First, we need to define the business problem and understand the analytical objectives. Company X wants to increase conversion rates and reduce wasted calls because they are not focused on the most potential customers. The solution to that is to build a logistics model and calculate the lead score of each customer, and the target conversion rate is 80%.

Step 1: Data Cleaning and Handling:

We need to clean the data by removing the variable if it contains more than 45% missing values, impute missing values, and 'select' values by appropriate values and outliers to make the model more accurate. With numerical variables, we impute outliers by the 99% percentile.

Step 2: Exploratory Data Analysis:

This is an important step to understand our variables and deliver valuable insights to the company. Moreover, we can define which variables should be removed from the dataset as they do not contribute to our model.

Step 3: Data Preparation:

In this step, we first drop the variables we mentioned in the EDA step to narrow down the important variables. Based on the remaining variables, we will create dummy variables from categorical variables, convert Yes/No to 1/0, and scale the numberical variables. Checking correlation will help us remove the variable between high-correlation variables, but in this case, we can keep all of them to process the next step. Finally, we split the training set and test set to build the logistic regression model.

Step 4: Model Building

As we have too many variables, it is better when we use RFE to keep important variables with n_features_selection = 15. After that, we can easily apply manual feature selection and the VIF method to remove variables with a high p-value and a high VIF. Based on the final model, we can predict the probability of each lead, assigning 1 if the probability is higher than 0.5, otherwise 0. However, this threshold is not as good as possible because the sensitivity is lower than we expected. The area under the curve is equal to 0.89, which indicates our model is good; we just need to find the optimal cut-off point. Both methods, including sensivitity-speciality and precision-recall, are applied, and we finally choose the cut-off point of 0.37 to meet the CEO’s expectation. Also, based on the probability, we can assign a lead score to each customer.

Step 5: Model Evaluation:

Applying our final model to a test set, we evaluate the accuracy, specificity, and sensibility of the test set. The values of them are close to those of the train set and as good as we expected.

The learnings:

- Following the assignment, I can understand how to solve a business problem step by step.
- The objective and expectation of analysis are crucial when we build and evaluate the model.
- Through the EDA step, much of the insight that company need can be given to improve performance.