Lead Scoring – Group Case Study: Summary

- By Anindo Mazumdar & Mukundan AP

The purpose of the document is to highlight steps taken to analyse the data and come up with the required solution.

Our analysis process was broken down into the below components –

1. **Analyzing the problem statement and requirement -**
    a. The first part was to understand what the problem statement was and defining high level steps which we'll take to resolve it.
    b. We noted that the aim is to provide a predictive model which gives 80% or more predictive power of successfully identifying a potential lead.
    c. Apart from creating the model, we were also tasked to summarize the finding in presentation format as well as answer a few questions related with the problem statement.
2. **Understanding the data and pre-processing on the data before model building -**
    a. Once we read through the data dictionary and got a feel for what the data is, we started off by importing the data and analysing various columns present.
    b. We eliminated the columns where we encountered a lot of missing values (>30% of the total rows)
    c. For rest of the columns, we looked and what unique values the column contains, and did missing value imputation/ column dropping based on that
3. **Visual cues and Outlier treatment -**
    a. Next step in analysis was to plot the data to try to see patterns emerging.
    b. We also treated outliers to reduce possibility of error being introduced by very high/very low values.
4. **Creating dummy variables and splitting data into test & train data -**
    a. We then started by looking at categorical values and creating dummy variable to consider them in further analysis.
    b. We also looked at Correlation matrix to get a feel for correlation between variables in the data model
    c. We then created to sub data set of the for our analysis purpose (namely test and train data) and continued our analysis on the train data set.
5. **Scaling variables -**
    a. To ensure all the variables in the model lie on the same scale, we used Standard Scaler to scale the variable values.
6. **Variable selection through RFE and manual observation –**
    a. We used RFE to limit the model to 15 most influential variables and then used manual elimination method to further reduce the predictor variables. We utilized VIF for analysis work here.
7. **Model Tuning and finalization**
    a. We then used various cut off values and calculated accuracy , precision and recall at each point to reach to an optimum value set.

b. The same value set was then tested on test data to derive final results.

Learnings Gathered from the assignment –

a) Importance of understanding the input data first before deep diving into solution building
b) Takes judgment call when it comes to variable elimination and null value imputation
c) Manual fine tuning of model is absolutely essential, cannot just rely on available feature elimination methodologies