Lead Score Case Study

SUBMITTED BY:

PRAVEEN MAURYA

SOURAV KARMAKAR

Problem Statement

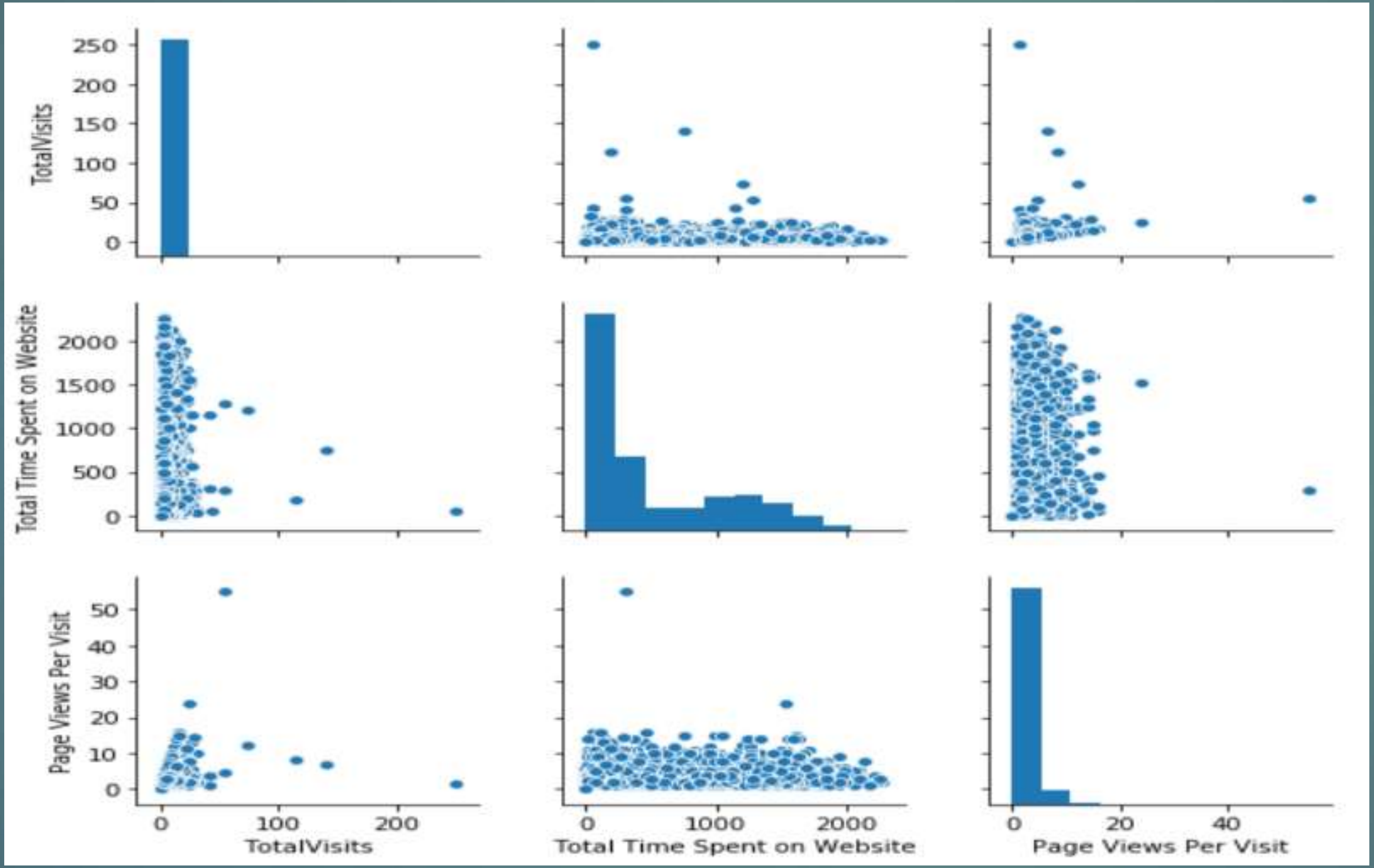
- An ‘X’ Education Company, offers online courses to industry professionals, gets many people landing on their website from the advertisements posted on various platforms and search engines like Google. The interested people provide their contact details to the company.
- These people are called as leads and are followed by the sales team to convert them, i.e. sell them the course. The conversion to lead ratio is currently 30%. The company wants to improve this ratio by following up only the potential leads that would most probably be interested in buying the course and filter out other.
- As a Data Scientist, we have to build a Machine Learning model using Logistic Regression that helps the company in achieving this and reducing the manpower in irrelevant follow ups by the sales team.

Reading, understanding and cleansing data

- The given data set has 37 columns and 9240 rows with two unique columns, ‘Lead Number’ and ‘Prospect ID’. It has few insignificant columns, ‘Tags’, 'Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Profile Score','Asymmetrique Activity Score‘.
- It has many non-responsive values with name ‘Select’ that we replaced with null values. It has 3 columns with more than 50% null values.
- After removing the insignificant variables, variables with high null values and records with more than 5 null values we are left with 28 variables and 9191 rows. 99.47% of rows have been retained.
- Imputed the null values in remaining variables with mean/median/mode.

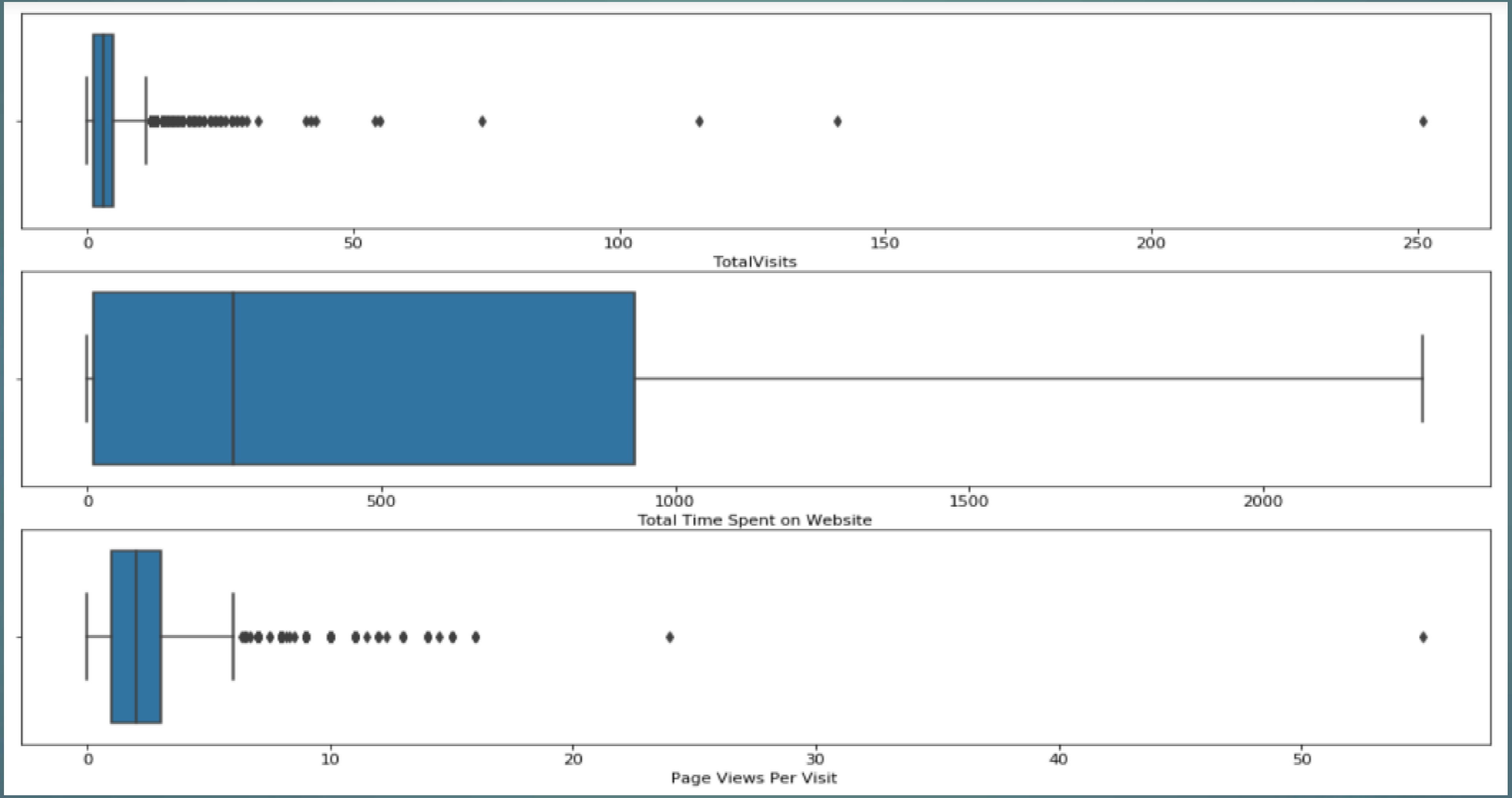
Visualizing data

- From the below pairplots we can that the variable are not collinear.

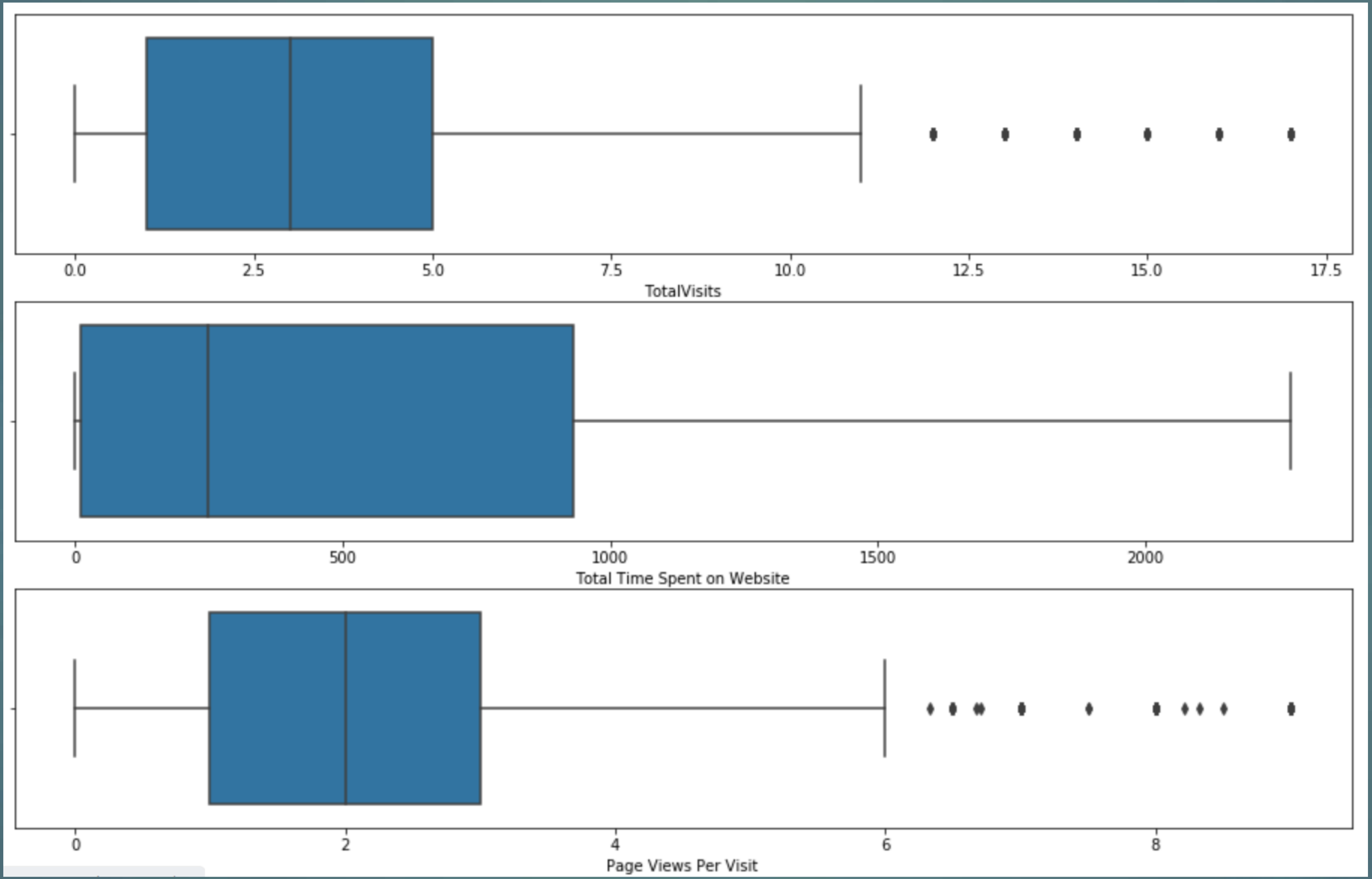


Outlier Analysis

- Outliers found in ‘Total Visits’ and ‘Page Views Per Visit’



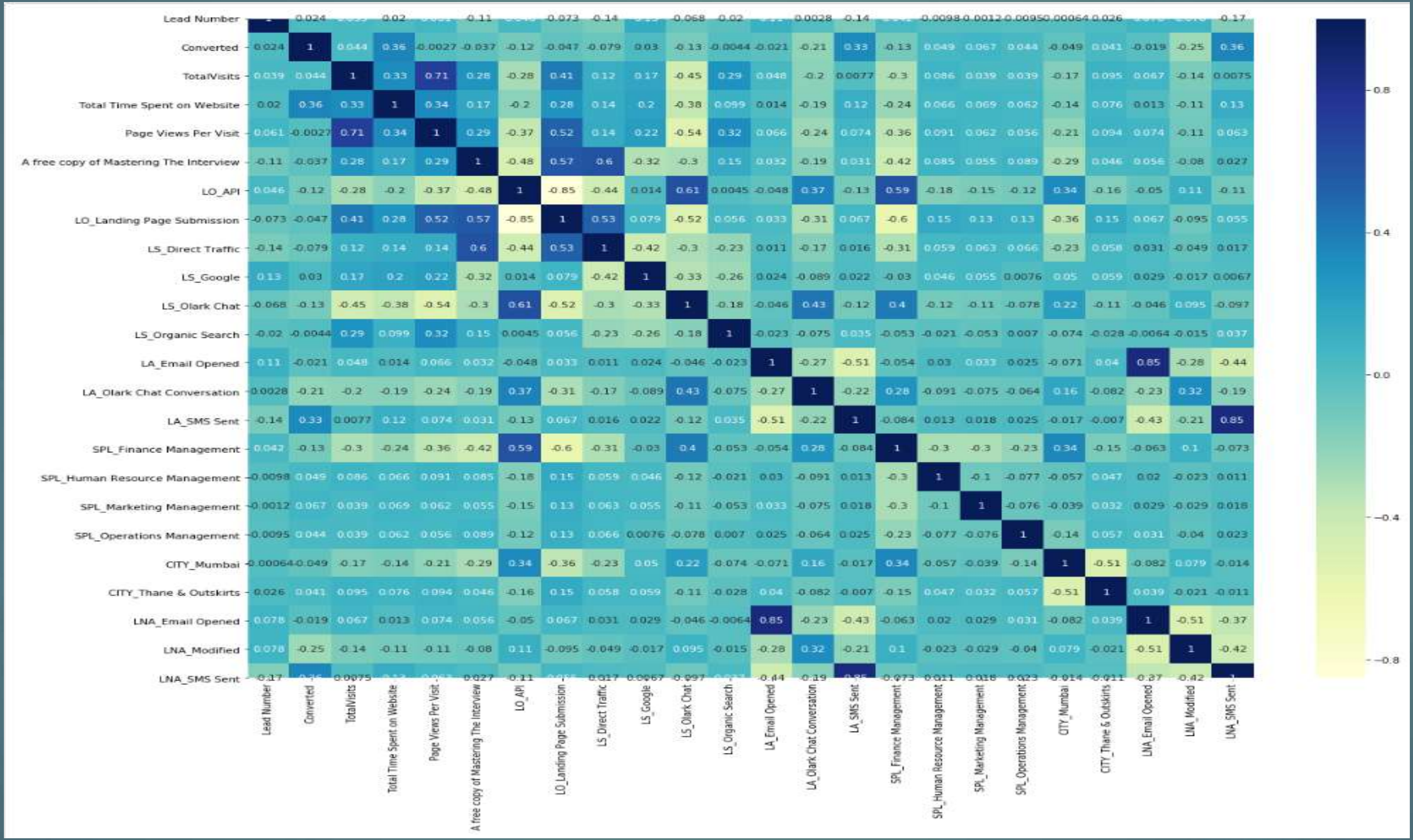
- Treated outliers by capping the data with an upper limit of 99%



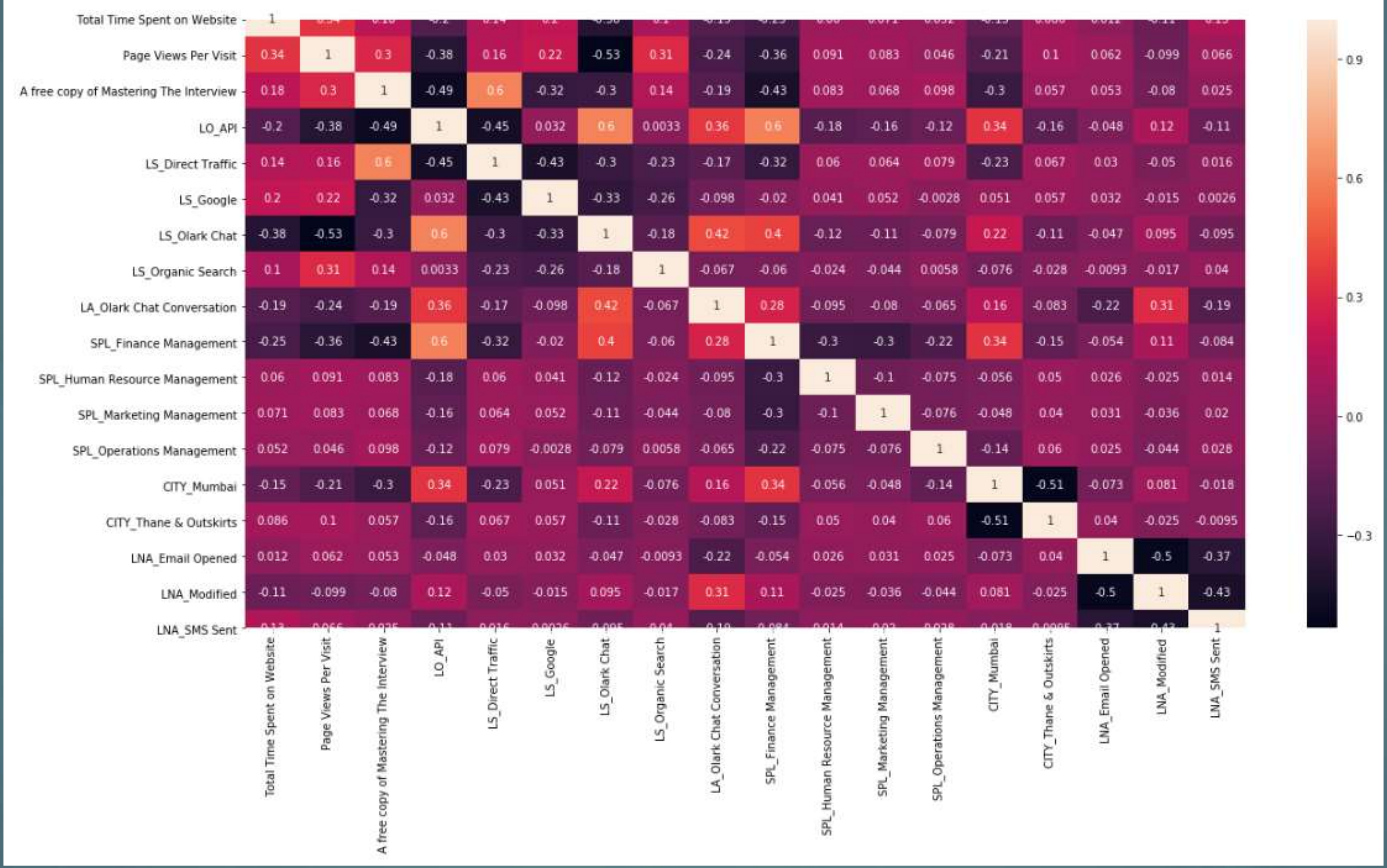
Data Preparation

- Removed 16 variables with very highly skewed values
- Mapped Yes/No values to binary 1/0
- Performed bucketing to keep most significant values in categorical variables and treated remaining as ‘Others’
- Created dummy variables for 6 columns and assigned them prefixes
- Dropped columns for which dummy variables have been created and concatenated the dummy columns with the original dataframe and called it ‘Lead’
- We have 24 columns and 9191 rows in Lead dataframe
- Created X=Predictor variables and y=Response variable(Converted)
- Split this data into train and test with a ratio of 70 : 30.
- Scaled the numerical variable using StandardScaler in X_train
- The lead conversion rate is 38.34%

- Plotted correlation matrix and dropped highly correlated variables from both train and test datasets



- Correlation matrix after dropping highly correlated variables.



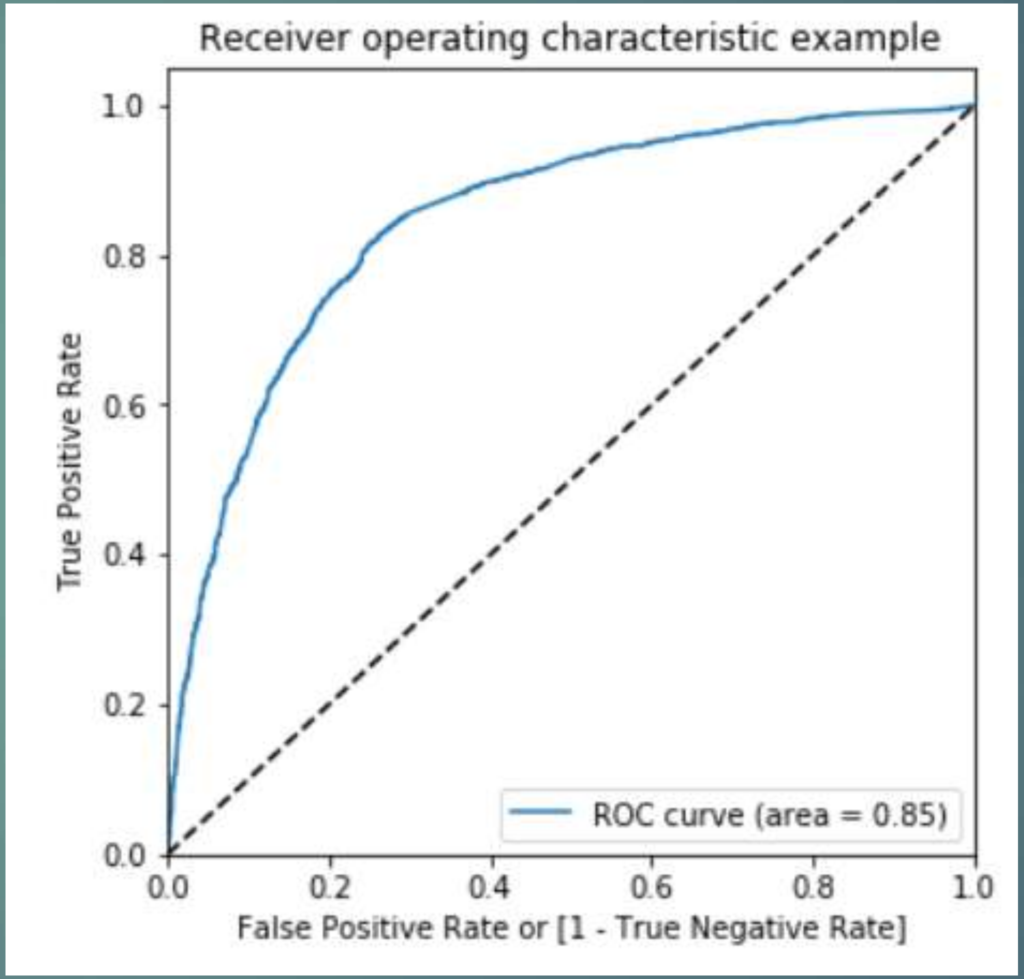
	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview	LO_API	LS_Direct Traffic	LS_Google	LS_Olark Chat	LS_Organic Search	LA_Olark Chat Conversation	SPL_Finance Management	SPL_Human Resource Management	SPL_Marketing Management	SPL_Operations Management	CITY_Mumbai	CITY_Thane & Outskirts	LNA_Email Opened	LNA_Modified	LNA_SMS Sent
Total Time Spent on Website	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Page Views Per Visit	0.34	1	0.3	-0.38	0.16	0.22	-0.53	0.31	-0.24	-0.36	0.091	0.083	0.046	-0.21	0.1	0.062	-0.099	0.066
A free copy of Mastering The Interview	0.18	0.3	1	-0.49	0.6	-0.32	-0.3	0.14	-0.19	-0.43	0.083	0.068	0.098	-0.3	0.057	0.053	-0.08	0.025
LO_API	-0.2	-0.38	-0.49	1	-0.45	0.032	0.6	0.0033	0.36	0.6	-0.18	-0.16	-0.12	0.34	-0.16	-0.048	0.12	-0.11
LS_Direct Traffic	0.14	0.16	0.6	-0.45	1	-0.43	-0.3	-0.23	-0.17	-0.32	0.06	0.064	0.079	-0.23	0.067	0.03	-0.05	0.016
LS_Google	0.2	0.22	-0.32	0.032	-0.43	1	-0.33	-0.26	-0.098	-0.02	0.041	0.052	-0.0028	0.051	0.057	0.032	-0.015	0.0026
LS_Olark Chat	-0.38	-0.53	-0.3	0.6	-0.3	-0.33	1	-0.18	0.42	0.4	-0.12	-0.11	-0.079	0.22	-0.11	-0.047	0.095	-0.095
LS_Organic Search	0.1	0.31	0.14	0.0033	-0.23	-0.26	-0.18	1	-0.067	-0.06	-0.024	-0.044	0.0058	-0.076	-0.028	-0.0093	-0.017	0.04
LA_Olark Chat Conversation	-0.19	-0.24	-0.19	0.36	-0.17	-0.098	0.42	-0.067	1	0.28	-0.095	-0.08	-0.065	0.16	-0.083	-0.22	0.31	-0.19
SPL_Finance Management	-0.25	-0.36	-0.43	0.6	-0.32	-0.02	0.4	-0.06	0.28	1	-0.3	-0.3	-0.22	0.34	-0.15	-0.054	0.11	-0.084
SPL_Human Resource Management	0.06	0.091	0.083	-0.18	0.06	0.041	-0.12	-0.024	-0.095	-0.3	1	-0.1	-0.075	-0.056	0.05	0.026	-0.025	0.014
SPL_Marketing Management	0.071	0.083	0.068	-0.16	0.064	0.052	-0.11	-0.044	-0.08	-0.3	-0.1	1	-0.076	-0.048	0.04	0.031	-0.036	0.02
SPL_Operations Management	0.052	0.046	0.098	-0.12	0.079	-0.0028	-0.079	0.0058	-0.065	-0.22	-0.075	-0.076	1	-0.14	0.06	0.025	-0.044	0.028
CITY_Mumbai	-0.15	-0.21	-0.3	0.34	-0.23	0.051	0.22	-0.076	0.16	0.34	-0.056	-0.048	-0.14	1	-0.51	-0.073	0.081	-0.018
CITY_Thane & Outskirts	0.086	0.1	0.057	-0.16	0.067	0.057	-0.11	-0.028	-0.083	-0.15	0.05	0.04	0.06	-0.51	1	0.04	-0.025	-0.0095
LNA_Email Opened	0.012	0.062	0.053	-0.048	0.03	0.032	-0.047	-0.0093	-0.22	-0.054	0.026	0.031	0.025	-0.073	0.04	1	-0.5	-0.37
LNA_Modified	-0.11	-0.099	-0.08	0.12	-0.05	-0.015	0.095	-0.017	0.31	0.11	-0.025	-0.036	-0.044	0.081	-0.025	-0.5	1	-0.43
LNA_SMS Sent	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1

Building Model

- Using Recursive Feature Elimination(RFE) selected top 13 most significant variables and created of list of these variables
- Added constant and build the model with 13 variables
- Variables ‘SPL_Operations Management’ and ‘LNA_Email Opened’ were statistically insignificant as their p-values were high. The VIF values were low
- Predicted using the above model and defined initial cut-off of 0.5 conversion probability. Lead Score(Conversion Probability > 0.5) ➔ 1
- Sensitivity = 66%, Accuracy = 78% ,based on confusion matrix created for above predicted values
- After dropping all insignificant variables over iteration we were left with 11 significant variables
- There is no significant difference in accuracy and sensitivity at this stage for the model

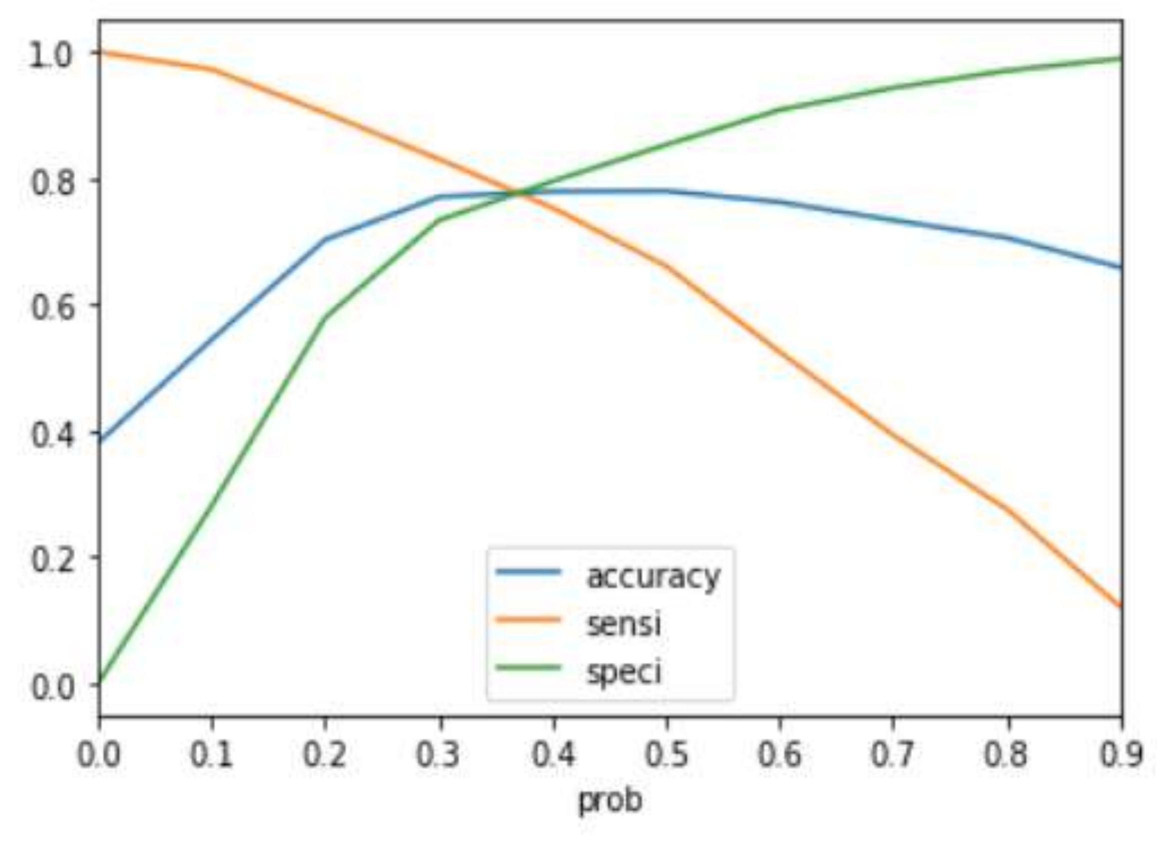
ROC Curve to get the best cut-off

- From the ROC curve we can see that significant amount of area is under the curve which suggests that the built model is a good model.



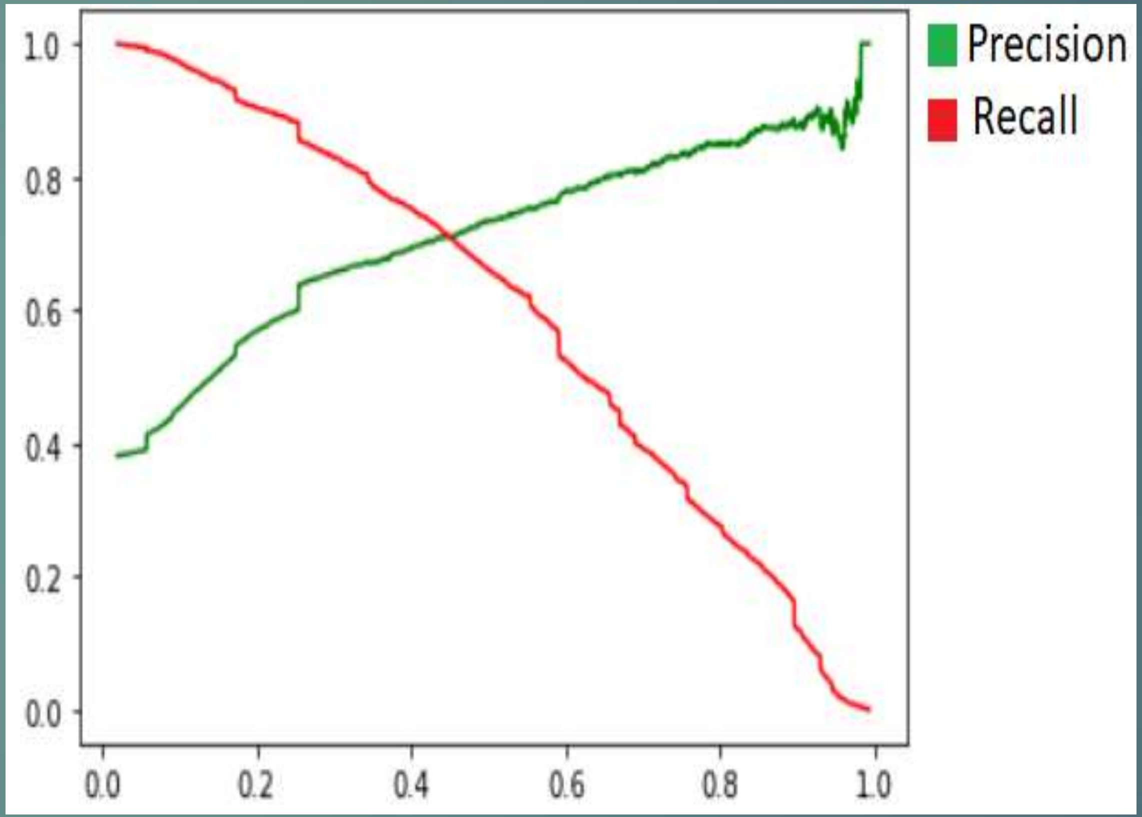
Optimum Cut-off ?

- From the plot between accuracy, sensitivity and specificity it is clear that the cut-off lies between 0.3 and 0.4
- Chose the cut-off of 0.3 and updated the final predicted values
- For this model:
  - Accuracy ➔ 77%
  - Sensitivity ➔ 73%
  - Specificity ➔ 83%



Precision – Recall trade off

- From the plot between precision and recall we can see that they intersect at around 0.4



Making Predictions

- Used the learnt model from training data set and applied on the test dataset
- Following are the quality metrics after making predictions on test dataset:
  - Accuracy → 78%
  - Sensitivity → 82%

Conclusion

- The top 3 variables affecting the lead conversion are:
  - Lead Source
  - Last Notable Action
  - Last Action
- The top 3 dummy variables affecting the lead conversion are:
  - LS_Direct Traffic
  - LS_Organic Search
  - LS_Google