X Education

(Data Analysis to increase the lead conversion using marketing team)

Project by:
Yogita Goswami
Chandana Joshi

X Education Case Study Analysis

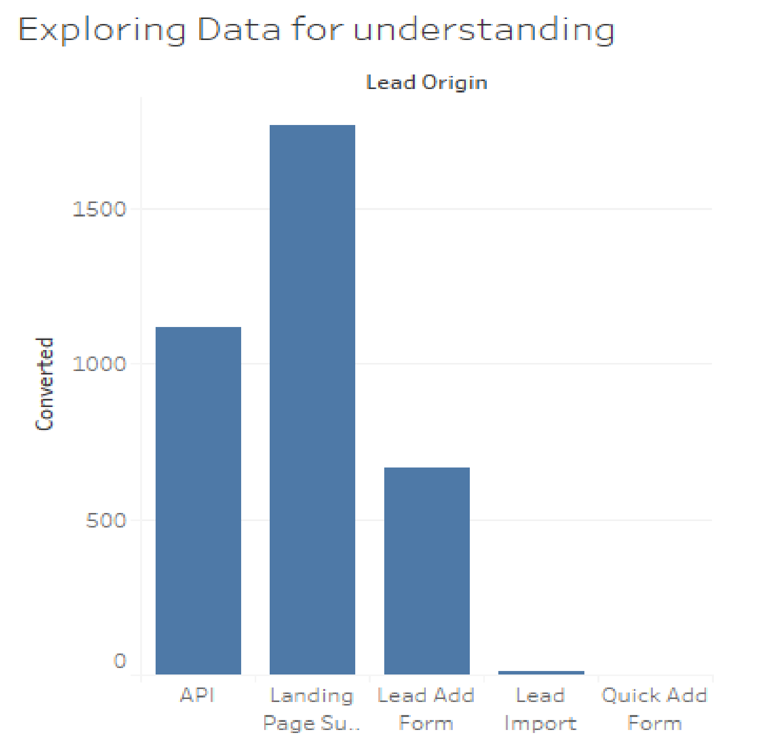
Problem Statement:

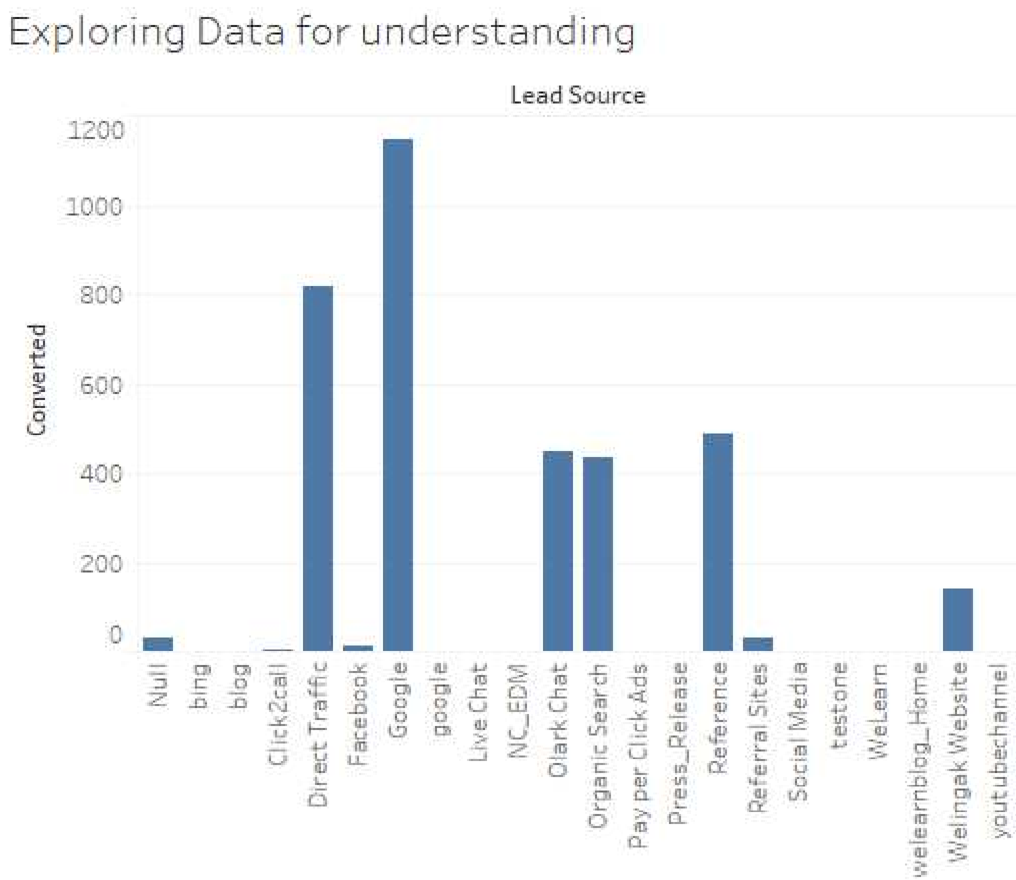
The company markets its courses on several websites and captures the browsing history and application filled by the interested candidates. Observation is only around 30% of the interested candidates actually proceed to take the course. Analyze the data and provide insights to increase the conversion to 80% using targeted marketing to interested candidates

Available Data Source for analysis: Leads dataset having 9000 potential leads information like source and origin from which they learnt about the course, if they have subscribed to free magazine, newsletter, updates on different courses, time they spent on website, # of visits etc.

Data Analysis Approach

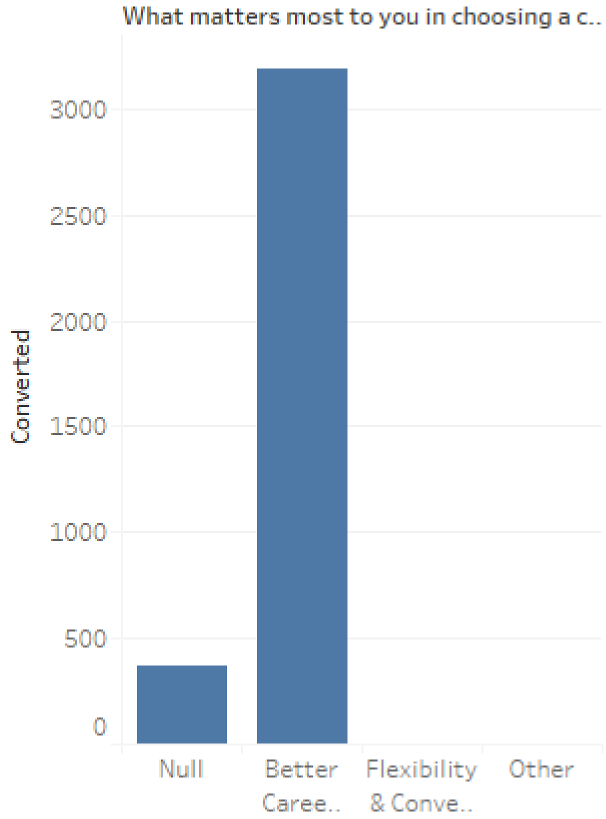
Step 1 :Understanding the Data:

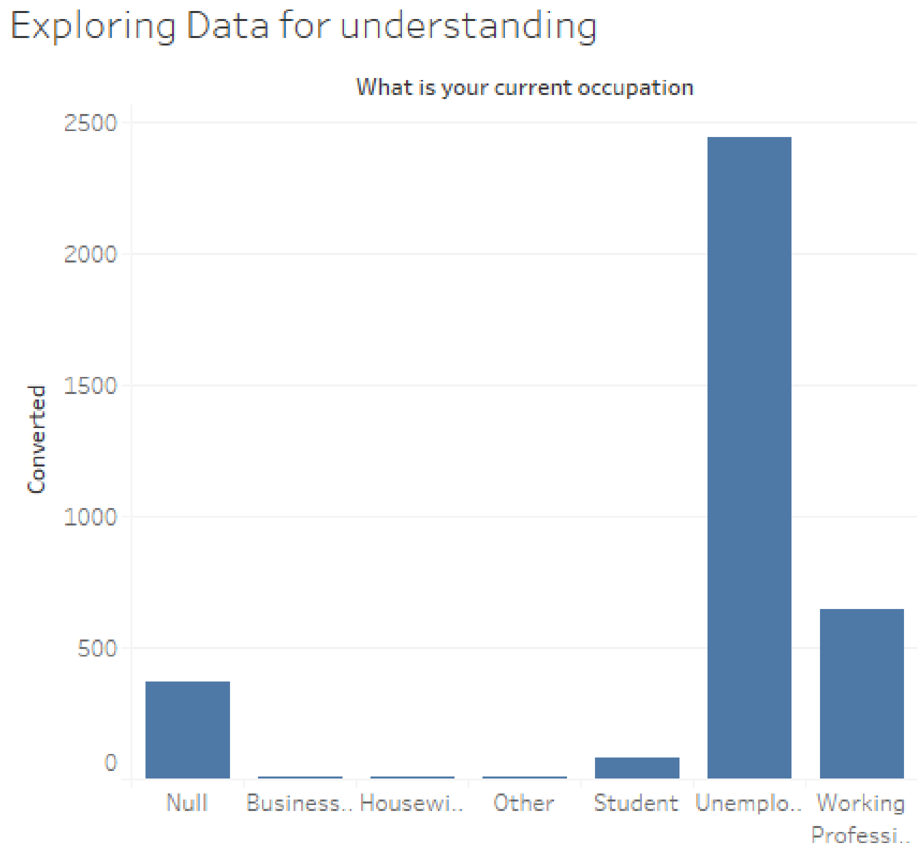




Data Analysis Approach

Step 1 :Understanding the Data:





Data Analysis Approach

- **Step 2: Handle Missing Values**

Remove columns having more than 30% missing values

Remove missing values or impute them with relevant values.

Address all the missing values.

Out[574]:

Prospect ID	0.00
Lead Number	0.00
Lead Origin	0.00
Lead Source	0.39
Do Not Email	0.00
Do Not Call	0.00
Converted	0.00
TotalVisits	1.48
Total Time Spent on Website	0.00
Page Views Per Visit	1.48
Last Activity	1.11
Country	26.63
Specialization	15.56
How did you hear about X Education	23.89
What is your current occupation	29.11
What matters most to you in choosing a course	29.32
Search	0.00
Magazine	0.00
Newspaper Article	0.00
X Education Forums	0.00
Newspaper	0.00
Digital Advertisement	0.00
Through Recommendations	0.00
Receive More Updates About Our Courses	0.00
Tags	36.29
Lead Quality	51.59
Update me on Supply Chain Content	0.00
Get updates on DM Content	0.00
Lead Profile	29.32
City	15.37
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65
I agree to pay the amount through cheque	0.00
A free copy of Mastering The Interview	0.00

Data Analysis Approach

Step 3 : Address categorical variables with Binary Mapping

Replace YES and NO with 1 and 0 using binary mapping function

Step 4: Address other categorical variables with get dummies for variables like Lead Origin, Lead source, Specialization, Last notable activity etc.

You will now have final Data frame where missing values and categorical variables are handled.

Step 5: Do the test and training split with the above data frame

Data Analysis Approach

- **Step 6 : Building a Model with GLM**
- **Step 7:Feature elimination with RFE**
- **Step 8: Check VIF and P values to eliminate features. VIFs are less than 5, so we proceed to look at p values (to be less than 0.05)**

	Features	VIF
11	Country_unknown	2.65
14	Specialization_dont know	2.18
4	LeadOrigin_Lead Add Form	1.89
3	LeadOrigin_Landing Page Submission	1.63
18	MattersMost_dont know	1.61
22	LastNotableActivity_SMS Sent	1.41
7	LeadSource_Welingak Website	1.37
1	Total Time Spent on Website	1.33
0	Do Not Email	1.19
17	CurrentOccupation_Working Professional	1.18
24	LastNotableActivity_Unsubscribed	1.07
6	LeadSource_Unknown	1.06
20	LastNotableActivity_Olark Chat Conversation	1.06
12	Specialization_Hospitality Management	1.02
13	Specialization_Services Excellence	1.01
15	HearAboutX_Email	1.01
23	LastNotableActivity_Unreachable	1.01
10	Country_Qatar	1.00
9	Country_Nigeria	1.00
8	Country_Italy	1.00
16	CurrentOccupation_Housewife	1.00
5	LeadSource_NC_EDM	1.00

< 5

Data Analysis Approach

Iterate and remove features having higher p value, after final iteration we get below.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3913	0.116	-3.369	0.001	-0.619	-0.164
Do Not Email	-1.5154	0.178	-8.532	0.000	-1.864	-1.167
Total Time Spent on Website	1.1025	0.041	27.151	0.000	1.023	1.182
LeadOrigin_Landing Page Submission	-0.8830	0.121	-7.278	0.000	-1.121	-0.645
LeadOrigin_Lead Add Form	2.4056	0.233	10.319	0.000	1.949	2.862
LeadSource_Welingak Website	2.3710	0.757	3.131	0.002	0.887	3.855
Country_unknown	1.0378	0.119	8.757	0.000	0.806	1.270
Specialization_dont know	-0.9764	0.122	-7.993	0.000	-1.216	-0.737
CurrentOccupation_Working Professional	2.3871	0.193	12.395	0.000	2.010	2.765
MattersMost_dont know	-1.1061	0.088	-12.560	0.000	-1.279	-0.933
LastNotableActivity_Had a Phone Conversation	3.1165	1.183	2.634	0.008	0.798	5.435
LastNotableActivity_SMS Sent	1.6237	0.080	20.390	0.000	1.468	1.780
LastNotableActivity_Unreachable	1.9850	0.524	3.788	0.000	0.958	3.012
LastNotableActivity_Unsubscribed	1.4439	0.519	2.780	0.005	0.426	2.462

$P < 0.05$

Data Analysis Approach

- **Step 9 : List of features after removing features from RFE, look at top 3 co-efficients for getting significant 3 variables**

Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2607.3
Date:	Mon, 10 Jun 2019	Deviance:	5214.5
Time:	14:49:02	Pearson chi2:	6.30e+03
No. Iterations:	7	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3913	0.116	-3.369	0.001	-0.619	-0.164
Do Not Email	-1.5154	0.178	-8.532	0.000	-1.864	-1.167
Total Time Spent on Website	1.1025	0.041	27.151	0.000	1.023	1.182
LeadOrigin_Landing Page Submission	-0.8830	0.121	-7.278	0.000	-1.121	-0.645
LeadOrigin_Lead Add Form	2.4056	0.233	10.319	0.000	1.949	2.862
LeadSource_Welingak Website	2.3710	0.757	3.131	0.002	0.887	3.855
Country_unknown	1.0378	0.119	8.757	0.000	0.806	1.270
Specialization_dont know	-0.9764	0.122	-7.993	0.000	-1.216	-0.737
CurrentOccupation_Working Professional	2.3871	0.193	12.395	0.000	2.010	2.765
MattersMost_dont know	-1.1061	0.088	-12.560	0.000	-1.279	-0.933
LastNotableActivity_Had a Phone Conversation	3.1165	1.183	2.634	0.008	0.798	5.435
LastNotableActivity_SMS Sent	1.6237	0.080	20.390	0.000	1.468	1.780
LastNotableActivity_Unreachable	1.9850	0.524	3.788	0.000	0.958	3.012
LastNotableActivity_Unsubscribed	1.4439	0.519	2.780	0.005	0.426	2.462

Data Analysis Approach

- Final list of variable

Do Not Email

Total Time Spent on Website

LeadOrigin_Landing Page Submission

LeadOrigin_Lead Add Form

LeadSource_Welingak Website

Country_unknown

Specialization_dont know

CurrentOccupation_Working Professional

MattersMost_dont know

LastNotableActivity_Had a Phone Conversation

LastNotableActivity_SMS Sent

LastNotableActivity_Unreachable

LastNotableActivity_Unsubscribed

Important features

Top 3 variables which are contributing for Lead conversion:

Check the co-efficient to find the most contributing 3 features

1. LastNotableActivity_Had a Phone Conversation

2. LeadOrigin_Lead Add Form

3. CurrentOccupation_Working Professional

Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2607.3
Date:	Mon, 10 Jun 2019	Deviance:	5214.5
Time:	14:49:02	Pearson chi2:	6.30e+03
No. Iterations:	7	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3913	0.116	-3.369	0.001	-0.619	-0.164
Do Not Email	-1.5154	0.178	-8.532	0.000	-1.864	-1.167
Total Time Spent on Website	1.1025	0.041	27.151	0.000	1.023	1.182
LeadOrigin_Landing Page Submission	-0.8830	0.121	-7.278	0.000	-1.121	-0.645
LeadOrigin_Lead Add Form	$\boxed{2.4056}$	0.233	10.319	0.000	1.949	2.862
LeadSource_Welingak Website	2.3710	0.757	3.131	0.002	0.887	3.855
Country_unknown	1.0378	0.119	8.757	0.000	0.806	1.270
Specialization_dont know	-0.9764	0.122	-7.993	0.000	-1.216	-0.737
CurrentOccupation_Working Professional	$\boxed{2.3871}$	0.193	12.395	0.000	2.010	2.765
MattersMost_dont know	-1.1061	0.088	-12.560	0.000	-1.279	-0.933
LastNotableActivity_Had a Phone Conversation	$\boxed{3.1165}$	1.183	2.634	0.008	0.798	5.435
LastNotableActivity_SMS Sent	1.6237	0.080	20.390	0.000	1.468	1.780
LastNotableActivity_Unreachable	1.9850	0.524	3.788	0.000	0.958	3.012
LastNotableActivity_Unsubscribed	1.4439	0.519	2.780	0.005	0.426	2.462

Top 3 variables which are contributing for Lead conversion: check co- efficients to find most contributing

1. **LastNotableActivity**

2. **LeadOrigin**

3. **CurrentOccupation _Working Professional**

Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2607.3
Date:	Mon, 10 Jun 2019	Deviance:	5214.5
Time:	14:49:02	Pearson chi2:	6.30e+03
No. Iterations:	7	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3913	0.116	-3.369	0.001	-0.619	-0.164
Do Not Email	-1.5154	0.178	-8.532	0.000	-1.864	-1.167
Total Time Spent on Website	1.1025	0.041	27.151	0.000	1.023	1.182
LeadOrigin_Landing Page Submission	-0.8830	0.121	-7.278	0.000	-1.121	-0.645
LeadOrigin_Lead Add Form	2.4056	0.233	10.319	0.000	1.949	2.862
LeadSource_Welingak Website	2.3710	0.757	3.131	0.002	0.887	3.855
Country_unknown	1.0378	0.119	8.757	0.000	0.806	1.270
Specialization_dont know	-0.9764	0.122	-7.993	0.000	-1.216	-0.737
CurrentOccupation_Working Professional	2.3871	0.193	12.395	0.000	2.010	2.765
MattersMost_dont know	-1.1061	0.088	-12.560	0.000	-1.279	-0.933
LastNotableActivity_Had a Phone Conversation	3.1165	1.183	2.634	0.008	0.798	5.435
LastNotableActivity_SMS Sent	1.6237	0.080	20.390	0.000	1.468	1.780
LastNotableActivity_Unreachable	1.9850	0.524	3.788	0.000	0.958	3.012
LastNotableActivity_Unsubscribed	1.4439	0.519	2.780	0.005	0.426	2.462

Data Analysis Approach

- **Step 10: Predict the values , create confusion matrix and evaluate the accuracy of the model (accuracy= 0.81)**

```
# Confusion matrix
from sklearn import metrics
confusion = metrics.confusion_matrix(y_train_pred_final.Convert, y_train_pred_final.predicted )
print(confusion)
```

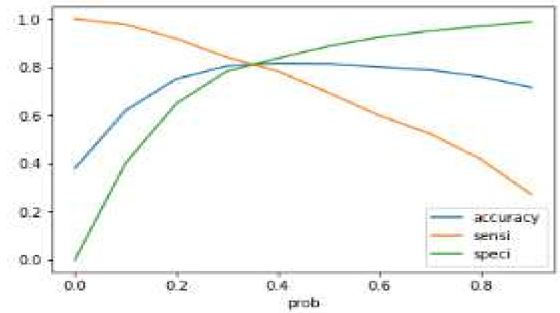
```
[[3506  447]
 [ 741 1678]]
```

```
# checking the  accuracy.
print(metrics.accuracy_score(y_train_pred_final.Convert, y_train_pred_final.predicted))
```

```
0.8135593220338984
```

- **Step 11: Find the cut off with specificity sensitivity and accuracy cut off=0.3**

```
In [62]: # Let's plot accuracy sensitivity and specificity for various probabilities.
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```




```
In [63]: y_train_pred_final['final_predicted'] = y_train_pred_final.Convert_Prob.map( lambda x: 1 if x > 0.3 else 0)
y_train_pred_final.head()
```

Step 12: Assigning Lead score

Assigning lead score from 0 to 100

In [98]:
```
y_pred_final['Score'] = y_pred_final.Convert_Prob.map(lambda x: round(x*100))
y_train_pred_final['Score'] = y_train_pred_final.Convert_Prob.map(lambda x: round(x*100))
```

We can see the equation is:

Logit(P) = - 0.3913 - 1.5154 x Do Not Email is Yes + 1.1025 x Total Time Spent on Website - 0.8830 x Lead Origin of 'Landing Page Submission' + 2.4056 x Lead Origin of 'Lead Add Form' + 2.3710 x Lead Source of 'Welingak Website' + 1.0378 x Country is Unknown - 0.9764 x Specialization dont know + 2.3871 x CurrentOccupation is Working Professional - 1.1061 x What MattersMost is dont know + 3.1165 x LastNotableActivity is 'Had a Phone Conversation' + 1.6237 x LastNotableActivity is 'SMS Sent' + 1.9850 x LastNotableActivity is 'Unreachable' + 1.4439 x LastNotableActivity is 'Unsubscribed'

Score = Logit(P) x 100 (score between 0 and 100)

Suggestions/Recommendations

- 1. Marketing Teams can call the all the candidates who are unemployed or working professionals to explain about the course and try to give some special offers to attract the prospects.

- 2. Marketing Teams can call the candidates whose LeadOrigin is LeadAddform or Landingpage.

- 3. Marketing Teams can call candidates with LastNotableActivity as who had phone call, SMS sent or email opened.

Thank You