Case Study Partners:

Amith T Umesh & Soubhik Bandhyopadhyay (IIIT-B / DS 26)

Summary Report
Lead Score Case Study

- It is very common these days that most of the people from professional working background who are seeking for their career growth or transition, Students who are looking for solid platform to up skill themselves are finding a way to opt courses along with existing responsibilities.
- In order to serve this purposes various online education service providers started forming their business. They target people who are industry professionals and change seekers to join their various available courses to upgrade themselves from the current role.
- In our case study, we have a similar business problem where X-Education is an education company referred above who sells online courses to industry professionals.
- The company is marketing its various courses on several websites and search engines like Google. It is also focusing on past referrals. The acquired leads are tried to reach through calls or emails. In this process some are converted and most of them are not.
- The claim or challenge given by X-Education team is that the amount of leads are much higher and the conversion of the leads into sale are very much poor.
- Now, X-Education is asking us to go through their data and analyze to help them to select the most promising leads which are having higher probability of getting converted into successful sales.
- Based on these, we are supposed to build prescribed statistics model to understand the influencing factor for the positive business results and also to help them eliminate the dragging factor which is resulting in negative impact in their business.

- **Data Analysis** :

  - Now that we have been provided with data and which is not completely clean and needs to be treated with following steps and procedures.
  - Basically data contains around 9000 data points and the data information provides us that it has numerous null values inside it. By notice, we also know that the customers who had not selected any of the options while filling the application are left as ‘Select’. Which is also considered as Null values.
  - So, in order to treat/Clean that we had to convert first all the select values to null values. Later, we had to involve in dealing with individual variables to check its null values and its rating against the total number of variables. Finally, to drop or replace the null values with the most suitable value of that variable.
  - Further, we started with getting our hands on data by executing Exploratory Data Analysis (EDA). Here, we plotted plots for almost each of the variables to visualize the information compounded in it.
  - We had to bucket most of the least counted values in the variables to the most common or into other set of that variable. Which will help is minimizing the mass of the data for further analysis.
  - Then we dropped the insignificant variables assuming that it has nothing much to do inside our analysis.
  - We created dummy variables for getting all our set of values contained in the variables.
  - Then we split the entire data into 0.7 and 0.3 for train data and test data respectively.
  - We also used standard scalar tool to fit and transform the train data variables had a different ranges.
  - Since the number of feature variables were very huge and it was not practically possible to deal with all the variables manually. Hence we decided to use feature RFE selection mode to select top 15 reasonable variables.
  - With the selected feature variables, we usually dealt with it manually by looking into stats summary and Variance Inflation Factor (VIF).
  - We dropped very few variables which had p-values more than 0.05 and VIF more than 5.
  - Along with that we had also taken care of model accuracy by monitoring it every time.
  - We created confusion matrix and calculated various factors such as sensitivity, specificity, false

positive rate, positive and negative predictive values.

- Our train set had around 90% for all accuracy, sensitivity and specificity.
- Alongside we also plotted ROC curve where the curve was close to one by attaining 0.97 and which represents very good predictive model.
- Each time the metrics were calculated and accuracy were monitored for the variations.
- The prediction on the test data was made and the sensitivity, specificity and accuracy from the confusion matrix were again very close to 90%.

○ The Final Model Statistics Summary :

- Based on our final model, following feature variables are having positive impact in the line of business. Those are as follows :

Feature Variables	coef
Tags_Closed by Horizzon	8.8727
Tags_Lost to EINS	8.7737
Tags_Will revert after reading the email	3.976
Lead Source_Welingak Website	3.7757
Tags_Busy	3.6679
Last Notable Activity_SMS Sent	2.7326
Lead Origin_Lead Add Form	2.474
Total Time Spent on Website	1.1325
Lead Source_Olark Chat	1.036

- Based on our final model, few feature variables are also having positive impact in the line of business. Those are as follows :

Feature Variables	coef
Tags_Ringing	-1.9883
Tags_switched off	-2.0853
Lead Quality_Not Sure	-3.5101
Lead Quality_Worst	-3.5338

- Considering all the above factors, X-Education Company has higher possibilities to attain to reach all the potential buyers to convince the buyers and convert the sales.

****