Lead Scoring Case Study Report

Logistic Regression is a classification model which is used for making prediction when out variables are categorical.

As per the business requirement it a classification problem so we will be building a Logistic Regression model to assign a lead score between 0 and 1 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps involved in the analysis:

1. Data Understanding.
   We have leads dataset from the past with around 9000 records. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
2. Data Cleaning.
   a. Missing value treatment
      As we saw there were few columns have missing values more than 35% which is too large so we dropped those columns.
   b. Handling invalid values
      Since it is an online course, we are not much concerned
      about Country and City columns so we will be dropping these columns. Many of the categorical variables had a level called 'Select' which is similar to Null values so we have treated those columns accordingly.
3. Data Preparation.
   a. Dummy variable creation for categorical features such as Lead Origin, Lead Source etc.
   b. Train and Test data split.
   c. Re-scaling the continuous features using MinMax Scaler which brings the features in standard range.
4. Model creation
   a. Feature elimination using RFE method to get top 15 vital features.
   b. Manual feature elimination by observing p-value and VIF to further optimize the model.
5. Model Evaluation
   We evaluated the final model by determining the optimal cut-off using Accuracy, Sensitivity, Specificity. Also we further verified it on basis of Precision Recall trade-off.
6. Making predictions on Test set
   We used our final model to predict the values for test dataset and found out that the accuracy of test data is same as that of train data.

As per our final model we can draw conclusions as follows:

- Customers who have visited and spent more time on the website are potential Hot leads.
- The X education company should focus on customers who have reached to them via Welingak Website and Olark Chat.
- The company should reach out to customers who have provided their details using Lead add form option.
- In order to get good job opportunities, students or unemployed customers are always looking for high market demand skills like courses offered by the X education company. So, these customers should be considered as potential Hot leads.
- Customers who had phone conversation with sales team and also enquired about the courses through SMS might be considered as Hot leads since these activities shows customers interest in courses.