SUMMARY

Goal of the Assignment:
Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Description:
The assignment was started with reading the data and cleaning the data of all the missing values. Some columns with more than 70% missing values were dropped, in all other columns the missing values were imputed with a logical value that was found out by visualising each column.

After missing value treatment univariate analysis was performed on each column to see which columns are relevant and will impact our model building. The columns with more than 90 % same value were dropped because they did not contribute anything to the analysis or model building.
Before proceeding to model building the data was prepared by binary mapping of categorical variable and dummy encoding and then split into train and test data and the train set was scaled using standard scaler.
For modelling we chose 15 variables for our initial model through RFE and dropping two variables got us to our final model with significant p-values and VIFs.
The ROC curve for our model was plotted after constructing the confusion matrix.
The area under curve (AUC) was 95%. The curve plotted between Accuracy, Sensitivity and Specificity gave us the optimal cut-off probability to be 0.2. The Lead score were assigned to each of the 9000 leads and then precision and recall were calculated.
The prediction was made on the test set and we got the following result for train set and test set:

Train Set:
Accuracy: 91%
Sensitivity: 85%
Specificity: 94%
Precision: 90%
Recall: 85%
F1 Score: 88%

Test Set:
Accuracy: 90%
Sensitivity: 85%
Specificity: 94%
Precision: 89%
Recall: 84%
F1 Score: 86%

85% of the values (actual converted) are predicted by the model
94% of the values (actual not converted) are predicted by the model