Lead Scoring Case Study

Submitted by:

Saikat Chowdhury

Aman Diwakar

Contents

PROBLEM STATEMENT

1. X Education sells online courses to industry professionals.
2. 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.
3. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
4. 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.

Lead Conversion Process -Demonstrated as a funnel

BUSINESS OBJECTIVE

1. Increasing the Lead Conversion rate from around 30% to around 80%.
2. Current Lead conversion is around 30%.
3. Building the right model to identify and classify the most potential leads tagged as ”Hot Leads”.
4. To help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.
5. The conversion rate from the “Hot Leads” should be around 80%.
6. The model should be adjustable to include company’s requirement changes.
7. To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads.

SOLUTION APPROACH

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 large amount of missing values and 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 variable etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

Feature Scaling & Dummy Variables and encoding of the data.

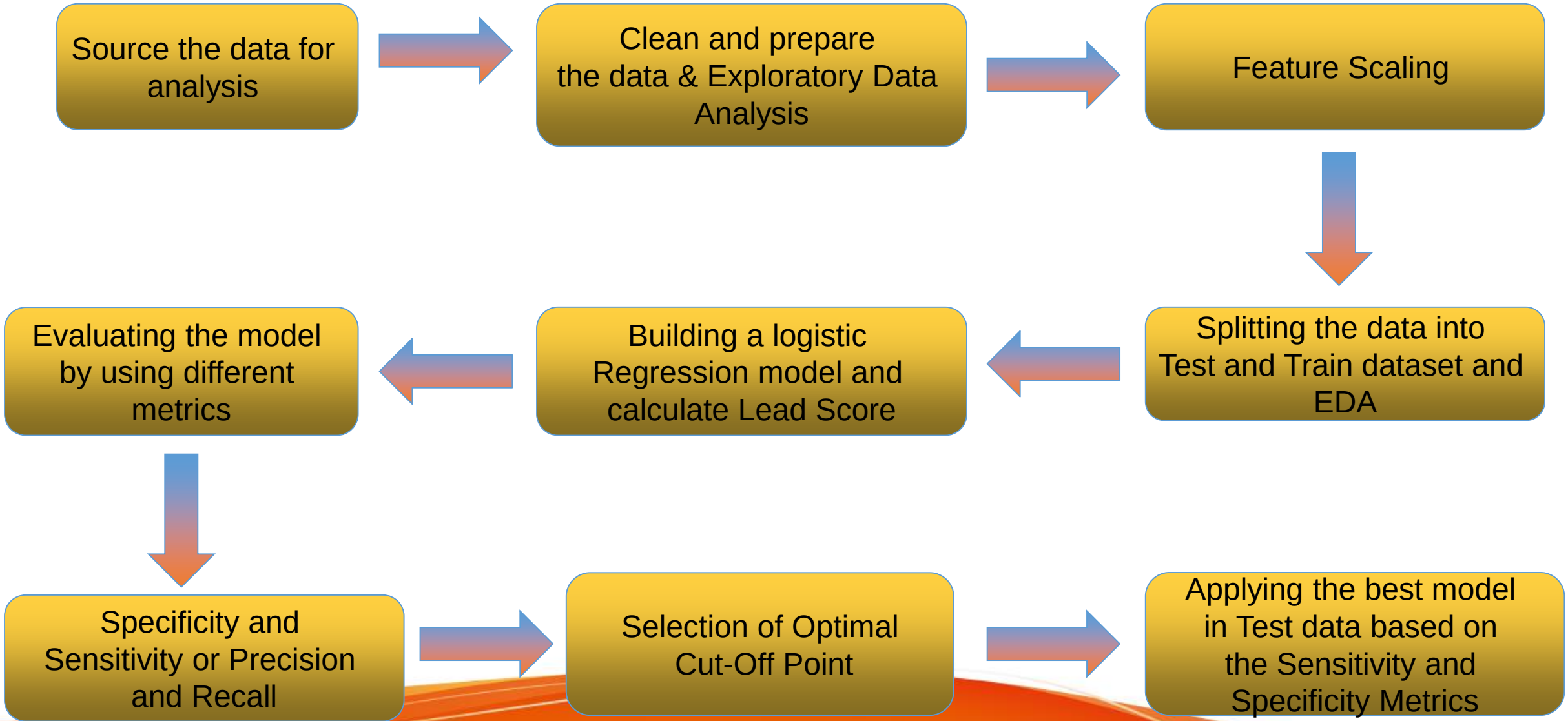
Classification technique: logistic regression used for the model making and prediction.

Validation of the model.

Model presentation.

Conclusions and recommendations.

METHODOLOGY



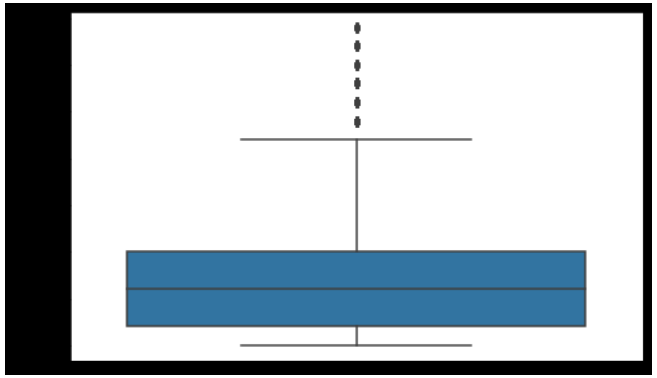
DATA SOURCING, CLEANINGAND PREP

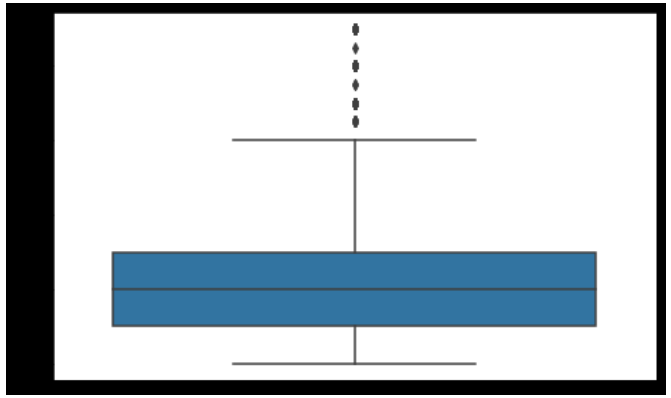
- The data available to us was in a csv format and had 9240 entries with 37
- columns each.
- Python libraries were used for the same.
- Data was checked for duplicate entries and none were found.
- Missing values in the data were dealt with using various methods like
  1. Replacing the missing values with mode for categorical data and median or mean for numerical data.
  2. Columns with very high percentage of data which did not have much business relevance were dropped.

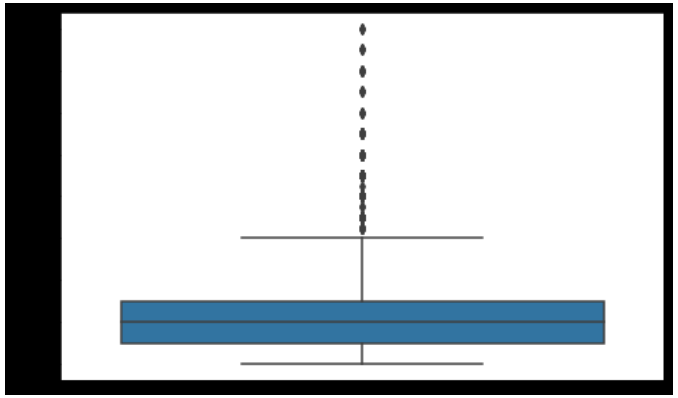
DATA SOURCING, CLEANINGAND PREP

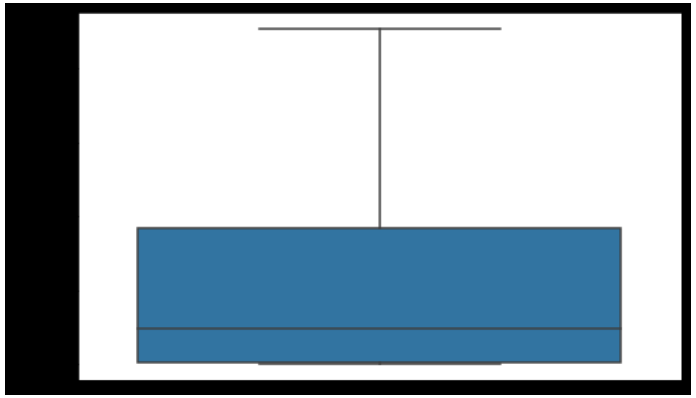
- Outlier treatment of data was done on the basis of percentiles of the data range.
- Boxplot was made to see this graphically.
- Example can be seen here -

1. It can be seen that outliers exist in the columns TotalVisits and Page Views Per Visit columns.



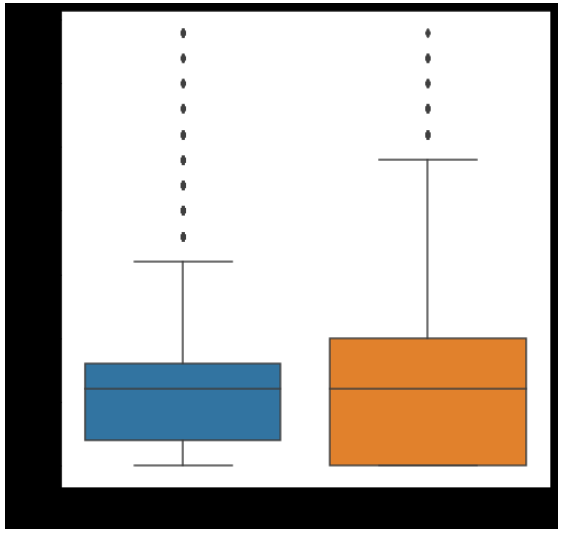


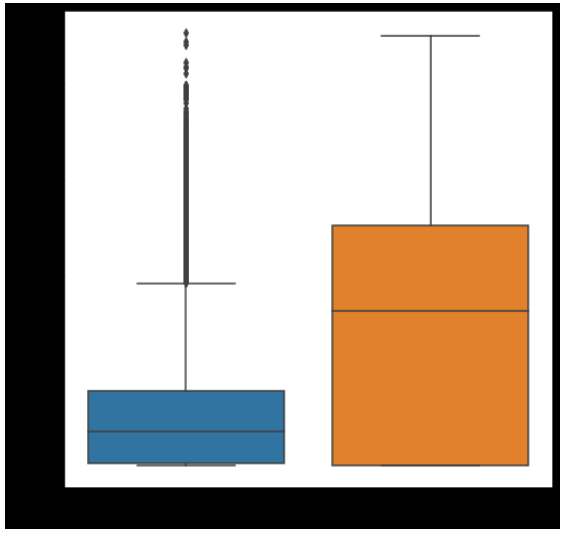


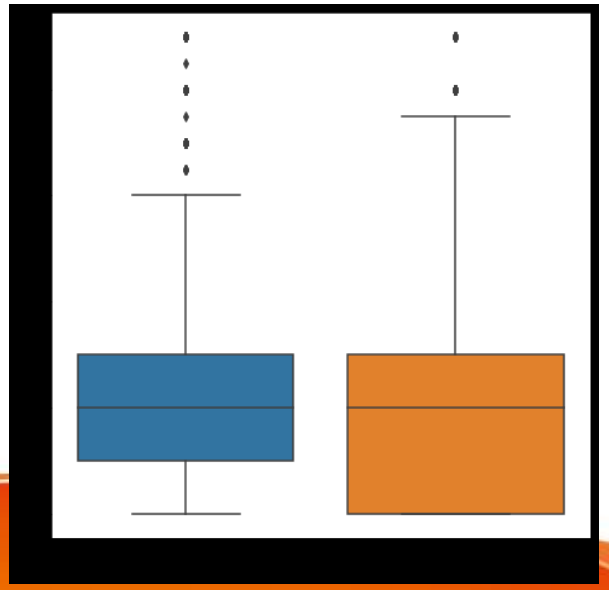


DATA SOURCING, CLEANINGAND PREP

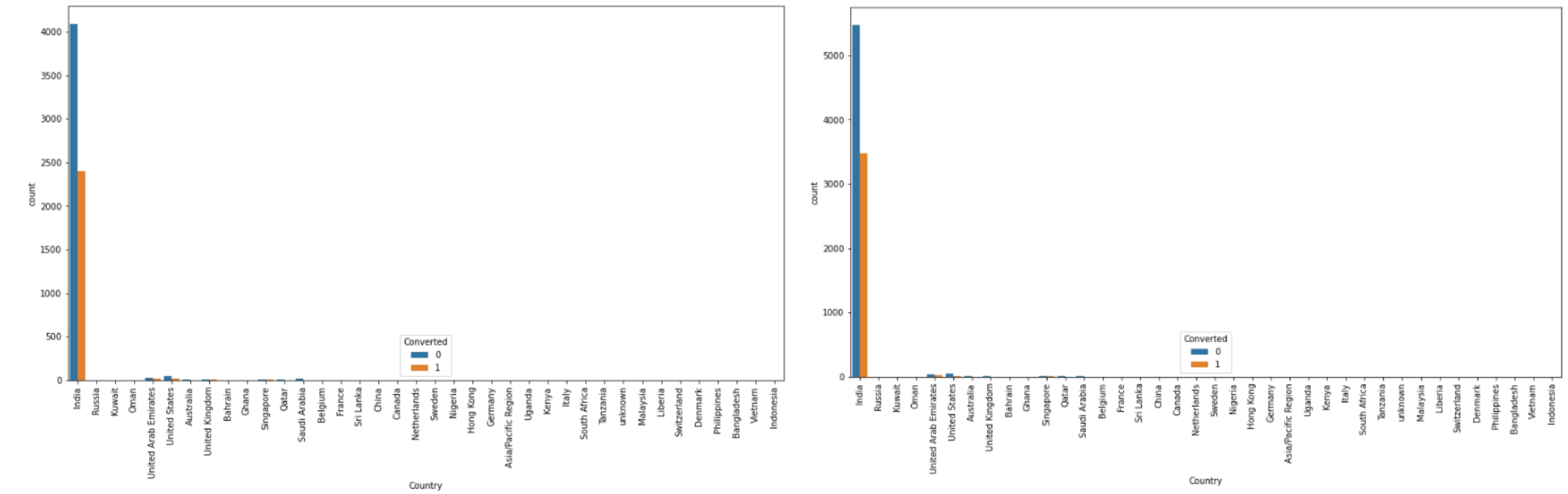
- Some outlier data is deleted which may skew our results by making our model less accurate.
- Some columns which will not contribute to our analysis have been dropped.
- Example - What matters most to you in choosing a course, columns that have more than 40% null values have been dropped, etc.
- After cleaning data 98% data has been retained.
- There are 8953 entries and 14 columns.





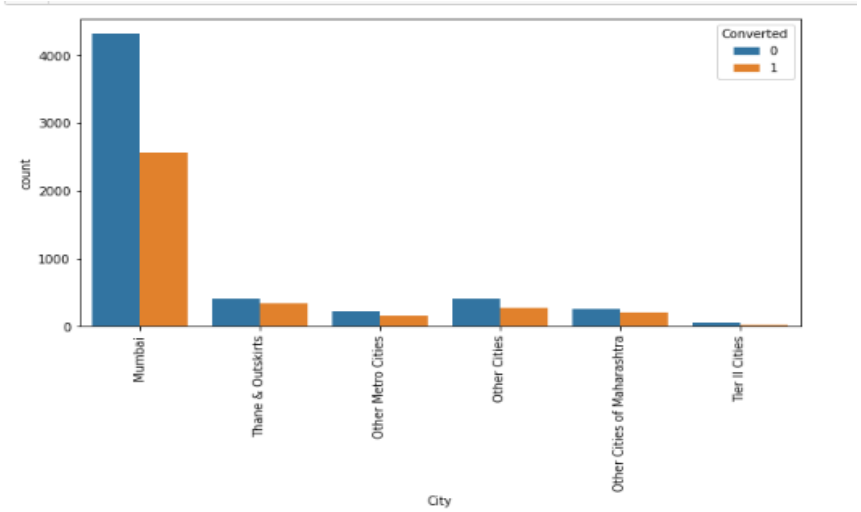


EXPLORATORY DATA ANALYTICS

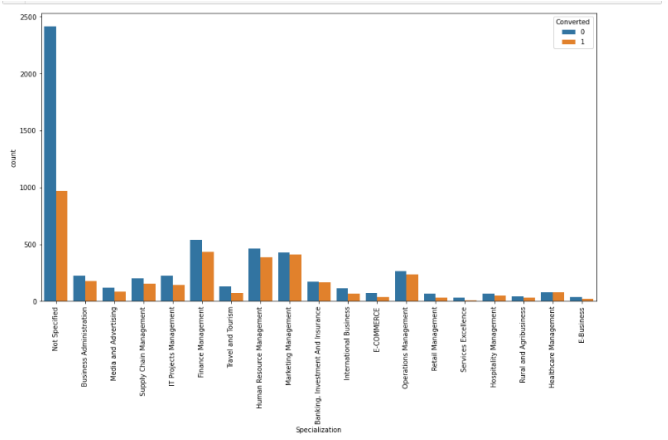


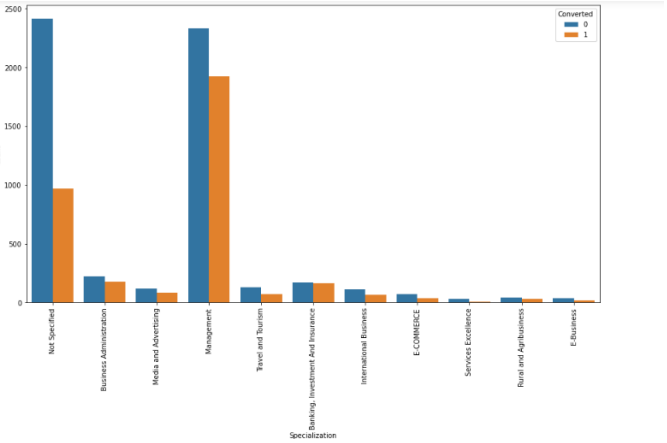
Visualizing the Country column

UNIVARIATE & BIVARIATE ANALYSIS



Visualizing City column after replacing NaN values



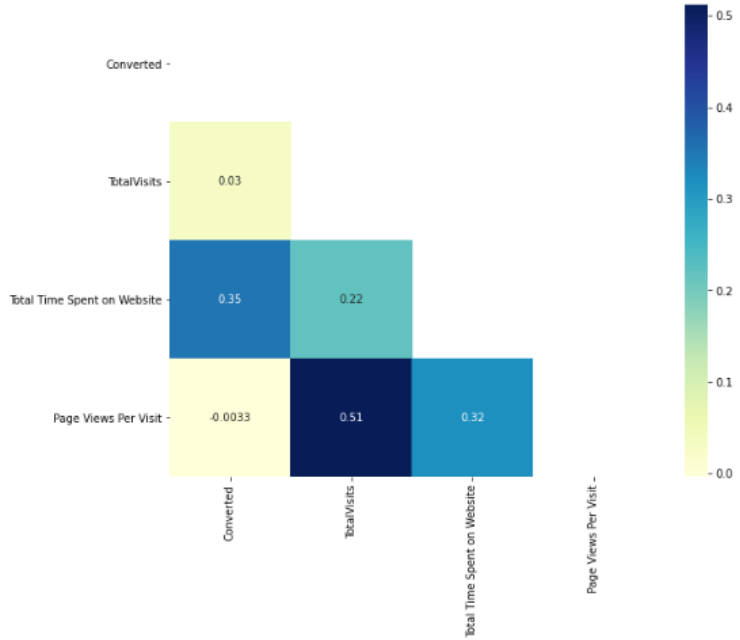


Visualizing the Specialization columnn

MULTIVARIATE ANALYSIS

-Pairplots and heatmaps were created to understand the correlation between the numerical variables.

- Total Visits and Page Views Per Visit has a strong correlation. In such a case, one may be dropped before analysis starts.



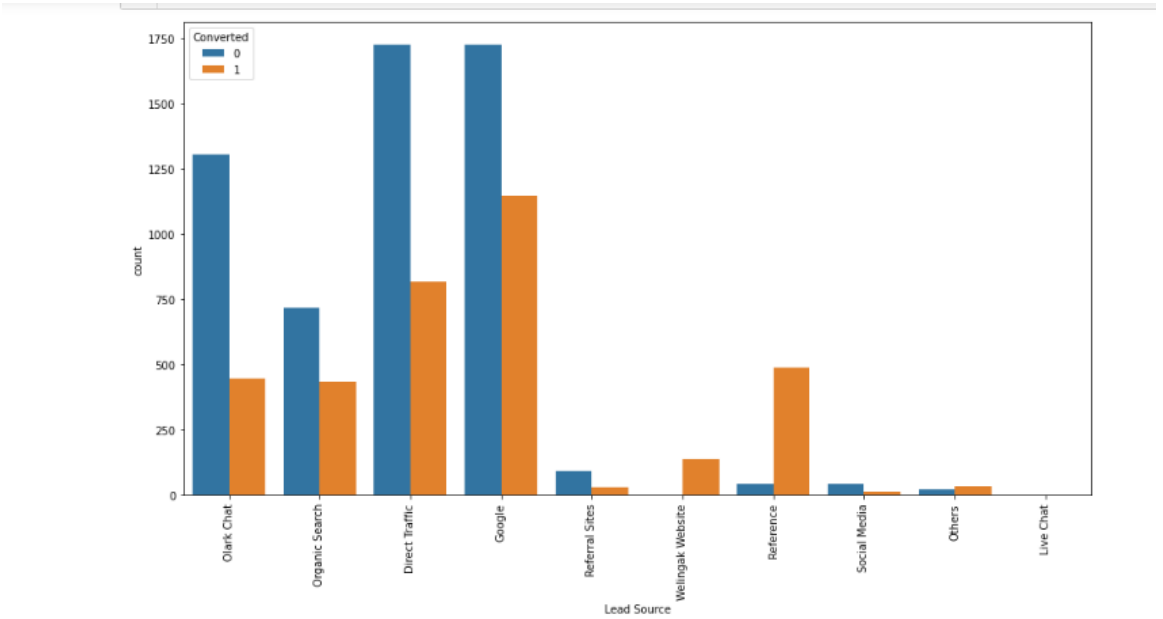
Checking correlations of numeric values

DATA CLEANING BASED ON EDA RESULTS

CONVERTED VARIABLE -

-This is also our target variable. Indicates whether a lead has been successfully converted or not.

- The count of leads converted is almost half of the leads that are not converted



Google and Direct traffic generates Maximum number of leads
Reference and welingak website lead source have high conversion rate.

MODEL BUILDING

-After EDA, Logistic Regression Model is built in python using GLM() function, under statsmodel library.
-The model contained all the variables, some of which had insignificant coefficients.
-Such variables are removed using Automated Approach: RFE (Recursive feature elimination) with number of features = 15.
-Manual approach based on VIFs and p values.
-The final tally of variables with their respective values
Significant p-values near to zero
VIFs < 3

Out[101]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6267
Model:	GLM	Df Residuals:	6251
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1254.7
Date:	Sun, 07 Mar 2021	Deviance:	2509.3
Time:	14:09:28	Pearson chi2:	8.34e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1899	0.088	-13.480	0.000	-1.363	-1.017
Total Time Spent on Website	0.8970	0.053	16.999	0.000	0.794	1.000
Lead Origin_Lead Add Form	1.6712	0.450	3.714	0.000	0.789	2.553
Lead Source_Direct Traffic	-0.8320	0.129	-6.471	0.000	-1.084	-0.580
Lead Source_Referral Sites	-0.5284	0.465	-1.138	0.255	-1.439	0.382
Lead Source_Welingak Website	3.9043	1.110	3.518	0.000	1.729	6.079
Last Activity_SMS Sent	1.2373	0.223	5.555	0.000	0.801	1.674
Last Notable Activity_Modified	-1.2839	0.150	-8.532	0.000	-1.579	-0.989
Last Notable Activity_Olark Chat Conversation	-1.7123	0.490	-3.496	0.000	-2.672	-0.752
Last Notable Activity_SMS Sent	1.0151	0.257	3.943	0.000	0.511	1.520
Tags_Closed by Horizzon	6.9834	1.019	6.853	0.000	4.986	8.981
Tags_Interested in other courses	-2.1641	0.407	-5.321	0.000	-2.961	-1.367
Tags_Lost to EINS	5.7302	0.608	9.419	0.000	4.538	6.923
Tags_Other_Tags	-2.4417	0.210	-11.633	0.000	-2.853	-2.030
Tags_Ringing	-3.5858	0.243	-14.752	0.000	-4.062	-3.109
Tags_Will revert after reading the email	4.4263	0.185	23.989	0.000	4.065	4.788

[101]:

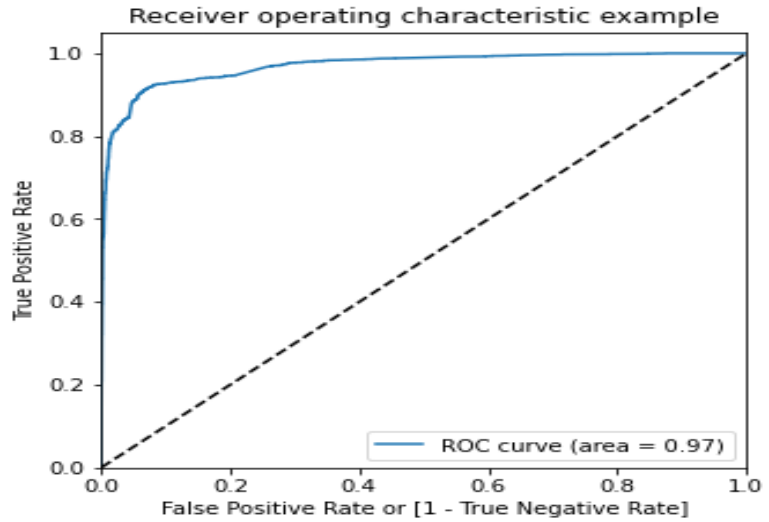
	Features	VIF
1	Lead Origin_Lead Add Form	1.82
12	Tags_Will revert after reading the email	1.56
4	Last Activity_SMS Sent	1.46
5	Last Notable Activity_Modified	1.40
2	Lead Source_Direct Traffic	1.38
3	Lead Source_Welingak Website	1.34
10	Tags_Other_Tags	1.25
0	Total Time Spent on Website	1.22
7	Tags_Closed by Horizzon	1.21
11	Tags_Ringing	1.16
8	Tags_Interested in other courses	1.12
9	Tags_Lost to EINS	1.06
6	Last Notable Activity_Olark Chat Conversation	1.01

VIF's values looks good now.

Model building 1

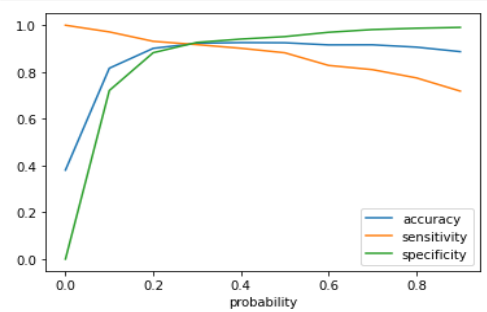
ROC CURVE

ROC Curve demonstrates tradeoff between sensitivity and specificity.
Closer the curve follows the left-hand border and then the top border of ROC space, the more accurate the test
Closer the curve comes to 45° diagonal of the ROC space, the less accurate the test.
For our model, ROC curve is towards the upper left corner, and area under the curve is more as displayed in figure. Thus, our model is an optimal choice to move forward with the analysis

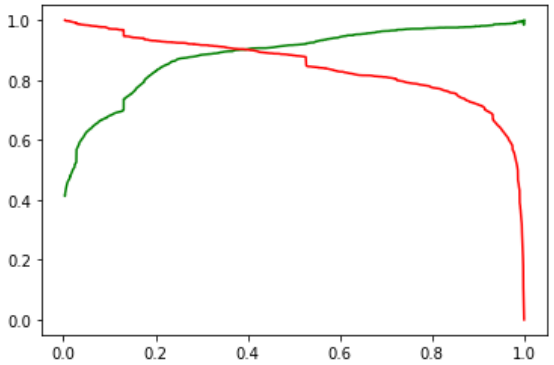


ROC curve(area = 0.97), which indicates a good predictive model.

Finding optimal cut off point



plotting the accuracy, sensitivity and specificity for various probabilities.



Our final model gave us the following performance metrics.

• We used Recall/Precision trade-off graph to derive the optimal threshold value.

cut-off point 0.3

	over all model accuracy	precision	Recall/ Sensitivity	Specificity
Train	92.29%	88%	91.69%	92.65%
Test	92.77%	89.15%	91.98%	93.25%

MODEL EVALUATION, COMPARISON & CONCLUSION

- Accuracy, Sensitivity and Specificity values of test set are around 92%, 91% and 93% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 92%
- Hence, overall this model seems to be good.

BUSINESS RECOMMENDATIONS - 1

The top three variables in model which contribute most towards the probability of a lead getting converted are -

1. Lead Origin_Lead Add Form
2. Tags_Will revert after reading the email
3. Last Activity_SMS Sent

BUSINESS RECOMMENDATIONS - 2

Aggressive workflow for converting leads -

1. High sensitivity implies that our model will correctly identify almost all leads who are likely to Convert.
2. It will do that by overestimating the Conversion likelihood.
3. To follow an aggressive workflow choose a lower threshold value for Conversion Probability. This will ensure the Sensitivity rating is very high which in turn will make sure almost all leads who are likely to Convert are identified correctly and the agents can make phone calls to as much of such people as possible
4. Phone calls must be done to people

CONCLUSION

1. The model is prepared for prediction of the conversion of the leads.
2. The probability values are generated by the model.
3. The cut off decided for the model is 0.3. All leads whose probability is generated above this threshold value can be classified as Hot Lead.

4. Main Variables that contribute to analysis are

    1. Total Time Spent on Website
    2. Lead Origin_Lead Add Form
    3. Lead Source_Direct Traffic

5. Specialization and Total time spent also predict the conversion rate.
6. Concrete conclusion cannot be made but suggestions can be given as the data is very less

Thank You