Case Study for Logistic Regression

Submitted by: Nikhil M Kanchan
Nikhil Kumar

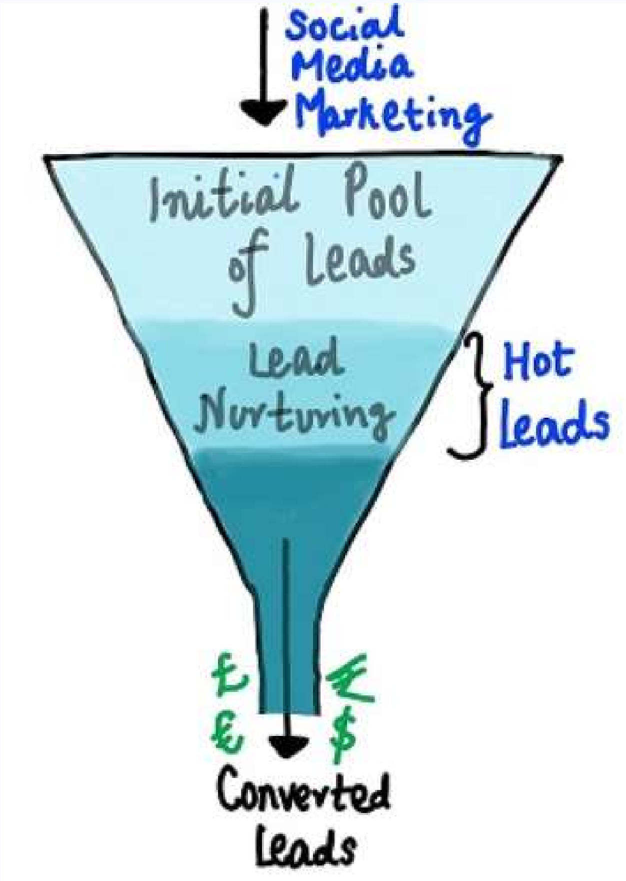
OUTLINE

1. Objective
2. Problem Statement
3. Modelling Approach
4. Modelling Approach - Detailed
5. The Model
6. Inferences for the Business
7. Interesting Visuals

OBJECTIVE

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.

Through various channels, the company receives lead – various attributes of people who showed initial interest or enquired about the course. This is a bulk data that has a very low conversion rate. Based on the historic data of leads provided by the company, the objective here is to define a process to assign a Lead Score to all the leads in a way that higher Lead Score is assigned to a lead that has high probability of conversion. The company also needs suggestions on various deviations it can take based on the Target it has for the quarter.



Problem Statement

- The leads received on a daily basis is structured but high in number due to which only some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- Efforts of sales team are not channelized as they work on bulk number of leads.
- X Education wants to make the process to identify leads more efficient. It wishes to identify the most potential leads, also known as ‘Hot Leads’. With the improved process the lead conversion rate should go up and hence the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
- The above mentioned solution needs to be in form of the logistic binary model that can identify probability of lead conversion and assign a Lead Score such that higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The target conversion rate set by X Education is 80%.
- Variations in the model is required to quickly increase or decrease the number of Hot Leads (probable conversions) based on quarterly target and current conversion status.

Modelling Approach

- Data sourcing
- Data preparation (Cleaning, Imputing and Univariate & multivariate analysis)
- Data Split (Train and Test)
- Data Modelling
- Model Evaluation
- Model Metrices

MODELLING APPROACH - DETAILED

- Sourcing of data from the CSV file provided by X Education.
- Upon importing the data, it was analyzed to identify firstly the variables of least importance for modelling and then EDA was performed on the data.
- During EDA, the following was performed:-
  - Renamed the columns to python standards
  - Identifying columns with Null
  - Define approach to handle outliers for numeric variables
  - Imputing data for null and outlier values in numeric variables based on details analysis of data
  - Approach to bin numeric data for time related variables to make it a categorical variable was rejected as it would impact the accuracy of model. Using this approach we could have categorized info like time spent on website into values such as High, Medium, Low. The trade-off here was data and accuracy.
  - Handling nulls in categorical variables

MODELLING APPROACH – DETAILED ...CONTD

- For categorical variables, Re-binning the data into lesser number of categories where the count in a given category was very less. This approach reduced the number of dummy variable columns to be defined, and hence the complexity of model.
- Dropping the columns with high imbalance. This was required as there were columns with imbalance of more than 95%:5%. Such data provides a incorrect picture to the model and can work as a redundant variable in the model.
- Creation of Dummy Variables for the categorical variables. For a categorical variable with 'N' distinct values 'N-1' dummy variables are created and hence the approach of re-binning reduced the number of Dummy Variables. This re-binning was done in a way that Others (values of like nulls, Others, Unknown and also values with very less coverage) are always assigned a value of 0. Hence, 'drop_first = True' parameter always dropped this column and the columns with definitive values were retained.

- Splitting the data into train & test in 70%-30% ratio
- Using Standard Scaler, scale the train dataset for the numerical – continuous columns
- Creating the initial model using the RFE approach with 13 variables
- Removal of variables with high P-value or high VIF as these variables are either less significant or have correlation with other variables

MODELLING APPROACH – DETAILED ...CONTD

- Creation of Confusion Matrix for the train dataset
- Evaluation of model on train dataset using various metrices – accuracy, sensitivity, specificity, precision & recall.
- Defining output dataset that provides a unique lead ID along with lead score in the range of 0-100 along with other required data
- Identifying the best fit metric for our data – sensitivity and specificity
- Creating the ROC curve with the final model to see its area
- Finding the optimal cut-off of probability above which the lead should be considered as hot lead
- Executing the model on test dataset
- Determine various metrices for the test result
- Assigning lead score against lead ID

THE MODEL

The sigmoid function for the multivariate logistic regression is defined as:-

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\ldots+\beta_nX_n)}}$$

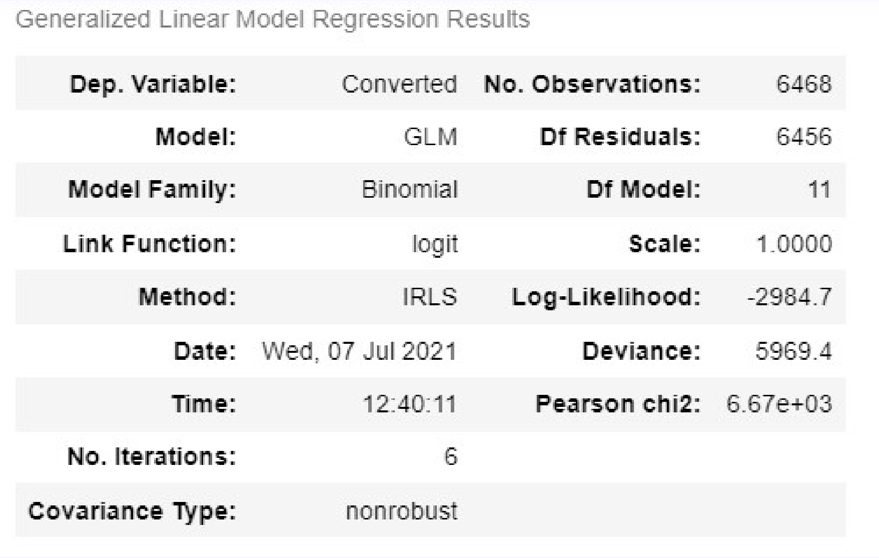
Here, $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots. + \beta_nX_n$

For our Lead Scoring project for X Education, this Y can be defined as follows:-

Y = 1.6764 + 1.1084 * Time_Spent_Website + 0.2877 * Occupation -3.4205 * Origin_lps -3.2979 * Origin_api -0.2833 * Source_dirTraffic + 0.936 * Source_chat + 0.5179 * LA_email + 1.5462 * LA_sms -1.0314 * LA_chat -0.7249 * LA_cnv_lead -1.5505 * LA_email_Bounce

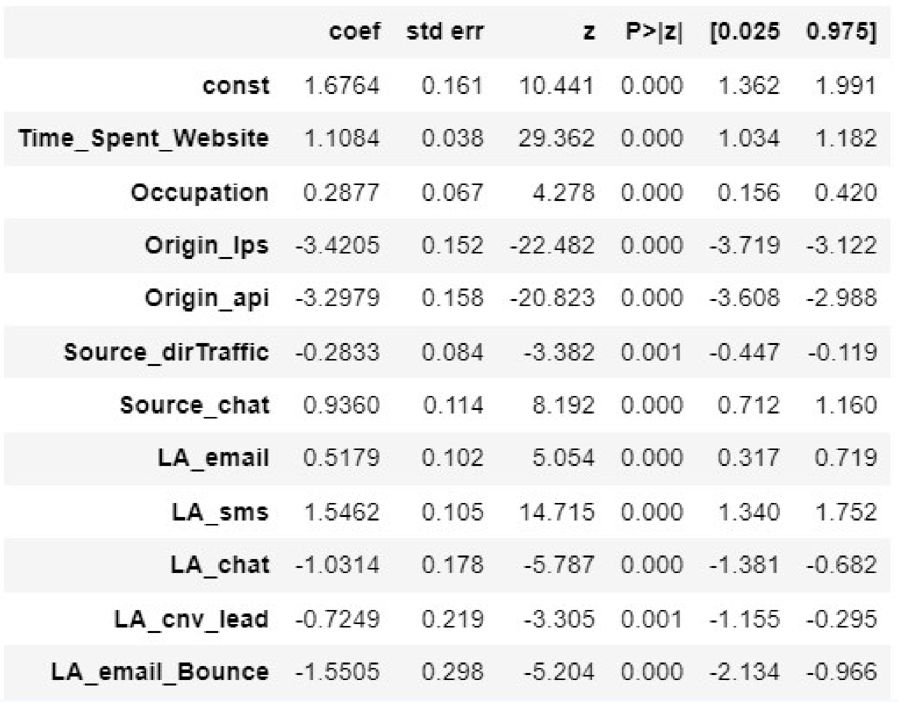
THE MODEL

….CONTD



Generalized Linear Model Regression Results

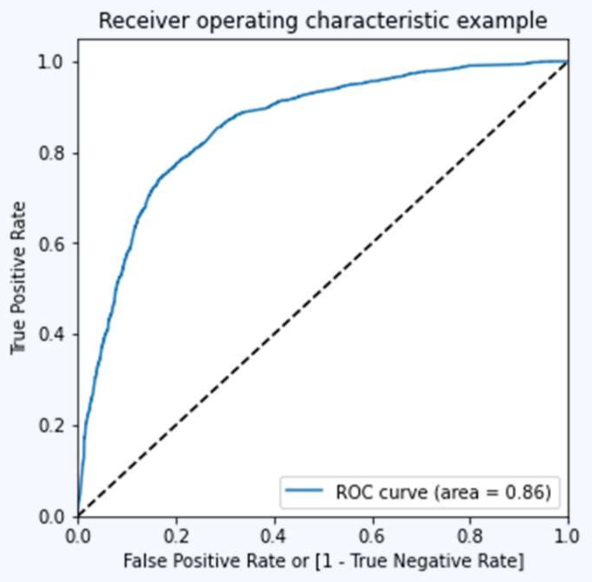
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6456
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2984.7
Date:	Wed, 07 Jul 2021	**Deviance:**	5969.4
Time:	12:40:11	**Pearson chi2:**	6.67e+03
No. Iterations:	6		
Covariance Type:	nonrobust		



	coef	std err	z	P>|z|	[0.025	0.975]
const	1.6764	0.161	10.441	0.000	1.362	1.991
Time_Spent_Website	1.1084	0.038	29.362	0.000	1.034	1.182
Occupation	0.2877	0.067	4.278	0.000	0.156	0.420
Origin_lps	-3.4205	0.152	-22.482	0.000	-3.719	-3.122
Origin_api	-3.2979	0.158	-20.823	0.000	-3.608	-2.988
Source_dirTraffic	-0.2833	0.084	-3.382	0.001	-0.447	-0.119
Source_chat	0.9360	0.114	8.192	0.000	0.712	1.160
LA_email	0.5179	0.102	5.054	0.000	0.317	0.719
LA_sms	1.5462	0.105	14.715	0.000	1.340	1.752
LA_chat	-1.0314	0.178	-5.787	0.000	-1.381	-0.682
LA_cnv_lead	-0.7249	0.219	-3.305	0.001	-1.155	-0.295
LA_email_Bounce	-1.5505	0.298	-5.204	0.000	-2.134	-0.966

THE MODEL

….STATS

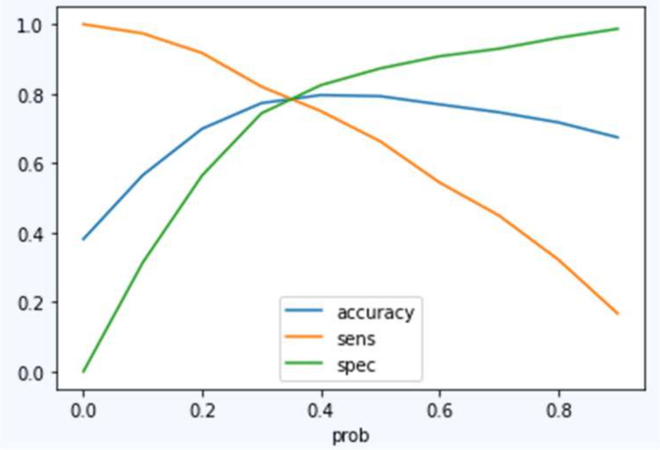


ROC Curve

- It shows the tradeoff between sensitivity and specificity
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

Probability Cut-off

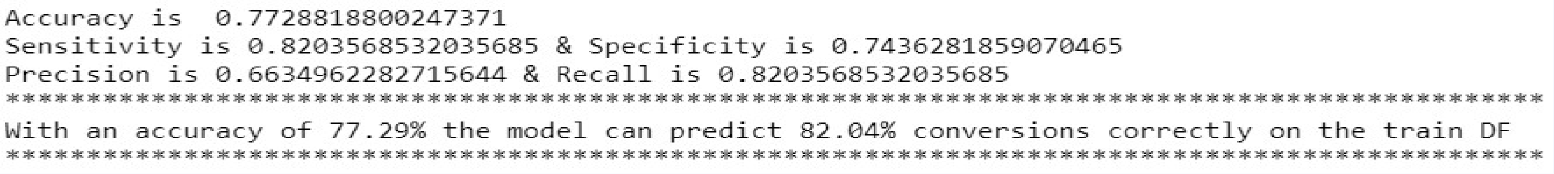
Visualization of accuracy, sensitivity & Specificity at different values of probability cut-off



THE MODEL

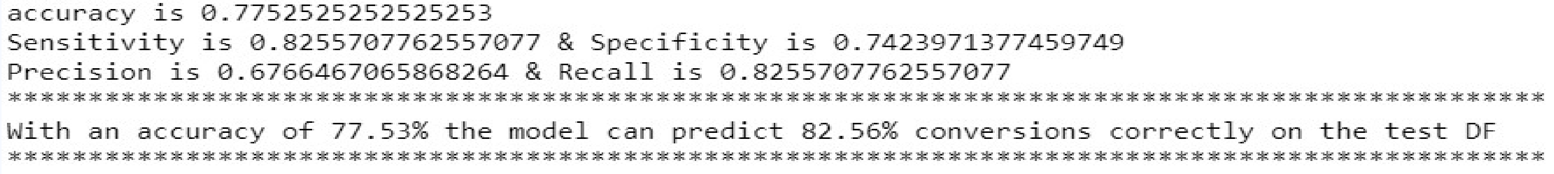
….STATS

Train Data



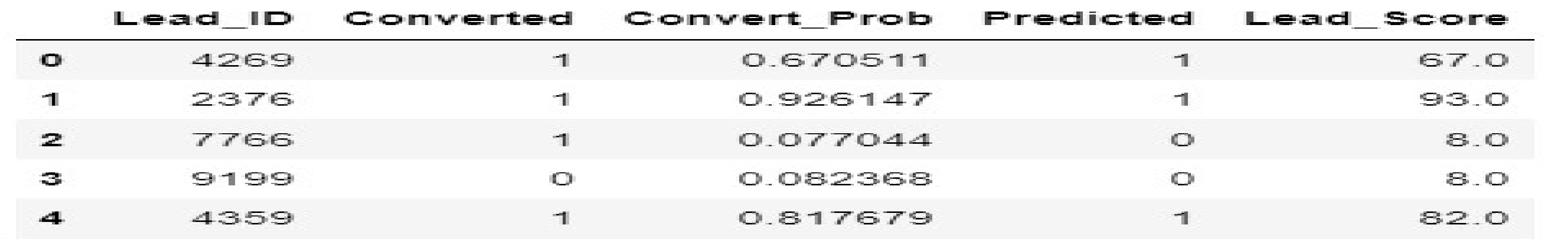
```
Accuracy is  0.7728818800247371
Sensitivity is 0.8203568532035685 & Specificity is 0.7436281859070465
Precision is 0.6634962282715644 & Recall is 0.8203568532035685
***********************************************************************************************
With an accuracy of 77.29% the model can predict 82.04% conversions correctly on the train DF
***********************************************************************************************
```

Test Data



```
accuracy is 0.7752525252525253
Sensitivity is 0.8255707762557077 & Specificity is 0.7423971377459749
Precision is 0.6766467065868264 & Recall is 0.8255707762557077
***********************************************************************************************
With an accuracy of 77.53% the model can predict 82.56% conversions correctly on the test DF
***********************************************************************************************
```

Output Data



	Lead_ID	Converted	Convert_Prob	Predicted	Lead_Score
0	4269	1	0.670511	1	67.0
1	2376	1	0.926147	1	93.0
2	7766	1	0.077044	0	8.0
3	9199	0	0.082368	0	8.0
4	4359	1	0.817679	1	82.0

INFERENCES FOR THE BUSINESS

The most important variables affecting the chances of Lead Conversion are as follows. Higher their value, better the chance of conversion:-

- Last Activity (email or chat)
- Time Spent on Website
- Source (chat)

Also, the following variables impact the chances of Lead Conversion negatively hence lower their value better the chance of Conversion:-

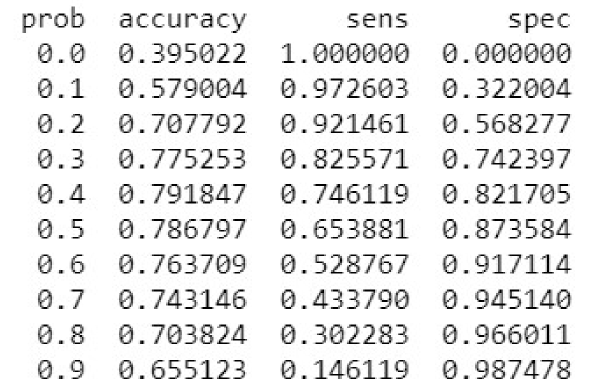
- Lead Origin Type of Landing Page Submission or API
- Last Activity as email bounce or chat
- Lead Source of Direct Traffic

INFERENCES FOR THE BUSINESS

The suggestion for business is to use the Model with conversion probability Cut-off as 0.3 or Lead Score >= 30.

The time of year when the company has more interns to work on lead, this cut-off can be changed to 0.2 (Lead Score of >= 20). This way the number of Leads to be contacted will be increased and so is the Sensitivity.

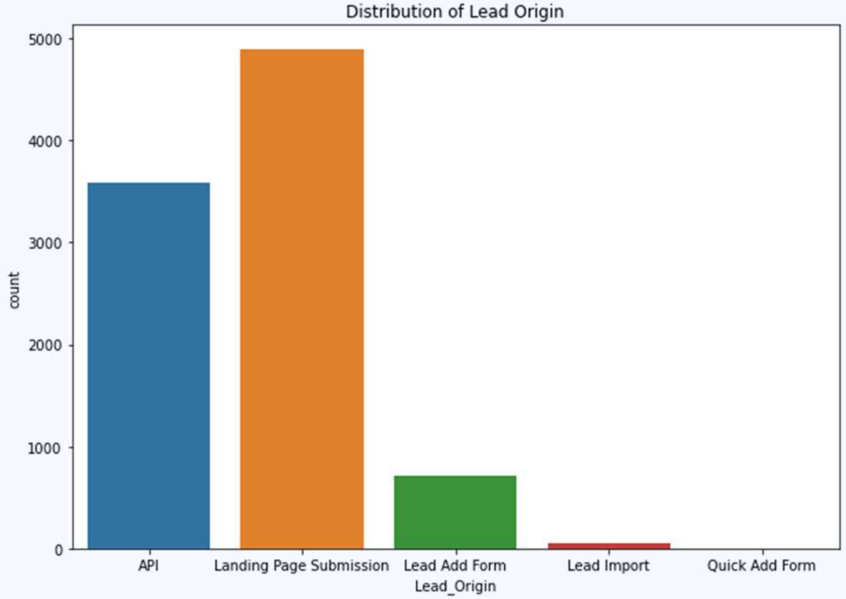
When the quarterly target is met and business wants to use the workforce for other tasks, the cut-off can be made stringent at 0.4. This will reduce the Sensitivity but the conversion ratio (Accuracy will increase).

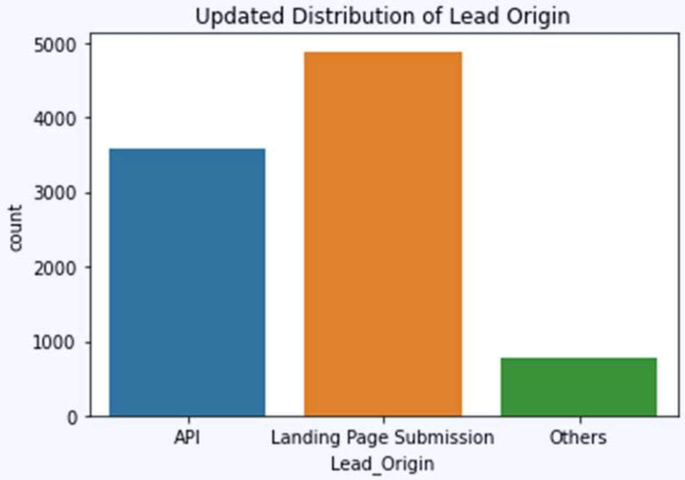


```
prob  accuracy      sens      spec
 0.0  0.395022  1.000000  0.000000
 0.1  0.579004  0.972603  0.322004
 0.2  0.707792  0.921461  0.568277
 0.3  0.775253  0.825571  0.742397
 0.4  0.791847  0.746119  0.821705
 0.5  0.786797  0.653881  0.873584
 0.6  0.763709  0.528767  0.917114
 0.7  0.743146  0.433790  0.945140
 0.8  0.703824  0.302283  0.966011
 0.9  0.655123  0.146119  0.987478
```

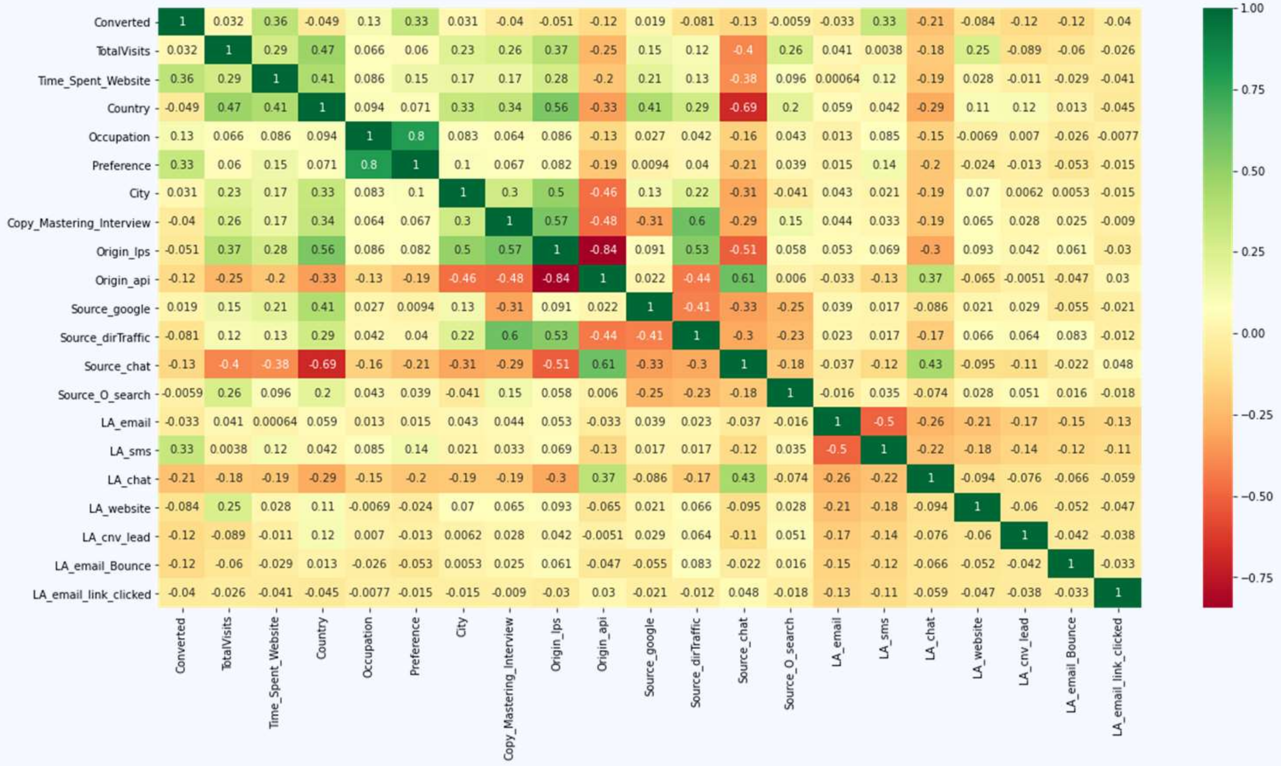
INTERESTING VISUALS

Many variables had multiple values some of which had very low occurrence. Clubbing them as others reduced the number of values (and hence the Dummy Variables) and provided significant numbers for Model building. One such example shown above.





INTERESTING VISUALS



The heat map suggested correlation among various variables. Though we processed with the model with such variables, but most of them got dropped as the model evolved except for the ones with lower correlation.

Thank You