Summary:

Lead Scoring Assignment is an ML project where we have to find the features that contribute most to the conversion rate. Initially, only 30% of calls were getting converted to leads but this was inefficient, hence with the help of logistic regression and EDA we were able to find some of the best features that we need to focus on which contribute a lot towards the conversion into leads and thus increase the chances of our accuracy and other metrics around 80%.
EDA and Data Cleaning were carried out on the given data set. Most of the values were dropped because of a large number of null values in those columns. Even univariant analysis was done to treat Outliers. Analysis like Univariant, Bivariant, and multivariate analysis was carried out. Scaling and one hot encoding is also done.
After getting the data into the most suitable format data was split into Training and Testing data in the ratio 70:30 and the model was built on the training set by carrying both automatic and manual (based on p-value and VIF Score) approach. Recursive Feature Elimination is used for the automated approach where we initially used 15 features. A further selection is done manually by looking at the multicollinearity and statistical significance of features and the overall fit of the model. The 12 most significant features are 'Total Time Spent on Website', 'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Welingak Website', 'Last Activity_Email Opened', 'Last Activity_Others', 'Last Activity_SMS Sent', 'Occupation_Working Professional', 'Specialization_Banking, Investment and Insurance', 'Specialization_Finance Management', 'Specialization_IT Projects Management', 'Specialization_Rural and Agribusiness'.

Note:
1. Data is cleaned wherever necessary
2. One hot encoding on categorical data.
3. EDA is carried out (univariant, bivariant, and multivariate analysis)
4. RFE is carried out to deduce the features to 15
5. Manual approach is used to deduce the features to 12
6. Functions are created wherever there is repetitive work.
7. Evaluation of the model is done.
8. Graph is plotted for accuracy, sensitivity, and specificity to get the point of merge where the threshold value results in a balanced output of various scores.
9. precision_recall_curve is used to confirm the same.
10. Accuracy, Sensitivity, Specificity, and Precision are close to 80% for training data set by only using 12 variables.
11. Out of 12 Features only Lead Origin_Landing Page Submission is impacting the model negatively.
12. Rest all the Features that impact negatively the model.
13. People spending more time on websites have more conversion rate
14. People with Working professionals and the ones who prefer the sms have a high conversion rate.
15. For Training data set for cutoff 0.331

Accuracy: 0.81
Sensitivity: 0.8
Specificity: 0.81
16. For the Test data set
Accuracy: 0.8
Sensitivity: 0.8
Specificity: 0.8
Since Accuracy, Sensitivity, and Specificity for both the training and test data are the same we can state that our Logistic Regression is performing well even on unseen data. Thus, meeting the business requirement.