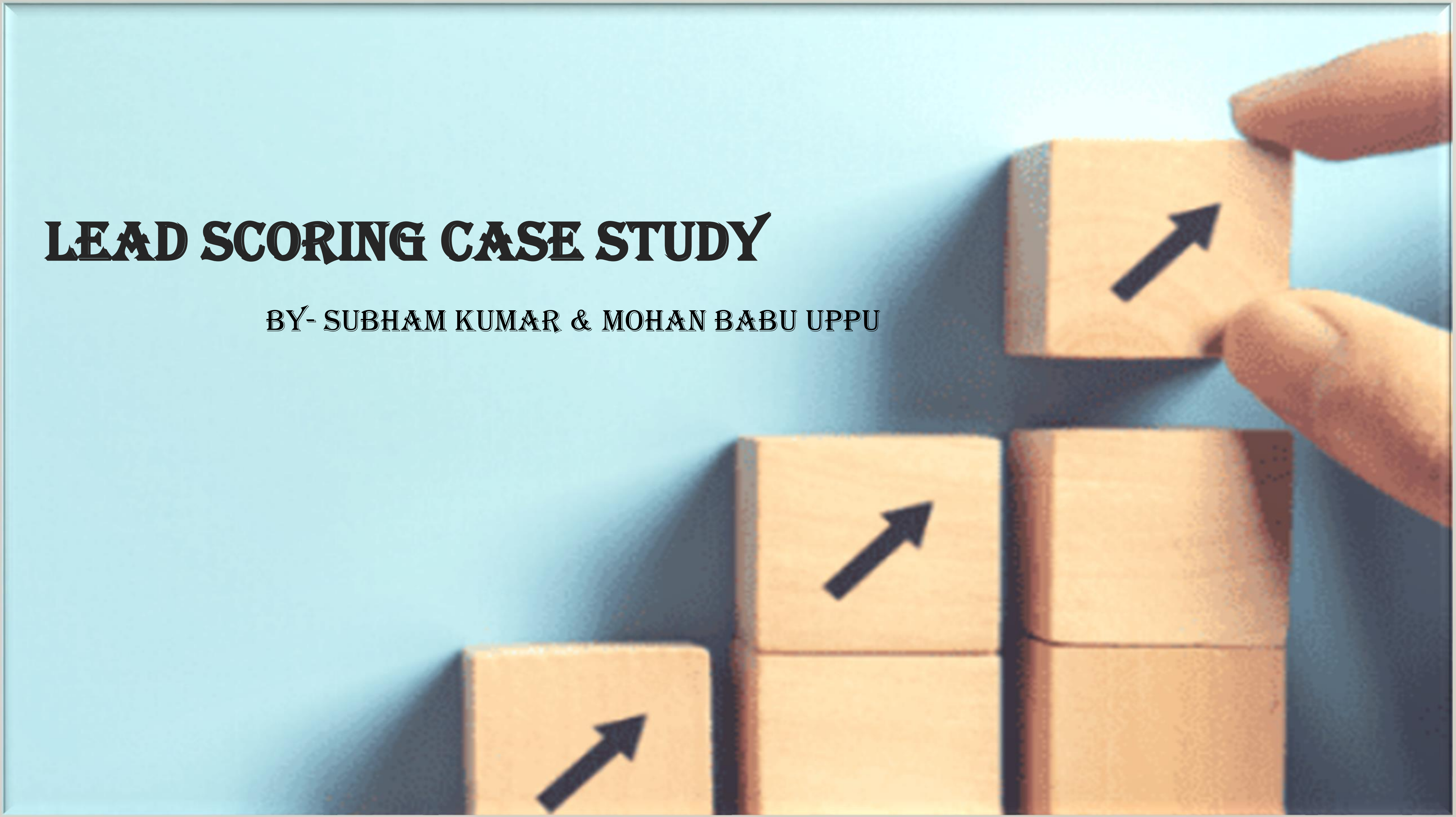
LEAD SCORING CASE STUDY

BY- SUBHAM KUMAR & MOHAN BABU UPPU



Problem statement

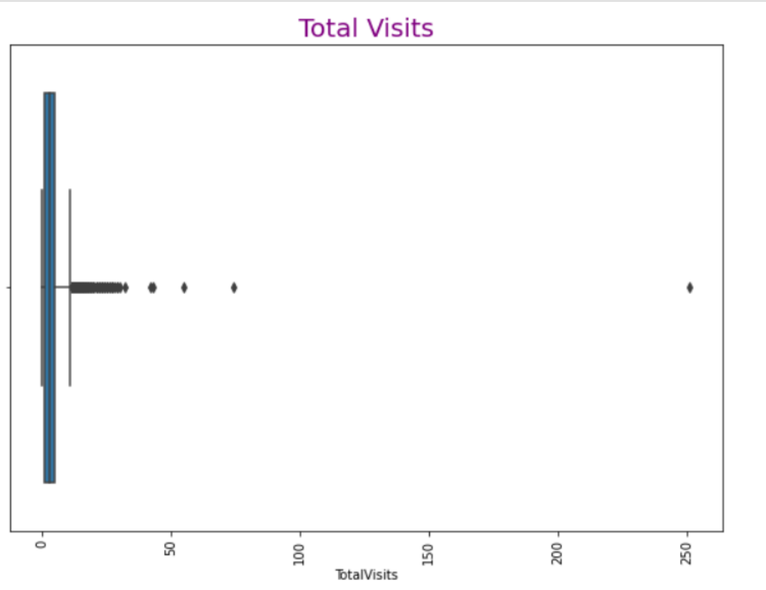
- Building a logistic regression model for X education by assigning a lead score between 0 and 100 for targeting particular leads to be converted or not. Here leads refer to the individuals finally enrolling for a course by giving contact details like phone number and email address
- The CEO has given a ballpark of the target lead conversion rate to be around 80%
- The proposed model should be able to adjust to if the company's requirement changes in the future as well

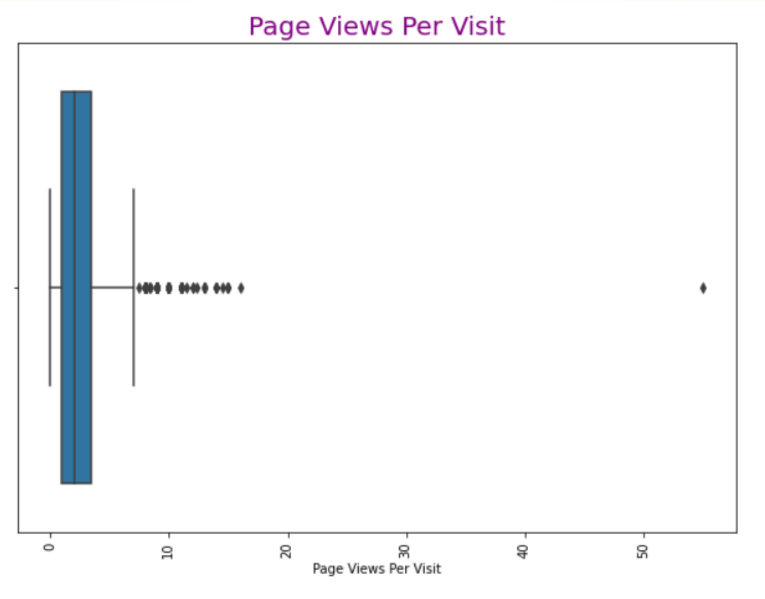
Procedure

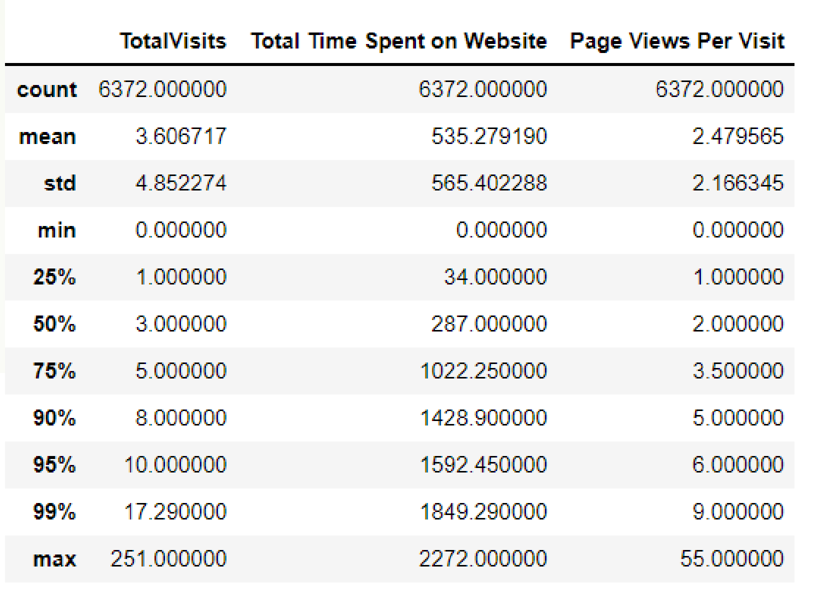
- **Data inspection and cleaning-** Checking the data types of different columns, investigating for null values and removing unwanted columns from the data frame
- **EDA and data preparation-** Data visualization ,Checking outliers, Creating dummy variables for categorical variables, Splitting the data into train and test data set, Scaling of the data and RFE
- **Building a logistic regression model-** Building a model on train data and dropping columns based on VIF and high p value
- **Matrix score test-** Finding confusion matrix, plotting ROC and finding the accuracy along with sensitivity and specificity
- **Prediction on Test data-** The final prediction of test data conveying evaluation on the basis of model accuracy, sensitivity and specificity

Checking for outliers

From the box plot we can observe that there are some outliers in TotalVisits and Page views per visit . However, if we see the summary of these variables at each percentile, we can find that the numbers are gradually increasing and these outliers are actually the individuals who are potential lead and should be contacted.



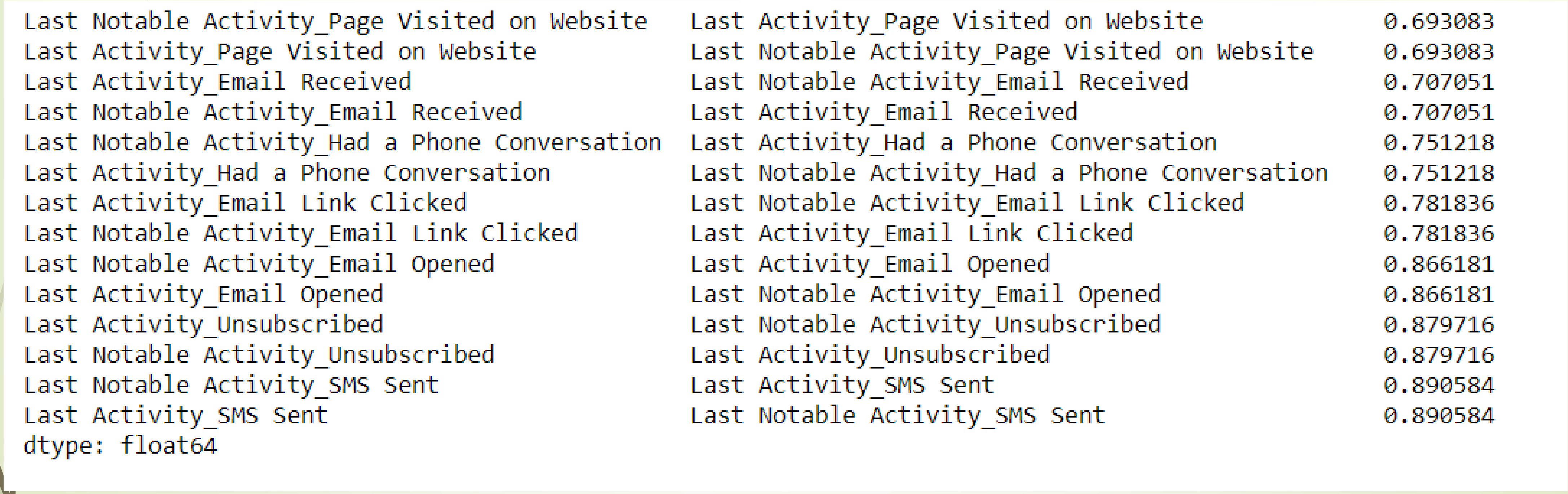




	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	6372.000000	6372.000000	6372.000000
mean	3.606717	535.279190	2.479565
std	4.852274	565.402288	2.166345
min	0.000000	0.000000	0.000000
25%	1.000000	34.000000	1.000000
50%	3.000000	287.000000	2.000000
75%	5.000000	1022.250000	3.500000
90%	8.000000	1428.900000	5.000000
95%	10.000000	1592.450000	6.000000
99%	17.290000	1849.290000	9.000000
max	251.000000	2272.000000	55.000000

Correlation

In this step, we find the correlations between variables which are more than 0.6 and drop them . Below are the variables which are dropped.



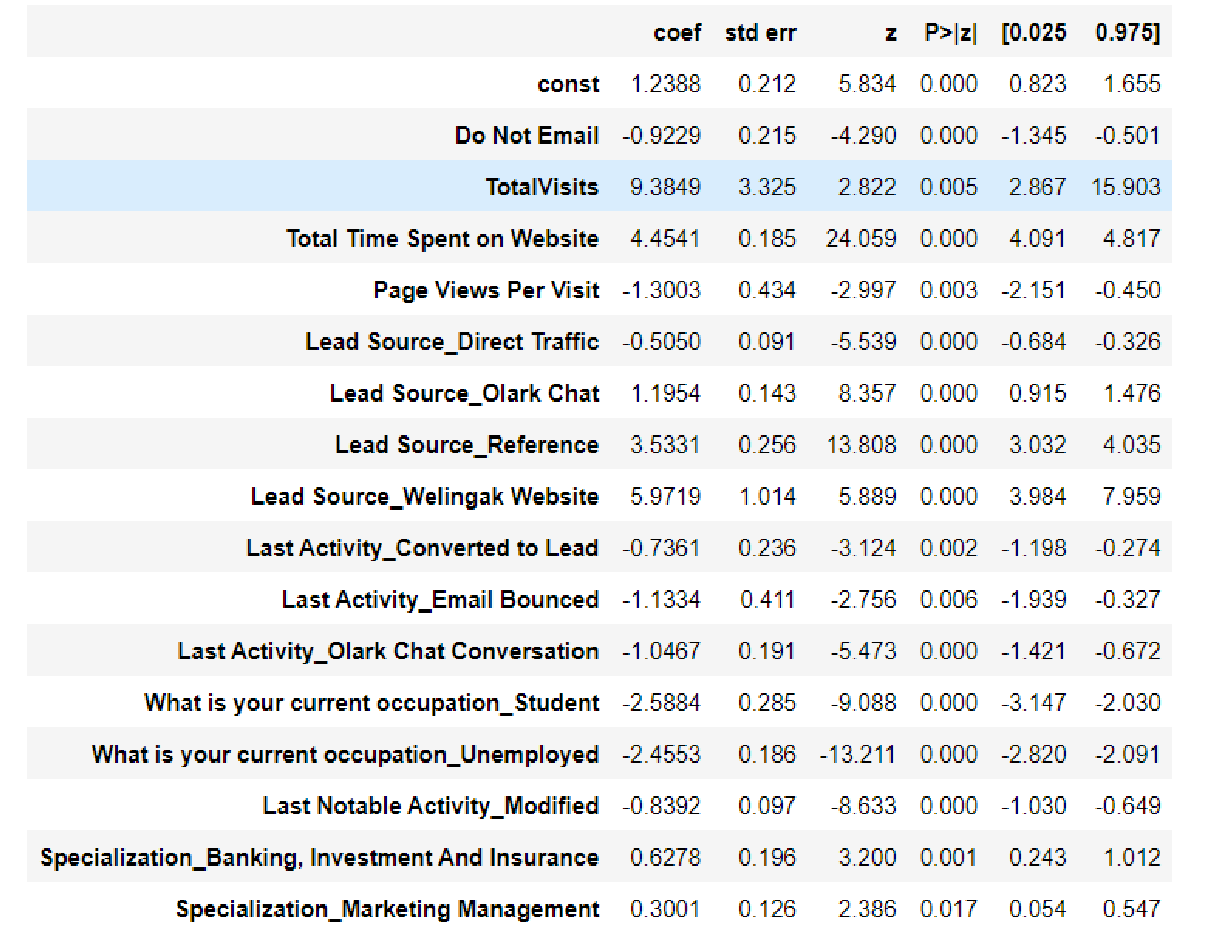
```
Last Notable Activity_Page Visited on Website   Last Activity_Page Visited on Website           0.693083
Last Activity_Page Visited on Website           Last Notable Activity_Page Visited on Website   0.693083
Last Activity_Email Received                    Last Notable Activity_Email Received            0.707051
Last Notable Activity_Email Received            Last Activity_Email Received                    0.707051
Last Notable Activity_Had a Phone Conversation  Last Activity_Had a Phone Conversation          0.751218
Last Activity_Had a Phone Conversation          Last Notable Activity_Had a Phone Conversation  0.751218
Last Activity_Email Link Clicked                Last Notable Activity_Email Link Clicked        0.781836
Last Notable Activity_Email Link Clicked        Last Activity_Email Link Clicked                0.781836
Last Notable Activity_Email Opened              Last Activity_Email Opened                      0.866181
Last Activity_Email Opened                      Last Notable Activity_Email Opened              0.866181
Last Activity_Unsubscribed                      Last Notable Activity_Unsubscribed              0.879716
Last Notable Activity_Unsubscribed              Last Activity_Unsubscribed                      0.879716
Last Notable Activity_SMS Sent                  Last Activity_SMS Sent                          0.890584
Last Activity_SMS Sent                          Last Notable Activity_SMS Sent                  0.890584
dtype: float64
```

Model Building

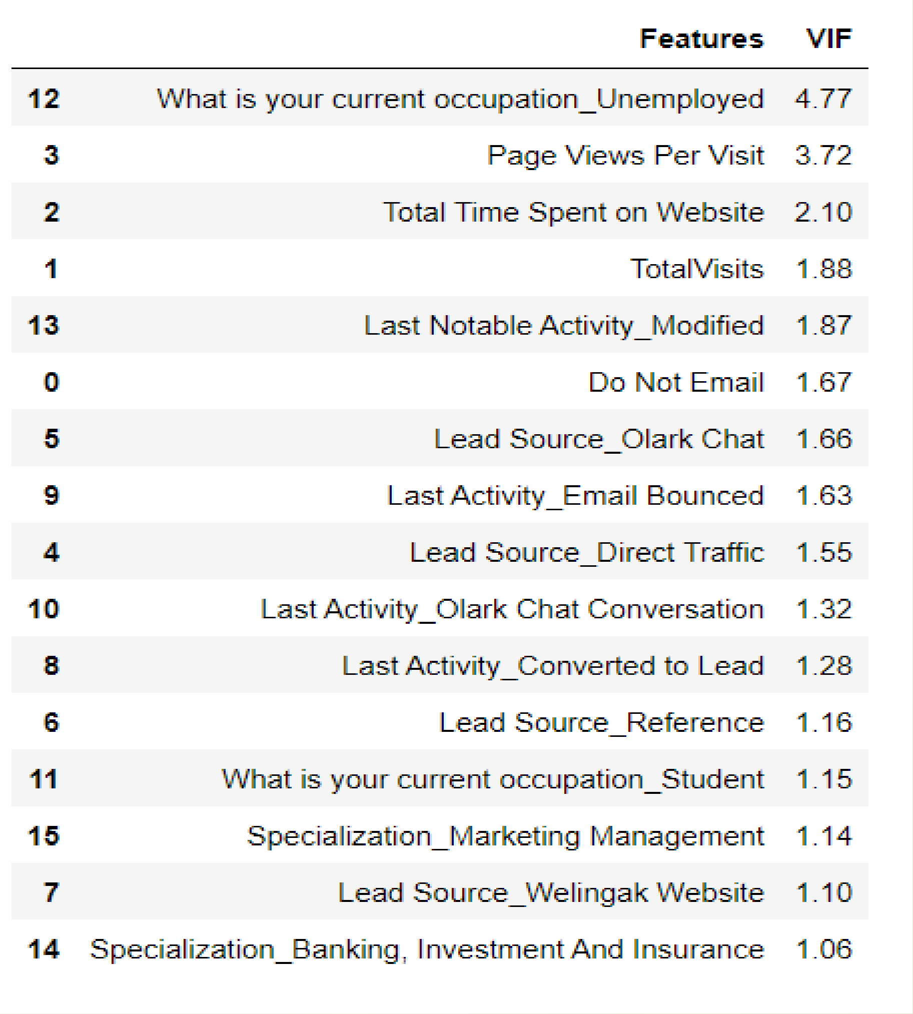
- We have built a logistic regression model with RFE feature selection and VIF to check for collinearity
- We have removed irrelevant variables and also variable with high collinearity and built the model starting with 34 variables
- After checking the p value at each stage, we removed the variable having p value more than 0.05 and finally used 16 variables for the final model
- VIF and RFE was also checked for the model to ensure only relevant features have been used for the model and there is no multicollinearity
- The final model was used for further predictions

Final model summary and VIF summary :

We can see the p values are less than 0.05 and the VIF values are less than 5.



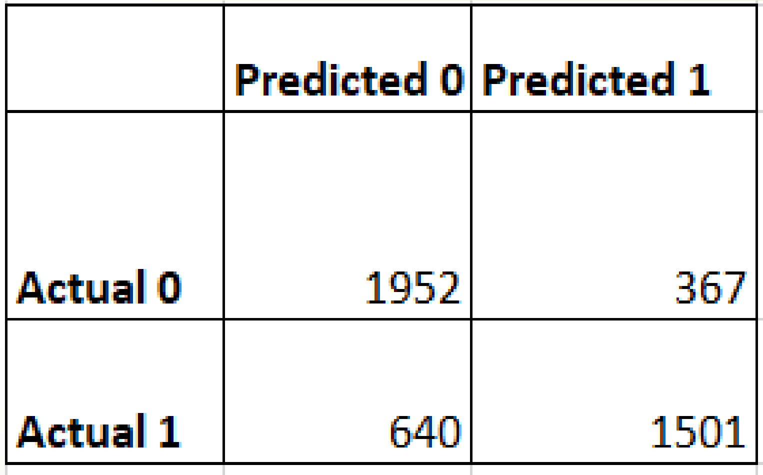
	coef	std err	z	P>|z|	[0.025	0.975]
const	1.2388	0.212	5.834	0.000	0.823	1.655
Do Not Email	-0.9229	0.215	-4.290	0.000	-1.345	-0.501
TotalVisits	9.3849	3.325	2.822	0.005	2.867	15.903
Total Time Spent on Website	4.4541	0.185	24.059	0.000	4.091	4.817
Page Views Per Visit	-1.3003	0.434	-2.997	0.003	-2.151	-0.450
Lead Source_Direct Traffic	-0.5050	0.091	-5.539	0.000	-0.684	-0.326
Lead Source_Olark Chat	1.1954	0.143	8.357	0.000	0.915	1.476
Lead Source_Reference	3.5331	0.256	13.808	0.000	3.032	4.035
Lead Source_Welingak Website	5.9719	1.014	5.889	0.000	3.984	7.959
Last Activity_Converted to Lead	-0.7361	0.236	-3.124	0.002	-1.198	-0.274
Last Activity_Email Bounced	-1.1334	0.411	-2.756	0.006	-1.939	-0.327
Last Activity_Olark Chat Conversation	-1.0467	0.191	-5.473	0.000	-1.421	-0.672
What is your current occupation_Student	-2.5884	0.285	-9.088	0.000	-3.147	-2.030
What is your current occupation_Unemployed	-2.4553	0.186	-13.211	0.000	-2.820	-2.091
Last Notable Activity_Modified	-0.8392	0.097	-8.633	0.000	-1.030	-0.649
Specialization_Banking, Investment And Insurance	0.6278	0.196	3.200	0.001	0.243	1.012
Specialization_Marketing Management	0.3001	0.126	2.386	0.017	0.054	0.547



	Features	VIF
12	What is your current occupation_Unemployed	4.77
3	Page Views Per Visit	3.72
2	Total Time Spent on Website	2.10
1	TotalVisits	1.88
13	Last Notable Activity_Modified	1.87
0	Do Not Email	1.67
5	Lead Source_Olark Chat	1.66
9	Last Activity_Email Bounced	1.63
4	Lead Source_Direct Traffic	1.55
10	Last Activity_Olark Chat Conversation	1.32
8	Last Activity_Converted to Lead	1.28
6	Lead Source_Reference	1.16
11	What is your current occupation_Student	1.15
15	Specialization_Marketing Management	1.14
7	Lead Source_Welingak Website	1.10
14	Specialization_Banking, Investment And Insurance	1.06

Prediction

- The final model was used to predict values in the train data set
- We created the dataframe with actual churn flag and predicted probabilities . Columns such as Conversion, Convertion_prob and Prospect ID were created.
- Column 'predicted' was created with 1 if Convertion_Prob > 0.5 else 0
- The confusion matrix is as per the adjacent table and the overall accuracy is 77 %

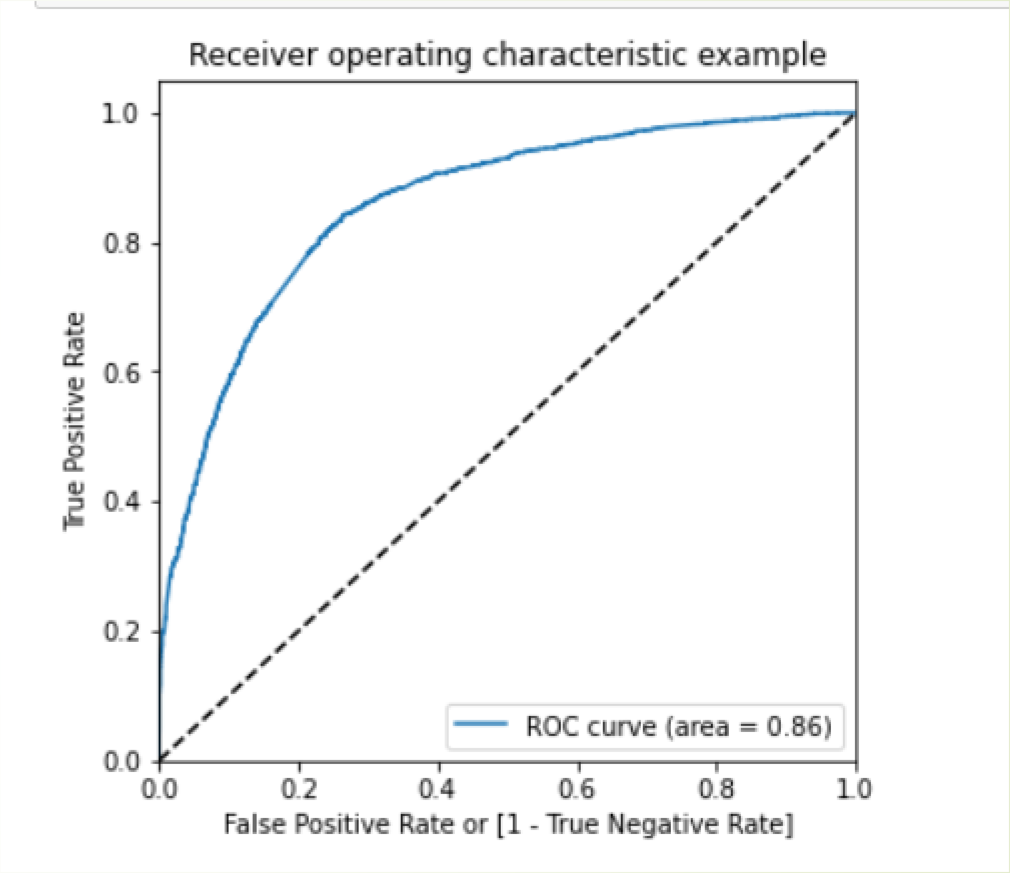


	Predicted 0	Predicted 1
Actual 0	1952	367
Actual 1	640	1501

ROC Curve

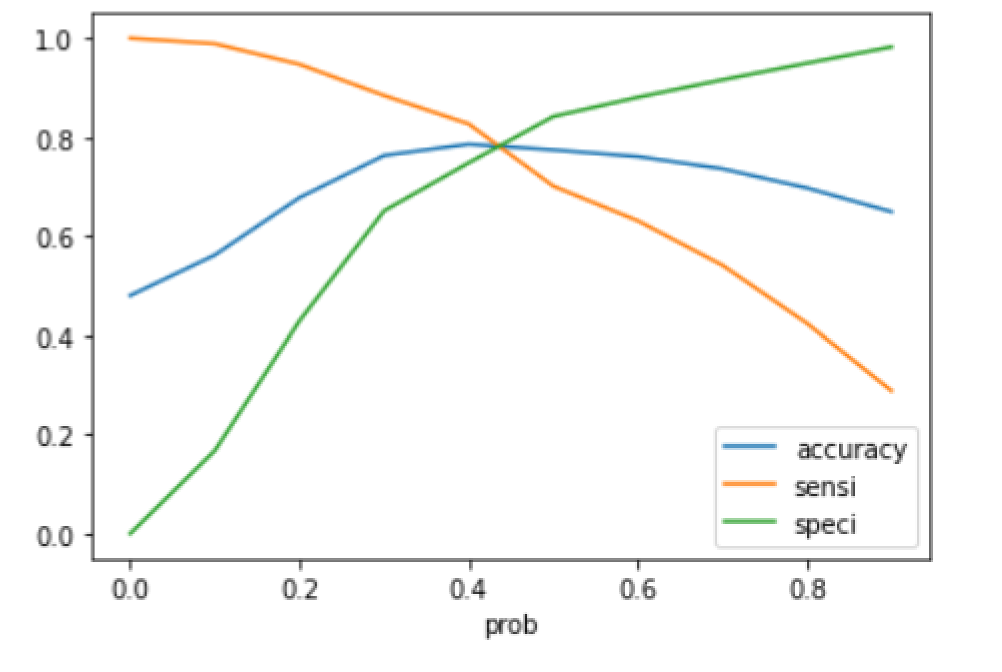
An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- As per the graph, the ROC curve covered almost 86% of the total area.



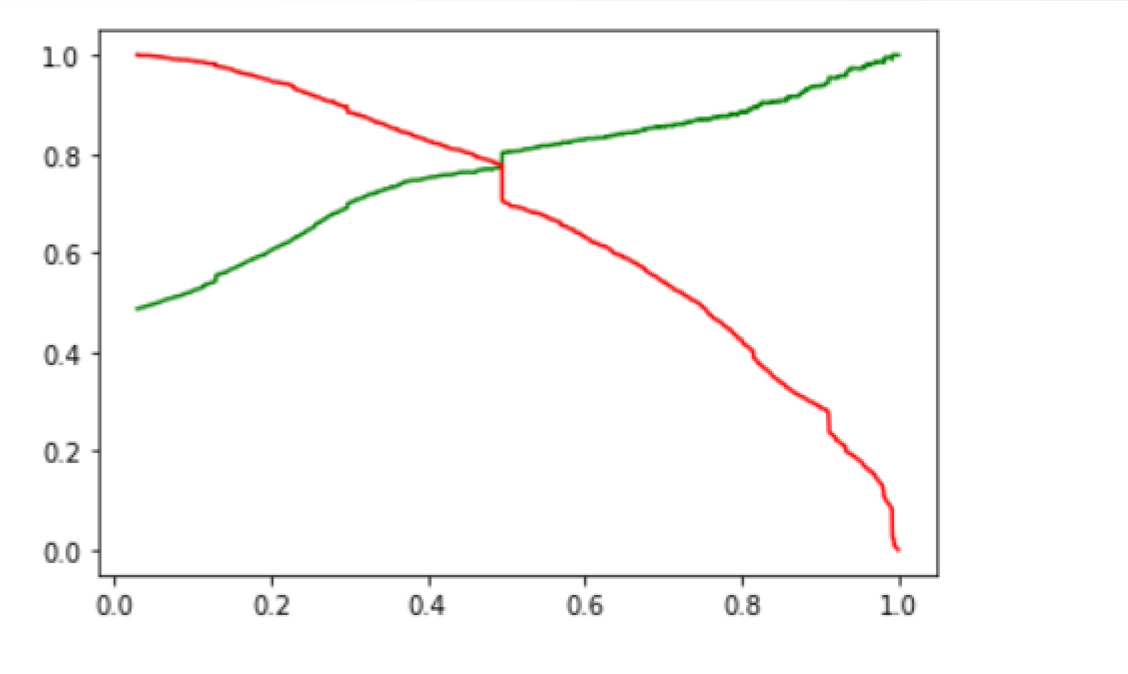
Optimal Threshold

- Optimal cutoff probability is that probability where we get balanced sensitivity and specificity
- From the adjacent curve, we can conclude that 0.4 is the optimum point to take it as a cut off probability



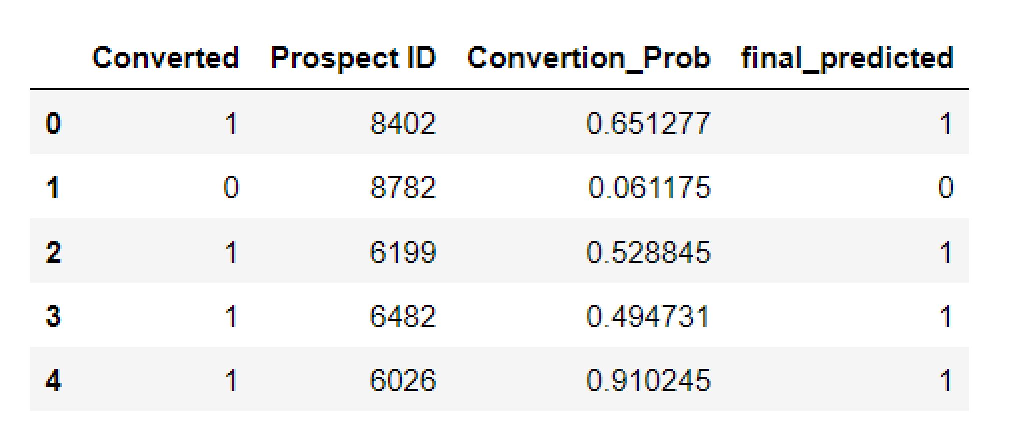
Precision and Recall

- Precision is a metric that qua the number of correct positiv predictions and recall is as sc sensitivity used for quantifying positive predictions by findin positive and false negative.
- The adjacent graph shows c precision. Both curves meet



Prediction on the test data

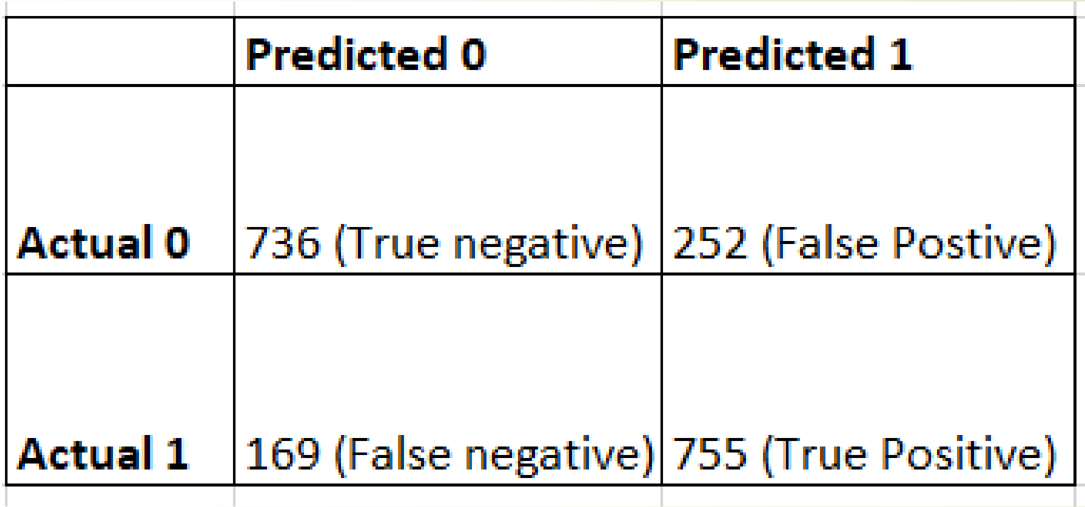
Below are the final predictions after doing the model evaluation for test data:



	Converted	Prospect ID	Convertion_Prob	final_predicted
0	1	8402	0.651277	1
1	0	8782	0.061175	0
2	1	6199	0.528845	1
3	1	6482	0.494731	1
4	1	6026	0.910245	1

Confusion Matrix

- **Model Accuracy (Correctly predicted labels / Total no. of labels):** 77%
- **Sensitivity (TP / TP + FN) :** 0.81
- **Specificity (TN / TN + FP) :** 0.74
- **false postive rate (FP / TN+FP):** 0.25
- **Positive predictive value (TP/ TP+FP):** 0.74
- **Negative predictive value (TN/ TN+FN):** 0.81



	Predicted 0	**Predicted 1**
Actual 0	736 (True negative)	252 (False Postive)
Actual 1	169 (False negative)	755 (True Positive)

Recommendations

The key features that contribute most towards the lead getting converted:

- TotalVisits
- Total Time Spent on Website
- Lead Source_Welingak Website

Potential features which can contribute in the conversion of leads and can be focused more:

- Lead Souce_Olark Chat
- Lead Source_Reference

The conversion of students and unemployed candidates is very less hence working professionals with specialization like banking, investment and insurance and marketing management can be focused more for conversion.

Since the total visits and total time spent on website is more , the company can advertise more on the website and also introduce chat support for addressing doubts of the individuals. Referrals by existing students can be encouraged to increase the lead conversion as well since it is one of the potential features.

Thank you