SUMMARY REPORT

LEADS SCORING CASE STUDY

An education company named X education provides info and markets courses for various fields by collecting data and using various websites and search engines as sources for the customers to view. It's leads data has been provided in this case study as it turns out that the company wishes to focus on a better and higher lead conversion rate through communicating leads with more potential for active results. We create an ML logistic regression model to figure out if a customer will return with the intention to buy/proceed with the course after viewing the source and courses marketed by the company.

The data includes several factors from previous leads such as total time spent, total visits, types of lead origins, sources, if requests to call and email back are granted, specialisation, occupation, country and remarks of the marketing team as to which customer has potential to take up the course and which does not.

We further go along with the data cleaning aspect of the analysis by figuring out null values and dropping those columns or replacing with optimal values. From the initial explorative data analysis, we understand that the specialisations from finance management, banking, investment and insurance have shown more promise of taking up a course and occupations including unemployed and working professionals have opted in for better career prospects. A decent number of customers also require flexibility and convenience as a factor to choose the course.

Finally, we attempt at building the model through a train-test split of 70-30, then standardise the continuous variables. Using RFE set as 25 variables and run the first features training set model. Checking for p-value lower than 0.5 we remove other features one after the other that higher in p-value, further re-calculating the model. We use the VIF aka Variance Inflation Factor to check the multi-collinearity in the model. After using a VIF of 3, we further remove those features and get a final model. After calculating the accuracy, sensitivity and specificity of the model we apply the prediction on the test set and finally calculate the ROC curve. In the final prediction the lead score is calculated. The evaluation metrics using a cut-off of 0.3 are calculated as accuracy of 90%, sensitivity of 96%, specificity of around 88%, precision score of 83% and recall score of 96% which makes for a good logistic regression model.

We finally conclude that the company should focus on leads that will revert back with email, leads closed by 'Horizzon' and to avoid customers that have been busy to respond, marked as switched off, not interested in further education or are already a student.