Problem Statement:

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%.**

Lead conversion process: 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.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.**

Data :

You have been provided with a leads dataset from the past with around 9000 data points. 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. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Another thing that you also need to check out for are the levels present in the categorical variables.

Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.

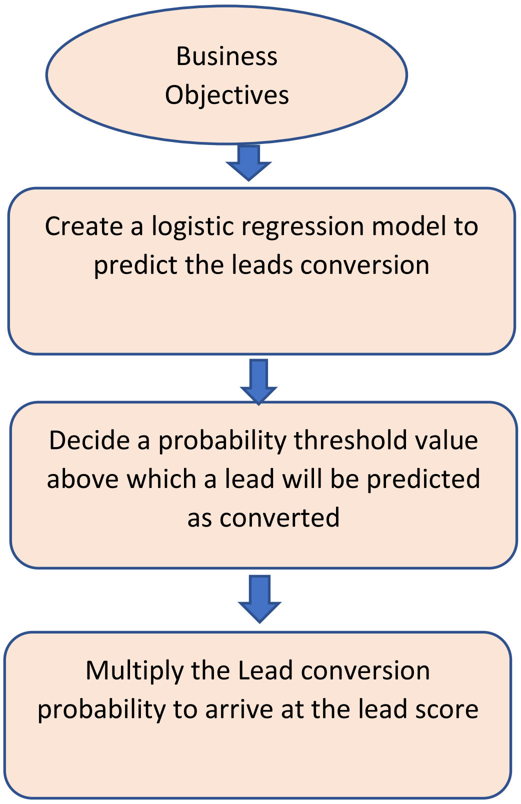
Goal :

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted

Summary:

As per the Business requirement we need to build a logistic regression model , which can help X Education system to identify the potential leads which can be converted as paying customer and also help them to reduce making calls to each and every leads which may or may not be getting converted

I sub divided the Business objective into 3 categories



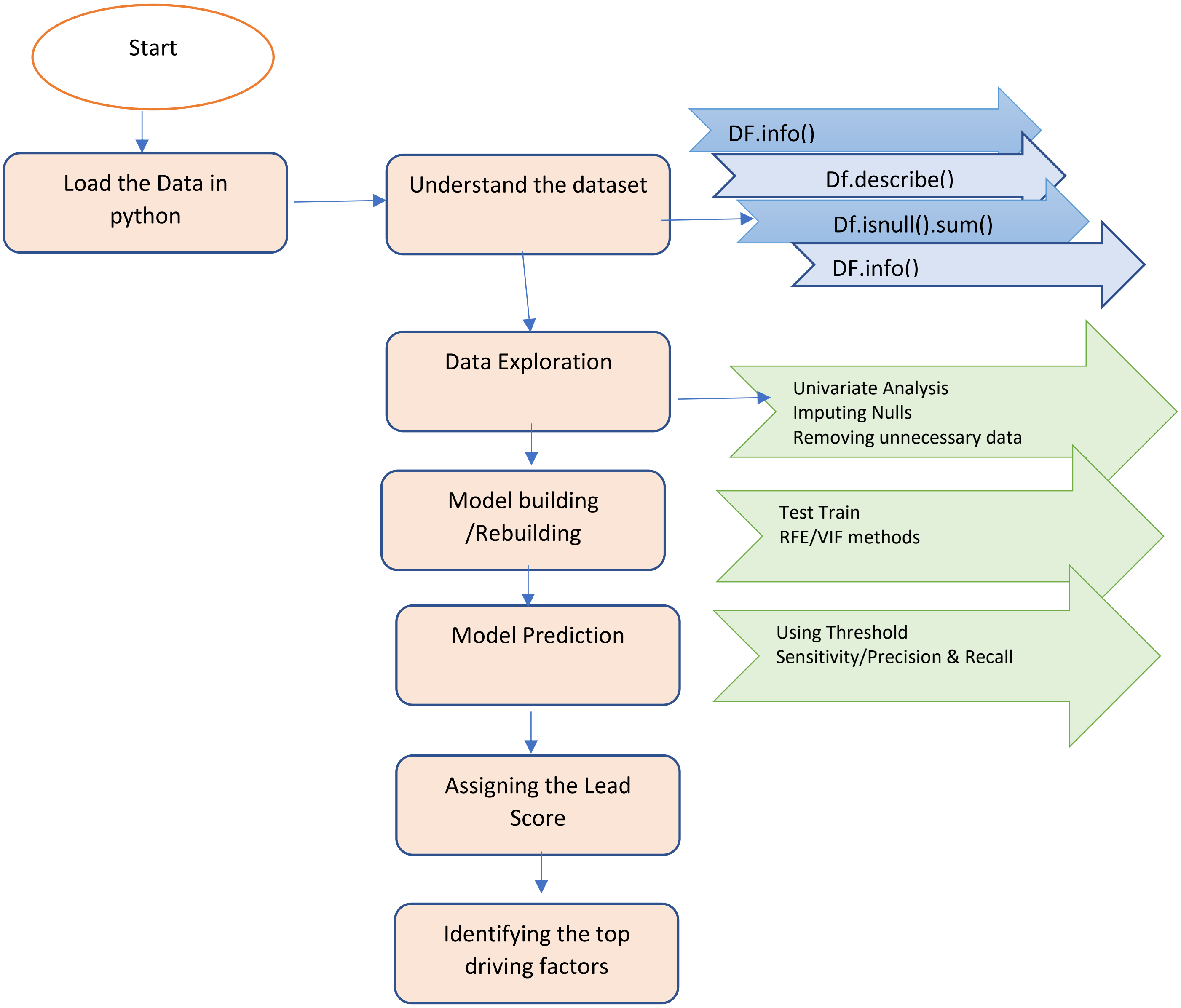
Business Objectives

Create a logistic regression model to predict the leads conversion

Decide a probability threshold value above which a lead will be predicted as converted

Multiply the Lead conversion probability to arrive at the lead score

Data Analysis Process Flow



Brief description of steps:

:- Loaded the csv file inspected the rows /column details.

:- Look for the Null values

:- Imputed the null values using Median, Mode and by dropping few rows

:- Performed exploratory data analysis, using univariate and bivariate variable analysis

:- Using EDA identified the unwanted variables and dropped those columns

:- Imputed the “Select” value for City column using most frequent variable, in Specialization columns created a new variables and imputed the values , for country using the median method to impute it.

:- Changed the Categorical columns having ‘Yes/No’ values using ‘0’ & ‘1’

:- Created dummy variable

:- Preformed model evaluation using VIF /RFE score method after calculating the p- value and z-score.

:- Using Threshold value tried evaluating the model to get the best fit model

:- Merged the train dataset with the original and assigned the Lead score for 0 to 100.

:- Used ROC curve to identify the threshold

:- Identified the top 3 variables which can derive the results.

Conclusion : A person who always uses or browse to internet and keep checking the email will be able to get converted as lead.