

# **Lead Scoring Case Study**

## **Presentation**

### **Presenters:**

- 1. Mohammedrafik Mansoori**
- 2. Saurabh Dongare**

## 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.
- Company markets the courses on several platforms so that professional could opt the online courses, though the marketing is good and they receive fair amount of leads in converting a customer to buy their product, but the conversion rate is not as expected rather poor, despite of the leads.
- Now looking the problem we are suppose to find some criteria using which the above mentioned issue can be rooted out.

### Business Objectives:

- Our Objective 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.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- Then suggest this to CEO so that these leads could be converted.

# Analysis Approach

Data Understanding (Understanding Problem statement, Columns provided , data dictionary, Variable identification, uniqueness of variables. )

Data Cleaning/Standardizing (Dropping unnecessary variables or variables with imbalance, Renaming columns, duplicates handling, Imputing missing values.

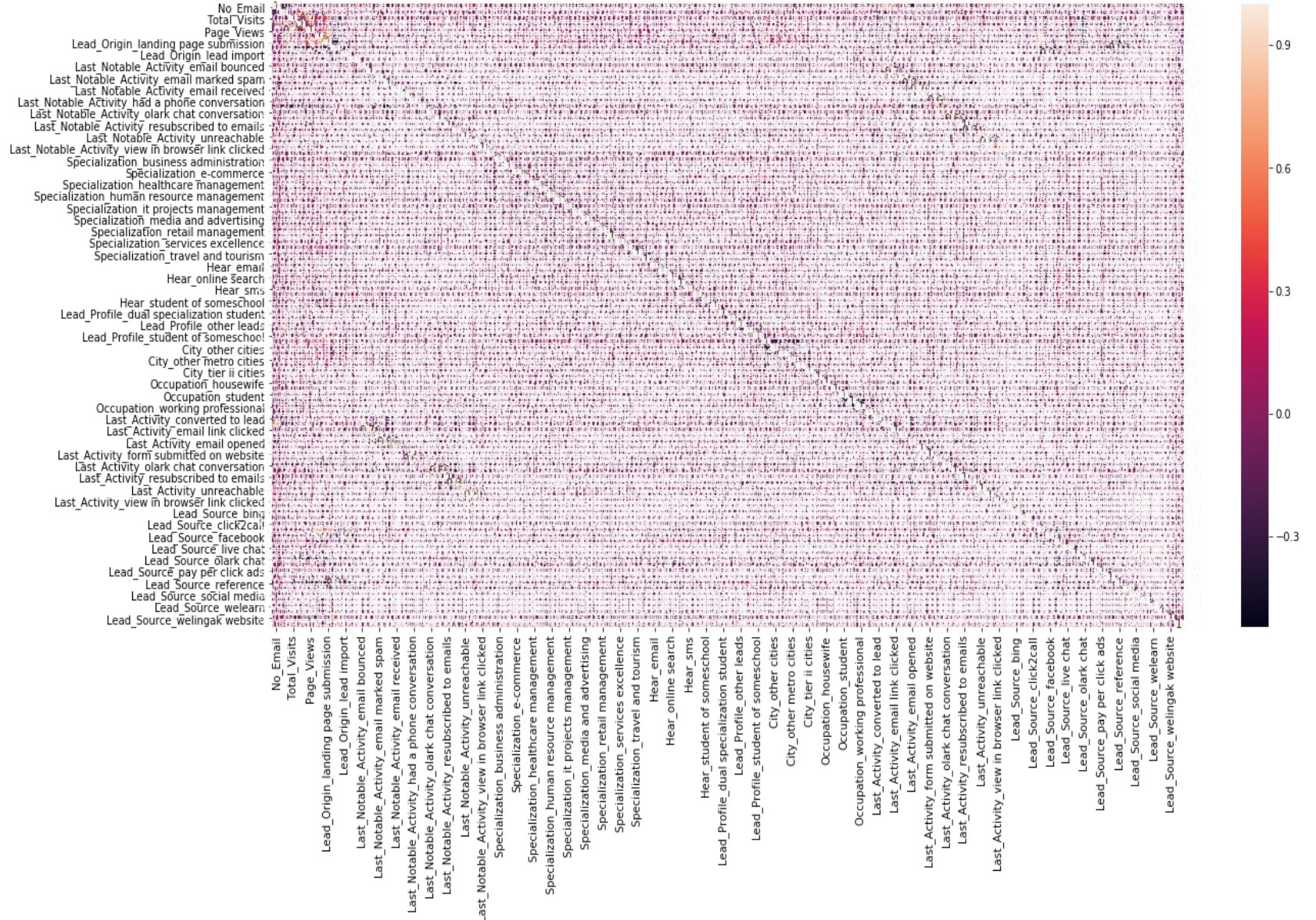
Creating dummy variables and Outlier treatments

Model creation -- Train Test split, feature scaling, Eliminating multi co-linearity on highly correlated data making feature selection using RFE.

Model building – On train set building model and scaling it to the finest using P values and VIFs. Getting the predicated values, plotting ROC , metrics finding

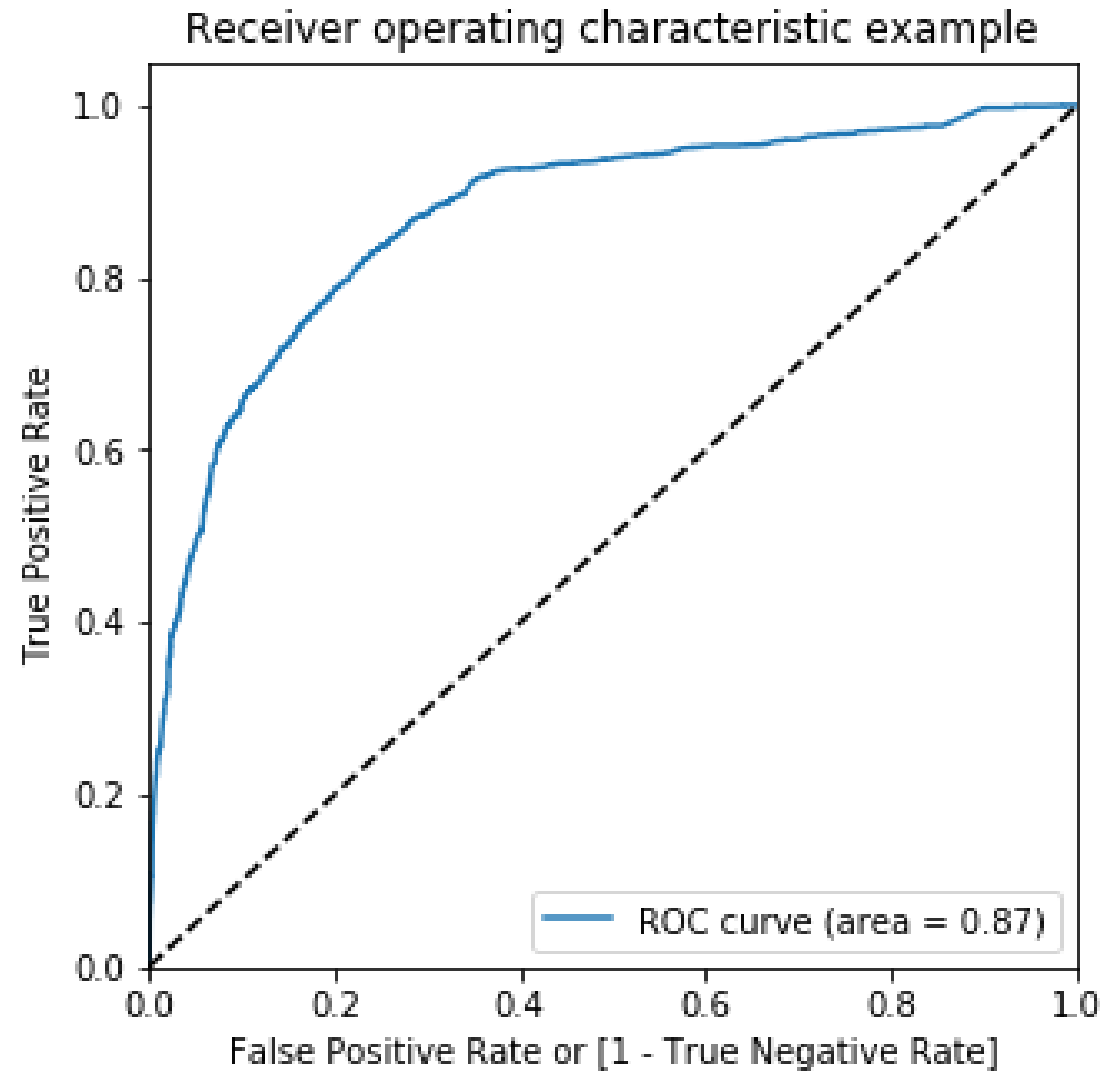
Making predictions on Test set – evaluating both train and test set, finding metrics.

1. We cannot see well correlated variables from the matrix, so we let RFE decide what to choose.
2. We performed feature selection using RFE (with 15 variables) and built the model.
3. We manually dropped collinear variable Using VIF and p-values.



# ROC Curve

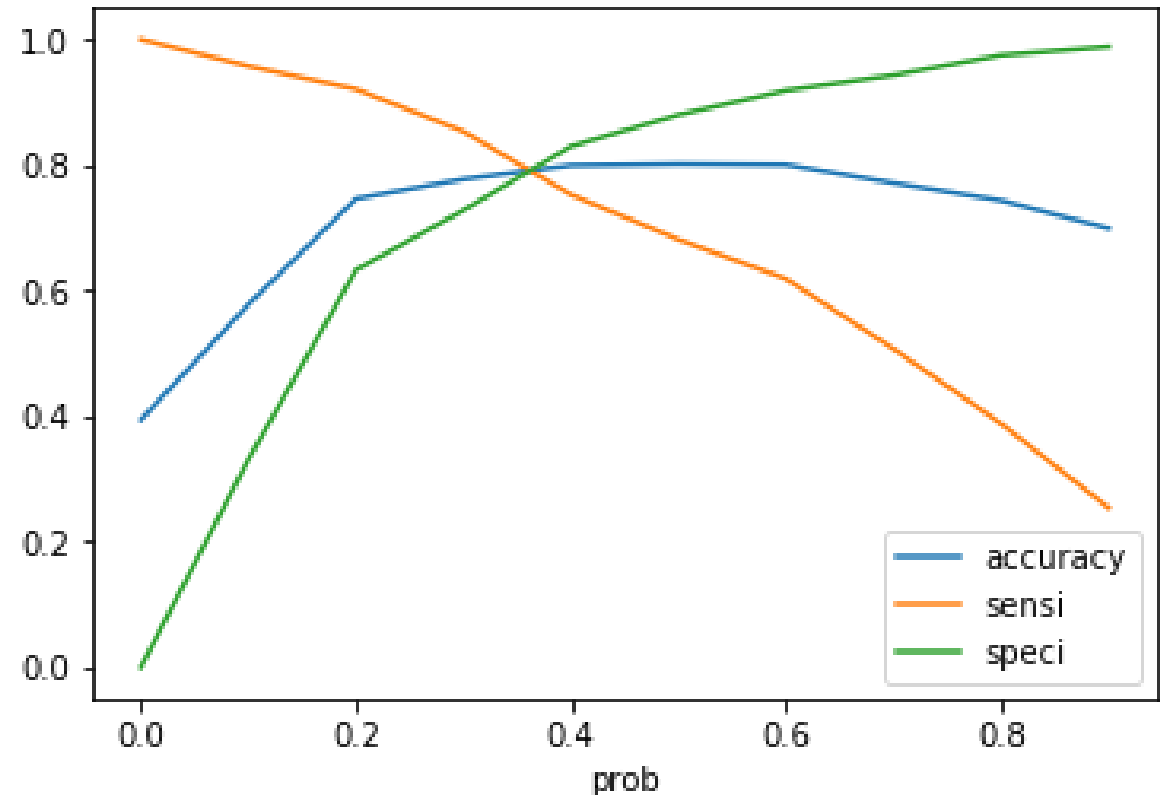
The AUC is 0.87 which is decent and the curve is not close to the diagonal





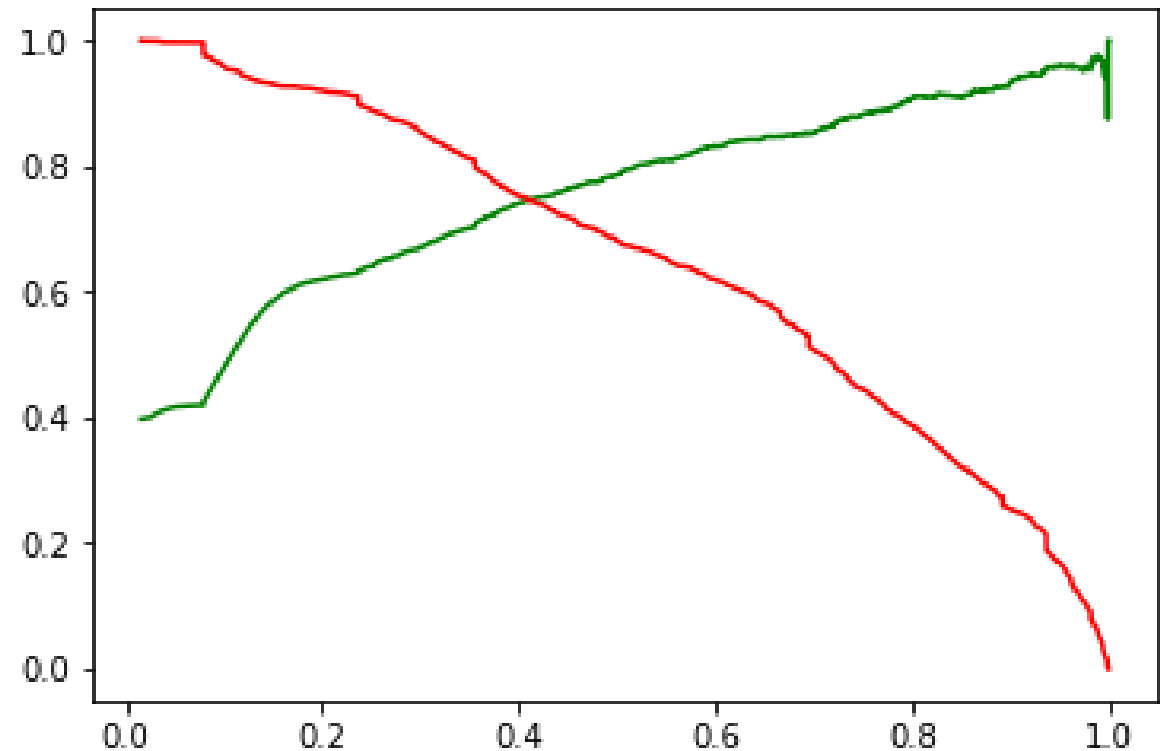
## Plot showing Accuracy, Sensitivity and Specificity for various probabilities

From the plot above, 0.35 is the optimum point to take it as a cutoff probability



# Precision and Recall Tradeoff

From the curve above, 0.42 is the optimum point to take it as a cutoff probability



## Conclusion And Recommendations

Looking at the model built, we conclude that firm should focus on the customers with below attributes

- Lead Profile is “potential” or “student of some school”
- Lead origin is “Lead Add Form”
- Occupation is “Working Professional”
- Last Activity is “SMS sent”,
- “Do Not Email” is set to “Yes”
- “Total Time Spent on Website” is high



# Model Scalability Scenario 1



## Scenario

X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.



## Solution

The cut-off threshold for conversion probability need to be brought down so that the **Sensitivity** (True Positive Rate) value increases thereby providing a higher number of Hot Leads to be followed up on. The flip side of this, ideally, is that the False Positive Rate ( $1 - \text{Specificity}$ ) increases which means that some Hot Leads predicted by the model may not actually be the Hot ones. We are ok to take this trade-off in this specific case since we can afford to have more people following up for conversion.

# Model Scalability Scenario 2



## Scenario

Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.



## Solution

The cut-off threshold for conversion probability need to be increased so that the False Positive Rate ( $1 - \text{Specificity}$ ) value decreases thereby providing a lower number of Hot Leads to be followed up on thereby reducing the need for sales agents to make a greater number of calls which is exactly what is needed in this specific case.

# Thank You!!