Lead Scoring Case Study-Summary

By: Nicky Paul Eapen & Deepak Krishna AR

This case study aims to help X Education identify and target the most promising leads. Our goal is to assign a score for each lead by building a logistic regression model and selecting the features which are most important in determining lead conversion.

We started off with inspecting the data for any data quality issues. Few notable issues were:

1. There were a lot of features with missing value. Few of the columns were dropped. And for some important features, the rows of missing observations were removed.
2. There were a few features like 'Do not Call' , 'Get updates on DM content', 'Update me on supply chain Content', etc, having a very few to no variance. This data is insignificant and thus removed.

A total of 24 features were removed and we retained approximately 70% of the original dataset.

Outlier Analysis was done to identify any skewness. There were two numerical features with extreme outliers: 'TotalVisits' and 'Page Views Per Visit". These outliers were handled by capping the extreme value to the 0.99th quantile.

Dummy Variable Creation was done for categorical variables. Binary variables were converted to 0 and 1.

The Dataset is split into train and test data in 70:30 ratio. The numeric values are scaled using 'MinMax Scaler' from 'sklearn'. Correlation between variables was inspected by listing down the top correlated features and removing them thus avoiding any issues of Multicollinearity.

Feature Selection is done using Recursive Feature Elimination method in the sklearn library. Using RFE ranking, 15 best features were shortlisted out of 92 features. The statistics of these features were further inspected by generating Generalized Linear Model Regression Result using statsmodel. A total of 3 models were generated by eliminating a feature with high P value in each step. The final model was used to generate a prediction of the lead conversion probability and a random cut off of 50% probability was chosen to predict if the lead is converted or not.

To further evaluate our efficiency of the model we calculate metrics such as Accuracy, Sensitivity and Specificity using the confusion matrix and plot an ROC curve to test the accuracy of the model.

An optimum cut off probability value was found out by plotting Accuracy, Sensitivity and Specificity and we obtained a value of 0.42 as cut off. The predictions are now recalculated and validated again by calculating Accuracy, Sensitivity and Specificity.

Finally, on the test data set prediction was done based on our final model and the optimum cut off of 0.42. The evaluation metrics were recalculated and the values were very close to what we obtained in the train dataset.

The top 3 features obtained from the model are:

1. **Lead Source_Welingak Website**
2. **Total Time Spent on Website**
3. **Lead Source_Reference**

X Education should focus on leads that come through Reference and ‘welingak’ website.
The total time a prospect spends on the company's portal is a strong indicator of a Hot Lead.