LEAD SCORING CASE STUDY

Gunaseelan P
Sandeep Pradhan
Manish Kumar

Problem Statement

1. Help an online education company named X Education to identify the potential customers(Hot Leads) from a large dataset.

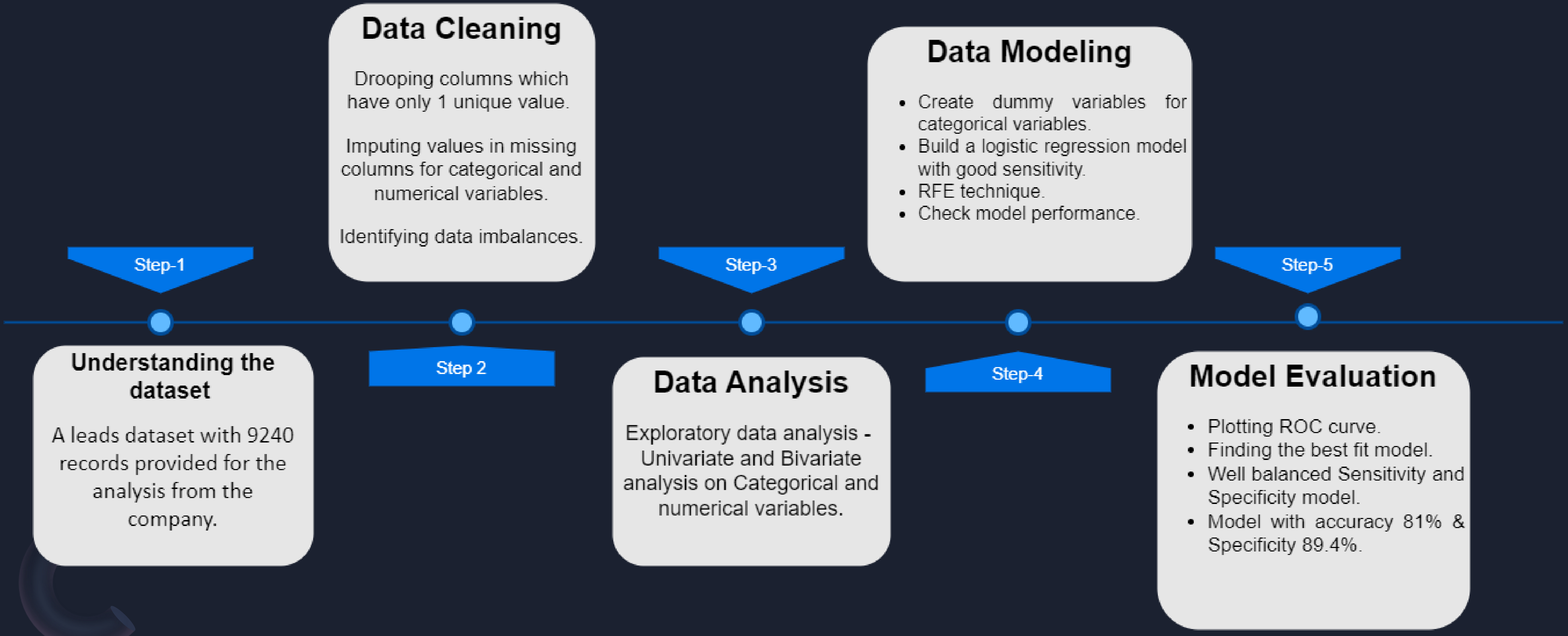
2. The company requires to create a model where in a lead score assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Goal

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

2. Model should be able to adjust to the company's requirement changes in the future also should keep the accuracy also.

Steps Taken for the Analysis and Model Building



Step-1

Understanding the dataset

A leads dataset with 9240 records provided for the analysis from the company.

Step 2

Data Cleaning

Drooping columns which have only 1 unique value.

Imputing values in missing columns for categorical and numerical variables.

Identifying data imbalances.

Step-3

Data Analysis

Exploratory data analysis - Univariate and Bivariate analysis on Categorical and numerical variables.

Step-4

Data Modeling

- Create dummy variables for categorical variables.
- Build a logistic regression model with good sensitivity.
- RFE technique.
- Check model performance.

Step-5

Model Evaluation

- Plotting ROC curve.
- Finding the best fit model.
- Well balanced Sensitivity and Specificity model.
- Model with accuracy 81% & Specificity 89.4%.

1. Read & Understand the given Dataset

- A leads dataset with 9240 records provided for the analysis from the company.
- Total 37 columns are present in the dataset.
- Each record assigned with a unique_ID ('Prospect ID') for identification.
- The current conversion rate of leads is 38.54%
- Initial Analysis show around 18 columns are having Null values, so Data cleaning is required before creating a model.

2. Data Cleaning

- Dropped the columns ('Magazine', 'Receive More Updates About Our Courses', 'I agree to pay the amount through cheque', 'Get updates on DM Content', 'Update me on Supply Chain Content') which are having only 1 unique values.

- Replaced the missing values of the columns with Mean/Mode data- ('Specialization', ''Lead Source'', ''What is your current occupation'', 'TotalVisits', 'Page Views Per Visit’).

- Dropped the columns which are having very less unique values and high data imbalance-('Tags', 'Country', ''What matters most to you in choosing a course'', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations’).

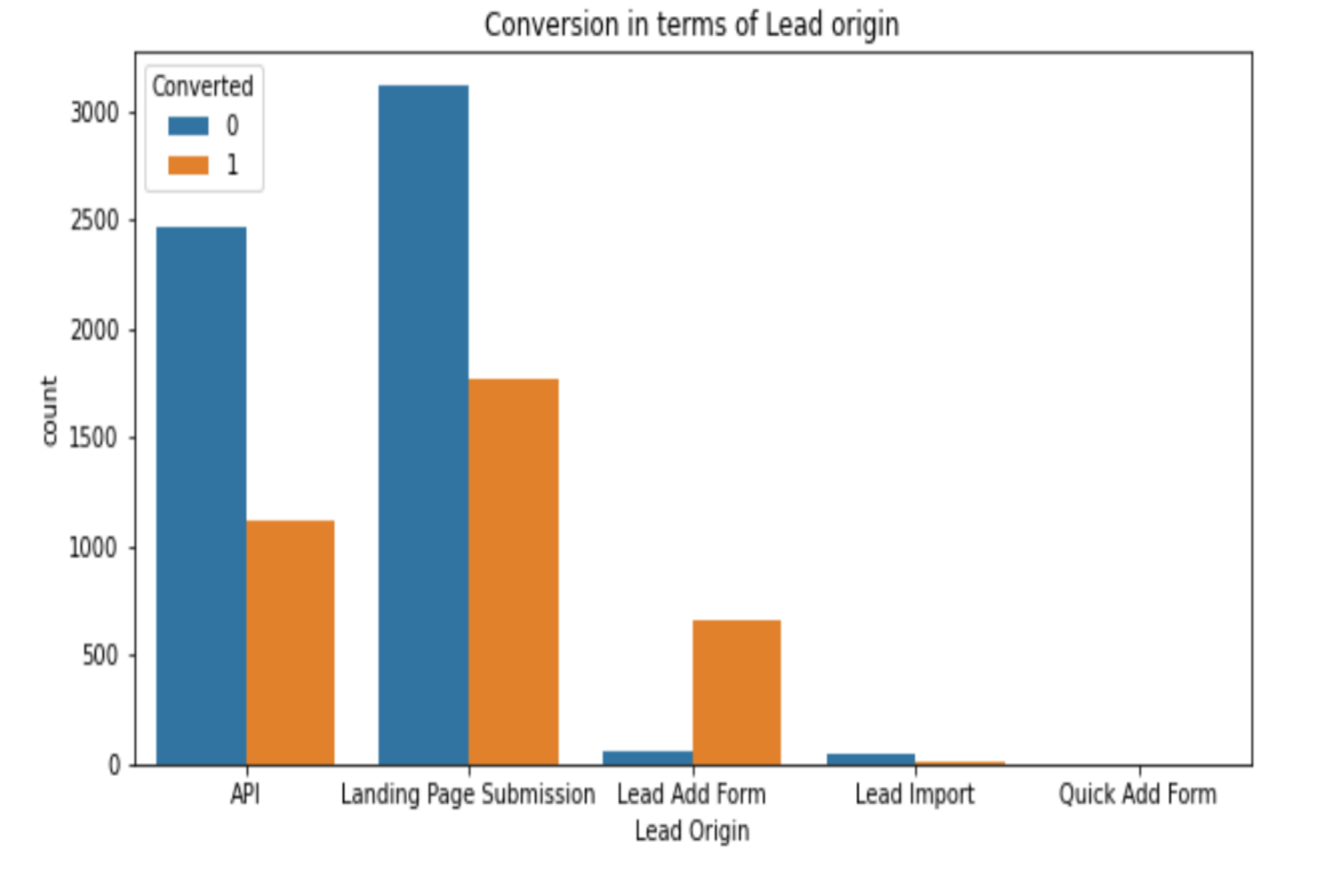
- Replaced/removed all the null values/incorrect data from the lead dataset. This would help creating a proper model for analysis.

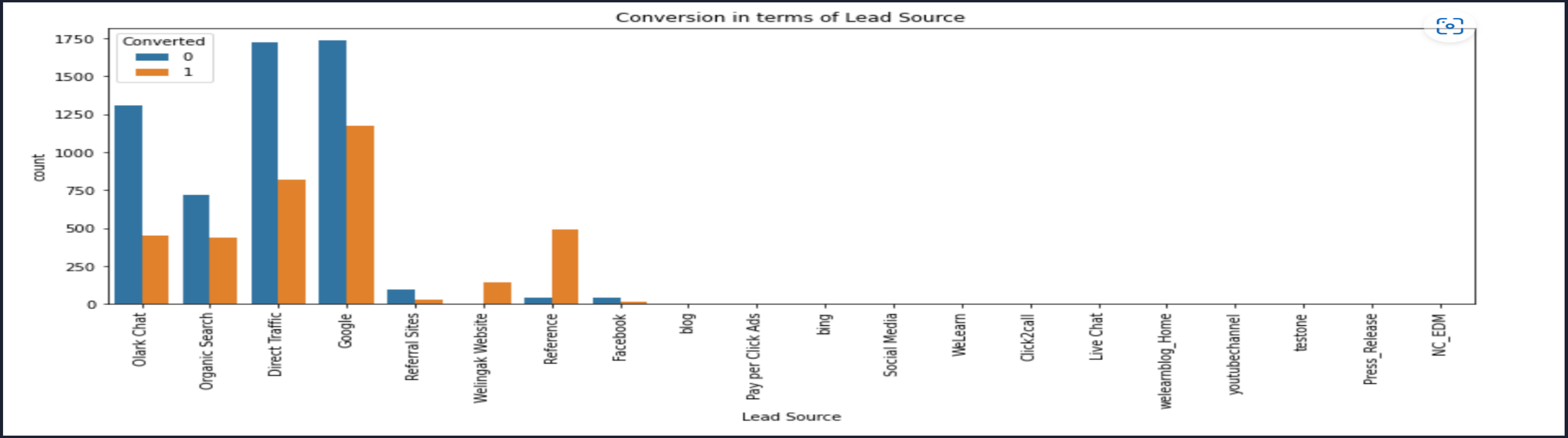
3. Exploratory Data Analysis

3.1 Analysis on Categorical Variables

i. Conversion in terms of Lead Origin

- Lead Add Form has the highest percentage of conversions
- API and Landing Page Submission have less conversion rate but has maximum number of leads counts.
- Lead Import has the least conversions and leads count.
- To improve overall lead conversion rate, focus should be on improving lead conversion rate of API and Landing Page Submission. Also, generate more leads from Lead Add form since they have a very good conversion rate.





ii. Conversion in terms of Lead Source

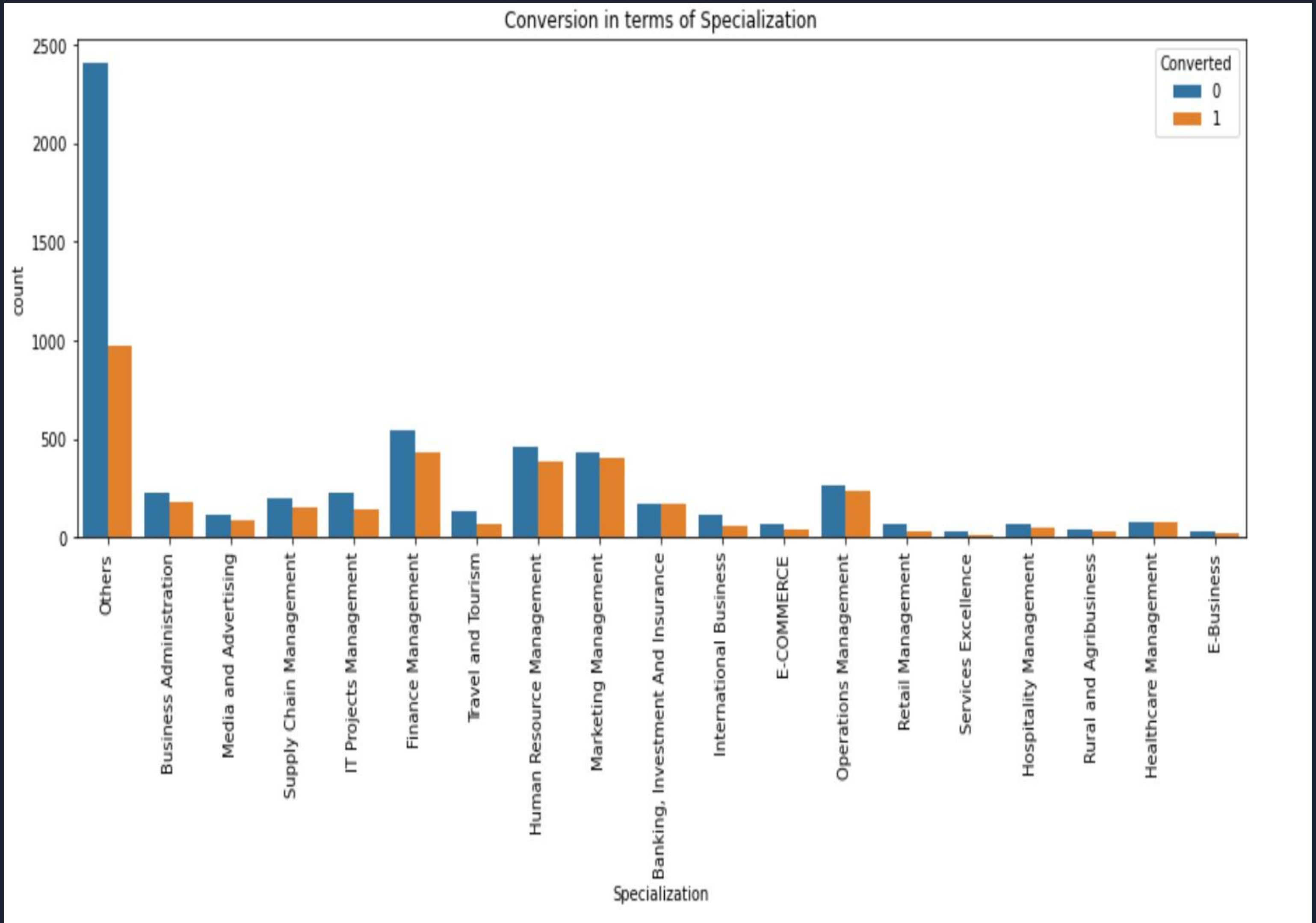
- Google and direct traffic has maximum number of leads but their conversion is less.
- Welingak website and References has highest conversions but the number of leads through that source is very less.
- Olark chat and organic search has significant number of leads but their conversion rate is less.
- Lead source in other category such as Click2call', 'Live Chat', 'NC_EDM', 'Pay per Click Ads', 'Press_Release', 'Social Media', 'WeLearn', 'bing', 'blog', 'testone', 'welearnblog_Home', 'youtubechannel' has very less leads.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic and google lead source .Also , generate more leads from reference and welingak website since they have a very good conversion rate.

iii. Specialization

- We can see that most of the people who have not mentioned the specialization did not convert which shows they are less interested.

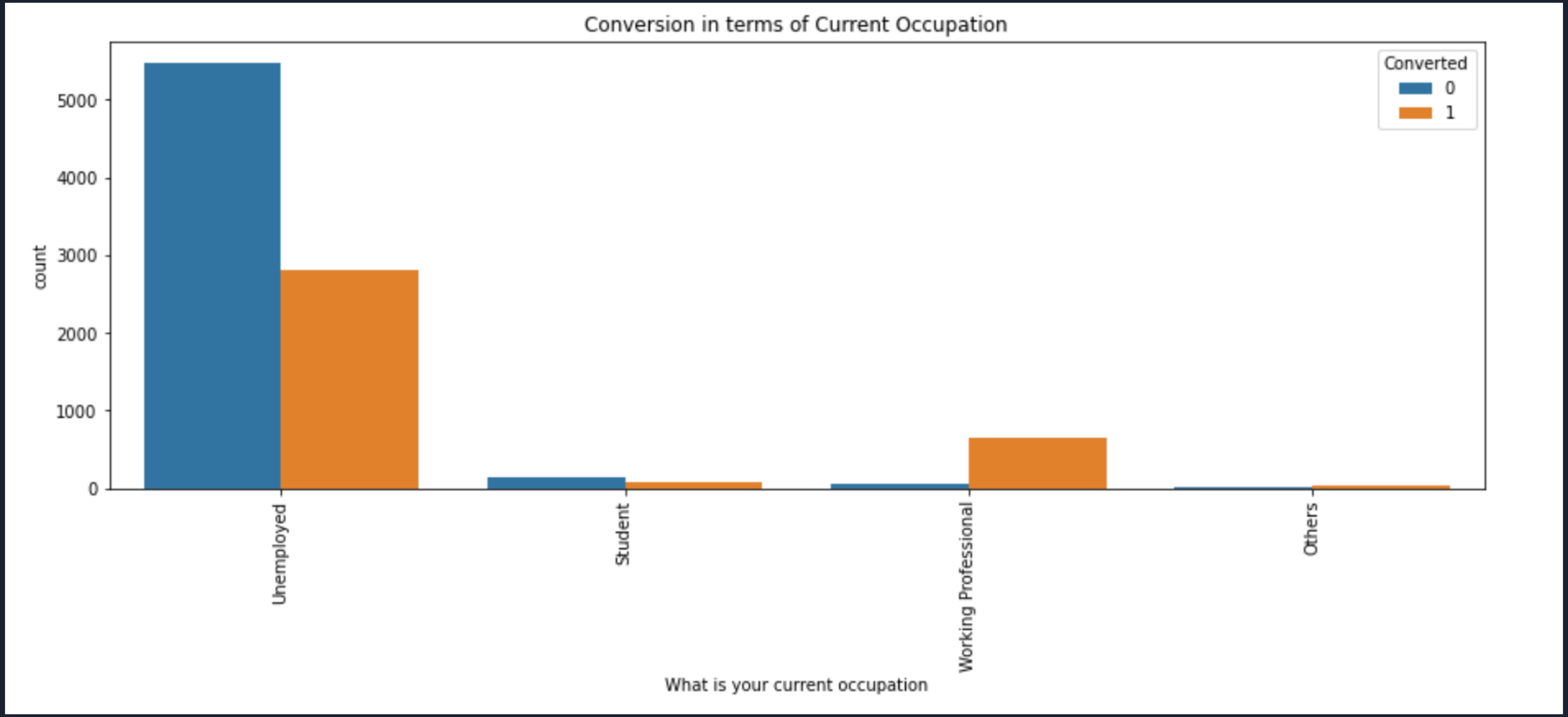
- People who have mentioned the specializations have better conversions.

- In that Banking, Investment & insurance have almost 50% conversion rate. Other specializations like Finance, HR, Operations, Marketing also have comparatively better conversion though it is less.

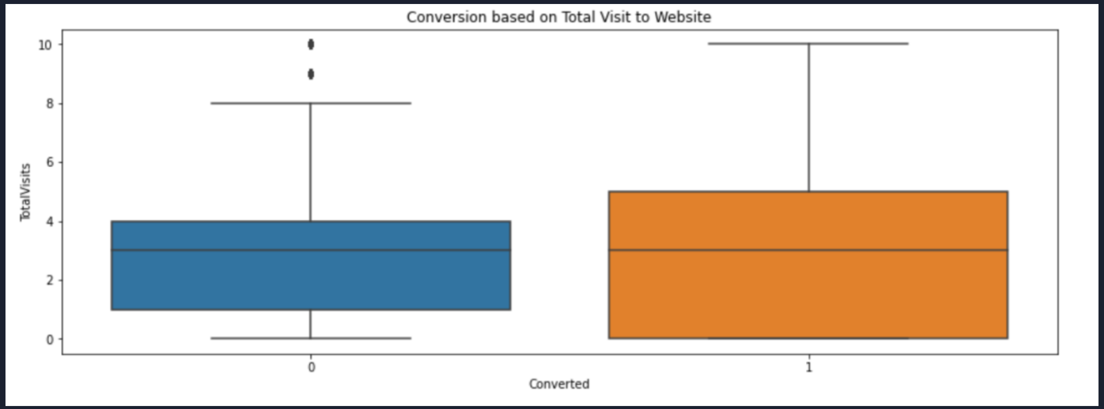


iv. Current Occupation

- It shows that unemployed people are the larger group of leads but conversion is higher in Professional group which is a smaller group.
- This can be because of the higher fees for the course which unemployed people cannot afford and also professional people always look for upskilling courses

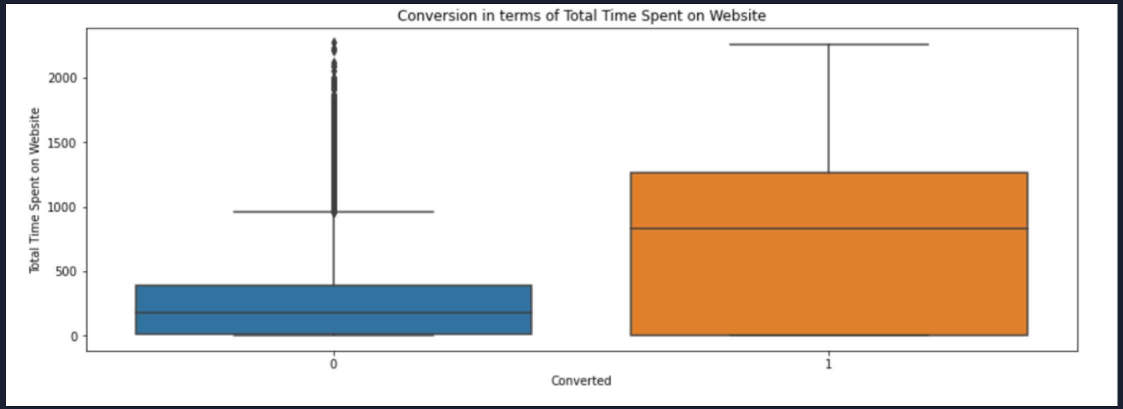


3.2 Analysis on Numerical Variables



i. Total Visits to Website

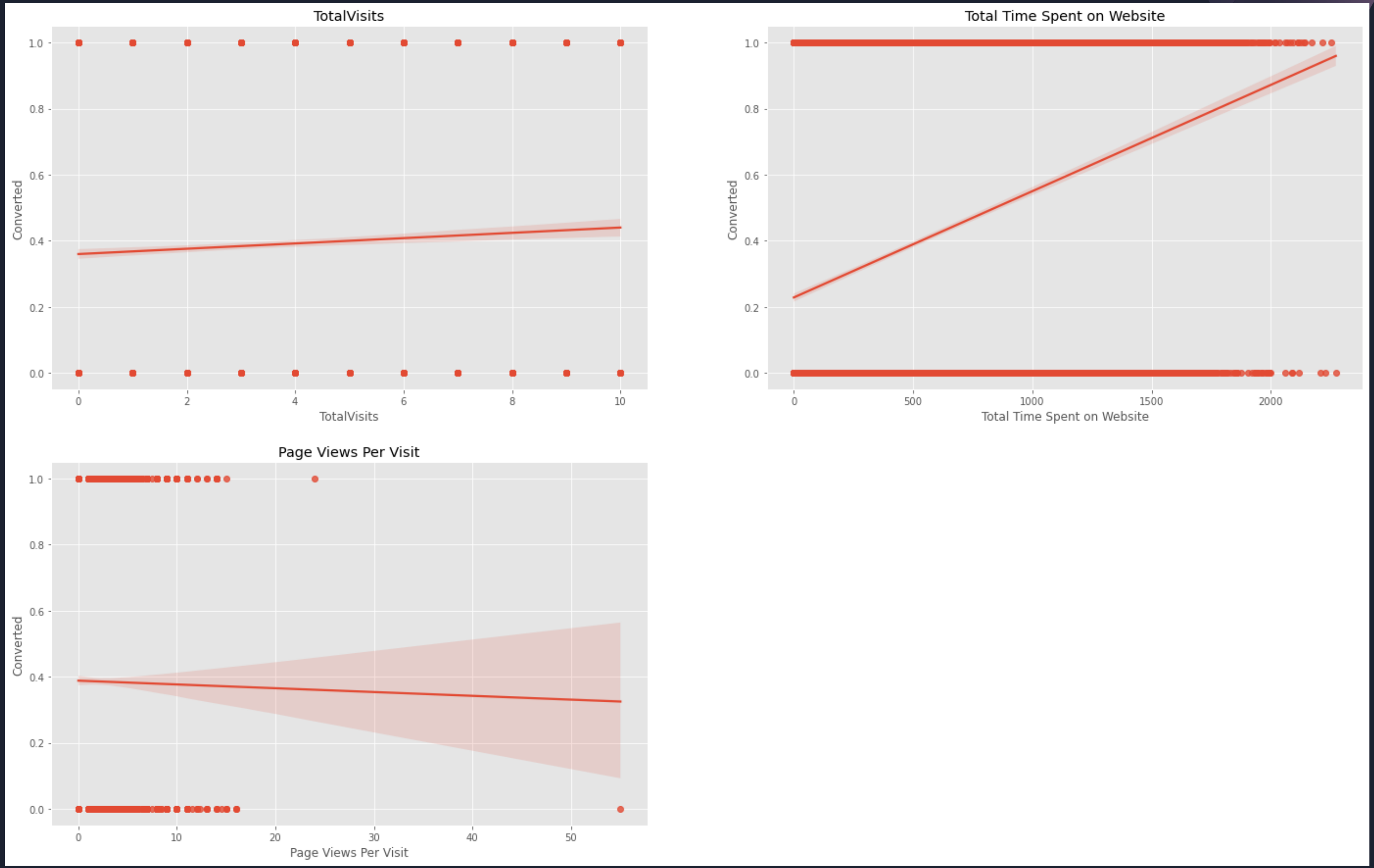
- People who visits the platform have almost similar chances of getting converted and not getting converted as the median for both plots is along the same line.



ii. Total Time spent on the Website

- This shows that people who spend more time on the website have higher chances of getting converted to users.

- This shows that people who spend more time on the website have higher chances of getting converted to users.



3. Data Preparation and Model Creation

After the EDA Analysis we have to prepare the data for effective model creation.

Steps Required for Model creation

1. Data conversion – Covert the Binary Fields(Yes/No) to Numeric (0/1)
2. Create Dummy Values for the Category Variables and Drop the original variables
3. Split the Dataset in to Train and Test Dataset (70%/30% split) Model build based on the train dataset only. Test Dataset reserved for Testing the model created using Train set.
4. Scale and transform the numeric variables for a perfect fit.
5. Identify the correlated values and build a model using RFE

3. Data Preparation and Model Creation

- Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached.

- Ran the RFE as 15 features to select , after multiple iteration and removing unwanted variables from set, We have identified the below features are required for an effective model creation

  'Do Not Email', 'Total Time Spent on Website',

  'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',

  'Lead Source_Olark Chat', 'Lead Source_Welingak Website',

  'Last Activity_Converted to Lead',

  'Last Activity_Had a Phone Conversation',

  'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent',

  'Specialization_Hospitality Management', 'Specialization_Others',

  'What is your current occupation_Working Professional'

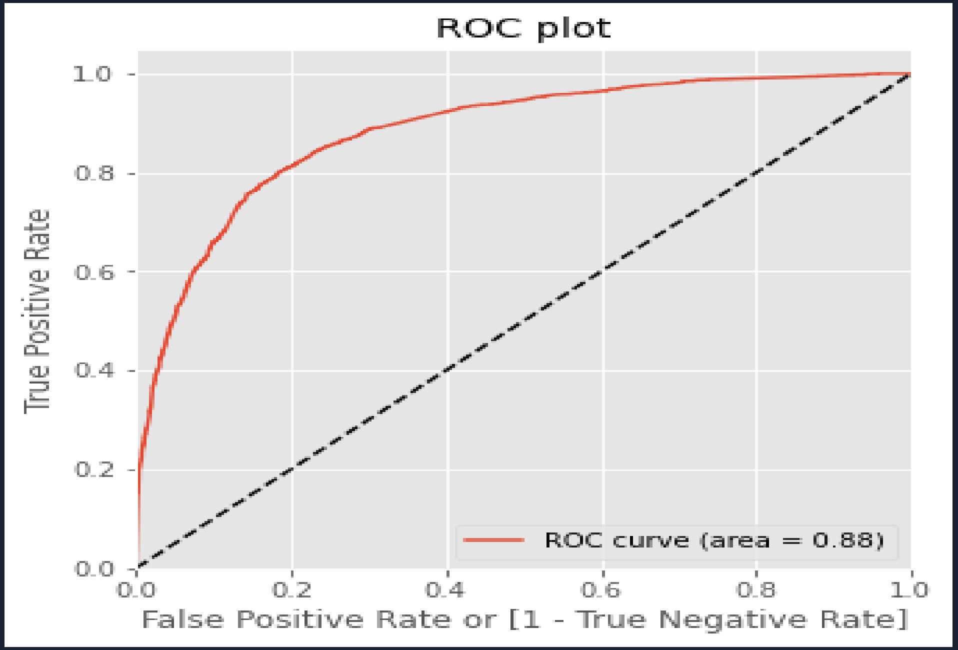
Final Model Statistics :

1. VIF values less than 5(indicates a low correlation of that predictor with other predictors),
2. P values less than 0.05

4. Model Evaluation

ROC Plot

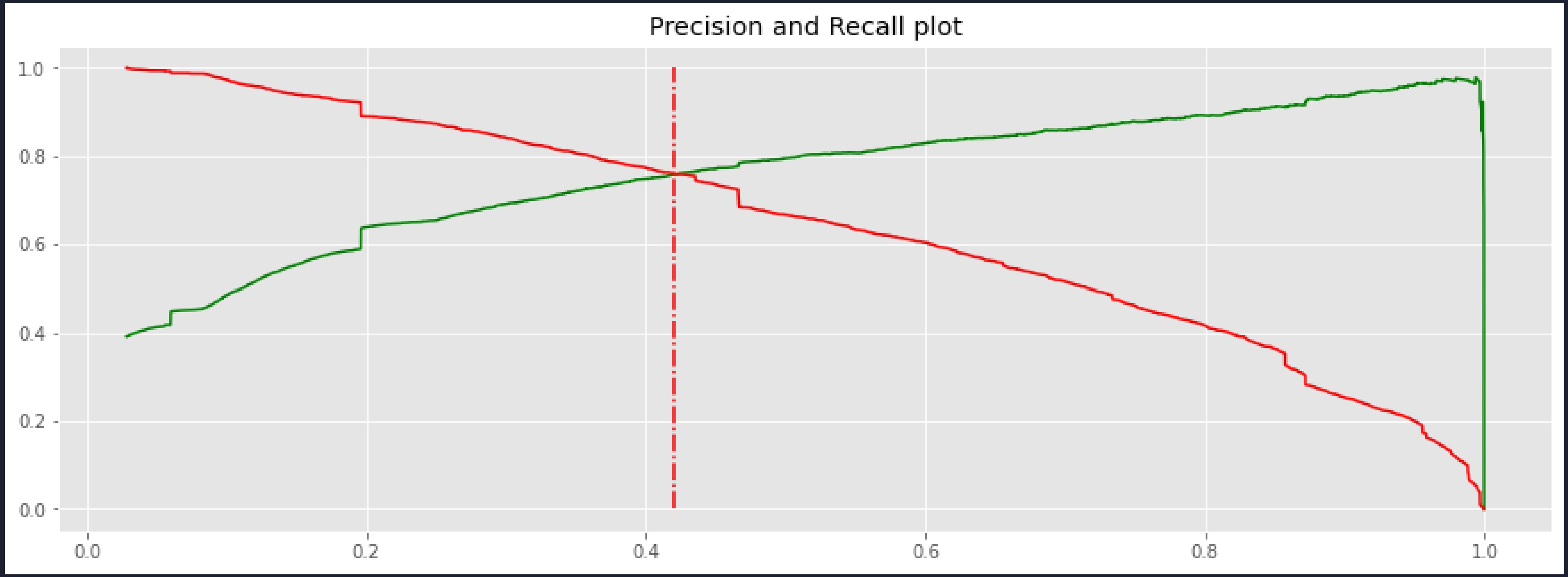
- Receiver Operating Curve shows the trade off between sensitivity and specificity of a logistic regression model.
- If area under the curve is higher, the model is considered to be a good model.
- In our analysis, the area under ROC curve is 0.88 which is a good value for a model.



Precision and Recall

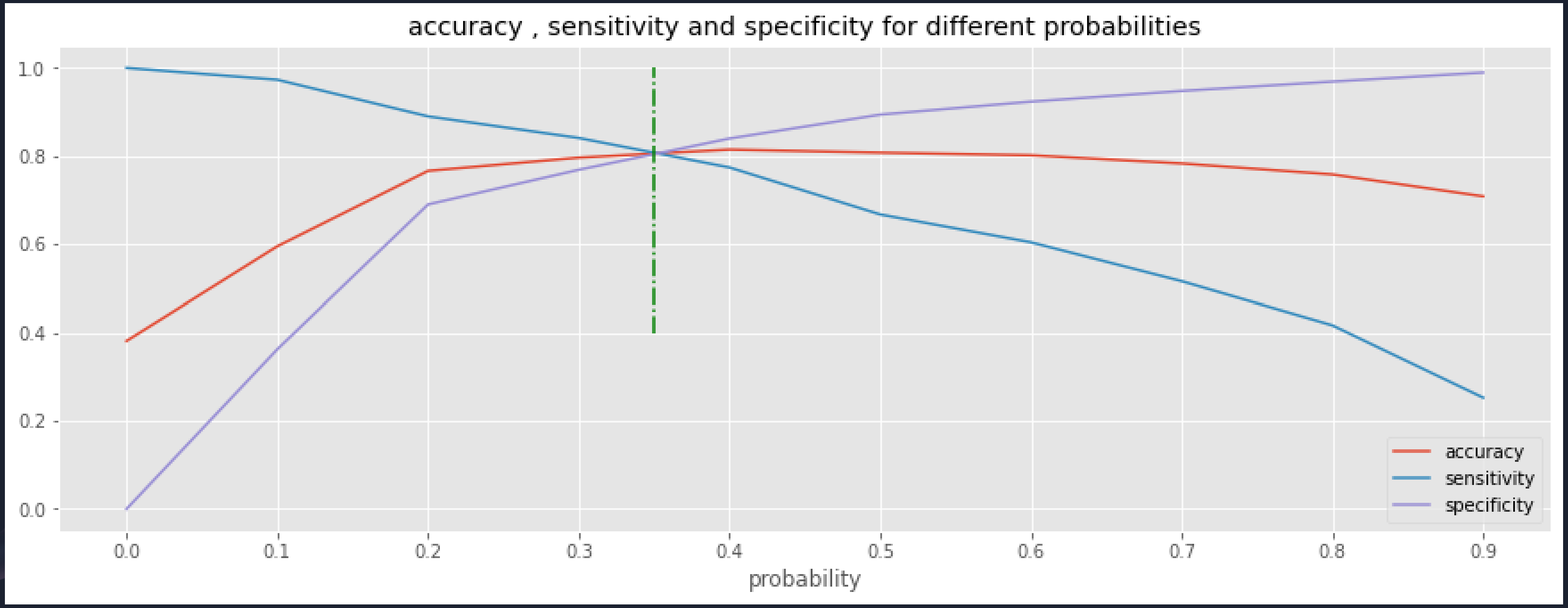
- Precision – It is the ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples (either correctly or incorrectly).

- Recall – It is the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples. It measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.



Final Model Evaluation

- Accuracy : 81%
- Sensitivity : 68%
- Specificity : 89.4%
- Optimal Cut Off point : 0.35 for a well-balanced Sensitivity and Specificity values



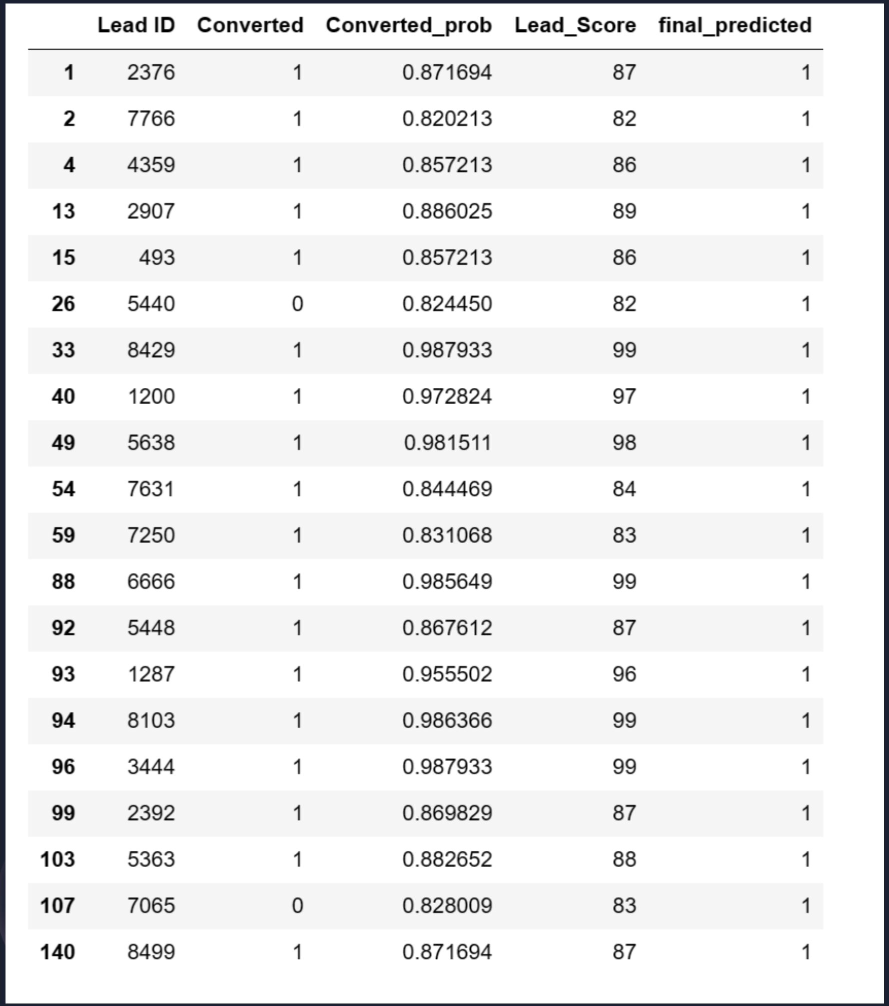
5. Model Testing

The Final Model tested with the test to data to verify the Fit and Finalize the variables

- Final Model Evaluation with test Data :
  - Accuracy : 81%
  - Sensitivity : 74.1%
  - Specificity : 85.2%

- The Accuracy , Sensitivity and Specificity values are looking good, this model can be used now for identifying the ‘Hot Lead'

Lead Score – Hot Leads



	Lead ID	Converted	Converted_prob	Lead_Score	final_predicted
1	2376	1	0.871694	87	1
2	7766	1	0.820213	82	1
4	4359	1	0.857213	86	1
13	2907	1	0.886025	89	1
15	493	1	0.857213	86	1
26	5440	0	0.824450	82	1
33	8429	1	0.987933	99	1
40	1200	1	0.972824	97	1
49	5638	1	0.981511	98	1
54	7631	1	0.844469	84	1
59	7250	1	0.831068	83	1
88	6666	1	0.985649	99	1
92	5448	1	0.867612	87	1
93	1287	1	0.955502	96	1
94	8103	1	0.986366	99	1
96	3444	1	0.987933	99	1
99	2392	1	0.869829	87	1
103	5363	1	0.882652	88	1
107	7065	0	0.828009	83	1
140	8499	1	0.871694	87	1

- The records in the Hot_Leads are the potential customers.

- From the lead score table, we can consider the top leads whose score is above 80% and they have higher chance of getting converted.

- There is unique Lead ID through which the sales team can communicate with the Hot Leads and provide more information on the courses.

- The lead conversion rate would go up if the sales team focus more on communicating with the Hot Leads rather than making calls and communicating to everyone.

6. Observations and Conclusions

- According to the model analysis, following variables are the potential features to identify the ‘Hot Leads'
  - The total time spend on the Website
  - Current Occupation
  - Lead origin
  - Do not Email Flag
  - Last Activity
  - Lead Source
- Customers who comes under below category are more likely to be turn to potential buyer
  - 'Total Time Spent on Website' is high
  - Customers Current Occupation is 'Working Professional'
  - Lead Sources are from - Welingak Website & Reference
  - Lead Origin in 'Landing Page Submission' or 'Lead Add Form'
- If the Last activity of the customer is one of the category below, then the customer is a ‘Hot Lead'
  - Converted to Lead
  - Had a Phone Conversation
  - Olark Chat Conversation
  - SMS Sent