature Scaling
- Correlations
- Model Building
  - Feature Selection Using RFE
  - Model Improvisation
- Model Finalisation
- Model Evaluation
- Recommendation

OUTLIER DETECTION

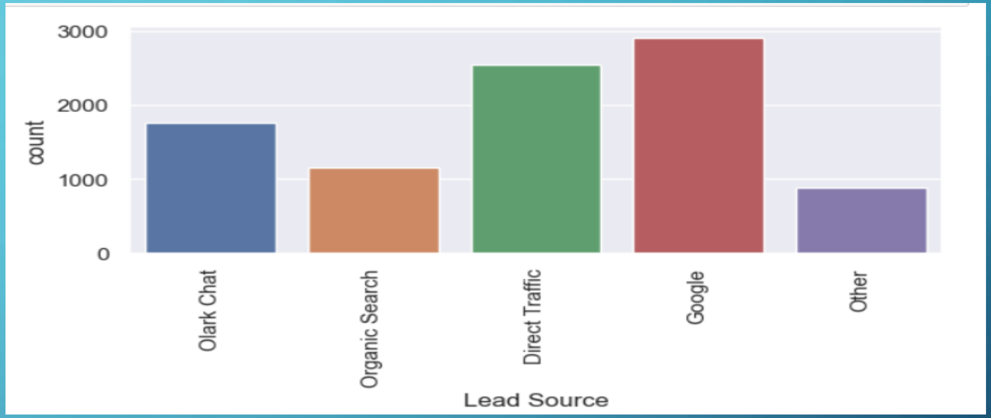
To detect the outliers we performed univariate analysis on numerical dtype columns.

Outliers are Present in both the Variables ‘TotalVisits’ and ‘Page Views Per Visit’ it should be treated and the value are spreaded above median highly in ‘Total Time Spent on Website’.

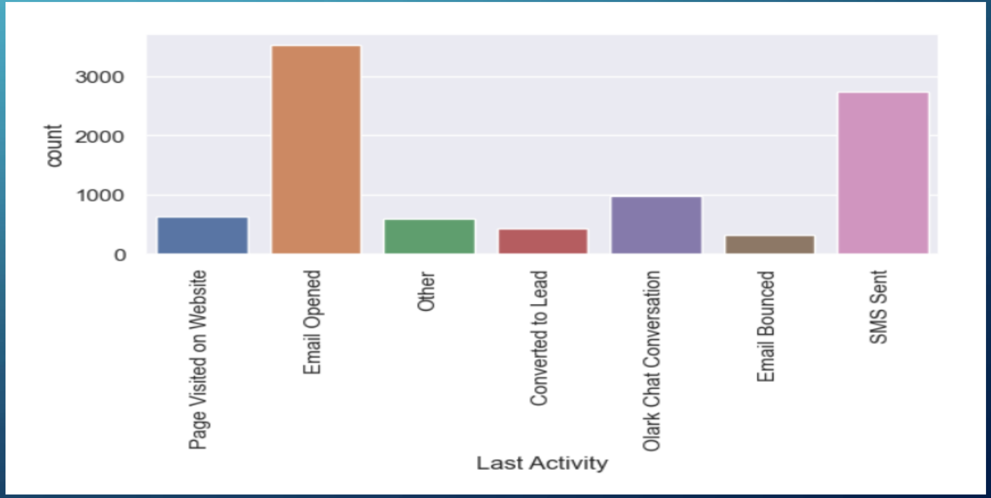


UNIVARIATE ANALYSIS

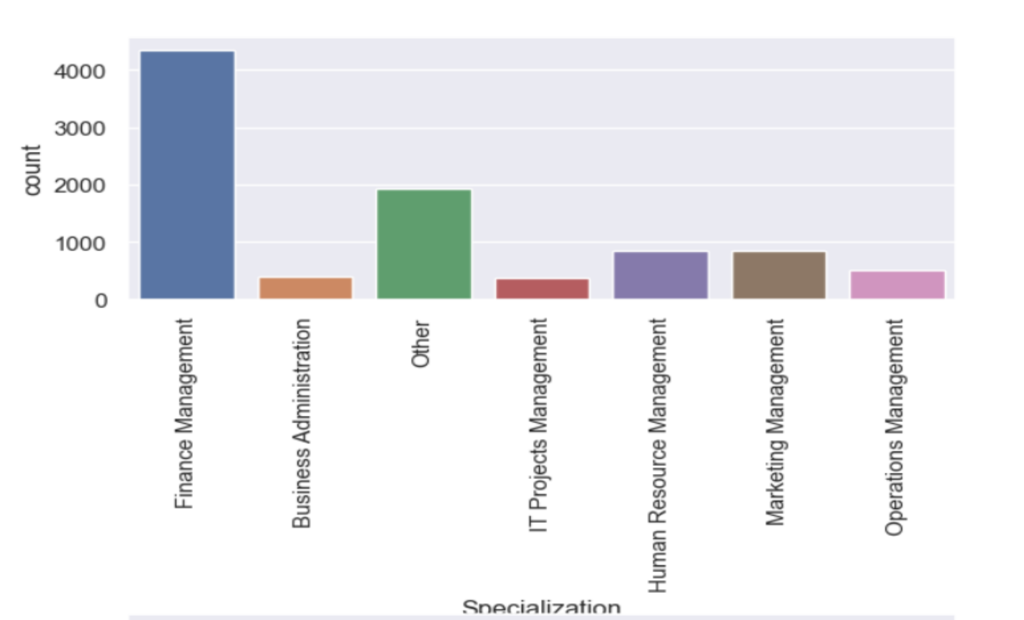
- Lead Source Vs Converted
  - In Lead Source Direct Traffic and Google are the two main source for Leads



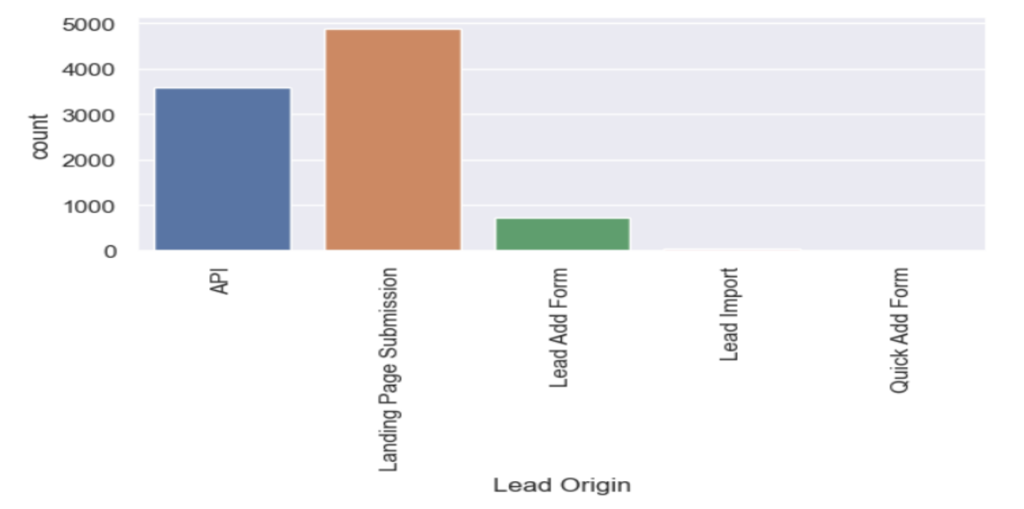
- Last Activity Vs Converted
  - The Number of values is High in Email Opened and SMS Sent in Last Activity



- Specialization Vs Coverted
  - Most of the people chooses Finance Management Specialization rather than other Specialization
  - The IT Project management have very lees so that most of the People not prefered this Specialization

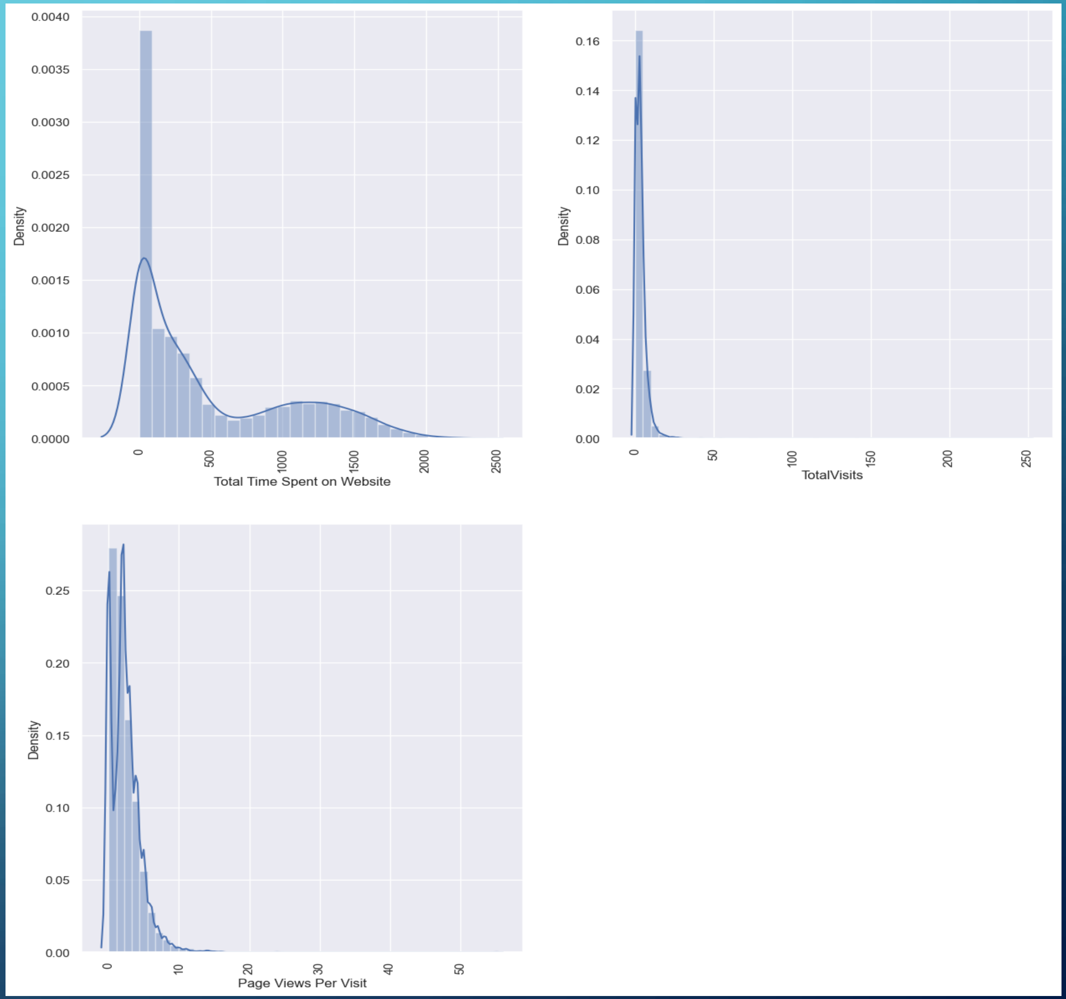


- Lead Source Vs Converted
  - Landing Page Submission and API to be a promising Lead Source for high conversion.



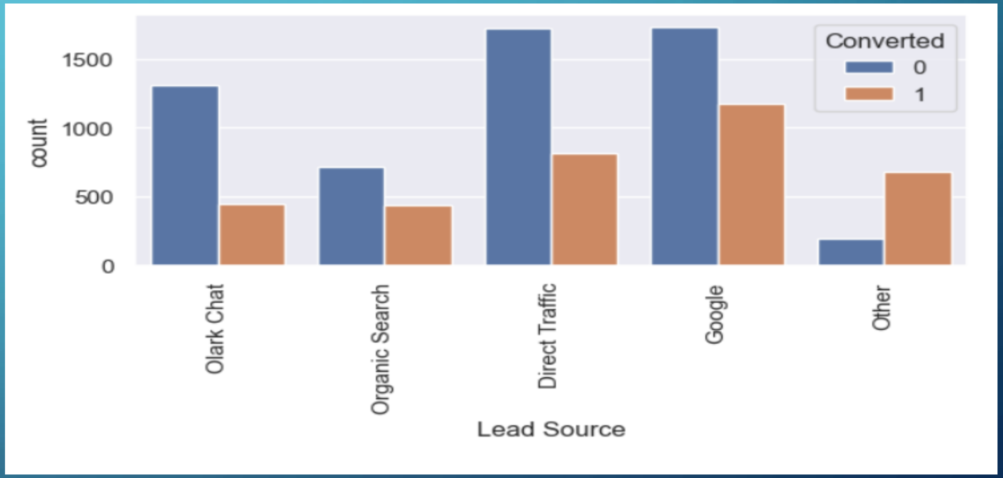
CONTINUOUS UNIVARIATE ANALYSIS

- None of the Continueous Variables are in Normal distribution
- Presence of Outliers in Total Visits and Page Views Per Visit
- In total visits more values is between 0-50 and page views per visits 0-20

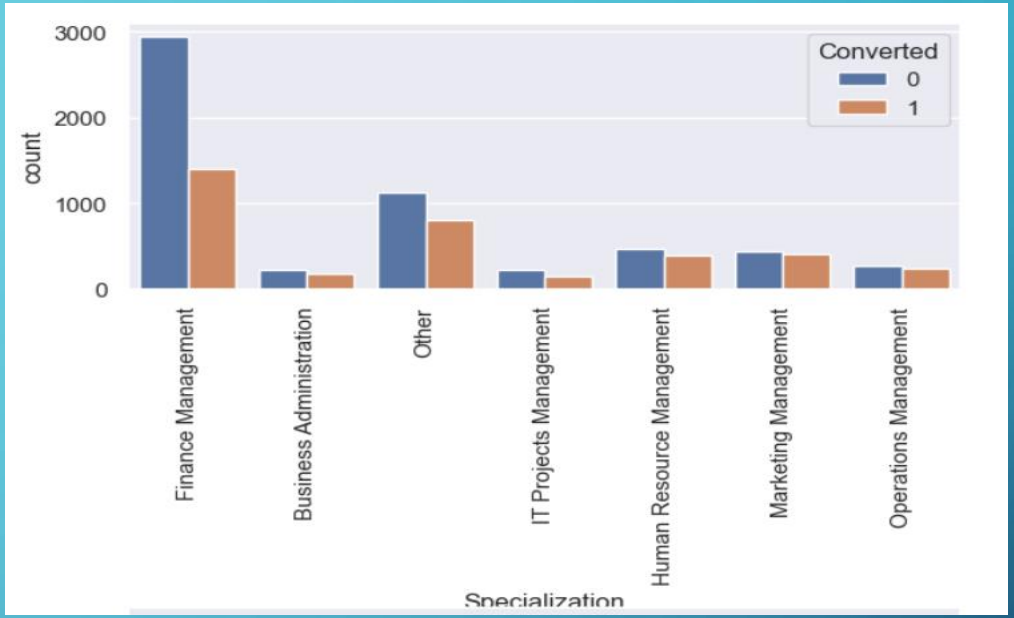


BIVARIATE ANALYSIS
ANALYSIS WITH RESPECT TO TARGET COLUMN CONVERTED

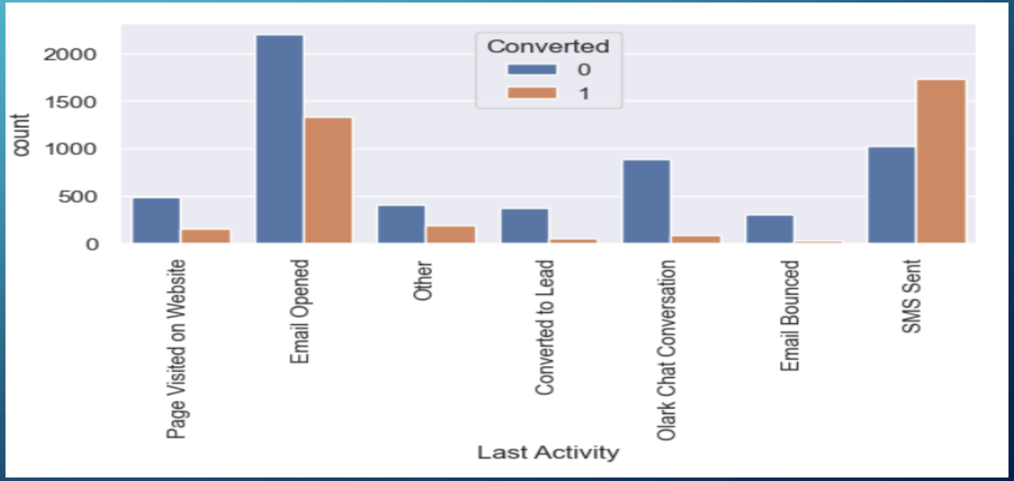
- Lead SourceVs Converted
  - In Lead Source The number of Hot leads is higher in Direct Traffic and Google less in Other Category



- Specialization Vs Converted
  - In Specialization the most of the leads are comes from Finance management but here Hot leads are lesseer than Cold leads.



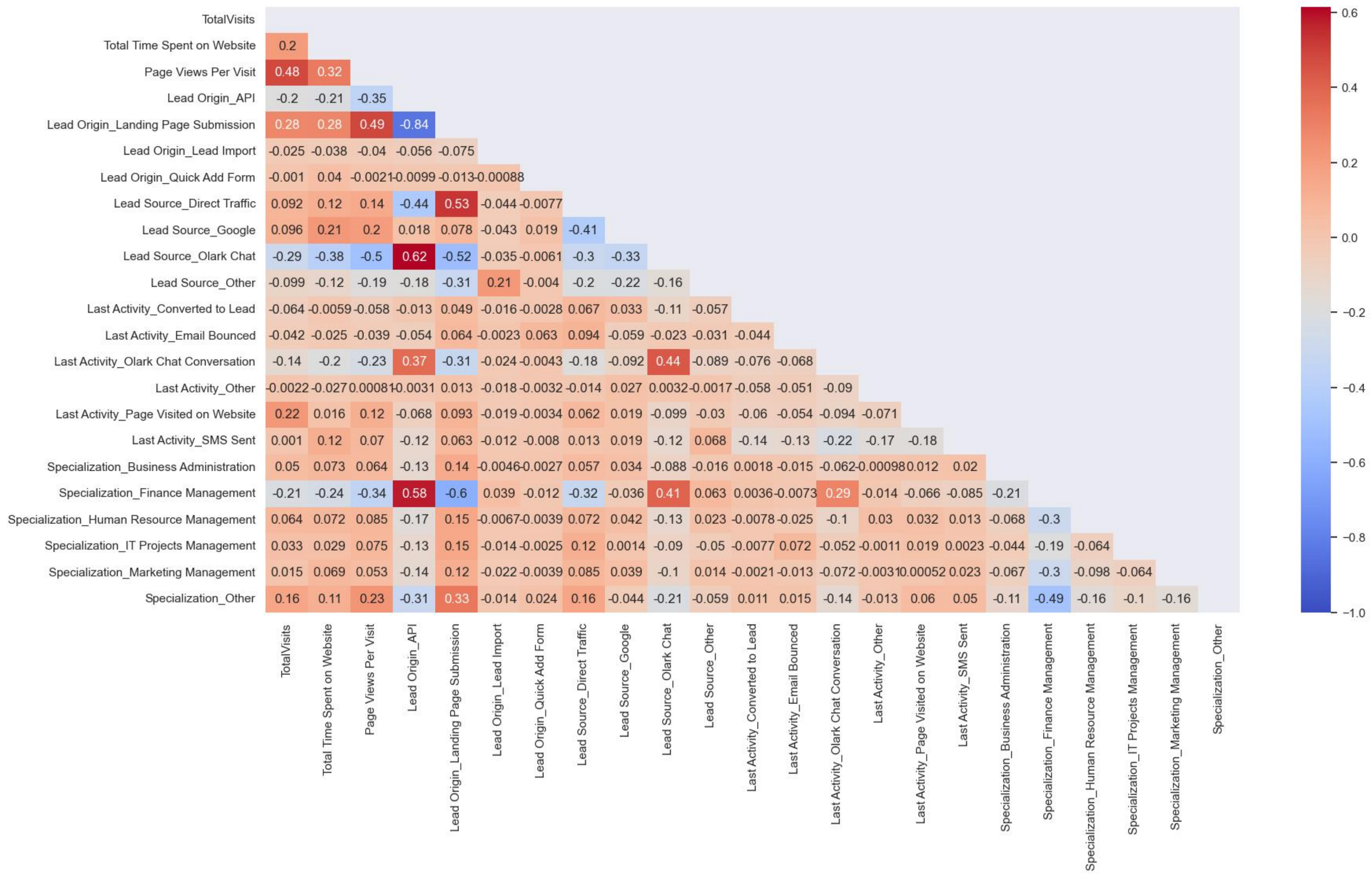
- Last Activity Vs. Converted
  - In Last Activity the number of Hot leads is higher in SMS and in EMAIL cold leads is higher than hot leads.
  - In Last Notable Activity it's mostly same as Last Activity.



FINDING THE CORRELATION USING HEATMAP

- We can see some columns are highly correlated

- We will let RFE to decide to drop columns or not



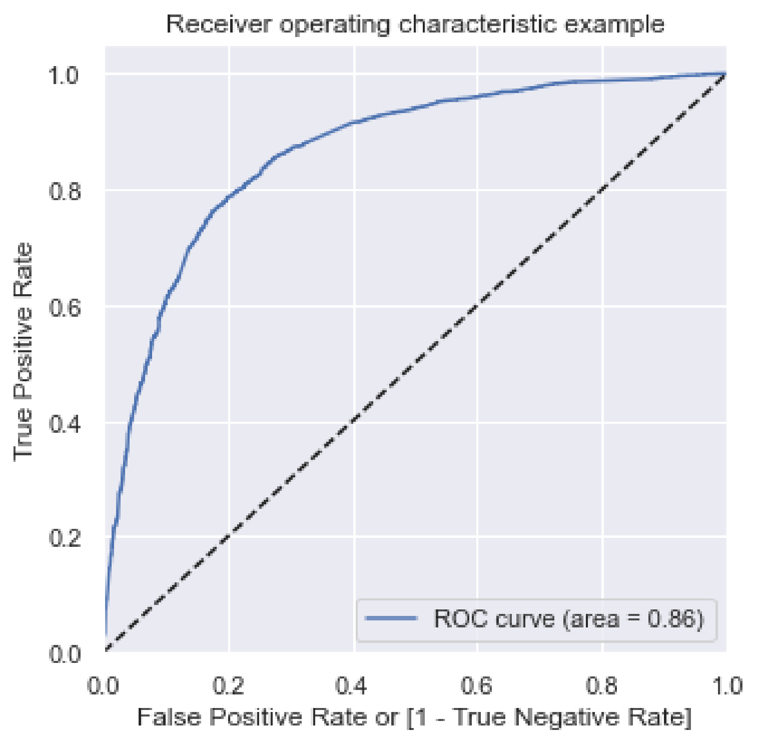
	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_API	Lead Origin_Landing Page Submission	Lead Origin_Lead Import	Lead Origin_Quick Add Form	Lead Source_Direct Traffic	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Other	Last Activity_Converted to Lead	Last Activity_Email Bounced	Last Activity_Olark Chat Conversation	Last Activity_Other	Last Activity_Page Visited on Website	Last Activity_SMS Sent	Specialization_Business Administration	Specialization_Finance Management	Specialization_Human Resource Management	Specialization_IT Projects Management	Specialization_Marketing Management	Specialization_Other
TotalVisits																							
Total Time Spent on Website	0.2																						
Page Views Per Visit	0.48	0.32																					
Lead Origin_API	-0.2	-0.21	-0.35																				
Lead Origin_Landing Page Submission	0.28	0.28	0.49	-0.84																			
Lead Origin_Lead Import	-0.025	-0.038	-0.04	-0.056	-0.075																		
Lead Origin_Quick Add Form	-0.001	0.04	-0.0021	-0.0099	-0.013	-0.00088																	
Lead Source_Direct Traffic	0.092	0.12	0.14	-0.44	0.53	-0.044	-0.0077																
Lead Source_Google	0.096	0.21	0.2	0.018	0.078	-0.043	0.019	-0.41															
Lead Source_Olark Chat	-0.29	-0.38	-0.5	0.62	-0.52	-0.035	-0.0061	-0.3	-0.33														
Lead Source_Other	-0.099	-0.12	-0.19	-0.18	-0.31	0.21	-0.004	-0.2	-0.22	-0.16													
Last Activity_Converted to Lead	-0.064	-0.0059	-0.058	-0.013	0.049	-0.016	-0.0028	0.067	0.033	-0.11	-0.057												
Last Activity_Email Bounced	-0.042	-0.025	-0.039	-0.054	0.064	-0.0023	0.063	0.094	-0.059	-0.023	-0.031	-0.044											
Last Activity_Olark Chat Conversation	-0.14	-0.2	-0.23	0.37	-0.31	-0.024	-0.0043	-0.18	-0.092	0.44	-0.089	-0.076	-0.068										
Last Activity_Other	-0.0022	-0.027	0.00081	-0.0031	0.013	-0.018	-0.0032	-0.014	0.027	0.0032	-0.0017	-0.058	-0.051	-0.09									
Last Activity_Page Visited on Website	0.22	0.016	0.12	-0.068	0.093	-0.019	-0.0034	0.062	0.019	-0.099	-0.03	-0.06	-0.054	-0.094	-0.071								
Last Activity_SMS Sent	0.001	0.12	0.07	-0.12	0.063	-0.012	-0.008	0.013	0.019	-0.12	0.068	-0.14	-0.13	-0.22	-0.17	-0.18							
Specialization_Business Administration	0.05	0.073	0.064	-0.13	0.14	-0.0046	-0.0027	0.057	0.034	-0.088	-0.016	0.0018	-0.015	-0.062	-0.00098	0.012	0.02						
Specialization_Finance Management	-0.21	-0.24	-0.34	0.58	-0.6	0.039	-0.012	-0.32	-0.036	0.41	0.063	0.0036	-0.0073	0.29	-0.014	-0.066	-0.085	-0.21					
Specialization_Human Resource Management	0.064	0.072	0.085	-0.17	0.15	-0.0067	-0.0039	0.072	0.042	-0.13	0.023	-0.0078	-0.025	-0.1	0.03	0.032	0.013	-0.068	-0.3				
Specialization_IT Projects Management	0.033	0.029	0.075	-0.13	0.15	-0.014	-0.0025	0.12	0.0014	-0.09	-0.05	-0.0077	0.072	-0.052	-0.0011	0.019	0.0023	-0.044	-0.19	-0.064			
Specialization_Marketing Management	0.015	0.069	0.053	-0.14	0.12	-0.022	-0.0039	0.085	0.039	-0.1	0.014	-0.0021	-0.013	-0.072	-0.0031	0.00052	0.023	-0.067	-0.3	-0.098	-0.064		
Specialization_Other	0.16	0.11	0.23	-0.31	0.33	-0.014	0.024	0.16	-0.044	-0.21	-0.059	0.011	0.015	-0.14	-0.013	0.06	0.05	-0.11	-0.49	-0.16	-0.1	-0.16	

MODEL BUILDING

- Splitting into Train-Test Set
- Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variable
- Build next model
- Eliminate variables based on high p-value
- Check VIF value for all existing columnsPredict using train set
- Evaluate accuracy and other metric
- Predict using text set
- Precision and Recall analysis on test predictions

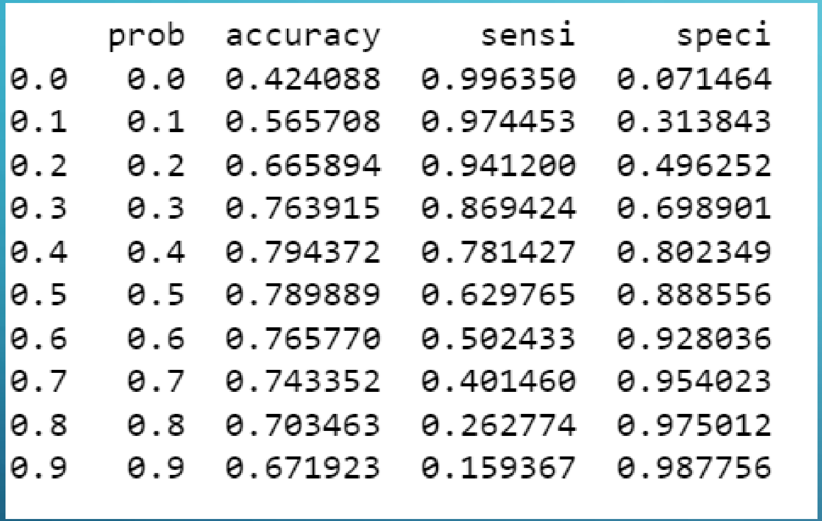
ROC CURVE

- An ROC curve shows tradoff between sensitivity and specificity (increase in one will cause decrease in other). The closer the curve follows the y-axis and the top border of the ROC space then it means more area under the curve and the more accurate the test prediction will be. The Closer the curve comes to the 45degree diagonal of the ROC space the lesser the area under the curve and less the test prediction will be.

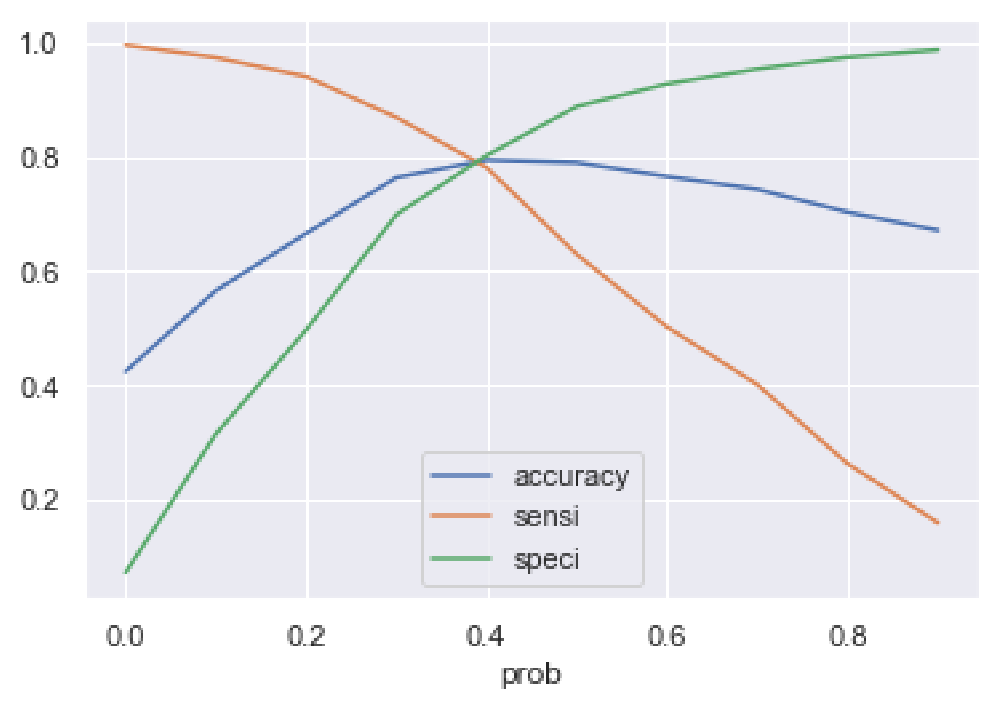


MODEL EVALUATION

- Accuracy, sensitivity and specificity for various probability cutoffs



```
     prob  accuracy     sensi     speci
0.0   0.0  0.424088  0.996350  0.071464
0.1   0.1  0.565708  0.974453  0.313843
0.2   0.2  0.665894  0.941200  0.496252
0.3   0.3  0.763915  0.869424  0.698901
0.4   0.4  0.794372  0.781427  0.802349
0.5   0.5  0.789889  0.629765  0.888556
0.6   0.6  0.765770  0.502433  0.928036
0.7   0.7  0.743352  0.401460  0.954023
0.8   0.8  0.703463  0.262774  0.975012
0.9   0.9  0.671923  0.159367  0.987756
```

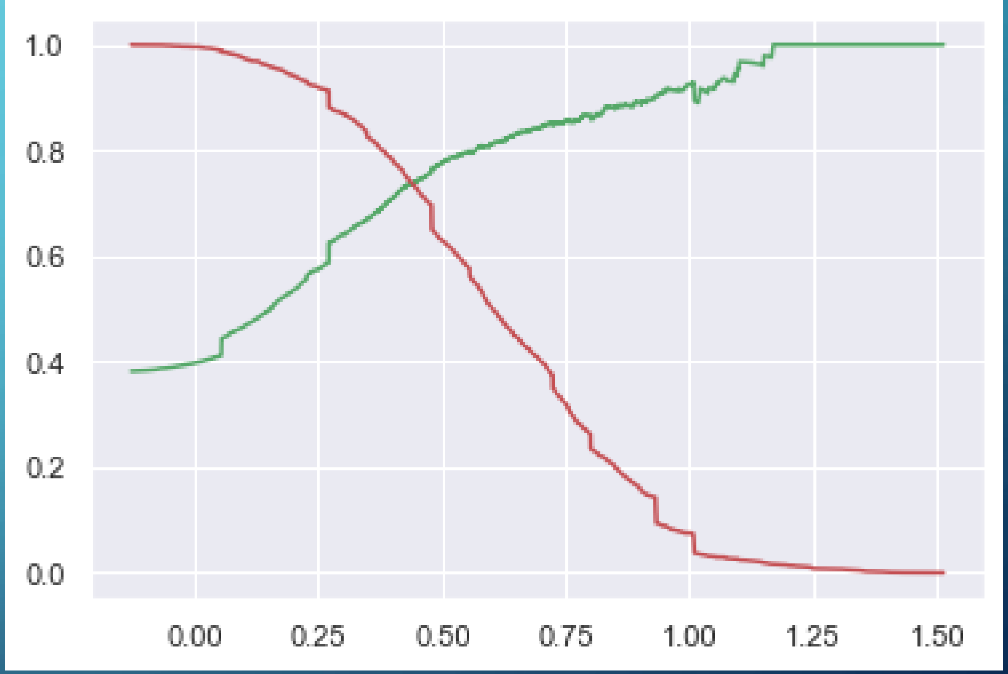


- From this we can have value 3.7 as the cutoff value

PRECISION VS RECALL

- Green-Precision
- Red-Recall

- Here we got 0.37 as the Cut-off as Precesion-Recall Threshold



ACCURACY, SENSITIVITY, SPECIFICITY

- **Sensitivity** (true positive rate) is the probability of a positive test result, conditioned on the individual truly being positive.
- **Specificity** (true negative rate) is the probability of a negative test result, conditioned on the individual truly being negative.

CONFUSION MATRIX

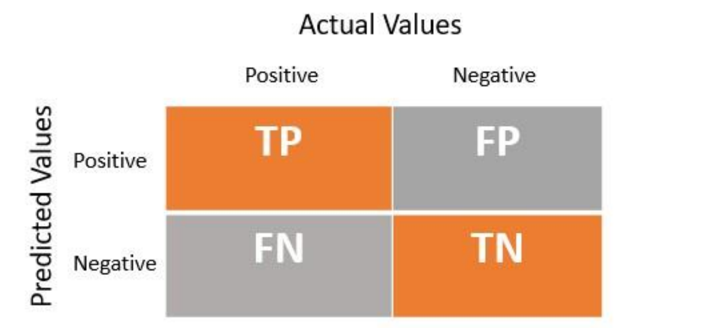
For our model

3556	446
913	1553

Accuracy: 0.7857142857142857

Sensitivity 0.8102189781021898

Specifitiy 0.7706146926536732



		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

Figure 1

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

CONCLUSION FROM LOGISTIC REGRESSION MODEL

- We can see that our model is doing well in test set also
- Sensitivity means how our model is telling that actually converted and model prdecited them as as converted.
- We can see that our model is giving about .80 sensitivity.
- it means that 80 percent time our model is able to predict (actually)converted as (prdicited)converted.