Summary Report

In this case study we have performed the several processes to build the logistic regression model for the X Education company. The main objective of this analysis and model building for the X Education company is, the X Education company wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. To identify such a customer, they will convert into a paying customer. We have done the following steps; the steps are explained in the detail.

In the first step, we have understood the data values and the variables of the dataset. After this we have performed the **data cleaning** process. In this process we have identified, many variables are consisting of the null, duplicate, meaning less values. In addition, in the dataset some variables are imbalanced. Some variables are not making much meaning in the dataset, if we keep such a variable in the dataset then these variables are affecting our model accuracy rate, due to this we have removed these variables from the dataset. The variables are namely Asymmetrique Activity, Asymmetrique Activity Score, Asymmetrique Profile Score, Country, Lead Profile, Tags, Lead Quality, 'How did you hear about X Education, City, etc.

In the second step, we have performed the Exploratory Data Analysis (**EDA**) Process. While performing this analysis we have illustrated several insights from the dataset. Also, this step helped us to understand the dataset variables very clearly and the relationships between each other. In addition, we have identified which variables are more important to build a very effective model. By keeping in mind that the X Education main objective is to convert the **80% of leads** that visited to the company into the paying customer.

Some numerical variables consisted of very high values as compared to their respective means. So, we have created the graphical chart by using boxplot to illustrate this pattern. In this we observed that these variables are having very high **outliers** and needed to be treated. Hence, we retained 99% quantile of data and removed the max values.

In the third step, we were clear that the data is now ready for modeling, but all the variables were not the numerical, and we needed to convert such variables into the numeric values to build out the logistic regression model. For that we created the dummy variables and dropped the extra non meaningful variables from the dataset. Now the data was converted into the numerical form. So, we have split the dataset into the train and test data frames 70%, 30% respectively.

In this fourth step, before starting the **model building** process. We have checked the correlation between the variables with the help of a heat map. And it was very difficult to understand the correlation between the variables. Due to this we have performed the Recursive Feature Elimination (**RFE**) method to take the top 15 relative variables from the train dataset to train the model. With this we started the model building process. We iterated through 3 models, and evaluated each model based on p-value and VIF. Some variables had a very high "**p value**" or "**VIF**". We had to remove such variables from the model and rebuild the model again. After doing the same process twice, we observed that the third model has variables with low P and VIF values, and that is very good for our model.

Once we were happy with our model, we performed predicting converted leads followed by the model evaluation. In the model evaluation we have observed that model giving 74% of accuracy score.

To better understand the effectiveness of the model and optimize the cut off we plotted the **ROC Curve** and calculated the Precision - Recall metrics on the train dataset. In the ROC curve, we observed that the model generated has 87 % **AUC** which is a decent value for an effective and successful model. We also checked the **sensitivity**, **specificity** and **accuracy** of the model, which were 74%, 83%, and 80% respectively. We plotted the different probabilities of our model [.1 to .9] against accuracy, sensitivity and specificity and got the cut off at **0.35** by taking this threshold we have created our final prediction on the conversion probability of the customer towered by the X Education Company. Also, we have checked the precision, recall for the model, and the model gives the 88%, 35% respectively. When we plotted this we got a cut off for 0.4 but we decided not to use it as recall was only 35%.

In the very final step we have made the prediction on the test dataset. In this, the model gives the sensitivity, specificity and accuracy of the model 74%, 83%, 81% respectively. We have also assigned a lead score for each lead in the test data set which can be leveraged to identify hot leads in changing business environments.

The conclusion of this report, the X Education company has to focus mostly on the following variables to achieve the 80% lead conversion rate toward their company. These variables have a high potential to understand the customer profile and whether that customer will potentially buy the courses from the company or not. Following are the variables are:

1. Total time Spent on website.
2. Wherever the Lead Origin was:
    - Add format
3. Whenever the Lead Source was:
    - Direct Traffic
    - Google
    - Organic Search
    - Referral Sites
    - Welingak Website
4. When last notable activity:
    - Olark Chat Conversation
    - Unreachable
5. When the email does not send yes.
6. When the Last Activity:
    - Had a Phone Conversation
    - SMS Sent
7. When the customer current occupation is working professional

END