Summary

1. The given dataset had a few columns that were from the sales team, as we are making a model to predict the sales team, who are good leads, so at that stage, this data does not exist (as this data is added after approaching), so we did not include such features (e.g., Tags)

2. The dataset had a lot of yes/no type columns, which look like a great thing to have, especially the columns related to the advertisement, whether the customer has seen the ad on one platform or other, but it came out to be useless information as most of them were too skewed, (99%+ answer being NO), so we dropped such columns as it does not add any 'learning' to us predictive model.

3. The geographic location of the customer can be a good indicator to target the customers, if there are a huge number of customers from the USA, we can treat them separately and service them according to their time zone. We checked such cases and found out that the majority of customers are from India only, and we further dug down and checked whether cities match with the countries. The results were not very good, there were a lot of customers who had put the cities, that were not even part of their native country, so we dropped the idea of doing imputations and simply made 3 categories of city columns for the simplicity of the model.

4. Used semi-automated or mixed approach to model building, tried different input combinations of techniques and found out that RFE is going to be the coarse tuning method, in that I also, tried 15,20,25 limit and found out that 20 is the sweet spot, as it does not delete some of the most crucial features from a business suggestion perspective. The Statsmodel is used for further fine-tuning as it has a better statistical summary and it is easy to make better decisions about which features to keep and which not.

5. After juggling through the VIF and p-value confirmation we finally came up with the best 18 features, and the model evaluation was done for the train data, the parameters were good enough but the sensitivity was not on par, (62%) which is not something, that decreased the threshold from 0.5 to 0.4 to 0.3 and found out that 0.3 or 30% gives the best results (ROC curve 82% area, accuracy ~77%, sensitivity ~75% etc).

   These were the major hurdles and learnings from the case study.