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 starts 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%.

You have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

To Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Analysis Approach:

1.Loading the dataset and all the necessary libraries.
2.Data Cleaning/Preparation:
    2.1 Checking for missing values. If present , dropping columns that have missing value percentage greater than 45%.
    2.2 Plotting graphs for the columns that have missing value percentage less than 45% to check how to impute the missing values.
    2.3 'Lead quality' column had more than 50% missing value percentage but did not drop it because it can tell us more about a lead.

2.4 Many of the categorical variables have a level called 'Select' that is as good as a null value. So replacing all 'Select' with Null values and then checking for missing value percentage again.If greater than 45% , the column is dropped.

3.Data Analysis:

3.1 Comparing all the columns with the Converted column:

3.2 All columns that had unique values were dropped as not much information could be derived from them.(except ProspectID).

3.3 Plotted graphs for all the columns to see the relationship between the column levels and the converted column.

3.4 After plotting for all columns some more columns have been dropped as they were providing very less information.(Some columns had high amount of outliers).

4.Creating dummy variables for the categorical and mapping Yes/No to 1/0.

5.Splitting the data into training set and test set.

6.Applying Scaling function to scale numerical column.

7.Building the Logistic Regression Model.

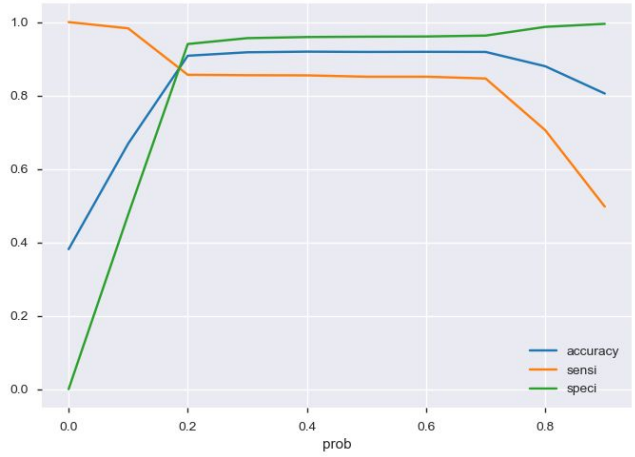
8.Applying recursive feature elimination(RFE) to identify the best variables.

9.Dropping all features that have p-value greater than 0.05.

10.When all features have p-value less than 0.05,proceeding to model evaluation.

11.Calculating Model evaluation parameters such as Accuracy,Sensitivity,Specificity,Precision and recall.

12.Finding Optimal Cutoff Point by plotting accuracy sensitivity and specificity for various probabilities.



13. From the curve above, 0.18 is the optimum point to take it as a cutoff probability.

14. Making predictions on the test set.

15.Using the probability threshold value of 0.18 on the test dataset to predict if a lead will convert.

16.Added a Lead Score column to the final model by multiplying the conversion probability with 100.

17.Evaluting the final model by calculating accuracy,specificity,sensitivity,precision and recall.