Summary

- The business requirement is to identify leads weather they will convert or not, So using Logistics regression would be best in the scenario.
- There are various attributes where the field is '**select**' , Replacing them with null will make the analysis Easier.
- There are various fields where 30% of nulls are present. Dropping all these values will be the option as replacing them with mean/median will create bias in the data.
- There are 5 attributes where values are unique (only one value). These attributes will not contribute to impact outcomes, so dropping these values will be better options.
- For values where missing value is less than 30%, imputing them with median (for continuous variables) and mode (for categorical variables).
- There are various attributes where 99% of values belongs to one segment,so dropping them as well as they are not useful for outcomes.
- For outlier treatment,I have clipped all outlier data to 95%ile to remove outliers.
- After 5 repeatation the final model has 80% of accuracy along with 80% each of specificity and sensitivity.
- Selecting 0.4 as cutoff frequency proves more optimum as it produced above results.
- The results of test set also produces similar output proves model is very stable.