Lead Score Case Study: Summary
Submitted by Anuj Sharma

The Case study is summarized into the following steps of action taken to perform the case study.

1) Understanding the business problem
   The case study is with X Education company. X Education company is trying to sell their courses to multiple potential students. The job here is to identify the leads with higher chances of opting for the course and provide the information to sales team for them to focus on increasing their conversion.
2) Breaking it into data science problem
   The required challenge can be solved by training a Logistics regression model on the available information and predict the students with higher potential of opting for the course.
3) Collecting data set
   A dataset is available with more than 9000 students' entry consisting of both converted and not converted leads.
4) Reading the data set
   Reading the dataset using Python language in Juypter Notebook environment.
5) Correcting the dataset into readable formats
   a. Imputing NA Variables
      Substituting the NAN values with either mod or mean values.
   b. Modifying variables
      Modifying categorical variables with making combinations and concatenation. Modifying categorical variables with yes and no value to 1 & 0.
   c. Dropping Empty rows
      Removing the rows with multiple NAN values.
6) Exploratory Data Analysis
   a. Univariate EDA
      Analyzing individual variable on their distribution and frequencies.
   b. Bivariate EDA
      Analyzing relations between two variables and their dependencies on each other. Mostly analyzing variables with their relations with conversion.
   c. Multivariate EDA
      Analyzing relationship between multiple variables using correlation and pair plots.
7) Modifying Variables
   a. Dropping Irrelevant Variables
      There are some columns that have all same values so dropping them. Like 'I agree to pay the amount through cheque', 'A free copy of Mastering The Interview', and other columns like Prospect ID and Lead number, that are just serially generated unique numbers.
   b. Categorical Variable to Dummy Variables
      creating dummy variables from multiple categorical variables that we have.
   c. Numeric Variables- Scaling
      Scaling is not done in this case study, as the number gap was not big.
8) Train-Test Split
   Splitting the new modified complete dataset into two datasets, train and test.
9) Recursive Feature Elimination Method- Top 30 Choice of variables
   Since in the new dataset there are more than 80 columns, so using RFE method to get the top 30 contributing variables directly impacting Target Variable- Converted.
10) Logistics Regression Model Training
    In this case study used statsmodel to train logistics regression model on the training dataset. Used Generalized linear model training method with statsmodel as it provides summary and p-values directly to select the variables.

11) Backward Approach- Removing variables.
    a. Using p-Value

       Since the null hypothesis is that the independent variables have no linear relation with the dependent variable. So any value of lower than 5% ie lower than 0.05 suggest that we can reject the null hypothesis and can conclude that the Independent variable has a direct relation with dependent variable. And for any p-value higher than 0.05, we continue with the null hypothesis and hence remove the variable from staying in the model training. The same practice is used in the model training for selecting the variables in training the model.

    b. VIF

       Variance Inflation Factor defines the multicollinearity between variables, any VIF higher than 5, is not good to select the variable.

    c. Iteration of point a & b

       Point a & b are iterated till the final selection of variables is achieved.

    d. Finalizing a Model

       By repeating point c, a final model with 15 iterations of model training an optimum selection of variables is achieved in the model training.

12) Predicting Values with Logistics Regression Model

    With the final model, values are predicted from the model.

13) Setting Cut-Off, Parameters Checking

    The model needs high accuracy and high sensitivity to find out students with higher probability to opt for the course.

    a. Accuracy
    b. Specificity
    c. Sensitivity
    d. False Positive Rate
    e. False Negative Rate

14) ROC Curve- Finding Optimum Cut Off

    Using the ROC curve, an optimum cut-off of 0.28 is decided.

15) Precision & Recall, Trade off- if Required

    We wanted a higher recall and with the trade-off achieved from precision recall curve is a cut off of 0.42

16) Prediction on Test Dataset- Model Evaluation

    The predictions on test results are really good and an accuracy of 84% is achieved on test data.