The problem to be solved is to use the data and evaluate Z score for each lead, which will be an indictor for probability of conversion. The Z score will assist the sales team to prioritize the customer with higher probability first rather than reaching all leads.

The data was imported with memory optimized data by down casting the datatypes. While preparing the data, columns had missing values, in this business case, there was ‘SELECT’ option which is equivalent to a missing value. Missing values were handled by dropping and imputing some values with ‘UNKNOWN’ as imputing with a median or mode will modify the raw data and alter the result.

By performing this imputation with ‘UNKNOWN’, it was understood that maximum null entries are coming from the Olark Chat platform and corrective action need to be taken. Two numerical columns had outliers which were treated by capping. Processed data was then Standardized and dummy variable were created for modelling.

Using RFE, features were selected and model building, evaluating against p-value and VIF, dropping feature continued until desired results were achieved. The optimum cut-off value was selected as 30% considering in this business case, sensitivity, it is ok even if non-converted value has conversion, as we will not miss out any.

The probabilities were added against each lead number and now the sales team can prioritize the leads to pursue based on these probabilities. The final model had ROC value of 0.89, sensitivity value of 0.84 in test data and a recall value of 0.59.