Summary

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.

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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Let's see how we can handle this scenario

Firstly, lets see how to prepare the dataset for analysis and draw inferences from it to make a better business decision.

- First step to import all the necessary libraries like pandas, numpy, seaborn matplotlib etc. Rest can be imported as in when required.
- Read and understand the data
- Do the EDA for model building
- Create dummy variables and concat with the original dataset for further analysis
- Divide the dataset into train and test with 70:30 ratio
- Start building models
- Check for high VIF and P value variables and start dropping them one by one
- Once the VIF value and the P values are less than 4 and 0.5 respectively consider it as final model
- Calculate the Accuracy, specificity and sensitivity values for the training and test set
- Make sure that for both training and test set the values for above mentioned parameters should be at least 80% as this is the ball park set by CEO
- Visualize an ROC for better understanding
- Need to find the optimal threshold cutoff as an arbitrary cutoff is considered at the beginning of the model building
- Once it is done plot the thresholds points to finalise on the optimal point
- Model evaluation is to be done post this step using confusion matrix and calculating the accuracy, sensitivity and specificity
- We also need to check the precision and recall values
- Precision means percentage of results which are relevant and Recall means percentage of total relevant results correctly classified by the algorithm
- In the last few step, we have to check for the model evaluation on the test dataset and check for the model effectiveness on test dataset.
- Need to compare both and see how the model is doing.

- If the difference between the training set and test set is minimal it is said to be a good model built.

Learnings gathered from the assignment are as follows:

- The company should focus more on calling leads from reference and welingak website as they have a high chance of converting
- Since this course is for professionals, they should call working professionals on priority
- The company should call leads who spend more time on their website so that they don't go to some other website
- The company should contact leads whose last activity was SMS Sent
- The company need not focus on Olark chat conversation as they are less likely to convert
- The company need not make calls to leads whose origin is Landing page submission as they might not convert
- Leads with specialization as Others need not to be contacted as they are not likely to get converted
- Leads who have selected 'yes' for Do not email need not to be contacted as they don't want to be contacted.