Lead scoring case study

Bandi Samuel
Pallawi Jyoti

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

The company requires to build a model and need 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 CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goal:

Build a logistic regression model to assign a lead score 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.

Approach taken for Analysis

Inspecting and cleaning the dataset if required.

Dropping NA values and having a complete data, without Null values

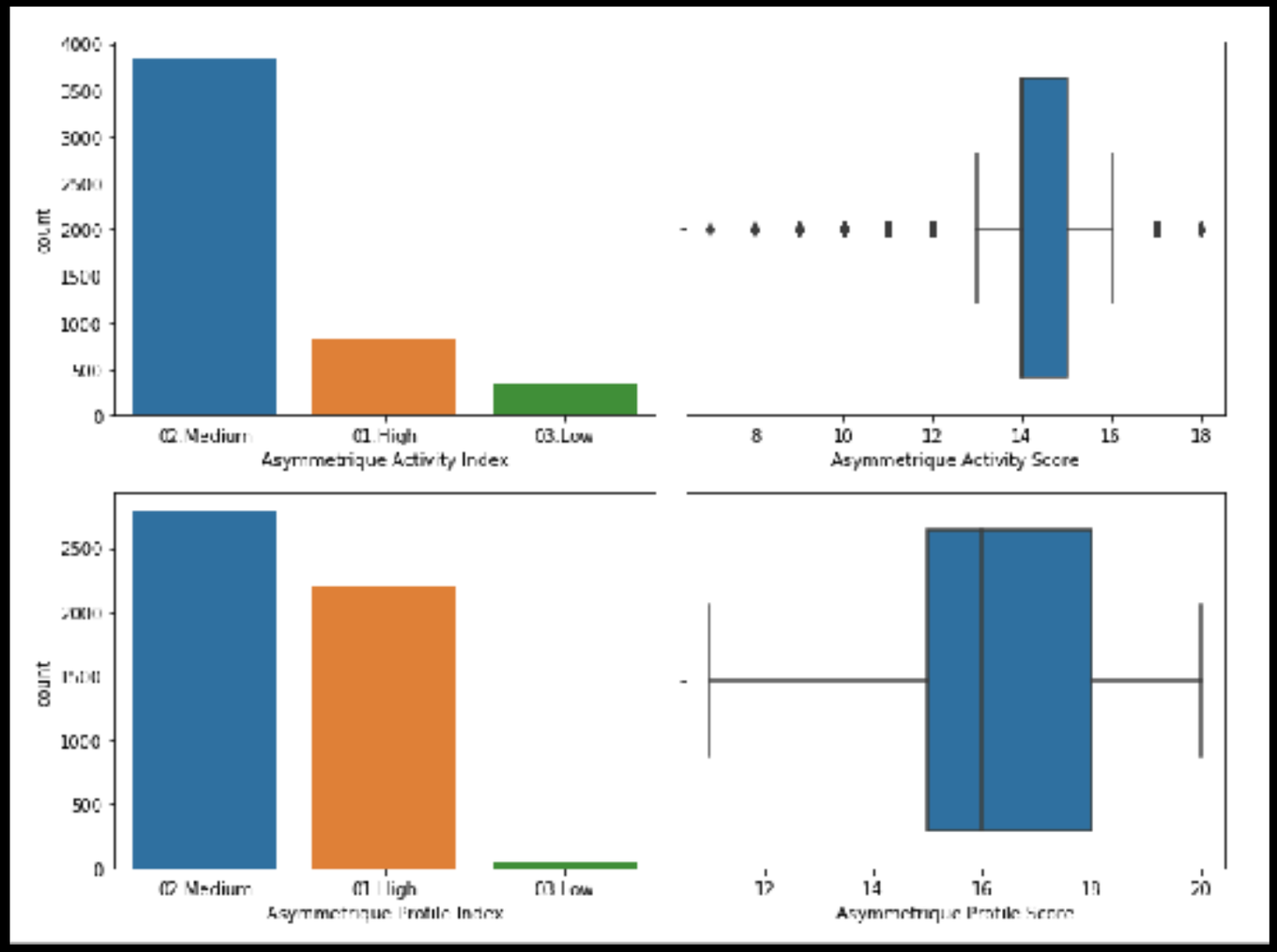
An index and score assigned to each customer based on their activity and their profile

Exploratory Data Analytics:
Univariate Analysis

Feature Selection Using RFE

Metrics beyond simply accuracy

An index and score assigned to each customer based on their activity and their profile



There is too much variation in these parameters so it’s not reliable to impute any value in it.

Exploratory Data Analytics:

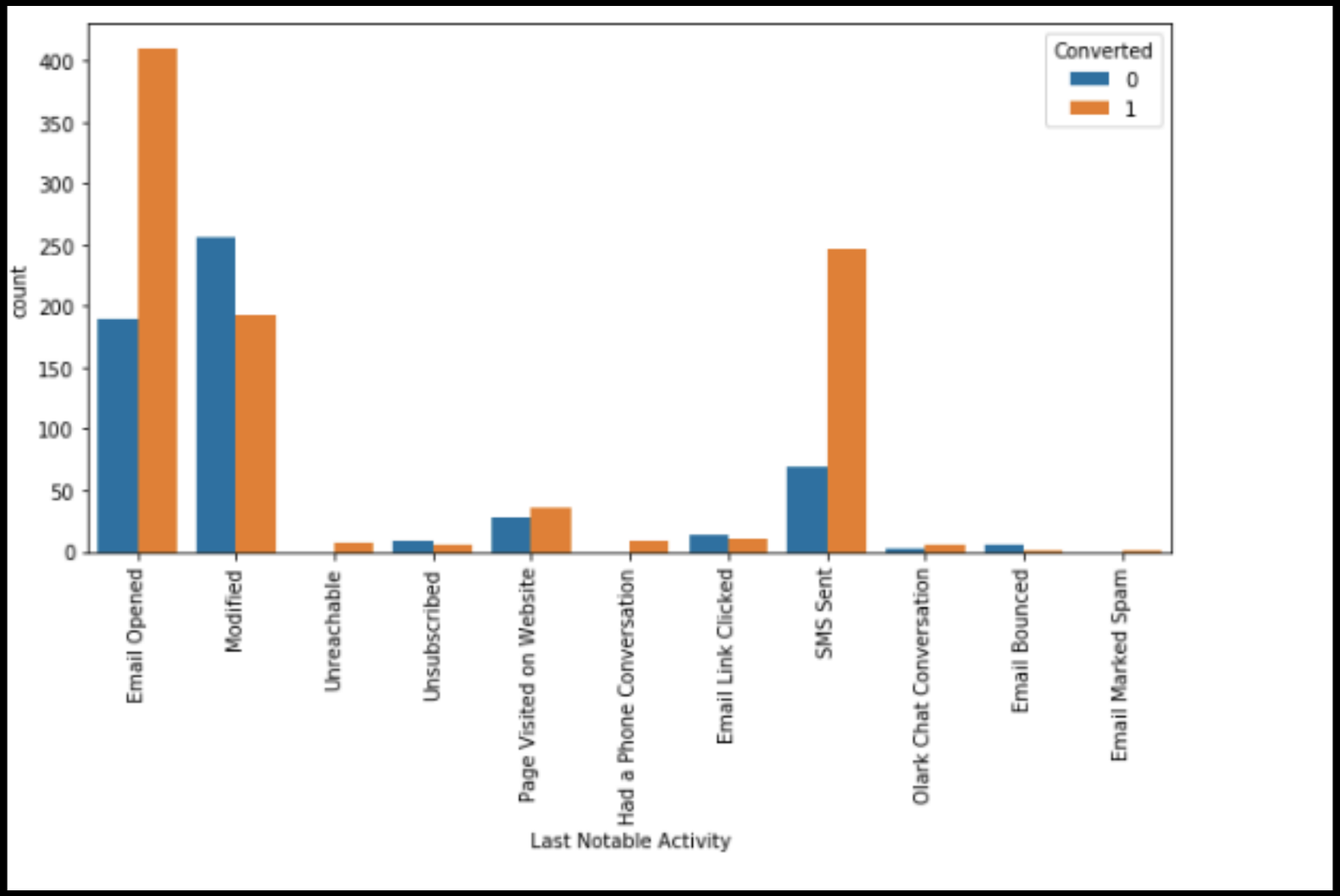
Univariate Analysis

Converted is the target variable, Indicates whether a lead has been successfully converted (1) or not (0).

```
Converted = (sum(data['Converted'])/len(data['Converted'].index))*100
Converted
```

```
61.82183523107837
```

To improve overall lead conversion rate, we need to focus more on improving lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



Based on the univariate analysis we have seen that many columns are not adding any information to the model

Creating a dummy variable for some of the categorical variables and dropping the first one

	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	Lead Source_Others	Lead Source_Reference	Lead Source_Referral Sites	Last Activity_Email Bounced
k to scroll output; double click to hide			0	0	0	0	0	0	0
6	1	0	1	0	0	0	0	0	0
22	1	0	1	0	0	0	0	0	0
24	0	0	1	0	0	0	0	0	0
26	1	0	0	0	1	0	0	0	0

Model Building:

Feature Selection Using RFE

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	1045
Model Family:	Binomial	**Df Residuals:**	1031
Link Function:	logit	**Df Model:**	13
Method:	IRLS	**Scale:**	1.0000
Date:	Sun, 09 Jun 2019	**Log-Likelihood:**	-190.94
Time:	22:59:37	**Deviance:**	381.88
No. Iterations:	23	**Pearson chi2:**	1.45e+03
		Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5344	0.190	2.810	0.005	0.162	0.907
Do Not Email	-1.7324	0.674	-2.571	0.010	-3.053	-0.412
Last Activity_SMS Sent	0.4356	0.383	1.137	0.255	-0.315	1.186
Specialization_IT Projects Management	1.6034	0.731	2.192	0.028	0.170	3.037
Tags_Busy	3.1699	0.663	4.782	0.000	1.871	4.469
Tags_Interested in full time MBA	-23.0968	2.22e+04	-0.001	0.999	-4.35e+04	4.35e+04
Tags_Lost to EINS	5.8322	1.438	4.056	0.000	3.014	8.650
Tags_Not doing further education	-23.3094	1.64e+04	-0.001	0.999	-3.22e+04	3.22e+04
Tags_Ringing	-1.5655	0.588	-2.662	0.008	-2.718	-0.413
Tags_Will revert after reading the email	3.3126	0.340	9.733	0.000	2.645	3.980
Tags_switched off	-1.1790	0.904	-1.304	0.192	-2.951	0.593
Lead Quality_Not Sure	-2.4295	0.372	-6.526	0.000	-3.159	-1.700
Lead Quality_Worst	-5.8313	1.065	-5.477	0.000	-7.918	-3.745
Last Notable Activity_Olark Chat Conversation	-1.7444	1.161	-1.503	0.133	-4.020	0.531

Metrics beyond simply accuracy:

```
# Let's check the overall accuracy.
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
0.9375
```

```
TP = confusion2[1,1] # true positive
TN = confusion2[0,0] # true negatives
FP = confusion2[0,1] # false positives
FN = confusion2[1,0] # false negatives
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.8589420654911839
```

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.9814814814814815
```