SUMMARY

The first step in the procedure was to comprehend the current company issue. After the issue was identified, data understanding became the main priority. In order to do this, we had to thoroughly familiarise ourselves with the words and definitions of each feature as well as their application to the issue at hand by digging into the data dictionary.

The next step after understanding the data was data preparation. This included activities like importing the data and carrying out data cleansing processes. Outlier identification was done, missing values were taken care of, and data types were repaired. Additionally, characteristics like "prospect id" and "lead number" that were unnecessary and did not advance the research were removed.

Exploratory data analysis (EDA) was carried out after the data had been produced. This involved using univariate, bivariate, and multivariate analysis approaches to look at the data. To glean insightful expertise and a thorough comprehension of the dataset, data imbalances were also taken into consideration.

The data was further prepared for model development after EDA was finished. Highly associated variables were found using correlation analysis, and these were then eliminated using knowledge of the business area. To facilitate their inclusion in the model, categorical variables were also changed into dummy variables.

For the purpose of evaluating the model, the data was then divided into a training set (70%) and a test set (30%). In order to prevent any bias in the model's performance, standard scaling was used to make sure the variables were on a same scale.

The data had been correctly prepared; therefore, the modelling step had begun. The modelling method in this instance was logistic regression. The model was tested on the training set with a cutoff value of 0.5 after being trained using statsmodel.

Several metrics, including accuracy, sensitivity, specificity, recall, precision, and the ROC curve, were used to evaluate the model's performance. To find the ideal cutoff range, accuracy, sensitivity, specificity, recall, and precision curves were also investigated.

A threshold value of 0.35 was chosen after thorough consideration and will be used to any future analysis or decision-making procedures. In this thorough procedure, the business challenge was understood, the data was prepared and explored, a logistic regression model was built, its performance was assessed, and an ideal cutoff point for decision-making was established.

The model produced an accuracy, sensitivity, and specificity of 81% on the training set after using the cutoff value of 0.35. On the same set, the model also earned recall scores of 79% and precision scores of 72%. The model had a recall score of 79%, a precision score of 75%, a sensitivity score of 79%, and an accuracy score of 81% for the test set.

The model consistently performs well across various assessment measures for both the training and test datasets. This shows that the model is trustworthy, prevents overfitting, and effectively generalises to new data. Its strong performance is further supported by the absence of substantial biases in its predictions. The model routinely achieves excellent accuracy, in conclusion.