**Summary : Lead Scoring Case Study - Rajat Patil & Todd Ryman**

**Problem Statement**

**An education company named X Education sells online courses to industry professionals. Although X Education gets a lot of leads, its lead conversion rate is very poor. The company requires a model wherein a lead score is assigned to each lead such that a high score indicates a hot lead and low score indicates a cold lead. The CEO has given a target lead conversion rate of 80%.**

1. **Build a logistics Regression model from the dataset to assign a value from 0 to 100 for a lead score for each lead to show the probability of the lead converting to student.**
2. **Lead conversion rate to be around 80%**

**Solution Methodology**

1. **Data Understanding**

**We were given a dataset containing 9000 leads and their conversion success. The dataset provides the choices that the customer made filling in the form on X Education website. There is also some internal data in the dataset.**

1. **Data Preparation**

* **Convert each object value in each object column to lower case string,**
* **Changed the value of “select” to null. As the value “select” is the default value meaning the user did not select any value.**
* Then we remove the columns which have a null value 30% or more.
* Dropping Prospect ID as it is a unique identifier and we already have Lead number. Dropping, I agree to pay the amount through cheque because we are only concerned with if they convert not how they pay
* For countries columns, filling the null values with “unknown”. As 95% of known countries values are “india” we converted other countries to “outside india”.
* Removed columns in which 99% of the customers responded in the same way and categorical variables with one unique value as this does not give any additional information for our analysis.
* Some categorical variables multiple values with small percentages, we converted these values into a single category “other” for ease of analysis.
* Mapped the categorical variables with binary values of yes and no to 1 and 0.
* Remaining categorical variables we made dummy variables and removed the original columns.

**Outliers Analysis:**

The outliers will not affect the data analysis.

1. **Model Building**

We split the data into train and test data and built our model using the train dataset. We apply feature scaling and used RFE to reduce the number of variables to twenty. Using statistical analysis of the remaining variables we checked all the variables VIF and p-values. Removed variables with the high VIF or p-values. We do this one variable at a time as each removal can change the results.

1. **Model Evaluation**

Plotted the ROC curve with resulting score of 0.89. Ploted graph of sensitivity, specificity and accuracy to allow us to find the point of intersection (0.35) which is the optimal cutoff probability of for our analysis. Using the cutoff value, we assign a binary value of 1 or 0 for the predicted conversion. The model had the following metrics, Accuracy= 80.8%, Sensitivity = 84.2 %, Specificity = 78.7%.

We run the test data through our model to check performance. The results were, Accuracy= 80.5%, Sensitivity = 83.5 %, Specificity = 78.8%.

The result is in line with the train dataset result and meets the requirements of CEO of X Education.