LEAD SCORE CASE STUDY

TEAM MEMBERS –

MANISHA DASH, SANJEEBANI SWAIN, JAYANTI KUMARI

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. 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 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%.

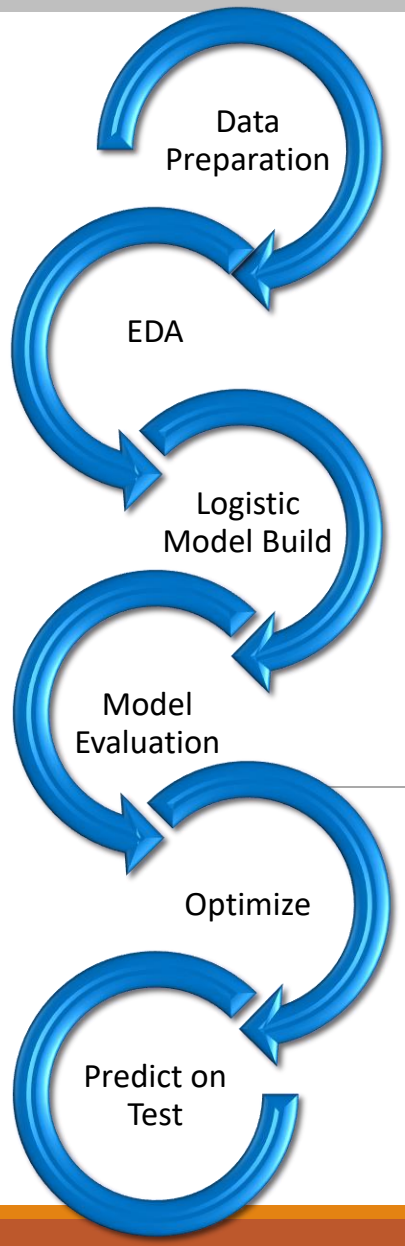
OBJECTIVE

THE MAJOR OBJECTIVE OF THE CASE STUDY IS TO -

1.BUILD AN OPTIMAL LOGISTIC MODEL THAT WOULD ASSIGN A LEAD SCORE BETWEEN 0 AND 100 TO EACH LEAD GENERATED , WHICH COULD BE IN TURN USED BY THE COMPANY TO DETERMINE THE POTENTIAL LEAD TO BE TARGETTED .A HIGH SCORE WOULD MEAN HIGHER POTENTIAL OF THE LEAD BEING CONVERTED .THIS WOULD INVOLVE A LOGISTIC REGRESSION ALGORITHM AS IT INVOLVES GENERATION OF SCORES

2.BASED ON THE MODELING IDENTIFY THE IMPORTANT VARIABLES AND PROVIDE FEASIBLE RECOMMENDATIONS

APPROACH



- ✓ Data ingestion through Python from the csv file
- ✓ Data cleansing activity and prepping for analysis

- ✓ Univariate Analysis
- ✓ Categorical variable analysis

- ✓ Initial feature selection through RFE
- ✓ Python libraries to run the logistic models
- ✓ Iteratively remove attributes and check p values & VIF

- ✓ Confusion matrix
- ✓ Accuracy score, Sensitivity & specificity

- ✓ Optimize Cut off
- ✓ Re-evaluate sensitivity ,Specificity

- ✓ Rerun the model on the test data
- ✓ Confusion matrix - Accuracy score, Sensitivity & specificity

DATA MANIPULATION

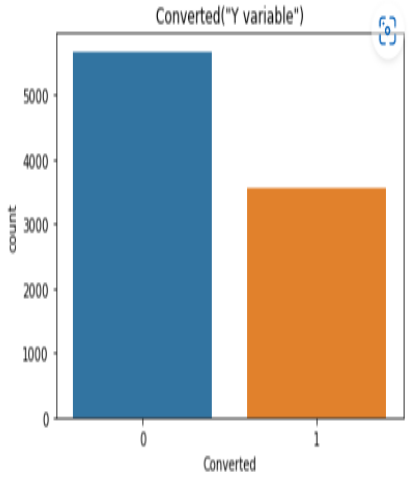
- There are 37 attributes and 9240 records in the ‘leads.csv’ file for analysis purpose.
- Attributes that has single value in it doesn’t make sense for analysis , hence dropped .Ex- Magazine
- Converting ‘Select’ to null as it doesn’t hold any value
- Attributes with >35% nulls in it were removed, as they may not account more into the EDA & models
- Even after doing the previous steps , there were huge number of attributes which had null values in it

but removing them was not logical at this stage .However they will be removed from model if found insignificant

- converting values to 'not provided' instead of null in the attributes which has >20% nulls.
- Country field was categorised as India, not provided and outside India for better readability
- Dropping ID fields(prospect ID & Lead number) as they don’t matter in model
- Few other fields were dropped which did not have enough variance

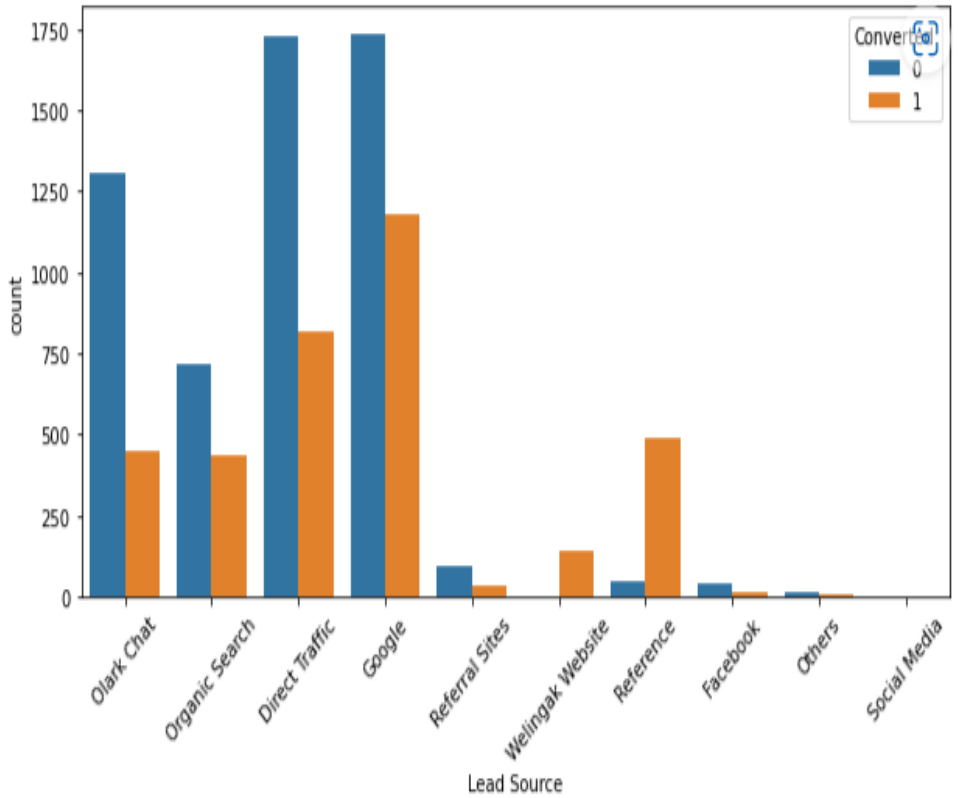
EDA Summary –Lead source & Origin

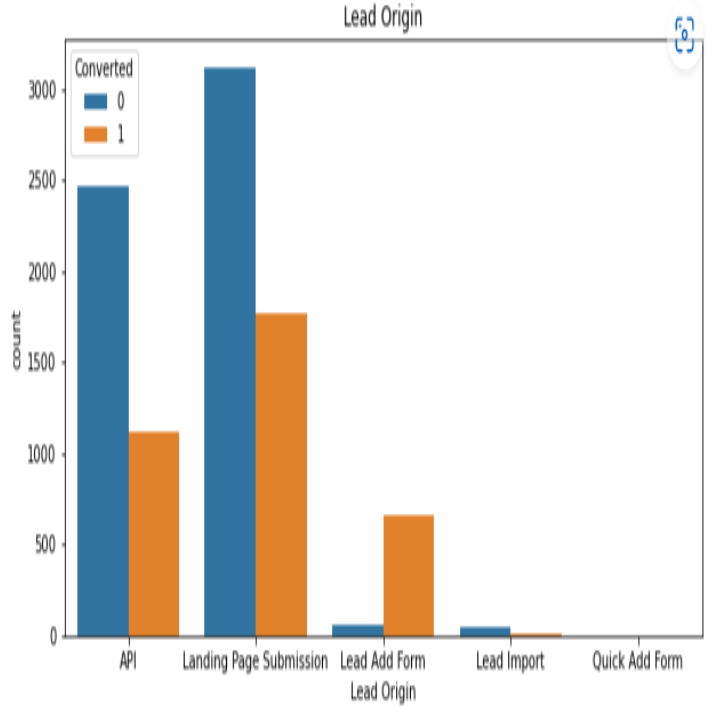
```
sns.countplot(leads['Converted'])
plt.title('Converted("Y variable")')
plt.show()
```




```
# Finding out conversion rate
Converted = (sum(leads['Converted'])/len(leads['Converted'].index))*10
Converted
```

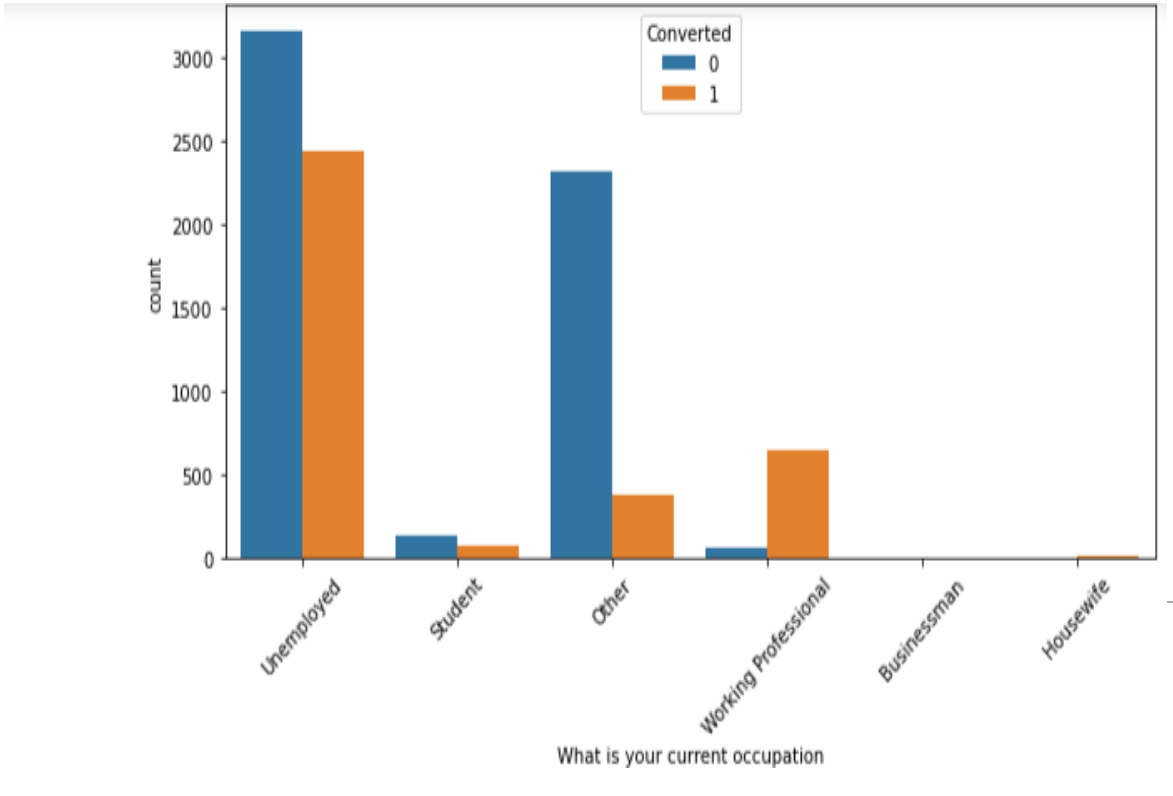
38.53896103896104

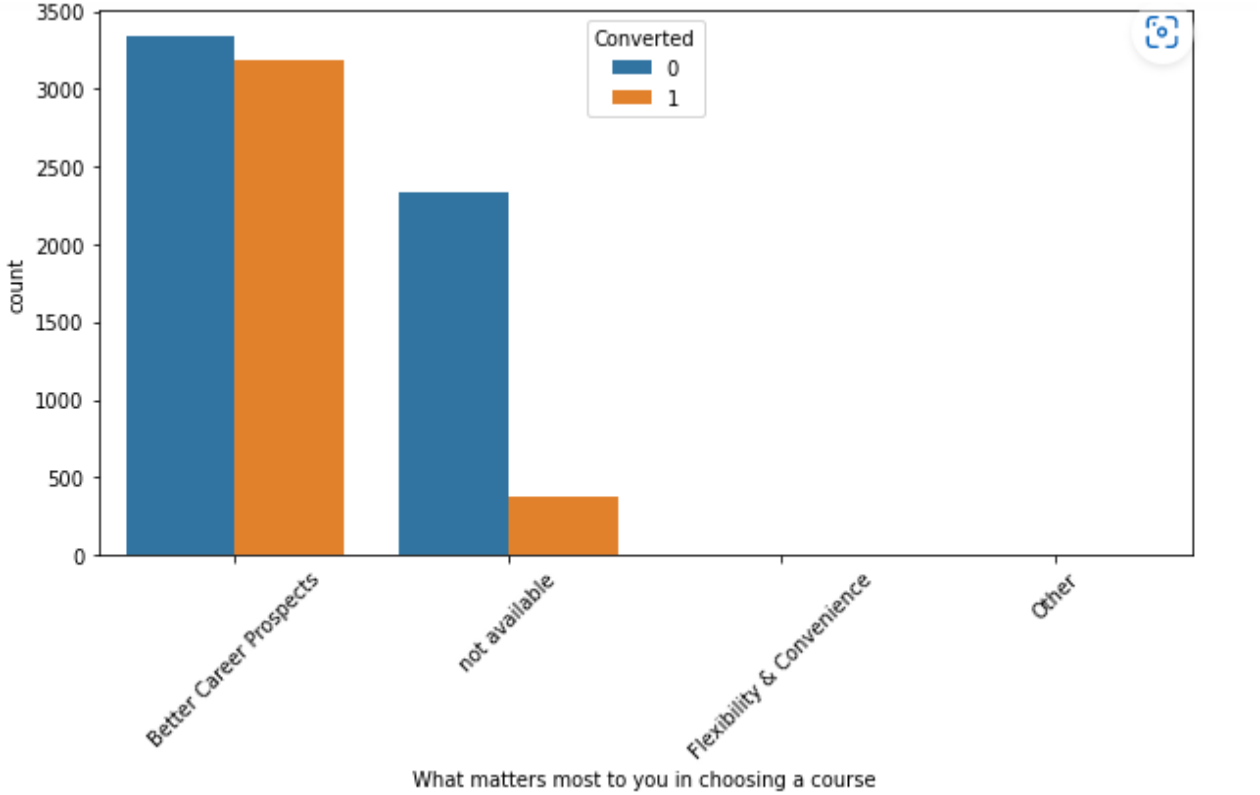




- Current conversion rate is around 39%
- Maximum leads are generated from Google, Olark chat, Organic search ..
- Most leads were generated at the landing page submission

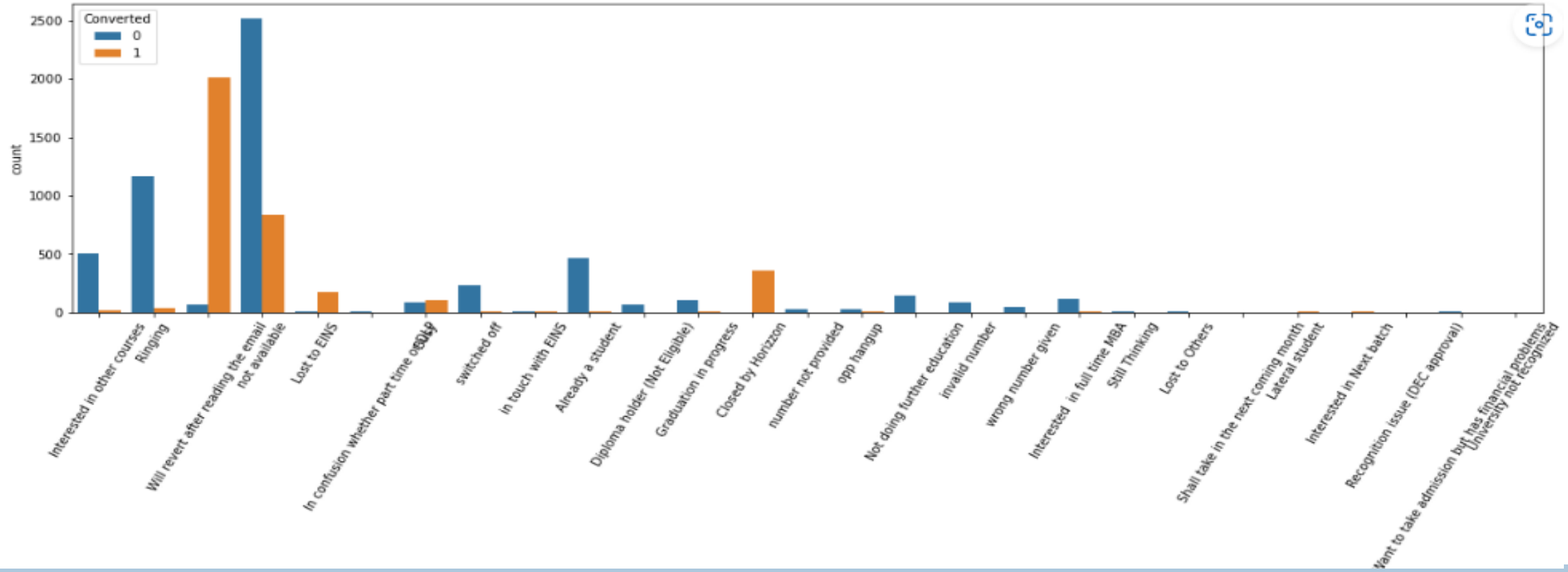
EDA Summary- contd





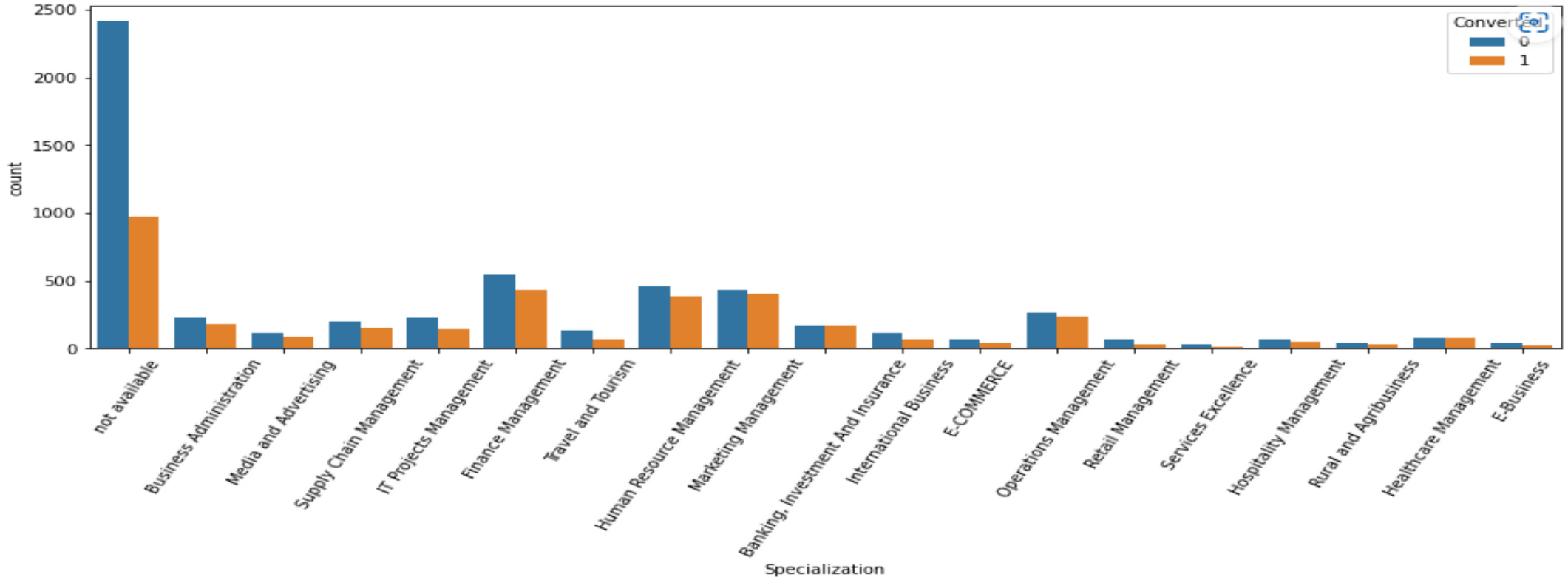
- Maximum leads generated are unemployed and their conversion rate is more than 50%. Conversion rate of working professionals is very high.
- Most leads are looking for better career prospects

EDA Summary- Tags



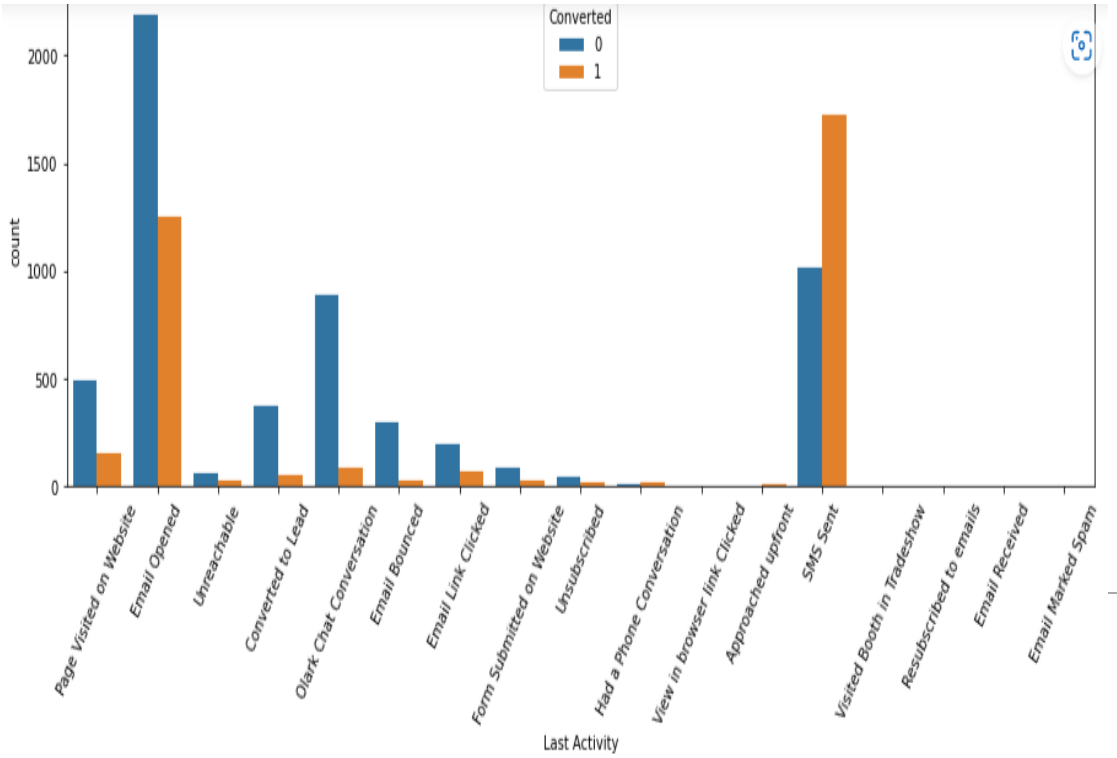
- Profiling on tags do not show much relevant behaviour

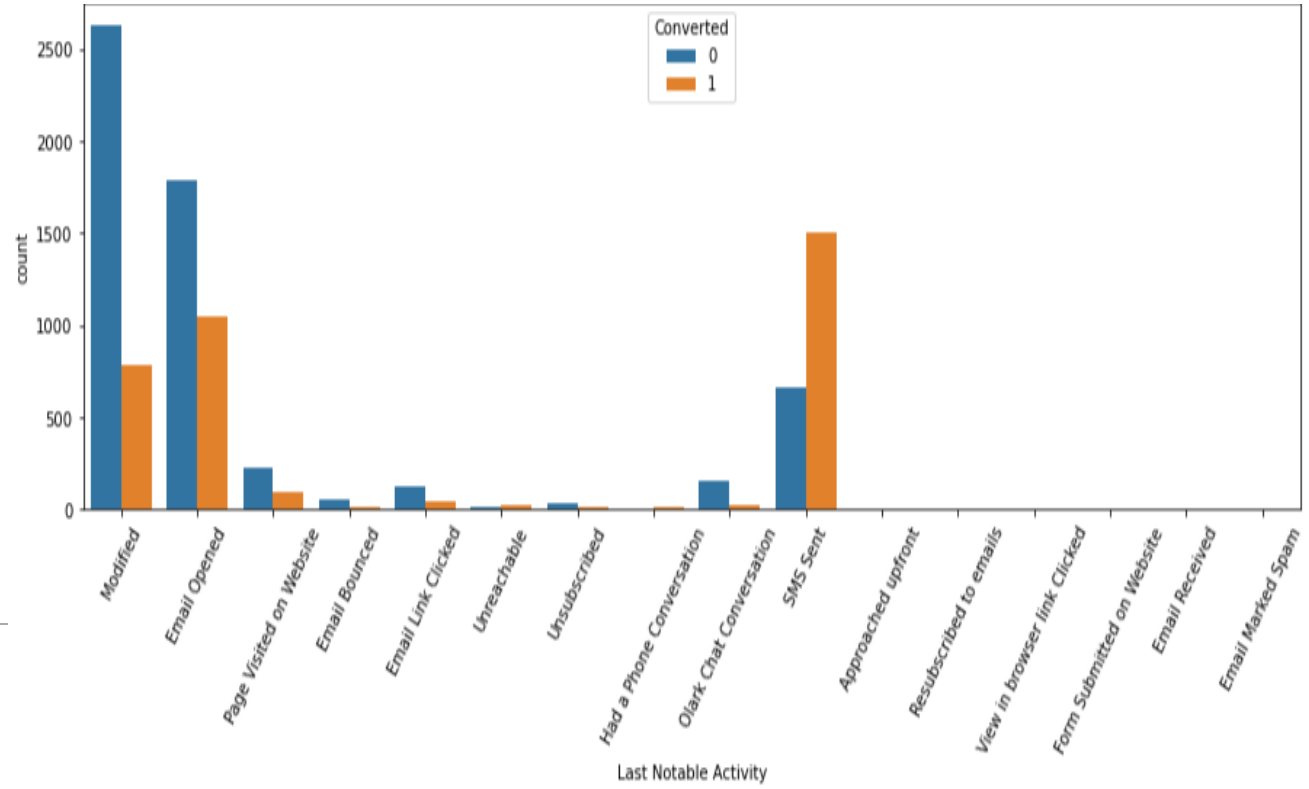
EDA Summary- Specialization



- For most leads , the specialization is not captured. Next best specialization is mostly IT project mgt followed by Travel & tourism very similar to Human Resource Management,

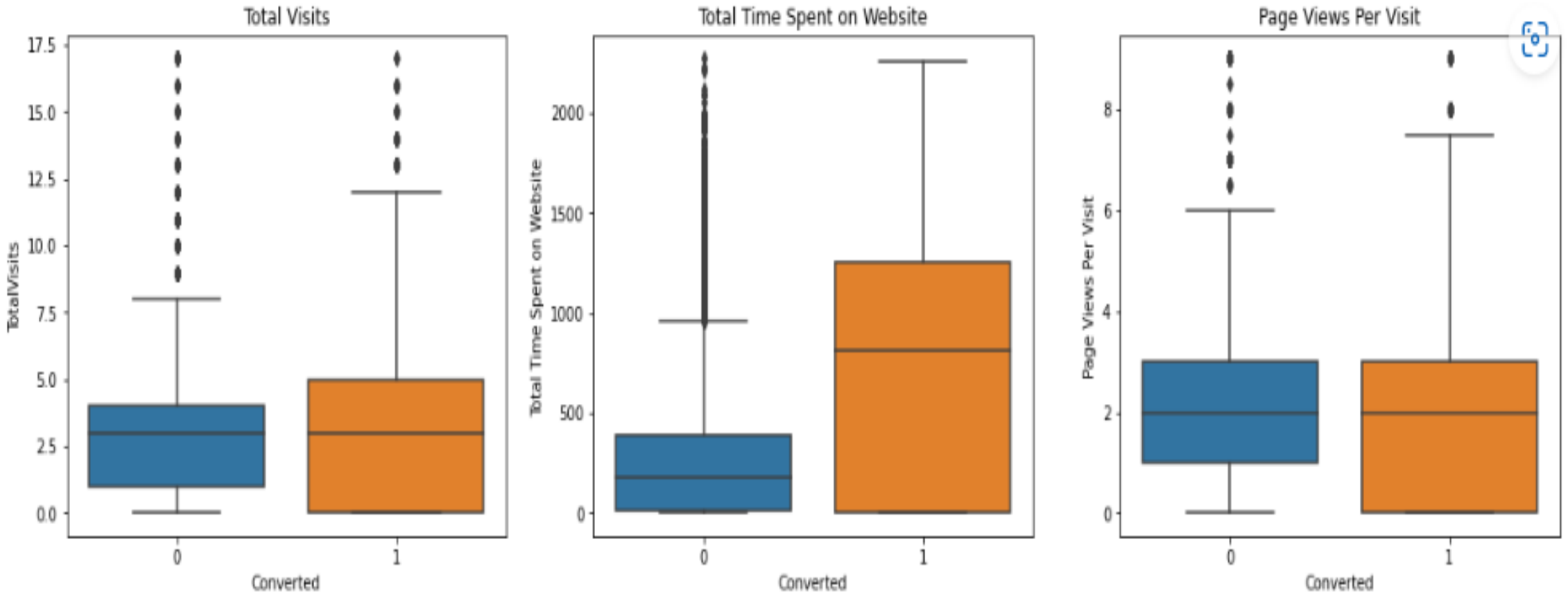
EDA Summary- Last Activity





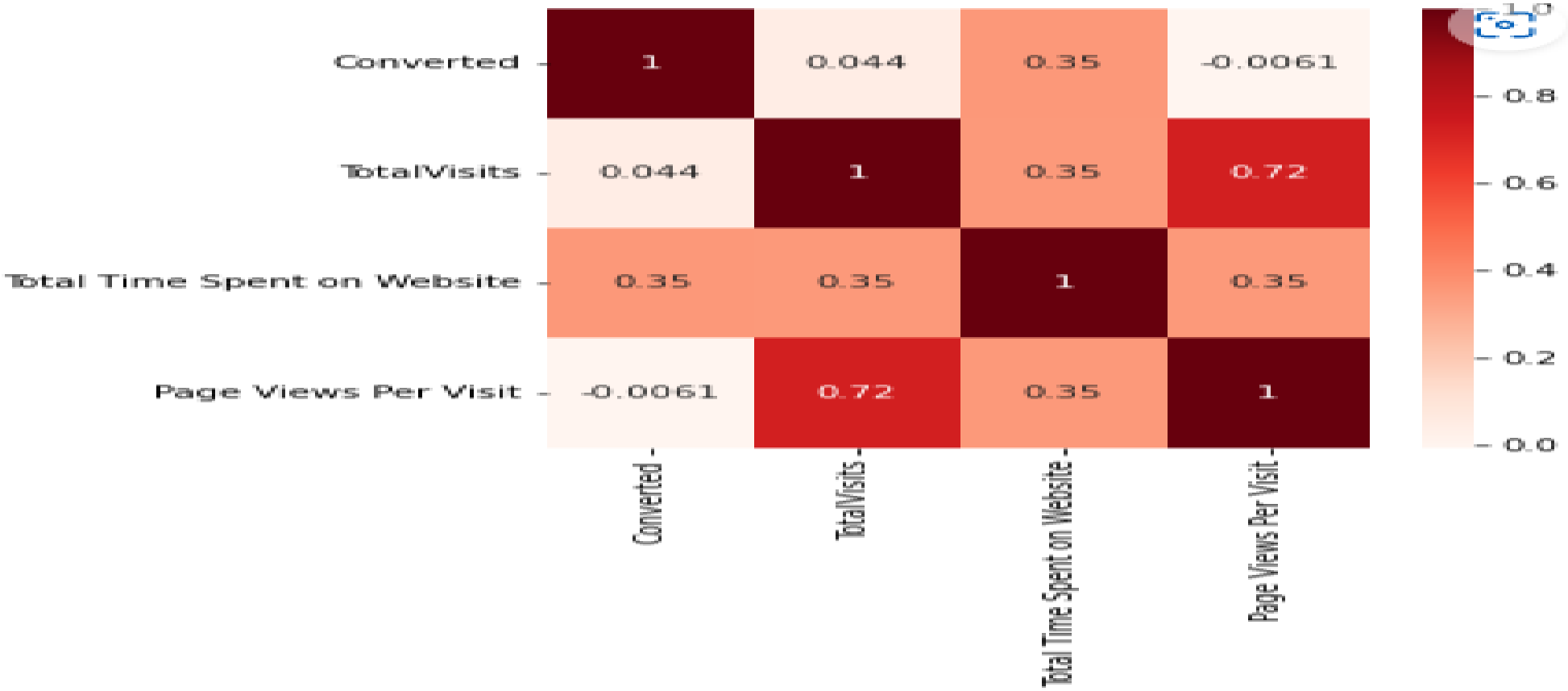
- Maximum leads are generated having last activity as Email opened but conversion rate is not too good. SMS sent as last activity has high conversion rate.
- Also observed , Maximum leads are generated having last activity as sms sent.

EDA Summary- Numerical Variables



- The numerical variables needed outlier treatment
- Leads spending more time on the website tend to be converted.

EDA Summary- Numerical Variables (correlation metrix)



- The numerical variables needed outlier treatment
- Leads spending more time on the website tend to be converted.

Model Building

- Importing the necessary python libraries to perform logistic regression
- Splitting the data set into Train and Test (30-70 ratio) using the train_test_split() function
- Feature Scaling through Min max scaler
- Finding the most relevant features for further analysis through RFE using sklearn with 15 variables as output

```
col # these are the list of attributes as selected as part of RFE

Index(['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit',
       'Lead Origin_lead add form', 'Lead Source_direct traffic',
       'Lead Source_welingak website', 'Do Not Email_yes',
       'Last Activity_olark chat conversation',
       'What is your current occupation_housewife',
       'What is your current occupation_working professional',
       'Last Notable Activity_email link clicked',
       'Last Notable Activity_email opened', 'Last Notable Activity_modified',
       'Last Notable Activity_olark chat conversation',
       'Last Notable Activity_page visited on website'],
      dtype='object')
```

- Running the logistic model on the test data set for the iteratively and checking the p-value & VIF in each step until optimal is reached (variables are iteratively removed whose p value>0.05 or VIF >5)
- Predictions are made on the test data set . With the current cut off as 0.5 we have around 81% accuracy, sensitivity of around 70% and specificity of around 87%. But this may not be optimal . Hence use the ROC curve to determine the optimal threshold for predictions

Confusion Matrix and Accuracy

- **'Confusion Matrix' -** shows a comparison of the predicted and actual labels
- Metrics used to evaluate the performance of the logistic regression -
  - ✓ Accuracy
  - ✓ Sensitivity, specificity and the ROC curve
  - ✓ Precision and Recall

- TP = confusion[1,1] # true positive
- TN = confusion[0,0] # true negatives
- FP = confusion[0,1] # false positives
- FN = confusion[1,0] # false negatives

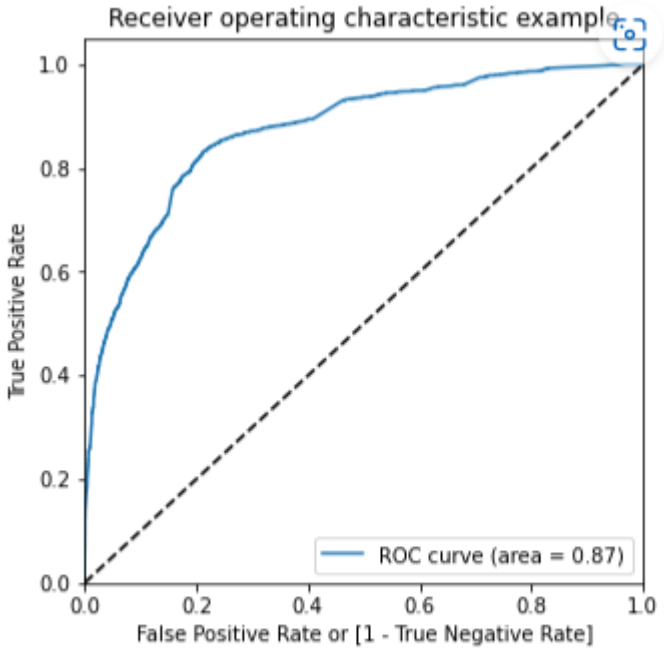
- Sensitivity =TP / float(TP+FN)
- Specificity - TN / float(TN+FP)
- Precision =TP / TP + FP
- Recall -TP / TP + FN

- ROC Curve –

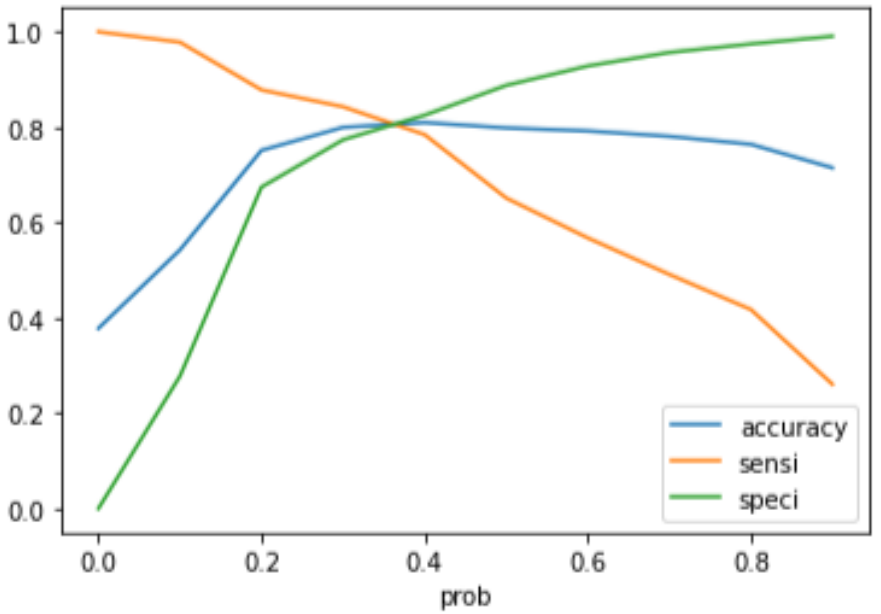
✓ It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
✓ The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
✓ The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity

ROC Curve- Find Optimal Probabilities



Note - Area under ROC is 0.87 ,pretty decent value




```
Optimal cutoff is 0.35
```

- Optimal cutoff for our problem statement was 0.42 instead of 0.5
- Now the same cut off is used and the model is run against the train data set to find the optimal results
- On running the model on train data set , with 0.42 as probability cut off , we observed a Precision around 72% and Recall around 78% on the test dataset

Key Take Aways

1. The Optimum cutoff is around 0.42 with precision of nearly 72% and recall of around 78%
2. Attributes that mattered the most for lead conversions -
   Total number of visits , total time spent on the website
   when the lead source came from Google, direct traffic , organic search .
   when the last activity was sms, olark chat conversion
   When the lead origin is Lead add format
   current occupation is working professional

3. Using such parameters , the company could actually set up there business model in order to have more leads converted

4. Targeted marketing/customized telephonic conversations could be made to the people who have done the following:

- Spent a lot of time in the website & keep returning back to the website multiple times
- Last activity through SMS or Olarkchat
- Targeted advertisements for working professionals
- Discount coupon codes for who has at least initiated the chat or sent SMS