Summary

1. Data Understanding:
    a. Checked Null Values and imputed the Values with reasonable data Values
2. Data Cleaning:
    a. Few Columns have ‘Select’ value in the data and this as same as null
    b. Removed columns with 45% of null values
    c. Imputed null values with mode
    d. Dropped missing values in Categorical features
    e. Removed columns irrelevant for model building (Prospect_ID, Lead_Number and Last Notable_Activity)
3. EDA:
    a. performed a data imbalance check on the "Converted" target variable. Compared to records that were not converted (61.5%), it had less converted records (38.5%).
    b. carried out univariate and bivariate analysis to derive conclusions. To avoid managing several categories during modelling, categorical columns such as Lead Source and Specialisation were combined to form the "Others" category because they had different values with low counts.
4. Data Preparation:
    a. Prepare train and test data by splitting the data set into 70:30 ratio
    b. Scale the features using Standard scaler
5. Model Building:
    a. Created Logistic regression model with RFE for 15 features. And dropped the features having high p value and VIF. Final data have 10 features.
6. Model Validation:
    a. Checked Confusion matrix, Accuracy and other factors like ROC Curve
    b. Assigned Lead score on the training data
    c. Performed probability prediction
7. Making Prediction:
    a. Performed Scaling and predicted using final model
8. Model Evaluation:
    a. Created Confusion matrix and ROC curve
    b. Performed lead score Assignment on test data
    c. Top feature for prediction:
        i. Lead Origin_Lead Add Form
        ii. • Lead Source_Welingak Website
        iii. • Working Professional

Accuracy of the model is 80% which is very good to use in real word data.