Case Study: Logistic Regression

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

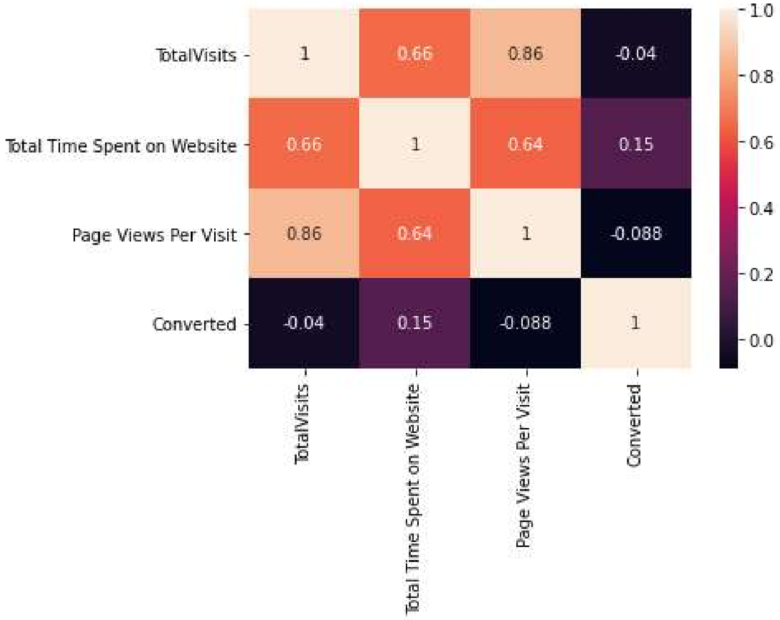
So based on various factors which are leads the company wants to predict the probability of leads converting into customers

Data Cleaning:

- We checked for null values in columns and removed the columns which had more than 3000 null values.
- We removed the columns which were not usefull for our analysis
- Then we checked columns with very low variances and dropped them
- There were some categorical variables which had attribute called "select" which means that student has not selected for any option for that column,so this was as good as missing values and we dropped the columns which have more "select" values

Transforming the data

- We analysed numerical colums by plotting paiplot and found that their result was bit skewed,so we transformed and we got better correlation and plots were more better



	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Converted
TotalVisits	1	0.66	0.86	-0.04
Total Time Spent on Website	0.66	1	0.64	0.15
Page Views Per Visit	0.86	0.64	1	-0.088
Converted	-0.04	0.15	-0.088	1

Dummy variable creation

- We created dummies for categorical variables
- Created dummies for specialization column separately because it had some attributes of value “select”.We dropped the dummies column which had “select”

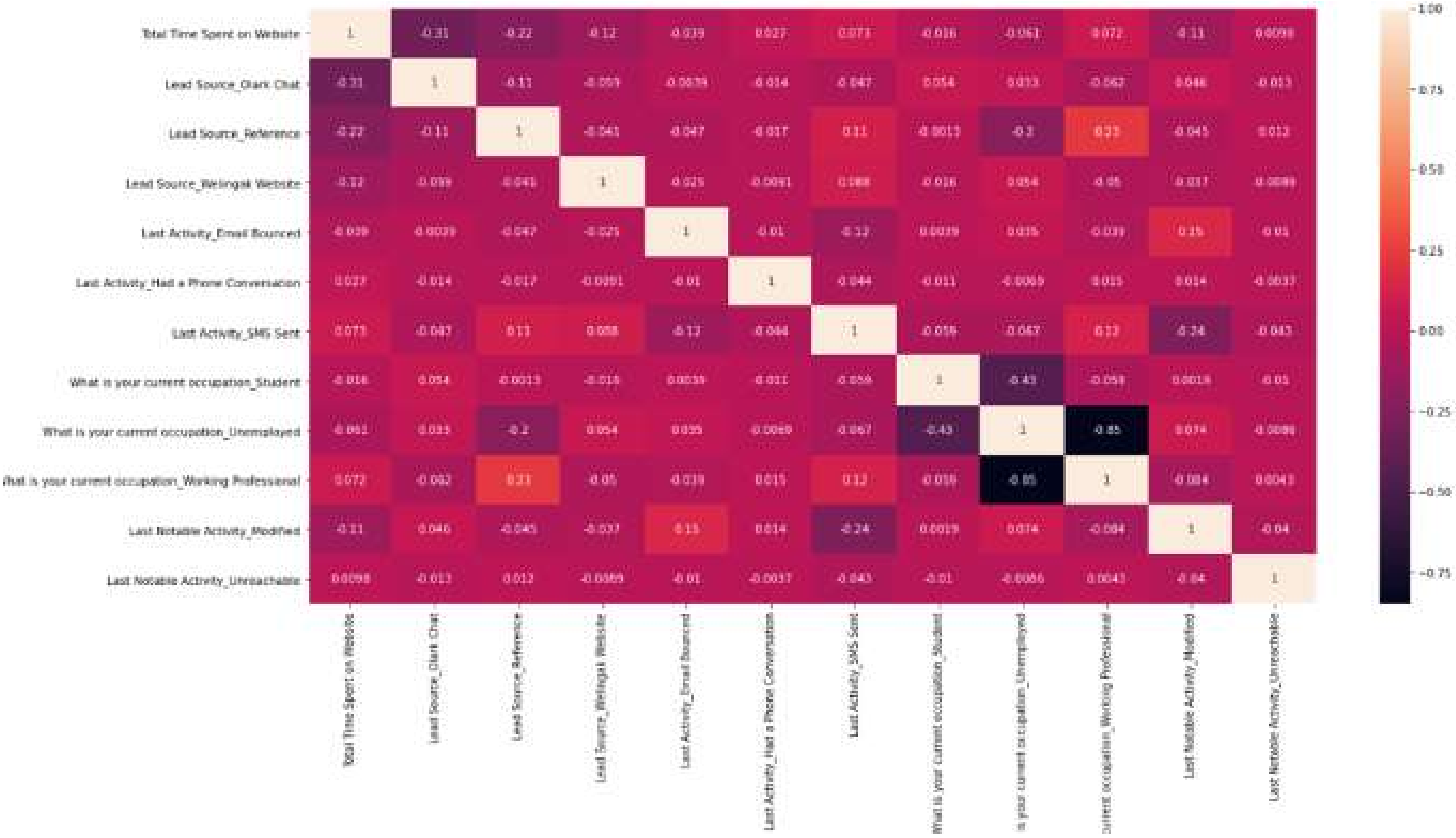
Test Train Split

- We split the data into training and testing for our model
- Put all features except target to variable x and target to y
- We split the data into 70% training and 30% test data
- We then scaled the data for better results.We used standardscaler.

Model Building¶

- We used Recursive feature elimination to select features for our model.We selected 15 models for this
- We used confusion matrix to analyse our model.Our model was found to be 78.8% accurate

- We then used variance inflation factor (VIF) to check multicollinearity and the result we got after few iteration was found to be 78.66
- We checked validated the model and collinearity using heatmap


Total Time Spent on Website
Lead Source_Olark Chat
Lead Source_Reference
Lead Source_Welingak Website
Last Activity_Email Bounced
Last Activity_Had a Phone Conversation
Last Activity_SMS Sent
What is your current occupation_Student
What is your current occupation_Unemployed
/hat is your current occupation_Working Professional
Last Notable Activity_Modified
Last Notable Activity_Unreachable
1 -0.31 -0.22 -0.12 -0.039 0.027 0.073 -0.016 -0.061 0.072 -0.11 0.0098
-0.31 1 -0.11 -0.059 -0.0039 -0.014 -0.047 0.054 0.033 -0.062 0.046 -0.013
-0.22 -0.11 1 -0.041 -0.047 -0.017 0.11 -0.0013 -0.2 0.23 -0.045 0.012
-0.12 -0.059 -0.041 1 -0.025 -0.0091 0.088 -0.016 0.054 -0.05 -0.037 -0.0089
-0.039 -0.0039 -0.047 -0.025 1 -0.01 -0.12 0.0039 0.035 -0.039 0.15 -0.01
0.027 -0.014 -0.017 -0.0091 -0.01 1 -0.044 -0.011 -0.0069 0.015 0.014 -0.0037
0.073 -0.047 0.11 0.088 -0.12 -0.044 1 -0.056 -0.067 0.12 -0.24 -0.043
-0.016 0.054 -0.0013 -0.016 0.0039 -0.011 -0.056 1 -0.43 -0.058 0.0019 -0.01
-0.061 0.033 -0.2 0.054 0.035 -0.0069 -0.067 -0.43 1 -0.85 0.074 -0.0086
0.072 -0.062 0.23 -0.05 -0.039 0.015 0.12 -0.058 -0.85 1 -0.084 0.0043
-0.11 0.046 -0.045 -0.037 0.15 0.014 -0.24 0.0019 0.074 -0.084 1 -0.04
0.0098 -0.013 0.012 -0.0089 -0.01 -0.0037 -0.043 -0.01 -0.0086 0.0043 -0.04 1
Total Time Spent on Website
Lead Source_Olark Chat
Lead Source_Reference
Lead Source_Welingak Website
Last Activity_Email Bounced
Last Activity_Had a Phone Conversation
Last Activity_SMS Sent
What is your current occupation_Student
is your current occupation_Unemployed
current occupation_Working Professional
Last Notable Activity_Modified
Last Notable Activity_Unreachable
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

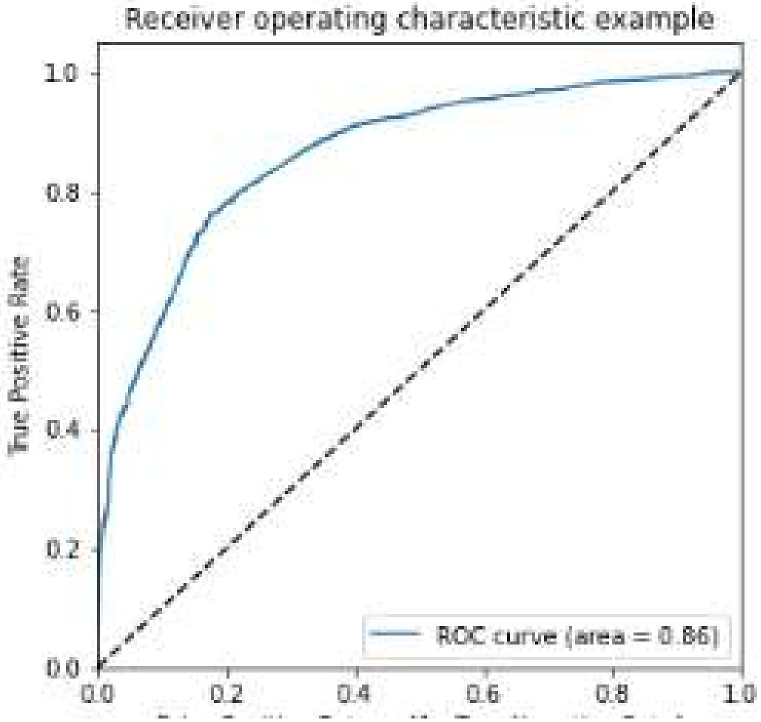
Metrics beyond simply accuracy¶

- We checked various other factors to measure our model

- Sensitivity : TP / (TP+FN) = 73.5 %
- Specificity : TN /(TN+FP) = 83.4 %
- False positive rate : FP/ (TN+FP)=16 %
- Positive predictive value: TP / (TP+FP) = 80%

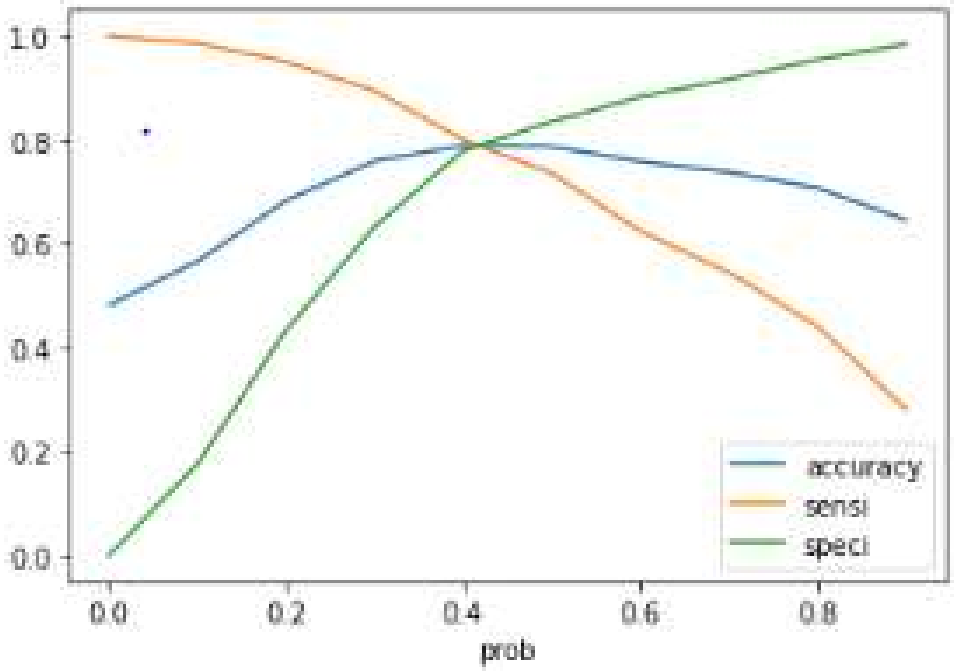
ROC Curve

- 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.



Finding Optimal Cutoff Point

- Optimum cutoff was fount to be around .42
- overall accuracy:78.9%

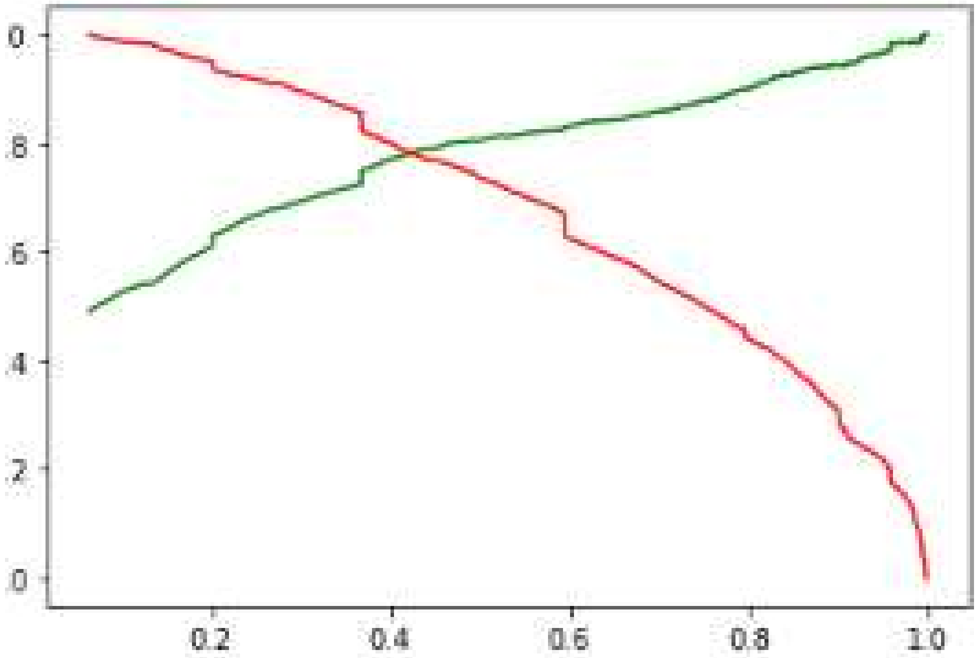


Precision and Recall

- Precision = 80.4
- Recall = 73.5

Precision and Recall Tradeoff

- Cutoff between precision is around .42 which is same

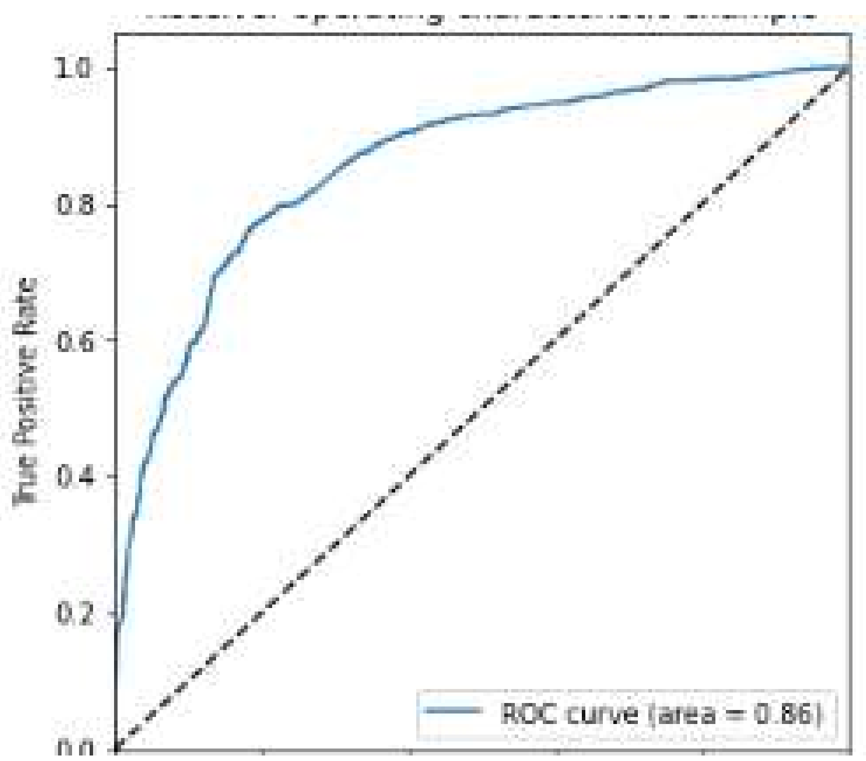


Making predictions on the test set

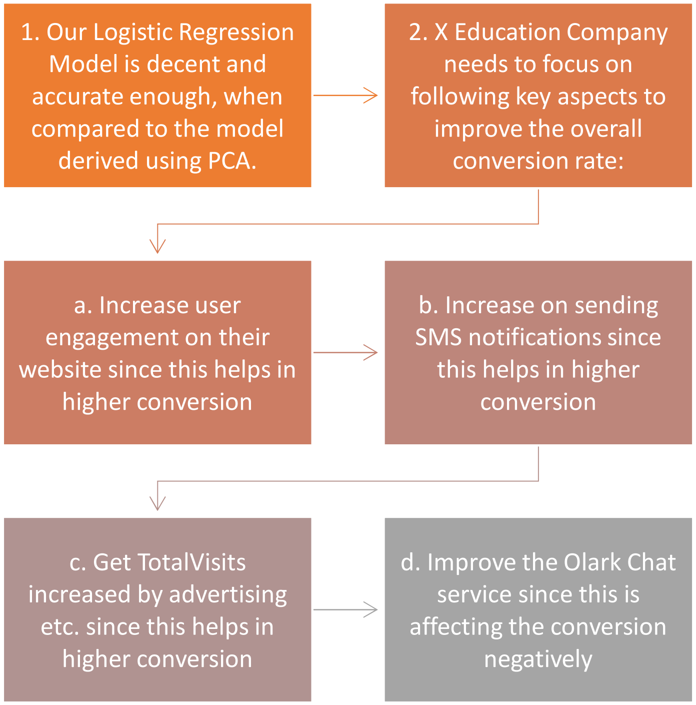
- Precision : 77.48
- Recall : 78.16
- Sesitivity: 78.16
- Specificity : 79.11
- False positive rate: 77.48

ROC curve on test set

- Its similar to train set ROC curve



Conclusion



1. Our Logistic Regression Model is decent and accurate enough, when compared to the model derived using PCA.

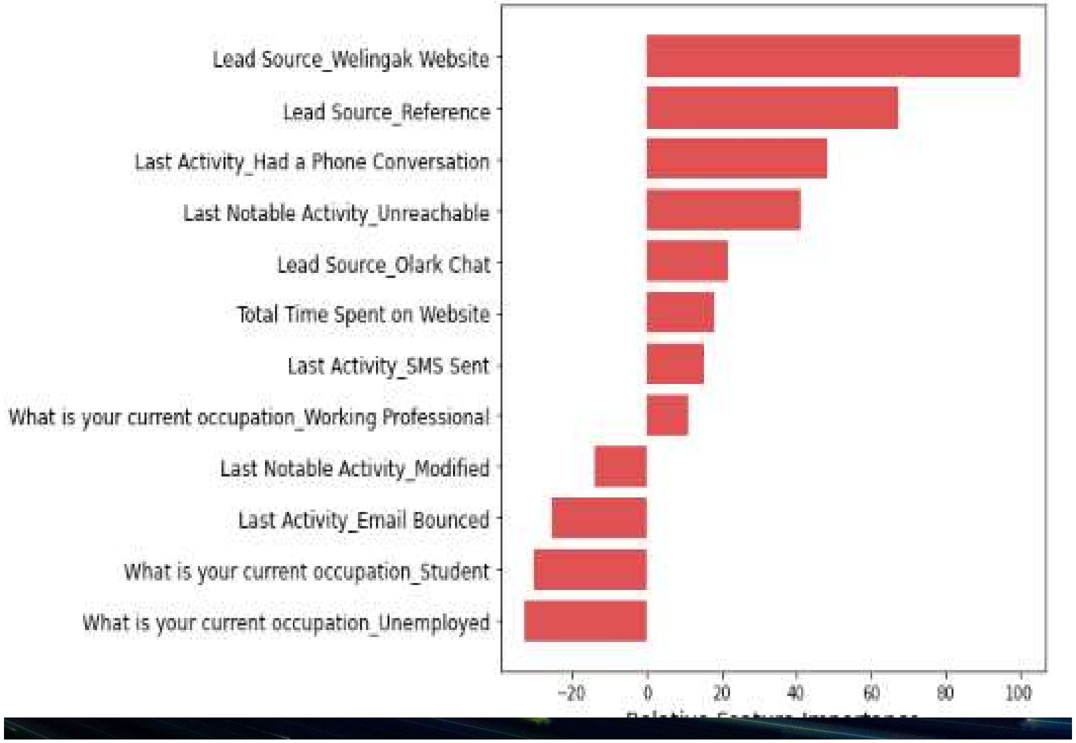
2. X Education Company needs to focus on following key aspects to improve the overall conversion rate:

a. Increase user engagement on their website since this helps in higher conversion

b. Increase on sending SMS notifications since this helps in higher conversion

c. Get TotalVisits increased by advertising etc. since this helps in higher conversion

d. Improve the Olark Chat service since this is affecting the conversion negatively


Lead Source_Welingak Website
Lead Source_Reference
Last Activity_Had a Phone Conversation
Last Notable Activity_Unreachable
Lead Source_Olark Chat
Total Time Spent on Website
Last Activity_SMS Sent
What is your current occupation_Working Professional
Last Notable Activity_Modified
Last Activity_Email Bounced
What is your current occupation_Student
What is your current occupation_Unemployed
Relative Feature Importance