Summary

The Lead Score Case Study aims at improving the lead conversion rate from 30% to 80% or more. And for the same we have gone through very obvious steps as –

1. **Data understanding:**
   The learning starts with very important step of understanding the data, as we at first focused on the importance of existence of a single column in it and tried to get the essence of it's meaning in terms of our problem statement.

2. **Data Cleaning, Manipulation and Structuring:**
   We are removing and imputing the null values based on the importance of the data and clubbing the similar features.

3. **Data Analysis:**
   - We just kept on increasing our data understanding with a vigorous EDA. We came to various patterns, data imbalances, outliers and about the redundancy of the various columns and with our understanding of the data, we tried cleaning, imputing it as per the columns' nature and importance.
   - We have not left a single stone unturned to get the trend from the data as performed Univariate, Bivariate, multivariate analysis (Heatmap for Pearson's Correlation Coefficient.)

4. **Preparing the data for modelling:**
   - Once we are satisfied with the readiness of the data for model, we created the dummy variable for the non-binary categorical variable.
   - Standardized the numeric columns to scale the whole data set.
   - After this we proceeded with the next step of splitting the data in the ratio of 7:3 for train and test respectively so that we can experiment on train data and validate on test data.

5. **Model Building for Logistic Regression:**
   - Once it is done we moved to building Logistic Regression Model with all feature and encountered many insignificant value.
   - Thus we have taken the help of RFE and selected top20 features and then manually dropped the further features till a stable model is obtained on the basis of p-value and VIF.
   - We just tried with a random cutoff about the conversion rate and tried to understand the metrics.
   - Now with the help of ROC Curves and Recall Precision Trade off we tried to figure out optimal cut off value for lead conversion.
   - Based on the business understanding, we should focus more on sensitivity, as it is the accuracy of predicting the positive class and our business is to increase the lead Conversion rate, thus we focus on increasing the Sensitivity.

6. **Prediction and Evaluation of the Test data:**
   - The same model preparation is done on the test data with the only difference that we do not fit but only transform the data and we tried to do a validation of our model on the test that how efficiently it is predicting the lead.
   - Finally, we came out with important features such as Lead Source, Last Notable Activity, Time spent on website, Lead Origin, etc. on which Lead