LEAD SCORING CASE STUDY – SUMMARY REPORT

X Education wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires wants a model to 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

The required target lead conversion rate is around 80%.

Followed the below steps to arrive at the final result.

1 .Read and understood the given leads data set.

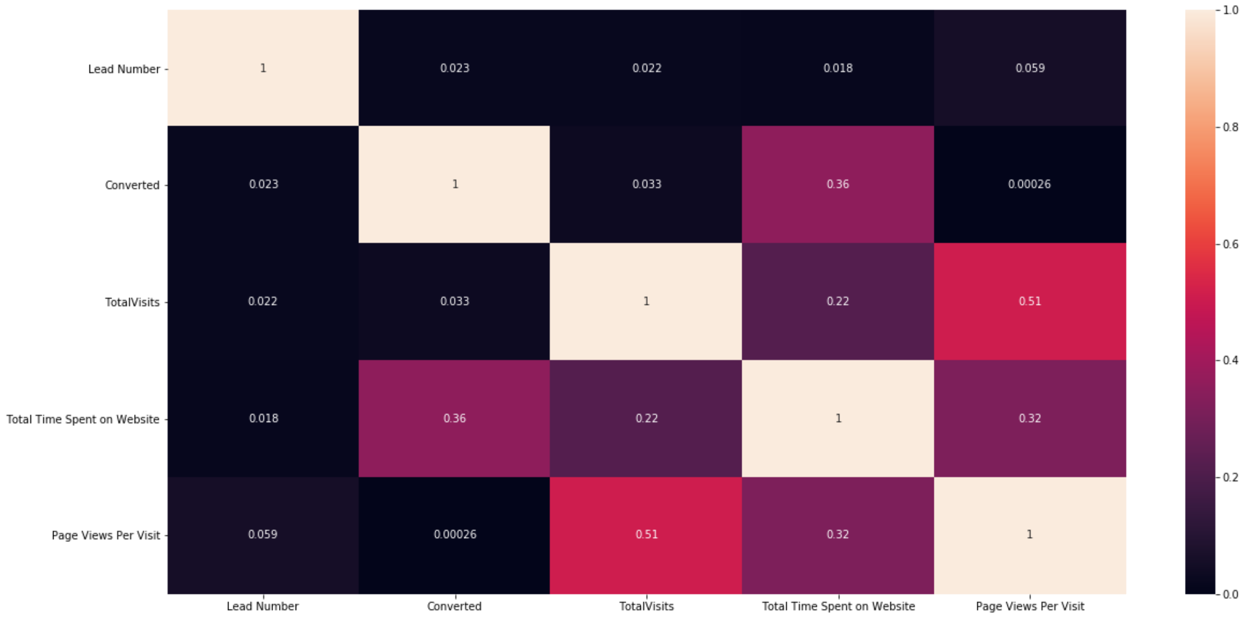
Number of features	37
Number of Records	9240
Presence of outliers	Yes
Presence of null values	Yes

2. Cleaned the data

- Checked for duplicate records
- Replaced 'Select' with null
- Dropped the columns with more than 30% of null values
- Dropped the columns which have only one value as it will not be useful for analysis or model building
- Dropped the columns with no significant variance in the values.
- Deleted the rows with null values

3. Visualized the data

- Created pair plots for all the numeric variables in the data set
- Created box plots for all the numeric features to check for the outliers in the data set
- Created a heat map to understand the correlation between the variables



4. Data Preparation

- For two level categorical variables assigned 1 to ‘Yes’ and 0 to ‘No’
- Created dummy variables for categorical variables with more than two levels
- The total number of columns after the above steps is 68

5. Split the data into train and test sets

- Split the data into train and test sets in the ratio of 70:30
- Removed the outliers in the data based on TotalVisits column

6. Scaling

- Scaled all the numeric variables using standard scaler

7. Conversion Rate

- The conversion rate of the given data set is 37.85541106458012

8. Model Building

- Divided the leads train data set into X and Y sets
- Built logistic regression model and checked the summary
- Selected 15 features using RFE
- Re-built the logistic regression model with 15 features and checked the summary
- Dropped the columns with high P and VIF values
- Re-built the logistic regression model and checked the summary

- Repeated above two steps till P and VIF values of the features are in the acceptable range

9.Making predictions on train set

- Made predictions on the train data set using the logistic regression model and with 0.5 as the cut off

10.Metrics

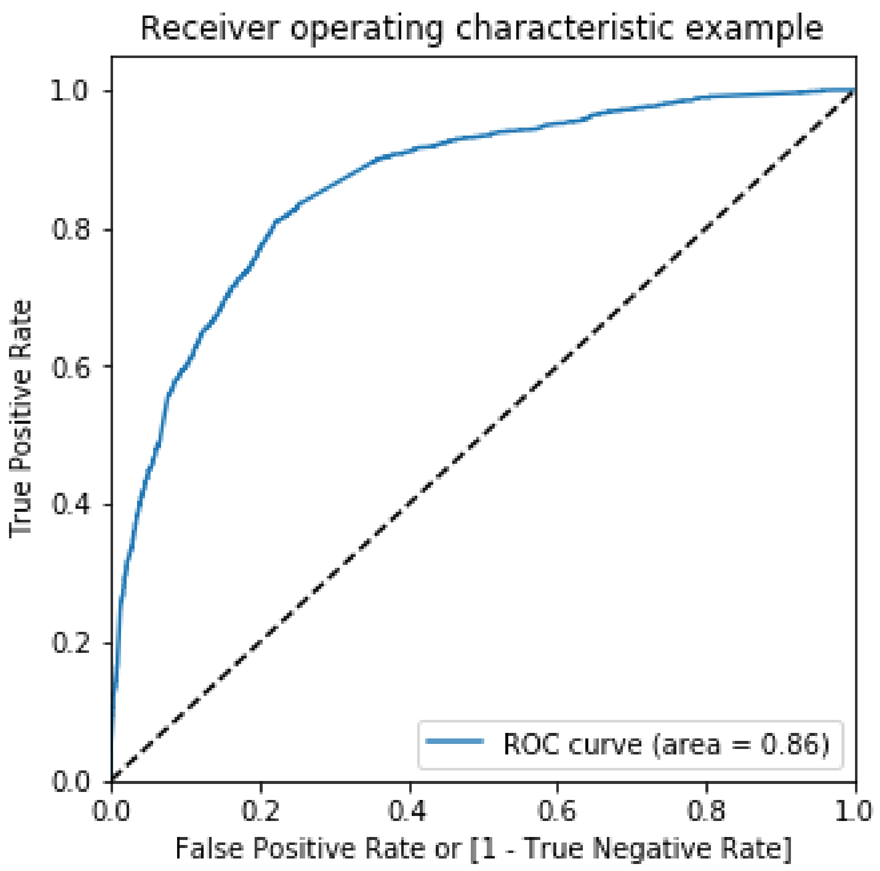
- Created confusion matrix

Predicted	Not Converted	Converted
Actual		
Not Converted	3328	481
Converted	794	1501

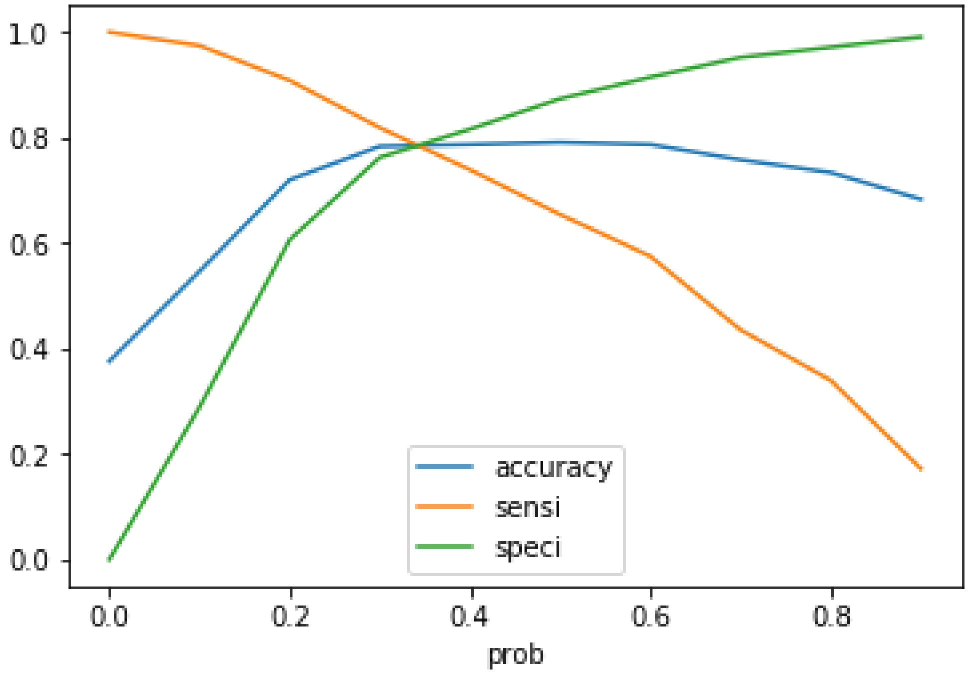
- Checked the following metrics

Accuracy	0.791
Sensitivity	0.654
Specificity	0.873
False postive rate	0.126
Positive predictive value	0.757
Negative predictive value	0.807

- Plotted ROC curve



- ROC curve was good.
- Found the optimal cut off point as 0.33,using Accuracy,Specificity and Sensitivity curve



- Made predictions using new cut off 0.33

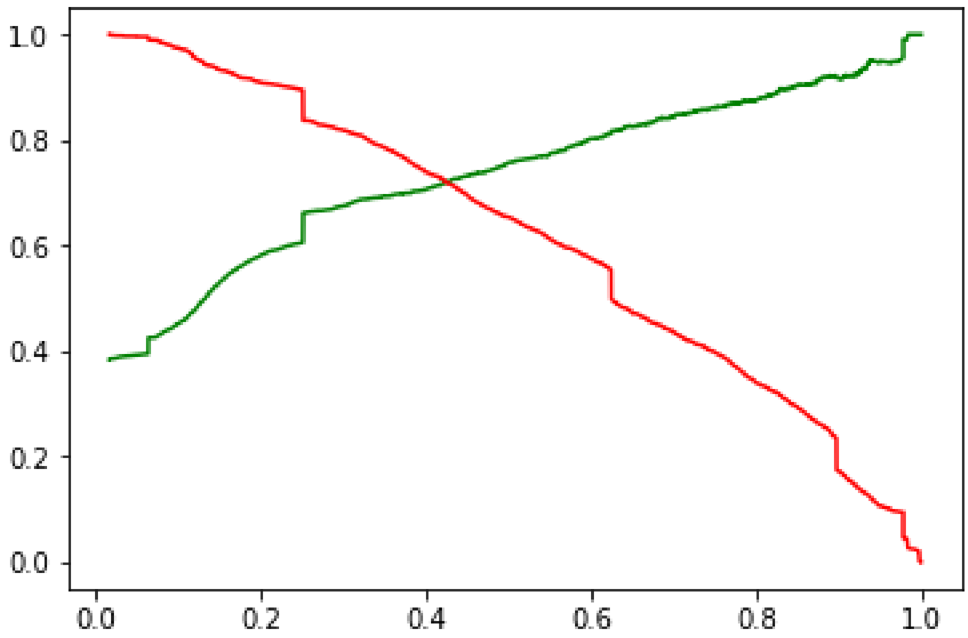
- Created confusion matrix

Predicted	Not Converted	Converted
Actual		
Not Converted	2980	829
Converted	456	1839

- Checked the following metrics

Accuracy	0.789
Sensitivity	0.801
Specificity	0.782
False positive rate	0.217
Positive predictive value	0.689
Negative predictive value	0.867
Precision	0.757
Recall	0.654

- Plotted precision and recall curve to find optimum cut off



- Found optimum cut off to be 0.42 from the above curve
- Made predictions using new cut off 0.42
- Created confusion matrix

Predicted	Not Converted	Converted
Actual		
Not Converted	3154	655
Converted	627	1668

- Checked the following metrics

Accuracy	0.789
Sensitivity	0.726
Specificity	0.828
False positive rate	0.171
Positive predictive value	0.718

11.Making predictions on test set

- Made predictions on the test data set using the logistic regression model and with 0.42 as the cut off from precision and recall curve for better results

- Created confusion matrix

Predicted	Not Converted	Converted
Actual		
Not Converted	1363	279
Converted	237	738

- Checked the following metrics

Accuracy	0.802
Sensitivity	0.756
Specificity	0.830
False positive rate	0.169
Positive predictive value	0.725
Negative Predictive value	0.851

12. Making predictions and calculating the lead score using original data.

- Made predictions on the original data set using the logistic regression model and with 0.42 as the cut off from precision and recall curve for better results
- Assigned scores from 0 to 100 to the leads based on the conversion probability, such that 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.

13. Finding the score to achieve 80% conversion rate

- Found from the above analysis that to achieve 80% conversion rate the sales team of X Education Company has to focus on the leads with score greater than equal to 73