Lead Score Case Study

By:

Mahima Bansal

Raviraj Gundety

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 38 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:

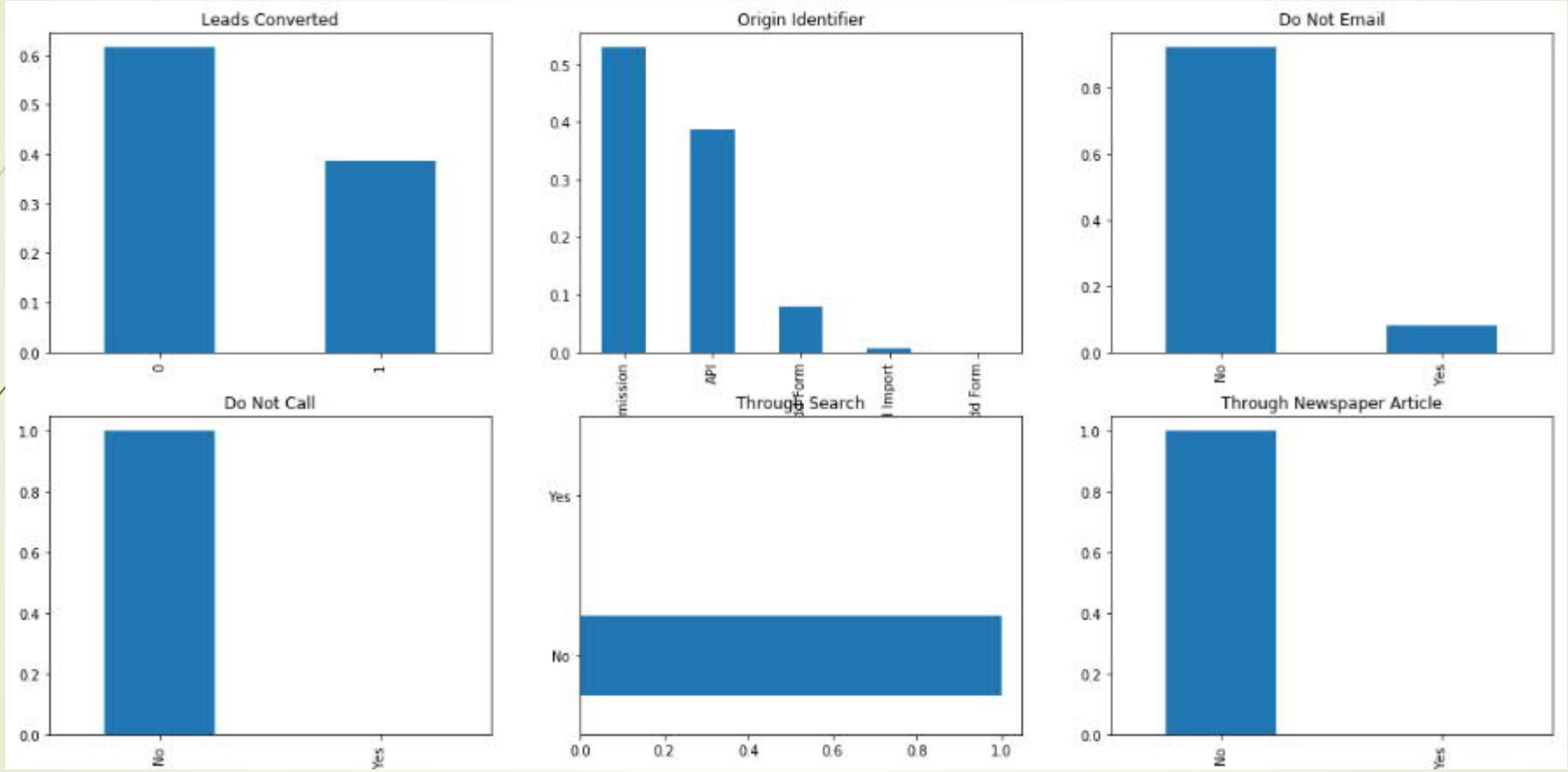
- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers 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.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”, “Get updates on DM Content”,” I agree to pay the amount through cheque” have single value for every lead. So we dropped them as felt they are irrelevant for modelling.
- Variables having more than 40% of missing values were dropped as right decision cannot be reached taking them into account.
- There were data values like select in some columns .. Where such values were more than 40% , deleted those variables and where the values were below it, handled them by replacing there values like “Mumbai” for column like “City”
- For some we corrected the spelling when a variable had same data value for 2 columns but spelling was different as in “Lead Source”
- Removed some variable as they were highly skewed as taking them for modelling would have given wrong result. Below are the names of few such features:

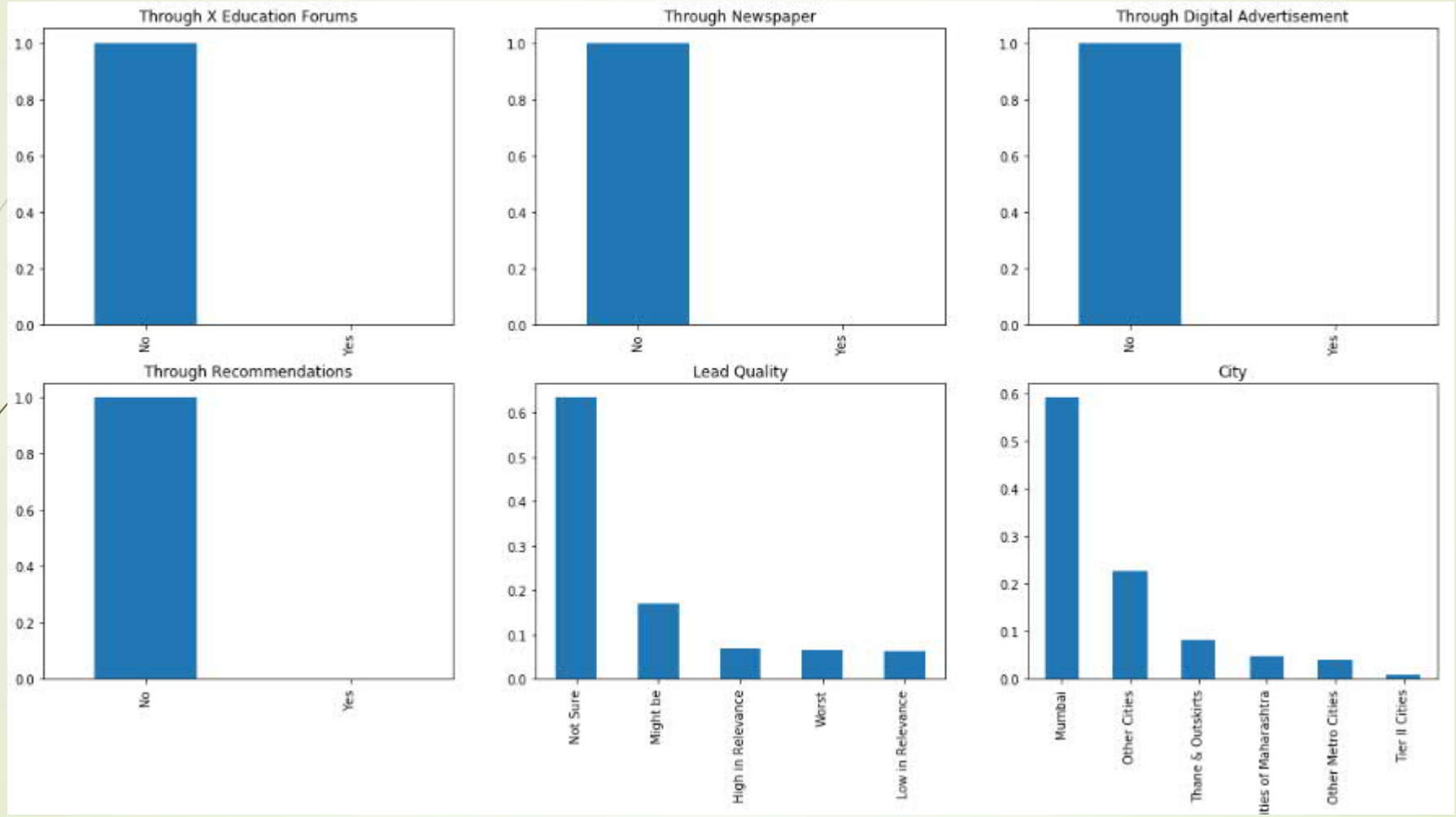
  'Do Not Email', 'Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', "What matters most to you in choosing a course", 'What is your current occupation'

Univariate Analysis:



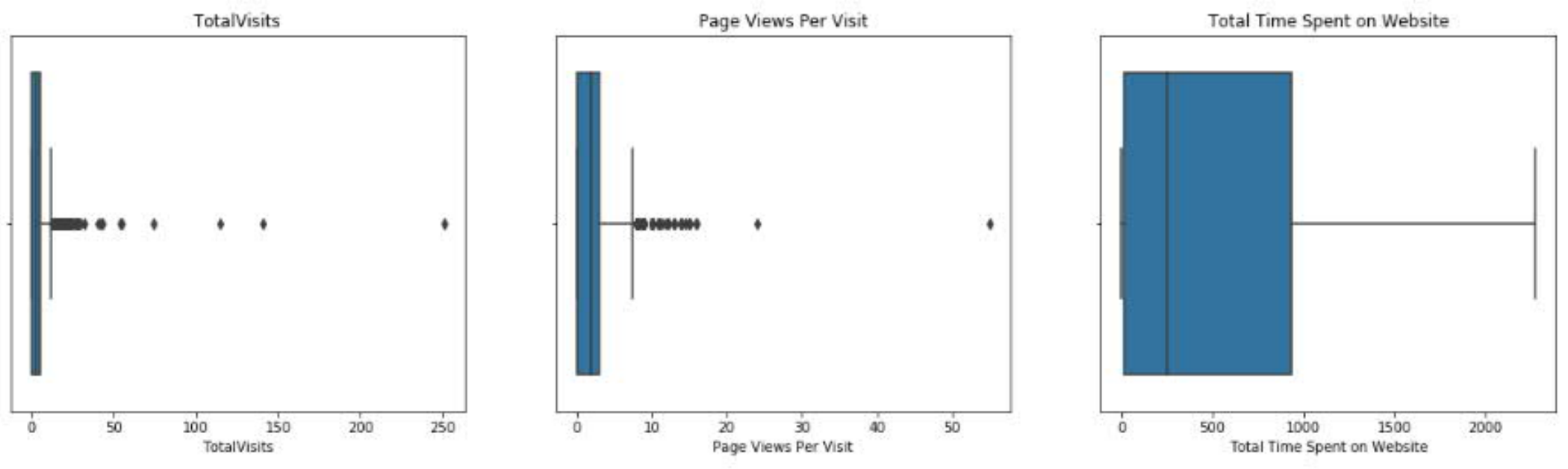
From the above plot, we can see that for features like 'Do Not Email', 'Do Not Call', 'Search', 'Newspaper Article', have highly imbaanced values and should not be taken for modelling.

Univariate Analysis:



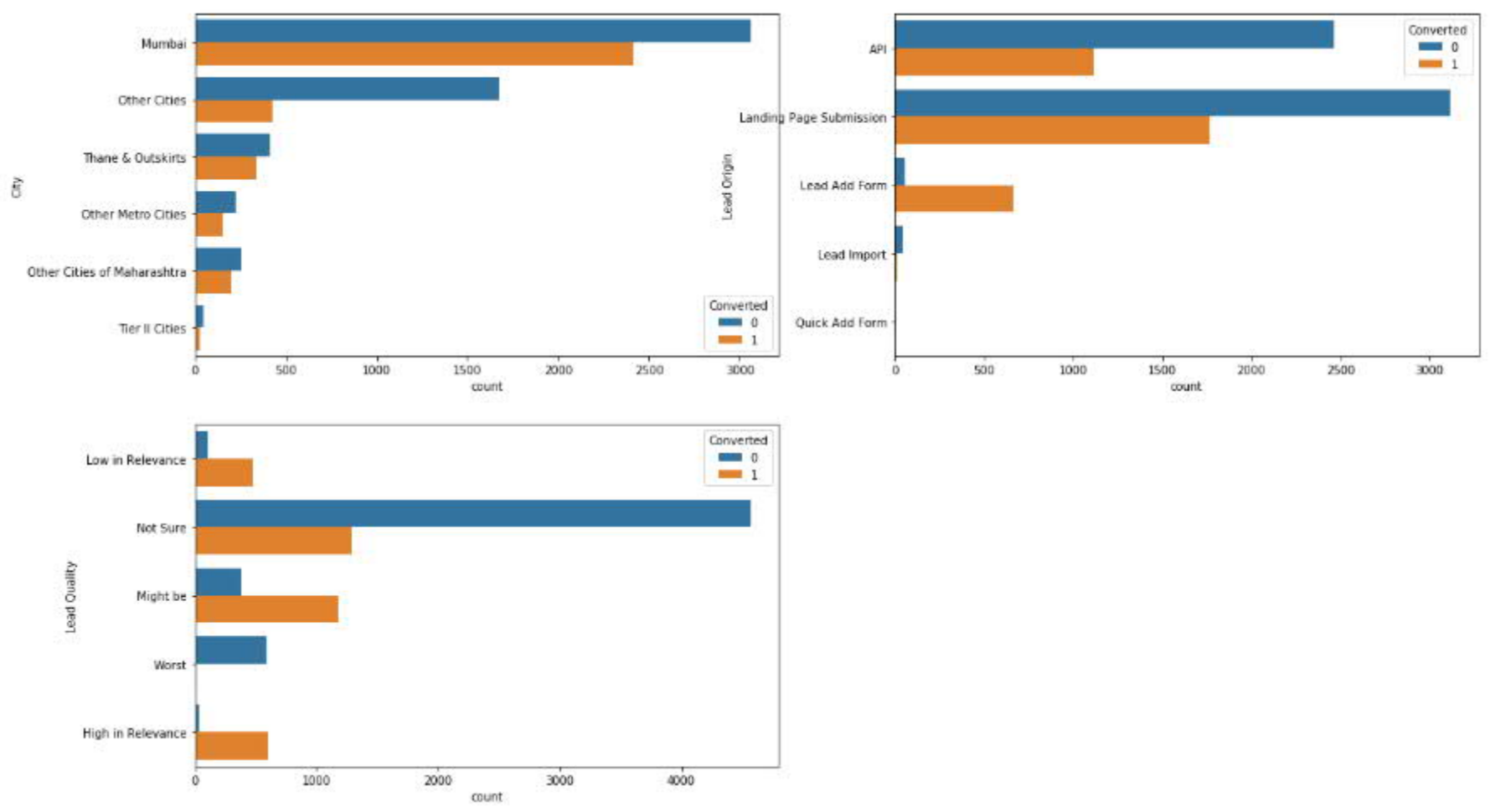
In this also, we can notice for features like 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations ', have highly imbalnced values and should not be taken for modelling.

Univariate Analysis:



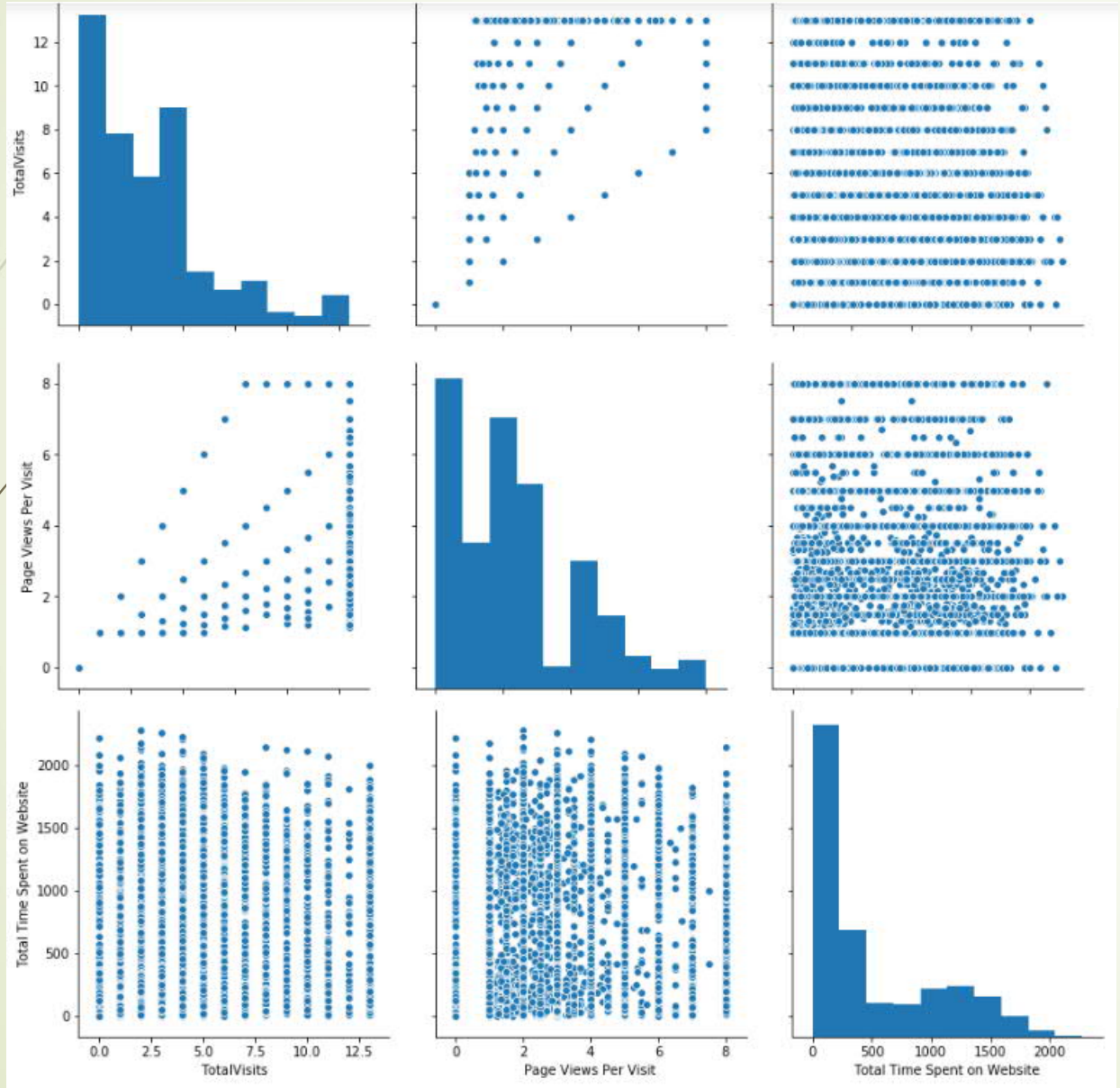
Here, for feature “Total Visit” and “Page Views per Visit” cleaned the features by handling outlying values.

Bivariate Analysis:



Here for visualizing data, in terms of “Converted Leads”, bar graph was plot for “Lead Origin”, “City” and “Lead Quality”. Here we can conclude, that weightage for “landing page submission” in “Lead Origin” was mainly tracked.

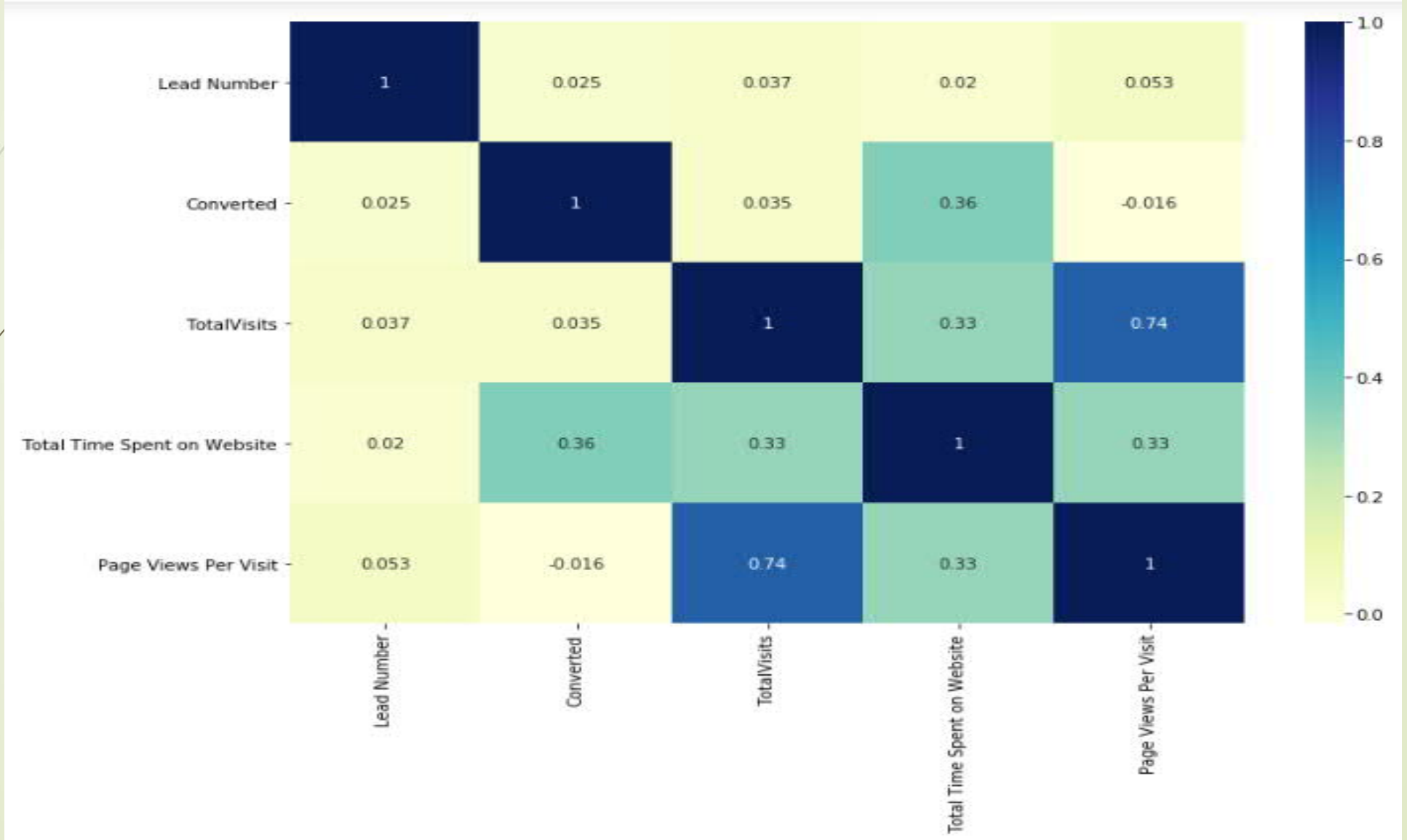
Bivariate Analysis:



Correlation was checked for features: ‘TotalVisits','Page Views Per Visit‘ and 'Total Time Spent on Website’.

Visualizing this plot we can conclude that there was not much correlation among above features.

Visualizing Correlation Through Heat-map:



Data Conversion:

- Numerical Variables are Normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 109

Model Building:

- Splitting the Data into Training and Testing Sets
- The first basic step for Logistic regression is performing a train-test split, we have chosen 70:30 ratio.
- Used RFE for Feature Selection
- Ran RFE with 20 variables as output
- Build Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5
- Predicted on test data set

- Overall accuracy obtained=84%

Initial Conversion Rate: 38.53%

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4662	0.131	11.201	0.000	1.210	1.723
Total Time Spent on Website	1.0760	0.043	24.886	0.000	0.991	1.161
Lead Origin_Lead Add Form	2.7497	0.203	13.517	0.000	2.351	3.148
Lead Source_Olark Chat	1.3109	0.110	11.956	0.000	1.096	1.526
Lead Source_Welingak Website	3.4694	0.749	4.633	0.000	2.002	4.937
Country_Saudi Arabia	-2.0788	0.805	-2.583	0.010	-3.656	-0.502
Specialization_Hospitality Management	-1.0217	0.367	-2.787	0.005	-1.740	-0.303
Lead Quality_Might be	-1.4175	0.151	-9.413	0.000	-1.713	-1.122
Lead Quality_Not Sure	-3.3871	0.135	-25.146	0.000	-3.651	-3.123
Lead Quality_Worst	-5.3268	0.362	-14.700	0.000	-6.037	-4.617
Last Notable Activity_Had a Phone Conversation	2.2212	1.230	1.806	0.071	-0.190	4.632
Last Notable Activity_Modified	-0.6570	0.093	-7.096	0.000	-0.838	-0.476
Last Notable Activity_Olark Chat Conversation	-1.1201	0.348	-3.215	0.001	-1.803	-0.437
Last Notable Activity_SMS Sent	1.3490	0.094	14.415	0.000	1.166	1.532
Last Notable Activity_Unreachable	1.3853	0.609	2.273	0.023	0.191	2.580

- Top 20 features were auto selected by RFE.
- Result summary showed 3 most important variables as:
  - Lead Source
  - Lead Origin
  - Last Notable Activity

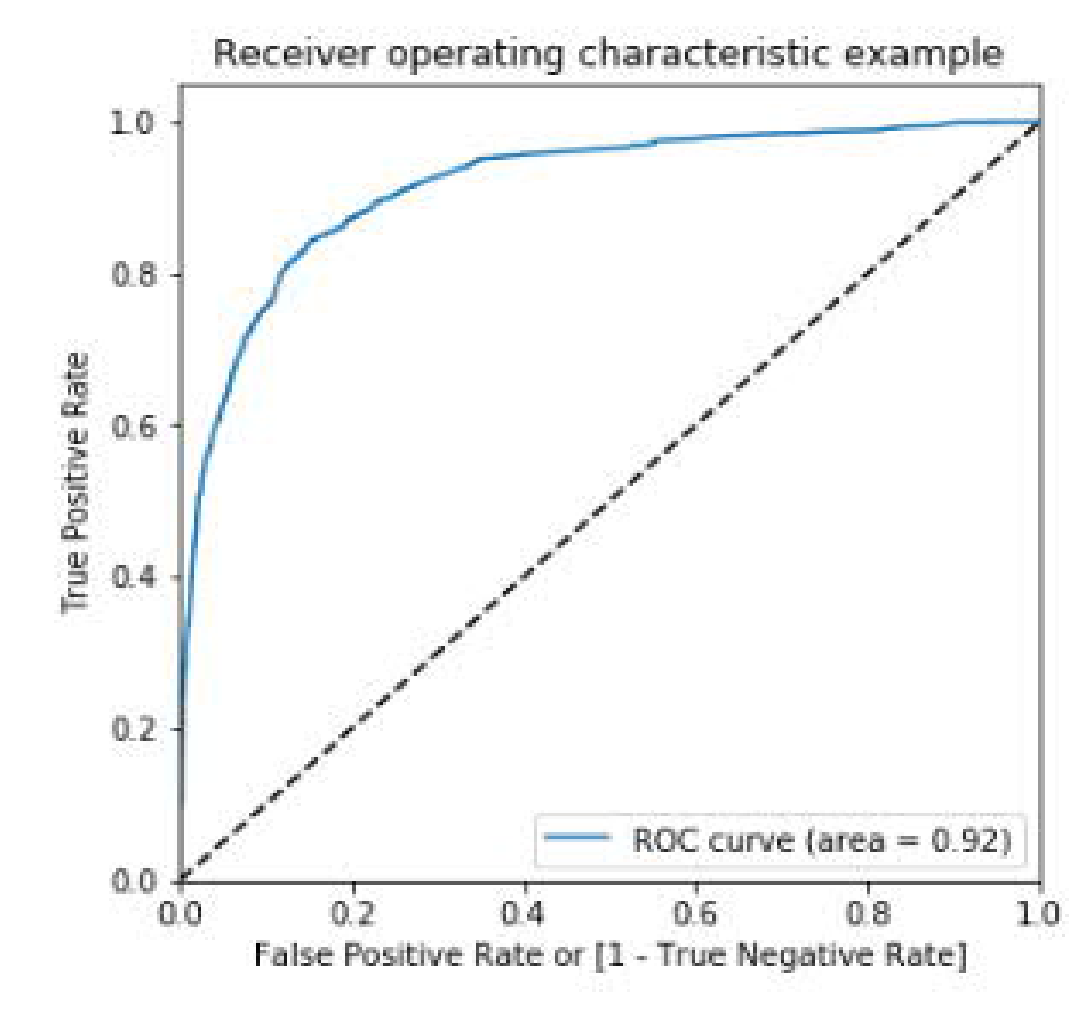
The accuracy of this model (train data)= 84.57%

Approx. No Multi-collinearity:

	Features	VIF
7	Lead Quality_Not Sure	2.17
10	Last Notable Activity_Modified	1.91
12	Last Notable Activity_SMS Sent	1.78
6	Lead Quality_Might be	1.55
2	Lead Source_Olark Chat	1.51
1	Lead Origin_Lead Add Form	1.39
3	Lead Source_Welingak Website	1.25
0	Total Time Spent on Website	1.23
8	Lead Quality_Worst	1.13
11	Last Notable Activity_Olark Chat Conversation	1.08
5	Specialization_Hospitality Management	1.02
4	Country_Saudi Arabia	1.00
9	Last Notable Activity_Had a Phone Conversation	1.00
13	Last Notable Activity_Unreachable	1.00

- This showed that our selected variables have satisfying collinearity.
- And P-value lower than .05 showed these features as significant one.

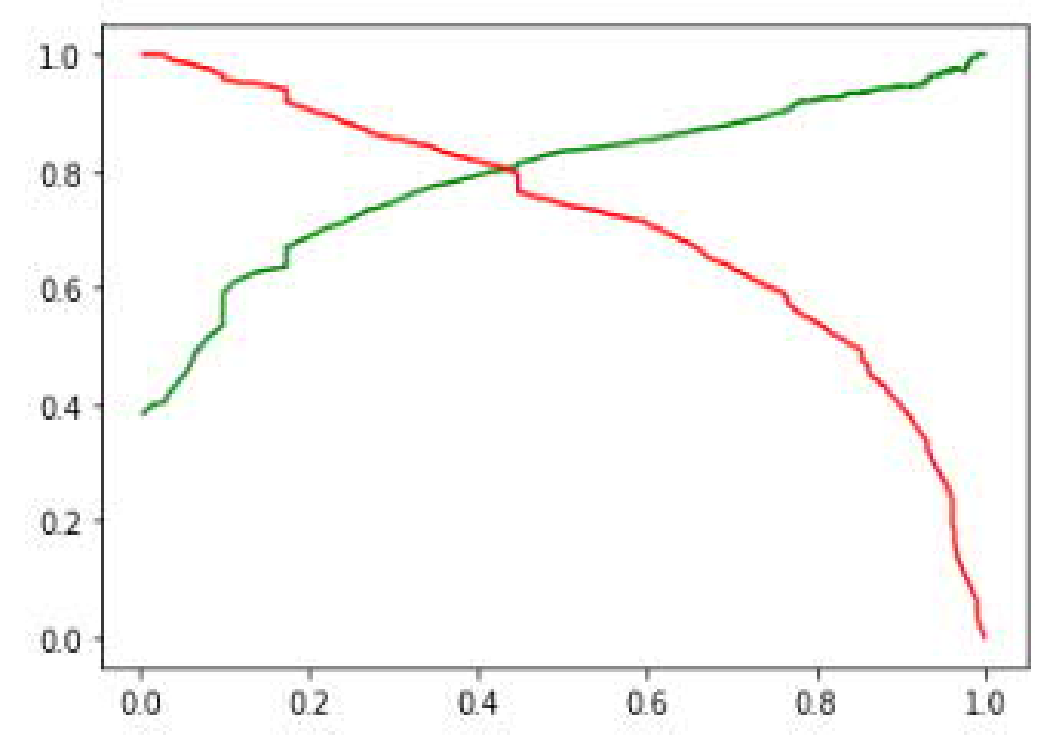
ROC Curve:



From this ROC curve, 0.35 is the optimum point to take it as a cutoff probability.

Metrics beyond accuracy: For Train data

Precision and Recall Tradeoff



Sensitivity: 83.8%
Specificity: 85.0%
Precision: 83.4%
Recall: 74.2%

Metrics obtained on Test data:

```
In [121]: # Let's check the overall accuracy.
          metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)

Out[121]: 0.8466810966810967

In [122]: confusion2 = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.final_predicted )
          confusion2

Out[122]: array([[1462,  215],
                 [ 210,  885]], dtype=int64)

In [123]: TP = confusion2[1,1] # true positive
          TN = confusion2[0,0] # true negatives
          FP = confusion2[0,1] # false positives
          FN = confusion2[1,0] # false negatives

In [124]: # Let's see the sensitivity of our logistic regression model
          TP / float(TP+FN)

Out[124]: 0.8082191780821918

In [125]: # Let us calculate specificity
          TN / float(TN+FP)

Out[125]: 0.8717948717948718
```

ACCURACY: 84.6%	SENSITIVITY:80.8%	SPECIFICITY:87.1%

Conclusion

It was found that the variables that mattered the most in the potential buyers are:

1. Lead Source_Welingak Website
2. Lead Origin_Lead Add Form
3. Last Notable Activity_SMS Sent
4. Lead Source_Olark Chat
5. Total Time Spent on Website

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

Thank You..