LEAD SCORING CASE-STUDY

- R.MADHAVAN
- VIDULA AROLKAR

PROBLEM STATEMENT

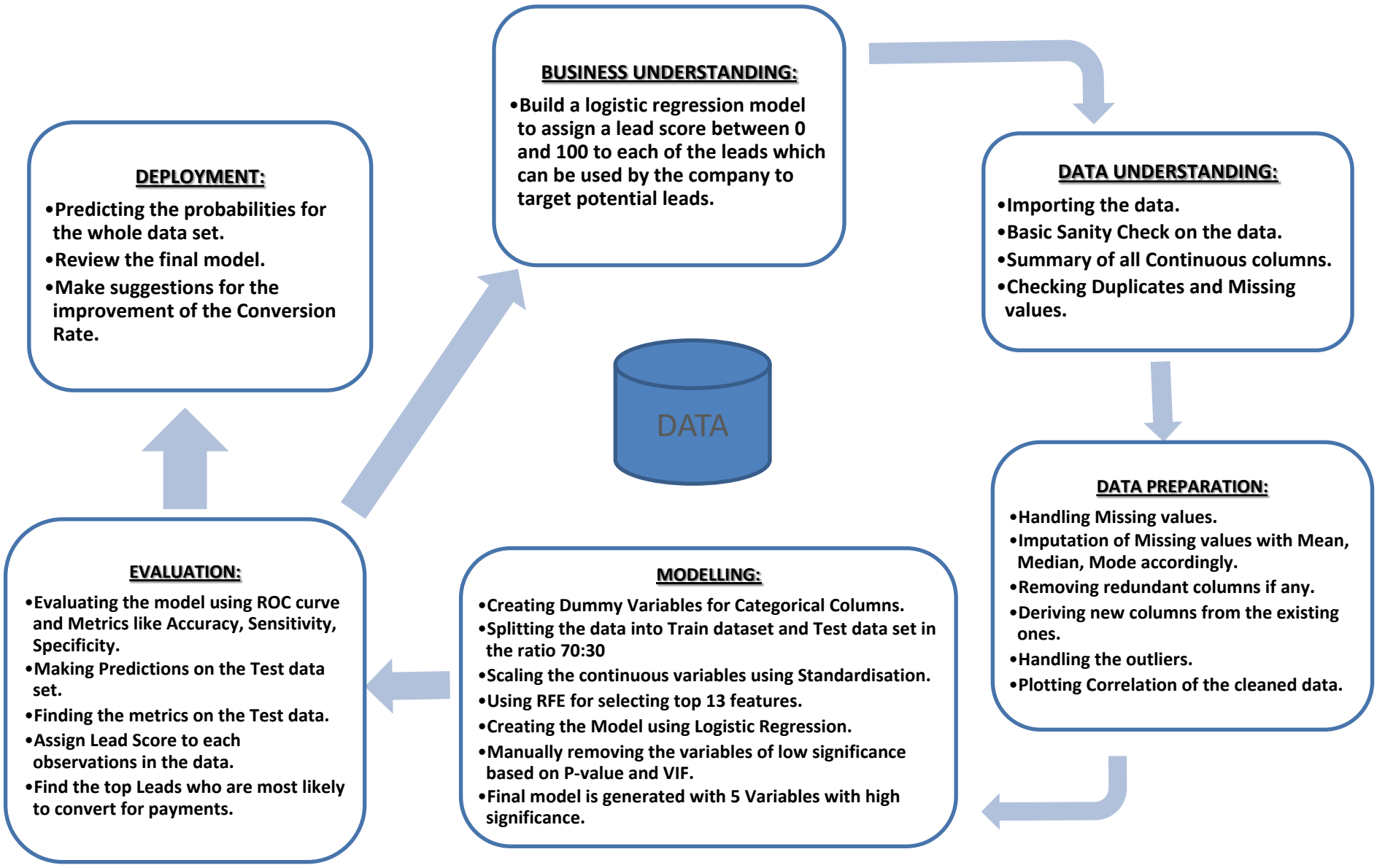
PROBLEM STATEMENT

- 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.
- But its lead conversion rate is very poor.
- To achieve efficiency in this process , they need to find out the potential leads, who may convert for payments.
- Hence they can target those leads rather than calling everyone.

OBEJCTIVES

- To 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.
- To help the company to select the most promising leads.
- The model should be able to adjust to if the company's requirement changes in the future.

SOLUTION METHODOLOGY

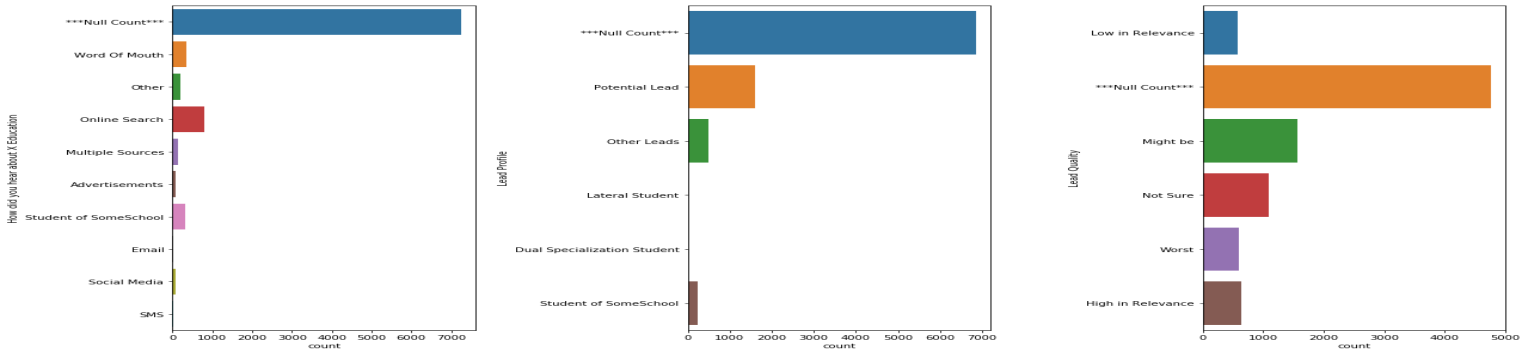


DATA PREPARATION

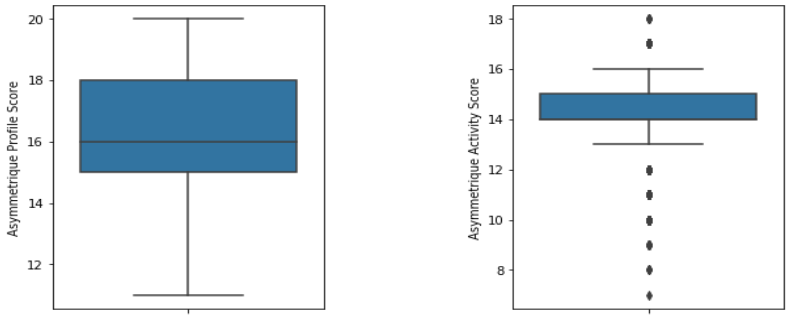
❑ HANDLING DATA QUALITY ISSUES:

- *It is found that there are many missing values in the data.*
- *Also, 4 columns have 'Select' as one of the value .*
- *After looking at the data dictionary and business perspective we can figure out that 'Select' values are actually null values . (a lead did not fill any of the options)*
- *We replaced all the 'Select Values with 'NaN'.*
- *Performed some Univariate Analysis on the Data and finally decided to drop those columns as it may create Bias.*

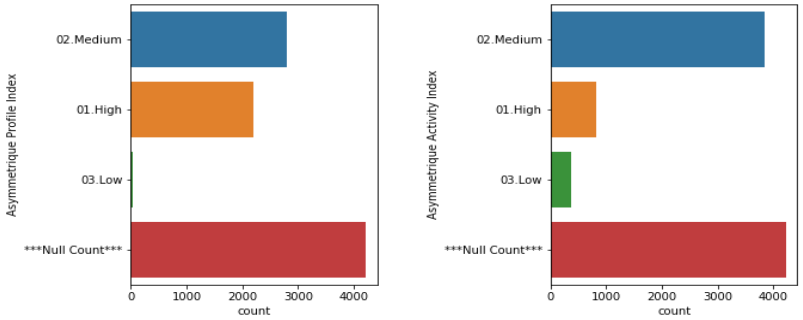
Univariate Categorical (Including null values) - For variables having above 50% values missing



Spread of the features related to Asymmetrique Profile and Activity

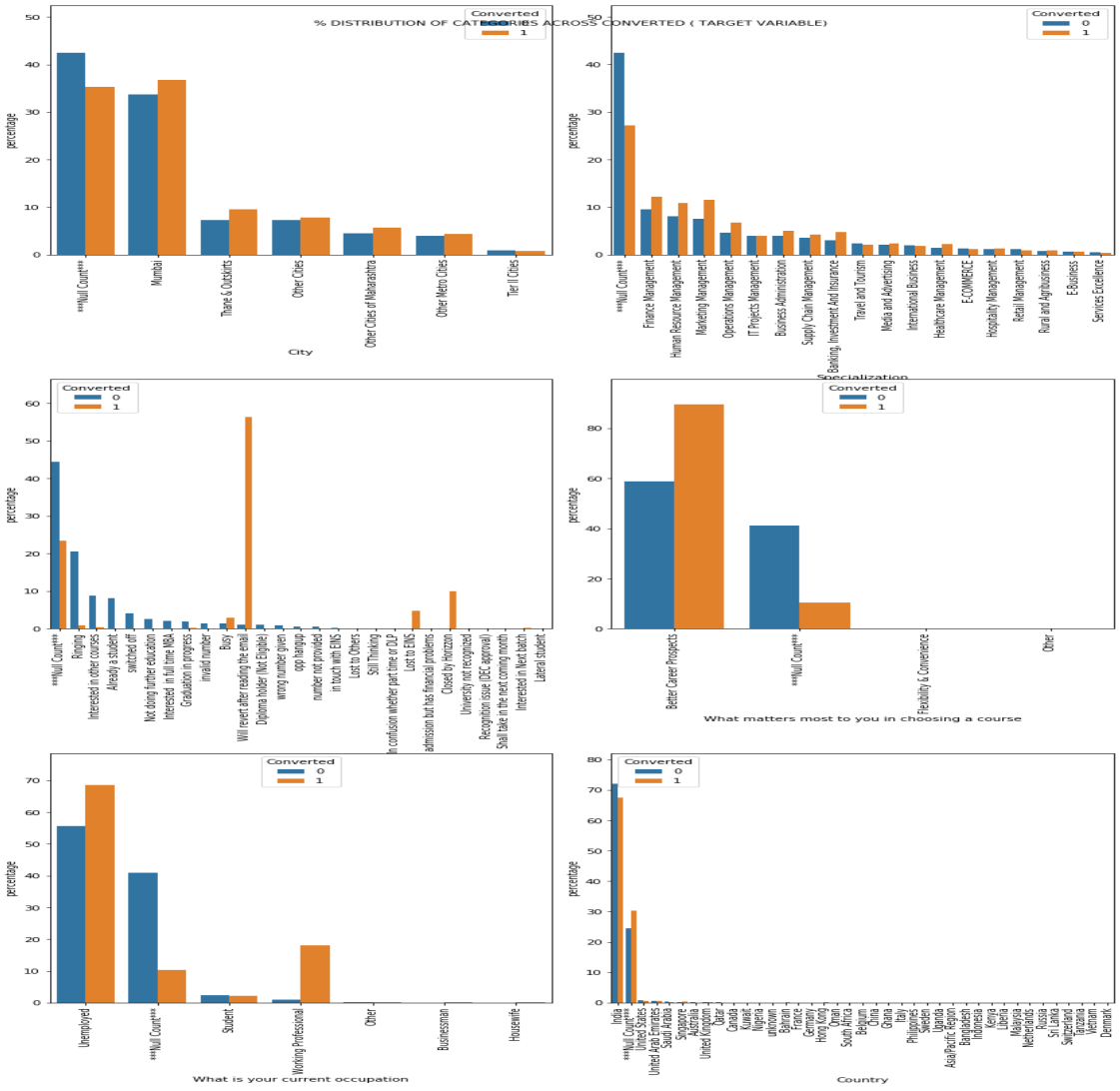


Univariate(categorical) - Asymmetrique Profile and Activity Index



❑ *SEGMENTED UNIVARIATE ANALYSIS:*

- ***Performed Segmented Univariate Analysis on data with more than 25% missing values.***
- ***This will help us to see how these categories are behaving for people who have 'Converted' and people who have not converted.***



From this analysis, we derived 2 new variables:

1. Status_reply after email.
This variables has only 2 sub-categories: 'Will revert after reply' and 'Others'

2.What_matters_BETTER_CAREER.
The Null Values and other sub-categories are combined as 'Others'.

3.

3. Rest all variables like City, Country, What is your current occupation , Specialization are dropped as missing values are more.

4. From the above plots , we get a lot of interesting insights. Almost 60% of the people who have 'Converted' replied back saying ' Will respond after reading the emails'.

ANALYSIS FOR BUILDING 2 MODELS :

- ***Tags' is a column which is generated after the sales team gets in touch with the lead . But it is observed to be the most crucial indicator for 'conversion'.***

- ***Also the chief goal of this project is to pick up 'hot leads' from the initial pool of leads so that the sales team can focus more on 'nurturing' the 'hot leads' rather than focusing on everyone. In such case we can use the power of the 'tags' variable to fit a better model and assign lead score to the current data.***

- ***But what if this model is used in the future? What if the sales team wishes to target the 'hot leads' directly without even getting in touch with the initial pool of leads. It is also important that we a build a model that can get adjusted to future requirements.***

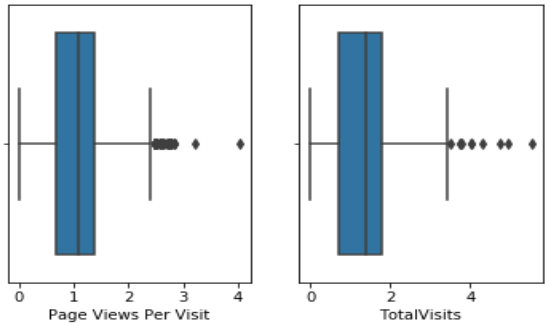
MODEL A focuses on prediction where we intend to build a model which has 'Tags' in it , see the variables which has positive effect on the lead conversion and assign lead scores for the current data.

MODEL B focuses on forecasting for the future . 'Tags' feature is not present in this. Here our goal is to project for the future. We are not concerned about how many features are being used (model can involve negative coefficients).

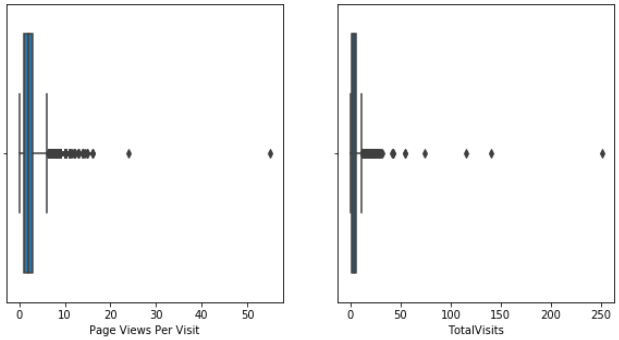
❑ *OUTLIERS TREATMENT:*

- ***It is found that there are some columns with almost 1 % of missing values.***
- ***When we plot the spread of these Numerical data, it is found that there are some outliers.***
- ***The distribution is Right Skewed.***
- ***Also removing those outliers may lead to loss of important data and hence we are scaling the data using Log- Scale.***

Spread of the features related to Visits - (AFTER LOG-TRANSFORMATION)



Spread of the features related to Visits - (RIGHT SKEWED)



❑ *REMOVING REDUNDANT COLUMNS:*

- ***Some columns have only 1 unique value and these columns does not show any variance. So dropping this column is a better decision.***
- ***Columns with 2 unique values are converted to 0 and 1 using Label encoder.***
- ***Based on the imbalance in this data, only 4 columns were found to be useful.***
  1. ***A free copy of Mastering The Interview***
  2. ***Do Not Email***
  3. ***Status_reply after email***
  4. ***What_matters_BETTER_CAREER***

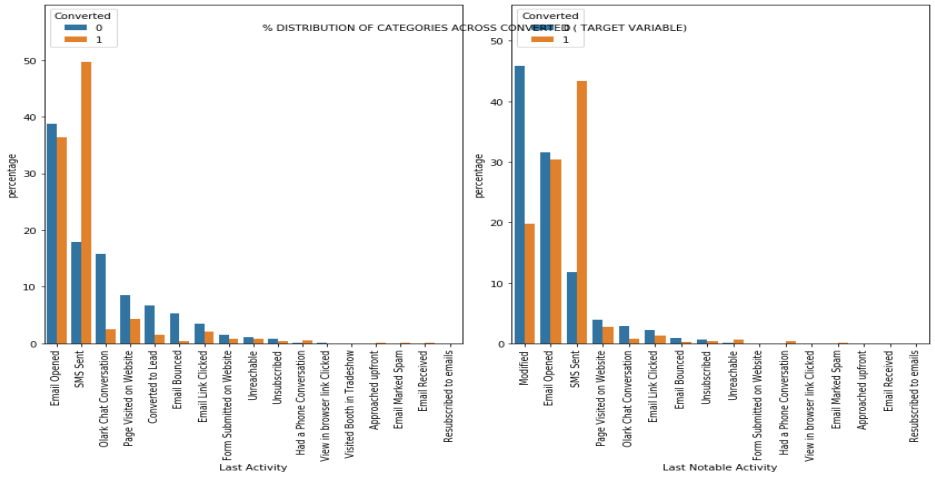
❑ GROUPING SUB-CATEGORIES:

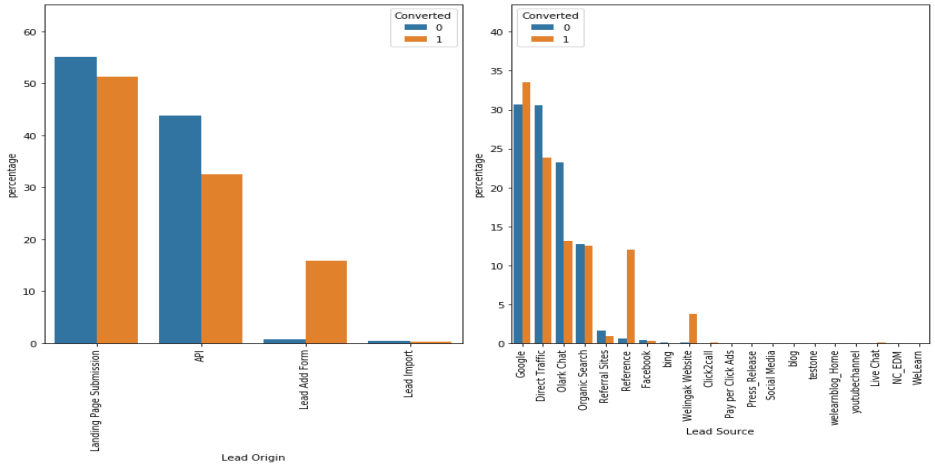
- *Some of the columns have various sub-categories with low frequency.*
- *It is better to group those sub-categories into a single one.*
- *Hence, for variables like Last Activity, Lead Source and last Notable Activity, we group the last sub-categories into a group called 'Others'.*
- *Now, There are no Outliers in 'Total Time spent on Website.*

❑ DATA PREPARATION FOR MODEL- B:

We created 3 new Variables : Temp 1 , Temp 2, Temp 3

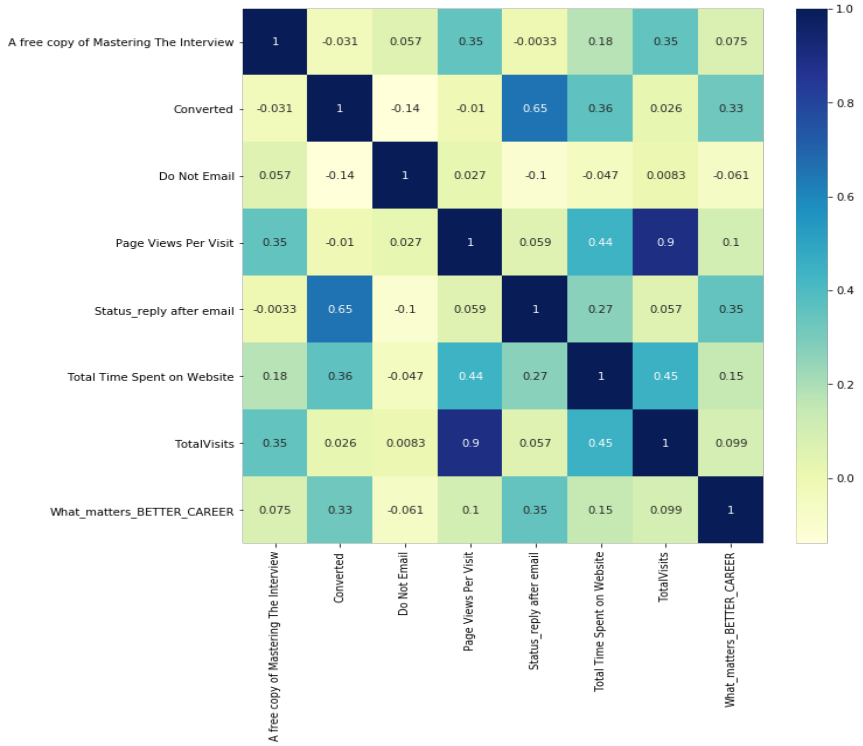
- *Temp 1: The Null values in Specialization are replaced with 'Info not provided.*
- *Temp 2: Replacing Null values of Last Activity with 'Others'*
- *Temp 3: Grouping this sub-categories ('SMS Sent',' Had a Phone Conversation', 'Unreachable') into 'sms_phone_contact '.*





DATA RETAINED AFTER CLEANING:

HEATMAP SHOWING THE CO-RELATION B/W ALL NUMERICAL FEATURES

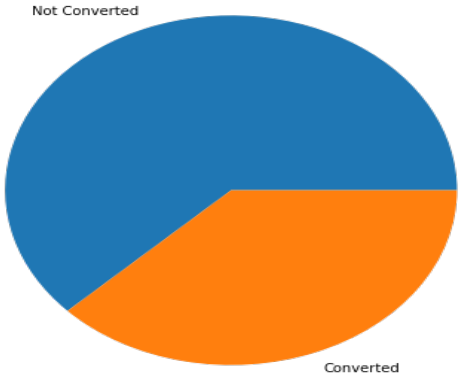


- *There were 9240 rows and 37 columns before cleaning.*
- *After data cleaning and preparation , we are left with 9035 rows and 14 columns.*
- *Hence , we have lost almost 1 % of the rows and some features as well.*
- *This is the Correlation plot of the remaining data.*

IMBALANCE IN THE TARGET VARIABLE:

- *We can see that , there is an imbalance in the Target Variable.*
- *Only 37% of the persons are Converted.*
- *We need to increase the conversion rate for growth of the company.*

Pie-Chart showing DATA IMBALANCE (target variable)



MODELLING- LOGISTIC REGRESSION

DUMMY VARIABLES

Dummy variables are created for all the Categorical Columns.

Remove the dummy column created for a sub-category where '1' is minimal comparing to all other sub-categories under a particular category(column)

TRAIN-TEST SPLIT

The data is split into Train and Test in the ratio 70:30.

The Model is trained on the Train data set and evaluated on the test data set.

SCALING

Scaling is performed for having all the variables on same scale.

Normalisation

(Min-Max Scaler) is used for Scaling.

FEATURE ELIMINATION USING RFE:

- *Recursive Feature Elimination is used to find the best performing features in the data set.*
- *We have a lot of variables now due to dummy creation.*
- *Through RFE , top 13 variables are selected , which are suitable for model building and which truly explains the model.*

```
from sklearn.linear_model import LogisticRegression  #importing Class LogisticRegression
logreg = LogisticRegression()
```

```
from sklearn.feature_selection import RFE
rfe = RFE(logreg, 13)             # running RFE with 13 variables as output
rfe = rfe.fit(X_train, y_train)
```

The following are the features picked using RFE

```
col = X_train.columns[rfe.support_]
col
```

```
Index(['Do Not Email', 'Page Views Per Visit', 'Status_reply after email',
       'Total Time Spent on Website', 'TotalVisits',
       'Last Activity_Converted to Lead', 'Last Activity_Email Bounced',
       'Last Activity_Olark Chat Conversation',
       'Last Notable Activity_Email Bounced',
       'Last Notable Activity_sms_phone_contact',
       'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
       'Lead Source_Welingak Website'],
      dtype='object')
```

PROCESS AND ANALYSIS:

- *We used Stats model and Generalised Linear Model (GLM) for creating the model.*
- *After creating the model with these 13 variables, we manually removed some variables on the basis of P-value and VIF.*
- *Any Variable having P-value more than 0.05 or VIF greater than 5 are dropped.*
- *Also some Variables had negative coefficients, which will have a negative impact on the final decision.*
- *Our goal is to get those factors which can improve the model conversion rate, hence those variables with negative coefficients are dropped.*

MODEL AND ROC CURVE- MODEL A

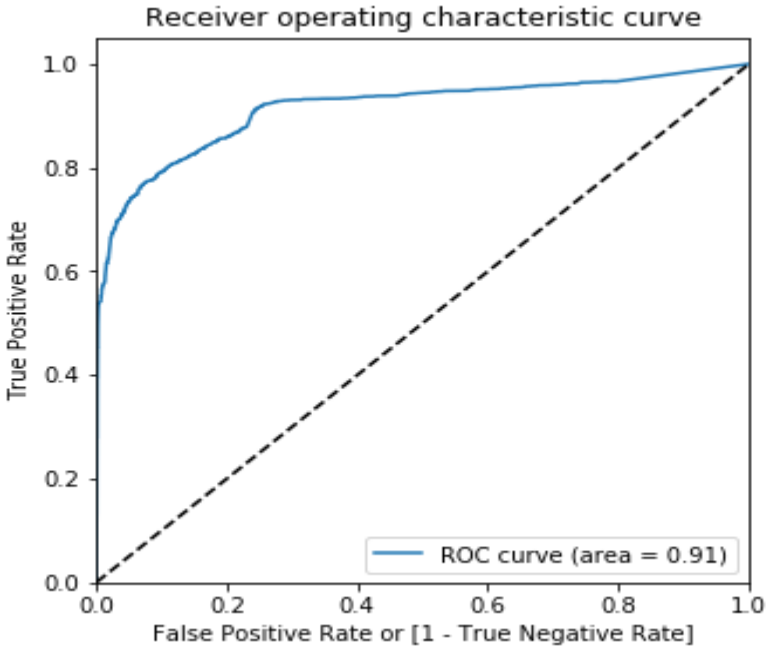
Here is our final model A with good significance value.
Also, there is no multicollinearity in the variables.
The final variables are :

- ***Last Notable Activity_sms_phone_contact***
- ***Lead Origin_Lead Add Form***
- ***Lead Source_Welingak Website***
- ***Status_reply after email***
- ***Total Time Spent on Website .***

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6324
Model:	GLM	**Df Residuals:**	6318
Model Family:	Binomial	**Df Model:**	5
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2043.5
Date:	Sun, 19 Apr 2020	**Deviance:**	4086.9
Time:	18:36:46	**Pearson chi2:**	7.39e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Last Notable Activity_sms_phone_contact	1.4554	0.091	16.023	0.000	1.277	1.633
Lead Origin_Lead Add Form	3.9643	0.240	16.521	0.000	3.494	4.435
Lead Source_Welingak Website	1.9497	0.756	2.579	0.010	0.468	3.431
Status_reply after email	4.5043	0.162	27.789	0.000	4.187	4.822
Total Time Spent on Website	3.7579	0.165	22.739	0.000	3.434	4.082
const	-2.7723	0.070	-39.448	0.000	-2.910	-2.635

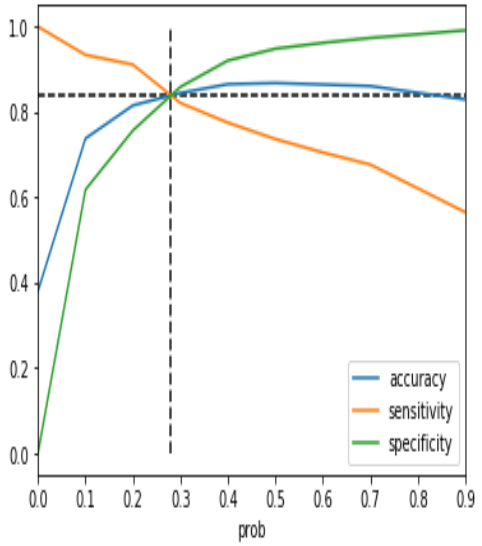


- ***After predicting the probabilities using the Final Model, the metrics like TPR and FPR are plotted.***
- ***Receiver Operating Characteristics (ROC) curve is used for this purpose.***
- ***We can see that the Area Under Curve(AUC) is 0.91, which is a very good estimate for Accuracy.***
- ***The Curve is quite steeper which is an indicator of a good model.***

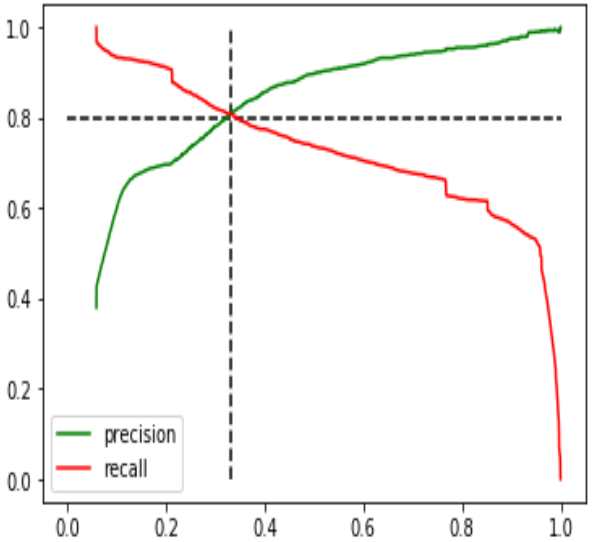
OPTIMAL THRESHOLD VALUE:

- *We calculated the Accuracy, Sensitivity, Specificity for different cut-off values ranging from 0 to 0.9.*
- *Based on the Sensitivity –Specificity trade off, the optimal threshold values comes out to be 0.28*
- *Based on Precision –Recall Trade-off , the optimal cut off is found to be 0.33.*
- *The goal of the case study is to filter 'hot leads' such that 80% of such lead have higher probability of conversion which will ensure that our conversion rate will go up*
- *Hence, lets choose our threshold value as 0.28 according to sensitivity-specificity tradeoff*

Plot showing optimal cut-off point based on Sensitivity-specificity tradeoff



Plot showing the Tradeoff between precision and recall



MODEL - A

- ***Threshold value: 0.28***
- ***Accuracy for Train Data: 0.838***
- ***Sensitivity for Train Data: 0.835***
- ***Specificity for Train Data: 0.839***

EVALUATING THE MODEL ON WHOLE DATA:

- *The complete data was scaled using .transform()*
- *Conversion Probabilities are predicted and multiplied by 100 to get the Lead Score.*
- *The Lead scores are mapped to YES (HOT LEAD) and NO (COLD LEAD)according to the threshold value(0.28)*
- *The final Hot Leads are identified by sorting the Lead Scores.*

LEAD SCORE=100*CONVERSION PROBABILITY

INFERENCE:

- *We can see that Lead Score with 99.99 are the SUPER HOT Leads and hence we need to focus on these leads first.*
- *The chances of these people to Convert are higher .*
- *After evaluating the model on Train and Test data, it is found that the Accuracy (84 %) is almost same for both the data sets. This means the Converted Persons are correctly identified.*
- *Sensitivity of Train data is 83% and Test data is 80% , both are closer values*

	Lead Number	Lead Score	HOT LEAD
6243	601868	99.99	YES
8120	587853	99.99	YES
2011	640191	99.99	YES
7234	593962	99.99	YES
6937	596446	99.99	YES
...	...	...	...
5411	608365	33.11	YES
6548	599567	33.07	YES
9145	580323	33.04	YES
116	659345	33.00	YES
3175	629593	33.00	YES

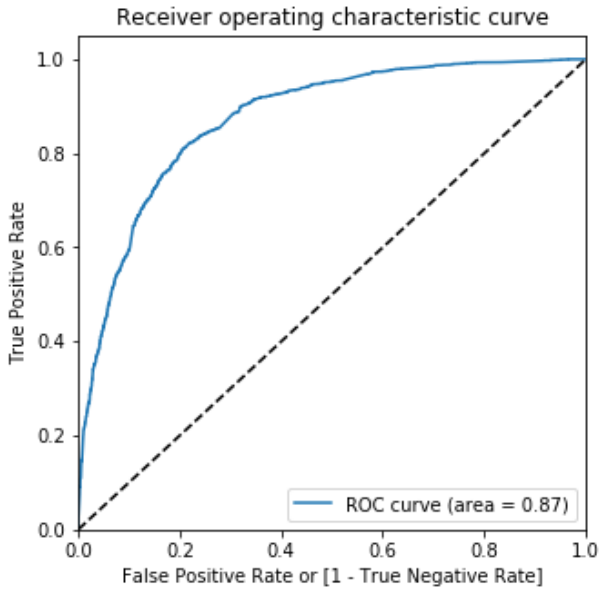
MODEL AND ROC CURVE- MODEL B

This is our final model B with good significance value.

Also, there is no multicollinearity in the variables.

The final variables are :

- *Do Not Email*
- *Last Activity_Other*
- *Last Notable Activity_Modified*
- *Last Notable Activity_ Olark Chat Conversation*
- *Lead Origin_Landing Page Submission*
- *Lead Source_Reference*
- *Lead Source_Welingak Website*
- *Total Time Spent on Website .*
- *What_matters_BETTER_CAREER*



Dep. Variable:	Converted	**No. Observations:**	6324
Model:	GLM	**Df Residuals:**	6314
Model Family:	Binomial	**Df Model:**	9
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2736.5
Date:	Mon, 20 Apr 2020	**Deviance:**	5473.1
Time:	16:48:00	**Pearson chi2:**	6.31e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

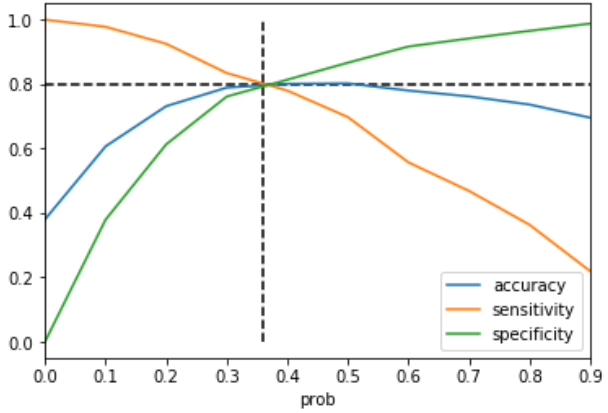
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0732	0.104	-10.303	0.000	-1.277	-0.869
Do Not Email	-1.4467	0.170	-8.486	0.000	-1.781	-1.113
Last Activity_Other	-1.2474	0.073	-17.106	0.000	-1.390	-1.105
Last Notable Activity_Modified	-0.9278	0.077	-12.034	0.000	-1.079	-0.777
Last Notable Activity_Olark Chat Conversation	-1.1570	0.338	-3.419	0.001	-1.820	-0.494
Lead Origin_Landing Page Submission	-0.6662	0.074	-8.975	0.000	-0.812	-0.521
Lead Source_Reference	3.4635	0.235	14.712	0.000	3.002	3.925
Lead Source_Welingak Website	4.5079	0.724	6.226	0.000	3.089	5.927
Total Time Spent on Website	4.2434	0.151	28.090	0.000	3.947	4.539
What_matters_BETTER_CAREER	1.3100	0.087	15.073	0.000	1.140	1.480

- *We used the same RFE technique to eliminate the features.*
- *Manually removed some variables on the basis of P-value and VIF.*
- *Used Normalization for scaling.*
- *Plotted the ROC curve for threshold value.*
- *The precision-recall trade off and Sensitivity-Specificity trade off is also plotted for MODEL-B*

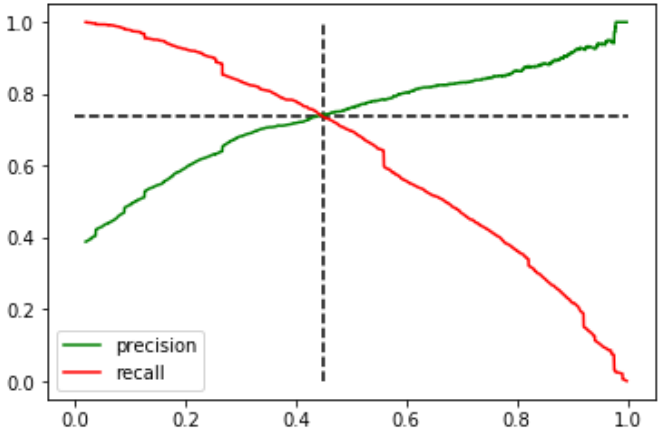
SOME METRICS FOR MODEL-B

- ***The goal of this model-B is building a stable model where all the metrics are to their optimum best.***
- ***We need to take care of accuracy score first. In a general case scenario the cut-off point can be determined using sensitivity and specificity tradeoff.***
- ***But if the requirements of the future change and according to business requirement at that period the cut-off point can be changed***

Plot showing optimal cut-off point based on Sensitivity-specificity tradeoff



Plot showing the Tradeoff between precision and recall



MODEL - B

- ***Threshold value: 0.36***
- ***Accuracy for Train Data: 0.80***
- ***Sensitivity for Train Data: 0.80***
- ***Specificity for Train Data: 0.79***

PREDICTIONS ON THE TEST DATA(MODEL A):

- *The Test data set is scaled using .transform() method.*
- *The probabilities are predicted for the Test Data.*
- *Using the same threshold value (0.28), the leads are assigned labels as 1 0r 0 (1= Converted, 0=Not Converted)*
- *Confusion Matrix is generated and the metrics are computed.*

PREDICTIONS ON THE TEST DATA(MODEL B)

- *The Test data set is scaled using .transform() method.*
- *The probabilities are predicted for the Test Data.*
- *Using the same threshold value (0.35)*
- *Confusion Matrix is generated and the metrics are computed to test the model efficiency*

MODEL A

Accuracy of Test Data: 0.84
Sensitivity of Test Data: 0.80
Specificity of Test Data: 0.85

MODEL B

Accuracy of Test Data: 0.78
Sensitivity of Test Data: 0.77
Specificity of Test Data: 0.79

CONFUSION MATRIX FOR TEST DATA(MODEL A):

	NOT CONVERTED	CONVERTED
NOT CONVERTED	1438	254
CONVERTED	200	819

INFERENCES & RECOMMENDATIONS

- ***Sensitivity is a measure which tells us the probability of positives being correctly identified out of the real positives. In this period we want to contact as many leads as possible without missing a slightest chance of conversion. So, here even if we call to a lead with low probability of conversion it is fine but we don't want to miss out anyone.***
- ***We decrease the probability as low as possible. According to our model 0.1 is a good cut off to target maximum leads and ensure high sensitivity. But, it also depends on how many leads are the interns capable of targeting in total and then we can decide the cut off accordingly.***
- ***In case the company does not intend to spend time on making calls. False Positive rate is a measure which tells the probability of a label being predicted as 'yes' when it is actually 'no'. We want False Positive rate to be as low as possible in other words we want specificity to be as high as possible***
- ***So, under such circumstances our specificity needs to be close to 1. This can be achieved by making the cut-off value of label assignment higher. Here, we may miss out some true potential leads but in this situation we can afford it***
- ***When a lead tells 'will reply after checking the email' there is a very high probability that he will be converted. The executives and company should take necessary measures to always follow up with marketing calls to such leads.***
- ***The team needs to keep the websites up to date or can make websites full of contents to attract customers.***
- ***The person who fills a 'form' is likely to get converted. Targeting them is necessary because the Lead Origin is Lead Add Form.***
- ***We need to focus on maintaining any Automated SMS or Emails for follow ups.***
- ***There are many important variables like city, specialization , occupation which can potentially explain Conversion better.***
- ***. It is important for the management to make few of these information mandatory to fill , so that we can use them better understand the 'customer' mindset and take build important decisions for the business***