Brief Summary for steps taken for the assignment & the learnings

Objective of the assignment – Create a logistic regression model to predict the lead conversion for each lead and decide the optimal cut off value for the probabilities of predictions for target variables 1 and 0.

Steps taken for the assignment –

1) Understanding the data set and then accordingly involved the steps for data cleaning i.e. removed columns who's having null values more than 60%, then imputed null values with median, mode and based on domain expertise.
2) Handled select category, as this was kept unchecked by many people. So changed it to null values for further analysis.
3) Checked for outliers and based on IQR range handled the outlier for few columns.
4) Then performed EDA, while performing EDA performed few data cleaning as well those columns which are unwanted for our analysis like found few of the columns had almost 98% of one category. So dropped those columns.
5) Then data preparation before passing the data for model building.
6) Data preparation like dealt with categorical variable. Created dummy variables for the categorical columns.
7) Then scaled the training data set which help the model to perform well. Used standard scaler for scaling.
8) Applied RFE (recursive feature elimination) method to select the best 15 independent variables to predict the target variables for model building.
9) Applied GLM algorithm from statsmodel library for logistic regression.
10) Built the model with the selected top 15 RFE variables.
11) Dropped the insignificant variables which had high p-value and high VIF.
12) After multiple model building (removing variables high p-value & high VIF) landed with the final model.
13) Performed model evaluation i.e. various metrics like Accuracy, TPR, FPR, Sensitivity, Specificity, Precision and Recall.

14) Then calculated the optimal cut off threshold point on training data for the probabilities to predict the target variables.

15) Applied AUC & ROC curve to determine the optimal threshold value for the model.

16) Then proceed with prediction on test data and performed model evaluation like various metrics which performed on training data.

17) Checked the accuracy and confusion metrics for both training and test data set.

18) Calculated the lead score on probabilities of predictions (on predictions of training & test data) for the overall data set based on the problem statement.

19) Assigned the lead score table to the overall final data set to help the management to get the most likely converted leads based on lead score.

<u>Learning Gathered</u> –

1) The business focus is towards the most likely converted leads to target the potential leads.
2) So, for that business needs to focus on the sensitivity, so that the potential leads should not be ignored.
3) So as per the business requirement the optimum threshold values selected was 0.3 where we found that the accuracy is 0.84, sensitivity is 0.86 and specificity is 0.83 which is best for the model.
4) While model building, we found that top 3 variables that contribute most towards the probability of a lead getting converted are:
    a) Lead Source_Welingak Website
    b) Lead Orogin_Lead Add Form
    c) What is your current occupation_Working Professional
