Lead Scoring Case Study

-Manikumar Raju

-Raghavi Premkumar

Objective

- To help X Education to select the potential leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.

- To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads

Methodology

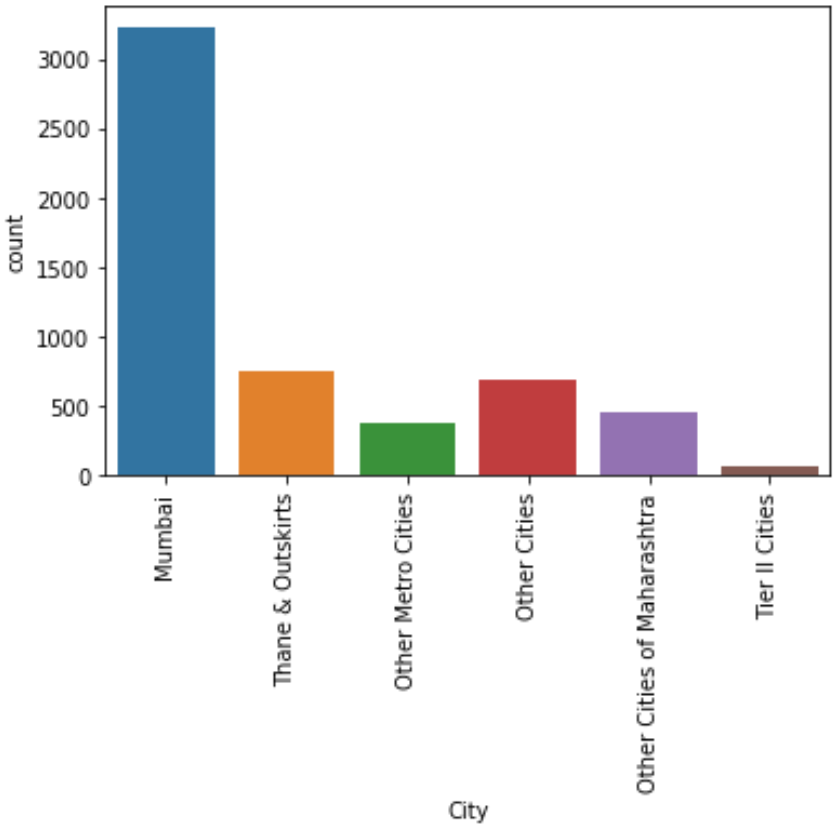
- Data Cleaning
- EDA
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification – Logistic Regression
- ROC
- Conclusions and recommendations.

Data Cleaning

- Initially duplicate values are handled
- Then NA values and their percentages are taken into consideration
- There are several values called ‘Select’ which are formed due to customers not filling the required fields.
- There are considered as NA and was allotted to relevant categories in each column
- Columns with higher percentage of missing values are dropped
- Outliers are handled.

Data Cleaning

- For Eg: City is considered and its Null values are reduced

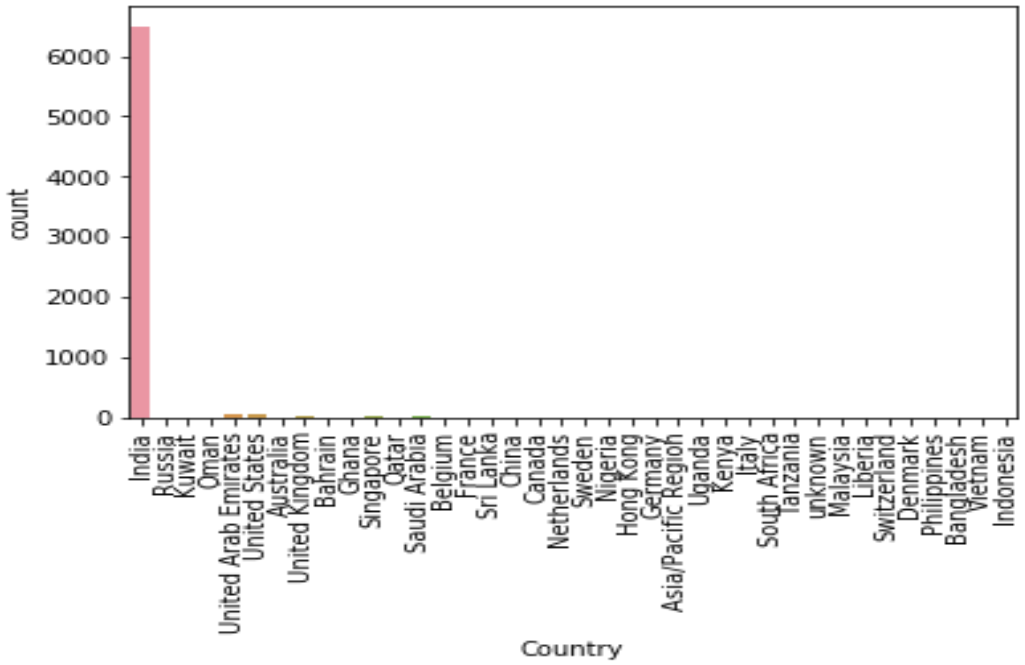


In this case since Mumbai has maximum count, null values are assumed as ‘Mumbai’ and percentage of NaN is reduced.

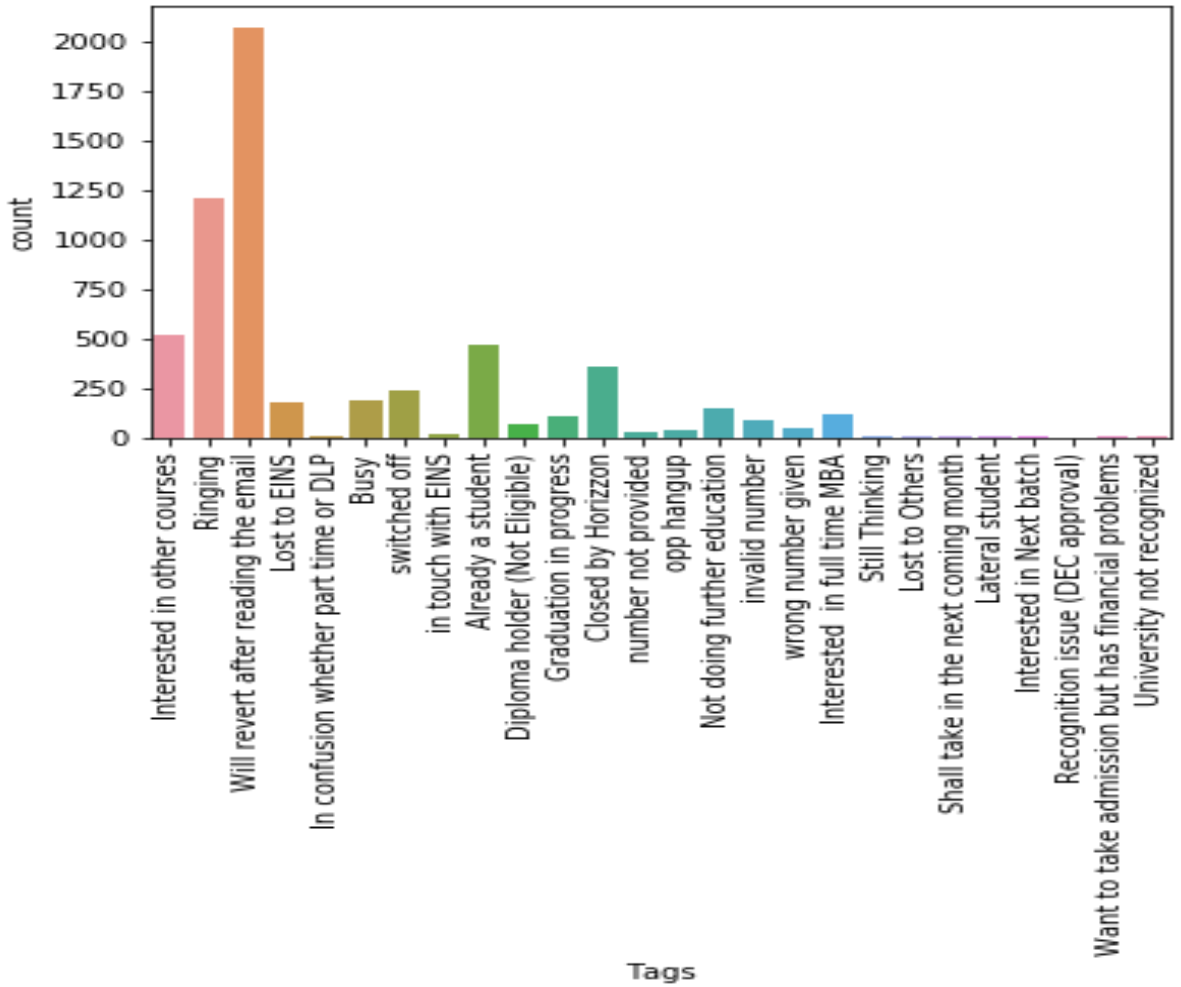
Exploratory Data Analysis

- Univariate Analysis are done.

If you consider country, most of the visitors irrespective of them converted or not belongs to INDIA. Since it doesn't provide any insights we can drop them



- Where as Tags is a significant feature to betaken into consideration.



Feature Scaling ,Dummy variables, Encoding

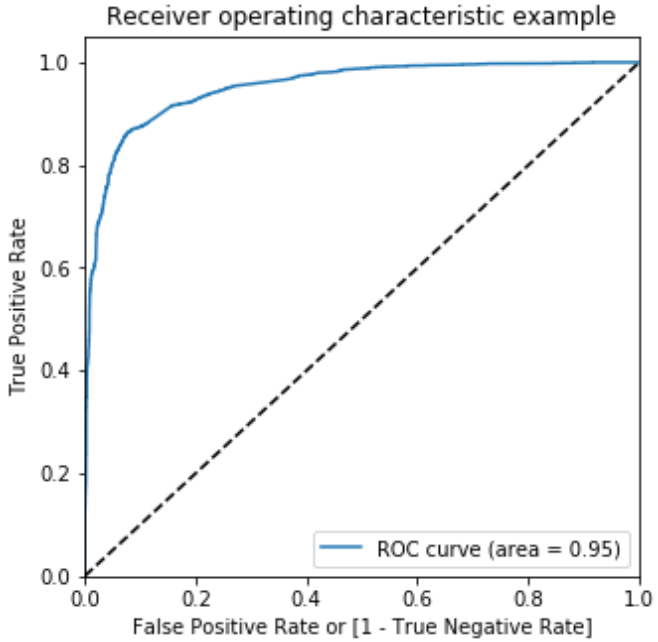
- Numerical Variables are Normalised
- Dummy Variables are created for categorical variables
- Converted binary variables (Yes/No) to 0/1.

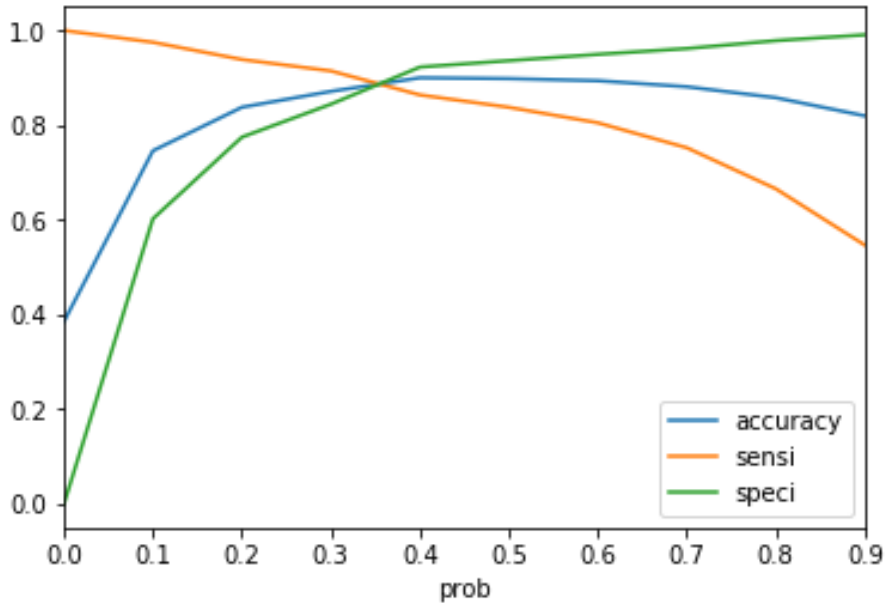
  Eg: 'Search','Do Not Email', 'Do Not Call',
- Scaling helps in interpretation.
- ‘Standardisation’ was used to scale the data for modelling. It basically brings all of the data into a standard normal distribution with mean at zero and standard deviation one.

Logistic Regression Model

- Data split into training and testing Sets
- The first basic step for regression is performing a train-test split, we have with 70:30 ratio.
- RFE is used for feature selection.
- RFE is run for 15variables as output.
- Model is built by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5
- Overall Accuracy = 0.89

ROC Curve





- ROC Curve is obtained
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion

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

- The total time spend on the Website.
- Total number of visits.
- When the lead source was Google, Direct traffic, Organic search
- When the last activity was SMS, Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.

X education has to focus on the above factors to convert the website visitors into potential students of the course.