X Education Data Analysis for Lead conversion Report

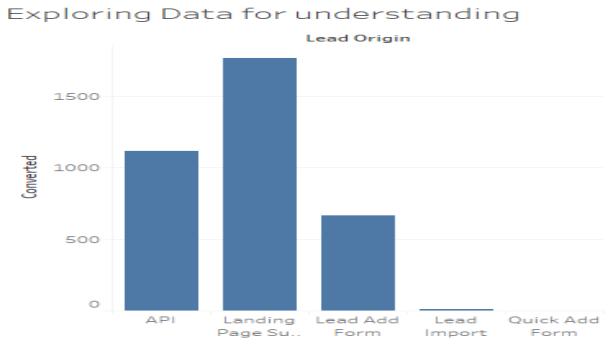
Problem Statement:

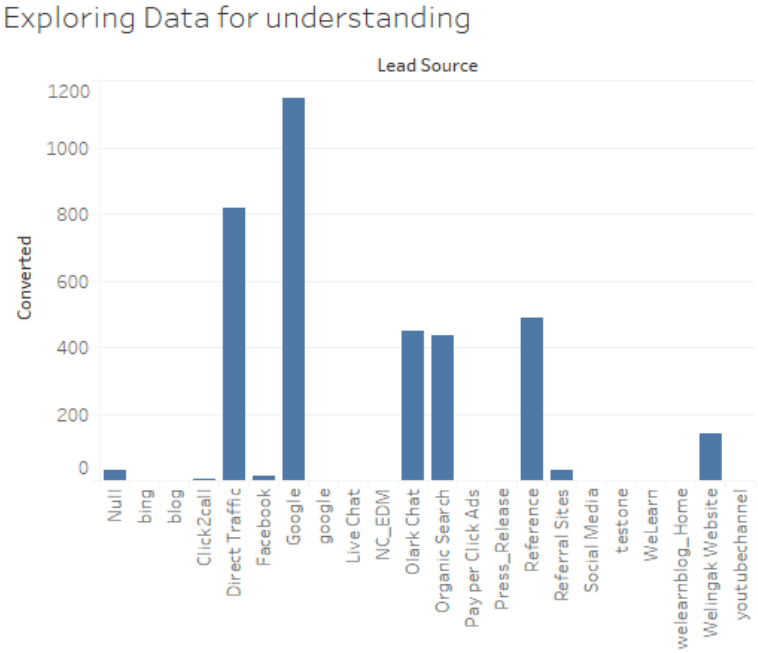
The company markets its courses on several websites and captures the browsing history and application filled by the interested candidates. Observation is only around 30% of the interested candidates actually proceed to take the course. Analyze the data and provide insights to increase the conversion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

Step 1: Understanding the Data:

Understand the different features and do some basic EDA to understand the spread of data





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.

Replacing missing values in country column by 'unknown' and replacing all 'select' value with 'don't know'. Other than that removing all null values.

Step 3 : Address categorical variables with Binary Mapping i.e. Data Preparation

Creating the binary mapping function and, then applying the binary map to the final dataset.

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, Country, How did you hear about X Education, What is your current occupation, What matters most to you in choosing a course, 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 we will be Putting feature variable to X and Putting response variable to y and then splitting the data into test and train.

Step 6 : Building a Model with GLM

Using logistic regression we will build the model.

Step 7: Feature elimination with RFE

Assessing the model with statsmodels

Check VIF and P values to eliminate features, since VIFs are less than 5, so we proceed

to look at p values (to be less than 0.05). from checking the p-values we removed columns LeadSource_NC_EDM, Newspaper, Country_Italy, Country_Nigeria, LastNotableActivity_Resubscribed to emails, Country_Qatar, CurrentOccupation_Housewife, HearAboutX_Email, LeadSource_Unknown, Specialization_Services Excellence, Specialization_Hospitality Management, LastNotableActivity_Olark Chat Conversation

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

- 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

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

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**

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

1. **LastNotableActivity**

2. **LeadOrigin**
3. **CurrentOccupation_Working Professional**

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

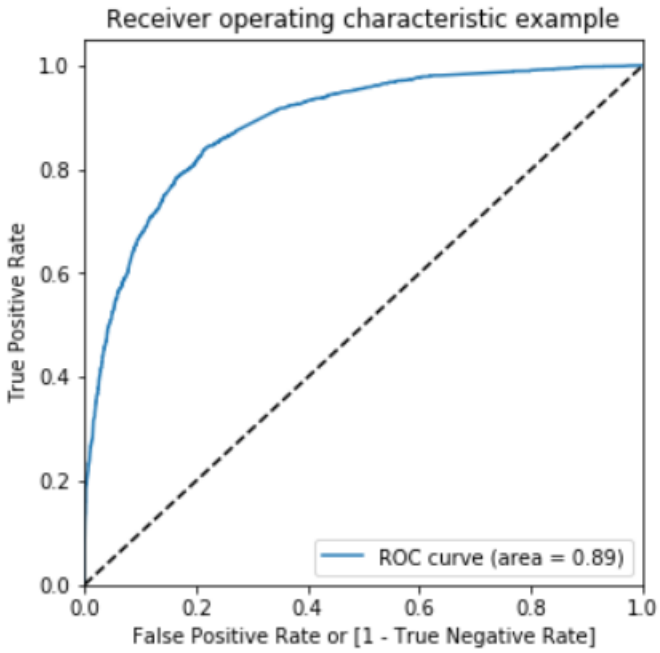
```
# Confusion matrix
from sklearn import metrics
confusion = metrics.confusion_matrix(y_train_pred_final.Convert, y_train_pred_final.predicted )
print(confusion)
```

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

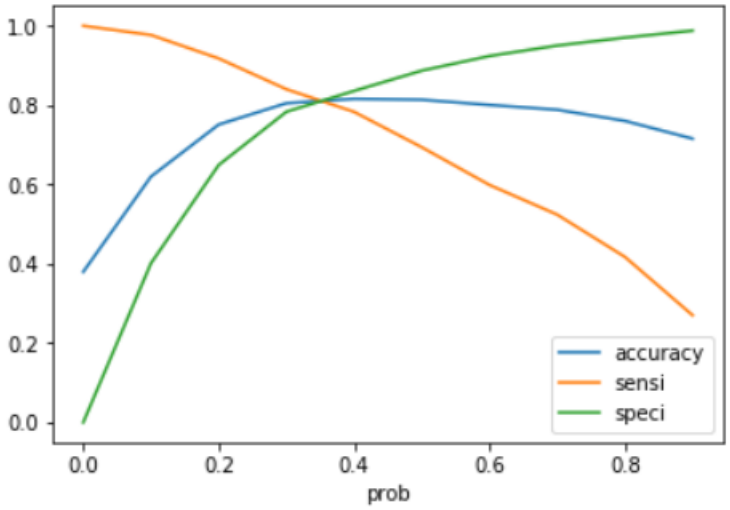
Step 11: Find the cut off with specificity sensitivity and accuracy cut off=0.3

Sensitivity of the model is: .6936
Specificity of the model is: .8869
false postive rate: .1130
positive predictive value : .7896
Negative predictive value : .8255

Step 12: ROC Curve : the area is 0.89



Let's plot accuracy sensitivity and specificity for various probabilities:



Step 13: Assign 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 to improve Lead conversion

- 1. 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. Call the candidates whose LeadOrigin is LeadAddform or Landingpage.
- 3.target candidates with LastNotableActivity as who had phone call, SMS sent or email opened.