**SUMMARY**

**OVERVIEW:**

We have done building a Machine learning model and Data Analysis for X Education Company to identify the ‘Hot Leads’ as well as to find more ways to improve potential lead conversion rate. We were provided with a dataset having information on the leads (both converted and non-converted) and their behavior before enrolling for one or more of the online courses.

We have followed few steps to do that, and they are:-

**a) Data Reading and Understanding** – Here, we looked at the data, its rows and columns, and understood our requirements.

**b) Data Cleaning** – Here, we checked for the missing values in the dataset and either imputed them with valid information or removed them if they were irrelevant. Also, we categorized the records into fewer levels where needed.

**c) Exploratory Data Analysis** – In this step, we did univariate and bivariate analysis on the features and found out the correlation between the variables using a heat map. We, also, identified if some of the features have outliers in it, and treated them. (Depending on the business scenario and requirements.)

**d) Data Preparation** – We performed outlier treatment, created dummy variables for the categorical variables, used Standard Scaler to scaling them and we also split the data set into Train and Test Dataset.

**e) Model Building** – At this stage, we used RFE to filter the 15 most important and relevant variables from 39 variables from the dataset. Then we created models, checked their P-value, VIF score, then delete the features with high p-value and high VIF score (>5) and repeat these steps until we reached a position where all these values looked stable and good.

**f) Model Evaluation** - We calculated evaluation metrics like Accuracy, Sensitivity, and Specificity etc. on train dataset first and then verified them on ROC curve. When everything looked good, we followed the same procedure on test data set again and evaluate all the metrics on the test dataset.

**OUR FINDINGS:**

When we promote our online education program via many ad campaigns, search engine optimization, blogging etc. we end up with a huge number of leads. However, **only 38% of the whole dataset have 'Converted'** meaning, only 38% of the total leads have enrolled with one or more of the online education programs.

We, also, saw there are customers with various kind of occupations, for example - Working professionals, Unemployed, Students etc. who have enrolled for our program. Among those conversions, **most of our leads reside in Mumbai and are Unemployed**. Furthermore, they **found out** about the program **via Google ads and opted for specialization which falls under 'Management' category**. In addition to that, we also noticed the potential leads who **spend maximum time on the website had a high rate of conversion**.

We would suggest, to make the Company's **social media presence more prominent** and keep working on the **website content** which would result in **higher engagement**. Also, the company could **assign high performing sales team** to be proactive on **Olark Chat** so that we can improve potential lead conversion on that platform as well. We can also, work on **targeting audience residing outside Maharashtra**, as company provides online education program and is suitable for everyone who wants to up skill in their career path.

Now, coming to our machine building and prediction part, if we compare the values obtained for Train & Test:

**Train Data:** Accuracy: 80% Sensitivity: 67% Specificity: 88% Precision: 67% Recall: 82%

**Test Data:** Accuracy: 77% Sensitivity: 83% Specificity: 74% Precision: 65% Recall: 83%

The **score of Accuracy, Sensitivity and Accuracy in Train and Test data doesn't differ much**. Which means our **model can predict the conversion rate pretty decently** and we should be able to **give the CEO confidenc**e in making good calls based on this model's **'Hot Lead' prediction**.

**Submitted by:** **Anugata Mahapatra** and **Bhavya Joshi**