Summary of Lead Score Case Study

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. Although X Education gets a lot of leads, its lead conversion rate is very poor so CEO asks to concentrate on hot leads which will get converted to their Student

Approach followed:

1. Understanding the data and reading the data using Pandas dataframe
2. Check the data types and missing values in the data
3. After seeing the data, found that lot of values are given as select , which doesn't have any meaning so replaced all select values with NULL
4. Verified if any unique values in the columns, if the value is unique it is not going to give any idea to us so dropped those columns
5. Identified missing values in each of the column and column with 70% missing values and columns with name Asymmetrique are dropped
6. Filled with not available for the missing values so that the model built is accurate instead of filling with any other values
7. Pair plots are drawn for numerical values to see the total visits, total time spent on website and page views per visit
8. Did Univariate Analysis and Bivariate Analysis to see how the lead conversion happened based on the city , Lead origin, Tags, Last Notable activity, specialization and Lead source
9. Created dummy variables for Categorical variables
10. Split the data by 70% Train and 30% Test with the target variable as Converted
11. Using logistic regression, started building the model, selected columns by RFE and calculated VIF
12. Dropped the columns(What is your current occupation_Unemployed & Tags_Will revert after reading the email) with high VIF , as they are highly correlated and accuracy of the model might get impacted
13. Started with prediction and model evaluation using confusion matrix, sensitivity and specificity
14. Prepared ROC curve and optimal cut off curve to see the area under the curve, area seems to 0.96 hence decided model is accurate
15. After doing all the above steps, Overall accuracy came out to be 86.15%
16. Hence concluded below points after doing the analysis:
    a. Total time spent on website as well as page views per visit has higher lead rate
    b. When the last activity is SMS, Olark Chart conversation lead rate is high
    c. People from Mumbai with occupation as working professional and Tags_ Will revert after reading the email
    d. If the lead source is Google, probability is high
    e. Lead quality with value as might be also achieving good lead