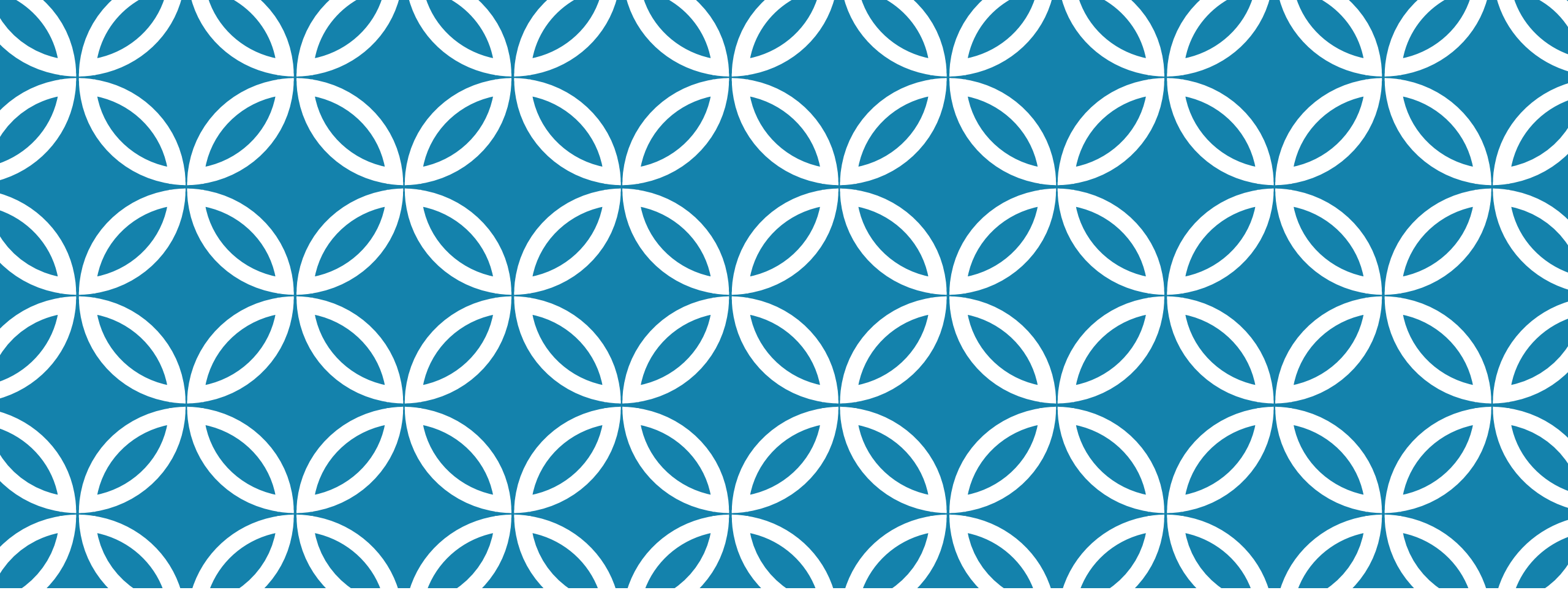


LEAD SCORING CASE STUDY

S S B Phani Pradeep Miriyala
Chandra Sekhar Reddy Kandimalla

AGENDA

Business goal and Approach

Data Preparation

Data Cleaning

Model Development

Deriving Optimal cut-off point

Final Model and Prediction on test data

Conversion Rate calculation

Adding a column for lead score

Problem Statement

An education company named X Education sells online courses to industry professionals. 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.

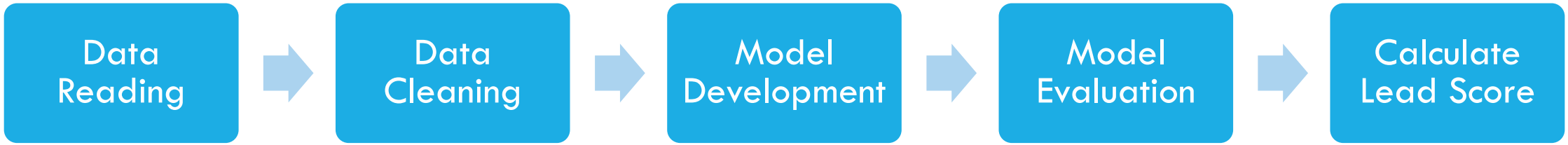
Now, although X Education gets a lot of leads, its lead conversion rate is very poor.

It is desired to calculate lead score ranging from 0 to 100 in order to identify hot leads.

BUSINESS GOAL AND APPROACH

Business Goal

1. Build a model which will assign a lead score 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
2. Help company to select the most promising leads, i.e. the leads that are most likely to convert into paying customers
3. Ballpark of the target lead conversion rate to be around 80%.



DATA READING

The raw data set has 9240 rows and 37 features with

- 30 Categorical features
- 8 Numeric and object (unique IDs) features

Imbalance percentage is 38% (Converted = 1)

There are some data columns in dataset that have values “Select”. These are as good as “Null” values. Before proceeding with data cleaning, this must be corrected.

These columns are:

- Specialization
- How did you hear about X Education
- Lead Profile
- City

```
In [4]: # Inspect data for null values
        data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object
 11  Country                                        6779 non-null   object
 12  Specialization                                 7802 non-null   object
 13  How did you hear about X Education             7033 non-null   object
 14  What is your current occupation                6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                         9240 non-null   object
 17  Magazine                                       9240 non-null   object
 18  Newspaper Article                              9240 non-null   object
 19  X Education Forums                             9240 non-null   object
 20  Newspaper                                      9240 non-null   object
 21  Digital Advertisement                          9240 non-null   object
 22  Through Recommendations                        9240 non-null   object
 23  Receive More Updates About Our Courses         9240 non-null   object
 24  Tags                                           5887 non-null   object
 25  Lead Quality                                   4473 non-null   object
 26  Update me on Supply Chain Content              9240 non-null   object
 27  Get updates on DM Content                      9240 non-null   object
 28  Lead Profile                                   6531 non-null   object
 29  City                                           7820 non-null   object
 30  Asymmetrique Activity Index                    5022 non-null   object
 31  Asymmetrique Profile Index                     5022 non-null   object
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object
 35  A free copy of Mastering The Interview         9240 non-null   object
 36  Last Notable Activity                          9240 non-null   object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

```
List of columns with entry "Select" :  ['Specialization', 'How did you hear about X Education', 'Lead Profile', 'City
']
```

DATA CLEANING

Begin by dropping all columns who have missing values of more than 40% of times. After dropping such columns, we are left with 30 columns.

Categorical variable cleaning steps

- If category has missing values, create new category for missing values.
- In categorical variable values, there are repeated values with minor spelling variations. Thus they can be clubbed together. E.g.: Clubbing “google”, “Google”
- After clubbing such entries, group categories whose number of occurrences of a category fall below certain threshold, we will group them to create new category named “Others”.

Columns with very low variance

- Some columns have almost no variance due to the column being almost constant across entire dataset. Such columns have no use in model preparation. Some of such columns are:
  “X Education Forums”, “Newspaper”, “Digital Advertisement”, “Do not call”, “Through Recommendations”, “Receive more updates about our courses” etc.,

Columns with very high variance

- Columns “Prospect ID” and “Unique ID” are unique for every row in data set. This high variance will not have any impact on conversion. It would be safe to ignore them for model preparation.

DATA CLEANING

Conversion of binary categorical variables to numeric

- Some categorical variables have just two levels “Yes”, “No”. These can be converted to numeric by replacing “Yes” with 1 and “No” with 0 respectively.

Outliers for numeric variables

- Outliers are removed from data set. Outliers are identified as values above 99 percentile of values from TotalVisits column. After removing outliers, number of rows left are 9157. There is not much data lost.

Imputing missing values for numeric variables

- The columns “TotalVisits” and “Page Views Per Visit” will be imputed with rounded integer value of mean level.

Creation of dummy columns for categorical columns

- Dummy columns are created for categorical columns
- Original columns are dropped after dummy columns are created.

Data size after cleaning operations is 9157 x 73 columns

DATA MODELING

Model is split in to train and test sets with ratio of 70:30.

Random state for model is chosen as 100.

Training data is scaled using StandardScaler

Build a logistic regression model using all features (73)

After reviewing the built model, some features have high P values. To eliminate these columns, we apply RFE technique.

Run RFE to identify which columns have support.

Rebuild model using the columns that have support.

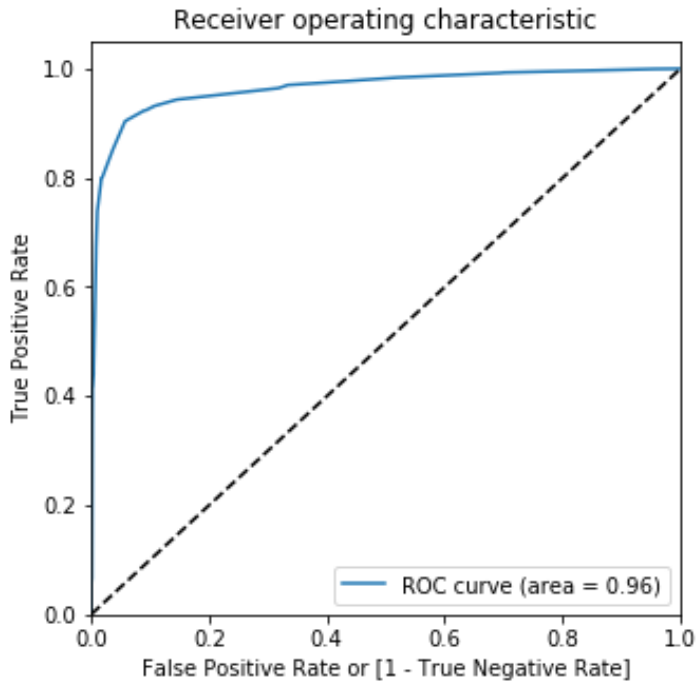
Select cut off point as 0.5

METRICS – TRAINING DATA, CUT-OFF = 0.5

Metrics

Metric	Value
Confusion Matrix	[3705 224] [242 2238]
Accuracy	92.73%
Sensitivity	90.24%
Specificity	94.3%
Precision	90.9%
Recall	90.24%

ROC Curve



OPTIMAL CUT-OFF POINT

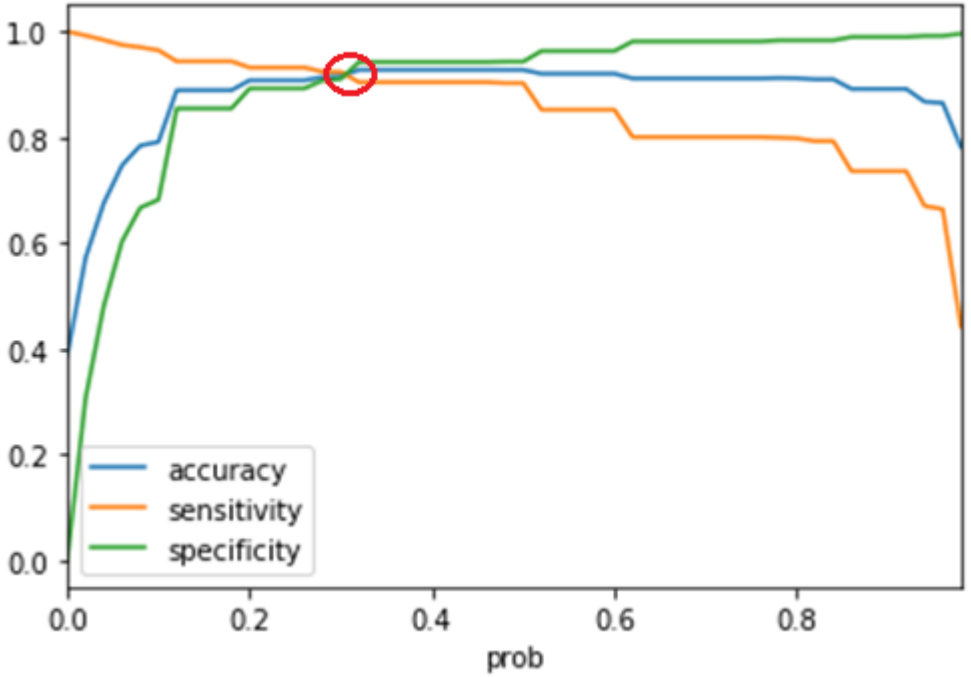
Procedure

In previous slide, the metrics were calculated using an arbitrary cut-off point 0.5.

This cut-off however is not accurate.

To select optimal cut-off point, we will plot Accuracy, Sensitivity, Specificity values for different cut-off probability values.

Accuracy – Sensitivity – Specificity Curve



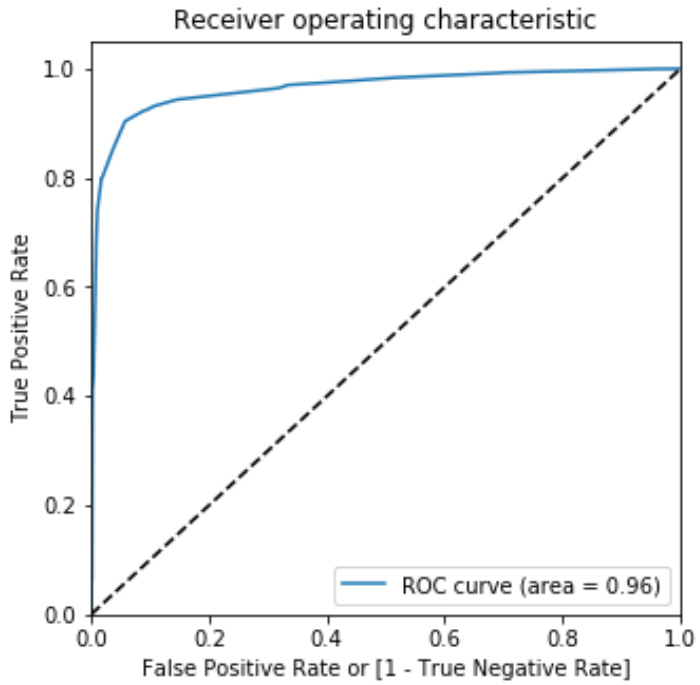
From the plot, optimal Cut-off point is approximately 0.3

METRICS – TRAINING DATA, CUT-OFF = 0.3

Metrics

Metric	Value
Confusion Matrix	[3576 353] [190 2290]
Accuracy	92.73%
Sensitivity	92.34%
Specificity	91.02%
Precision	86.64%
Recall	92.34%

ROC Curve

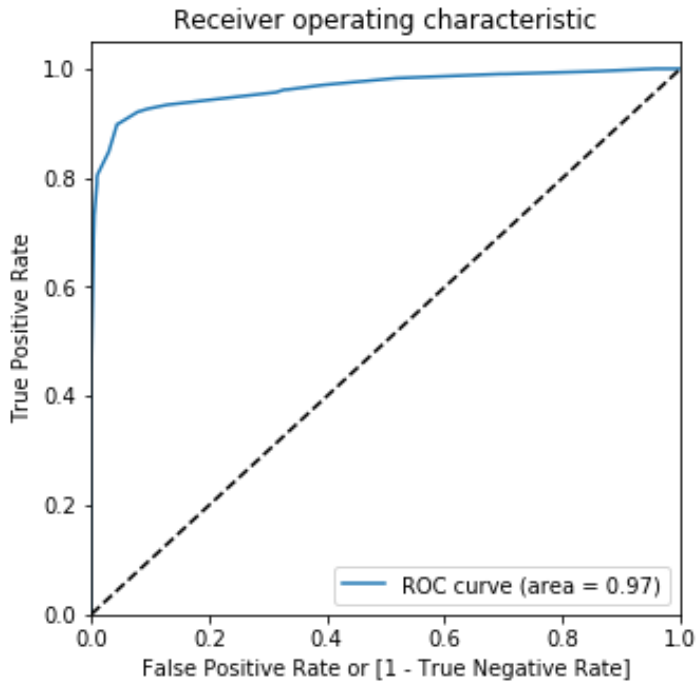


METRICS – TEST DATA

Metrics

Metric	Value
Confusion Matrix	[1567 136] [82 963]
Accuracy	92.07%
Sensitivity	92.15%
Specificity	92.01%
Precision	87.63%
Recall	92.15%

ROC Curve



CONVERSION RATE

Conversion rate will be ratio of number of people really converted to number of people predicted to be converted.

Thus this is ratio of True-Positives to Total Positives predicted.

- Converted people = True Positives
- Total hot leads = Total Positives (TP + FP)

Thus, conversion rate will be 87.6 %

LEAD SCORE CALCULATION

The regression model calculates a probability in range of 0 to 1.

Multiply the probability by 100 to get a lead score.

Thus, the cut-off point now will change to 30. (Cut off probability = 0.3)