Lead Scoring Case Study

By Mayuresh Madhukar Patankar

Problem statement

- An education company named X Education sells online courses to industry professionals. The typical lead conversion rate at X education is around 30%. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. The CEO has given a ballpark of the target lead conversion rate to be around 80%. The company wishes to identify the most potential leads, also known as ‘Hot Leads’ so that sales team will focus on communicating with the potential leads rather than making calls to everyone

Requirements

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps Need to Solve The Case Study

- ***Step1-Reading and Understanding Data***

- ***Step2-Data Cleaning***

- ***Step3-Exploratory Data Analysis***

- ***Step4-Data Preparation***

- ***Step5-Model Building***

- ***Step6-Step6-model evalutation train data set***

- ***Step7-conclusion***

Reading and Understanding Data

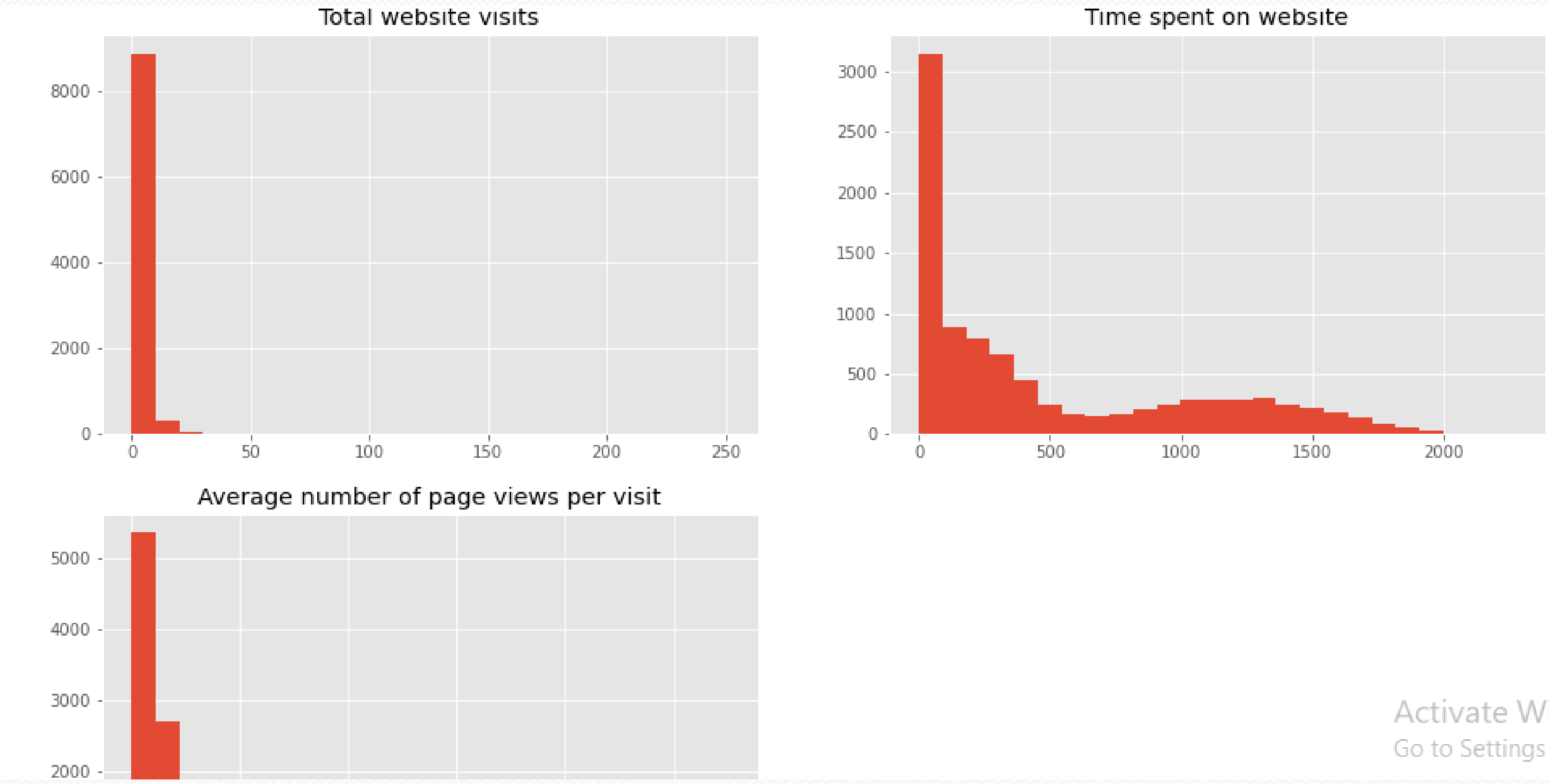
- Reading Data using pandas librerary.
- df = pd.read_csv('Leads.csv')
- Data shape=(9240, 37)
- There are no any duplicate file
- Data information-class 'pandas.core.frame.DataFrame'> RangeIndex: 9240 entries, 0 to 9239 Data columns (total 37 columns)

- Observations
- A large number of columns have null values. Those columns should ideally be dropped Prospect ID and Lead Number both serve the same purpose. They are both unique identifiers. We will drop Prospect ID Column names are just too long. We will modify the column names Few categorical columns have "Select" in their entries. Those select are essentially null values because Select appears when someone does not select anything from the dropdown

Data Cleaning

- # shorten column names
- # Select all non-numeric columns
- # select all the columns that have a "Select" entry
- # Calculate percentage of null values for each column
- Observation: As can be seen, there are quite a few columns with high number of missing data. Since there are no ways to get data back from reliable sources, we can drop all those columns that have missing values > 40%
- Drop columns that have null values > 40% or Sales generated columns
- # Calculate percentage of null values for each column
- Observations There are five columns that still have high null values: country, specialization, occupation, course_selection_reason, and city. We will look at them individually to see what can be done
- **combine low representing categories**
- **lead_origin column**
- **Handle Numerical columns**

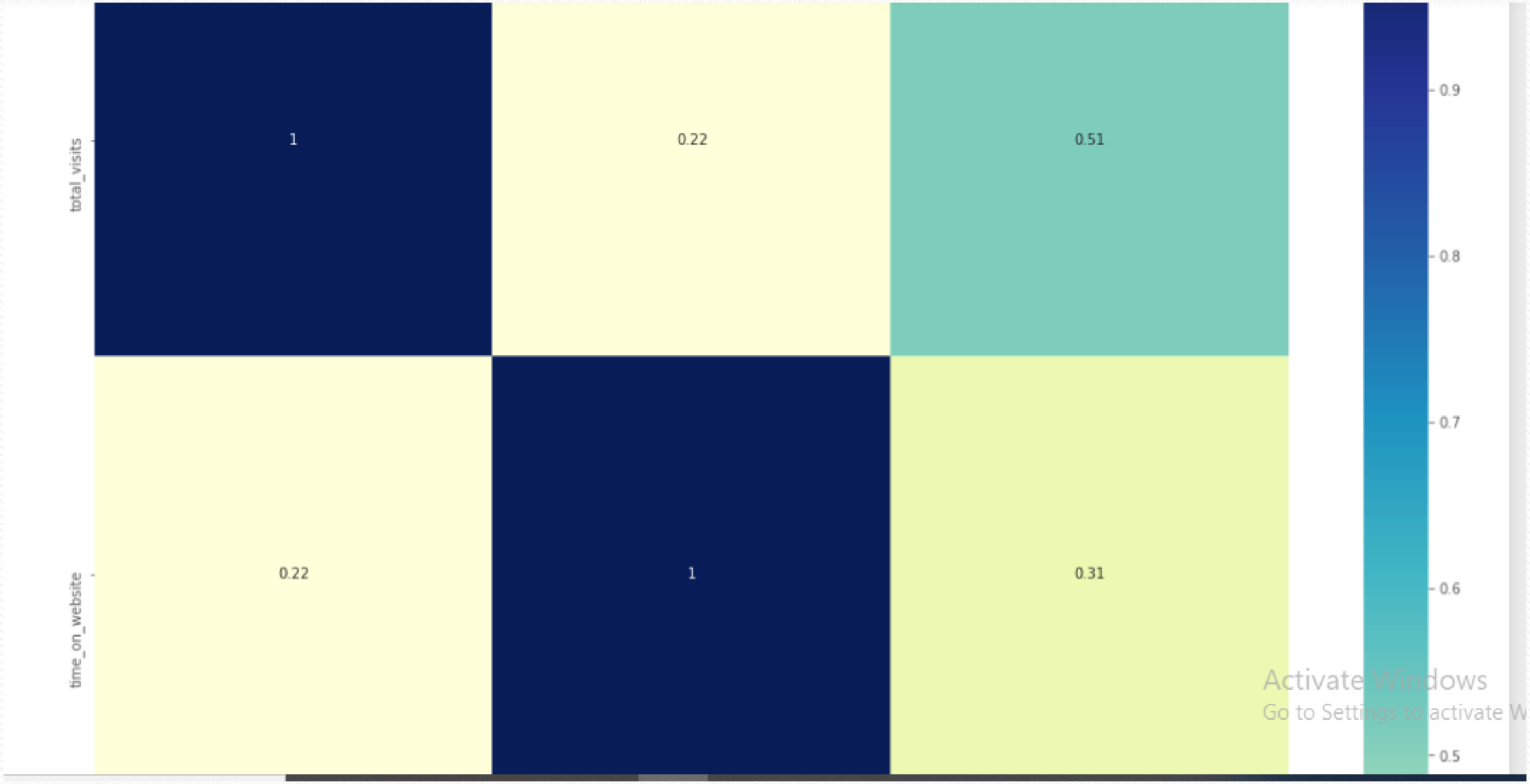
EDA


Total website visits

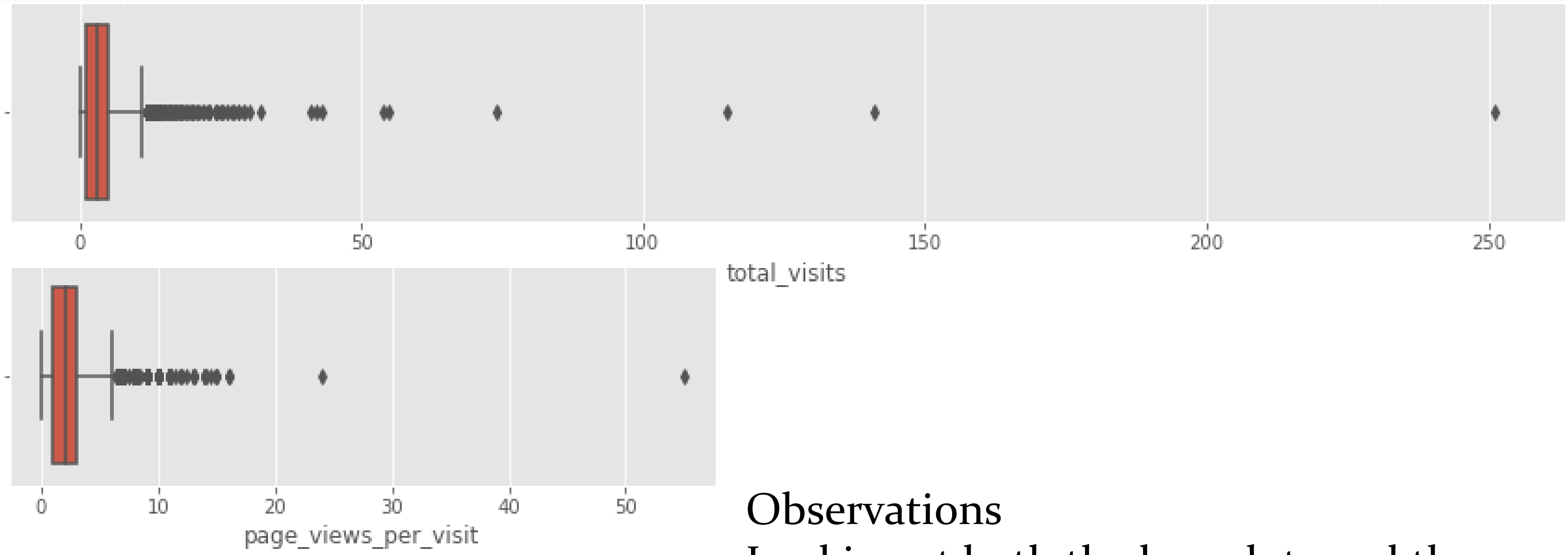
Time spent on website

Average number of page views per visit

Heatmap



Check for outliers

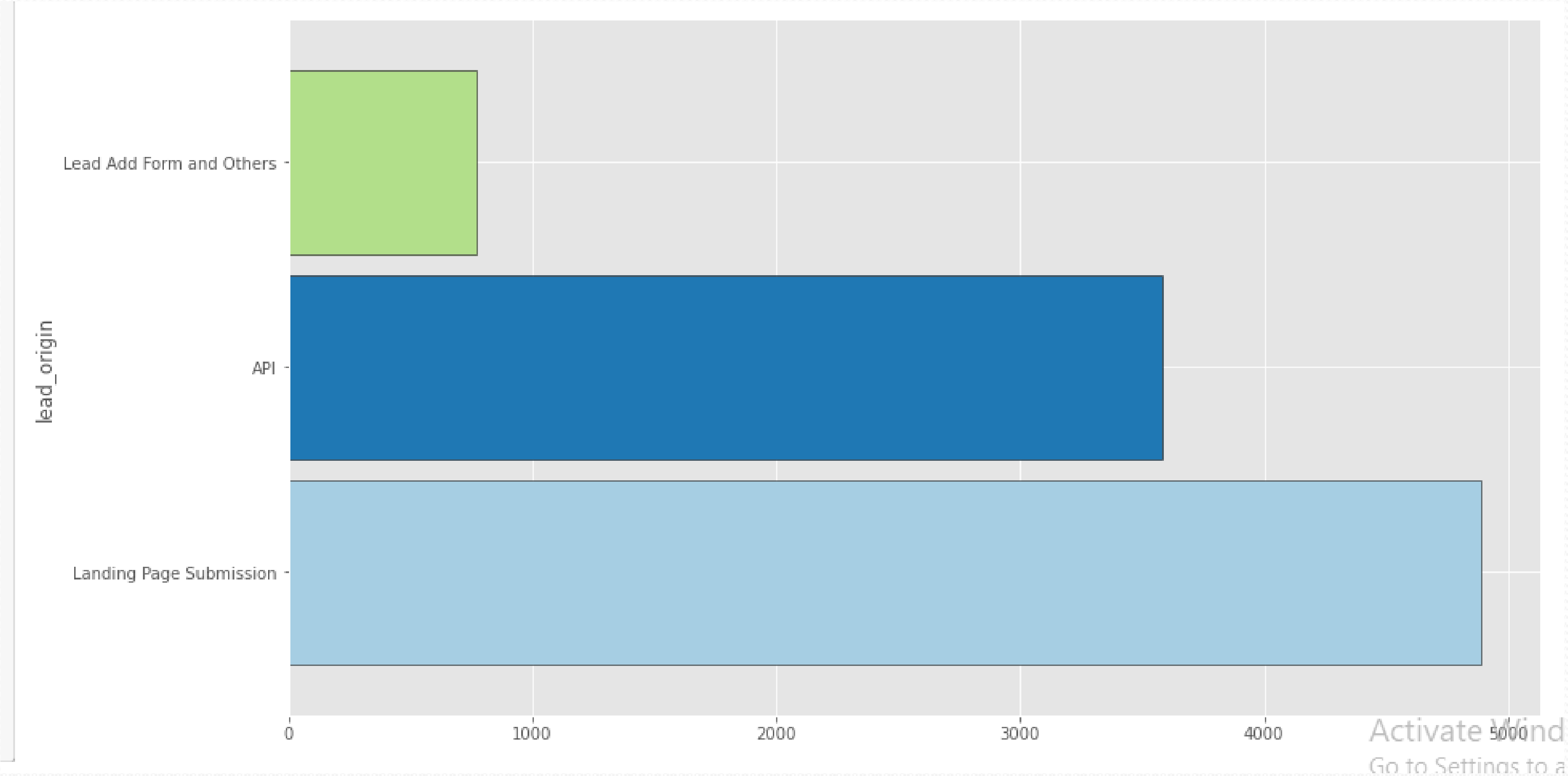


Observations

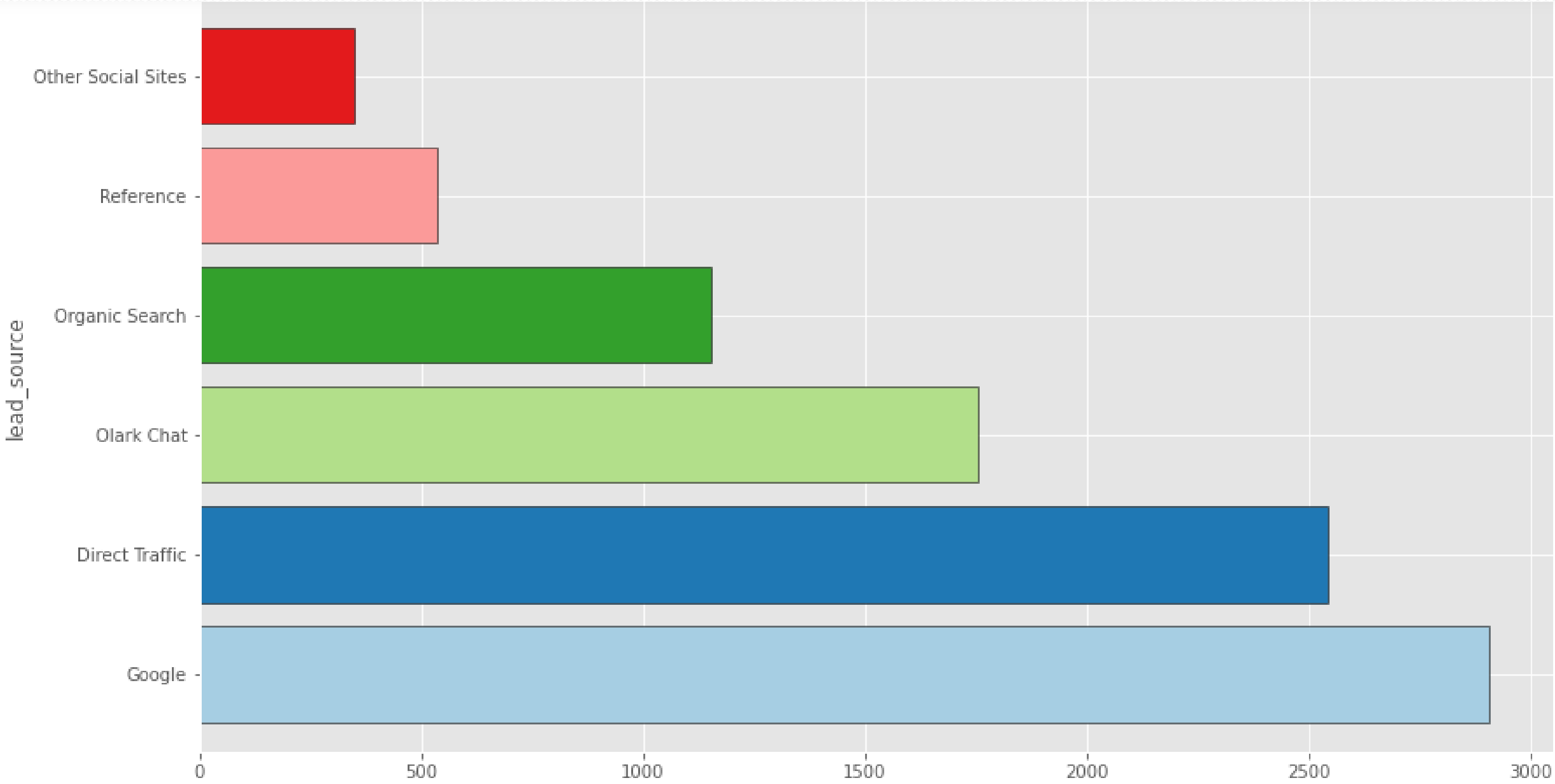
Looking at both the box plots and the statistics, there are upper bound outliers in both total_visits and page_views_per_visit columns. We can alsopercentile.

see that the data can be capped at 99

Categorical columns

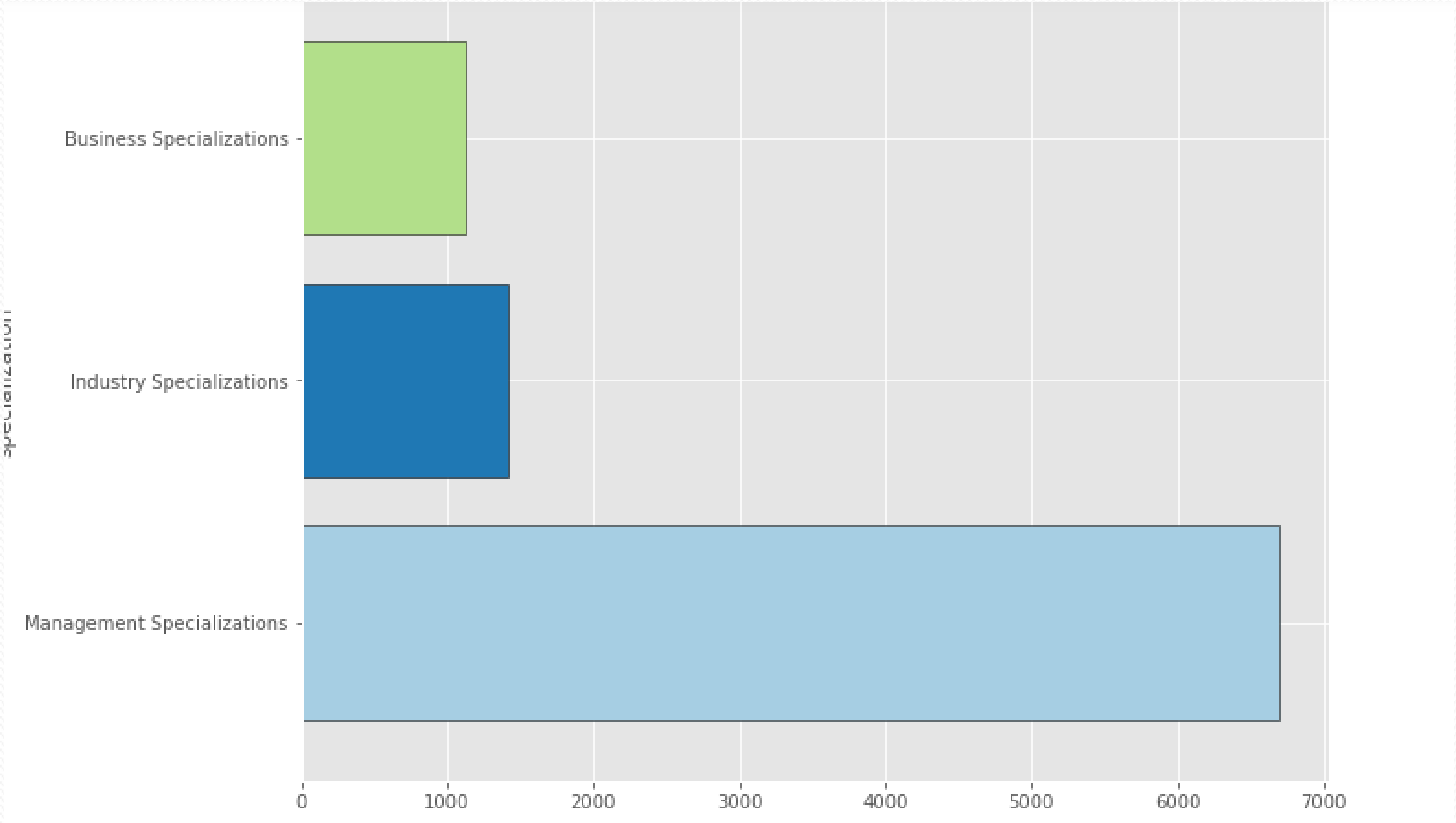


Lead Source

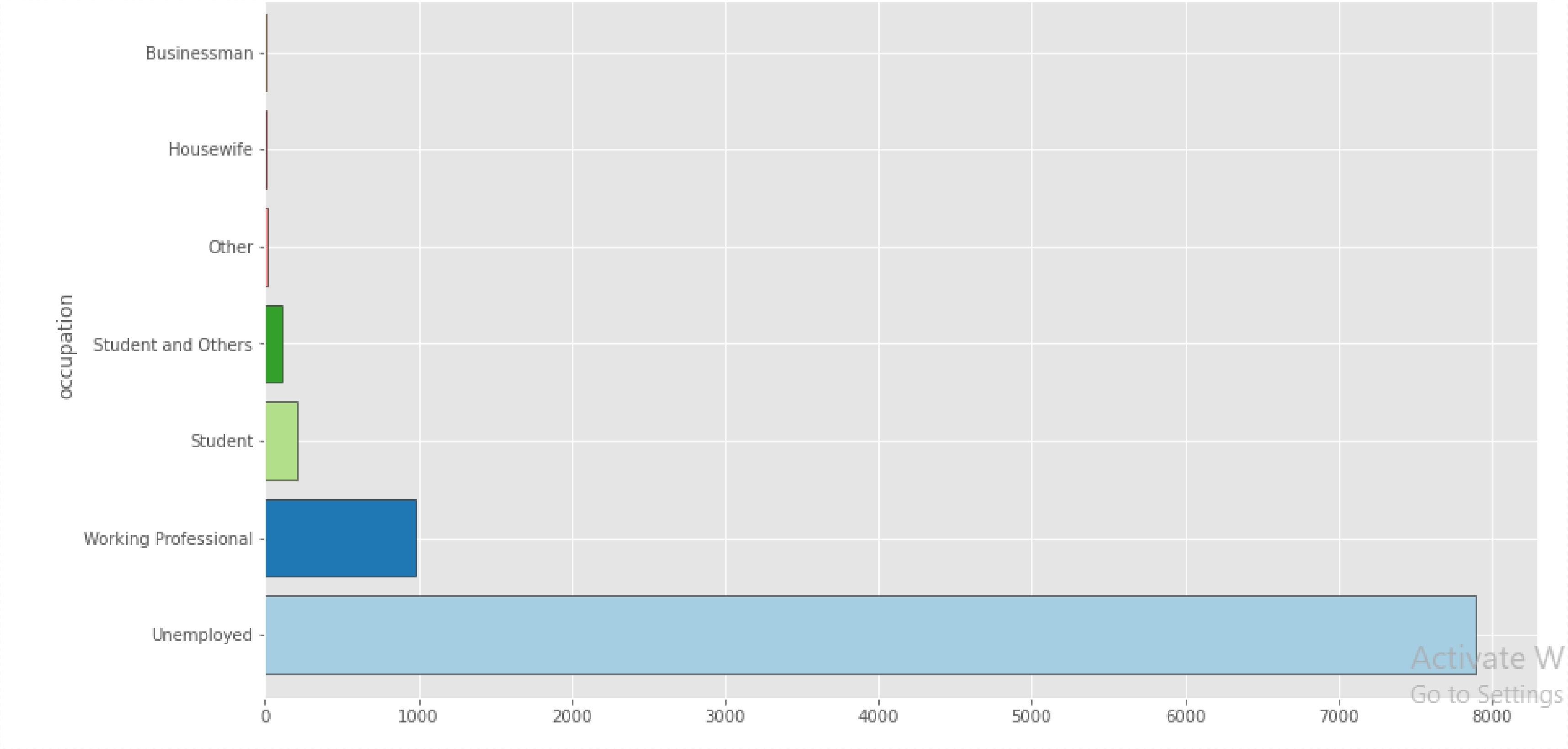


lead_source
Other Social Sites
Reference
Organic Search
Olark Chat
Direct Traffic
Google
0
500
1000
1500
2000
2500
3000

#Specialization

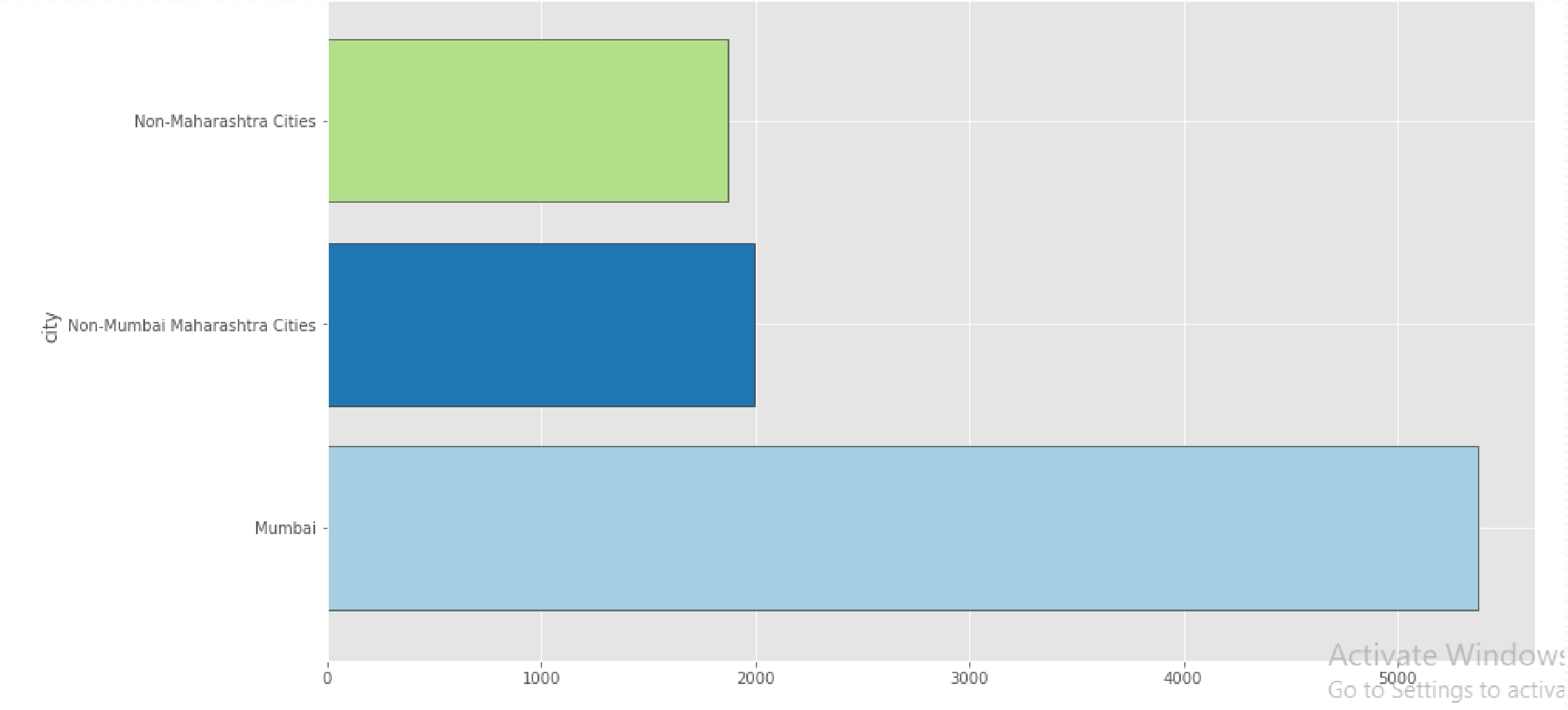


#Occupation



occupation
Businessman
Housewife
Other
Student and Others
Student
Working Professional
Unemployed
0
1000
2000
3000
4000
5000
6000
7000
8000

City



Mumbai in particular and Maharashtra in general dominates the lead. This is likely due to the fact that the courses are based in Mumbai

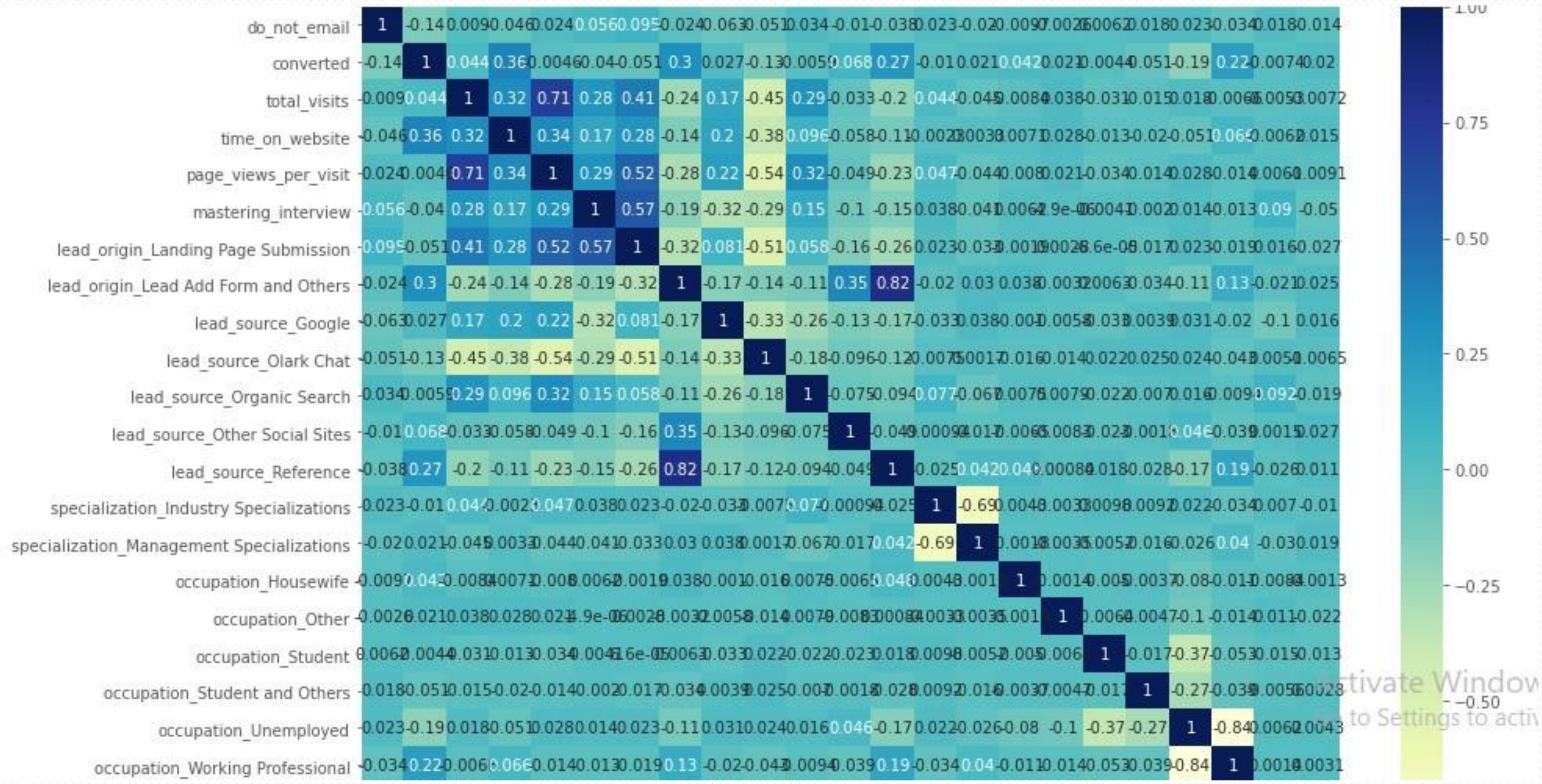
Data Preparation

- Converting Binary (Yes/No) to 0/1
- lead_number = 9240 lead_origin = 3 lead_source = 6 do_not_email = 2 specialization = 3 occupation = 7 city = 3 mastering_interview = 2
- We have two binary columns: do_not_email, mastering_interview
- Creating dummy variable for categorical columns Categorical columns are: lead_origin, lead_source, specialization, occupation, city
- # Dropping the columns for which dummies have been created
- capping at 99 percentile

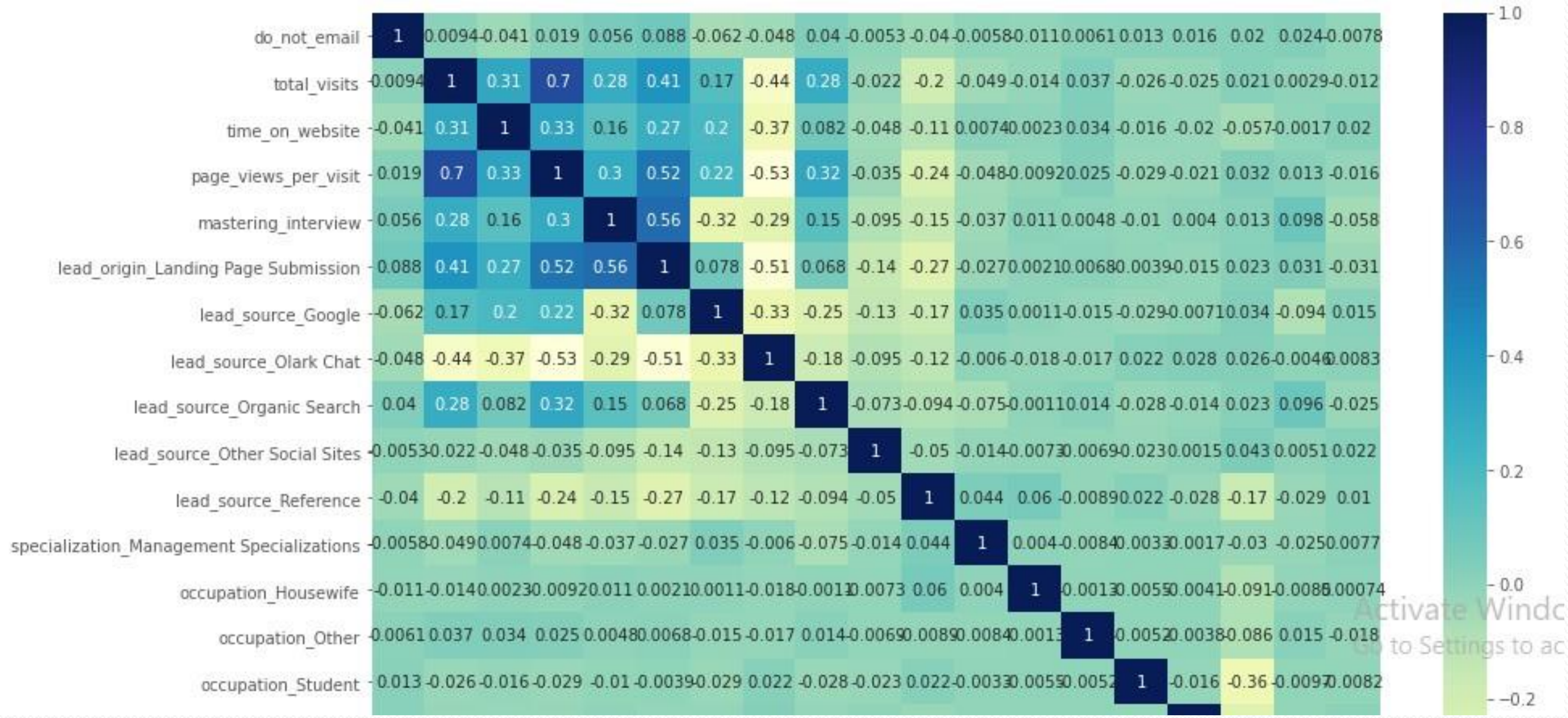
Test-Train Split

- # Putting feature variable to X
- # Putting response variable to y
- # Splitting the data into train and test
- Feature Scaling
- NOTE-The conversion rate is 38.5%
- see the correlation matrix
- Drop highly correlated dummy variables
- check the correlation matrix again

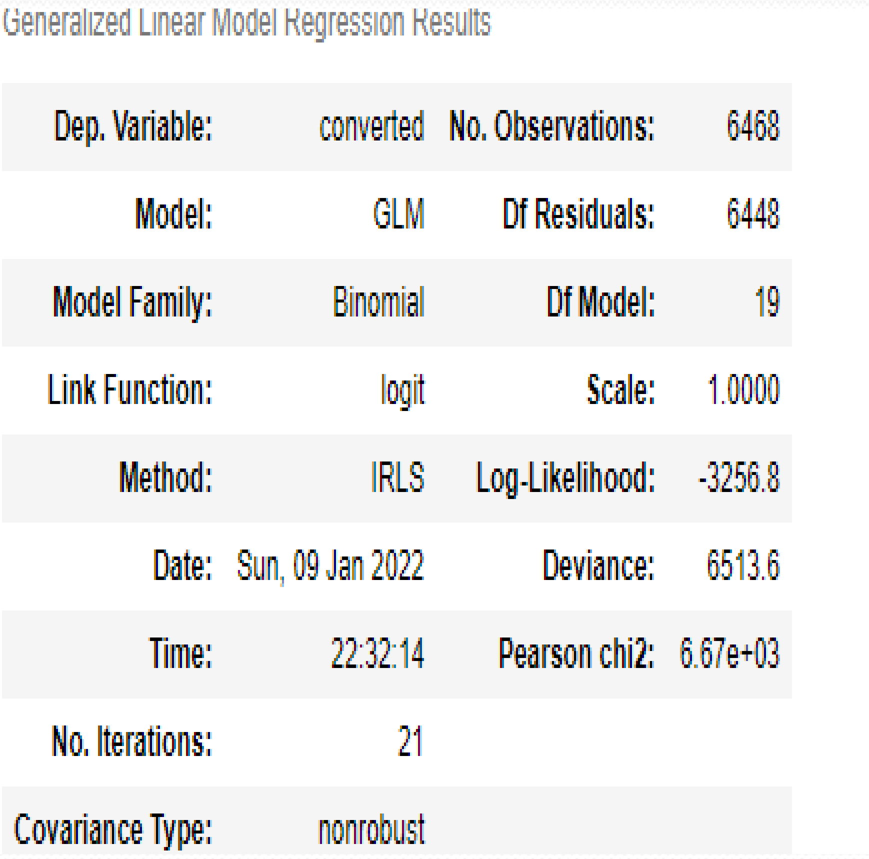
see the correlation matrix



check the correlation matrix again

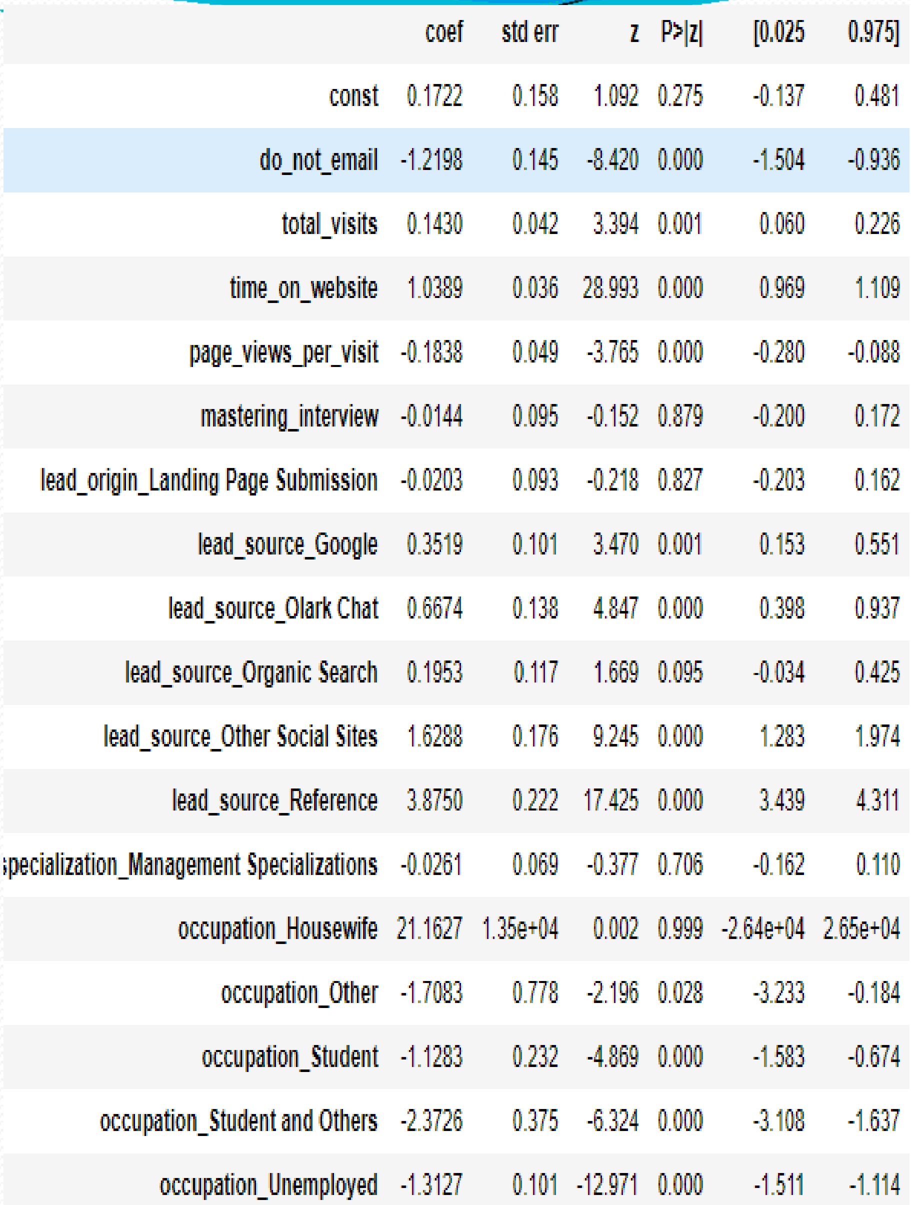


Model Building



Generalized Linear Model Regression Results

Dep. Variable:	converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6448
Model Family:	Binomial	Df Model:	19
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3256.8
Date:	Sun, 09 Jan 2022	Deviance:	6513.6
Time:	22:32:14	Pearson chi2:	6.67e+03
No. Iterations:	21		
Covariance Type:	nonrobust		



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1722	0.158	1.092	0.275	-0.137	0.481
do_not_email	-1.2198	0.145	-8.420	0.000	-1.504	-0.936
total_visits	0.1430	0.042	3.394	0.001	0.060	0.226
time_on_website	1.0389	0.036	28.993	0.000	0.969	1.109
page_views_per_visit	-0.1838	0.049	-3.765	0.000	-0.280	-0.088
mastering_interview	-0.0144	0.095	-0.152	0.879	-0.200	0.172
lead_origin_Landing Page Submission	-0.0203	0.093	-0.218	0.827	-0.203	0.162
lead_source_Google	0.3519	0.101	3.470	0.001	0.153	0.551
lead_source_Olark Chat	0.6674	0.138	4.847	0.000	0.398	0.937
lead_source_Organic Search	0.1953	0.117	1.669	0.095	-0.034	0.425
lead_source_Other Social Sites	1.6288	0.176	9.245	0.000	1.283	1.974
lead_source_Reference	3.8750	0.222	17.425	0.000	3.439	4.311
specialization_Management Specializations	-0.0261	0.069	-0.377	0.706	-0.162	0.110
occupation_Housewife	21.1627	1.35e+04	0.002	0.999	-2.64e+04	2.65e+04
occupation_Other	-1.7083	0.778	-2.196	0.028	-3.233	-0.184
occupation_Student	-1.1283	0.232	-4.869	0.000	-1.583	-0.674
occupation_Student and Others	-2.3726	0.375	-6.324	0.000	-3.108	-1.637
occupation_Unemployed	-1.3127	0.101	-12.971	0.000	-1.511	-1.114

Feature selection using RFE

- initiate logistic regression
- # initiate rfe
- # assign columns
- check what columns were not selected by Rfe
- Creating a function to get the column details
- # Do not email column
- details('Do Not Email')

- details('TotalVisits')
- and capping the column as mentioned earlier
- cap('TotalVisits')
- Lets look at Total Time Spent on Website column
- details('Total Time Spent on Website')
- details('Last Activity')

- details('Country')
- Observation-There are 27 percent null values in this column, and other values are mostly India, so this column would not be that useful, we could create a binary column using this like 'India' if there weren't so many null values here. but considering the null values, lets drop this column

Conclusion

- Conclusion: -
- The lead_score column can be used to identify the potential leads to focus first.
- Higher the score, higher are the chances for the lead to convert.
- In case, there are limited sales representatives, then the score cut-off should be higher to ensure a higher conversion probability people are contacted further to turn them into a potential customer.
- It is the same as increasing the precision value of the model by adjusting the cut-off point to a higher value.
- In case there are more resources available in the sales team (i.e., interns, etc.), then the score cut-off can be lowered.
- As there are more human resources, the company can afford a higher rate of False positives as it will increase the customer outreach and, in turn, increase the potential customer who will take the online courses.

- Thank You