Lead Scoring Assignment

Krishnan Vybhava Srinivasan

Nithya V

Our Understanding

Business Objective

- X Education sells online courses to industry professionals

Problem Statement

- Lead conversion rate is very poor (~30%)

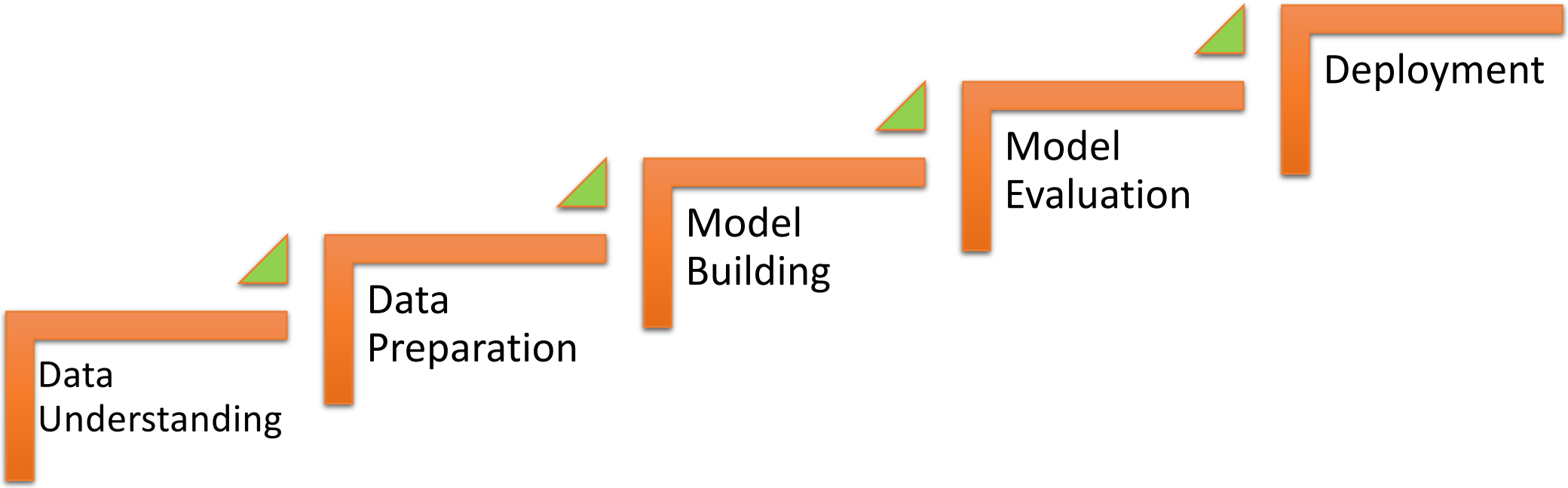
Business Understanding

- Lot of leads generated in the initial stage
- Lot of effort to nurture leads
- 30% of leads get converted to customers

Engagement Goal

- Model to assign a lead score to each of the leads to identify Hot Leads
- Target lead conversion rate to be around 80%.

Approach


Data Understanding
Data Preparation
Model Building
Model Evaluation
Deployment

Data Understanding

- Client Provided us with Leads Datasets Containing **9,240 data points** (rows)
- The dataset **consists 37 attributes** such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.,
- The variable **Converted**, tells if the lead was converted (‘**1**’) or not converted (‘**0**’).
- The Categorical variables contain data called **Select**, which indicates the data has not been provided and hence, equivalent to **Null**
- An overview of the Data provided as per **Table 1.1**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
Prospect ID                                    9240 non-null object
Lead Number                                    9240 non-null int64
Lead Origin                                    9240 non-null object
Lead Source                                    9204 non-null object
Do Not Email                                   9240 non-null object
Do Not Call                                    9240 non-null object
Converted                                      9240 non-null int64
TotalVisits                                    9103 non-null float64
Total Time Spent on Website                    9240 non-null int64
Page Views Per Visit                           9103 non-null float64
Last Activity                                  9137 non-null object
Country                                        6779 non-null object
Specialization                                 7802 non-null object
How did you hear about X Education             7033 non-null object
What is your current occupation                6550 non-null object
What matters most to you in choosing a course  6531 non-null object
Search                                         9240 non-null object
Magazine                                       9240 non-null object
Newspaper Article                              9240 non-null object
X Education Forums                             9240 non-null object
Newspaper                                      9240 non-null object
Digital Advertisement                          9240 non-null object
Through Recommendations                        9240 non-null object
Receive More Updates About Our Courses         9240 non-null object
Tags                                           5887 non-null object
Lead Quality                                   4473 non-null object
Update me on Supply Chain Content              9240 non-null object
Get updates on DM Content                      9240 non-null object
Lead Profile                                   6531 non-null object
City                                           7820 non-null object
Asymmetrique Activity Index                    5022 non-null object
Asymmetrique Profile Index                     5022 non-null object
Asymmetrique Activity Score                    5022 non-null float64
Asymmetrique Profile Score                     5022 non-null float64
I agree to pay the amount through cheque       9240 non-null object
A free copy of Mastering The Interview         9240 non-null object
Last Notable Activity                          9240 non-null object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

Table 1.1 – Overview of Data

Data Preparation

Data Cleaning

Inspection of Data Pre and Post Cleaning is as be under:

	Pre Cleaning	Post Cleaning	Lost in Cleaning
Row	9,240	9,074	166
Columns	37	31	6

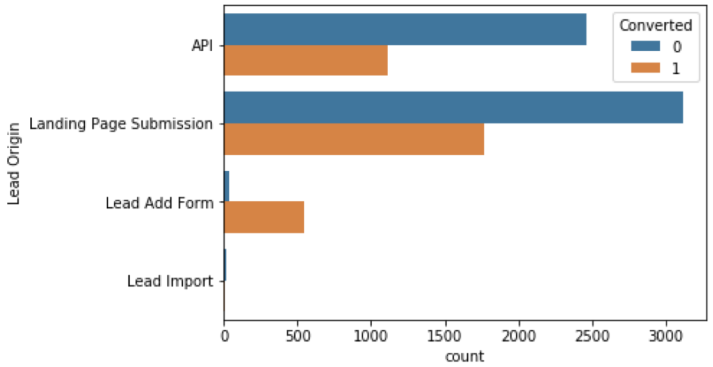
Post Cleaning and treatment of Null values, 'Page Views Per Visit', 'Last Activity ', 'TotalVisits' and 'Lead Source' have null values less than 2%, hence **decided to delete those rows**, as it would have minimum impact on analysis

Columns	Reasons
How did you hear about X Education	78% of Column Null, hence no meaningful analysis of column possible. Hence, **dropped**
Lead Profile	74% of Column Null, hence no meaningful analysis of column possible. Hence, **dropped**.
City	Where City is Null and Country is India, imputed city value as '**Mumbai**' as more than 58% of City is Mumbai. Balance City values where Null imputed as '**Other Cities**'
Country	Where Country is Null. Imputed them with India, since 95% of current value is **India**.
Specialization	There is no specific trend, hence it is assumed in the lead form the option was not available hence none was selected. Hence, imputing Null as '**Others**'
What is your current occupation	imputing instances where 'What is your current occupation' is Null as '**Unemployed**'
What matters most to you in choosing a course	Around 99% of values is '**Better Career Prospects**', hence imputing the same where value is Null
Tags	Assumed that Null values are to indicate will revert, hence imputing with **Will revert after reading the email**
Lead Quality	Assumed that Null values are to indicate **Not Sure**, hence imputing the same.
'Asymmetrique Activity Index' ,'Asymmetrique Activity Score','Asymmetrique Profile Index','Asymmetrique Profile Score'	No meaningfull imputation of values is possible. Also given 45% is Null it will not be suitable for any meaningful analysis. Hence, it is better to **drop these 4 columns.**

Data Preparation - Cont'd

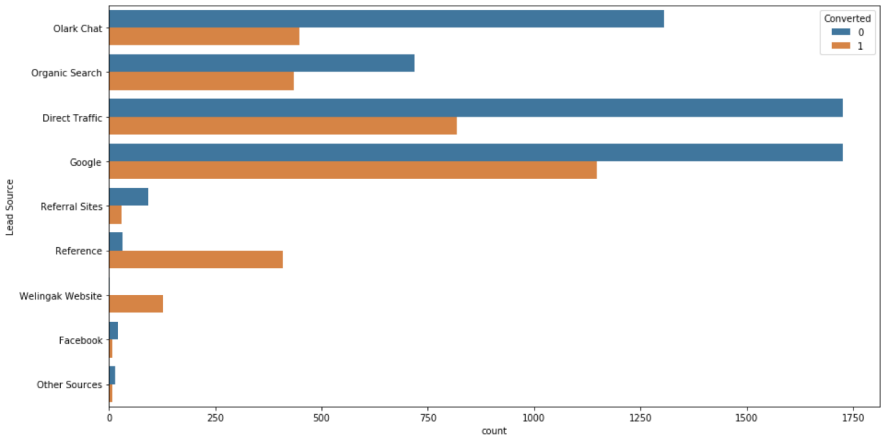
Data Analysis

Lead Origin



- Conversion for API and Landing Page Submission is only around 30%
- Conversion for Lead Add form is very high although overall volumes is very less

Lead Source

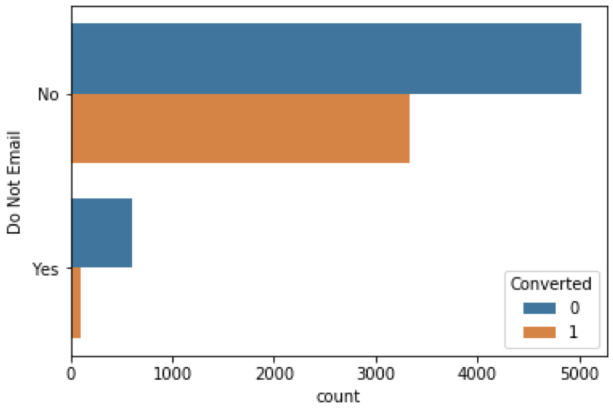


- Most of the source of lead is from Direct and Google
- Reference leads have very high conversion leads
- Wellingak website also has very high conversion leads but number of leads are low
- Other lead source with smaller counts was merged under Other Sources.

Data Preparation - Cont'd

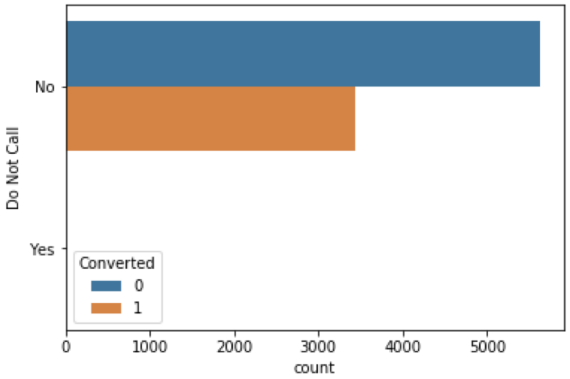
Data Analysis

Do Not Email



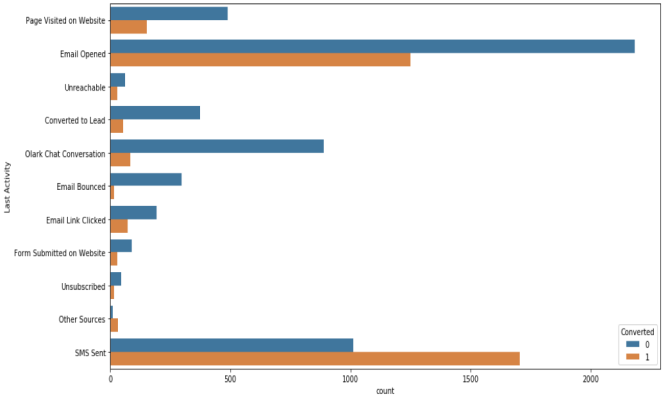
- People who have not opted for Do Not email seem to have a better conversion rate

Do Not Call



- People who have not opted for Do Not Call seem to have a better conversion rate

Last Activity

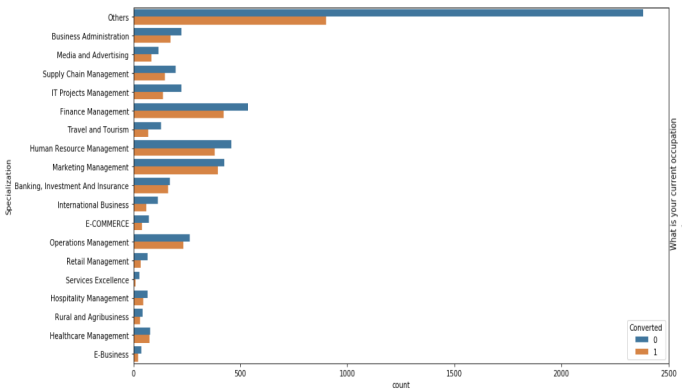


- SMS has the maximum conversion
- Email opened has lot of transactions but the conversion rate is around 30%
- Values with small counts was merged with Others

Data Preparation - Cont’d

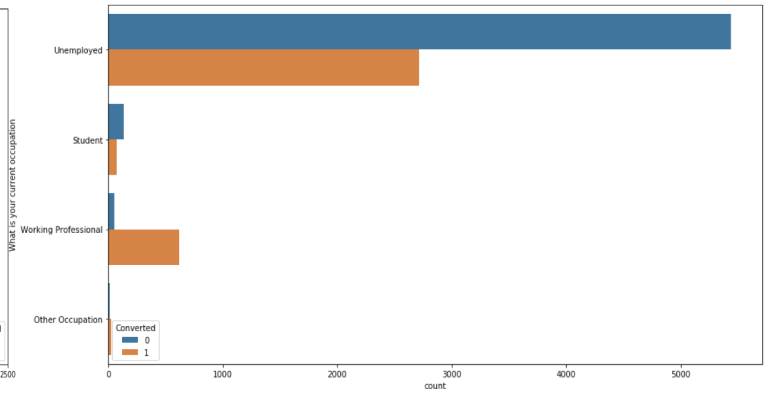
Data Analysis

Specialization



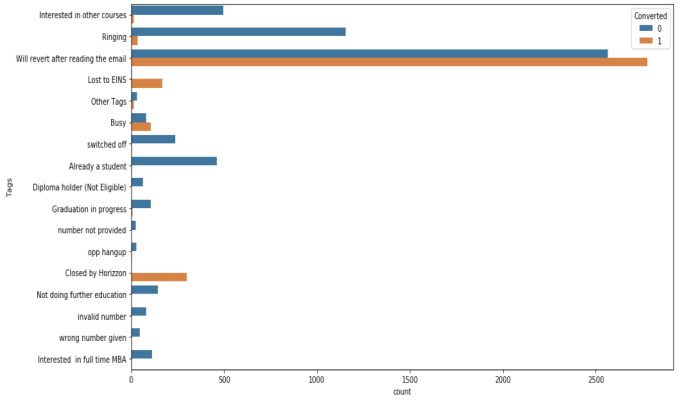
- Not much expect the Others category is the maximum, while many specialization have good conversion ratio

What is Your Current Occupation



- The maximum people enquiring are unemployed
- Working Professionals are most likely to join
- Occupation with smaller counts was merged with Other Occupation

Tags

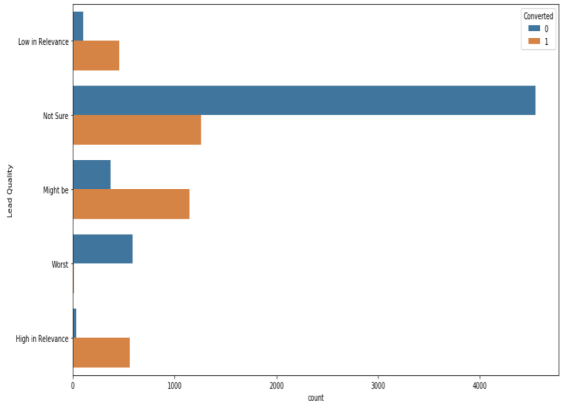


- Can be seen the Will revert to email has high conversion ratio
- Low transaction counts are clubbed together as Other Tags

Data Preparation - Cont'd

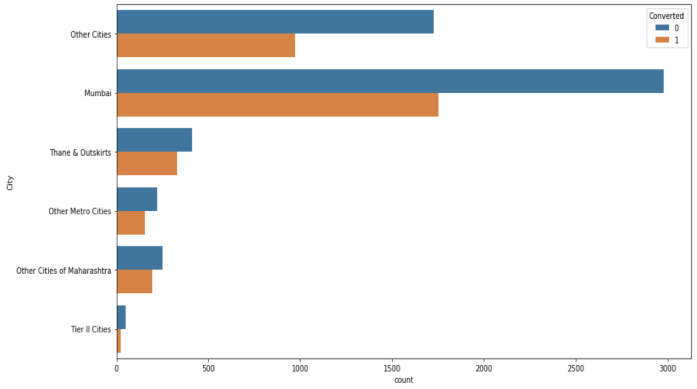
Data Analysis

Lead Quality



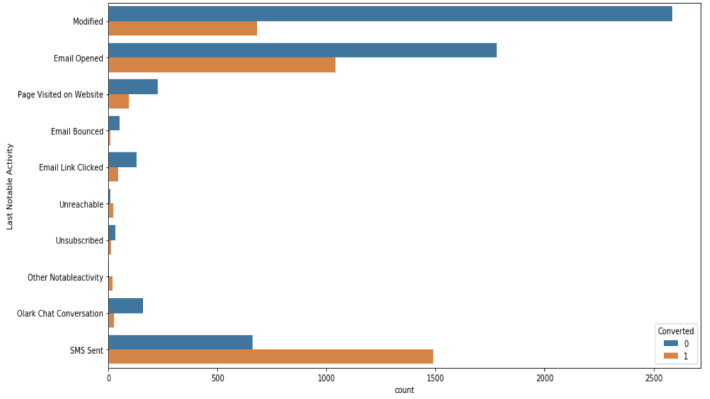
- Not sure has number of leads and 20% conversion ratio.
- Might be and low relevance have a high conversion ratio

City



- Mumbai maximum leads and good conversion ratio

Last Notable Activity



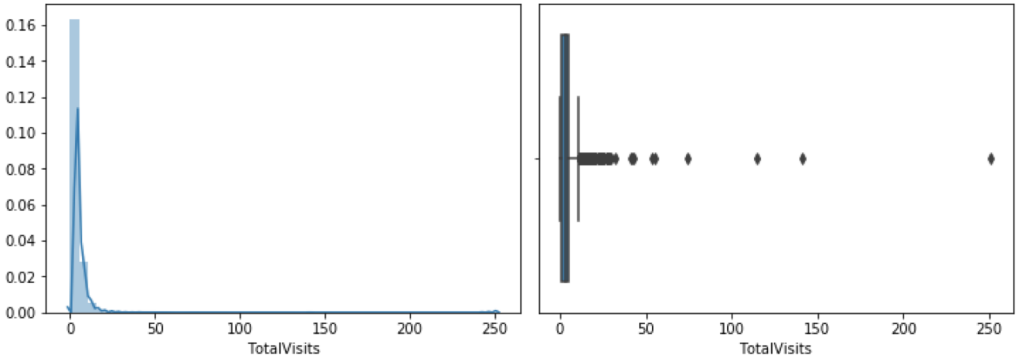
- SMS has great conversion
- Modified and EMail has most leads

Data Preparation - Cont'd

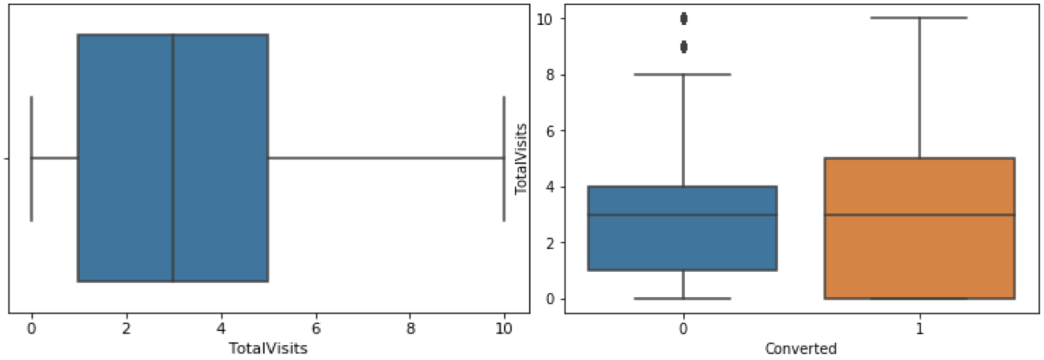
Data Analysis

Total Visits

Before Outlier Adjustment



After Outlier Adjustment



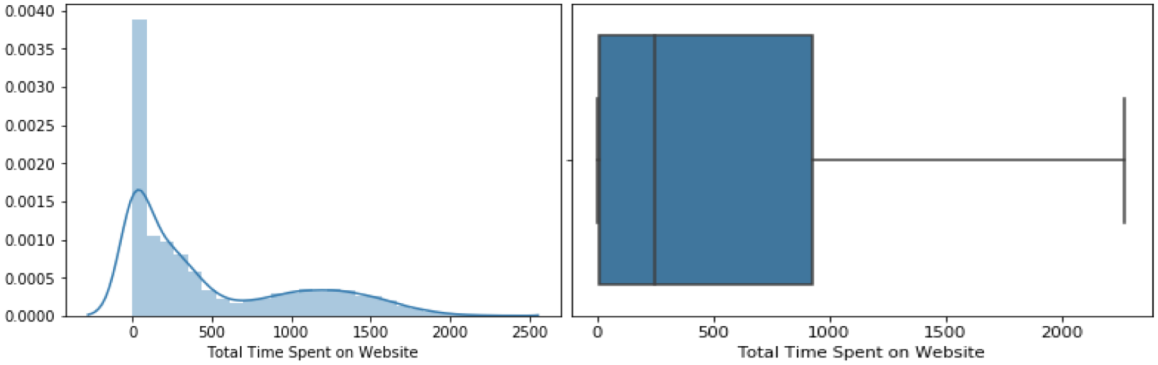
- Nothing specific based on total visits as in converted they spend less visits and more visits
- Mean of both are similar but Upper and Lower range is wider in converted as compared to not converted

Data Preparation - Cont'd

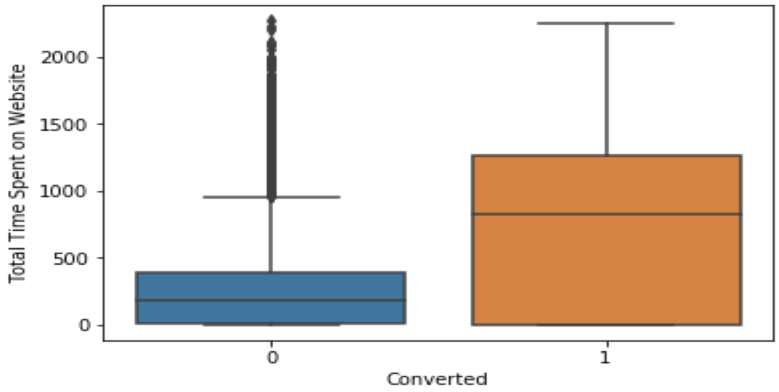
Data Analysis

Total Time Spent on Websites

Before Outlier Adjustment



After Outlier Adjustment



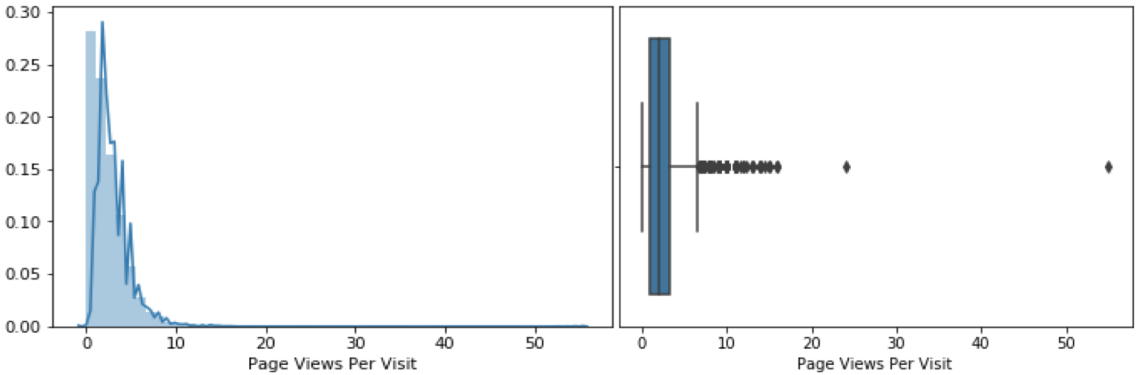
- It appears that anyone spending more time on website is likely to convert
- There are instances where people spend time on website and still did'nt convert

Data Preparation - Cont'd

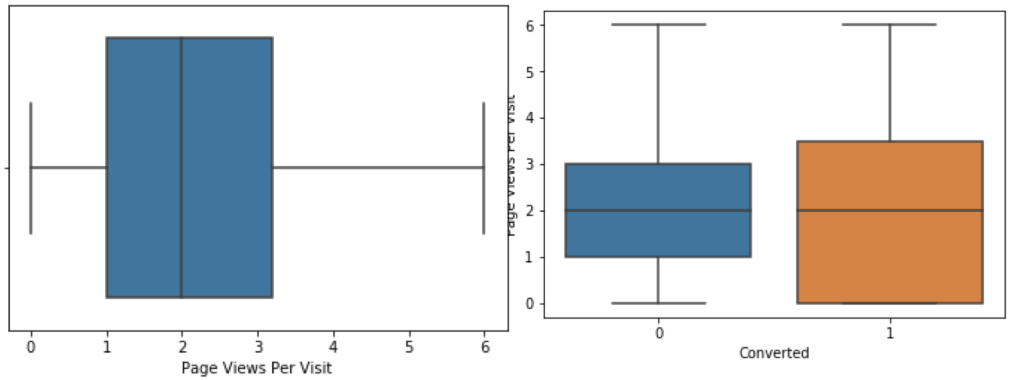
Data Analysis

Pages views per Websites

Before Outlier Adjustment



After Outlier Adjustment



- 'Page Views Per Visit'
- They are very similar not much can be concluded

Data Preparation - Cont'd

Feature Reduction

Feature to be Dropped	Reasons
1. A free copy of Mastering The Interview 2. I agree to pay the amount through cheque 3. Get updates on DM Content 4. Update me on Supply Chain Content 5. Receive More Updates About Our Courses 6. Through Recommendations 7. Digital Advertisement 8. Newspaper 9. X Education Forums 10. Newspaper Article 11. Magazine 12. Search 13. What matters most to you in choosing a course 14. Country	These features are being dropped as most of these categorical variables have only one level and beyond one level the count of other levels are insignificant

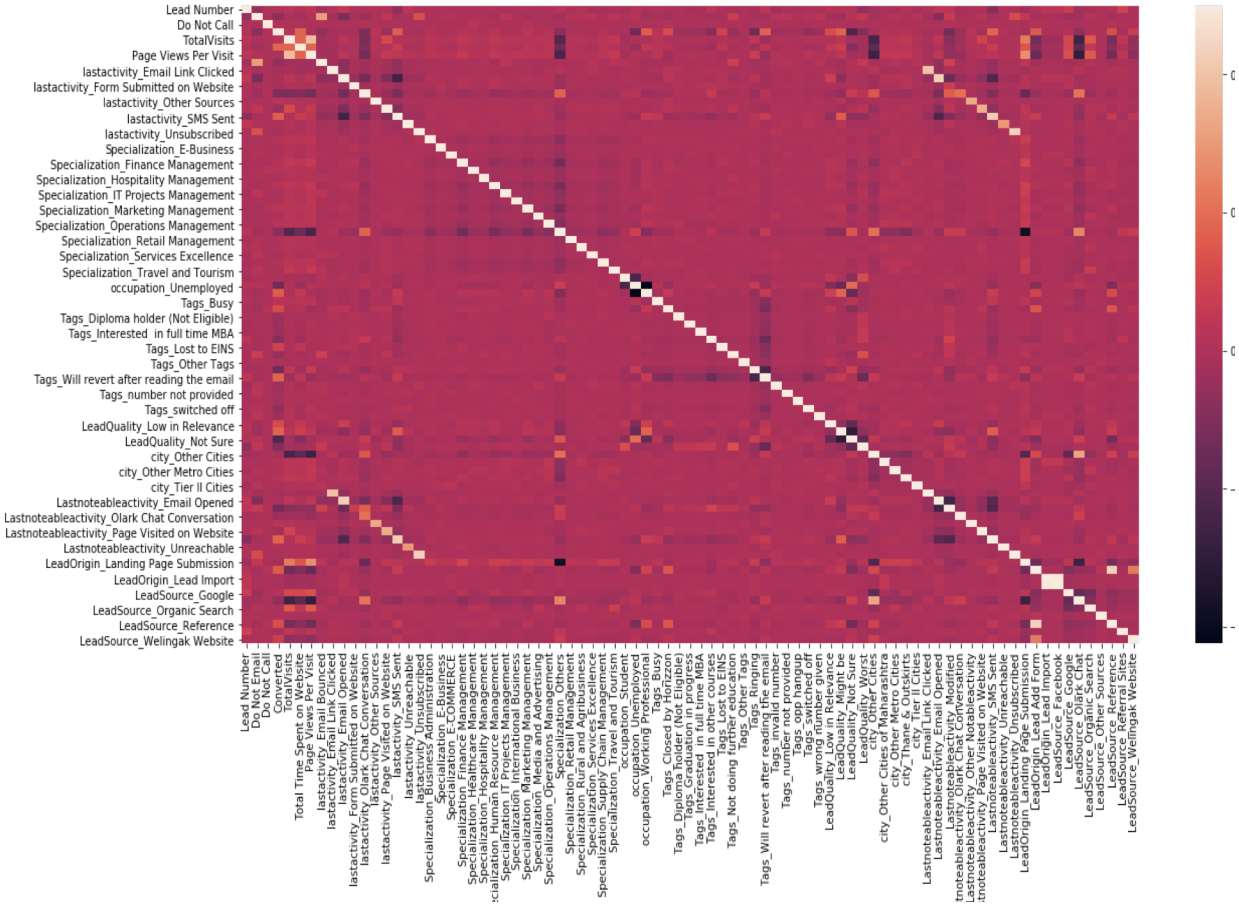
Inspection of Data Pre and Post feature reduction is as be under:

	Pre	Post	Lost in Cleaning
Row	9,074	9,074	-
Columns	31	17	14

Data Preparation - Cont'd

Dummy Variables, Train and Test Dataset

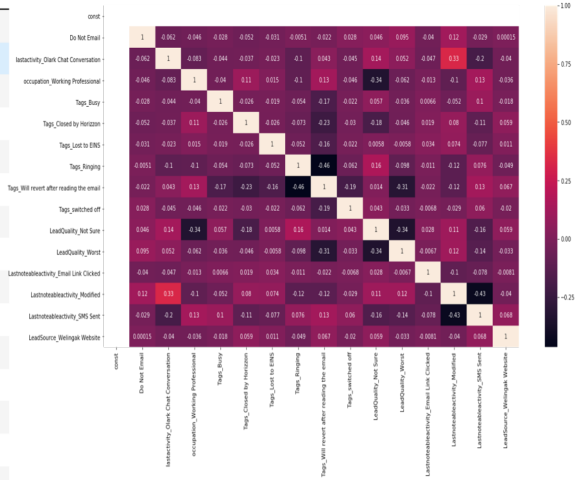
- Converted all the Categorical variables into Dummy variables
- Subsequent to conversion of Categorical variables into Dummy variables, the data now has 84 features (increase from 17 features prior to the process)
- Correlation matrix shows few have high correlation (based on light color)
- Dataset is Split into Train and Test Dataset ratio of 70:30
- Scaled the continuous variables on Train Dataset
- Dropped 2 Identifier Columns and Moved the "Response" variable to another Dataset



Model Building

- Based on RFE we picked 20 Features.
- After 5 iterations of dropping 1 feature at a time, since they were insignificant, we arrived at the model (illustrated aside)
- 15 features selected the VIF's were less than 5 indicating no Multicollinearity.

	Features	VIF
14	LeadSource_Welingak Website	3.46
9	LeadQuality_Not Sure	3.32
8	Tags_switched off	1.65
1	lastactivity_Olark Chat Conversation	1.33
5	Tags_Lost to EINS	1.27
2	occupation_Working Professional	1.25
0	Do Not Email	1.17
3	Tags_Busy	1.13
12	Lastnoteableactivity_Modified	1.13
7	Tags_Will revert after reading the email	1.12
4	Tags_Closed by Horizzon	1.11
6	Tags_Ringing	1.08
10	LeadQuality_Worst	1.05
13	Lastnoteableactivity_SMS Sent	1.03
11	Lastnoteableactivity_Email Link Clicked	1.01

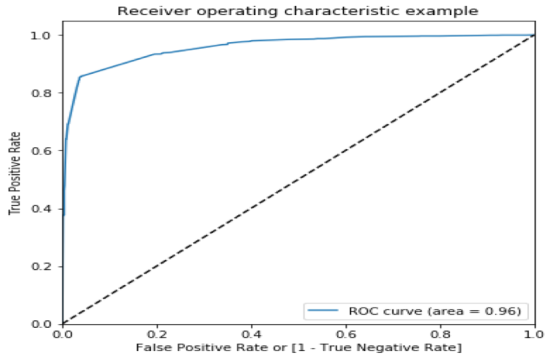


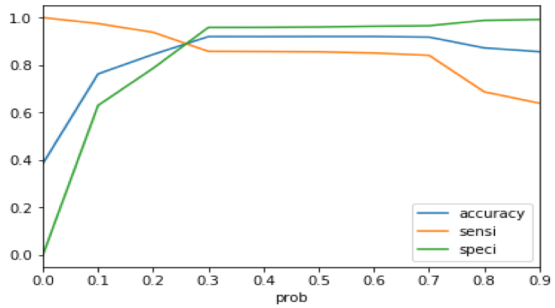
- Type of Model – Logistic Regression/ GLM
- Model Family – Bionomial
- API– Statsmodel
- Feature Selection Method – RFE

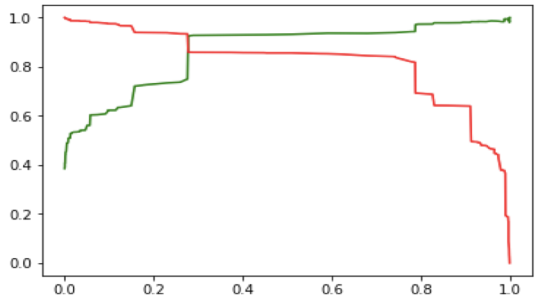
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5151	0.220	-6.872	0.000	-1.947	-1.083
Do Not Email	-1.2731	0.217	-5.879	0.000	-1.698	-0.849
lastactivity_Olark Chat Conversation	-1.0749	0.192	-5.610	0.000	-1.450	-0.699
occupation_Working Professional	1.3313	0.291	4.579	0.000	0.761	1.901
Tags_Busy	3.6020	0.326	11.065	0.000	2.964	4.240
Tags_Closed by Horizzon	8.2464	0.763	10.810	0.000	6.751	9.741
Tags_Lost to EINS	9.2286	0.757	12.190	0.000	7.745	10.712
Tags_Ringing	-1.8074	0.334	-5.415	0.000	-2.462	-1.153
Tags_Will revert after reading the email	3.8485	0.228	16.882	0.000	3.402	4.295
Tags_switched off	-2.4117	0.585	-4.125	0.000	-3.558	-1.266
LeadQuality_Not Sure	-3.2965	0.129	-25.551	0.000	-3.549	-3.044
LeadQuality_Worst	-3.8061	0.810	-4.696	0.000	-5.395	-2.218
Lastnoteableactivity_Email Link Clicked	-0.9865	0.372	-2.653	0.008	-1.715	-0.258
Lastnoteableactivity_Modified	-0.7709	0.115	-6.705	0.000	-0.996	-0.546
Lastnoteableactivity_SMS Sent	2.2747	0.127	17.966	0.000	2.027	2.523
LeadSource_Welingak Website	4.3047	0.742	5.802	0.000	2.851	5.759

Model Building/ Evaluation

- Based on the model obtained with 15 features, we determined the **optimal point as 0.3**:
  - ROC Curve shows trade-off between true positive rate against the false positive rate. Based on which a probability of 0.3 seems reasonable
  - Also, plotted the accuracy, sensitivity, and specificity on graph, which also points a probability of 0.3 seems reasonable
  - Plotting a graph between Precision and Recall also determines 0.30 as optimal point
  - We assigned a Lead Score based on probability. (Probability of Conversion *100 = Lead Score)
  - **Lead Score more than 30% was considered as "Hot Lead", else a "Cold lead"**







Model Evaluation

- Based on Optimal Cut off point of **0.30**, we populated the Lead Score and for Lead Score greater than 30, predicted lead to Convert.

- Based on that key metrics are as under:

Metrics	%	Explanation
Accuracy	92%	Accuracy of predicting conversion or non Conversion
Sensitivity/ Recall	86%	Probability correctly detecting of conversion
Specificity	96%	Proportion of actual 'Non Conversion' that are correctly identified as such
Positive Predictive Value/ Precision	93%	Probability that a predicted 'Yes' is actually a 'Yes'
Negative Predictive Value	91%	Probability that a predicted 'No' is actually a 'No'

The metrics look healthy and hence the model is considered good.

Model Evaluation

- Based on Optimal Cut off of 0.30, we made the prediction on Test Database and the results of Key Metrics are as under:

Metrics	Train Data	Test Data
Accuracy	92%	91%
Sensitivity/ Recall	86%	84%
Specificity	96%	96%
Positive Predictive Value/ Precision	93%	92%

The metrics look good and hence the model is considered good.

Recommendation

- Optimal Cut off Probability Point as 0.30
- If the probability of conversion is **more than 30%**, we consider that as **“Hot Lead”**
- If the probability of conversion is **Less than 30%,** we consider that as **“Cold Lead”**

Higher Probability of Conversion	Lower Probability of Conversion
Tagged as Lost to EINS	Lead Quality is Worst
Tagged as Closed by Horizzon	Lead Quality is Not Sure
Lead Source is from Welingak Website	Tagged as phone switched off
Tagged as Will revert after reading the email	Tagged as Phone Ringing
Tagged as Busy	Do Not Email is Opted