Lead Scoring Case Study

Group Members –
1. Shubham Patil
2. Pius Lobo
3. Abhishek D

Introduction

This CASE STUDY aims to give you an idea of BUILDING a MACHINE LEARNING MODEL in a real business scenario. In this CASE STUDY, apart from applying the techniques that we have learned in the MACHINE LEARNING module, we will also develop a basic understanding of MODLE BUILDING in real business scenarios, how data is used and model is built in order to get a higher lead conversion.

Problem Statement

- X Education sells online courses to industry professionals.
- 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.

Data Understanding

This dataset has 2 files as explained below:

1. **‘Leads.csv'** contains all the information of many professionals who are interested in the courses land on their website and browse for courses.

2. **'Leads Data Dictionary.csv'** is data dictionary which describes the meaning of the variables.

Solution Methodology

- Data cleaning and data manipulation.

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains a large number of missing values and is not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

- EDA

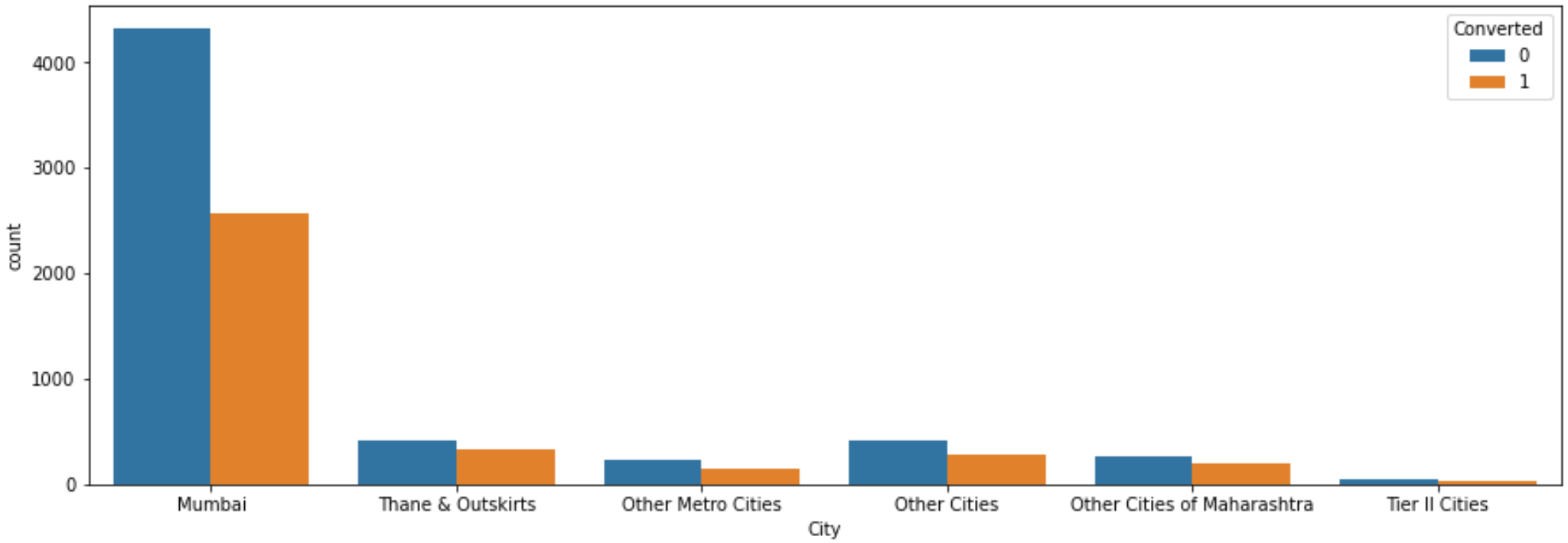
1. Univariate data analysis: value count, distribution of variables, etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression is used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

Data Manipulation

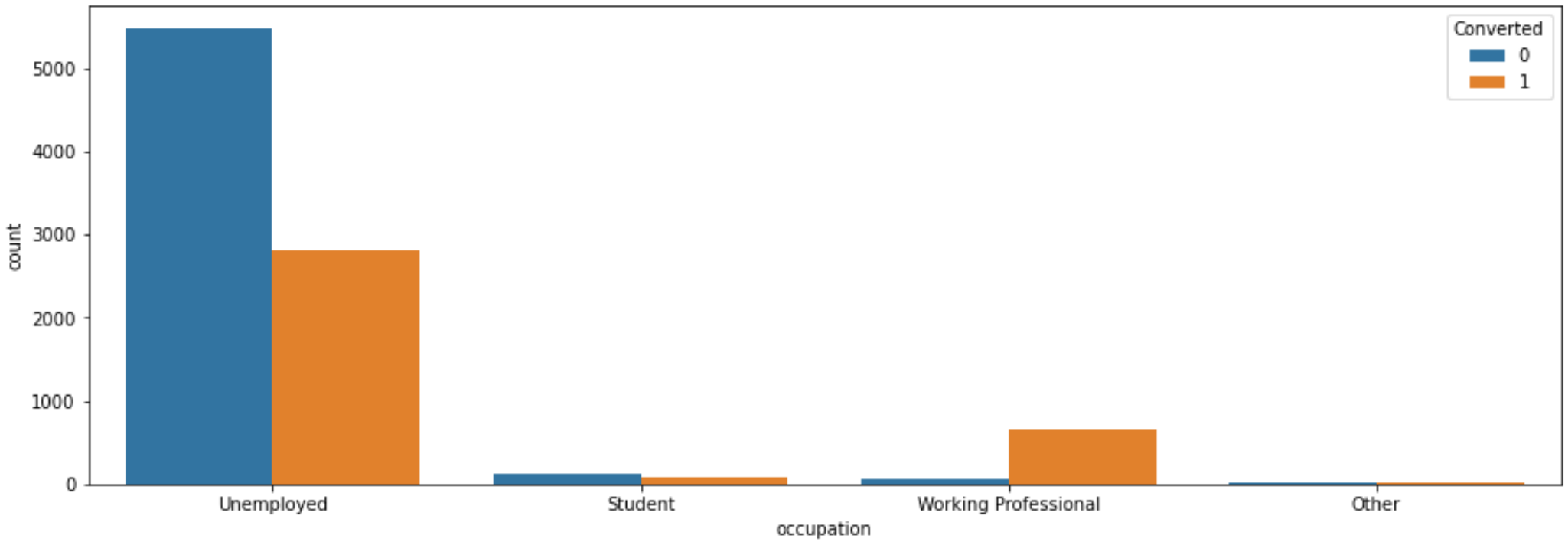
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply“, Chain Content", "Get updates on DM Content , I agree to pay the amount through, cheque" etc. have been dropped.
- Removing the "Prospect ID" and "Lead Number" which are not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are:
- "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement" etc.
- Dropping the columns having more than 35% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.

Analysis of ‘CITY’



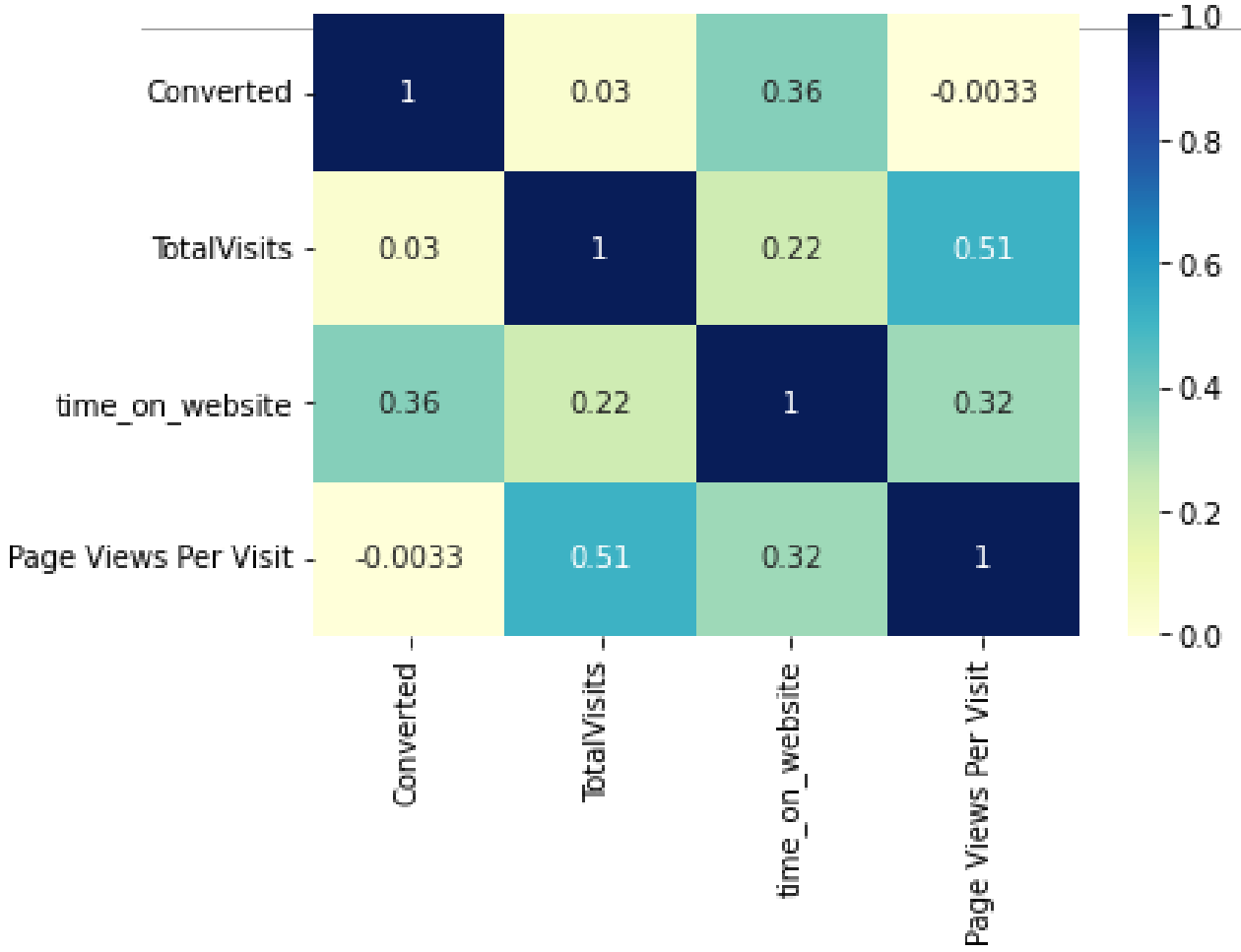
The plot shows the highest numbers who are interested from Mumbai compared to other cities

Analysis of ‘OCCUPATION’



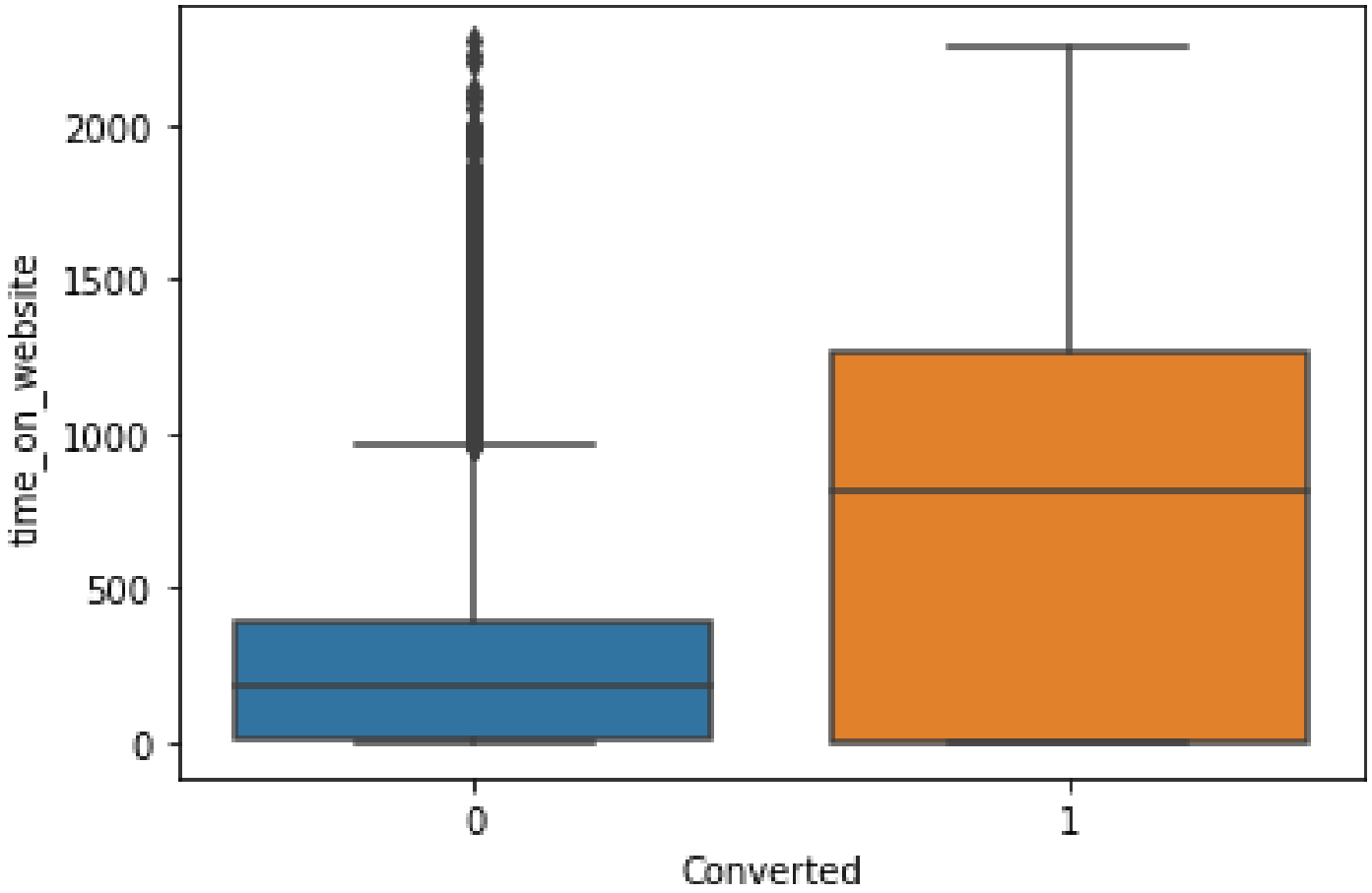
From this graph We can say that there is a positive response from the working professionals.

HEAT map



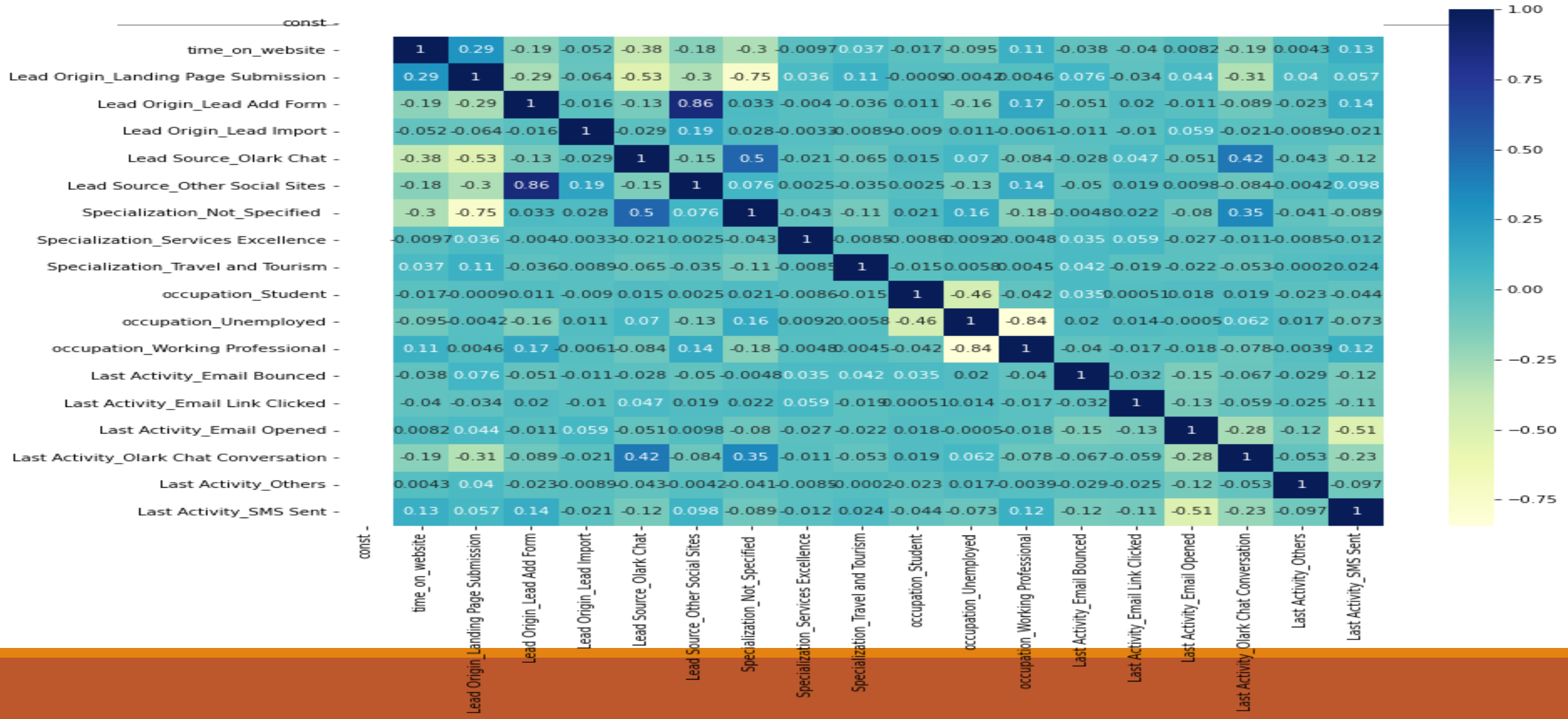
Looking at the heatmap we can say that Converted has good correlation with time_on_website.

Analysis of “time_on_website”



From plot we can see that average of who are Converted as yes are more.

This Heat map shows the correlation between variables



MODEL BUILDING

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6291
Model:	GLM	Df Residuals:	6272
Model Family:	Binomial	Df Model:	18
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2682.5
Date:	Mon, 02 Jan 2023	Deviance:	5365.0
Time:	23:51:22	Pearson chi2:	6.49e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3783
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4047	0.543	0.745	0.456	-0.660	1.469
time_on_website	1.0683	0.040	26.910	0.000	0.990	1.146
Lead Origin_Landing Page Submission	-0.9231	0.126	-7.326	0.000	-1.170	-0.676
Lead Origin_Lead Add Form	3.6553	0.339	10.785	0.000	2.991	4.320
Lead Origin_Lead Import	0.5472	0.591	0.926	0.354	-0.611	1.705
Lead Source_Olark Chat	1.0574	0.123	8.576	0.000	0.816	1.299
Lead Source_Other Social Sites	0.0368	0.277	0.133	0.894	-0.507	0.580
Specialization_Not_Specified	-1.0436	0.121	-8.635	0.000	-1.280	-0.807
Specialization_Services Excellence	0.5830	0.609	0.958	0.338	-0.610	1.776
Specialization_Travel and Tourism	-0.4636	0.236	-1.966	0.049	-0.926	-0.001
occupation_Student	-1.3398	0.577	-2.321	0.020	-2.471	-0.208
occupation_Unemployed	-1.6110	0.529	-3.044	0.002	-2.648	-0.574
occupation_Working Professional	0.9391	0.560	1.677	0.093	-0.158	2.036
Last Activity_Email Bounced	-1.3824	0.380	-3.638	0.000	-2.127	-0.638
Last Activity_Email Link Clicked	0.4473	0.235	1.902	0.057	-0.014	0.908
Last Activity_Email Opened	0.8315	0.121	6.870	0.000	0.594	1.069
Last Activity_Olark Chat Conversation	-0.7439	0.194	-3.825	0.000	-1.125	-0.363
Last Activity_Others	1.1739	0.237	4.947	0.000	0.709	1.639
Last Activity_SMS Sent	1.9671	0.124	15.822	0.000	1.723	2.211

This is the 1st model in which the **P** value of 'Lead Source_Other Social Sites' is high 0.894 so we dropped the variable. And build 2nd, 3rd models.

FINAL MODEL

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6291
Model:	GLM	Df Residuals:	6274
Model Family:	Binomial	Df Model:	16
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2710.0
Date:	Mon, 02 Jan 2023	Deviance:	5420.1
Time:	23:59:53	Pearson chi2:	6.49e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3728
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3484	0.524	-0.665	0.506	-1.375	0.678
time_on_website	1.0913	0.040	27.604	0.000	1.014	1.169
Lead Origin_Lead Add Form	4.1580	0.208	20.028	0.000	3.751	4.565
Lead Origin_Lead Import	1.1029	0.514	2.147	0.032	0.096	2.110
Lead Source_Olark Chat	1.3871	0.114	12.134	0.000	1.163	1.611
Specialization_Not_Specified	-0.4315	0.086	-5.014	0.000	-0.600	-0.263
Specialization_Services Excellence	0.5959	0.603	0.989	0.323	-0.585	1.777
Specialization_Travel and Tourism	-0.4864	0.233	-2.084	0.037	-0.944	-0.029
occupation_Student	-1.4329	0.567	-2.528	0.011	-2.544	-0.322
occupation_Unemployed	-1.6721	0.520	-3.217	0.001	-2.691	-0.653
occupation_Working Professional	0.9474	0.550	1.722	0.085	-0.131	2.026
Last Activity_Email Bounced	-1.4879	0.379	-3.929	0.000	-2.230	-0.746
Last Activity_Email Link Clicked	0.4520	0.234	1.934	0.053	-0.006	0.910
Last Activity_Email Opened	0.8315	0.121	6.899	0.000	0.595	1.068
Last Activity_Olark Chat Conversation	-0.6984	0.192	-3.633	0.000	-1.075	-0.322
Last Activity_Others	1.1782	0.236	4.985	0.000	0.715	1.641
Last Activity_SMS Sent	1.9393	0.124	15.691	0.000	1.697	2.182

This is the final model in which the **P** value of the variable are near zero

MODEL BUILDING

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

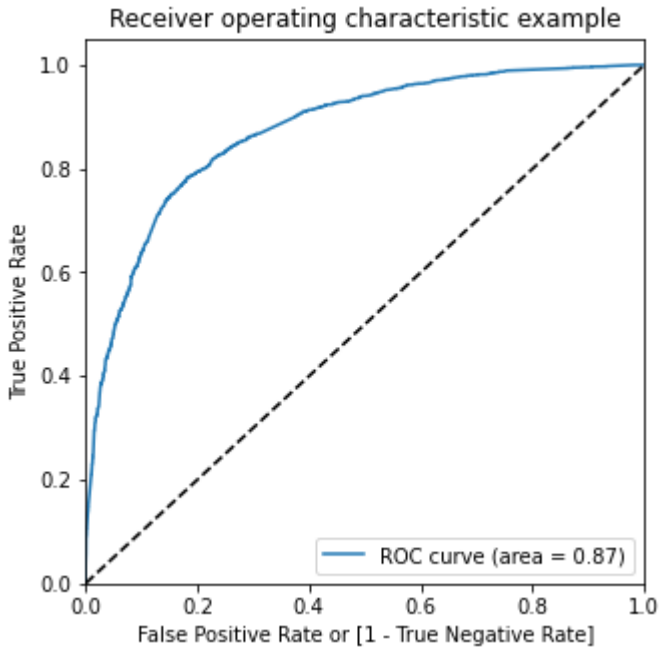
VIF Check

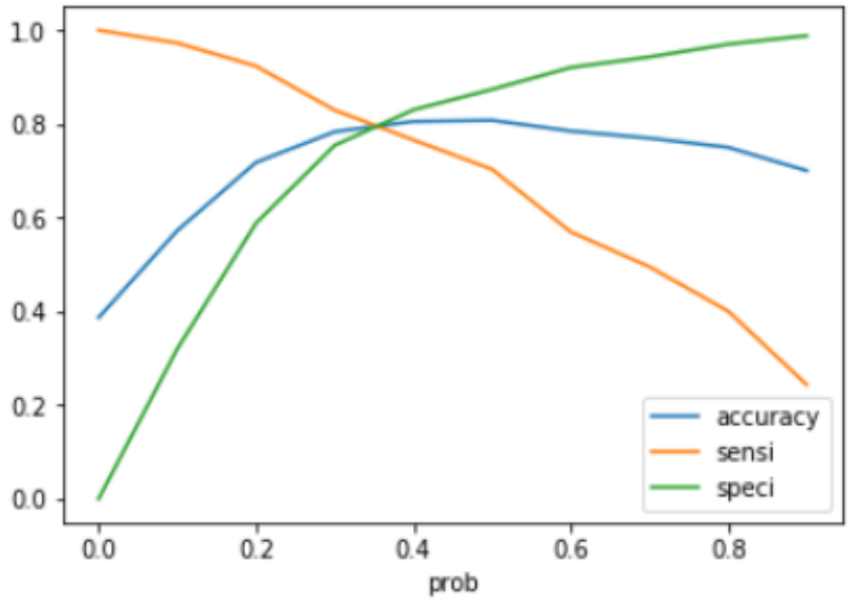
	Features	VIF
7	occupation_Unemployed	7.52
11	Last Activity_Email Opened	4.00
14	Last Activity_SMS Sent	3.40
4	Specialization_Not_Specified	2.32
12	Last Activity_Olark Chat Conversation	2.23
3	Lead Source_Olark Chat	2.11
8	occupation_Working Professional	1.60
0	time_on_website	1.32
1	Lead Origin_Lead Add Form	1.28
9	Last Activity_Email Bounced	1.28
10	Last Activity_Email Link Clicked	1.23
13	Last Activity_Others	1.17
6	occupation_Student	1.16
5	Specialization_Travel and Tourism	1.04
2	Lead Origin_Lead Import	1.02

VIF values are checked to confirm the multicollinearity

And it seems to be okay all the vif values are under limit.

ROC Curve





- ROC curve plot u to show the diagnostic ability of binary classifiers.
- From the second graph, it is clear that optimal cut-off point is 0.35.

Conclusion

The following factors affected potential purchasers the most (in descending order):

1. The total time spent on the Website.
2. Reason of course selection
3. When their current occupation is as a working professional.
4. When the lead source was:
    1. Google
    2. Direct traffic
    3. Organic search
    4. Welingar website
5. When the last activity was:
    1. SMS
    2. Olark chat conversation

Thank You