LEAD SCORE CASE STUDY

JITHIN PRAKASH K

ABHISHEK BHATTACHARYA

Introduction

This case study is performed in accordance with the data provided for an education company called X Education. The case study will help to analyze and filter the probable professional (leads) who can opt for a course in X education and become customers. There are various parameters which came into play during this analysis and with the help of statistical and modelling techniques we have tried to get a good amount of insight to find the potential leads who can become customers.

Since the dataset provided for contains numerous factors , the objective is to consider those variables which has a direct impact on the application of the customer also consider the variables with a correct amount of data.

The whole study has been done on the google colab IDE using Python coding utilizing the numpy, pandas, matplotlib, seaborn, stats models and sklearn libraries.

Approach

The case study is carried out by keeping several things in mind which includes different considerations like analyzing the missing values, imputation method with reasons to replace the missing values, plotting the data to check the correlations, visualizing the binary, numeric and categorical features.

The analysis has been done on the dataset by imputing/removing data for the sake of improving efficiency of data prediction. Identifying the variables as Categorical and Numerical and then deducing their correlation is also a part of this study. The variance is checked for the features and eliminated when found low. The analysis also includes visualization and model building techniques such as splitting data into train and test, feature elimination, logistic regression, plotting ROC curve, accuracy measures etc.

Problem Statement

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.

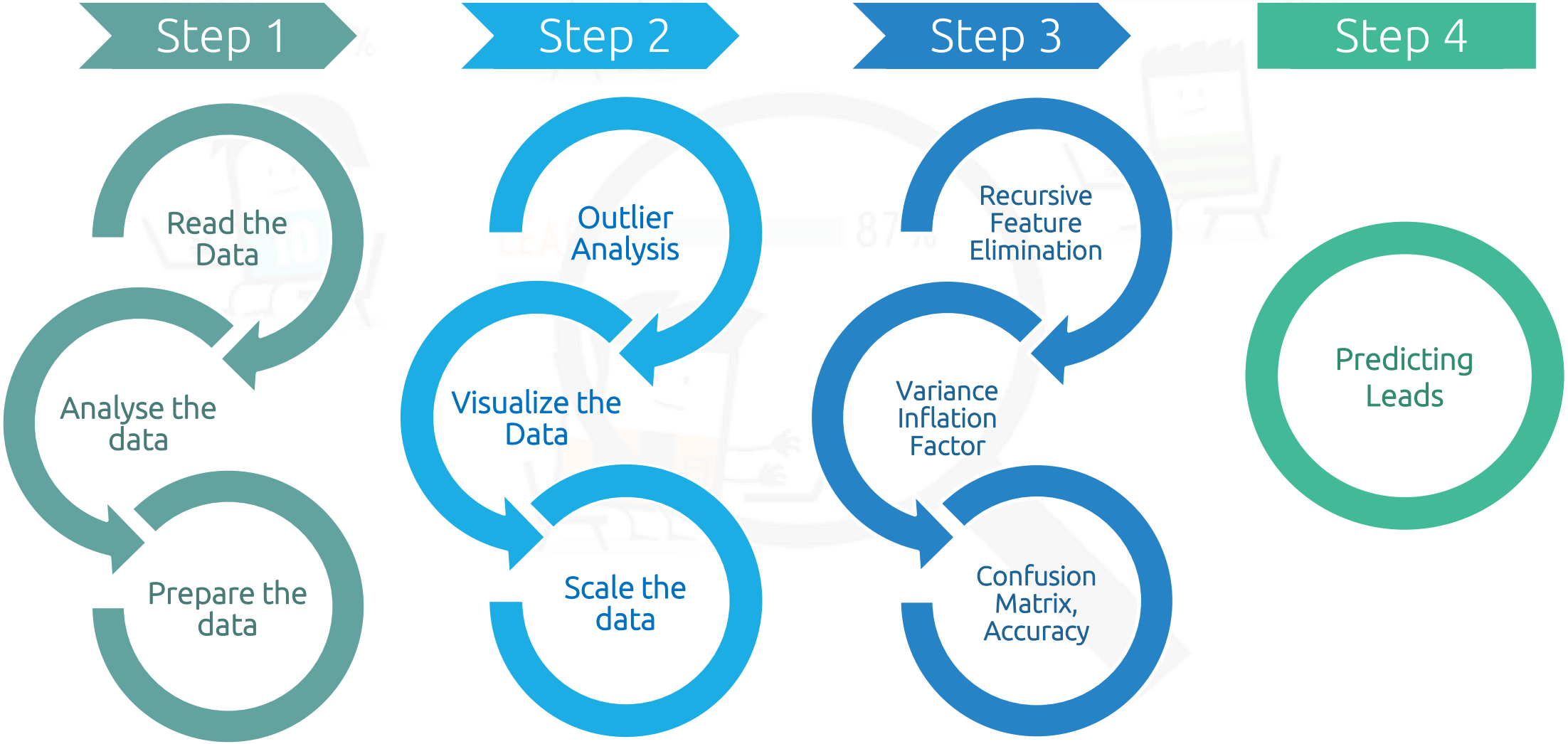
The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified 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 38.5%.

The requirement is to analyze the data in such a way so that potential leads can be identified easily which will make lead conversion rate to go up easily. Also to obtain the variables/features/categories that highly impact the lead conversion.

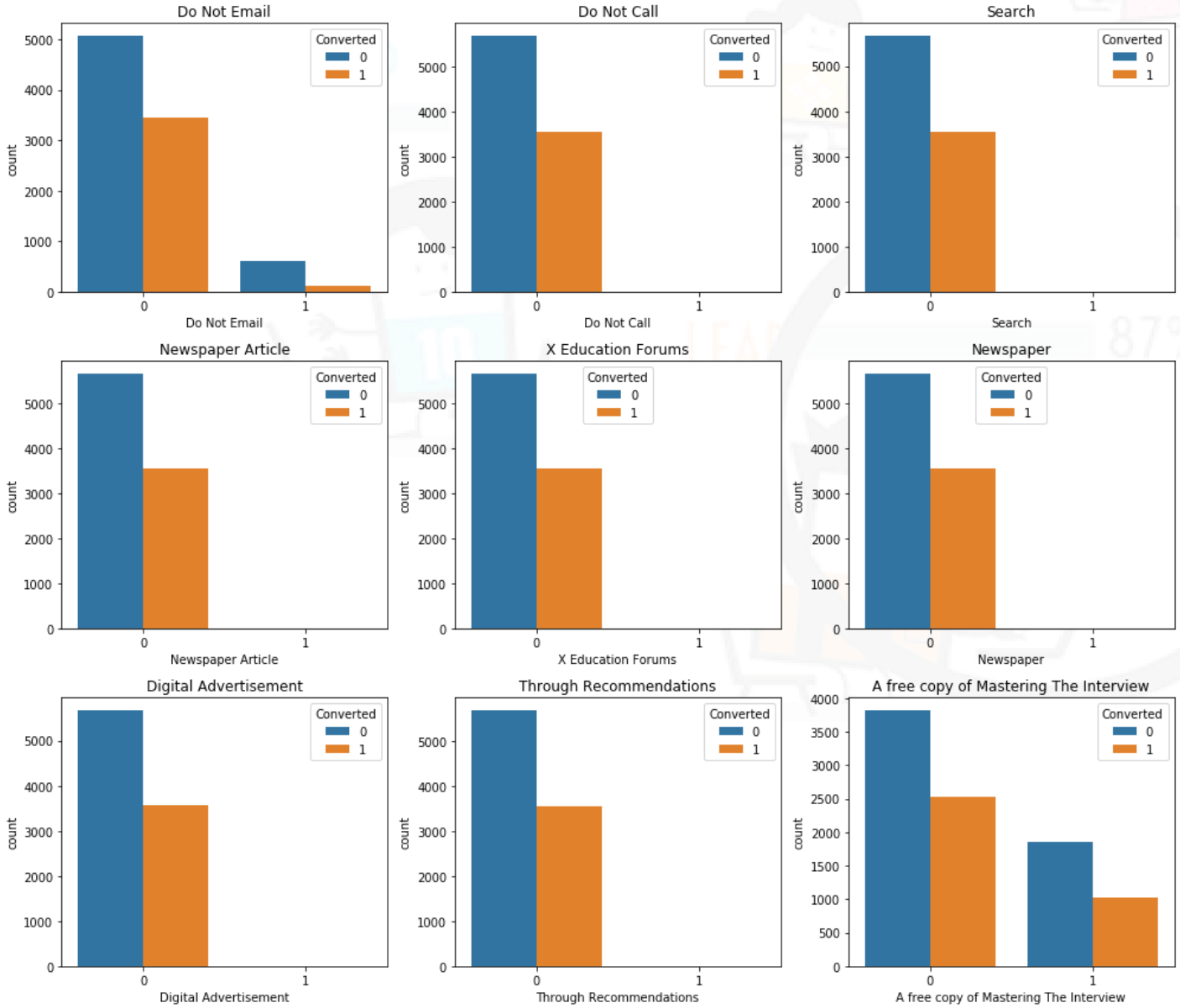
Key Steps Involved

- Reading and loading the data from source file to python pandas data frame.
- Analysing the data-frame to get the information of the columns, datatypes, their statistical values like mean, median, mode and percentiles.
- Applying data cleansing and data pre-processing techniques.
- Finding the missing values and imputing the same.
- Data transformation to convert some negative values in positive like no. of days, prices etc.
- Finding out the categorical and numerical variables.
- Analysing the numeric variables to detect the outliers.
- Visualising the categorical , numerical and binary variables.
- Converting the categorical features to dummy variables.
- Assessing the correlation between the dummy variables.
- Performing feature elimination to remove the insignificant variables.
- Data Modelling and logistic regression model building, assessing VIF and p-Values to refine the model.
- Plotting different charts for better insights and easy understanding.
- Concluding the insights.

Analysis Approach



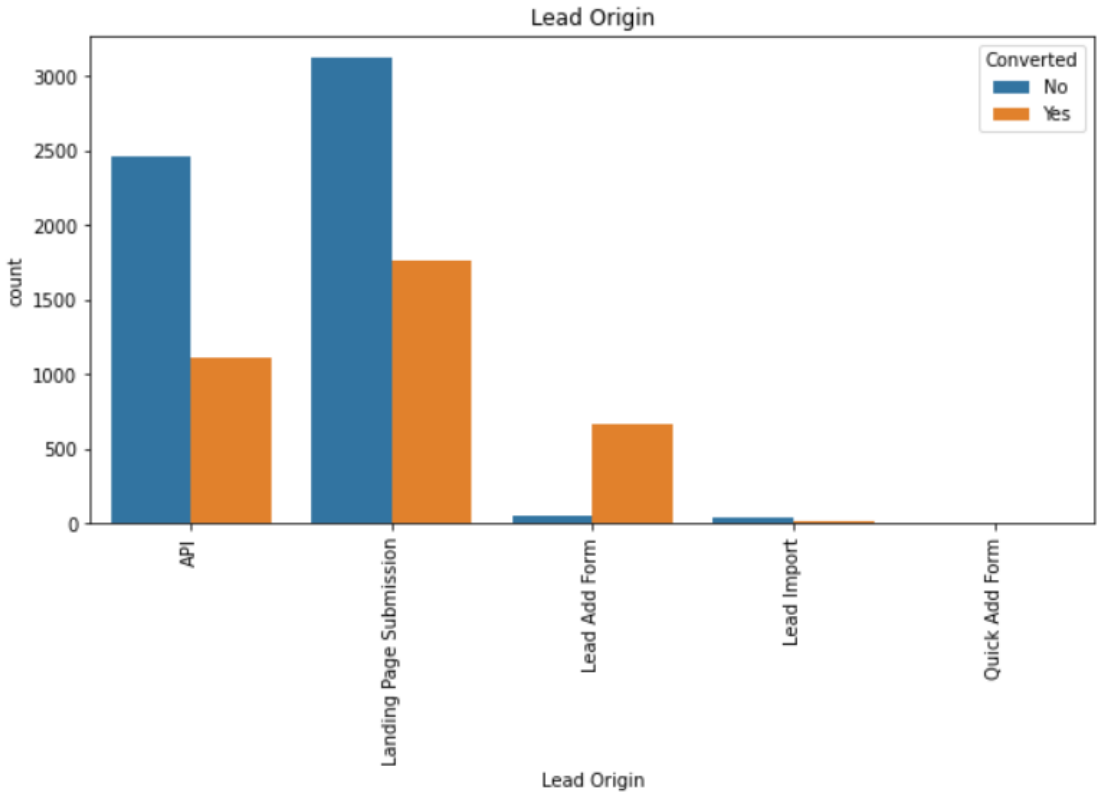
Binary Features Visualization

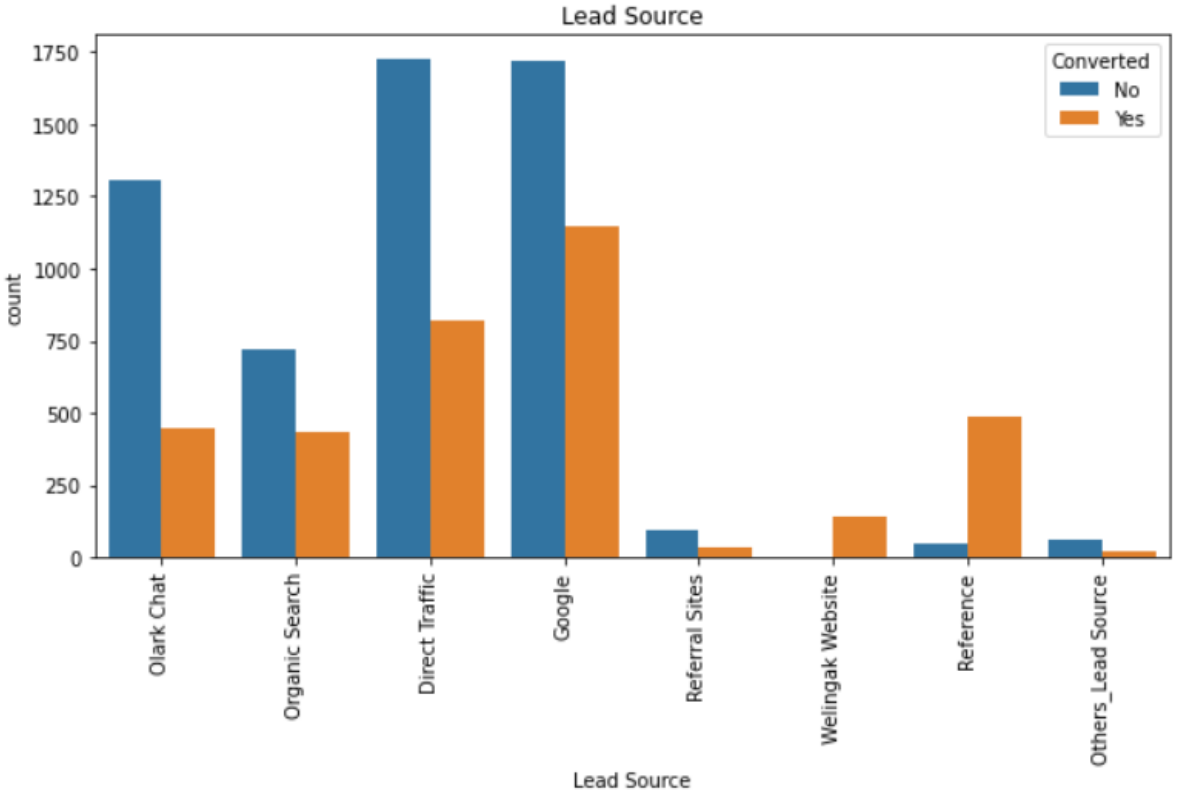


Most of these features show low variance, i.e. most of the data lies in 'No' category and these variables do not have an effect in the target variable since these are with low variance and hence dropped

A free copy of Mastering The Interview : Shows 69% of NOs and 31% YES.

Categorical Features Visualization (1/3)

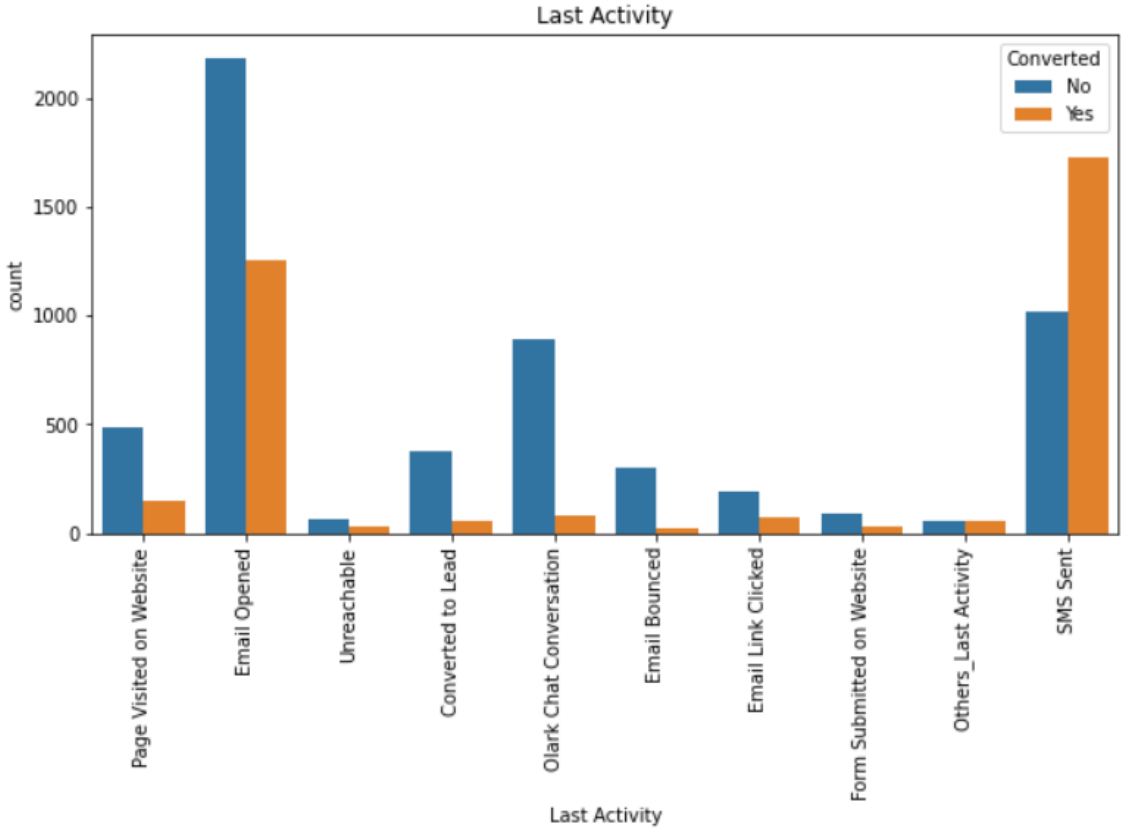




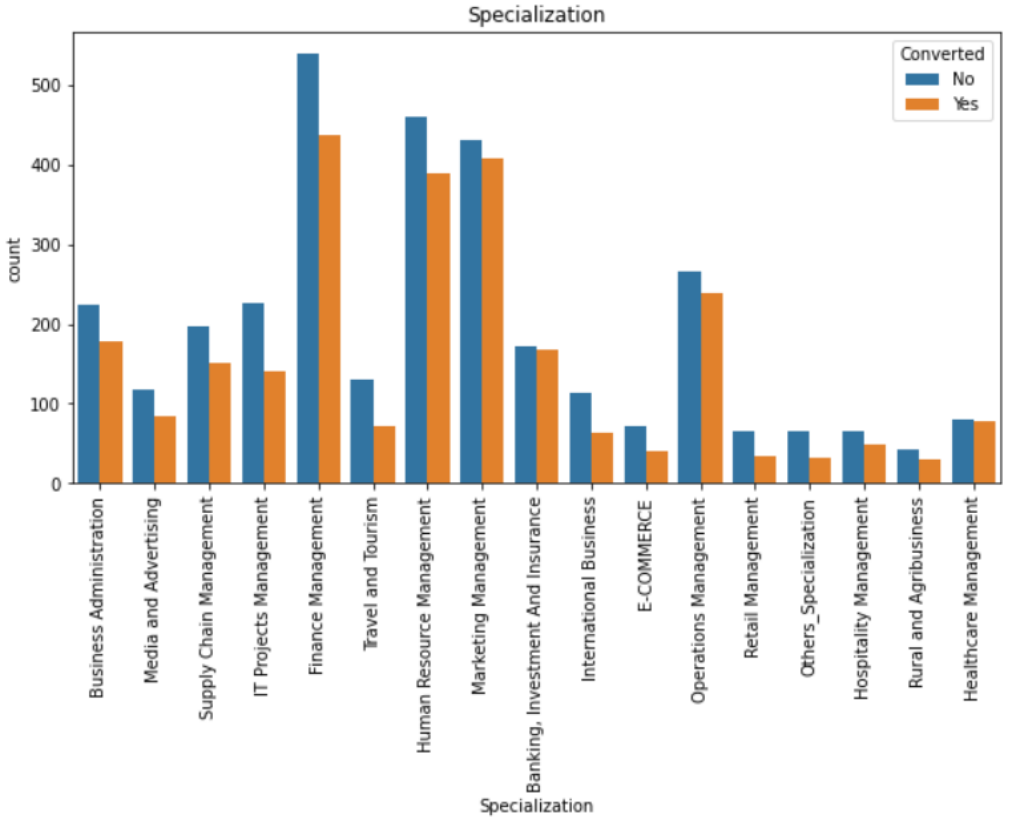
- Landing Page Submission, has higher number of leads, and higher number of conversion
- API, also has higher number of leads, and higher number of conversion but bit less compared to landing page submission.
- The Lead Add Form has higher number of conversion rate compared to other categories, however, the number of lead is less.
- Quick Add form has the least number of leads.

- The Direct Traffic and Google has more number of leads, however Google has a higher conversion compared to direct traffic.
- Lead conversion for Reference is Higher and for Welingak Website is also slightly higher.
- Olark Chat ratios conversion rate is lower.

Categorical Features Visualization (2/3)

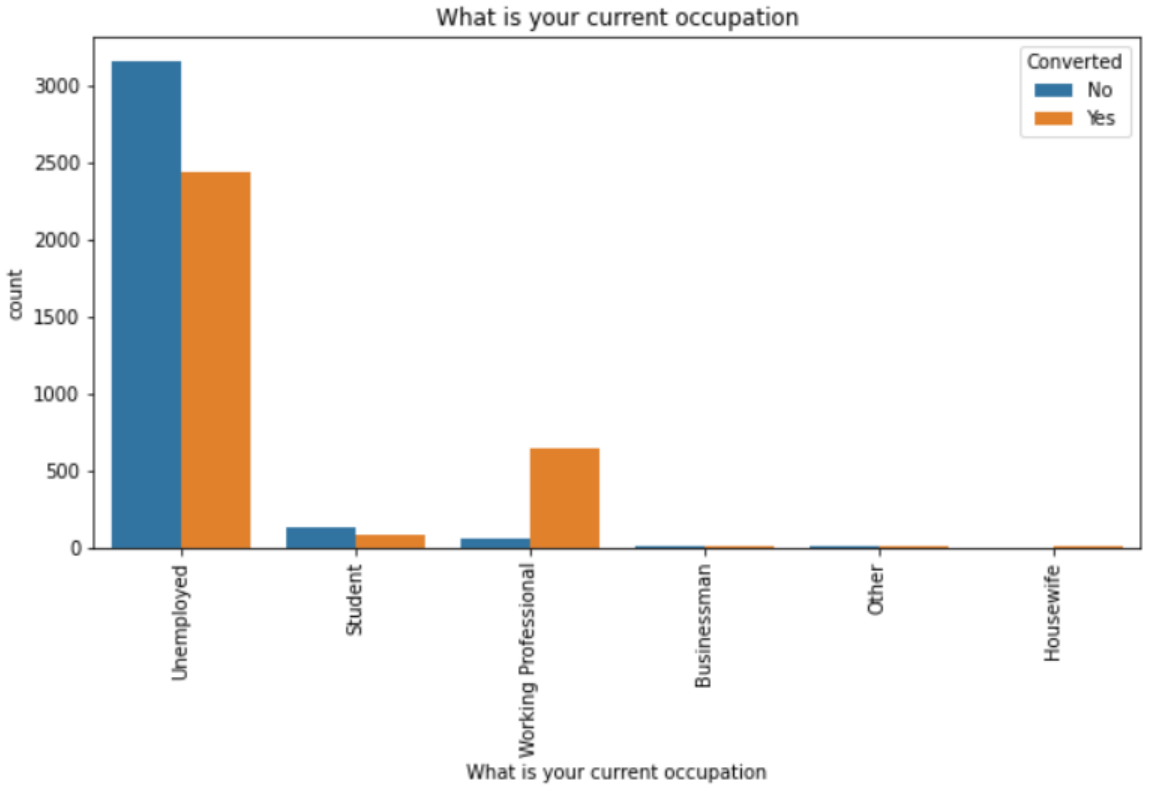


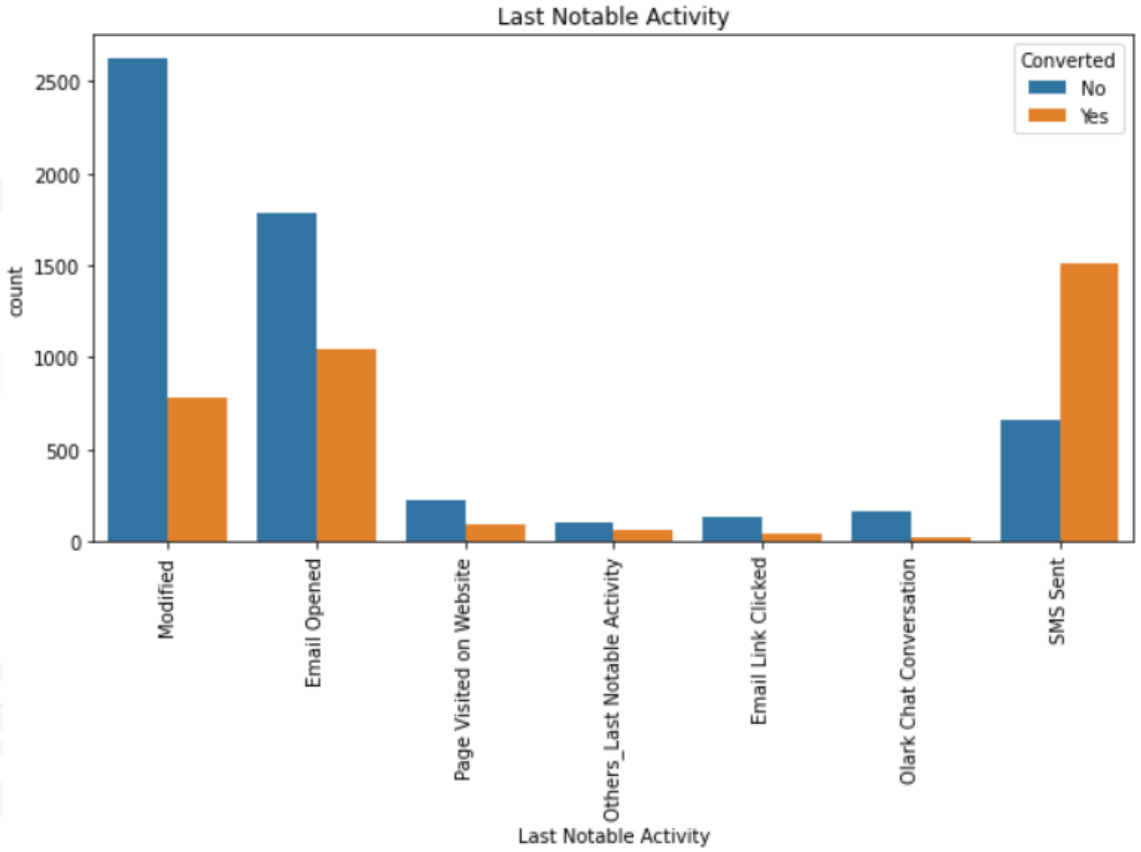
- SMS sent leads have highest conversion rate.
- Number of leads are more for email opened and so is the count of conversions.
- Olark chat conversation has less number of conversions



- More number of leads are from Finance Management specializations.
- Conversion rates are more for Banking, Investments, Marketing Management and Operation Management.
- Total Number of leads are less from Health care Management, however there is approximately 50% conversion rate so need to concentrate more on such group.

Categorical Features Visualization (3/3)

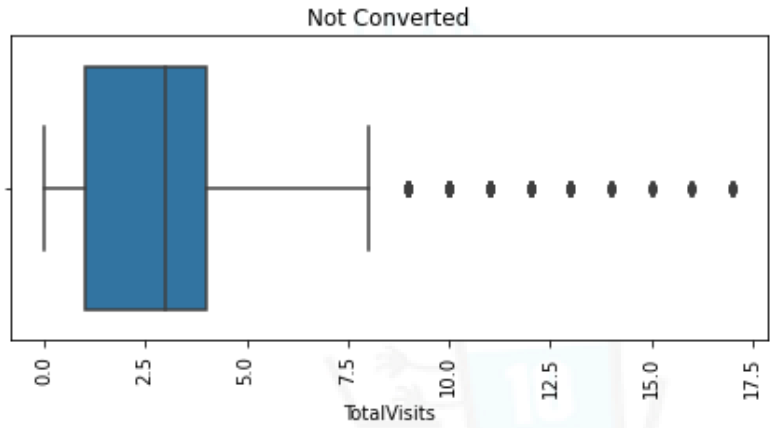


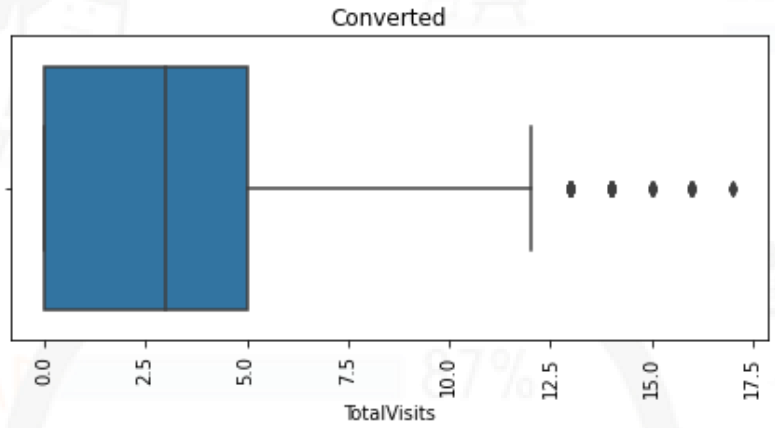


- Unemployed are high in both converted and non converted categories.
- The rate of leads conversion is higher for Working Professional.

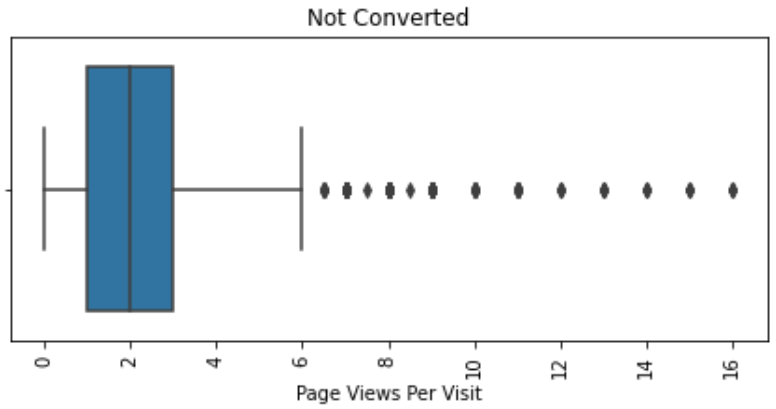
- Modified has more number of leads, however the conversion rate is low.
- Total number of leads are less from Others category
- SMS Sent has higher lead conversion rate and number.

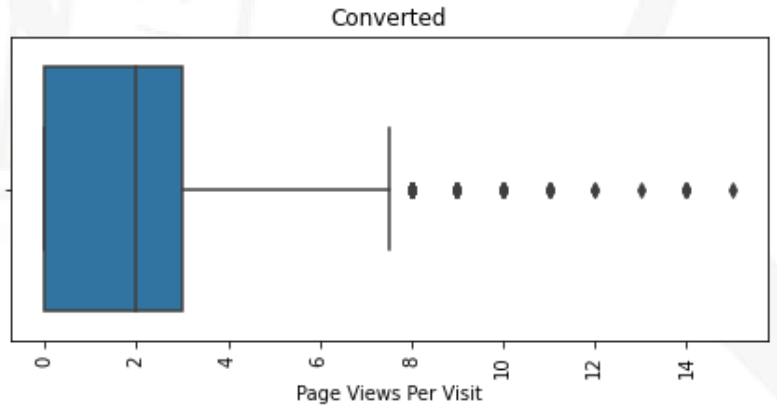
Numerical Features Visualization





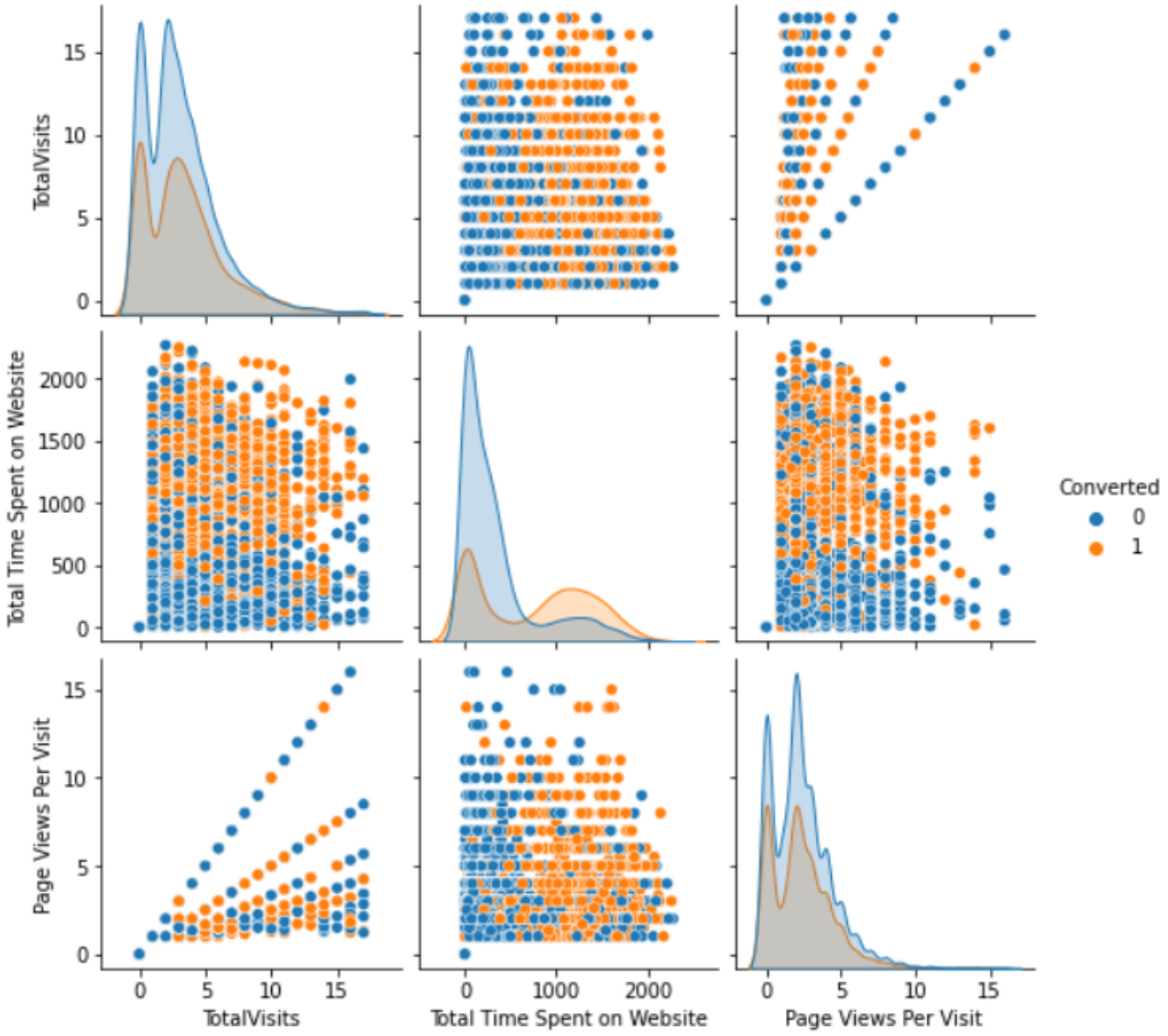
- The Median of total visits for non converted is 2.7 and max is at 7.5
- The Median of total visits for converted is around 2.7 and max is at 12.5.
- The spread is more for converted, with minimum and 25th percentile at 0.

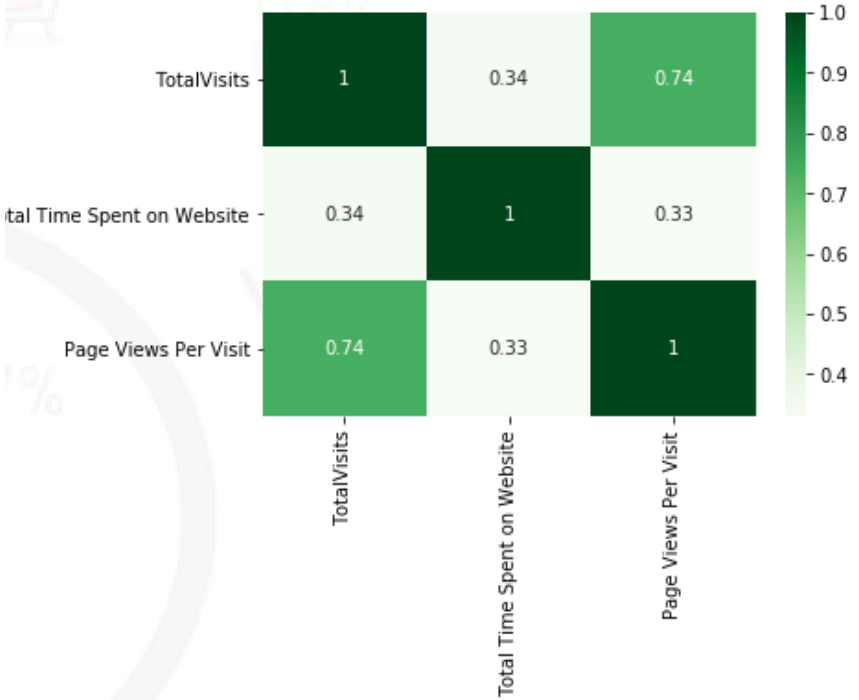




- The box plot has outliers at higher values.
- The Median of Page View per visit for non converted is 2 and max is at 6
- The Median of Page View per visit for converted is around 2 and max is at 7.8.
- The spread is more for converted, with minimum and 25th percentile at 0

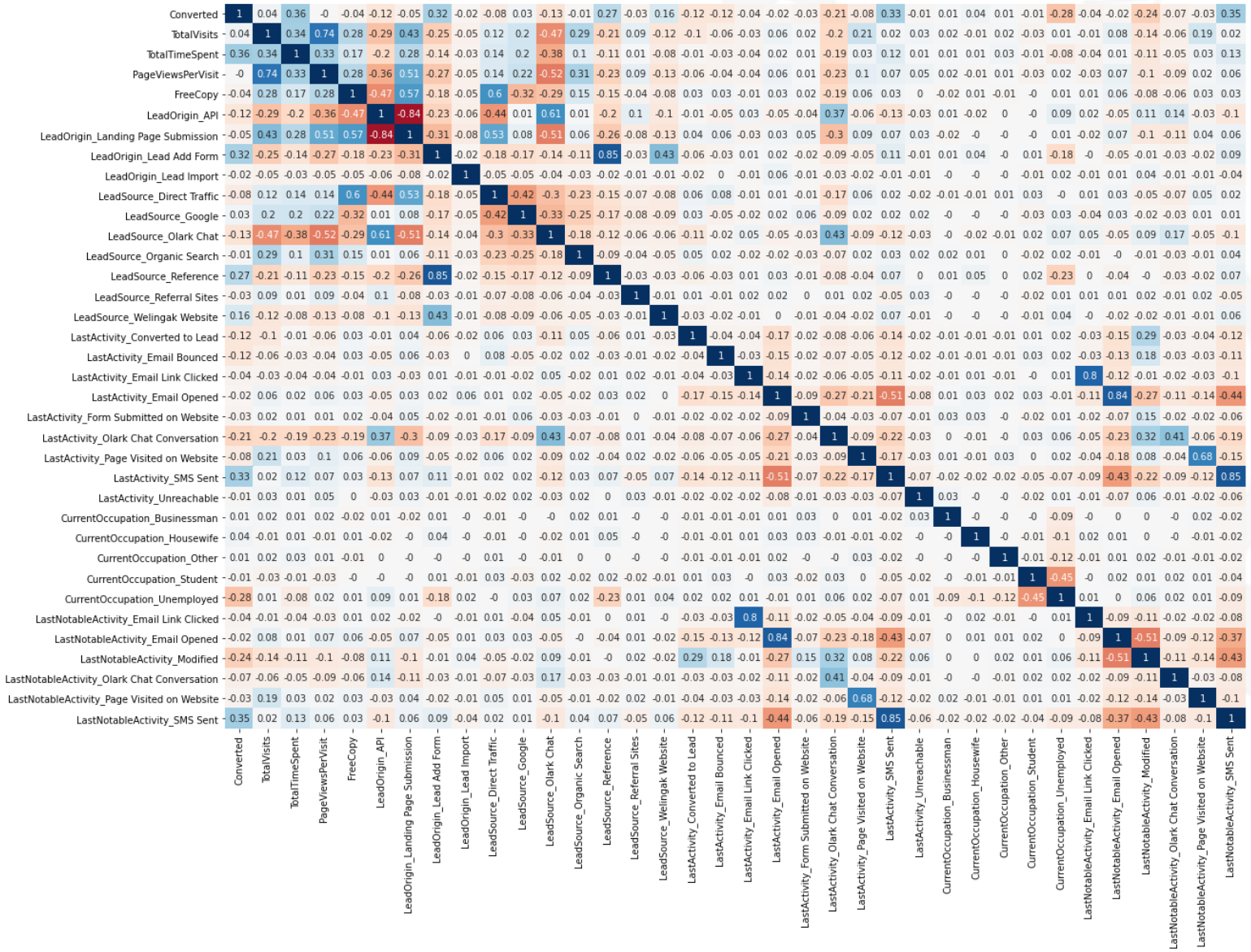
Numerical Features – Pair plot and Heatmap





- The distribution plot shows that all the numerical data is right skewed and the converted leads show a bimodal trend.
- Total Visits and Page Views per Visit are positively correlated

Correlation with dummy variables



- Lead Origin API and Lead Origin Landing Page Submission are highly negatively correlated.
- Lead Source Reference and Lead Origin Lead Add form are highly positively correlated.
- Last Notable activity email link clicked and Last activity email link clicked are highly positively correlated.
- Last Notable activity email opened and Last activity email opened are highly positively correlated.
- Last Notable activity SMS sent and Last activity SMS sent are highly positively correlated.

Model building with all features

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6409
Model:	GLM	Df Residuals:	6379
Model Family:	Binomial	Df Model:	29
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2681.3
Date:	Sat, 21 Nov 2020	Deviance:	5362.5
Time:	19:16:31	Pearson chi2:	7.14e+03
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2002	0.589	0.340	0.734	-0.955	1.356
TotalVisits	0.2600	0.052	4.958	0.000	0.157	0.363
TotalTimeSpent	1.1063	0.040	27.755	0.000	1.028	1.184
PageViewsPerVisit	-0.2475	0.055	-4.470	0.000	-0.356	-0.139
FreeCopy	-0.1447	0.104	-1.390	0.165	-0.349	0.059
LeadOrigin_API	0.0072	0.105	0.068	0.946	-0.199	0.213
LeadOrigin_Lead Import	-0.2246	0.652	-0.345	0.730	-1.502	1.053
LeadSource_Direct Traffic	-0.9103	0.519	-1.754	0.079	-1.928	0.107
LeadSource_Google	-0.5727	0.514	-1.114	0.265	-1.581	0.435
LeadSource_Olark Chat	0.3121	0.527	0.592	0.554	-0.720	1.344
LeadSource_Organic Search	-0.7968	0.523	-1.522	0.128	-1.823	0.229
LeadSource_Reference	3.2126	0.563	5.711	0.000	2.110	4.315
LeadSource_Referral Sites	-0.7408	0.596	-1.243	0.214	-1.909	0.427
LeadSource_Welingak Website	4.8274	0.888	5.437	0.000	3.087	6.567
LastActivity_Converted to Lead	-1.0470	0.359	-2.916	0.004	-1.751	-0.343
LastActivity_Email Bounced	-1.9390	0.408	-4.751	0.000	-2.739	-1.139
LastActivity_Email Link Clicked	-0.8607	0.352	-2.445	0.014	-1.551	-0.171
LastActivity_Email Opened	-0.5158	0.295	-1.748	0.081	-1.094	0.063
LastActivity_Form Submitted on Website	-0.7427	0.447	-1.660	0.097	-1.620	0.134
LastActivity_Olark Chat Conversation	-1.4789	0.348	-4.255	0.000	-2.160	-0.798
LastActivity_Page Visited on Website	-0.7418	0.364	-2.039	0.041	-1.455	-0.029
LastActivity_SMS Sent	0.6895	0.296	2.329	0.020	0.109	1.270
LastActivity_Unreachable	0.1365	0.430	0.317	0.751	-0.707	0.980
CurrentOccupation_Businessman	1.3308	0.956	1.391	0.164	-0.544	3.205
CurrentOccupation_Housewife	22.3643	1.58e+04	0.001	0.999	-3.1e+04	3.1e+04
CurrentOccupation_Student	0.5505	0.230	2.393	0.017	0.100	1.001
CurrentOccupation_Working Professional	2.7505	0.187	14.732	0.000	2.385	3.116
LastNotableActivity_Modified	-0.8389	0.096	-8.732	0.000	-1.027	-0.651
LastNotableActivity_Olark Chat Conversation	-0.7681	0.380	-2.023	0.043	-1.512	-0.024
LastNotableActivity_Page Visited on Website	-0.6656	0.306	-2.177	0.029	-1.265	-0.066

It is evident from the results that most of the features are having higher p-values and those can be eliminated using RFE method.

Model building after applying RFE

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0491	0.091	-0.541	0.589	-0.227	0.129
TotalTimeSpent	1.0968	0.039	27.997	0.000	1.020	1.174
LeadSource_Direct Traffic	-1.3341	0.113	-11.837	0.000	-1.555	-1.113
LeadSource_Google	-0.8865	0.107	-8.281	0.000	-1.096	-0.677
LeadSource_Organic Search	-1.1635	0.130	-8.920	0.000	-1.419	-0.908
LeadSource_Reference	2.9634	0.238	12.471	0.000	2.498	3.429
LeadSource_Referral Sites	-1.0220	0.313	-3.265	0.001	-1.636	-0.408
LeadSource_Welingak Website	4.5212	0.725	6.234	0.000	3.100	5.943
LastActivity_Email Bounced	-1.3021	0.288	-4.522	0.000	-1.867	-0.738
LastActivity_Olark Chat Conversation	-0.7925	0.194	-4.092	0.000	-1.172	-0.413
LastActivity_SMS Sent	1.2399	0.074	16.827	0.000	1.095	1.384
CurrentOccupation_Businessman	1.3636	0.939	1.452	0.147	-0.477	3.205
CurrentOccupation_Housewife	22.2383	1.59e+04	0.001	0.999	-3.11e+04	3.12e+04
CurrentOccupation_Working Professional	2.7218	0.185	14.702	0.000	2.359	3.085
LastNotableActivity_Modified	-0.9298	0.082	-11.403	0.000	-1.090	-0.770
LastNotableActivity_Olark Chat Conversation	-0.7941	0.372	-2.133	0.033	-1.524	-0.064

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0475	0.091	-0.523	0.601	-0.225	0.130
TotalTimeSpent	1.0975	0.039	28.025	0.000	1.021	1.174
LeadSource_Direct Traffic	-1.3322	0.113	-11.826	0.000	-1.553	-1.111
LeadSource_Google	-0.8841	0.107	-8.263	0.000	-1.094	-0.674
LeadSource_Organic Search	-1.1642	0.130	-8.927	0.000	-1.420	-0.909
LeadSource_Reference	2.9909	0.237	12.607	0.000	2.526	3.456
LeadSource_Referral Sites	-1.0235	0.313	-3.270	0.001	-1.637	-0.410
LeadSource_Welingak Website	4.5206	0.725	6.233	0.000	3.099	5.942
LastActivity_Email Bounced	-1.3078	0.288	-4.539	0.000	-1.872	-0.743
LastActivity_Olark Chat Conversation	-0.7969	0.194	-4.116	0.000	-1.176	-0.417
LastActivity_SMS Sent	1.2361	0.074	16.781	0.000	1.092	1.380
CurrentOccupation_Businessman	1.3589	0.940	1.445	0.148	-0.484	3.202
CurrentOccupation_Working Professional	2.7184	0.185	14.683	0.000	2.356	3.081
LastNotableActivity_Modified	-0.9274	0.081	-11.385	0.000	-1.087	-0.768
LastNotableActivity_Olark Chat Conversation	-0.7920	0.372	-2.127	0.033	-1.522	-0.062

	Features	VIF
13	LastNotableActivity_Modified	1.94
8	LastActivity_Olark Chat Conversation	1.72
9	LastActivity_SMS Sent	1.47
2	LeadSource_Google	1.42
1	LeadSource_Direct Traffic	1.36
14	LastNotableActivity_Olark Chat Conversation	1.33
4	LeadSource_Reference	1.21
3	LeadSource_Organic Search	1.19
12	CurrentOccupation_Working Professional	1.18
0	TotalTimeSpent	1.16
7	LastActivity_Email Bounced	1.12
6	LeadSource_Welingak Website	1.04
5	LeadSource_Referral Sites	1.01
11	CurrentOccupation_Housewife	1.01

The RFE methods helps in eliminating most of the features with collinearity.

Further, model is fine tuned by eliminating features having higher p-values and VIF.

Final Model

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6409
Model:	GLM	Df Residuals:	6396
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2720.5
Date:	Sat, 21 Nov 2020	Deviance:	5441.1
Time:	19:16:33	Pearson chi2:	7.00e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

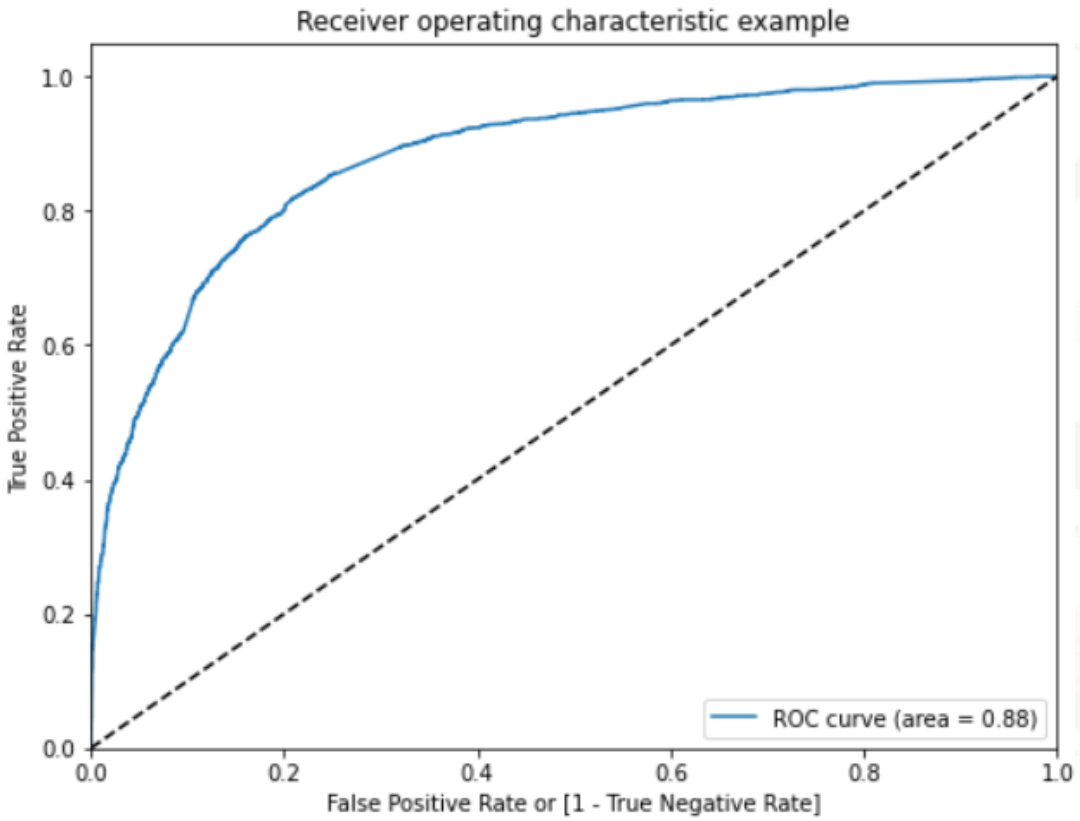
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0576	0.090	-0.637	0.524	-0.235	0.120
TotalTimeSpent	1.0935	0.039	28.016	0.000	1.017	1.170
LeadSource_Direct Traffic	-1.3289	0.113	-11.808	0.000	-1.549	-1.108
LeadSource_Google	-0.8829	0.107	-8.255	0.000	-1.093	-0.673
LeadSource_Organic Search	-1.1570	0.130	-8.890	0.000	-1.412	-0.902
LeadSource_Reference	2.9895	0.237	12.605	0.000	2.525	3.454
LeadSource_Referral Sites	-1.0078	0.313	-3.222	0.001	-1.621	-0.395
LeadSource_Welingak Website	4.4857	0.724	6.199	0.000	3.067	5.904
LastActivity_Email Bounced	-1.3246	0.288	-4.605	0.000	-1.888	-0.761
LastActivity_Olark Chat Conversation	-1.0323	0.169	-6.104	0.000	-1.364	-0.701
LastActivity_SMS Sent	1.2352	0.074	16.802	0.000	1.091	1.379
CurrentOccupation_Working Professional	2.7139	0.185	14.657	0.000	2.351	3.077
LastNotableActivity_Modified	-0.8850	0.079	-11.249	0.000	-1.039	-0.731

	Features	VIF
11	LastNotableActivity_Modified	1.80
9	LastActivity_SMS Sent	1.47
2	LeadSource_Google	1.41
1	LeadSource_Direct Traffic	1.35
8	LastActivity_Olark Chat Conversation	1.31
4	LeadSource_Reference	1.20
3	LeadSource_Organic Search	1.18
10	CurrentOccupation_Working Professional	1.18
0	TotalTimeSpent	1.16
7	LastActivity_Email Bounced	1.11
6	LeadSource_Welingak Website	1.04
5	LeadSource_Referral Sites	1.01

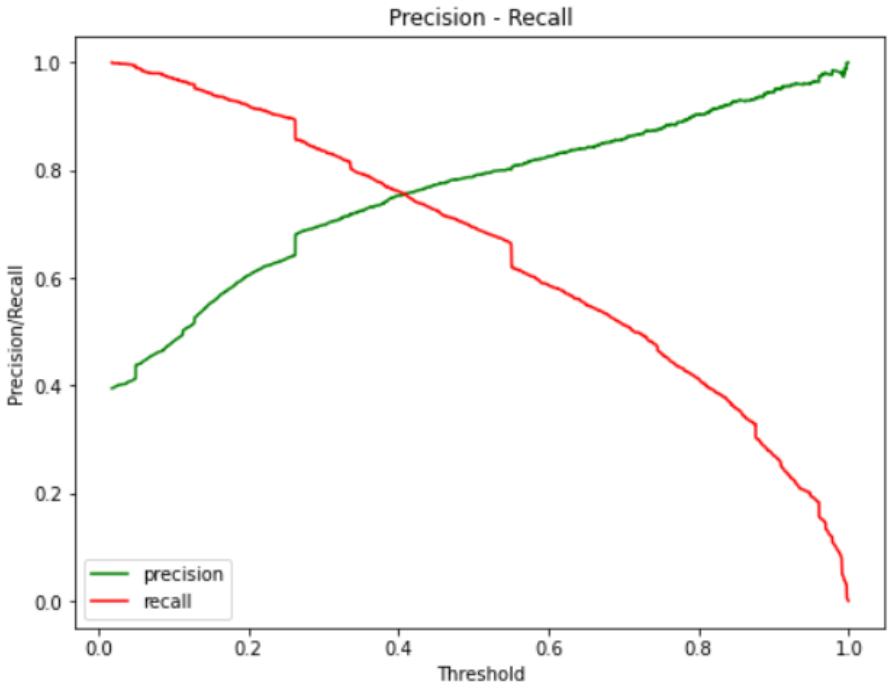
The final model clearly depicts that all the features are having p-value less than 0.05 and VIF < 5.

Accuracy Measures

ROC Curve



Precision - Recall

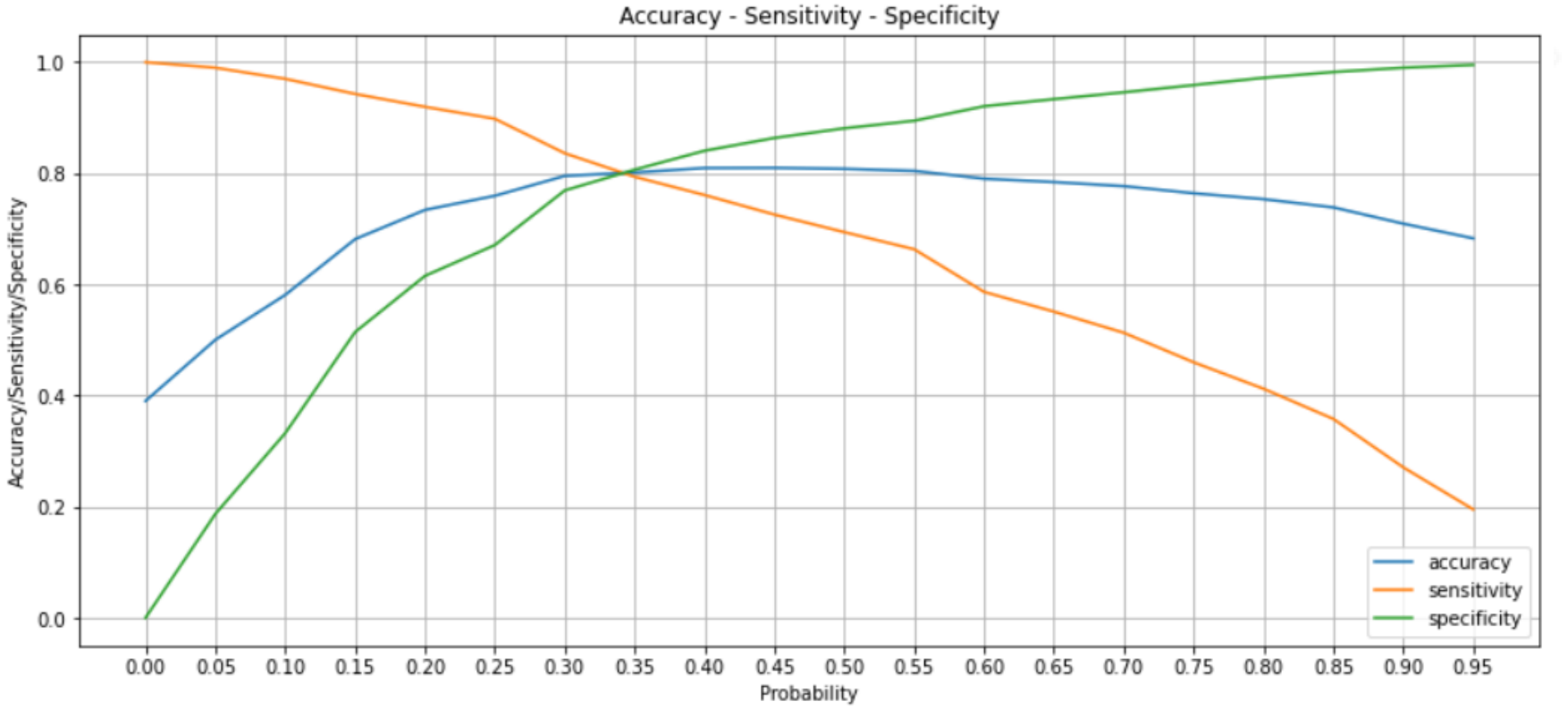


Confusion Matrix

	Predicted No	Predicted Yes
Actual No	3443	465
Actual Yes	765	1736

Accuracy	Precision	Recall Sensitivity	Specificity	F1 Score	Threshold
0.81	0.79	0.69	0.88	0.74	0.50

Optimal Cutoff Point



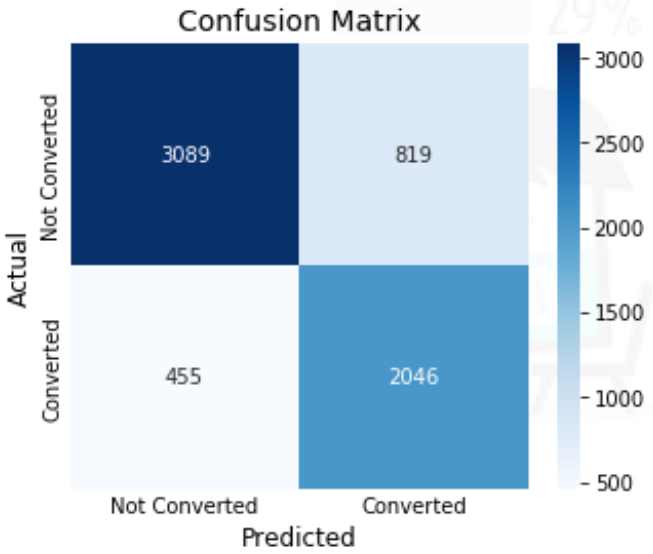

```
fpr, tpr, thresholds = roc_curve( y_train_pred_final.Converted,
                                  y_train_pred_final.Converted_prob, drop_intermediate = False )
print('Optimal cutoff : %.2f'%thresholds[np.argmax(tpr - fpr)])
```

```
Optimal cutoff : 0.33
```

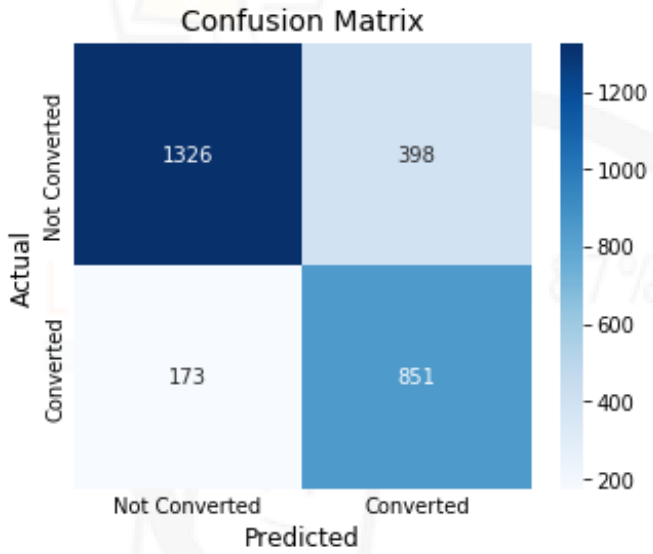
By plotting the accuracy ,sensitivity and specificity, we get 0.33 as the optimum point to take it as a cutoff

Accuracy Measures for Optimal cutoff

Confusion Matrix – Training set



Confusion Matrix – Testing set



	Accuracy	Precision	Recall Sensitivity	Specificity	F1 Score	Threshold
Training Set	0.80	0.71	0.82	0.79	0.76	0.33
Testing Set	0.79	0.68	0.83	0.77	0.75	0.33

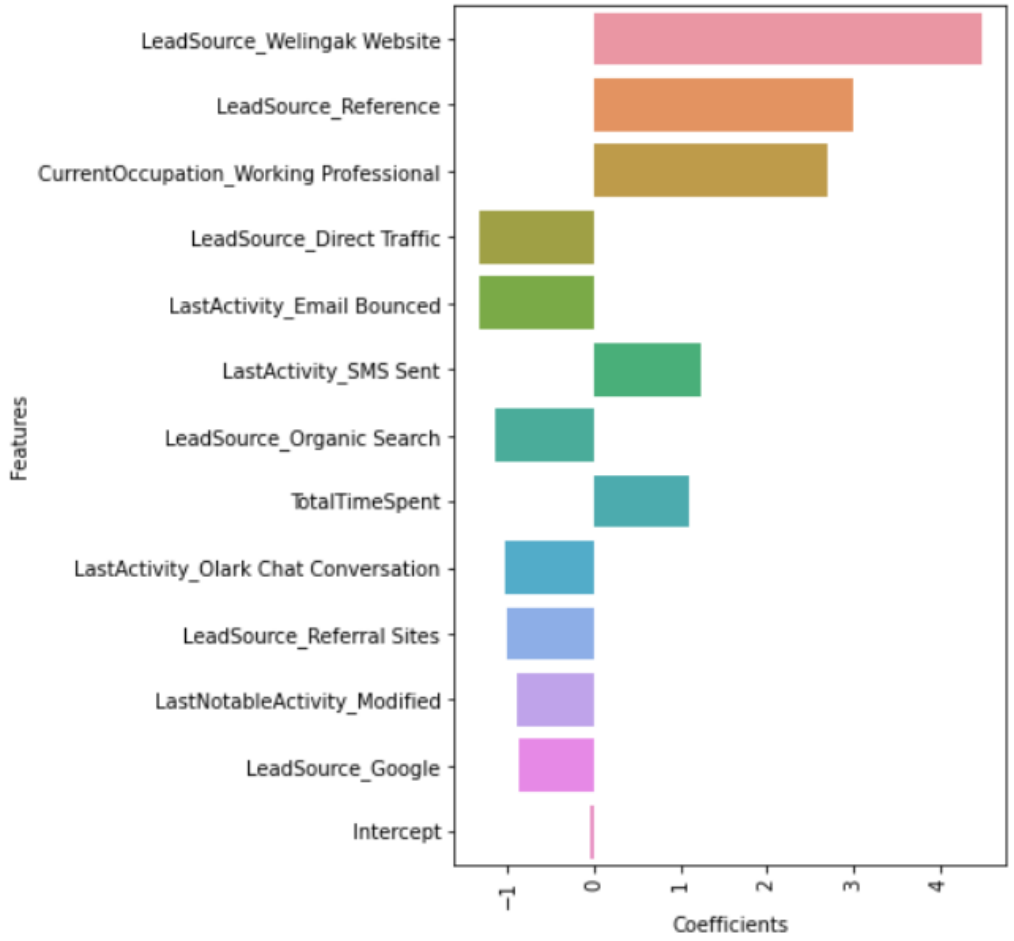
- With optimal cut-off threshold, we have a model accuracy of 0.80 for training set and 0.79 for testing set.
- The ability to predict conversions correctly by the model (precision) is 71% for training set and 68% for test set.
- Ability to detect the lead conversion, i.e. the sensitivity of the model is 82% for training set and 83% for test set. In this case to reduce the false negatives, i.e. model predicting a converted lead as not converted need to be reduced so that there is no risk and we don't lose lead conversion. A higher sensitivity in turn mean a lower false negative rate.

Conclusion

The following are the recommendation to the X Education

The features that have a larger impact on lead conversion are in the order of:

1. Lead Source
    - Welingak Website - Should be considered
    - Reference - Should be considered
    - Direct Traffic - Should be avoided
    - Organic Search - Should be avoided
    - Referral Sites - Should be avoided
    - Google - Should be avoided
2. Current Occupation
    - Working Professional - Should be considered
3. Last Activity
    - Email Bounced - Should be avoided
    - SMS Sent - Should be considered
    - Olark Chat Conversation - Should be avoided
4. Total Time Spent - Should be considered
5. Last Notable Activity
    - Modified - Should be avoided



Thank you