X Education - Lead Scoring Case Study

Identification of Hot Leads to focus more on them and thus enhancing the conversion ratio for X Education

Submitted by
Pooja Shakkarpude
Srinivasa Varadhan

Background

X Education Company

- X Education , An education company named sells online courses to industry professionals
- Many interested professionals land on their website
- The company markets its courses on several websites 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
- 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%

Problem Statement

X Education Company's Problem

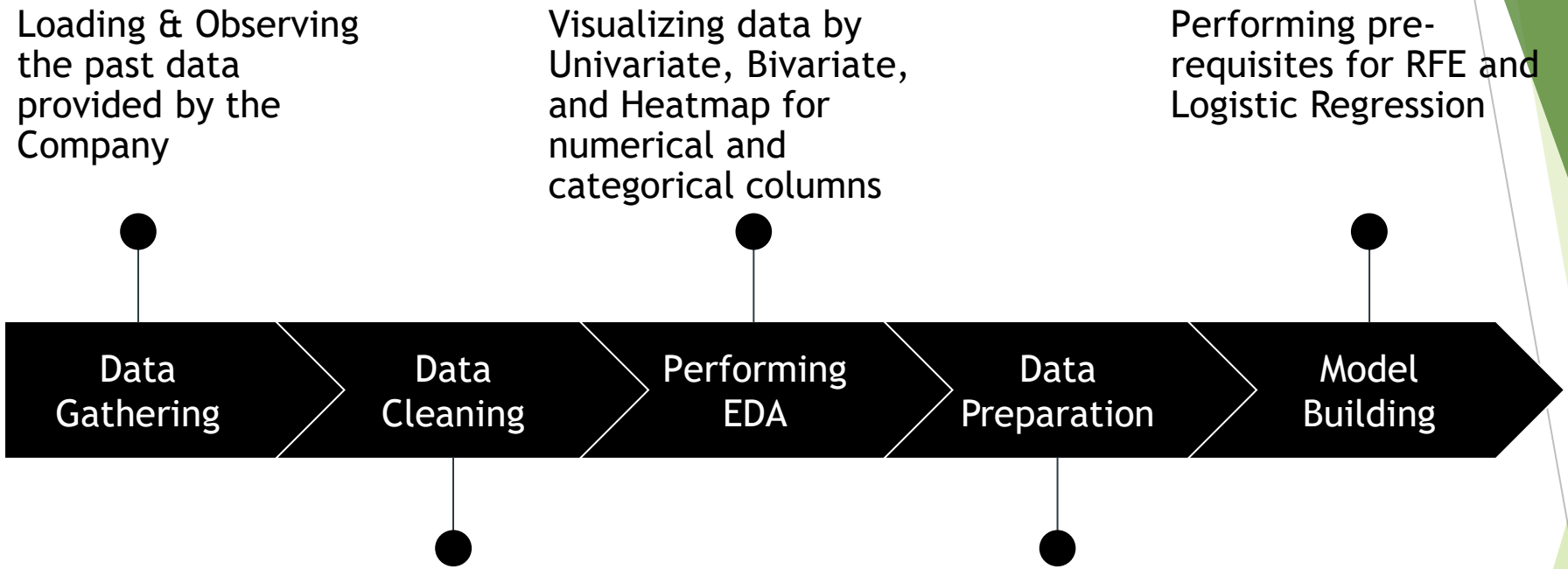
- X Education gets a lot of leads but its lead conversion rate is very poor
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone
- We will help them to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- We are required to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be 80%.

Solution

Selection of Hot Leads

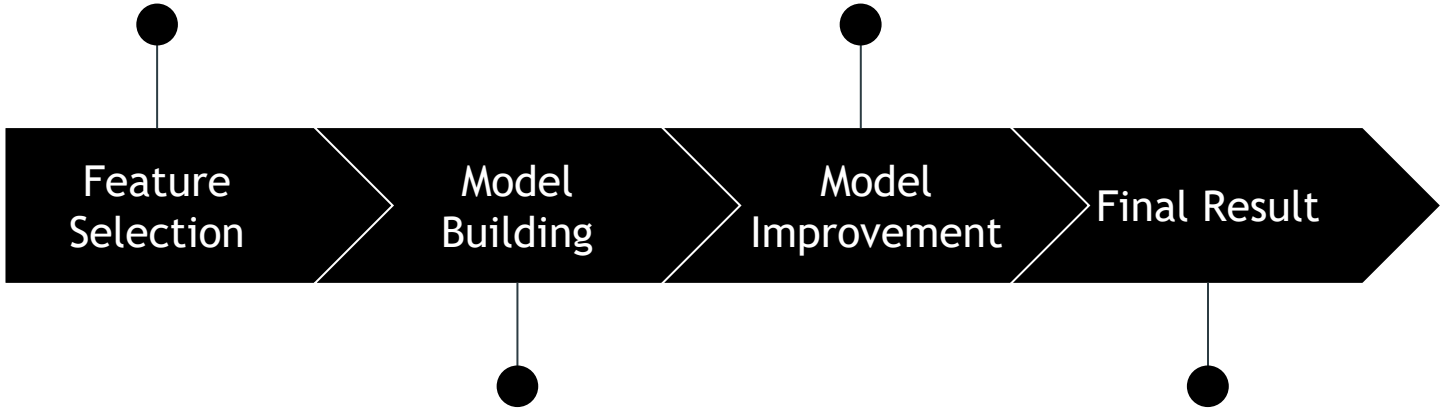
- For our Problem Solution, the crucial part is to accurately identify hot leads.
- The more sensitivity we obtain the hot lead, the more chance we get of higher conversion ratio.
- Since we have a target of 85% conversion rate, we would want to obtain a high accuracy in obtaining hot leads.

Implementation


Loading & Observing the past data provided by the Company
Visualizing data by Univariate, Bivariate, and Heatmap for numerical and categorical columns
Performing pre-requisites for RFE and Logistic Regression
Data Gathering
Data Cleaning
Performing EDA
Data Preparation
Model Building
Duplicate removal, null value treatment, unnecessary column elimination, etc.
Outlier Treatment, Feature – Standardization, Dummy Variable creation

Using RFE

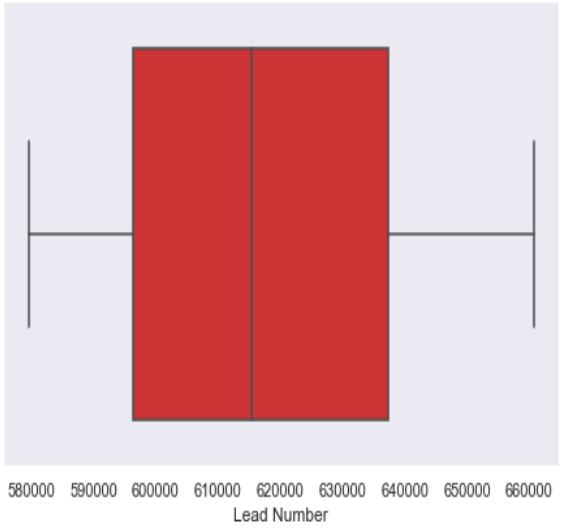
Reduction of columns and Model re-building with p value and VIF

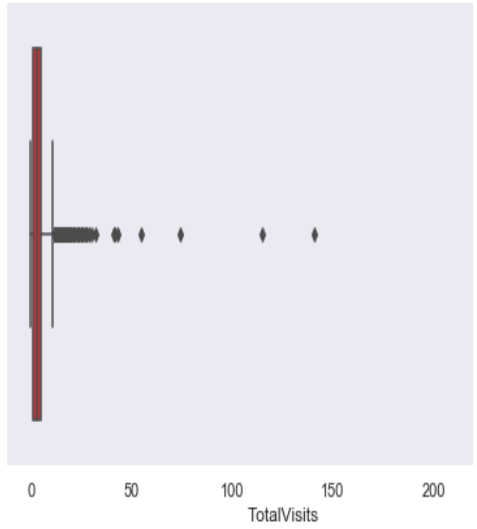


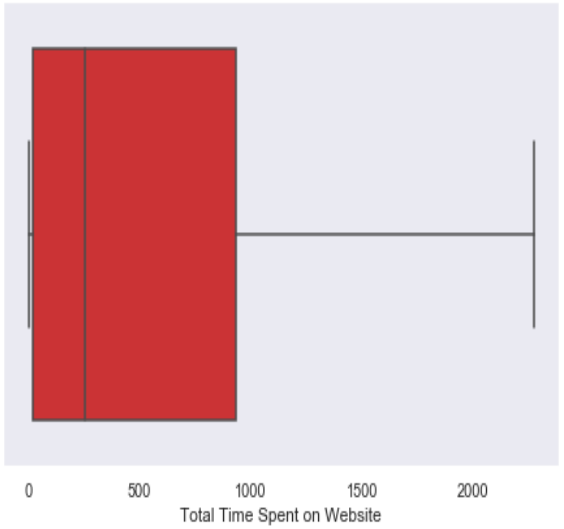
Model building using GLM for selected columns

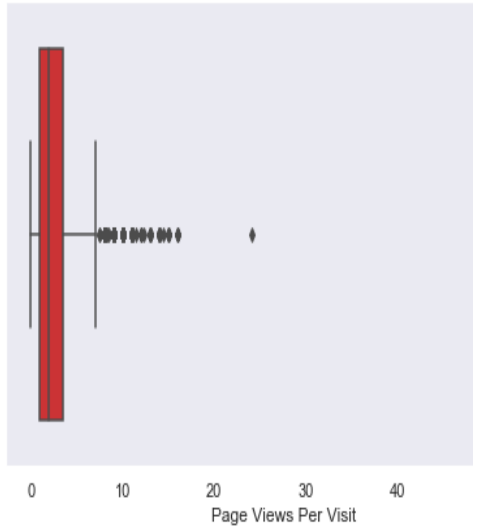
Final Model Analysis and performance on Test Data

Visualization



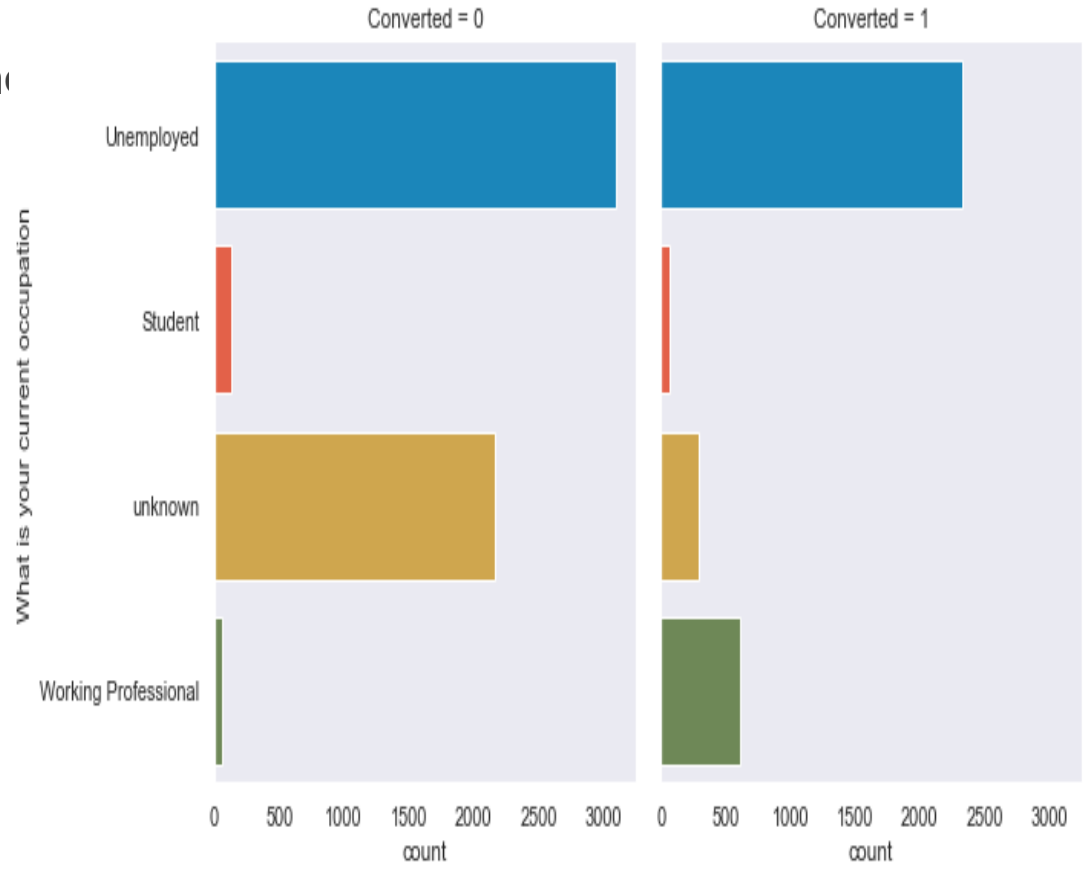




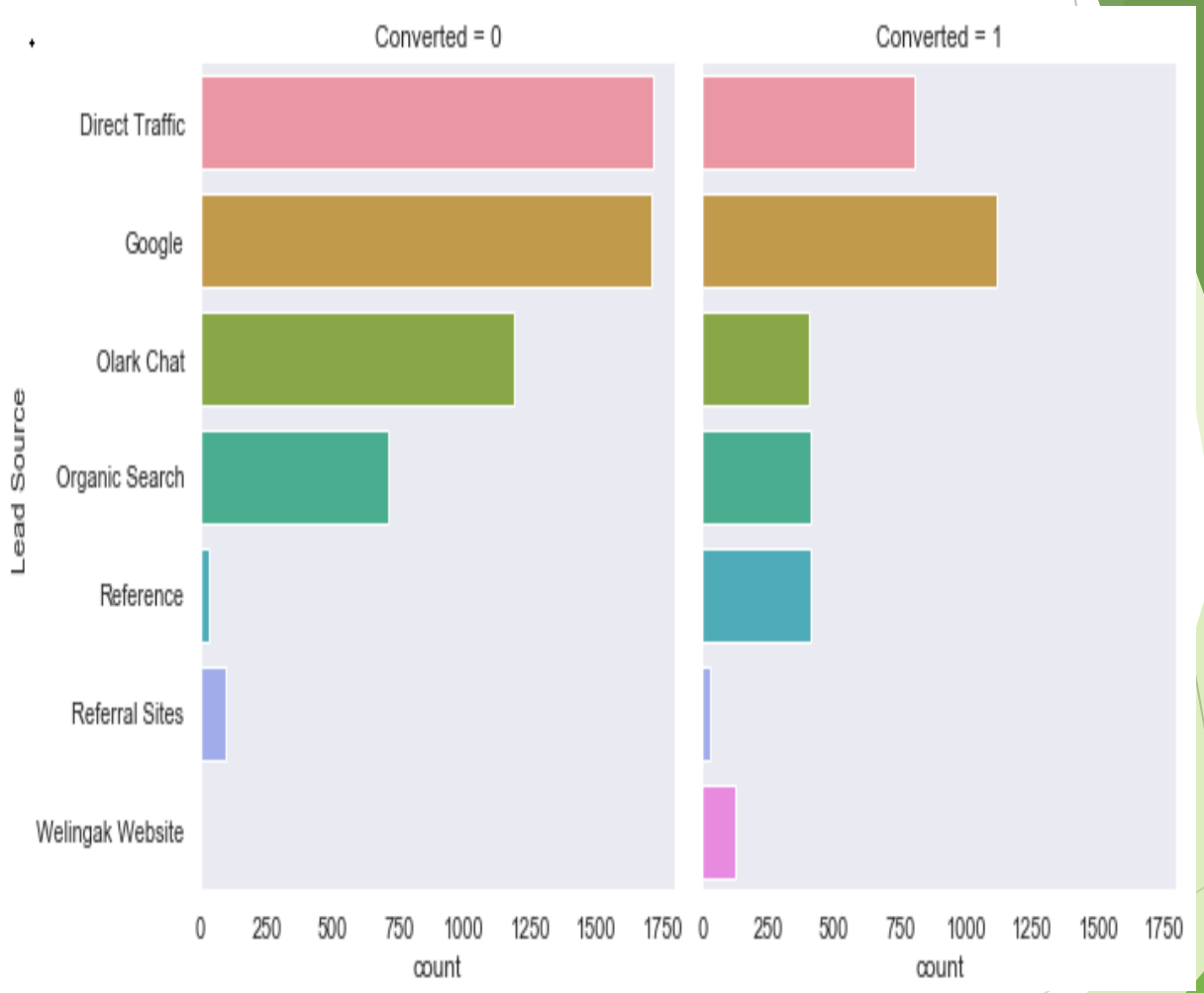


- In the numerical columns :
  Total Visits and *Page views Per visit* have outliers.

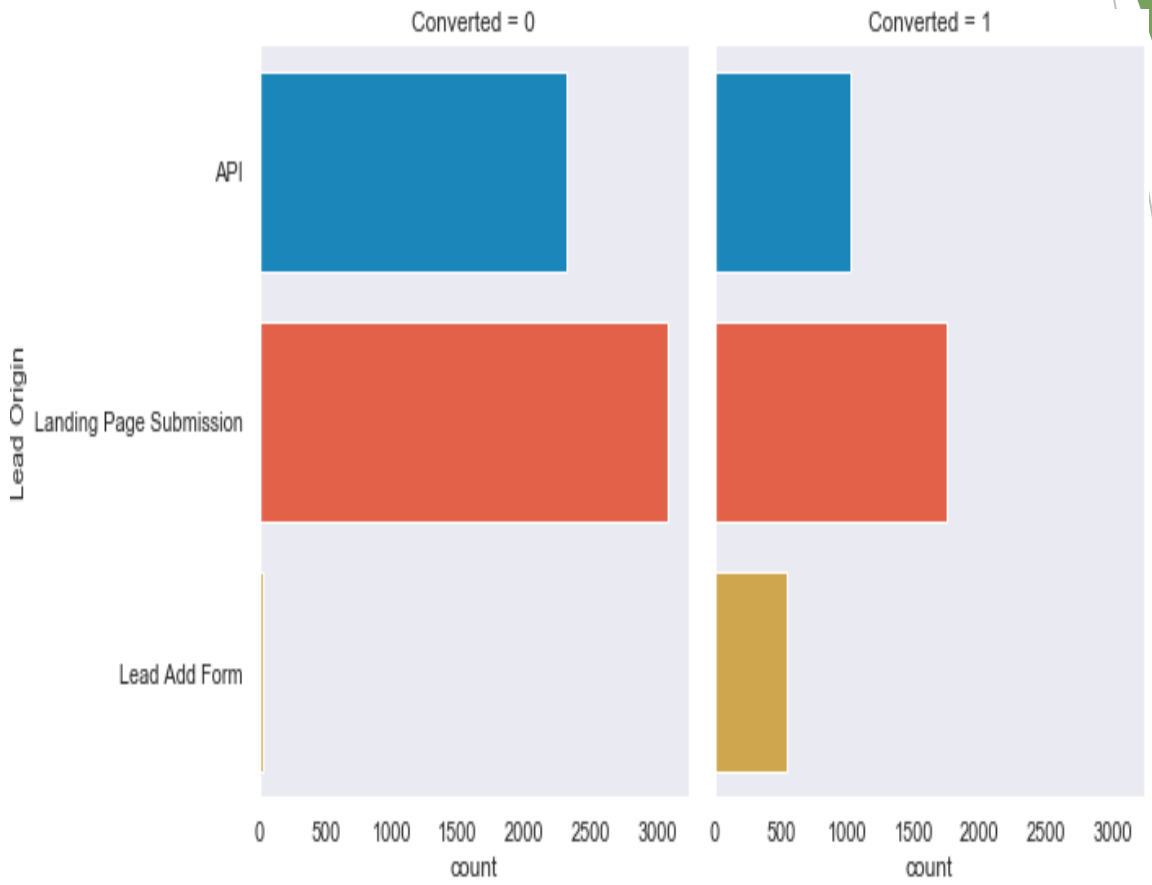
- EDA plots depicting variation in categorical columns for those wh Converted and those who didn't, and here we can see Working professional are more towards conversion.



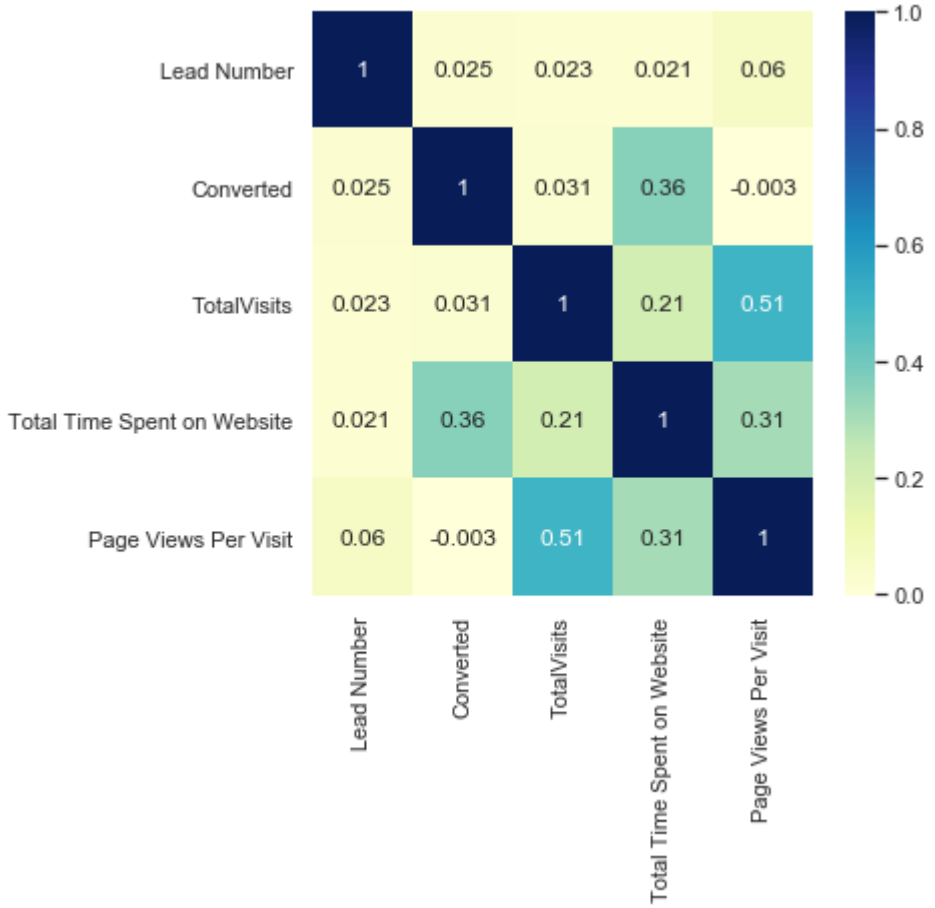
- EDA plots depicting variation in categorical column (Lead Source). Here we can see Reference and Welingak Website are more towards conversion and Olark chart is performing poorly.

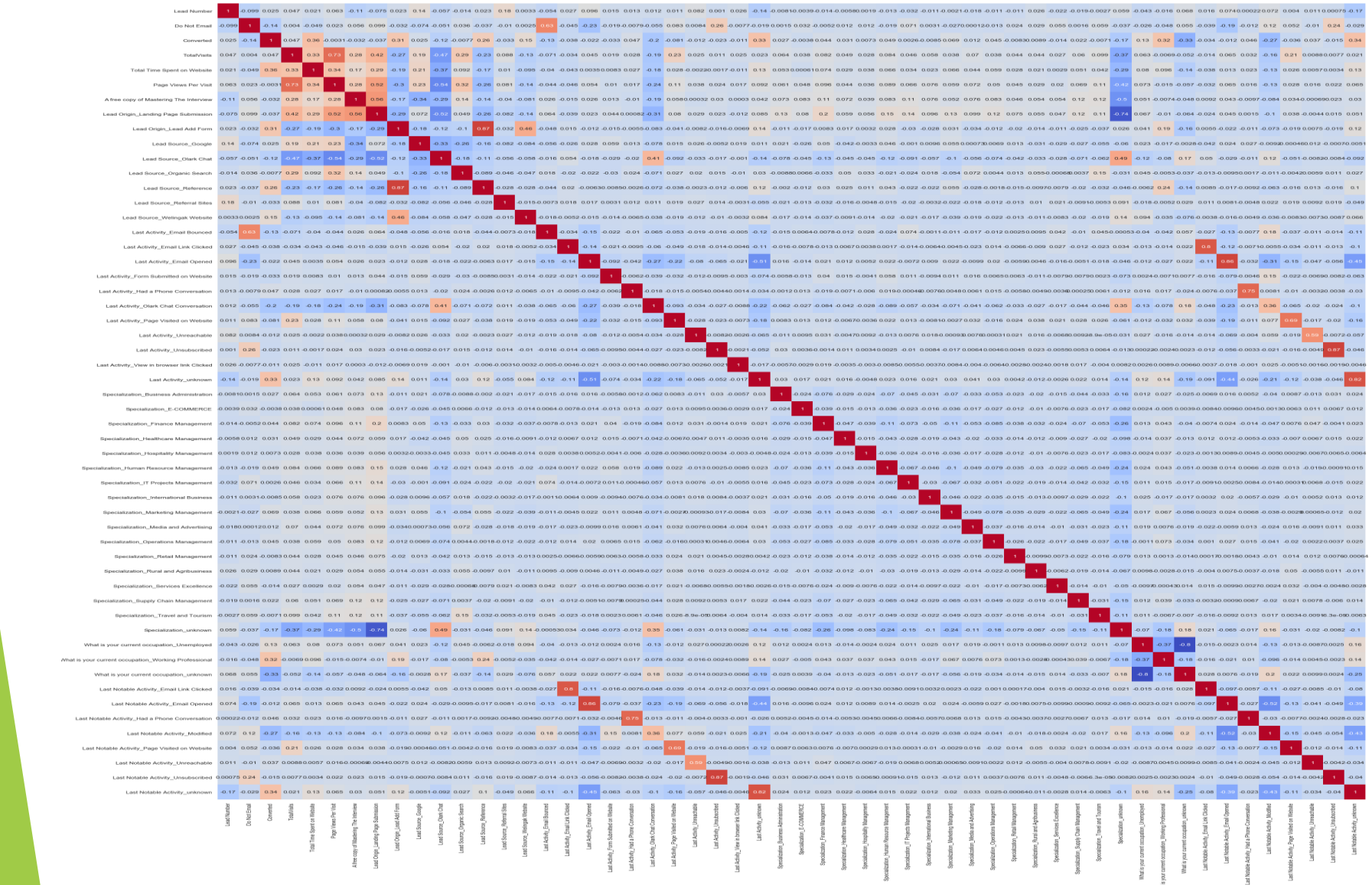


- EDA plots depicting variation in categorical column (Lead Origin) for converted and who didn't.
- Leads which are orginated from the Lead Add form are more towards conversion



- EDA plots depicting correlation (Heat Map) of all selected numerical columns.
- Total Time Spent on Website are highly correlated with Conversion.

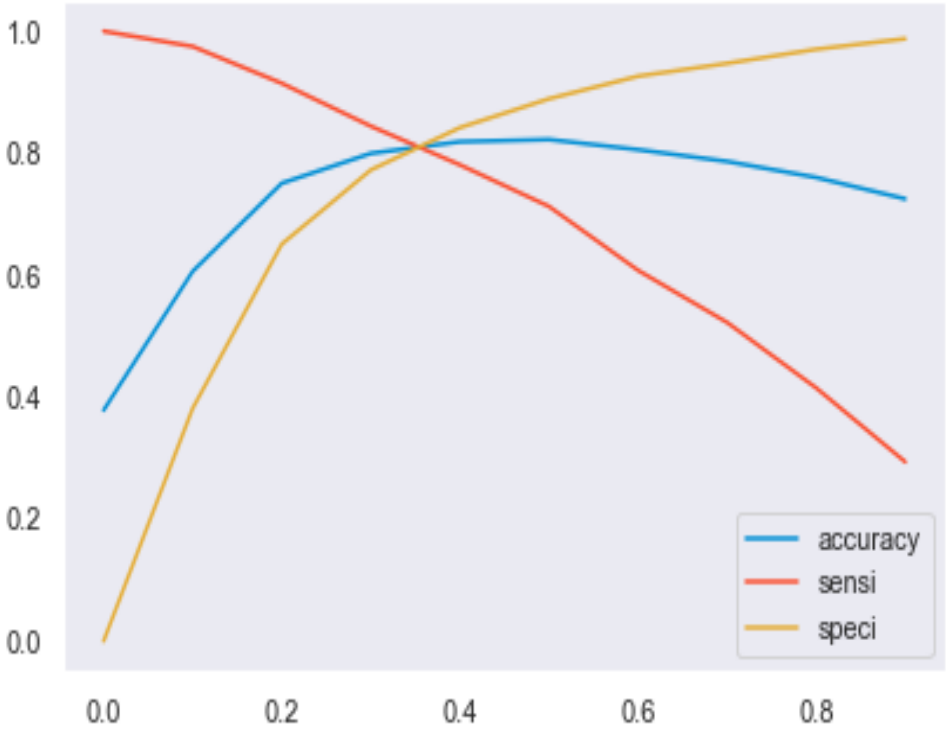


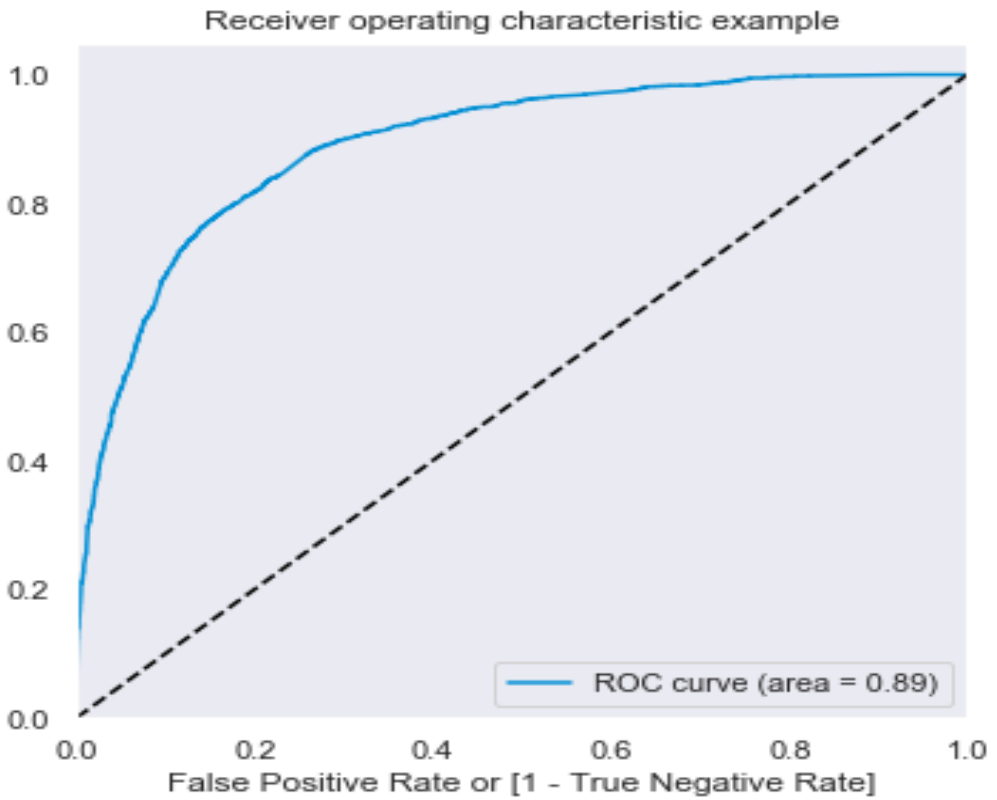


- EDA plots depicting correlation (Heat Map) of all selected columns numerical columns and dummy columns).
- Specilization details are not filled by all the leads from source the Olark Chat. Need to identify whats wrong in UI or input cells

Linear Regression Final Model Parameters

Area under ROC = 0.89





	prob	accuracy	sensi	speci
0.0	0.0	0.378511	1.000000	0.000000
0.1	0.1	0.606140	0.974978	0.381503
0.2	0.2	0.750653	0.914582	0.650815
0.3	0.3	0.799641	0.844262	0.772465
0.4	0.4	0.818746	0.780846	0.841829
0.5	0.5	0.822338	0.712683	0.889122
0.6	0.6	0.805683	0.607852	0.926169
0.7	0.7	0.786414	0.522433	0.947189
0.8	0.8	0.760124	0.415013	0.970310
0.9	0.9	0.725016	0.293788	0.987651

Inferences from Model

Business Insights Derived from our Model

Top 3 variables in model, that contribute towards lead conversion are:

a) Lead Source
b) Last Activity
c) What is your current occupation?

Inferences from Model

Business Insights Derived from our Model

Top 3 variables in my model, that should be focused are:

a) Lead Source – Welingak Website
b) Lead Source – Reference
c) What is your current occupation - Working Professional?

Conclusion

- The very hot leads are the leads which are having z-score greater then 60%. One group of the sales team should focus on these peoples for conversion.
- The other hot leads which are having z-score between 60%-30% and has to be focused as next priority.
- The leads which are having lesser than 30% z-score must called and directed to Welingak website or try to understand their challenges in getting them converted.
- Benchmark Welingak website and figure out what it is making it to be a major source of conversion.
- Olark Chat is worst performing, and most data is missing through this channel. Figure out if there is a challenge in UI, database issue or the profile of the users' is a reason for it. Fix this issue so that there will be no missing values.
- Benchmark Lead Add form and try to implement the reason for conversion same in the landing page submission.