# Lead Score Case Study

GROUP MEMBERS

1.KRUNAL SONI

2. PRATHMESH KULKARNI

# Problem Statement:-

An education company named X Education sells online courses to industry professionals. 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. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

**What you need to do?**

● X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

● 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 higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

# Assignment Steps:-

1. Data Cleaning  and transformation

check missing values . Null values, import those values , dropping columns which are not useful , handling outliers.

2. EDA

Univariate analysis, bivariate analysis . To get pattern between variables and get distribution of values using the matplotlib libreary and seaborn, correlation between variables using heat map method.

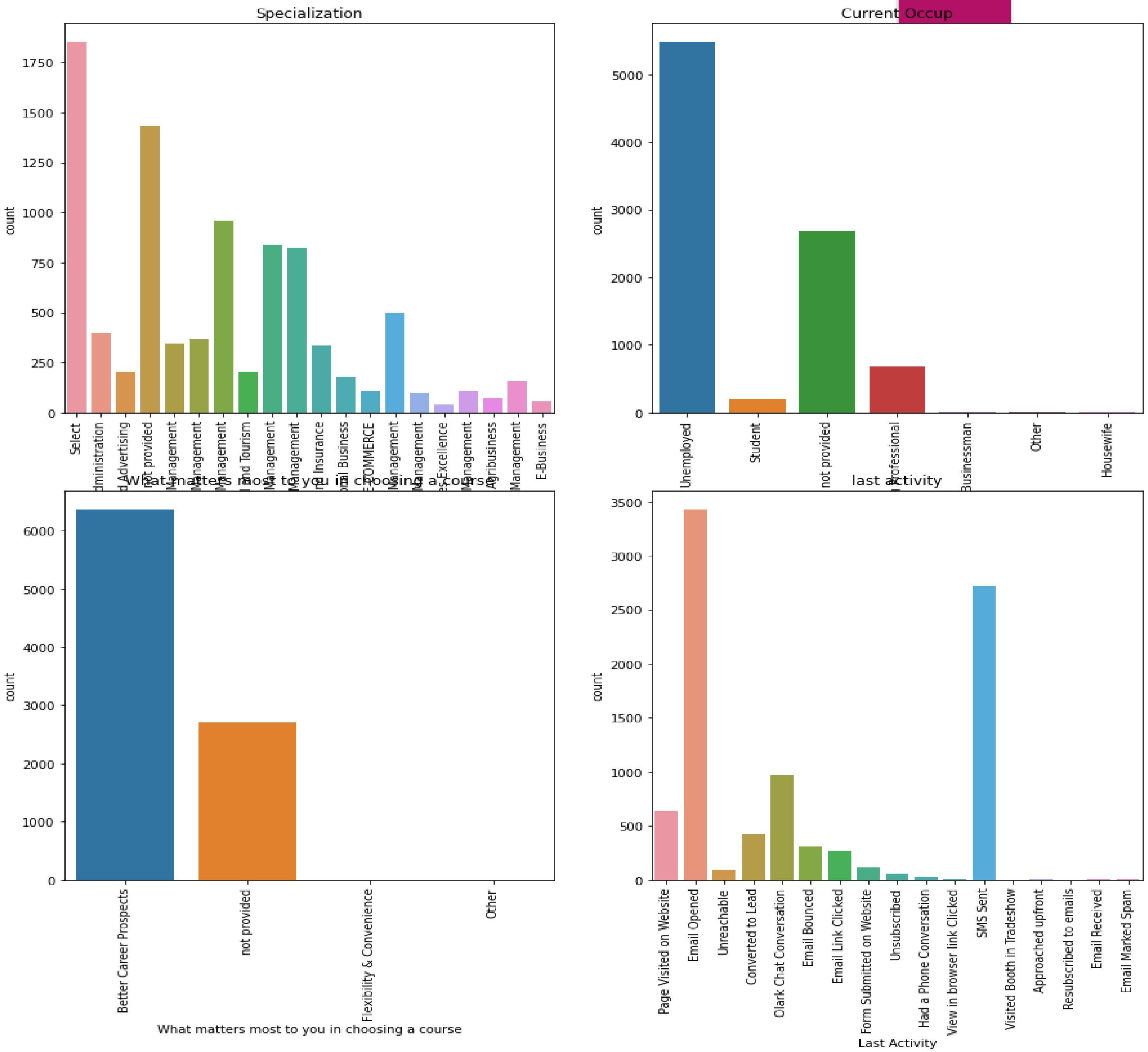
3. Divide data sets on train test split , Features scaling , dummy variables using minimax scalars .

4. Model building using Rfe , check sensitivity , precision ,recall , model performance on test data sets , optimal cut off , check p value and vif value

5. Model validation , and finding conclusion , recommendation.

# EDA

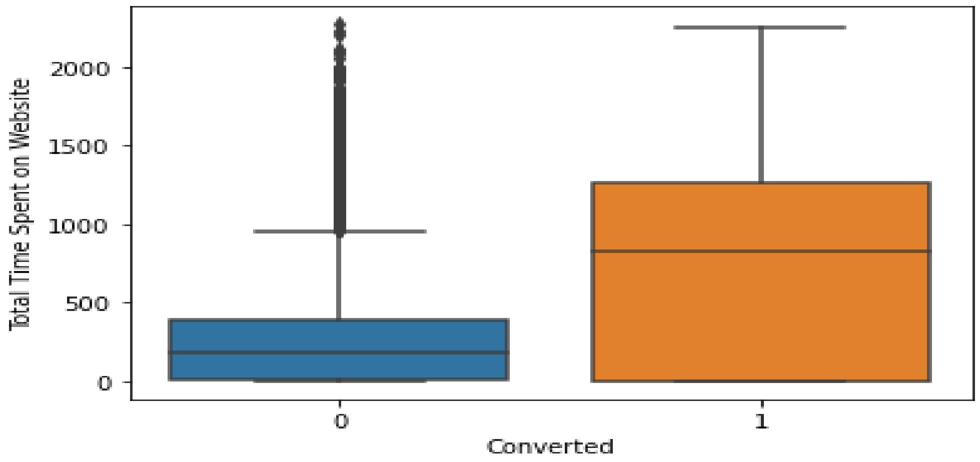
1. Most of the people choosing the course most of them they working professionals.
2. Based on updated data management side specialization are more preferable
3. People choosing the course for their better career prospects
4. Last activity of clients or most frequently useful tool for automated marketing based on last activity is sms, email , olark chat conversation

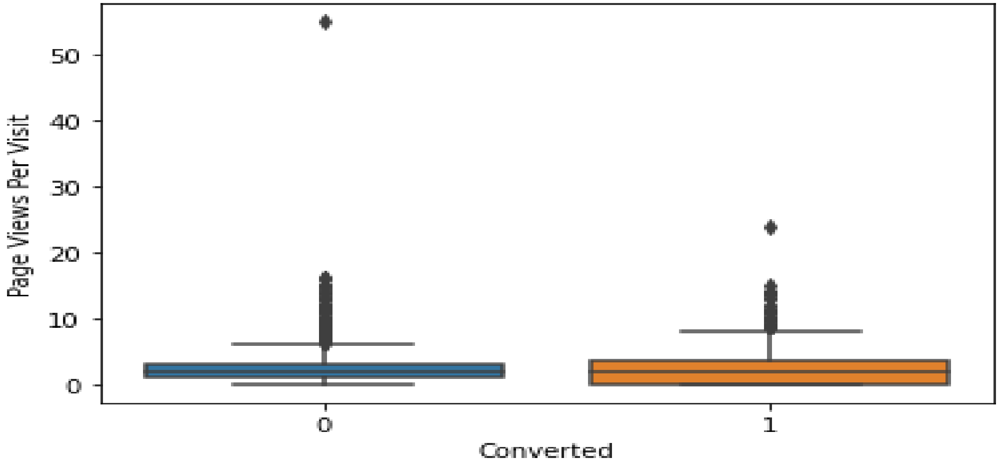


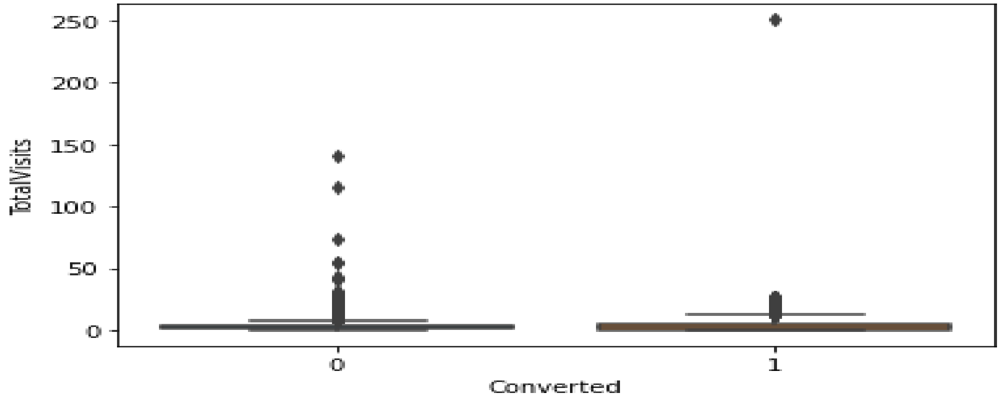
# Numerical data sets relation

1. Total time spent on websites on customer are high which is good tool for lead conversation.
2. Page views per visit on lead conversation rate are same

3. Total visits median value of boxplot datasets are same so this is also not much useful







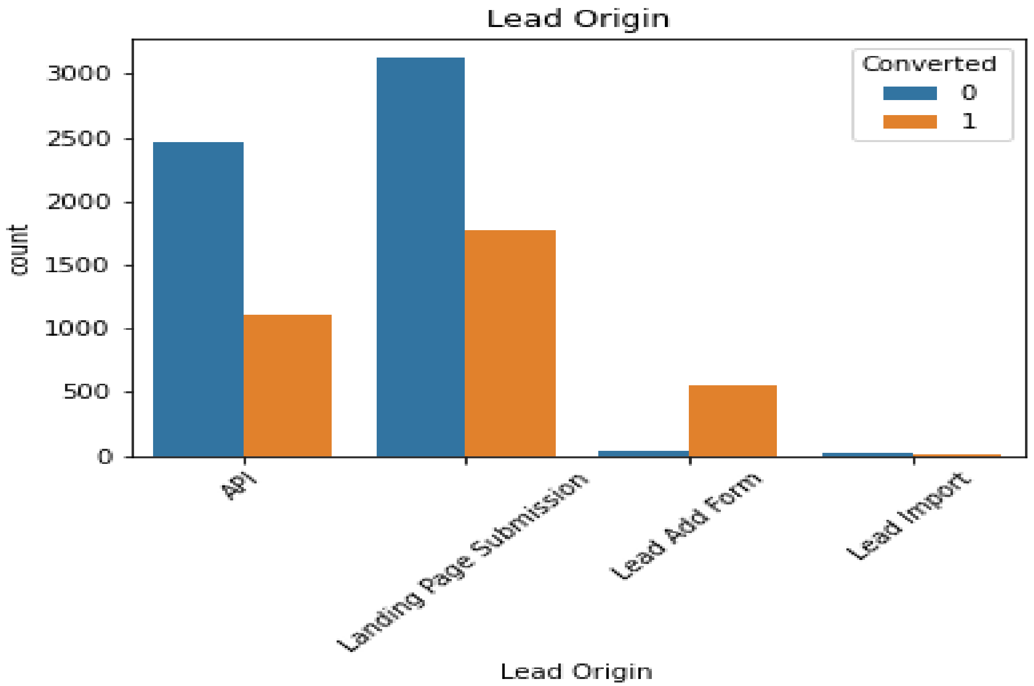
# Numerical variables relation on heat map & categorical variables of lead origin

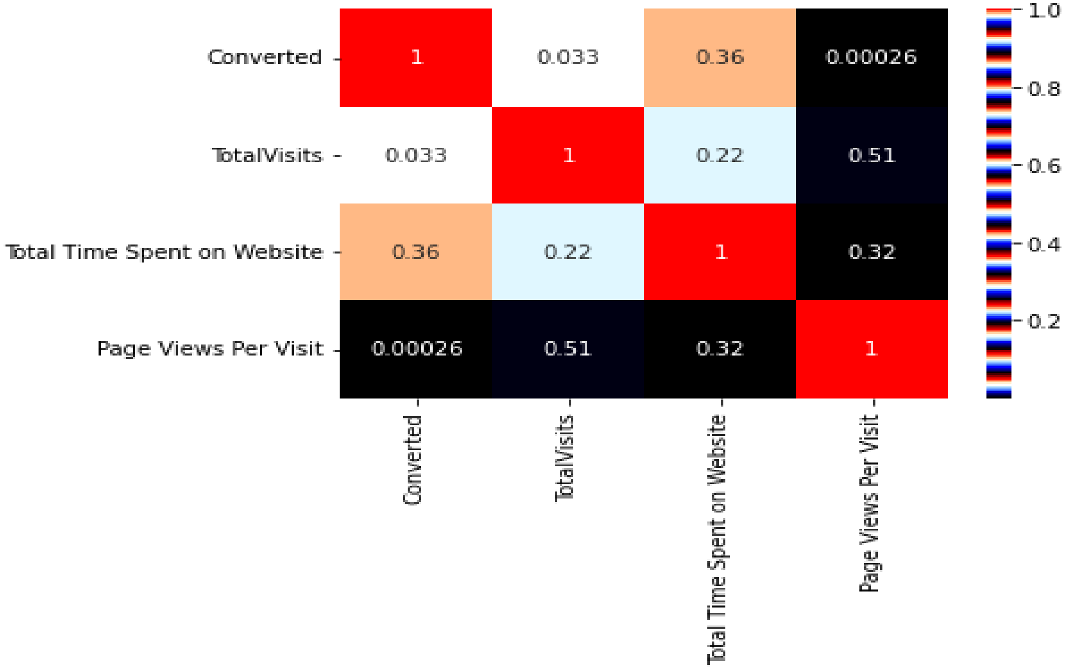
Lead origin :-

landing page submission we can generate many leads and converted in to hot leads

Heat map of numerical data :-

Total visits & pages views per visits are highly correlated , time spent on websites and page views per visit





| | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit |
|---|---|---|---|---|
| Converted | 1 | 0.033 | 0.36 | 0.00026 |
| TotalVisits | 0.033 | 1 | 0.22 | 0.51 |
| Total Time Spent on Website | 0.36 | 0.22 | 1 | 0.32 |
| Page Views Per Visit | 0.00026 | 0.51 | 0.32 | 1 |

# Categorical variables analysis

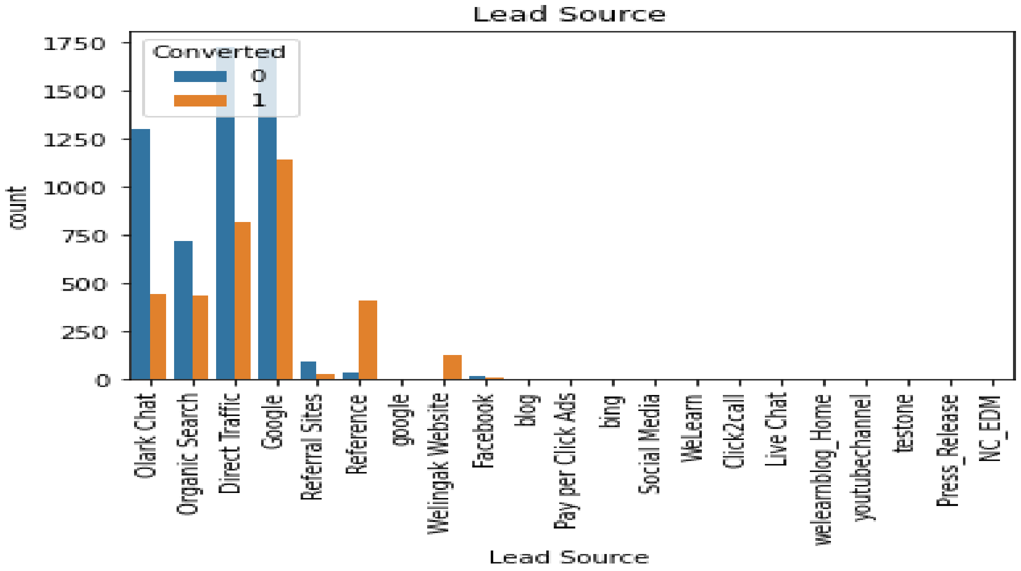
Lead source :-

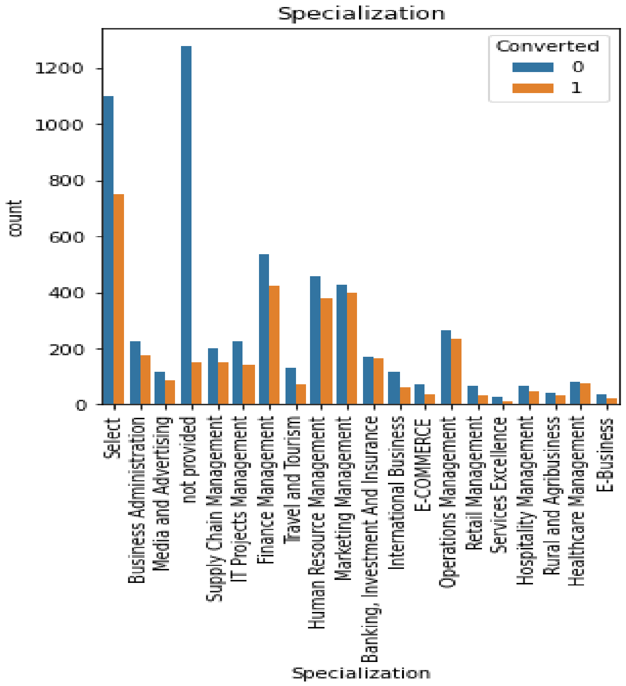
from google , direct traffic , organics search ,

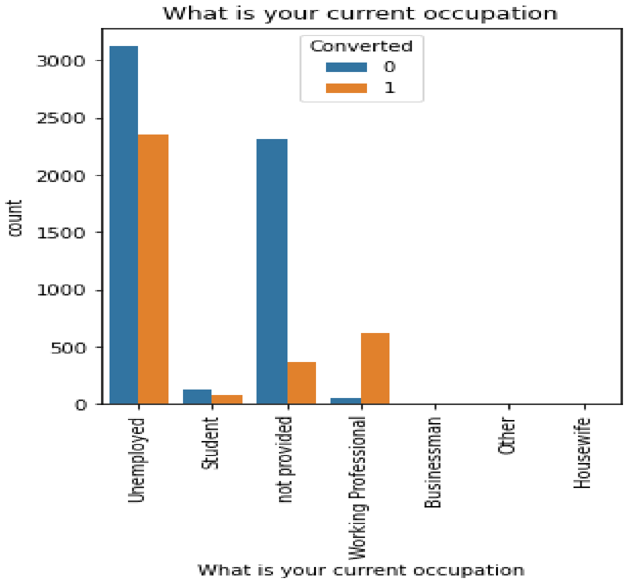
Olark chat those elements are for lead generating for company.

Specialization :- clients choose speclization are mostly finance management , human resource management , operation management , marketing management

Current occupation :- people are choosing the course are mostly unemployed and working professionals .



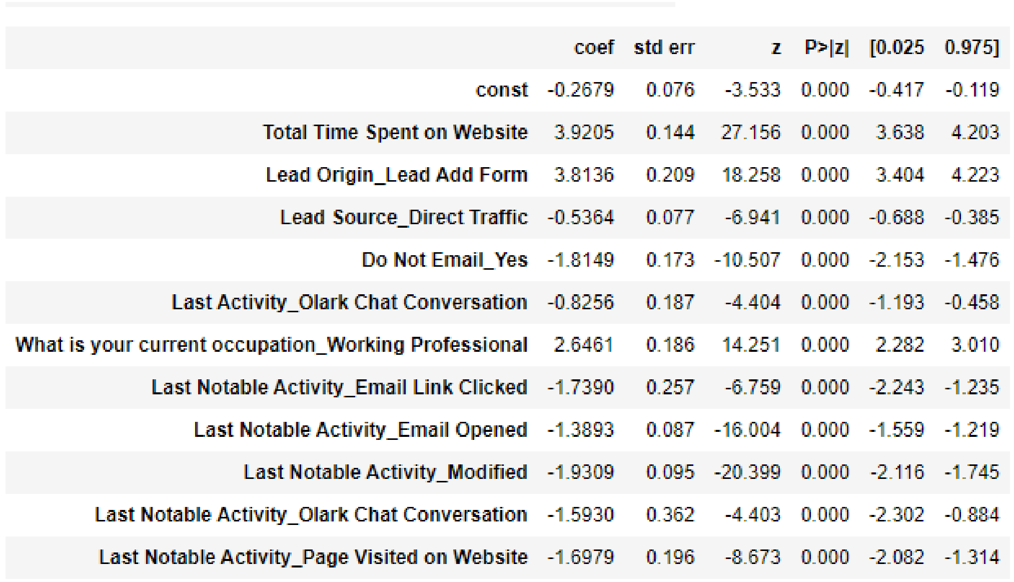




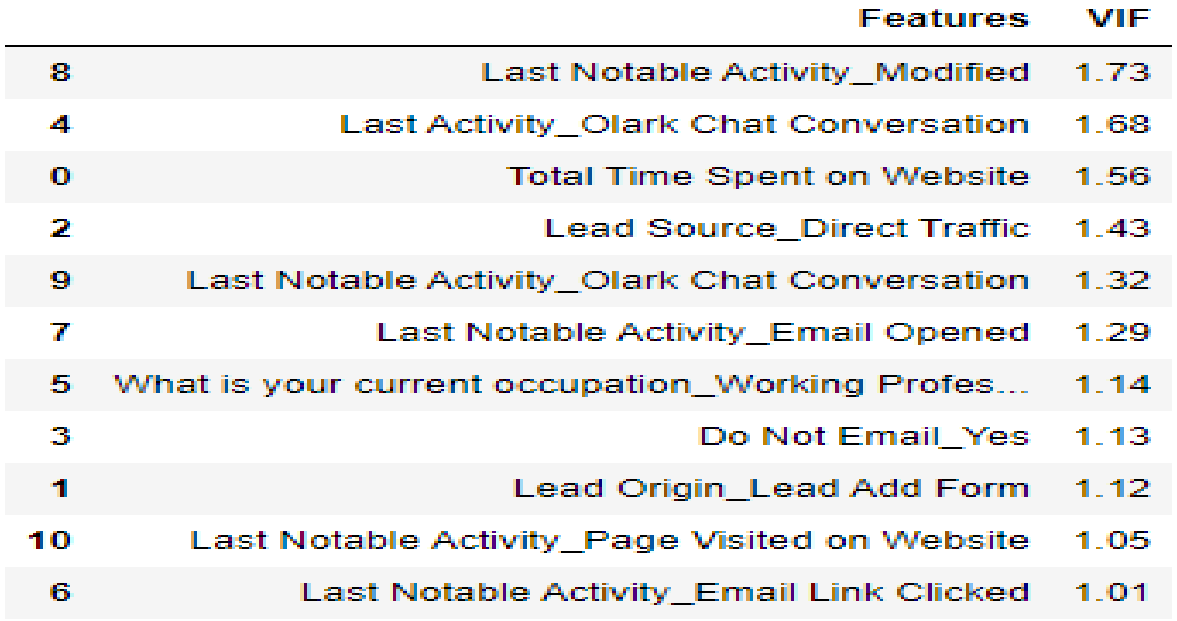
# Final Model

Here in all data sets p value is zero so they are significant and satisfy the technical conditions for good  model

Vif of all variables are less than 5 which is all satisfy the perfect model for further analysis



| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.2679 | 0.076 | -3.533 | 0.000 | -0.417 | -0.119 |
| Total Time Spent on Website | 3.9205 | 0.144 | 27.156 | 0.000 | 3.638 | 4.203 |
| Lead Origin_Lead Add Form | 3.8136 | 0.209 | 18.258 | 0.000 | 3.404 | 4.223 |
| Lead Source_Direct Traffic | -0.5364 | 0.077 | -6.941 | 0.000 | -0.688 | -0.385 |
| Do Not Email_Yes | -1.8149 | 0.173 | -10.507 | 0.000 | -2.153 | -1.476 |
| Last Activity_Olark Chat Conversation | -0.8256 | 0.187 | -4.404 | 0.000 | -1.193 | -0.458 |
| What is your current occupation_Working Professional | 2.6461 | 0.186 | 14.251 | 0.000 | 2.282 | 3.010 |
| Last Notable Activity_Email Link Clicked | -1.7390 | 0.257 | -6.759 | 0.000 | -2.243 | -1.235 |
| Last Notable Activity_Email Opened | -1.3893 | 0.087 | -16.004 | 0.000 | -1.559 | -1.219 |
| Last Notable Activity_Modified | -1.9309 | 0.095 | -20.399 | 0.000 | -2.116 | -1.745 |
| Last Notable Activity_Olark Chat Conversation | -1.5930 | 0.362 | -4.403 | 0.000 | -2.302 | -0.884 |
| Last Notable Activity_Page Visited on Website | -1.6979 | 0.196 | -8.673 | 0.000 | -2.082 | -1.314 |



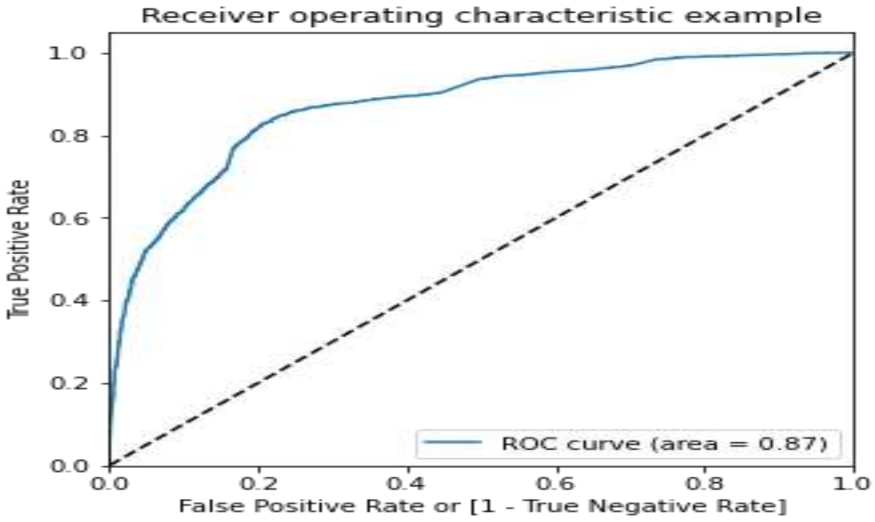
| | Features | VIF |
|---|---|---|
| 8 | Last Notable Activity_Modified | 1.73 |
| 4 | Last Activity_Olark Chat Conversation | 1.68 |
| 0 | Total Time Spent on Website | 1.56 |
| 2 | Lead Source_Direct Traffic | 1.43 |
| 9 | Last Notable Activity_Olark Chat Conversation | 1.32 |
| 7 | Last Notable Activity_Email Opened | 1.29 |
| 5 | What is your current occupation_Working Profes... | 1.14 |
| 3 | Do Not Email_Yes | 1.13 |
| 1 | Lead Origin_Lead Add Form | 1.12 |
| 10 | Last Notable Activity_Page Visited on Website | 1.05 |
| 6 | Last Notable Activity_Email Link Clicked | 1.01 |

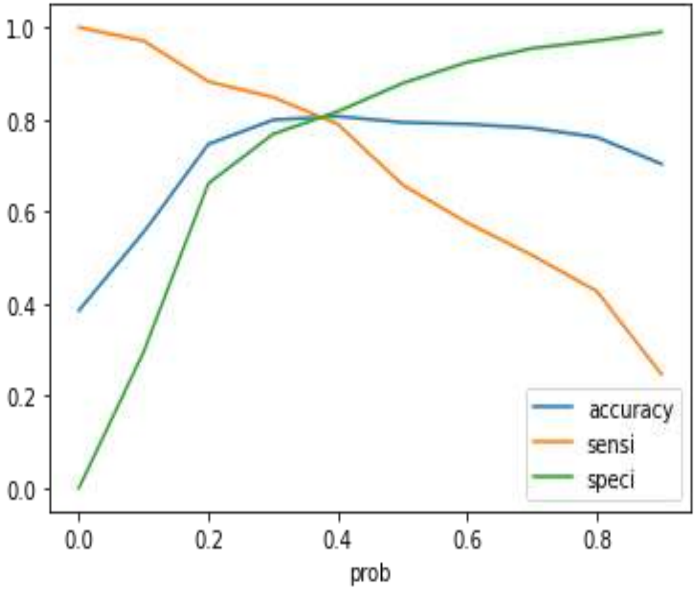
# ROC curve , intersection of accuracy, sensitivity, specificity

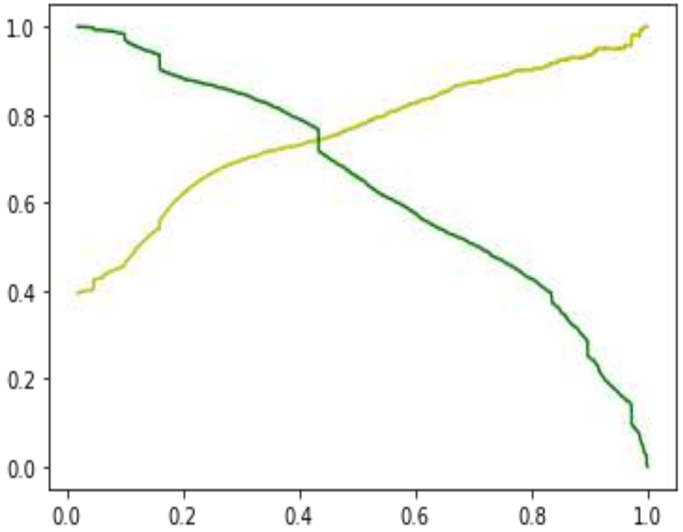
Roc curve of is under 0.87 which is good for data sets

Intersection of accuracy , sensitivity and specificity of data sets is 0.38 which is optimal cut-off

Trade off of precision and recall value is 0.42







# DATA Conversion

- Numerical variables are normalised
- Create dummy variables then convert it into float , int64,unit8 data types
- After cleaning part we have 9074 rows and 23 columns
- Remove data sets based on 40 % of missing values
- Used minimax scaler for creating dummies
- Dropping the unique value of columns because it is not useful for further analysis
- Outliers treatment and importing select and nan value in missing values because those columns can not be drop which contain useful data for company

# Conclusion & Recommendation

Conclusion:-
1.most of the time spends on websites
2.last activity sms , email, chat conversation of most of the lead
3.choosing course because of better career options
4.lead generation from google , olak chat , organic searches , references
5.landing page from submission
6.most are working professionals etc.

Recommendation :-
Company need marketing campaign and generate leads from google , olark chat , references , or use the automated tools like email, sms etc. for human less and in limited time for generating hot leads to achieve the target