Lead Scoring

(Case Study)

Submitted by Mr.A.Vivekanand & Mr.Abhishek Pawar

Objective:

To build a Logistic Regression Model to predict whether a lead for online courses for an education company named X Education would be successfully converted or not.

Problem Solving Strategy

Step 1:Understanding the Data Set & Data Preparation

Step 2:Applying Recursive feature elimination to identify the best performing subset of features for building the model

Step 3:Building the model with features selected by RFE. Eliminate all features with high p-values and VIF values and finalize the model

Step 4:Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.

Step 5:Decide on the probability threshold value based on Optimal cutoff point and predict the dependent variable for the training data.

Step 6:Use the model for prediction on the test dataset and perform model evaluation for the test set.

Business Objectives

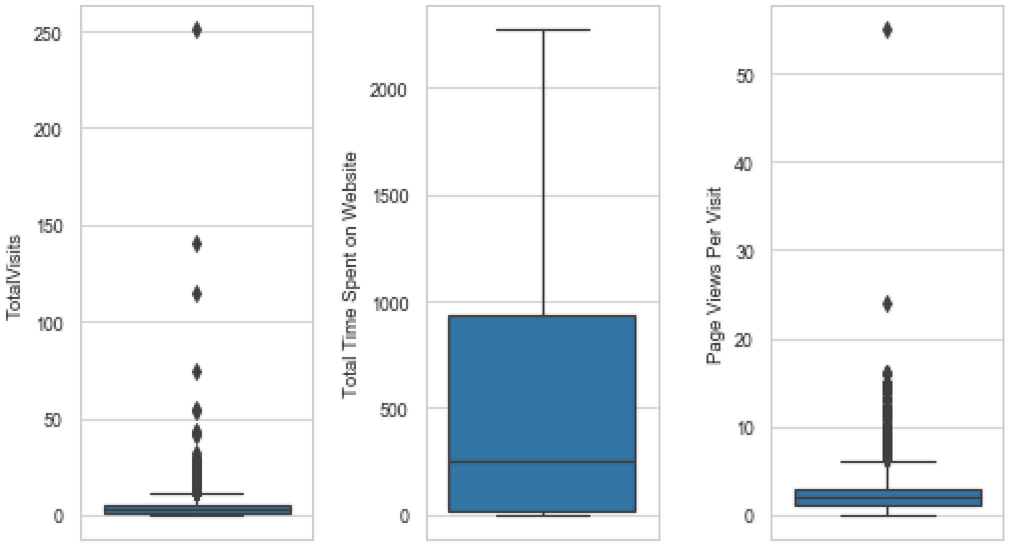
1. To help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.

2. 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.

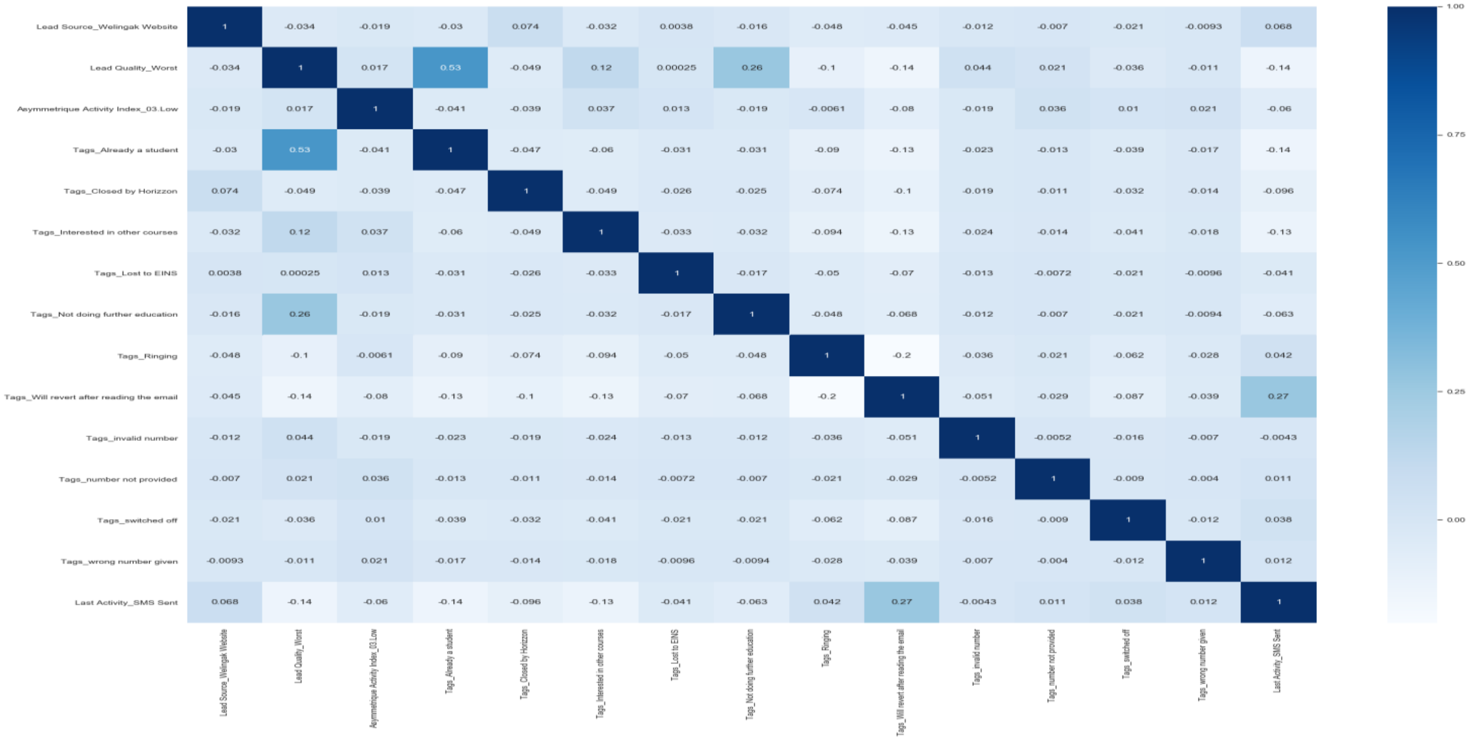
Strategy to meet the Objective

1.Create a Logistic Regression model to predict the Lead Conversion probabilities for each lead.
2.Decide on a probability threshold value above which a lead will be predicted as converted, whereas not converted if it is below it.
3.Multiply the Lead Conversion probability to arrive at the Lead Score value for each lead.
4.Handling 'Select' values in some columns
5.Assigning a Unique Category to NULL/SELECT Values
6.Outlier Treatment
7.Binary Encoding
8. Dummy Encoding
9. Test-Train Split
10. Feature Scaling

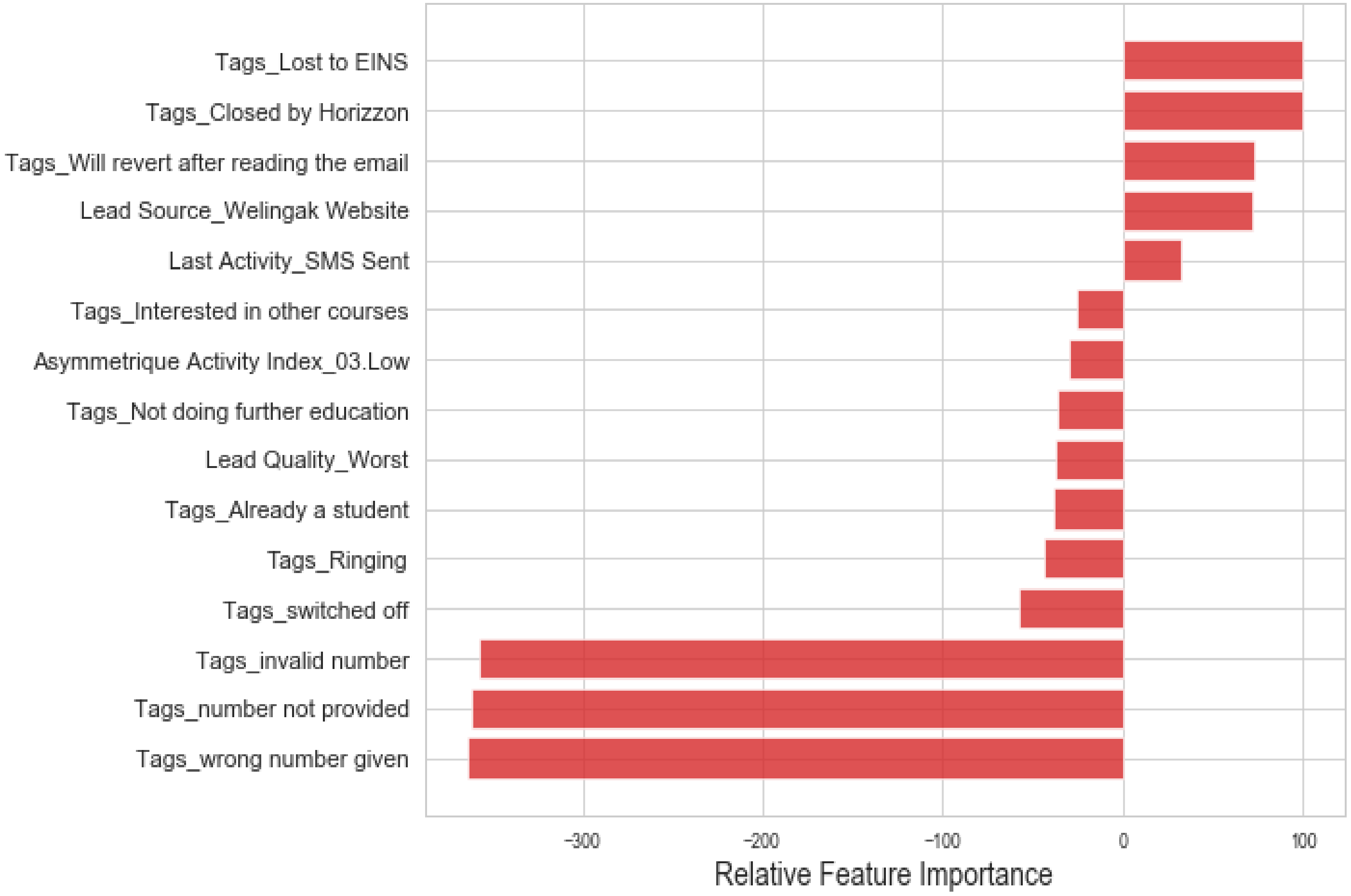
Outlier Analysis



Heatmap showing Correlations



Relative Feature Importance



Data Preparation & Feature Engineering

- Remove columns which has only one unique value
- Removing rows where a particular column has high missing values
- Imputing NULL values with Median
- Imputing NULL values with Mode

Recommendations & Problem Solutions

Which are the top three variables in your model that contribute most towards the probability of a lead getting converted?

Ans)

1. Tags_Lost to EINS
2. Tags_Closed by Horizzon
3. Tags_Will revert after reading the email