LEAD SCORE Case Study

- Group Member
- Geetika Phutela
- Prerit Jain
- Shweta Singh

Problem Statement

- An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- 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 typical lead conversion rate at X education is around 30%.

Business Objectives:

- 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.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

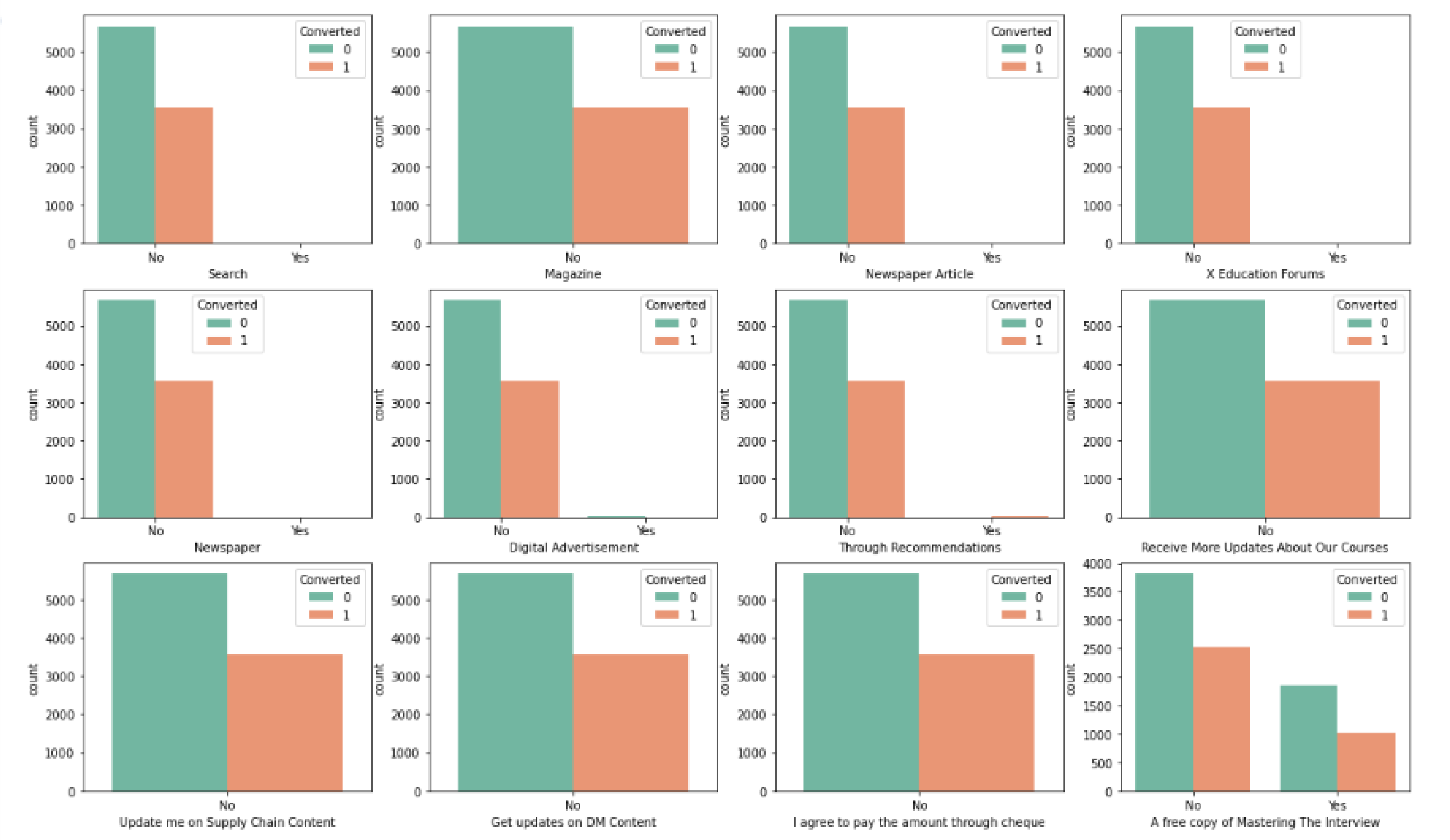
Solution Methodology

- Data cleaning and data manipulation.
    1. Check and handle duplicate data.
    2. Check and handle NA values and missing values.
    3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
    4. Imputation of the values, if necessary.
    5. Check and handle outliers in data.
- EDA
    1. Univariate data analysis: value count, distribution of variable etc.
    2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

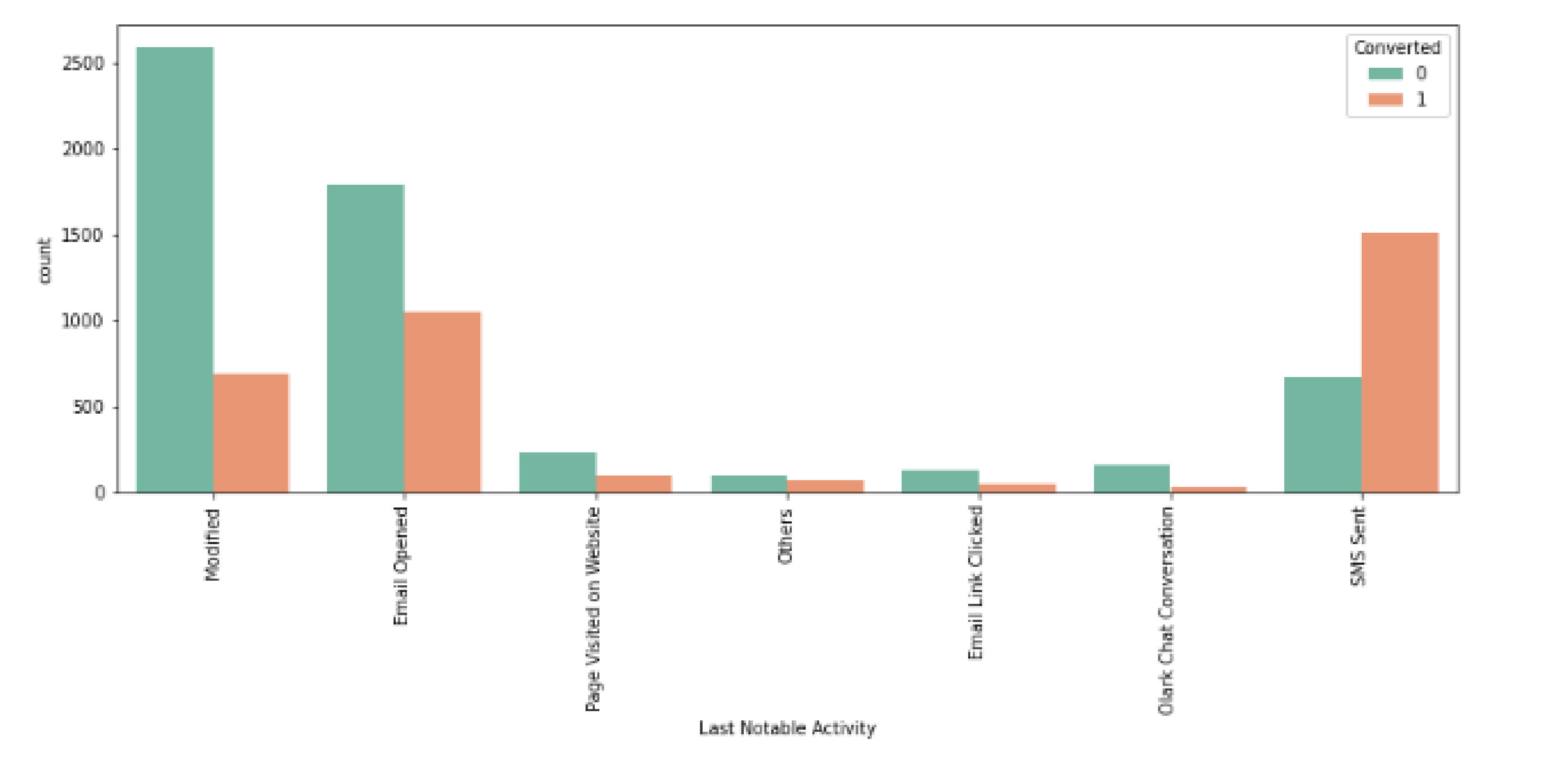
Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240
- Few columns observed with level called 'Select' which means that the customer had not selected the option for that column which is why it shows 'Select'. These values are as good as missing values and hence converted 'Select' values to Nan
- Dropped Columns with Missing Values >=35%
- Following columns have null values : - Country ,Lead Source ,Total Visits, Page Views Per Visit , Last Activity, What is your current occupation , What matters most to you in choosing a course.
- For columns except 'A free copy of Mastering The Interview' data is highly imbalanced, thus we dropped them
- "A free copy of Mastering The Interview" is a redundant variable, hence included in list of dropping columns.
- Google is having highest number of occurrences, hence the missing values imputed with label 'Google’
- Since, missing values are very high , we imputed all missing values with value 'not provided’
- As we can see that most of the data consists of value 'India', no inference can be drawn from this parameter. Hence, we dropped this column
- Since no information has been provided regarding occupation, we replaced missing values with new category 'Not provided’
- This column spread of variance is very low , hence dropped.

EDA

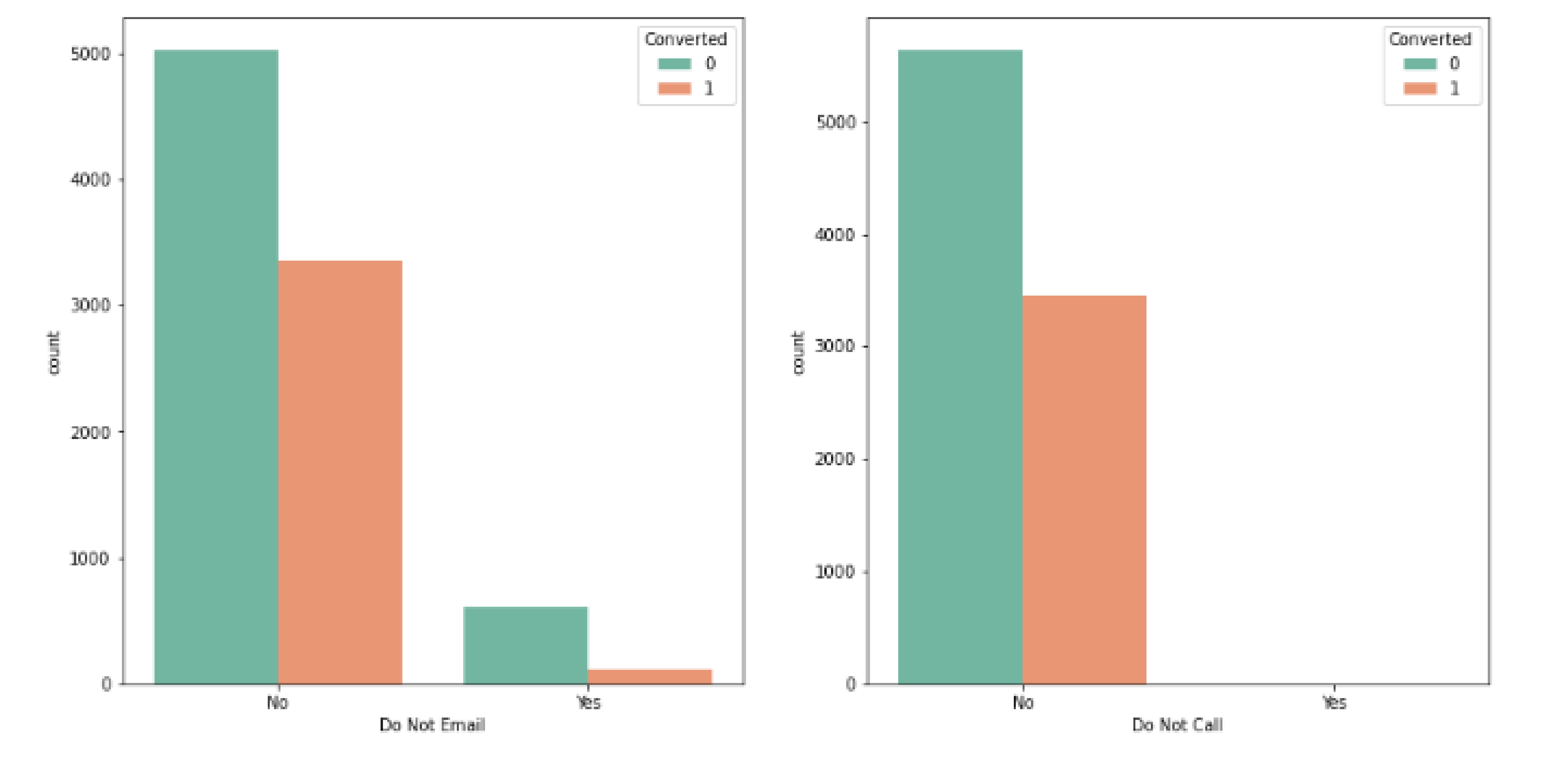

Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Yes
Search
Magazine
Newspaper Article
X Education Forums
Newspaper
Digital Advertisement
Through Recommendations
Receive More Updates About Our Courses
Update me on Supply Chain Content
Get updates on DM Content
I agree to pay the amount through cheque
500
1500
2500
3500
A free copy of Mastering The Interview

EDA

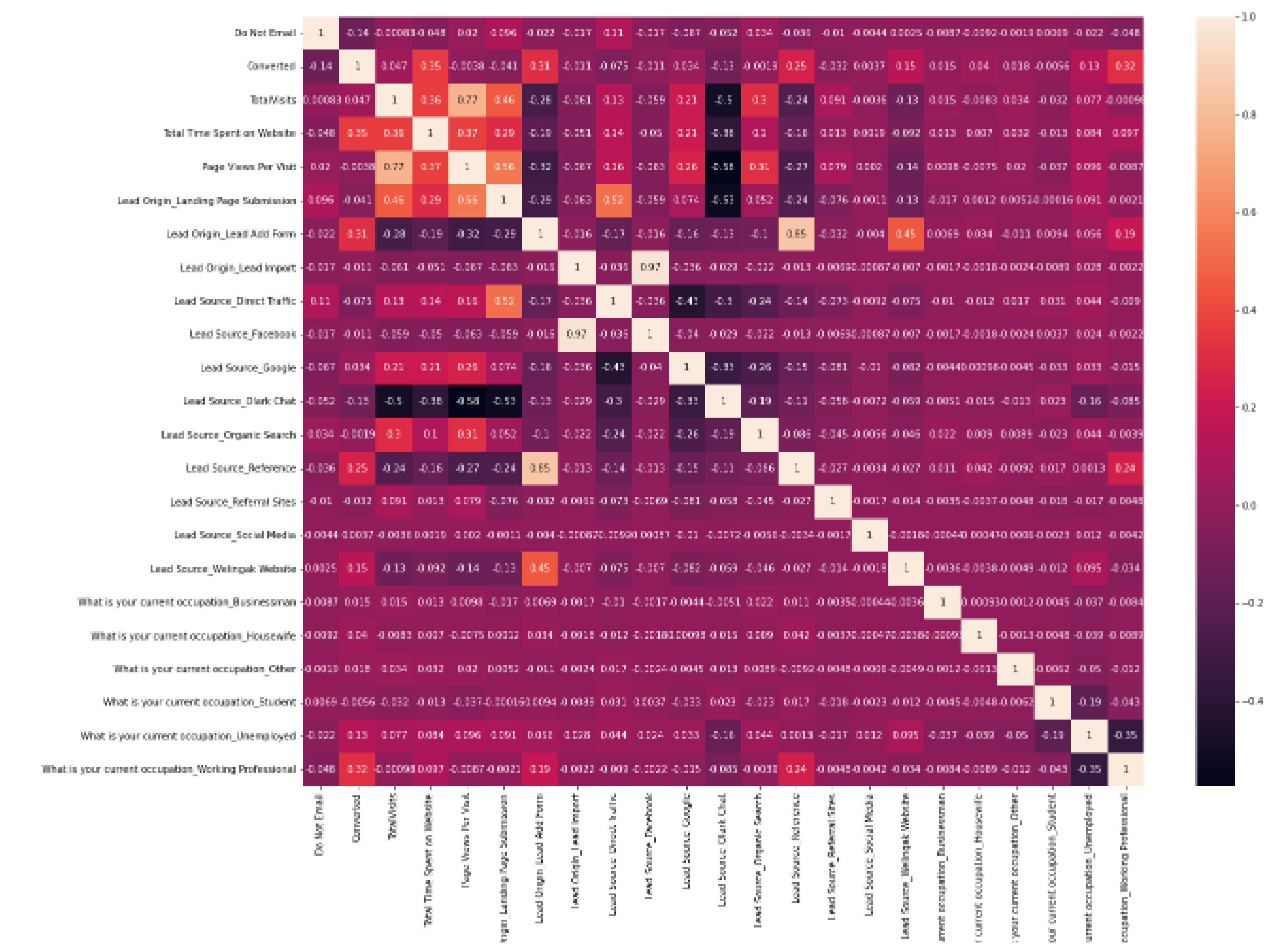


Converted
🟩 0
🟧 1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Others
Email Link Clicked
Olark Chat Conversation
SMS Sent
Last Notable Activity

Categorical Variable Relations



Correlation Matrix



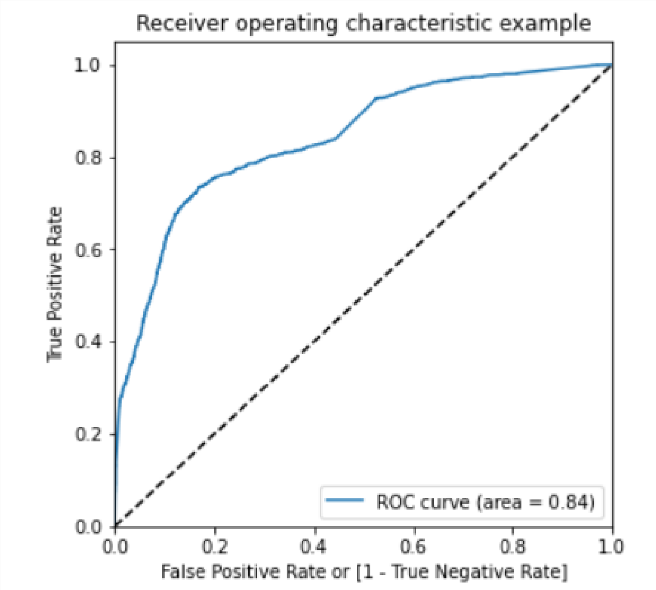
Data Conversion

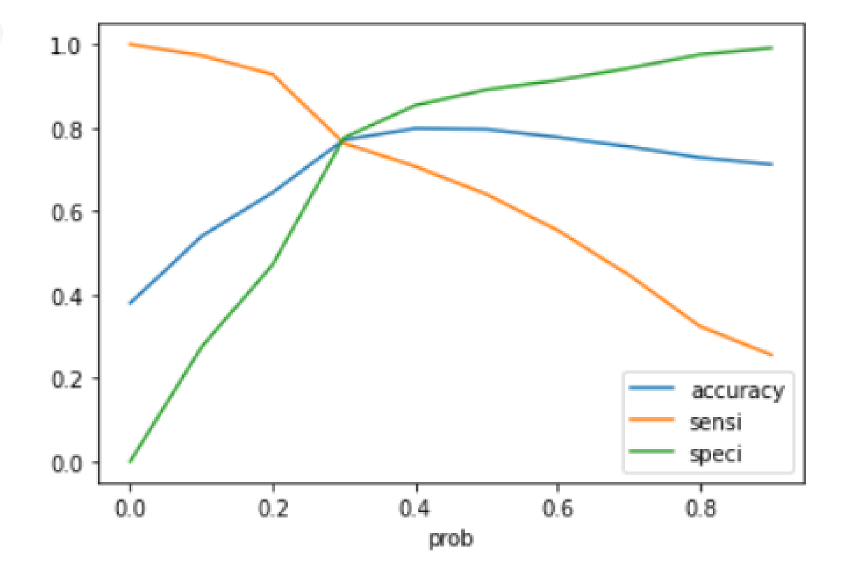
- Numerical Variables are Normalized
- Dummy Variables are created for object type variables

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 77%

ROC Curve





- Finding Optimal Cut off Point
- From the second curve above, 0.3 is the optimum point to take it as a cutoff probability.

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77%, 75% and 77% which are approximately closer to the respective values calculated using trained set.
- Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
    1. Lead Origin_Lead Add Form
    2. What is your current occupation_Working Professional
    3. Total Time Spent on Website