Lead Score_Case Study

Presented By Team:
M Veerabhadra Rao, R Venkata Sainath, Aishwarya Arun

Business Problem Understanding

Overview

- Industry professionals can purchase online courses from X Education, a company that provides education. Several experts who are interested in the courses visit their website on any given day and search for courses.
- On numerous websites and search engines like Google, the firm advertises its courses. When arriving at the website, these visitors may browse the courses, submit a form for the course, or watch some videos.
- These persons are categorized as leads when they fill out a form with their phone number or email address. Also, the business receives leads from earlier recommendations. Once these leads are obtained, sales team members begin calling, sending emails, etc.
- Some leads are converted during this procedure, but most are not. At X Education, the normal lead conversion rate is roughly 30%. In the beginning, a lot of leads are generated, but very few of them end up becoming paying clients

Problem Statement:

- You have been asked by X Education to assist them in choosing the leads that have the best chance of becoming paying clients. The business wants you to create a model in which you give each lead a lead score so that leads with higher lead scores have a better chance of converting, while leads with lower lead scores have a lesser chance of converting. The desired lead conversion rate has been estimated by the CEO to be in the range of 80%.

Goals and Objectives

- Create a logistic regression model to provide each lead a lead score between 0 and 100 that the business may use to target potential prospects. In contrast, a lower number would indicate that the lead is chilly and unlikely to convert, while a higher score would indicate that the lead is hot and most likely to convert.

Dataset characteristics & Approach

Leads.csv have the following fields:

Variables	Description
Prospect ID	A unique ID with which the customer is identified.
Lead Number	A lead number assigned to each lead procured.
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Do Not Email	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
Do Not Call	An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.
Converted	The target variable. Indicates whether a lead has been successfully converted or not.
TotalVisits	The total number of visits made by the customer on the website.
Total Time Spent on Website	The total time spent by the customer on the website.
Page Views Per Visit	Average number of pages on the website viewed during the visits.
Last Activity	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
Country	The country of the customer.
Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.
How did you hear about X Education	The source from which the customer heard about X Education.
What is your current occupation	Indicates whether the customer is a student, umemployed or employed.
What matters most to you in choosing this course	An option selected by the customer indicating what is their main motto behind doing this course.
Search	Indicating whether the customer had seen the ad in any of the listed items.
Magazine	
Newspaper Article	
X Education Forums	
Newspaper	
Digital Advertisement	
Through Recommendations	Indicates whether the customer came in through recommendations.
Receive More Updates About Our Courses	Indicates whether the customer chose to receive more updates about the courses.
Tags	Tags assigned to customers indicating the current status of the lead.
Lead Quality	Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead.
Update me on Supply Chain Content	Indicates whether the customer wants updates on the Supply Chain Content.
Get updates on DM Content	Indicates whether the customer wants updates on the DM Content.
Lead Profile	A lead level assigned to each customer based on their profile.
City	The city of the customer.
Asymmetrique Activity Index	An index and score assigned to each customer based on their activity and their profile
Asymmetrique Profile Index	
Asymmetrique Activity Score	
Asymmetrique Profile Score	
I agree to pay the amount through cheque	Indicates whether the customer has agreed to pay the amount through cheque or not.
a free copy of Mastering The Interview	Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.
Last Notable Activity	The last notable acitivity performed by the student.

Approach:

The framework of our analysis is as follows:

1. Reading and Understanding Data
2. Data Cleaning
3. Data Visualization and Outlier Treatment
4. Data Preparation and Feature Scaling
5. Model Building
6. Making Predictions on the Train Set
7. Making Predictions on the Test set
8. Conclusion

Data Sourcing, Reading, and Understanding Data

Data Sourcing:

- Leads.csv contains
  - RangeIndex: 9240 rows
  - Data columns: 37 entries from Prospect ID to Last Notable Activity
  - dtypes: float64(4), int64(3), object(30)

Reading and Understanding Data:

Here we tried to get the look and feel of the data, we observed the following things

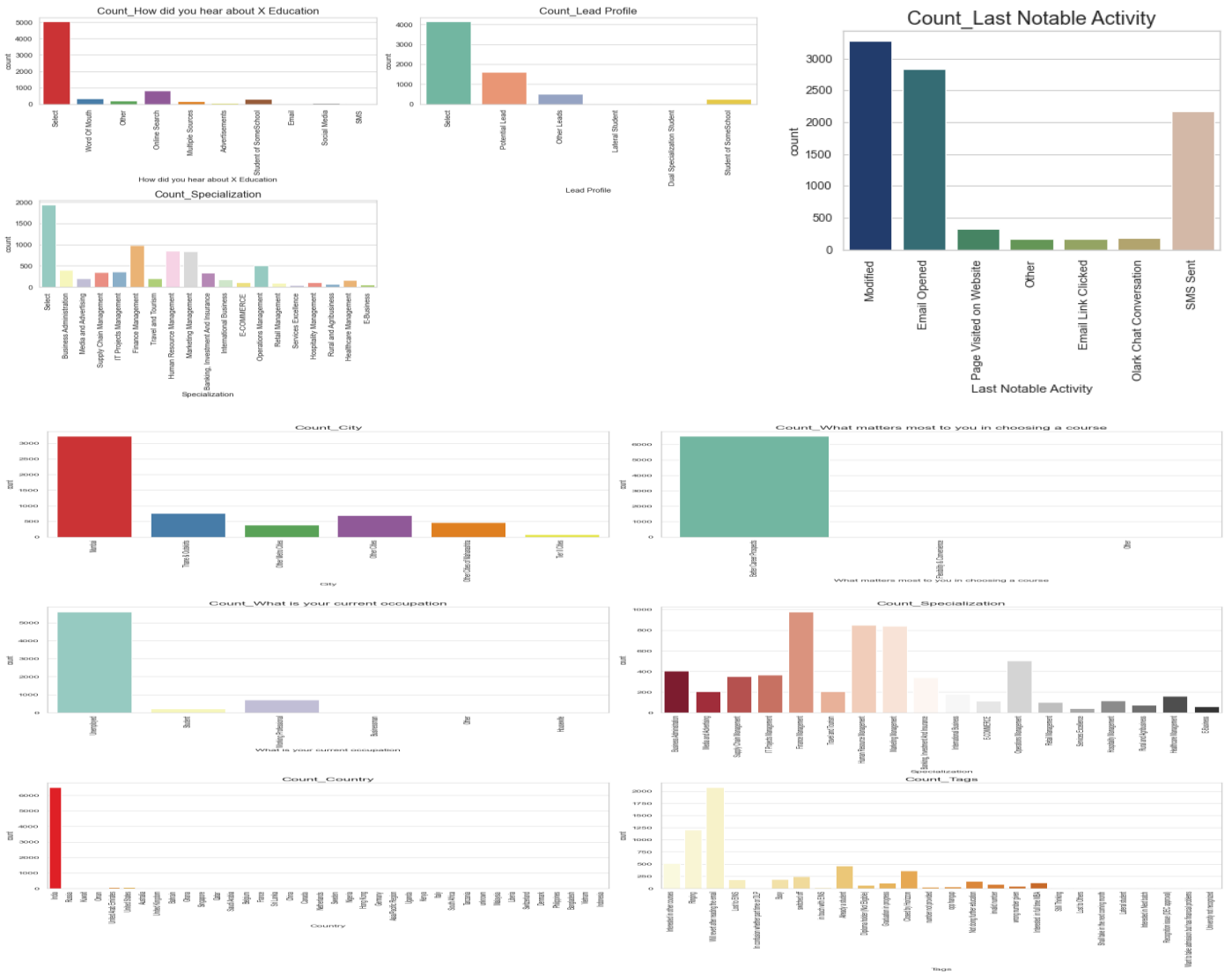
- The number of rows and columns
- Data types of each column.
- Checking the first few rows of how the data looks
- Checking how the data is spread and for duplicates if any.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object
 11  Country                                        6779 non-null   object
 12  Specialization                                 7802 non-null   object
 13  How did you hear about X Education             7033 non-null   object
 14  What is your current occupation                6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                         9240 non-null   object
 17  Magazine                                       9240 non-null   object
 18  Newspaper Article                              9240 non-null   object
 19  X Education Forums                             9240 non-null   object
 20  Newspaper                                      9240 non-null   object
 21  Digital Advertisement                          9240 non-null   object
 22  Through Recommendations                        9240 non-null   object
 23  Receive More Updates About Our Courses         9240 non-null   object
 24  Tags                                           5887 non-null   object
 25  Lead Quality                                   4473 non-null   object
 26  Update me on Supply Chain Content              9240 non-null   object
 27  Get updates on DM Content                      9240 non-null   object
 28  Lead Profile                                   6531 non-null   object
 29  City                                           7820 non-null   object
 30  Asymmetrique Activity Index                    5022 non-null   object
 31  Asymmetrique Profile Index                     5022 non-null   object
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object
 35  A free copy of Mastering The Interview         9240 non-null   object
 36  Last Notable Activity                          9240 non-null   object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

Data Cleaning

Here we checked for discrepancies in the dataset

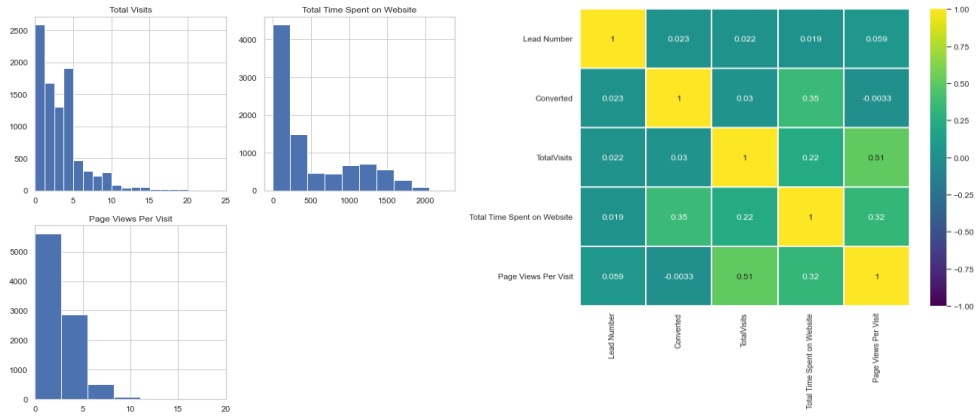
- The first step to clean the dataset we chose was to drop the variables having unique values.
- Then, there were a few columns with the value ‘Select’ which means the leads did not choose any given option. We changed those values to Null values.
- We dropped some columns having NULL values greater than 35%.
- Next, we removed the imbalanced and redundant variables. This step also included imputing the missing values as and where required with
  - We used mode imputation for categorical columns.
  - We used mean imputation for numerical columns if there is no skewness in the data.
  - We used median imputation for numerical columns if there is skewness in the data.

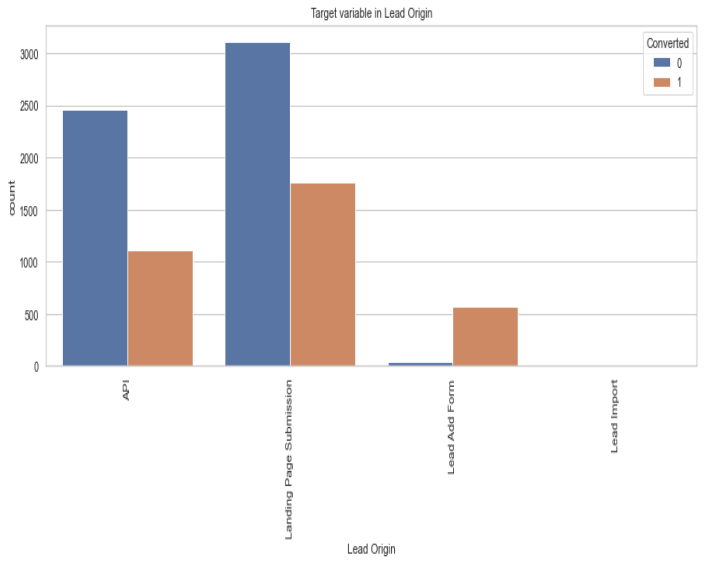


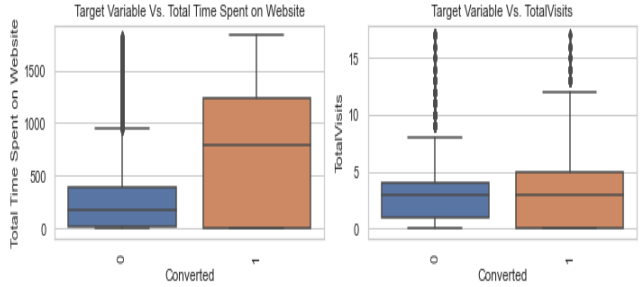
Data Visualization and Outlier Treatment

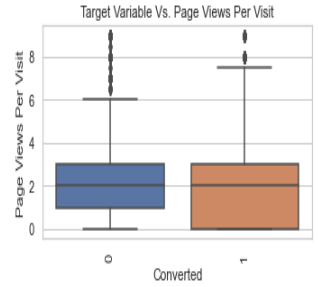
Here we have done Univariate and Bivariate analysis on the clean data for both categorical and numerical variables and checked for outliers in the dataset

- Here we checked for discrepancies in the dataset
- We used univariate analysis on the categorical column to determine which columns make more sense and eliminated those whose variance is almost zero.
- Additionally, a bivariate analysis was done on categorical columns to determine their variation w. r.t Converted column
- A univariate analysis was performed on numeric columns by drawing boxplots to see if there were outliers in the data. We used the IQR method to handle outliers in the dataset because only a few columns have outliers.
- To determine how the leads are related to these columns, we conducted a bivariate analysis on numerical columns with Converted columns.
- In this step, we also plotted the correlation matrix to identify the columns which are correlated.









Data Preparation and Feature Scaling

Feature Engineering:

At this stage, our data was very clean, and no outliers. We know that logistic regression takes the input parameters as numerical values. Hence, we converted all the categorical columns to numerical ones.

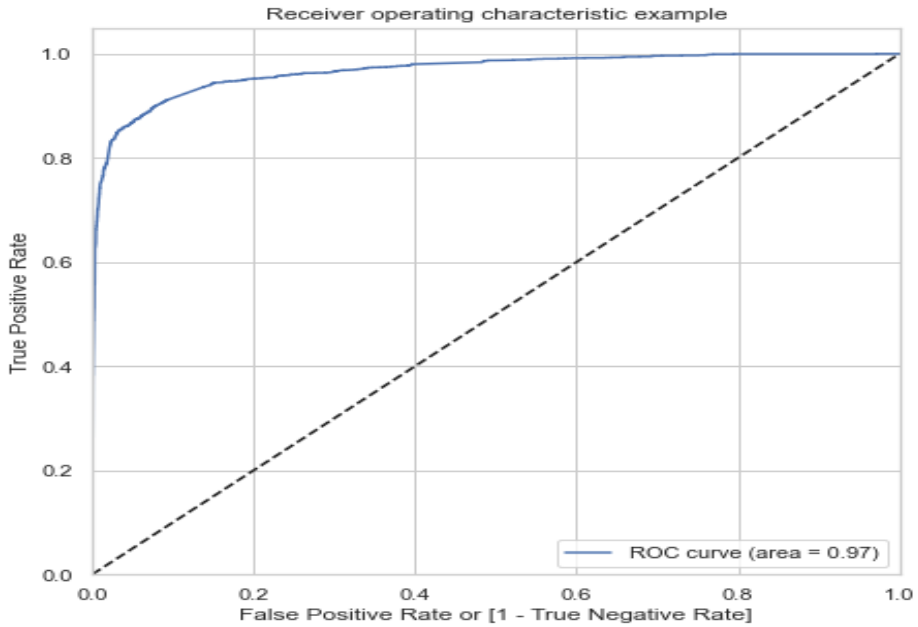
- Columns that have only two levels “Yes” and “No” were converted to numerical using binary mapping.
- Columns that have more than two levels were converted to dummies using the pd.get_dummies function.
- The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

	Do Not Email	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search
5428	0	-1.071483	-0.871984	-1.184151	0	0	0	0	0	1	0
8583	0	0.641520	2.066787	0.128349	0	1	0	0	1	0	0
4637	0	-0.386282	-0.740141	-0.134151	1	1	0	0	0	0	0
4468	0	-0.043681	-0.205124	0.390849	1	1	0	0	0	0	0
2058	0	1.326721	-0.583457	-0.449151	0	1	0	0	0	0	0

Model Building

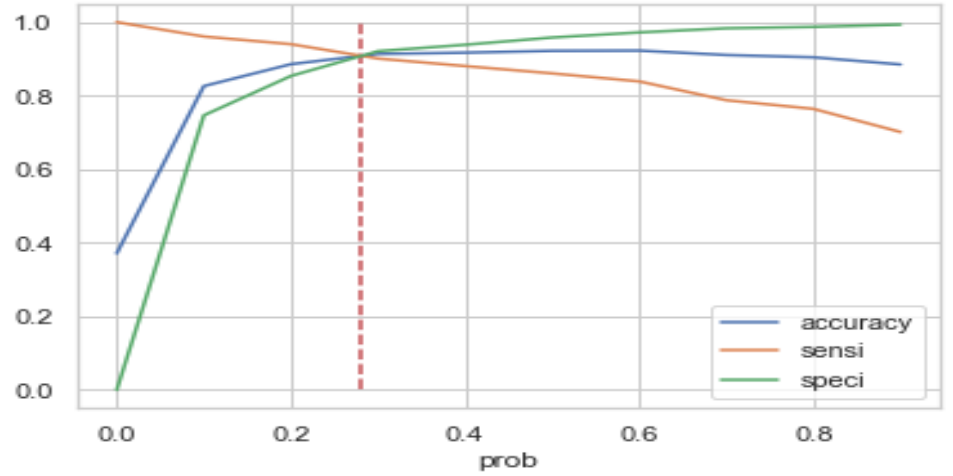
- Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features. RFE uses the model accuracy to identify which attributes (and
- combination of attributes) contribute the most to predicting the target attribute
- Then, we build three models using the glm() function, which is part of the formula submodule of (stats models). The glm() function fits generalized linear models, a class of models that includes logistic regression.
- Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
- Finally, we arrived at the 12 most significant variables. The VIFs for these variables were also found to be good. Variance inflation factor(VIF) is used to treat the multi-collinearity.
- Once the stable model was created, we predicted probabilities on the train set and created a new column predicted with 1 if the probability is > 0.5 else 0.
- We calculated the confusion matrix on this predicted column to the actually converted column and also calculated the metrics sensitivity, specificity, precision, recall, and accuracy.
- Finally, plotted the ROC curve to find the area under the curve.

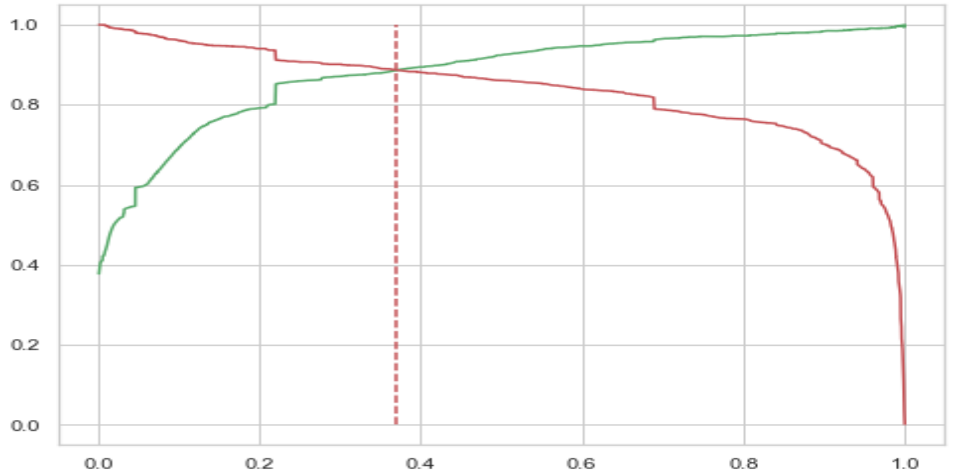
	Features	VIF
1	Lead Origin_Lead Add Form	1.942572
2	Lead Source_Olark Chat	1.641850
9	Tags_Others	1.612950
10	Tags_Will revert after reading the email	1.594053
0	Total Time Spent on Website	1.469263
5	Last Activity_SMS Sent	1.431213
11	Last Notable Activity_Modified	1.400635
4	Lead Source_Welingak Website	1.378624
6	Tags_Closed by Horizzon	1.224177
12	Last Notable Activity_Olark Chat Conversation	1.076704
7	Tags_Lost	1.064434
3	Lead Source_Others	1.064050
8	Tags_No phone number	1.016235



Making Predictions on the Train Set

- We used 0.5 as the cut-off in step 5. We computed the probability with several cut-offs to ensure that it was the best cut.
- We calculated the 3 metrics — accuracy, sensitivity, and specificity — for probabilities ranging from 0.0 to 0.9.
- An ideal cutoff of 0.28 was discovered from the intersection of sensitivity, specificity, and accuracy to generate predictions on the training dataset, as shown in the figure:
- The best cutoff was used to create predictions on the test dataset and was derived from the training dataset's Precision recall graph, as shown in the figure:
- Observably, the trade-off between precision and recall is 0.37. We can therefore confidently decide to classify any Prospect Lead with Conversion Probability greater than 37% as a hot lead.





Making Predictions on the Test set

- After finalizing the optimum cutoff and calculating the metrics on the train set, we predicted the data on the test data set. Below are the observations:
- Train Data:
  - Accuracy: 92%
  - Sensitivity: 90%
  - Specificity: 92%
  - Precision: 92%
  - Recall: 86%

- Test Data:
  - Accuracy: 92%
  - Sensitivity: 89%
  - Specificity: 94%
  - Precision: 90%
  - Recall: 89%

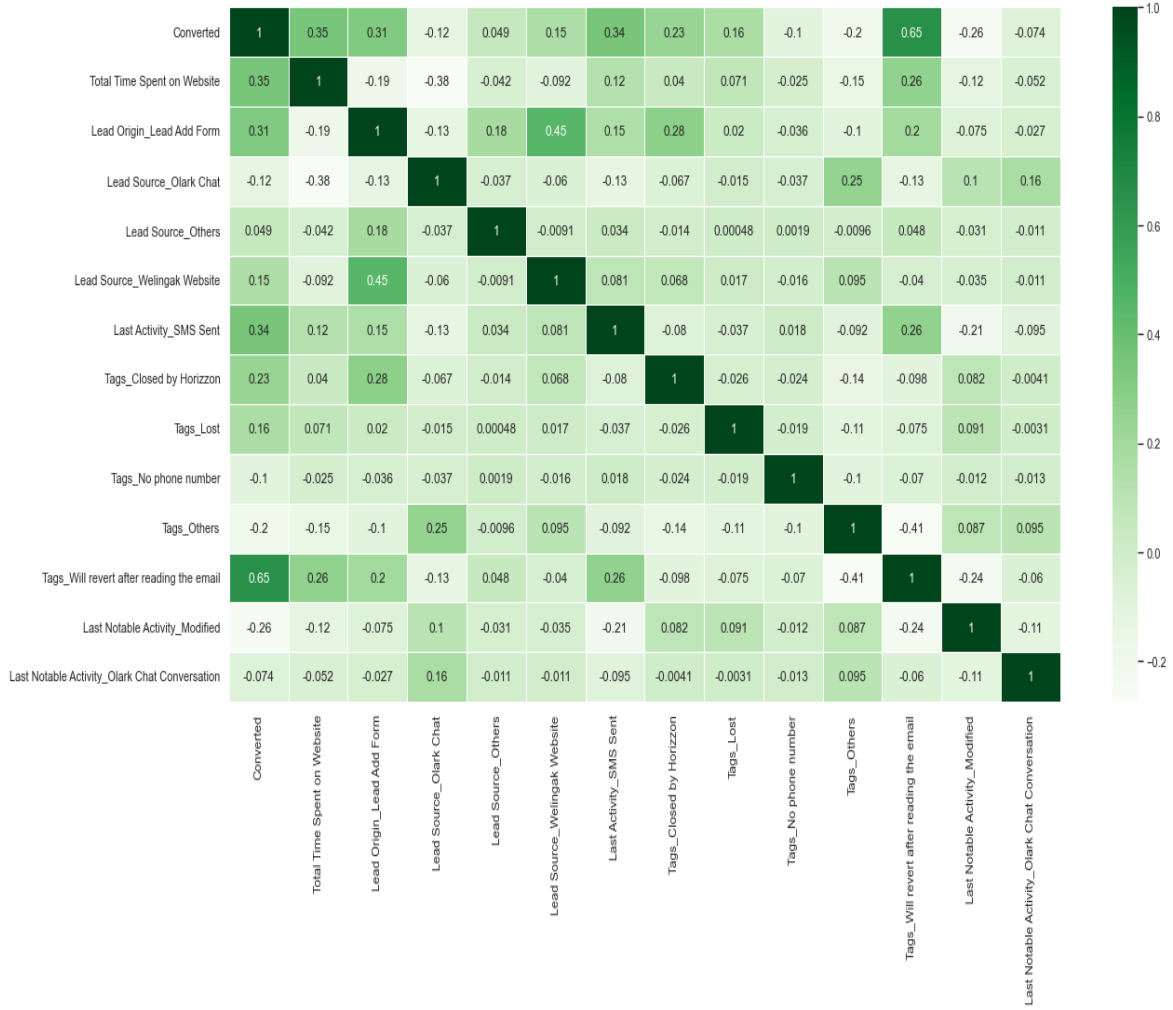
Conclusion

Conclusions:

- The Model seems to predict the Conversion Rate very well on the final predicted model which meets the expectation of the CEO and has given a ballpark of the target lead conversion rate to be around 80%.
- The good value of sensitivity of our model will help to select the most promising leads.

Recommendations:

- If the company has limited time and resources, you should turn to Hot_Leads i.e Get the most conversions and avoid wasted calls with prospects who have a conversion probability of 80% or more.
- If the company has enough resources and time, you should reach out to every prospective customer. But with plenty of time, you should also focus on your low-converting clients to improve your overall lead conversion rate.
- Heat shows the variables help in driving the hot leads




Thanks