# Lead Scoring Case Study Summary

**Problem Statement:**

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:


As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc. ) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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 and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

**Solution Summary**:

1. **Reading and understanding the data**

   Reading the dataset and analysing its shape and datatypes.

2. **Data Cleaning**

   I first dropped the variables having a high percentage of NULL values in them. Then I performed NULL value imputing according to the feature. Then I created 0 and 1 levels for binary categorical variables. At last, outliers were identified and removed.

3. **Data Analysis**

   Then the Exploratory data analysis started, where each feature was explored according to its impact on the target variable. At the end, unimportant features were dropped.

4. **Dummy Variable creation**

   Then dummy variables were created for different categorical variable with different levels.

5. **Train Test Split**

   The data set was then split into Train and Test data set with a ratio of 70:30.

6. **Feature Scaling**

   Then Standard Scaler was imported and used to scale the numeric variables to -1 to 1.

7. **Feature Selection**

   After first selecting the features through EDA, then RFE was used. Top 20 features were then selected.

8. **Model Building**

   An incremental model building approach was then used to eliminate features with high p-value and high VIF value. From the features finally selected the final model was built.

9. **ROC curve**

   After plotting the curve, the area under curve came out to be around 89, which meant that the model was a good fit.

10. **Finding the optimal probability threshold**

    From the accuracy, sensitivity and specificity trade off graph the optimal cut off was chosen as 0.34.

11. **Testing the model on both Train and Test set and comparing the metrics**

    Training Data: Accuracy: 81.0 % Sensitivity: 81.7 % Specificity: 80.6 %

    Test Data: Accuracy: 80.4 % Sensitivity: 80.4 % Specificity: 80.5 %

    Thus, we have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80%. The Model seems to predict the Conversion Rate according to the requirement. Business can be confident on using this model.