Assignment: Part III

Summary Report: Lead score Case Study

The case study was based on an education company named X Education selling online courses to industry professionals leveraging several websites and search engines. The company's objective is to improve the conversion rate for the leads. X Education has reached out to us to help them build a logistic regression model to assign a lead score to each of the leads. A higher lead score means that the lead has higher likelihood of conversion.

Data Clean-up: The first step was to understand the data with the help of given data dictionary. Based on the column understanding, we started analysis on the dataset by understanding the patterns within each column. There were large percentage of null values that needed to be fixed. It was also observed that many records where defaulted with "Select" which is assumed that the user have not selected any of the drop-down selection on the user interface. These were treated using the following steps:

- Treated *"Select"* as blank value.
- Dropped columns considering a threshold of 40% for blank values
- Dropped columns that are heavily skewed towards one value.
- Dropped ID columns that does not impact lead score
- Merged values with < 0.5% occurrence into one category named "Other"
- Merged all values with < 0.1% occurrence for *Last Notable Activity* into "Other"
- Converted binary variables (Yes/No) to 0/1

There were many score variables which are assigned to the case **after** the sales person or the lead has spoken with the customer. Hence these columns were not relevant since these values will not be available when the sales team is looking to speak with the customers. We decided to drop these columns. Such columns are

- Asymmetrique Activity Index
- Asymmetrique Profile Index
- Asymmetrique Activity Score
- Asymmetrique Profile Score
- Tags
- Last Activity

Individual analysis on remaining categorical columns was performed and observed that the value counts indicate a very heavy skewed values (almost 99%) towards one answer for the columns like Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations. So we dropped such columns as well.

Some columns in the continuous numerical variables like *TotalVisits*, *Total Time Spent on Website* and *Page Views per Visit* had outliers identified which needed to be treated. This was treated by capping the values to 99 percentile and imputing the missing values of total visits with 50% quantile value based on the pattern of the values in this column.

The dataset has almost 38% conversion ratio based and hence the data is good for modelling.

Data Preparation: The next step was to prepare the data for Logistic regression model. We created dummy features for categorical variables with multiple levels. Once the dummy variables were created, the original columns were dropped. Test-Train data split was performed using the 70 to 30 ratio and a random state of 100. ***Scaling*** was performed to standardize the dataset to reduce skewness of the model. Co-relation matrix was generated to identify the variables with high collinearity. Two variables 'Specialization_Unknown','Occupation_Other' where dropped from test and train data set based on their high co-relation with the other variables.

Model Building: The next step was to build the logistic regression model. First the feature selection using RFE with a set of 25 variables was performed to select the columns for the modelling exercise. The model was created using all the RFE recommended columns. Generalized Linear Model Regression Summary Results provided us the critical decision factors that enable to choose the variables that can be dropped. The criteria that was followed are VIF should be under 5, P-Value should be under 0.05, accuracy, sensitivity and specificity should be more than 75% and the model should not over fit. We continued to delete the columns and rebuilding the models based on the p values followed by VIF value comparisons. Finally, the 11th iteration model was finalised with 15 variables. All variables had a good value of VIF and P-Value. So, no more variables were dropped.

Model Evaluation: The next step was to evaluate different metrics and the cut-offs for the model. We generated the ROC curve found out that the ***area under the curve is at around 0.89***. The probability cut off curve using sensitivity, specificity and accuracy was drawn along with the Precision and recall roll-off. We then finalized the ***optimal cut off at 0.35***.

Validation against test set: We performed predictions against the test data set and reviewed the accuracy, sensitivity and specificity etc. The criteria was met with more than 80% sensitivity.

As a result of the model, the lead scores were generated that could be used for converting the leads. A lead score greater than 35 have a likelihood of around 80%+ conversion.

During the model generation, the learnings gathered are as follows:

- Current Occupation mentioned as working professionals is more likely to be converted as a lead
- The customer who had a phone conversation already with the sales team is likely to be converted as a lead
- The customers who have filled up the Lead application form is likely to be converted as a lead
- The customers who have assessed Welingak Website is likely to be converted as a lead

Sales team should consider all the above points to identify and improve conversion rates