LEAD SCORING CASE STUDY

SUBMITTED BY

CHANDANA SARANGI

SAI KRISHNA & UMESH PANDEY

Business Objective:

To Help X Education Select Most Promising Leads (Hot Leads), i.e. The Leads That Are Most Likely To Convert Into Paying Customers.

- **Selection of hot leads**
- **Focused marketing**
- **Higher lead conversion rate**

GOALS OF THE CASE STUDY:

- **To build a Logistic Regression model that assigns lead scores value between 0 and 100 to each of the leads which can be used by the company to target potential leads such that the customers with higher lead score usually have a higher conversion chance and vice versa.**
  - **Create a Logistic Regression model to predict the Lead Conversion probabilities for each lead.**
  - **Decide on a Probability Threshold value above which a lead will be predicted as converted, whereas not converted if it is below it.**
  - **Multiply the Lead Conversion probability to arrive at the Lead Score value for each lead.**

APPROACH:

- **Reading & understanding the data**
- Data cleaning
- EDA
- Feature scaling
- Splitting the data into test & train dataset
- Prepare the data for modelling
- Model building
- Model evaluation-specificity & Sensitivity or precision recall
- Making predictions on the test
- Assigning lead score
- Hot leads Determination
- Feature Importance Determination

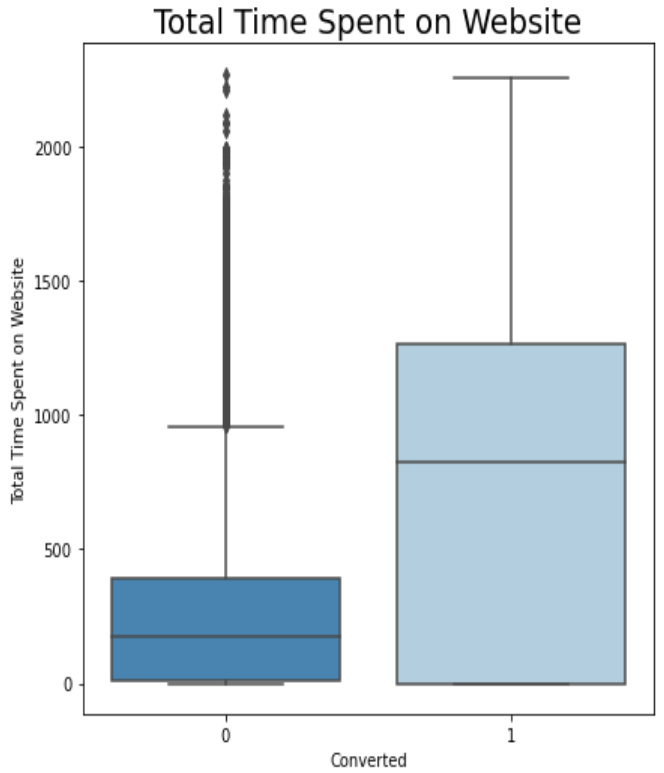
Data Cleaning:

- **There were few columns which were having a high number of null values which are dropped straight away. Further, we can not eliminate all columns which might be useful and were having a great impact on our model but are having a strong possibility of high number of null values so we can treat those columns by imputing with mean and median by observing the type of the variables i.e. (continuous or categorical). The outliers which were found during the analysis has also been removed. After those process up to 98% data has been retained and on this cleaned data the analysis was performed.**

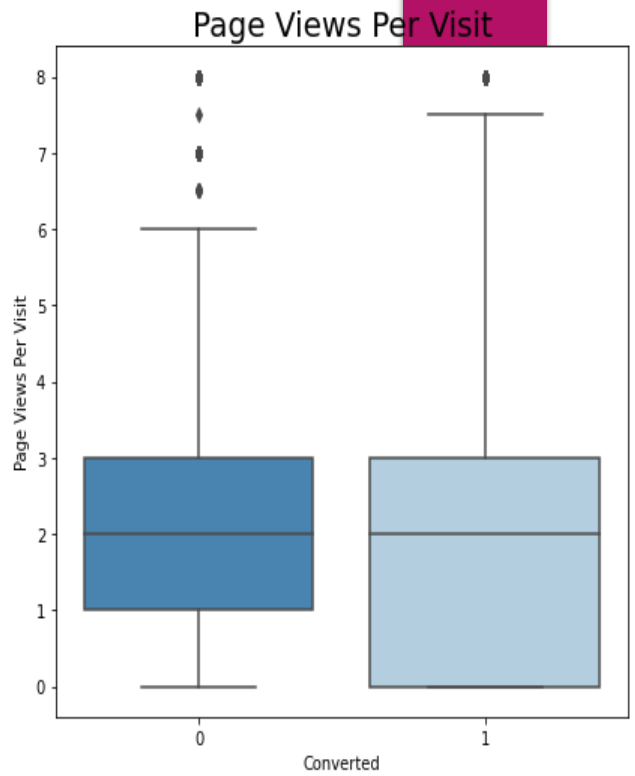
Exploratory Data Analysis (EDA):

- **EDA was performed on the cleaned data by plotting different types of plots and analysing both the variables which is continuous and categorical. Univariate analysis was done against the target variable for better understanding. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable and it also describes each variable on its own.**
- **Some of the insights are as follows: People spending more time are promising Leads, The Lead Origin- Landing Page Submission has the highest conversion rate among others, Google has the highest conversion rate, leads whose Last Activity was SMS sent had the best conversion rate, Lead from Specialization who are unknown/Select columns has the highest rate of conversion, Person who are working professional has the highest conversion rate comparatively to others.**

Numerical Variable Analysis



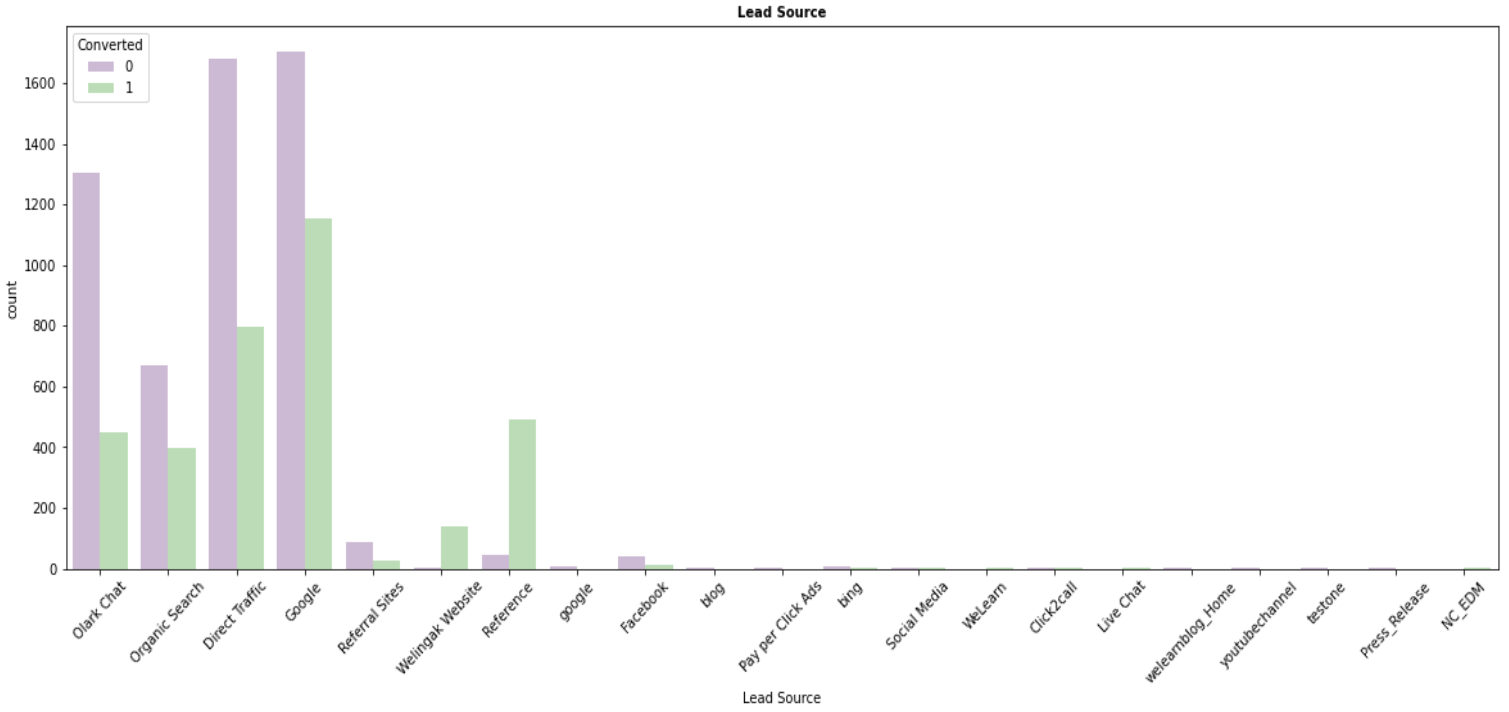
Total time spent on Website



Pages views per visit

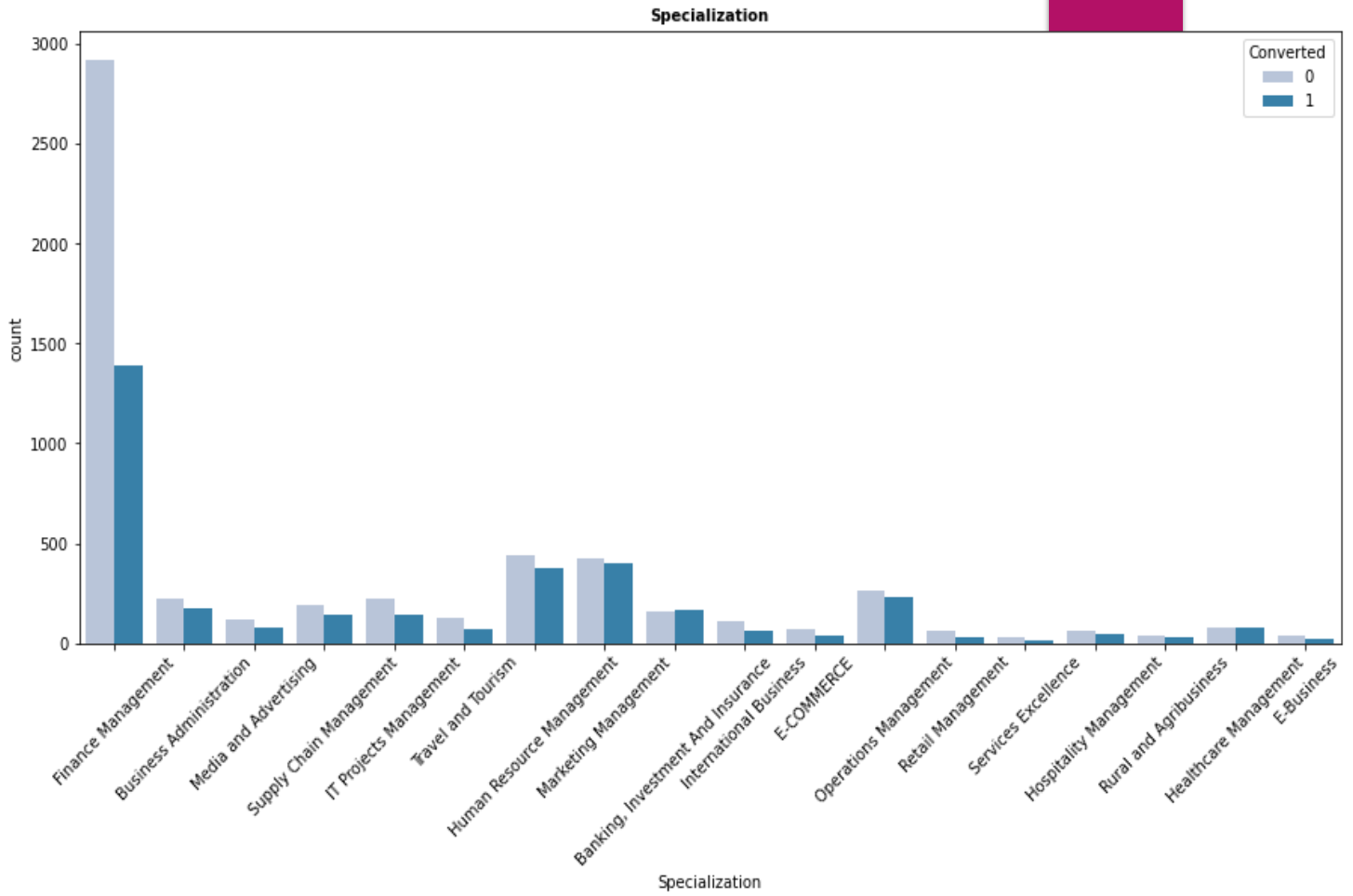
Lead Source

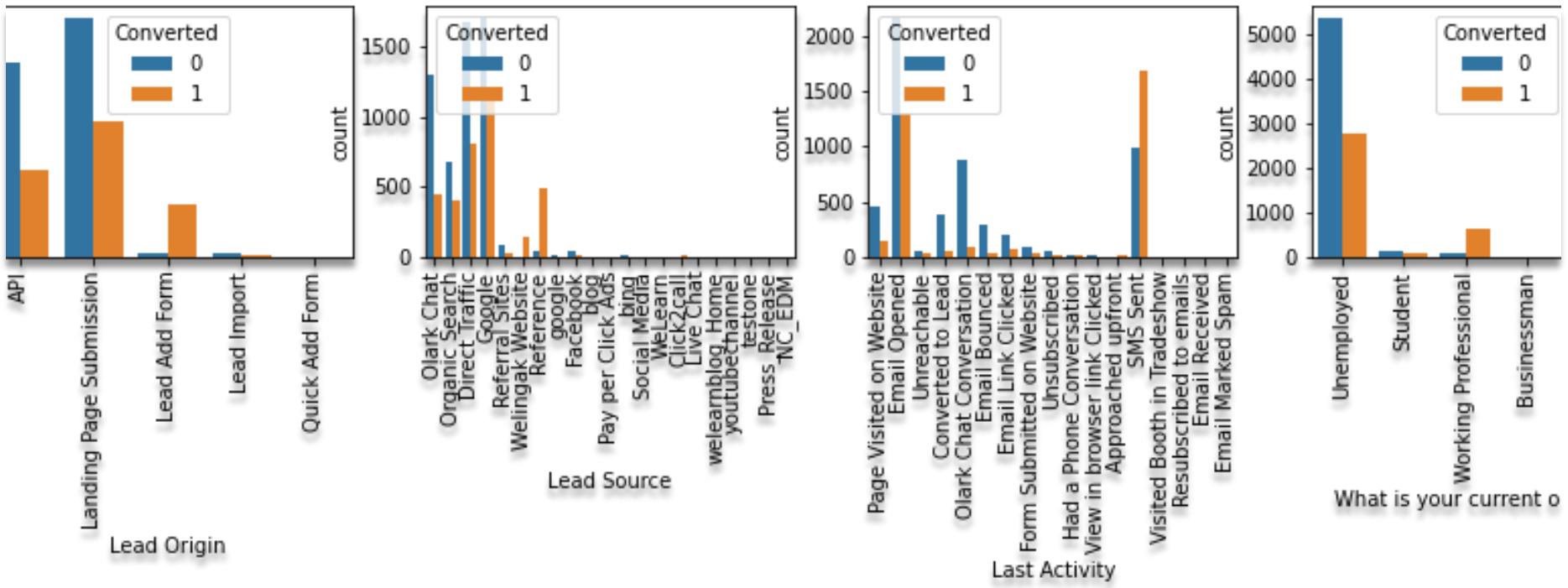
- Google has the highest conversion rate



Specialization

. Leads from Finance Management has the highest rate of conversion.



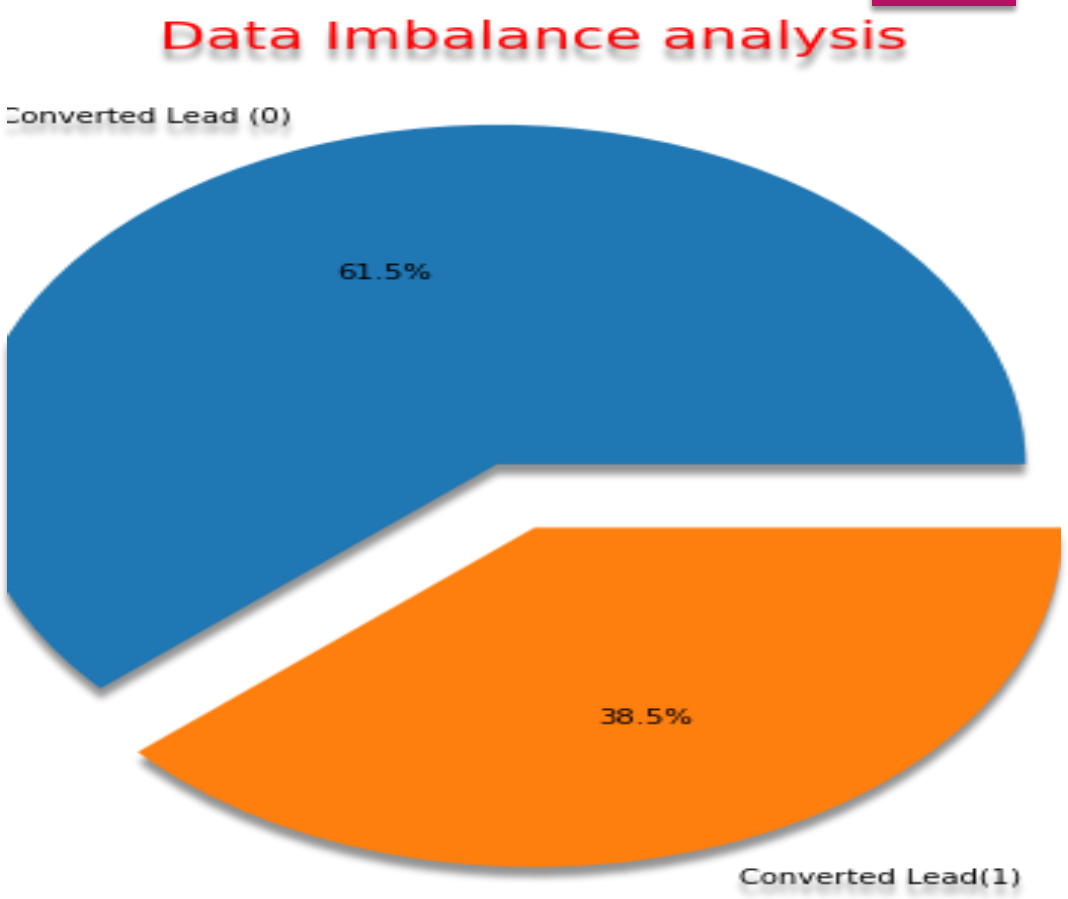


Categorical Column Analysis:

The majority of Converts are Professionals in the Workplace. Due to the job search and career aspects, people who are unemployed also have the highest conversion rates. The most leads are generated by "API" and "Landing Page Submission," although they have lower conversion rates. Less leads are generated through the "Lead Add Form," but the conversion rate is excellent.

Data Imbalance Analysis

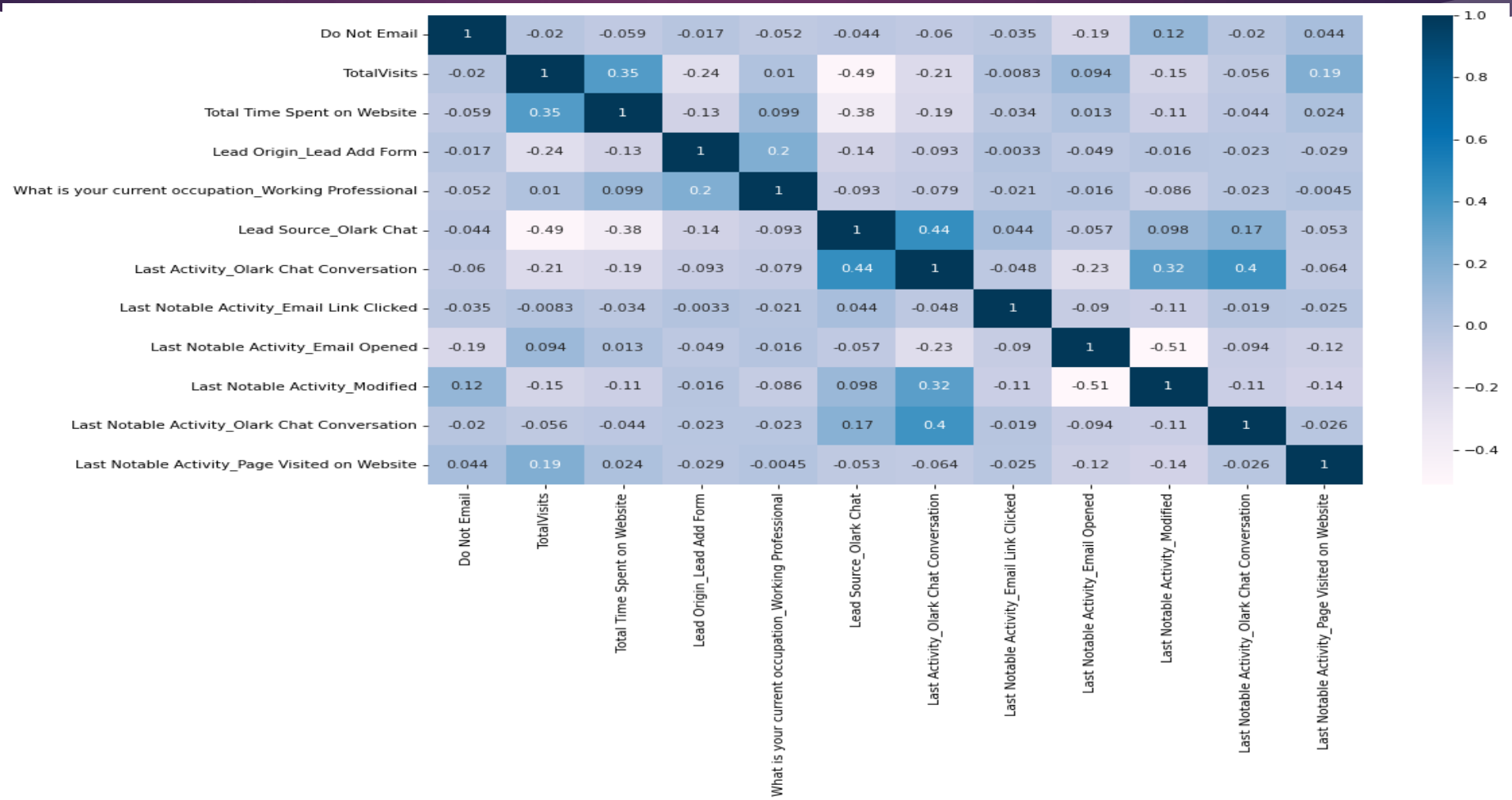
As per the study above, the data is not overly unbalanced. In the data, 61.5% of leads are not converted, whereas 38.5% of leads are converted.



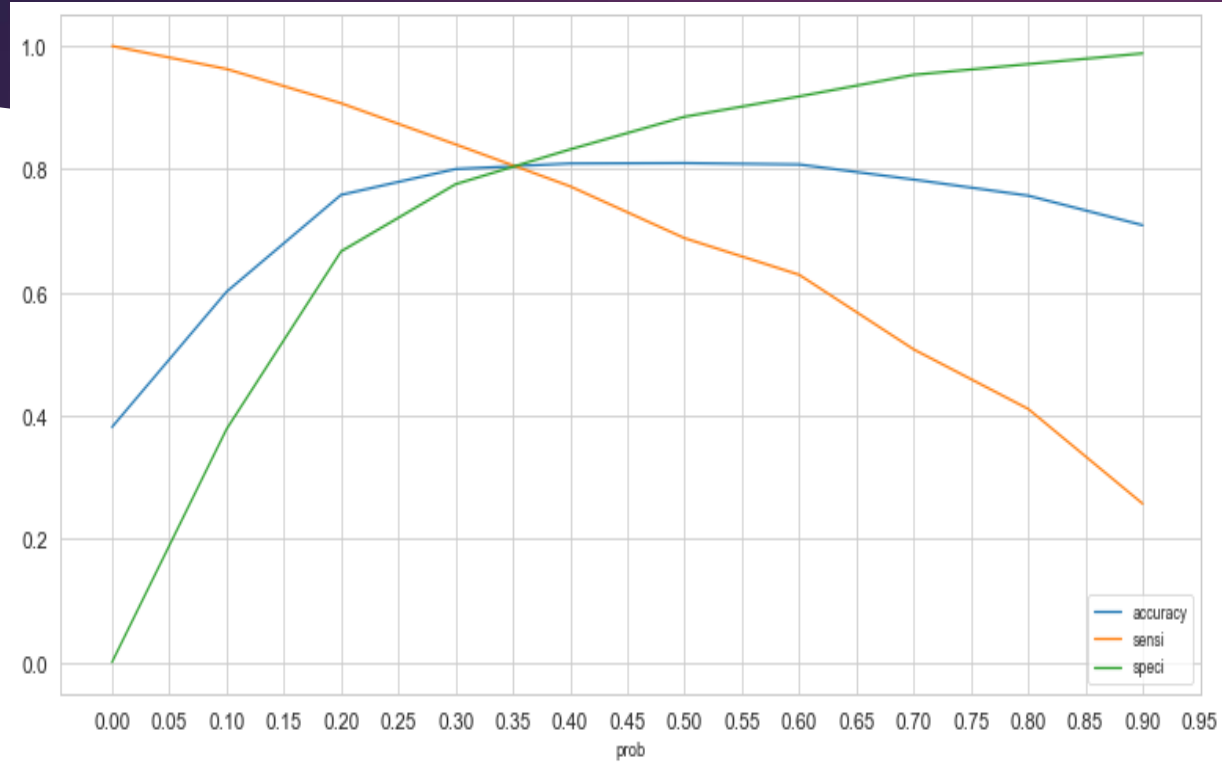
Model Building:

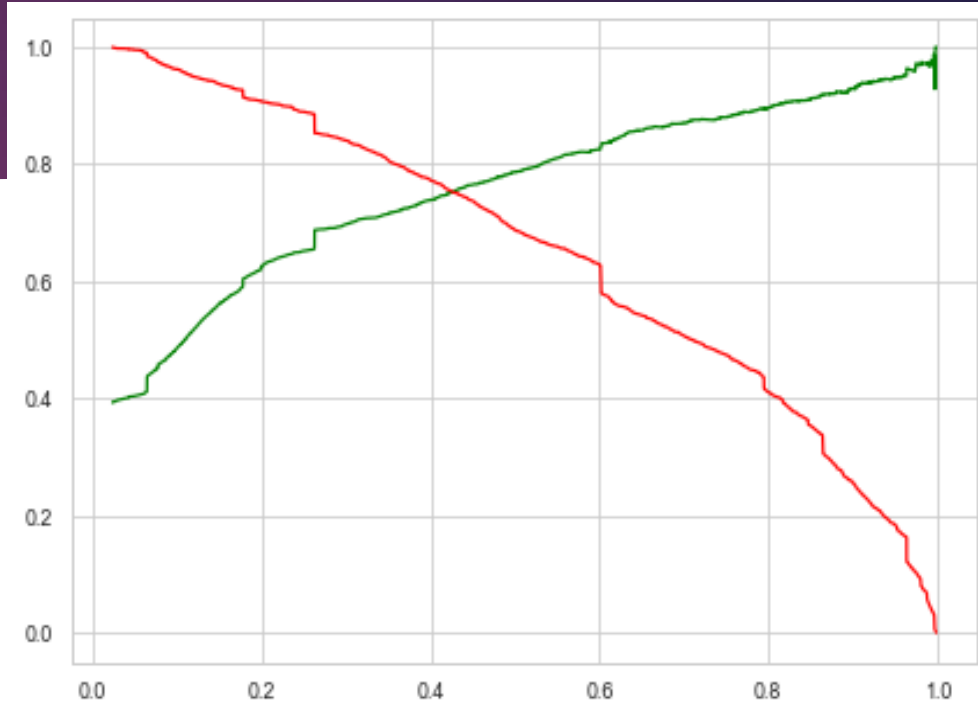
- Dummy variables were created for categorical columns and scaling was done. **In simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.** Scaling was carried out in order to bring all the features into a comparable range. Then we achieved the splitting of train and test dataset with 70% and 30%.
- Feature selection was applied using RFE technique and then the elimination was done according the steps followed for fetching the column having high P-value & VIF, a final model was obtained after occurrence of 4 times until both VIF and p-values reached under acceptable range.

Correlation:



MODEL EVALUATION (TRAIN):

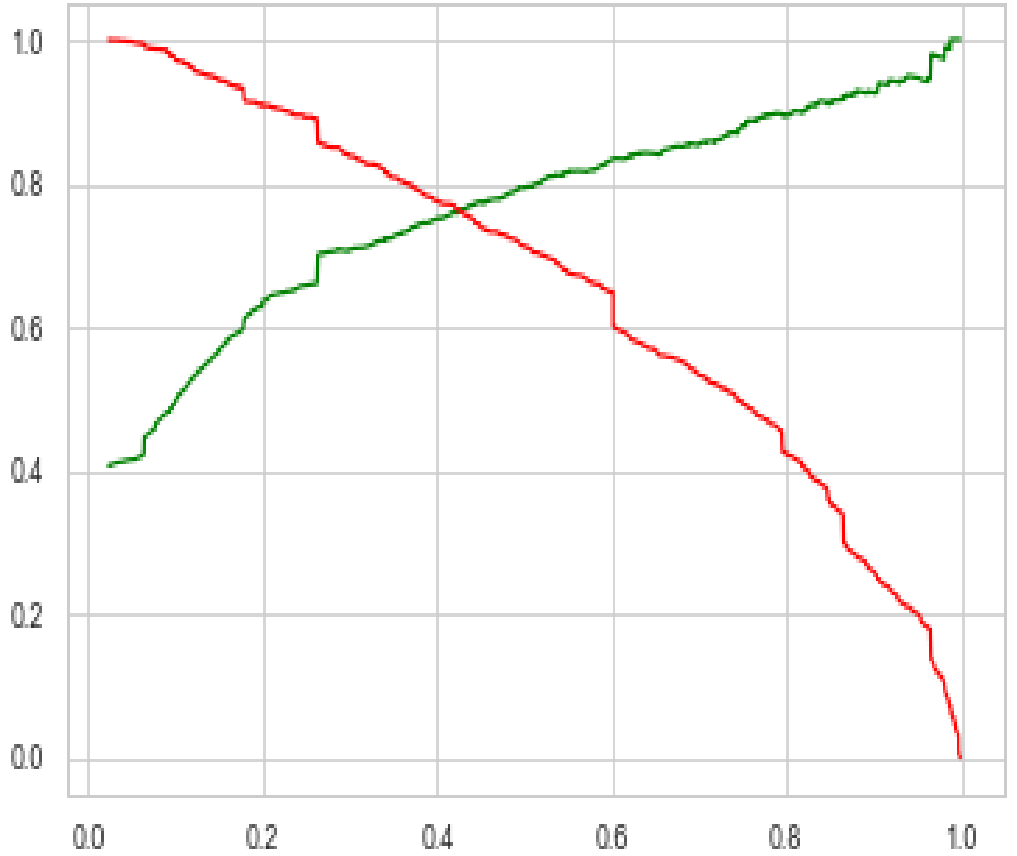




- **Accuracy : 80.33%**
- **Sensitivity : 81.66%**
- **Specificity : 79.50%**
- **Precision : 71.08%**
- **Recall : 81.66 %**

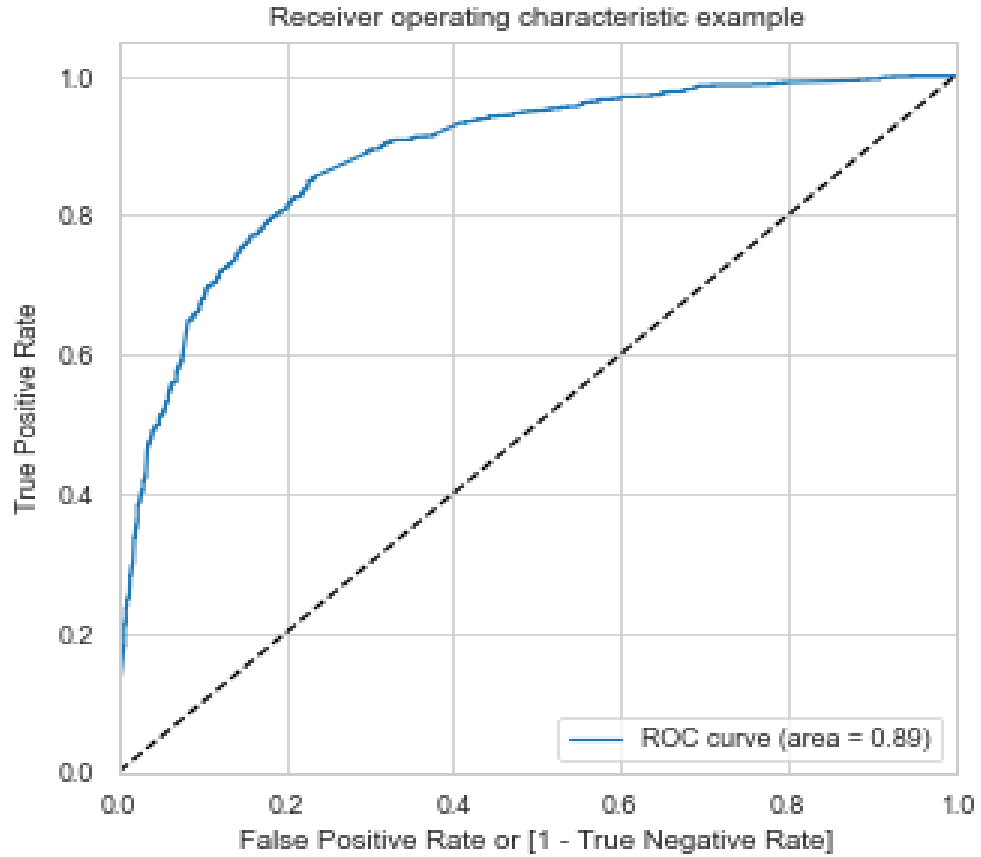
MODEL EVALUATION (TEST):

- Accuracy : 80.36%
- Sensitivity : 81.70%
- Specificity : 79.48%
- Precision : 72.10%
- Recall : 81.70 %



ROC Curve:

The ideal value for the ROC Curve is 1 which indicates a "GOOD Predictive Model," which is what we are obtaining.



Prospect Id	Converted	Converted_Prob	final_Predicted	final_predicted	Lead_Score
37	1	0.829137	1.0	NaN	83
64	1	0.962763	NaN	1.0	96
76	1	0.847543	NaN	1.0	85
77	1	0.977352	NaN	1.0	98
79	1	0.996222	NaN	1.0	100

Determining HOT LEADS with 89% Accuracy & more than 80% Conversion Rate!

Every lead in the Original Dataframe has their Lead Score calculated.

- **To obtain the complete list of leads available, the Train And Test Dataset is concatenated. The Lead Score for each lead is calculated by multiplying the Conversion Probability by 100.**
- **The likelihood that a lead will be converted increases with lead score, and vice versa. Considering that, we had chosen 0.34 as our ultimate Probability cutoff for determining if a lead will convert or not.**

RELATIVE IMPORTANCE OF FEATURES:

Each feature's relative importance is calculated on a scale of 100, with 100 being the highest possible score.

• The model parameters' Coefficient (Beta) Values for each of these features are used to establish their relative relevance.

• Features with High Positive Beta Values are those that have a major impact on a lead's likelihood of conversion.

• In a similar vein, traits with high negative beta values make the least contribution.

IMPORTANCE OF A FEATURE

• The top three factors that have the greatest effects on the likelihood of Lead Conversion

Total Time Spent on Website	100.00
Lead Origin_Lead Add Form	95.79
What is your current occupation_Working Professional	57.02
Lead Source_Olark Chat	32.21
TotalVisits	25.30
Last Activity_Olark Chat Conversation	-25.47
Last Notable Activity_Email Opened	-31.80
Do Not Email	-34.90
Last Notable Activity_Olark Chat Conversation	-38.36
Last Notable Activity_Email Link Clicked	-39.64
Last Notable Activity_Modified	-42.80
Last Notable Activity_Page Visited on Website	-43.38
dtype: float64	

SITUATION 1:
The company has two months of interns. They want to aggressively increase lead conversion. They want to call as many of these potential leads as they can since they want almost all of them to convert.

Solution:

- **The proportion of real conversions that were accurately predicted out of all actual conversions is known as sensitivity. As we previously observed, sensitivity declines as the threshold rises.**
- **High Sensitivity means that practically all leads who are likely to convert will be accurately predicted by our model. Additionally, some of the non-conversions can be overestimated and incorrectly categorised as conversions.**
- **It is a smart move to choose high sensitivity as the organisation has extra staff for two months and wants to increase the lead conversion's aggressiveness. We must select a LOW THRESHOLD VALUE in order to attain great sensitivity.**

SITUATION 2:

The company occasionally meets its quarterly target ahead of schedule. It desires that the sales team concentrate on some fresh work. Therefore, the company's goal is to avoid making phone calls during this time unless they are really necessary.

Solution:

- The percentage of actual non-conversions that were accurately predicted out of all actual non-conversions is known as specificity. As the threshold rises, it rises as well.
- High Specificity means that practically all leads who are not likely to convert will be predicted correctly by our model. While doing so, it can mistakenly categorise some conversions as non-conversions.
- High specificity is an excellent approach because the organisation has already surpassed its quarterly goal and doesn't want to make pointless phone calls.
- It will make sure that only consumers with a very high chance of converting will receive phone calls. We must select a Great THRESHOLD VALUE in order to attain high specificity.

The top 3 factors that Most Influence the Chance of Lead Conversion

1. Total Time Spent on Website
2. Lead Origin_Lead Add Form
3. What is your current occupation_Working Professional

The top three categorical/ dummy variables on which the greatest attention should be paid in order to raise the likelihood of lead conversion are:

1. Total Time Spent on Website
2. Lead Origin_Lead Add Form
3. What is your current occupation_Working Professional
4. Tag_ will revert after reading the email

Who are hot leads potential paying customers to consider ??

- The "Working Professionals" should receive calls from the company because they are more likely to convert.
- Who frequently visits websites or who spends a lot of time on websites can be attracted by making websites more user-friendly and informative.
- You can target people based on their most recent SMS and email opening activity.
- Individuals with the tag "Will revert after reading emails" may be potential targeted leads.
- Last Notable activity Had a phone conversation as the most recent noteworthy activity.

THANK YOU