Lead Scoring case study

We followed the below steps to build logistic model:
Step1. We started with data understanding using pandas inbuilt function like info(), describer(), shape(). This helped to understand the nature of the data, columns having null values and to identify categorical columns
Step2: We took the next step of handling the NULL values. So we dropped the columns with high NULL values as they will not impact the analysis/decision factors much
Step3: Next, we identified the skewed column(having high asymmetry) based on value_counts() and removed as they won't be useful and may impact the model by giving biased design
Step4: Created dummy variables for all the categorical variables. Converted the columns which had yes/no to corresponding 0-1 values.
Step5: We used the IQR method to identify the outlier and considered only 0.01 to 0.99 percentile of values
Step6: As a next step, we split the data into train and test for model building
Step7: Scaled the data using sklearn library of python to balance the impact of all variables
Step8: We built the first model by using stats model and analyzed the p-vales for each variable.
Step9: Since the number of variables was very high, we used RFE method to eliminate less important variables and used only 15 to build variables.
Step10: Then we analyzed the stats model for these 15 columns and removed the variables which had high p-value. We dropped 2 columns one by one having high p-value
Using the coefficients, we identified key features which were highly impactful in identifying the criteria impacting conversion.
Step11: VIF score for all the remaining columns was below 5 so there was no need to drop any extra column.
Step12: Then we calculated the probability of lead conversion by considering the 0.5 value as cut off for lead conversion.
Step13: Next, we checked the overall accuracy of the model which came out as 0.8113, which was good and acceptable. The other matrix like specificity gave good result
Step14: We then tried to plot ROC curve to get correct value of cut off. From the ROC we found the ideal cut off was 3.5.
Step15: The different matrixes generated gave us the accuracy of 0.80408 and sensitivity of 0.8103 which looks promising.
Step16: Then we ran the model on test data which gave accuracy of 0.801 and sensitivity of 0.8075.

Learning from this analysis:
We understood the important aspect of logistic regression model by developing it with data.
Use of logistic regression helped to build the model for categorical outcomes. For eg: Here we wanted to create model for 'convert' and 'no convert'.
Additionally, we learnt the use of different metrices like confusion, sensitivity, specificity to gain important insight from data . With this exercise, we were able to identify the features to increase the conversion rate of possible leads