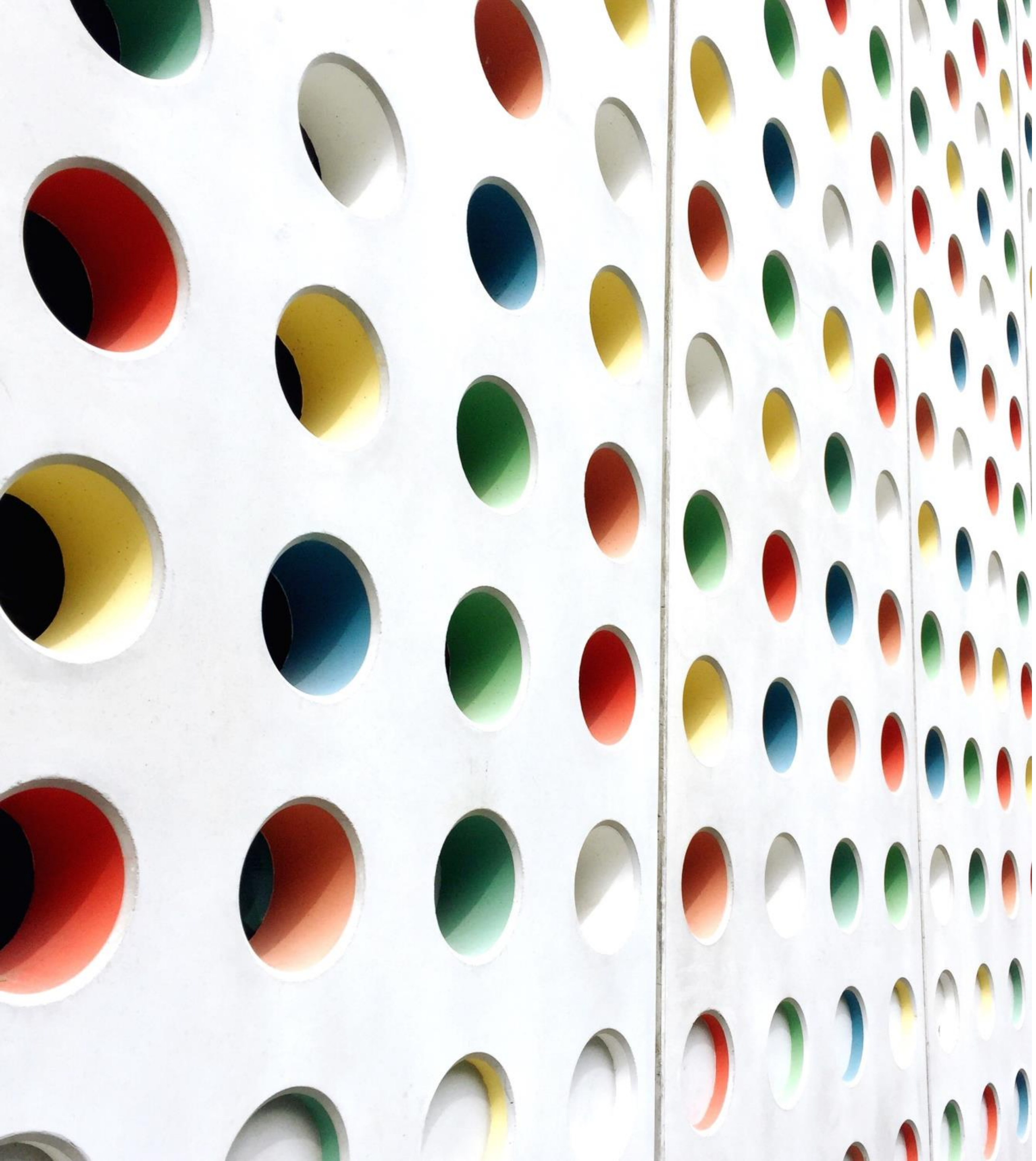


LEAD SCORING CASE STUDY

BANUPRIYA.R

Problem Statement

- X Education sells online courses to industry professionals
- 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
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

Business Objective:

- X education wants to know the most promising leads
- For that they want to build a Model which identifies the hot leads
- Deployment of the model for the future use.

Proposed Solution

- I. Data understanding, Data Cleaning and Visualisation
- II. Model Pre-processing
    - Data Encoding
    - Splitting Data-Train and Test
    - Feature Scaling
- III. Model building
- IV. Model Analysis with Training Data
- V. Adding Lead Score
- VII. Model Analysis with Test Data
- VII. Conclusion/Inference From Machine Learning Model

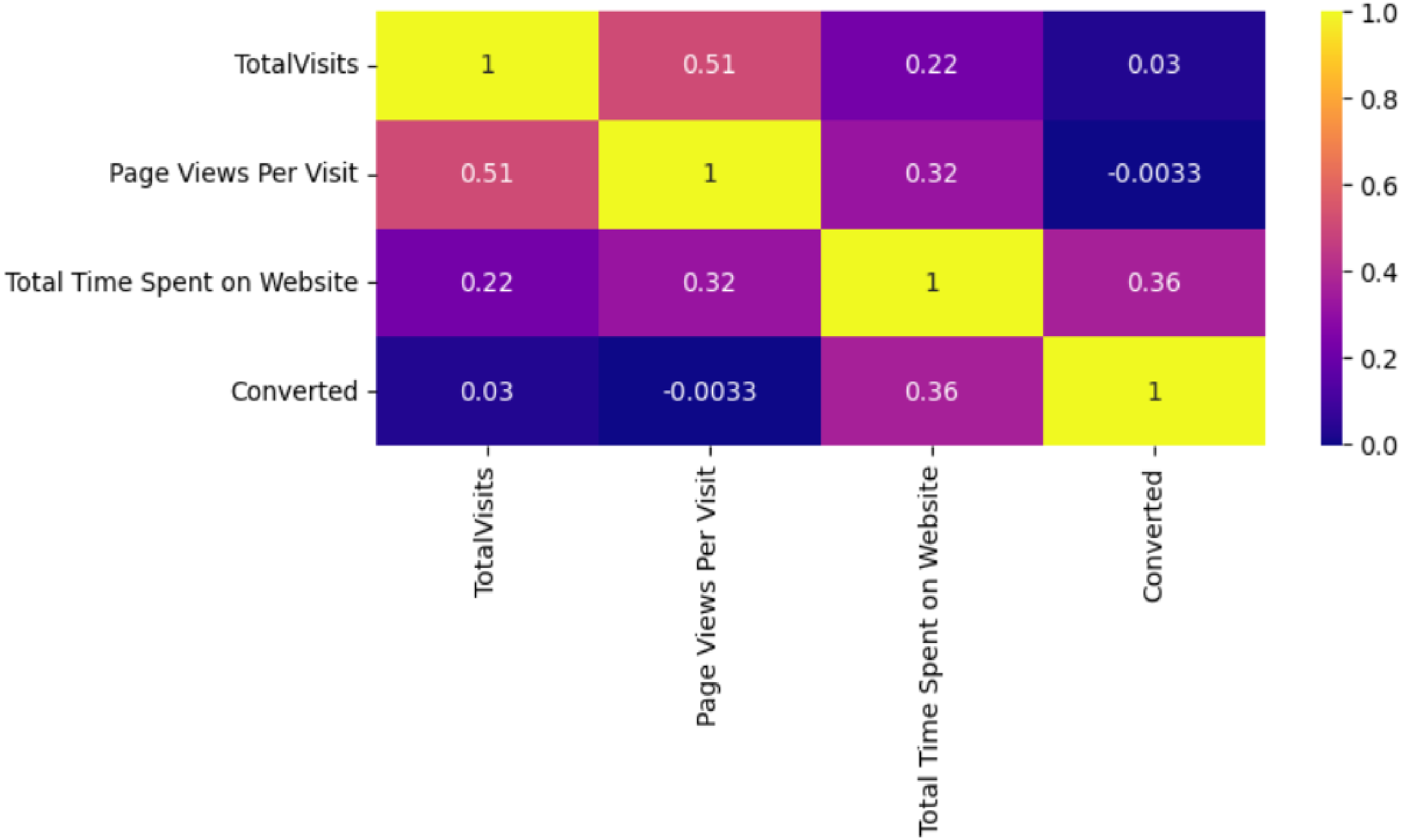
I. Data understanding, Data Cleaning and Visualisation

- This dataset has: 9240 rows, 37 columns
- There are few columns with quite a high number of missing/null values in the dataframe.
- There are no select value in the dataframe "data_leads“
- Columns 'Prospect ID','Lead Number' has no use for modelling so we are dropping these column
- These columns have same one unique value: Magazine, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque’. These columns are of no use as they have only one category of response from people and these can be dropped.

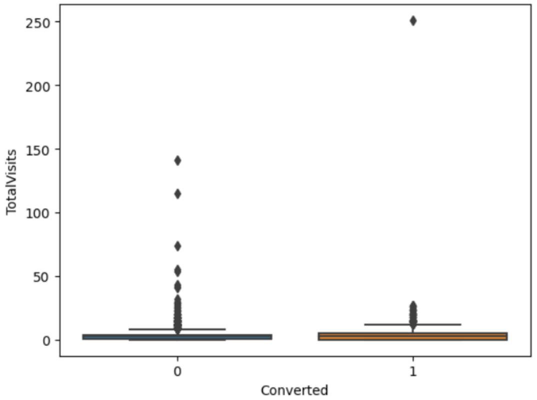
I.1. Categorical value Analysis

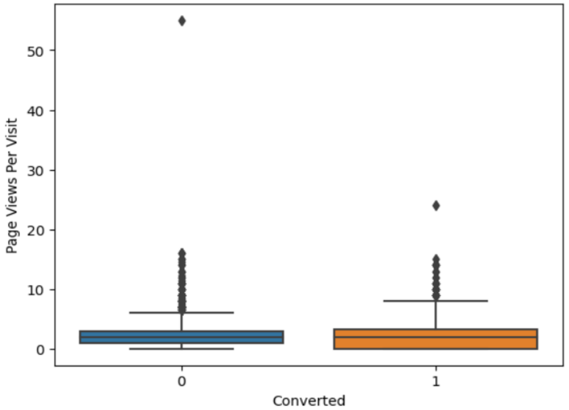
- “How did you hear about X Education” column and “Lead Profile” column have respectively more than 78% and 74% null values => So it’s better to be drop these columns.
- “City” has 39 % missing values. Imputing missing values with Mumbai will make the data more skewed. Hence “City” column can be dropped.
- More than 80% of the customers are from India.It does not make business sense to impute missing values with India.So, “Country” column can be dropped
- “Tags” has 36% missing values. Tags are assigned to customers indicating the current status of the lead. Since this is current status, this column will not be useful for modeling.so it can be dropped.
- “What matters most to you in choosing a course” has 29% missing values and More than 70% people have selected 'better career prospects’. This is massively skewed and will not provide any insight.
- “Specialization” with Management in them have higher number of leads. So this is a significant variable .Hence imputation or dropping is not a good choice. We need to create others category
- “What is your current occupation”: We can impute the missing values with 'Unemployed' as it has the most values. This seems to be a important variable from business context.
- “Lead Source”: "Google" having highest number of occurences ,hence we will impute the missing values with 'Google’
- “Last Activity”: "Email Opened" is having highest number of values so will impute the missing values with label 'Email Opened'.

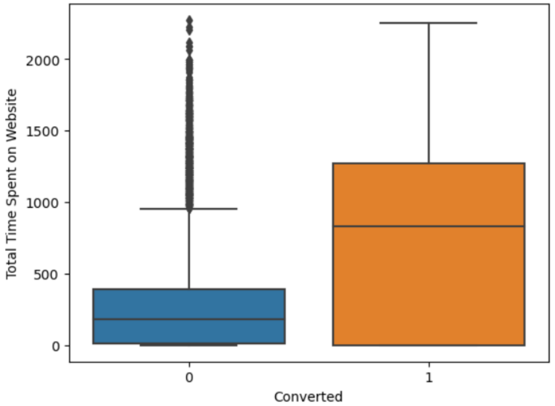
I.2. Numerical value Analysis



I.2. Numerical value Analysis

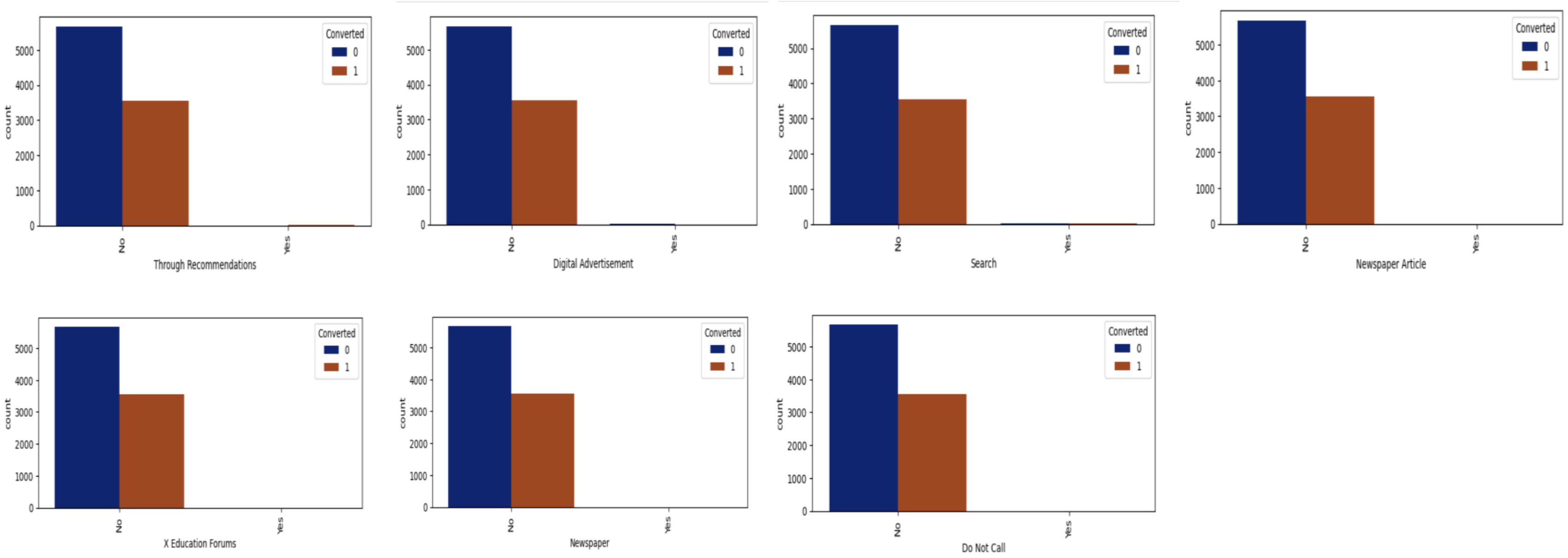






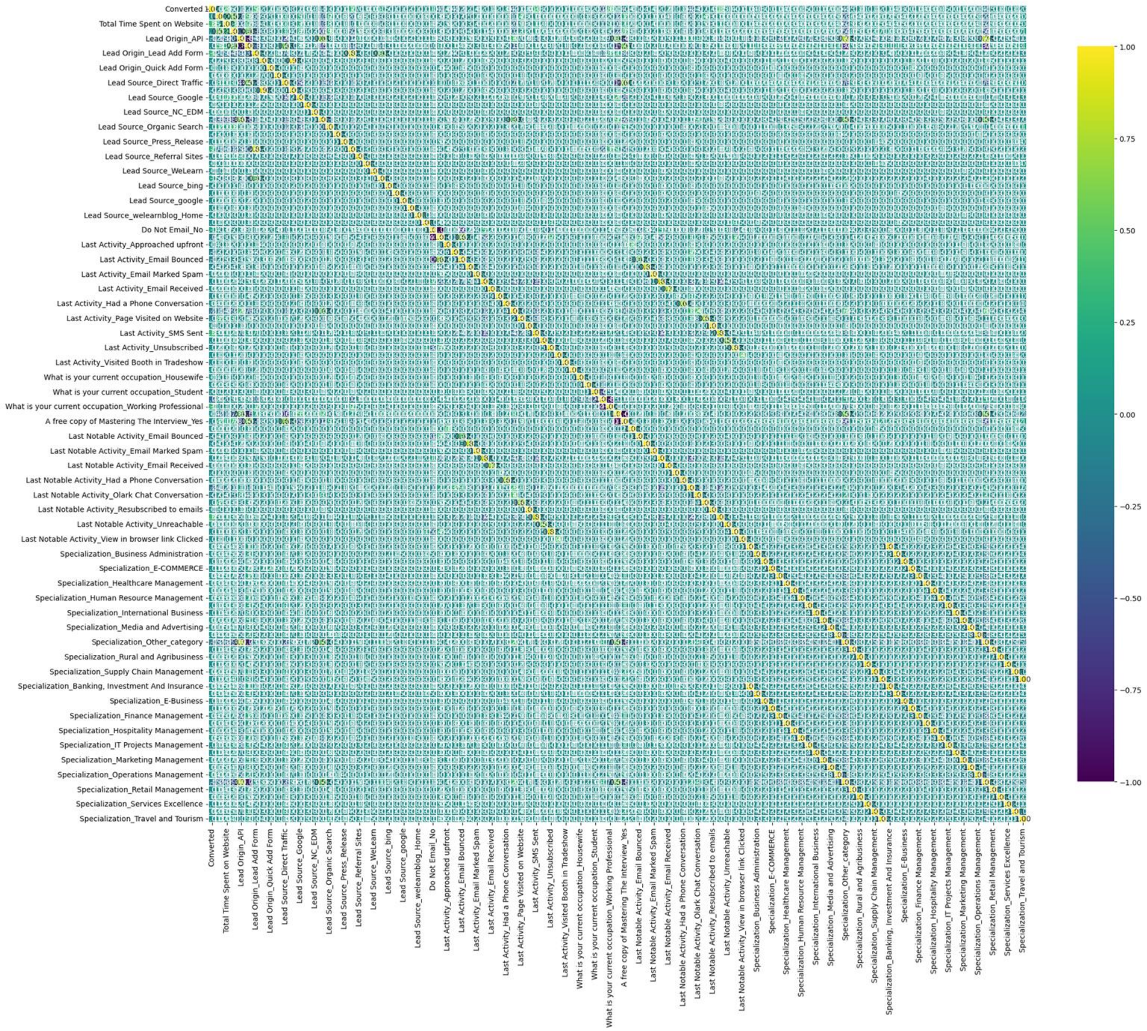
- Missing values in 'TotalVisits, 'Page Views Per Visit’, 'Total Time Spent on Website' can be imputed with mode

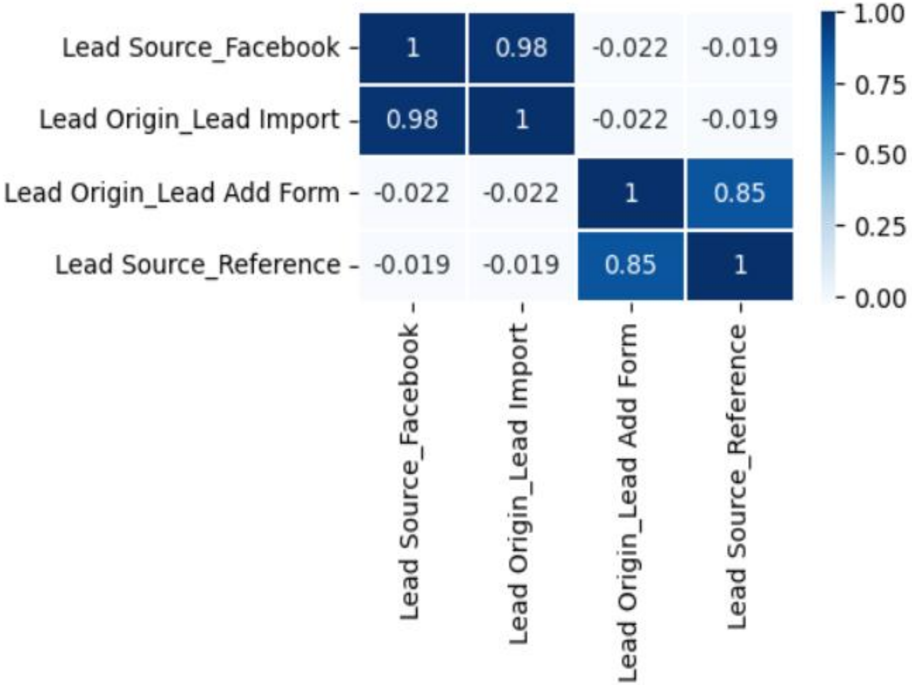
I.3. Leftover columns in categorical variables.



- “Through Recommendations”, “Digital Advertisement”, “Search”, “Newspaper Article”, “X Education Forums”, News Paper”, “Do not call” will not add any value to the model "No" has a much higher frequency indicating that the majority of the observations fall into the "No" category.
- => So we are dropping these columns.

II. Model Pre-processing



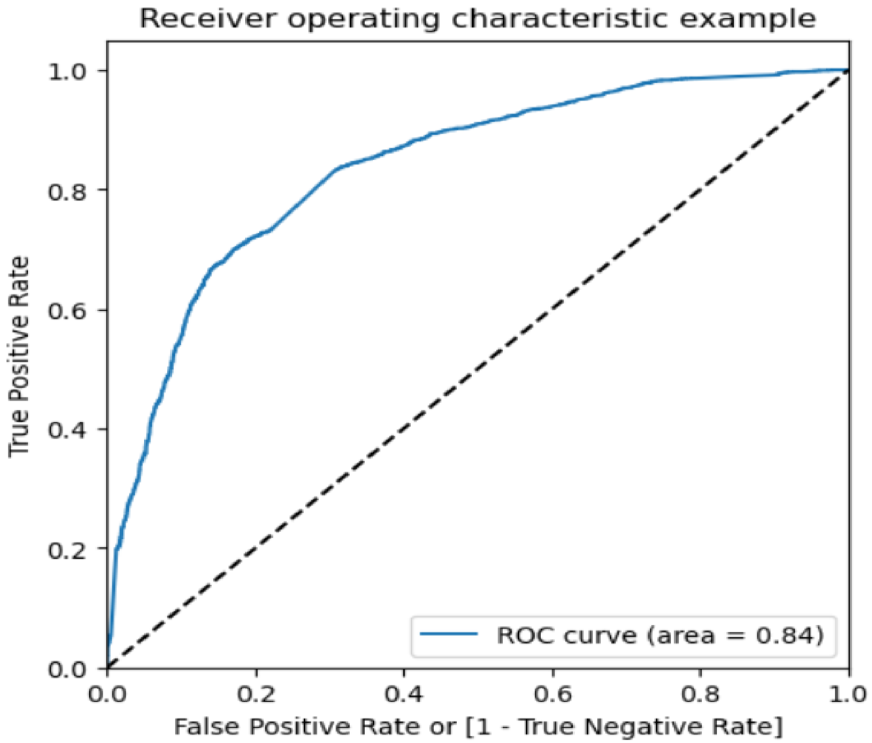


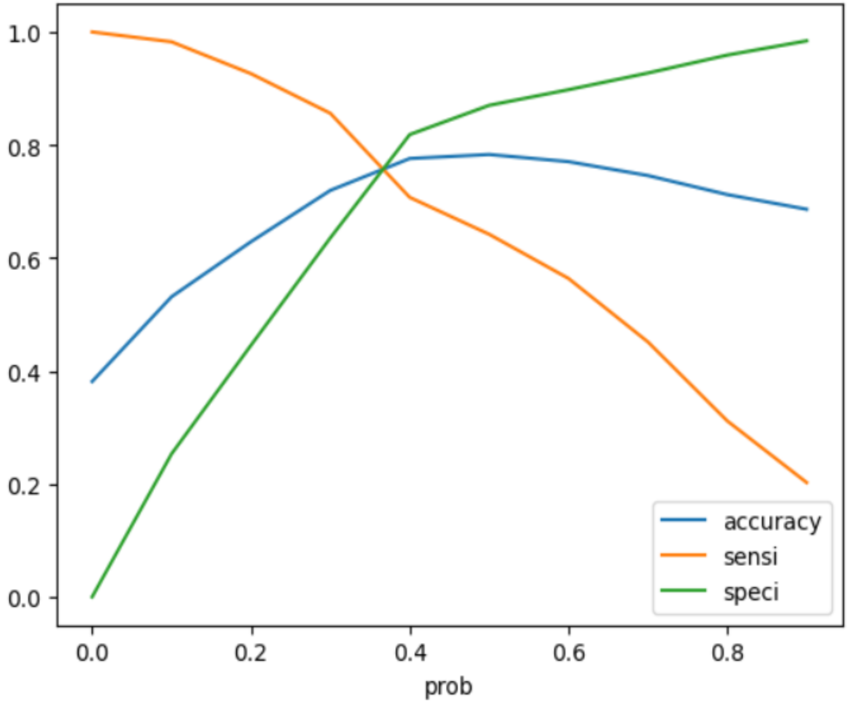
III. Logistic Regression

- Building Machine Learning Training Model :

- - Model 1: with the parameters emitted by RFE
- - Model 2: build the Model again after dropping the columns with High P-Value & High VIF
- - Model 3: lets build the Model again after dropping the columns with very High P-Value
- - Model 4: build the Model again after dropping Lead Source_Click2call parameter

IV. Model Analysis with Training Data





- The ROC curve's area under the curve is 0.84, indicating a promising model performance.

- 0.36 is the approximate point where all the curves meet

IV. Model Analysis with Training Data

- Confusion Matrix = array([[3146, 856],[670, 1796]])

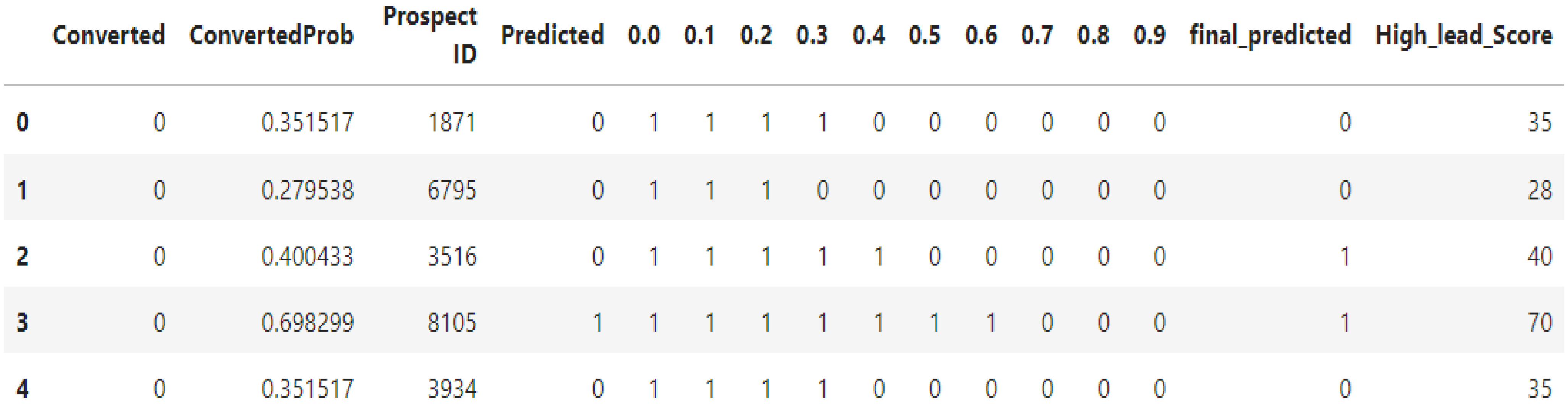
- Accuracy = 76.40%

- Sensitivity = 72.83%

- Specificity = 78.61%

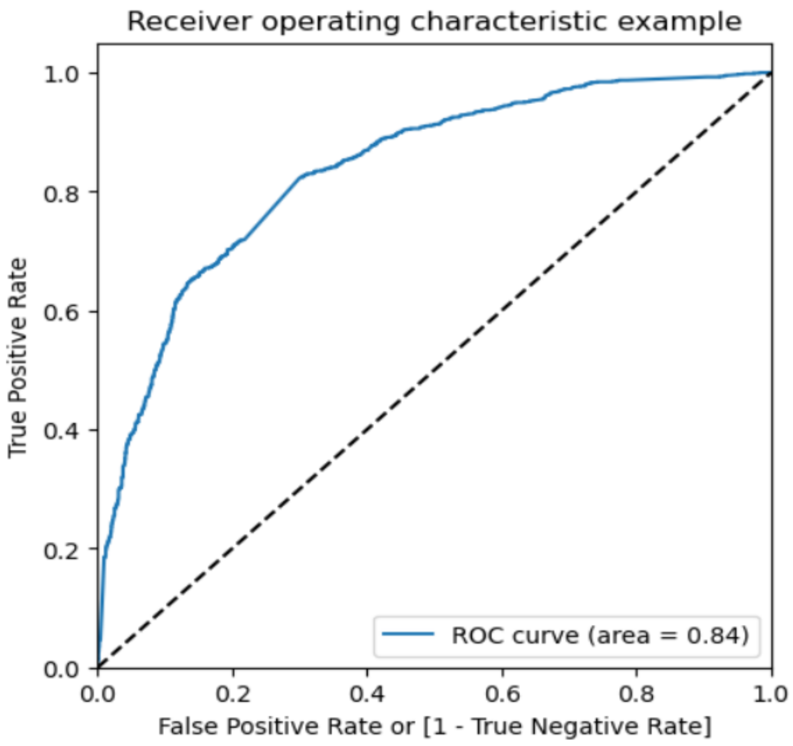
- F1 Score: 0.7018366549433372 => F1 Score of 0.702 Represents a reasonable balance between precision and recall

V. Adding Lead Score



	Converted	ConvertedProb	Prospect ID	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	High_lead_Score
0	0	0.351517	1871	0	1	1	1	1	0	0	0	0	0	0	0	35
1	0	0.279538	6795	0	1	1	1	0	0	0	0	0	0	0	0	28
2	0	0.400433	3516	0	1	1	1	1	1	0	0	0	0	0	1	40
3	0	0.698299	8105	1	1	1	1	1	1	1	1	0	0	0	1	70
4	0	0.351517	3934	0	1	1	1	1	0	0	0	0	0	0	0	35

VI. Model Analysis with Test data



Area under ROC curve is 0.84 out of 1 which indicates a good predictive model

- For Test data
- Accuracy: 76.11%
- Sensitivity: 72.83%
- Specificity: 78.61%
- These matrices are almost the same as the train set, so our final model 4 is performing with good consistency on both Train and test data
-

VII. Conclusion

- As seen from our final model, following are the Parameters which help us to predict the probablity of Leads conversion and hence increasing the Chances of Lead Conversion from 30% to higher than 70 %:

- Lead Origin_Lead Add Form 3.306135
- Last Activity_Had a Phone Conversation 1.519508
- Total Time Spent on Website 1.076119
- Lead Source_Olark Chat 0.340387
- Lead Source_Direct Traffic -0.322981
- Lead Origin_Landing Page Submission -0.629927
- Last Activity_Form Submitted on Website -1.055841
- Last Activity_Email Link Clicked -1.060514
- Last Activity_Converted to Lead -1.820172
- Last Activity_Olark Chat Conversation -2.120415
- Last Activity_Email Bounced -2.582260

VII. Conclusion

- Based on the Logistic Regression Model final Features with positive coefficients, such as "Lead Origin_Lead Add Form," "Last Activity_Had a Phone Conversation," and "Total Time Spent on Website," have a stronger positive influence on predicting the probability of leads converting to take the course. Higher values in these features increase the likelihood of conversion. And negative coefficients, like "Last Activity_Olark Chat Conversation" and "Last Activity_Email Bounced," have a negative influence on the conversion probability. Higher values in these features decrease the likelihood of conversion.

- We found that Overall Model Performance (Training Data): Accuracy: 76.40% Sensitivity (Recall): 72.83% Specificity: 78.61% F1 Score: 0.7018 Overall Model Performance Test Data: Accuracy: 76.11% Sensitivity (Recall): 72.83% Specificity: 78.61% The evaluation matrics for Test and Training Data are pretty close to each other which indicates that the model is performing consistently across different evaluation metrics. Also, the model's overall performance is good, as indicated by accuracy and F1 Score, emphasizes its effectiveness in predicting lead conversion outcomes.