Executive summary on lead scoring assignment

X Education provides online courses to various parties including working professionals, students etc. X education wants to understand the predictors for converting a lead. A logistic regression model is used to handle this problem. Below is a summary of the overall approach to this.

1. There were 9240 data rows with 37 columns.
2. Data were checked whether the types of data were in order or any amendment to be made. However, all data types were in correct format.
3. Data cleaning was initiated to see what actions to be taken against missing values. All missing values more than 40% were removed for further analysis. Other missing values were imputed based on mode or created a new field as others depend on the case.
4. For columns having multiple levels, the levels were reduced by clubbing together.
5. EDA was carried out by plotting graphs and then checked for any outliers. There were several fields with outliers which were removed for further analysis.
6. Then data were transformed prior to the logistics regression. Data with Yes and No were replaced with 1 and 0. All dummy variables were created for categorical variables.
7. After all adjustments, there were 8924 rows with 66 columns.
8. Then the data were split to train and test data sets. Numerical columns were standardized.
9. A RFE model was used with 15 columns to choose to see best predictor variables. Based on the p-value and VIF values insignificant columns were removed from the dataset. Final model had 11 columns.
10. Then model was evaluated based on accuracy, sensitivity and specificity. A ROC curve was drawn to see whether the model is good. AUC came to 93%.
11. The model used a cut-off value of 0.5. To see the best cut-off point the intersection of the 3 lines were used. The new cut-off value came to 0.34
12. Based on this cut-off value, lead scores were assigned to train data.
13. Predictions were made using the test data set. The same cut-off 0.34 was used. Model was evaluated using accuracy, sensitivity and specificity.

    Train data:

    Accuracy - 0.86 Sensitivity: 0.79 Specificity: 0.91

    Test data:

    Accuracy - 0.86 Sensitivity: 0.84 Specificity: 0.87

14. Seems the model working well on the test data as well. The top 3 predictor variables came out to be
    - Total time spent on website
    - Will revert after reading the email
    - Tags_others