LEAD SCORING CASE STUDY

Problem Statement

- An education company named X Education sells online courses to industry professionals
- On any given day, many professionals who are interested in the courses land on their website and browse for courses
- Any individual who shows interest in X Education’s online courses become a lead for the company
- Goal of this case study is to identify if a particular lead will get converted as a customer in future.
- By identifying such potential customers, company wants to approach them with some good offers and contents.

Analysis Approach

- **Data Exploration**
  - Reading data
  - Analyzing data size, data types etc.
- **Data Quality Checks**
  - Duplication checks
  - Missing values analysis
  - Outlier analysis
- **Data Cleansing**
  - Missing value imputation
  - Outliers treatment

Analysis Approach...

- **Data Preparation**
  - Binary and one-hot encoding
  - Categorization of low count data to *Others*
  - *Dummy encoding*
  - *Train-Test Split*
  - *Scaling using min max scalar*
  - *Drop correlated columns*

Analysis Approach...

- **Model building**
  - Feature selection using RFE
  - Validate correlation-ship between RFE selected variables
  - Assess model using StatsModel
    - Feature elimination using p-value and VIF score
  - Prediction on train set
    - Validate predicted values using different metrics such as sensitivity, specificity, accuracy etc.
    - Finding optimal cut-off point
  - Predictions on the test set
    - Validate similar metrics on test predicted values

Data Exploration

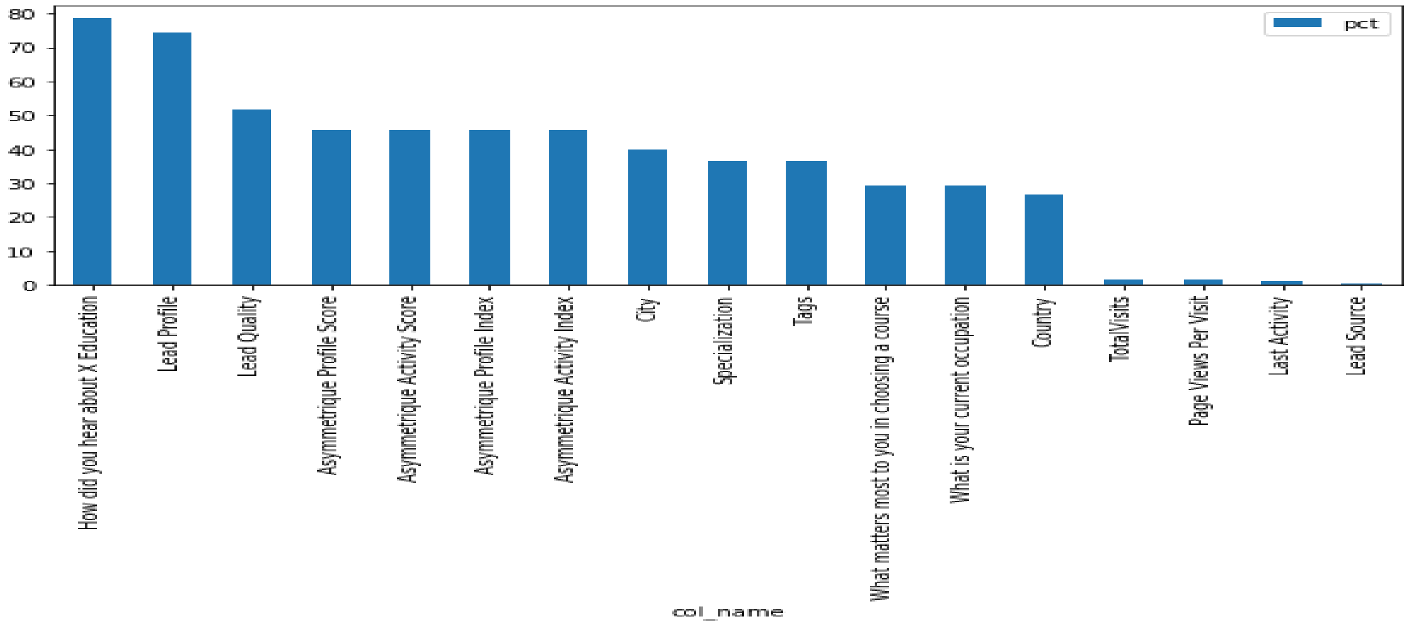
- Leads data contained around 9000+ records with 37 columns
- Many columns are with binary(Yes/No) answers
- Converted column shows whether lead was actually converted or not, i.e. lead is converted to active customer or not.

Data Quality Checks

- **Duplication checks**
  - ProspectID column is used to label individual customer. There are no duplicates found in this column.
- **Missing Value Checks**
  - ***Select*** value in any column is because customer did not select any option from dropdowns available on application form. So, this value can be treated as null as we are not sure about customer’s point of view

Data Quality Checks
Missing Value plot

Following plot shows missing values % in each column.



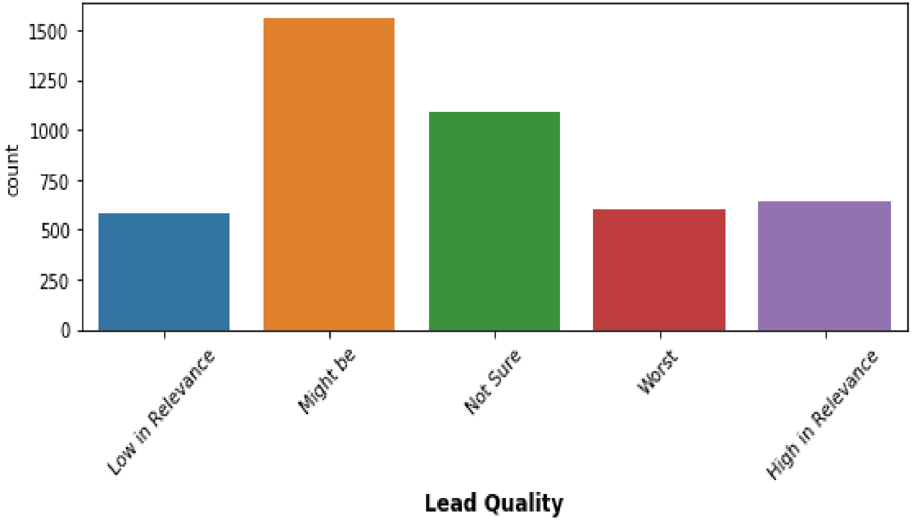
Data Cleansing
Missing value imputation

Here, we have checked value count distribution for each variable. Using this, we can decide if a particular value can be imputed in place of NULLs.

An example in explained below.

Adjacent plot is the ***Lead Quality*** column value count distribution.

Lead Quality column indicates the quality of lead based on the data and intuition the employee who has been assigned to the lead. So, We can impute null values in ***Lead Quality*** column with ***Not Sure*** value as we do not know quality of the lead. Also, this is possibly not a drop down column in the form. This column's data is derived on the intuition of other data points.

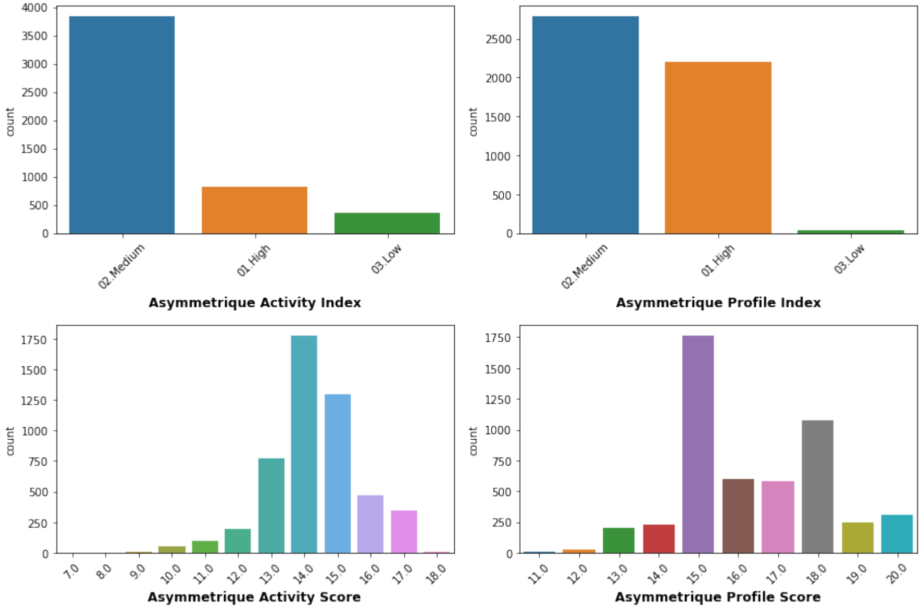


Data Cleansing
Missing value imputation

We have also analyzed the count distribution of following variables

- ***Asymmetrique Activity Index***
- ***Asymmetrique Profile Index***
- ***Asymmetrique Activity Score***
- ***Asymmetrique Profile Score***

An index and score assigned to each customer based on their activity and their profile. Also, there is no specific pattern how scores and indexes are calculated. So, It will be highly difficult to predict value to be imputed in null records. So, it is better to drop columns these columns



Data Cleansing
Missing value imputation

- Using similar approached, we imputed country as India in NULL country values, as customer visiting website are mostly Indian customers. So, we can follow this intuition
- NULL values in ***What matters most to you in choosing a course*** is imputed with Better Career Prospects which constitutes 99% of available values in this column
- Tags and City column is dropped as there is a lot of variance in both of these columns. So, it is difficult to impute any value in place of NULLs
- Also, we have dropped rows with more than 30% missing data

Data Cleansing

Outlier treatment

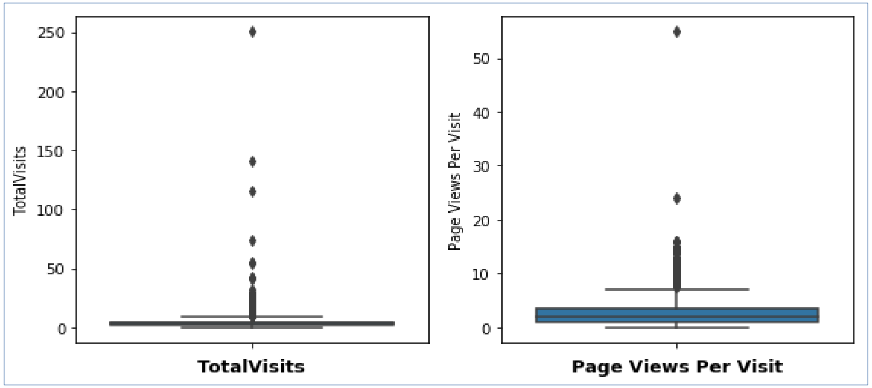
Out[30]:

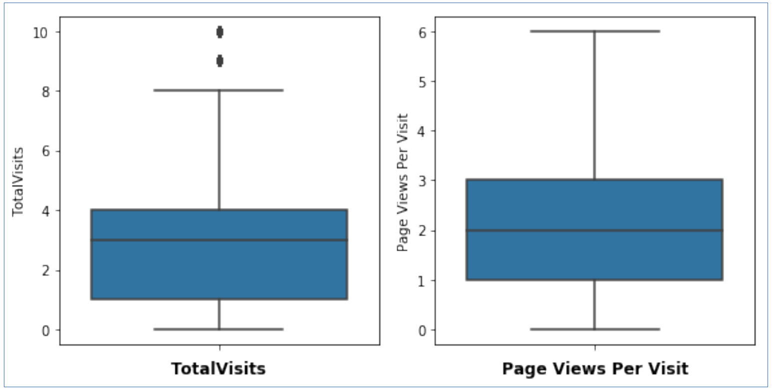
	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	8669.000000	8669.000000	8669.000000	8669.000000	8669.000000
mean	617121.545968	0.392433	3.613104	504.743454	2.479183
std	23765.460081	0.488320	4.911026	547.515388	2.149064
min	579533.000000	0.000000	0.000000	0.000000	0.000000
2.5%	581213.100000	0.000000	0.000000	0.000000	0.000000
50%	615326.000000	0.000000	3.000000	267.000000	2.000000
75%	638014.000000	1.000000	5.000000	953.000000	3.500000
80%	642029.600000	1.000000	5.000000	1104.400000	4.000000
90%	650789.200000	1.000000	7.000000	1387.000000	5.000000
95%	655496.600000	1.000000	10.000000	1566.000000	6.000000
99%	659609.240000	1.000000	17.000000	1843.320000	9.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000

As per statistical analysis of numerical column in dataset, we found outliers in ***TotalVisits*** and ***Page Views Per Visit*** columns.

Data Cleansing

Outlier treatment





- We removed outliers above 95th percentiles
- Above plots shows, distribution of both the variables, before and after removing outliers

Data Preparation

- In this section, we did binary encoding for following columns
  - 'Do Not Email'
  - 'Do Not Call'
  - 'Search'
  - 'Newspaper'
  - 'Digital Advertisement'
  - 'Through Recommendations'
  - 'A free copy of Mastering The Interview'
- Create dummy variables using remaining variables

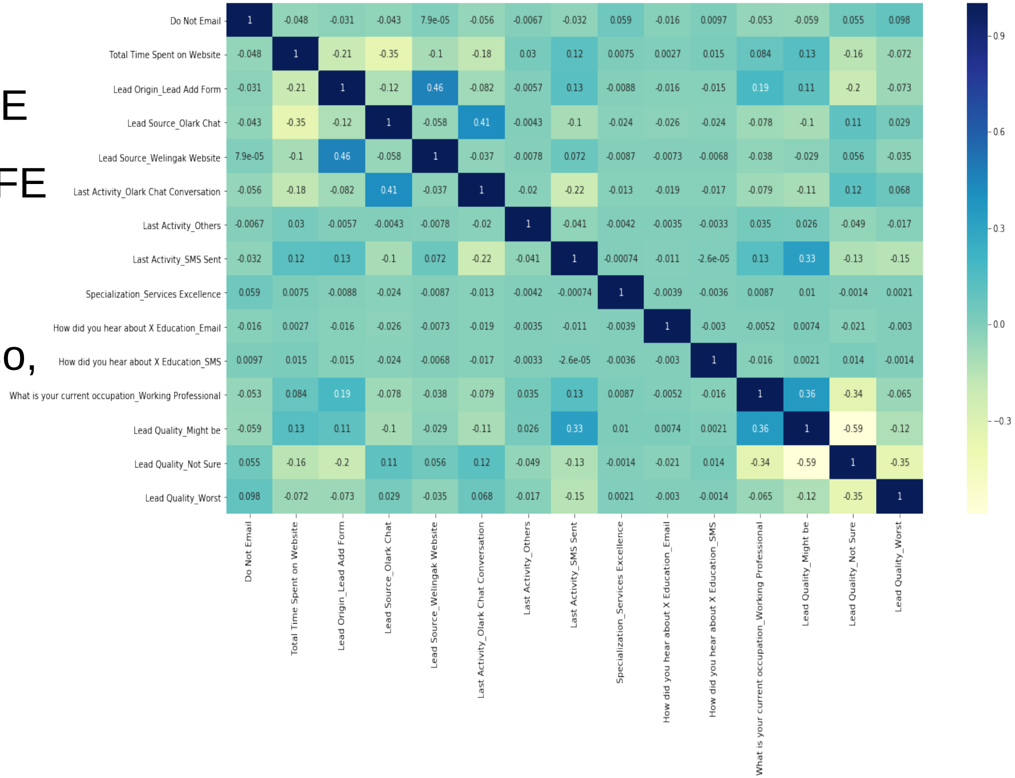
Data Preparation

- Next step is train-test split
- After train-test split, train variables are scaled to confine the values between 0 and 1 using **MinMaxScaler()**
- Same **MinMaxScaler** instance to be used to scale test dataset as well
- Check correlation-ship between new dummy variables created to check multicollinearity
- Drop highly correlated variables

Model Building

Feature selection using RFE

- Feature selection using RFE
  - Selected 15 top variables using RFE
  - Check multi-collinearity between RFE selected variables
  - There is no high multi-collinearity between RFE selected variables. So, no need to drop any variables



Model Building

Assessing the model with StatsModels

Following statistical summary is given by final model derived after iterations. Variables are dropped and model is re-created for validating results. Final model shows that p-value of all the independent variables is close to 0 and thus are significant.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 5628
Model:                            GLM   Df Residuals:                     5616
Model Family:                Binomial   Df Model:                           11
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2009.3
Date:                Mon, 26 Aug 2019   Deviance:                       4018.7
Time:                        14:07:37   Pearson chi2:                 5.92e+03
No. Iterations:                     7   Covariance Type:             nonrobust
========================================================================================================================
                                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                          0.0825      0.141      0.584      0.559      -0.195       0.360
Do Not Email                                  -1.2053      0.193     -6.253      0.000      -1.583      -0.828
Total Time Spent on Website                    4.1958      0.187     22.388      0.000       3.829       4.563
Lead Origin_Lead Add Form                      2.6717      0.251     10.633      0.000       2.179       3.164
Lead Source_Olark Chat                         1.8007      0.124     14.535      0.000       1.558       2.043
Lead Source_Welingak Website                   4.2081      1.041      4.043      0.000       2.168       6.248
Last Activity_Olark Chat Conversation         -1.1846      0.183     -6.458      0.000      -1.544      -0.825
Last Activity_SMS Sent                         1.2959      0.087     14.955      0.000       1.126       1.466
What is your current occupation_Working Professional     1.7316      0.216      8.016      0.000       1.308       2.155
Lead Quality_Might be                         -1.3326      0.161     -8.295      0.000      -1.648      -1.018
Lead Quality_Not Sure                         -3.0976      0.141    -21.978      0.000      -3.374      -2.821
Lead Quality_Worst                            -5.1948      0.374    -13.893      0.000      -5.928      -4.462
========================================================================================================================
```

Model Building

Assessing the model with StatsModels

- We have also validated VIF score of all the features to identify any multi-collinearity between chosen independent variables.
- Adjacent screenshot shows VIF score for all the features given by final model
- We can see that VIF score for all the features is below 5. Thus, there is no collinearity issue between selected features.

	Features	VIF
0	const	12.55
10	Lead Quality_Not Sure	2.59
9	Lead Quality_Might be	2.13
3	Lead Origin_Lead Add Form	1.63
11	Lead Quality_Worst	1.58
4	Lead Source_Olark Chat	1.39
2	Total Time Spent on Website	1.34
5	Lead Source_Welingak Website	1.33
6	Last Activity_Olark Chat Conversation	1.27
8	What is your current occupation_Working Profes...	1.23
7	Last Activity_SMS Sent	1.20
1	Do Not Email	1.03

Prediction on train data set and Model Validation

- Conversion probability is calculated by model with around 84% accuracy in train data set

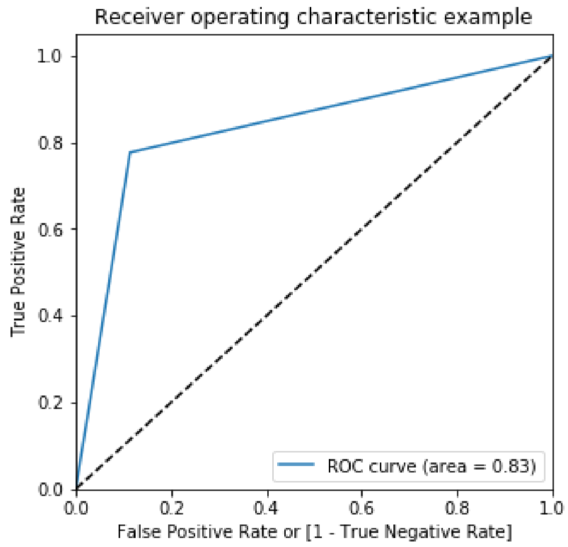
```
#accuracy
print(metrics.accuracy_score(y_train_dataframe.Converted,y_train_dataframe.Predicted))
```

```
0.8431058990760484
```

- **ROC curve:**

  Area under the curve is 0.83.

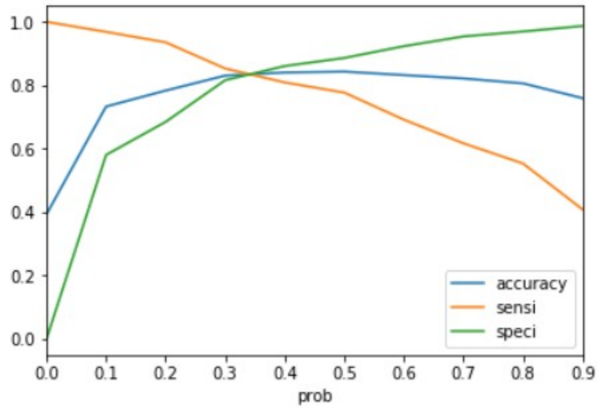
  This indicates that model has good measure of separability



Model Validation...

- We have also calculated accuracy,Sensitivity and Specificity measures for different cut-off points as mentioned in below table. As per the below table and plot, 0.3 seems to be the optimal cut-off point.

	prob	accuracy	sensi	speci
0.0	0.0	0.393213	1.000000	0.000000
0.1	0.1	0.732942	0.968369	0.580381
0.2	0.2	0.783227	0.935382	0.684627
0.3	0.3	0.830490	0.852689	0.816105
0.4	0.4	0.840263	0.808857	0.860615
0.5	0.5	0.843106	0.776774	0.886091
0.6	0.6	0.832090	0.691369	0.923280
0.7	0.7	0.821429	0.616810	0.954026
0.8	0.8	0.805792	0.553095	0.969546
0.9	0.9	0.759240	0.408043	0.986823



Predictions On The Test Data set

- Following are the metrics calculated for test dataset predictions
  - Accuracy Score

```
In [897]: metrics.accuracy_score(y_test_dataframe.Converted,y_test_dataframe.finalConverted)
Out[897]: 0.8329879817654372
```

  - Sensitivity & Specificity

```
In [899]: cf = metrics.confusion_matrix(y_test_dataframe.Converted,y_test_dataframe.finalConverted)
print("Sensitivity",cf[1,1]/(float(cf[1,1]+cf[1,0])))
print("Specificity",cf[0,0]/(float(cf[0,0]+cf[0,1])))
print("Positive predication",cf[1,1]/float(cf[1,1]+cf[0,1]))
print("Negtive prediction",cf[0,0]/float(cf[0,0]+cf[1,0]))

Sensitivity 0.8509249183895539
Specificity 0.821954484605087
Positive predication 0.7461832061068703
Negtive prediction 0.8996336996336997
```

All the above metrics have values very close to what we got with train data set So, model's behavior is consistent and it can be considered to be a good model

Observations & Recommendations

As per the model summary mentioned, we can say that following are few variables which we can use to optimize our conversion rate.

1) **Lead Source_Welingak Website** - Highest positive standardized coefficient value
2) **Total Time Spent on Website** - 2nd Highest standardized positive coefficient value
3) **Lead Origin_Lead Add Form** - 3rd highest standardized positive coefficient value
4) **Lead Quality_Worst** - Highest standardized negative coefficient value

Marketing team needs to concentrate on these variables to optimize our conversion rate.