Lead Scoring Case Study

-Ishan Singh Rawat

DSC-47

Assignment for upGrad & IIIT Bangalore

Problem Statement

- X Education sells online courses to industry professionals.
- The company markets its courses on several websites and gets a lot of leads, 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% which 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.

Business Objective

- X Education wants the information of the most promising leads.
- The company requires a model to be built wherein a lead score is assigned 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..
- ballpark of the target lead conversion rate is to be around 80%. which is set by CEO, in particular.

Methodology for Problem Solving

- Loading Datasets
- Data Understanding
  1. Checking Structure of datasets
  2. Checking other attributes like using info(), describe(), shape etc
- Data Cleaning
  1. Deleting all the Redundant columns and rows
  2. Analyzing all the missing values, improper data types, duplicated rows
  3. Imputing all the Missing values
  4. Rectifying the improper data types
- Outliers treatment
- Data Analysis (EDA)
  1. Imbalance data Analysis
  2. Defining functions for plotting
  3. Univariate analysis of each variable

Data Preparation for Model creation

1. Dummy Variable Creation
2. splitting the data into Train Test
3. Scaling feature
4. checking for correlations in data modelling

Model Evaluation

1. Calculating all the important metrics such as

   (sensitivity, specificity, recall, precision, F1-score, etc.)
2. Plotting ROC curve
3. Plotting precision and recall trade off
4. Finding optimal cut off probability
5. Making predictions using test dataset

Final results and Summary

Data Cleaning :

- Initial step was to find the duplicate values we found none
- Next we checked for missing values in rows and kept a threshold of 35% since many columns were having more than 40% of missing data
- Columns having missing values more than 35% were dropped since imputing these columns would certainly deprecate the data quality

data have

- For columns having less than 2% missing data we have dropped the rows, after which 98% of total data was retained.
- Since all the numeric columns had less than 2% data so we have handled them during the above process.
- The remaining categorical columns, which had null values between 15% - 30% of missing values but also were highly skewed data was also dropped.
- Those Columns not having highly skewed data were imputed using the modal value of that column.
- Columns having incorrect data type were also rectified.

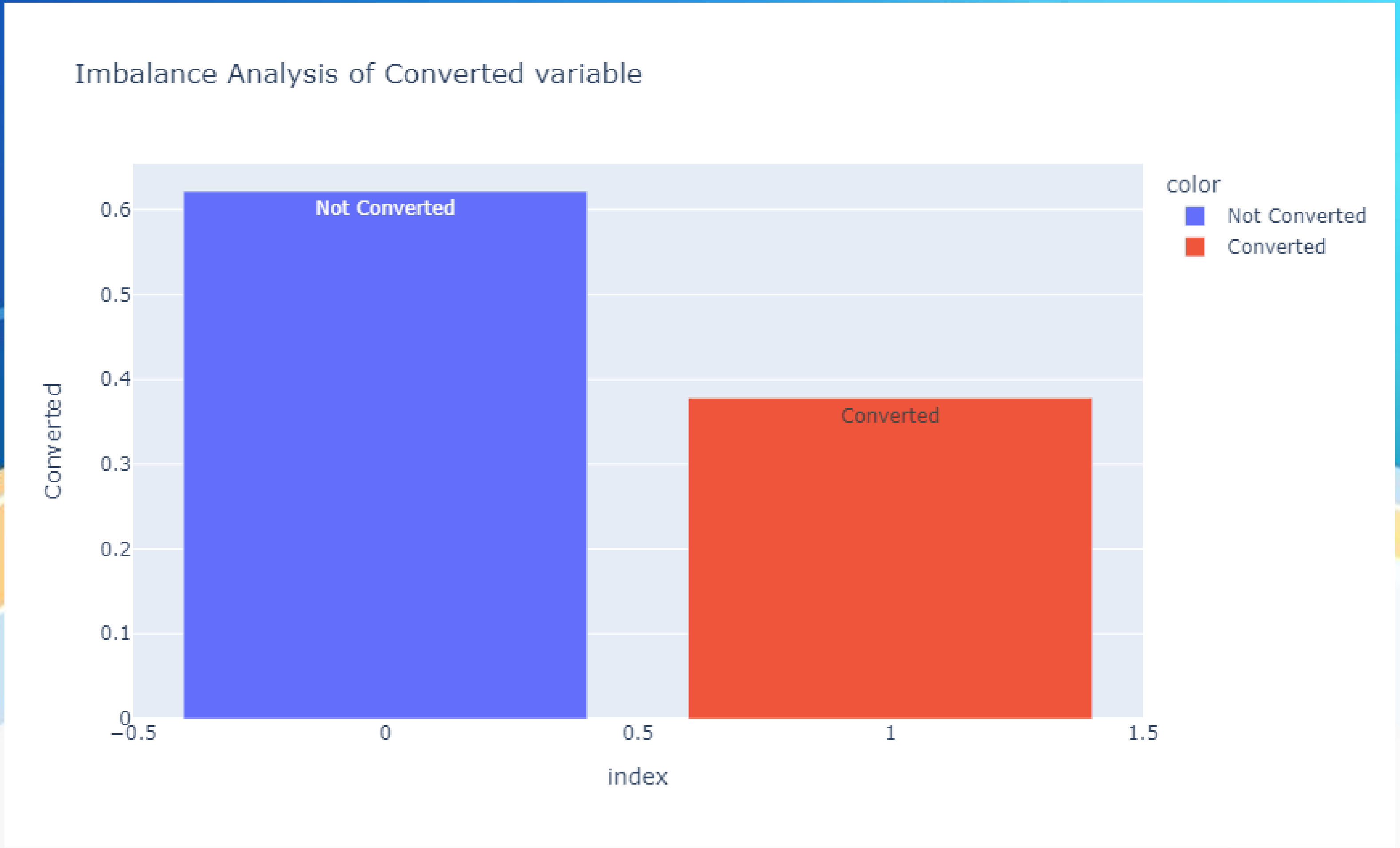
values have
rows)

Data Analysis (EDA)

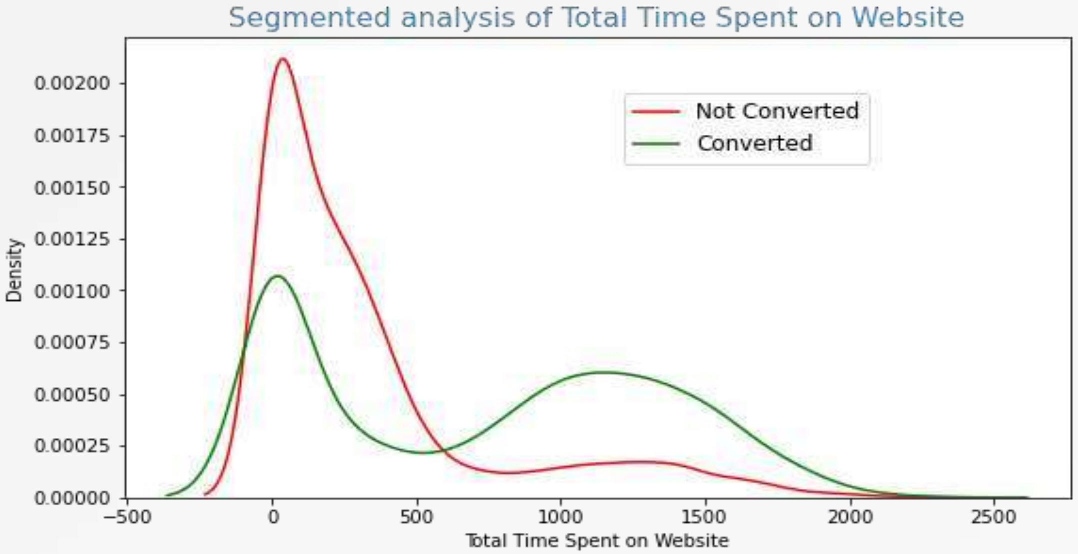
1. Analysis Of Imbalance for Converted Variable
2. Analysis Of Numerical Variables With Respect To Converted Variable
3. Analysis Of Categorical Variables With Respect To Converted Variable

Imbalance Analysis

Only 38% Of Total Leads Have Been Converted Whereas 62% Are Not Converted

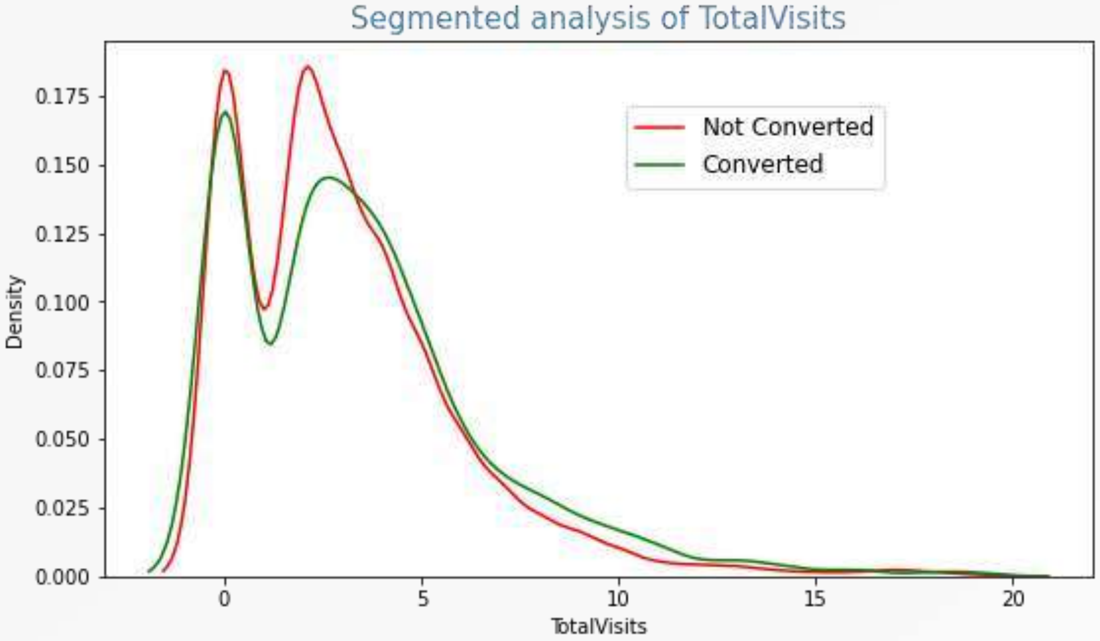


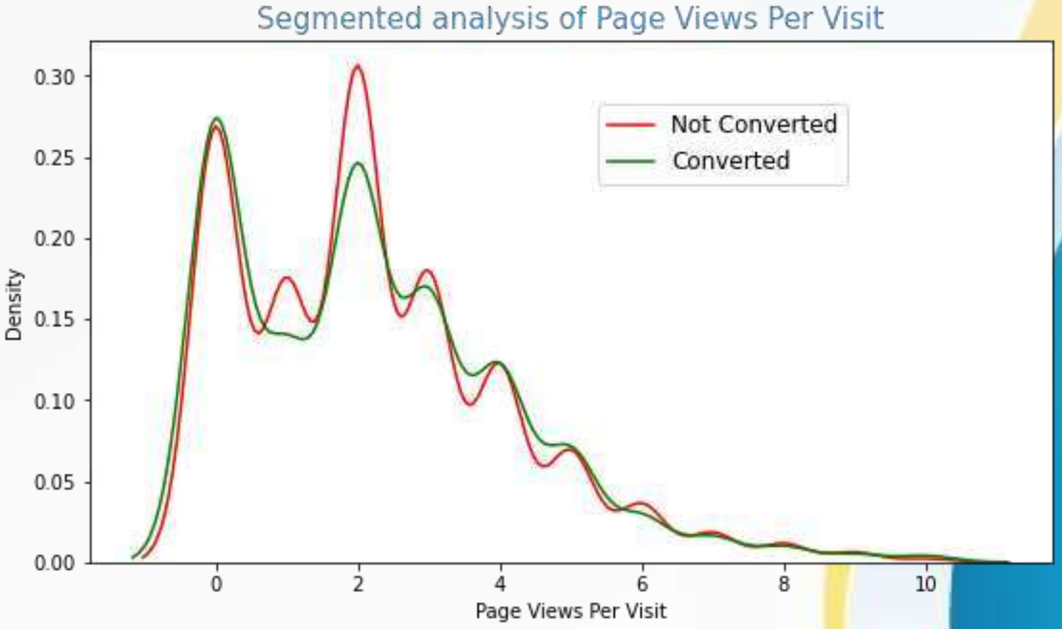
Numerical Variables



Analysis:

1. Those customers who spend more than 10 mins on the website are more likely to be converted leads than those who spends less than 10 mins surfing the website.
2. In case of other numerical features such as(`TotalVisits` and `Page Views Per Visit`), there is no proper distinction between converted and not-converted customers, but we can say that, those customers who view more pages as well as visit the website more times are likely to be converted leads

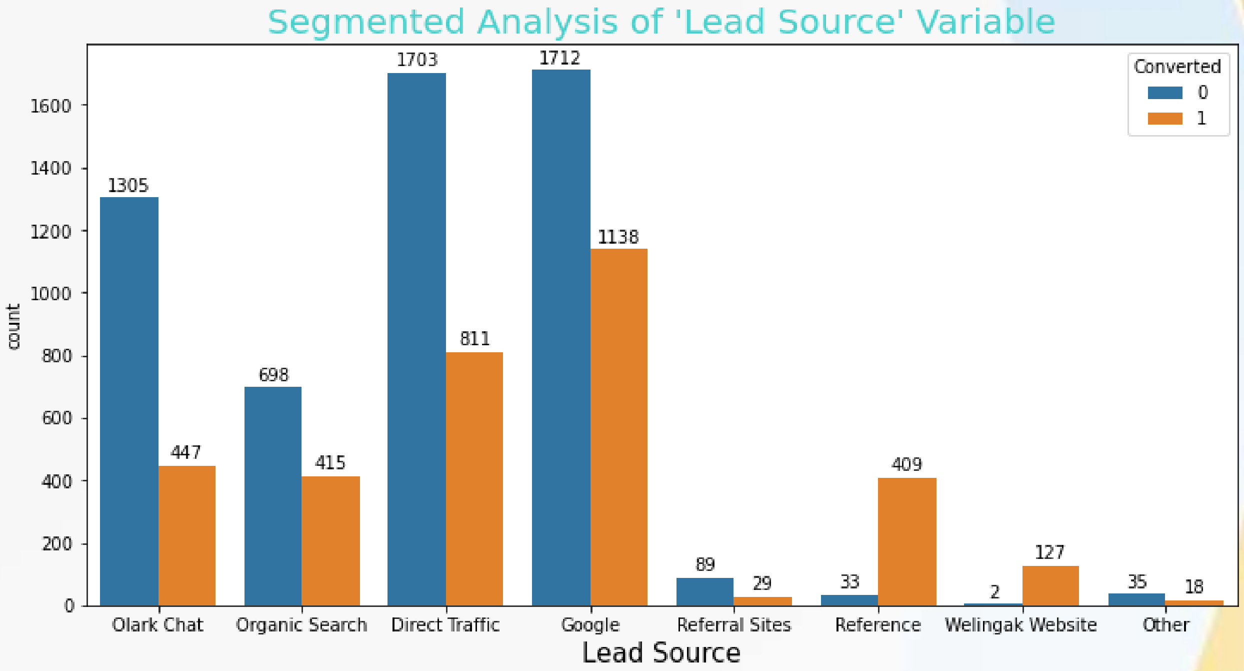


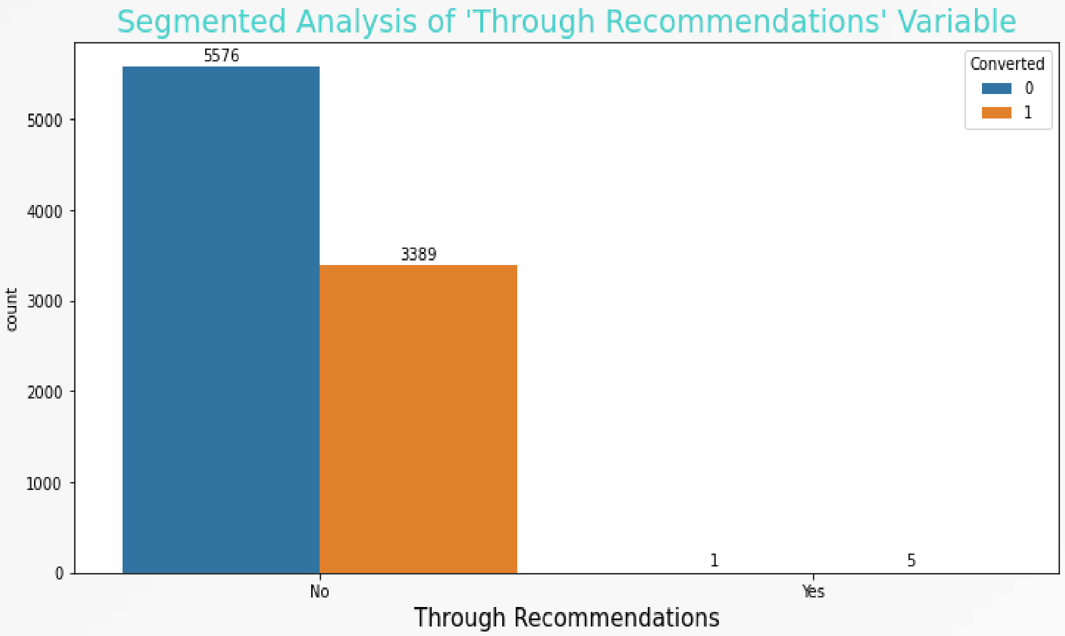


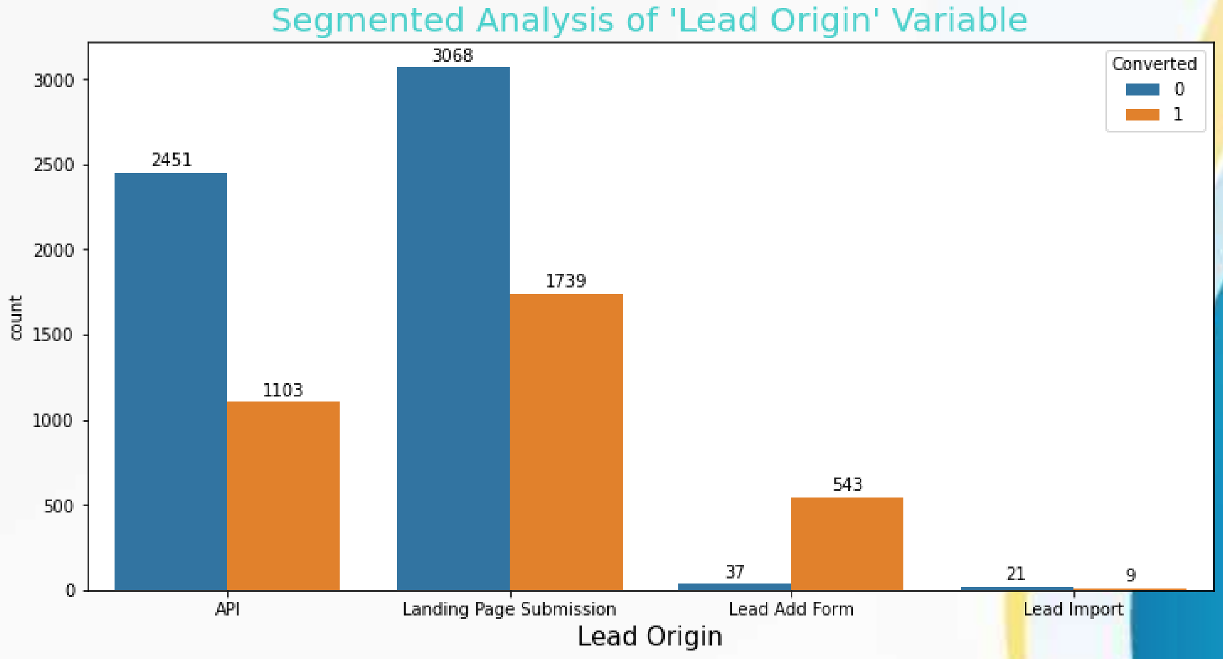
Categorical Variables

Analysis

- Leads which are generated from `reference`, `Welingak Website` and `Lead Add Form Lead Origin` can be considered as hot leads as the conversion rate is more than 90%
- From recommendation of other customers there is high probability of becoming converted lead of due to high conversion rate and can be termed as hot leads.
- In the column of `Through Recommendations’ the `Yes` category although having less entries, have highest conversion rate,which is more than 83%





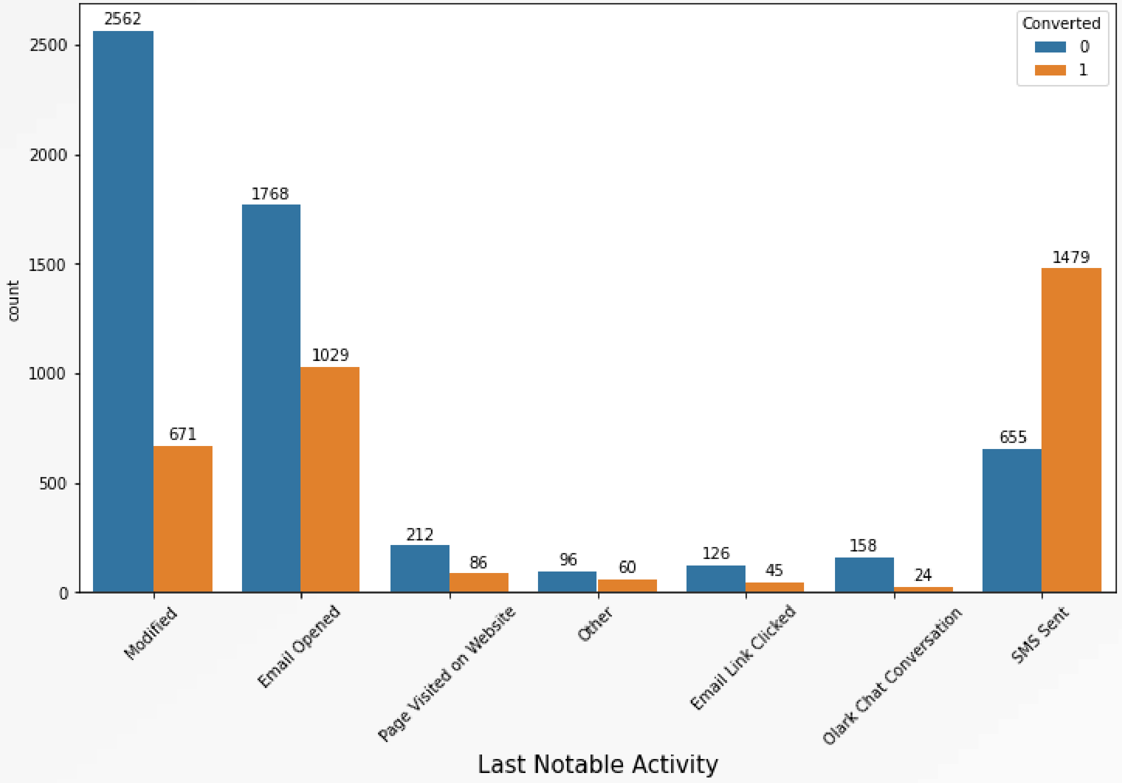


Categorical Variables

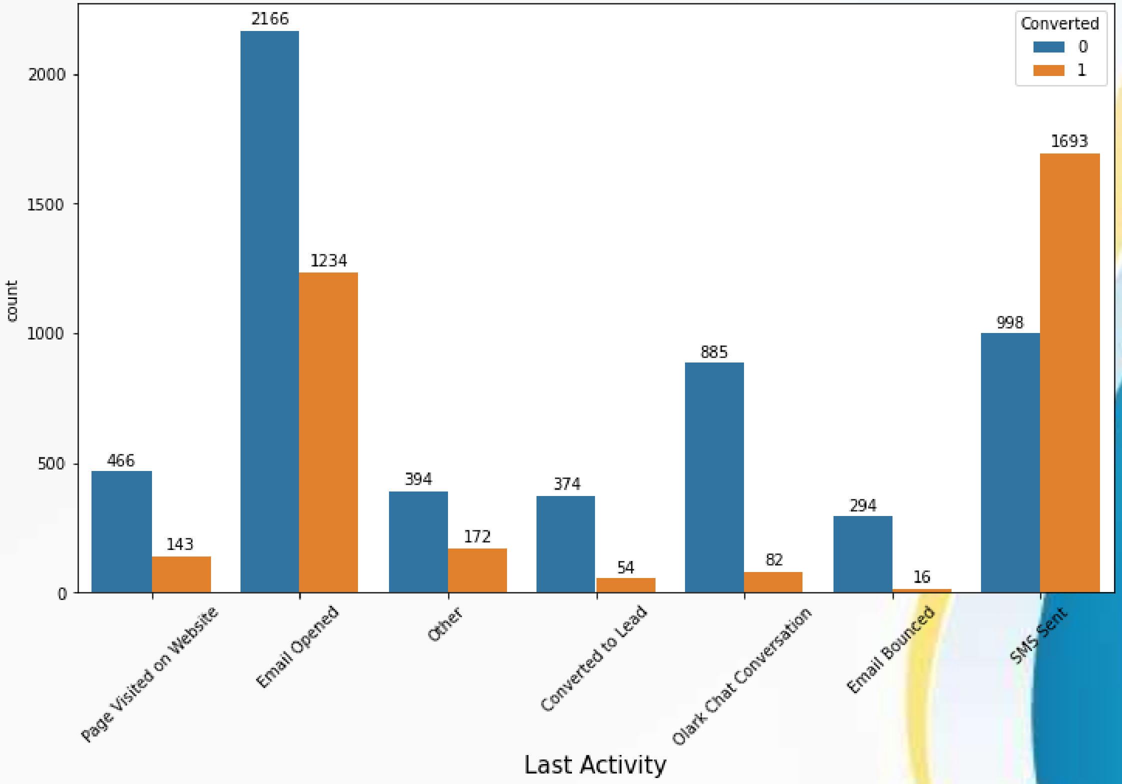
Analysis:

- In column `Last Activity` , `SMS Sent` category is the one with highest conversion rate of 63%
- In which categories namely `Olark Chat Conversation` and `Email Bounced` have the lowest conversion rates of 9% and 8% respectively.
- In the column `Last Notable Activity` , the highest conversion rate of 70% is in category ‘ SMS Sent
- The lowest conversion rate is in “Modified, and “Email opened”.

Segmented Analysis of Last Notable Activity Variable



Segmented Analysis of Last Activity variable

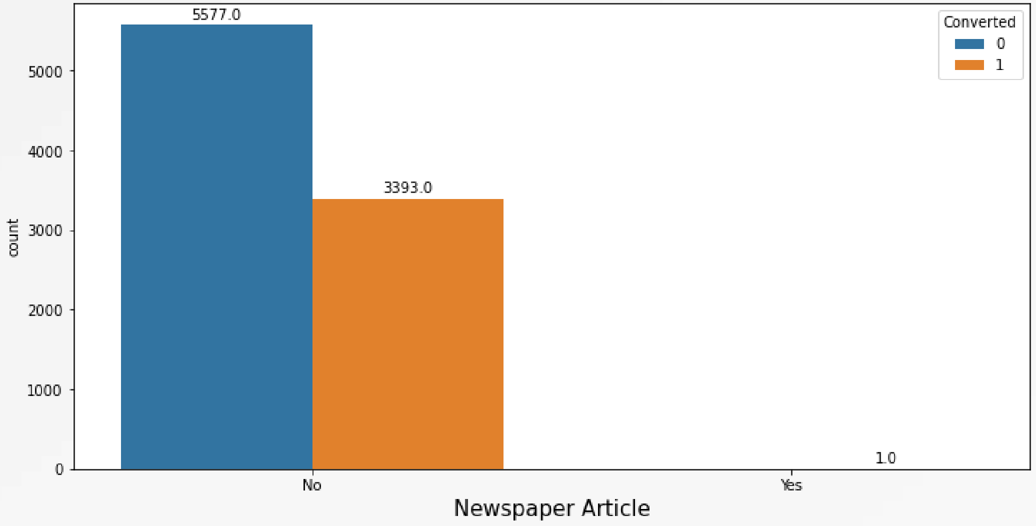


Categorical Variables

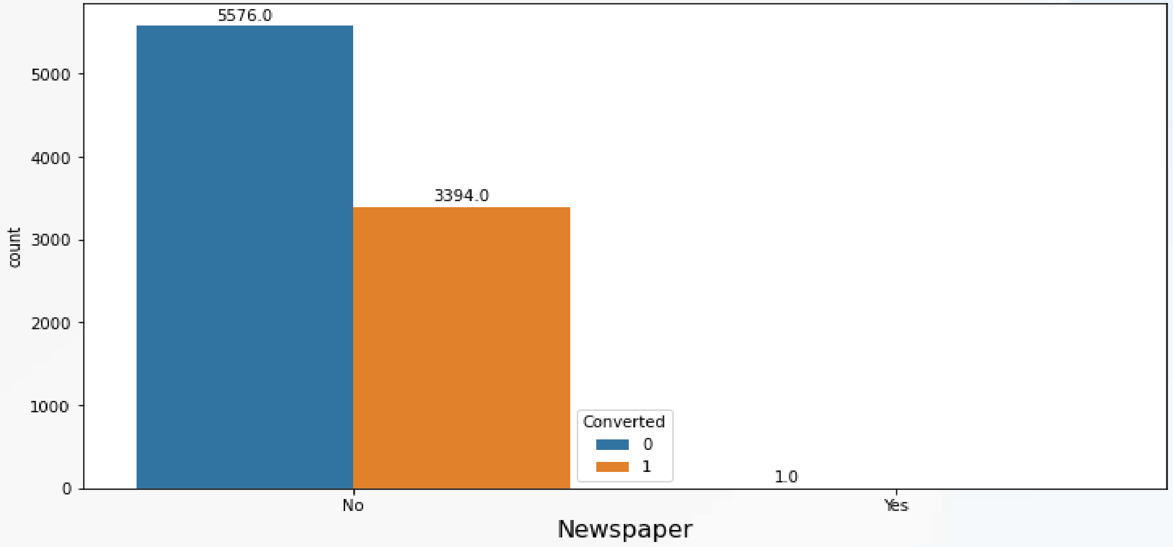
Analysis:

- Around 99% of customers haven’t seen any ad through the mentioned channels
- All the columns having category 'No' have around 38% of the conversion rate

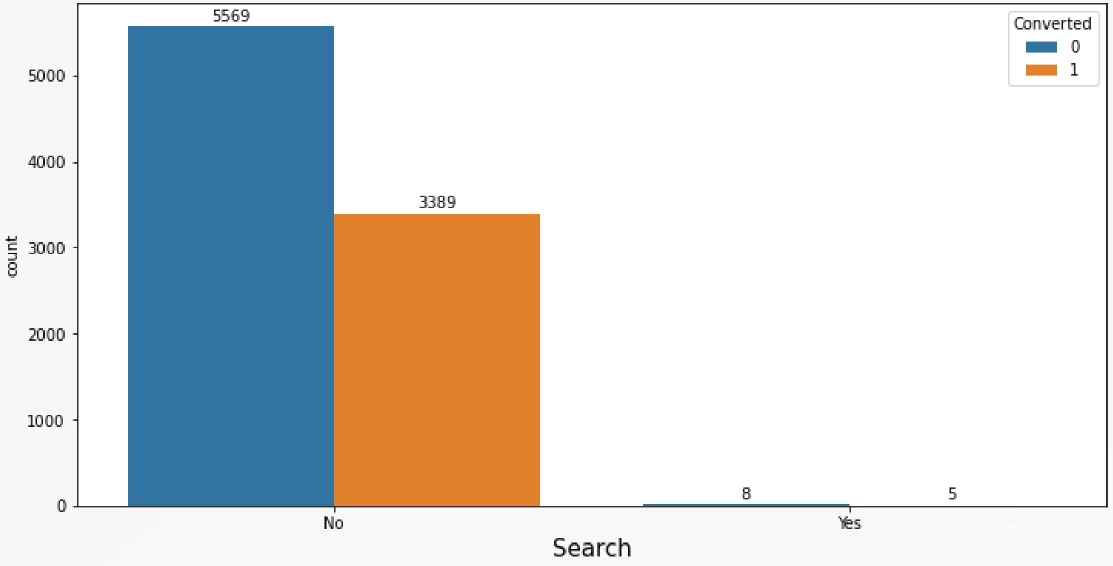
Segmented Analysis of 'Newspaper Article' Variable



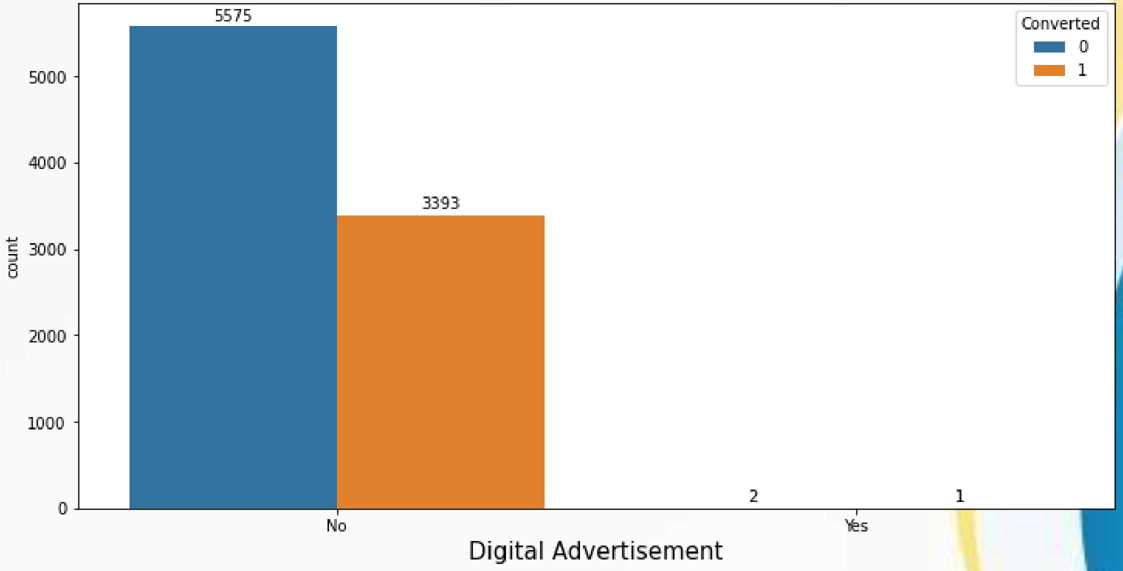
Segmented Analysis of 'Newspaper' Variable



Segmented Analysis of 'Search' Variable



Segmented Analysis of 'Digital Advertisement' Variable

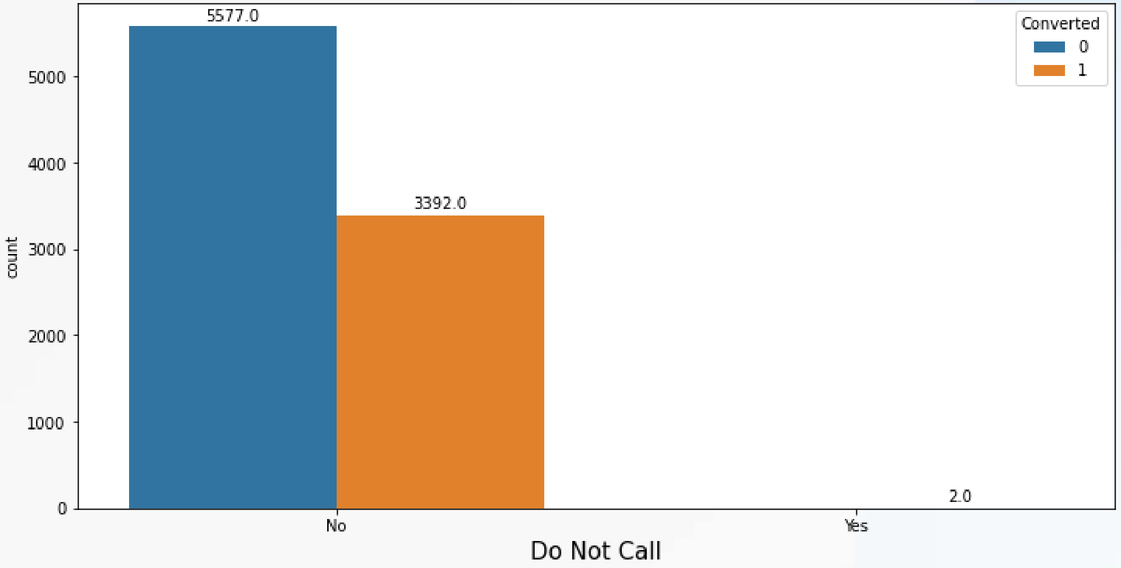


Categorical Variables

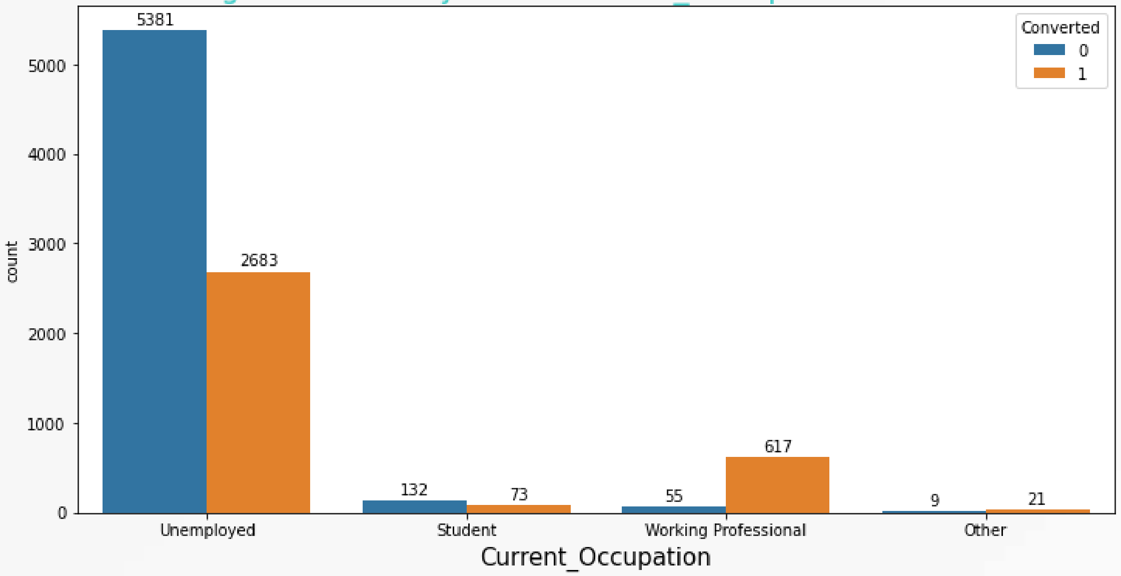
Insights

- In `Do Not Email` column major conversion(about 40%) has happened when customer opted for it
- In `Do Not call` column major conversion(about 38%) has happened when customer opted for it. Also though only 2 customers opted to not get a call but they both got converted
- In `Current_Occupation` column `Working Professional` category has the highest conversion rate, greater than 67% followed by `Other` category having conversion rate of 60%
- Working Professionals are most likely to convert

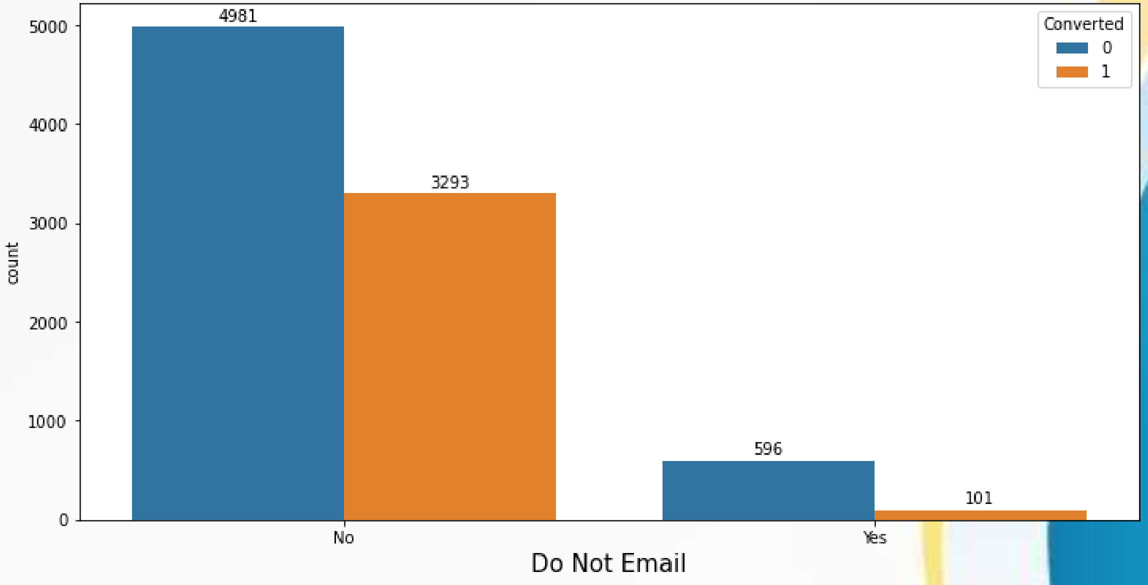
Segmented Analysis of 'Do Not Call' Variable



Segmented Analysis of 'Current_Occupation' Variable



Segmented Analysis of 'Do Not Email' Variable



Variables Impacting Conversion Rate

Customer Filled Variables Which Impact Conversion Rate

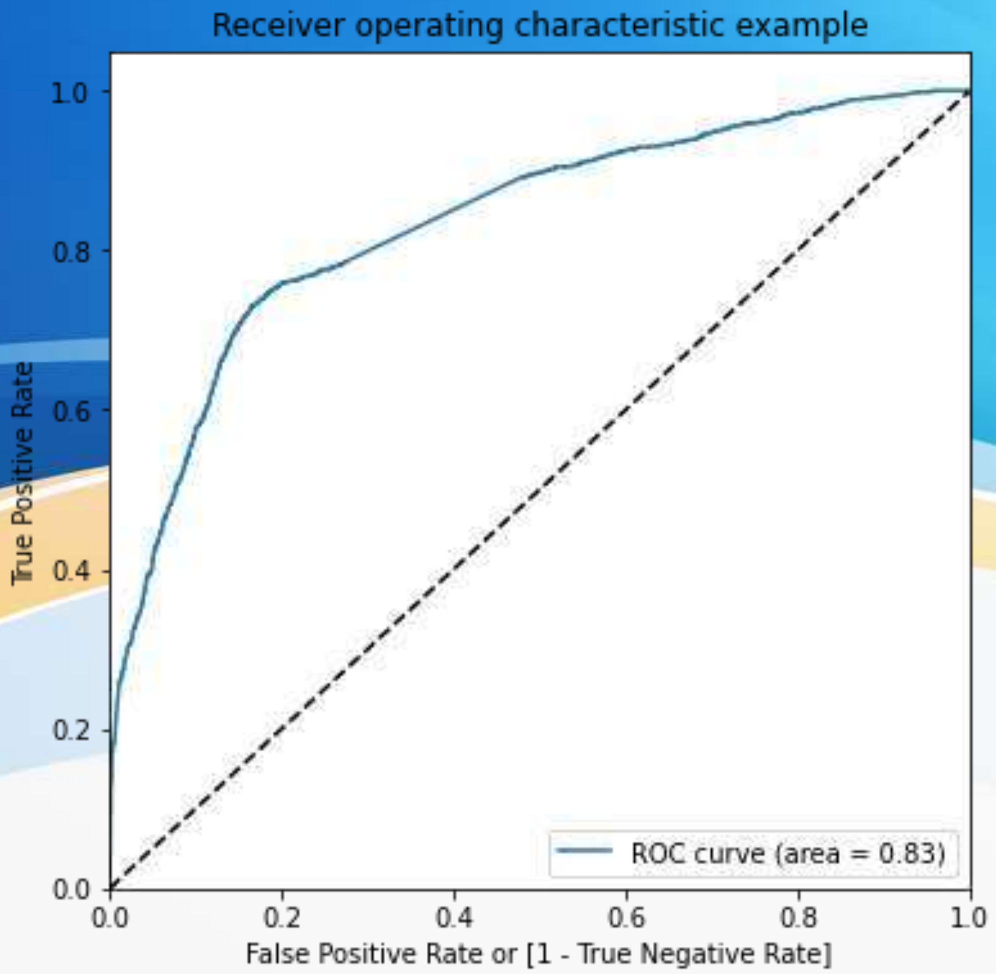
- Do Not Email
- Totalvisits
- Total Time Spent on Website
- Lead Origin_lead Add Form
- Lead Source_Direct Traffic
- Lead Source_Google
- Lead Source_Organic Search
- Lead Source_Referral Sites
- Lead Source_Welingak Website
- Current_Occupation_Working Professional

Sales Team Generated Variables which Impact Conversion Rate

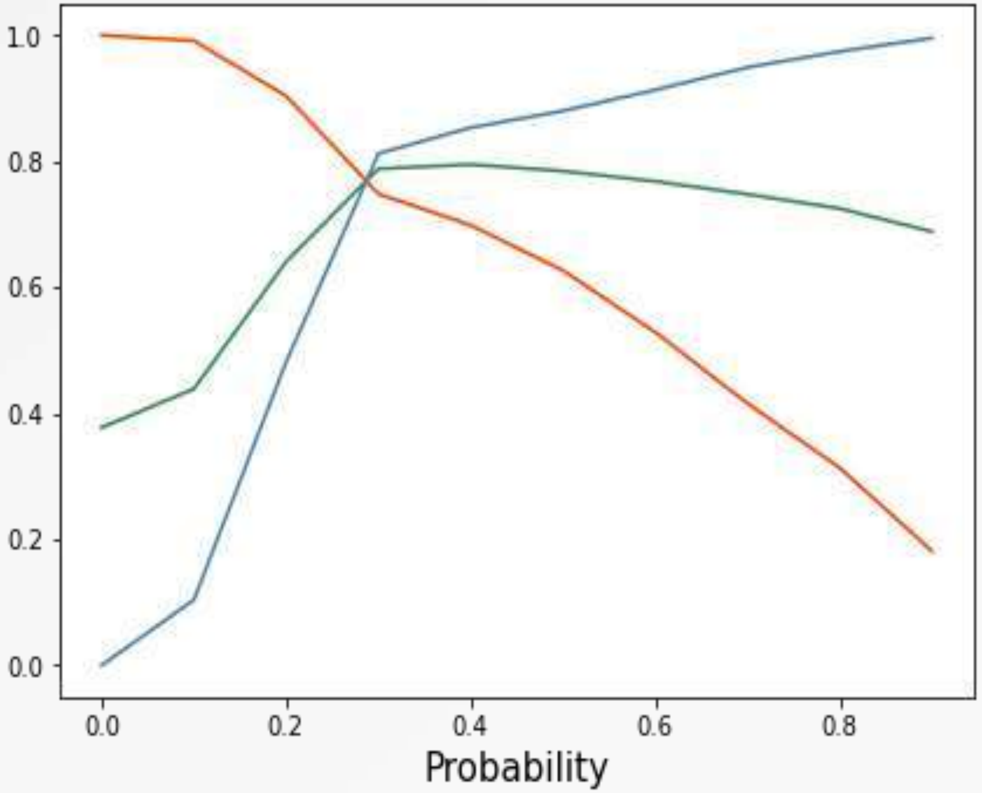
- Last Activity_SMS Sent
- Last Notable Activity_SMS Sent

Model Evaluation on Train Dataset

ROC Curve



Sensitivity-Specificity



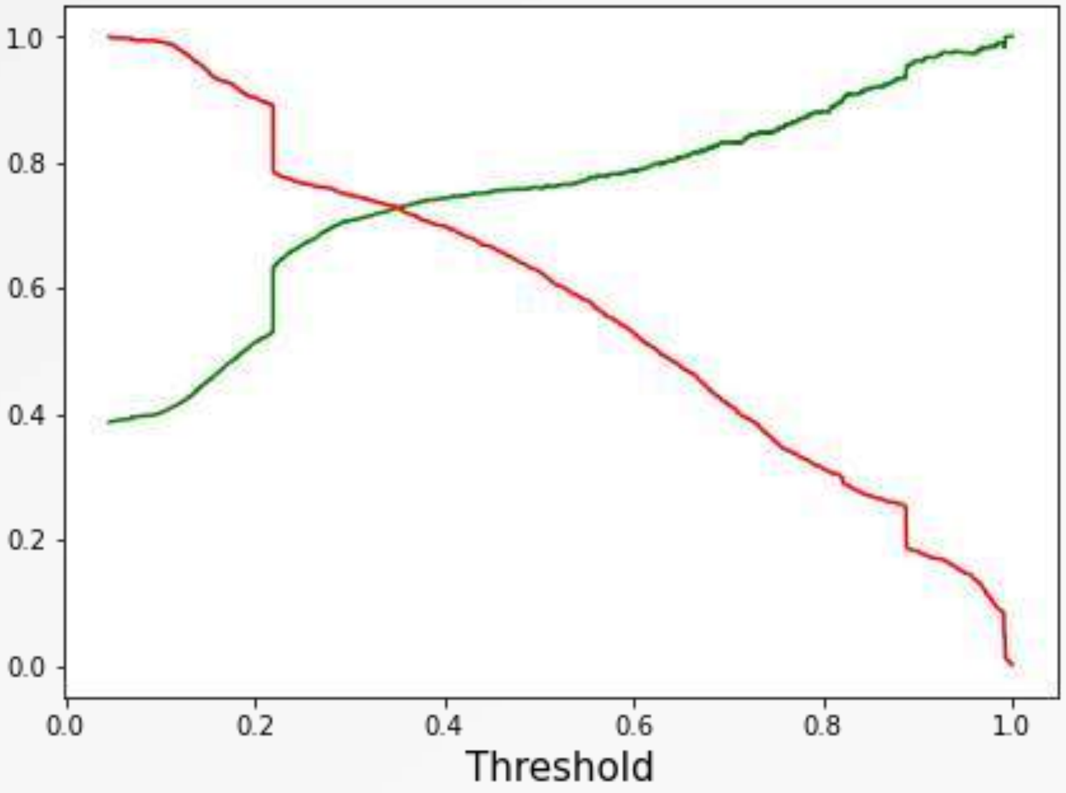
Confusion Matrix

3176	734
596	1773

- Accuracy Score of 78.8%
- Specificity = 81.2%
- Sensitivity = 74.8%
- False Positive Rate = 18%
- Positive Predictive Rate = 70.7%
- Negative Predictive Rate = 84%

When using the Sensitivity-Specificity-Accuracy plot, we found the optimal cutoff point to be 0.3.

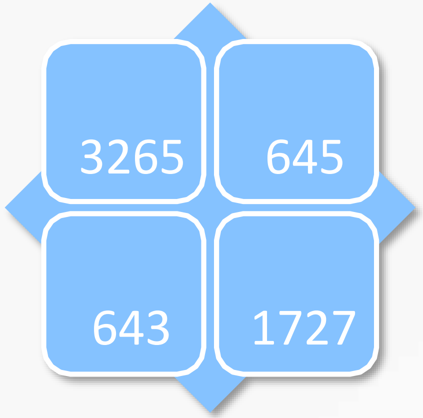
Precision-Recall Tradeoff



When using the Precision-Recall Tradeoff,

we found that the optimal cutoff point to be 0.35.

Confusion Matrix



3265	645
643	1727

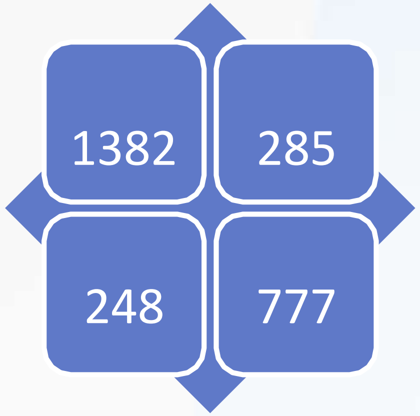
- **Accuracy Score of 79.5%**
- **Precision = 72.8%**
- **Recall = 72%**
- **Specificity = 83.5%**

Model Evaluation On Test Dataset

Using the sensitivity-specificity we found that the optimal cutoff point was 0.3. When we plotted the precision-recall tradeoff, we got the optimal cutoff point to be 0.35, since we pay more emphasis on sensitivity inorder to predict “hot leads” we will henceforth choose 0.3 as the final optimal cutoff point

- **Accuracy Score of 80.2%**
- **Specificity = 82.9%**
- **Sensitivity = 75.8%**

Confusion Matrix



1382	285
248	777

Considering the Cutoff of 0.3 we can infer that if any lead having a leads score of 30 or above then, that lead has high probability of getting converting and should be considered as a potential “hot lead”.

Conclusion

- The conversion rate before building the machine learning model was merely 38% but post building the model we were able to predict 80% of leads which could converted making a conversion rate of about 80%, i.e. an incrementof 42%.

  Any lead possesing a Lead Score which is greater than 30 should be considered as a Hot Lead

- Customers who spend about 10 minutes or more on the website or have Lead Origin as 'Lead add form' should be considered as potentail hot leads since they have a high chances of getting converted.
- Clients having 'Reference' or 'Wellingak Website' as Lead Source or customers who are working professional generally have a high probability of getting converted and must be considered as potential hot Leads
- Customers who visit the website more often (5 times or more) have a good chance of getting converted , we can target them inorder to increase the conversion rate

Recommendations

- While surveying we have found that a lot of leads which are generated in the initial stage out of which very few of them come out as paying customers. In the intermediate stage, we have to cater these potential customers well (i.e via. sharing the knowledge about the product, constantly communicating etc.) which will ultimately lead us to increase the conversion rate.

- Customers having a lead score of (30 or more)are more likely to join , so sales team has to be actively looking after these leads , so that we don’t miss any potential customer.

- Once all the customers having a lead_score of 30 or more have been catered, then the sales team should also consider and target customers with lead_score of 20 and above as there is a possibility of these customers to convert too.

- One strategy to minimize the rate of useless call is to pick those customers having a lead score of 60 or above, since they are most likely to join the course, regular follow up would increase the chances of conversion .

If a customer is having a lead score lower than 60 or 70 but have a combination of attributes which are, multiple visits to website, spending more than 10 minutes of their time on the website surfing for information regarding the course, customer is a working professional or has Lead Source as ‘Welingak Website’ and has Lead Origin as ‘Lead Add Form’ or visited the site due to recommendation provided by fellow students then those customers are most likely to convert and should be our potential target.