# Lead Scoring Case Study

# Problem Statement, Business Problem and Methodology

# 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. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

# Business Goal

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. The company requires to build a model wherein they 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

# Methodology

1. Data Cleaning and Manipulation: The given leads data is needed to be uploaded and then be cleaned by carefully looking for duplicate data (if any) and deleting it, for null values and filling them with appropriate values, deleting columns having null values greater than 35% and handle the outliers.
2. EDA: Create appropriate univariate, bivariate and multivariate analysis graphs to understand the data well and find insights like correlation etc.
3. Model Building: First create dummy variables for all the categorical variables, then split the data into train and test set, scale the data appropriately, and finally build a models it you satisfy the condition of $p<=0.05$ and $VIF<5$.
4. Model Validation: Find the optimum cut off point and after that find the the accuracy, precision sensitivity and specificity for the final model on test set.

# Data Cleaning & Manipulation and EDA (Exploratory Data Analysis)

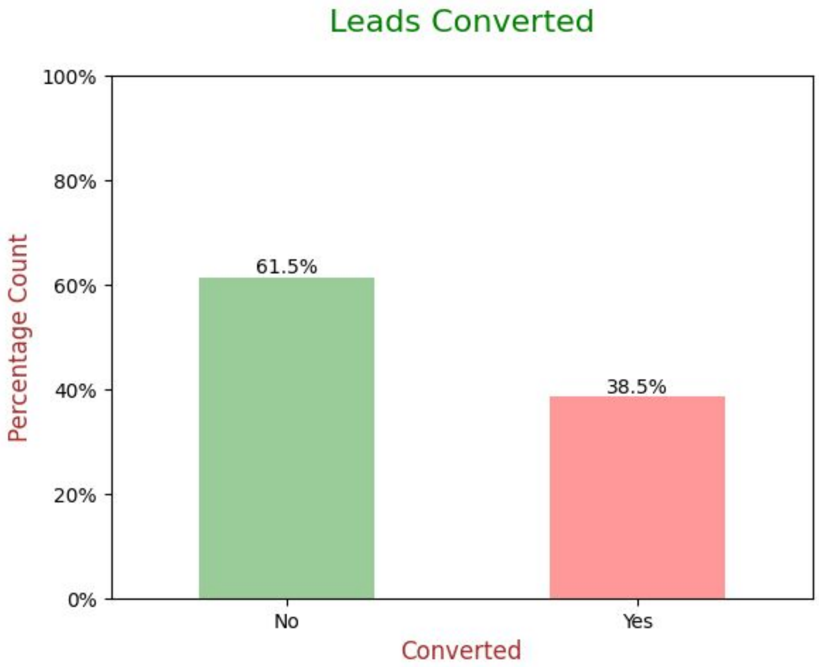
# Data Cleaning & Manipulation

- There were total of 9240 rows and 37 columns in total.
- The null values were checked and columns having null values greater than 35% were dropped (Except specialization because those can be filled with ‘Not Provided’)
- Columns having null values less that 35% were filled appropriately using mode or any other suitable method.
- Columns such as Prospect ID, Lead number etc who did not had any effect on analysis were dropped.
- Remaining columns were checked for their skewness, highly skewed ones were dropped. Outliers were checked for the numerical columns.

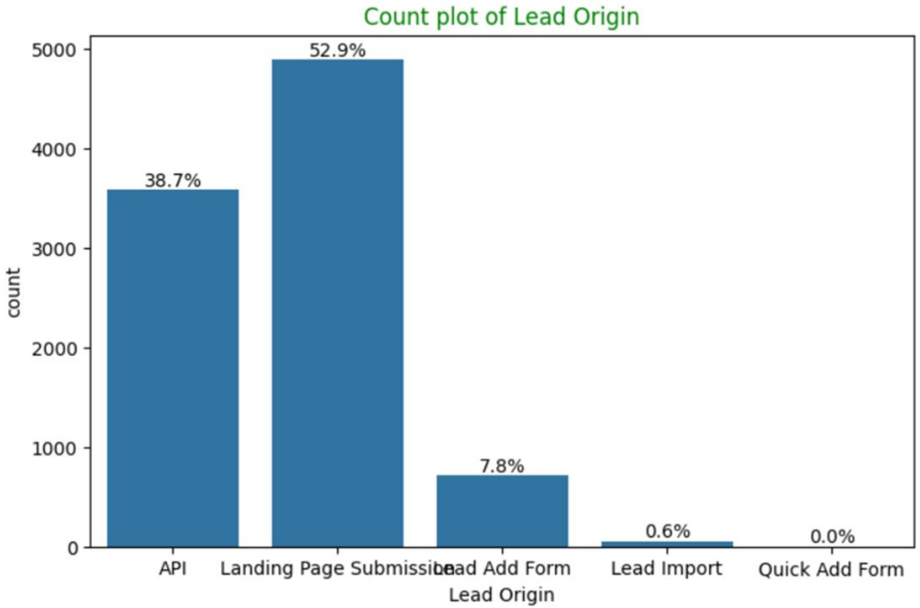
# EDA (Exploratory Data Analysis)

Insights:

Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority). While 61.5% of the people didn't convert to leads. (Majority)

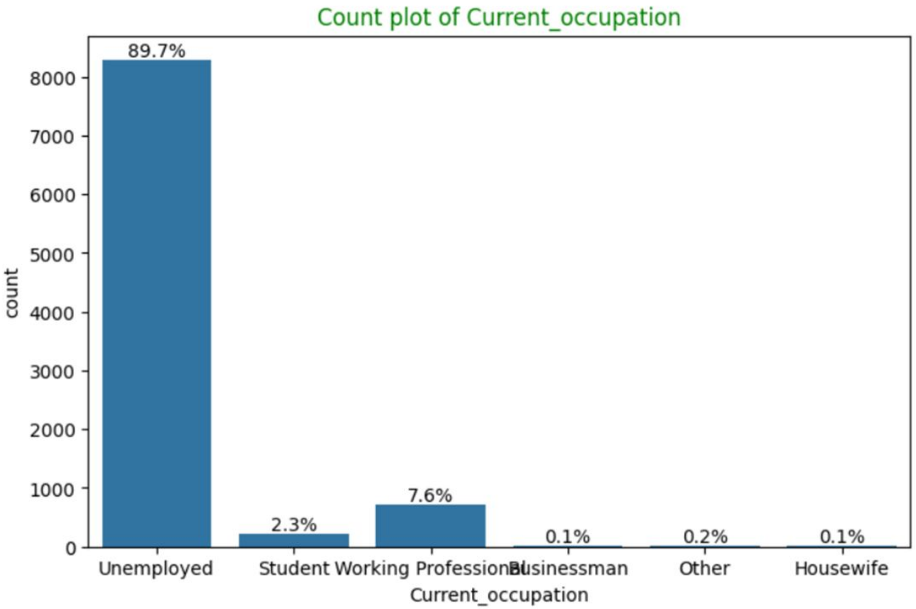


# Univariate



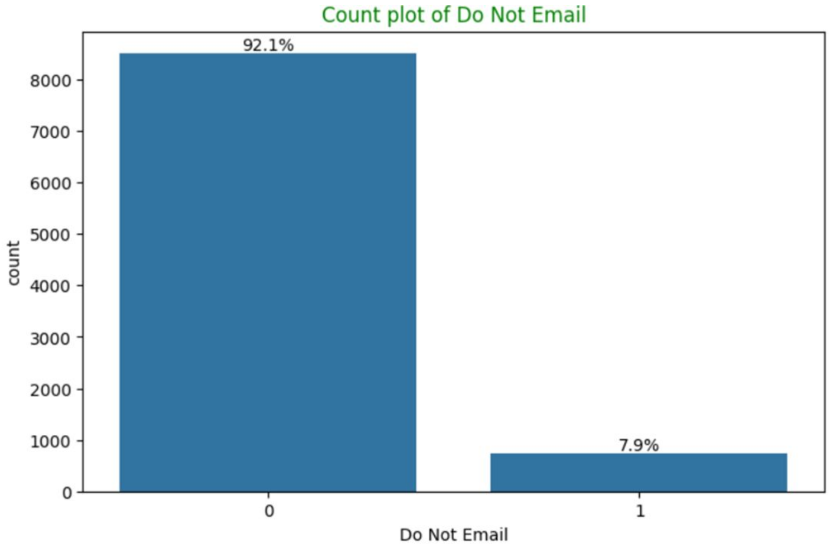
Insights :

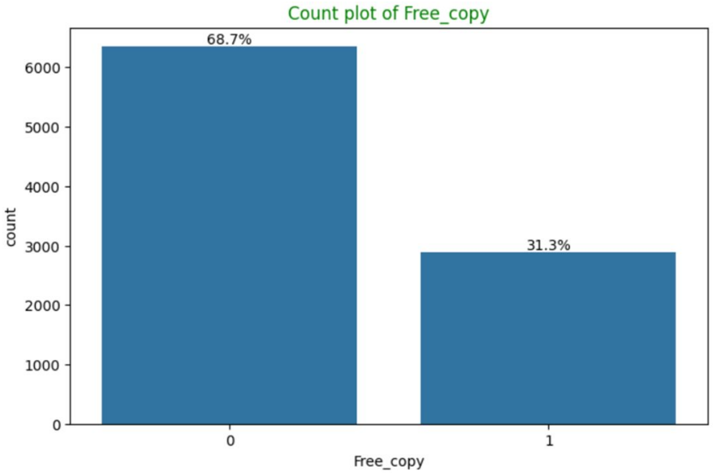
The maximum number of lead origin is from landing page and least from quick add form



Insight:

The maximum number of leads are unemployed and the minimum number of leads are housewife.

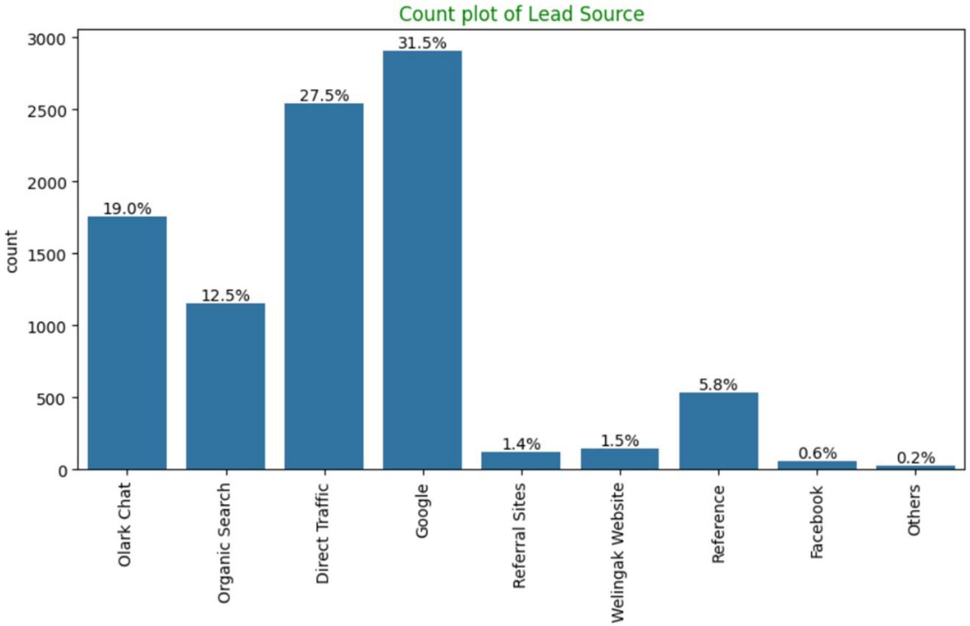




Insight:  Maximum number of leads opted for Do not email and for free copy.

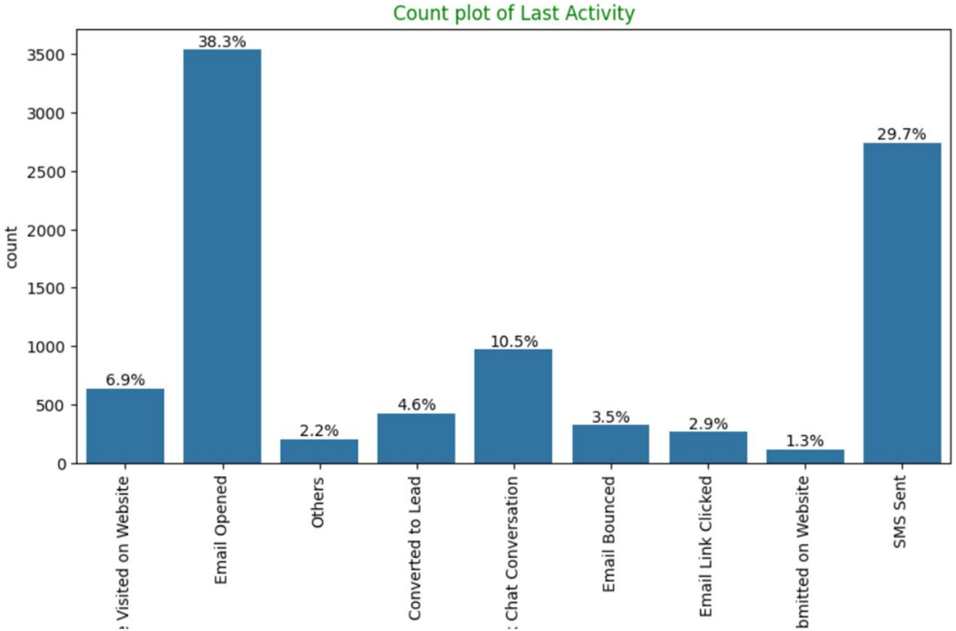
Insights:

The source for maximum leads are from google and the least are from others.

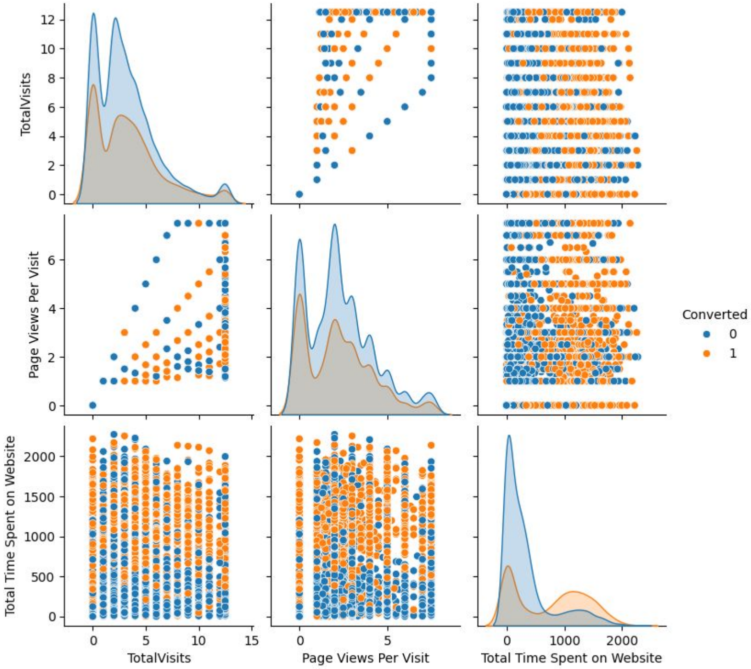


Insights:

The maximum leads had email opened as last activity and minimum leads had form submitted on website as their last activity.

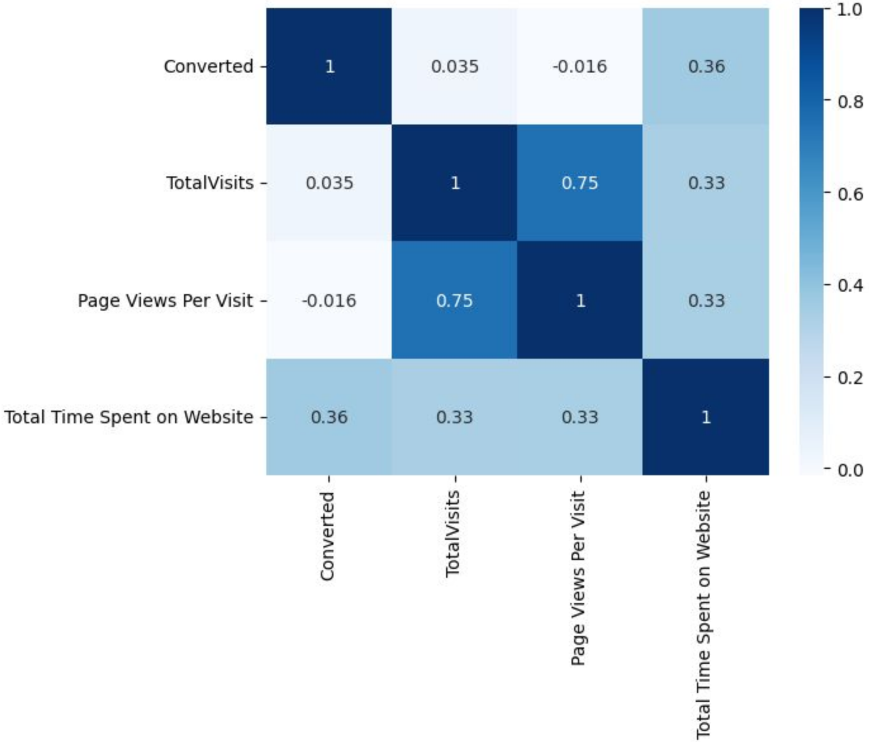


# Bivariate



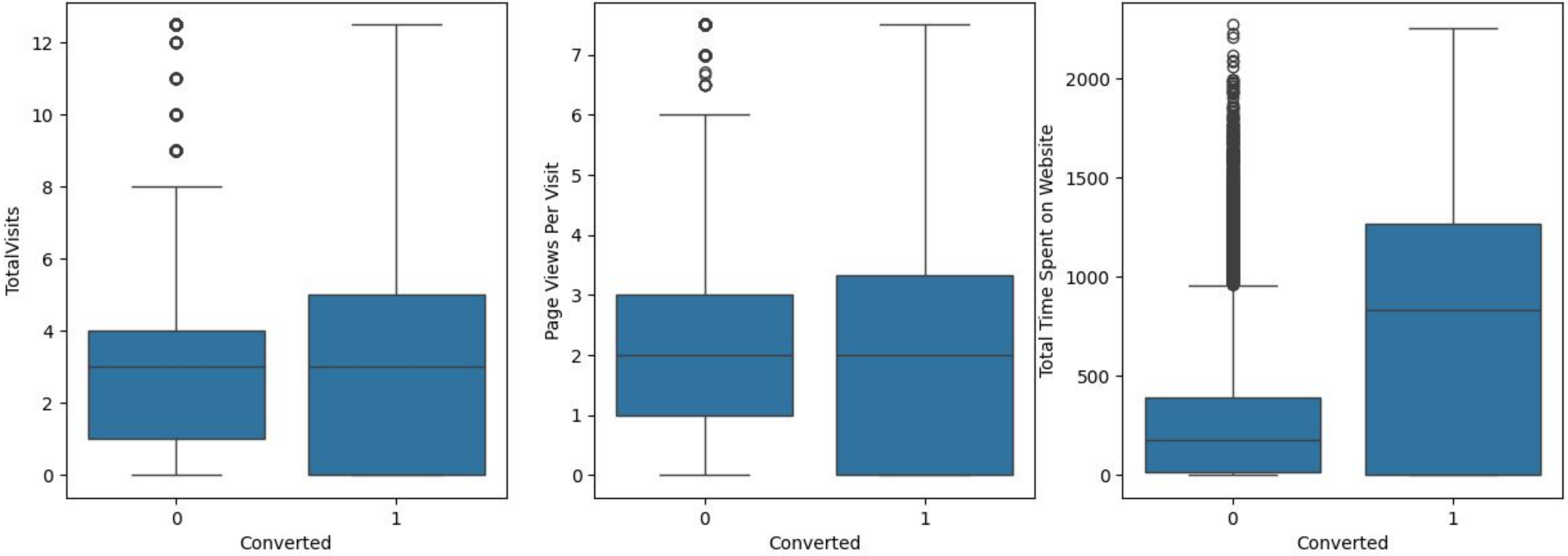
Insight:

Page views per visit and Total visits have somewhat a linear combination.



Insight:

Page views per visit and Total visits have somewhat a linear combination.



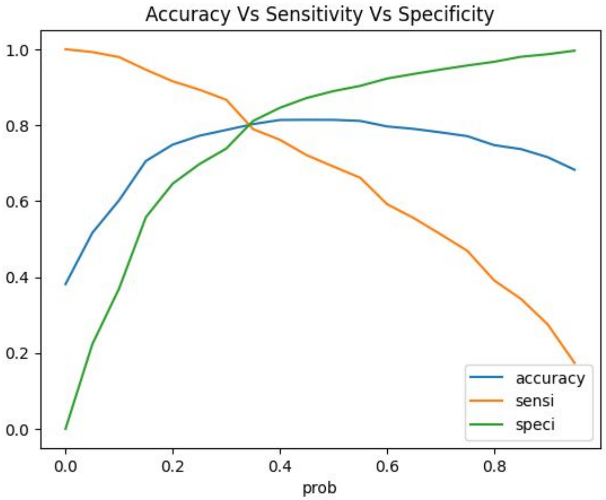
Insights: Past Leads who spends more time on Website are successfully converted than those who spends less as seen in the boxplothere are quite a few outliers but we will leave them for now.

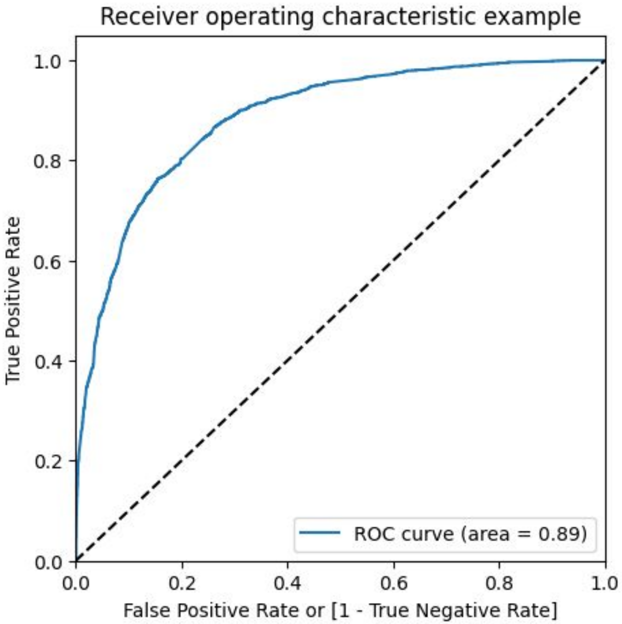
# Model Building and ROC Curve

# Model Building

- Created dummy variables for categorical data.
- The cleaned data is split into train set (70%) and test set (30%).
- We scale the numerical data of train data set to start making model.
- We use RFE for feature selection.
- We build the model using condition that p value should be less than 0.05 and VIF should be less than 3. We eliminate columns one by one till we satisfy both the condition
- After finding the final model, we find the optimum cutoff point, which in this case is 0.345.
- We test the model on test data and find accuracy, specificity, and precision.

# ROC Curve





The optimum cut- off is 0.369

# Conclusion and Recommendations

# Conclusion

- The variables that can be considered are : Do Not Email, Total Time Spent on Website, Lead Source_Olark Chat, Lead Source_Others, Lead Source_Reference, Lead Source_Welingak Website, Last Activity_Email Opened, Last Activity_Olark Chat Conversation, Last Activity_Others, Last Activity_SMS Sent, Specialization_Hospitality Management, Specialization_Others, Current_occupation_Other, Current_occupation_Student, Current_occupation_Unemployed
- Leads with score of more than 28 have more than 87% chances of conversion and can be targeted by the X Education. Anyone having a score of >= 28 can be considered a Hot Lead.
- Since the accuracy comes out to be 80.05% which surpasses the threshold by the CEO.
- Evaluation of our final model at our final cutoff of 0.345 on test set: Accuracy: 80.05%, Sensitivity/Recall: 79.73% = 80%, Specificity: 81%, Precision: 73.3%

# Recommendations

- If they wish to make the lead conversion more aggressive. The following strategies can be adopted: They can lower the probability cut-off so that more leads can be classified as hot leads. They should contact the leads whos source is Welingak website or reference, has last activity as SMS Sent or as they have a high chances of conversion. Apart from that they can also look for people who spend a lot of time on their website as they can be potential hot leads.
- If they wish to make the lead conversion less aggressive. The following strategies can be adopted: The probability cutoff can be raised so that only those leads are contacted those have very high chance of conversion. A personalised SMS or email should be sent to the leads and only the ones responding positively should be contacted. The only priority should be given to customers who spent a considerable amount of time on website because they have a very high chances of getting converted.

# Thank You!