SUMMARY

X Education, a company that sells online courses to industry professionals, has many potential clients that visit their webpage, looking for a course to buy. These potential customers, when they fill up any form providing their email address or phone number, are then classified as 'leads'.

However, the conversion rate of Leads at X Education is very poor, about 30%. The company requires us to build a model wherein we need to assign a lead score to each of the customers wherein the lead score represents the probability if the lead will be converted or not.

That is to say, A score of 1 would mean a hot lead and a score of 0 would mean a dead lead.

Data Cleaning :-

- After looking at the data, it was evident that grouping of similar features was required, so we have grouped all the similar columns such as "interested in other courses" and "Not doing further education" into a category called "doesn't show interest. And many others as described in the notebook.
- There were many select entries in columns, indicating potential blanks or unassigned values, we changed it to "Not Provided" for better understanding of the data
- Columns with high null values (>40%) were dropped and so were columns which did not provide any value to the dataset.
- Split the dataset to train and test to maintain data integrity.

EDA:-

- Checked data imbalance
- Performed univariate and bivariate analysis for categorical and numerical values, found the most influential variables such as time spent on website, lead source etc.
- Identified further highly correlated columns / features.

Data Preparation:-

- Created one-hot encoded dummy variables for categorical variables.
- Numerical columns were imputed with KNN imputer
- Missing values were considered as separate categories, as imputing them with mode would skew the data and result in worse model performance.
- Outliers were capped at 95th percentile, as logistic regression is very sensitive to those.
- Feature scaling was done and columns that were highly correlated to other columns were dropped.
- Created a Data Pipeline to ensure further scalability and solve multiple business problems at once.

Model Building:-

- A total of three models were built. They are logistic regression, random forest and decision tree.
- Decision Tree model overfit to our data in our case, so we did not continue with it.

- Logistic regression and random forest both performed well and above the required mark.
- Using the pipeline, feature reduction was automated and nothing was done manually. This will save countless manhours in the company.

Model Evaluation:-

- Both models performed equally well, although logistic regression performed marginally better.
- Logistic regression scored 94.5% in test accuracy.
- Random Forest Classifier Scored 94.15% in test accuracy.
- The model assigns a value from 0 to 1, the more the value the more the probability of the lead converting.
- The Model identified the most influential variables to be ‘last notable activity’, ‘asymmetrique activity score’ and ‘specialization’ as the top most influential variables in the dataset.

Conclusion:-

- Customers spending more time on the website were very highly correlated to be hot leads.
- Lead sources, such as referrals and others were very influential too, boosting this sector can give good returns.
- Working professionals are the ideal customerbase for X education, as most of them have positively correlated features.
- Implementing a pipeline in the model process makes the task very efficient and enables the business to react quickly on changing market trends.