Lead Scoring Assignment

Vivek Venugopal and Thousif Ahmed

October 2021

We began this assignment by getting an intuitive feel of the data using pandas functions such as .info() and .describe(). We then analysed the variables for null values. Initially we dropped the variables with more than 60% null values. For the other variables we analysed each variable individually to decide on further action - some were dropped and some were imputed. Once all the null values were addressed, we began our EDA where we analysed each variable and addressed the skew in data. For example the Country variable was heavily skewed as most of the customers were from India.

After we were satisfied with this, we began preparing the data for modelling. The first step was to create dummy variables for categorical variables; this was done through mapping and using the pandas get_dummies() function. Post this the data was split into train and test data using the scikit learn library. Before building the model, the data was scaled using the standard scaler from scikit learn.

The model was built using the statsmodel library. To improve the model, top 20 variables were selected using RFE. Variables were dropped one by one till all the p-values were below .05. To further improve the model VIF values were analysed to ensure there is no multi co-linearity. The final step was to ensure the correct cut-off of the model. This was done by using ROC analysis.

One of the important lesson we learnt was about handling skewed data. Initially we did not address this and we could see that the model output did not make any business sense at all. When we started to investigate, we realised that we had to tackle the skewness in the data first. Once we addressed this our model started to behave in a more realistic manner.