Lead Score Case Study

Name – Samiksha Zagade

Sahil Amale

Shivendra Singh

Data

- 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. 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%. We have built a model where a lead score is assigned to each of the leads such that the customers with a higher lead score have a higher conversion to increase the conversion rate.

Approach

- Data Reading and Understanding
- Data Cleaning
- EDA
- Data Preparation
- Model Building
- Model Evaluation
- Optimal Cutoff(ROC Curve)
- Predictions on Test set
- Conclusion

Step 1 – Getting the data ready for Analysis

- We first imported the required libraries and started reading and understanding the data.
- The data has 9240 rows and 37 columns where17 columns had null values and 6 columns had null values greater than 30% which were dropped later on.
- After dropping the unnecessary variables, handling the null values and outliers we were left 14 final columns.
- The final Columns are Lead Origin, Lead Source, Do Not Email, Total visits, Total time spent on website, Pages view per visit, specialisation, city, What is your current occupation, What matters most to you in choosing a course, A free copy of Mastering The Interview, Lead Profile and Last Notable Activity.

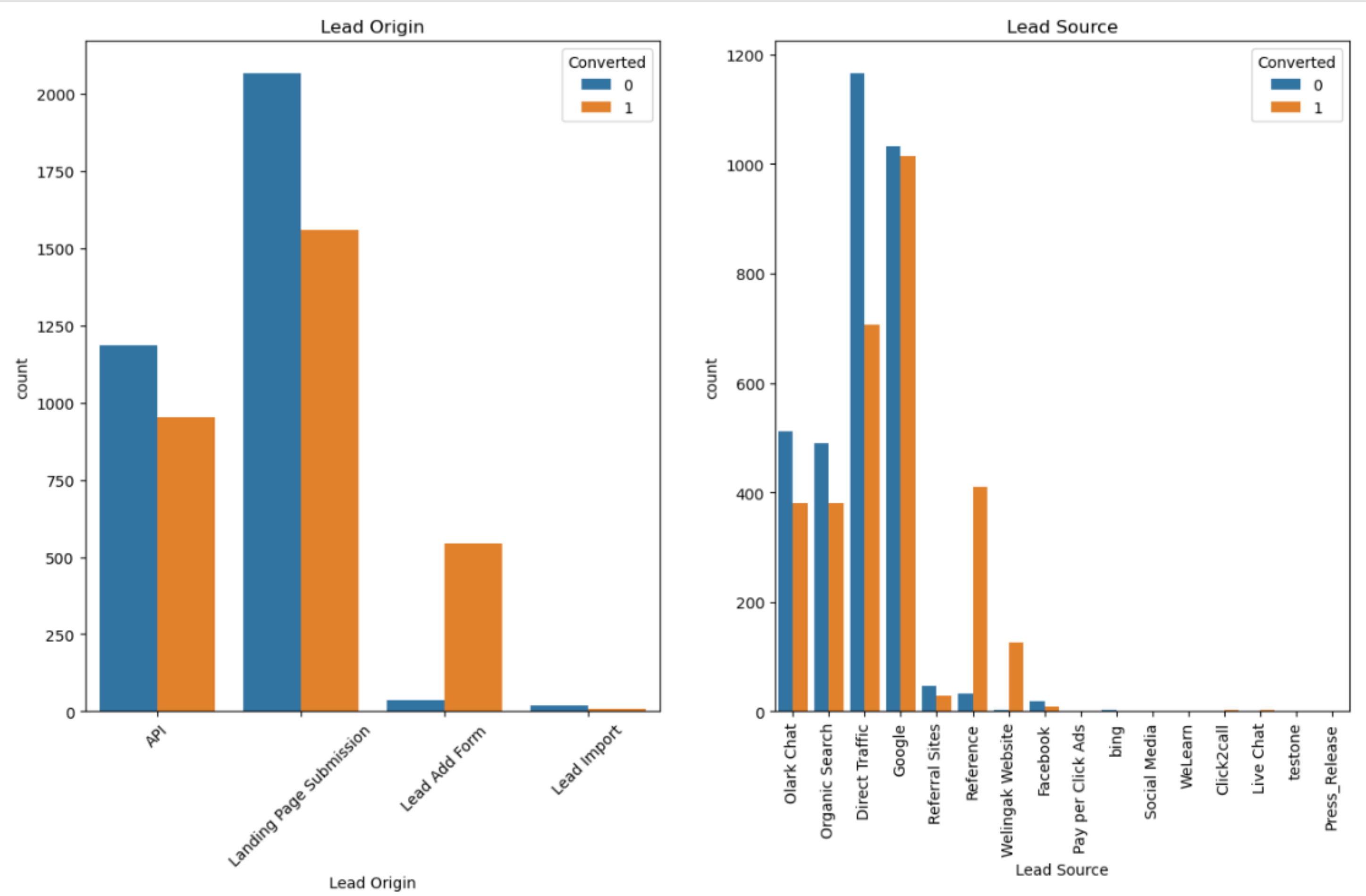
Visualisation

Lead Origin-

- Landing page submissions has highest conversion and lead add form has more conversion than non conversion.

Lead Source-

- Google has highest conversion.
- Direct Traffic has high non conversion rate.



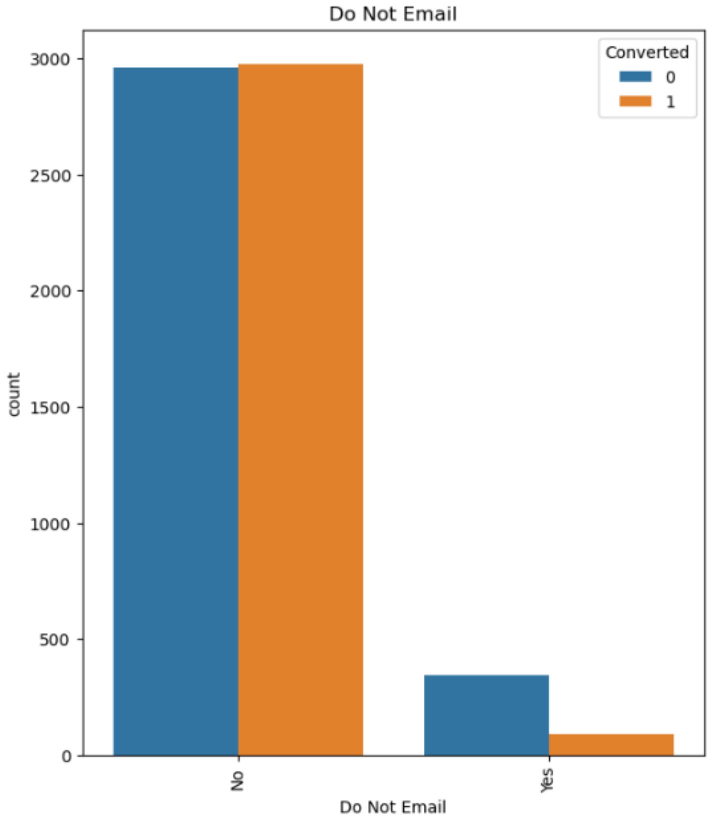
Visualisation

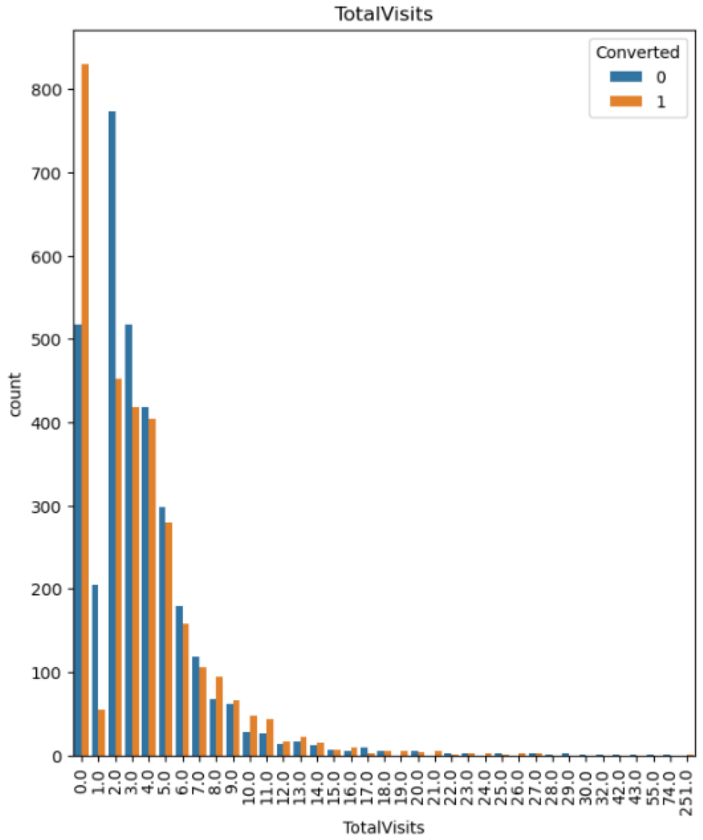
Do not Email-

- The people who are converted have not preferred getting emails.
- The conversion rate and non conversion rate is almost equal for the people who do not want email.

Total Visits-

- 0 visits has highest conversion rate.
- People who have visited the site 2 times have not joined the courses it seems.





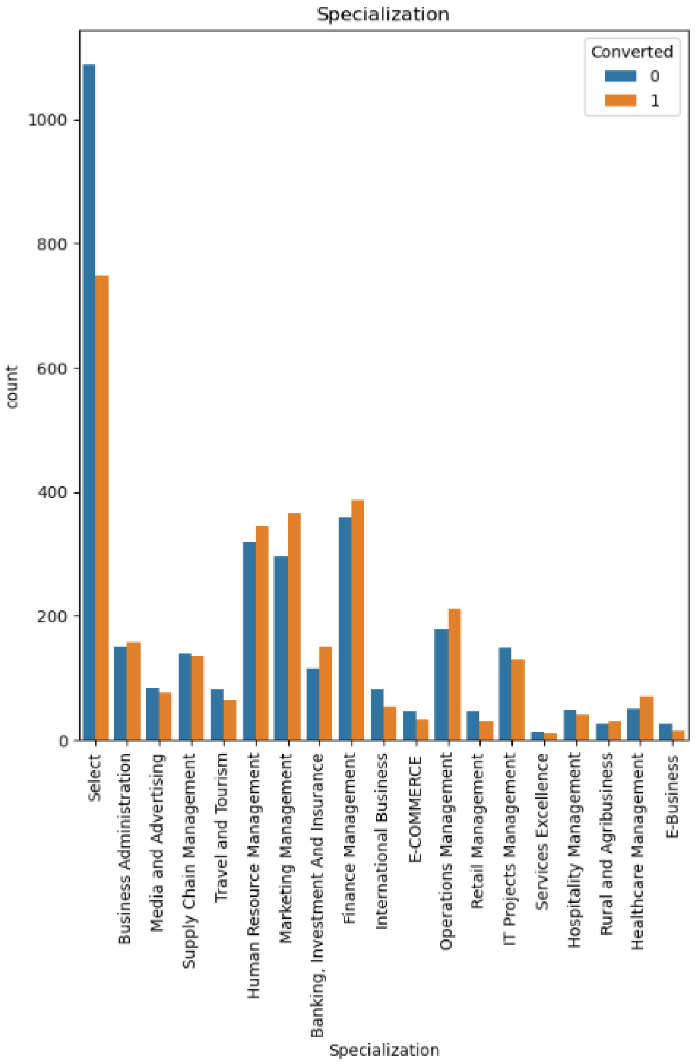
Visualisation

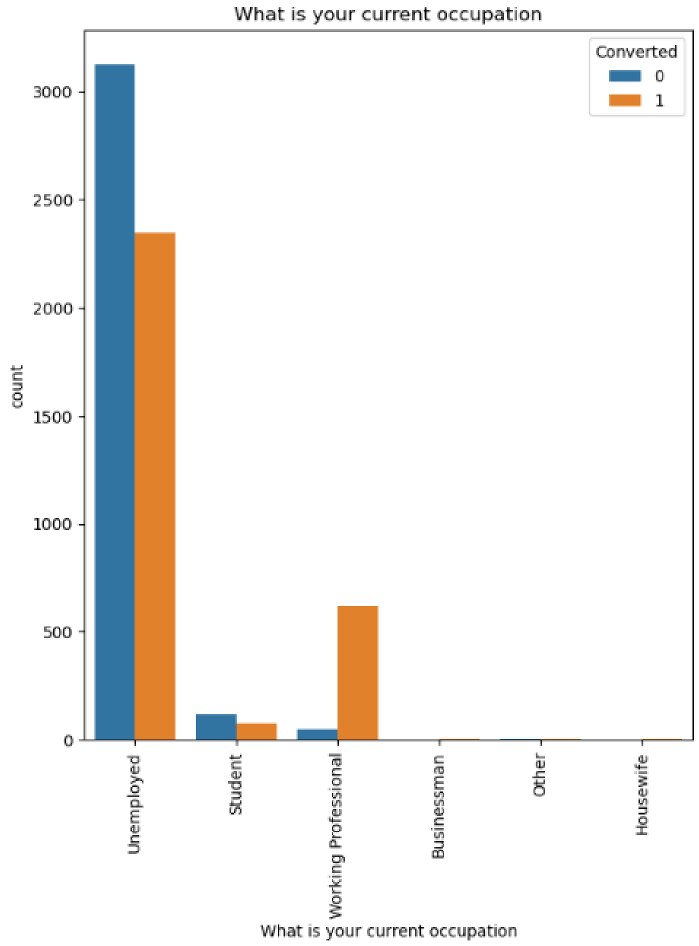
Specialization-

- Most of the leads have no information about specialisation.
- Whereas Finance Management, Banking, marketing management and Human Resource Management can be promising leads.

What is your current occupation-

- Working Professional opt for the course more whereas unemployed do not convert.





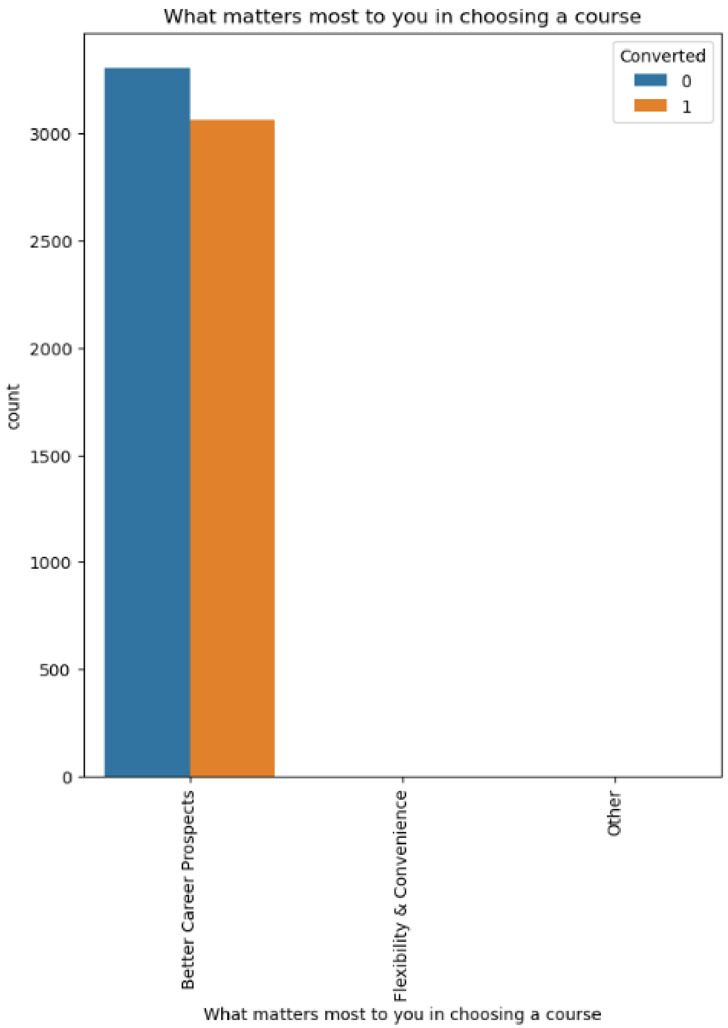
Visualisation

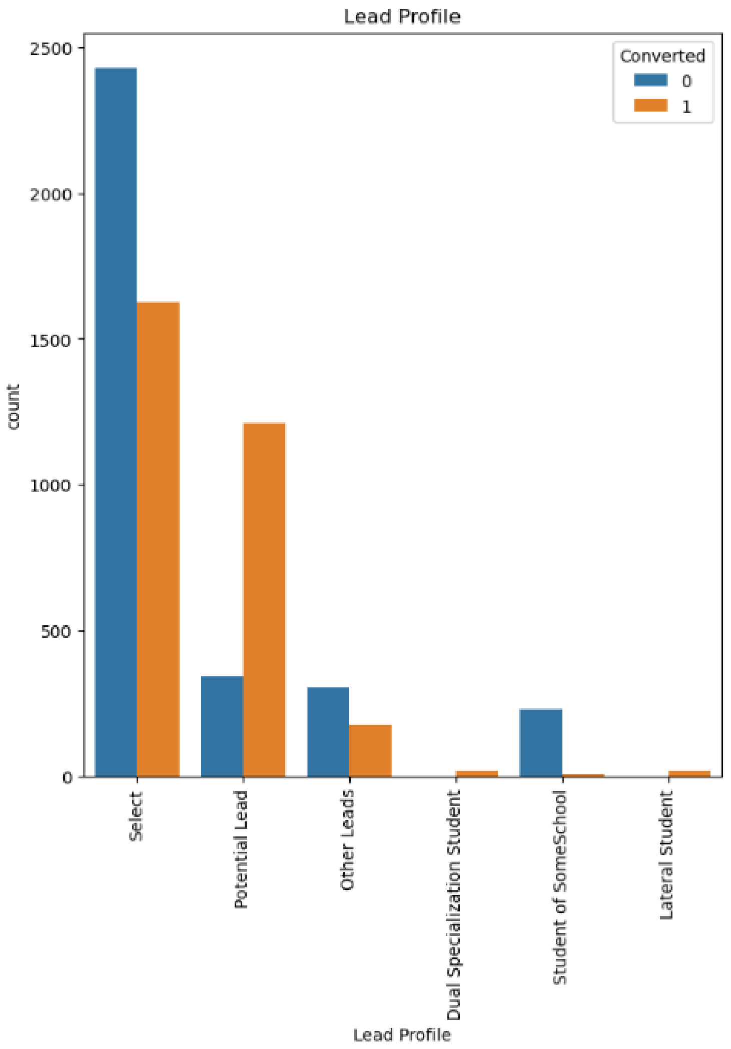
What matters most to you in choosing a course –

- Better Career Prospects has high non conversion rate.

Lead Profile-

- Potential lead is very important feature as it has high conversion rate which shows that the predicted potential leads are most of the time correct.
- Select that is null value has most of the data so most of the data is not present.





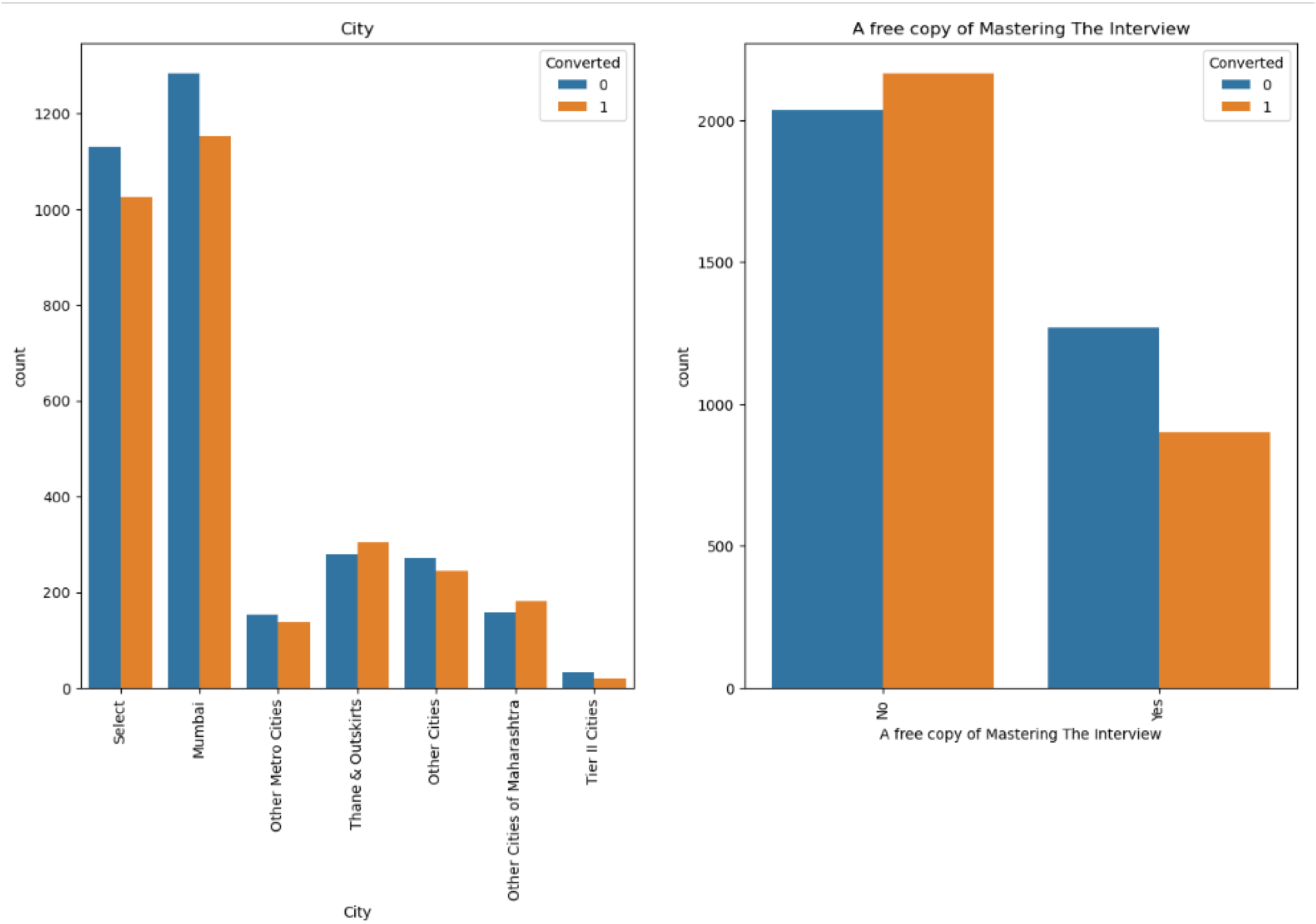
Visualisation

City-

- Thane and outskirts cities has high conversion rate.
- Whereas people from Mumbai don’t prefer taking the course.

A free copy of mastering interview-

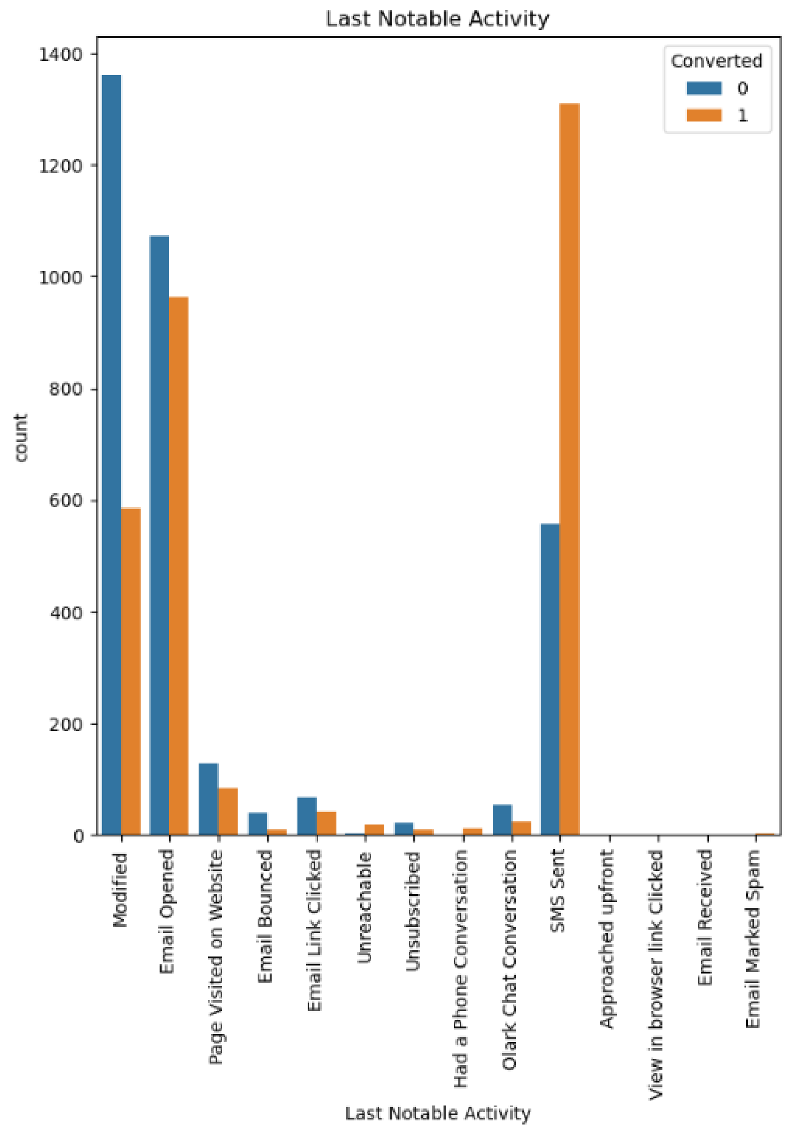
- People who have not asked for the copy have high conversion rate.



Visualisation

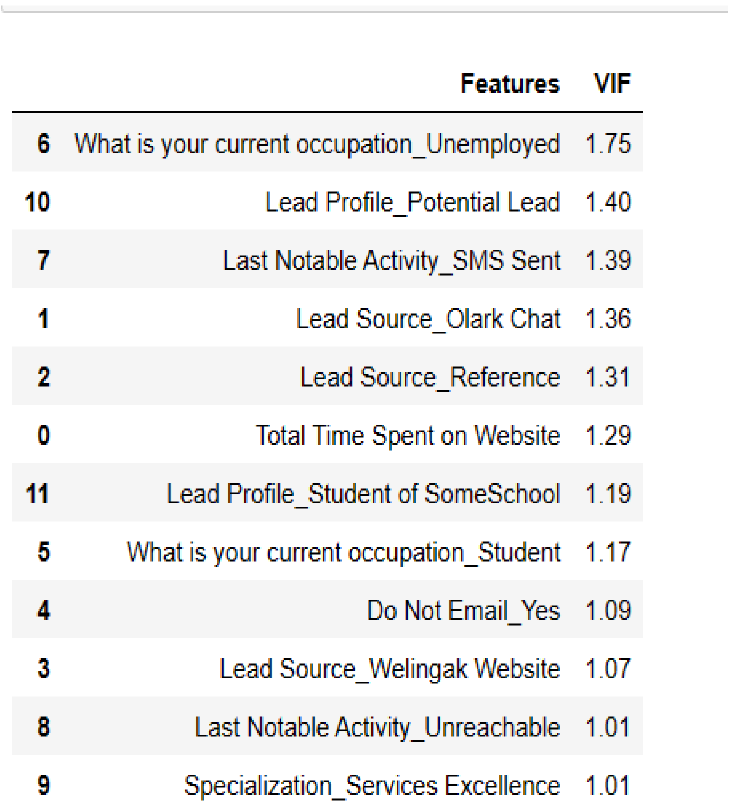
Last Notable Activity-

- People who received messages have very high conversion rate.
- Modified have very high rate of not conversion rate and Email opened also have less people who opted for the course.
- So messages can be a very effective way of converting the potential leads.



Model Building

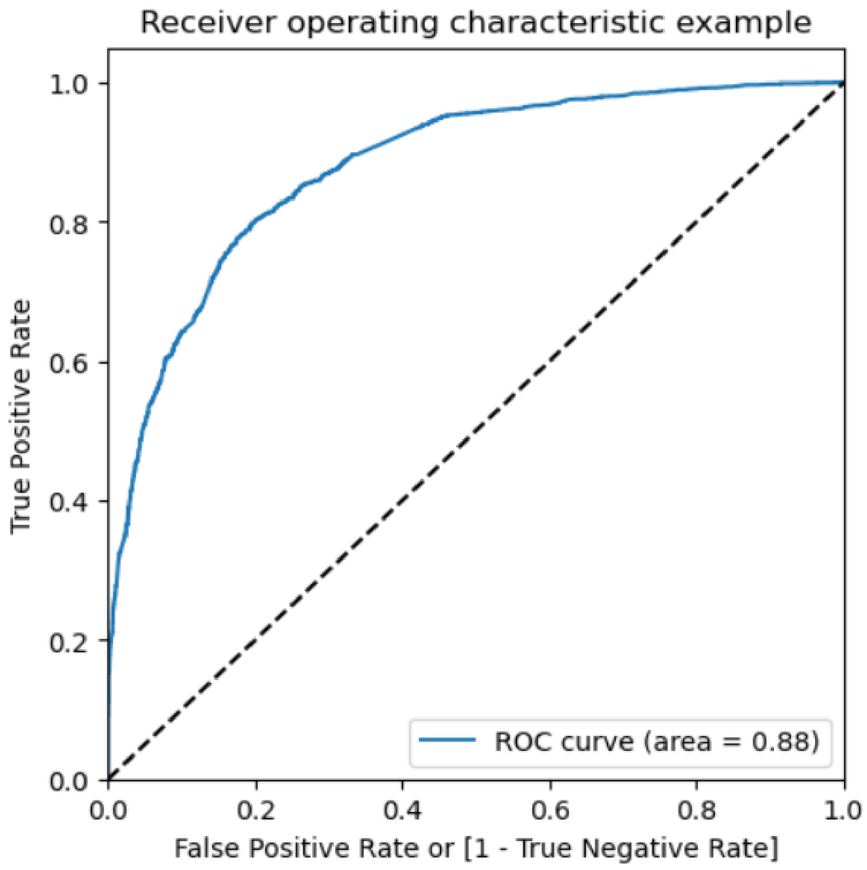
- Once the data is prepared for modelling, we do the train-test split where training data is 70% of the data and 30% is testing data.
- The model created with 15 variables using RFE was fitted using GLM().
- Then after checking the summary statistics and VIFs, the final model had 12 columns left.
- The model was evaluated with cutoff taken to be 0.5, the accuracy, sensitivity and specifivity were 0.80, 0.76 and 0.83 respectively.



	Features	VIF
6	What is your current occupation_Unemployed	1.75
10	Lead Profile_Potential Lead	1.40
7	Last Notable Activity_SMS Sent	1.39
1	Lead Source_Olark Chat	1.36
2	Lead Source_Reference	1.31
0	Total Time Spent on Website	1.29
11	Lead Profile_Student of SomeSchool	1.19
5	What is your current occupation_Student	1.17
4	Do Not Email_Yes	1.09
3	Lead Source_Welingak Website	1.07
8	Last Notable Activity_Unreachable	1.01
9	Specialization_Services Excellence	1.01

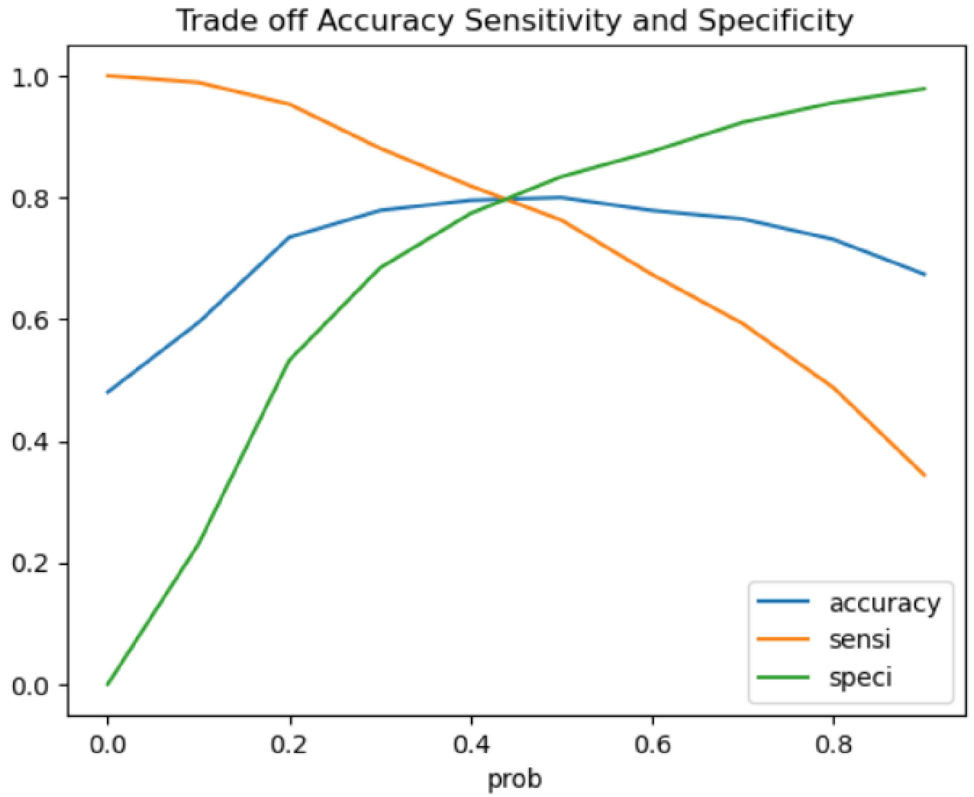
Optimal Cutoff by ROC curve

- The Receiver Operating Characteristics(ROC) Curve is the trade off between the sensitivity and specificity.
- Greater the area under the curve, the model is more accurate.
- Sensitivity is given by *TP/(TP+FN)*
- Specificity is given by *TN/(TN+FP)*
- Based on this curve we can see that near 0.4 can be the optimal cutoff but we will also plot tradeoff between accuracy, sensitivity ans specificity.



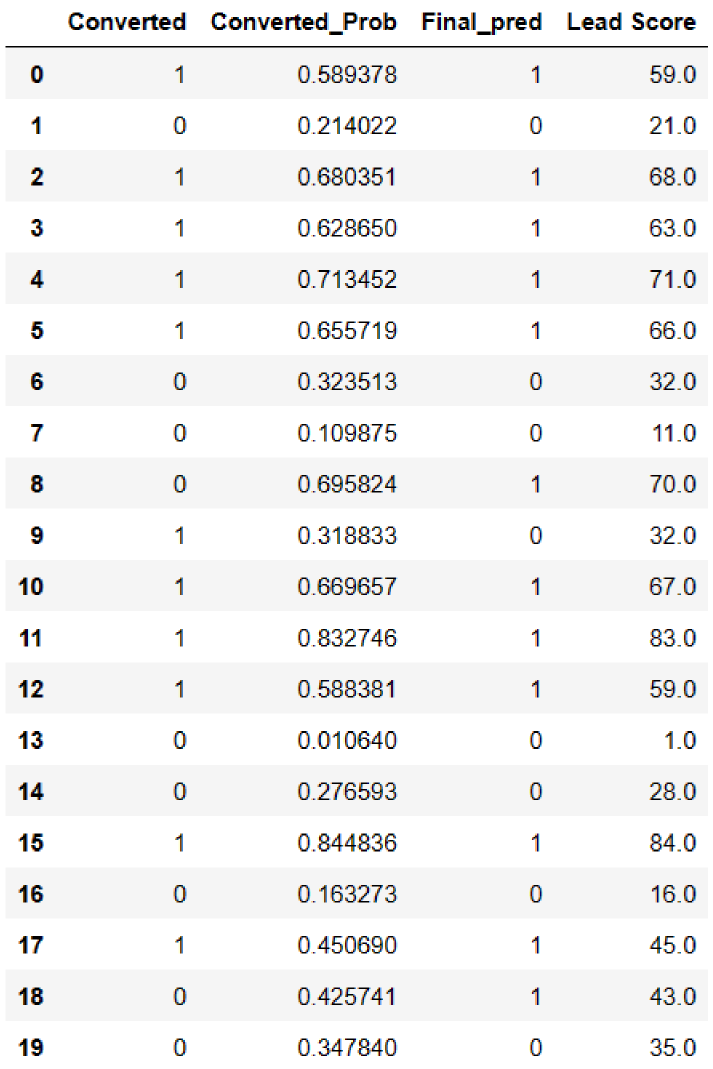
Optimal Cutoff

- Here the tradeoff is plotted by the points calculated for all the possible cutoff between 0 to 1 for accuracy, sensitivity and specificity .
- The intersection is at 0.42 in the line graph so approximately the optimal cutoff is taken to be around 0.4.
- Evaluating the model with cutoff as 0.4, the accuracy was 0.79, sensitivity and specificity was 0.81 and 0.78 respectively.
- Here we can see that the model is not overlapping the data which is good.



Test Model

- Now that the model is ready we predict on the test set, we fit the model using GLM() on test set.
- The accuracy of test set is 0.78 which is very close to the training set.
- Sensitivity = 0.81
- Specificity = 0.76
- Very similar to the training set.
- The lead conversion rate increased by 21% which previously was 30% and now is 51%.



	Converted	Converted_Prob	Final_pred	Lead Score
0	1	0.589378	1	59.0
1	0	0.214022	0	21.0
2	1	0.680351	1	68.0
3	1	0.628650	1	63.0
4	1	0.713452	1	71.0
5	1	0.655719	1	66.0
6	0	0.323513	0	32.0
7	0	0.109875	0	11.0
8	0	0.695824	1	70.0
9	1	0.318833	0	32.0
10	1	0.669657	1	67.0
11	1	0.832746	1	83.0
12	1	0.588381	1	59.0
13	0	0.010640	0	1.0
14	0	0.276593	0	28.0
15	1	0.844836	1	84.0
16	0	0.163273	0	16.0
17	1	0.450690	1	45.0
18	0	0.425741	1	43.0
19	0	0.347840	0	35.0

Conclusion

- The final model had 15 columns of high relevance after performing EDA.
- The model showed high accuracy of 78%.
- The optimal cutoff was found to be 0.4 using the ROC Curve and tradeoff between accuracy, sensitivity and specificity.
- The model correctly finds the most of the high leads and the leads that are not as important.
- Overall the model is accurate.
- Total time spent on the website can have an impact on lead conversion as people who more than average time spent on the website can be hot leads.
- Last Notable Activity and what matters most when choosing a course is also good indicator of lead conversion as it tells more about the mindset of the student.
- Marketing managers and human resources management have high conversion rate.
- References and offers for referring a lead show high conversion rate.

Insights and Results

- Landing page submissions has high conversion.
- The people who have got the information from the source google are mostly likely to convert so we can initiate advertisements and pop ups on google.
- Most leads do not prefer getting emails about the course but leads who are converted by the sms messaging have high lead conversion rate.
- Thus sms messages must be focused on so that the sms messages reach maximum audience which will increase the number of potentials leads.
- Most of the leads have no information about specialisation.
- Unemployed sector do not convert a lot so we can create enrolment offers and referral discounts to attract the leads to the course .

Insights and Results

- Potential lead is very important feature as it has high conversion rate which shows that the predicted potential leads are most of the time correct.
- So the feature potential lead has a very high accuracy in identifying the leads who will most likely convert.
- Mumbai has low conversion rate so we can focus on this city more while targeting the customers so that we can the conversion rate in that region.
- Region like metro cites and tier 2 cities have very few number of people who have explored about the course so we can maybe increase our advertisement of the course in this region which will help in increasing the leads and later check if the response is still not good then we can drop these from our list.

Thankyou