Lead Score Case Study

Submitted by

1.Aswathi P

2.Shilpi

Problem Statement

X Education sells online courses to industry professionals

Although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted

To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Steps

1. Importing necessary libraries

2. Read and inspect the data

2.1 Data Cleaning

3.EDA(Exploratory Data Analysis)

3.1 Univariate Analysis

3.1.1Categorical variable

3.1.2 Numerical Variable

3.2.Multivariate analysis

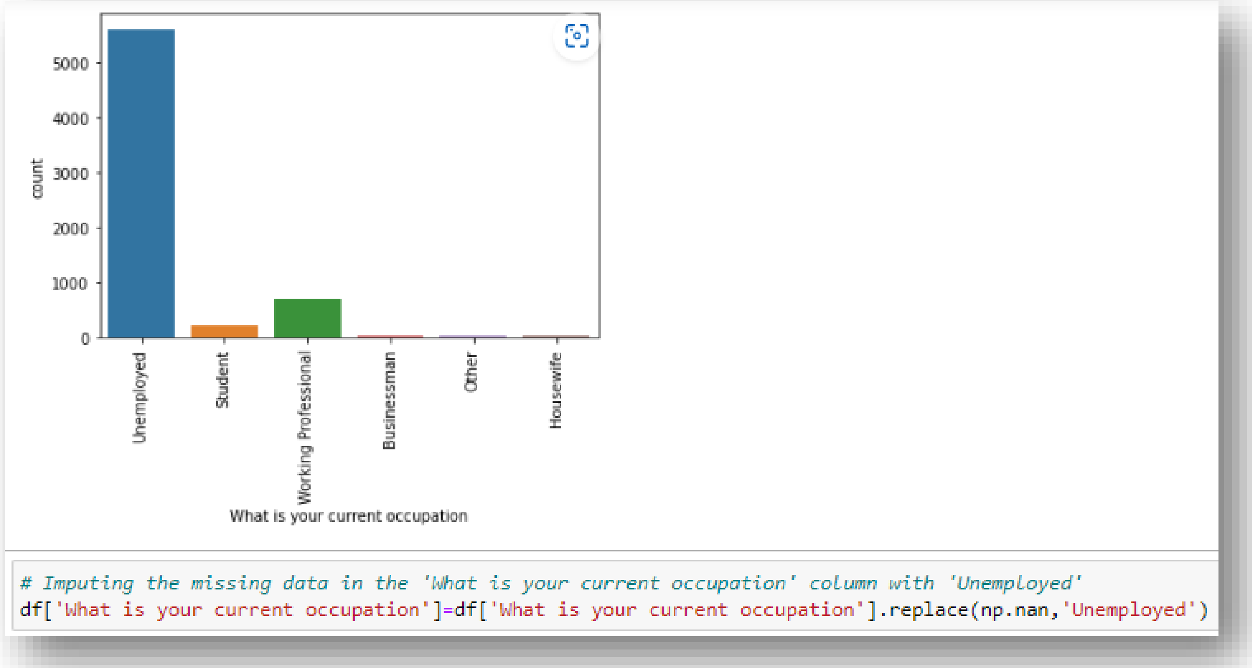
- 4.Data Preparation
- 4.1.Creating Dummy Variable for Categorical columns
- 4.2.Data Split
- 4.3.Scaling
- 5.Model building And Prediction on train set
- 6. Model Evaluation
- 7. Prediction
- 8. Conclusion

Data Cleaning

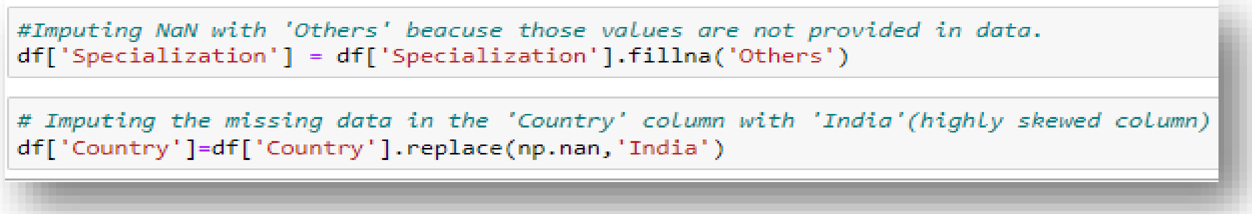
No duplicates in the data

Converted "Select" to Null

Imputed some categorical columns



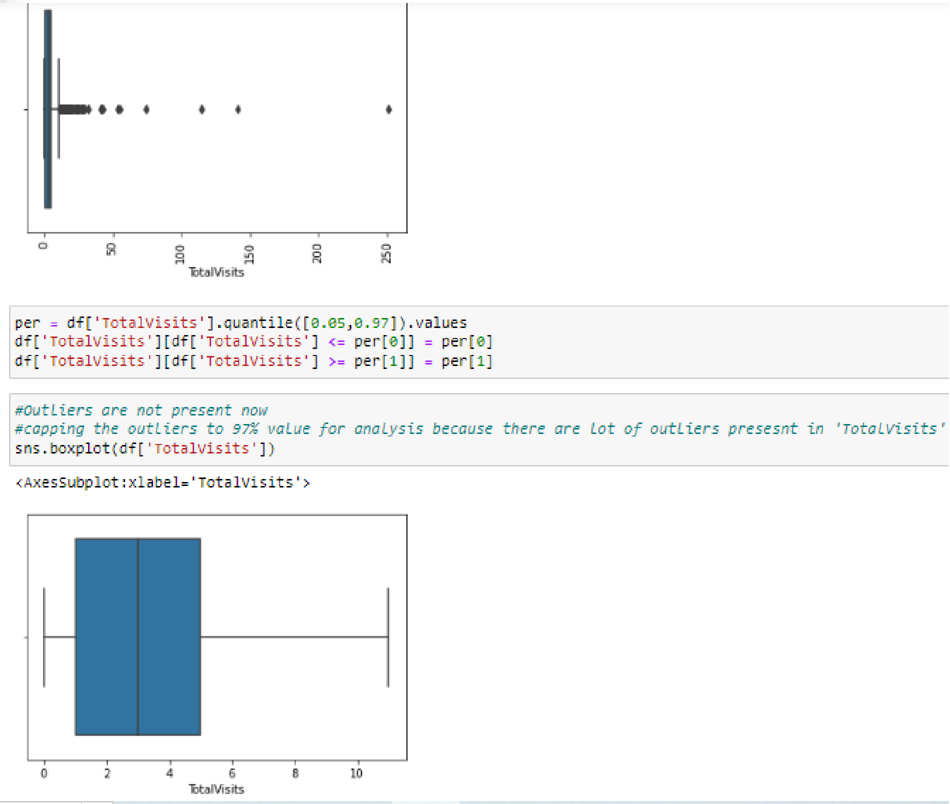
```
# Imputing the missing data in the 'What is your current occupation' column with 'Unemployed'
df['What is your current occupation']=df['What is your current occupation'].replace(np.nan,'Unemployed')
```



```
#Imputing NaN with 'Others' beacuse those values are not provided in data.
df['Specialization'] = df['Specialization'].fillna('Others')
```

```
# Imputing the missing data in the 'Country' column with 'India'(highly skewed column)
df['Country']=df['Country'].replace(np.nan,'India')
```

- Removed columns with 40% null values
- Verified row null values greater than 70%
- Removed outliers in the numerical columns



```
per = df['TotalVisits'].quantile([0.05,0.97]).values
df['TotalVisits'][df['TotalVisits'] <= per[0]] = per[0]
df['TotalVisits'][df['TotalVisits'] >= per[1]] = per[1]
```

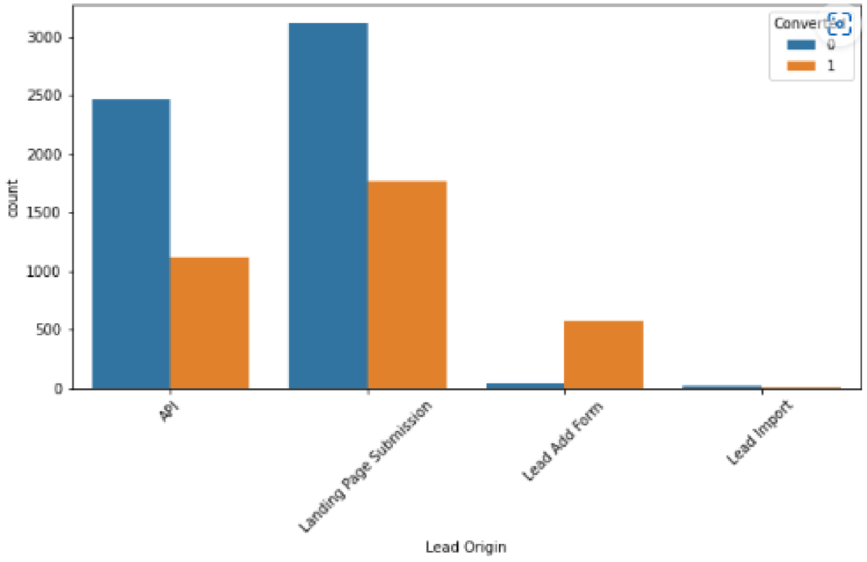
```
#Outliers are not present now
#capping the outliers to 97% value for analysis because there are lot of outliers presesnt in 'TotalVisits'
sns.boxplot(df['TotalVisits'])
```

```
<AxesSubplot:xlabel='TotalVisits'>
```

EDA

Categorical Variable

Lead Origin



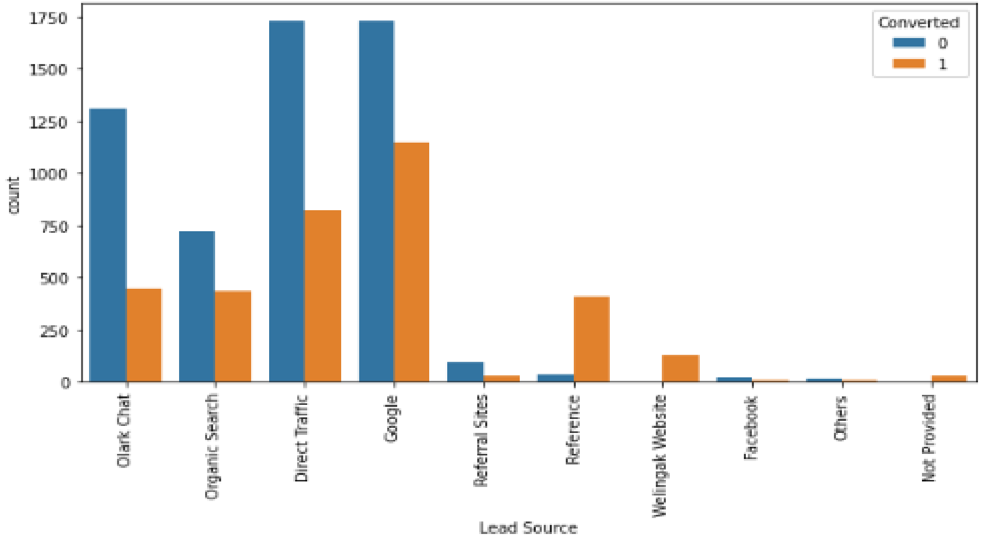
API and Landing Page Submission have 30-35% conversion rate

Generate more leads from Lead Add Form.

Lead Add Form has more than 90% conversion rate but count of lead are not very high

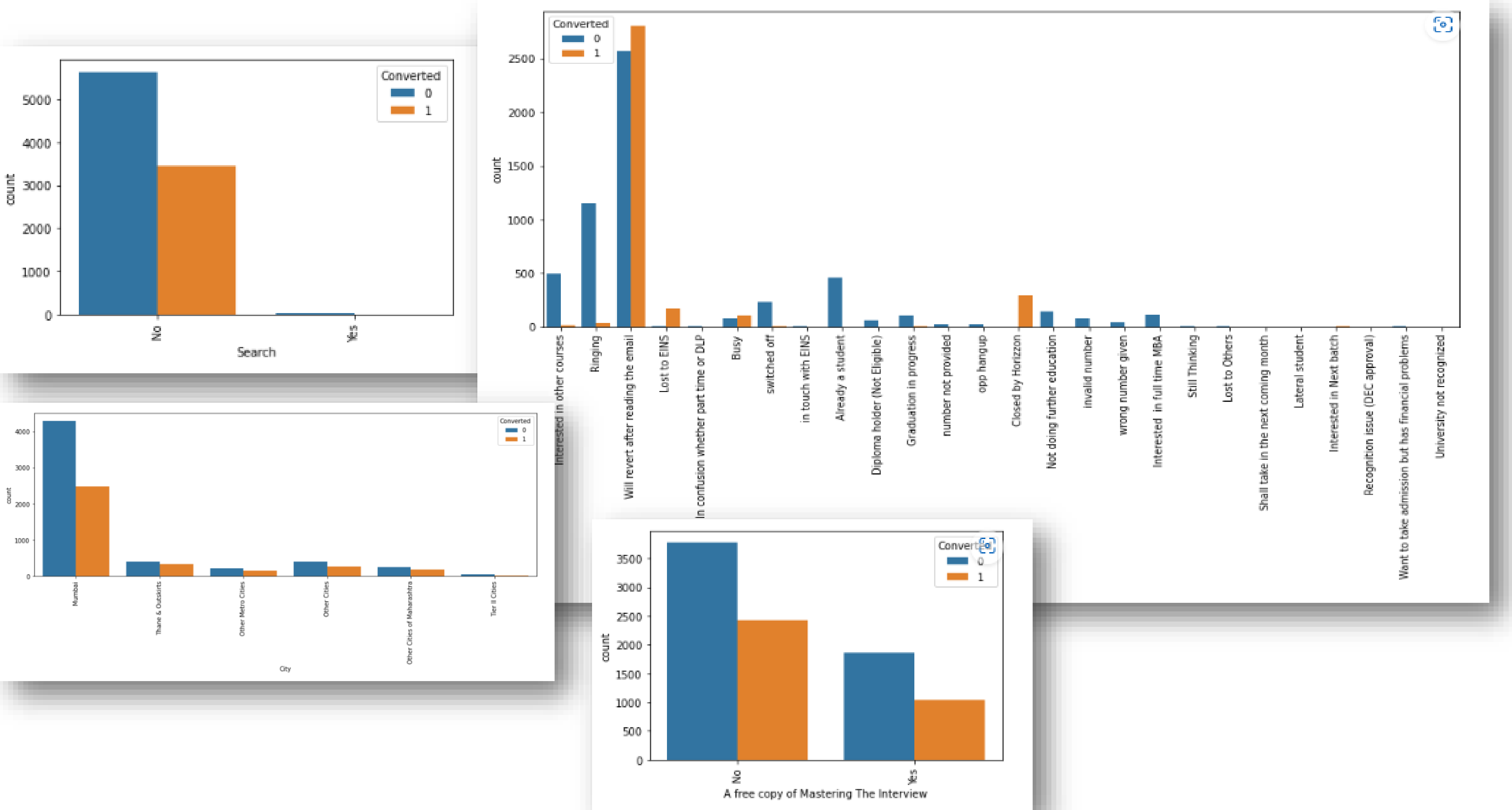
Lead Import are very less in count.

Lead Source



- Reference and Welingak Website have more than 90% conversion rate
- Organic Search and Google have more than 40% conversion rate
- Olark Chat and facebook have small rate of conversion

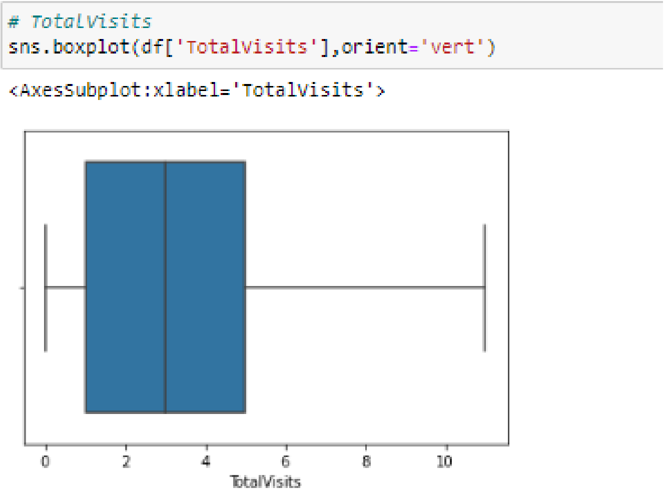
EDA


Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Search
Converted
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
number not provided
opp hangup
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Converted
0
1
count
4000
3000
2000
1000
0
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
City
Converted
0
1
count
3500
3000
2500
2000
1500
1000
500
0
No
Yes
A free copy of Mastering The Interview

EDA

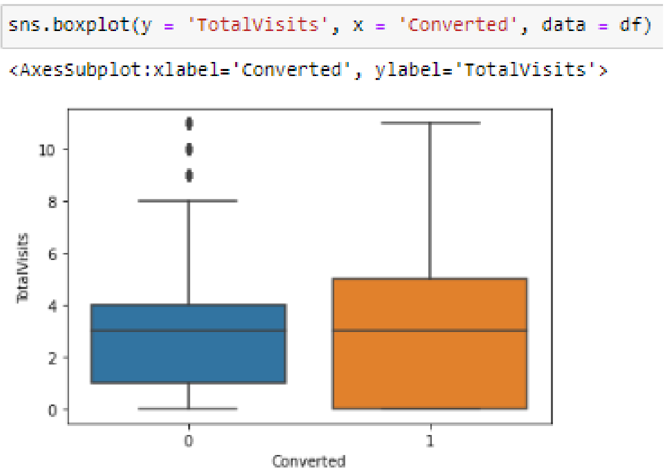
Numerical

- TotalVisits



```
# TotalVisits
sns.boxplot(df['TotalVisits'],orient='vert')
```

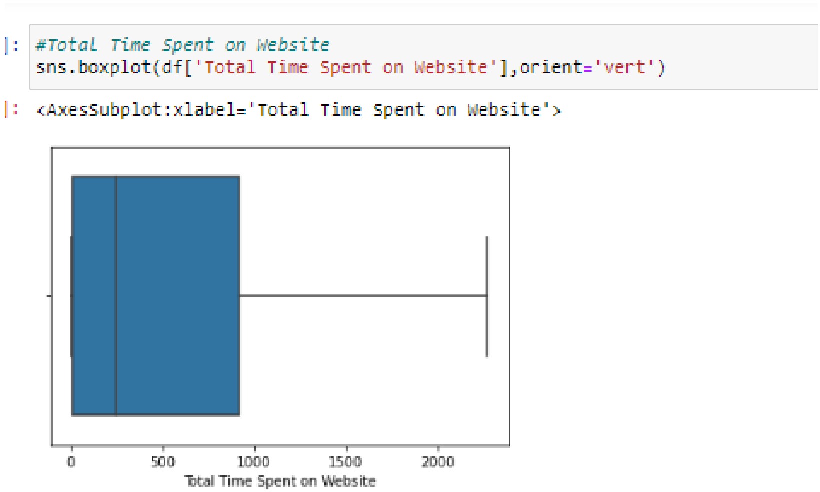
```
<AxesSubplot:xlabel='TotalVisits'>
```



```
sns.boxplot(y = 'TotalVisits', x = 'Converted', data = df)
```

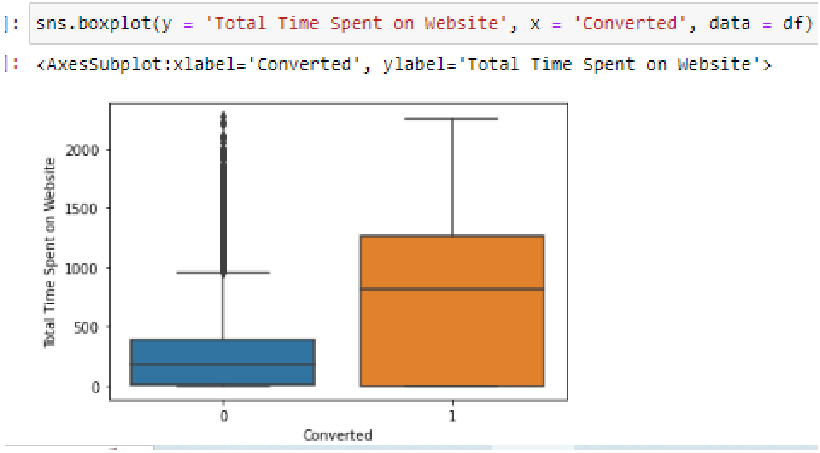
```
<AxesSubplot:xlabel='Converted', ylabel='TotalVisits'>
```

- Total Time Spent on Website



```
]: #Total Time Spent on Website
   sns.boxplot(df['Total Time Spent on Website'],orient='vert')
```

```
]: <AxesSubplot:xlabel='Total Time Spent on Website'>
```



```
]: sns.boxplot(y = 'Total Time Spent on Website', x = 'Converted', data = df)
```

```
]: <AxesSubplot:xlabel='Converted', ylabel='Total Time Spent on Website'>
```

Insights from EDA

- Leads spending more time on the weblise are more likely to be converted. Website should be made more engaging to make leads spend more time.
- we have seen that many columns are not adding any information to the model, hence we can drop them for further analysis
- Dropped columns :Lead Number ,Tags, Country , Search , Magazine , 'Newspaper Article, X Education Forums, Newspaper , Digital Advertisement , Through Recommendations, Receive More Updates About Our Courses, 'Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque, free copy of Mastering The Interview

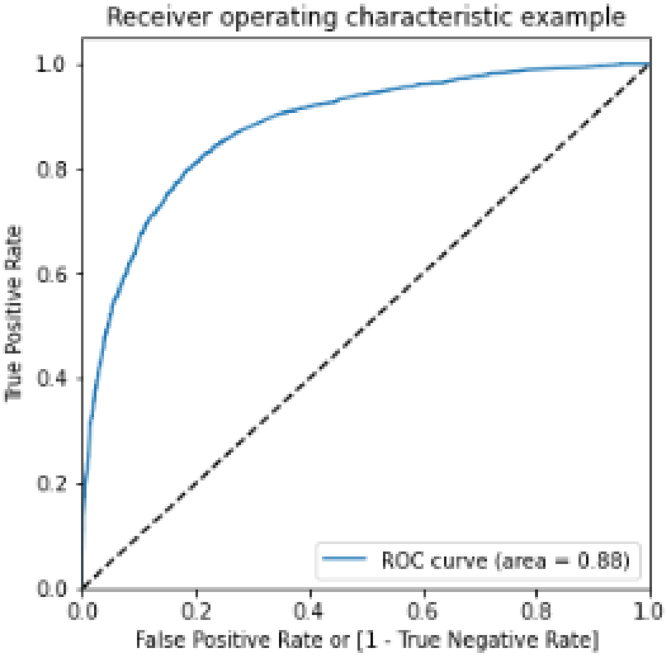
Data Preparation

- Created Dummy Variable for Categorical columns: Lead Origin, Lead Source, Last Activity, and Specialization, What is your current occupation, City, and Last Notable Activity.
- Number of Rows : 9103
- Number of columns: 70
- Split the data with 70% -train and 30% -test.
- Scaled the numerical variable used standard Scaler

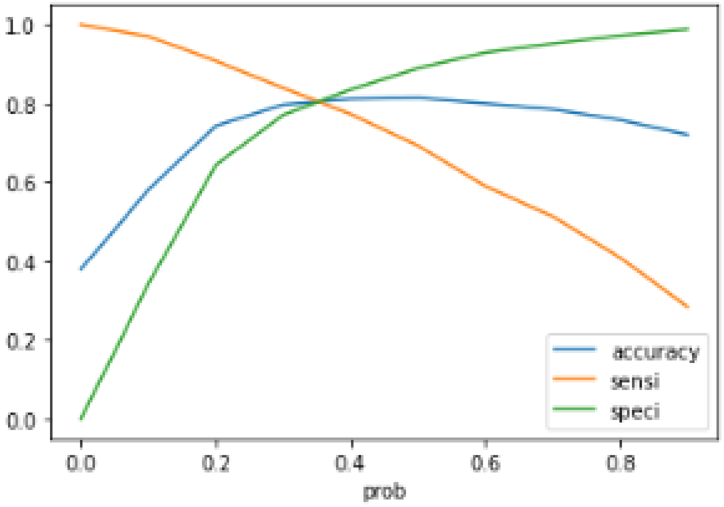
Model Building And Evaluation

- RFE is used feature selection
- Ran RFE to attain the top 20 relevant variables.
- Model retrained 9 times to achieve best model with VIF<5 and p value <0.05

Train data prediction done



- From the curve, 0.34 cut off probability.
- With 0.34 cut off Accuracy is 80% , Sensitivity is 81% and Specificity is 79%.



- A trade-off curve between precision and recall

Prediction And Conclusion

- Predicted test data with optimum cut off of 0.34 and got Accuracy: 81.2 %, Sensitivity: 82.6 %, Specificity: 80.3 %
- This help to achieve the goal of getting a ballpark of the target lead conversion rate to be around 80%
- The company should make calls to the leads coming from below categories:

  Lead Origin - Lead Add Form
  What is your current occupation- Working Professional
  Lead Source - Welingak Website
  Last Notable Activity-SMS Sent
  Last Activity- Other_Activity
  Total Time Spent on Website
  Lead Source- Olark Chat
  Last Activity -Email Opened

- The company should not make calls to the leads coming from below categories:

  Lead origin is "Landing Page Submission“
  Specialization was "Others"
  last activity was "Olark Chat Conversation"
  chose the option of "Do not Email" as "yes"