LEAD SCORING CASE STUDY

PRESENTED BY

M CHAITANYA
G AAKASH

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

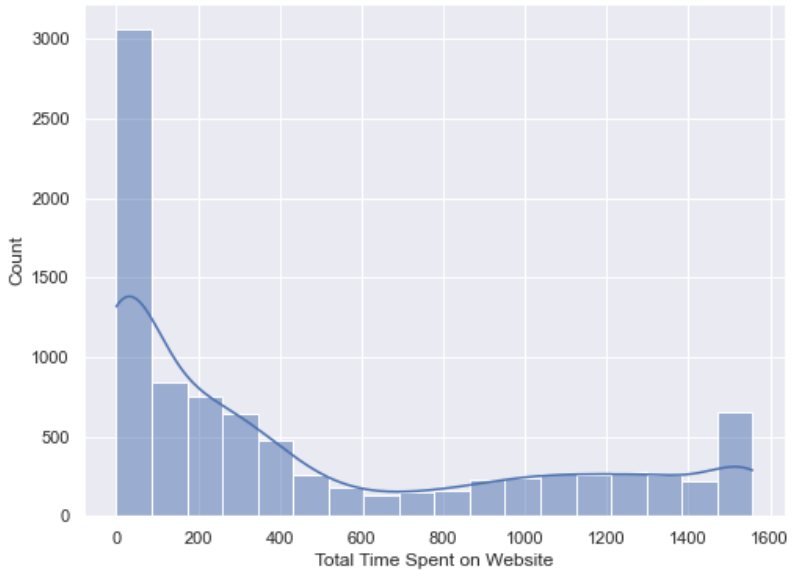
SOLUTION METHODOLOGY

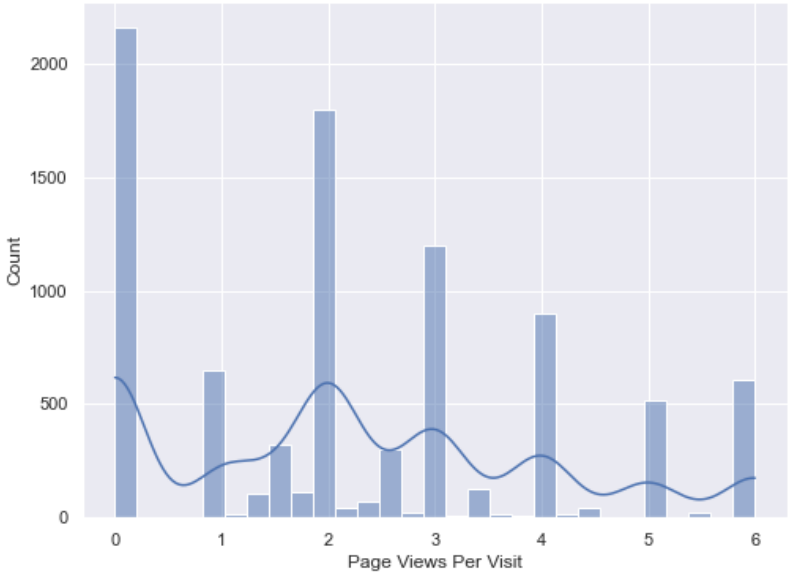
- Data cleaning and data manipulation.
  - Check and handle duplicate data.
  - Check and handle NA values and missing values.
  - Drop columns, if it contains large amount of missing values and not useful for the analysis.
  - Imputation of the values, if necessary.
  - Check and handle outliers in data.
- EDA
- Univariate data analysis: value count, distribution of variable etc.
- 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

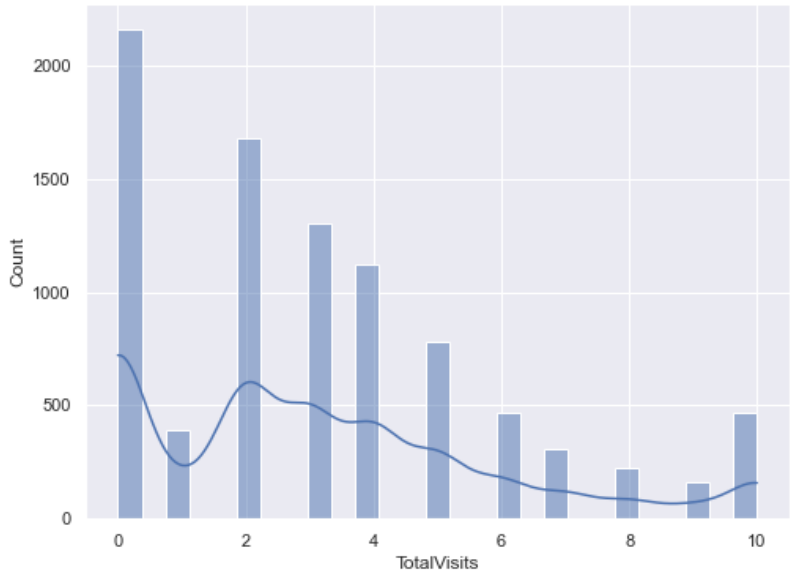
DATA MANIPULATION

- Total Number of Rows =37, Total Number of Columns =9240.
- Removing the “Prospect ID” which is not necessary for the analysis and set “Lead Number” as index.
- Converted “Select” value in columns to null values.
- Converted Yes / No values to 1 and 0 respectively for analysis.
- Single value features like “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” etc. have been dropped.
- After checking for the value counts for some of the object type variables, City, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score, Lead Profile, Lead Quality, Tags, What matters most to you in choosing a course, What is your current occupation. How did you hear about X education, Country, have more than 25% as missing value and are therefore dropped.

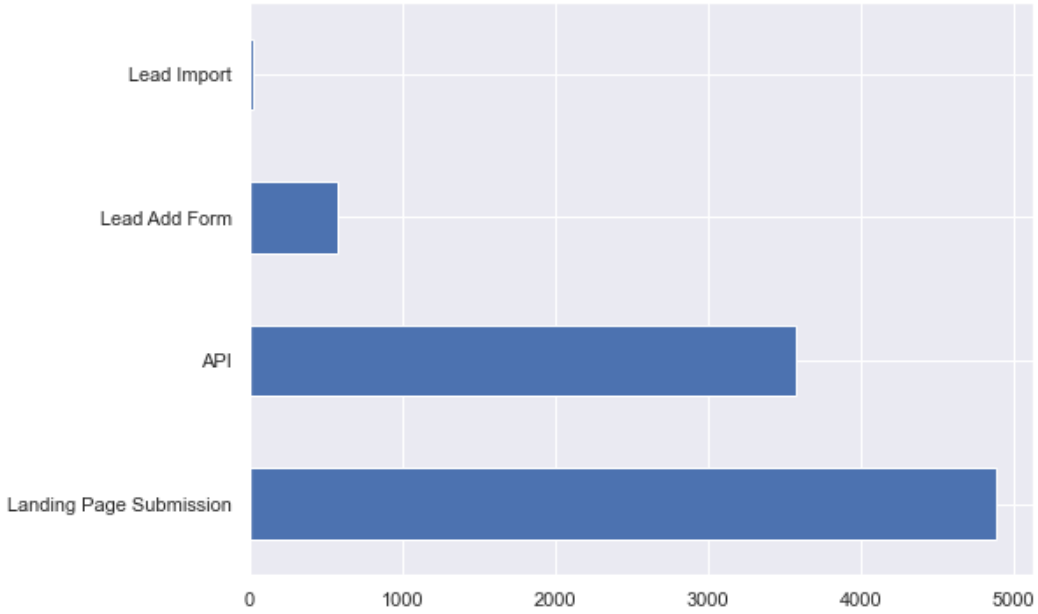
UNIVARIATE NUMERICAL COLUMN ANALYSIS

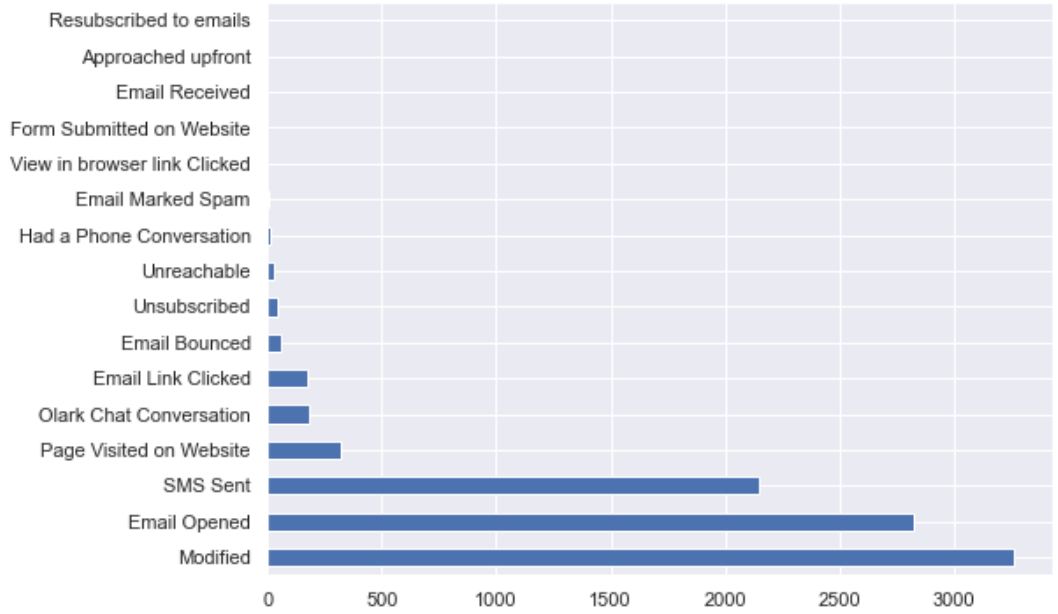




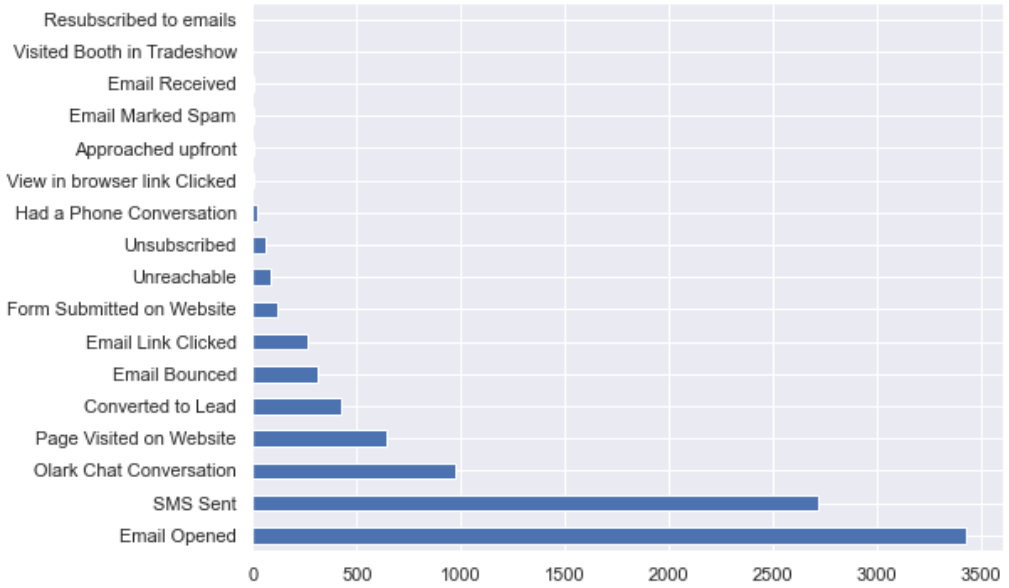


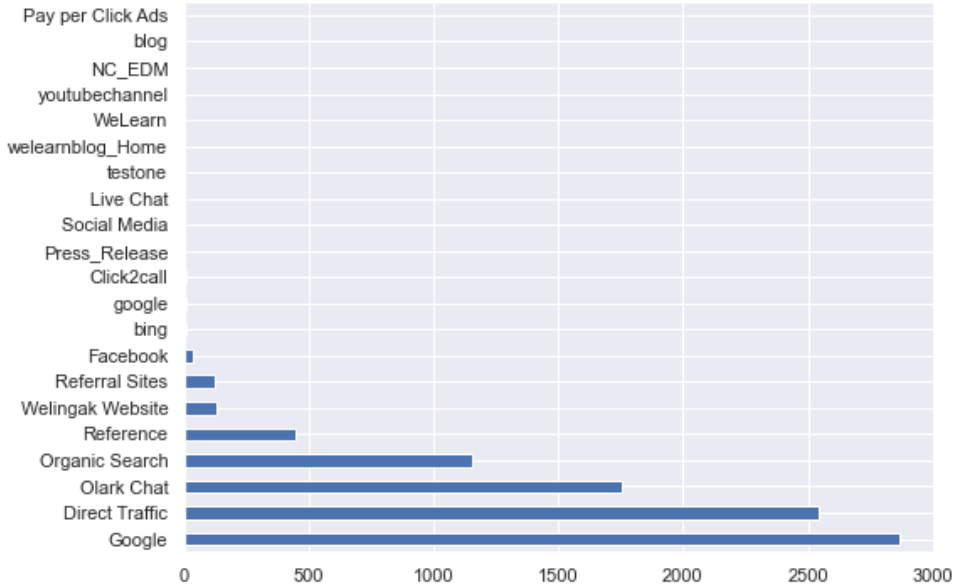
CATEGORICAL VARIABLE COUNT PLOT - 1



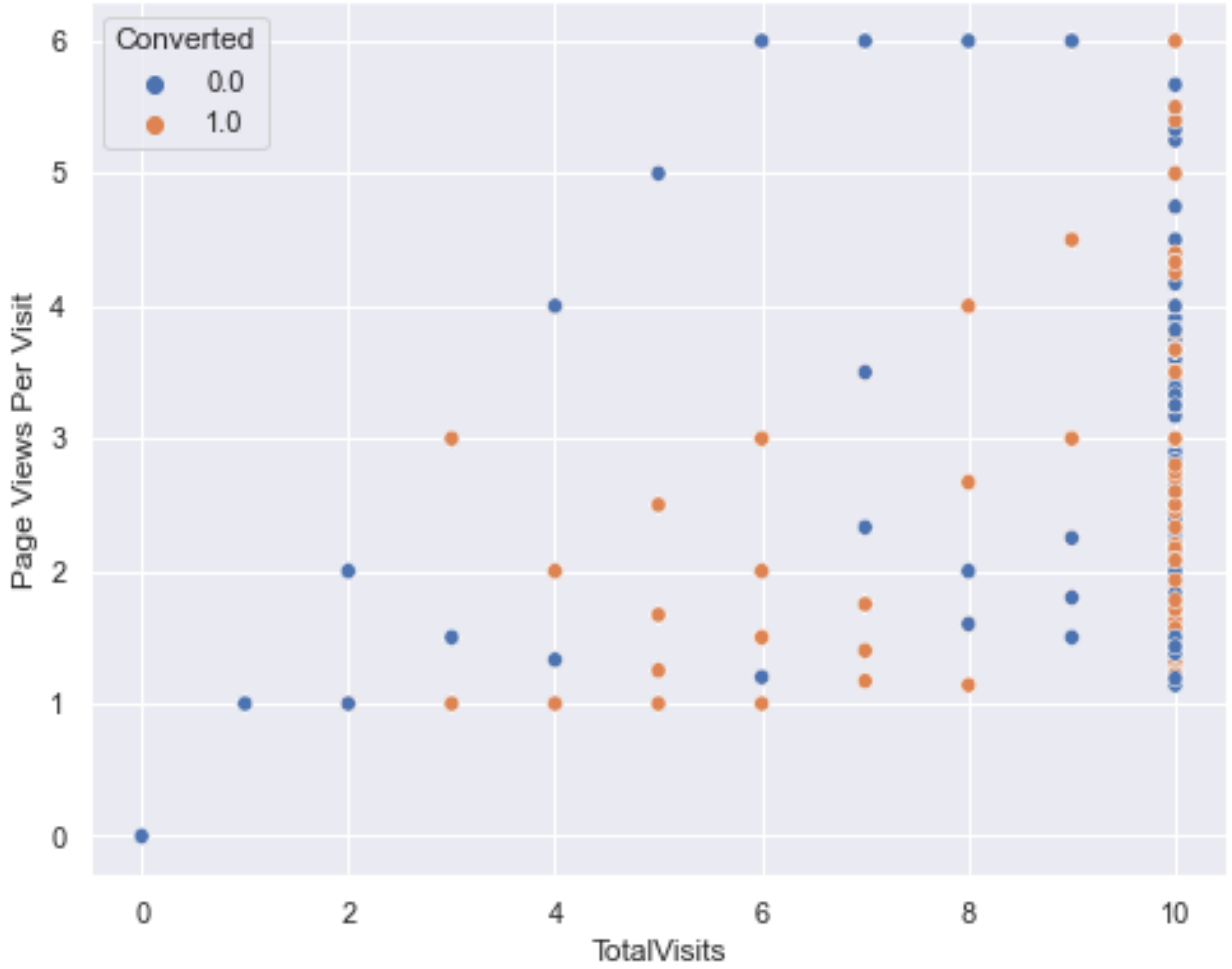


CATEGORICAL VARIABLE COUNT PLOT - 2

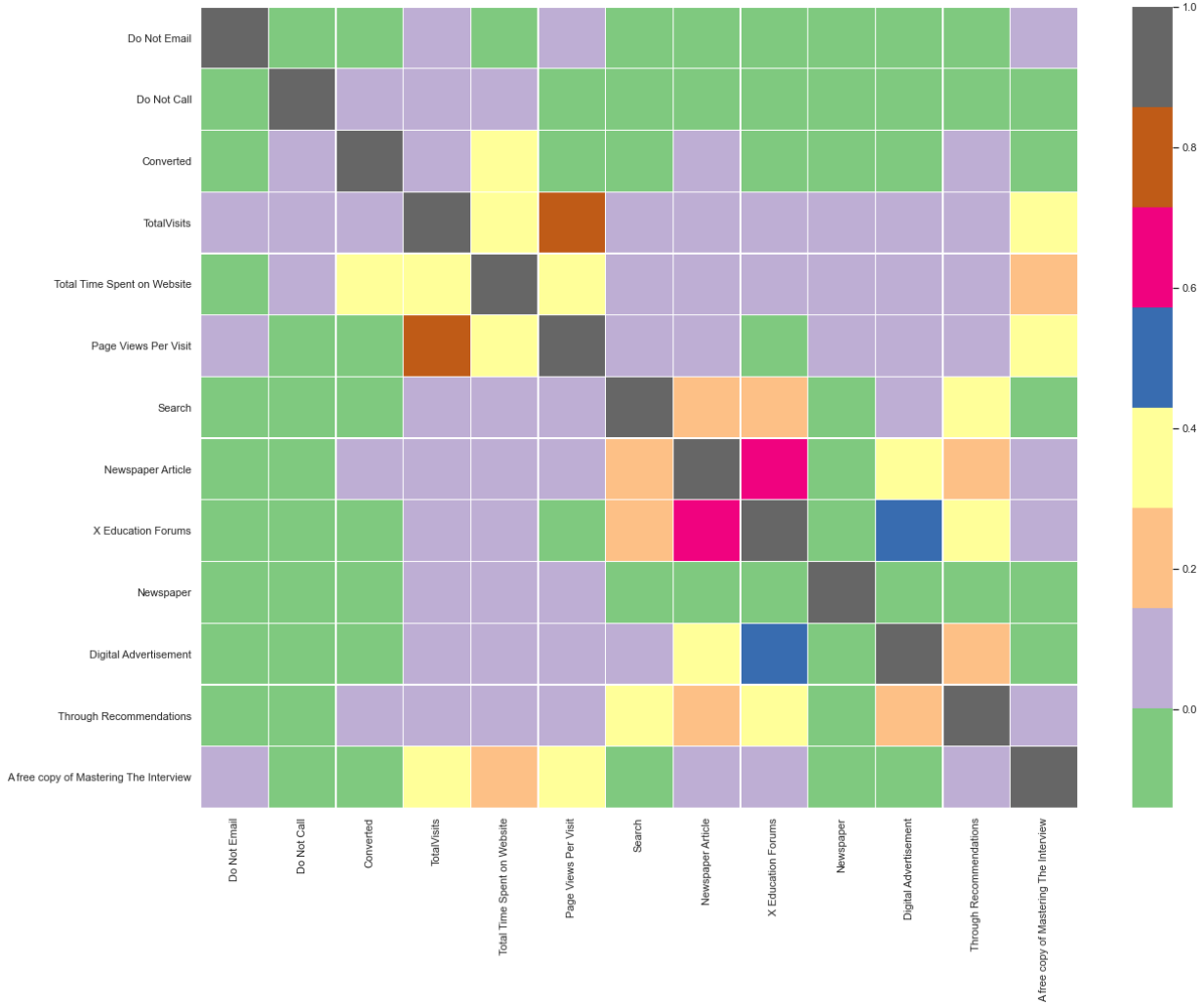




BIVARIATE ANALYSIS



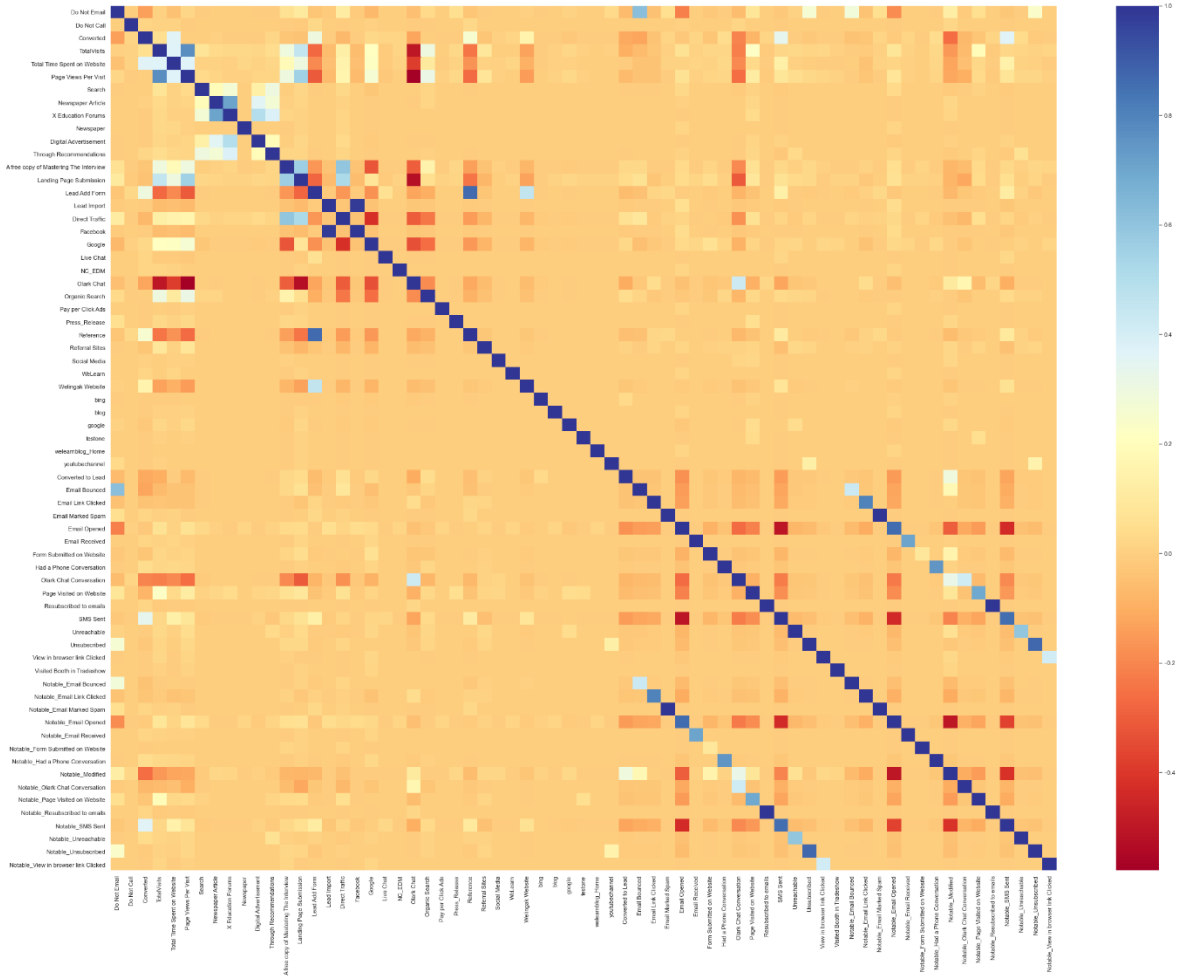
COLLINEARITY HEATMAP



DATA CONVERSION

- Numerical Variables like “Total Time Spent on Website”, “Total Visits” and “Total Page Views per Visit” are scaled and normalized.
- Dummy Variables are created for object type categorical variables.
- Total Rows for Analysis: 9074
- Total Columns for Analysis: 67

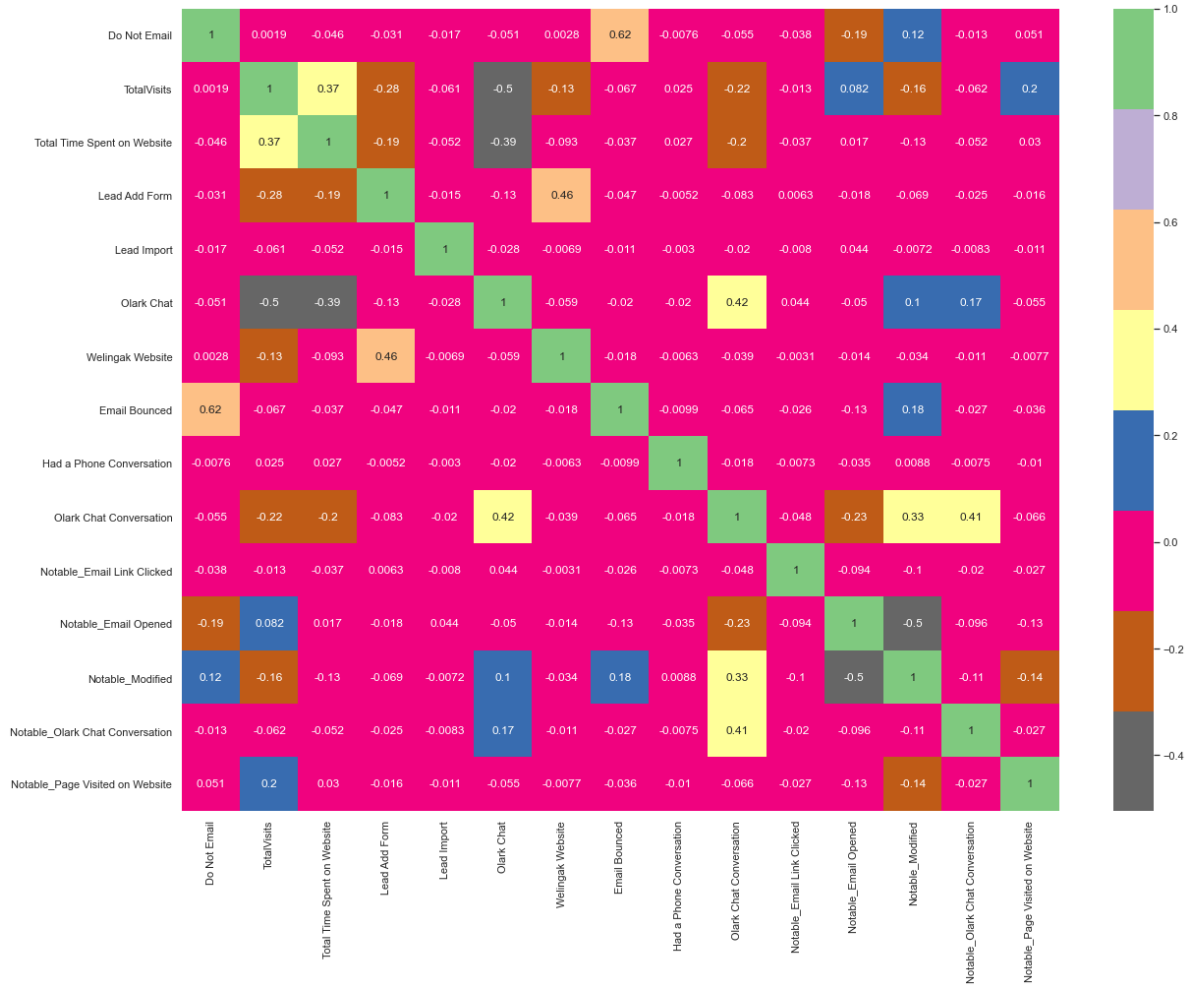
FINAL COLUMN COLLINEARITY HEATMAP



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
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- With the current cut-off of 0.4, we have Accuracy of 79%, Precision of around 72% and Recall of around 76% on our training set.

FINAL MODEL PARAMETERS COLLINEARITY



MODEL PREDICTIONS

- With the model being finalised and obtaining a decent score on the training set on all three parameters, we continue to use the model to make predictions on our test set.
- Using the same cut-off of 0.4 probability score, we have managed to get Accuracy of 80%, Precision of around 73% and Recall of around 75%.
- This score is inline with our training set score, making our model stable.

CONCLUSION

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

1) Lead origin is Lead add format
2) Total time spent on Website
3) Lead Source is Welingak Website
4) Last Activity was Had a Phone Conversation
5) Lead Origin as Lead Import
6) Last Activity was Olark Chat
7) Total Visits to the website
8) Page Views per Visit
9) Did not choose Do not Email option
10) Did not have the last activity as Olark Chat conversation
11) Did not have last notable activity as Email Opened
12) Did not have Email bounced
13) Did not Selinga have last notable activity as Modified
14) Did not have last notable activity as Olark Chat Conversation
15) Did not have last notable activity as Page Visited on Website
16) Did not have last notable activity as Email Link Clicked.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.