Summary

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 through different search engine like Google, through referrals, videos fill up the forms when browsing the courses etc.

The analysis is done as a part of X Education to help them understand how can we classify different leads and convert them to make sure they are choosing the course.

We have followed different steps in this analysis:

- We started with importing different libraries which are used for analysis.
- The second step was importing the data and looking for the basic details in the data like its shape, info, describing it, checking for missing values etc.
- **Data Cleaning**: Once we have checked for the missing values, we are trying to impute the missing values. We have removed the missing values where there are values > 30%. Then we have started with imputing the missing values. We have chosen mode for categorical values, and median for numeric values. Also, we have noticed a lot of categorical values with a lot of categories so we have used pd. get_dummies to convert them to numeric values and standardised the numerical values.
- **EDA**: A lot of categorical variables were having very insignificant values because there are hardly any values in it but we can remove them so we have still considered them while building the model. And in the case of numeric values, we have checked for outliers and there are no significant outliers that we found.
- **Data Visualisation**: We have used different visualisations and tried to come with interpretation that could help with model building. Like in the Country we have notice majority of the values are from India and very few from other places.
- **Model Building**: We have used RFE to get important features for model building and tried to reduce them in number to get the best fitted model.
- **Evaluating the mode**l: We have evaluated the model on accuracy, sensitivity, specificity on both train and test data.

When we were evaluating the model and predicting the values, we have used different metrics:

- We have confusion matrix for both the train data and test data.
- We have also used 0.5 as a initial cut-off while predicting the data
- Accuracy refers to the level of agreement between the actual measurement and the absolute measurement.
- We have got an accuracy of 79% on train data and 78.5% on the test data.
- Sensitivity is calculated as the number of correct positive predictions (TP) divided by the total number of positives (P). It is also called as Recall.
- The sensitivity in the train and test data are 83 % in both of the cases.
- Precision is the ratio of true positives to the total of the true positives and false positives.
- The precision in the train and test date 66% and 67%.
- We have also used specificity as one of the measures.
- The specificity, with formula TN / (TN+FP), tells us the true negative rate – the proportion of people that have not been converted and are correctly given a negative result.
- The specificity in the case of train and test data are 76% and 75%.
- We have also used ROC curve for identifying the cut off point.
- We have also plotted accuracy, sensitivity and specificity together to see where they intersect with each other.
- We have finally used the cut off as 0.3 to correctly predict the values.

While we used different evaluation metrics, we have come with few variables that we think we can focus on for the business:

- Do Not Email
- Lead Source - Olark Chat
- Last Activity - Email Bounced
- Specialization - Finance Management
- Last Activity – Olark Chat Conversation
- Total Time Spent On Website
- Last Activity – SMS Sent
- Last Activity – Unsubscribed
- Last Activity – Converted to Lead
- Lead Origin – Lead Add Form

- Last Activity – Email Link Clicked
- Lead Source – Facebook
- Last Activity – Submitted on Website

We can use more variables for analysis and come up with more models that could help the business. We can also use different analysis techniques like where there are more categorical values trying them to limit them and simplify them for better analysis.