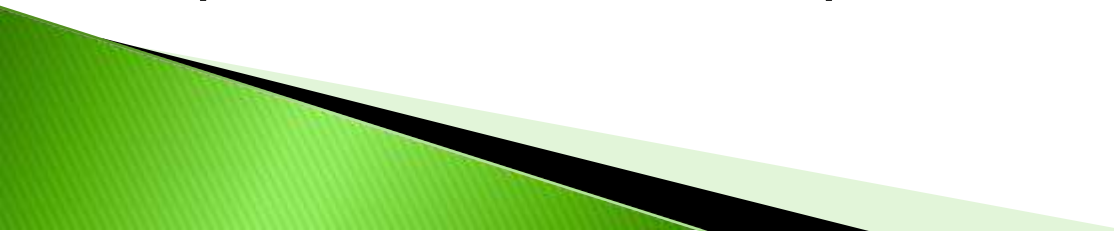


Lead Scoring Case Study

Presented by
Sayoni Paul
Sandip Sahu
Saurabh Bharatbhai Amarseda

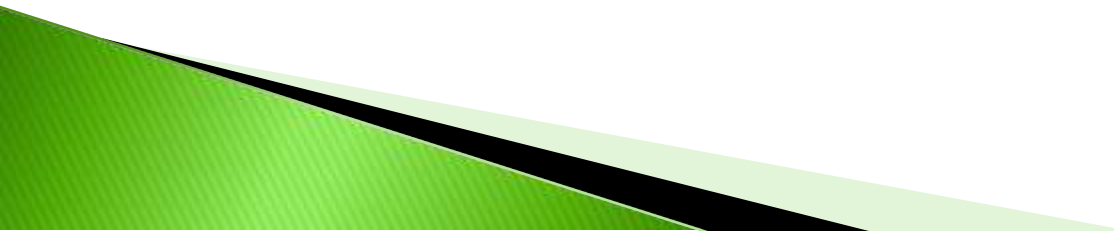
Goals of the Case Study

- ▶ 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.
 - ▶ There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.
- 

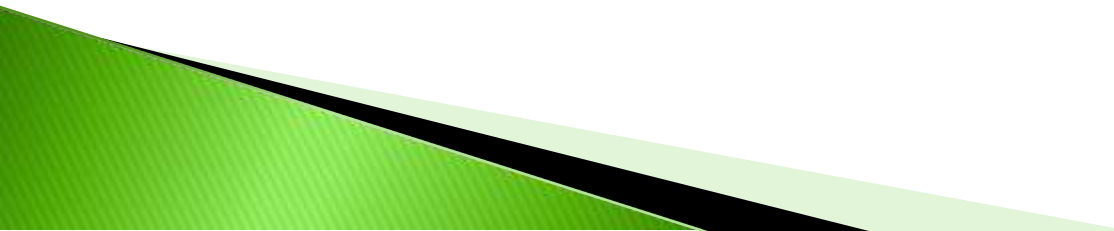
Problem Statement

- ▶ 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.

Approach

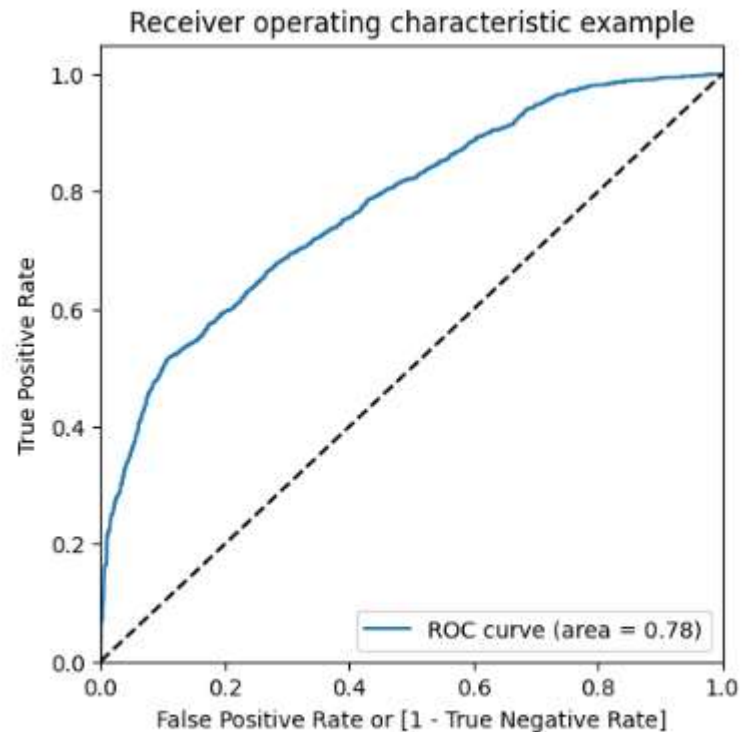
- ▶ Source of data
 - ▶ Reading and Understanding the data
 - ▶ Data Cleaning
 - ▶ Test-Train Split
 - ▶ Model Building
 - ▶ Creating prediction
 - ▶ Model evaluation
 - ▶ Optimize cut off
 - ▶ Prediction of Test set
 - ▶ Precision recall
 - ▶ Precision and recall tradeoff
 - ▶ Prediction on Test set
- 

Data Sourcing, Cleaning and Preparation

- ▶ Read data from CSV File
 - ▶ Define Outlier
 - ▶ Data Cleaning– Missing values and Null values
 - ▶ Dropping columns
- 

Optimise Cut off

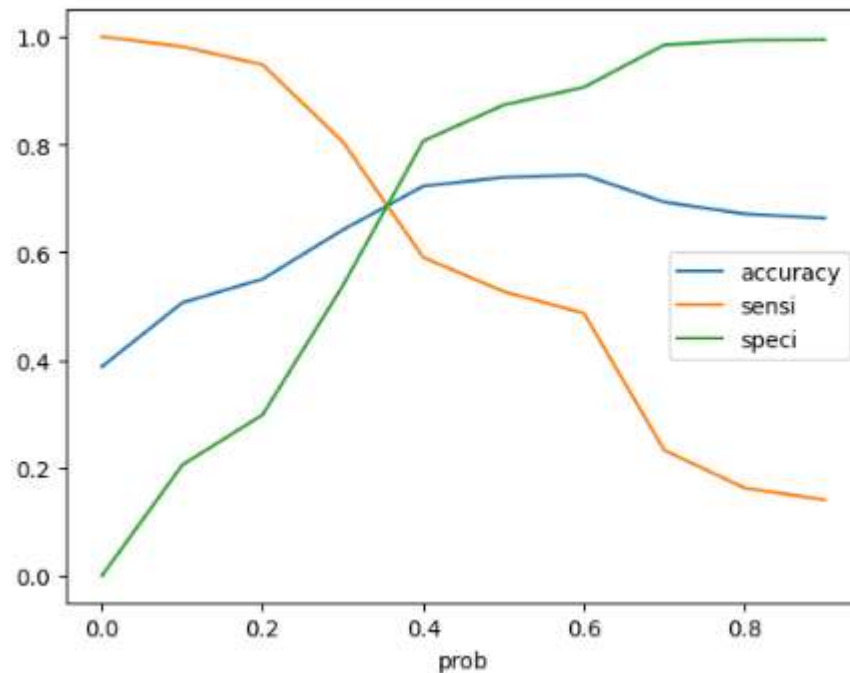
- The area under ROC curve is 0.78 which is a decent value



The area under ROC curve is 0.78 which is a decent value.

Model Evaluation: Sensitivity and Specificity of Train Data Set

- From the graph: optimal cut off is at 0.35



Model Evaluation: Sensitivity and Specificity of Test Data Set

- ▶ Accuracy=0.71
- ▶ Specificity=0.68
- ▶ Sensitivity=0.71

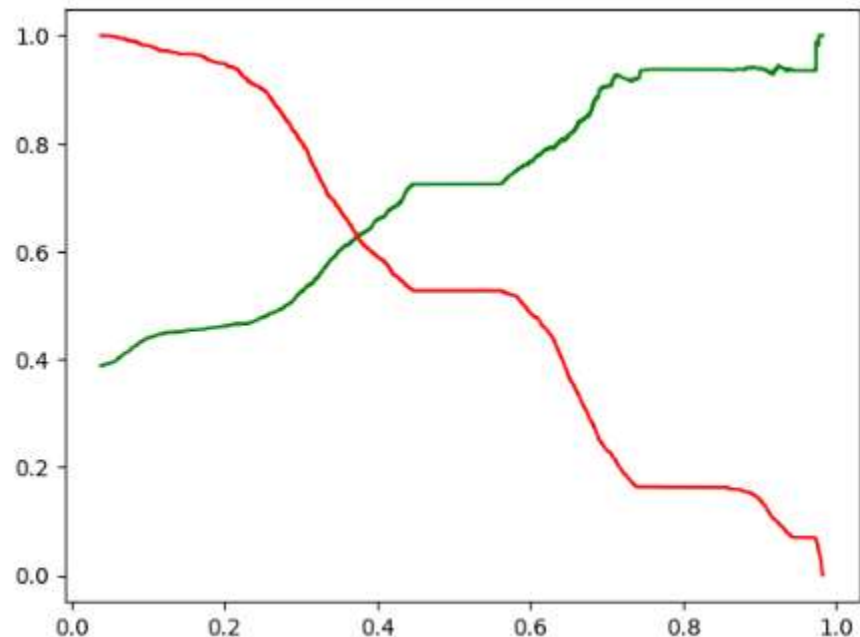
Model Evaluation: Precision and Recall of Train Data Set

- ▶ Precision = 0.72
- ▶ Recall = 0.52

Current cut off is 0.35, Precision around 72% and Recall around 52%

Model Evaluation: Precision and Recall tradeoff

- ▶ Accuracy= 0.72
- ▶ Precision= 0.67
- ▶ Recall= 0.58



Conclusion

- ▶ The optimal cut off is based on sensitivity and specificity of final prediction
 - ▶ For Test data set:
 1. Accuracy=0.76
 2. Sensitivity=0.68
 3. Specificity=0.71
- 