SUMMARY

Problem statement demanded us to find the hot leads among all the leads generated through various sources and origins. We need to bring the conversion rate of leads from 37% to nearly 80%.

For building the model we had to clean the data to select only the best features. There were certain columns with single category and skewed behaviour due to which they were dropped. Outliers were removed for numerical variables and dummy variables were created for multi-level categorical variables. EDA was performed to gain insights between various features and converted column. We find that Landing page submission lead to the greatest number of leads. Also google and direct traffic on website generate most leads. Leads who spend more time on website are likely to get converted. SMS communication tend to show better response in terms of conversion rate. Management related profiles show best conversion rates. Mumbai city generate most leads.

For building the model we split the cleaned dataset to training set and test set. Standard scaling was performed to scale the numerical variables. We built an initial model with all the features. After that we have used RFE technique to select top 20 features and rebuild the model. Then we check the p-value for the coefficients and eliminate the features with p-value greater than 0.05. We also calculate VIF and use a threshold of 5 and eliminate the features having VIF more than 5. By repeating these steps, we arrive at a final model with p-value and VIF withing out threshold and we use it for further analysis.

We predict the values for the train set using the model built. We then use the evaluation metrics (accuracy, sensitivity, specificity) to evaluate our model. Initially we use a default cut-off value of 0.5 to predict the score as the result from logistic model is probability. So, probabilities higher than 0.5 are considered as converted and less than 0.5 are not converted.

We then use ROC curve and draw graph for sensitivity, specificity and accuracy for various probabilities and at the intersection of 3 lines we get the optimum cut-off probability for our model. We get optimum cut-off value of 0.35. We obtain accuracy of 84.7%, sensitivity of 84.2% and specificity of 85.07% on training set. We get a precision of 84.4% and recall of 78.6% on training data.

After this we perform the evaluation on test data. Again, we scale the numerical variables first. Then we perform the predictions on the test set using the model we created. We use the cut-off probability of 0.35 for deciding the hot leads. We get an accuracy of 83.4%, sensitivity of 82.3% and specificity of 84.02% on the test set.

After this we assign the lead score by multiplying the probability with 100 to get a lead score out of 100. Higher the score obtained higher is the chance of lead getting converted.