# Lead Scoring case study

# Problem statement

- Create a model in such a way that the customers with high lead score have higher conversion chance and low lead score have lower conversion chance. The ballpark of the target lead conversion rate is around 80%.
- Also the model should be able to adjust if the company's requirement changes in near future.

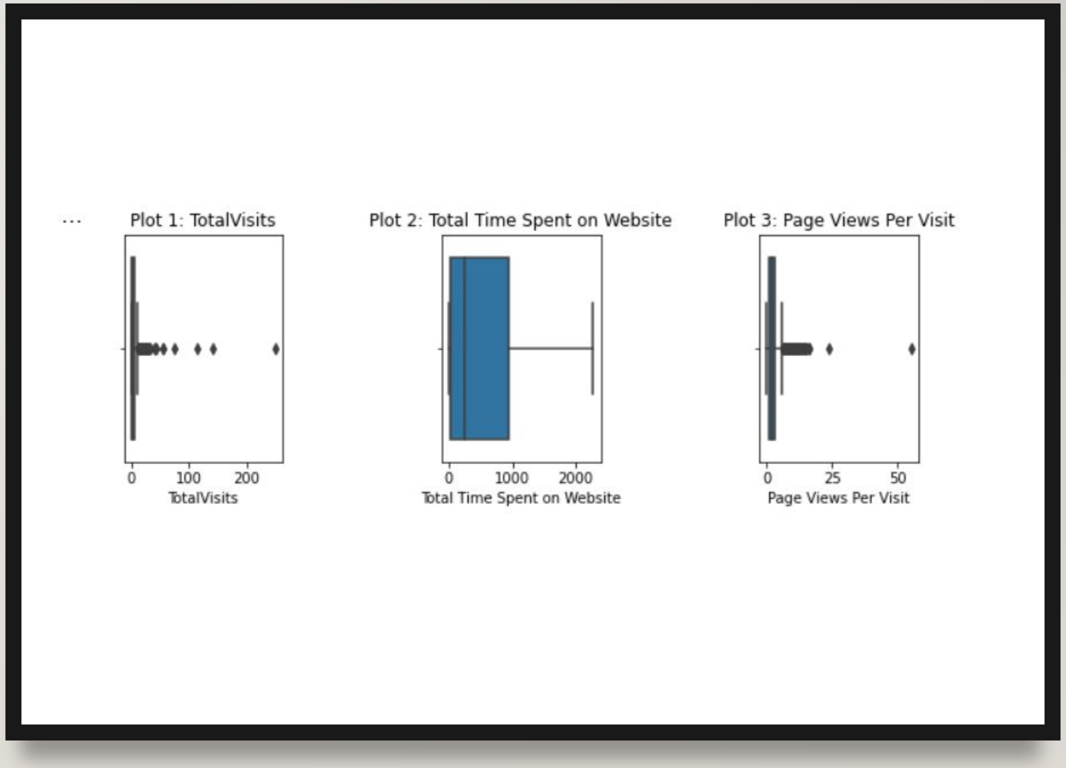
# Exploratory Data Analysis

- Data balance check
- Data cleaning
- Univariate and Bivariate Analysis of categorical variables
- Exploring Numeric variables and  Outlier Handling
- Dummy Variable Creation

We did EDA with all the above mentioned steps and with the help of those insights we created Train and Test split .

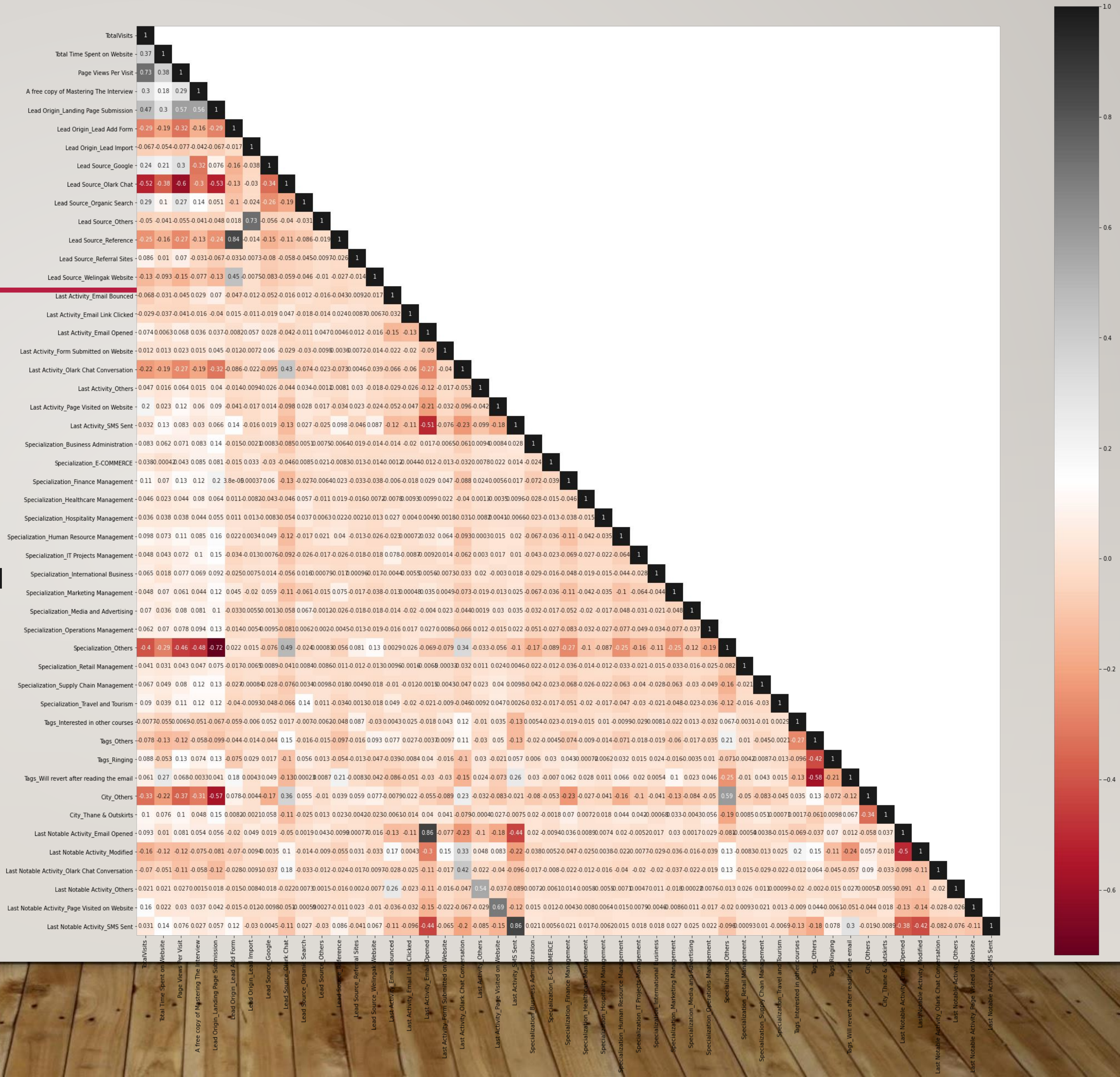
# Analysis approach

- We started our analysis with our cleaned dataset by converting all the binary variables to ‘0’ and ‘1’ and multiple categories into dummy variables.
- Next, we checked the outliers of the dataset. The visualization of those outliers we can see on the graph attached on the right side.
- Outliers in logistic model is very sensitive hence we need to deal with it without losing our valuable information. This can be achieved by creating bins. Hence, we did it



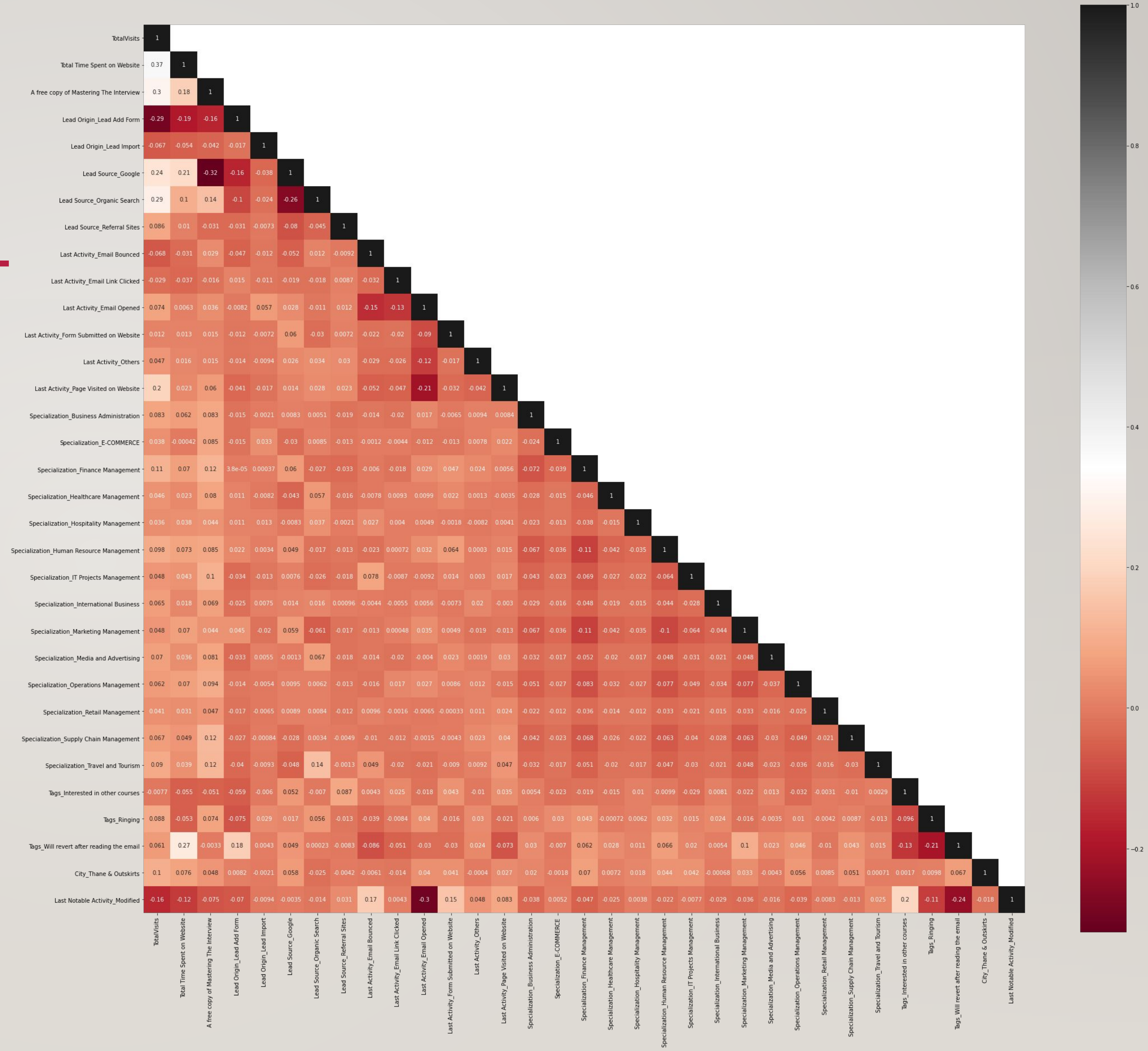
# Correlation

- After fixing the outliers and dummy creation we proceed with our next step of analysis which is data preparation.
- We split the dataset into train and test set and do standardization on the features.
- Standardization is required in order to keep all the variables in same scale which will help us in computation in more efficient way.
- Checked the correlation of the dataset. Attached heatmap is showing the correlation of all features present in the dataset. d) There are some high correlations in the heatmap which we dropped.



# Correlation

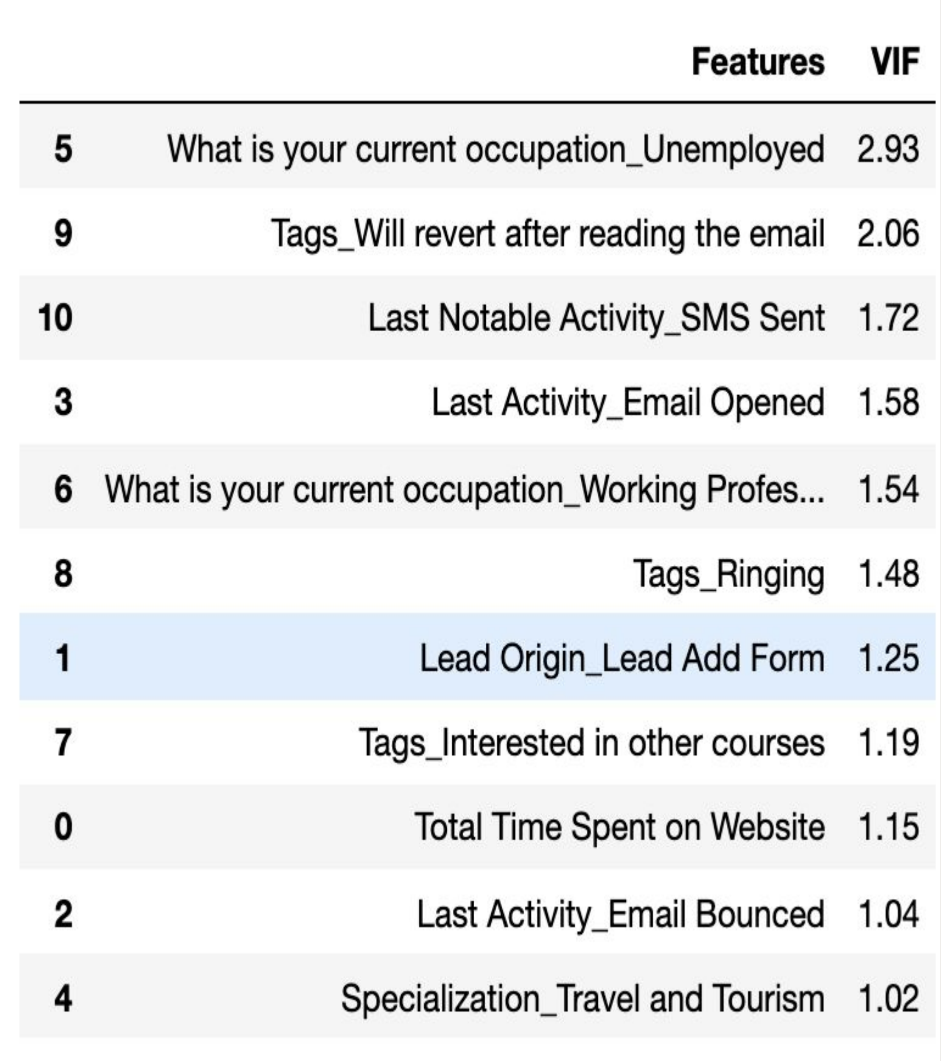
- After dropping those high correlations features, we plotted again a heatmap to check and it was confirmed that those highly correlated variables were dropped.
- 
- There are still few left, but we will check them after creating our model to verify how much they are impacting, as from the plot on the right it is not quite understandable which variable is having high correlation



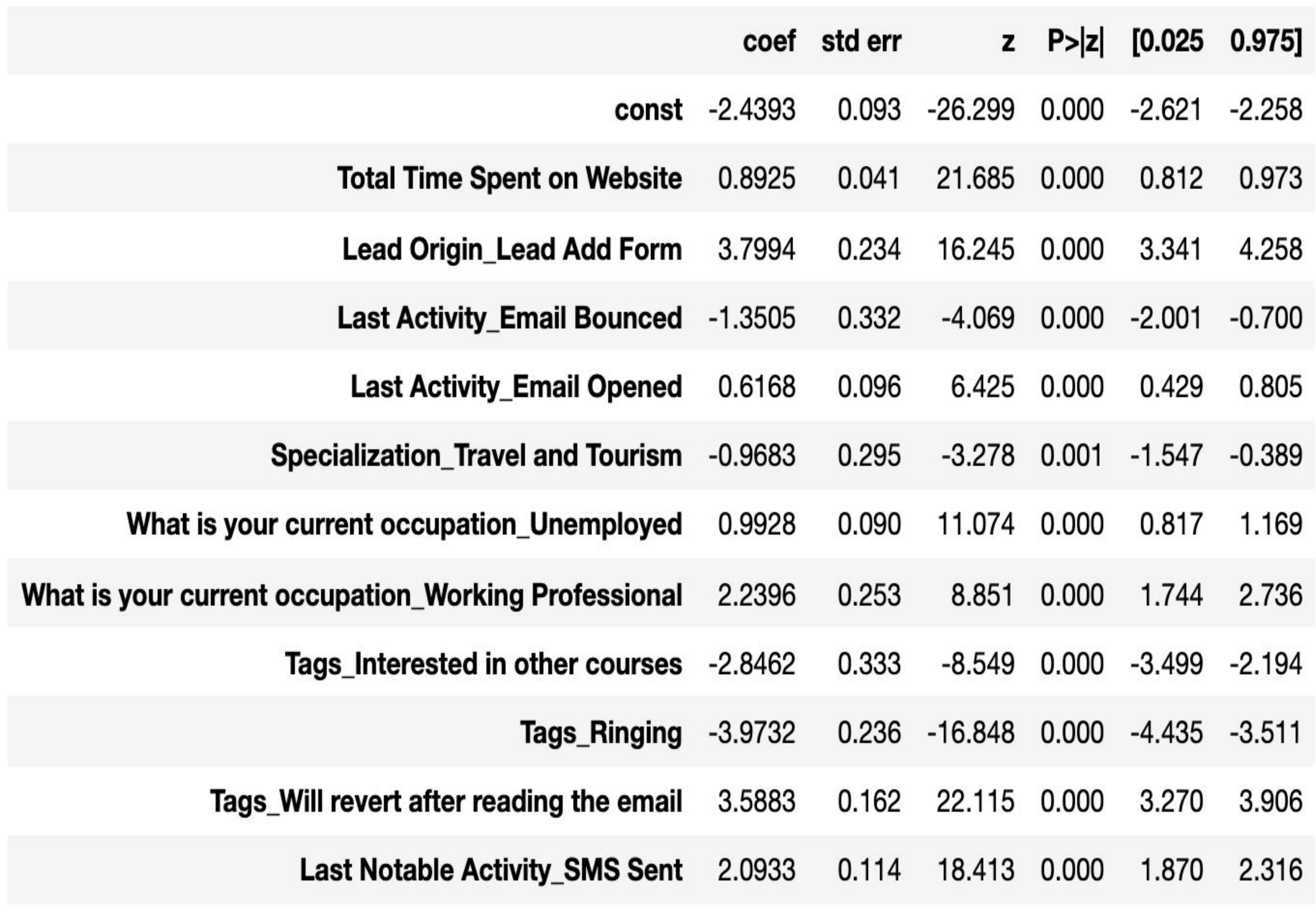
# Building a Model - RFE

- We build a model with all the features included and found there were many insignificant variables present in our model.
- We need to drop them, but we can't do it one by one as it is time consuming and not an efficient way to do so.
- Hence, we started with RFE method to deduct those insignificant variables. We choose with RFE count 15.
- We started creating our model with rfe count 15 and went dropping variables one by one until we reach the point where the model is having all significant variables and low VIF values.
- Now we evaluated our model by first predicting it. We created new dataset with original converted values and the prediction values.

# Final Model Visualization with VIF



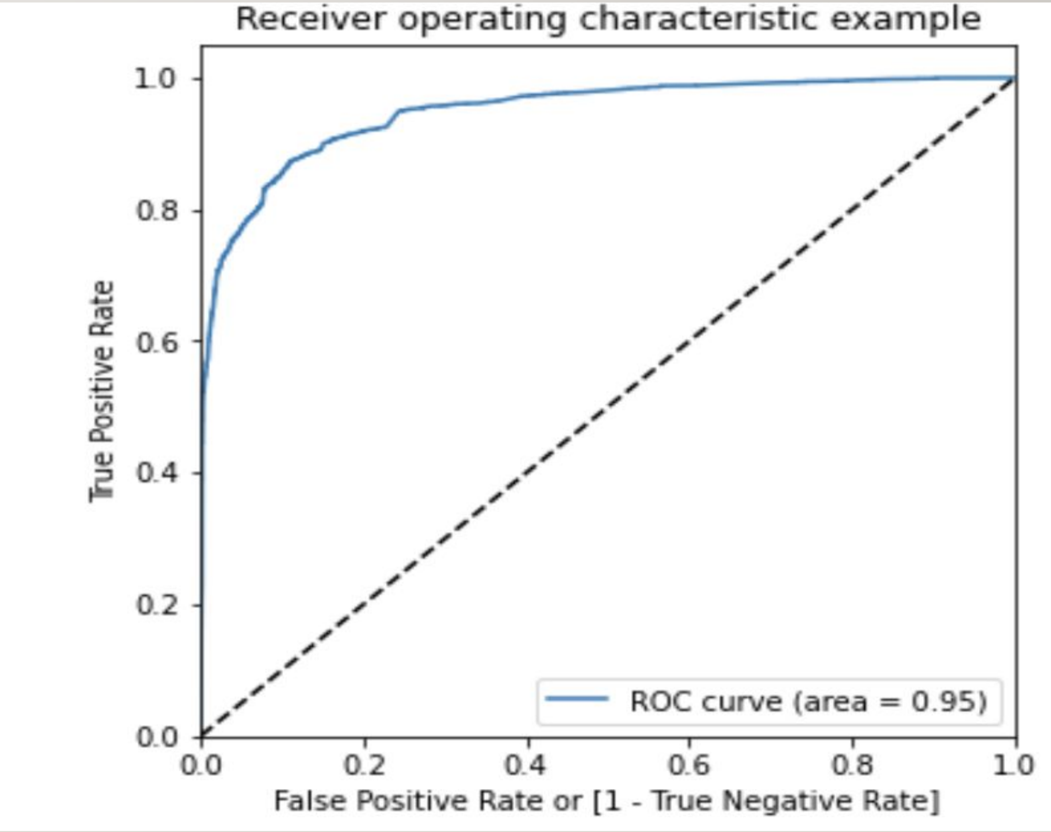
| | Features | VIF |
|---|---|---|
| 5 | What is your current occupation_Unemployed | 2.93 |
| 9 | Tags_Will revert after reading the email | 2.06 |
| 10 | Last Notable Activity_SMS Sent | 1.72 |
| 3 | Last Activity_Email Opened | 1.58 |
| 6 | What is your current occupation_Working Profes... | 1.54 |
| 8 | Tags_Ringing | 1.48 |
| 1 | Lead Origin_Lead Add Form | 1.25 |
| 7 | Tags_Interested in other courses | 1.19 |
| 0 | Total Time Spent on Website | 1.15 |
| 2 | Last Activity_Email Bounced | 1.04 |
| 4 | Specialization_Travel and Tourism | 1.02 |



| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -2.4393 | 0.093 | -26.299 | 0.000 | -2.621 | -2.258 |
| Total Time Spent on Website | 0.8925 | 0.041 | 21.685 | 0.000 | 0.812 | 0.973 |
| Lead Origin_Lead Add Form | 3.7994 | 0.234 | 16.245 | 0.000 | 3.341 | 4.258 |
| Last Activity_Email Bounced | -1.3505 | 0.332 | -4.069 | 0.000 | -2.001 | -0.700 |
| Last Activity_Email Opened | 0.6168 | 0.096 | 6.425 | 0.000 | 0.429 | 0.805 |
| Specialization_Travel and Tourism | -0.9683 | 0.295 | -3.278 | 0.001 | -1.547 | -0.389 |
| What is your current occupation_Unemployed | 0.9928 | 0.090 | 11.074 | 0.000 | 0.817 | 1.169 |
| What is your current occupation_Working Professional | 2.2396 | 0.253 | 8.851 | 0.000 | 1.744 | 2.736 |
| Tags_Interested in other courses | -2.8462 | 0.333 | -8.549 | 0.000 | -3.499 | -2.194 |
| Tags_Ringing | -3.9732 | 0.236 | -16.848 | 0.000 | -4.435 | -3.511 |
| Tags_Will revert after reading the email | 3.5883 | 0.162 | 22.115 | 0.000 | 3.270 | 3.906 |
| Last Notable Activity_SMS Sent | 2.0933 | 0.114 | 18.413 | 0.000 | 1.870 | 2.316 |

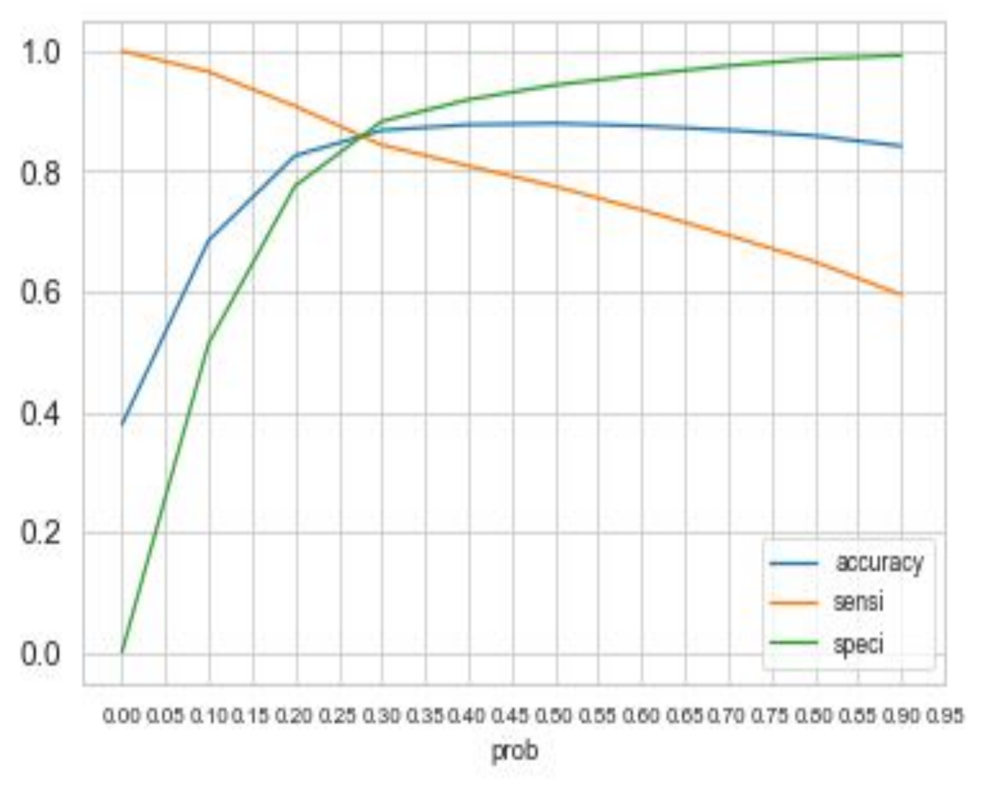
# Evaluating the model

- After building the final model making prediction on it(on train set), we created ROC curve to find the model stability with auc score(area under the curve) As we can see from the graph plotted on the right side, the area score is 0.95 which is a great score.
- And our graph is leaned towards the left side of the border which means we have good accuracy



# Finding the optimal cutoff point

- Now, we have created range of points for which we will find the accuracy, sensitivity and specificity for each points and analyze which point to chose for probability cutoff.
- We found that on 0.27 point all the score of accuracy, sensitivity and specificity are in a close range which is the ideal point to select and hence it was selected.
- To verify our answer we plotted this in a graph – line plot which is on the right side and we stand corrected that the meeting point is close to 0.27 and hence we choose 0.27 as our optimal probability cutoff

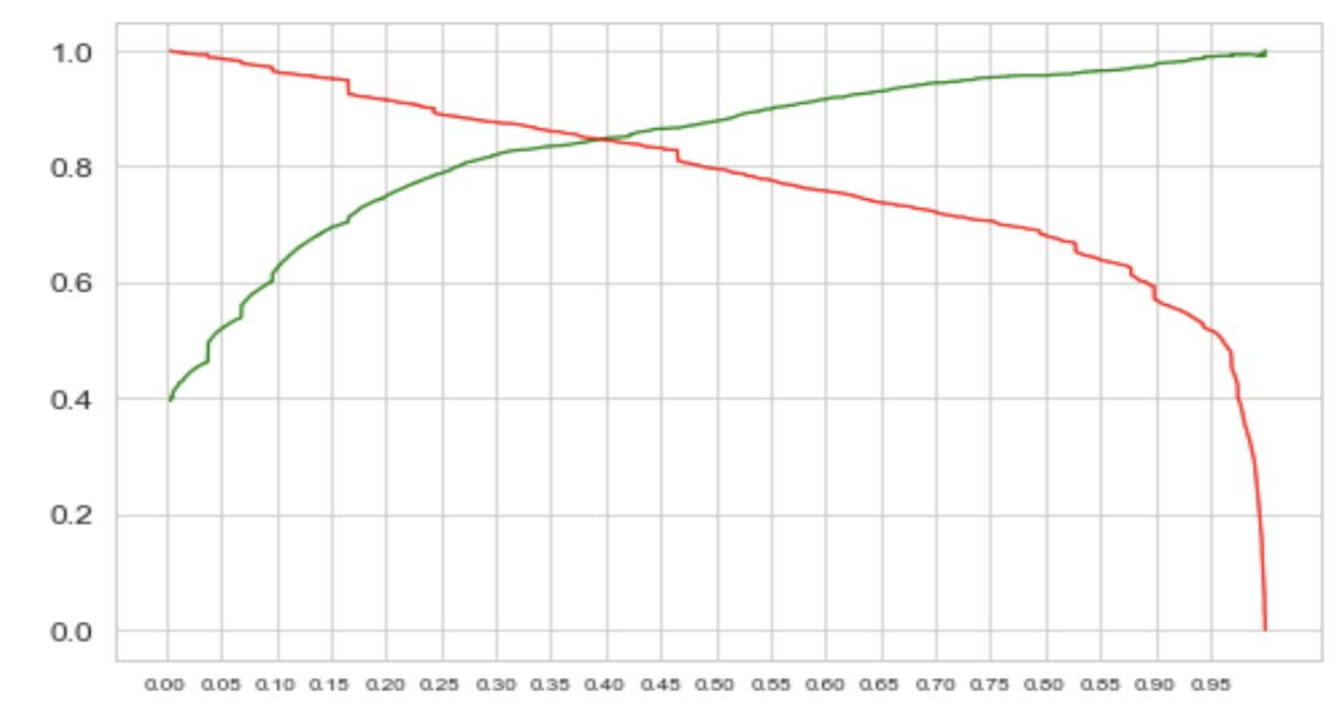


# Precision and Recall

- We used this cutoff point to create a new column in our final dataset for predicting the outcomes.
- After this we did another type of evaluation which is by checking Precision and Recall
- As we all know, Precision and Recall plays very important role in build our model more business oriented and it also tells how our model behaves.
- Now, recall our business objective - the recall percentage I will consider more valuable because it is okay if our precision is little low which means less hot lead customers but we don't want to left out any hot leads which are willing to get converted hence our focus on this will be more on Recall than Precision.
- i.e We get more relevant results - as many as hot lead customers from our model .

# Precision and Recall tradeoff

- We created a graph which will show us the tradeoff between Precision and recall. We found that there is a trade off between Precision and Recall and the meeting point is approximately at 0.4.



# Prediction on test set

- Before predicting on test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set and the new predictions values were saved in new dataframe.
- After this we did model evaluation i.e. finding the accuracy, precision and recall.
- The accuracy score we found was 0.88, precision 0.85 and recall 0.86 approximately.
- This shows that our test prediction is having accuracy , precision and recall score in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of converted, low the lead score lower the chance of getting converted.

# Conclusion

- Valuable Insights - The Accuracy, Precision and Recall/Sensitivity are showing promising scores in test set which is as expected after looking the same in train set evaluation steps. Means the recall is having high score value than precision which is acceptable for business needs.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state. Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
- a) Total time spend on website
- b) Lead Origin_Lead Add Form
- c) What is your current occupation _ working professional