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

The data provided had many lead behaviour data along with Identifiers, survey and demographic data. The data had to be treated for missing values where some of the given features were removed based on more than 30% of missing values within the distribution. Some features were removed where there was one predominant value or a constant value. Others were imputed with mode/ median values.

The next stage was EDA where features were analysed against the target to understand the influence and some features were dropped based on little to no effect. In the EDA in was found that,

- API and Landing Page submission brings most of the leads whereas the conversion rate is higher for Lead Add Form.
- Lead Import and Quick Add Form brings very few leads and Quick Add Form is having a zero-conversion rate.
- In lead source, Direct Traffic and Olark Chat brings in huge number of leads but suffers very low conversion whereas Google has good lead inputs and decent conversion.
- In lead source, reference shows the highest conversion rate.
- In Last Activity, had a Phone Conversation and SMS sent seems to generate hot leads having good conversion rate.
- Unemployed people seem to be making up for most of the leads but with low conversion of about half.
- Businessman and Working Professional contribute for higher conversions.
- Housewives are having less lead generation percentage, but the generated leads tend to be converted.

Then data formatting was carried out which included Standard Scaling of numerical and categorical encoding of categorical variables.

Finally, the data was split as train and test in a 70:30 proportion and the train data was given to fit the model. The model building involved filtering relevant features and fitting the model with the optimal features based on the p-value where features with p-values more than 0.05 are removed. Also VIF factor was taken to check multicollinearity.

Prediction was done on train and test data and performance metrics were calculated which included Accuracy, Sensitivity, Specificity etc. The AUC-ROC curve was also plotted to understand the model performance. The accuracy was obtained near to 80% as per requirement. Then lead score was computed for each customer.