# Lead Scoring Case Study : Summary

An education company named X Education sells online courses to industry professionals. The problem statement is to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

**Approach :**

**Data understanding and cleaning** : Read the metadata to understand the meaning of every data variables. Cleaned the data using methods like deletion of columns where huge chunk of data was missing, imputation for 'Select' (to be replaced as null), deleting rows with null after cleaning the entire database.

**Exploratory data analysis :** Univariate and multivariate data visualization exercise is performed. Outliers were treated.

**Dummy variable creation**: Dummy variable creation for the categorical variables.

**Train-Test split :** 7:3 ratio is used for train-test split.

**Scaling :** Numerical variables are scaled.

**Feature selection** : Used RFE for feature selection with 20 variables to selected. Removed variables with p-values higher than 0.05 and VIF higher than 5.

**Model evaluation :** With the cut off set at 0.5 we have accuracy of 82% with Sensitivity equal to around 70% and Specificity 90%.

**Prediction :** With the Cutoff set to 0.35 we have Accuracy, Sensitivity and Specificity around 80%.

**Precision & Recall :** Precision is around 72% and Recall is around 80%.

**Conclusion**

Following variables matter the most in deciding the potential lead in descending order.

- TotalVisits
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Direct Traffic
- Lead Source_Google
- Lead Source_Organic Search
- Lead Source_Referral Sites
- Lead Source_Welingak Website
- Last Activity_Email Bounced
- Last Activity_Olark Chat Conversation
- Last Activity_Unsubscribed