Lead Scoring – Logistic Regression

KAVITHA MAHESH

JEYA BALAJI

Business Understanding

- 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.

- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Problem Statement

- Now, 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'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:

Expected outcome

- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

- 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%.

Inspecting the data on a data frame,

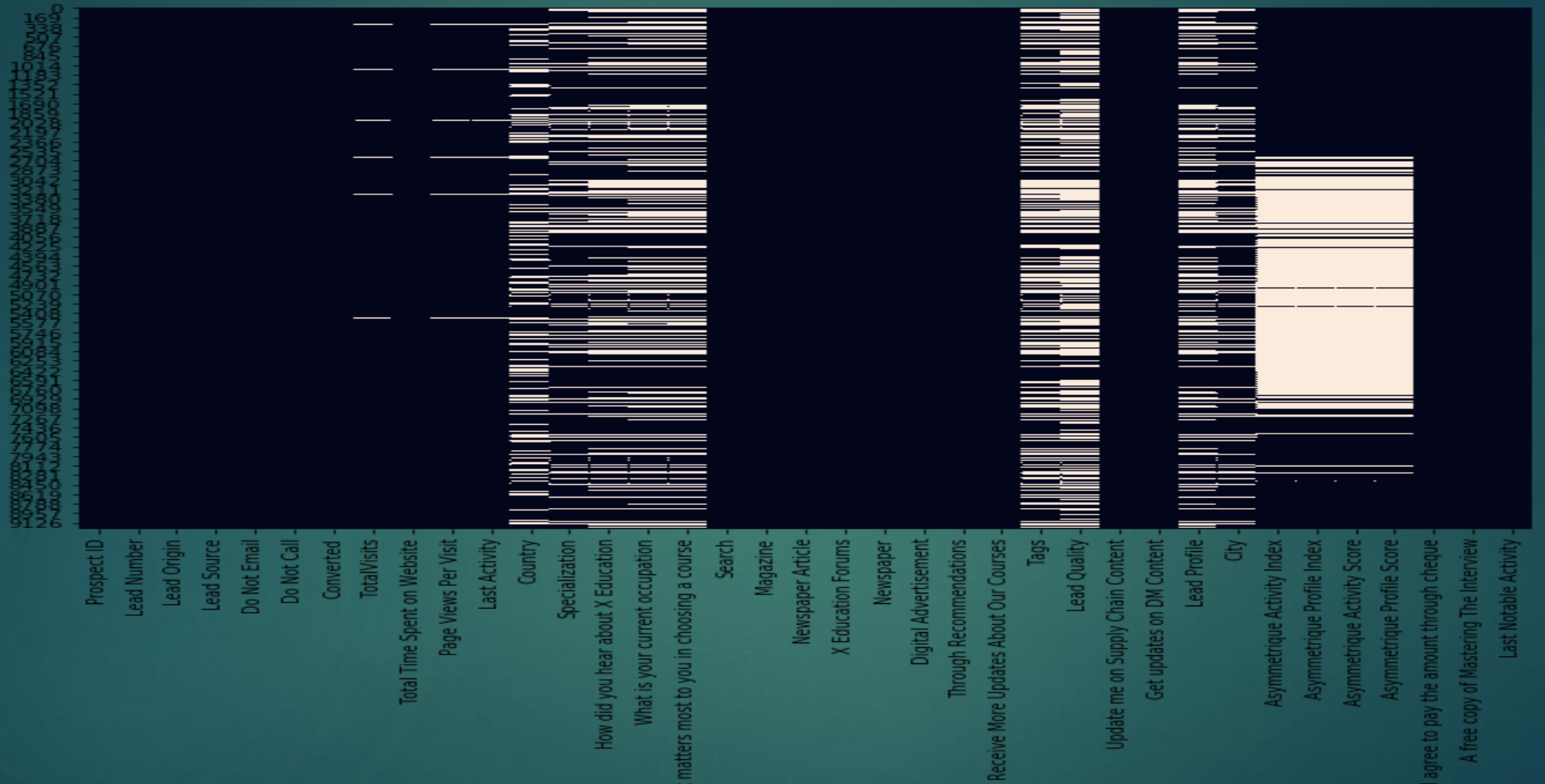
```
Prospect ID                                      object
Lead Number                                       int64
Lead Origin                                      object
Lead Source                                      object
Do Not Email                                     object
Do Not Call                                      object
Converted                                         int64
TotalVisits                                     float64
Total Time Spent on Website                       int64
Page Views Per Visit                            float64
Last Activity                                    object
Country                                          object
Specialization                                   object
How did you hear about X Education               object
What is your current occupation                  object
What matters most to you in choosing a course    object
Search                                           object
Magazine                                         object
Newspaper Article                                object
X Education Forums                               object
Newspaper                                        object
Digital Advertisement                            object
Through Recommendations                          object
Receive More Updates About Our Courses           object
Tags                                             object
Lead Quality                                     object
Update me on Supply Chain Content                object
Get updates on DM Content                        object
Lead Profile                                     object
City                                             object
Asymmetrique Activity Index                      object
Asymmetrique Profile Index                       object
Asymmetrique Activity Score                     float64
Asymmetrique Profile Score                      float64
I agree to pay the amount through cheque         object
A free copy of Mastering The Interview           object
Last Notable Activity                            object
dtype: object
```

Preparing the data for a model, looking for null values,.

	Total	Percentage
Lead Quality	4767	51.59
Asymmetrique Profile Score	4218	45.65
Asymmetrique Activity Score	4218	45.65
Asymmetrique Profile Index	4218	45.65
Asymmetrique Activity Index	4218	45.65
Tags	3353	36.29
What matters most to you in choosing a course	2709	29.32
Lead Profile	2709	29.32
What is your current occupation	2690	29.11
Country	2461	26.63
How did you hear about X Education	2207	23.89
Specialization	1438	15.56
City	1420	15.37
TotalVisits	137	1.48
Page Views Per Visit	137	1.48
Last Activity	103	1.11
Lead Source	36	0.39

Do Not Email	0	0.00
Do Not Call	0	0.00
Converted	0	0.00
Total Time Spent on Website	0	0.00
Lead Origin	0	0.00
Lead Number	0	0.00
Last Notable Activity	0	0.00
Newspaper Article	0	0.00
Search	0	0.00
Magazine	0	0.00
A free copy of Mastering The Interview	0	0.00
X Education Forums	0	0.00
Newspaper	0	0.00
Digital Advertisement	0	0.00
Through Recommendations	0	0.00
Receive More Updates About Our Courses	0	0.00
Update me on Supply Chain Content	0	0.00
Get updates on DM Content	0	0.00
I agree to pay the amount through cheque	0	0.00
Prospect ID	0	0.00

Visualizing occurrence of Null values in the columns based on rows



Dropping Unnecessary Columns NOT needed for Analysis

```
# Identifying if any column exists with only null values
leads.isnull().all(axis=0).any()
```

```
False
```

```
# Dropping all columns with only 0 values
leads.loc[:, (leads != 0).any(axis=0)]
leads.shape
```

```
(9240, 37)
```

```
#Remove columns which has only one unique value

"""
Deleting the following columns as they have only one unique value and hence cannot be responsible in predi
cting a successful lead case

Magazine
Receive More Updates About Our Courses
Update me on Supply Chain Content
Update me on Supply Chain Content
I agree to pay the amount through cheque

"""
leads= leads.loc[:,leads.nunique()!=1]
leads.shape
```

```
(9240, 32)
```

```
# Deleting the columns 'Asymmetrique Activity Score' & 'Asymmetrique Profile Score'
# as they will be represented by their corresponding index columns
leads = leads.drop('Asymmetrique Activity Score', axis=1)
leads = leads.drop('Asymmetrique Profile Score', axis=1)
leads.shape
```

```
(9240, 30)
```

```
# Deleting the columns 'Prospect ID' as it will not have any effect in the predicting model
leads = leads.drop('Prospect ID', axis=1)
#leads = leads.drop('Lead Number', axis=1)
leads.shape
```

```
(9240, 29)
```

```
# Deleting the columns 'What matters most to you in choosing a course' as it mostly has unique values and
 some null values.
leads = leads.drop('What matters most to you in choosing a course', axis=1)
leads.shape
```

```
(9240, 28)
```

```
# Deleting the columns 'How did you hear about X Education' as it mostly has null values or 'Select' value
s
# that contribute to the 'Converted' percentage.
leads = leads.drop('How did you hear about X Education', axis=1)
leads.shape
```

```
(9240, 27)
```

Removing rows where a particular column has high missing values

```
leads['Lead Source'].isnull().sum()
```

```
36
```

```
# removing rows where a particular column has high missing values because the column cannot be removed bec
ause of its importance
leads = leads[~pd.isnull(leads['Lead Source'])]
leads.shape
```

```
(9204, 27)
```

Imputing with Median values because the continuous variables have outliers

```
leads['TotalVisits'].replace(np.NaN, leads['TotalVisits'].median(), inplace =True)
```

```
leads['Page Views Per Visit'].replace(np.NaN, leads['Page Views Per Visit'].median(), inplace =True)
```

Imputing with Mode values

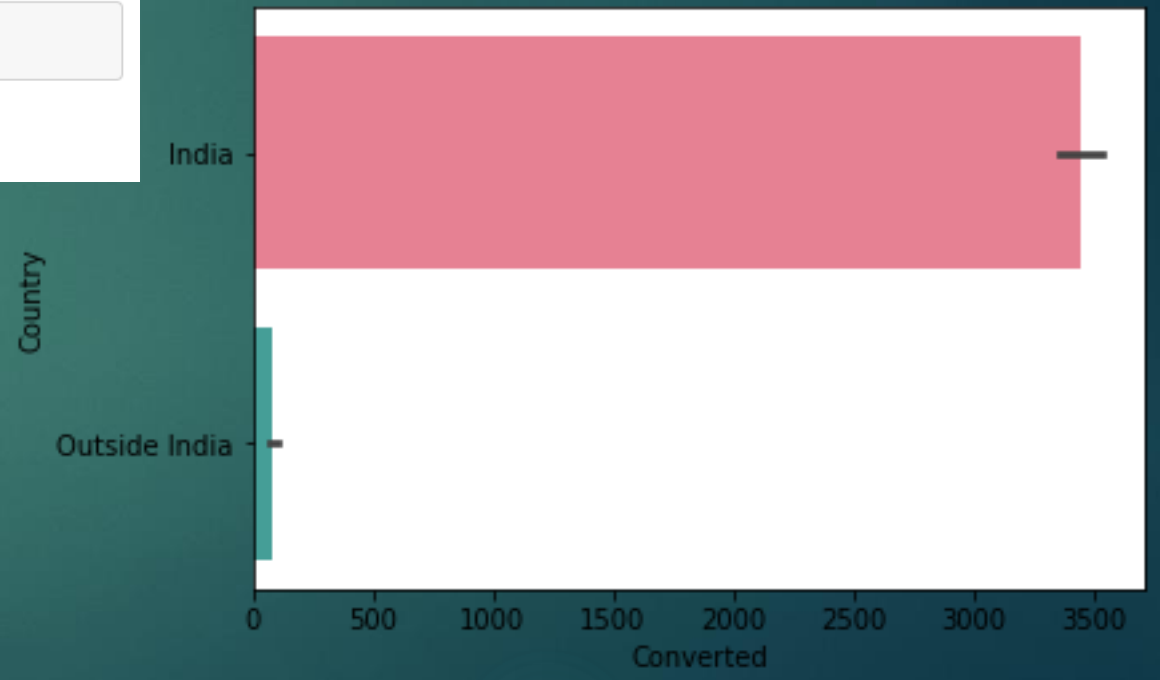
```
leads['Country'].mode()
```

```
0    India
dtype: object
```

```
leads.loc[pd.isnull(leads['Country']), ['Country']] = 'India'
```

```
leads['Country'] = leads['Country'].apply(lambda x: 'India' if x=='India' else 'Outside India')
leads['Country'].value_counts()
```

```
India            8917
Outside India     287
Name: Country, dtype: int64
```



Assigning An Unique Category to NULL/SELECT values

- Instead of deleting columns with huge null value percentage(which results in loss of data), this strategy adds more information into the dataset and results in the change of variance.
- Creating a new category consisting on NULL/Select values for the field Lead Quality
- There are some columns in dataset which have a level/value called 'Select'. This might have happened because these fields in the website might be non mandatory fields with drop downs options for the customer to choose from. Amongst the dropdown values, the default option is probably 'Select' and since these aren't mandatory fields, many customer might have chosen to leave it as the default value 'Select'

```
leads['Lead Quality'].value_counts()
```

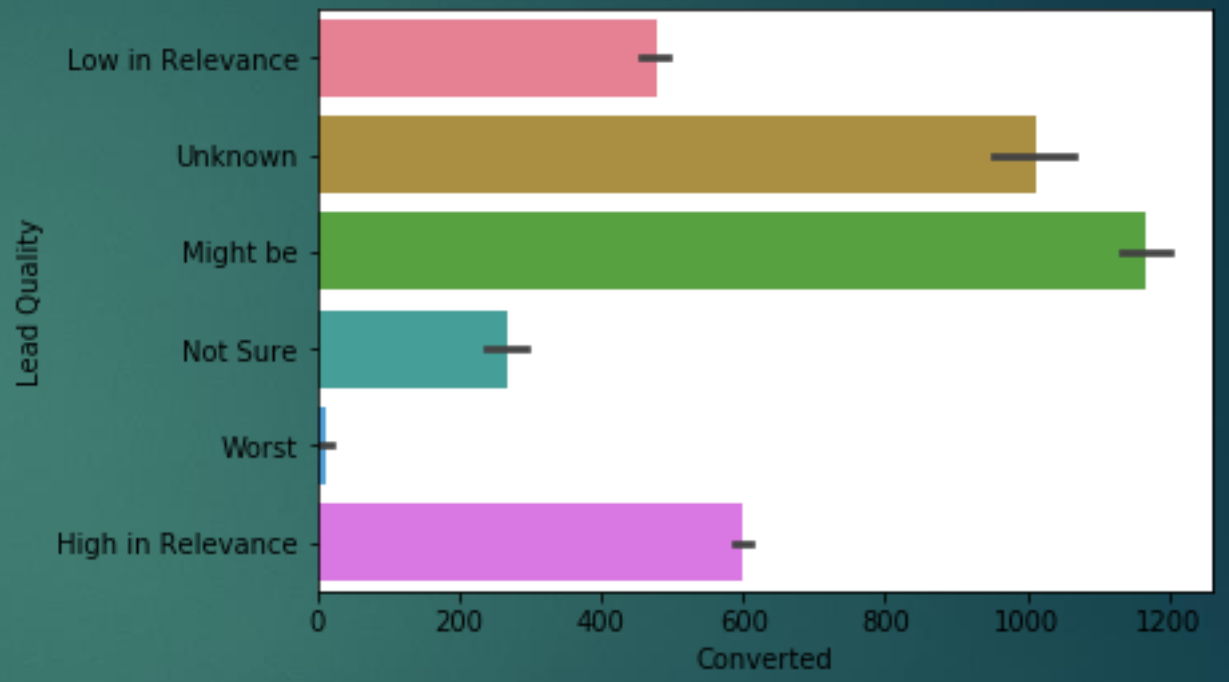
```
Might be              1545
Not Sure              1090
High in Relevance      632
Worst                  601
Low in Relevance       583
Name: Lead Quality, dtype: int64
```

```
leads['Lead Quality'].isnull().sum()
```

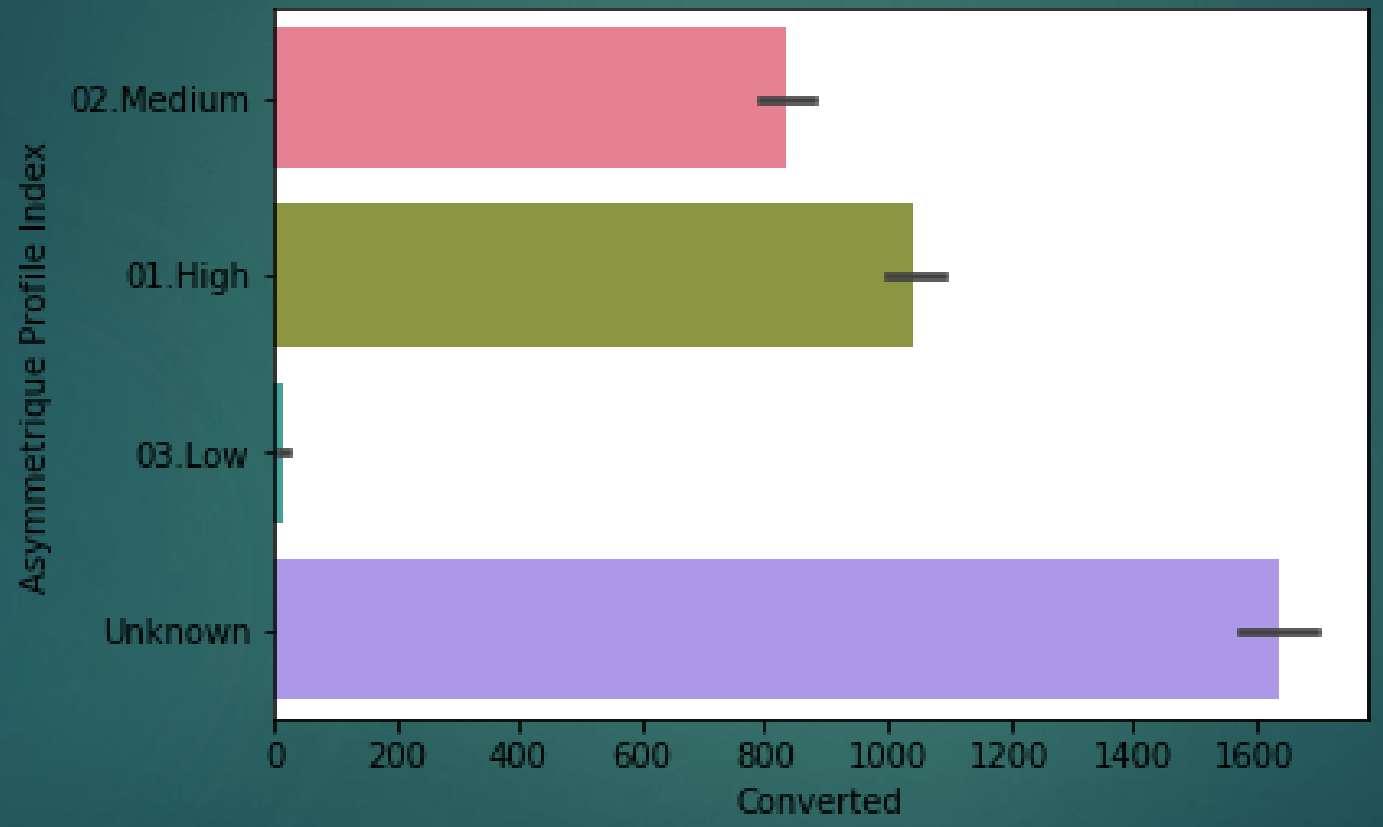
```
4753
```

```
leads['Lead Quality'].fillna("Unknown", inplace = True)
leads['Lead Quality'].value_counts()
```

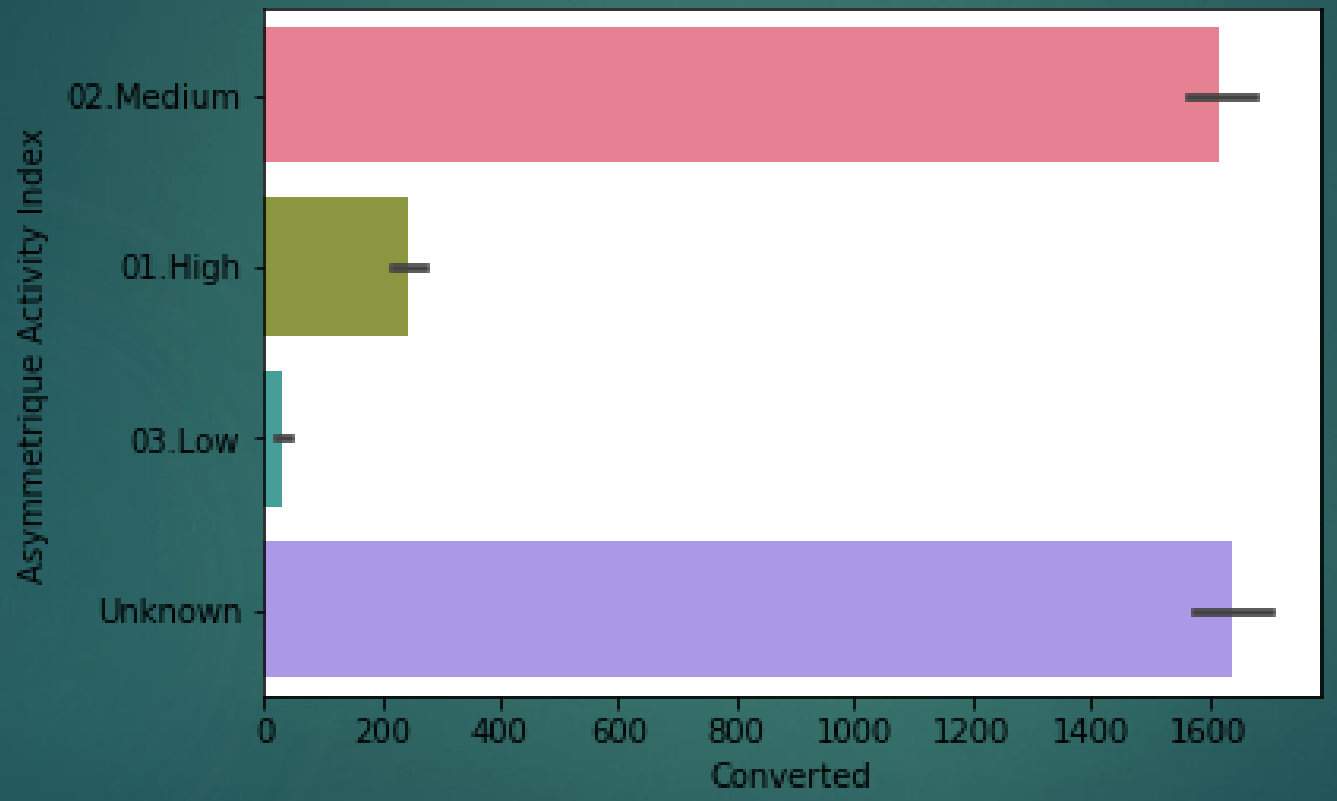
```
Unknown               4753
Might be              1545
Not Sure              1090
High in Relevance      632
Worst                  601
Low in Relevance       583
Name: Lead Quality, dtype: int64
```



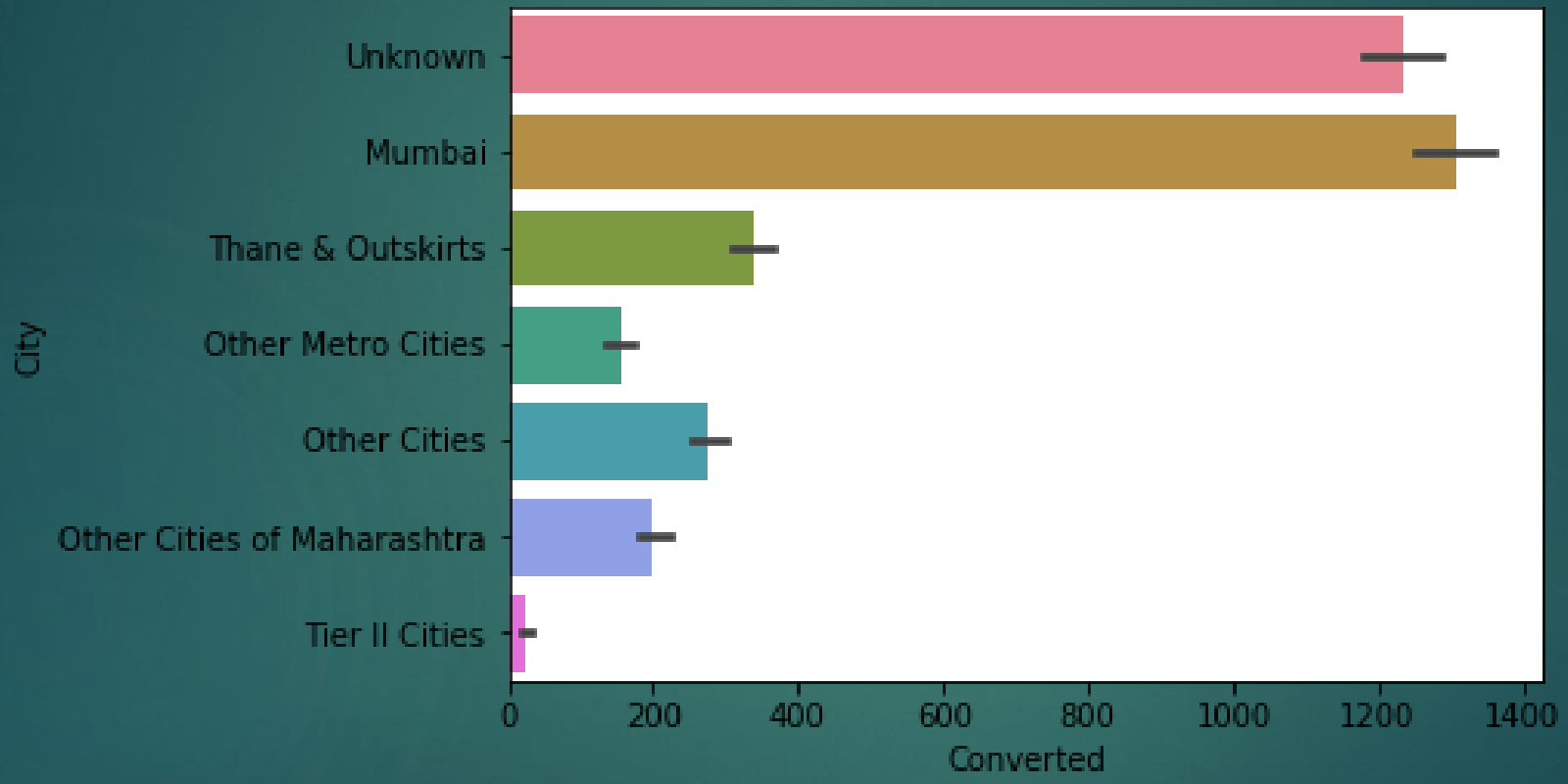
Creating a new category consisting on NULL/Select values for the field Asymmetrique Profile Index



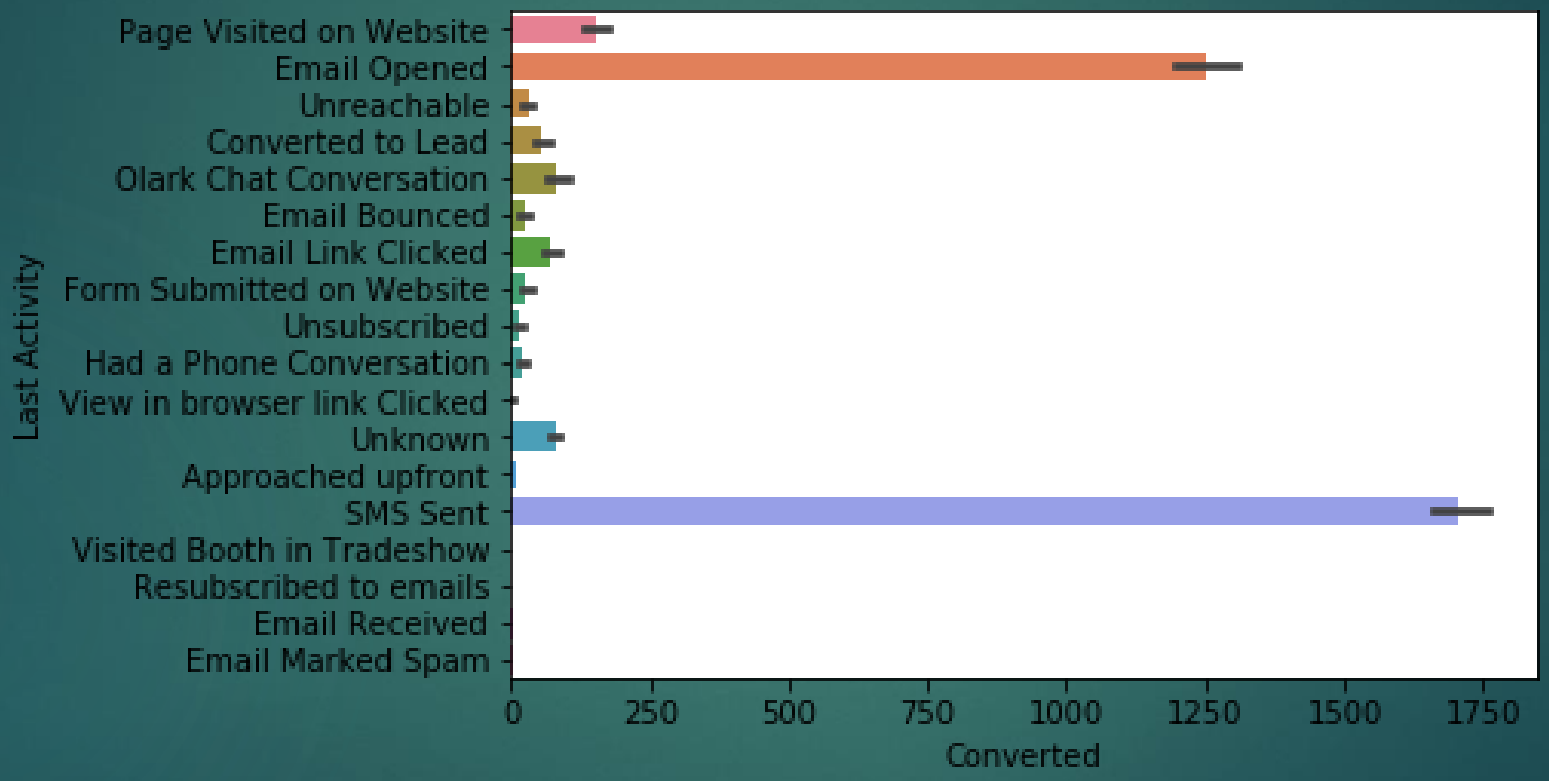
Creating a new category consisting on NULL/Select values for the field Asymmetrique Activity Index



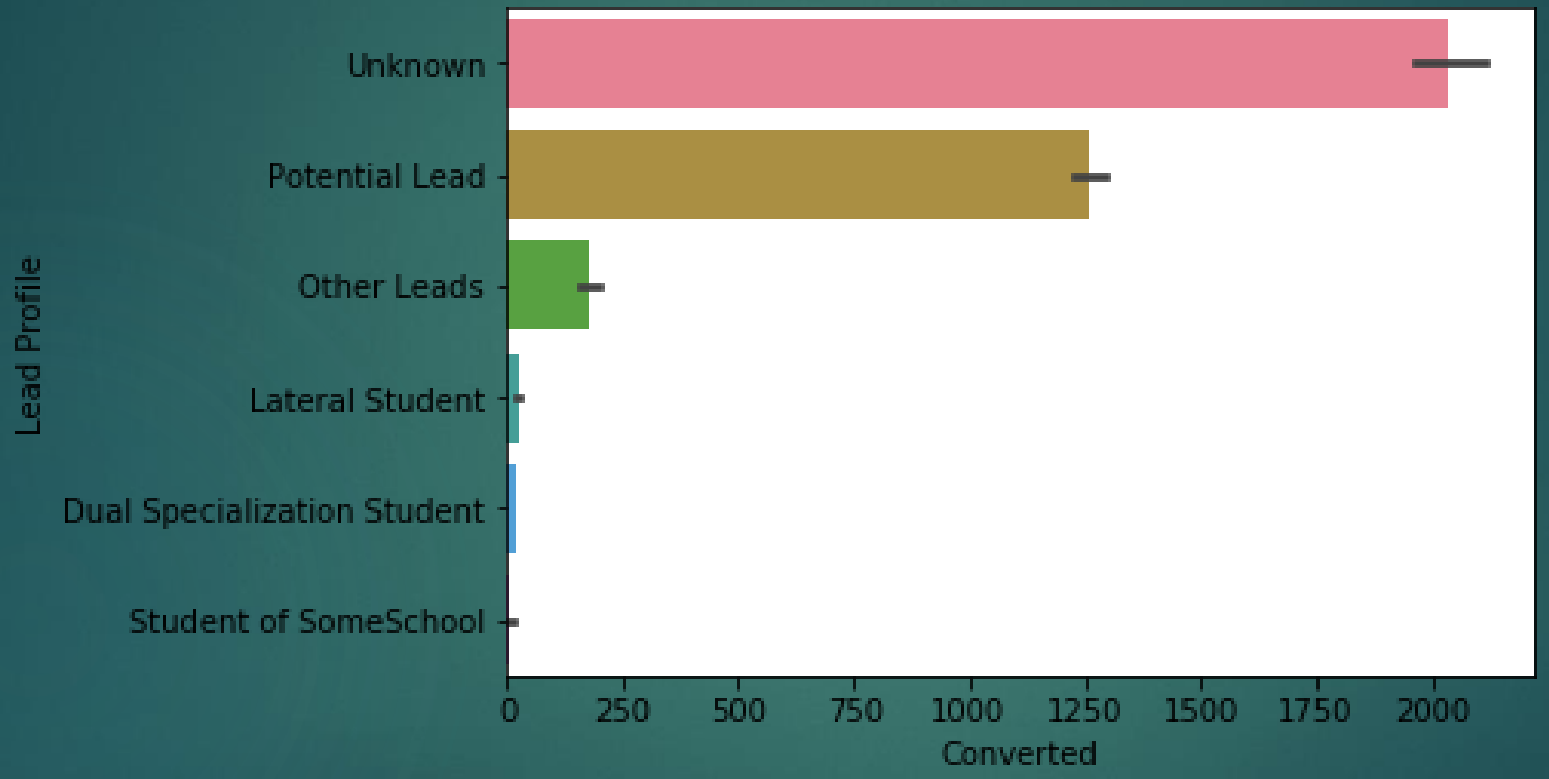
Creating a new category consisting on NULL/Select values for the field City



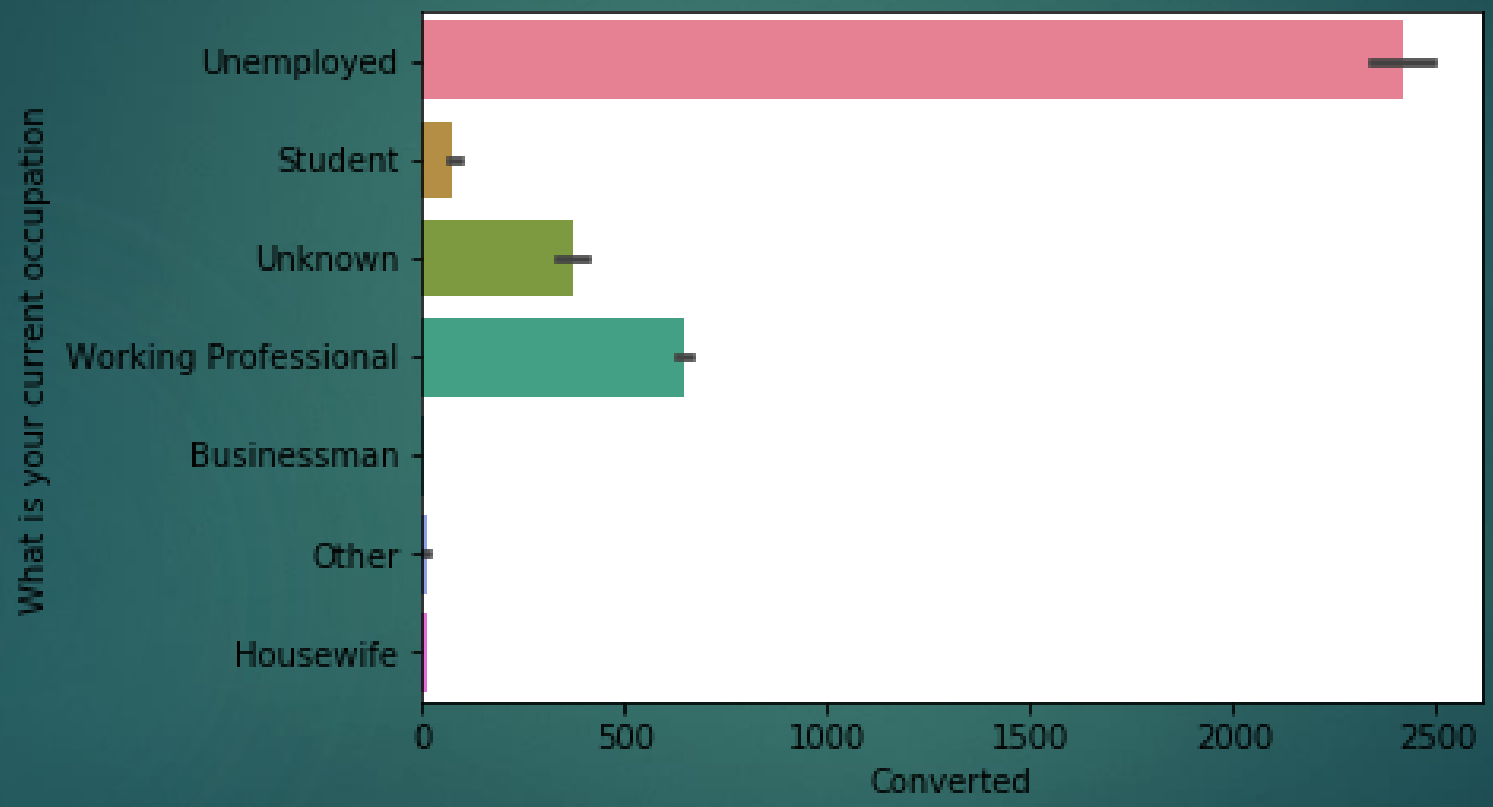
Creating a new category consisting on NULL/Select values for the field Last Activity



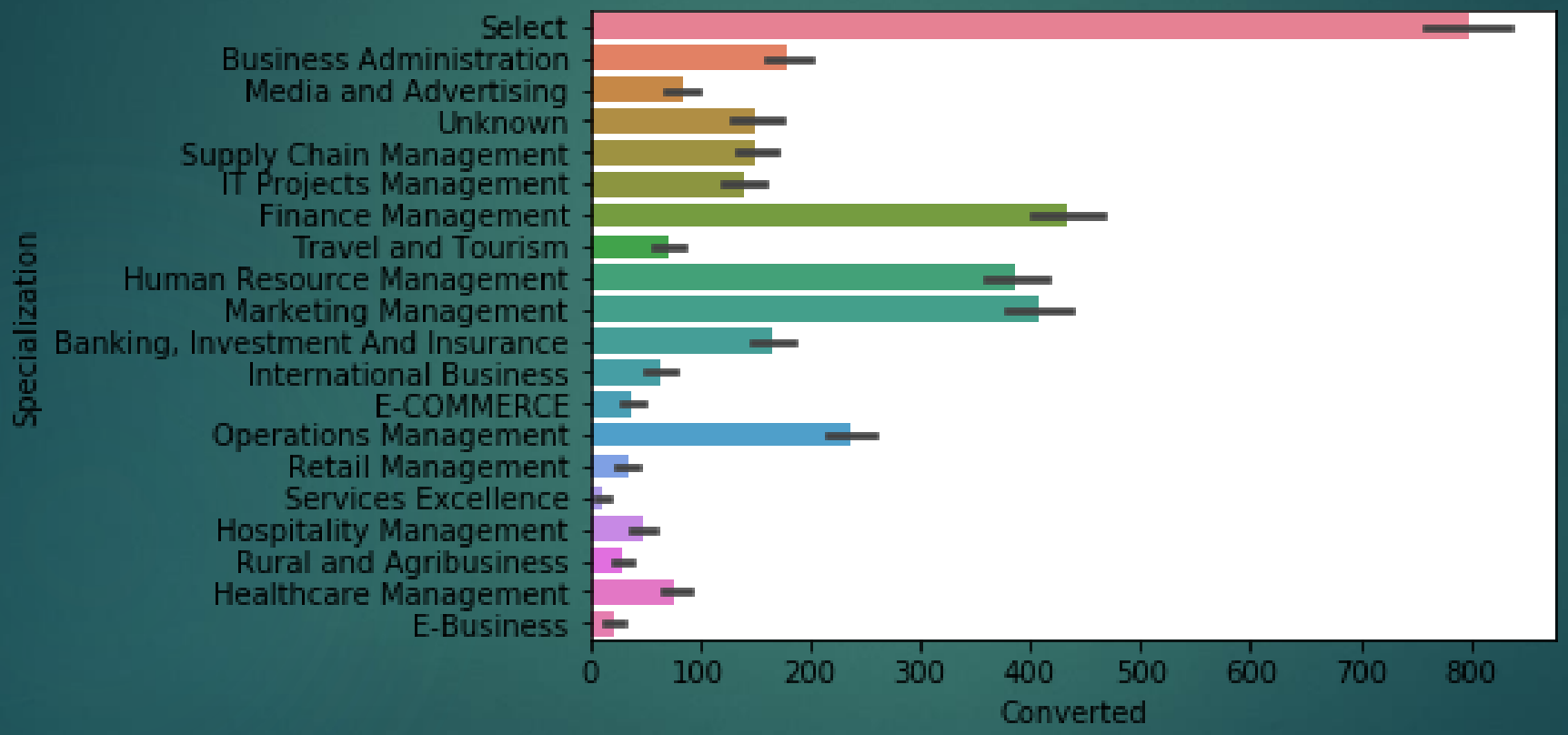
Creating a new category consisting on NULL/Select values for the field Lead Profile



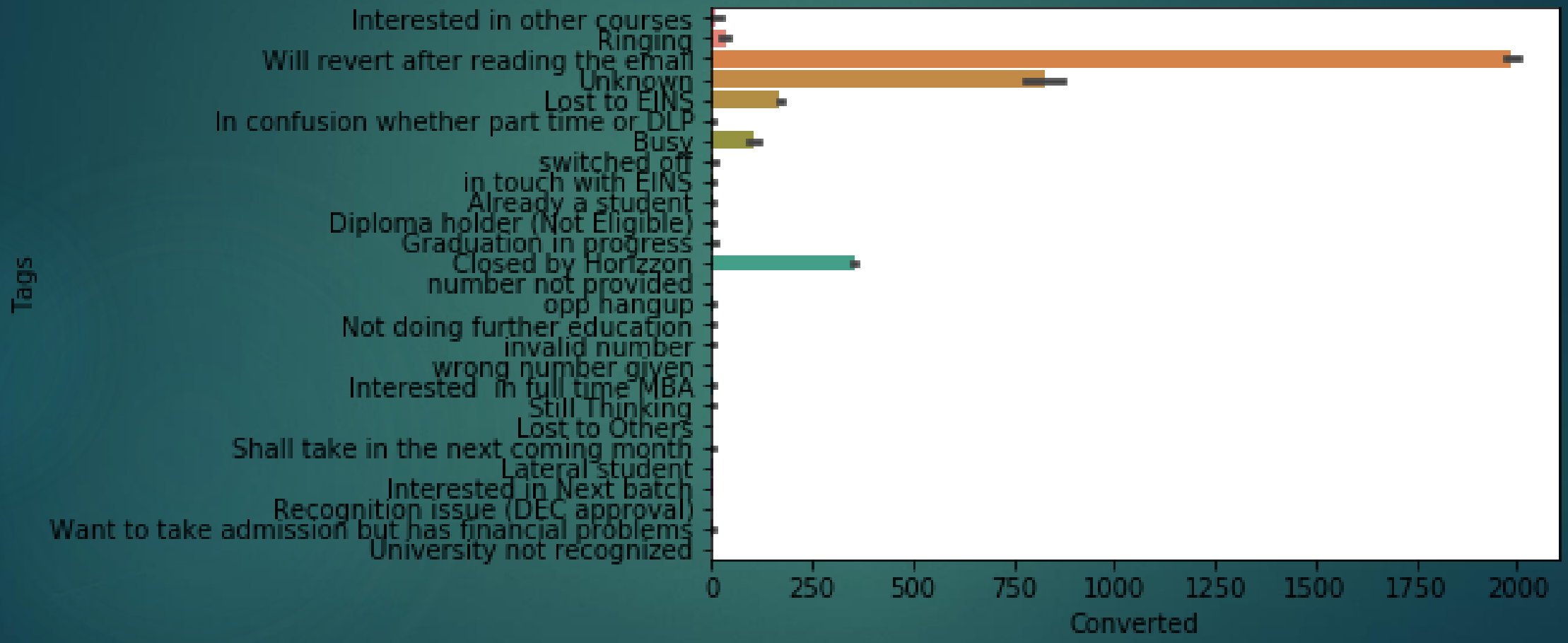
Creating a new category consisting on NULL/Select values for the field “What is your current occupation”



Creating a new category consisting on NULL/Select values for the field Specialization

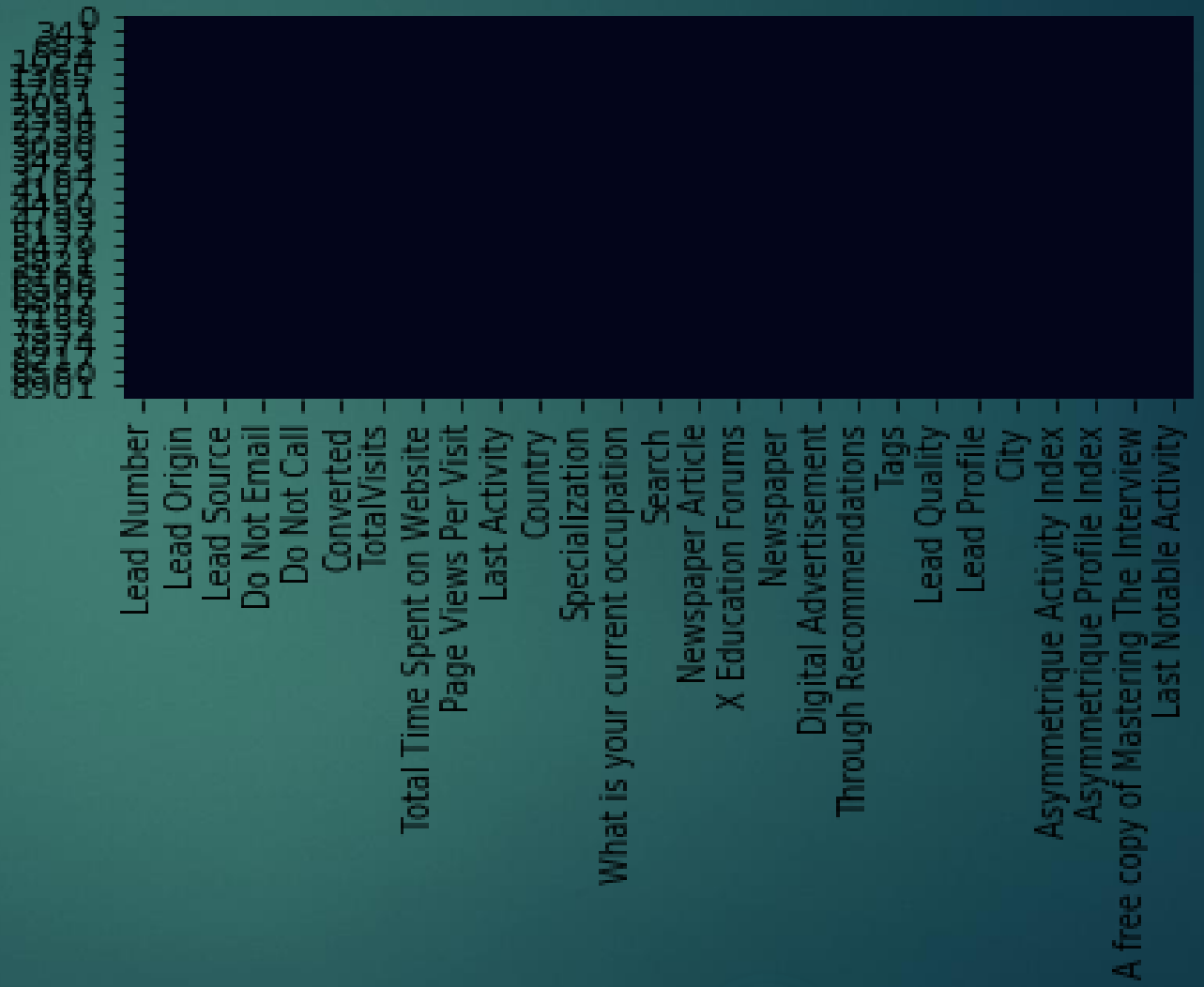


Creating a new category consisting on NULL/Select values for the field Tags

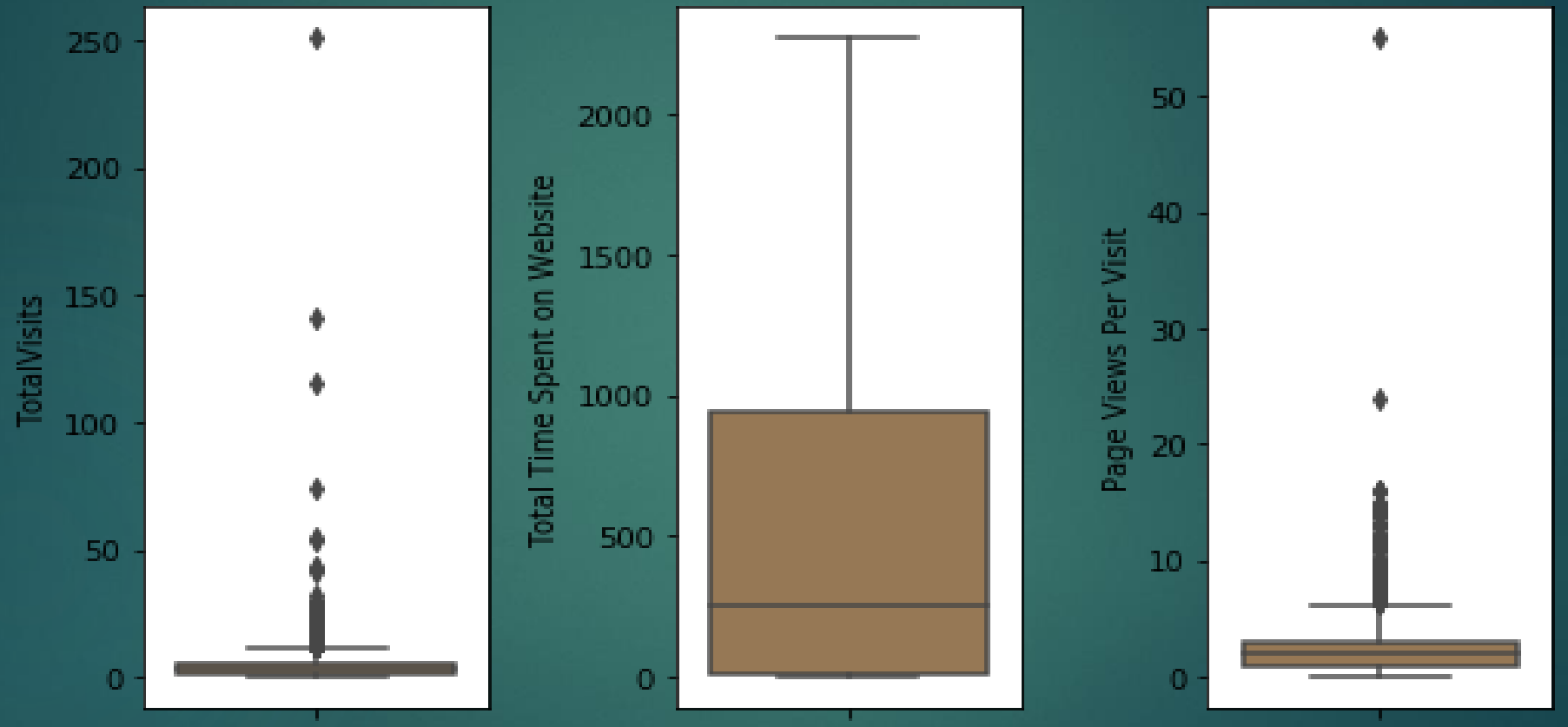


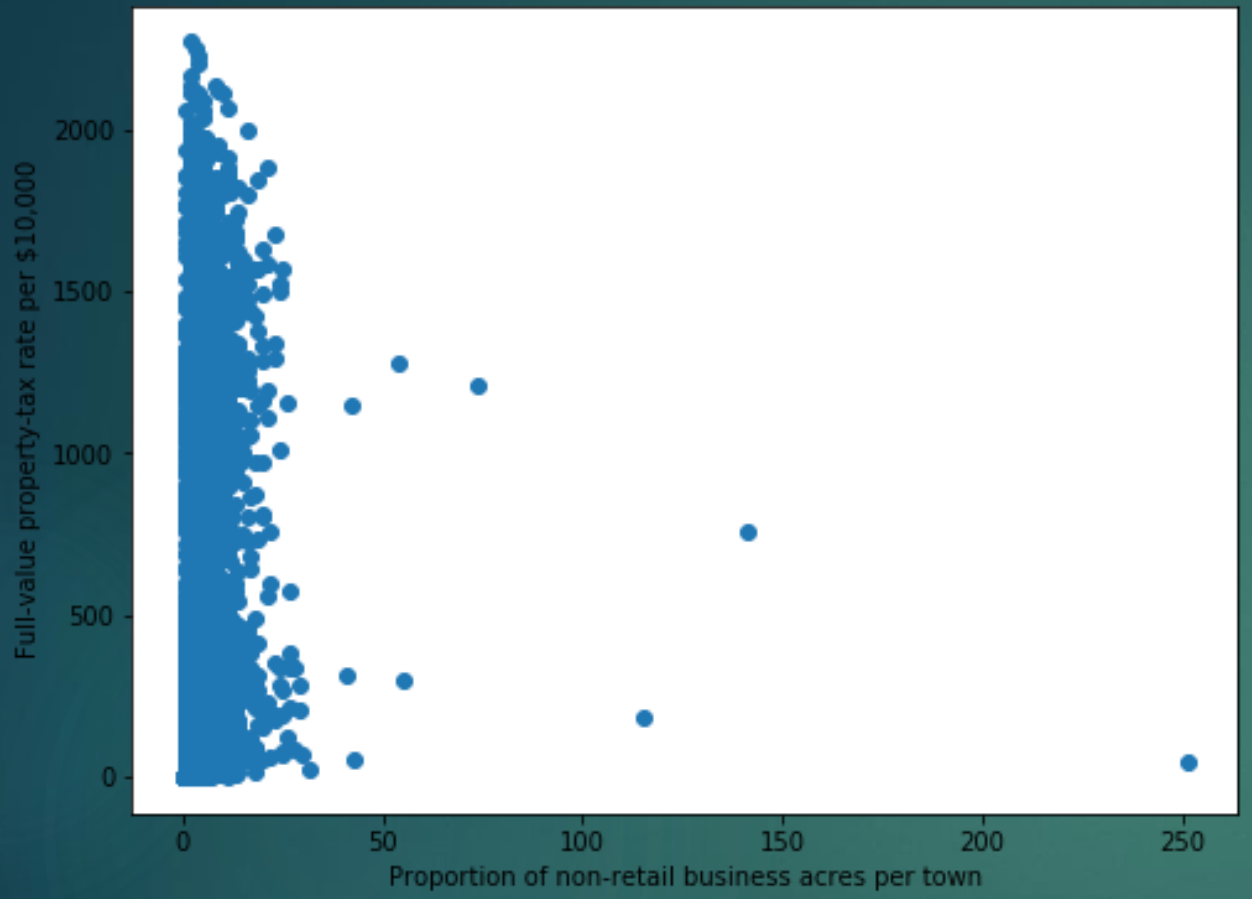
Re-inspecting null values..

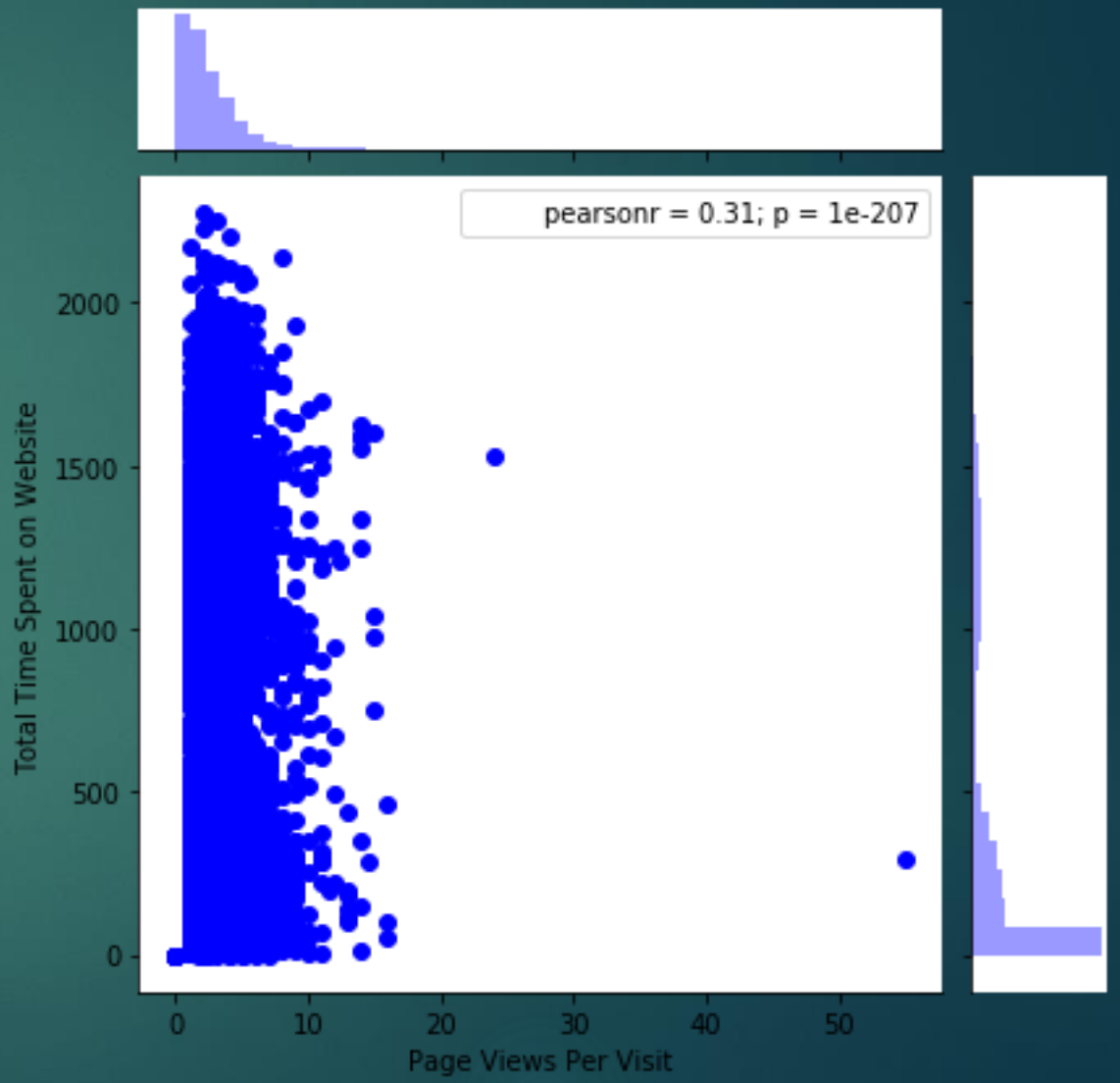
	Total	Percentage
Last Notable Activity	0	0.0
What is your current occupation	0	0.0
Lead Origin	0	0.0
Lead Source	0	0.0
Do Not Email	0	0.0



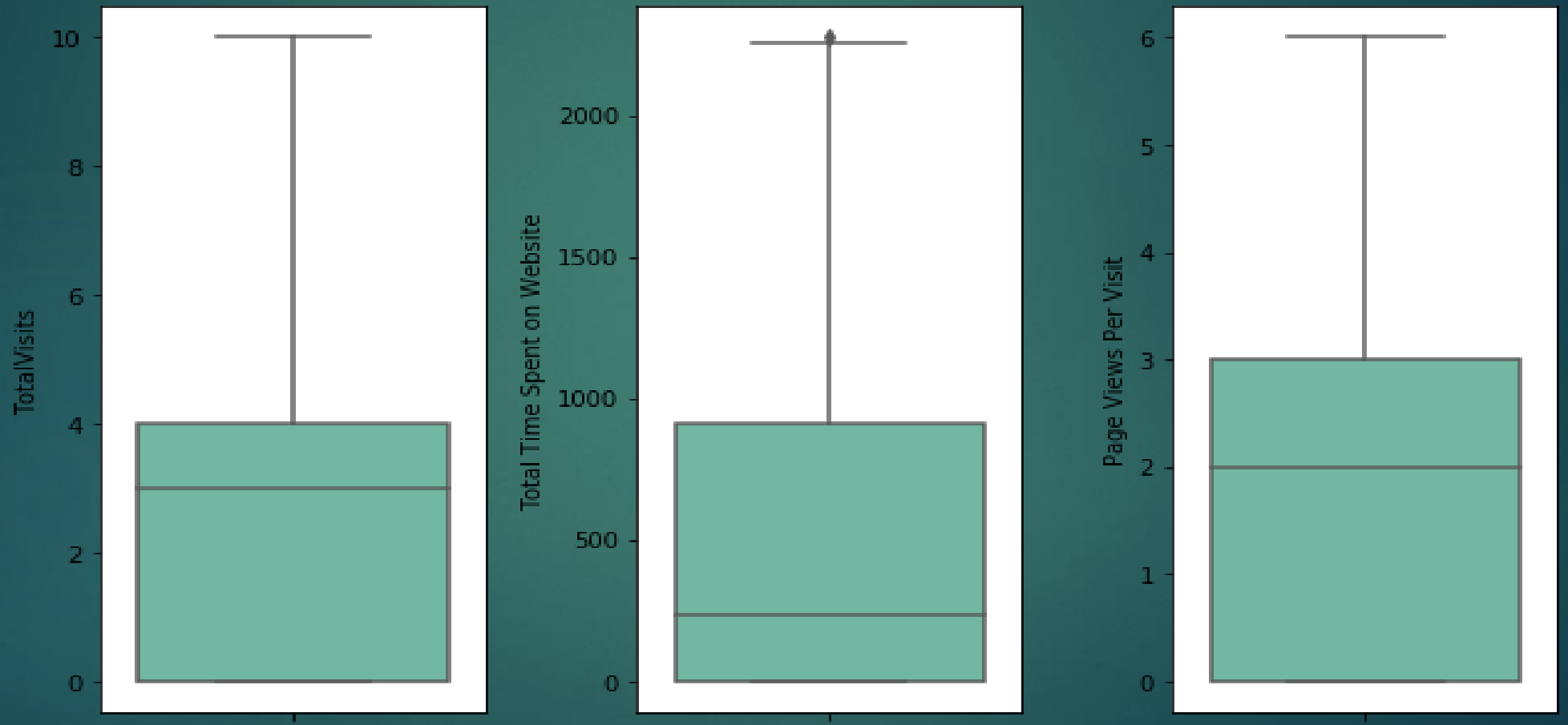
Checking for outliers..




Full-value property-tax rate per $10,000
2000
1500
1000
500
0
0
50
100
150
200
250
Proportion of non-retail business acres per town


pearsonr = 0.31; p = 1e-207
Total Time Spent on Website
2000
1500
1000
500
0
0
10
20
30
40
50
Page Views Per Visit

Removing outlier values based on the Interquartile distance for few continuous variables..



Manipulating rest of the data to fine tune it for model

Converting some binary variables (Yes/No) to 0/1

```
# List of variables to map

varlist =  ['Search','Do Not Email', 'Do Not Call', 'Newspaper Article', 'X Education Forums', 'Newspaper'
,
            'Digital Advertisement','Through Recommendations','A free copy of Mastering The Interview']

# Defining the map function
def binary_map(x):
    return x.map({'Yes': 1, "No": 0})

# Applying the function to the housing list
leads[varlist] = leads[varlist].apply(binary_map)
leads.head()
```

For categorical variables with multiple levels, creating dummy features (one-hot encoded)

```
# Creating a dummy variable for some of the categorical variables and dropping the first one.
dummy1 = pd.get_dummies(leads[['Country', 'Lead Source','Lead Origin','Last Notable Activity']], drop_firs
t=True)

# Adding the results to the master dataframe
leads = pd.concat([leads, dummy1], axis=1)
leads.shape
```

```
(8575, 66)
```

Dropping the repeated variables ¶

```
# We have created dummies for the below variables, so we can drop them
leads = leads.drop(['Lead Quality','Asymmetrique Profile Index','Asymmetrique Activity Index','Tags','Lead
Profile',
                    'Lead Origin','What is your current occupation', 'Specialization', 'City','Last Activi
ty', 'Country',
                    'Lead Source','Last Notable Activity'], 1)
leads.shape
```

```
(8575, 143)
```

Splitting train and test dataset, feature scaling and checking for lead conversion rate

Checking the Lead Conversion Rate

```
### Checking the Lead Conversion Rate
converted = (sum(leads['Converted'])/len(leads['Converted'].index))*100
converted
```

```
38.04081632653061
```

We have almost 38% lead conversion rate

Building the model..

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.865e+15	1.08e+08	-3.56e+07	0.000	-3.86e+15	-3.86e+15
Do Not Email	-2.322e+14	4.66e+06	-4.98e+07	0.000	-2.32e+14	-2.32e+14
Do Not Call	-28.8800	2.23e-06	-1.29e+07	0.000	-28.880	-28.880
TotalVisits	6.117e+13	1.51e+06	4.04e+07	0.000	6.12e+13	6.12e+13
Total Time Spent on Website	1.056e+14	1.07e+06	9.87e+07	0.000	1.06e+14	1.06e+14
Page Views Per Visit	-7.88e+13	1.64e+06	-4.81e+07	0.000	-7.88e+13	-7.88e+13
Search	-7.853e+14	2.9e+07	-2.7e+07	0.000	-7.85e+14	-7.85e+14
Newspaper Article	14.9691	1.59e-06	9.4e+06	0.000	14.969	14.969
X Education Forums	3.1361	1.64e-06	1.91e+06	0.000	3.136	3.136
Newspaper	-2.455e+15	6.76e+07	-3.63e+07	0.000	-2.45e+15	-2.45e+15
Digital Advertisement	-1.129e+14	4.85e+07	-2.33e+06	0.000	-1.13e+14	-1.13e+14
Through Recommendations	1.672e+15	5e+07	3.34e+07	0.000	1.67e+15	1.67e+15
A free copy of Mastering The Interview	3.07e+13	2.94e+06	1.05e+07	0.000	3.07e+13	3.07e+13
Country_Outside India	1.086e+14	4.99e+06	2.18e+07	0.000	1.09e+14	1.09e+14
Lead Source_Direct Traffic	1.666e+15	7.95e+07	2.09e+07	0.000	1.67e+15	1.67e+15
Lead Source_Facebook	7.147e+14	4.01e+07	1.78e+07	0.000	7.15e+14	7.15e+14
Lead Source_Google	1.688e+15	7.95e+07	2.12e+07	0.000	1.69e+15	1.69e+15
Lead Source_Live Chat	3.106e+15	6.31e+07	4.92e+07	0.000	3.11e+15	3.11e+15
Lead Source_NC_EDM	6.546e+15	1.04e+08	6.28e+07	0.000	6.55e+15	6.55e+15
Lead Source_Olark Chat	1.84e+15	7.94e+07	2.32e+07	0.000	1.84e+15	1.84e+15

Dep. Variable:	Converted	**No. Observations:**	6002
Model:	GLM	**Df Residuals:**	5871
Model Family:	Binomial	**Df Model:**	130
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Sun, 03 Mar 2019	**Deviance:**	nan
Time:	17:43:16	**Pearson chi2:**	2.09e+18
No. Iterations:	100	**Covariance Type:**	nonrobust

Lead Source_Organic Search	1.558e+15	7.96e+07	1.96e+07	0.000	1.56e+15	1.56e+15
Lead Source_Pay per Click Ads	-2.212e+14	1.04e+08	-2.12e+06	0.000	-2.21e+14	-2.21e+14
Lead Source_Press_Release	8.5926	1.38e-06	6.22e+06	0.000	8.593	8.593
Lead Source_Reference	6.49e+14	4.15e+07	1.56e+07	0.000	6.49e+14	6.49e+14
Lead Source_Referral Sites	1.488e+15	7.99e+07	1.86e+07	0.000	1.49e+15	1.49e+15
Lead Source_Social Media	-2.137e+15	1.06e+08	-2.01e+07	0.000	-2.14e+15	-2.14e+15
Lead Source_WeLearn	-11.7045	4.87e-07	-2.41e+07	0.000	-11.705	-11.705
Lead Source_Welingak Website	1.112e+15	4.2e+07	2.65e+07	0.000	1.11e+15	1.11e+15
Lead Source_bing	8.754e+14	9.28e+07	9.43e+06	0.000	8.75e+14	8.75e+14
Lead Source_blog	1.893e+15	1.04e+08	1.81e+07	0.000	1.89e+15	1.89e+15
Lead Source_google	-8.717e+14	9.31e+07	-9.37e+06	0.000	-8.72e+14	-8.72e+14
Lead Source_testone	-2.328e+15	1.04e+08	-2.23e+07	0.000	-2.33e+15	-2.33e+15
Lead Source_welearnblog_Home	2.053e+15	1.04e+08	1.97e+07	0.000	2.05e+15	2.05e+15
Lead Source_youtubechannel	-0.7858	1.58e-06	-4.98e+05	0.000	-0.786	-0.786
Lead Origin_Landing Page Submission	-2.119e+13	4.28e+06	-4.95e+06	0.000	-2.12e+13	-2.12e+13
Lead Origin_Lead Add Form	9.363e+14	6.77e+07	1.38e+07	0.000	9.36e+14	9.36e+14
Lead Origin_Lead Import	7.147e+14	4.01e+07	1.78e+07	0.000	7.15e+14	7.15e+14
Last Notable Activity_Email Bounced	2.328e+15	7.42e+07	3.14e+07	0.000	2.33e+15	2.33e+15
Last Notable Activity_Email Link Clicked	1.334e+15	7.39e+07	1.81e+07	0.000	1.33e+15	1.33e+15
Last Notable Activity_Email Marked Spam	2.971e+15	4.39e+07	6.77e+07	0.000	2.97e+15	2.97e+15
Last Notable Activity_Email Opened	1.849e+15	7.32e+07	2.53e+07	0.000	1.85e+15	1.85e+15

Last Notable Activity_Email Received	2.536e+13	1.2e+08	2.11e+05	0.000	2.54e+13	2.54e+13
Last Notable Activity_Form Submitted on Website	-2.721e+15	9.98e+07	-2.73e+07	0.000	-2.72e+15	-2.72e+15
Last Notable Activity_Had a Phone Conversation	1.6e+15	8.18e+07	1.95e+07	0.000	1.6e+15	1.6e+15
Last Notable Activity_Modified	1.592e+15	7.31e+07	2.18e+07	0.000	1.59e+15	1.59e+15
Last Notable Activity_Olark Chat Conversation	1.462e+15	7.35e+07	1.99e+07	0.000	1.46e+15	1.46e+15
Last Notable Activity_Page Visited on Website	1.625e+15	7.35e+07	2.21e+07	0.000	1.62e+15	1.62e+15
Last Notable Activity_Resubscribed to emails	29.4125	6.4e-07	4.59e+07	0.000	29.412	29.412
Last Notable Activity_SMS Sent	1.878e+15	7.32e+07	2.56e+07	0.000	1.88e+15	1.88e+15
Last Notable Activity_Unreachable	2.122e+15	7.54e+07	2.81e+07	0.000	2.12e+15	2.12e+15
Last Notable Activity_Unsubscribed	1.039e+15	7.68e+07	1.35e+07	0.000	1.04e+15	1.04e+15
Last Notable Activity_View in browser link Clicked	-1.461e+15	1.2e+08	-1.21e+07	0.000	-1.46e+15	-1.46e+15
Lead Quality_High in Relevance	-2.301e+14	5.63e+06	-4.09e+07	0.000	-2.3e+14	-2.3e+14
Lead Quality_Low in Relevance	-2.669e+14	5.45e+06	-4.89e+07	0.000	-2.67e+14	-2.67e+14
Lead Quality_Might be	-1.605e+14	4.06e+06	-3.96e+07	0.000	-1.61e+14	-1.61e+14
Lead Quality_Not Sure	-2.779e+13	3.68e+06	-7.54e+06	0.000	-2.78e+13	-2.78e+13
Lead Quality_Worst	-4.553e+14	5.57e+06	-8.17e+07	0.000	-4.55e+14	-4.55e+14
Asymmetrique Profile Index_01.High	-1.018e+14	3.86e+06	-2.64e+07	0.000	-1.02e+14	-1.02e+14
Asymmetrique Profile Index_02.Medium	-8.199e+12	3.34e+06	-2.46e+06	0.000	-8.2e+12	-8.2e+12
Asymmetrique Profile Index_03.Low	-1.459e+14	1.44e+07	-1.01e+07	0.000	-1.46e+14	-1.46e+14
Asymmetrique Activity Index_01.High	1.725e+14	4.13e+06	4.18e+07	0.000	1.73e+14	1.73e+14
Asymmetrique Activity Index_02.Medium	9.322e+13	3.34e+06	2.79e+07	0.000	9.32e+13	9.32e+13

Asymmetrique Activity Index_03.Low	-5.217e+14	5.07e+06	-1.03e+08	0.000	-5.22e+14	-5.22e+14
Tags_Already a student	-1.566e+15	6.49e+06	-2.41e+08	0.000	-1.57e+15	-1.57e+15
Tags_Busy	-8.087e+14	7.61e+06	-1.06e+08	0.000	-8.09e+14	-8.09e+14
Tags_Closed by Horizzon	6.206e+14	7.01e+06	8.85e+07	0.000	6.21e+14	6.21e+14
Tags_Diploma holder (Not Eligible)	-4.643e+15	1.11e+07	-4.19e+08	0.000	-4.64e+15	-4.64e+15
Tags_Graduation in progress	-9.442e+14	9.08e+06	-1.04e+08	0.000	-9.44e+14	-9.44e+14
Tags_In confusion whether part time or DLP	-2.595e+15	3.04e+07	-8.54e+07	0.000	-2.59e+15	-2.59e+15
Tags_Interested in full time MBA	-1.154e+15	8.87e+06	-1.3e+08	0.000	-1.15e+15	-1.15e+15
Tags_Interested in Next batch	3.661e+15	3.92e+07	9.33e+07	0.000	3.66e+15	3.66e+15
Tags_Interested in other courses	-1.068e+15	5.13e+06	-2.08e+08	0.000	-1.07e+15	-1.07e+15
Tags_Lateral student	4.176e+15	4.79e+07	8.72e+07	0.000	4.18e+15	4.18e+15
Tags_Lost to EINS	1.139e+15	7.42e+06	1.53e+08	0.000	1.14e+15	1.14e+15
Tags_Lost to Others	-3.588e+15	3.08e+07	-1.17e+08	0.000	-3.59e+15	-3.59e+15
Tags_Not doing further education	-1.25e+15	8.38e+06	-1.49e+08	0.000	-1.25e+15	-1.25e+15
Tags_Recognition issue (DEC approval)	-4.325e+15	6.89e+07	-6.27e+07	0.000	-4.32e+15	-4.32e+15
Tags_Ringing	-1.729e+15	4.4e+06	-3.93e+08	0.000	-1.73e+15	-1.73e+15
Tags_Shall take in the next coming month	4.608e+15	6.78e+07	6.8e+07	0.000	4.61e+15	4.61e+15
Tags_Still Thinking	-1.623e+15	3.42e+07	-4.75e+07	0.000	-1.62e+15	-1.62e+15
Tags_University not recognized	-2.409e+15	4.79e+07	-5.03e+07	0.000	-2.41e+15	-2.41e+15
Tags_Want to take admission but has financial problems	-3.546e+14	4.15e+07	-8.55e+06	0.000	-3.55e+14	-3.55e+14
Tags_Will revert after reading the email	5.366e+14	5.07e+06	1.06e+08	0.000	5.37e+14	5.37e+14
Tags_in touch with EINS	-1.108e+15	2.42e+07	-4.58e+07	0.000	-1.11e+15	-1.11e+15

Tags_invalid number	-5.177e+15	9.98e+06	-5.19e+08	0.000	-5.18e+15	-5.18e+15
Tags_number not provided	-2.815e+15	1.66e+07	-1.7e+08	0.000	-2.81e+15	-2.81e+15
Tags_opp hangup	-1.613e+15	1.62e+07	-9.98e+07	0.000	-1.61e+15	-1.61e+15
Tags_switched off	-1.871e+15	6.61e+06	-2.83e+08	0.000	-1.87e+15	-1.87e+15
Tags_wrong number given	-4.914e+15	1.27e+07	-3.87e+08	0.000	-4.91e+15	-4.91e+15
Lead Profile_Dual Specialization Student	2.939e+15	2.16e+07	1.36e+08	0.000	2.94e+15	2.94e+15
Lead Profile_Lateral Student	1.037e+15	1.79e+07	5.8e+07	0.000	1.04e+15	1.04e+15
Lead Profile_Other Leads	2.281e+14	4.7e+06	4.85e+07	0.000	2.28e+14	2.28e+14
Lead Profile_Potential Lead	1.225e+14	3.28e+06	3.74e+07	0.000	1.23e+14	1.23e+14
Lead Profile_Student of SomeSchool	-4.562e+13	8.03e+06	-5.68e+06	0.000	-4.56e+13	-4.56e+13
What is your current occupation_Businessman	-3.365e+14	4.82e+07	-6.98e+06	0.000	-3.36e+14	-3.36e+14
What is your current occupation_Housewife	3.036e+15	2.45e+07	1.24e+08	0.000	3.04e+15	3.04e+15
What is your current occupation_Other	9.483e+14	1.95e+07	4.87e+07	0.000	9.48e+14	9.48e+14
What is your current occupation_Student	1.032e+15	7.46e+06	1.38e+08	0.000	1.03e+15	1.03e+15
What is your current occupation_Unemployed	1.038e+15	4.32e+06	2.4e+08	0.000	1.04e+15	1.04e+15
What is your current occupation_Working Professional	1.118e+15	5.71e+06	1.96e+08	0.000	1.12e+15	1.12e+15
Specialization_Banking, Investment And Insurance	2.771e+13	6.78e+06	4.09e+06	0.000	2.77e+13	2.77e+13
Specialization_Business Administration	8.143e+13	6.5e+06	1.25e+07	0.000	8.14e+13	8.14e+13
Specialization_E-Business	-6.415e+13	1.29e+07	-4.96e+06	0.000	-6.41e+13	-6.41e+13
Specialization_E-COMMERCE	-1.305e+14	9.61e+06	-1.36e+07	0.000	-1.3e+14	-1.3e+14
Specialization_Finance Management	8.757e+13	5.75e+06	1.52e+07	0.000	8.76e+13	8.76e+13
Specialization_Healthcare Management	-4.925e+13	8.91e+06	-5.53e+06	0.000	-4.92e+13	-4.92e+13

Specialization_Hospitality Management	-2.862e+13	9.42e+06	-3.04e+06	0.000	-2.86e+13	-2.86e+13
Specialization_Human Resource Management	2.274e+13	5.74e+06	3.96e+06	0.000	2.27e+13	2.27e+13
Specialization_IT Projects Management	-6.794e+12	6.98e+06	-9.73e+05	0.000	-6.79e+12	-6.79e+12
Specialization_International Business	7.207e+13	8.12e+06	8.87e+06	0.000	7.21e+13	7.21e+13
Specialization_Marketing Management	5.332e+13	5.67e+06	9.4e+06	0.000	5.33e+13	5.33e+13
Specialization_Media and Advertising	1.1e+13	7.95e+06	1.38e+06	0.000	1.1e+13	1.1e+13
Specialization_Operations Management	4.443e+13	6.22e+06	7.14e+06	0.000	4.44e+13	4.44e+13
Specialization_Retail Management	-9.17e+13	1.02e+07	-9.02e+06	0.000	-9.17e+13	-9.17e+13
Specialization_Rural and Agribusiness	-2.008e+14	1.12e+07	-1.79e+07	0.000	-2.01e+14	-2.01e+14
Specialization_Select	-1.098e+13	4.18e+06	-2.63e+06	0.000	-1.1e+13	-1.1e+13
Specialization_Services Excellence	1.248e+14	1.66e+07	7.5e+06	0.000	1.25e+14	1.25e+14
Specialization_Supply Chain Management	-3.066e+13	6.74e+06	-4.55e+06	0.000	-3.07e+13	-3.07e+13
Specialization_Travel and Tourism	-7.874e+13	8.3e+06	-9.49e+06	0.000	-7.87e+13	-7.87e+13
City_Mumbai	-1.472e+14	4.61e+06	-3.19e+07	0.000	-1.47e+14	-1.47e+14
City_Other Cities	-1.375e+14	5.4e+06	-2.54e+07	0.000	-1.38e+14	-1.38e+14
City_Other Cities of Maharashtra	-7.228e+12	5.87e+06	-1.23e+06	0.000	-7.23e+12	-7.23e+12
City_Other Metro Cities	-2.81e+14	6.29e+06	-4.46e+07	0.000	-2.81e+14	-2.81e+14
City_Thane & Outskirts	-1.988e+14	5.26e+06	-3.78e+07	0.000	-1.99e+14	-1.99e+14
City_Tier II Cities	4.113e+14	1.1e+07	3.75e+07	0.000	4.11e+14	4.11e+14
Last Activity_Approached upfront	4.777e+15	2.92e+07	1.64e+08	0.000	4.78e+15	4.78e+15
Last Activity_Converted to Lead	-1.565e+14	1.06e+07	-1.48e+07	0.000	-1.57e+14	-1.57e+14
Last Activity_Email Bounced	-3.581e+14	1.17e+07	-3.05e+07	0.000	-3.58e+14	-3.58e+14
Last Activity_Email Link Clicked	2.393e+14	1.25e+07	1.92e+07	0.000	2.39e+14	2.39e+14
Last Activity_Email Marked Spam	2.971e+15	4.39e+07	6.77e+07	0.000	2.97e+15	2.97e+15
Last Activity_Email Opened	-9.667e+13	9.92e+06	-9.74e+06	0.000	-9.67e+13	-9.67e+13
Last Activity_Email Received	4.179e+15	6.8e+07	6.15e+07	0.000	4.18e+15	4.18e+15
Last Activity_Form Submitted on Website	4.846e+13	1.21e+07	4e+06	0.000	4.85e+13	4.85e+13
Last Activity_Had a Phone Conversation	1.575e+14	2.31e+07	6.81e+06	0.000	1.58e+14	1.58e+14
Last Activity_Olark Chat Conversation	-2.084e+14	1.01e+07	-2.07e+07	0.000	-2.08e+14	-2.08e+14
Last Activity_Page Visited on Website	-1.667e+13	1.06e+07	-1.58e+06	0.000	-1.67e+13	-1.67e+13
Last Activity_Resubscribed to emails	0	0	nan	nan	0	0
Last Activity_SMS Sent	1.688e+14	1e+07	1.68e+07	0.000	1.69e+14	1.69e+14
Last Activity_Unreachable	-1.374e+14	1.43e+07	-9.62e+06	0.000	-1.37e+14	-1.37e+14
Last Activity_Unsubscribed	6.919e+14	2.28e+07	3.03e+07	0.000	6.92e+14	6.92e+14
Last Activity_View in browser link Clicked	-2.15e+15	6.81e+07	-3.16e+07	0.000	-2.15e+15	-2.15e+15
Last Activity_Visited Booth in Tradeshow	1.354e+15	6.9e+07	1.96e+07	0.000	1.35e+15	1.35e+15

Feature Selection using RFE

Dep. Variable:	Converted	**No. Observations:**	6002
Model:	GLM	**Df Residuals:**	5981
Model Family:	Binomial	**Df Model:**	20
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1264.7
Date:	Sun, 03 Mar 2019	**Deviance:**	2529.4
Time:	17:43:35	**Pearson chi2:**	8.56e+03
No. Iterations:	24	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4929	0.090	-27.836	0.000	-2.668	-2.317
Lead Source_Welingak Website	3.2281	0.731	4.414	0.000	1.795	4.662
Lead Quality_Worst	-2.5504	0.761	-3.354	0.001	-4.041	-1.060
Asymmetrique Activity Index_03.Low	-2.4592	0.358	-6.869	0.000	-3.161	-1.758
Tags_Already a student	-3.8785	0.726	-5.344	0.000	-5.301	-2.456
Tags_Closed by Horizzon	5.1421	0.722	7.120	0.000	3.727	6.558
Tags_Diploma holder (Not Eligible)	-24.1871	2.82e+04	-0.001	0.999	-5.52e+04	5.52e+04
Tags_Interested in full time MBA	-3.0545	0.742	-4.117	0.000	-4.509	-1.600
Tags_Interested in other courses	-3.0288	0.330	-9.183	0.000	-3.675	-2.382
Tags_Lost to EINS	6.3792	0.831	7.677	0.000	4.751	8.008
Tags_Not doing further education	-3.7904	1.032	-3.674	0.000	-5.813	-1.768
Tags_Ringing	-4.2659	0.249	-17.107	0.000	-4.755	-3.777
Tags_Will revert after reading the email	3.5963	0.194	18.561	0.000	3.217	3.976
Tags_invalid number	-25.7192	2.7e+04	-0.001	0.999	-5.3e+04	5.29e+04
Tags_number not provided	-25.9733	4.5e+04	-0.001	1.000	-8.82e+04	8.82e+04
Tags_opp hangup	-3.5152	1.063	-3.308	0.001	-5.598	-1.433
Tags_switched off	-5.1620	0.724	-7.126	0.000	-6.582	-3.742
Tags_wrong number given	-26.1206	3.49e+04	-0.001	0.999	-6.84e+04	6.84e+04
What is your current occupation_Unemployed	2.0649	0.119	17.357	0.000	1.832	2.298
What is your current occupation_Working Professional	2.1458	0.364	5.903	0.000	1.433	2.858
Last Activity_SMS Sent	2.0390	0.112	18.174	0.000	1.819	2.259

Creating a data frame with the actual churn flag and the predicted probabilities

	Converted	Conversion_Prob	LeadID
0	0	0.065692	8529
1	0	0.009069	7331
2	1	0.833555	7688
3	0	0.076360	92
4	0	0.076360	4908

Creating new column 'predicted' with 1 if Churn_Prob > 0.5 else 0

	Converted	Conversion_Prob	LeadID	predicted
0	0	0.065692	8529	0
1	0	0.009069	7331	0
2	1	0.833555	7688	1
3	0	0.076360	92	0
4	0	0.076360	4908	0

Creating Confusion Metrics

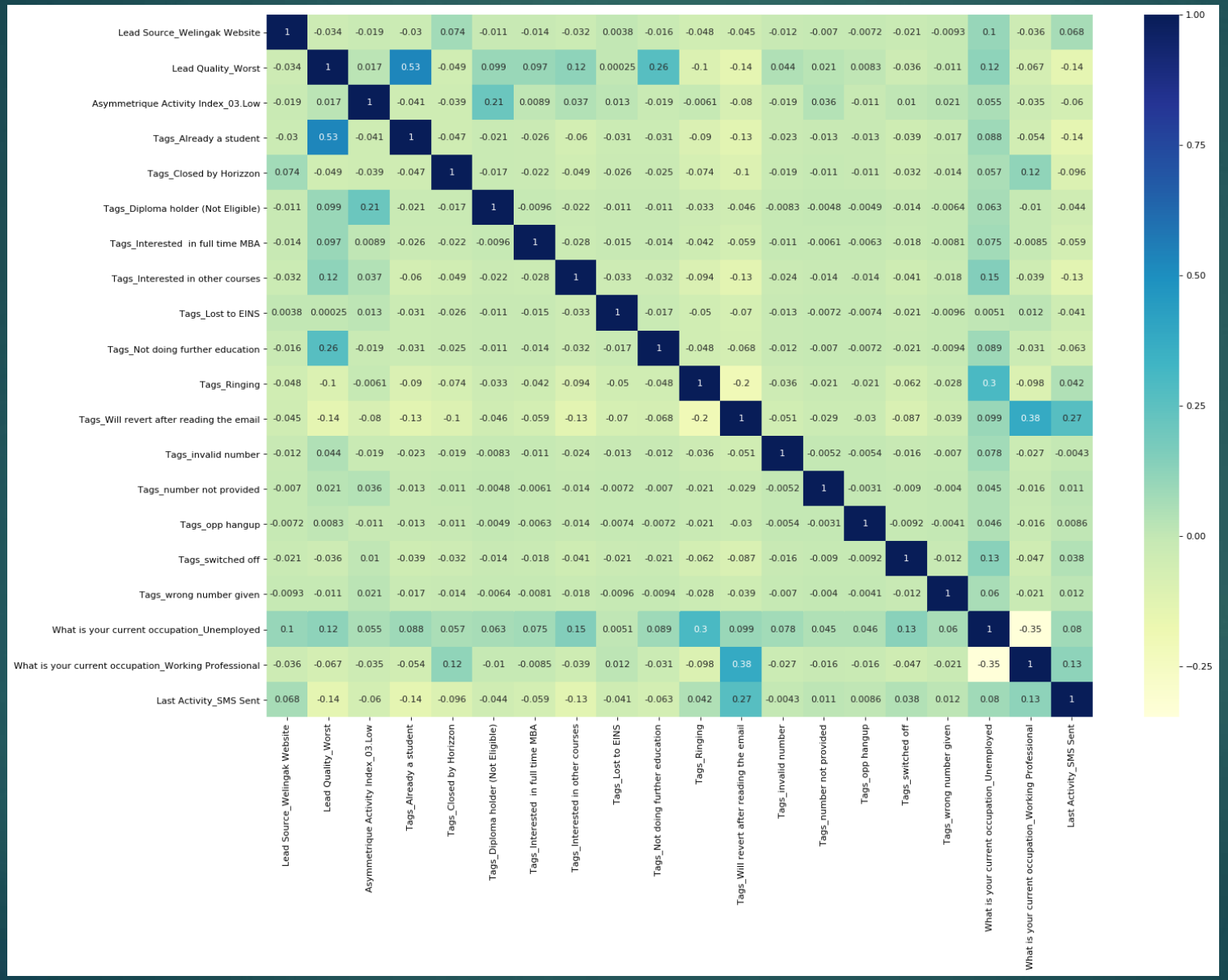
```
# Confusion matrix
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.predicted )
print(confusion)
```

```
[[3647   89]
 [ 409 1857]]
```

Checking VIFs

	Features	VIF
4	Tags_Closed by Horizzon	1.30
9	Tags_Not doing further education	1.27
15	Tags_switched off	1.20
5	Tags_Diploma holder (Not Eligible)	1.12
6	Tags_Interested in full time MBA	1.12
2	Asymmetrique Activity Index_03.Low	1.11
0	Lead Source_Welingak Website	1.09
12	Tags_invalid number	1.08
8	Tags_Lost to EINS	1.07
16	Tags_wrong number given	1.04
14	Tags_opp hangup	1.03
13	Tags_number not provided	1.03
18	What is your current occupation_Working Profes...	0.80
1	Lead Quality_Worst	0.69
10	Tags_Ringing	0.62
7	Tags_Interested in other courses	0.40
3	Tags_Already a student	0.38
11	Tags_Will revert after reading the email	0.09
17	What is your current occupation_Unemployed	0.01
19	Last Activity_SMS Sent	0.00

Checking correlation between features using Heat Map



Dropping the Variable and Updating the Model

Dep. Variable:	Converted	No. Observations:	6002
Model:	GLM	Df Residuals:	5982
Model Family:	Binomial	Df Model:	19
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1278.7
Date:	Sun, 03 Mar 2019	Deviance:	2557.4
Time:	17:44:02	Pearson chi2:	8.49e+03
No. Iterations:	24	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4804	0.089	-27.881	0.000	-2.655	-2.306
Lead Source_Welingak Website	3.2918	0.731	4.503	0.000	1.859	4.725
Lead Quality_Worst	-2.7112	0.739	-3.668	0.000	-4.160	-1.263
Asymmetrique Activity Index_03.Low	-2.4342	0.357	-6.817	0.000	-3.134	-1.734
Tags_Already a student	-3.8015	0.724	-5.247	0.000	-5.221	-2.382
Tags_Closed by Horizzon	5.1851	0.722	7.184	0.000	3.770	6.600
Tags_Diploma holder (Not Eligible)	-24.1120	2.81e+04	-0.001	0.999	-5.51e+04	5.51e+04
Tags_Interested in full time MBA	-2.9855	0.741	-4.028	0.000	-4.438	-1.533
Tags_Interested in other courses	-2.9603	0.329	-8.996	0.000	-3.605	-2.315
Tags_Lost to EINS	6.4382	0.838	7.684	0.000	4.796	8.080
Tags_Not doing further education	-3.7070	1.031	-3.596	0.000	-5.727	-1.687
Tags_Ringing	-4.1829	0.248	-16.855	0.000	-4.669	-3.696
Tags_Will revert after reading the email	3.6368	0.193	18.834	0.000	3.258	4.015
Tags_invalid number	-25.6348	2.7e+04	-0.001	0.999	-5.3e+04	5.29e+04
Tags_opp hangup	-3.4305	1.062	-3.231	0.001	-5.512	-1.349
Tags_switched off	-5.0770	0.724	-7.013	0.000	-6.496	-3.658
Tags_wrong number given	-26.0375	3.49e+04	-0.001	0.999	-6.85e+04	6.84e+04
What is your current occupation_Unemployed	1.9949	0.118	16.969	0.000	1.764	2.225
What is your current occupation_Working Professional	2.1030	0.363	5.788	0.000	1.391	2.815
Last Activity_SMS Sent	2.0063	0.111	18.069	0.000	1.789	2.224

Creating a data frame with the actual churn flag and the predicted probabilities

	Converted	Conversion_Prob	LeadID
0	0	0.065249	8529
1	0	0.009300	7331
2	1	0.820658	7688
3	0	0.077242	92
4	0	0.077242	4908

Creating new column 'predicted' with 1 if Churn_Prob > 0.5 else 0

	Converted	Conversion_Prob	LeadID	predicted
0	0	0.065249	8529	0
1	0	0.009300	7331	0
2	1	0.820658	7688	1
3	0	0.077242	92	0
4	0	0.077242	4908	0

Checking VIFs again..

	Features	VIF
4	Tags_Closed by Horizzon	1.29
9	Tags_Not doing further education	1.27
14	Tags_switched off	1.19
6	Tags_Interested in full time MBA	1.12
5	Tags_Diploma holder (Not Eligible)	1.12
2	Asymmetrique Activity Index_03.Low	1.11
0	Lead Source_Welingak Website	1.09
12	Tags_invalid number	1.08
8	Tags_Lost to EINS	1.07
15	Tags_wrong number given	1.04
13	Tags_opp hangup	1.03
17	What is your current occupation_Working Profes...	0.79
1	Lead Quality_Worst	0.69
10	Tags_Ringing	0.62
7	Tags_Interested in other courses	0.39
3	Tags_Already a student	0.38
11	Tags_Will revert after reading the email	0.09
16	What is your current occupation_Unemployed	0.01
18	Last Activity_SMS Sent	0.00

Dropping the Variable and Updating the Model

Dep. Variable:	Converted	No. Observations:	6002
Model:	GLM	Df Residuals:	5983
Model Family:	Binomial	Df Model:	18
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1305.1
Date:	Sun, 03 Mar 2019	Deviance:	2610.1
Time:	17:44:18	Pearson chi2:	8.25e+03
No. Iterations:	23	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4653	0.088	-27.969	0.000	-2.638	-2.293
Lead Source_Welingak Website	3.4161	0.731	4.676	0.000	1.984	4.848
Lead Quality_Worst	-2.7568	0.728	-3.787	0.000	-4.184	-1.330
Asymmetrique Activity Index_03.Low	-2.3688	0.357	-6.637	0.000	-3.068	-1.669
Tags_Already a student	-3.6760	0.724	-5.080	0.000	-5.094	-2.258
Tags_Closed by Horizzon	5.2742	0.721	7.314	0.000	3.861	6.687
Tags_Diploma holder (Not Eligible)	-22.9881	1.71e+04	-0.001	0.999	-3.35e+04	3.35e+04
Tags_Interested in full time MBA	-2.8602	0.740	-3.866	0.000	-4.310	-1.410
Tags_Interested in other courses	-2.8332	0.328	-8.641	0.000	-3.476	-2.191
Tags_Lost to EINS	6.4558	0.839	7.692	0.000	4.811	8.101
Tags_Not doing further education	-3.5698	1.030	-3.467	0.001	-5.588	-1.552
Tags_Ringing	-4.0320	0.246	-16.378	0.000	-4.515	-3.550
Tags_Will revert after reading the email	3.7184	0.192	19.386	0.000	3.342	4.094
Tags_invalid number	-24.4886	1.64e+04	-0.001	0.999	-3.22e+04	3.21e+04
Tags_opp hangup	-3.2794	1.061	-3.092	0.002	-5.358	-1.201
Tags_switched off	-4.9237	0.723	-6.809	0.000	-6.341	-3.506
What is your current occupation_Unemployed	1.8623	0.115	16.189	0.000	1.637	2.088
What is your current occupation_Working Professional	2.0226	0.363	5.570	0.000	1.311	2.734
Last Activity_SMS Sent	1.9628	0.109	17.982	0.000	1.749	2.177

Creating a data frame with the actual churn flag and the predicted probabilities

	Converted	Conversion_Prob	LeadID
0	0	0.064635	8529
1	0	0.009613	7331
2	1	0.795734	7688
3	0	0.078329	92
4	0	0.078329	4908

Creating new column 'predicted' with 1 if Churn_Prob > 0.5 else 0

	Converted	Conversion_Prob	LeadID	predicted
0	0	0.064635	8529	0
1	0	0.009613	7331	0
2	1	0.795734	7688	1
3	0	0.078329	92	0
4	0	0.078329	4908	0

Checking VIFs again..

	Features	VIF
4	Tags_Closed by Horizzon	1.29
9	Tags_Not doing further education	1.26
14	Tags_switched off	1.19
6	Tags_Interested in full time MBA	1.12
5	Tags_Diploma holder (Not Eligible)	1.12
2	Asymmetrique Activity Index_03.Low	1.11
0	Lead Source_Welingak Website	1.09
12	Tags_invalid number	1.08
8	Tags_Lost to EINS	1.06
13	Tags_opp hangup	1.02
16	What is your current occupation_Working Profes...	0.79
1	Lead Quality_Worst	0.69
10	Tags_Ringing	0.61
7	Tags_Interested in other courses	0.39
3	Tags_Already a student	0.38
11	Tags_Will revert after reading the email	0.09
15	What is your current occupation_Unemployed	0.01
17	Last Activity_SMS Sent	0.00

Dropping the Variable and Updating the Model

Dep. Variable:	Converted	No. Observations:	6002
Model:	GLM	Df Residuals:	5984
Model Family:	Binomial	Df Model:	17
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1313.2
Date:	Sun, 03 Mar 2019	Deviance:	2626.4
Time:	17:44:36	Pearson chi2:	8.42e+03
No. Iterations:	23	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4750	0.088	-28.020	0.000	-2.648	-2.302
Lead Source_Welingak Website	3.4678	0.731	4.747	0.000	2.036	4.900
Lead Quality_Worst	-2.8883	0.706	-4.092	0.000	-4.272	-1.505
Asymmetrique Activity Index_03.Low	-2.4330	0.351	-6.931	0.000	-3.121	-1.745
Tags_Already a student	-3.6149	0.723	-4.999	0.000	-5.032	-2.198
Tags_Closed by Horizzon	5.3212	0.721	7.382	0.000	3.908	6.734
Tags_Interested in full time MBA	-2.8081	0.740	-3.794	0.000	-4.259	-1.357
Tags_Interested in other courses	-2.7838	0.328	-8.493	0.000	-3.426	-2.141
Tags_Lost to EINS	6.5606	0.846	7.757	0.000	4.903	8.218
Tags_Not doing further education	-3.5144	1.030	-3.412	0.001	-5.533	-1.496
Tags_Ringing	-3.9921	0.246	-16.235	0.000	-4.474	-3.510
Tags_Will revert after reading the email	3.7631	0.192	19.646	0.000	3.388	4.138
Tags_invalid number	-24.4442	1.64e+04	-0.001	0.999	-3.22e+04	3.21e+04
Tags_opp hangup	-3.2379	1.061	-3.052	0.002	-5.317	-1.159
Tags_switched off	-4.8845	0.723	-6.756	0.000	-6.302	-3.467
What is your current occupation_Unemployed	1.8184	0.114	15.893	0.000	1.594	2.043
What is your current occupation_Working Professional	1.9876	0.362	5.486	0.000	1.277	2.698
Last Activity_SMS Sent	1.9808	0.109	18.198	0.000	1.767	2.194

Creating a data frame with the actual churn flag and the predicted probabilities

	Converted	Conversion_Prob	LeadID
0	0	0.064888	8529
1	0	0.009483	7331
2	1	0.789866	7688
3	0	0.077629	92
4	0	0.077629	4908

Creating new column 'predicted' with 1 if Churn_Prob > 0.5 else 0

	Converted	Conversion_Prob	LeadID	predicted
0	0	0.064888	8529	0
1	0	0.009483	7331	0
2	1	0.789866	7688	1
3	0	0.077629	92	0
4	0	0.077629	4908	0

Checking VIFs again..

	Features	VIF
4	Tags_Closed by Horizzon	1.28
8	Tags_Not doing further education	1.25
13	Tags_switched off	1.18
5	Tags_Interested in full time MBA	1.11
0	Lead Source_Welingak Website	1.08
11	Tags_invalid number	1.07
2	Asymmetrique Activity Index_03.Low	1.07
7	Tags_Lost to EINS	1.06
12	Tags_opp hangup	1.02
15	What is your current occupation_Working Profes...	0.78
1	Lead Quality_Worst	0.67
9	Tags_Ringing	0.59
6	Tags_Interested in other courses	0.38
3	Tags_Already a student	0.37
10	Tags_Will revert after reading the email	0.09
14	What is your current occupation_Unemployed	0.01
16	Last Activity_SMS Sent	0.00

Dropping the Variable and Updating the Model

Dep. Variable:	Converted	No. Observations:	6002
Model:	GLM	Df Residuals:	5985
Model Family:	Binomial	Df Model:	16
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1342.4
Date:	Sun, 03 Mar 2019	Deviance:	2684.8
Time:	17:44:56	Pearson chi2:	8.52e+03
No. Iterations:	8	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4751	0.088	-28.144	0.000	-2.647	-2.303
Lead Source_Welingak Website	3.6135	0.730	4.949	0.000	2.182	5.044
Lead Quality_Worst	-3.1794	0.670	-4.742	0.000	-4.494	-1.865
Asymmetrique Activity Index_03.Low	-2.3401	0.354	-6.605	0.000	-3.035	-1.646
Tags_Already a student	-3.4492	0.722	-4.776	0.000	-4.865	-2.034
Tags_Closed by Horizzon	5.4435	0.720	7.559	0.000	4.032	6.855
Tags_Interested in full time MBA	-2.6565	0.740	-3.591	0.000	-4.106	-1.207
Tags_Interested in other courses	-2.6347	0.327	-8.060	0.000	-3.275	-1.994
Tags_Lost to EINS	6.7102	0.862	7.786	0.000	5.021	8.399
Tags_Not doing further education	-3.3472	1.030	-3.250	0.001	-5.366	-1.329
Tags_Ringing	-3.8360	0.244	-15.709	0.000	-4.315	-3.357
Tags_Will revert after reading the email	3.8695	0.190	20.331	0.000	3.497	4.243
Tags_opp hangup	-3.0789	1.061	-2.903	0.004	-5.158	-1.000
Tags_switched off	-4.7274	0.722	-6.544	0.000	-6.143	-3.311
What is your current occupation_Unemployed	1.6711	0.112	14.926	0.000	1.452	1.891
What is your current occupation_Working Professional	1.8944	0.363	5.221	0.000	1.183	2.606
Last Activity_SMS Sent	1.9687	0.107	18.383	0.000	1.759	2.179

Creating a data frame with the actual Converted flag and the predicted probabilities

	Converted	Conversion_Prob	LeadID
0	0	0.064688	8529
1	0	0.009566	7331
2	1	0.762190	7688
3	0	0.077626	92
4	0	0.077626	4908

Creating new column 'predicted' with 1 if Conversion_Prob > 0.5 else 0

	Converted	Conversion_Prob	LeadID	predicted
0	0	0.064688	8529	0
1	0	0.009566	7331	0
2	1	0.762190	7688	1
3	0	0.077626	92	0
4	0	0.077626	4908	0

Checking VIFs again..

	Features	VIF
4	Tags_Closed by Horizzon	1.26
8	Tags_Not doing further education	1.23
12	Tags_switched off	1.17
5	Tags_Interested in full time MBA	1.10
0	Lead Source_Welingak Website	1.08
2	Asymmetrique Activity Index_03.Low	1.07
7	Tags_Lost to EINS	1.06
11	Tags_opp hangup	1.02
14	What is your current occupation_Working Profes...	0.77
1	Lead Quality_Worst	0.67
9	Tags_Ringing	0.58
6	Tags_Interested in other courses	0.38
3	Tags_Already a student	0.36
10	Tags_Will revert after reading the email	0.09
13	What is your current occupation_Unemployed	0.01
15	Last Activity_SMS Sent	0.00


Lead Source_Welingak Website 1 -0.034 -0.019 -0.03 0.074 -0.014 -0.032 0.0038 -0.016 -0.048 -0.045 -0.0072 -0.021 0.1 -0.036 0.068
Lead Quality_Worst -0.034 1 0.017 0.53 -0.049 0.097 0.12 0.00025 0.26 -0.1 -0.14 0.0083 -0.036 0.12 -0.067 -0.14
Asymmetrique Activity Index_03.Low -0.019 0.017 1 -0.041 -0.039 0.0089 0.037 0.013 -0.019 -0.0061 -0.08 -0.011 0.01 0.055 -0.035 -0.06
Tags_Already a student -0.03 0.53 -0.041 1 -0.047 -0.026 -0.06 -0.031 -0.031 -0.09 -0.13 -0.013 -0.039 0.088 -0.054 -0.14
Tags_Closed by Horizzon 0.074 -0.049 -0.039 -0.047 1 -0.022 -0.049 -0.026 -0.025 -0.074 -0.1 -0.011 -0.032 0.057 0.12 -0.096
Tags_Interested in full time MBA -0.014 0.097 0.0089 -0.026 -0.022 1 -0.028 -0.015 -0.014 -0.042 -0.059 -0.0063 -0.018 0.075 -0.0085 -0.059
Tags_Interested in other courses -0.032 0.12 0.037 -0.06 -0.049 -0.028 1 -0.033 -0.032 -0.094 -0.13 -0.014 -0.041 0.15 -0.039 -0.13
Tags_Lost to EINS 0.0038 0.00025 0.013 -0.031 -0.026 -0.015 -0.033 1 -0.017 -0.05 -0.07 -0.0074 -0.021 0.0051 0.012 -0.041
Tags_Not doing further education -0.016 0.26 -0.019 -0.031 -0.025 -0.014 -0.032 -0.017 1 -0.048 -0.068 -0.0072 -0.021 0.089 -0.031 -0.063
Tags_Ringing -0.048 -0.1 -0.0061 -0.09 -0.074 -0.042 -0.094 -0.05 -0.048 1 -0.2 -0.021 -0.062 0.3 -0.098 0.042
Tags_Will revert after reading the email -0.045 -0.14 -0.08 -0.13 -0.1 -0.059 -0.13 -0.07 -0.068 -0.2 1 -0.03 -0.087 0.099 0.38 0.27
Tags_opp hangup -0.0072 0.0083 -0.011 -0.013 -0.011 -0.0063 -0.014 -0.0074 -0.0072 -0.021 -0.03 1 -0.0092 0.046 -0.016 0.0086
Tags_switched off -0.021 -0.036 0.01 -0.039 -0.032 -0.018 -0.041 -0.021 -0.021 -0.062 -0.087 -0.0092 1 0.13 -0.047 0.038
What is your current occupation_Unemployed 0.1 0.12 0.055 0.088 0.057 0.075 0.15 0.0051 0.089 0.3 0.099 0.046 0.13 1 -0.35 0.08
What is your current occupation_Working Professional -0.036 -0.067 -0.035 -0.054 0.12 -0.0085 -0.039 0.012 -0.031 -0.098 0.38 -0.016 -0.047 -0.35 1 0.13
Last Activity_SMS Sent 0.068 -0.14 -0.06 -0.14 -0.096 -0.059 -0.13 -0.041 -0.063 0.042 0.27 0.0086 0.038 0.08 0.13 1
Lead Source_Welingak Website
Lead Quality_Worst
Asymmetrique Activity Index_03.Low
Tags_Already a student
Tags_Closed by Horizzon
Tags_Interested in full time MBA
Tags_Interested in other courses
Tags_Lost to EINS
Tags_Not doing further education
Tags_Ringing
Tags_Will revert after reading the email
Tags_opp hangup
Tags_switched off
What is your current occupation_Unemployed
What is your current occupation_Working Professional
Last Activity_SMS Sent
1.00
0.75
0.50
0.25
0.00
−0.25

Our latest model have the following features:

- All variables have p-value
- All the features have very low VIF values, meaning, there is hardly any multicollinearity among the features. This is also evident from the heat map.
- Overall accuracy of 0.9125 – highly acceptable

So we need not drop any more variables and we can proceed with making predictions using this model only

Calculating Metrics beyond Accuracy

```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.8195057369814651
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.9689507494646681
```

```
# Calculate false postive rate - predicting churn when customer does not have churned
print(FP/ float(TN+FP))
```

```
0.031049250535331904
```

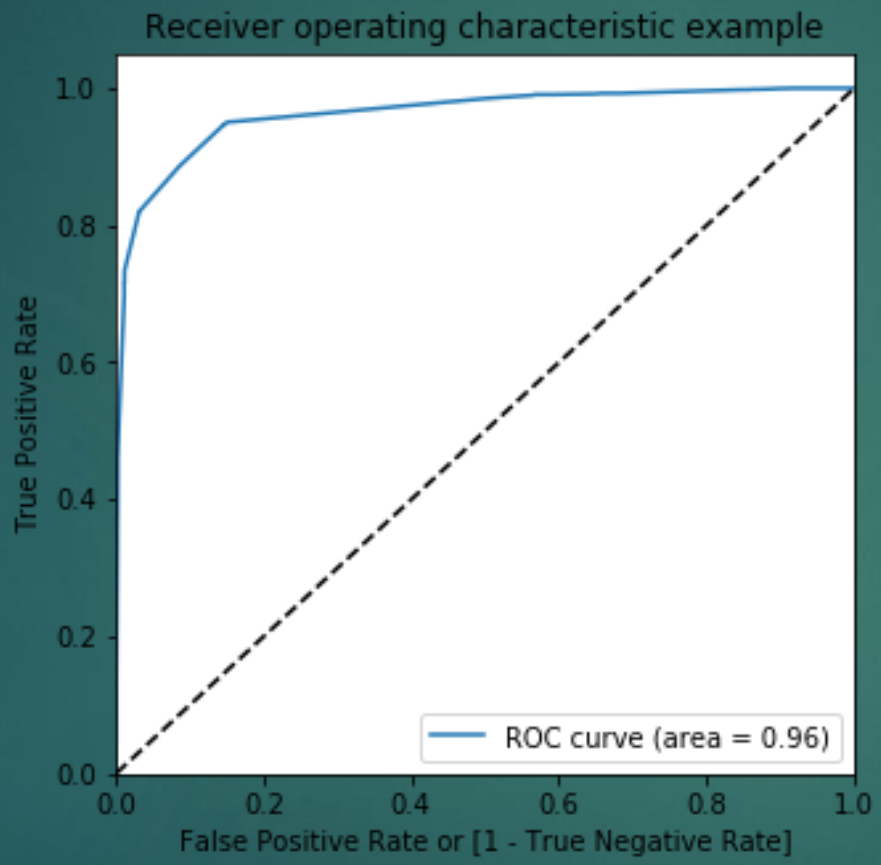
```
# positive predictive value
print (TP / float(TP+FP))
```

```
0.941206284845413
```

```
# Negative predictive value
print (TN / float(TN+ FN))
```

```
0.8984859766691486
```

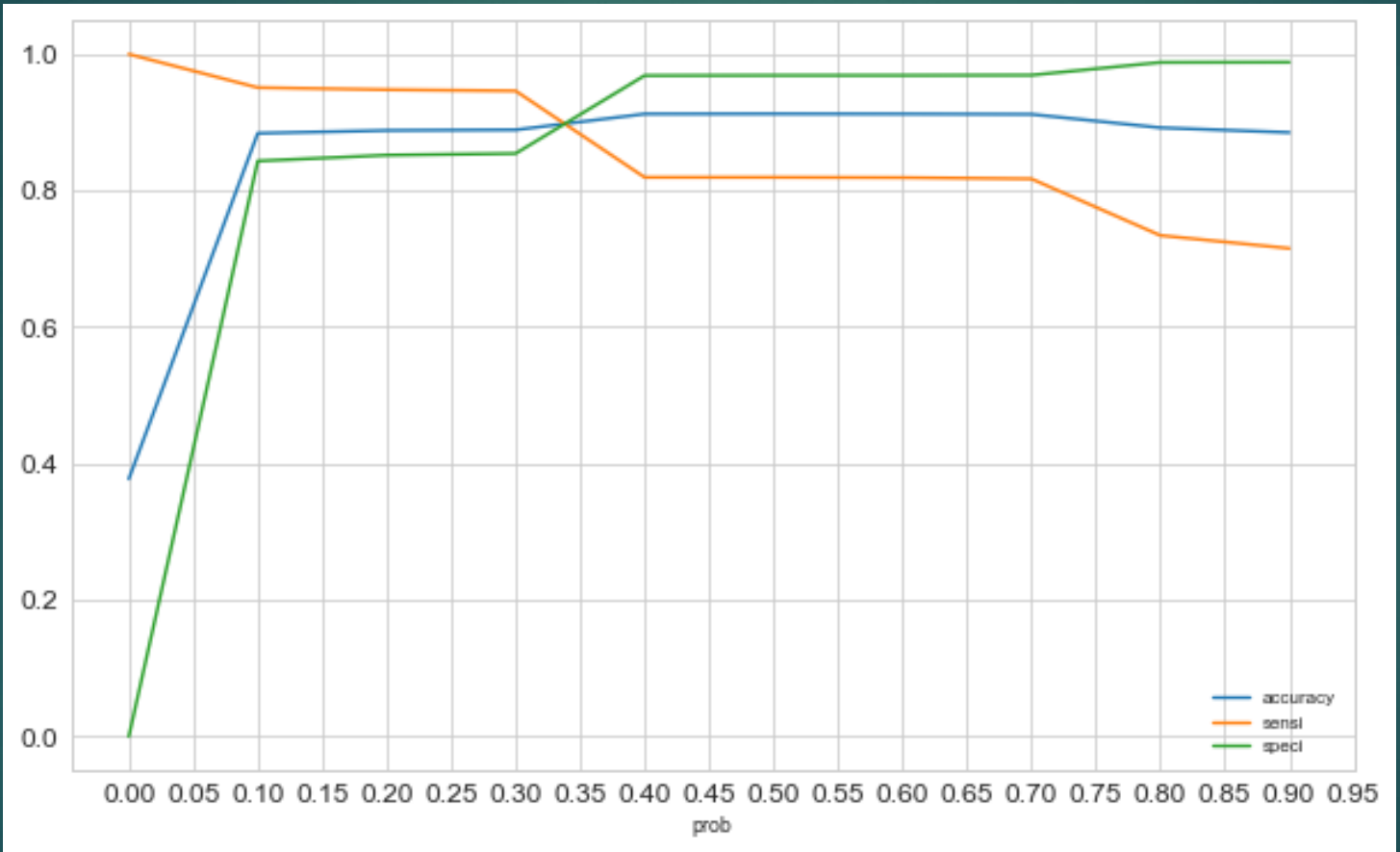
Plotting the ROC Curve



Finding Optimal Cutoff Point

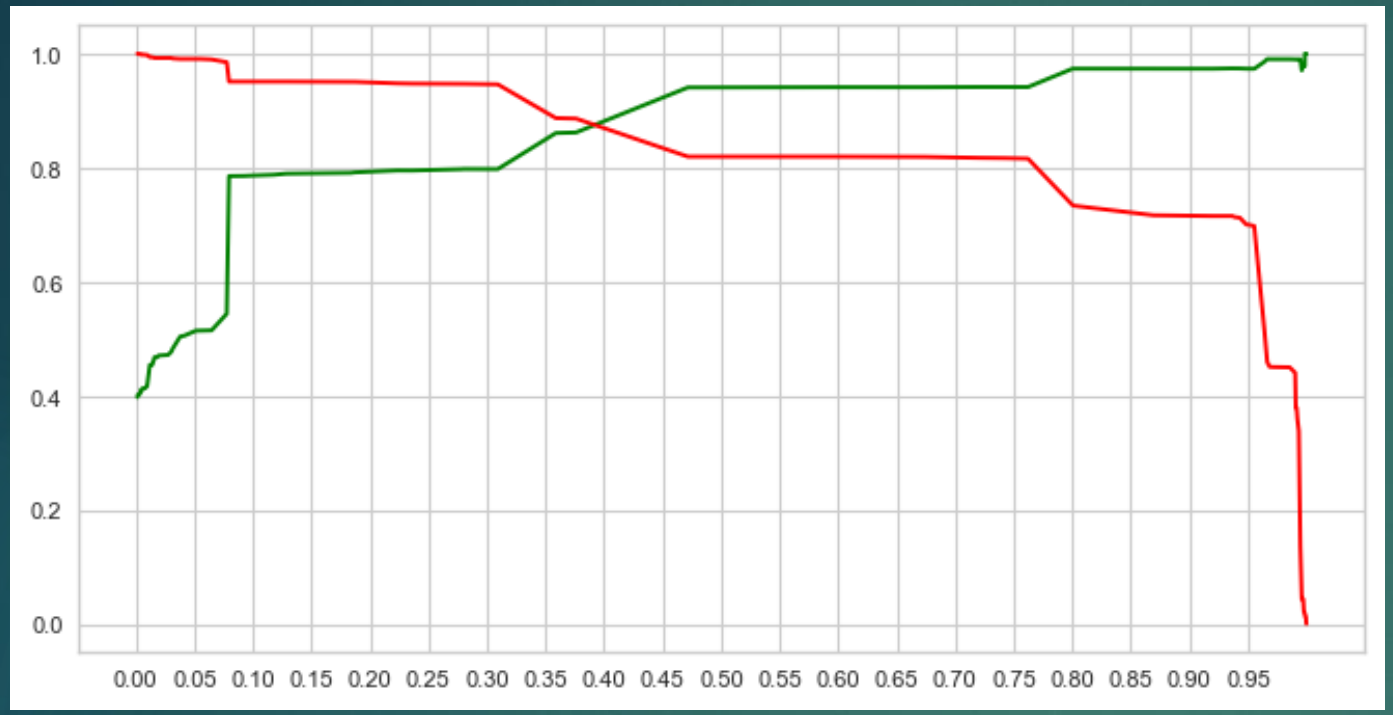
	Converted	Conversion_Prob	LeadID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0	0.064688	8529	0	1	0	0	0	0	0	0	0	0	0
1	0	0.009566	7331	0	1	0	0	0	0	0	0	0	0	0
2	1	0.762190	7688	1	1	1	1	1	1	1	1	1	0	0
3	0	0.077626	92	0	1	0	0	0	0	0	0	0	0	0
4	0	0.077626	4908	0	1	0	0	0	0	0	0	0	0	0

Plotting accuracy sensitivity and specificity for various probabilities



From the curve above, 0.33 is the optimum point to take it as a cutoff probability.

Precision and Recall



From the precision-recall graph above, we get the optical threshold value as close to .37. However our business requirement here is to have Lead Conversion Rate around 80%.

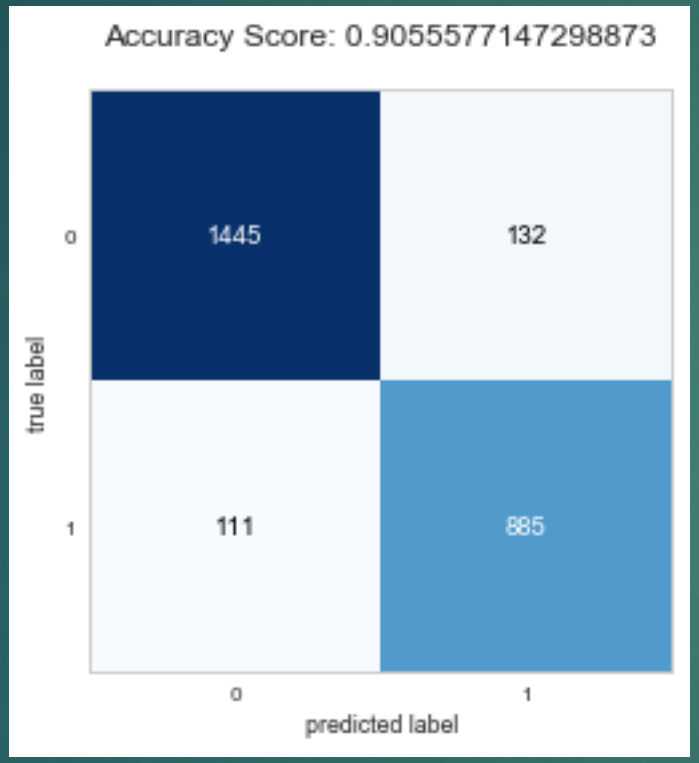
This is already achieved with our earlier threshold value of 0.33. So we will stick to this value.

Calculating the F1 score

```
F1 = 2*(precision*recall)/(precision+recall)
F1
```

```
0.8737231036731146
```

Confusion Matrix in Visuals



Different metrics beyond accuracy on the test dataset

Sensitivity ¶

TP / TP + FN

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.8885542168674698
```

Specificity

TN / TN + FP

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.9162967660114141
```

F1 = 2×(Precision*Recall)/(Precision+Recall) ¶

```
F1 = 2*(Precision*Recall)/(Precision+Recall)
F1
```

```
0.879284649776453
```

False Postive Rate

FP / TN + FP

```
# Calculate false postive rate - predicting churn when customer does not have churned
print(FP/ float(TN+FP))
```

```
0.08370323398858592
```

Positive Predictive Value ¶

TP / TP + FP

```
# Positive predictive value
print (TP / float(TP+FP))
```

```
0.8702064896755162
```

Negative Predictive Value

TN / TN + FN

```
# Negative predictive value
print (TN / float(TN+ FN))
```

```
0.9286632390745502
```

Precision

TP / TP + FP

```
Precision = confusion_test[1,1]/(confusion_test[0,1]+confusion_test[1,1])
Precision
```

```
0.8702064896755162
```

Recall

TP / TP + FN

```
Recall = confusion_test[1,1]/(confusion_test[1,0]+confusion_test[1,1])
Recall
```

```
0.8885542168674698
```

Classification Report

```
from sklearn.metrics import classification_report
print(classification_report(y_pred_final.Converted, y_pred_final.final_predicted))
```

```
             precision    recall  f1-score   support

          0       0.93      0.92      0.92      1577
          1       0.87      0.89      0.88       996

avg / total       0.91      0.91      0.91      2573
```

Cross Validation Score

To avoid overfitting, let us calculate the Cross Validation Score to see how our model performs

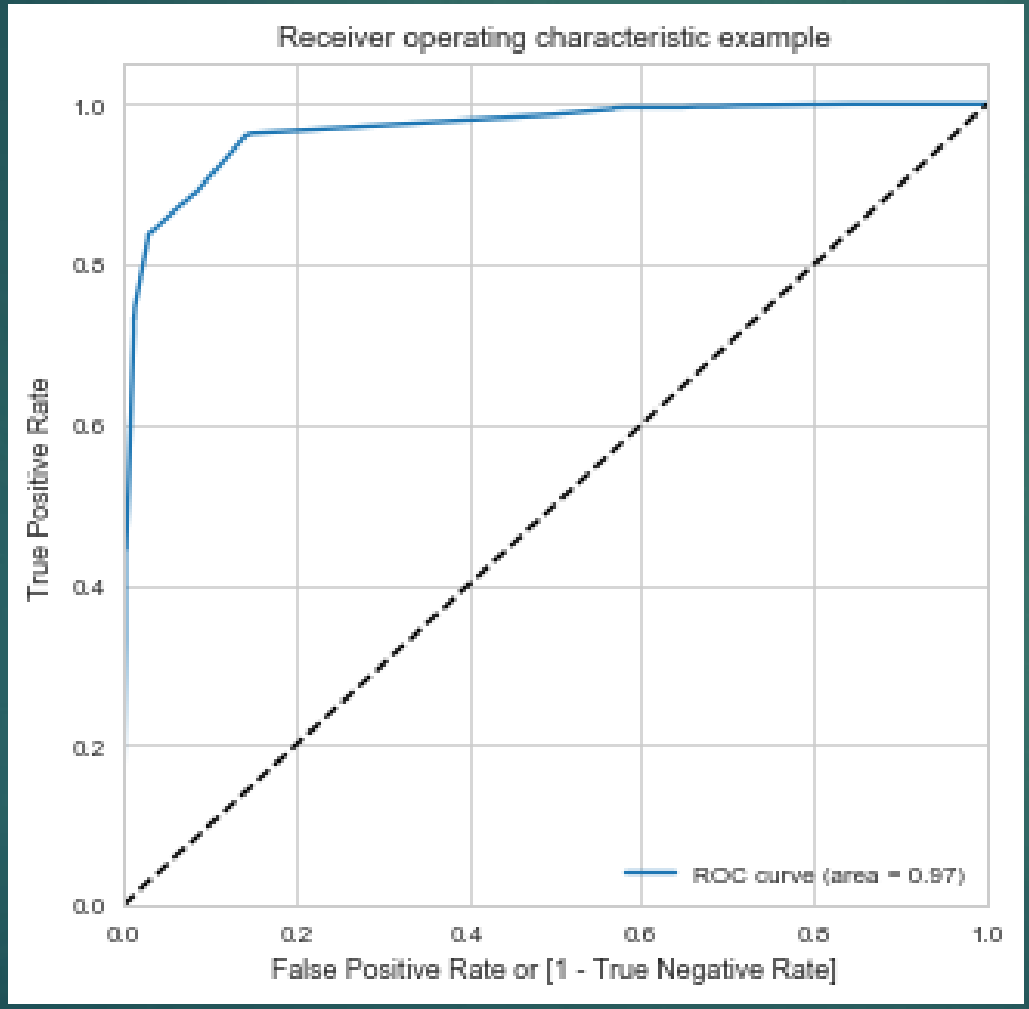
```
from sklearn.model_selection import cross_val_score

lr = LogisticRegression(solver = 'lbfgs')
scores = cross_val_score(lr, X, y, cv=10)
scores.sort()
accuracy = scores.mean()

print(scores)
print(accuracy)
```

```
[0.84364061 0.89731622 0.90898483 0.91248541 0.91501746 0.92424242
 0.92532089 0.92898719 0.92998833 0.9369895 ]
0.9122972868451112
```

Plotting the ROC Curve for Test Dataset



Calculating the Area Under the Curve(GINI)

```
def auc_val(fpr,tpr):
    AreaUnderCurve = 0.
    for i in range(len(fpr)-1):
        AreaUnderCurve += (fpr[i+1]-fpr[i]) * (tpr[i+1]+tpr[i])
    AreaUnderCurve *= 0.5
    return AreaUnderCurve
```

```
auc = auc_val(fpr,tpr)
auc
```

```
0.9678947241088641
```

As a rule of thumb, an AUC can be classed as follows,

- 0.90 - 1.00 = excellent
- 0.80 - 0.90 = good
- 0.70 - 0.80 = fair
- 0.60 - 0.70 = poor
- 0.50 - 0.60 = fail

Since we got a value of 0.9678, our model seems to be doing well on the test dataset.

Calculating Lead score for the entire dataset

Lead Score = 100 * Conversion Probability

test dataset

	LeadID	Converted	Conversion_Prob	final_predicted
0	6190	0	0.000591	0
1	7073	0	0.077626	0
2	4519	0	0.309185	0
3	607	1	0.999825	1
4	440	0	0.077626	0

train dataset

	LeadID	Converted	Conversion_Prob	final_predicted
0	8529	0	0.064688	0
1	7331	0	0.009566	0
2	7688	1	0.762190	1
3	92	0	0.077626	0
4	4908	0	0.077626	0

Concatenating the train and the test dataset with the Conversion Probabilities

	LeadID	Converted	Conversion_Prob	final_predicted
0	8529	0	0.064688	0
1	7331	0	0.009566	0
2	7688	1	0.762190	1
3	92	0	0.077626	0
4	4908	0	0.077626	0

Calculating the lead score value

	LeadID	Converted	Conversion_Prob	final_predicted	Lead_Score
0	8529	0	0.064688	0	6
1	7331	0	0.009566	0	1
2	7688	1	0.762190	1	76
3	92	0	0.077626	0	8
4	4908	0	0.077626	0	8

Associating the lead ids and score with respective leads

LeadID	Lead Number	Converted	Conversion_Prob	final_predicted	Lead_Score
0	660737	0	0.031109	0	3
1	660728	0	0.009566	0	1
2	660727	1	0.801308	1	80
3	660719	0	0.009566	0	1
4	660681	1	0.955452	1	96
5	660680	0	0.077626	0	8
6	660673	1	0.955452	1	96
7	660664	0	0.077626	0	8
8	660624	0	0.077626	0	8
9	660616	0	0.077626	0	8

Determining Feature Importance

```
Lead Source_Welingak Website                          3.61
Lead Quality_Worst                                   -3.18
Asymmetrique Activity Index_03.Low                   -2.34
Tags_Already a student                               -3.45
Tags_Closed by Horizzon                               5.44
Tags_Interested  in full time MBA                    -2.66
Tags_Interested in other courses                     -2.63
Tags_Lost to EINS                                     6.71
Tags_Not doing further education                     -3.35
Tags_Ringing                                         -3.84
Tags_Will revert after reading the email              3.87
Tags_opp hangup                                      -3.08
Tags_switched off                                    -4.73
What is your current occupation_Unemployed            1.67
What is your current occupation_Working Professional  1.89
Last Activity_SMS Sent                                1.97
dtype: float64
```

Getting a relative coefficient value for all the features wrt the feature with the highest coefficient

```
Lead Source_Welingak Website                          53.85
Lead Quality_Worst                                   -47.38
Asymmetrique Activity Index_03.Low                   -34.87
Tags_Already a student                               -51.40
Tags_Closed by Horizzon                               81.12
Tags_Interested  in full time MBA                    -39.59
Tags_Interested in other courses                     -39.26
Tags_Lost to EINS                                    100.00
Tags_Not doing further education                     -49.88
Tags_Ringing                                         -57.17
Tags_Will revert after reading the email              57.67
Tags_opp hangup                                      -45.88
Tags_switched off                                    -70.45
What is your current occupation_Unemployed            24.90
What is your current occupation_Working Professional  28.23
Last Activity_SMS Sent                                29.34
dtype: float64
```

Sorted order of features

```
Lead Source_Welingak Website                                  12
Lead Quality_Worst                                             9
Asymmetrique Activity Index_03.Low                             3
Tags_Already a student                                         8
Tags_Closed by Horizzon                                        1
Tags_Interested  in full time MBA                             11
Tags_Interested in other courses                               5
Tags_Lost to EINS                                              6
Tags_Not doing further education                               2
Tags_Ringing                                                  13
Tags_Will revert after reading the email                      14
Tags_opp hangup                                               15
Tags_switched off                                              0
What is your current occupation_Unemployed                    10
What is your current occupation_Working Professional           4
Last Activity_SMS Sent                                         7
dtype: int64
```

Plotting the relative coefficient variable values

Top 3 features that contribute to a lead conversion are,

	index	0
7	Tags_Lost to EINS	100.00
4	Tags_Closed by Horizzon	81.12
10	Tags_Will revert after reading the email	57.67

Concluding..

- All variables have p-value <0.05
- All the features have very low VIF values, meaning, there is hardly any Multicollinearity among the features.
- The overall accuracy of 0.9056 at a probability threshold of 0.33 on the test dataset is also very acceptable.

The conversion probability of a lead increases with increase in values of the following features in descending order

Features with Positive Coefficient Values
Tags_Lost to EINS
Tags_Closed by Horizzon
Tags_Will revert after reading the email
Lead Source_Welingak Website
Last Activity_SMS Sent
What is your current occupation_Working Professional
What is your current occupation_Unemployed

The conversion probability of a lead increases with decrease in values of the following features in descending order

Features with Negative Coefficient Values
Tags_switched off
Tags_Ringing
Tags_Already a student
Tags_Not doing further education
Lead Quality_Worst
Tags_opp hangup
Tags_Interested in full time MBA
Tags_Interested in other courses
Asymmetrique Activity Index_03.Low