Lead Scoring Case Study

- Manmitha Malineni
- Sowmya Cherukuwada

Problem Statement

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

- Objective here is to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Goal:

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
2. A higher score would mean that the lead is hot, i.e. is most likely to convert
3. A lower score would mean that the lead is cold and will mostly not get converted.

Steps Used to approach the problem

1. Importing Libraries
2. Loading the dataset
3. Handling Missing Values
4. Handling Class Imbalance.
5. Correlations
6. Outlier treatment
7. Dummy Variable Creation
8. Standardization of the data
9. Building a Model
    a. Feature Selection using RFE
    b. Applied Stats Models
10. Finding the optimal Cutoff Value -
    a. Using ROC Curve
    b. Precision-Recall Curve
11. Evaluating model performance on Test Data.
12. Conclusions

Handling Missing Values

1. Explicit Missing Values: There are columns with more than 30% of missing values. Identified all such columns and dropped them.
    a. For all the columns with less than 30% of missing data, dropped all such rows with missing / NA values.
2. Implicit Missing Values: There are a few columns where the option was not selected by the user. The column value is “Select”.
    a. Dropped all such rows which has the value as “Select”.

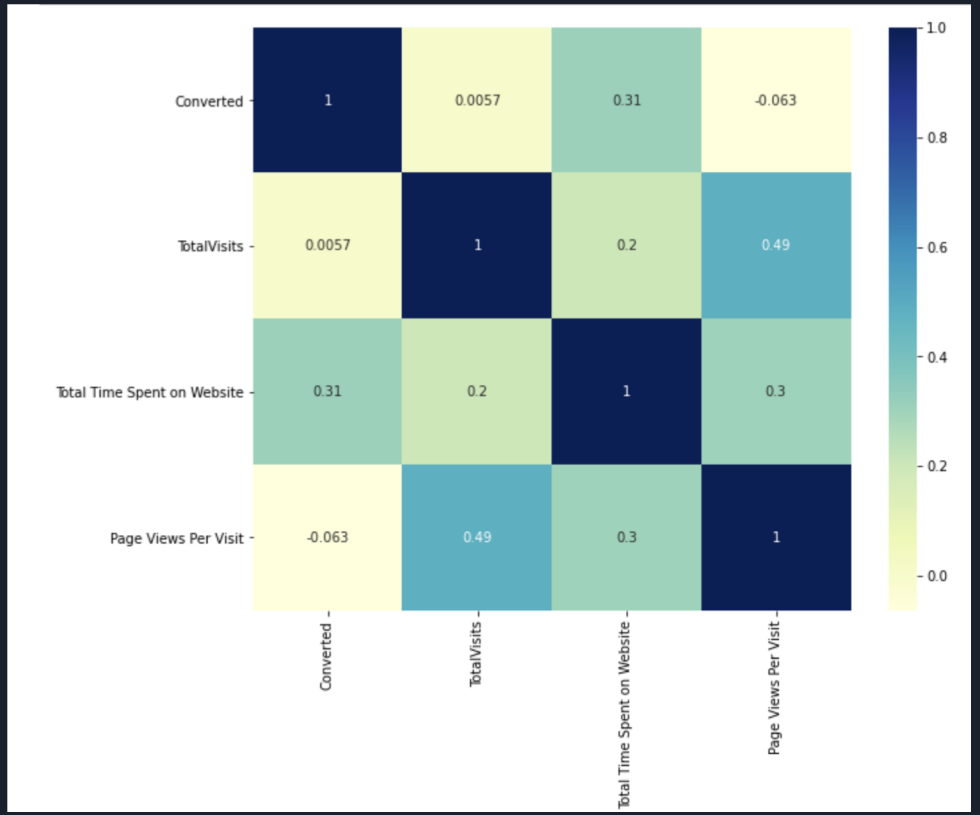
Class Imbalance:

There are a couple of categorical columns with a huge imbalance in the data. Dropped these columns as these could impact the model.

a. For eg: - The variable 'What matters most to you in choosing a course' has the value 'Better Career Prospects' 99.9% of the times.
b. Since this could impact the model performance, dropped this.

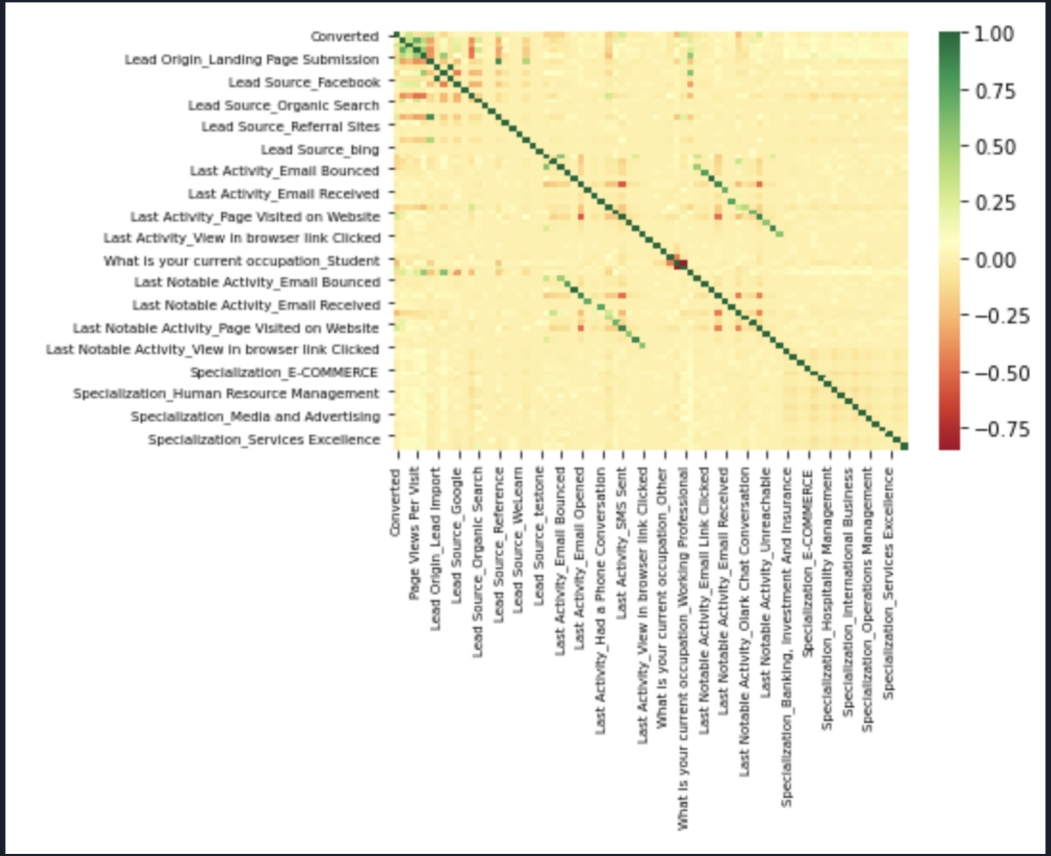
Correlation Between Numeric Variables

1. Total Time Spent on Website is having a positive correlation with “Converted” variable.
2. The columns “TotalVisits” and “page Views Per Visit” are highly correlated.



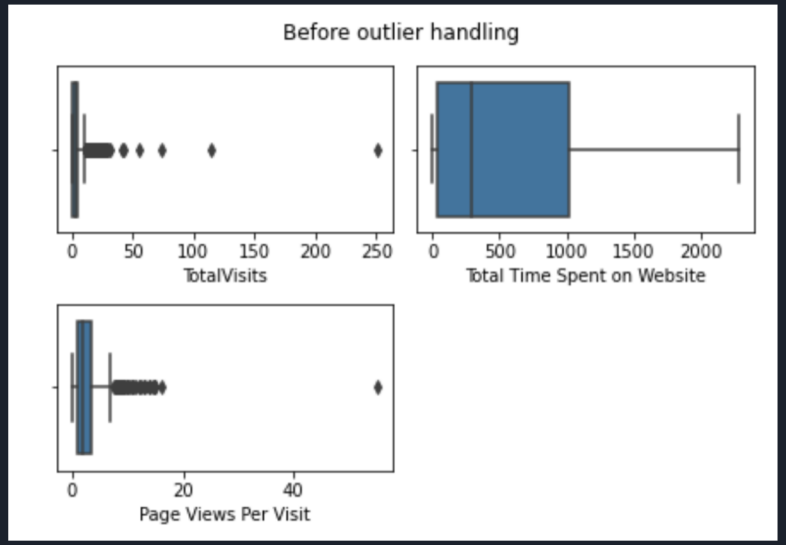
From this, it could be seen that there are high correlations in the dataset.

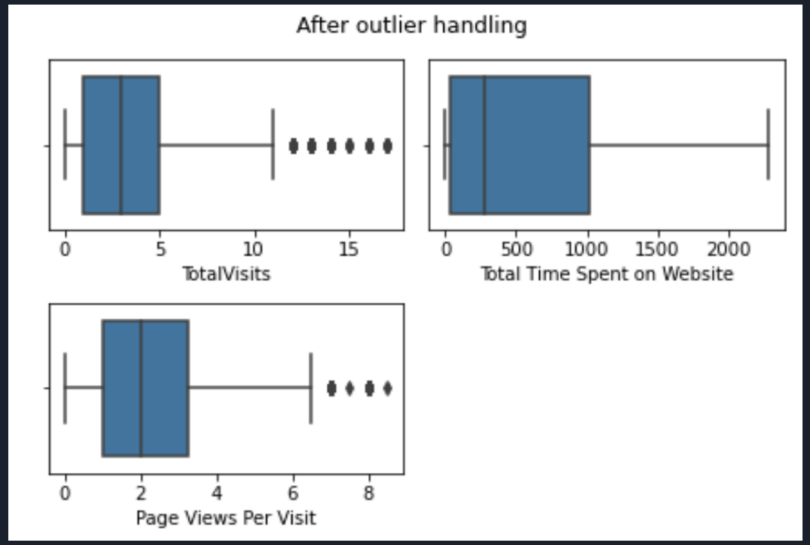
This would be taken care of - during model building, as we check for VIF's.



Outlier Treatment

1. The columns “TotalVisits”, “Page Views per visit” have outlier values.
2. Outliers in a logistic model are very sensitive, and could impact the model performance.
3. Hence, Removing 0.1% of the values from the upper quartile region





Dummy Variable Creation / One-Hot Encoding

1. Binary Variables: Converted all the binary columns to 0/ 1 encoding.
2. Categorical Variables: For all the categorical variables, created dummy variables, and dropped the first column.
    a. For eg: If there are n categories in a variable, this procedure gives n-1 variables.

Standardization of Data

1. The numeric variables present in the dataset, are not all in the same range.
2. So this could take more time for the logistic model to converge.
3. Used MinMaxScaler to get all the values in the same range.

Model Building

1. **Feature Selection using RFE:**
    a. After creating the dummy variables (from the previous step), the dataset contains around 74 columns.
    b. Since it would be time-consuming to check and drop each variables manually - we are using RFE to select the 15 significant features from the dataset.

1. **Applying Logistic Regression** (StatsModels Library):
    a. On the 15 features selected by the RFE, we then started applying the following steps -
        i. Iteratively ran the logistic regression model, verified and dropped the columns until all the p-values and VIFs are in range

Final Model Visualization with VIF

Dep. Variable:	Converted	**No. Observations:**	4358
Model:	GLM	**Df Residuals:**	4346
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1986.7
Date:	Tue, 07 Sep 2021	**Deviance:**	3973.4
Time:	10:12:04	**Pearson chi2:**	4.56e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5063	0.111	-22.669	0.000	-2.723	-2.290
TotalVisits	1.4897	0.280	5.322	0.000	0.941	2.038
Total Time Spent on Website	4.5709	0.191	23.909	0.000	4.196	4.946
Lead Origin_Lead Add Form	4.2107	0.251	16.768	0.000	3.719	4.703
Lead Source_Olark Chat	1.9077	0.138	13.833	0.000	1.637	2.178
Lead Source_Welingak Website	1.6097	0.759	2.121	0.034	0.123	3.097
Do Not Email_Yes	-1.5384	0.199	-7.749	0.000	-1.927	-1.149
Last Activity_Had a Phone Conversation	3.7392	1.187	3.149	0.002	1.412	6.067
Last Activity_Olark Chat Conversation	-1.0973	0.185	-5.927	0.000	-1.460	-0.734
Last Activity_SMS Sent	1.2052	0.085	14.116	0.000	1.038	1.372
What is your current occupation_Working Professional	2.4614	0.189	12.989	0.000	2.090	2.833
Last Notable Activity_Unreachable	2.7808	0.805	3.455	0.001	1.203	4.359

	Features	VIF
1	Total Time Spent on Website	2.05
0	TotalVisits	1.98
8	Last Activity_SMS Sent	1.53
2	Lead Origin_Lead Add Form	1.45
4	Lead Source_Welingak Website	1.28
3	Lead Source_Olark Chat	1.23
9	What is your current occupation_Working Profes...	1.22
7	Last Activity_Olark Chat Conversation	1.20
5	Do Not Email_Yes	1.04
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

Evaluating model on the training data using 0.5 as the cut-off.

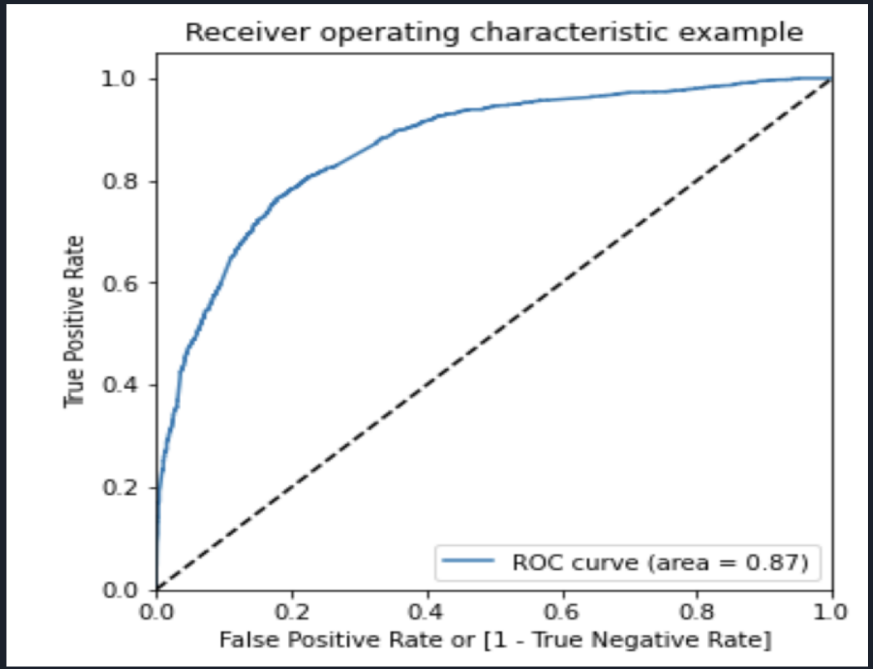
- Accuracy - 79.18%
- Confusion Matrix -

	Not Converted	Converted
Not Converted	1863	383
Converted	524	1588

- Sensitivity - 75.18
- Specificity - 82.94
- Precision - 80.56
- Recall - 75.18

Evaluating Model Using ROC Curve

1. After building the model, we evaluated the model stability using ROC Curve
2. The area under ROC Curve is 0.87
3. Since the line is closer to upper left side of the margin, we could conclude this is a good model.



Finding the Optimal Cut-Off Value:

Using 0.5 as the cut-off may not always provide us with the optimal solution. So we are using the following metrics to get the optimal cut-off value

1. Sensitivity-Specificity Trade-Off
2. Precision-recall Curve

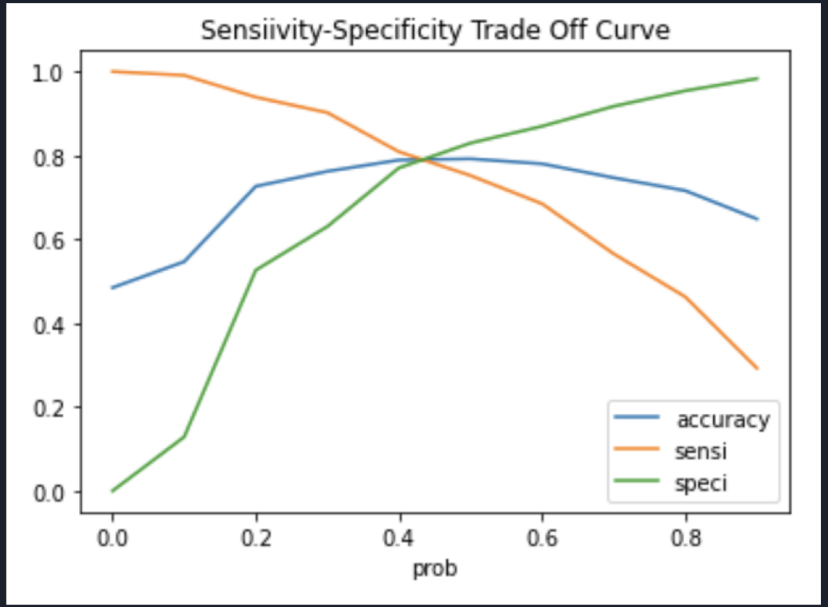
Sensitivity-Specificity Trade-Off:

As part of this, we have evaluated the sensitivity, specificity, and accuracy for the following cut-off values - 0.1, 0.2, 0.3, … 0.9

Using these values, we plotted the following graph.

From the graph, we could observe that for the value 0.40 - the accuracy, sensitivity and specificity are giving better results.

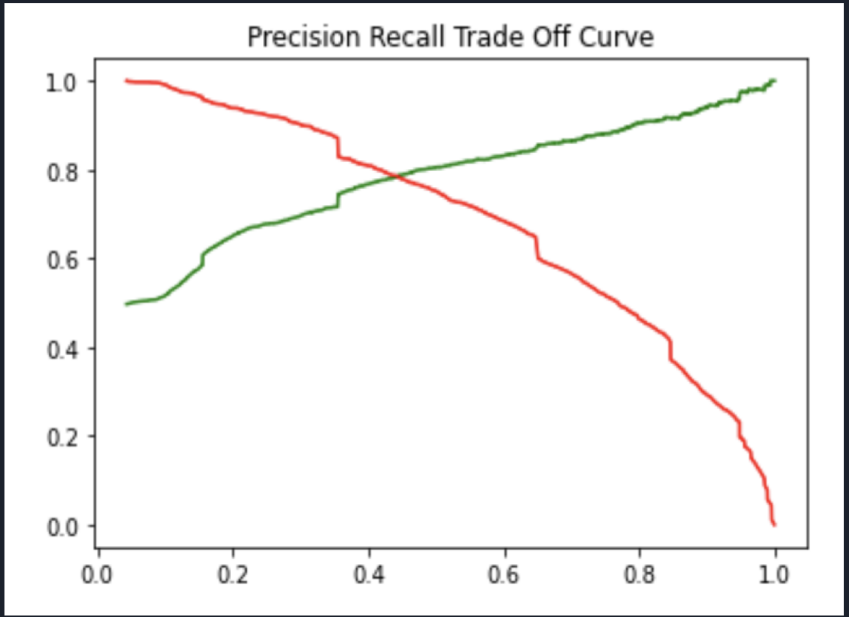
Optimal Cut-Off Value here - 0.4



Precision-Recall Curve

The following graph shows the trade-off between precision, recall.

We could see that - at the point 0.42, there is a good balance between precision and recall.



Evaluating Train Data performance using 0.40 as cut-off

a. Accuracy - 79.0%
b. Confusion Matrix -

	Not Converted	Converted
Not Converted	1760	486
Converted	426	1686

a. Sensitivity - 80.87
b. Specificity - 77.02
c. Precision - 76.79
d. Recall - 80.87

From the above metrics, it could be seen that sensitivity of the model has improved on using 0.40 as cut-off, instead of 0.5

Evaluating Model Performance on the test data:

1. The test data set contains additional columns than the one considered for model Building.
2. As part of this step, we have removed these columns.
3. Standardized the test data using the MinMaxScaler that was fitted with the train data set.
4. Using these values, we started evaluating the results to check the performance of our model.
5. The evaluation metrics can be seen below -
    a. Accuracy - 78.53
    b. Sensitivity - 79.90
    c. Specificity - 77.30
    d. Precision - 75.86
    e. Recall - 79.90

From these, we could conclude that the model is performing better and is able to predict on the unseen data set.

Conclusions:

- The Accuracy, Precision and Recall score we got from test set in acceptable range
- We have high recall score than precision score which we were exactly looking for.
- In business terms, this model can predict according to the business requirements in coming future.
- This concludes that the model is working well.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

- Lead Origin_Lead Add Form
- Total Time Spent on Website
- Last Notable Activity_Had a Phone Conversation
- What is your current occupation_Working Professional
- Last Notable Activity_Unreachable

Thank You