Summary For Lead Score Case Study

The approach for this project has been to divide the entire case study into various checkpoints to meet each of the sub-goals.

- Reading and Understanding the Data
- Data Cleansing.
- Data preparation and feature engineering.
- Applying Recursive feature elimination to identify the best performing subset of features for building the model.
- Building the model with features selected by RFE. Eliminate all features with high p-values and VIF values and finalize the model
- Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.
- Decide on the probability threshold value based on Optimal cutoff point and predict the dependent variable for the training data.
- Use the model for prediction on the test dataset and perform model evaluation for the test set.

The problem we faced in this case study and the solution are

1. **Feature Selection.** :We need to have a better business understanding to drop the feature which are having high null values and not necessary for our analysis.
2. **Outlier Treatment:** Depending upon the size of dataset we need to drop or impute the outliers for numerical variable.
3. **Manual/Automated feature selection :** we have not PCA for the feature selection instead we have used RFE for the feature selection as we have to describe the features that affects the target variable.
4. **Final Features For Mode:** we have taken 11 variables for our final model selection. This depends upon the no of features that are available for our analysis.
5. **Model Evaluation:** we have taken min 80% sensitivity with threshold of .28 as we wanted to cover all the positive cases that might occur. Again this depends upon the project requirement.
6. **NullValues Treatment:** To remove or to impute the null values for all the categorical and numerical variables depends upon the size of the data set. If the size is less then we can impute the dataset otherwise we can remove the null values from our analysis.