Identify Hot Leads for Company_X

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.
The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead.
The problem we have to solve by our model is to increase Hot Leads by providing leads having a high rate of conversion.
We start our model by importing data and performing EDA on this.

- The first step is to drop variables having more than 40% nulls since imputing with mean, median or mode might bias the model we have
- Next, we will identify categorical columns having more variability and see how to treat them so that it will be easy for our analysis. One example is grouping Tags column into 4 categories so it's easy to interpret the model, instead of looking at 10 different tags for the same column
- Handling missing data is where we see different techniques for categorical and numerical columns. Since we have select as one of the categories in many columns we imputed this with the new category. For continuous variables, outliers were few and it constituted close to 1% of data, so capping them wouldn't hurt our model

Once the above steps are taken care, we have our data ready to go ahead with modeling steps.

- Here we start modeling by creating dummy variables, removing highly correlated variables based on VIF and correlation matrix.
- We need not drop any variables if VIF is good and not crossing threshold (say 5).
- After this, if we plot ROC, we see the area under the curve is more representing model has high True positive rate
- To find the optimal cut-off point we plot accuracy, sensitivity, and specificity to see the point where all three meet.

Calculate the Sensitivity and Specificity of the model to look at the predictive power for True and False cases. Predict a test set and then see the predictive power again on the test set to see how well the model has trained.

Conclusion:

Through this exercise we get to know important factors such as time spent on the website, reverting to emails and place we live in effects a lot for our model. Once we have a Business understanding of the importance of variables, we can again tune this model for better results. For now if we look at importance of variable predicted by our model, we can say it’s closer to real life situation and can easily address the problem we have here.
Result we got from our model is people who live in metropolitan cities like Mumbai and who are not student are the ones looking to enroll to this program which makes sense. By looking at this we can say the model we built is pretty good and can predict Hot Leads for more than 85% of the time.