LEAD SCORING CASE STUDY

BY: POOJITHA PARUPALLY, NIRAJKUMAR PITHVA, PRAJWALA

DS C52 BATCH

CONTENT

- Problem Statement
- Available Data & Steps Followed
- Data understanding and Missing value treatment
- Outlier Analysis and treatment
- Univariate Analysis
- Bivariate Analysis-1
- Bivariate Analysis-2
- Model Building
- Model Evaluation – Metrics & ROC
- Summary
- Recommendataion

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- The company generates a large number of leads from various sources such as website visitors, form fill-ups, and referrals, but its lead conversion rate is only 30%.
- To improve this conversion rate, X Education wants to identify the most potential leads, also known as "Hot Leads." The company requires a lead scoring model that assigns a lead score to each lead based on the likelihood of conversion.
- The CEO has set a target lead conversion rate of 80%. The lead scoring model should help the sales team to prioritize potential leads that have a higher conversion chance and enable them to focus on communicating with them. By nurturing these potential leads, X Education can increase their chances of converting them into paying customers and achieve their target conversion rate.

AVAILABLE DATA & STEPS FOLLOWED

Available Data:

- A leads dataset from the past with around 9000 data points has been provided. This dataset consists of various attributes which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Steps followed:

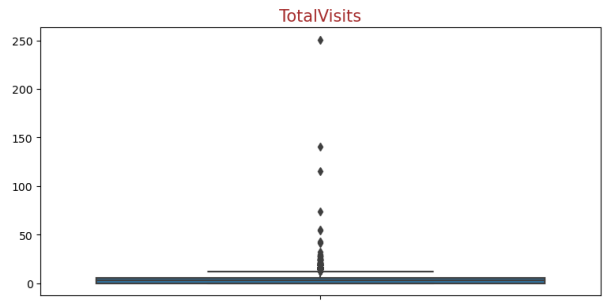
- Data Understanding, Missing value treatment
- Outlier analysis & Treatment
- Univariate and Bivariate analysis
- Model building using logistic regression
- Evaluation of the model using various metrics
- Conclusion

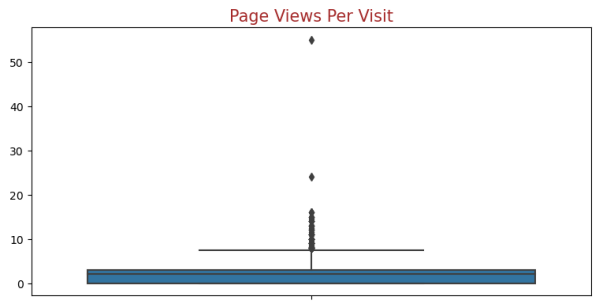
DATA UNDERSTANDING, MISSING VALUE TREATMENT

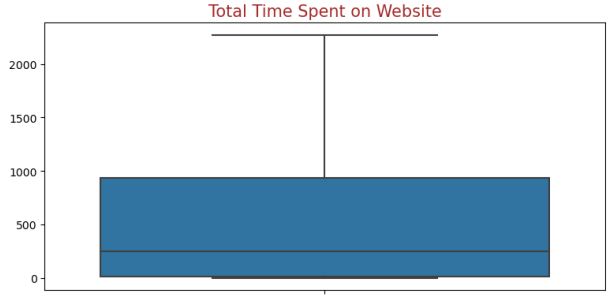
- The columns >40% missing values are straight away dropped
- City: 39.71% missing values. Imputing missing values with Mumbai(57.84%) will later cause bias in the model. Hence City column can be dropped.
- Specialization: 36.58% missing values. The specialization selected is evenly distributed. Dropping is not a good choice. Creating another category as 'Others' to replace.
- Tags, What matters most to you in choosing a course, Country, I agree to pay the amount through cheque, Get updates on DM Content, Update me on Supply Chain Content, Receive More Updates About Our Courses, Magazine Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations - This columns are dropped as it has relatively high missing values and/or has mostly one unique values. Hence, this does not give any insights.
- What is your current occupation, Last Activity, Lead Source, TotalVisits, Page Views Per Visit – These has very few missing values and are imputed with mode

OUTLIER ANALYSIS & TREATMENT

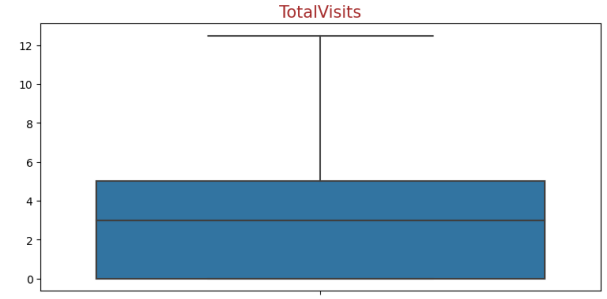
Before

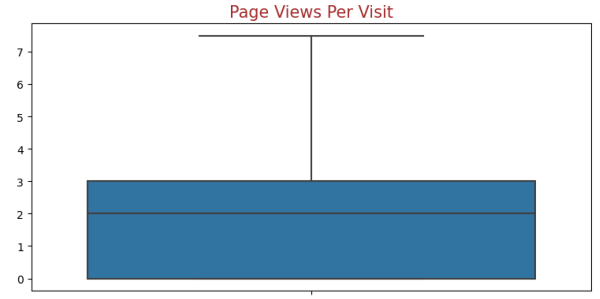


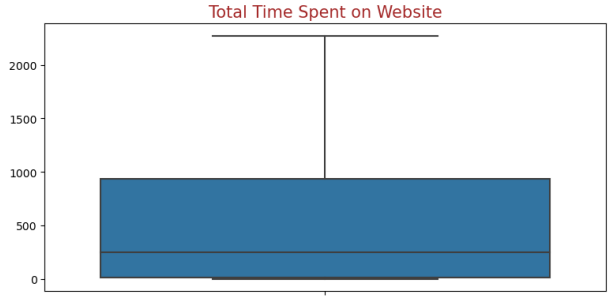




After

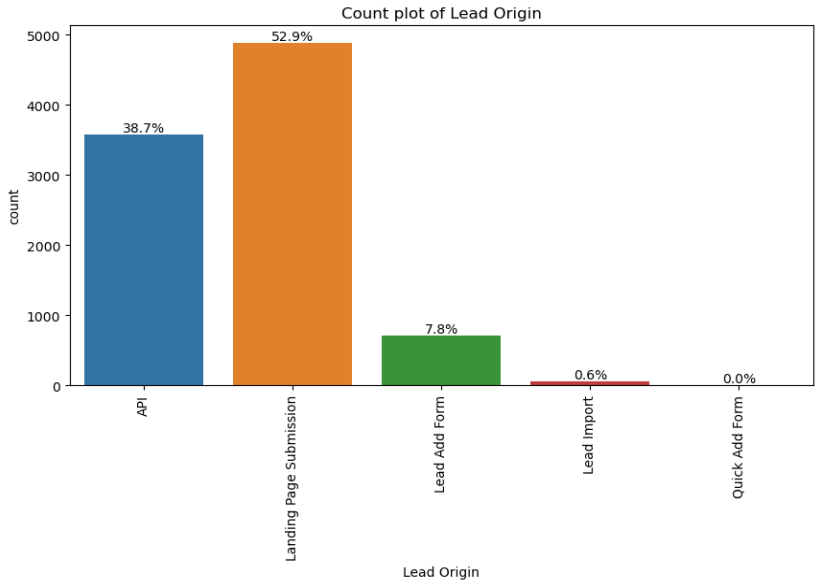


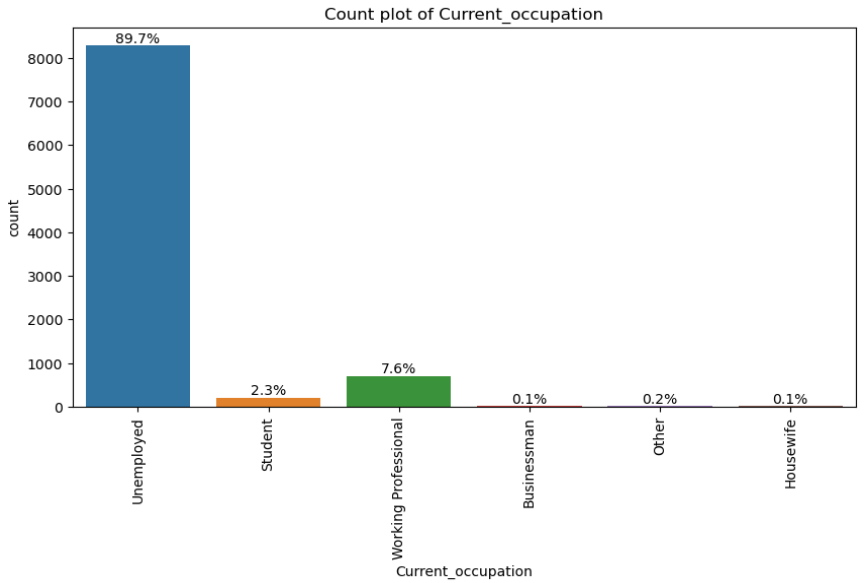


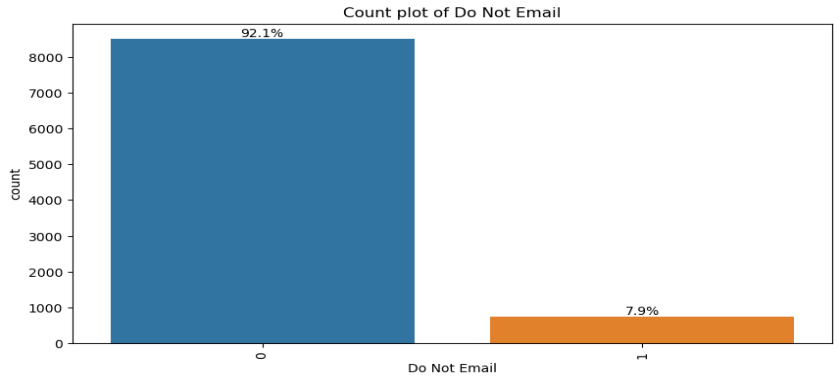


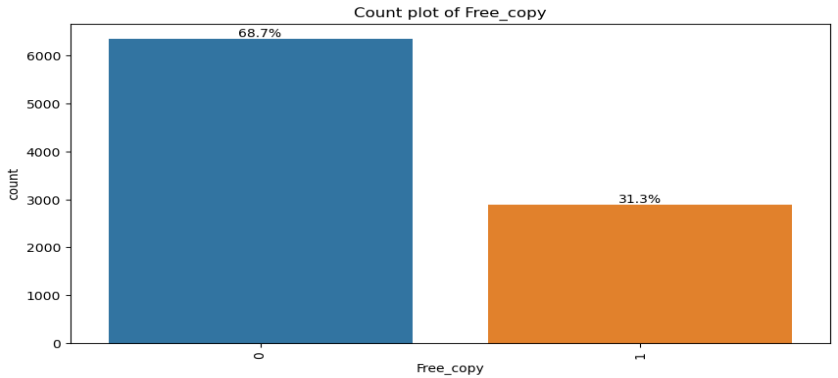
"TotalVisits","Page Views Per Visit":Both these variables contain outliers as can be seen in the boxplot So, these variables with outliers have been treated

UNIVARIATE ANALYSIS



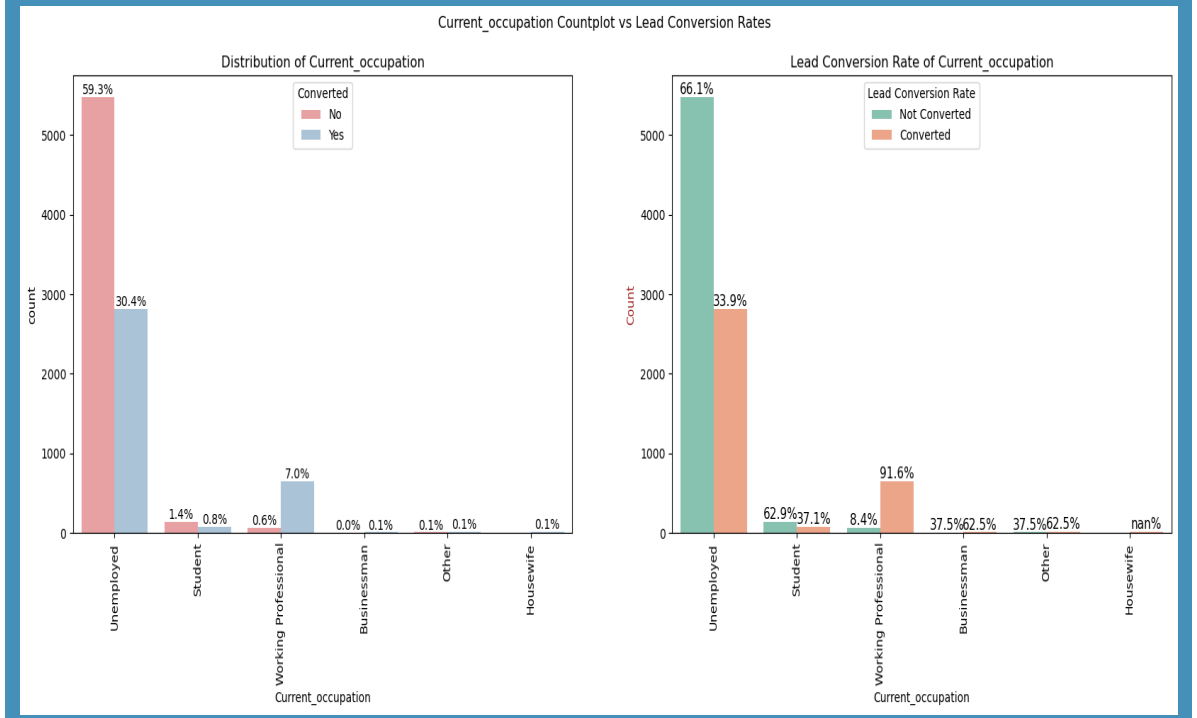






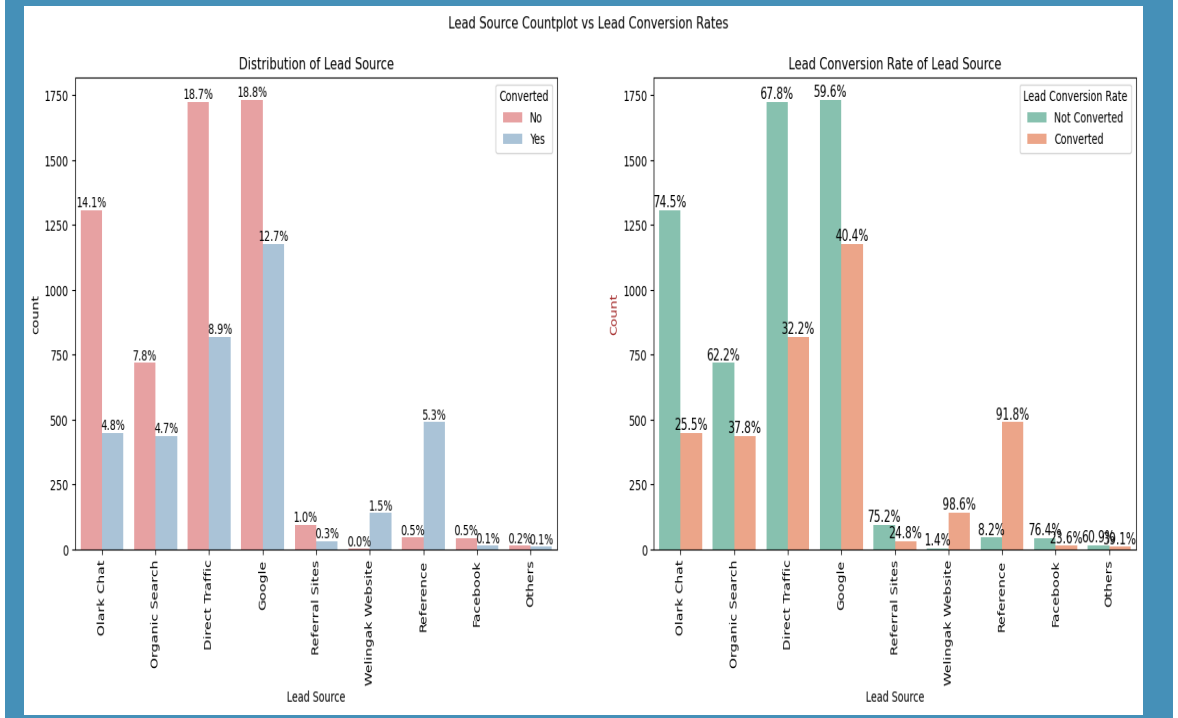
- Lead Origin: Landing Page Submission identified 53% customers; API identified 39%.
- Current occupation: It has 90% of the customers as Unemployed
- Do Not Email: Around 8% of the people has opted that they dont want to be emailed about the course.
- Free_copy : 68% of the people did not opt for the free copy of mastering interview.
- Lead Source: 31% and 28% Lead source is from Google & Direct Traffic combined
- Last Activity: 30 % and 38% last activity of customers from SMS and 38% of Emails.

BIVARIATE ANALYSIS



Almost 80% Leads are from Unemployed, whereas leads from working professional are under 10%.

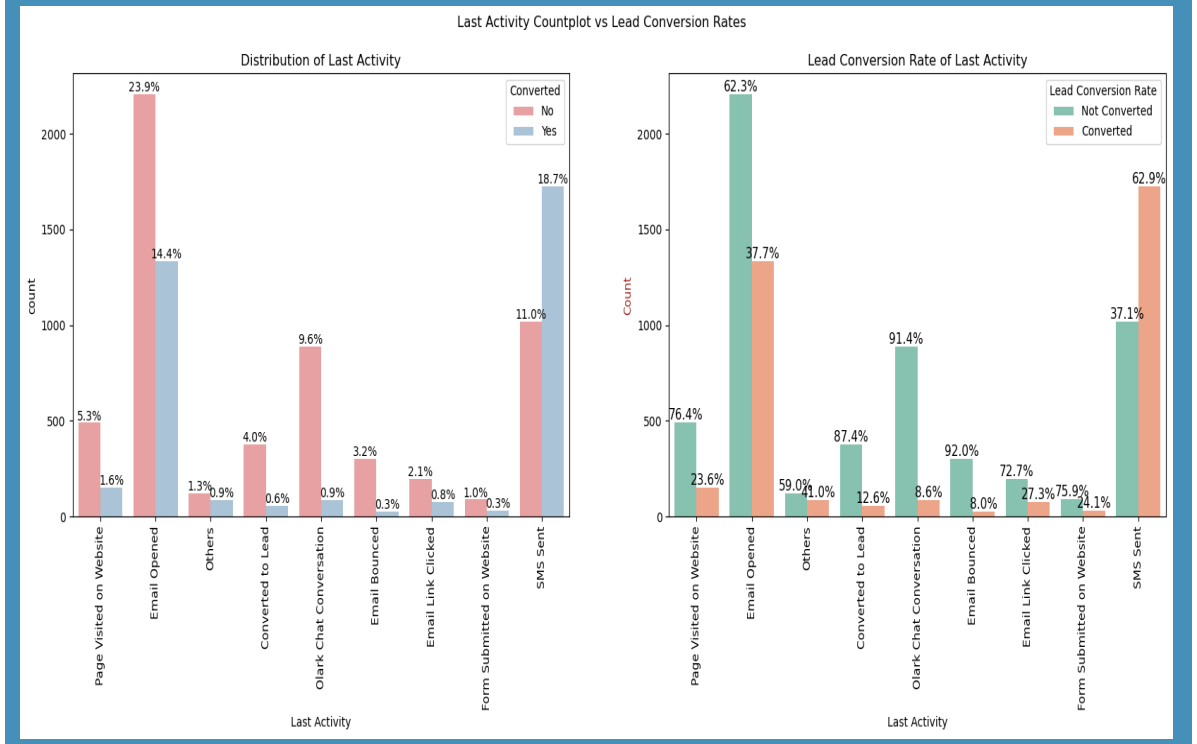
However, the lead conversion rates is very high (>90%) from Working professional



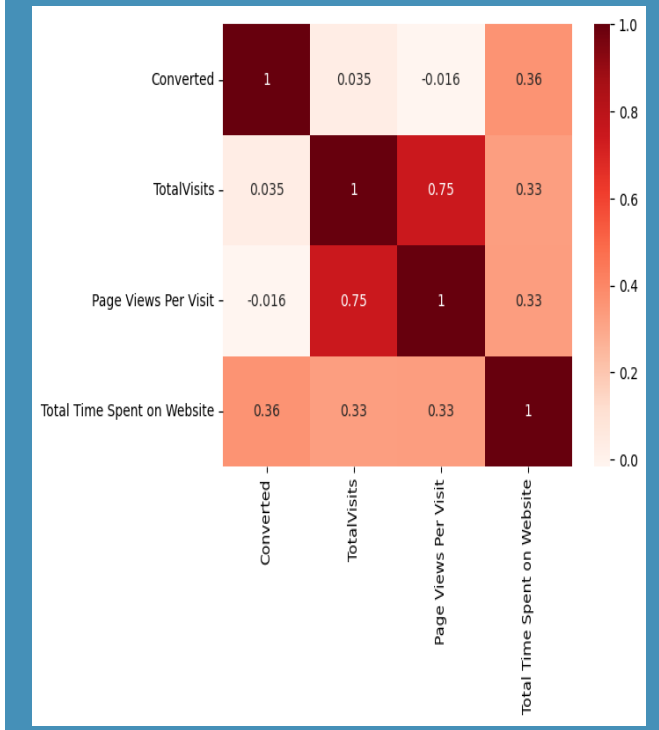
Direct traffic and Google generates very high % of leads, however Google has comparatively higher conversion rates.

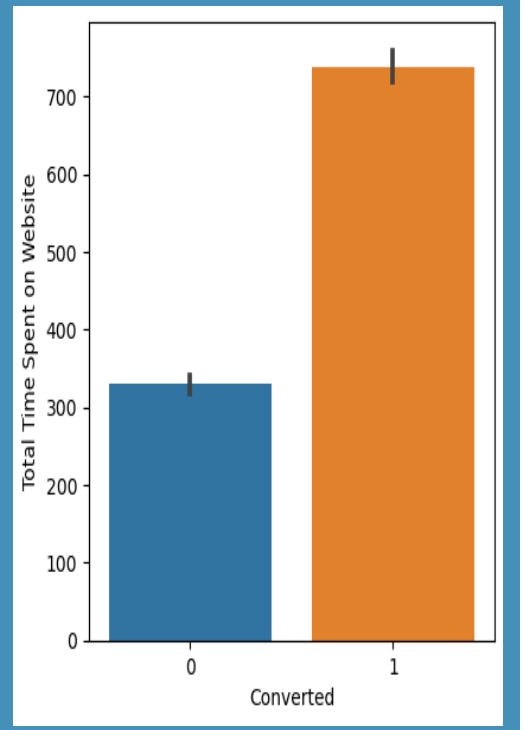
Another point to note is lead conversion from Welingak website and Reference is >90%, however lead generation is very low.

BIVARIATE ANALYSIS - 2



- If we observe Last activity; Email Opened and SMS Sent are highest lead generators and has high lead conversion, however SMS Sent clearly leads the race.





The Heatmap & barplot suggest relatively high correlation between "Converted" and "Total time spent on website"

MODEL BUILDING

- Once, the data is cleaned, dummy variables were generated for all categorical variables.
- The standard scaler was used for feature scaling.
- The train and test data split is set at 70% and 30% respectively.
- The hybrid approach consisting of Recursive Feature Elimination (RFE) and manual feature selection was adopted.
- The insignificant variables with p-values > 0.05 were removed one by one resulting in model 4. This was also showing VIF under 5, suggesting very low multi collinearity.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0236	0.143	-7.145	0.000	-1.304	-0.743
Total Time Spent on Website	1.0498	0.039	27.234	0.000	0.974	1.125
Lead Origin_Landing Page Submission	-1.2590	0.125	-10.037	0.000	-1.505	-1.013
Lead Source_Olark Chat	0.9072	0.118	7.701	0.000	0.676	1.138
Lead Source_Reference	2.9253	0.215	13.615	0.000	2.504	3.346
Lead Source_Welingak Website	5.3887	0.728	7.399	0.000	3.961	6.816
Last Activity_Email Opened	0.9421	0.104	9.022	0.000	0.737	1.147
Last Activity_Olark Chat Conversation	-0.5556	0.187	-2.974	0.003	-0.922	-0.189
Last Activity_Others	1.2531	0.238	5.259	0.000	0.786	1.720
Last Activity_SMS Sent	2.0519	0.107	19.106	0.000	1.841	2.262
Specialization_Hospitality Management	-1.0944	0.323	-3.391	0.001	-1.727	-0.462
Specialization_Others	-1.2033	0.121	-9.950	0.000	-1.440	-0.966
Current_occupation_Working Professional	2.6697	0.190	14.034	0.000	2.297	3.042

	Features	VIF
0	Specialization_Others	2.47
1	Lead Origin_Landing Page Submission	2.45
2	Last Activity_Email Opened	2.36
3	Last Activity_SMS Sent	2.20
4	Lead Source_Olark Chat	2.14
5	Last Activity_Olark Chat Conversation	1.72
6	Lead Source_Reference	1.31
7	Total Time Spent on Website	1.24
8	Current_occupation_Working Professional	1.21
9	Lead Source_Welingak Website	1.08
10	Last Activity_Others	1.08
11	Specialization_Hospitality Management	1.02

MODEL EVALUATION – METRICS & ROC

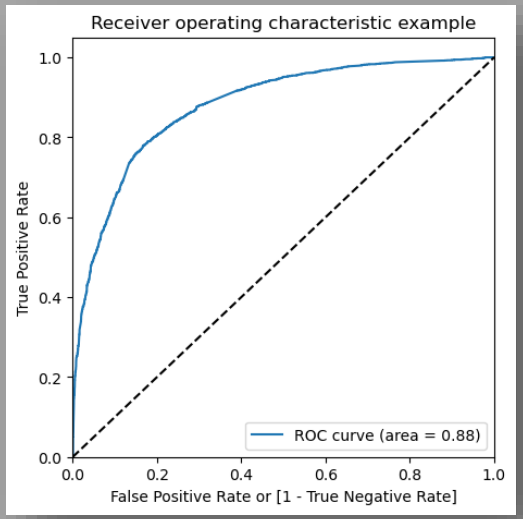
Considering, Probability Threshold = 0.5

Confusion Matrix		
Pred / Act	Negative	Positive
Negative	3588	414
Positive	846	1620

Accuracy = 80.5%

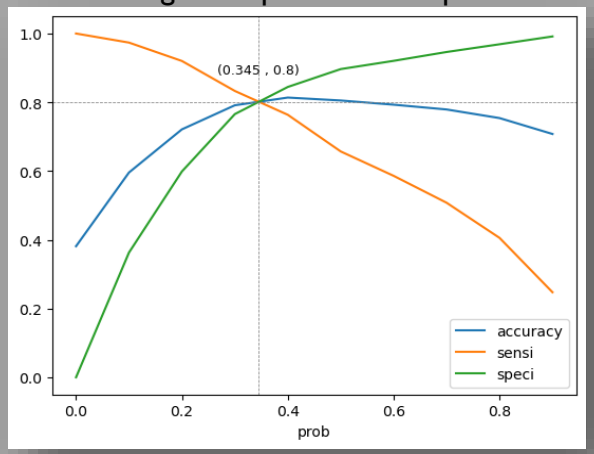
Sensitivity = 65.7%

Specificity = 89.7%



Train Dataset

Finding the Optimal cut-off point



Considering, optimal cut-off = 0.345 from above

Confusion Matrix		
Pred / Act	Negative	Positive
Negative	3230	772
Positive	492	1974

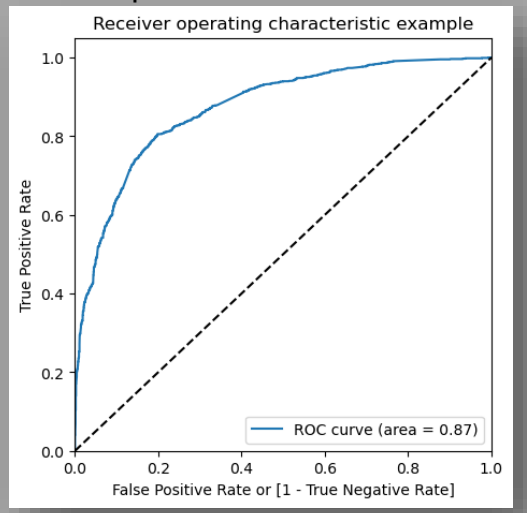
Accuracy = 80.45%

Sensitivity = 80.05%

Specificity = 80.71%

Test Dataset

Based on predicted value from Model 4



Confusion Matrix		
Pred / Act	Negative	Positive
Negative	1353	324
Positive	221	874

Accuracy = 80.34%

Sensitivity = 79.82%

Specificity = 80.68%

- The Area under ROC curve is >0.85 in case of Train and Test data set, which suggest the good model.
- The optimal cut-off is found to be 0.345, based on which further Metrics are derived.
- Also, other evaluation metrics like Accuracy, Sensitivity and specificity are very close in case of train and test data set.

SUMMARY

- Majority of Lead origins are Landing Page Submission and API
- More than 90% leads are from unemployed section; however, the working professional are more likely to get converted
- Google and Direct Traffic is the major Lead Source. Out of these, Google has relatively higher conversion rate.
- Lead conversion from Welingak website and Reference is >90%, however lead generation is very low.
- Under Last activity; Email Opened and SMS Sent are highest lead generators and has high lead conversion as well. Here, SMS Sent clearly leads the race.
- There is relatively high correlation between "Converted" and "Total time spent on website"
- Looking at the coefficient of finalized model 4, following are the top 3 variables contributing to lead conversions: Lead Source, Last Activity, Current Occupation
- The top 3 dummy variables contributing for the lead conversion are: Lead Source_Welingak Website, Lead Source_Reference, Current_occupation_Working Professional.

RECOMMENDATION

- **Strategy to follow for aggressive lead conversion**

a) Since Lead Source_Reference is one of the top contributor. A call between potential lead and someone who has completed a course can be arranged.

b) More efforts should be targeted towards leads from Working Professionals

c) Hospitality Management should not be promoted as specialization

d) Olark Chat has positive correlation when it comes to Lead generation, however Olark Chat has negative impact on conversation. Hence, the conversation experience of potential lead should be improved.

- **Strategy to follow when target achieved before deadline:**

a) Increase targeted advertise on Welingak Website

b) Send emails and messages with quick option to express interest.

c) Engage with major companies to explore possibility of getting in to contract as learning partner.

d) Referral reward program may be rolled out.

e) Work on new and interactive webpage, so that total time spend on webpage increase drastically.

THANK YOU