Lead Scoring Assignment

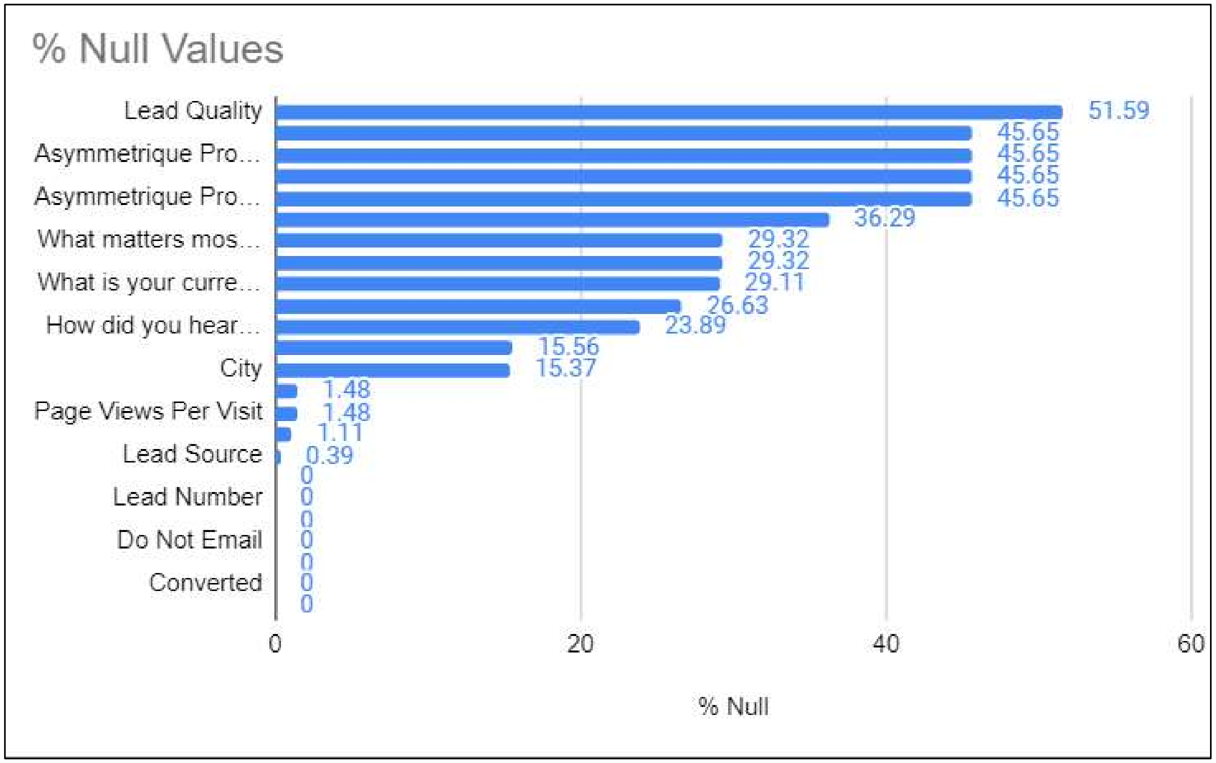
Team members:

1. Jeeva Sundhararajan
2. Nishant Kumar Sinha
3. Shivendu Kumar

Initial checks

- Checked different columns, the data stored in those columns, shape of dataframe and info to understand the importance of each columns
  - Initial data had 9240 rows with 37 columns

- Performed null analysis check the number of null values in each columns



Data cleaning

- Dropped columns namely city/country/prospect_id /lead numbers as they would not add many information to the model.
  - "Asymmetrique Activity Index","Asymmetrique Profile Index“ were also removed as we had the numerical score for the similar columns named : “Asymmetrique Profile score”/” Asymmetrique Activity score”
  - Last notable activity was also dropped as it had similar value with Last activity column

- Dropped the rows containing null values as part of data cleaning

- Post dropping rows with null values we observe that we are left with 4k data points which have 35-65 % split between sample which converted to lead vs which did not convert to lead

- Converted the categorical values to dummy variables so that relevance of each category can be measured on the model
  - For columns which had yes or no were converted to 1 and 0s while keeping the same column using map function
  - For columns which had multiple categories we converted them to dummy variables using “get_dummies”

    for ref olark chart had a phone conversation were part of lad orgin column which have been converted to dummy variables

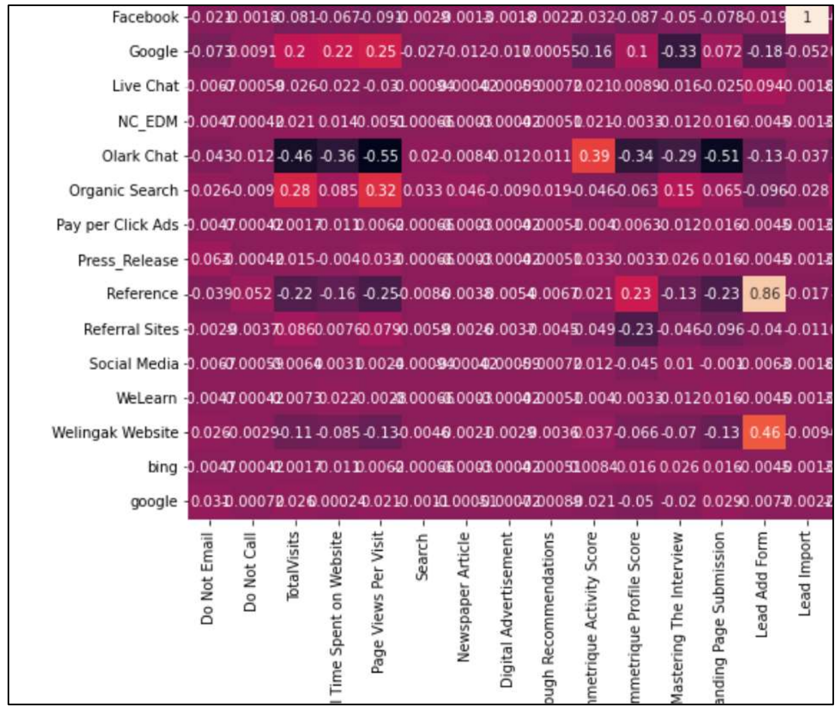
	Prospect ID	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Search	Magazine	...	Email Received	Form Submitted on Website	Had a Phone Conversation	Olark Chat Conversation
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	Olark Chat	0	0	0	0.0	0	0.0	0	0	...	0	0	0	0
1	2a272436-5132-4136-86fa-dcc88c88f482	Organic Search	0	0	0	5.0	674	2.5	0	0	...	0	0	0	0
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	Direct Traffic	0	0	1	2.0	1532	2.0	0	0	...	0	0	0	0
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	Direct Traffic	0	0	0	1.0	305	1.0	0	0	...	0	0	0	0
4	3256f628-e534-4826-9d63-4a8b88782852	Google	0	0	1	2.0	1428	1.0	0	0	...	0	0	0	0

- Outlier were also removed from the dataframe to remove their impact

Model preparation

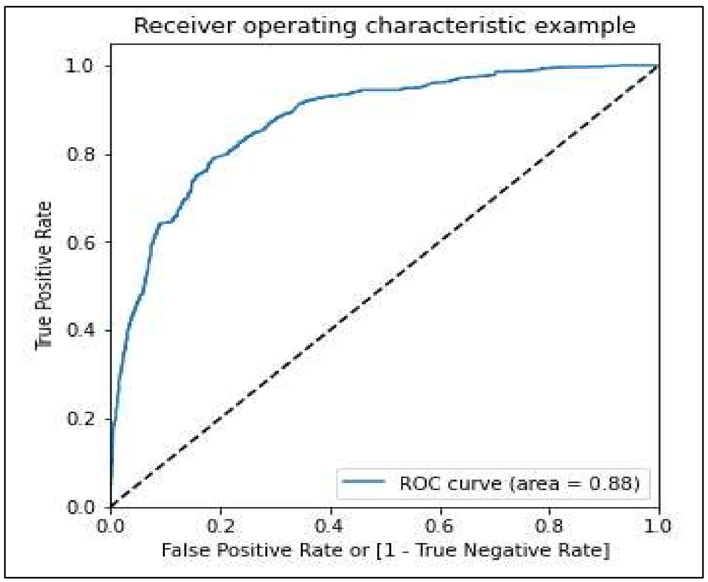
- Logistic regression model was built using the cleaned data.
  - The dependent variable was “converted” from the dataframe

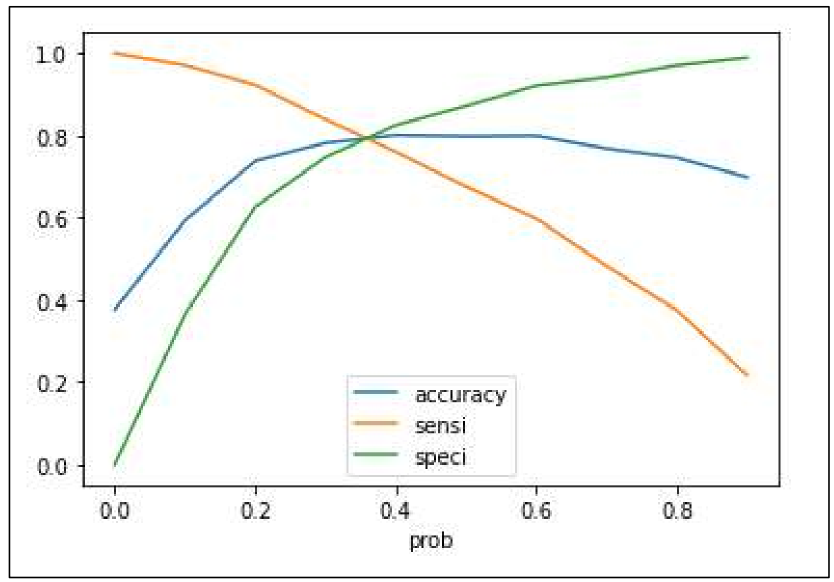
- The data was split 70-30 between train and test respectively
  - The data was scaled using standardscaler so that coefficient generation is stable for all variables
  - Multicollinearity was also checked by checking the correlation heatmap and highly corelated values were removed
  - We see high collinearity between facebook and lead import



- Following columns were removed as they had negligible values as 1 and donot show up in correlation heat map: 'Magazine','X Education Forums', 'Newspaper','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','Visited Booth in Tradeshow','blog’
- Some are part of dummy variable and some were found to have high correlation
- Initial model w/o RFE had many varaiables with greater P values
- RFE was then used to reduce the components to 15 variables
- Final model was then used to classify the users on the basis of cut off score of 0.5 and accuracy was found to be ~80% but there were certain variables with higher p value

- VIF values were also checked and post analysis, following columns/variables were also dropped : "Email Bounced","Social Media“
- ROC curve and accuracy vs sensitivity vs specificity curve was plotted to find better cut off value.
  - Since no additional information on any specific requirement by the client was provided the most accurate value was found to be at 0.38





Final model characteristics

- The built model had the following characteristics
  - Accuracy: 80.3%
  - Precision: 71.95%
  - Recall: 78.1%
  - Sensitivity: 78.1%
  - Specificity: 81.59%

  ** as per requirement one or more of these parameter will have to be traded off for the other by shifting the probability cut off

Business aspect

- Since we want to increase the hot leads a lower cut off less than 40% would give good number of users who can form a potential lead.
- Since we have provided a probability score as well, users with highest score will have the highest chance of conversion to a hot lead and the business should start with targeting these users as they turns out to be the most potential customers
- If the requirements comes at a time when the number of calls/contacts cannot be high we will have to increase the cut off so that calls are made only to the highest scored customers only maybe top 30% [cut off : 70%] as we would have to increase