**Lead scoring case study**

**Summary**

The analysis is done for an X-Education to select most promising leads i.e. the leads that are most likely to convert into paying customers. The data give to us includes information about how customers visit the website, the time spent there, how they reached the site and conversion rate.

The following steps are used to analyze the data and predict the lead score for each customer:

1. **Loading and Cleaning Data:**

* The data given to us has lot of columns which explains whether a lead is converted or not.
* There are total 9240 rows and 37 columns.
* Few columns consists of select values which indicates the customer left this as select without updating the result. We have replaced this select values as Nan values

1. **Exploratory Data Analysis:**

* For null values, we have replaced most of them with mode, few columns like tags and specialization as ‘not specified’.
* We have also combined the low frequency values of each column in order to decrease the dummies.
* We have also dropped most of the columns which are skewed towards only one category.

1. **Data Visualization:**

* Used count plots and histograms for univariate analysis
* Used count plots with converted variable to check which variable gives more conversion.
* Used pair plot to check the relationship between numerical variables
* Used heat map to check the correlation between each numerical variables and target variable

1. **Outlier Handling:**

* We have used the boxplot to check the outlier in the data and removed outliers by hard capping lower and upper quantiles as 0.05 and 0.95.

1. **Train test split:**

* The split was done at 70 % train data and 30 % test data

1. **Scaling:**

* Used MinMaxScaler to scale the numerical variables with different range to 0 to 1.

1. **Model Building and evaluation:**

* We have used the Stats models to build the model and used recursive feature elimination to eliminate the variables which are insignificant.
* We have also used p-value and variance inflation factor (VIF) to eliminate more features from the model which are insignificant. (P-value <0.05 and VIF < 2)
* We have used confusion matrix, sensitivity, specificity to plot the ROC (Receiver operating characteristic) curve to find the optimal cutoff of 0.3 which finally resulted in 90% each on train data

1. **Precision and recall:**

* This method is also used to check the cutoff and found out that cutoff is around 0.41 and found with precision and recall is around 88% for test data

1. **Predictions:**

* Prediction made on test data with 0.3 cutoff with accuracy, sensitivity and specificity of 90 % each

We found that variables that mattered most in conversion are:

* 1. Total Time Spent on Website
  2. Tags Lost to EINS
  3. Lead Origin Lead Add Form
  4. Tags Will revert after reading the email
  5. Tags Ringing
  6. Tags Interested in other courses
  7. Lead Source Welingak Website
  8. Last Notable Activity SMS Sent
  9. Last Activity Olark Chat Conversation
  10. Lead Source Olark Chat