Summary

Lead Scoring Case Study (18 Oct 22)

Submitted By : Naveen Upadhyay & Yerra Reshmitha (DSC-43)

1. Model building and prediction has been undertaken for company X Education to find more industry professionals to join their courses. The basic data provided to us gives us an insight into their behavior viz. the time they spent on the website, frequency of visit, conversion rate etc.

2. The following steps were used towards construction of the model:

(a) **Sourcing and Understanding data**. The dataset was imported and converted into a dataframe with the help of pandas library. The provided data dictionary was understood and the data shape, missing value checks, outlier checks and duplicate data checks were undertaken.

(b) **Data Cleaning**. During this step the columns with more than 40% missing data were dropped. The remaining columns having null values in significant numbers were changed to 'not provided' to prevent loss of data. 'Select' was replaced with 'NaN'. Rows still having missing values were dropped.

(c) **Data Visualization (Exploratory Data Analysis)**. Thorough EDA was undertaken to check the condition of the data. A number of categorical variables were found to be irrelevant and thus, were dropped. Outliers were observed in few numerical variables and were appropriately dealt with.

(d) **Handling Sales team and highly imbalanced data**. The columns populated by the sales team as well as those columns having highly imbalanced data were dropped.

(e) **Creating Dummy Variables**. The dummy variables were created for categorical columns.

(f) **Train Test Split**. The data was split in the ratio 70:30 for train and test data respectively. MinMaxscaler() was used for scaling the numeric features.

(g) **Building Logistic Model**. RFE was undertaken to achieve the top 15 relevant variables. Thereafter, the variables were removed manually depending on VIF > 5 and p values >0.05.

(h) **Model Evaluation**. A confusion matrix was created. An optimum cut-off value=0.32 was selected based on ROC curve. The vales of Accuracy, Sensitivity and Specificity were observed to be around 80% each.

(i) **Prediction**. Prediction was done on test data with optimum cut-off value=0.32 and the Accuracy, Sensitivity and Specificity were again observed to be around 80% each.

(j) **Precision recall**. This method was also used to recheck the model and the values obtained for train and test data were found to be similar with an accuracy of around 80%.

3. **Result**. The optimum cut-off was observed to be at 0.32. On train data we achieved an accuracy of 77.47%, Sensitivity of 82.79% and Specificity of 74.14%. For test data we achieved an accuracy of 77.08%, Sensitivity of 81.19% and Specificity of 74.74%. This indicates that the model is able to predict the dependent variable with around 80% accuracy which meets the required criteria.

4. **Conclusion**. Following personnel are most likely to convert and generate leads for the company:

(a) Whose last activity was through SMS or Olark chat conversation.

(b) Who has a management specialization.

(c) Who are working professionals.

(d) Who are visiting website repeatedly.

(e) Who are spending a lot of time on the website.