
# LEAD SCORING CASE STUDY

BY
SEYED JAVIDH

# Problem statement

- X Education is an education company that sells course. When people fill up a form providing their email address or phone number, they are classified to be a lead. Leads are acquired from various sources
- Lead conversion rate at X education is around 30%, which was poor.
- In order to identify the most potential or hot leads, we have to build a model and assign lead score.
- So that the sales team can be more focused towards the most promising leads.
- CEO’s target of conversion is around 80%

## Objective

- Built a most efficient and simple model to identify the hot leads and assign lead score.
- Deployment of the model for the future use.


# Approach / Methodology

01 Data Cleaning And Manipulation

02 Exploratory Data Analysis

03 Data Transformation

04 Model Building

05 Model Evaluation

06 Conclusions / Recommendations


# Data Cleaning & Manipulation

# DATA CLEANING & MANIPULATION

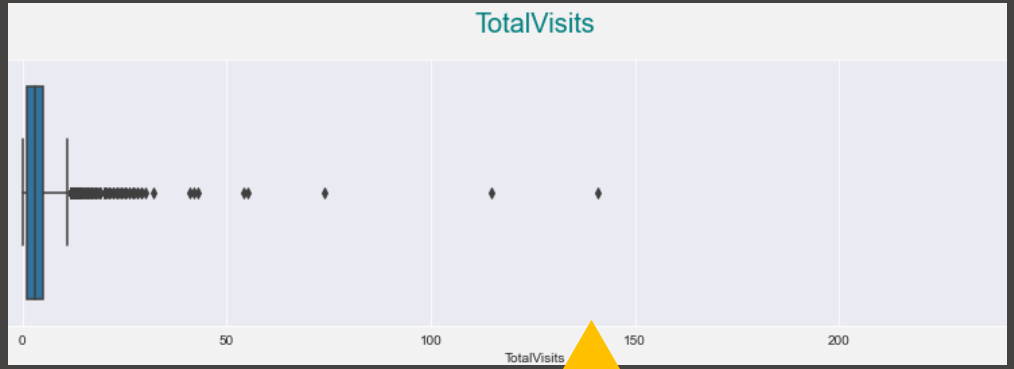
- ✓ Initial number of Records = 9240 ; Initial number of columns = 37
- ✓ No duplicates were found.
- ✓ All the columns with single unique value for all the records were dropped as it wont help in modeling.
- ✓ Replaced ‘Select’ with Nan.
- ✓ Columns with more than 45% were dropped.
- ✓ Columns like 'Prospect ID', 'Lead Number' with unique values for all the records were dropped.
- ✓ Columns with NAN were either replaced with mode value or named as new category ‘unknown’.
- ✓ Based on value counts, categories with negligible values are grouped as ‘others’.
- ✓ Dropping extreme outliers of numerical columns.
- ✓ Columns with imbalanced data were dropped.
- ✓ The values of ‘Yes’ & ‘No’ were converted to ‘1’ & ‘0’ respectively.


# Exploratory Data Analysis

# TOTAL VISITS

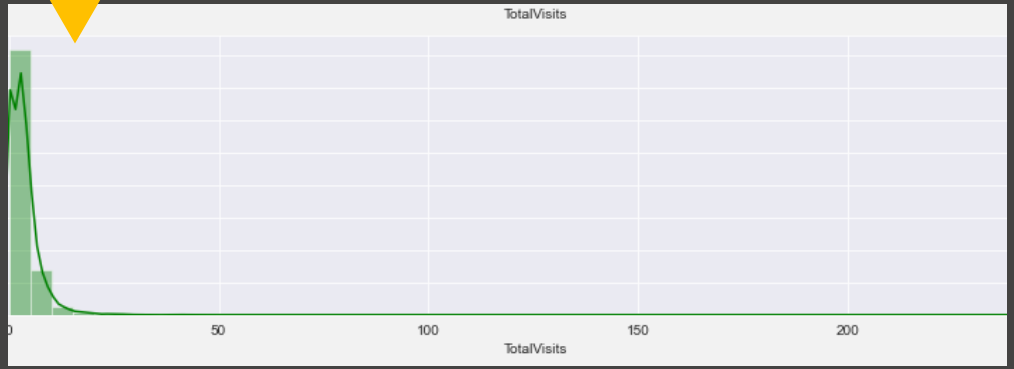
## INFERENCE

- ✓ If we observe the boxplot we can see that the there are definitely some outliers in the range of 250. It show that people are visiting the page for 250 times.
- ✓ It can also be observed from the histogram that most of the visits are in the range of 0 to 25. There are very less leads who have visited the page for more than 25 times.




**Box plot**

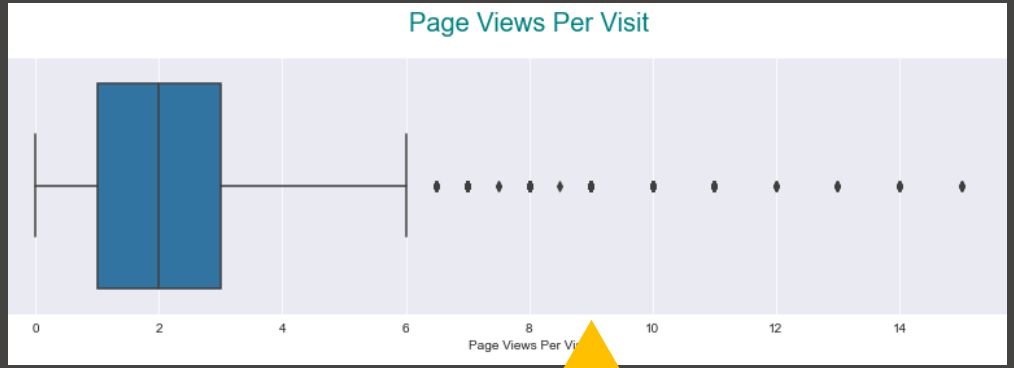
**Distribution**



# PAGE VIEWS PER VISIT

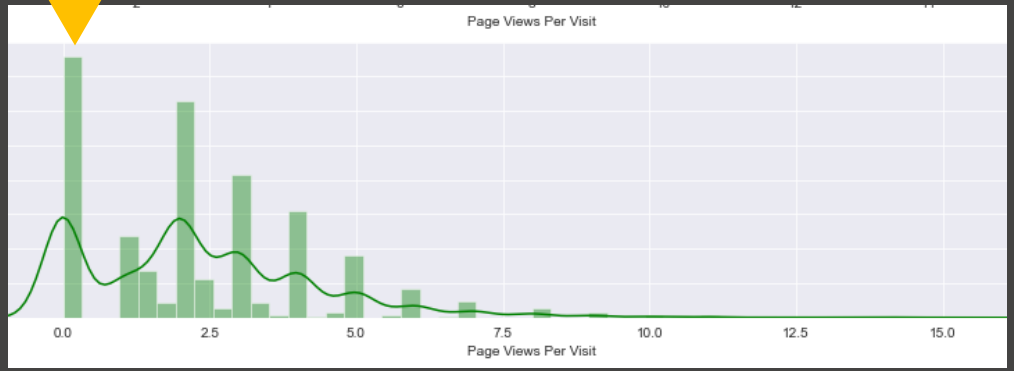
## INFERENCE

- ✓ From the boxplot we can definitely see that there are outliers in the data.
- ✓ And on the other hand, from the histogram we can see that the data is definitely skewed. With most of the data near the 0 to 10 bin.




**Box plot**

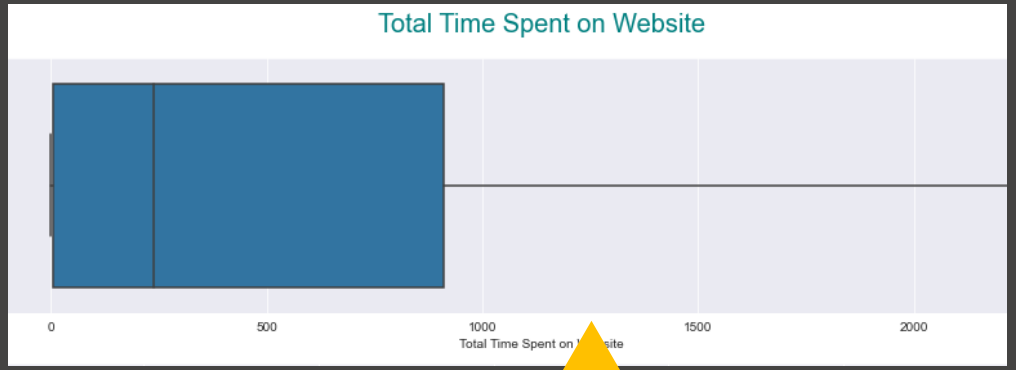
**Distribution**



# TOTAL TIME SPENT ON WEBSITE

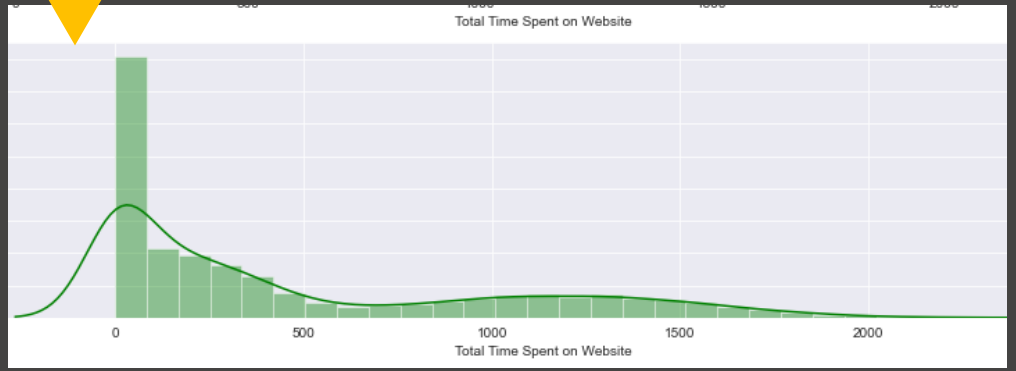
## INFERENCE

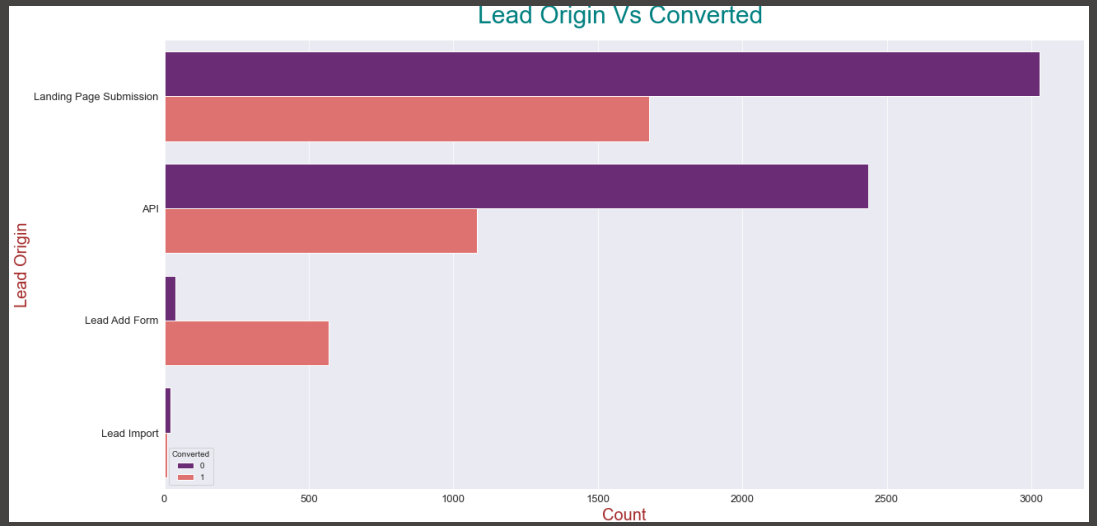
- ✓ Here, we can see from the boxplot that the mostly people spend about 1000 seconds on the website.
- ✓ Also, we can see from the histogram that it is skewed and most people spend near about 500 seconds on the website.




**Box plot**

**Distribution**



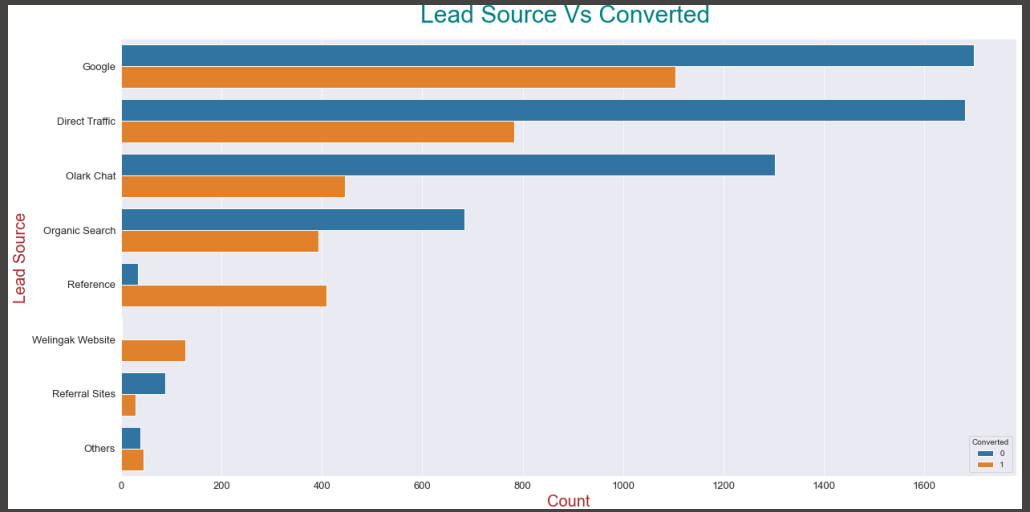


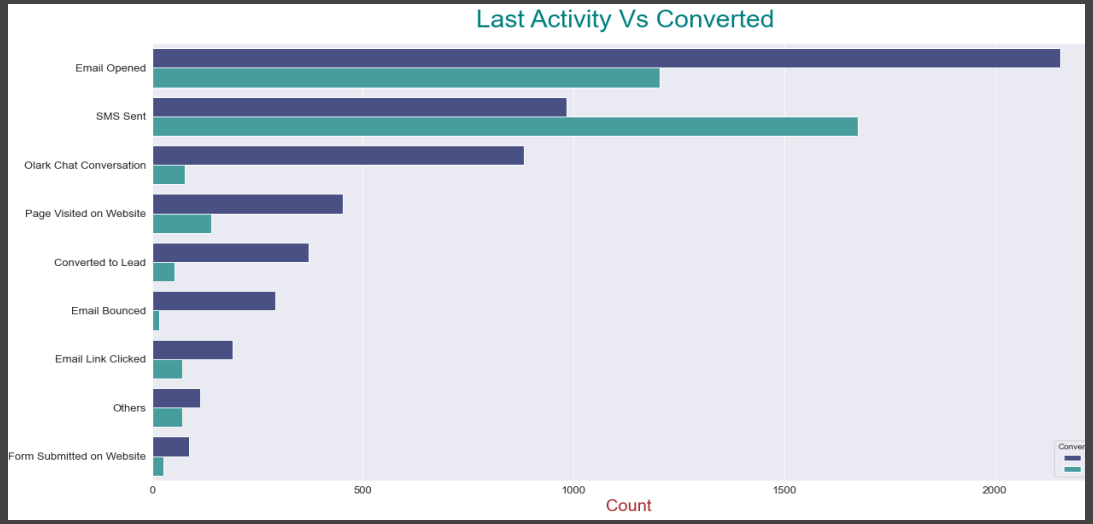

# INFERENCE

- ✓ Here, we can see that Leads, who's origin is from the Add Form section, are more likely to get converted later on.
- ✓ The ratio of Leads converted from the Landing Page Submission and API looks okayish, however not as great as that of Lead Add Form.

# INFERENCE

- ✓ From the above graph we can see that Leads who come through refrence or from Wellingak website, or any other sources are more likely to get converted.
- ✓ Leads from Google are also quite likely to get converted.



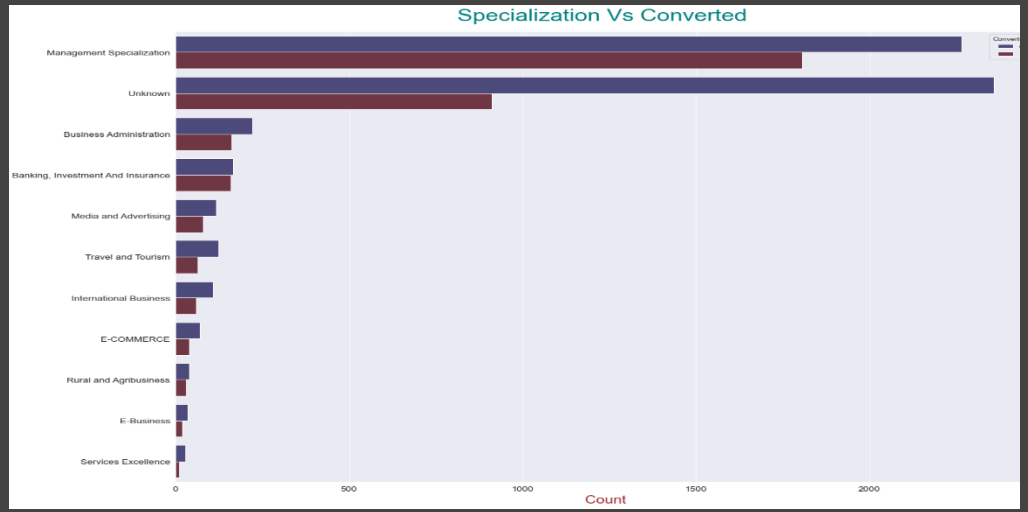


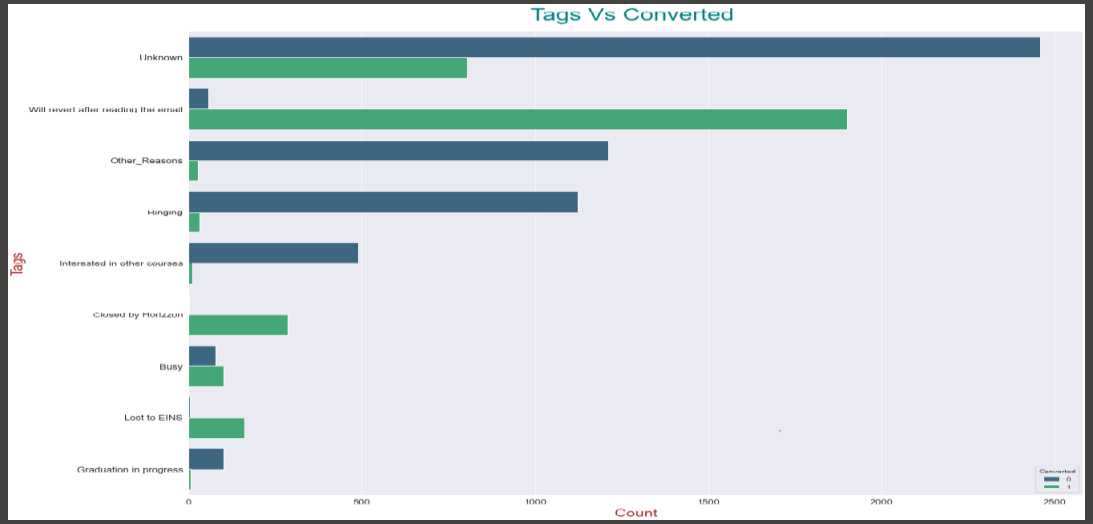

# INFERENCE

- ✓ Here, we can notice that the Leads who's last activity is sending SMS are really good to target, as they are more likely to get converted.
- ✓ However, we should avoid leads who's last activities are- Olark Chat Conversation, Email Bounced or already converted leads.

# INFERENCE

- ✓ From the graph we can understand that leads from Management, Business Administration, Banking investment and insurance are more likely to get converted.
- ✓ However, people who do not mention their specialization are less likely to be converted.



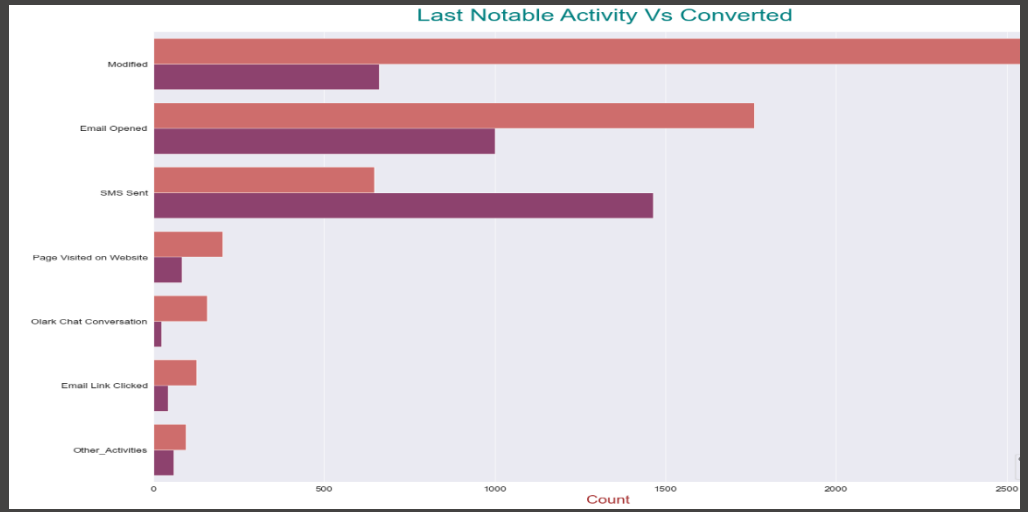


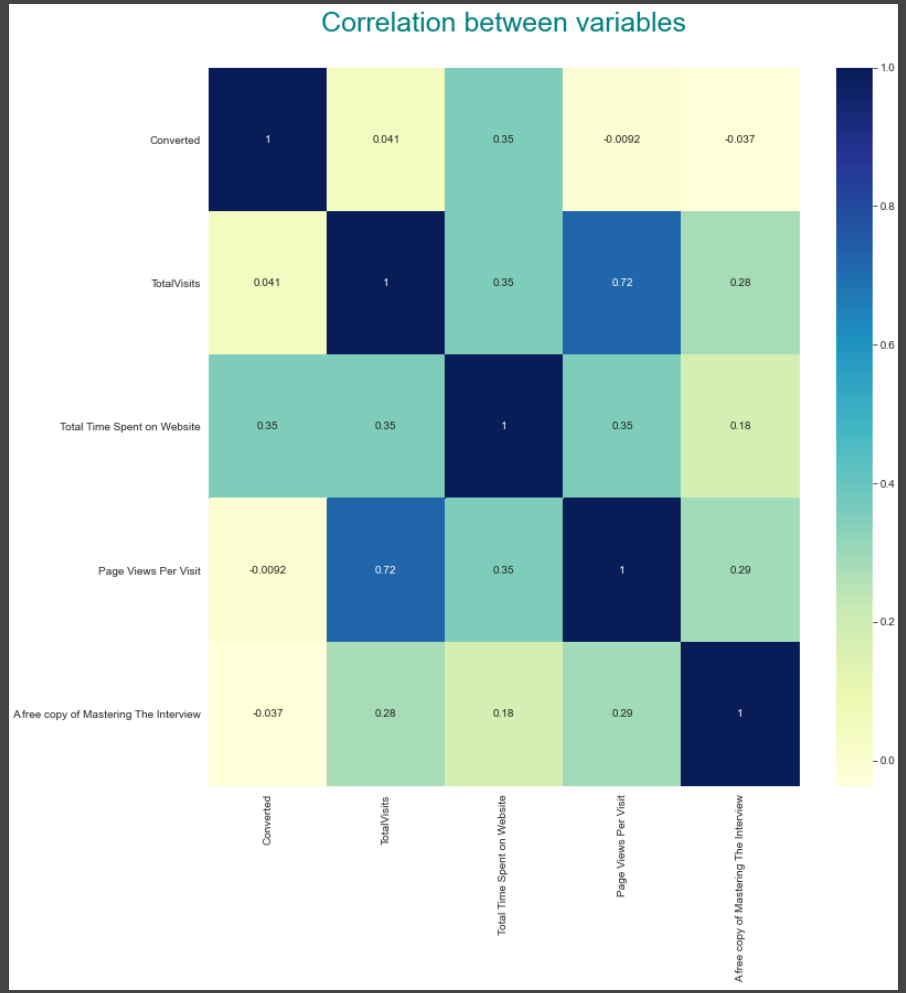

# INFERENCE

- ✓ It can be observed from the plot above that Leads who are tagged as "Will revert back after reading the email" are more likely to be converted followed by "Closed by Horizzon" and "Lost to EINS".
- ✓ Leads who are still graduating, interested in other course or their phone ringing and not picking up are less likely to get converted.

# INFERENCE

- ✓ The result here is very similar to that of Last Activity performed by the Lead.






# CORRELATION

## INFERENCE

- ✓ The heatmap clearly shows us that there is a strong correlation between "Page Views Per Visit" and "Total Visit" column.
- ✓ Similar positive correlations can be identified between "Total Time Spent on Website" against the "Converted" value.
- ✓ There is also a positive correlation between "Total Time Spent on Website" with both "Total Visit" and "Page Views Per Visit".


# Data Transformation

# DATA TRANSFORMATION

- ✓ Dummy variables are created
- ✓ Standard scaler was used to standardize the numerical columns
- ✓ Number of rows after EDA = 8868
- ✓ Number of columns after EDA = 50


# Model Building

# Model Building Process

## Train Test Split

### Using scikit learn

Splitting Train And Test Data

In The Ratio Of 70 : 30

For The Purpose Of

Evaluation

## Recursive Feature Elimination

### Using RFE (sklearn)

We Ran RFE To

Select Top 15

Predictor Variables

From Total 50 Variables

## Logistic Regression Model Building

### Using Stats Model

Removed Variables One By One

Which Has

P-value Grater Then 0.05

And VIF Above 5

# Final Model

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Converted | No. Observations: | 6204 |
| Model: | GLM | Df Residuals: | 6189 |
| Model Family: | Binomial | Df Model: | 14 |
| Link Function: | logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -1196.3 |
| Date: | Mon, 12 Jul 2021 | Deviance: | 2392.6 |
| Time: | 15:32:27 | Pearson chi2: | 8.50e+03 |
| No. Iterations: | 8 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -4.3841 | 0.218 | -20.133 | 0.000 | -4.811 | -3.957 |
| Total Time Spent on Website | 1.1248 | 0.064 | 17.500 | 0.000 | 0.999 | 1.251 |
| Lead Origin_Lead Add Form | 3.0066 | 0.459 | 6.550 | 0.000 | 2.107 | 3.906 |
| Lead Source_Olark Chat | 1.5997 | 0.154 | 10.357 | 0.000 | 1.297 | 1.902 |
| Lead Source_Welingak Website | 2.6224 | 0.870 | 3.015 | 0.003 | 0.918 | 4.327 |
| Last Activity_Email Bounced | -1.7582 | 0.498 | -3.531 | 0.000 | -2.734 | -0.782 |
| Last Activity_Olark Chat Conversation | -1.2890 | 0.232 | -5.560 | 0.000 | -1.743 | -0.835 |
| Tags_Busy | 3.1471 | 0.299 | 10.527 | 0.000 | 2.561 | 3.733 |
| Tags_Closed by Horizzon | 9.1264 | 1.033 | 8.831 | 0.000 | 7.101 | 11.152 |
| Tags_Lost to EINS | 8.6147 | 0.758 | 11.363 | 0.000 | 7.129 | 10.101 |
| Tags_Ringing | -0.9941 | 0.309 | -3.220 | 0.001 | -1.599 | -0.389 |
| Tags_Unknown | 2.6034 | 0.210 | 12.390 | 0.000 | 2.192 | 3.015 |
| Tags_Will revert after reading the email | 6.9711 | 0.266 | 26.158 | 0.000 | 6.449 | 7.493 |
| Last Notable Activity_Modified | -0.7132 | 0.139 | -5.135 | 0.000 | -0.985 | -0.441 |
| Last Notable Activity_SMS Sent | 2.2050 | 0.137 | 16.090 | 0.000 | 1.936 | 2.474 |

| | Features | VIF |
|---|---|---|
| 2 | Lead Source_Olark Chat | 1.85 |
| 1 | Lead Origin_Lead Add Form | 1.80 |
| 12 | Last Notable Activity_Modified | 1.77 |
| 11 | Tags_Will revert after reading the email | 1.72 |
| 10 | Tags_Unknown | 1.63 |
| 13 | Last Notable Activity_SMS Sent | 1.62 |
| 5 | Last Activity_Olark Chat Conversation | 1.57 |
| 0 | Total Time Spent on Website | 1.50 |
| 3 | Lead Source_Welingak Website | 1.30 |
| 7 | Tags_Closed by Horizzon | 1.24 |
| 9 | Tags_Ringing | 1.13 |
| 4 | Last Activity_Email Bounced | 1.10 |
| 8 | Tags_Lost to EINS | 1.07 |
| 6 | Tags_Busy | 1.05 |


# Model Evaluation

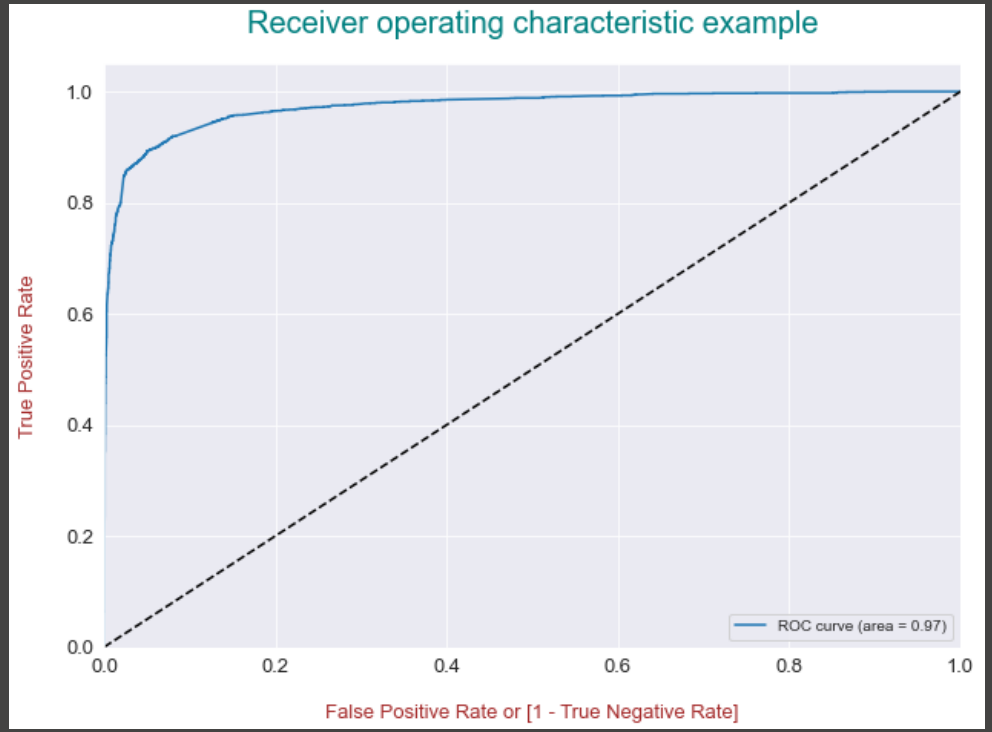
# Metrics – Train Dataset

| Precision | Recall |
|---|---|
| 88% | 91% |

| Accuracy | Sensitivity | Specificity |
|---|---|---|
| 92% | 88% | 96% |

| Positive Predictive Rate | Negative Predictive Rate | False Positive Rate |
|---|---|---|
| 93% | 93% | 4% |

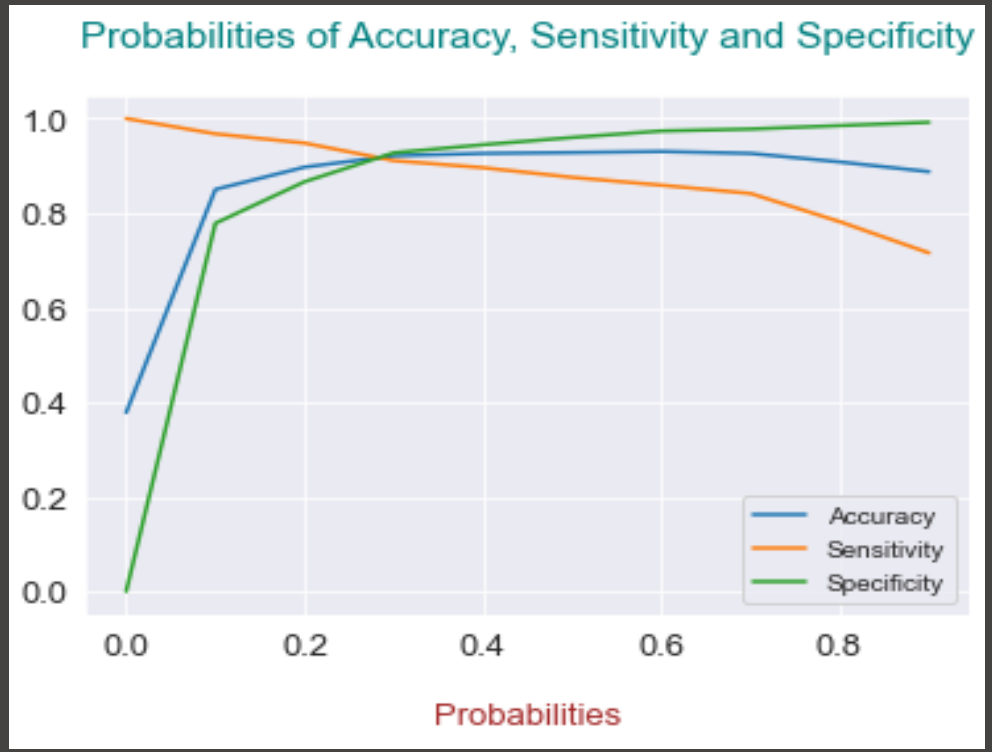
# ROC - Curve




## INFERENCE

- ✓ From the ROC curve, we can see that the area under the curve is very high (0.97).
- ✓ A high area under the curve indicates that the model is very good.

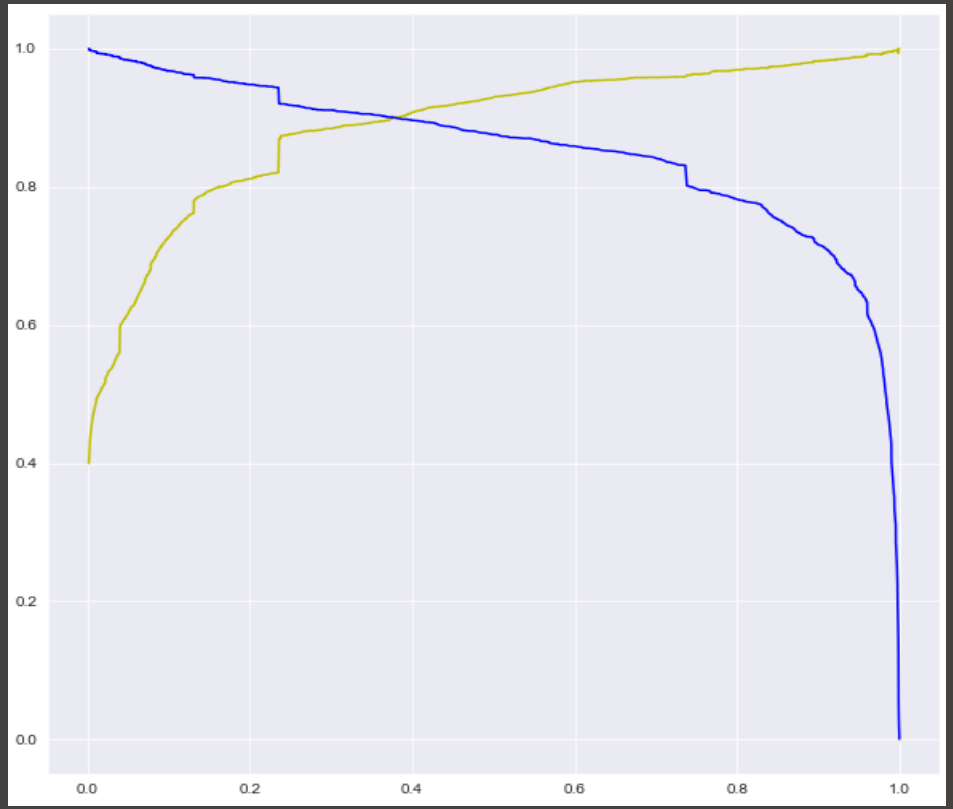
# Optimal Cut-Off




## INFERENCE

- ✓ From the above graph, we can make out that the optimal cut-off for our model will be 0.28.
- ✓ This is the point where the sensitivity, accuracy and specificity co-exist.

# Precision – Recall Trade Off




## INFERENCE

- ✓ From the graph, we can see that there is a trade off between Precision and Recall and the meeting point is near to 0.4

# Metrics – Test Dataset

**Precision**

88%

**Recall**

93%

**Sensitivity**

93%

**Specificity**

92%

# Conclusions And Recommendations

- ✓ Tags_Closed by Horizzon has the highest coefficient of 9.1264, which means keeping other variable constant an unit increase in temp results in 9.1264 unit increase in Probability of conversion.
- ✓ Tags_Closed by Horizzon, Tags_Lost to EINS and Tags_Will revert after reading the email are the top 3 variables having strong coefficients.
- ✓ Last Activity_Olark Chat Conversation, Tags_Ringing and Last Notable Activity_Modified have negative coeeficient, which means increase in values of these variables would result in decrease in value of Probability of conversion.
- ✓ Probability of conversion increases if Tags_Busy, Lead Origin_Lead Add Form, Lead Source_Welingak Website, Tags_Unknown, Last Notable Activity_SMS Sent, Lead Source_Olark Chat, Total Time Spent on Website increases as these variables have positive coefficients.
- ✓ Constant value - when all other variables are zero the Probability of conversion value will still be -4.3841
- ✓ Comparing Precision, Recall and other metrics value for both train and test. Our model performs well on test set as well.

# Conclusions continued…

- ✓ This model explains how exactly the Probability of conversion vary with different features. management can accordingly manipulate the business strategy to meet the conversion target and meet the business expectations.
- ✓ In business terms, this model can be deployed in the upcoming future to meet the X education's requirements.
- ✓ Focusing on the features of the model will increase their chances of contacting most of the potential buyers for the course.
- ➢ The Marketing team and evaluate the leads based on the top 3 variables and make sound business decisions.
- ➢ The Marketing team can also chase after leads, who spend longer time on their website, originate from Add form.
- ➢ The team can also come with interesting courses and offers that attract people with specialization in banking, investment and insurance.
- ➢ They can also keep a close watch on Leads originating from Olark Chat.


# Thank you