Lead Score Case Study's Summary Report

X Education company sells online courses to industry professionals. 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.

The steps followed to solution is as follows:

1. **Reading and Understanding the data**: We read the data and noted that there were null values in the data provided and also the outliers were found when we used the describe function
2. **Cleaning Data**: We began by checking for duplicate values and noted no duplicates. Dropped columns having more than 40% of null values and imputed the other values with either median or mode depending on whether it was numerical or categorical values. We also created new values where there were high percentage of missing values
3. **Exploratory Data Analysis**: Handled outliers in the numerical columns and dropped few columns which had only no values and didn't add any values to the data.
4. **Data Preparation (Dummy Variables):** We created dummy variables for the categorical variables.
5. **Train-Test-Split**: We then split the data set into train and test set with a chosen ratio of 70% for train and 30% for Test
6. **Rescaling the Feature**: We used Min Max Scalar method to scale the numerical features
7. **Model Building:** We then selected the top 15 variables using the RFE (Recursive Feature Selection) technique. We then built models depending on the variables selected by RFE and simultaneously checking for P-values and VIFs. We made sure to drop columns having P-values greater than 0.05 and VIFs which are more than 5. After building 4 models we finally got our final model with balanced VIFs and P-values
8. **Model Evaluation:** A confusion matrix was build and calculated the overall accuracy. We also calculated sensitivity and specificity to understand how reliable the model is
9. **Plotting the ROC Curve:** We then plotted the ROC Curve and the curve was pretty decent with an area coverage of 87%
10. **Finding Optimal Cut-Off Point:** We then plotted accuracy sensitivity and specificity for various probabilities. The intersecting point for considered as the Optimal probability cutoff point.
11. **Precision and Recall Values:** We then found the precision and recall values and noted that there is high rate of recall value when compared to Precision value which is ideal in this case. We also plotted precision and recall trade-off on trainset
12. **Making predictions on the test set**: We then made prediction on the test data which was 30% data and noted that our model was able to predict Sensitivity, Specificity, Accuracy, Precision and Recall percentages on a test set which was almost equal to train set. We then concluded our analysis by specifying top variables present in the data set