X-Education – Generate Lead Score using Logistic Regression

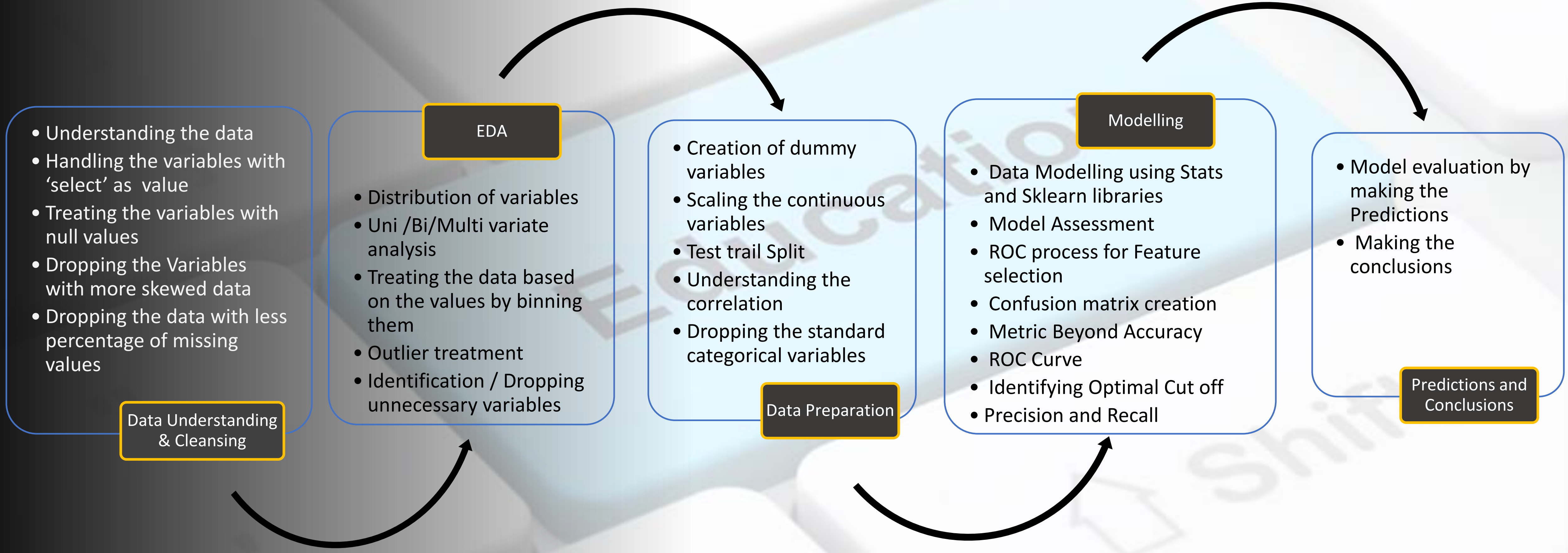
Srinivas Avala

Problem Statement

X-Education would like to assign a lead score between 0 and 100 to each of the leads which can be used by to target potential leads, using **logistic regression** model. The higher the score, the higher the potential of the lead to get converted.

Additionally, model should be able to adjust in case X-Education's requirement changes in the future.

Analysis Approach



Data Understanding & Cleansing

- Understanding the data
- Handling the variables with 'select' as value
- Treating the variables with null values
- Dropping the Variables with more skewed data
- Dropping the data with less percentage of missing values

EDA

- Distribution of variables
- Uni /Bi/Multi variate analysis
- Treating the data based on the values by binning them
- Outlier treatment
- Identification / Dropping unnecessary variables

Data Preparation

- Creation of dummy variables
- Scaling the continuous variables
- Test trail Split
- Understanding the correlation
- Dropping the standard categorical variables

Modelling

- Data Modelling using Stats and Sklearn libraries
- Model Assessment
- ROC process for Feature selection
- Confusion matrix creation
- Metric Beyond Accuracy
- ROC Curve
- Identifying Optimal Cut off
- Precision and Recall

Predictions and Conclusions

- Model evaluation by making the Predictions
- Making the conclusions

Understanding Data & Data Cleansing

Understanding Data:-

- Underlying data contained 37 features for 9240 customers
- Most of the variables are object type
- Many variables are holding the null values
- Prospect ID is the unique ID provided to each customer
- Converted is the Target variable to be focused on
- Lead Quality has the highest number of Null Values

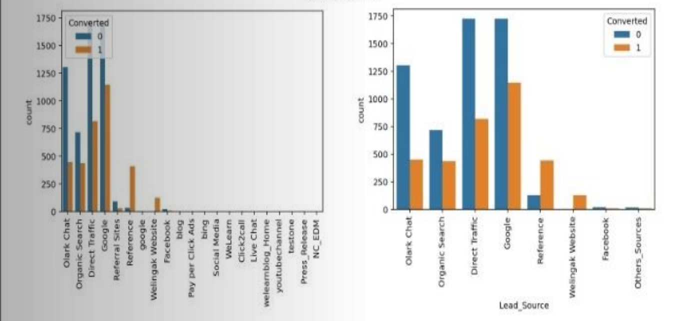
Data Manipulation(Cleansing):-

- Observed that there are 4 variables with ‘Select’ value which is same as the null value
- Identified that there are 10 variables has 35% of missing data, hence dropped it
- Identified that there are 3 fields with less than 2% of data hence dropped the records

This step provided 27 Variables and 9074 records without *Null* values

Exploratory Data Analysis - (EDA)

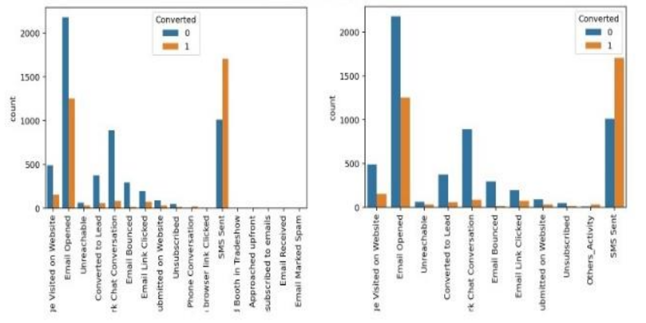
Lead Score



Before

After

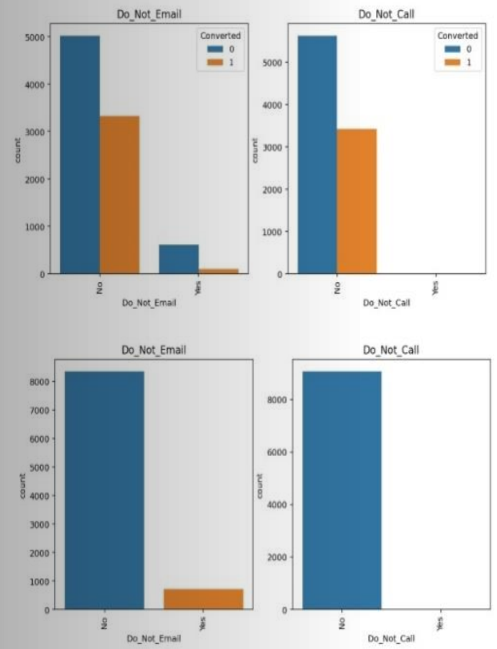
Lead Activity

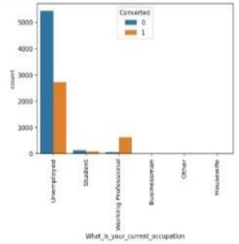


Before

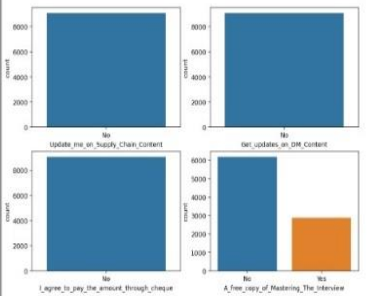
After

Do not Email/Do not Call

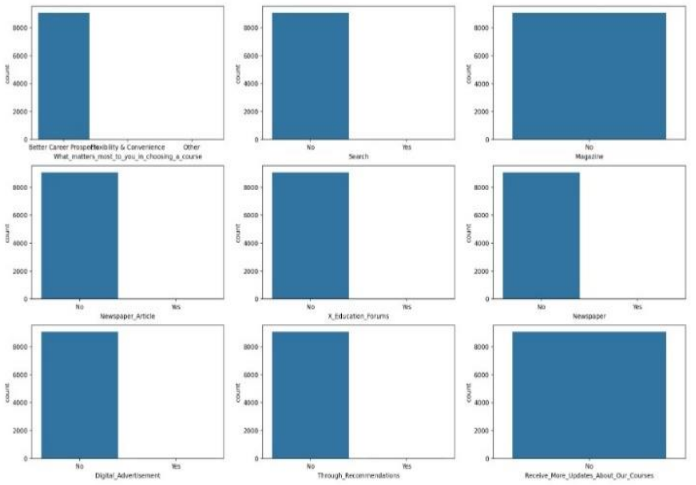




What is your current Occupation



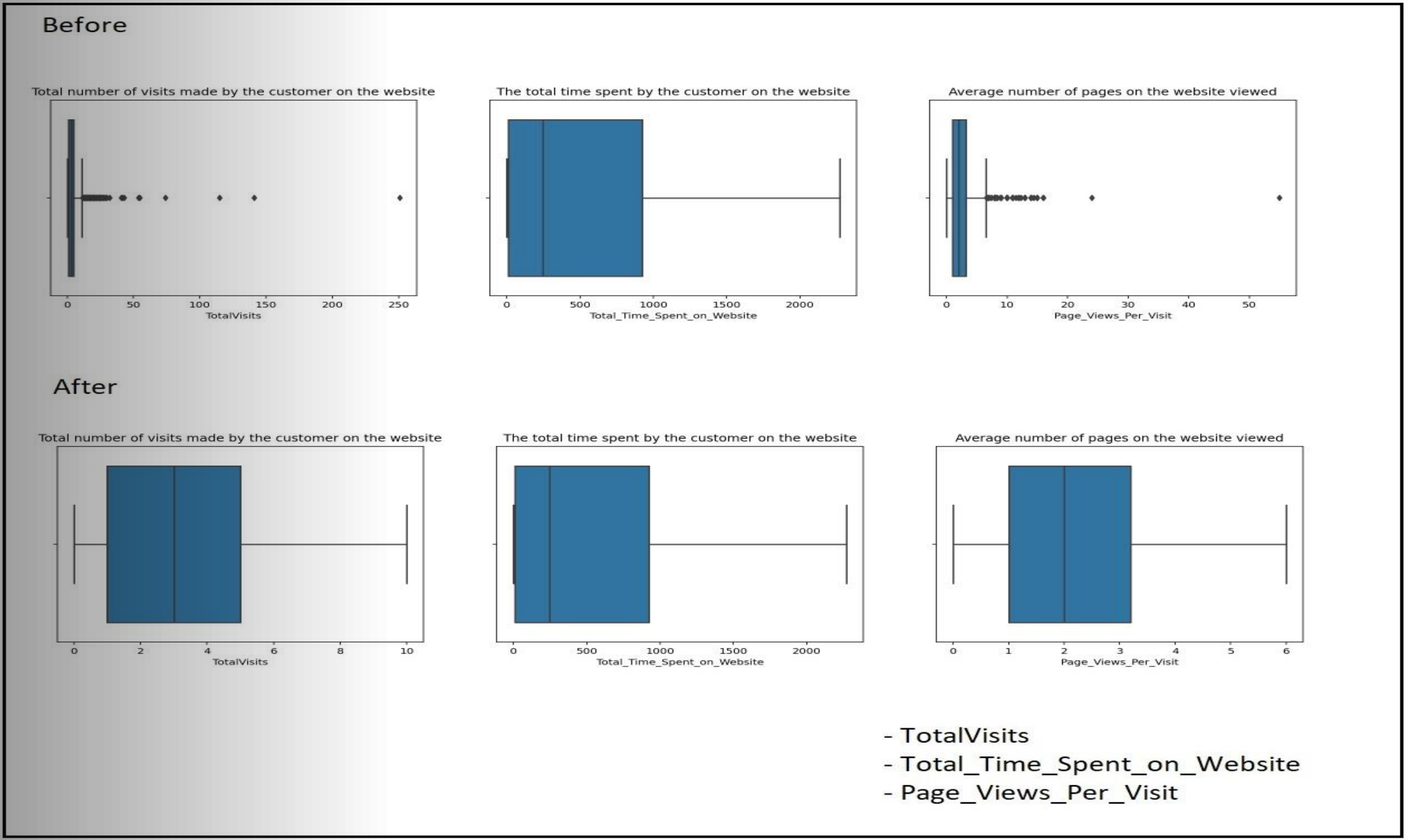
Update_me_on_Supply_Chain_Content,
Get_updates_on_DM_Content,
I_agree_to_pay_the_amount_through_cheque,
A_free_copy_of_Mastering_The_Interview



What_matters_most_to_you_in_choosing_a_course,Search,Magazine,Newspaper_Article,X_Education_Forums,Newspaper,Digital_Advertisement,Through_Recommendations,Receive_More_Updates_About_Our_Courses

EDA – Outlier Treatment

Outliers in continuous variables were treated using capping technique. All outliers were capped at 95th percentile of the values.



- TotalVisits
- Total_Time_Spent_on_Website
- Page_Views_Per_Visit

Observations

Treating the Skewed Data:-

Most of the variables had skewed data. Thus, dropped all such variables not adding any value to our Model

Combining the data:-

Variables like *Lead Score, Lead Activity & What is your current occupation* have multiple levels with most having lower number of data points. Thus, combined such levels as *Other*, so as to limit number of dummy variables created and increasing statistical explanatory power of these variables

Skewed Variables:-

- Update me on Supply Chain Content
- Get updates on DM Content
- I agree to pay the amount through cheque
- What matters most to you in choosing a course
- Search
- Magazine
- Newspaper Article
- X Education Forums
- Newspaper
- Digital Advertisement
- Through Recommendations
- Country
- Receive More Updates About Our Courses
- Do not email
- Do not Call

Conclusion:-

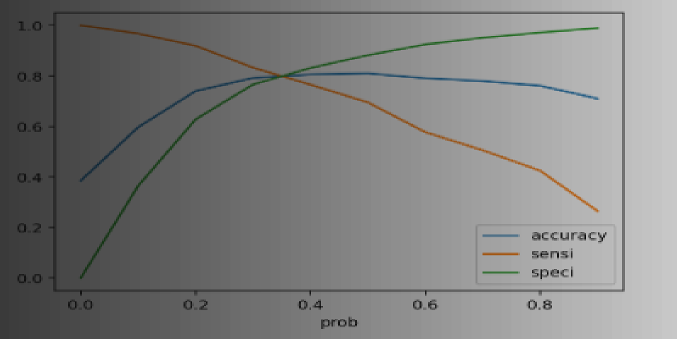
- 15 variables with skewed data were dropped
- 2 Variables(*Prospect ID and Lead Number*) having unique values were dropped

Only 10 Variables remained at the end of EDA process

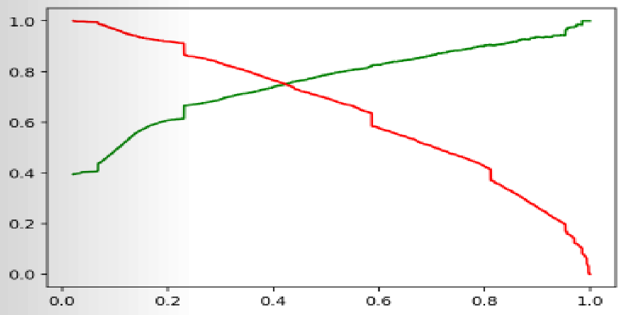
Data Preparation

- Numerical Variables were Normalized using MinMaxScaler
- Dummy Variables were created for all the categorical variables
- Total Number of rows left 9074
- Total Number of Variables after Dummy variables creation - 43

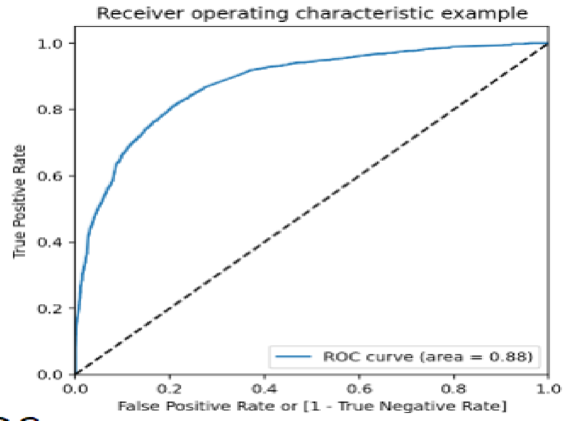
Data Modelling



prob vs accu,sensi,spec

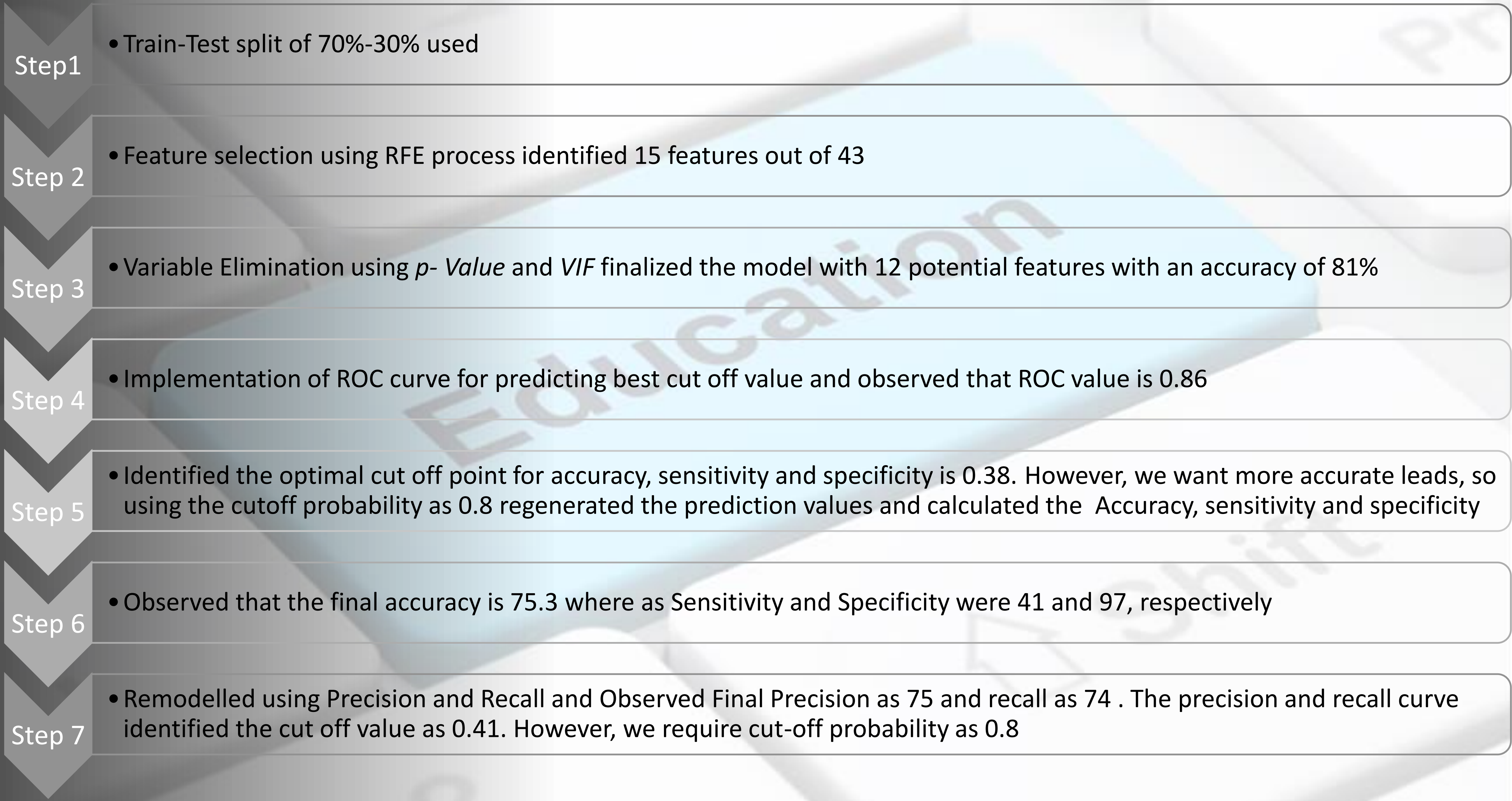


Precision vs Recall



ROC

Data Modelling Observations



Step1

- Train-Test split of 70%-30% used

Step 2

- Feature selection using RFE process identified 15 features out of 43

Step 3

- Variable Elimination using *p- Value* and *VIF* finalized the model with 12 potential features with an accuracy of 81%

Step 4

- Implementation of ROC curve for predicting best cut off value and observed that ROC value is 0.86

Step 5

- Identified the optimal cut off point for accuracy, sensitivity and specificity is 0.38. However, we want more accurate leads, so using the cutoff probability as 0.8 regenerated the prediction values and calculated the Accuracy, sensitivity and specificity

Step 6

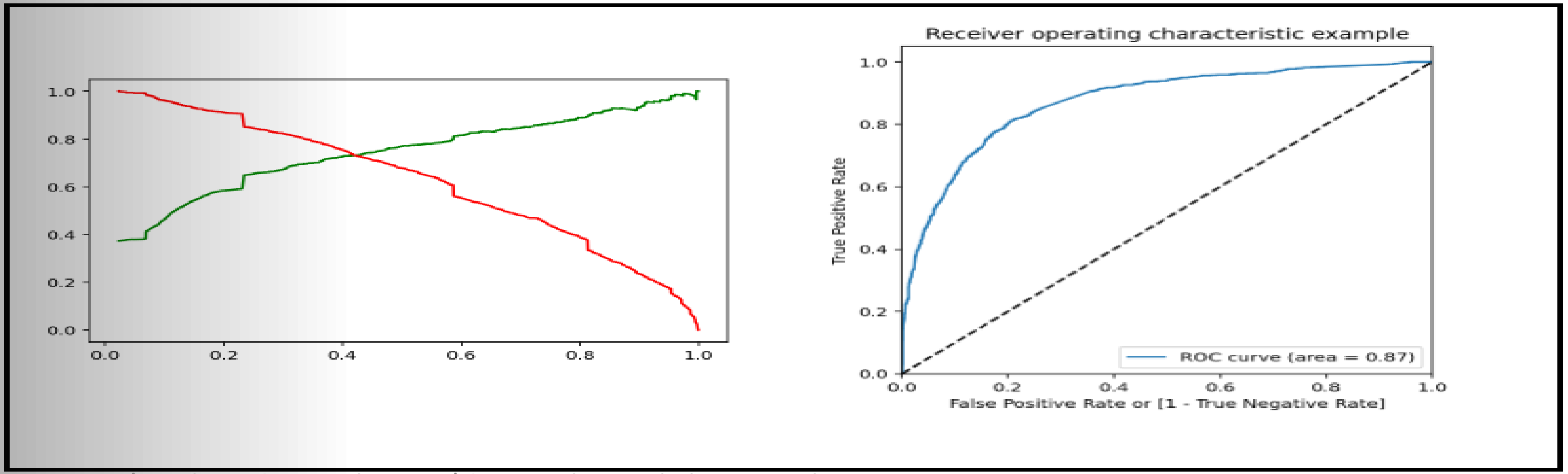
- Observed that the final accuracy is 75.3 where as Sensitivity and Specificity were 41 and 97, respectively

Step 7

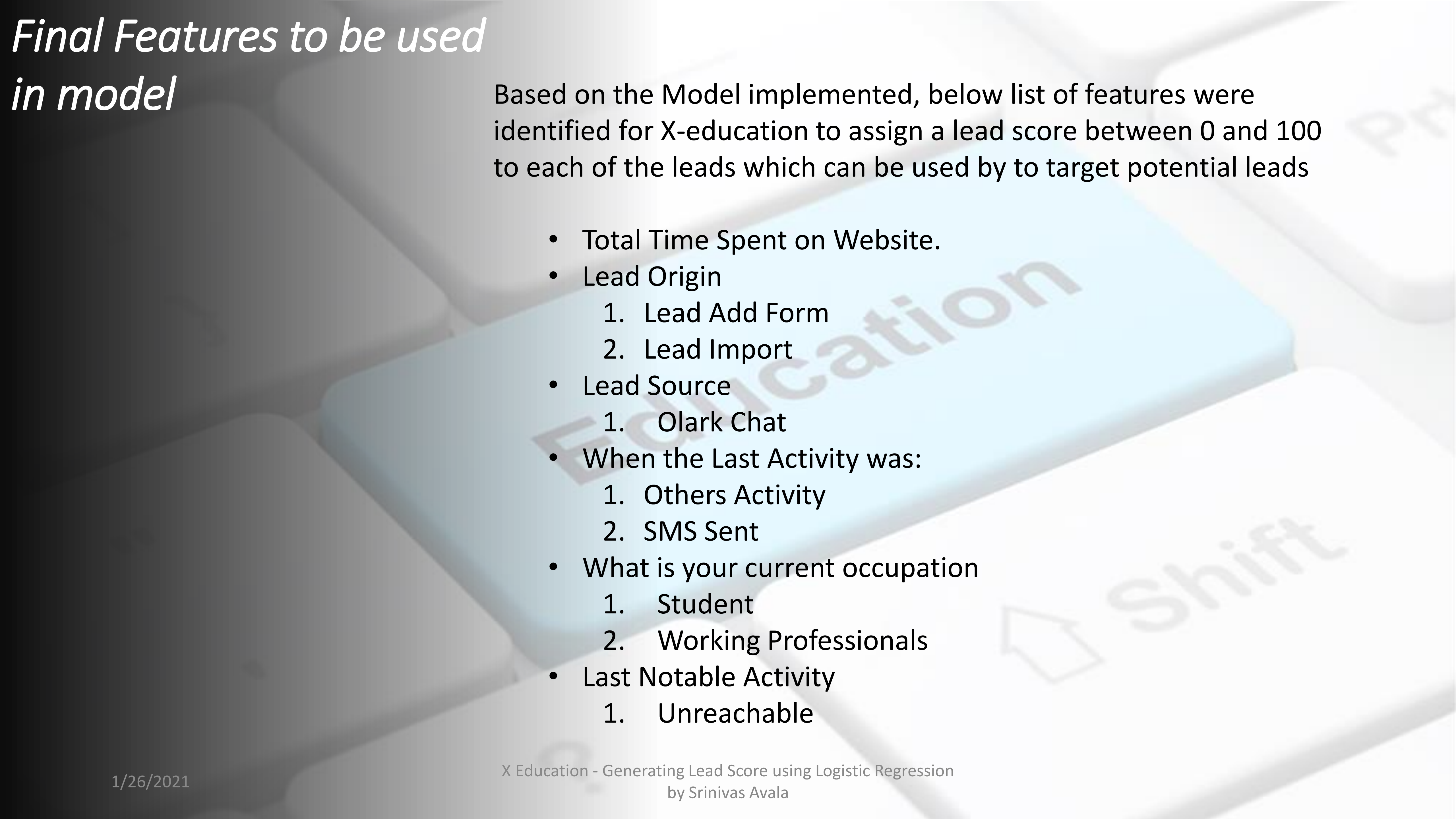
- Remodelled using Precision and Recall and Observed Final Precision as 75 and recall as 74 . The precision and recall curve identified the cut off value as 0.41. However, we require cut-off probability as 0.8

Predictions on Test Data

- Evaluated the Train model using the test data split
- Observed that Accuracy is evaluated as 75.3 where as Sensitivity is 78 and Specificity s 83 which is closer to the Train data set
- Observed that Precision and recall is 88% and 37% which is closer to the Train dataset. The ROC is 0.87 and Optimal cut off value is 0.80
- Precision vs Recall curve for test data and ROC curve is displayed as below



Final Features to be used in model



Based on the Model implemented, below list of features were identified for X-education to assign a lead score between 0 and 100 to each of the leads which can be used by to target potential leads

- Total Time Spent on Website.
- Lead Origin
  1. Lead Add Form
  2. Lead Import
- Lead Source
  1. Olark Chat
- When the Last Activity was:
  1. Others Activity
  2. SMS Sent
- What is your current occupation
  1. Student
  2. Working Professionals
- Last Notable Activity
  1. Unreachable

Calculating Lead Score

Lead Score can be calculated in 2 steps:

Step: 1

```
X = Leads_data_With_LeadScore['Lead_score'] = np.exp(-2.7167 +
(4.6870 * Leads_data_With_LeadScore['Total_Time_Spent_on_Website_Scaled']) +
(4.3440 * Leads_data_With_LeadScore['Lead_Origin_Lead_Add_Form']) +
(1.7033 * Leads_data_With_LeadScore['Lead_Origin_Lead_Import']) +
(1.0729 * Leads_data_With_LeadScore['Lead_Source_Olark_Chat']) +
(2.2155 * Leads_data_With_LeadScore['Last_Activity_Others_Activity']) +
(1.4820 * Leads_data_With_LeadScore['Last_Activity_SMS_Sent']) +
(0.4266 * Leads_data_With_LeadScore['What_is_your_current_occupation_Student']) +
(2.8027 *
Leads_data_With_LeadScore['What_is_your_current_occupation_Working_Professional'])
+ (2.0688 * Leads_data_With_LeadScore['Last_Notable_Activity_Unreachable']))
```

Step 2:

Probability (**OR** Lead Score) = X / (1 +X)

Conclusion

So, based on the model we developed, if we were to implement it on the given dataset, it would have generated 1513 leads (out of 9074) with Lead Score of greater than or equal to 80 (or probability of 80% or more).

NOTE: Lead score ranges between 0-100

Thank You…!!!