Lead Scoring Case Study Summary

—By Sankalp Seksaria and Vaibhav Parakh

The case study presented a problem statement towards building a model to assign lead score to prospective customers such that a customer with a higher lead score has a higher probability of conversion. The lead score to be assigned was on a scale of 0 to 100. The analysis began with reading data and understanding the nuances of each column. We began with Data Wrangling, under which we checked for unique customers and lead number. Post which, we accounted for missing values in the dataset. We had 17 columns in total which had missing data; out of which 7 columns accounted for more than 45% of missing data which were immediately dropped; and for the rest of the columns, suitable imputation was performed. Post imputation of data, we did an outlier treatment followed by grouping of various values under each feature under distinct heading. This also included checking for class imbalance for each feature and dropping those which did have a very high-class imbalance since it would not aid the analysis of the data.

Once we were through preparation of data, we undertook Univariate and Bivariate analysis, followed by correlation heatmap which clearly helped us understand what all features were contributing to higher conversion. Post visualisation, we undertook variable transformation which were of two types:

1. Creation of Dummy Variables followed by dropping the original variable.
2. Creating a binary map for variables which had Yes/No as values

After variable transformation, we split the data into training and test set and proceeded with scaling of features. For scaling of features, we undertook StandardScaler which scaled the data by centering the mean at 0 and standard deviation at 1.

We first began with building a general model using Statsmodels and took stock of the p-values of all the features. To help us with the feature selection, we called for Recursive Feature Selection from sklearn library and selected 15 top features that could reasonably help us in building a logistics model. We built a second model with these top 15 features and removed one feature that had a higher p-values in comparison to the rest of the features. After taking stock of the p-values and VIF values, we arrived at the final model.

We calculated all the metrics keeping threshold probability at 0.5 and also found out optimum cut-off point (threshold probability of 0.3) and recalculated all the metrics. We then went on to run the model on the test dataset and eventually created a table that had a column lead score which assigned value to all the probabilities of lead conversion on scale of 100.

To help us answer the business problem, we also graphed out which features contributed to higher probability of conversion.

Major challenges were encountered in data wrangling and learnings were in terms of which columns to be dropped and how to deal with data imputations and how to club variables and deal with data imbalance.