LEAD SCORING CASE STUDY

GROUP :

- Ayyub Mohammad
- Srabani Dutta
- Gaurav Kumar Singh

PROBLEM STATEMENT

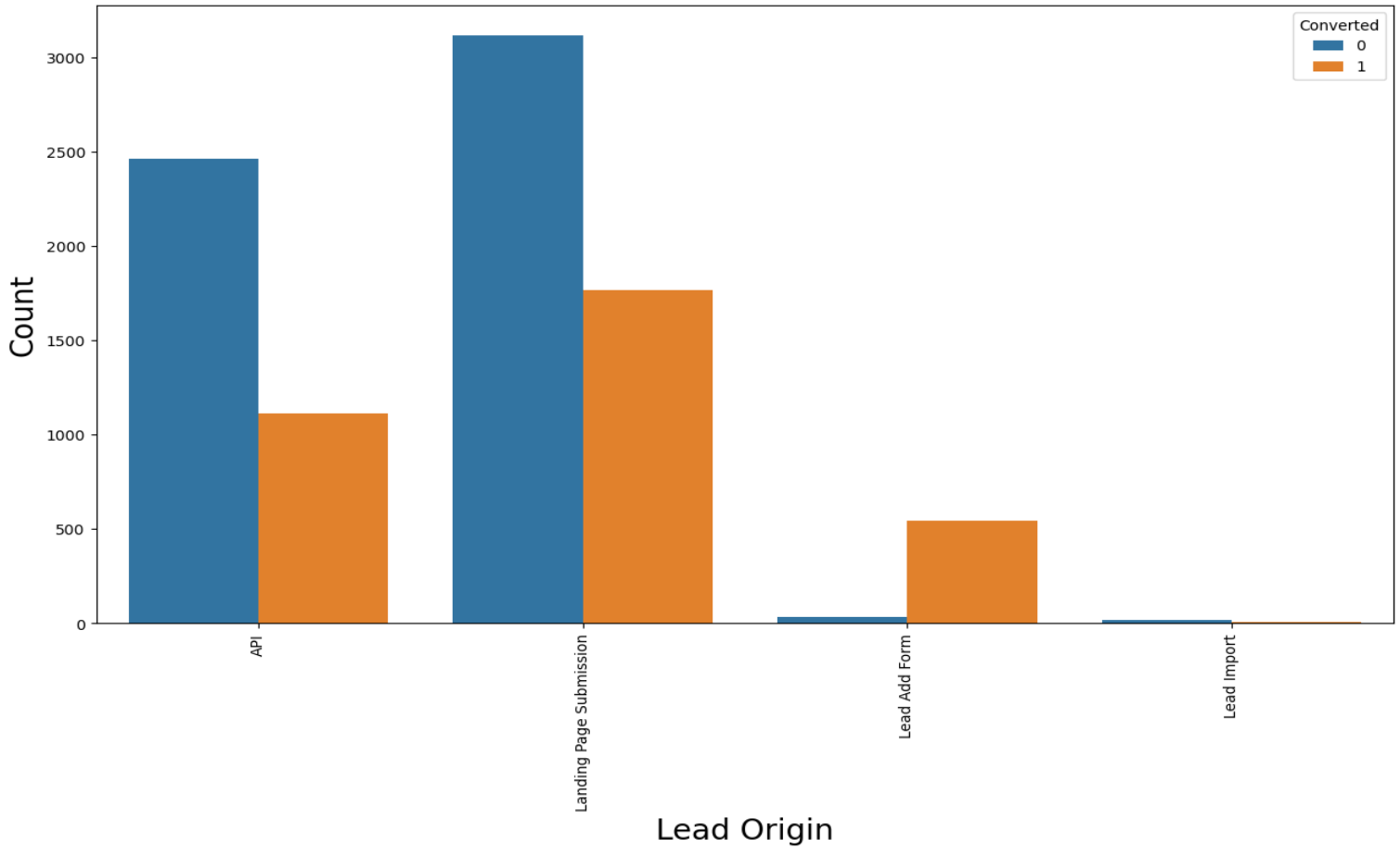
- An education company named X Education sells online courses to industry professionals. X Education company requires to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company wants to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Steps:

- Data cleaning by handling missing values
- Data preparation for model building
- Splitting of data set into train and test data set in 70:30 ratio
- Creation of dummy variables for categorical variables
- Feature scaling
- Logistic regression model building
- Calculation of VIF and P values
- Finding of optimal probability cut-off
- Checking the model performance over test data
- Creation of score variable to assign score to leads from 0 to 100.

Heatmap of numerical variables to understand correlation

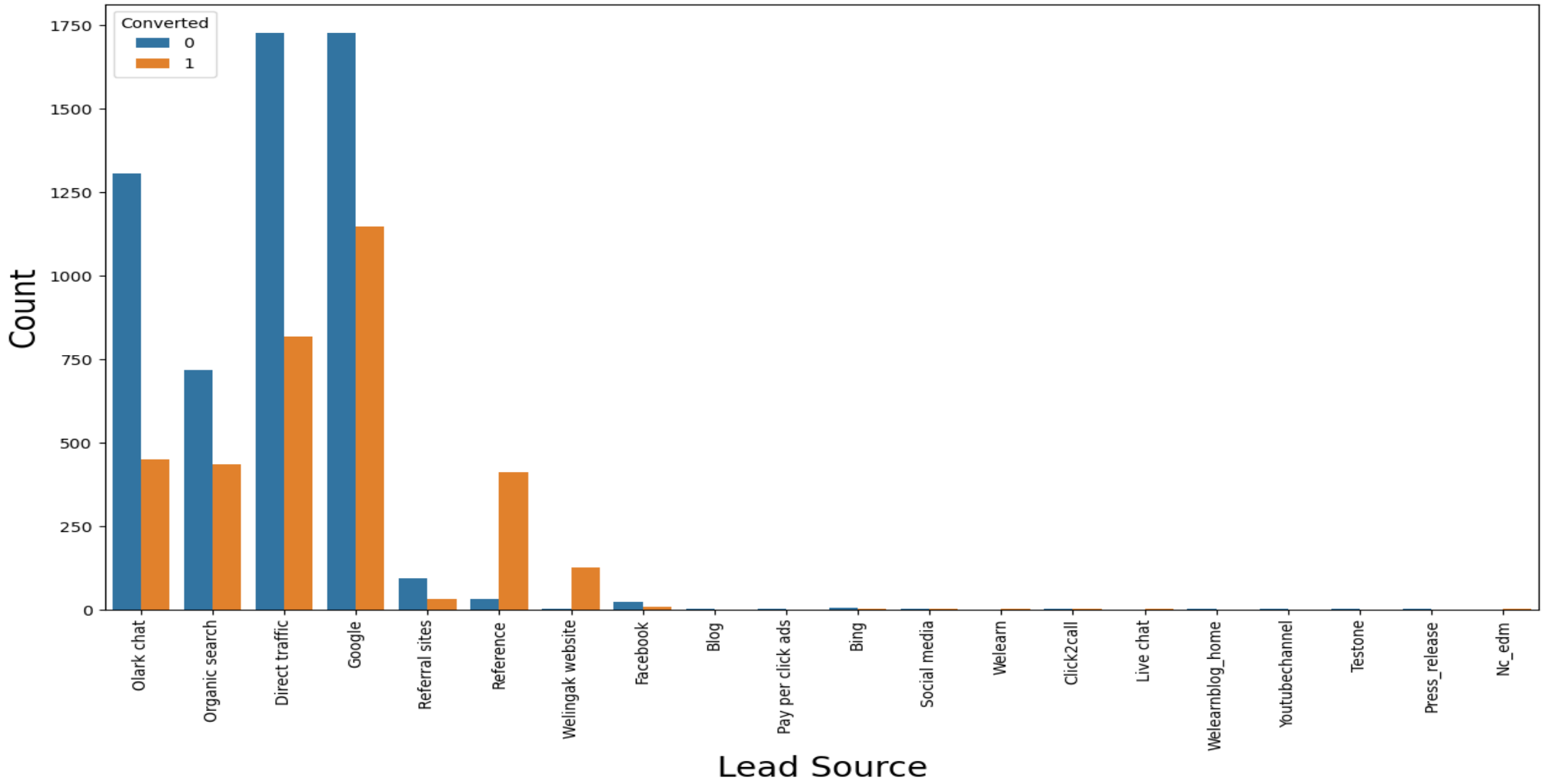
Visualization of the categorical variable ‘Lead Origin’



Insights:

- Landing page submission category has highest numbers of converted leads.
- Lead Add Form category has more converted numbers than not converted.

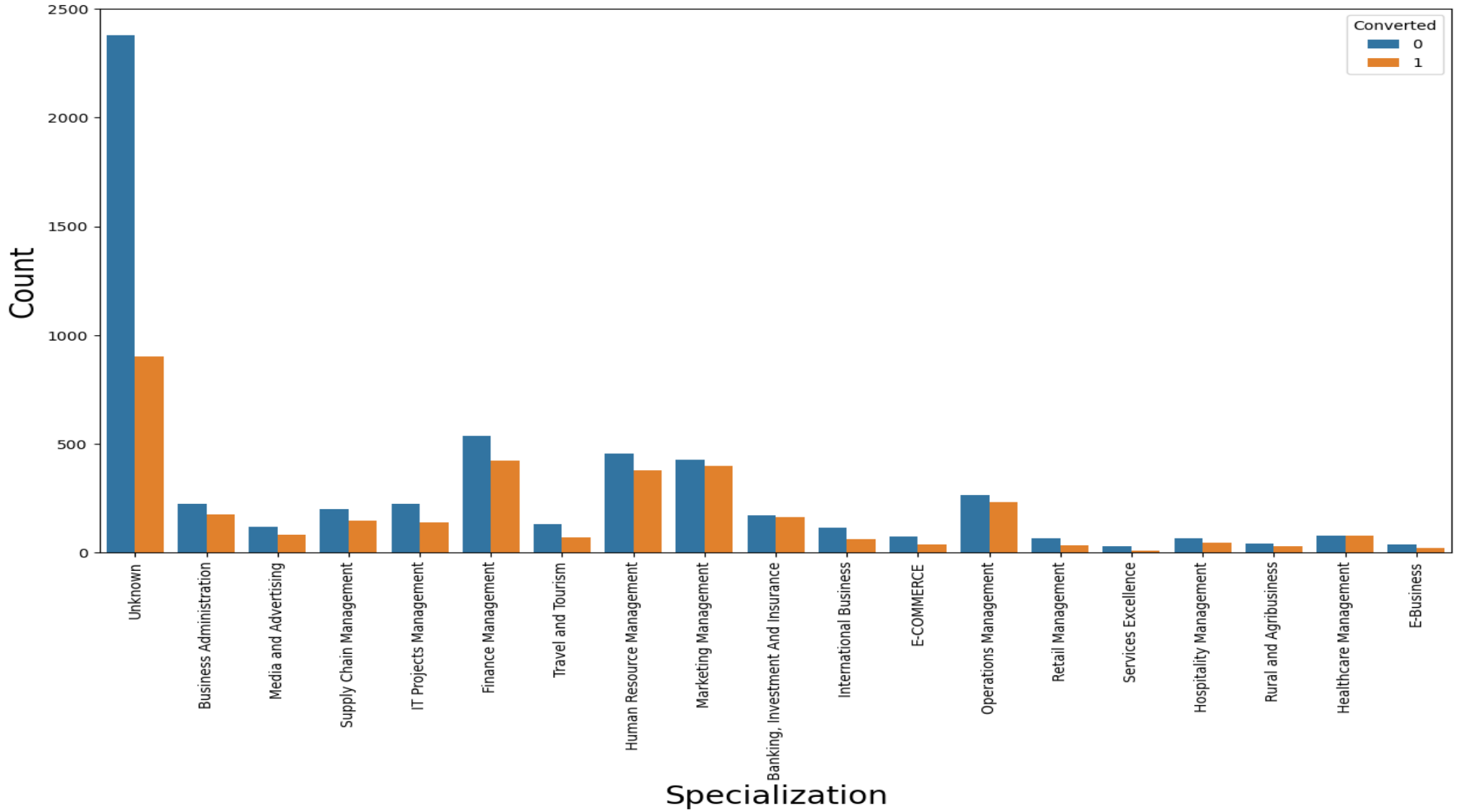
Visualization of the categorical variable ‘Lead Source’



Insight:

- Google as a Lead Source Category has highest numbers of converted people among all the rest of the categories.

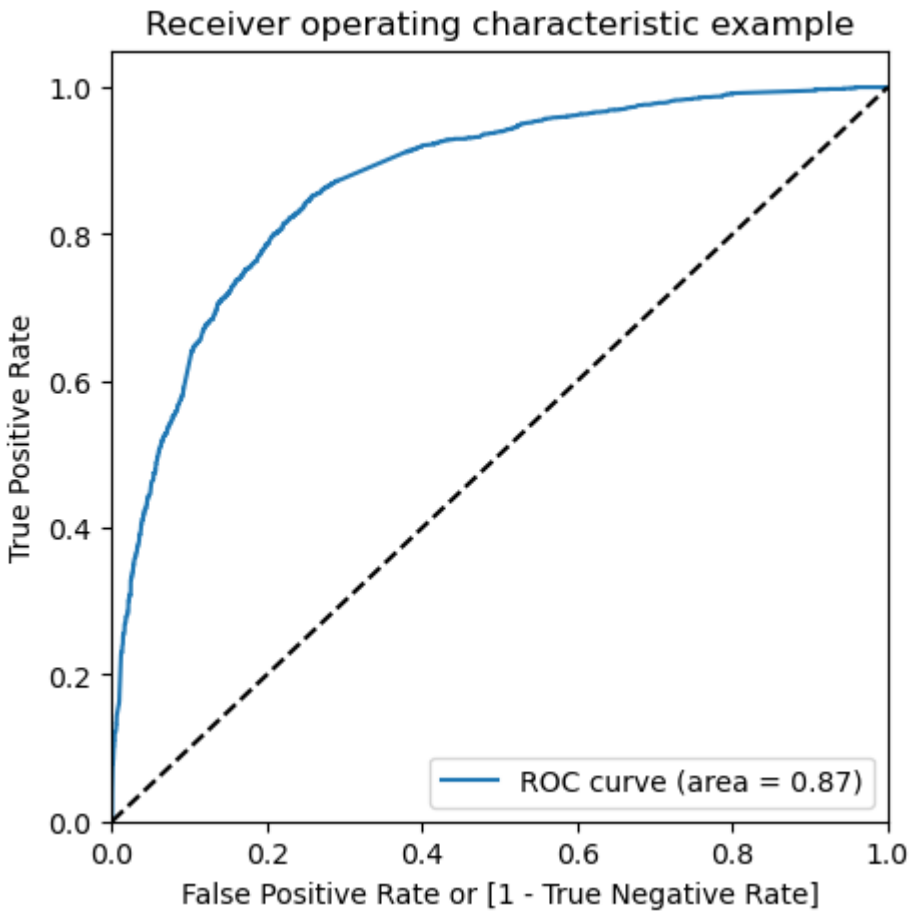
Visualization of the categorical variable ‘Specialization’



Insight:

- Marketing, HR and Finance Management people have the highest conversion rate, hence should be focused.

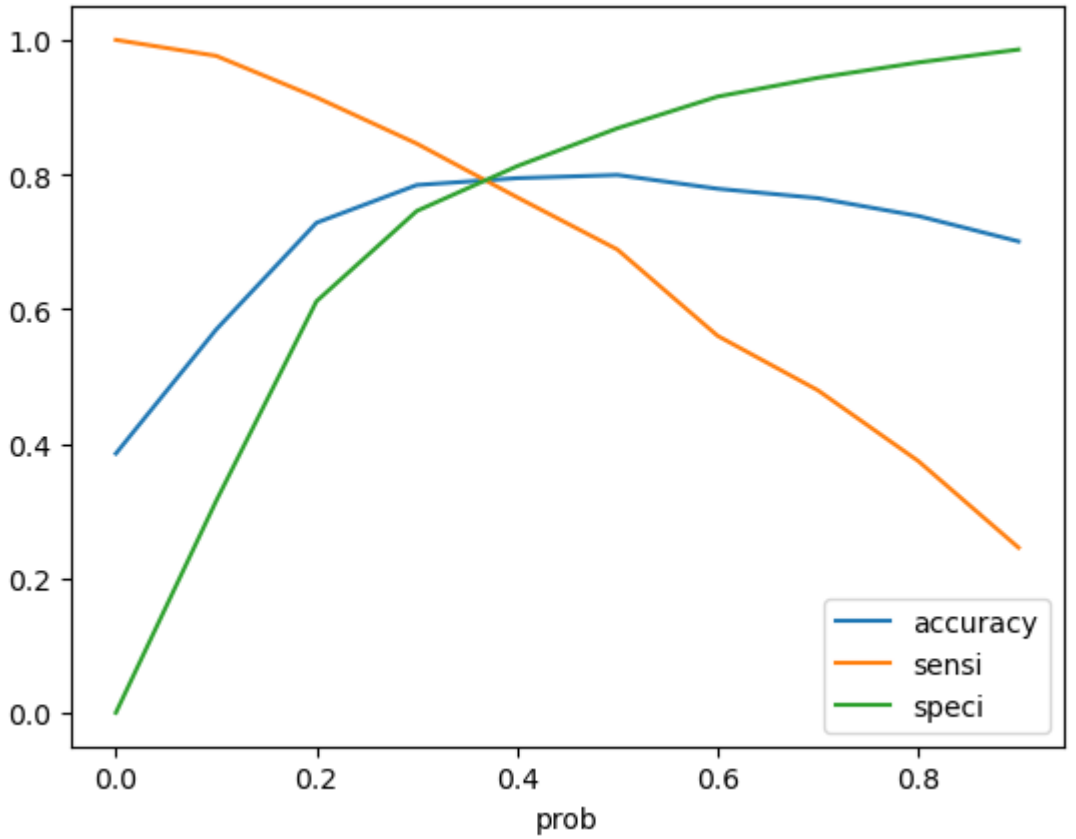
ROC CURVE TO SHOW THE TRADE OFF BETWEEN SENSITIVITY AND SPECIFICITY



Insights:

- ROC curve is towards left & has area covered up to 0.87.
- Hence it is a good predictive model.

Finding the optimal cut-off



Insight:

- From above graph and the table it can be seen that optimal cutoff is around 0.35

Confusion Matrix on test data Set Vs Train data set

Actual/Predicted	Not Converted	Converted
Not Converted	1366	368
Converted	200	789

- Accuracy = 79.14%
- Sensitivity= 79.77%
- Specificity= 78.77%
- Precision = 68.19%
- Recall = 79.77%

Actual/Predicted	Not converted	Converted
Not converted	3046	859
Converted	462	1984

- Accuracy = 79.20%
- Sensitivity= 81.11%
- Specificity= 78.00%
- Precision = 69.78%
- Recall = 81.11%

- According to our business goals, the recall rate is more useful because, even though our precision may be a little low and result in fewer hot lead customers, we don't want to miss any hot leads who are ready to sell.

- Hence our focus on this will be more on Recall than Precision.

Final Observations:

Important features (in descending order) responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

- Last Notable Activity_Had a Phone Conversation
- Lead Source_Welingak website
- Lead Origin_Lead Add Form
- Last Notable Activity_Unreachable
- Last Notable Activity_SMS Sent
- Lead Source_Olark chat
- Total Time Spent on Website
- Lead Origin_API
- Last Activity_Olark Chat Conversation
- Specialization_Unknown
- Last Activity_Email Bounced
- Lead Origin_Landing Page Submission

Valuable Insights:

- The Accuracy, Precision and Recall score we got from test set in aceptable range.
- We have high recall score than precision score which we were exactly looking for.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.

THANK YOU