Summary

The goal of this assignment was to build a model through logistic regression for company X's marketing team. The marketing team would then use it for prioritizing the prospect customers who can actually buy their product in the future. To start with model building, first, we have to check and clean the data i.e. remove null values from data set if they are above threshold. Then we need to replace missing values with the most occurring/reliable value in the column to remove any inconsistency. Once data is cleaned, we can proceed for data analysis. We need to do univariate, multivariate analysis and plot graphs between variables to check useful insights.

Then, we further on to build our model. Here we are dividing our data into two parts namely train and test data. Train (70% of data) data is used to build model and test (remaining 30% of data) is used test model built on train data. If the number of features is more than 15 then it is safely assumed that top 15 only will have a major impact on the model's decision. So, with RFE with those 15 are selected and further GLM is applied where the target value is expected to be a linear combination of the features. All features which shows highly nonlinear behaviors (Higher 'P') are removed from data set as they won't have much effect on the lead score.

Since the collinearity is the state where two variables are highly correlated and contain similar information about the variance within a given dataset. These conditions can be checked with the vif factor. Once this is done, specificity and sensitivity should be checked to understand model built till now. Then we need to find the trade-off between sensitivity and specificity through ROC curve. If curve follows the left-hand border and then the top border of the ROC space, the more accurate the test, which exactly what we can see our case also. Further, we need to find cutoff value in converted probability where we can see most of the 0's turning to 1. From plotting accuracy sensitivity and specificity for various probabilities, we get cut off as 0.2. To proceed further we need to add lead score column to a data set which is nothing but selecting all rows with probability greater than 0.2 and multiplying and rounding off converted probability. Once we are done with 'train' data, we need to check and follow the same procedure for 'test' data to check the correctness of our model.

To conclude, our built model has an accuracy of around 90% which is really good. The company can use this model to contact prospective customers. Further, they can increase and decrease probability, sort and pick the top or lower lead scores if they want to contact more or less interested customers. That will depend on their business needs. In this case study, we learned about how to apply linear regression with its nuances in real life examples. This will help us to deal with future aspects of LR.