# Lead Scoring Case Study

**Praveer Tiwari**
**Prerna Raviraj**

# PROBLEM STATEMENT

In the following case study a company named X Education wanted to increase their lead conversion rate in order to increase the efficiency and reduce wasting energy of the workforce.
Earlier the conversion rate was around **30%** which is needed to be pushed to **80%.**
Hence we need to analyse the data of leads that was collected of about 9000 data point and find those leads which have high chances of conversion called “hot leads”
The final leads filtered out from the data must have a 80% chance of conversion.

# ANALYSIS APPROACH

**Data Cleaning and Preparation:**

Deleted those columns which have only one unique value and hence are not useful in prediction

'Magazine', 'Receive More Updates About Our Courses' , 'Update me on Supply Chain Content' , 'Update me

Some null rows were removed

Imputing of null values with median in TotalVisits column was done

**Dealing with select and null values:**

Created a separate category for null and select values: called unknown

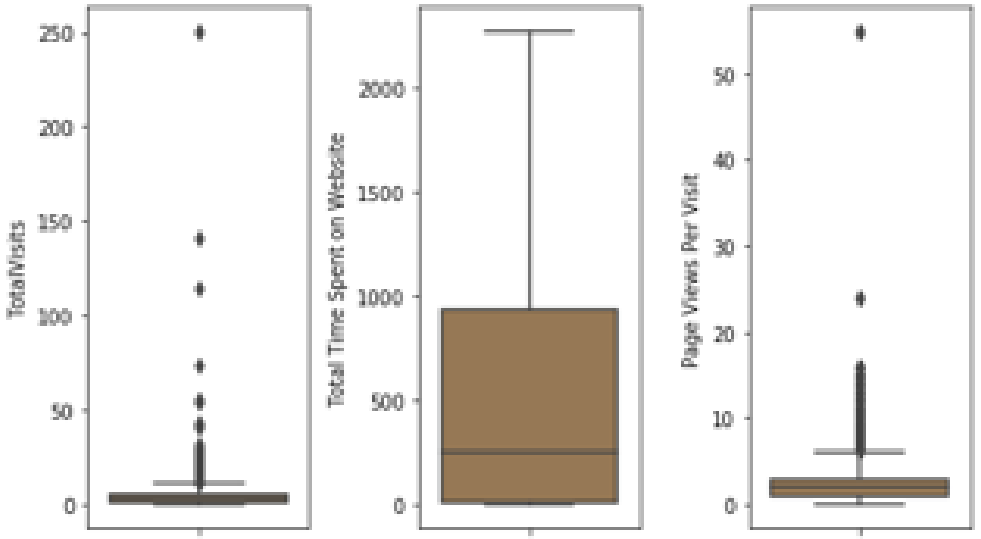
Hence increased the data recovery in columns like

Asymmetrique Activity Score, Asymmetrique Profile Score etc.

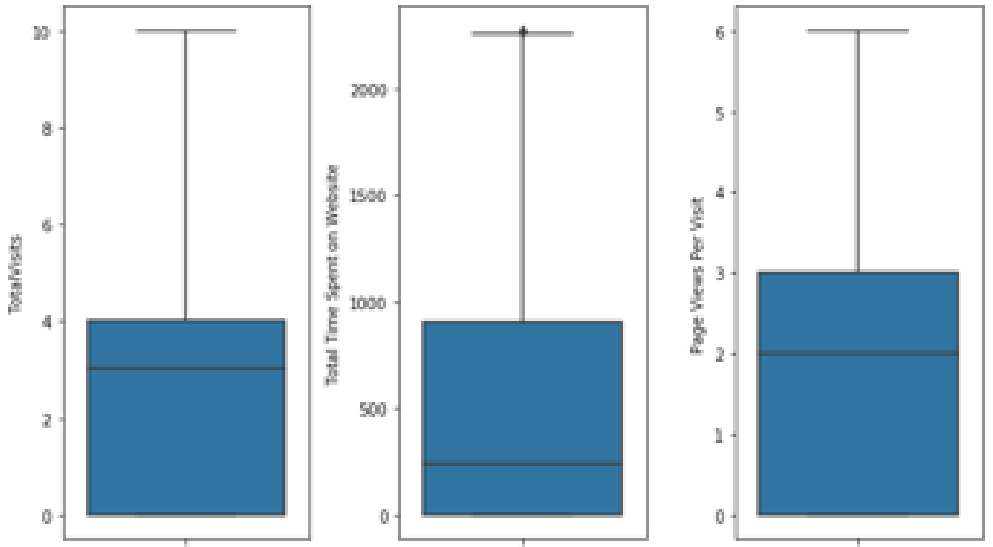
Binary encoding was done to make features easy for calculation

**Treating the outliers:**

Outliers were removed from columns like 'TotalVisits','Total Time Spent on Website','Page Views Per Visit'




before

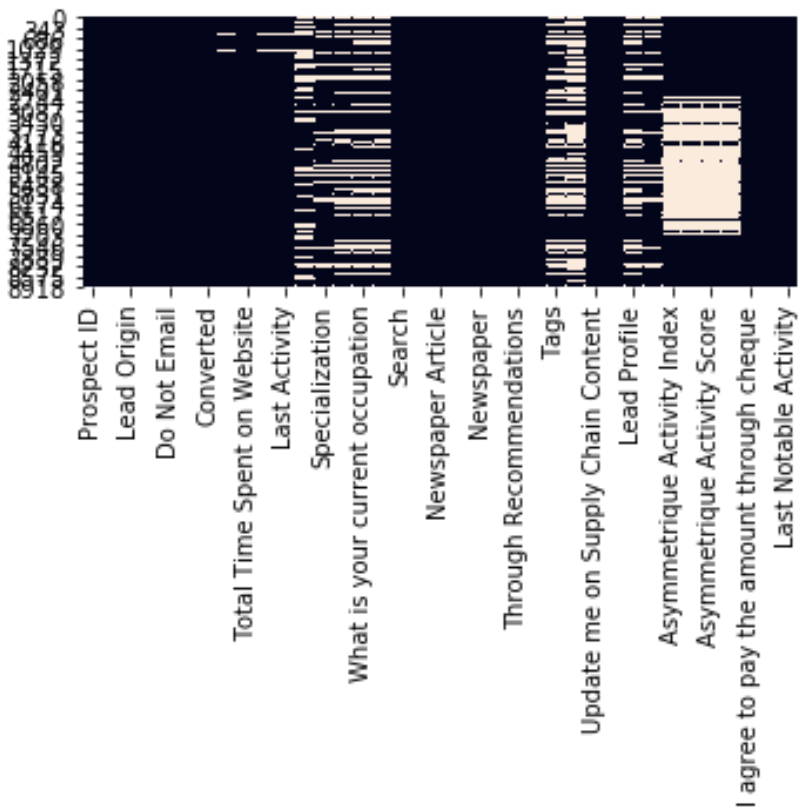



after

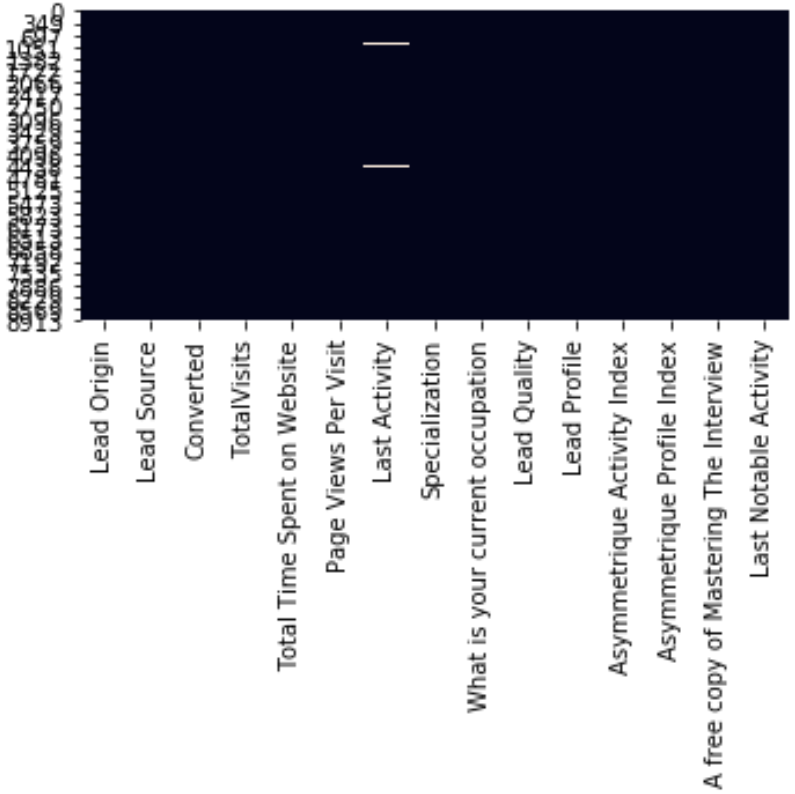
# VISUALIZATION OF THE NULL VALUES AFTER CLEANING AND DATA PREPARATION

White lines represent null values

## BEFORE




## AFTER



## FEATURE SCALING:

Scaling of these features were done using standard scaler
‘TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit’

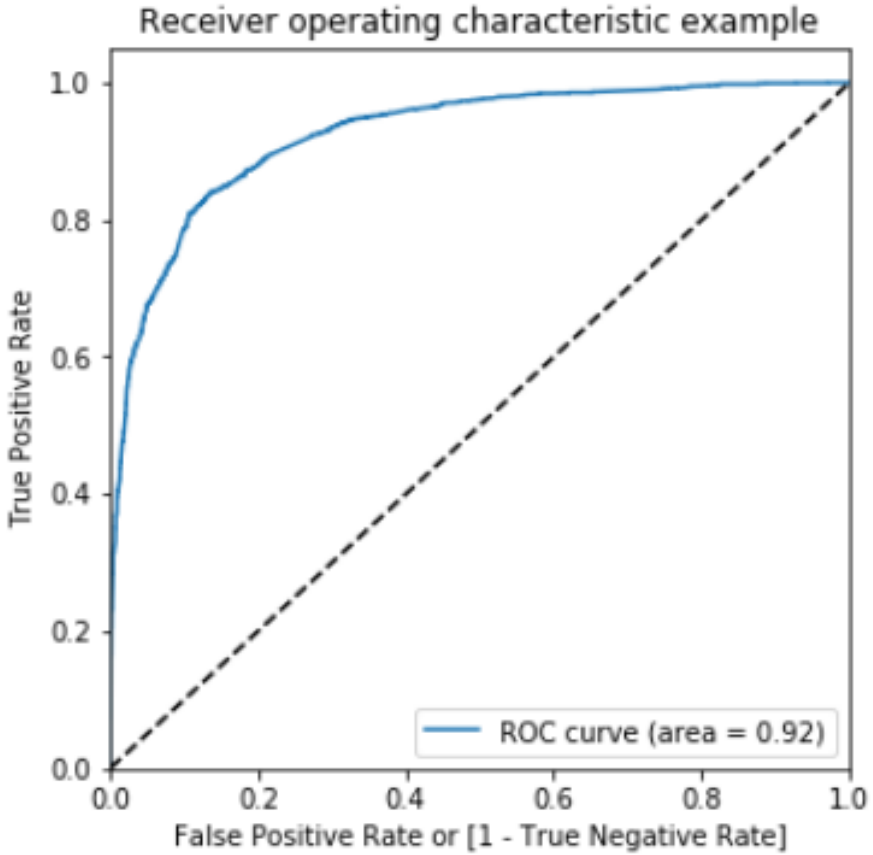
## DUMMY VARIABLES:

Firstly columns with object datatype were selected then , columns with high value of UNKNOWNS were selected and the unknown values were removed from it as there was no need for that data

# MODEL BUILDING

TEST –TRAIN split was performed on the data(70-30%)
We have used Logistic Regression model to analysis the data.
RFE was performed on x train and y train datasets to get the most useful features .We selected 15 features for our analysis.
After filtering columns using RFE and fitting this data in the model,
VIF (variance inflation factor) was calculated.
After getting desired low VIF Values we First checked the model accuracy using 0.5 as optimal threshold then moved to ROC for finding the appropriate optimal value .

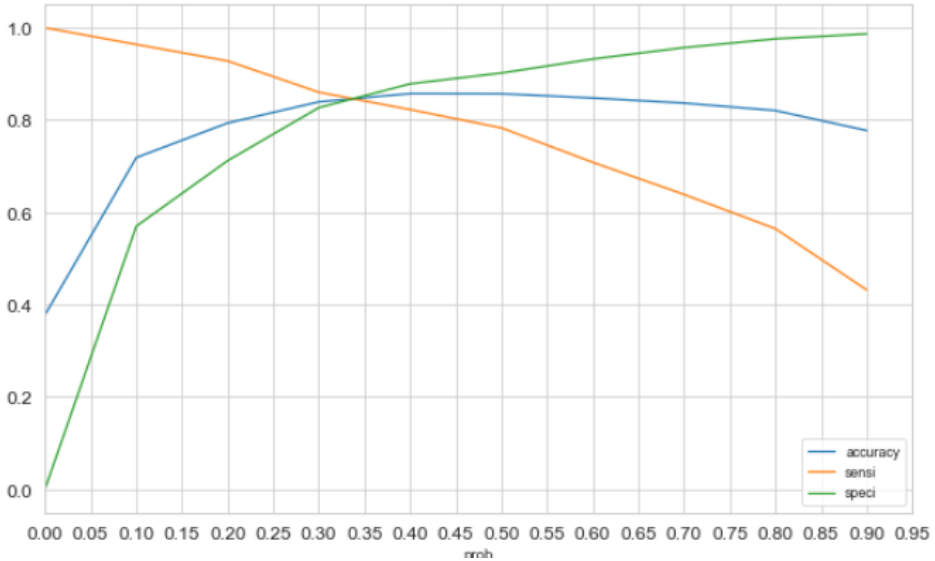
# ROC Curve and AUC Value



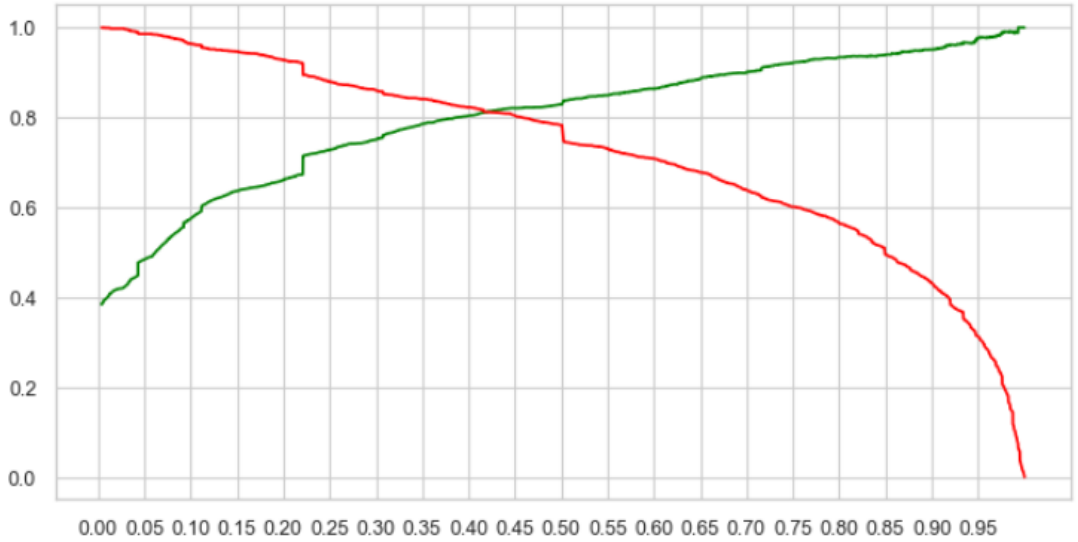

The ROC curve gives us the tradeoff between sensitivity and specificity. This is a good model as the ROC curve almost touches the upper left corner. When we calculate the Area Under the Curve (AUC), we get its value to be 0.92 which is very high. Hence this is a good model.

# OPTIMAL PROBABILITY THRESHOLD

## Using Sensitivity Specificity tradeoff




## Using Precision Recall tradeoff




The optimal threshold is calculated to be 0.33 with the sensitivity specificity tradeoff and around 0.41 with the precision recall tradeoff. We choose 0.33 as the final threshold value as the sensitivity specificity as well as accuracy are above 85%.

# EVALUATING MODEL ON TRAIN DATA SET

CONFUSION MATRIX

| Actual \ Predicted | Not converted | Converted |
|---|---|---|
| Not converted | 3142 | 550 |
| Converted | 347 | 1892 |

The model was first trained on the train data set. Then prediction were made on the test data set. We used the threshold value as 0.33 which was calculated.

The parameters when calculated giving the following values

Sensitivity= 0.84

Specificity=0.85

Accuracy=0.84

# PREDICTIONS AND EVALUATION OF TEST DATA SET

CONFUSION MATRIX

| Actual \ Predicted | Not converted | Converted |
|---|---|---|
| Not converted | 1566 | 34 |
| Converted | 525 | 418 |

The parameters when calculated giving the following values

Sensitivity= 0.44
Specificity=0.97
Accuracy=0.78

Final Test Data Set

| | Converted | Conversion_Prob | final_predicted |
|---|---|---|---|
| 0 | 0 | 0.27 | 0 |
| 1 | 0 | 0.02 | 0 |
| 2 | 0 | 0.02 | 0 |
| 3 | 1 | 0.42 | 1 |
| 4 | 0 | 0.00 | 0 |

Given above is the final prediction made by the model. We can see that in the test model there is a high difference between sensitivity and specificity. These values can be adjusted by changing the threshold based in requirements. Accuracy of the model is good at 78%

# LEAD SCORE CALCULATION

Lead scores for top ten records

| | Lead Number | Conversion_Prob | Converted | final_predicted | Lead_Score |
|---|---|---|---|---|---|
| 0 | 660737 | 0.17 | 0 | 0 | 17 |
| 1 | 660728 | 0.33 | 0 | 1 | 33 |
| 2 | 660727 | 0.96 | 1 | 1 | 96 |
| 3 | 660719 | 0.05 | 0 | 0 | 5 |
| 4 | 660681 | 0.72 | 1 | 1 | 72 |
| 5 | 660680 | 0.05 | 0 | 0 | 5 |
| 6 | 660673 | 0.82 | 1 | 1 | 82 |
| 7 | 660664 | 0.05 | 0 | 0 | 5 |
| 8 | 660624 | 0.08 | 0 | 0 | 8 |
| 9 | 660616 | 0.13 | 0 | 0 | 13 |

The train and test data set along with the Lead numbers are concatenated into one data frame.
The formula used to calculate lead score is as follows:
Lead score= conversion probability x 100
Any lead score above 34 will be predicted as a yes and below it will be a no, as our threshold value is 0.33.

# INFERENCE

Our final model has the following features-
15 features have been used for prediction of lead conversion
All variables have p values lesser than 0.99
All VIF values are below 1.79 meaning hardly any multicollinearity
The probability threshold is 0.33
The overall accuracy of the model is 78%

The value of beta predicts which variable has maximum influence on lead conversion. The graph on the left shows
that the variables ‘what is your current occupation_housewives’, lead source_welingak website and lead quality_high relevance have the highest positive effect on lead conversion. The variables  asyymetrique activity index_03 low, lead quality_worst and last activity_email bounced have maximum negative influence in lead conversion