Summary Report on Lead Score Assignment

The provided data set was first analyzed and it was noticed that there were a lot of categorial columns with the value as 'Select'. These values were updated to Nulls as they were understood to be the cases were the dropdown option was not selected and the default populated value of 'Select' was stored in the database. A few columns in which all columns had the same value ('No'/'Yes') were dropped. After that the numerical values in the data set were analyzed and were imputed with their median values. The cleaned and column reduced dataset had very few missing categorical values in just 2 columns. This was considered acceptable.

Outlier analysis was then done for the numerical columns using box plots. No treatment was done as the outlier values were reasonable. Dummy variables were created for the categorical columns and then highly correlated columns (absolute value of correlation > 0.8) were dropped. The dataset was then split into train and test sets to create a model and test. The continuous features were then normalized and a linear regression model was built. Feature selection was then done through RFE and a regression model was built again using the highly ranked feature arrived at using RFE.

The RFE model was then built again iteratively by dropping columns that have a high p-value one per iteration. A final model was arrived at and the VIF for these columns were calculated and observed to be within acceptable ranges. The accuracy of the model was calculated to be around 80% using the train model. An ROC curve was drawn and the area under the curve was found to be 0.86.

The optimal cutoff probability for the logistic regression was calculated to be 0.36 using accuracy, sensitivity & specificity. Using precision and Recall, the optimum cutoff probability was plotted and found to be 0.415. An overall accuracy of 80% was received here too. The model was then applied to the test dataset and again we received an accuracy of 80%. As we've consistently hit 80% accuracy, our model seems to be a good one and the Lead scores were calculated from the logistic regression probabilities and the score was appended to the original dataset.

It was observed that Leads that originated from the Lead Add form contributed most to a high Lead score and for leads whose emails have bounced, the lead score seemed to reduce. Improving the Lead score can be done easily by following up with the leads who had their emails bounced and whose last activity was olark chat conversation. Moving out from this phase of the flow can improve the lead score drastically and increase the chances of getting more leads converted. It was observed that if the last activity was having a phone conversation the lead score was higher. So following up with the leads via phone calls can improve the chances of conversion a lot.