Lead Scoring Case Study

Riya Tyagi

Anmol Agarwal

Problem Statement:-

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Data-set:-

We have a leads datasets from the past around 9000 datapoints. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted. Our dataset file name is “Leads.csv”.

Data Cleaning:-

- After Exploring the data “Leads.csv”, we explore the detail about the columns and identify the some irrelevant columns which has no use in our case study. So, we dropped these type of columns from our dataset.

- After Exploring we get some rows which needs to replace with specific value, mean or null values like in this dataset we have categorical variables have a level called 'Select' which needs to be handled

- Now , we calculate the percentage of null values in each columns and drop those columns which has more than 70% null values.

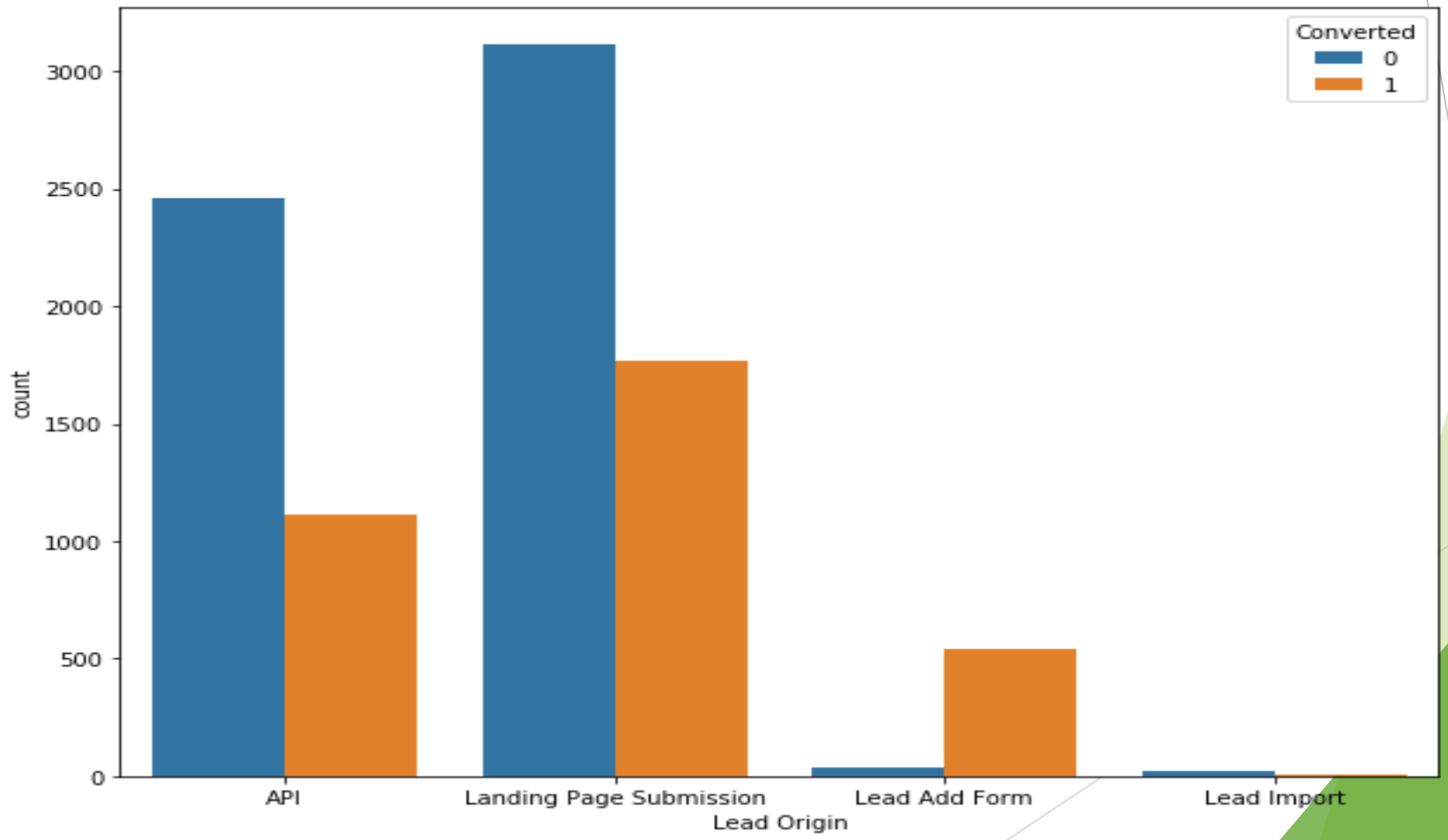
- As a resultant, now our dataset has 30 columns.

Data Analysis:-

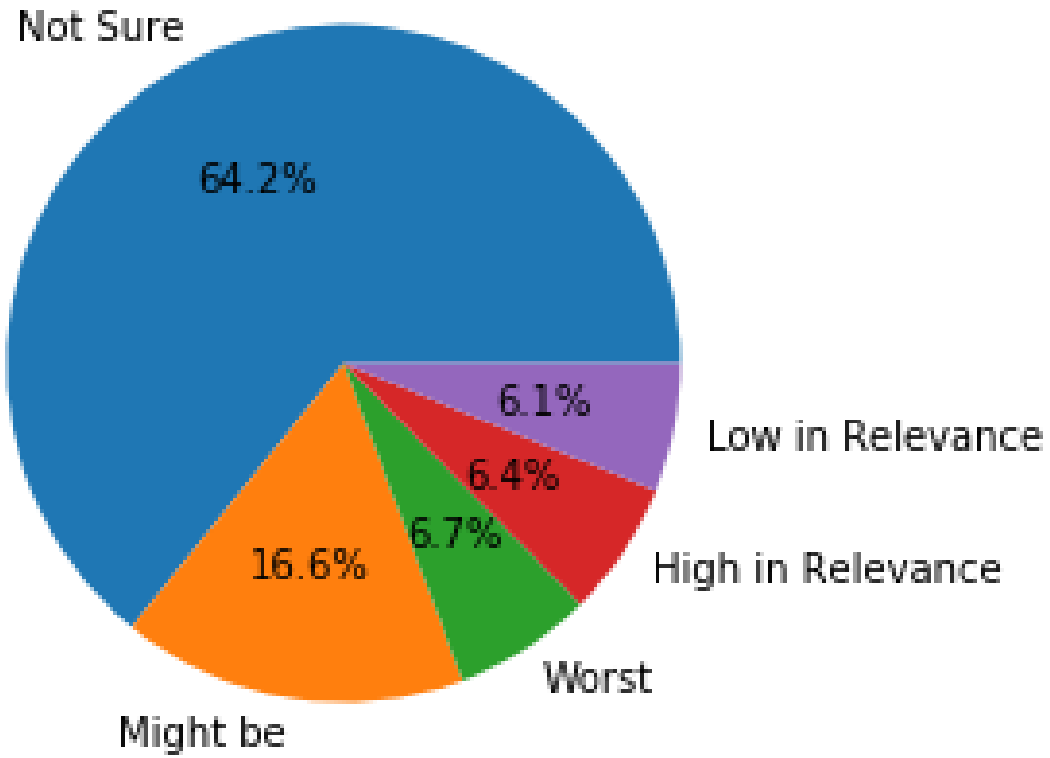
In this case study we use different approaches to handle and analysis the specific data. First we identify irrelevant columns which are not five any useful information in our data-set and dropped them by using univariate analysis and figure out some null values in our data-set and replace them with some value. After this we did some bi-variates Analysis on the given data set at the ends comes to the Modeling part where we Predicted conversion on the basis of target variable .Logistic Regression is the main part of this case study with the help of this We assign a lead score to each of the leads such that the customers With higher lead score have a higher conversion chance and the Customers with lower lead score have a lower conversion rate this Is the main aim of this case study that we did focus.

Univariate Analysis:-

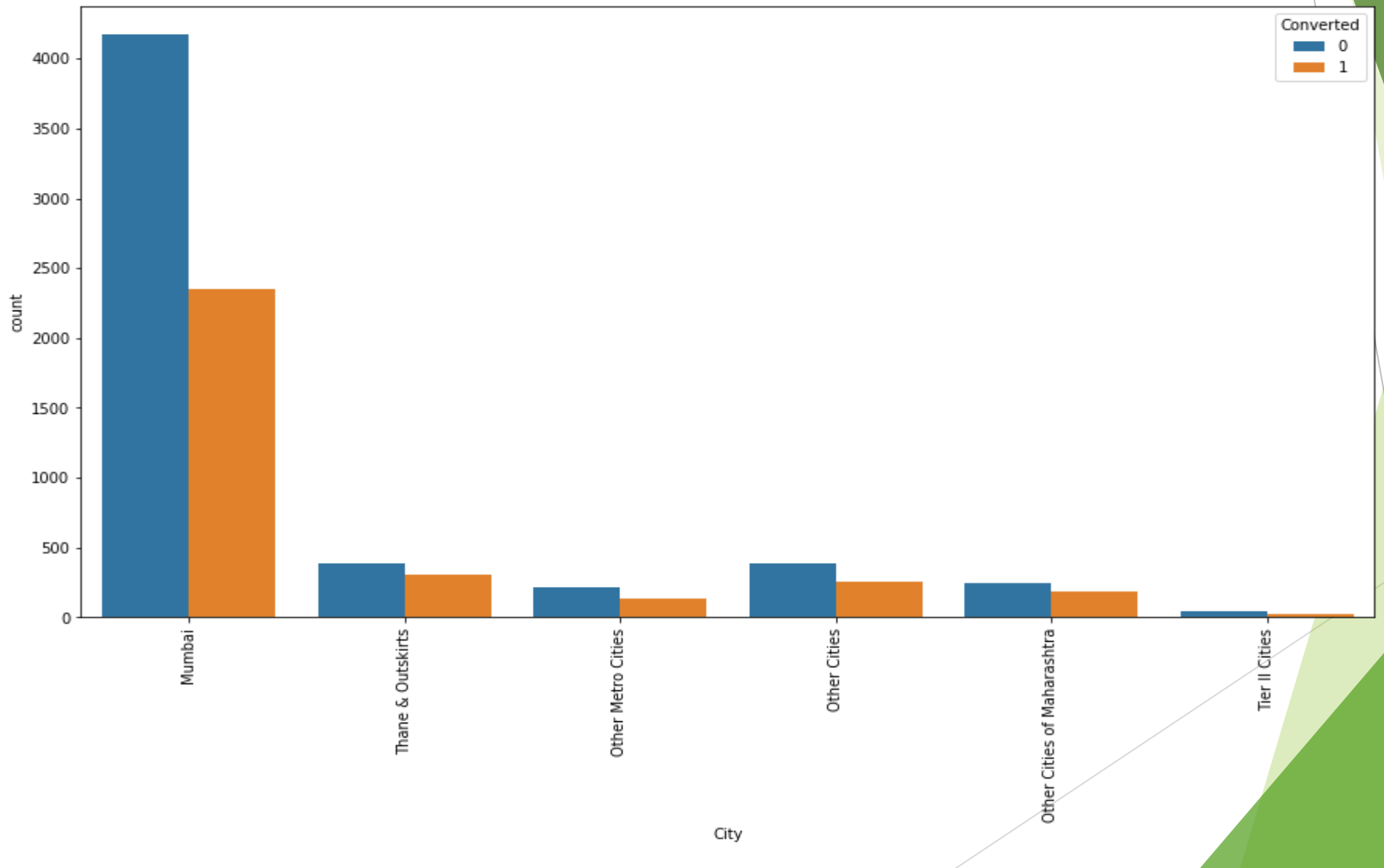
Lead Origin:- “Landing Page Submission shows the highest conversion rate.



Lead Quality:- we have almost 65% values with not sure.

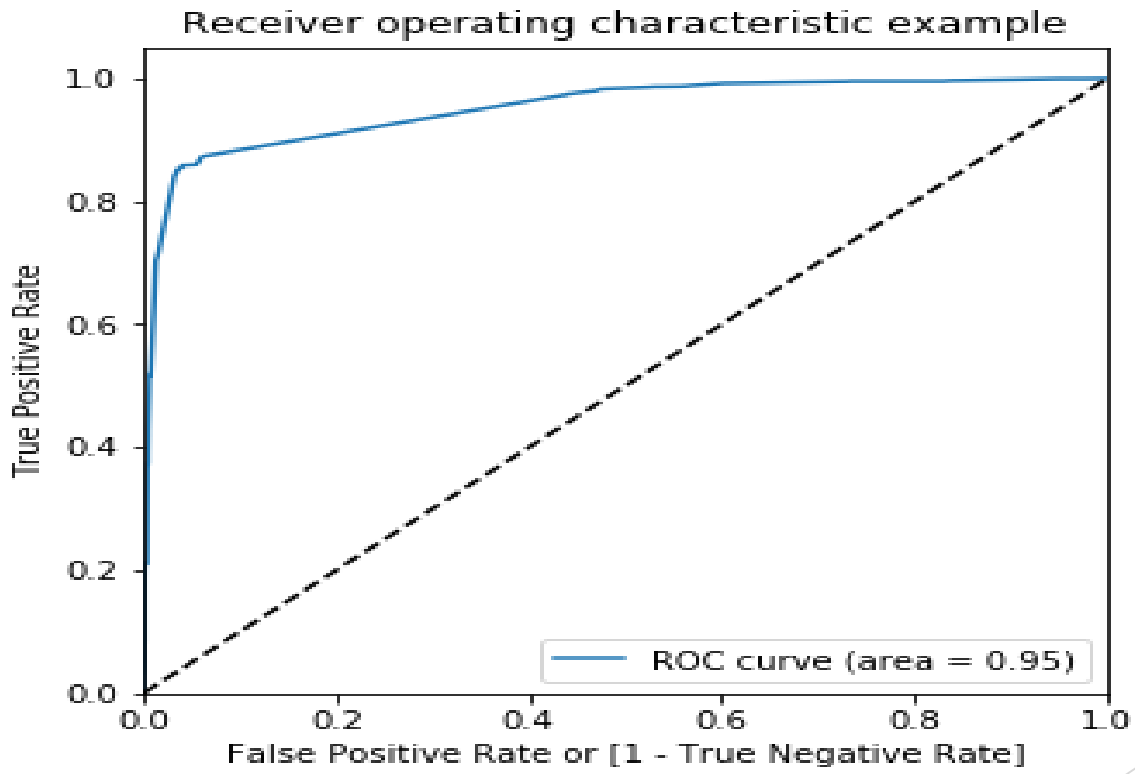


City:- Maximum conversion rate is from “Mumbai City”.

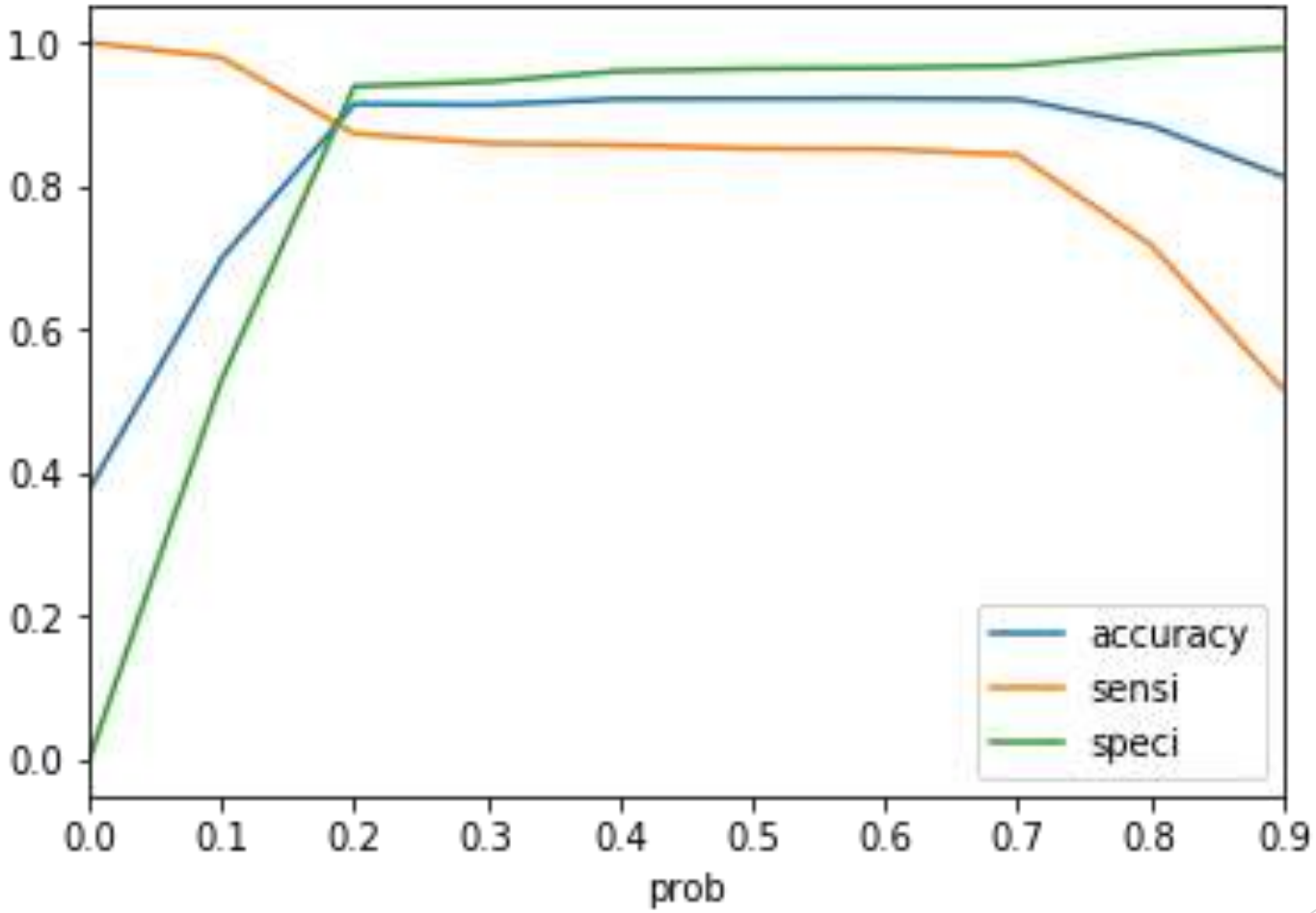


Model Building:-

We build a logistic regression model with our resultant columns after the analysis and with the help of RFE feature selection method we select the relevant features from our data-set. As resultant we got 15 columns. After this we dropped those columns which have high p value(>0.05) like “Tags_wrong number given”, “Tags_invalid number” etc. and got this ROC curve.



Plot between accuracy sensitivity and specificity for various probabilities:-



Assigning lead_score to the leads between 0-100

	Converted	Converted_Prob	Prospect ID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	Lead_Score
0	1	0.869010	7220	1	1	1	1	1	1	1	1	1	1	0	1	87
1	0	0.771785	3383	1	1	1	1	1	1	1	1	1	0	0	1	77
2	0	0.771785	770	1	1	1	1	1	1	1	1	1	0	0	1	77
3	1	0.998831	5785	1	1	1	1	1	1	1	1	1	1	1	1	100
4	1	0.832857	5252	1	1	1	1	1	1	1	1	1	1	0	1	83

Conclusion:-

Our model has the 90% accuracy of prediction which means we have the conversion rate as 90% of our model. We also assign the “lead_score” to each lead to identify the best lead in all data points, From our model we got this conversion equation-

Conversion = Do Not Email * (-1.3150) +
Lead Origin_Lead Add Form * 0.9184 +
Lead Source_Welingak Website * 4.0304 +
What is your current occupation_Working Professional * 1.1813 +
Tags_Busy * 4.2337 +
Tags_Closed by Horizzon * 8.5736 +
Tags_Lost to EINS * 8.6738 +
Tags_Ringing * (-1.5122) +
Tags_Will revert after reading the email * 3.8461 +
Tags_switched off * (-2.8432) +
Lead Quality_Not Sure * (-3.4320) +
Lead Quality_Worst * (-3.9863) +
Last Notable Activity_SMS Sent * 2.7582 +
Constant * (-1.9539)