**EXECUTIVE REPORT**

1. 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.
2. Now, although X Education gets a lot of leads, its lead conversion rate is very poor at 30%. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:

Lead Conversion Process - Demonstrated as a funnel

1. As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc. ) in order to get a higher lead conversion.
2. X Education wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company wants to build a model which assigns a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.
3. We utilized Python language for carrying out EDA and building a Logistic Regression Model to assign Lead Score for each lead. Previous Leads conversion data containing 9240 data points was provided by the company. After carrying out initial data sanity checks lot of Null values were found. After going through the data dictionary and data thoroughly, all columns having more than 40% missing values were dropped. Remaining feature having missing values were dealt with by imputing ‘Unknown’ for categorical variables. Rows were dropped for features having less than 2% missing values. This was followed by Outlier treatment of numerical features, this was done by dropping rows containing outliers. We were left with approximately 91% of data points remaining after data cleaning.
4. We also removed certain features like Tags because they are the tags given by calling executive based on their intuition and experience. This is removed to avoid any prejudices moreover this data will not be available before a lead is contacted.
5. During EDA Total Time Spent on Website and TotalVisits appeared to contribute towards conversions.
6. Logistic Regression model from Sklearn module of python was used to build Logistic Regression model and RFE is used for initial selection of 20 features. By eliminating features based on p-values and VIF a stable model was built. On building a ROC curve we obtained 0.89 which is a good score indicating that the model is close to ideal.
7. Now we needed to set a threshold value for deciding Hot Leads based on the probabilities. Since CEO wanted a conversion rate of 80% we used Precision vs Recall view to find the optimum threshold. Based on the Precision vs Recall graph we need a Precsion of 0.8, so threshold came close to 0.52. With a threshold of 0.52 we get a Precision of 0.8 and False Positive Rate was only 0.1 or 10%. So we decided that a threshold of 0.52 is ideal for our requirements of reaching a target of 80% conversion.
8. Based on the Logisitic Regression Model used, Total Time Spent on Website, Lead Originated through Add Froms and Occupation Working Professionals have higher probability to drive conversion. So company should work towards making the leads spend more time on the website with engaging content, ensure most leads fill the Add Forms for getting more details from them and try to contact working professionals more by having multiple calls by getting an appointment or contacting after working hours.