Lead Score Summary:

There are quite a few goals for this case study.

a. 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.

b. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

- We have first cleanse the raw data as dealt with all the missing values.
- Then univariate analysis is done to check the relationship between the target variables and other variables. And remove all the variables which are not showing any relationships.

- Then preparation of data takes place where we convert yes/no variables to binary variables and create dummy variable for other categorical variables.
- We will split the data afterwards so that we can first work on train data and when the model is finazied we can use same for test data and check accuracy and metrices beyond accuracy.
- Once model is build using RFE and all the unnecessary variables are deleted based on pvalue we will calculate VIF and check the values for all the variables and remove variable which have more that 10 as a vif value. (only n1 variable would get removed in one run)
- We will re iteratre this process till all the p values are less than 0.05 and vif less than 5.
- Once we get the final model we have plotted ROC curve to find the optimal threshold as the old threshold used was 0.5.Once we get the optimal threshold we can find all the others metrices like sensitivity, specificity, precision and recall values.

- We have use sklean and statsmodels library to build this model which has high sensitivity and less specificity which will result in covering almost all the true positives.
- We can change threshold value as per the client requirement. As mentioned in the goal of this case study the model should work even if the client requirement changes. For example: If client don’t want

to make unnecessary calls because the target is already achieved and it's the end of the quarter so we can simply increase the value of threshold which will result in increase of specificity. Increase in specificity will increase the true negatives in the model which will only show the smaller number of hot leads as more negatives conversion will result in the output.

- In this model the threshold use was 0.34.
- Overall accuracy of the model is also good hence we can consider that this model fits perfect with pvalue less than 0.05 and vif less than 5.