Summary Report of Lead scoring case study

- We have been appointed by X Education to help in selecting the most promising leads. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance.
- So, as I started studying the data it consisted of 9240 rows and 37 columns. Many attributes have 'Select' as a data, which is nothing but customer leaving the data field unanswered.

  This is more of a null value and has been treated in the same way.
- The important thing that has been done before modelling is Exploratory Data Analysis.
  We have started checking for null values. There are multiple columns having more than 40% of null values. Those columns have been dropped.
  The null values in other columns have been imputed with mean/median or mode of the column.
- Few columns which have 99% of records as a single value haven been dropped as these columns do not provide any additional information to the model.
- Multiple plots have been plotted to understand the behaviour of attributes.
  Plots taking in account of converted variable have been plotted to determine which variables favour in conversion of the lead.
- Dummies have been created. If a categorical column has n levels, n-1 dummies have been created.
- Split the data into Train and Test data in the ratio of 70-30 respectively.
  Scaled the numerical columns to make the data lie in the same range.
- We built the Linear Regression model using Statsmodels. This is the first and raw model built with all the columns present in the data. To increase the performance of the model, we have selected attributes using automated feature selection through RFE.
- We found Threshold value by plotting the metrics i.e., Accuracy, Sensitivity, Specificity. The threshold value is the value where Accuracy, sensitivity and Specificity lie almost in the same range.
- Using the threshold value, we found the lead score which is the probability of lead conversion. The leads having a score greater than threshold value are the leads who are likely to pay and take up the course from X Education when nurtured well.

- As we are satisfied with the model which has an accuracy of 87%, we fed test data to the model for validation.

  The numerical data in test data is scaled and fed to the model.

  Upon assessing the metrics of the model on test data, we achieved an accuracy of 88%.

  ----Train Data Metrics-----

Train Accuracy score:	0.869
Train Sensitivity:	0.863
Train Specificity:	0.873

  ----Test Data Metrics-----

Test Accuracy score:	0.881
Test Sensitivity:	0.864
Test Specificity:	0.892

- Few attributes that play an important role in determining the behaviour of the customer.

  Tags_Will revert after reading email
  TotalVisits
  Total Time Spent on Website
  Lead_source_Welingak website
  Lead_origin_Lead Add Form