LEAD SCORING CASE STUDY

Problem Statement: X Education the online education company sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Among the people who land up on the website, people who fill up the form providing their email address or phone number turns out to be a lead. Our aim here is to help increase the Lead Conversion Rate which is at a low 30% now. Targeting the right leads, is the solution here, for which we will 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.

APPROACH: We need to find the score for each lead whether the lead will turn out to join or not. Of course we will go with Logistic Regression modeling and find the probability for each lead and finally assign a score from 1-100 for each lead to mark the lead as important or not.

- We started with importing the necessary libraries and then went with Data understanding part to get a look and feel of the dataset. We observed the initial few records of the dataset; initial statistics to see what each feature has got to offer and the type of each column and most importantly the dimensions of the dataset.
- We could see many Null values for the features, hence started Data cleaning with Null value treatment. We went ahead by imputing null values with the mean or mode appropriately for some of the features where as introduced new level as 'unknown' in some cases. We went this approach, so as to retain the major chunk of the dataset and made sure we are neither introducing bias nor changing the distribution of the feature. In times we have dropped unnecessary columns either based on business intuitions or the value it can provide to our model.
- As part of Data Cleaning drive, we also had Data Imbalance checks and Outlier Detection and treatment. A call was taken based on the check on the features (either distribution or the counts) whether to retain the feature or not.
- After the Cleaning drive, we could see many categorical columns in our dataset. Either One-hot encoding was done for different levels present for a feature or else replaced the different levels with different rankings inorder to convert all categorical values to continuous.
- Train-Test split was done next with random_state of 100 in a ratio of 70% to 30%.
- Scaling was done for the features in train dataset inorder to bring down all the features in a common range. We went ahead with the Standardisation for these features.

- Since we had around 126 features in our last dataset, we used RFE to cut short all the features to only relevant few features and went ahead with model building.
- From the initial 25 features which RFE suggested, we used a backward approach to drop the insignificant features which had either high p-value or high Variance Inflation factor.
- Final model we got was evaluated not only just accuracy but for various features such as Sensitivity, Specificity, True Positive Rate, False Positive Rate.
- Use of ROC and AUC also to evaluate the model in-hand.
- Optimal Cut-off point was found out which we finalized and based on which we checked the model evaluation in both train and test datasets.
- Final dataset was produced which contains just the Lead Number and the Lead score which was found out based on the probability given by the final Logistic Regression Model.

WHAT THE FINAL DATASET INDICATES: The customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. Lead nurturing and strategic planning can be done in a way to contact the leads with high lead score so that we can get a good conversion rate.