Lead Scoring Assignment

Group Name:

1. Alok Singh
2. Nishan Patel
3. Sridhar Kumar
4. Harshal Zele

Business Objective

- ✓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.
- ✓X Education gets a lot of leads, but its lead conversion rate is 30% which is very poor.

Objective -

- ✓Successfully identify set of leads, which can increase the lead conversion rate as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
- ✓The company wants to build a model wherein a lead score need to assign to each of the leads, such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

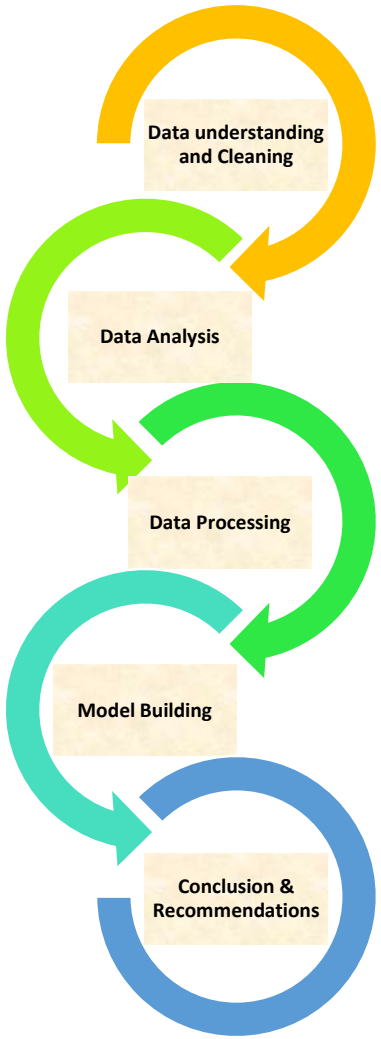
Problem solving methodology

Understanding the datasets

- ✓ 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.
- ✓ There are 9240 records of from the past.
- ✓ This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- ✓ The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Model Building-

- ✓ Using RFE for feature elimination.
- ✓ Logistic Regression using Stats Model



Data Cleaning and Analysis -

- ✓ Many columns dropped because of more than 3000 null values such as Asymmertrique score and index, Lead Quality, Tags etc.
- ✓ Many columns dropped because of no variation such as Magazine, Newspaper, News Paper Article.
- ✓ Imputations are required in variables such as Lead Source, Last Activity, Total Visites etc.
- ✓ Checking for Outliers as it impacts the modelling.

Conclusion –

- ✓ Need to select cut-off value of ~0.54 for 80% lead conversion.
- ✓ Based on business requirement, need to adjust the cut-off value for aggressive lead conversion.
- ✓ Based on business requirement, need to adjust the cut-off value for less aggressive lead conversion.

Data Cleaning:

- ✓ There are 5 types of columns dropped for model building.
  1. Variables which are unique identifiers such as **Prospect ID** and **Lead Number**

  2. Variables which has high number of null values such as **Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score, Lead Quality, Tags.** These columns has more than 3000 null values.

  3. Variables which has no variation in values, such as **Magazine, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque.**

  4. Variables which has 2 or 3 variations such as **X Education Forums, Newspaper Article, Newspaper**

  5. Variables which has more than 30% combined null and Select values such as **Lead Profile, How did you hear about X Education, How did you hear about X Education, Specialization**

Data Preparation:

- ✓ Imputed Columns –
  - ✓ **What is your current occupation & Last Activity –** Replace Null values with Others.
  - ✓ **TotalVisits, What matters most to you in choosing a course, & Page Views Per Visit** – Imputing null values with median.
  - ✓ **Lead Source** – Imputing null values with Google as it is very popular.

- ✓ Columns converted for model building.
  - ✓ Converting variables which has values as Yes and No. Such as **Search, Digital Advertisement, Through Recommendations, Do Not Email, Do Not Call, A free copy of Mastering The Interview.** Replaces values with 1= Yes and 0=No.

- ✓ Outlier Treatment –
  - ✓ Variables such as **TotalVisits, Page Views Per Visit, & Total Time Spent on Website** has considerable outliers.
  - ✓ We treat them such that our data set reduces to 8679 rows.

Feature Elimination - RFE

- ✓ Dummy Variable Creation –
  - ✓ There are Categorical variables which needs dummy variable creation such as **Lead Origin, Lead Source, Last Notable Activity, Last Activity, What is your current occupation, What matters most to you in choosing a course**
  - ✓ Post Dummy Variable creations we have 73 columns.

- ✓ Rescaling the Variables-
  - ✓ There are variables which needs scaling such as **Total Visits, Total Time Spent on Website, Page Views Per Visit**
  - ✓ We are using MinMax Scaler for scaling.

- ✓ We used RFE for feature elimination.
  - ✓ Using Logistic regression and 15 features selection.

Model Building

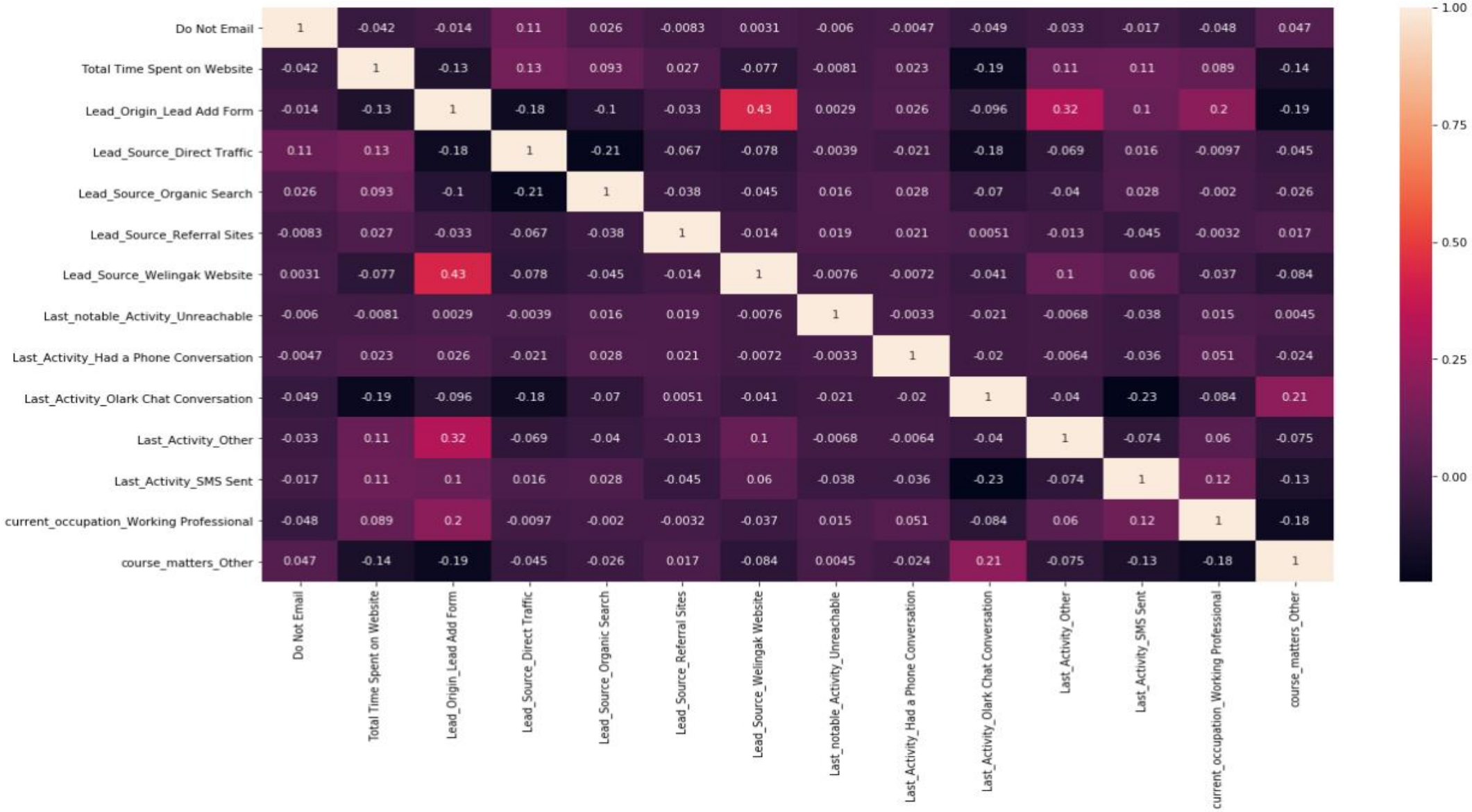
- ✓ Using StatsModel GLM for model building.
  - ✓ It is used to check the statistics for model fine tuning.
  - ✓ Using Binomial family, since predictor is binary.

- ✓ Model-1 Analysis
  - ✓ P-value of **current_occupation_Housewife** has 0.999 which is more than 0.05.
  - ✓ Hence it is insignificant in presence of other variables. Hence we need to drop it.
  - ✓ Accuracy of model is 80.41

- ✓ Model-2 Analysis
  - ✓ Accuracy of Model-2 after dropping insignificant variable marginally changed to 80.36.
  - ✓ P-value of **all variables is less than 0.05.**
  - ✓ Hence we need to check the multi collinearity. (VIF) of model-2.
  - ✓ VIF values of all variables are less than 5 except constant. We will drop `const` variable now.

- ✓ Conclusion
  - ✓ We conclude that Model-2 has better accuracy, P-values and VIF statistics.

Model Variables- Heatmap

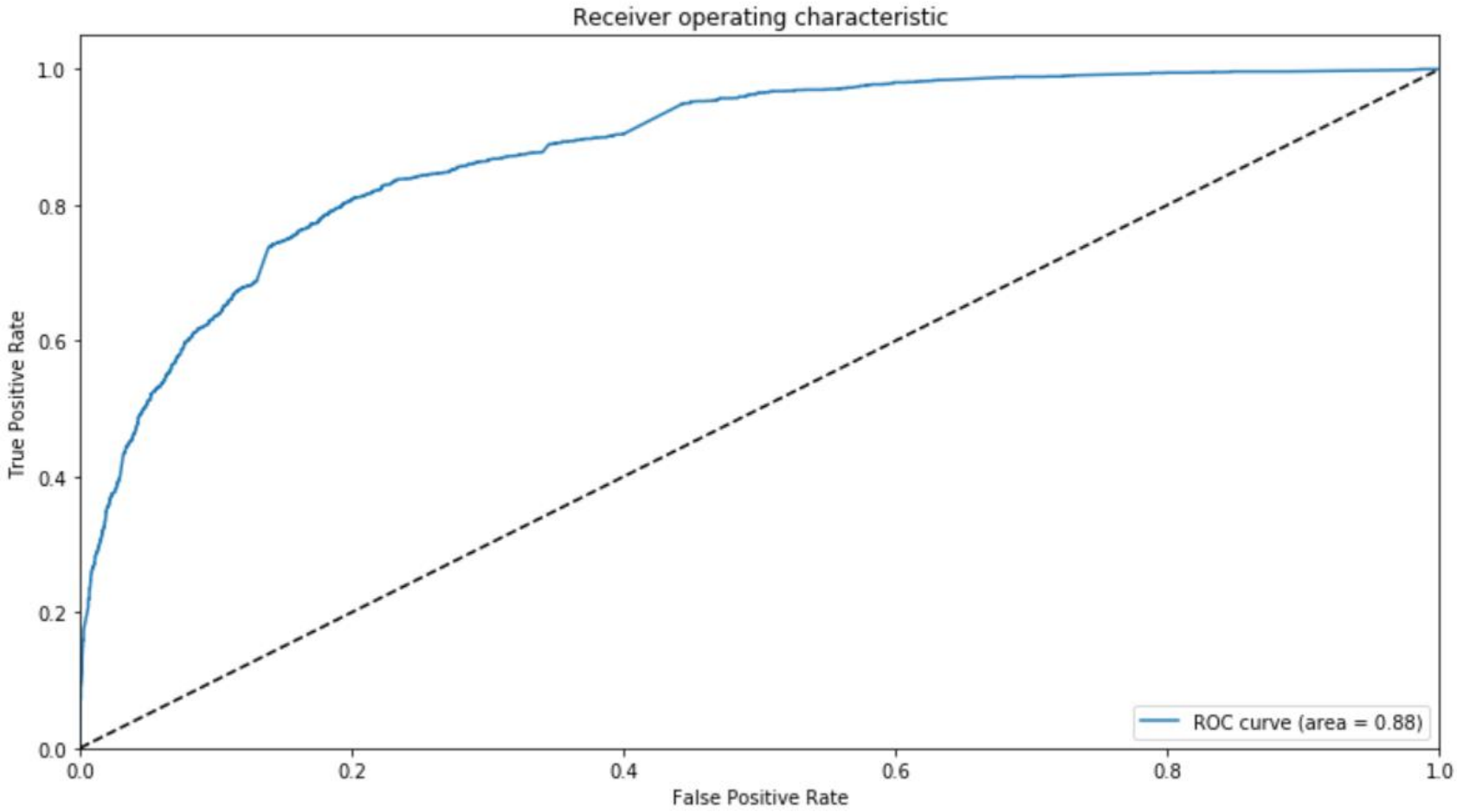


Top Variables Of Model

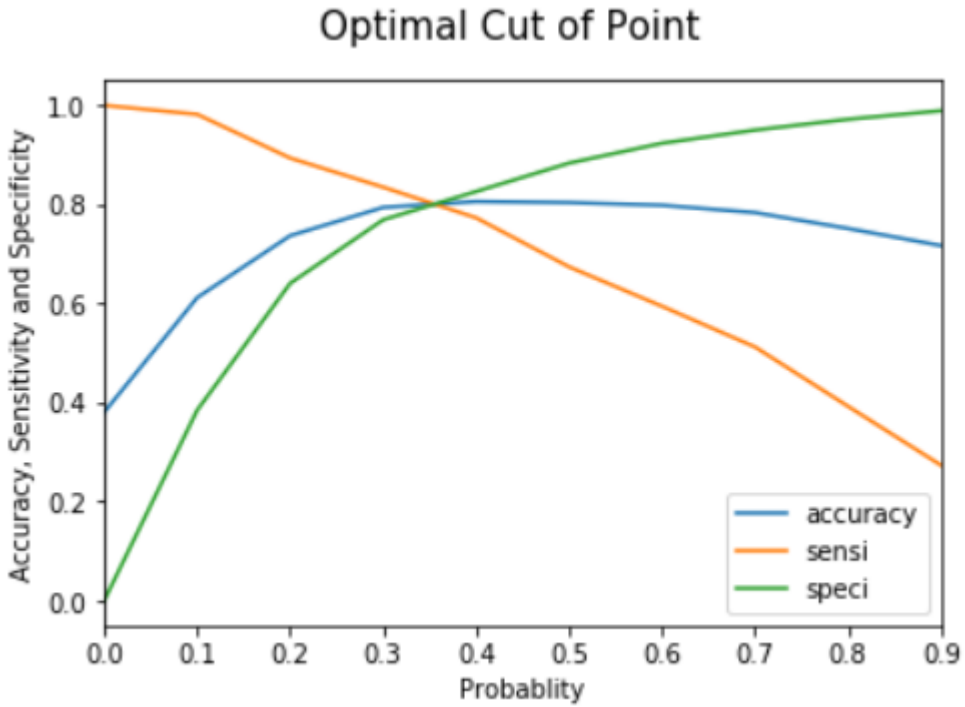
- ✓ **Top 2 Numeric variables**
  - ✓ Total time spent on Website.
  - ✓ Total Visits

- ✓ **Top 3 Categorical/Dummy variables**
  - ✓ Lead_Origin_Lead Add Form
  - ✓ Current_occupation_Working Professional
  - ✓ Lead_Source_Welingak Website

ROC of Model (0.88)

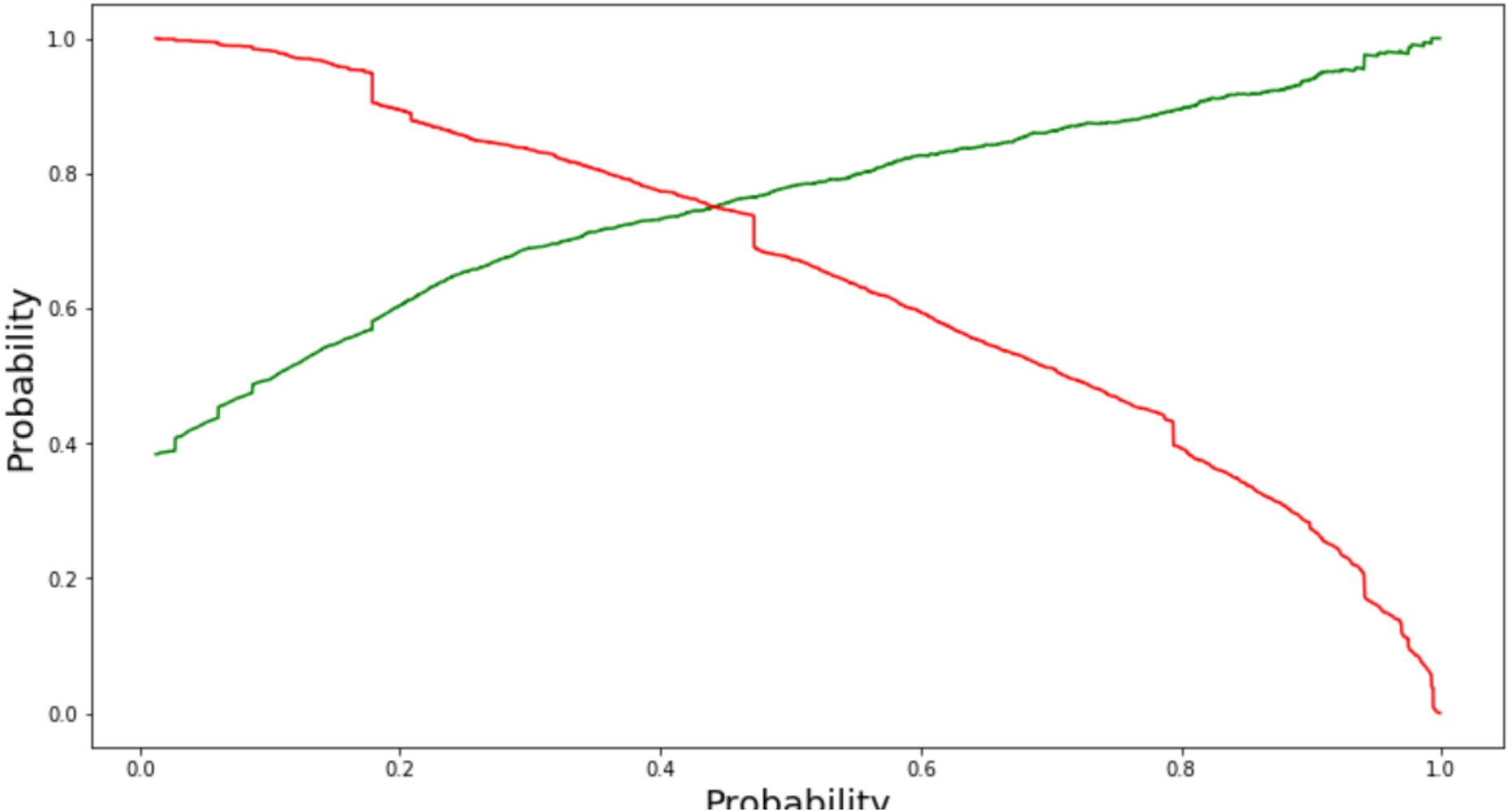


Model - Optimal Cut-off



- ✓ ~0.35 Probability is optimum point where Accuracy, Sensitivity and Specificity is optimum.
- ✓ X-Education CEO requirement is to have ball park 80% conversion Rate
- ✓ To achieve it, we need set the Cut-off Probability to ~0.55

Precision and Recall Trade Off



Model Evaluation beyond accuracy

- ✓ Our Model Accuracy is 79.56%. Now Let's check other Metrics when we want to achieve 80% lead conversion rate (i.e. precision). This is on test data.

- ✓ Confusion Matrix

Predicted / Actual	Not Converted	Converted
Not Converted	1429	154
Converted	378	643

- ✓ Metrics

Metrics	
Sensitivity	62.97
Specificity	90.27
FPR	9.72
Precision	80.67

Subjective Questions - 3

- ✓ X-Education is hiring 10 interns. So that they want to make lead conversion more aggressive.
- ✓ Based on this, we will set the Cut-off to 0.10

- ✓ Confusion Matrix

Predicted / Actual	Not Converted	Converted
Not Converted	664	919
Converted	23	998

- ✓ Metrics

Metrics	
Sensitivity	97.74
Specificity	41.94
FPR	58.06
Precision	52.06

Subjective Questions - 4

- ✓ X-Education want to go slow when target is achived.
- ✓ Based on this, we will set the Cut-off to 0.75

- ✓ Confusion Matrix

Predicted / Actual	Not Converted	Converted
Not Converted	1520	63
Converted	533	488

- ✓ Metrics

Metrics	
Sensitivity	47.79
Specificity	96.02
FPR	3.97
Precision	88.56

Factors Causing Conversion

- ✓ Factors Causing Conversion positively
  - ✓ People spending time on website looking for courses.
  - ✓ People who add form to show interest in courses.
  - ✓ People who visit Wellingak Website.
  - ✓ People with whom had a phone conversation to explain about courses.
  - ✓ People with whom had a conversation over SMS.
  - ✓ People who are working Professionals.

- ✓ Factors Causing Conversion negatively
  - ✓ People who opts out of mail promotions.
  - ✓ People having conversation on Olark Chat.
  - ✓ People directly coming to X-Education.
  - ✓ People who do not respond to SMS or phone Calls.

Recommendations

- ✓ Recommendations.
  - ✓ Improve X-Education Website
    - ✓ To get right information easily.
    - ✓ More Engaging.
    - ✓ Encourage to submit the form.
  - ✓ Constant communications
    - ✓ Telephonic conversation to explain about courses.
    - ✓ Follow up SMS
  - ✓ People who are working Professionals.
    - ✓ Design the courses which suits to working professionals.
  - ✓ People who are showing less interest, can be characterized,
    - ✓ Opts out of promotions.
    - ✓ Stop communications.
    - ✓ Chatting with Olark

Conclusions

- ✓ X-Education uses 0.35 cut-off to obtain model accuracy of 79.56%
- ✓ Lead conversion rate of 80.67 percent (Precision) is achieved, as requested by CEO
- ✓ Lead score (or hotness) is assigned to each row of data, such that a higher lead score implies a higher conversion chance and vice versa.
- ✓ Model is stable, which can be seen in correlation matrix between features, which shows correlation is very less among different features. Thus multicollinearity is handled.
- ✓ Model is able to handle different requirements of low Type-II error or low Type-I error using different cut-off points respectively.
- ✓ People spending time on website,