LEAD SCORING CASE STUDY

<SHREELEKHA>

<SHRINJOY>

<SHRAVANI>

THE PROBLEM

What is the problem?	• The company has very low lead conversion of only 30%.
Who has this problem?	• X Education (Online course platform).
Why should this problem be solved?	• This will increase the efficiency, productivity and overall sales of X education.
How will I know this problem has been solved?	• If the converted percentage has increase to closer to 80% conversion.

BUSINESS OBJECTIVE

- X Education online course selling platform needs to distinguish between hot and cold customers
- This distinction will help the company to reach the customers where there is maximum probability for conversion
- So, a model needs to be built and deployed that can easily identify hot leads, that X Education could target more effectively.

WORKABLE SOLUTIONS

Step #1

Data Cleaning and Data Manipulation

- Handling missing data
  - Columns containing >40% null values are dropped
- Deleting records
- Handling duplicate data
- Outliers handling

Step #2

EDA and Data Imputation

- Univariate & Bivariate Analysis
- Missing data handling in selected columns
- Dummy columns creation for categorical variables
- Charting various categorical data.
- Plotting heatmap to find correlation between variables

Step #3

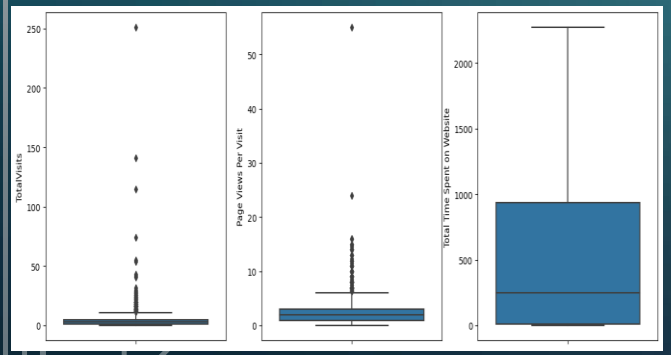
Model Building

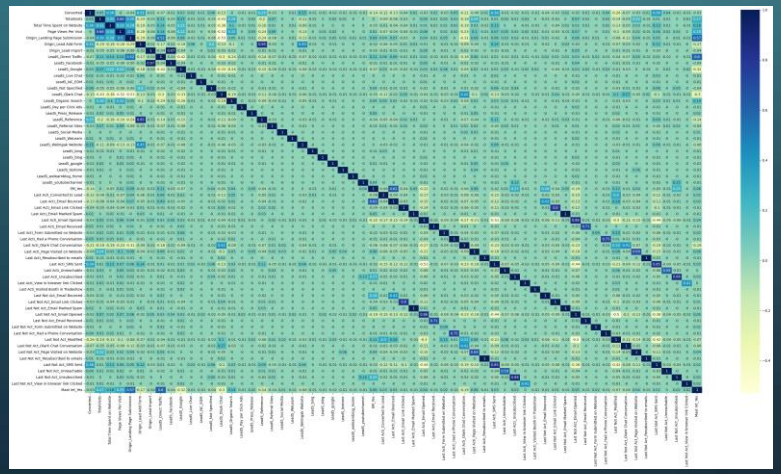
- Rescaling the data
- RFE to get 15-20 features
- Predicting probability of conversion
- Cutoff analysis
  - ROC Curves
    - Finding Sensitivity, Specificity, Accuracy scores
- Prediction on test data and comparison

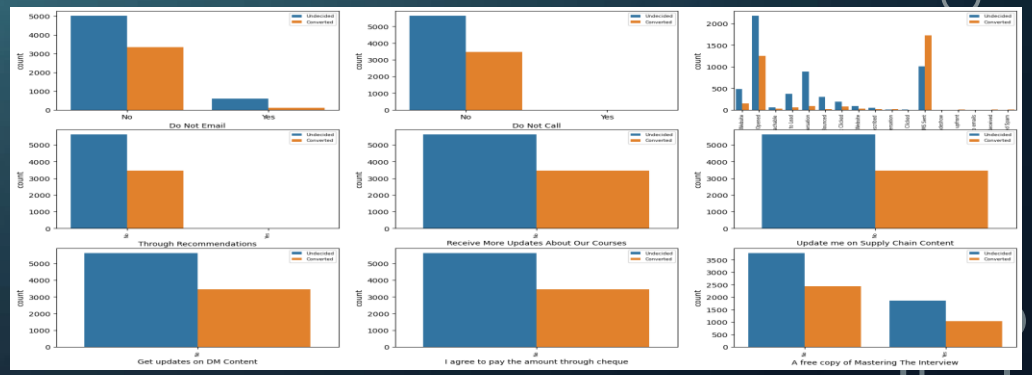
Note: Step 1 and 2 have some overlaps as it was necessary to understand the data before we took actions.

EDA AND DATA MANIPULATION

- Out of the total 37 variables or columns provided in the data 27 were dropped and only 10 were selected. Columns were dropped based on following criteria-
  - Very high volume of missing data (over 40%)
  - Even in the ones that were around 30% missing data, we dropped columns, because there was no clear way to assign some value/ label to missing data. Responses were too fragmented to assign any value to the column.
  - Some columns were dropped because there was only one specific response from all respondent (like India was the only country listed by everyone)
- About 1.5% of the records/ rows were also dropped because of-
  - Missing data
  - Outliers
- EDA helped us understand, key features of each column. Also in identifying broad patterns of correlations.







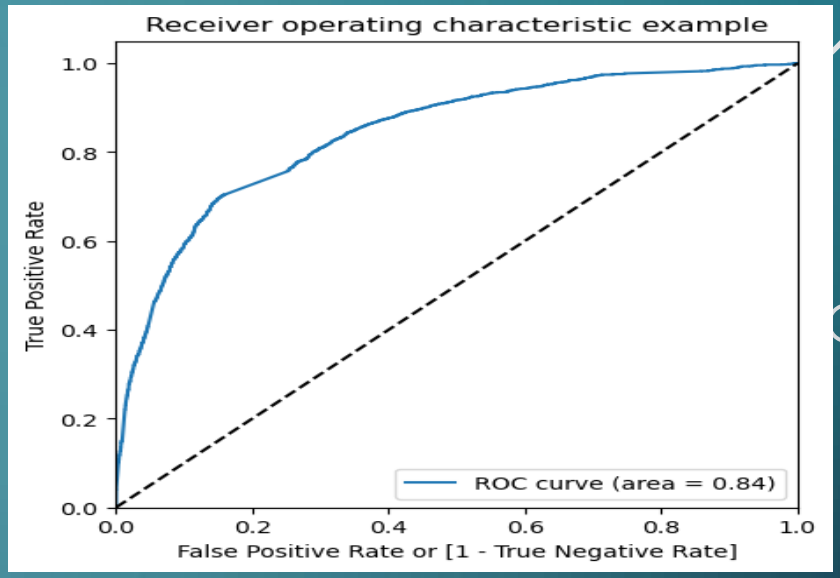
MODEL BUILDING

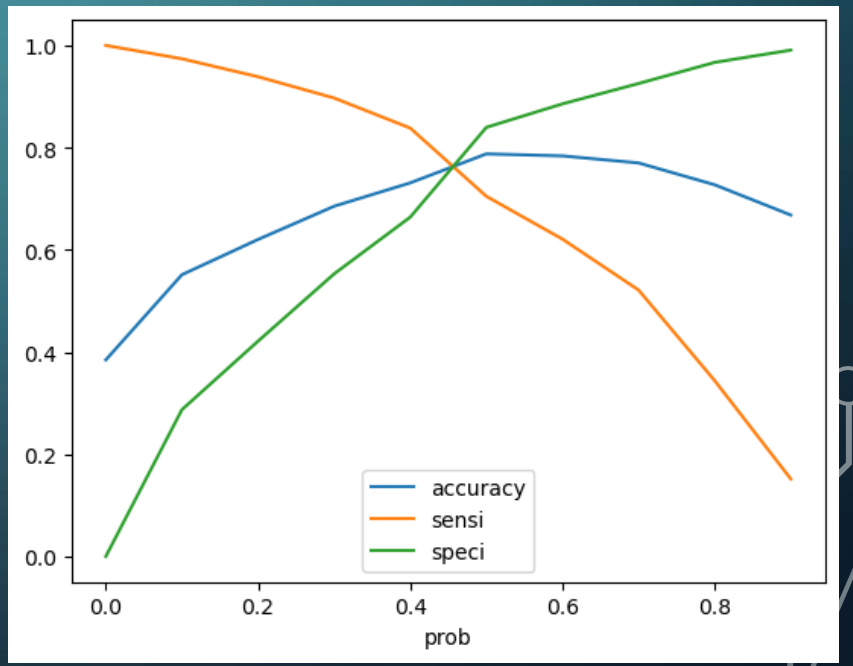
- Getting X and Y for train and test data
- After running RFE, we selected 15 out of the 61 columns, finalizing the model features by running VIF
- The accuracy with a 0.50 cutoff of 78.7%
- With the current model we are able to predict 70% (sensitivity) of the conversions correctly. Which is a good prediction. Although, it also means that we will be missing out on 30% of customers who are likely to convert. This is a bigger challenge in this particular case. Since conversions, per se, are so low, it is important that we do not miss out on any key customers here.

	Features	VIF
2	Page Views Per Visit	4.00
0	TotalVisits	3.53
1	Total Time Spent on Website	1.98
8	EM_Yes	1.82
10	Last Acti_Email Bounced	1.74
4	LeadS_Direct Traffic	1.41
13	Last Not Act_SMS Sent	1.37
3	Origin_Lead Add Form	1.33
6	LeadS_Welingak Website	1.27
9	Last Acti_Converted to Lead	1.05
5	LeadS_Referral Sites	1.03
11	Last Acti_Olark Chat Conversation	1.02
14	Last Not Act_Unreachable	1.01
7	LeadS_google	1.00
12	Last Not Act_Had a Phone Conversation	1.00

ROC CURVE

- The curve shows a sharp initial plot curve. This shows the ratio between TPR and FPR. The ROC curve area of 0.84 is an indication of a good fit of the model.
- We observe that a cutoff closer to 0.45 is more feasible. In this case we have taken 0.46 which is visible in 2nd graph.
- At this point, the correct prediction of conversions is as high as 79%.
- So, we have improved the prediction of conversions
- The overall accuracy of the model drops to 74.5% from 78.7%. Which is a better situation to be in as we are able to predict conversions better.





PREDICTION ON TEST DATA

- It was important to predict on the test data set, to know if we were getting a good fit even there.
- All the steps were followed and we started predicting the test set. The new prediction values were saved in a new dataframe.
- After this we did model evaluation i.e. finding the accuracy, precision, and recall
- This also shows that our model is stable and good accuracy, sensitivity/specificity

TESTING THE PROTOTYPE

Train Data:

- Sensitivity: 78.5
- Specificity: 72
- Accuracy: 78.9

Test Data:

- Sensitivity: 77
- Specificity: 71
- Accuracy: 73.3

FINAL RESULT

1. Firstly, the Model will help X Education with a score against each Lead. This will be key to identifying a lead as Hot or Cold. Leads with >0.46 cutoff, could be considered as hot Leads.
2. Secondly, based on the key drivers, X Education can focus on very specific marketing activity on certain websites or / and also decide means of targeting these lead (like specifically only tele -calling and no charting allowed)
   - The top 3 variables in our model which contribute highly to the lead generation are

     Total time spent on Website (3.4)

     Last Notable Activity of a Phone Conversation (2.9)

     Lead Source being Welingak Website (2.3)

With this, we can infer that more emphasis should be provided if there is more time spent on website and lead source as Welingak website. Also, its evident that only interested and high chance of conversion students opt for phone calls.

More promotions and campaigning can be concentrated for such students.