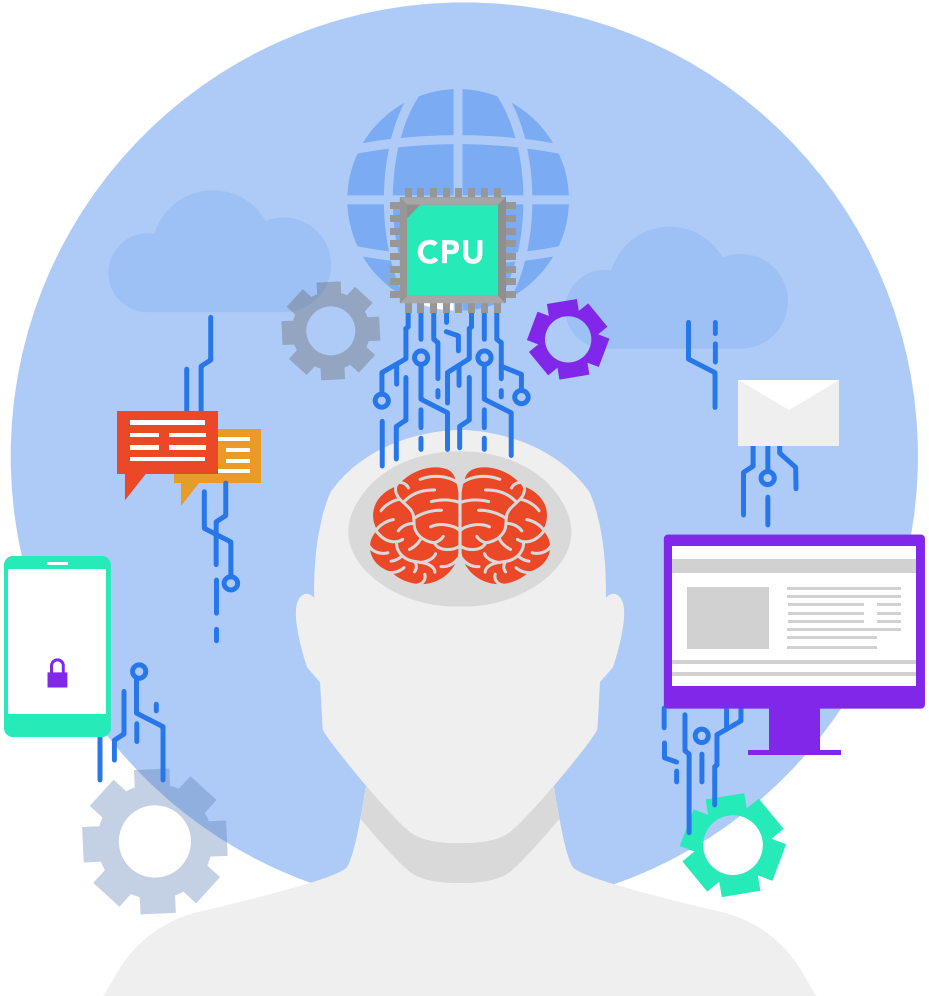


Lead Scoring Case study

By Meghana, Hemant & Chandana

Problem Statement & Goal

Problem Statement

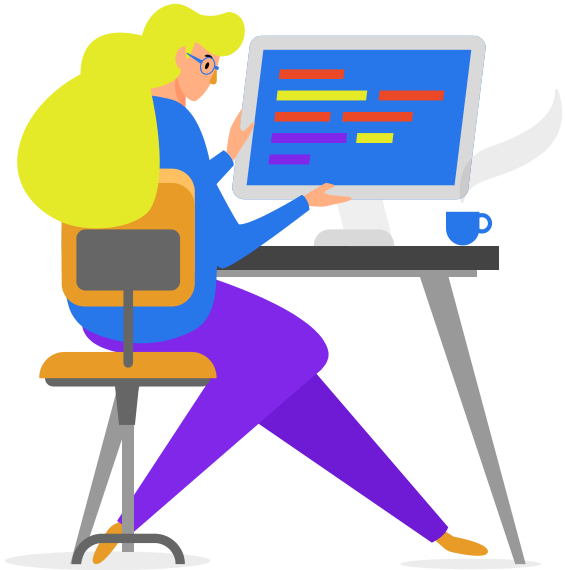
The X Education wanted to identify the potential leads which can convert in to payment and focus on them. They wanted us to build a model wherein each lead need to be assigned a score such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business Goal

Building a logistic regression model to assign a lead score between 0 and 100 to each of the leads. A lead with higher score would mean Hot Lead is most likely to convert and lower score lead would be the cold and will mostly not get converted

There are some more problems presented by the company with our model they should be able to adjust to if the company's requirement changes in the future

Approach



01 Data Cleaning & Imputing missing Values

02 EDA : Univariate , Bivariate & Multivariate Analysis

03 Feature Scaling & Creating Dummy Variables

04 Model Building : Logistic Regression

05 Model Evaluation : Accuracy , Specificity , Sensitivity, Precision & ReCall

06 Conclusion & Recommendation

Process flow

Data Cleaning & Preparation

1. Cleaning the data suitable for analysis
2. Remove duplicate data
3. Outlier treatment
4. Exploratory data analysis

01

Data Dividing & Feature scaling

1. Diving the data into train & test dataset
2. Feature scaling of numerical variables

02

Model Building

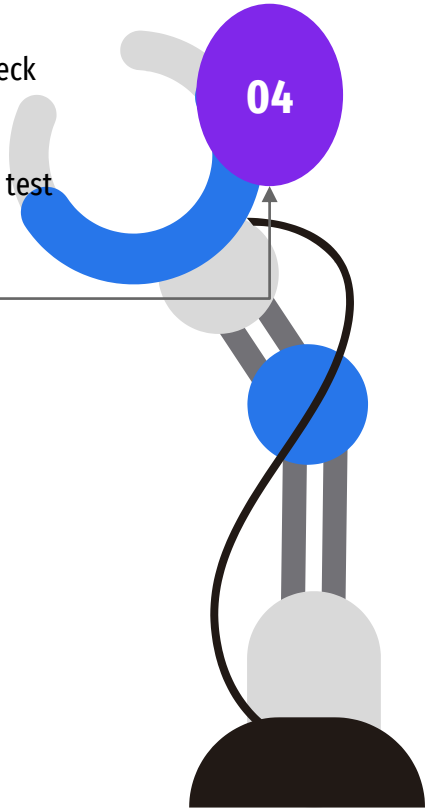
1. Feature selection using RFE, VIF and p-value
2. Determine optimal model using Logistic Regression
3. Calculate various evaluation metrics

03

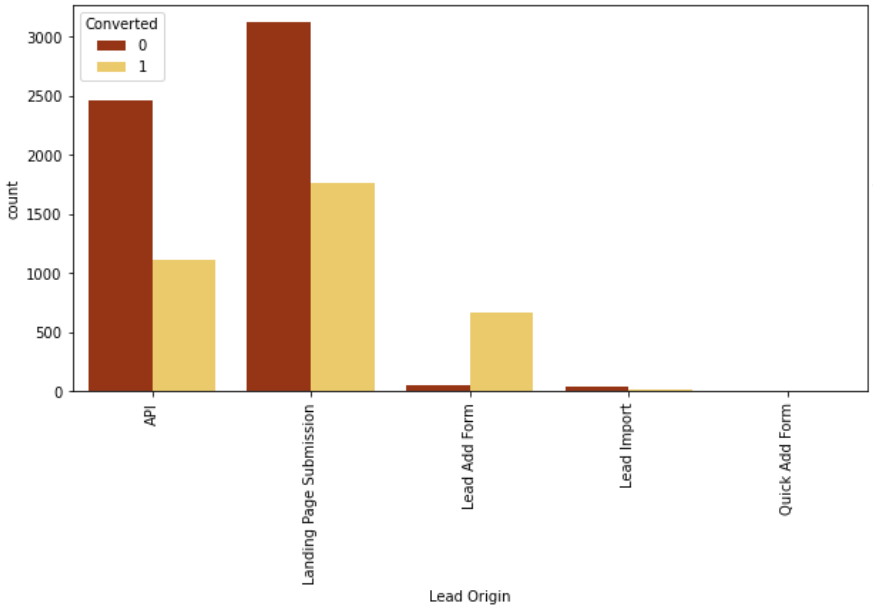
Result

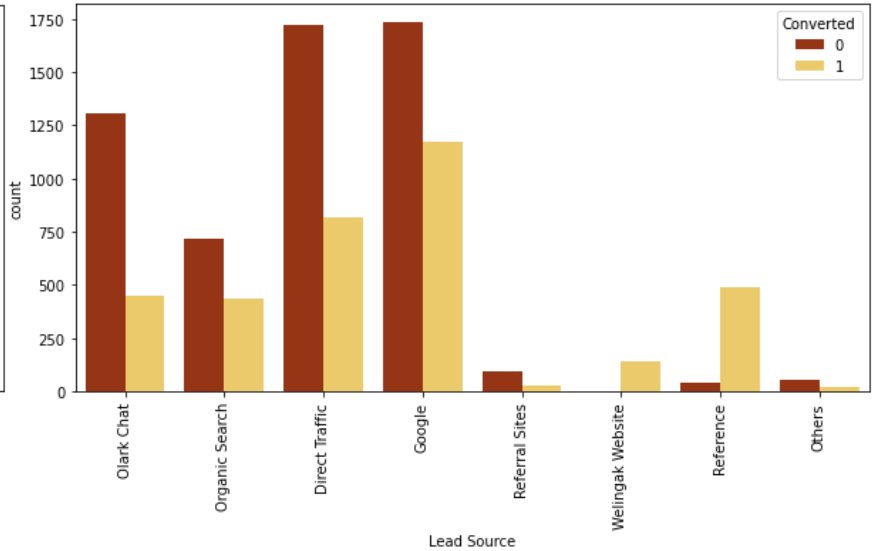
1. Determine Lead score & check if target final prediction is greater than 80%
2.Evaluate final prediction on test set

04



Exploratory Data Analysis

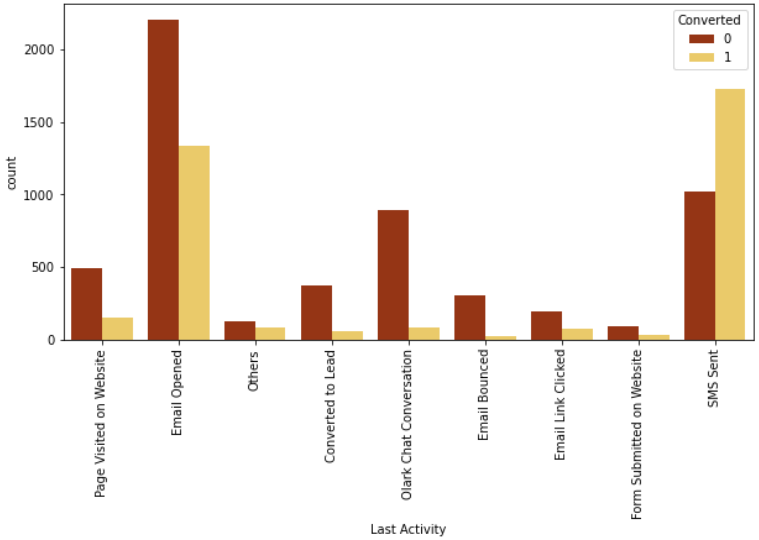


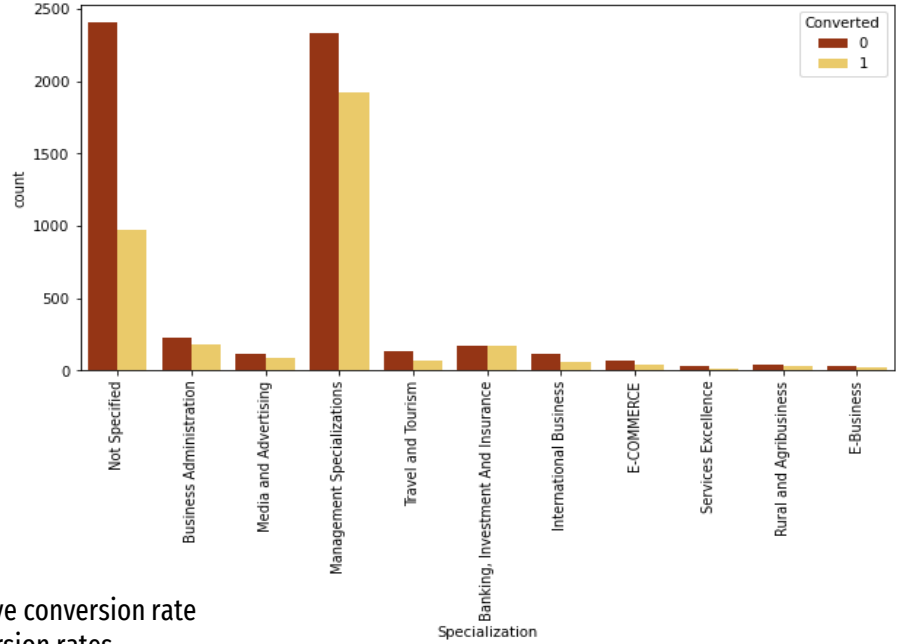


Inferences drawn from "Lead Origin" & "Lead Source":

1. Customers identified from Lead Add Form have the highest conversion rate as compared to other Lead Origins.
2. As per the count plot leads from **Direct Traffic** & **Google** have highest negative conversion rates.
3. **Reference** & **Welingak website** show positive conversion rate.
4. Interesting fact is that the Leads generated through **Welingak website** do not have any negative conversion.

Exploratory Data Analysis



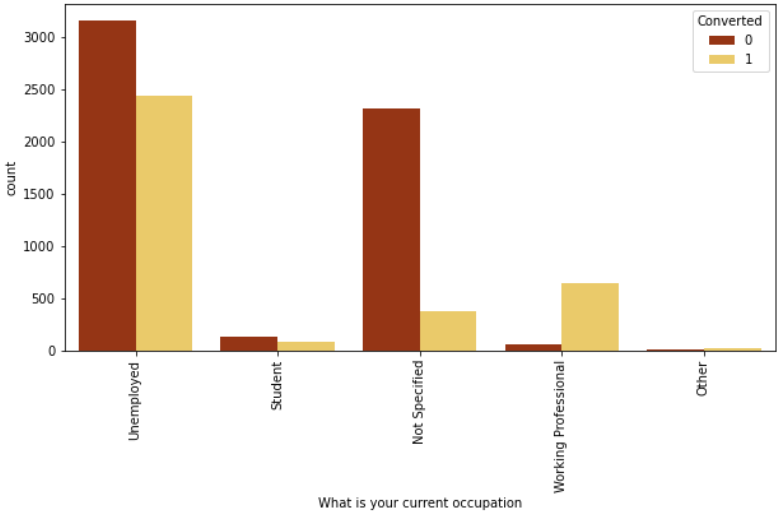


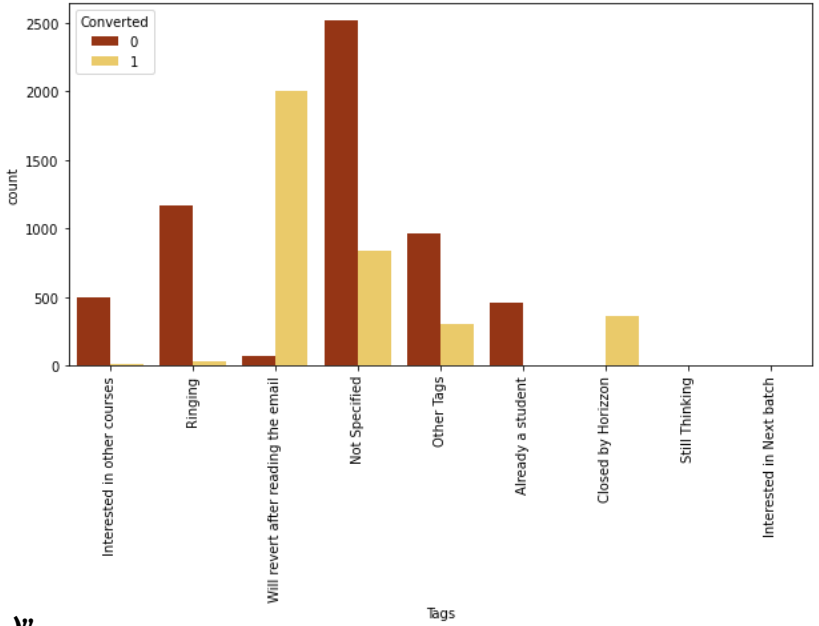
Inferences Drawn from "Last Activity" & "Specialization":

- Last activity performed by customers is SMS sent & it has the highest positive conversion rate
- Customers with Management specialization have the highest positive conversion rates.
- Customers who specialize in Services Excellence have the lowest positive conversion rate
- Customers whose specialize is not specified also have lower positive rate.

*** The activity SMS Sent has the highest positive conversion rate out of all the activities**

Exploratory Data Analysis



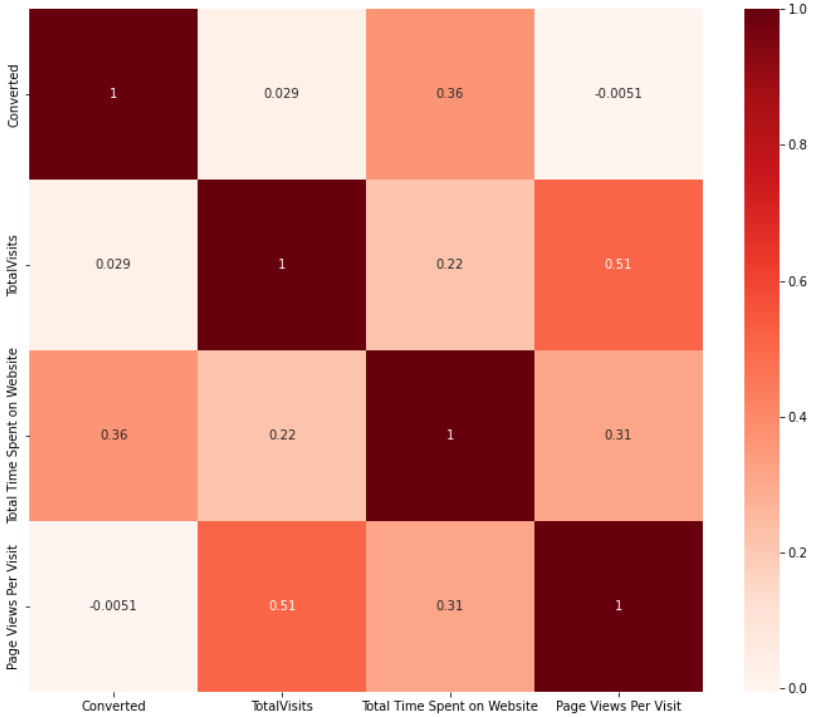


Inferences Drawn from "Occupation" & "Tag(current status)":

- The working professionals have high positive conversion rates & It is clearly evident from the plot above.
- The Unemployed customers have the highest negative conversion rates
- The status tags 'Will revert after reading the email' & 'Closed by Horizzon' have the highest positive conversion rate.
- Also, the tag 'Ringing', 'Not Specified', Interested in other courses" and 'Other Tags' have high negative conversion rates

. No of unemployed leads are more than other catagories

ANALYSIS OF NUMERICAL VARIABLES

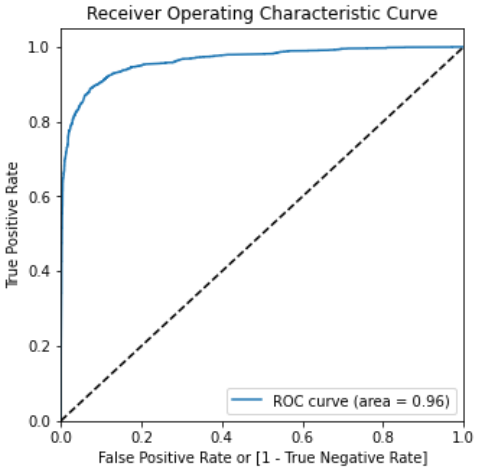


INFERENCE:

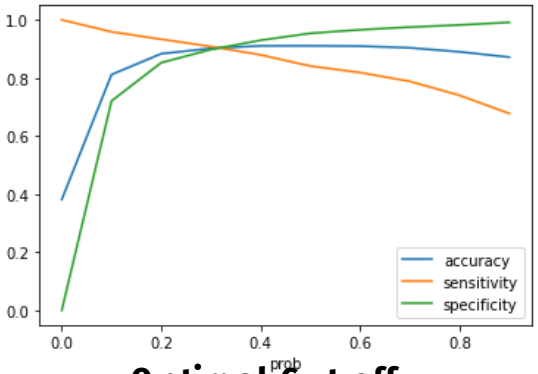
- As we can see from the heatmap, Total Visits and Page Views Per Visit have the highest correlation.

- The target variable i.e, Converted and Page Views Per Visit have negative correlation.

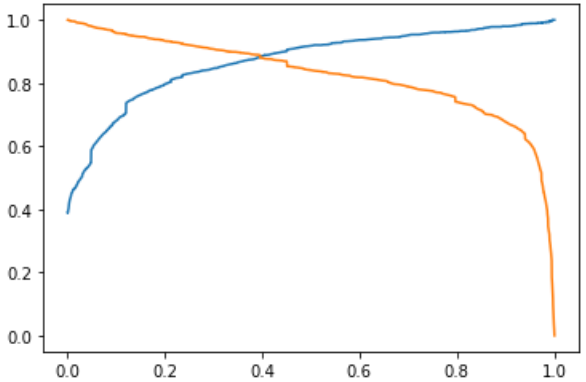
Model Building



ROC Cure



Optimal Cut off point achieved 0.3

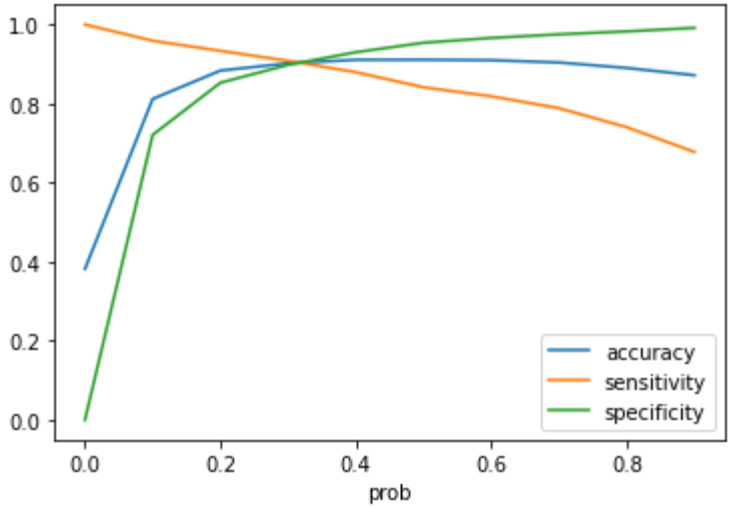


Precision and Recall Trade of

- We have chosen the train test Split Ratio As 70:30
- Using RFE to choose top 15 Variables
- Building Model by dropping the variables with P-Value > 0.05 And VIF> 5
- Predictions On Test Dataset
- Overall Accuracy Is 90.0 % & subject to change
- ROC cure has a value of 0.96 is a good one

Accuracy, Sensitivity & Specificity

	prob	accuracy	sensitivity	specificity
0.0	0.0	0.381581	1.000000	0.000000
0.1	0.1	0.811567	0.958814	0.720712
0.2	0.2	0.883388	0.933278	0.852605
0.3	0.3	0.901776	0.908155	0.897840
0.4	0.4	0.910577	0.879325	0.929860
0.5	0.5	0.910734	0.841021	0.953748
0.6	0.6	0.909634	0.818369	0.965947
0.7	0.7	0.903819	0.788303	0.975095
0.8	0.8	0.889989	0.740115	0.982465
0.9	0.9	0.871601	0.677512	0.991360



- From the curve, we can see that 0.3 is our optimal cutoff point

Model Evaluation

Training Data	
Accuracy	0.901
Sensitivity	0.908
Specificity	0.897

Test Data	
Accuracy	0.909
Sensitivity	0.913
Specificity	0.906

LEAD SCOREING FOR TESTING DATA

	Prospect ID	Converted	Converted_Prob	Final_Predicted	Lead_Score
0	6906	1	0.998022	1	100
9	7008	1	0.968745	1	97
11	3074	1	0.977028	1	98
13	6163	1	0.972033	1	97
16	6482	1	0.845816	1	85
...	...	...	...	...	...
2716	309	1	0.893479	1	89
2717	9234	1	0.931237	1	93
2718	8028	1	0.939947	1	94
2719	5807	1	0.994227	1	99
2723	1540	1	0.923401	1	92

There are 787 leads which can be contacted and have a high chance of getting converted whose lead score is more than 85.

Conclusion & Recommendation

1. Logistic regression model is used to predict the probability of conversion of a customer.
2. Lead Score & conversion rate of final predicted model is over 90% in test data as well as Training Data
3. Overall this model is compatible to adjust with the company's future requirements as well
4. Top 3 Variables that contributes for leads getting converted in the model are:

1.Tags_Will revert after reading the email

2. Tags_Closed by Horizzon

3. Last_Activity_SMS Sent