SUMMARY

Submitted by:

1. **Sudhanshu Singh**
2. **Sweta Singh**

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 start 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%.

Although X Education gets a lot of leads, its lead conversion rate is very poor. X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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.

Our Whole solving strategy involves the following steps.

• **Importing and cleaning data** - We first need to clean and prepare data. We imported the dataset and do basic checks on our dataset, checking the dataset for the amount of NaNs present. Also we would drop the columns with more than 30% NaNs.

• **Creating Dummy Variables-** We had to deal with the categorical feature in this case we are using dummy columns so that we can pass this created columns in our model.

• **Splitting the data into train and test**- We split the dataset into train and test dataset so that whatever model we build on train dataset we can test it on our test dataset.

- **Scaling the data**- Now for scaling the data, we applied ‘MinMaxScaler’ to the data.

- **Perform RFE & GLM to the test data**- Now we proceed with Feature Selection using RFE. After one iteration we get descent P-values and VIF values are also in control.

- **Selecting the cutoff**- We tested metrics that includes Confusion Metrics, Sensitivity Specificity, Precision and Recall. For our model we have selected 0.4 as our cutoff.

- **Apply the learning to test dataset-** After applying all the leanings on the test data our accuracy has decreased but our main target is to increase the sensitivity .Our model is doing quite a good job to identifying around 80% of the hot leads.