Lead Scoring Case Study

Case study team

- Suraj
- Abhilasha
- Aishwarya Ram
- Anurag

Background – Lending Club case Study

Background

- Education company named X Education sells online courses to industry professionals. And wants to maximize the leads who visit the company website through various means.

Objective

- The typical lead conversion rate at X education is around 30%. We as a analysts have to find the prominent leads and increase the efficiency.

Data understanding

Types of variables

- Lead sources
- Total visits
- Time spent
- City

Lead Sources like Google, Direct Traffic, Olark Chat, Organic Search, Reference, Welingak Website and many more.

Total visits a whole number having large variation from person to person.

Time spent a numerical value having large variation from person to person.

City the location from which they are staying.

Lead score

On doing univariate analysis on “Lead Score” column

```
In [15]: ##Analyzing the column based on lead source
         ##Lets list number leads listed in the dataset.
         lead_df["Lead Source"].value_counts()

Out[15]: Google              2868
         Direct Traffic      2543
         Olark Chat          1753
         Organic Search      1154
         Reference           443
         Welingak Website    129
         Referral Sites      125
         Facebook            31
         bing                6
         google              5
         Click2call          4
         Press_Release       2
         Live Chat           2
         Social Media        2
         youtubechannel      1
         testone             1
         blog                1
         NC_EDM              1
         WeLearn             1
         Pay per Click Ads   1
         welearnblog_Home    1
         Name: Lead Source, dtype: int64
```

Do not e-mails

On doing univariate analysis on “Do Not Email” column

```
In [16]: #Analysis of vcolumn do not email
         lead_df["Do Not Email"].value_counts()

Out[16]: No    8358
         Yes    716
         Name: Do Not Email, dtype: int64
```

NUMERICAL LEADS DESCRIPTION

On doing analysis on 3 numerical columns

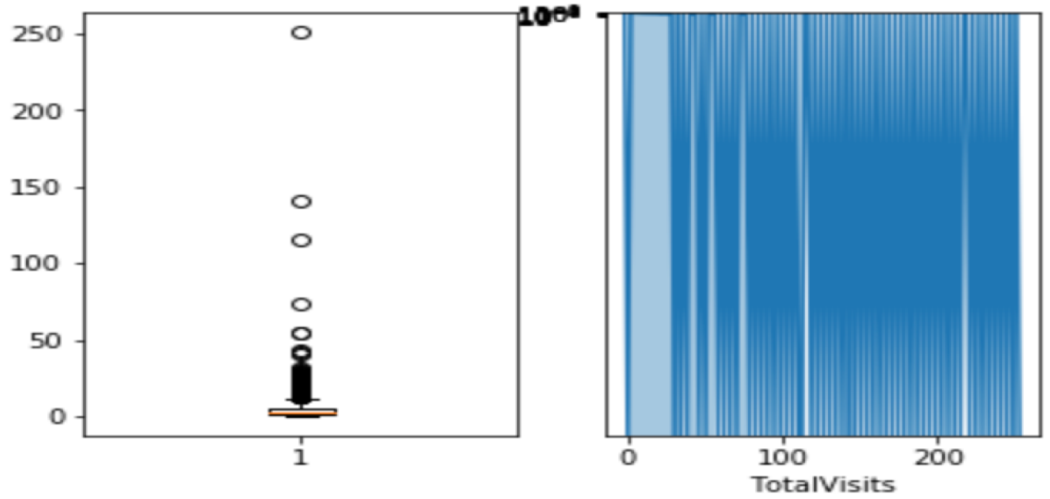
```
# Checking outliers at 25%,50%,75%,90%,95% and 99%
num_lead_df.describe(percentiles=[.25,.5,.75,.90,.95,.99])
```

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9074.000000	9074.000000	9074.000000
mean	3.456028	482.887481	2.370151
std	4.858802	545.256560	2.160871
min	0.000000	0.000000	0.000000
25%	1.000000	11.000000	1.000000
50%	3.000000	246.000000	2.000000
75%	5.000000	922.750000	3.200000
90%	7.000000	1373.000000	5.000000
95%	10.000000	1557.000000	6.000000
99%	17.000000	1839.000000	9.000000
max	251.000000	2272.000000	55.000000

Numerical analysis of total visits

In the sub plot 121 the outliers are seen and a box plot is used. A log graph is shown in subplot 122 to show the density of users having different total visits.

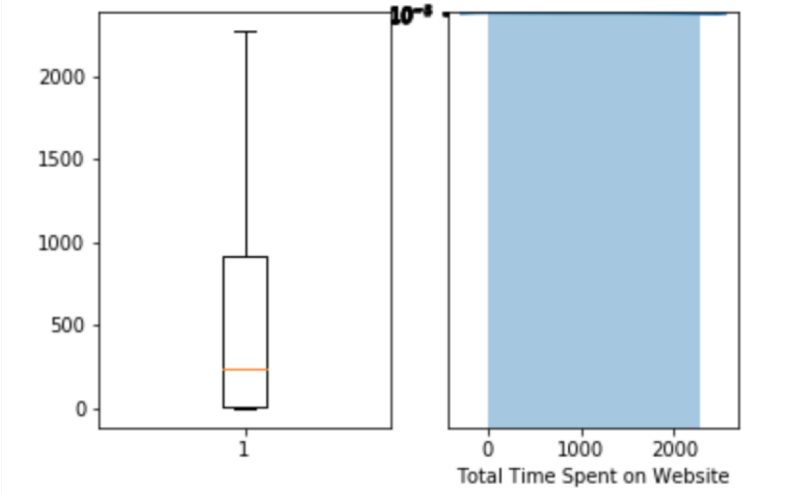
```
# simple density subplot
plt.subplot(1, 2, 1)
plt.boxplot(num_lead_df['TotalVisits'])


# log density subplot
plt.subplot(1, 2, 2)
sns.distplot(num_lead_df['TotalVisits'])
plt.yscale('log')
plt.show()
```



Numerical analysis of total time spent

In the sub plot 121 the outliers are seen and a box plot is used. A log graph is shown in subplot 122 to show the density of users having different time spent .

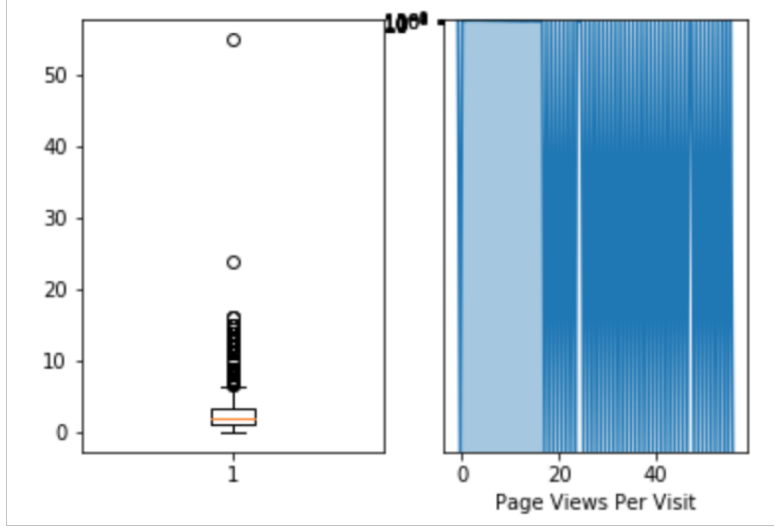
```
# simple density subplot
plt.subplot(1, 2, 1)
plt.boxplot(num_lead_df['Total Time Spent on Website'])


# log density subplot
plt.subplot(1, 2, 2)
sns.distplot(num_lead_df['Total Time Spent on Website'])
plt.yscale('log')
plt.show()
```



Numerical analysis of page Views per visit

In the sub plot 121 the outliers are seen and a box plot is used. A log graph is shown in subplot 122 to show the density of users having different page views per visit.

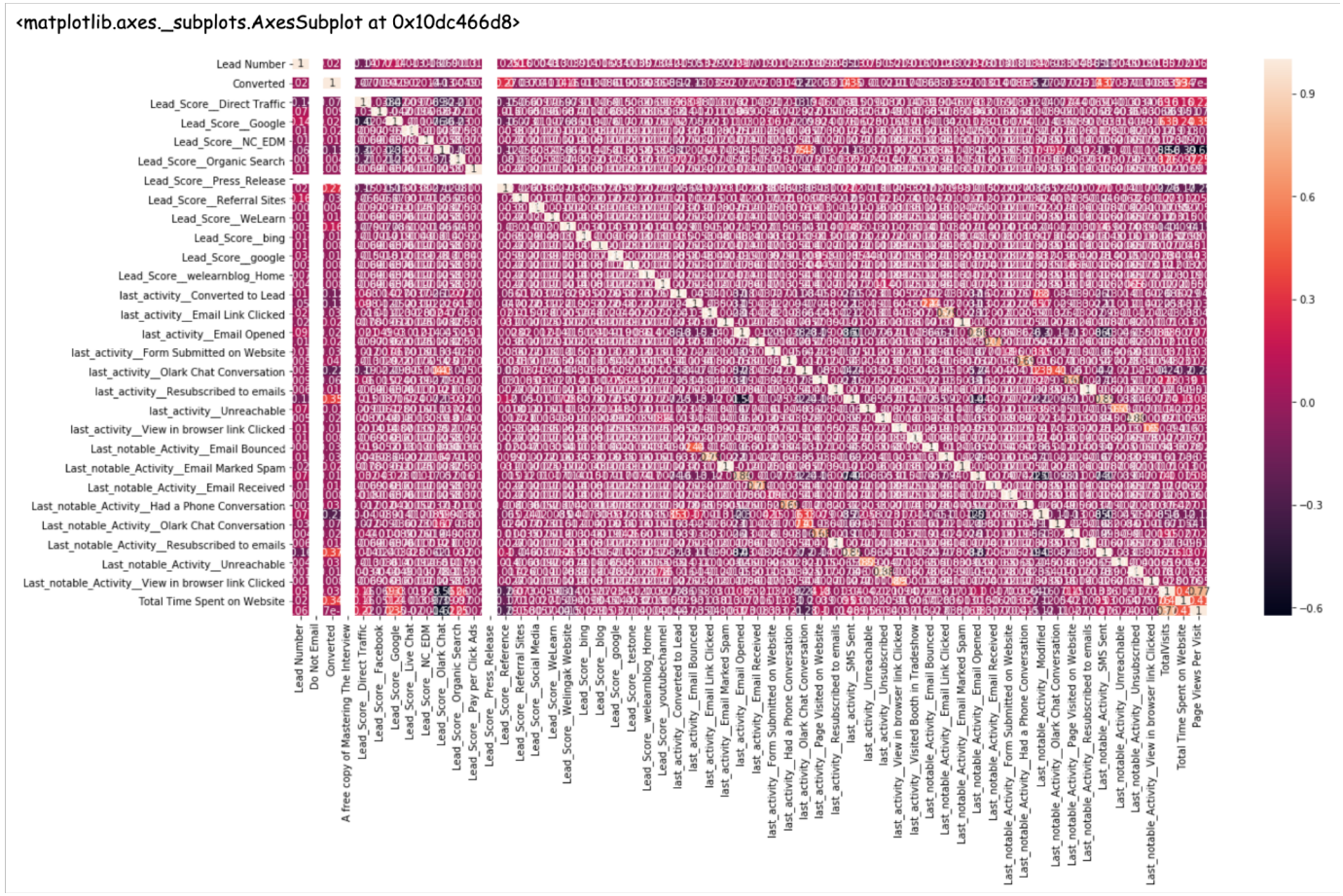
```
# simple density subplot
plt.subplot(1, 2, 1)
plt.boxplot(num_lead_df['Page Views Per Visit'])


# log density subplot
plt.subplot(1, 2, 2)
sns.distplot(num_lead_df['Page Views Per Visit'])
plt.yscale('log')
plt.show()
```



Co-relation matrix

In the following co-relation matrix we can see the co-relation between all the variables. And the blanks in the matrix are due to null values.

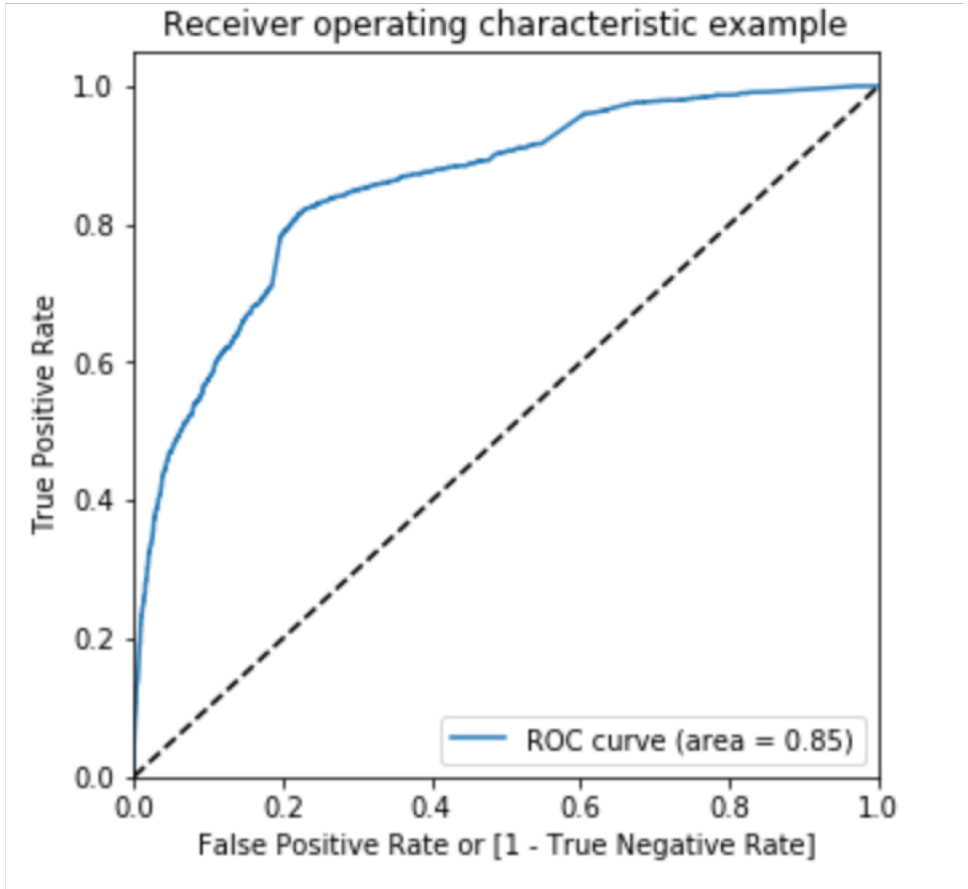
All the variables which have less co-relation are removed and the co-relation matrix is obtained.



Updated co-relation matrix

ROC curve

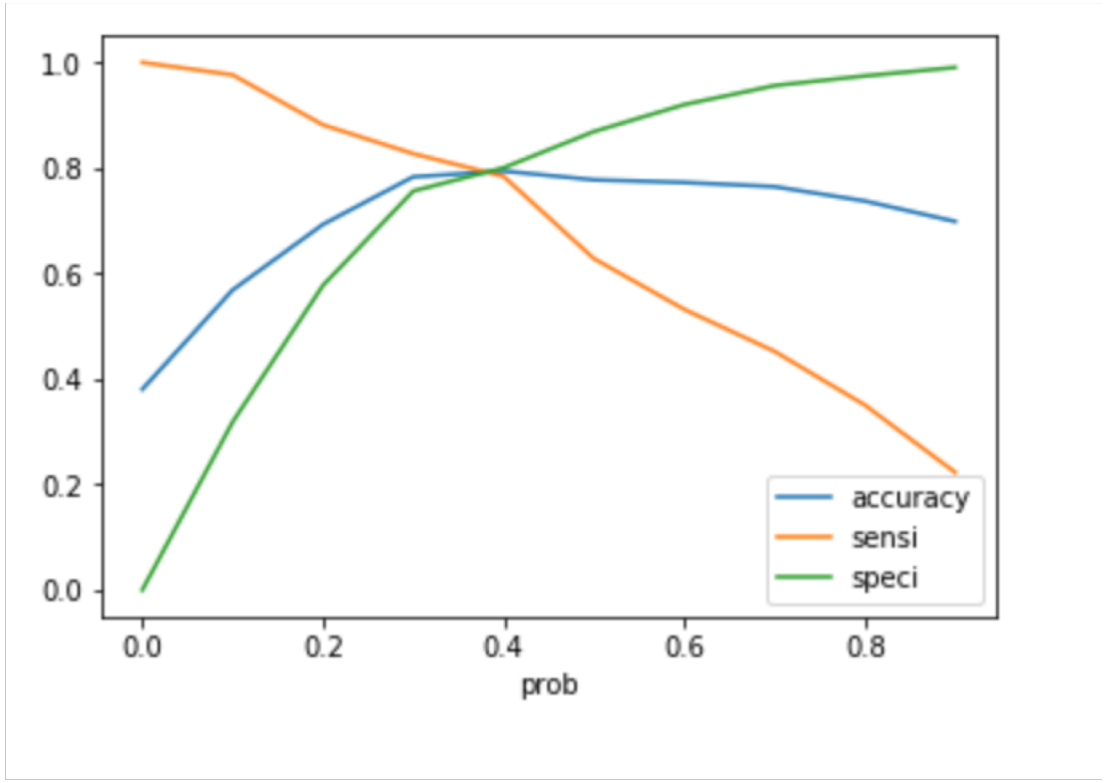
An ROC curve demonstrates several things:
It shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Plotting the ROC curve to know the accuracy of the model and how good the model is.

Finding the optimal cut-off point

Optimal cut-off on **Logistic** Regression probabilities. Reference line: intersection point at which there is a balance between sensitivity and specificity; it corresponds to the **optimal cut-off** on **logistic** regression probabilities



Conclusion

- The final model has a accuracy of 80% and it help the institute to increase the accuracy from nearly 30% to 80%.
- This help the company to concentrate on prominent leads to increase the profits of the company.
- And this would require less work from the company and less man power is required.