Lead Conversion Case Study

NareshKumar Sappati
Yaswanth Sai Rama Krishna Gampa

Agenda:

- Problem statement
- Goals of case study
- Data understanding and cleaning
- Data Outliers handling
- Exploratory Data Analysis
- Correlation and its implecations
- Logistic Model building.
- Model validations.

Problem statement

X Educations is one of the leading online courses provider for industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

We have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes. Some of those may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not. You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page.

Goals of case study

- 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. 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.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Data understanding and cleaning

- The given dataset named “Leads.csv” is of a size of 9240*37.
- The given dataset has no duplicate values.
- Data balance of target field looks good.
- There are few data imbalances in independent variables:

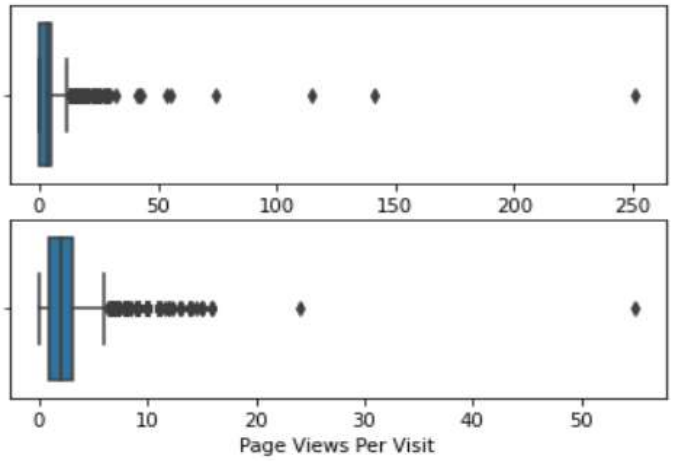
Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement etc.. are biased (more than 99.9%) towards only one category there won't be much use of these variables in the data set; these columns can be deleted.

Missing values

- Lead Quality, Tags : Columns with more than 30% missing values can be droped from data frame.
- Lead Profile, What matters most to you in choosing a course, What is your current occupation, Country, How did you hear about X Education, Last Activity: We can fill these nulls with mode of the columns.
- Asymmetrique Activity Index, Asymmetrique Profile Score, Asymmetrique Activity Score, Asymmetrique Profile Index : lets keep the scores and delete indexes; replace score nulls as 0
- Specialization, City : fill null with 'Select'
- Page Views Per Visit: fill null with mean value
- TotalVisits: full null with 0

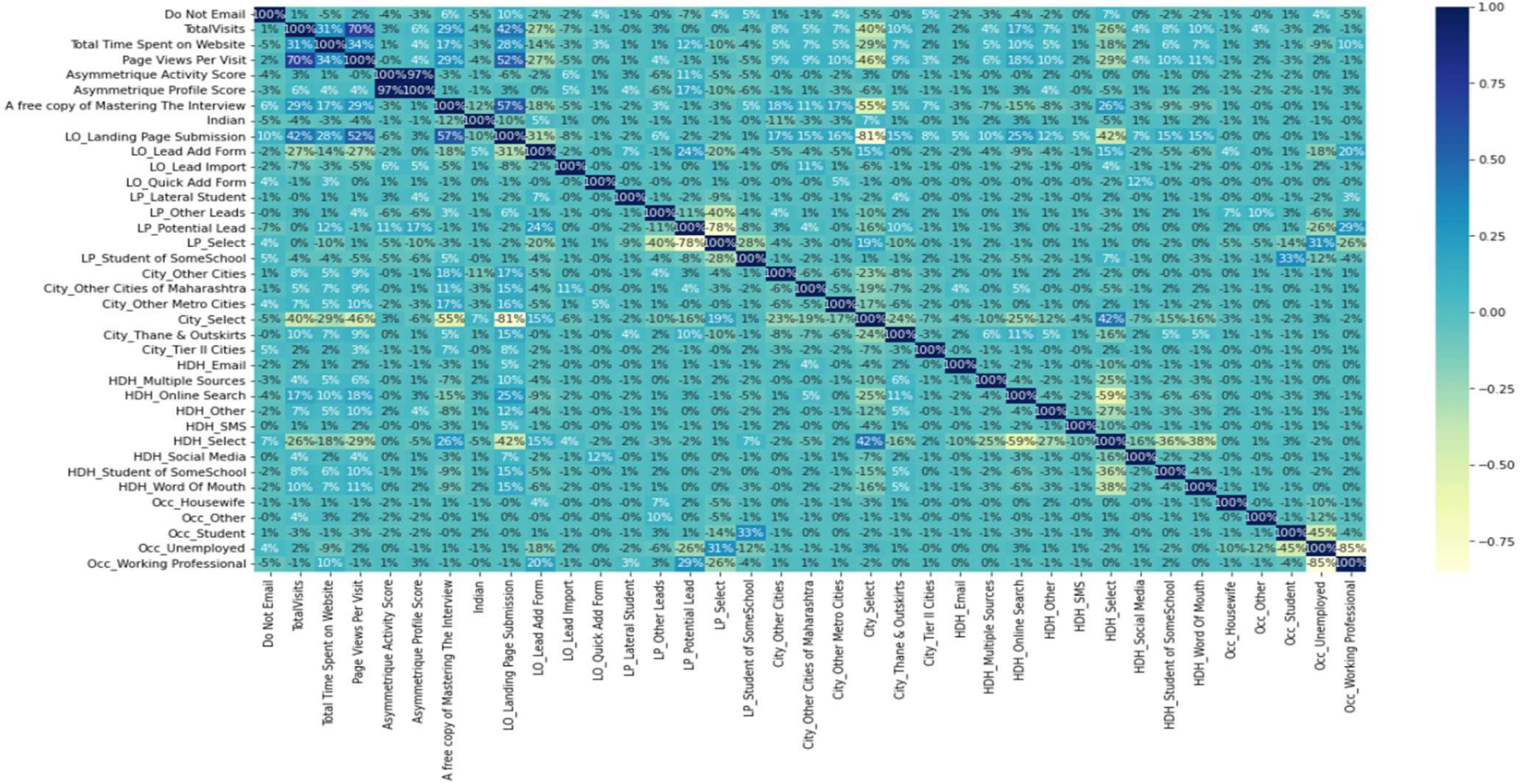
Data Outliers

```
plt.subplot(2,1,1)
sns.boxplot(df_leads['TotalVisits'])
plt.subplot(2,1,2)
sns.boxplot(df_leads['Page Views Per Visit'])
plt.show()
```



There are some Outliers for TotaVisits and Page Views per Visit. We have to handle it by capping the value of those variables.

Correlation



Correlation

Above Correlation matix show the one on one relation between X variables.

- Asymmetrique Activity Score is highlu correlates to Asymmetrique Profile Score: Profile Score is more business understandable than Activity Score, so we can drop Asymmetrique Activity Score.
- LP_Select, City_Select, HDH_Select, LO_Landing Page Submission are highly correlated (-ve or +ve) to their categorical siblings.
- TotalVisits is highly correlated to some of the other variables.

Based on above observations we can drop some of those variables from dataset.

Model building

- Create logistic model with all available X variables. and found some high P values.
- We did choose RFE for feature elimination process.
- We got top 15 features to build the model.
- Checked VIF for Multicollinearity and surprisingly there is no issue with VIF values all are in good range.
- Based on p-value we did eliminate some features from model and finalized logml5

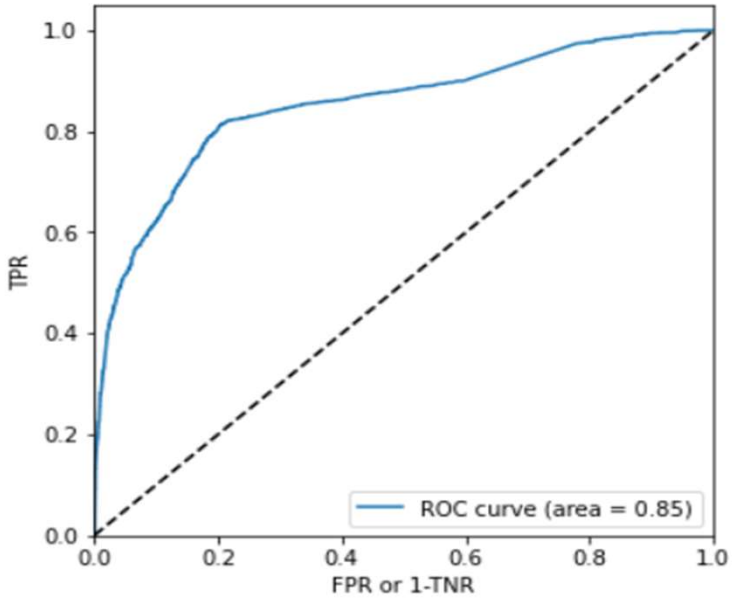
We have to check ROC curve of the model and decide on cutoff value in the next step.
Following is the summary of final model.

Model building

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6452
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2853.2
Date:	Mon, 17 Oct 2022	**Deviance:**	5706.4
Time:	17:20:18	**Pearson chi2:**	1.03e+04
No. Iterations:	21	**Pseudo R-squ. (CS):**	0.3605
Covariance Type:	nonrobust		

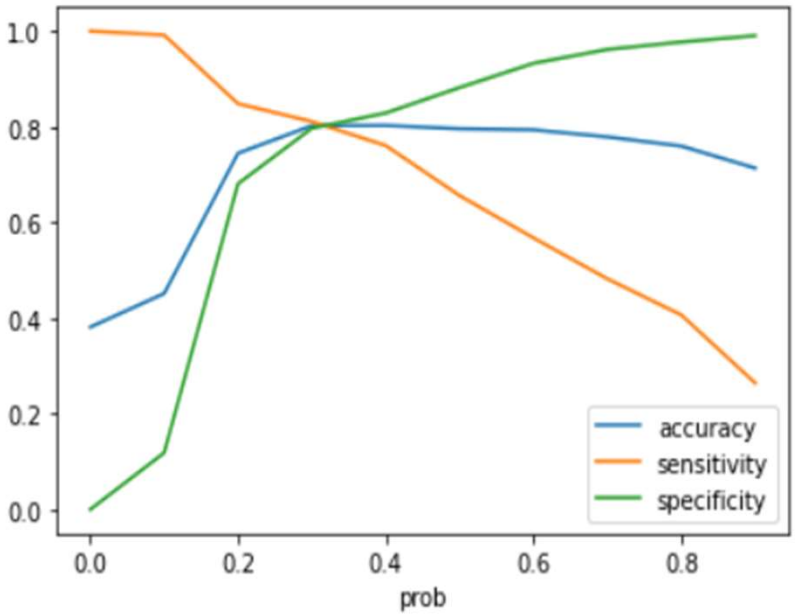
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3505	0.207	-6.533	0.000	-1.756	-0.945
Do Not Email	-1.2615	0.157	-8.025	0.000	-1.570	-0.953
Total Time Spent on Website	0.9652	0.034	28.275	0.000	0.898	1.032
A free copy of Mastering The Interview	-0.3239	0.074	-4.386	0.000	-0.469	-0.179
Indian	0.3412	0.205	1.668	0.095	-0.060	0.742
LO_Lead Add Form	3.4195	0.180	19.013	0.000	3.067	3.772
LO_Lead Import	-0.5804	0.538	-1.079	0.281	-1.635	0.474
LP_Lateral Student	2.4227	1.088	2.227	0.026	0.291	4.555
LP_Potential Lead	1.6218	0.092	17.550	0.000	1.441	1.803
LP_Student of SomeSchool	-2.2359	0.436	-5.131	0.000	-3.090	-1.382
HDH_Email	0.5473	0.569	0.962	0.336	-0.568	1.663
HDH_Multiple Sources	-0.3822	0.250	-1.528	0.126	-0.872	0.108
HDH_SMS	-0.9192	0.740	-1.243	0.214	-2.369	0.531
Occ_Housewife	22.5771	1.31e+04	0.002	0.999	-2.57e+04	2.57e+04
Occ_Student	0.6209	0.255	2.433	0.015	0.121	1.121
Occ_Working Professional	2.5777	0.186	13.844	0.000	2.213	2.943

Receiver Operating Characterstics (ROC) Curve



Above ROC curve looks good and plotting the values as expected.

Finding optimal threshold



In Order to select cutoff value we have to consider relatively good metics values.

As you can see, at about a threshold of 0.3, the curves of accuracy, sensitivity and specificity intersect, and they all take a value of around 80%.

Modle Validation

Confusion Matrix, Accuracy are first choice of validation metrics.
Precission and Recall should also be checked for cutoff value provided.

Following are the validation values for train set.

```
++++++++++++++ Confusion Matrix ++++++++++++++++++++++++++++++++++++++++++++
[[3190  812]
 [ 466 2000]]

++++++++++++++ Accuracy Score, Precission & Recall ++++++++++++++++++++++++
Accuracy Score:   0.8024
Precission:       0.7112
Recall:           0.811
```

Modle Validation

Following are the validation values for test set.

```
++++++++++++++ Confusion Matrix ++++++++++++++++++++++++++++++++++++++++++++
[[1344  333]
 [ 218  877]]
++++++++++++++++++++++++++++++++
Accuracy            : 0.8012
Specificity         : 0.8014
Precision           : 0.7248
Recall              : 0.8009
```

Metric values are looking good and are in line with train set.