Lead Scoring Case Study- Summary Report

Data reading and cleaning: Initially, presence of missing values has been inspected. It has been observed that most of the variables are having missing values. However, despite imputing the missing values. It has been concluded to drop the columns that have missing values **greater than 30%.** It has also been determined to remove columns that has no contribution for further analysis such as 'Prospect ID', Lead Number' and so on.

Similarly, for attributes which have missing values less than 30%, column wise analysis is performed. Some categorical variables have values as '**Select'** which is no different than a null or missing value. These values are replaced with null and the missing value percentage for all the columns was analyzed again. Now, few more columns showed missing values greater than 30% and all these columns have been dropped. For other attributes, despite dropping the records, it has been preferred to drop the rows with missing values. The reason for dropping off the missing values rather than imputing is to not introduce any **bias** in the dataset. After dropping off the columns, the dataset is inspected to find the ratio of values in each column to ensure if the attributes are **not biased**. But, there are about 10 variables which not having enough distributed values. Hence, these columns are dropped as well.

Univariate and Bivariate analysis: Once, the data cleaning was performed on the dataset, univariate and bivariate analysis is performed. From the univariate analysis, outliers in numeric variables have been identified. However, rather than removing those row, all the outliers have been capped to **95 percentile**. The reason is that in the data cleaning step, a large portion of data has been dropped. Thus, to avoid further loss of data, this method of outlier treatment has been considered. Likewise, in bivariate analysis, only one numerical variable showed a strong positive correlation with the target variable.

Data Preparation and modelling: After creating the dummy variables followed by the standard scaling of numeric variables. There are in total **101 attributes** available for developing the first logistic regression model. The first model was built by considering the top **15 variables** using the RFE technique. The feature engineering has been performed by analyzing the statistics summary and multi-collinearity of the attributes. Once the final model has been attained, the probabilities for each record has been classified based on the arbitrary value (0.5) and evaluated. But, to finalize the optimal threshold value, the ROC curve and precision-recall curve has been analyzed and the resulting cutoff value for a balanced metric was observed to be **0.42**.

Final evaluation: Finally, considering the resulting cutoff value being 0.42, the model evaluation has been performed on the train data and the **accuracy** was approximately 80%. Similarly, the model evaluation was performed on the test data and the resulting metrics were almost closer to as that of train data, proving that the model is **stable** and classifying the records to satisfy the business needs