Lead Score Case Study

By: Ojen Shrestha

Table of Content

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
- The typical lead conversion rate at X education is around 30%. The typical lead conversion rate at X education is around 30%. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- The company requires you to build a model wherein you 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%.

Dataset

- The dataset give to us is Leads dataset(Leads.csv)
- In this dataset we have 9240 rows and 37 columns.
- In this dataset ‘Converted’ is our target variable.

Dataset

- First, we started looking at the dataset using info, describe, head
- We saw that there are few column values as 'Select' for the ones which the customers have not selected any values, so we converted these values to null.
- Then we checked the missing/null values in each column
- We handled all missing value column one by one.

```
Prospect ID                                      0.00
Lead Number                                      0.00
Lead Origin                                      0.00
Lead Source                                      0.39
Do Not Email                                     0.00
Do Not Call                                      0.00
Converted                                        0.00
TotalVisits                                      1.48
Total Time Spent on Website                      0.00
Page Views Per Visit                             1.48
Last Activity                                    1.11
Country                                         26.63
Specialization                                  36.58
How did you hear about X Education              78.46
What is your current occupation                 29.11
What matters most to you in choosing a course   29.32
Search                                           0.00
Magazine                                         0.00
Newspaper Article                                0.00
X Education Forums                               0.00
Newspaper                                        0.00
Digital Advertisement                            0.00
Through Recommendations                          0.00
Receive More Updates About Our Courses           0.00
Tags                                            36.29
Lead Quality                                    51.59
Update me on Supply Chain Content                0.00
Get updates on DM Content                        0.00
Lead Profile                                    74.19
City                                            39.71
Asymmetrique Activity Index                     45.65
Asymmetrique Profile Index                      45.65
Asymmetrique Activity Score                     45.65
Asymmetrique Profile Score                      45.65
I agree to pay the amount through cheque         0.00
A free copy of Mastering The Interview           0.00
Last Notable Activity                            0.00
dtype: float64
```

Data Cleaning

- First, we started with ‘Country’ column. Here found out that almost all the values are ‘India’ we replace the null values to India
- Next, we looked at ‘Specialization’. We found out that the missing value count was a bit higher so we replaced the missing value as a new category called ‘Others’.
- Next, we looked at ‘How did you hear about X Education’ column which had 78% missing value so we can drop this column.
- Next column we have is 'What is your current occupation' with 29% missing value. In this column ‘Unemployed’ has the highest value count so we replace the null values to ‘Unemployed’.
- Next, we looked into the column 'What matters most to you in choosing a course' which has 29% missing value. Here almost all value is 'Better Career Prospects', so we replaced null values to the same.

Data Cleaning

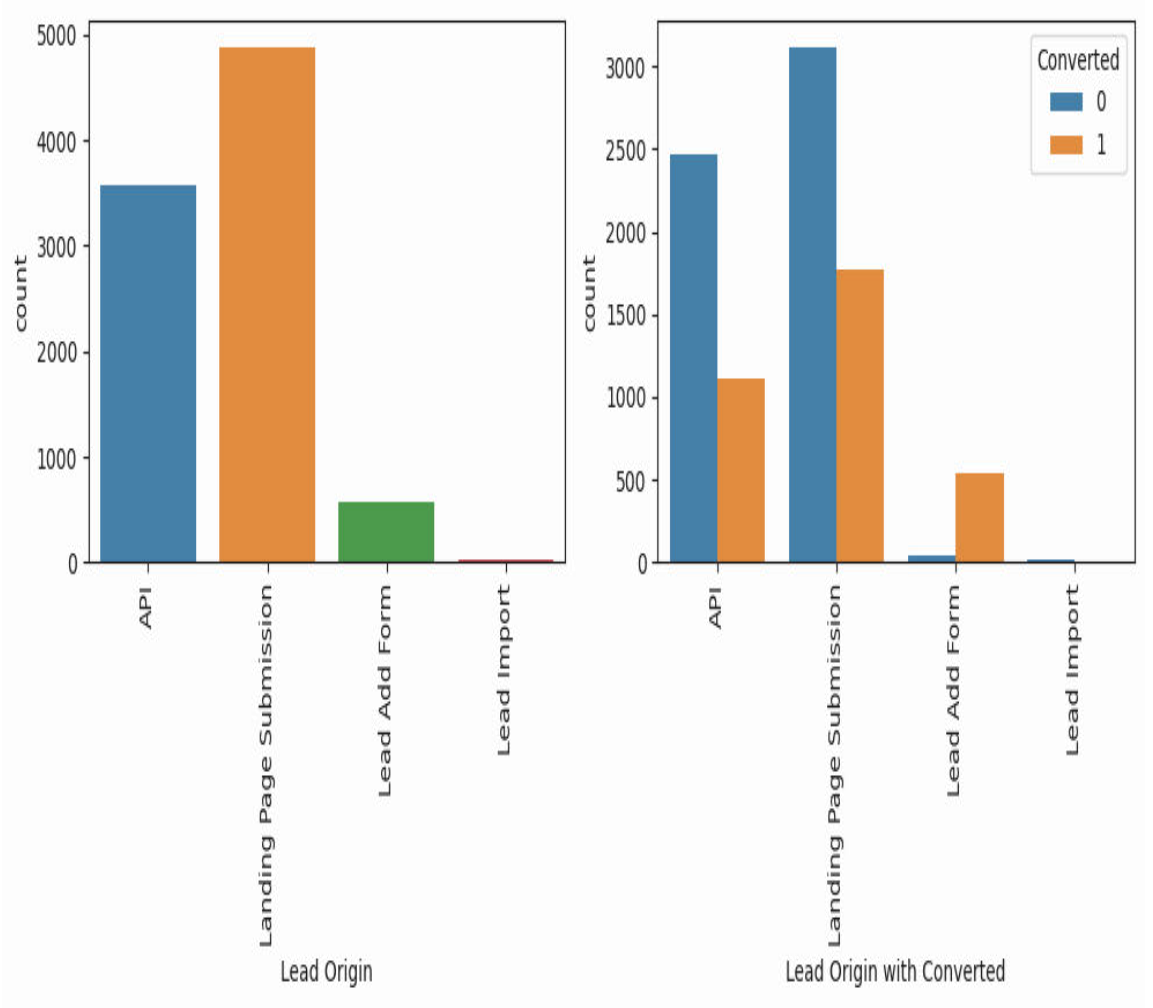
- Next, missing column was ‘Tags’ which had 36% missing values. There was a high amount of missing value to replace with any other category so we replaced with new category ‘Others’.
- Next, we looked at column 'Lead Quality' which had 51% missing value. Here there was a high amount of missing value to replace with any other category so we replaced with new category ‘Others’.
- Next, we looked at column 'Lead Profile' which had 74% missing value. The missing value percent was very high so we had to drop this column.
- Next, we looked at column 'City' which had 39% missing values. According to the spread of above data we can see the ‘Mumbai’ has the highest count, so we replaced it with Mumbai.
- Next we had 4 columns 'Asymmetrique Activity Index','Asymmetrique Activity Score','Asymmetrique Profile Index','Asymmetrique Profile Score’ with 45 % missing value. In these columns missing values are high plus there is variation in data, so we cannot replace the values so we had drop the columns.
- Now at least we had columns with very less <2 % missing values, so for these we just dropped the rows as it wont have any significant difference.
- We completed handling missing values then we moved to EDA.

Exploratory Data Analysis

- First, we checked the ‘Converted’ ratio which 37.8% which is pretty decent.
- Now we will look into each column one by one and analyze the column.

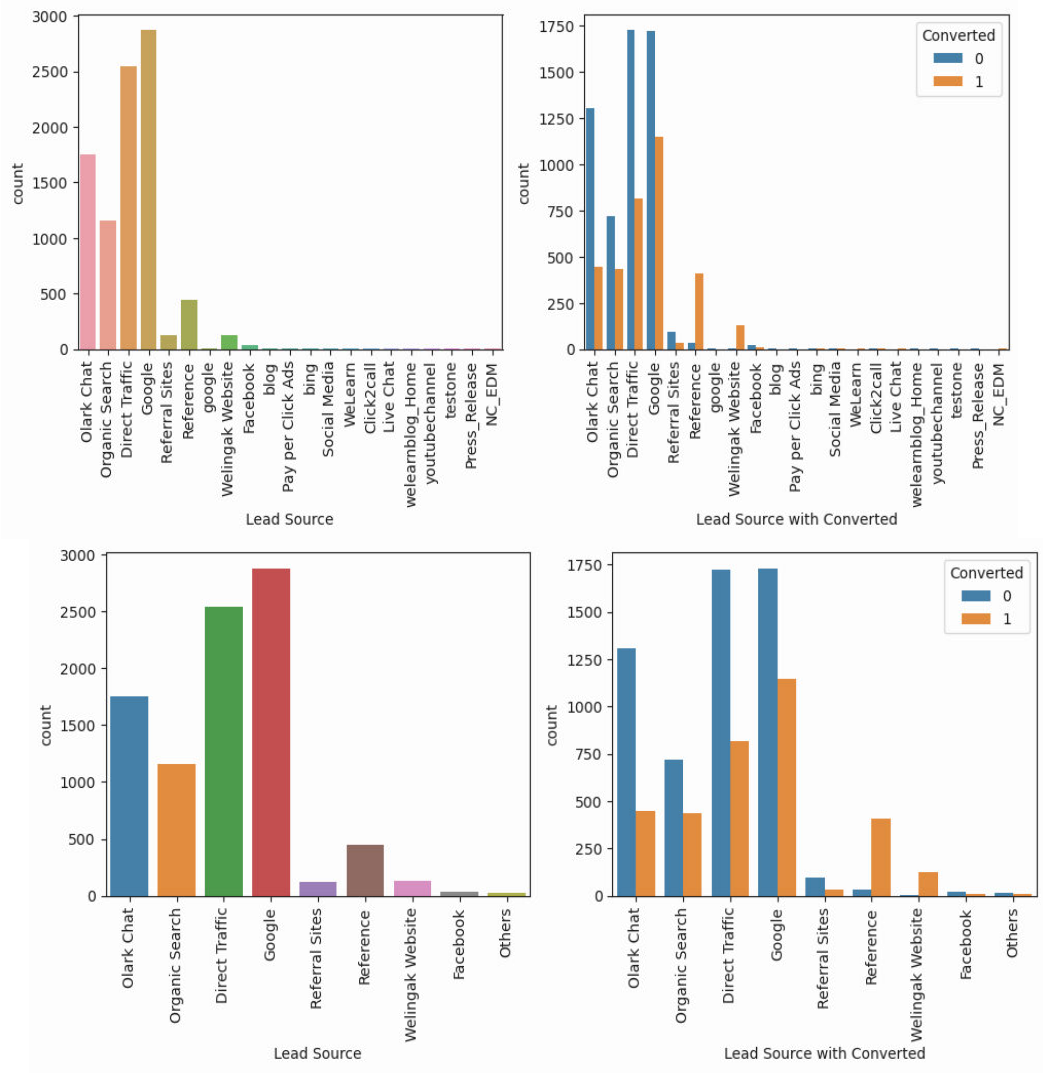
Exploratory Data Analysis

- **Lead Origin:**
  - We can see that for API and Landing Page Submission the conversion rate is around 40%-50%.
  - Lead Add Form has very less data but almost 90% of data is converted.
  - Lead Import has very less amount of data.
  - So, we need to increase the conversion rate of API and Landing Page Submission and generate from leads from Lead Add Form



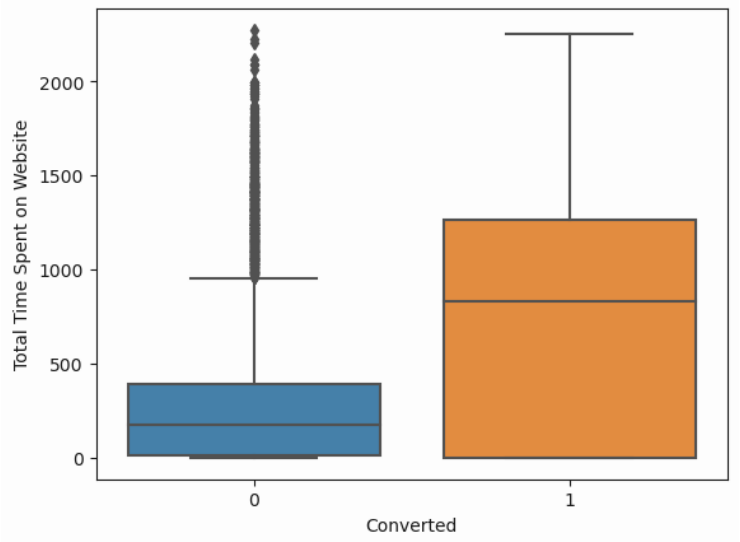
Exploratory Data Analysis

- **Lead Source:**
  - Here we can see that, first there is more data points so we need to combine few and also Google is written twice 'google' and 'Google', so we need to fix this.
  - We can see the google has the highest Lead Source and around 70% of Google Lead Source are converted.
  - Next Direct Traffic also has high Lead Source but its conversion rate is little less compared to Google, approx. 40%-50%
  - We can see the Reference conversion and Welingak Website is approx. 90%
  - In order to increase the conversion more Reference, Welingak Website to increase their Lead Source and and try to increase conversion rates of google, direct traffic, olark chat, organic search



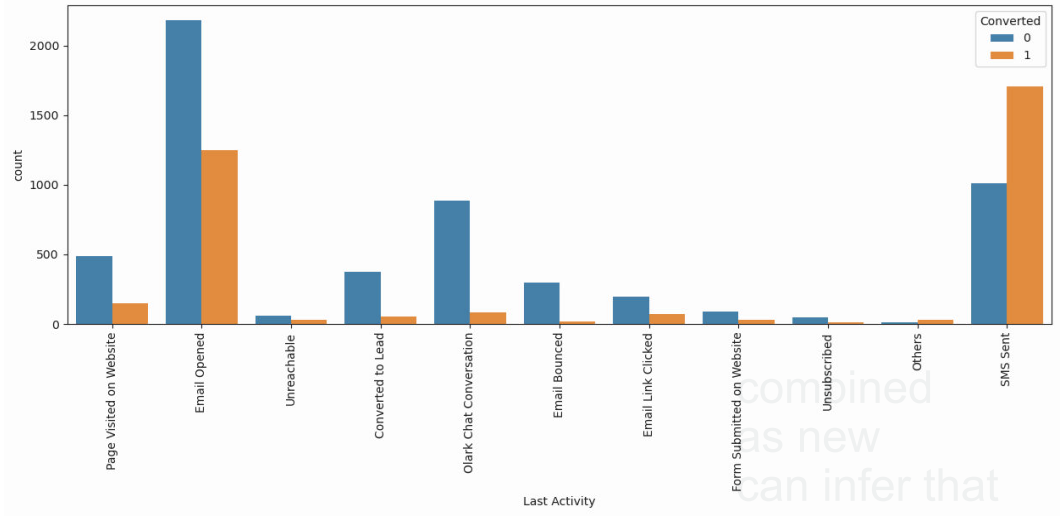
Exploratory Data Analysis

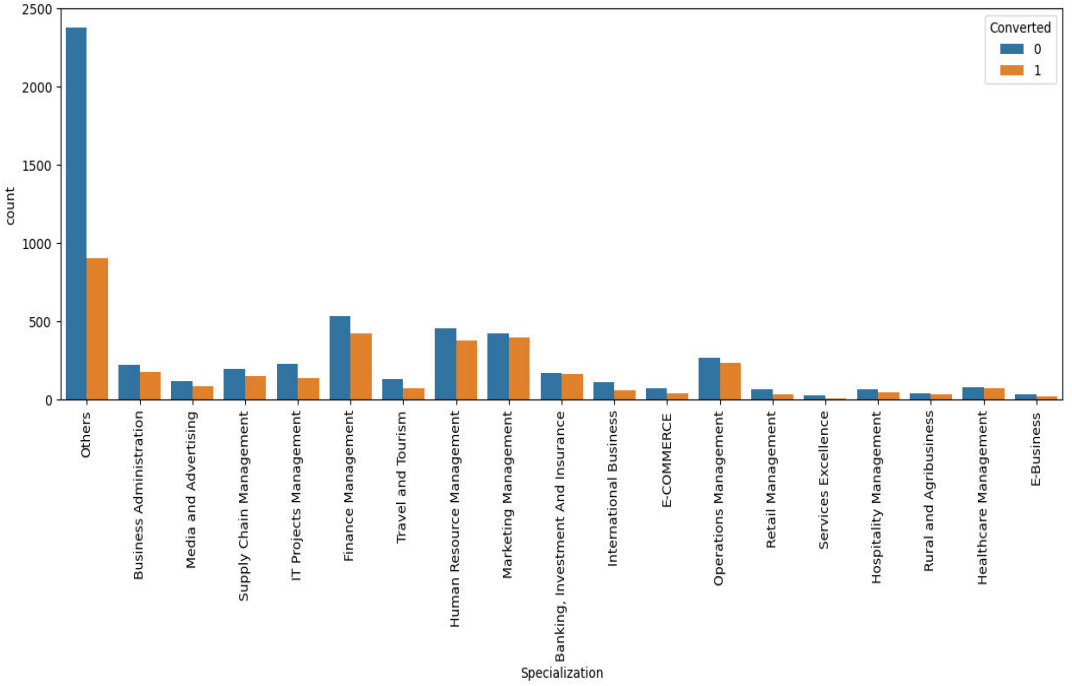
- **Do Not Email and Do Not Call:** Here we saw that 'No' volume is high in both columns and the conversion rate is approx 70%.
- **TotalVisits:** This column had a lot of outliers so we took only from 5% to 95% and visualized a boxplot and found the median of Converted as ‘0’ and ‘1’ were same for this column.
- **Total Time Spent on Website:** In the graph we can see that there is more conversion rate if you spend more time on the website. Therefore, the company should build a more appealing website.



Exploratory Data Analysis

- **Last Activity:** In this column we have a lot of categories with very less data, so we these into one for better analysis category ‘Others’. From the graph we there is alot of email opened has the highest conversion rate
- **Country:** 96% of the data is from India
- **Specialization:** This column spread in different categories. From the second graph 'Finance Management','Human Resource Management’, Insurance’, category so we can

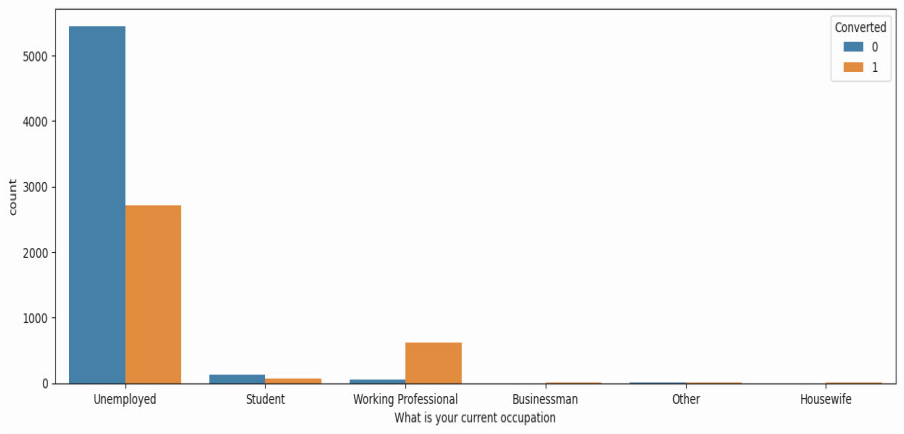


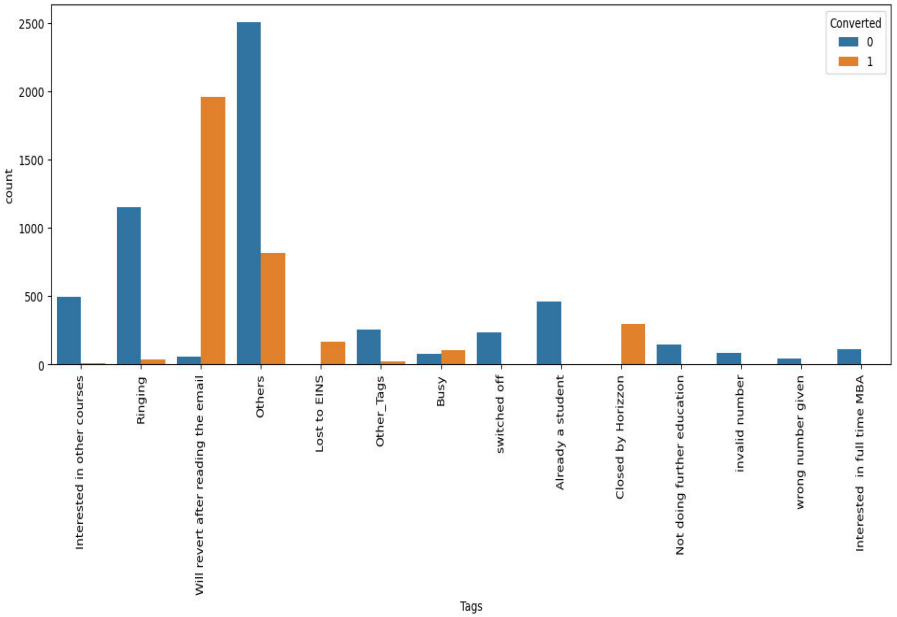


Exploratory Data Analysis

What is your current occupation: From the first graph we can see that there is very high volume of Unemployed category with around 50% conversion rate and Working Professional have a very high conversion rate.

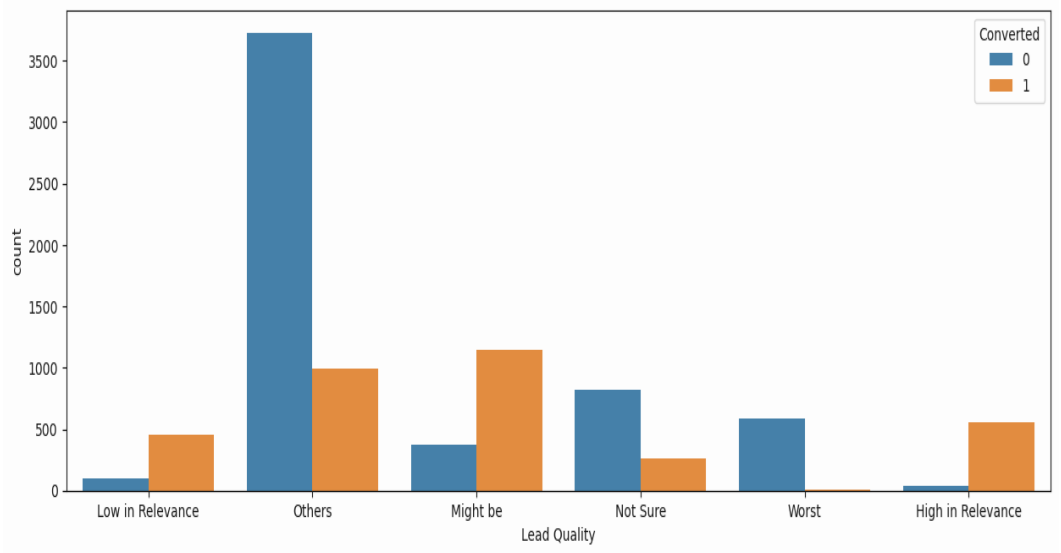
Tags: Here we could see a lot of small categories so we combined these columns as ‘Other_tags’. From the second graph we can infer that 'Interested in other courses','Ringing','Switched off','Already a student' has a very less conversion rate. Will revert after reading the email has a very good amount of data plus the conversion rate is also high

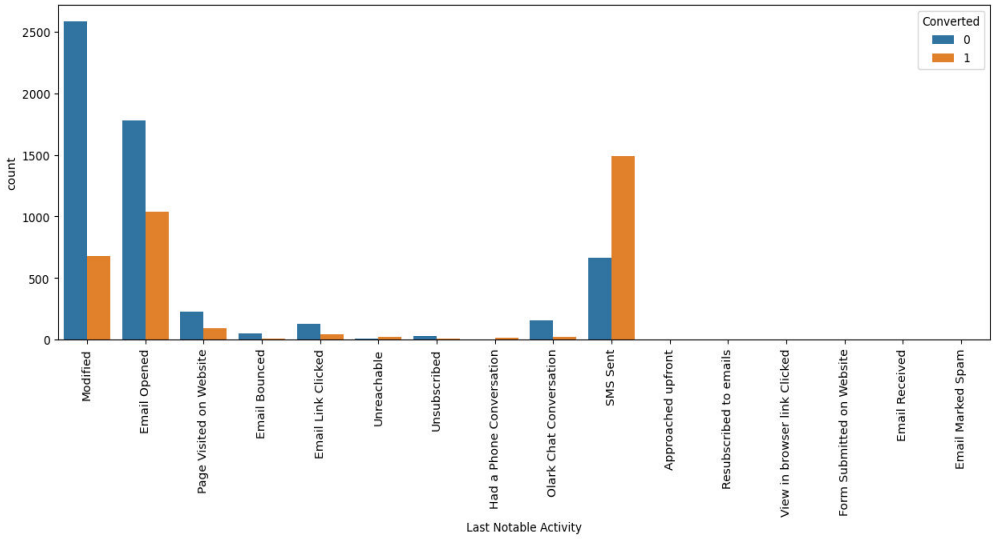




Exploratory Data Analysis

- **Lead Quality:** From the first graph we can infer that both 'Low in Relevance' and 'High in Relevance' has a high converison rate. 'Might be' category also has a good conversion rate.
- **Last Notable Activity:** From the second graph we can infer that Modified has the highest volume with very less conversion rate. Second Email Opened is also high with approx. 50% conversion rate. SMS Sent has the highest conversion rate.





Exploratory Data Analysis

- Now we have few other columns like ‘What matters most to you in choosing a course’, ‘Search’, ‘Magazine’, ‘Newspaper Article’, ‘X Education Forums’, ‘Newspaper’, ‘Digital Advertisement’, ‘Through Recommendations’, ‘Receive More Updates About Our Courses’, ‘Update me on Supply Chain Content’, ‘Get updates on DM Content’,’I agree to pay the amount through cheque’, ‘A free copy of Mastering The Interview’, ‘City’ which does not have a proper spread of data. These columns categories is focused on just one particular category, therefore these columns does not provide much information while model building, so we dropped these columns.

Data Preparation

- First, we mapped columns 'Do Not Email', 'Do Not Call' to ‘0’ and ‘1’.
- Next, we created dummies for our categorical variables and concatenated to the data and dropped the original categorical columns.
- Next, we assigned the target variable ‘Converted’ to ‘y’ and rest of columns excluding ‘Prospect ID’ to ‘X’
- Next we created a train test split of the data with 70% of data as train set and 30% as test set.
- Next we scaled our numerical variables using StandardScaler.

Model Building

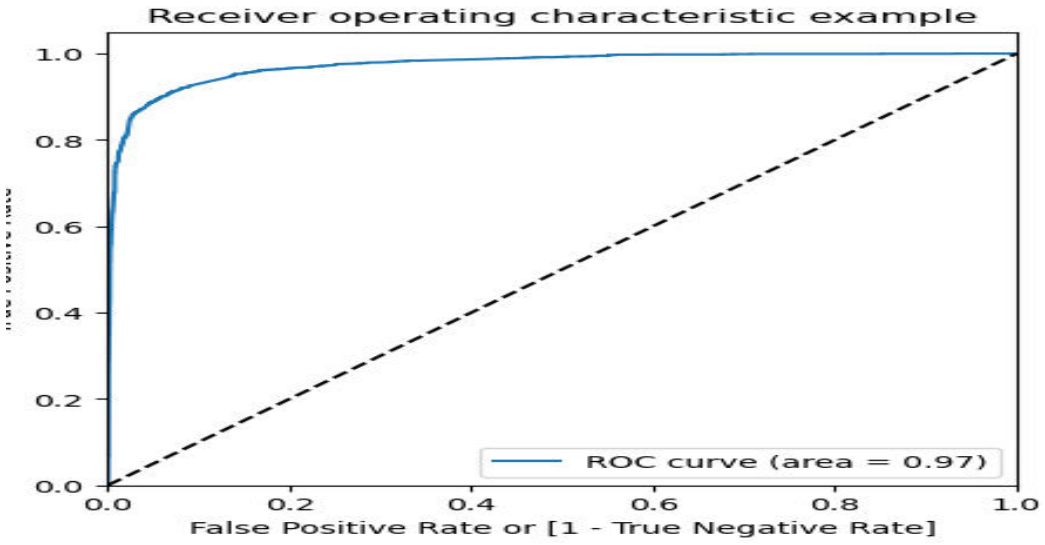
- We used Logistic Regression and for feature selection we used RFE. We took 20 columns with the help of RFE.
- Then we used stats model to analyze the how the columns were doing in our model by looking at the p-value. We found few columns whose p-value was greater than 0.05 so we dropped these columns and again created model.
- Next, we checked the VIF of the columns and found a column whose VIF was greater than 5 so we dropped the column and again created the model. After this the model looked good so we started predicting our y_train

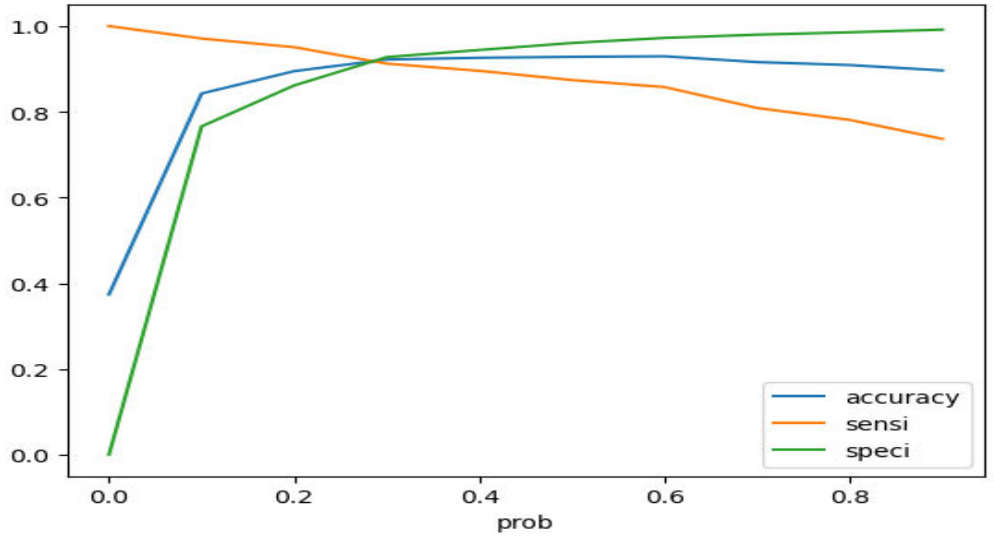
Model Building

- After predicting we found the probability of Converted. At first, we took a random cut-off of 0.5 to predict Converted.
- With this we got the accuracy of our model to be 92% and confusion matrix as
```
[[3820  157]
 [ 299 2075]]
```
- Next, we calculated new other metric like:
  - Specificity: 0.96
  - Sensitivity: 0.87
  - False Positive Rate: 0.03
  - Positive Predictive Value: 0.929
  - Negative Predictive value: 0.927

Model Building

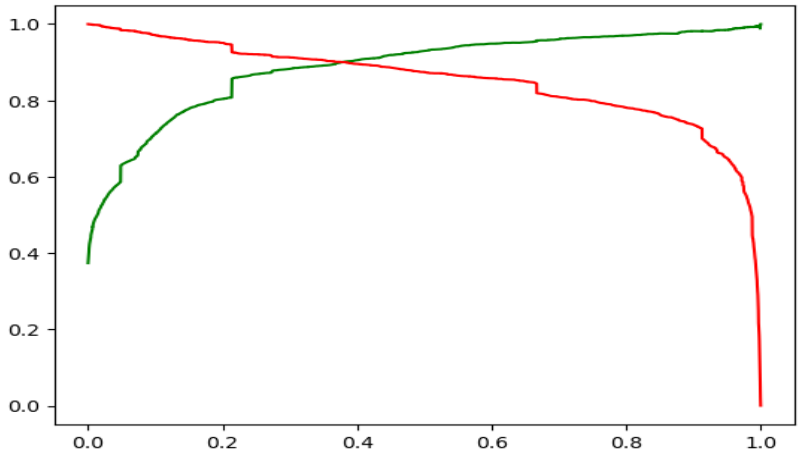
- Next, we created the ROC curve. We know that in ROC Curve that greater the area under the curve the better.
- Next, we found the optimal threshold (cut-off) by calculating accuracy sensitivity and specificity for various probability cutoffs. From the second curve,0.3 is the optimum point to take it as a cutoff probability. After this we look 0.3 as the cut-off and calculated the new predicted value.
- Next we calculated the Lead Score





Model Building

- Next, we calculated the metric again whose score came out to be quite similar the previous analysis. Accuracy 92%, confusion metric `[[3689, 288], [ 208, 2166]]`, Specificity: 0.927, Sensitivity:0.91, False Positive Rate:0.07, positive predictive value: 0.88, Negative predictive value:0.94
- Next, we calculated the Precision and Recall values where in we got Precision as 0.88 and Recall as 0.91
- Next, we calculated the Precision and Recall Trade-off.



Making Prediction on Test Dataset and Model Evaluation

- First, we scaled transformed the columns, then added the constant and predicted the y_test. We got the probability of Converted and we used 0.3 as cut-off to calculate the final prediction
- Next, we calculated the accuracy which came out to be 92% and confusion matrix
- Next, we calculat[[1540, 122], [83, 978]]sitivity: 0.921 and Specificity: 0.926
- Overall, we have created a good model with 92% accuracy with 0.3 cut-off

Conclusion

- We have built a logistic regression model and assigned 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.
- The top three variables in your model which contribute most towards the probability of a lead getting converted are Total Time Spent on Website, Lead Source, Lead Origin.
- The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are Last Notable Activity_SMS Sent, Do Not Email, Last Notable Activity_Modified