The use case was to build a logistic regression model in order to assign a lead score to each of the leads that can be used by X Education to help target potential leads which will help the company to get more customers.

The data had multiple lead behaviour data along with the Identifiers, survey and demographic data.

The data was to be treated for missing values in which some of the given features were removed based on more than 30% of the missing values within the distribution.

Some features are removed where seen to be one predominant value /constant value.

Rest of the features imputed with mode/ median values.

The next stage was EDA where features were analysed against to the target to help understand the influence. Few features were dropped based on little to no effect.

In the EDA it was found that

- API, the Landing Page submissions bring maximum leads whereas the conversion rate is higher on the Lead Add Form.

- Lead Import, Quick Add Form brings minimum leads and Quick Add Form having a zero-conversion rate.

- In lead source, Direct Traffic and Olark Chat brings maximum leads but suffers from low conversion whereas Google has good lead inputs and decent conversion.

- In lead source, references showcase maximum conversion rate.

- In Latest Activity, Phone Conversations and SMS seems to generate hot leads having good/decent conversion rate.

- Unemployed people seem to be making up for most of the leads but with minimum conversion of almost about half.

- Businessman and Working Professional contribute to higher conversions.

- Housewives are having lower lead generation percentage, but the generated leads tend to be converted.

Data formatting was carried out which includes Standard Scaling of numerical, categorical encoding of categorical variables.

Finally, the data was split into train and test in a 70:30 proportion and the train data was found to fit the model.

The model building involves filtering relevant features and also fitting the model with optimal features based on the p-value. The features of p-values being more than 0.05 are removed. Also, the VIF factor was taken to check multicollinearity. Prediction was also done on the train and test data and performance metrics were calculated which included Accuracy, Sensitivity and Specificity etc.

The AUC-ROC curve was plotted to understand the model performance. The accuracy was obtained near to 80% as per requirement. Then lead score was computed for each customer