Lead scoring Case Study

Shreegowri Shetty

Problem Statement

1. X education sells online course to industry professionals.
2. X education gets a lot of leads, its lead conversion rate is very poor. For example if say the acquire 100 leads in a day only about 30 of them are converted.
3. To make this process more efficient, the company wishes to identify the most potential leads also known as “hot leads”.
4. 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 from promising leads.
- For that they want to build a model which identifies the hot leads.
- Deployment of the model for the future use.

Data manipulation

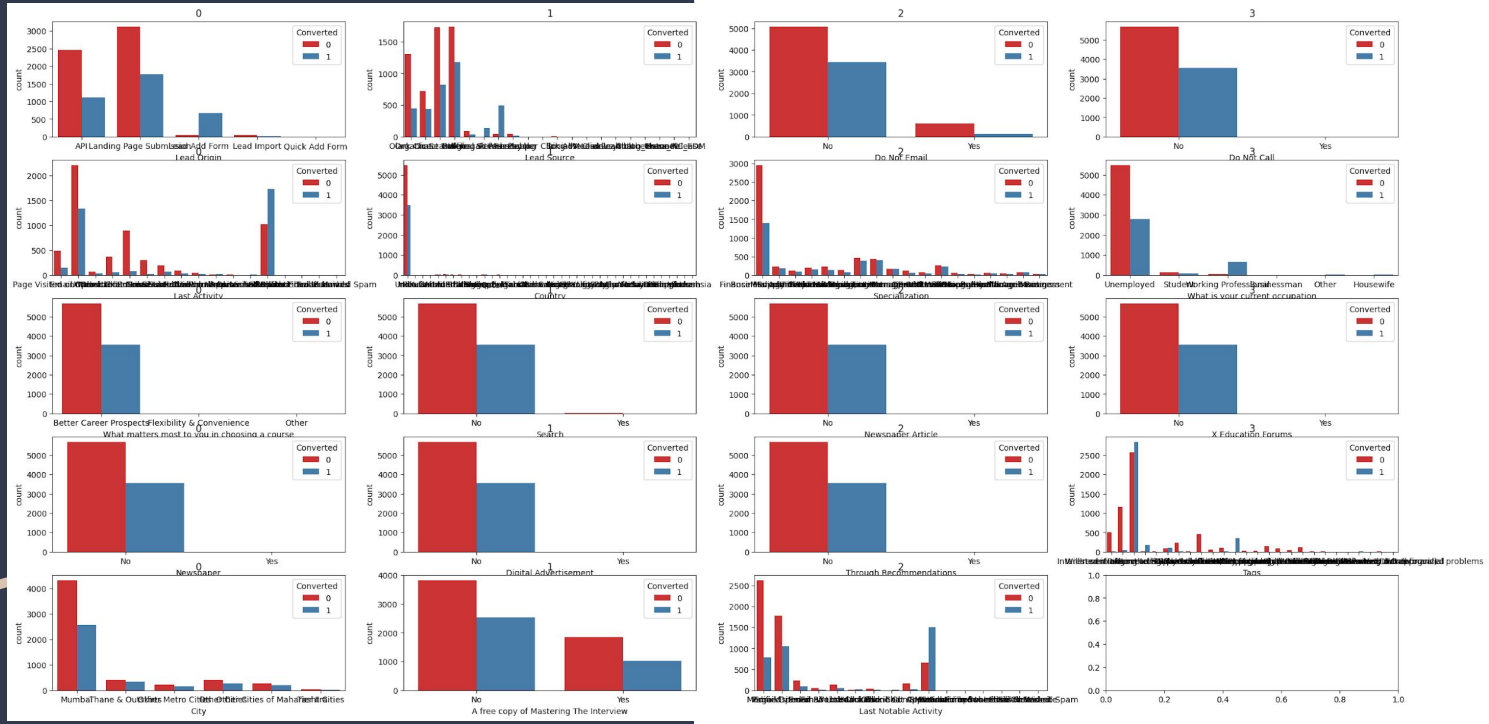
- Total number of rows = 9214, columns = 37.
- Few single value features will dropped as they were just increasing the complexity of the model
- Prospect ID and lead numbers where dropped as they content only unique values which would not help in the model
- Few additional columns like “do not call”,“search”, “what matters most to you in choosing course” etc where dropped as they were not giving any additional information.
- All the columns having more than 35% null values where dropped.

Solution Methodology

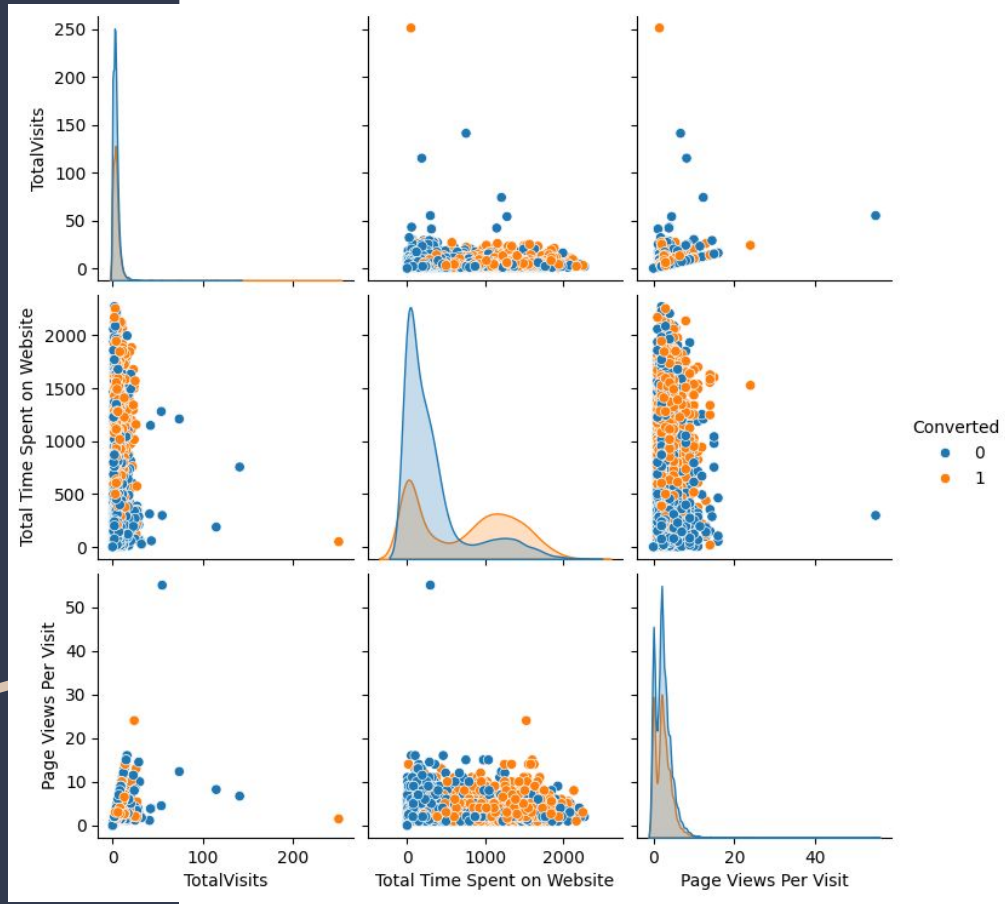
- Data cleaning and data manipulation.
  - Checking and handling duplicated data.
  - Checking and handling null values and missing values.
  - Dropping the columns which contains large amount of missing values and also the columns which are not important for the prediction.
  - Checking for the outliers and handling them.
- EDA
  - Univarient data analysis like value count, distribution variable etc was done.
  - By-variant data analysis like correlation coefficient, pattern between the variables was looked into
- Features scaling and creating the dummy variables was performed
- Logistic regression was used as a classification technique to build a model for prediction.
- Validation of the model was done.
- Conclusions and few recommendations for suggested.

EDA

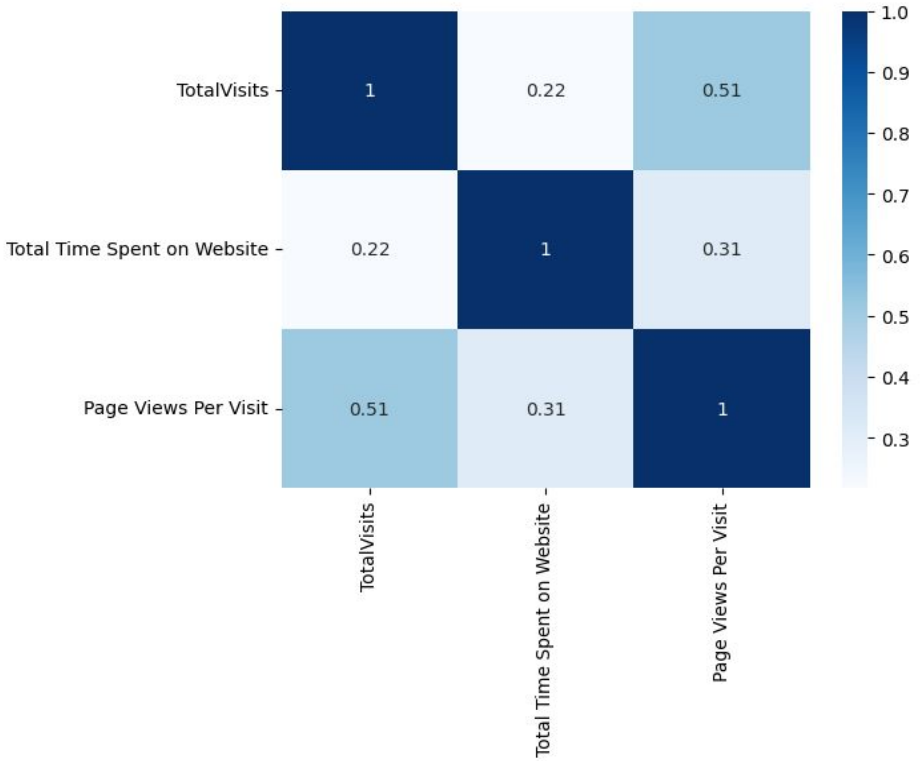
Categorical Features



Numerical Features



Heat Map


TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.22
0.51
0.22
1
0.31
0.51
0.31
1
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3

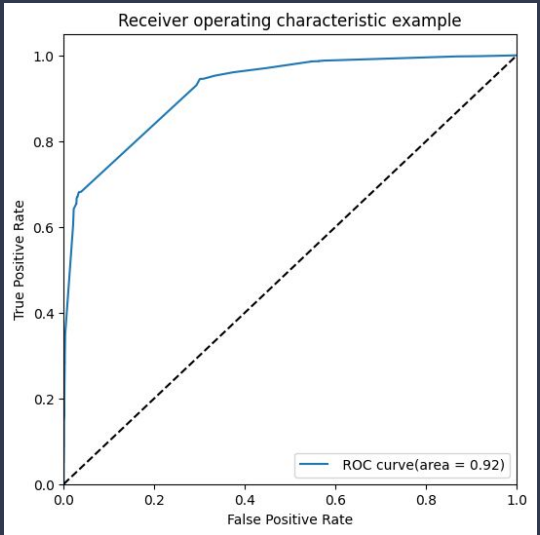
Data Conversion

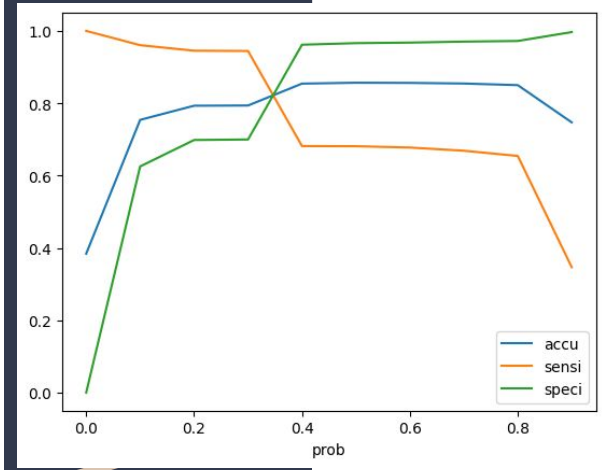
- Numeric values were normali
- Created dummy variables for the object type variables.

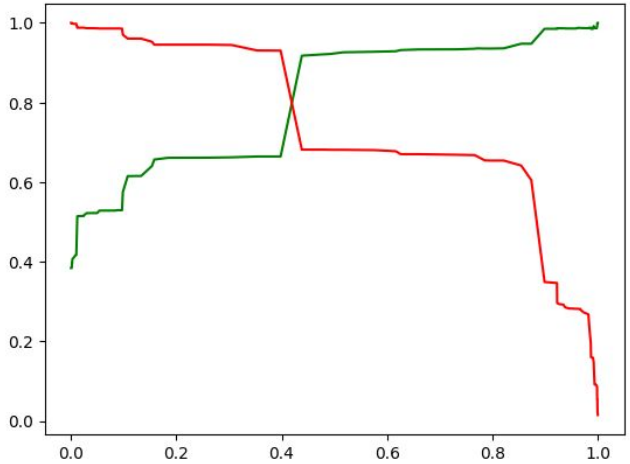
Model building

- For building the model data is split into training set and testing set in the ratio 70:30
- RFE was used for feature selection selected top 15 variables which help in production.
- Manually removing the variables whose p-value is greater than 0.05 and VIF > 5
- Once the model is built the model is tested on test data to get the prediction.
- Test data accuracy is 79.18% and for the train data 79.28%

ROC







Finding optimal cutoff point

Using the above graph the optimal cut of point was found to be 0.25 altho 0.38 also gives similar accuracy but for our model we choose 0.35 as the cutoff value

Conclusion

Following are the variables that matter the most.

- Lead origin
  - Lead add form
- Do not email
  - Yes
- Last activity
  - Olark Chat
  - SMS sent
  - Converted to lead
- Tags
  - In touch with EINS
  - Will revert after reading the email
  - Last to EINS
  - Busy
  - Closed by Horizon
- Current occupation
  - Professional
  - Unemployed