Summary of Lead Scoring case study

By- Subham Kumar and Mohan Babu Uppu

OBJECTIVE:

- Building a logistic regression model for X education by assigning a lead score between 0 and 100 for targeting particular leads to be converted or not. Here leads refer to the individuals finally enrolling for a course by giving contact details like phone number and email address
- The CEO has given a ballpark of the target lead conversion rate to be around 80%
- The proposed model should be able to adjust to if the company's requirement changes in the future as well

METHODOLOGY:

- **Data inspection and cleaning-**
  Data was loaded along with the required libraries and shape, size, data types of different columns was checked. The columns were checked for null values and unwanted columns as well as rows/columns with high percentage of null values were removed.
- **EDA and data preparation-**

  Data visualization (box plots) were analyzed for checking outliers. Dummy variables for categorical variables were created and the initial as well as repeated columns were removed. The data was then split into train and test data set followed by data scaling.
- **Building a logistic regression model-**

  A logistic regression model was built with RFE feature selection and VIF to check for collinearity. The irrelevant variables and also variable with high collinearity were removed and model was built with 34 variables. After checking the p value at each stage, the variable having p value more than 0.05 were removed and finally used 16 variables for the final model.

Finally, the VIF and RFE was also checked for the model to ensure only relevant features have been used for the model and there is no multi-collinearity.

CONFUSION MATRIX AND OTHER TESTS FOR CHECKING ACCURACY:

- Further process like Confusion matrix, Model Accuracy, Sensitivity and specificity been done for the final model.
- ROC curve shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. As per the graph, the ***ROC curve covered almost 86%*** of the total area for our final model.
- The values of accuracy, sensitivity, and specificity were used for finding out the optimal threshold and 0.4 was found out to be the optimum point to take as a cut-off probability from the graph
- Precision and recall were also plotted and both the curves met at 0.5

PREDICTION ON TEST DATA:

Lastly, the model evaluation of the test data is done by doing the confusion matrix after making predictions, checking the model accuracy, sensitivity and specificity.

Below are the final values for model evaluation:

- **Model Accuracy (Correctly predicted labels / Total no. of labels):** 77%
- **Sensitivity (TP / TP + FN) :** 0.81
- **Specificity (TN / TN + FP) :** 0.74
- **false postive rate (FP / TN+FP):** 0.25
- **Positive predictive value (TP/ TP+FP):** 0.74
- **Negative predictive value (TN/ TN+FN):** 0.81

CONCLUSION:

The key features that contribute most towards the lead getting converted:

- TotalVisits
- Total Time Spent on Website
- Lead Source_Welingak Website

Potential features which can contribute in the conversion of leads and can be focused more:

- Lead Souce_Olark Chat
- Lead Source_Reference

The conversion of students and unemployed candidates is very less hence working professionals with specialization like banking, investment and insurance and marketing management can be focused more for conversion.

Since the total visits and total time spent on website is more , the company can advertise more on the website and also introduce chat support for addressing doubts of the individuals. Referrals by existing students can be encouraged to increase the lead conversion as well since it is one of the potential features.