LEAD-SCORING CASE STUDY

By -
Swarup Das (dasswarup53@gmail.com)
Shubham Bammi (shubhambammi27@gmail.com)

Problem Statement

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.

Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

The company wants to identify the hot leads / potential customers so that the sales team can nurture their interest and help them convert and contribute to company's revenue .

Objective

The objective of this analysis is to select the most promising leads, i.e. the leads that are most likely to convert into paying customers, also known as ‘Hot Leads’ for X Education. If we successfully identify this set of leads, the lead conversion rate would go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

- To build a logistic regression 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.
- Target is to get lead conversion rate to be around 80%.

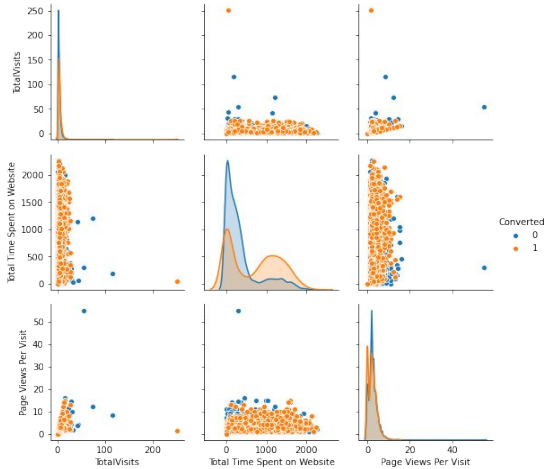
Steps Involved in Analysis

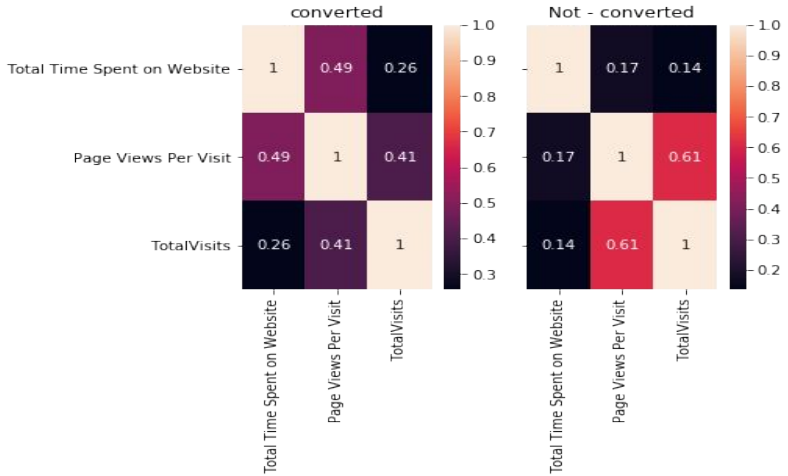
1. Load the data for analysis
2. Clean and prepare the data
3. Exploratory Data Analysis
4. Data Processing (Feature Scaling , One Hot Encoding)
5. Splitting the data into Test and Train dataset
6. Building a logistic Regression model and calculate Lead Score
7. Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall
8. Applying the best model in Test data based on the Sensitivity and Specificity Metrics

Load & Clean the Data for Analysis

- Import the data from Leads dataset & loaded in data frame.
- Convert data into clean format suitable for analysis .
- Remove duplicate data.
- Many of the categorical variables had a level called 'Select' . This has been handled by replacing it by null value or removing them.
- Dropped the columns which had missing values.
- Dropped the columns that don’t change as there was no variance.
- Removed & corrected the dirty data by handling null /NaN values.
- Identified the variables which needs to be scaled.

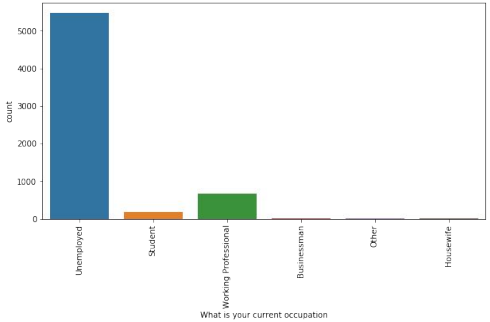
Exploratory Data Analysis

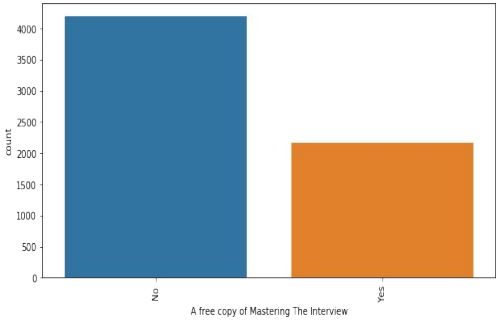


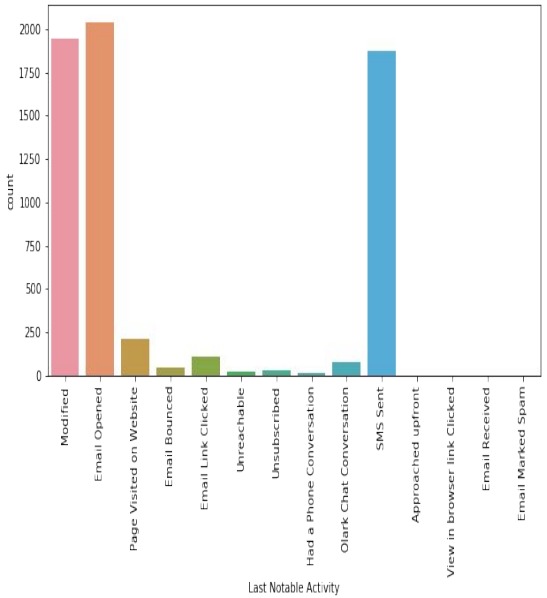


We can see some positive amount of correlation between Page view per and Total Time Spent On Website for converted folks . Whereas we can see strong correlation between total visits and Page Views Per Visit for not converted customers

Exploratory Data Analysis







- Most of the people voted "NO" for A free copy of Mastering the Interview
- Email Opened , Modified and SMS Sent have large counts for the Last notable activity .
- Most of the people in the dataset (5000+) are unemployed
- Lead Source are mostly coming from Google , Direct Traffic , Olark Chat , Organic Search etc (decreasing order of lead counts)

Feature Scaling , Splitting the Data & Dummy Encoding

- Feature Scaling of Numeric data using MinMaxScaler on the below variables:
  - Total Visits
  - Page Views Per Visit
  - Total Time Spent on Website
- Splitting data into train and test set where 70% is train set and 30% is test set.
- After splitting the data , we performed one hot encoding on the categorical data that was consisted in these columns :

  ['Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity', 'What is your current occupation','A free copy of Mastering The Interview', 'Last Notable Activity','Specialization']

Model Building

- We have not used correlation metrics (heat map) due to high numbers of variables / columns. Due to this we dropped multi – correlated variables once we build our first model.
- After downloading required packages (Stats Model) for Logistic Regression:
- Top 15 features, selected by RFE. (here is the 1 st model using 15 features)

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4445
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2072.8
Date:	Fri, 20 Nov 2020	**Deviance:**	4145.5
Time:	22:53:46	**Pearson chi2:**	4.84e+03
No. Iterations:	22		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0061	0.600	-1.677	0.094	-2.182	0.170
TotalVisits	11.3439	2.682	4.230	0.000	6.088	16.600
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	2.9483	1.191	2.475	0.013	0.614	5.283
Lead Source_Olark Chat	1.4584	0.122	11.962	0.000	1.219	1.697
Lead Source_Reference	1.2994	1.214	1.070	0.285	-1.080	3.679
Lead Source_Welingak Website	3.4159	1.558	2.192	0.028	0.362	6.470
Do Not Email_Yes	-1.5053	0.193	-7.781	0.000	-1.884	-1.126
Last Activity_Had a Phone Conversation	1.0397	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6492	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1544	0.630	-1.831	0.067	-2.390	0.081
What is your current occupation_Unemployed	-1.3395	0.594	-2.254	0.024	-2.505	-0.175
What is your current occupation_Working Professional	1.2743	0.623	2.045	0.041	0.053	2.496
Last Notable Activity_Had a Phone Conversation	23.1932	2.08e+04	0.001	0.999	-4.08e+04	4.08e+04
Last Notable Activity_Unreachable	2.7868	0.807	3.453	0.001	1.205	4.369

We have multiple variables it is important to eliminate some variables, to make this model more sustainable and actionable by business. Though p – values for most of the variable was close to 0.

Model Building - Feature Elimination Using RFE & VIF

- Build the model , evaluate the significance using p-values and VIF and eliminate the variables . This steps were repeated until we reached a stable state where we had low P-values and low VIF values .
- Final model had 11 variables and coefficients are as follows.

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4449
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.1
Date:	Fri, 20 Nov 2020	**Deviance:**	4158.1
Time:	22:54:02	**Pearson chi2:**	4.80e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

Model Building - Confusion Matrix & Accuracy %

- Cut off we used = 0.5
- To check accuracy of model, created confusion matrix and accuracy%.
- So we can see few misclassification as well such as 383 and 506.
- However, accuracy% for this model is 78.86%

	Converted	Conversion_Prob
0	0	0.300117
1	0	0.142002
2	1	0.127629
3	1	0.291558
4	1	0.954795

	Converted	Conversion_Prob	Predicted
0	0	0.300117	0
1	0	0.142002	0
2	1	0.127629	0
3	1	0.291558	0
4	1	0.954795	1

```
# lets create the confusion matrix for the same
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.Predicted )
confusion
```

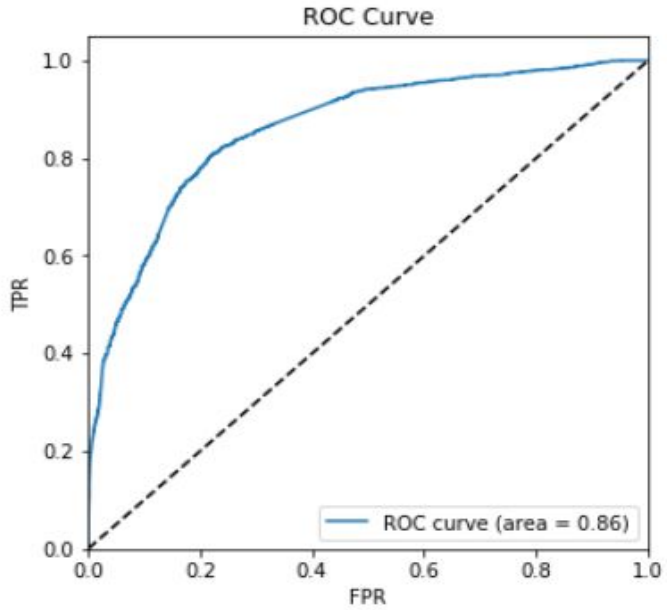
```
array([[1929,  383],
       [ 560, 1589]], dtype=int64)
```

```
#checking for accuracy
metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.Predicted)
```

```
0.7886124187401928
```

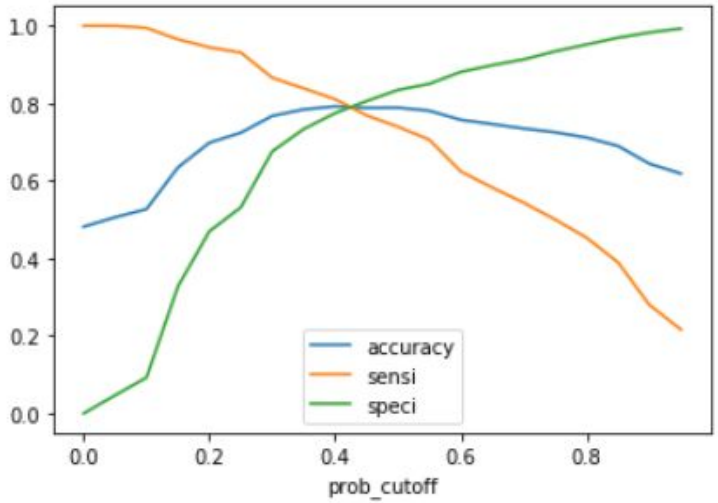
Model Building - Accuracy & ROC Curve

- ROC Curve: Trade off b/w TPR and FPR. As we know higher the area under the curve of an ROC, the better is model.
- With an arbitrary threshold value of 0.5, we have the following metrics i.e. Accuracy is 78.8%.
- ROC - AUC : 0.86



Model Building - Finding the Optimal Threshold

- We calculated values of Accuracy, Sensitivity & Specificity at different cut off values and stored them in a data frame.
- As we can see from chart, 0.42 is optimum probability cut off for our dataset.
- Accuracy with 0.42 cut off is = 79.08



Conclusion

- While we have checked both Sensitivity-Specificity Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 78.4%, 77.9% and 78.9% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 79% in train set and 78% in test set.
- The top 3 variables that contribute for lead getting converted in the model are :
    - Total Time Spent on Website
    - Lead Add Form (from Lead Origin)
    - Had a Phone Conversation (from Last Notable Activity)
- Hence overall this model seems to be good.

Thank You