Lead Score Case Study

Problem Statement

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%

There are quite a few goals for this case study.

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company that we need to answer to also.

Steps taken to build the model

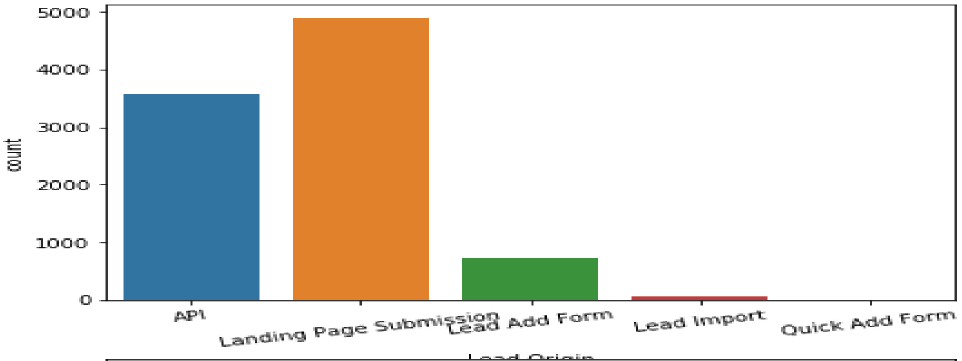
- Data Inspection
- Data Cleaning and Preparation
- EDA using Univariate, Bivariate and Multivariate Analysis
- Model Building
- Conclusion

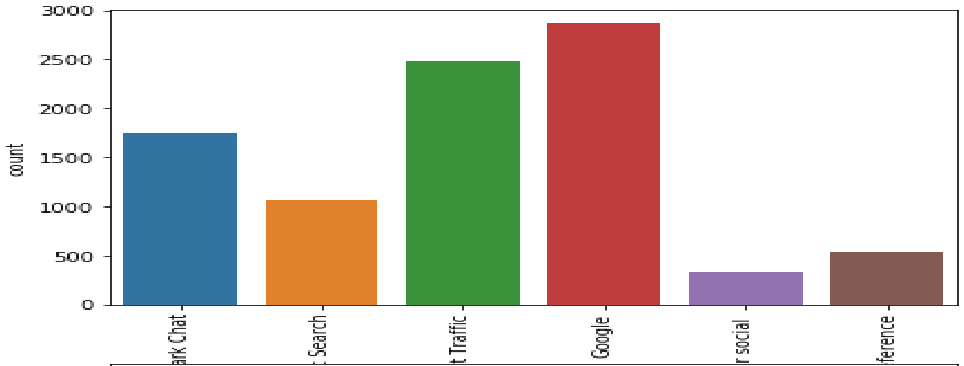
Data Cleaning

- We first inspected the data using dtypes, shape, describe and info fucntions.
- We replaced ‘Select’ with null values in the necessary columns as it means no other value was selected so can be considered as null value.
- In data cleaning, we removed the columns which had 40% or more missing values.
- After that, we deleted the columns that had unique values as they don’t contribute towards the modelling.
- We also dropped the columns that were skewed meaning that had just one value for more than 97% times.
- We also combined categories in the categorical column that had very less values into one category.
- And finally in the outlier treatment, removed the rows for numerical columns that had valued beyond 99th percentile.
- At this stage, we were still able to retain 98% of the data.

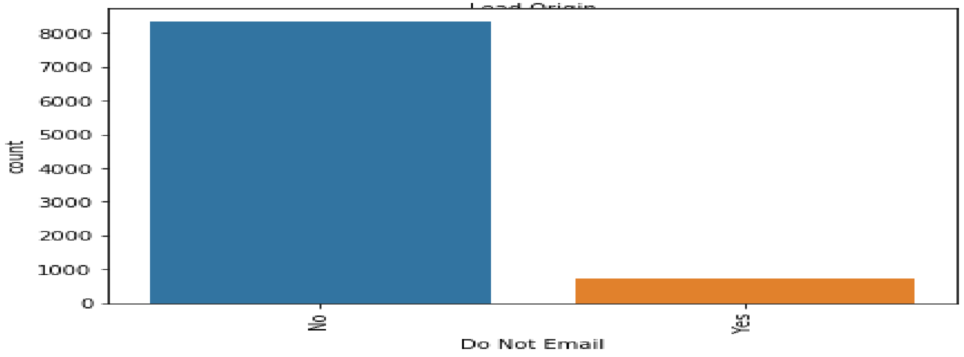
Exploratory Data Analysis: Univariate

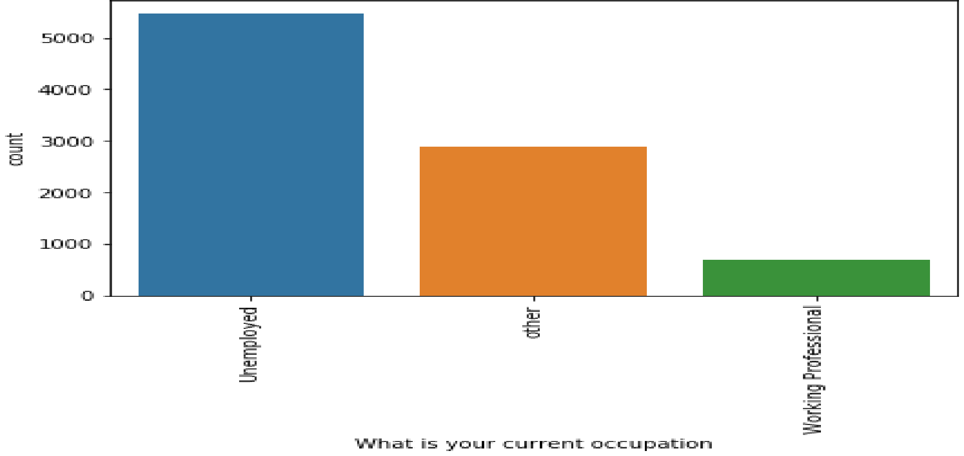
Lead Origin

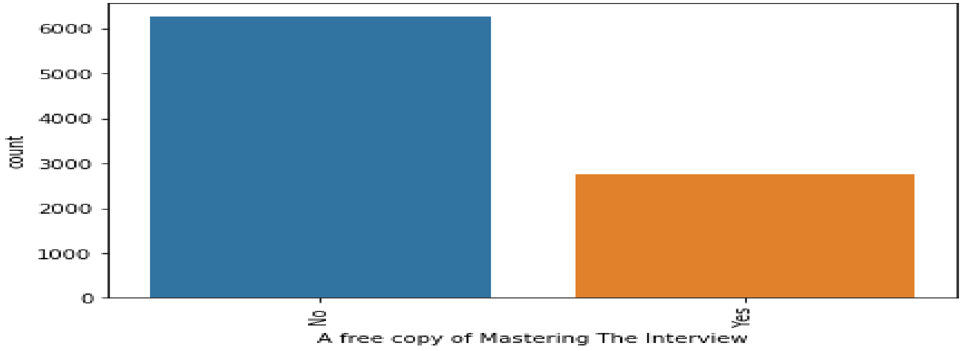




Lead source

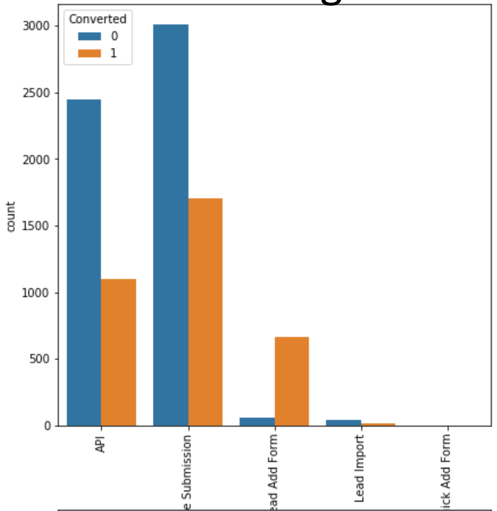




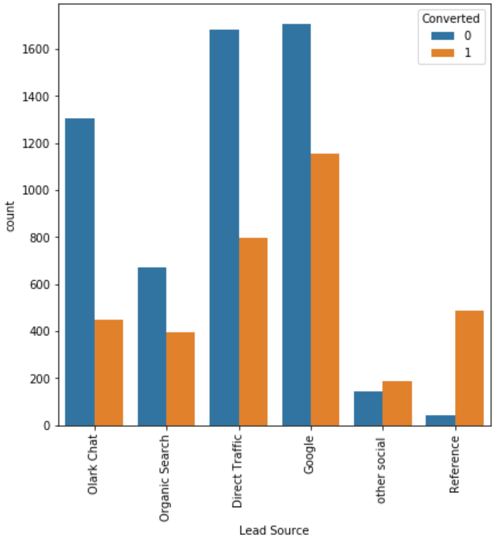


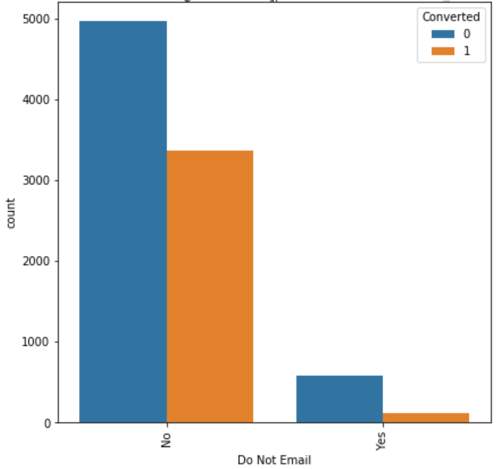
Exploratory Data Analysis: Bivariate

Lead Origin

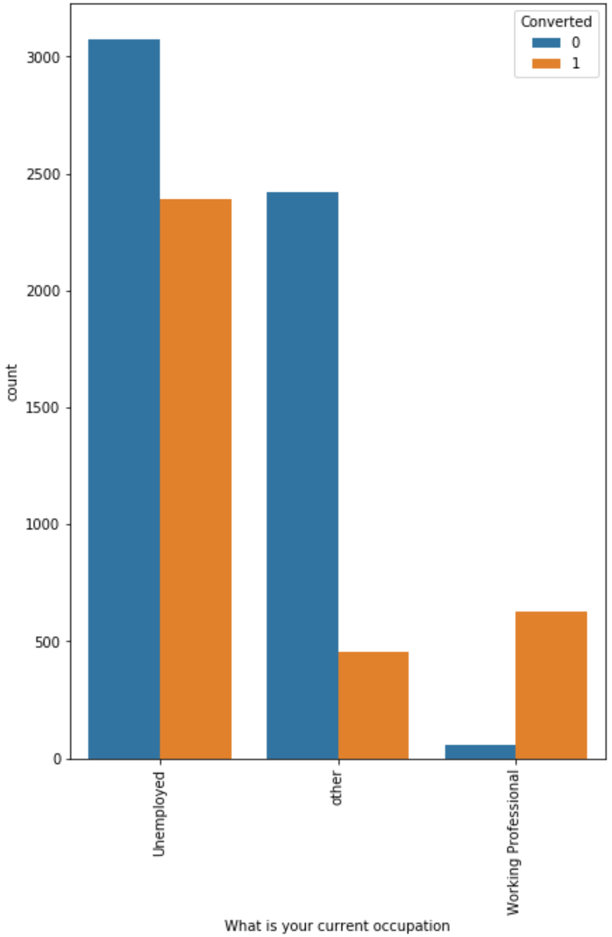


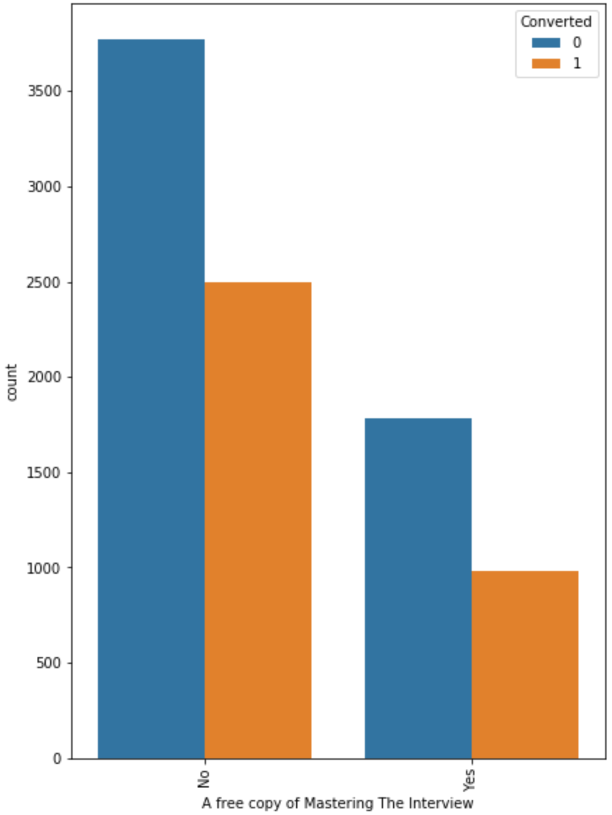
Lead Source



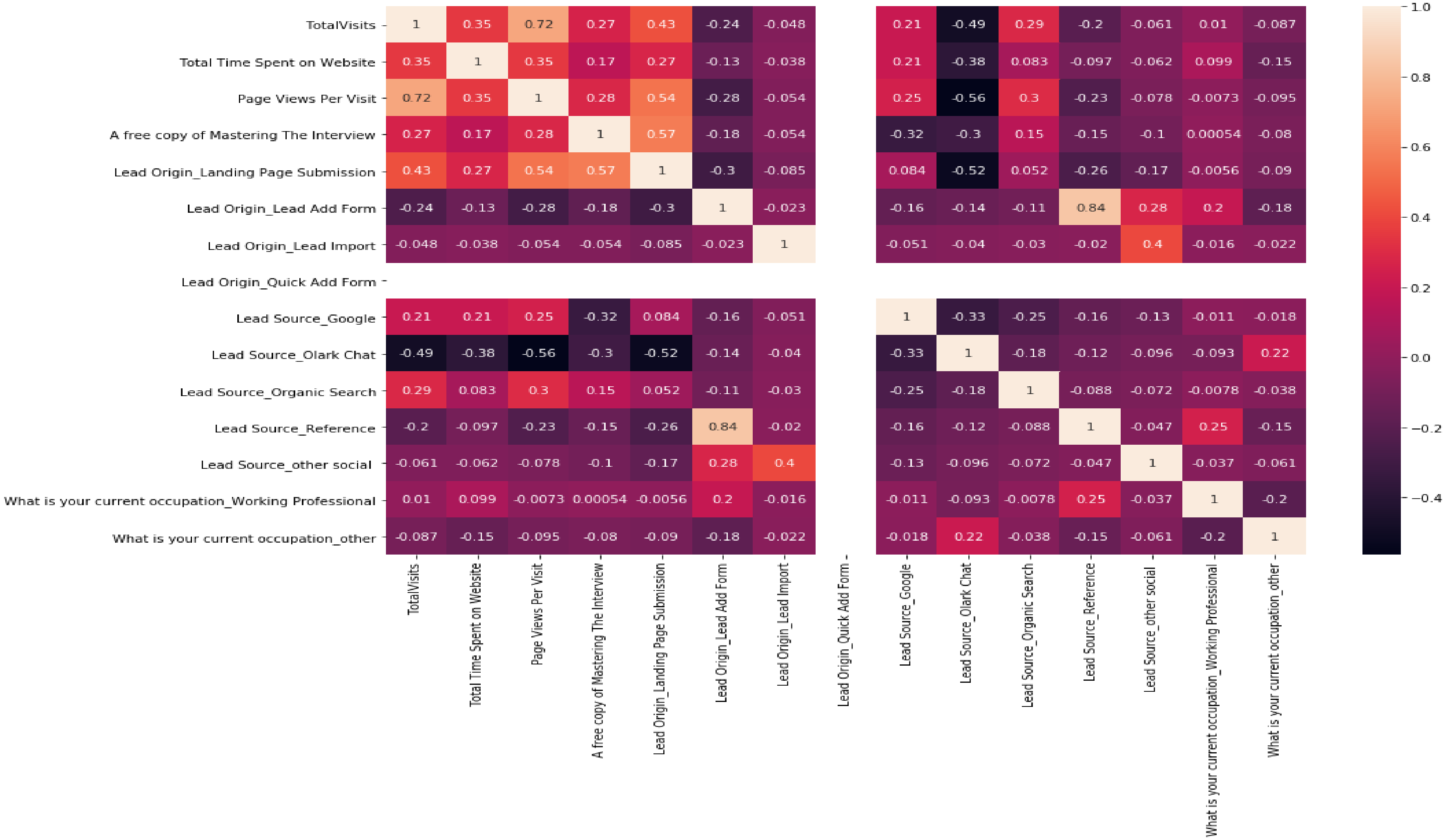


Do Not Email





Exploratory Data Analysis: Multivariate



	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Origin_Quick Add Form	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	Lead Source_Reference	Lead Source_other social	What is your current occupation_Working Professional	What is your current occupation_other
TotalVisits	1	0.35	0.72	0.27	0.43	-0.24	-0.048		0.21	-0.49	0.29	-0.2	-0.061	0.01	-0.087
Total Time Spent on Website	0.35	1	0.35	0.17	0.27	-0.13	-0.038		0.21	-0.38	0.083	-0.097	-0.062	0.099	-0.15
Page Views Per Visit	0.72	0.35	1	0.28	0.54	-0.28	-0.054		0.25	-0.56	0.3	-0.23	-0.078	-0.0073	-0.095
A free copy of Mastering The Interview	0.27	0.17	0.28	1	0.57	-0.18	-0.054		-0.32	-0.3	0.15	-0.15	-0.1	0.00054	-0.08
Lead Origin_Landing Page Submission	0.43	0.27	0.54	0.57	1	-0.3	-0.085		0.084	-0.52	0.052	-0.26	-0.17	-0.0056	-0.09
Lead Origin_Lead Add Form	-0.24	-0.13	-0.28	-0.18	-0.3	1	-0.023		-0.16	-0.14	-0.11	0.84	0.28	0.2	-0.18
Lead Origin_Lead Import	-0.048	-0.038	-0.054	-0.054	-0.085	-0.023	1		-0.051	-0.04	-0.03	-0.02	0.4	-0.016	-0.022
Lead Origin_Quick Add Form															
Lead Source_Google	0.21	0.21	0.25	-0.32	0.084	-0.16	-0.051		1	-0.33	-0.25	-0.16	-0.13	-0.011	-0.018
Lead Source_Olark Chat	-0.49	-0.38	-0.56	-0.3	-0.52	-0.14	-0.04		-0.33	1	-0.18	-0.12	-0.096	-0.093	0.22
Lead Source_Organic Search	0.29	0.083	0.3	0.15	0.052	-0.11	-0.03		-0.25	-0.18	1	-0.088	-0.072	-0.0078	-0.038
Lead Source_Reference	-0.2	-0.097	-0.23	-0.15	-0.26	0.84	-0.02		-0.16	-0.12	-0.088	1	-0.047	0.25	-0.15
Lead Source_other social	-0.061	-0.062	-0.078	-0.1	-0.17	0.28	0.4		-0.13	-0.096	-0.072	-0.047	1	-0.037	-0.061
What is your current occupation_Working Professional	0.01	0.099	-0.0073	0.00054	-0.0056	0.2	-0.016		-0.011	-0.093	-0.0078	0.25	-0.037	1	-0.2
What is your current occupation_other	-0.087	-0.15	-0.095	-0.08	-0.09	-0.18	-0.022		-0.018	0.22	-0.038	-0.15	-0.061	-0.2	1

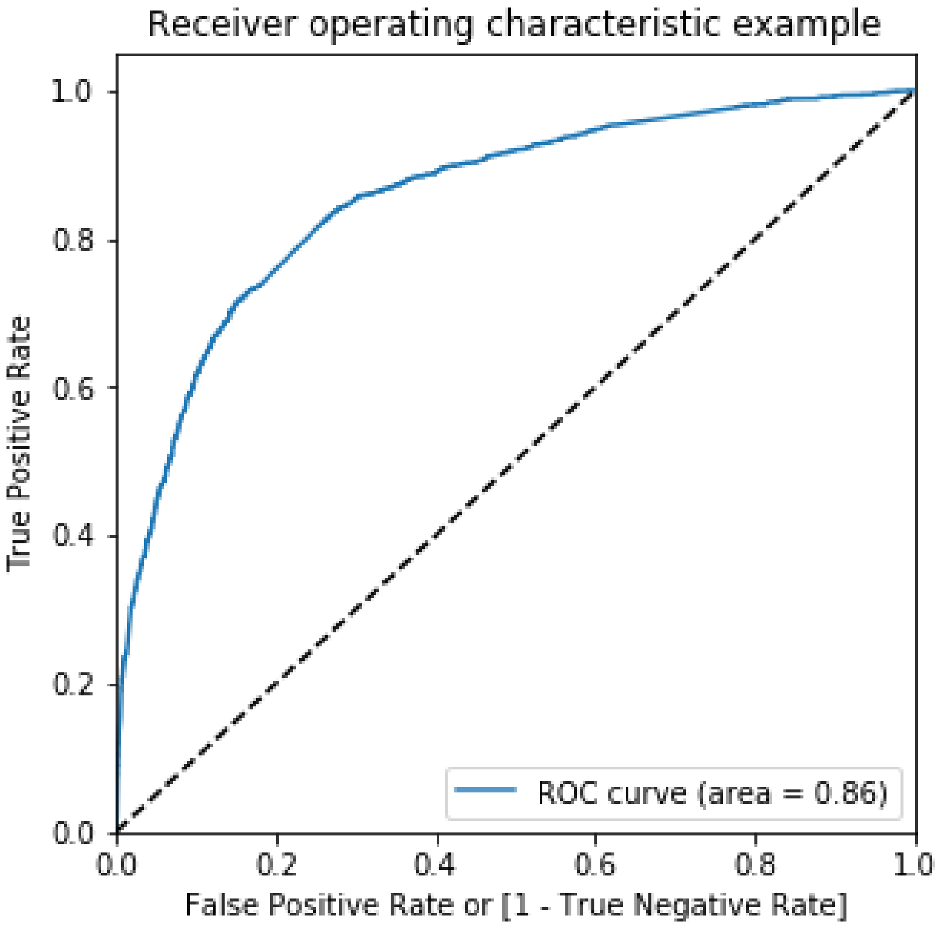
Findings in EDA

1) Here we found that in Lead Origin, the major contributor is Landing Page Submission followed by API.
2) In Lead Source, most of the leads were coming from Google and least were coming from Other social forums.
3) Most of the people opted for Do not Email.
4) Majority of the leads are Unemployed followed by Others.
5) Also, most of the people opted for No in case of A free copy of Mastering the Interview.
6) Most of the leads converted are from Google however Direct Traffic also contributes a lot in conversion.
7) Lead Origin from Lead Ad Form is higher.
8) For Working Professionals, conversion rate is much higher in comparison to others.

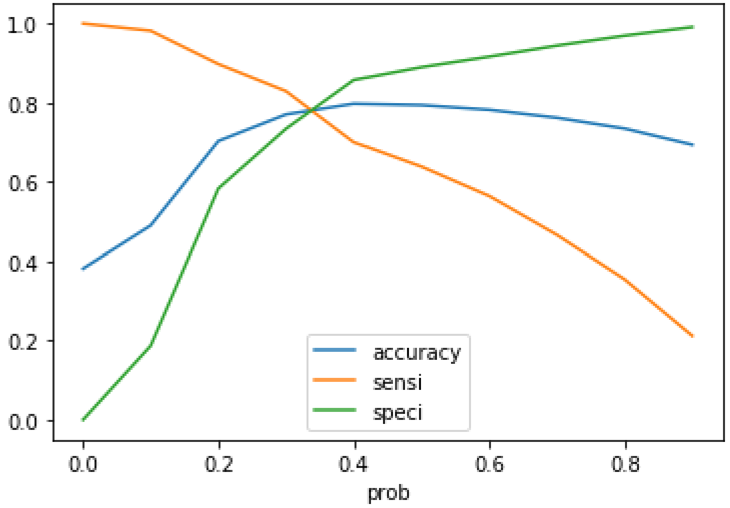
Model Building Steps

- We first divided our data in training and test sets.
- Feature selection was done using RFE with initially 15 features.
- Build the model. Predicted the values using that model.
- Checked p-values and VIF for all the features and basis that we dropped some features having high p- values.
- Finally we were left with 9 features.
- We then created Confusion Matrix.
- Accuracy of the model came out to be 79.38%
- We also calculated other parameters: Sensitivity- 64%, Specificity- 89%
- Then we plotted, ROC curve and basis that found the optimal cut-off point which came out to be 0.3
- Finally, the percentage of final_predicted conversions came out to be 83% with overall accuracy of 77%
- Sensitivity and Specificity of the final model on our test data came out to be 83% and 73% respectively.
- Precision came out to be 67% while Recall came out to be 85%.

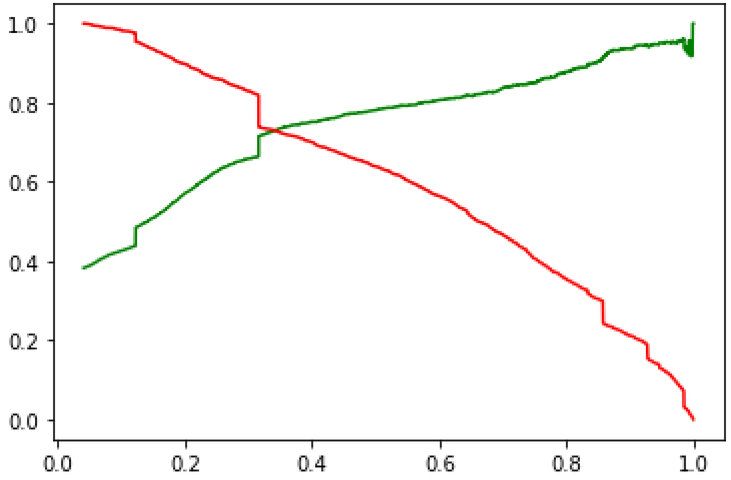
Curves in the model



ROC Curve



Sensitivity, Specificity
And Accuracy Curve



Recall and Precision Curve

Precision- Green
Recall- Red

Conclusion

The parameters that contributes the most towards a hot lead are as follows-

a) Total Visits

b) Total Time Spent on Website

c) Lead Origin through Lead Add Form

d) Lead Source through -

i) Google

ii) Olark Chat

iii) Reference

e) What is your Current Occupation-

i) Working Professionals

ii) Other

Hence, in order to convert most leads X Education should calls these leads.

As suggested in the subjective answers, they should use Social Media in order to promote the business when they don’t make phone calls and should use other means as well like text messages and emails.