Summary Report

X Education, an education company which wants to sell its online course saw that the conversion rate for the leads it received is as low as around 30%. This is considered low by the company and therefore, they wish to identify ‘Hot Leads’. ‘Hot Leads’ refer to the leads which have features that would suggest that there is a high chance of converting these leads in comparison to the others. For this, they intend to assign each lead a score and then focus on leads with a higher score which helps them reach a conversion rate of 80%.

For this, we created a binary classification model using logistic regression technique on the data of previous leads. Binary classification is a supervised machine learning algorithm which helps provide a probability to each lead which can be used to categorize it as ‘Hot Lead’ or otherwise based on a cutoff point. To implement this, logistic regression is one of the techniques which uses log of odds and maps a linear graph on the multivariate data to classify it.

To prepare the data for the data modelling, we had to do some preparatory steps on the data provided to us:

1. We started with confirming that the data types of each column and updating, wherever required.
2. We cleaned the data by removing default values like ‘Select’ which are as good as missing records.
3. We continue thereafter by handling all the other missing values by removing either the columns, rows or through imputation.
4. Then we handle the outliers in the numerical columns through either dropping such records or through limiting these extremes.
5. Then we move to Exploratory Data Analysis where we identified the following:
    a. Identified certain important attributes with high converted customers.
    b. Drop columns which have only one value in majority records
    c. Bucket values in the columns where too many values in the column exist but with low frequency.
6. Encoding the categorical columns into numerical columns with binary values so that a mathematical technique (logistic regression) can be applied on top of it.
7. Checking correlation and remove highly correlated columns to avoid redundancy.
8. Scaling the numerical values so that the columns have similar value to the model.

Now that the data was ready, we built a model on the data by removing insignificant columns and reducing multicollinearity. We were able to train our model to an accuracy of 93% and evaluated it to similar percentages as well.

It was found that certain variables such as Tags, Lead Source, Last Activity, Time Spent on Website were having higher significance to decide if the leads will convert.

Through this project, we were able to practice EDA techniques on the data provided and put into practice a classification model creation with logistic regression technique. We learnt how to train the data with high accuracy and evaluate it with various metrics. We were able to use them to highlight recommendations for the X Education company.