Lead Score Assignment Summary

1. **Understanding the business objective:**

   First and foremost we have to gather the requirements from the business team, about what needs to be predicted, based on the inputs provided we can decide on which ML algorithms are most suitable for our Prediction analysis. As for the Lead assignment we are given with the dataset and data dictionary files.

   We have to build a ML model for an An Education company selling online courses to professionals. The company gets leads in two ways. One way is Professionals landing on the company website who either fill up the form with email and phone and watch relevant videos. The other way is through referrals. The company receives lots of leads but the lead conversion rate is very poor (30%). The objective is to identify the hot leads who are likely to be converted, so the lead conversion rate gets better by building a classification model (logistic regression model) and assign a lead score. Higher the lead score, higher the chances of it being a hot lead. The target lead score conversion rate must be around 80%.

2. **Data Cleaning:**

   First try to find if we have any missing rows or unwanted rows like header rows and footer rows, summary total row or subtotal rows. Check for any column descriptor rows in the data set, if any exists drop such rows. Then we have to deal with missing column names. If any we have to rename with an appropriate column name, or any columns that we feel like are not important for analysis we can drop such columns. Check if columns can be splitted into multiple columns for better analysis. Or multiple columns that can be merged into one for better analysis. ALso if any missing values we have to either input them or we can remove columns with missing values or we can remove records that have missing values.

   In the case of the leads dataset, there is no need in treating missing columns and rows as everything looks fine. But we see a lot of missing values in the dataset and on top of that we see Select value in a few columns which has no meaning in our analysis hence we can treat them as missing values or NaN values. So we drop columns whose missing values percentage is close to around 30%. After dropping with such columns we are left with around 23 columns

3. **Data Quality Check:**

Here we have to find if there are any outliers that exist within the numerical continuous variables. If any outliers exist then we have to thoroughly analyze whether the outliers are valid and well within the range or the outliers are falling continuously. If not then we have dealt with them by either removing such rows or imputing such values. Also check whether the data present in the dataset are of correct data types else we need to correct them to correct data types.

In the case of the Leads dataset, we see 3 columns which are continuous numerical variables. Out of which one have no outliers and the other two have outliers but they are well within the limit and cannot be treated as outliers. APart from that we don't see any incorrect data types within the columns.

4. **EDA analysis:**

Here we can do two types of analysis, one is univariate analysis, bivariate analysis and multivariate analysis, in univariate analysis we analyze a single column and in bivariate we analyze one column on X-axis and other column on y-axis and in multivariate analysis we analyze two or more columns and interpret the insights. Once we have the insights we draw the conclusions for analyzed columns in the dataset

In the case of the Leads dataset, we have done univariate analysis on three columns and bivariate analysis on 5 columns with the target variable “Converted”. Based on this, these were the insights we can draw from it. The majority of the leads have originated from landing page submission and API. If 92% of the leads have opted for No to "Do not Email". Email opened (37.62%) and SMS sent (30%) contribute to the majority of 'Last Activity' by the leads. If the lead originates from an API source then they are most likely to be converted. If the lead is sourced from Olark chat or Organic Search then they are most likely to be converted. if the lead didn't opt for No to Email, they are most likely to be converted. If the lead is unemployed, then they are more likely to be converted. if the lead didn't opt for A free copy of Mastering The Interview, they are most likely to be converted

5. **Model Building:**

We are going to build a logistic regression model, for that first we need to preprocess the data, we have to convert all categorical columns into dummy variables. Once we have the dummy variables we can remove the actual categorical columns and append the dummy columns to the actual dataset. Once it is done we have to split the dataset into a train set and a test set with 7:3 split. Then we build a logistic regression model with the train set and then predict the results with the test set. We can compare the accuracy of the results for both train set and test set and check for confusion matrix to find the false

positives and false negatives, if we see a huge difference then we have to perform RFE analysis and perform feature reduction and rebuild the model, and compare accuracies, confusion matrix etc. if we don't see much improvement, we can go further on reducing the feature variables until we get the desired output

In the case of the Leads dataset, the first model has better accuracy with accuracy of 81% for train data and 79.73% for test data when compared to second model with accuracy of 76.44% for train data and 75.93% for test data. The first model has less False Positives (FP) 387 when compared to the second model which has False Positives (FP) of 472. The first model and second model has almost very similar False Negatives (FN).