Summary Report of Lead Scoring Case Study

In this case study the main objective is to build a model based on Logistic Regression to predict Leads who essentially are visitors of a website of an education company which provides education online, name X Education. Their problem is that they have Leads (the visitors of the website). They come and visit their website from different platform such as Google, Organic Search, through past referrals, through social media marketing where they market their offerings or land directly on the company's website to look what courses they are offering and related queries. Once these people land on the website, they might browse the courses or fill up a form for the course

When these people fill up a form providing their email address or phone number, they are classified to be a lead. 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. Only 30% of the total visitors are enrolling themselves. So, X Education wants to know how to detect the promising leads so that they can target and convert leads into a paying customer.

Model Building: -

We built a logistic regression model to solve the business problem of the company. We built a model on the dataset they have provided to us. We built our model in following steps: -

1. **Reading and Understanding the data: -** The dataset contains more than 9000 datapoints spreading across 37 features. The dataset was not a cleaned dataset and a lot of missing values were present. We identified those columns where null values are present.
2. **Missing Value Treatment:** - We tried to omit missing value to clean the data. We tried by dropping columns, by dropping rows and also by dropping subset of the data. But in each process, we lost high amount of data. So, impute all the categorical variable by their mode or some suitable value and numerical values by their mean.
3. **Visualizing Data: -** We also visualize each column to understand the distribution of data. By visualization it was easier to understand. Also, visuals indicate which value is good to impute missing part of the column.
4. **Exploratory Data Analysis:** - We performed EDA on each of the column to get the idea of importance of variables in building the model. We also observed that how these variables were converted.
5. **Outlier Treatment:** - Some outliers were also there in the data. We also treated them.
6. **Dropping irrelevant column: -** We also dropped some columns which looked irrelevant in building the model.
7. **Conversion of data:** - We convert Yes/No columns into 1/0 and created dummies for other non-numeric columns. Since, a numerical dataframe is required to build a regression model.
8. **Splitting the Data:** - We split data into train and test set in ratio of 70:30. 70% of the data for training and 30% for test.

9. **Feature Scaling:** - In the next step we scaled the data to bring all the continuous variables on to the same comparable scale.
10. **Model Building –** Model was built on the training set using logistic regression.

11. **Feature Selection –**
    a. Features were selected based on RFE by removing the ones having P-score of more than 0.5.
    b. VIF score was also checked to eliminate features with score greater than 5.
12. **Accuracy check –**
    a. Prediction was made upon cutoff of 0.5 and accuracy came out to be 0.923.
    b. ROC curve was plotted and we could see the curve followed the left-hand border and then the top border of the ROC space very closely and hence proving high accuracy.
13. **Cutoff calculation –**
    a. Accuracy, sensitivity and specificity were calculated for various cutoffs.
    b. Plotting them together gave us the optimal cutoff point of 0.2.
14. **Precision and Recall –**
    a. Precision and recall values were calculated with recall value being more than 80%, the model was serving its purpose.
15. **Testing the model –**
    a. The model was tested on the dataset split earlier and the recall was again greater than 80%.
16. **Final conversion rate –**
    a. The overall dataset was again split based on our model predicting a successful conversion to the actual result, the conversion rate was calculated at 88%, on par with the expected by Mr. CEO.