LEAD SCORE: CASE STUDY

An **ed-tech company** by the name 'X-education' wants to assign a lead score to each of its leads based on the data available. The data is populated from the customer, as well as the sales team. Customers put in a lot of details about themselves, however, only some details indicate a hot lead. The primary goal of this case study is to find the lead score and find the pattern in the data which suggests the most important details available in our dataset.

In this case study, we implemented **logistic regression** to classify if a customer can be a hot lead or not. The CEO of X-Education has suggested that the conversion rate needs to be **around 80%.** Since, we cannot take the risk of predicting potential leads as non-potential, we decided to reduce the false negatives. Thus, the primary metric to be considered was sensitivity/recall.

We had **9240 observations** and **37 features** in the dataset. The dataset had features in both categorical and numerical form. Pre-processing was required on the data where it was supposed to be cleaned and then prepared for modelling.

During data cleaning, the most important steps that were taken are:

- Columns had a category 'Select' which was basically telling that the particular field was left blank. Thus, these categories were replaced with NULL values.
- Categories which had more than **39%** missing/NULL values were dropped.
- Columns where the categories had **below 5%** (on an average) observations were clubbed into a common category '**Others**'.
- Features with binary categories were replaced with numerical elements – 1 and 0.
- Numerical features were checked for outliers and **top 0.5** percentile values were dropped.
- Sales columns were dropped after performing EDA.
- The dataset was divided into training and testing for robust validation of our model.
- Numerical features were then scaled between 1 and 0 using MinMaxScaler.
- Categorical columns were then converted to dummy columns for each of the categories.

With the above steps, the dataset was ready for modelling. RFE was then employed to decide the **top 15 features**. By using these features, we performed manual selection to understand the significance of the features and it's multicollinearity, by using P-Values and VIF-values. This was done with the help of Generalized Linear Models API from Statsmodels library.

The quality of the model was assessed with the help of ROC-AUC Score, **0.87** score indicated that the model had good performance. Then 2 different plots were used to decide the most optimal cut-off threshold, which were sensitivity-specificity-accuracy and precision-recall plots.

Using the threshold of **0.32**, predictions were made using test set. The metrics were compared for both train and test set, since the scores were almost equal it was conclusive evidence that the model was able to generalize well. Then using this model 'Lead Scores' were generated, by predicting probability and multiplying with 100, for all the leads. Based on the model's summary, the final recommendations were also provided to business.