Brief Summary Report

Lead Scoring Case Study

Harshvardhan Pareek & Shujaatali Badami

Strategy

1. Data Gathering
2. Data Cleaning
3. Data Analysis
4. Building Logistic Regression model and calculating Lead score
5. Model Evaluation
6. Conclusion

- **Data Gathering**

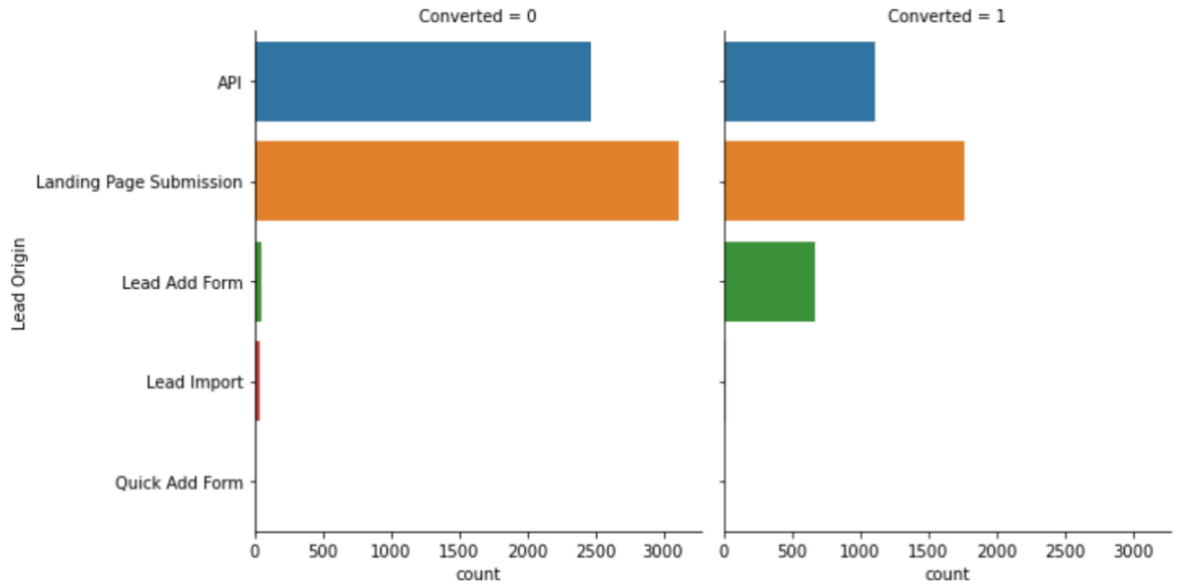
  Initially, the dataset was loaded into the jupyter file, and the data was analysed for things like the dataset's shape, the datatypes of the columns, and some statistical information about the data, such as the mean and mode of the data, the median, and outliers.

- **Data Cleaning**

  We chose to eliminate certain unnecessary columns from the dataset after seeing them. Certain columns displayed a 'Select' label, indicating that the client did not select any options. It was preferable to leave it as a null value since there were no viable alternatives for the consumer to pick. Due to the structure of the data, outliers were identified in two columns and were managed by upper capping them.
  We eliminated columns with more than 30% of their values missing. We substituted missing values in the remaining category columns with the mode value. Two columns had similar names, which was resolved by renaming the columns in a single format. Due to the nature of the data, we opted to impute missing values in the numerical columns by their corresponding modes after outlier treatment and subsequent analysis.

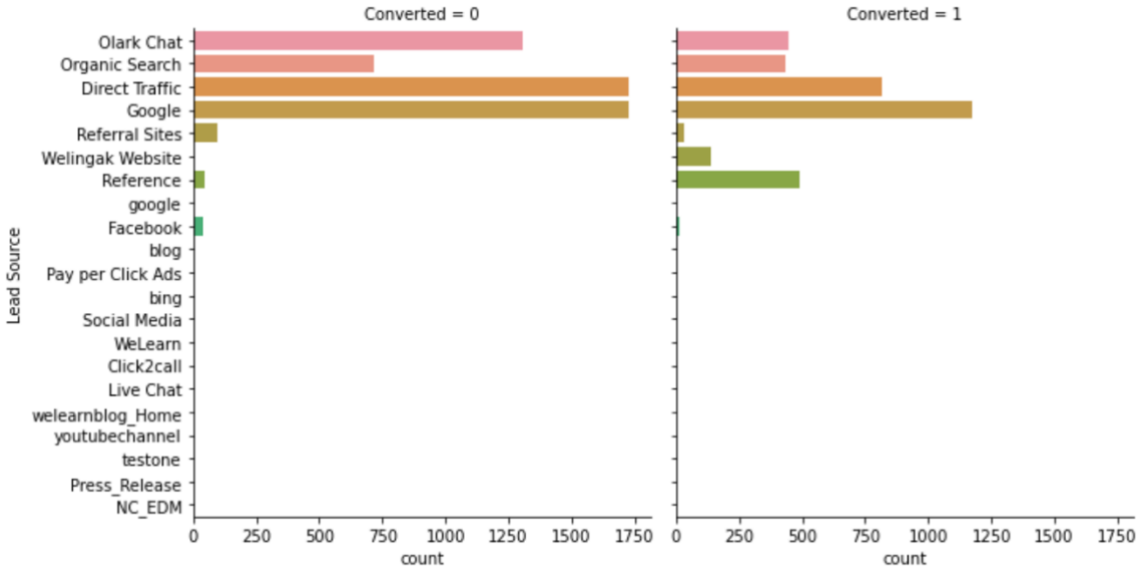
- **Data Analysis**

Lead Origin



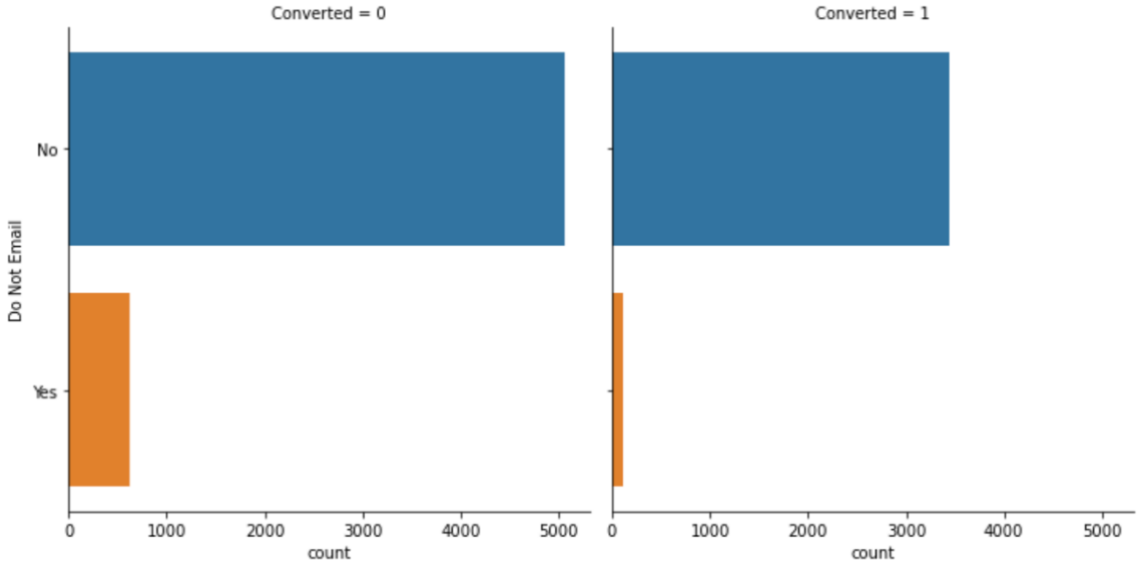
As can be observed, the highest conversion rate occurred during Landing Page Submission. Additionally, there was just one conversion from a fast add form.

Lead Source

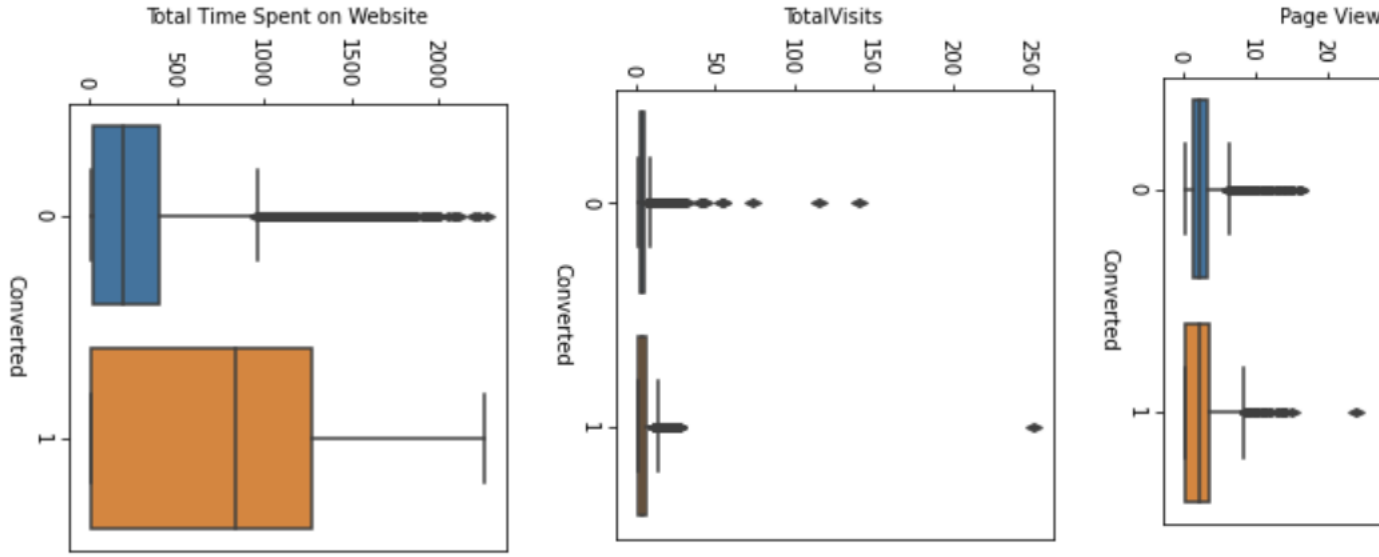


As can be observed, the primary source with the highest conversion rate is Google.

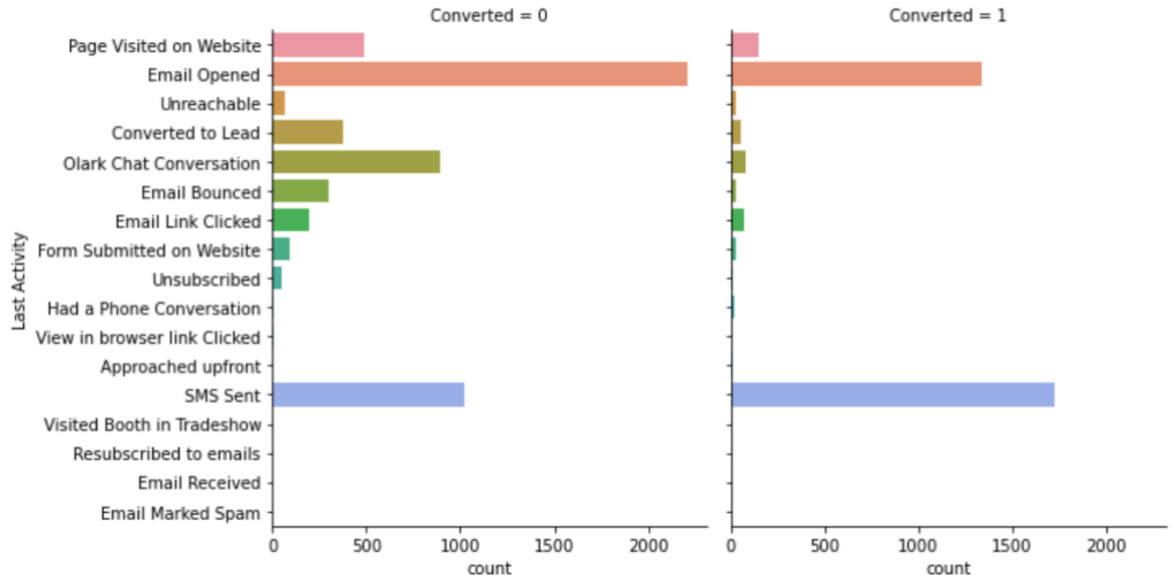
Do Not Email



According to the above graph, significant conversion occurred as a result of the emails sent.



The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit.



According to the following graph, the 'SMS Sent' Last Activity value had the most conversions, followed by the 'Email Opened' value.

Building Logistic Regression Model and calculation of Lead score

Numerous variables with high p-values were detected in the logistic regression model. We utilised RFE to choose features since the number of variables is pretty large and manually examining each one would be inefficient.

Following RFE, all columns were picked depending on their ranking and modelling was repeated.

Each feature with a p-value larger than 0.05 was eliminated one by one, and the model was constructed again.

All the features now had p-value less than 0.05 and the final model obtained was as below:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1608	0.115	-10.096	0.000	-1.386	-0.935
Lead Source_Google	0.3106	0.093	3.338	0.001	0.128	0.493
Lead Source_NC_EDM	-1.301e-16	6.32e-17	-2.059	0.040	-2.54e-16	-6.24e-18
Lead Source_Olark Chat	1.4259	0.168	8.501	0.000	1.097	1.755
Lead Source_WeLearn	5.944e-16	6.87e-17	8.646	0.000	4.6e-16	7.29e-16
Last Notable Activity_Email Bounced	3.49e-16	2.68e-17	13.005	0.000	2.96e-16	4.02e-16
Total Time Spent on Website	1.1751	0.048	24.285	0.000	1.080	1.270
TotalVisits	0.3676	0.058	6.330	0.000	0.254	0.481
Page Views Per Visit	-0.2671	0.065	-4.118	0.000	-0.394	-0.140
Lead Source_Social Media	0	0	nan	nan	0	0
Last Notable Activity_SMS Sent	1.6294	0.088	18.549	0.000	1.457	1.802
Lead Origin_Lead Add Form	0	0	nan	nan	0	0
Lead Origin_Landing Page Submission	-0.1649	0.109	-1.516	0.130	-0.378	0.048
Lead Source_bing	0	0	nan	nan	0	0
Last Activity_Olark Chat Conversation	-1.7910	0.186	-9.605	0.000	-2.156	-1.426
Lead Origin_Lead Import	0	0	nan	nan	0	0
Lead Source_Live Chat	0	0	nan	nan	0	0
Lead Source_Pay per Click Ads	0	0	nan	nan	0	0
Lead Origin_Quick Add Form	0	0	nan	nan	0	0

Following that, we examined the model for concerns of multicollinearity using VIF values.

```
Computing VIF values to keep track of multicollinearity
================================================================================
                                    Features   VIF
7                        Page Views Per Visit  2.76
2                      Lead Source_Olark Chat  2.43
6                                 TotalVisits  2.18
11    Lead Origin_Landing Page Submission  1.79
9           Last Notable Activity_SMS Sent  1.43
13  Last Activity_Olark Chat Conversation  1.40
0                         Lead Source_Google  1.37
5               Total Time Spent on Website  1.36
1                         Lead Source_NC_EDM   NaN
3                        Lead Source_WeLearn   NaN
4     Last Notable Activity_Email Bounced   NaN
8                  Lead Source_Social Media   NaN
10               Lead Origin_Lead Add Form   NaN
12                        Lead Source_bing   NaN
14                 Lead Origin_Lead Import   NaN
15                   Lead Source_Live Chat   NaN
16           Lead Source_Pay per Click Ads   NaN
17              Lead Origin_Quick Add Form   NaN
================================================================================
Overall accuracy: 0.7849796699354221
================================================================================
```

Each variable had a VIF value less than seven. As a result, there were no significant multicollinearity difficulties. We did not eliminate 'Page Views Per Visit' and 'Total Visits' since, from a business standpoint, these variables may be critical and, by deleting them, we risk overfitting the model.

As a result, we refined this model and utilised it for further research and prediction. Additionally, a receiver operating characteristic (ROC) curve was generated to ensure the model's stability, and an ideal cut-off point was computed taking accuracy, sensitivity, and specificity into account.

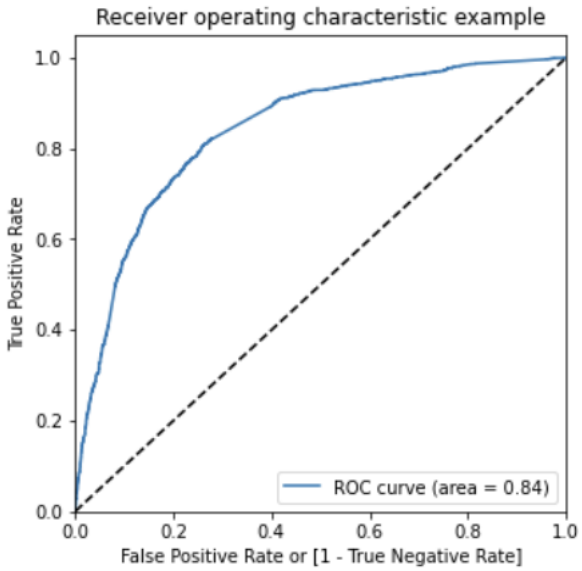
ROC Curve

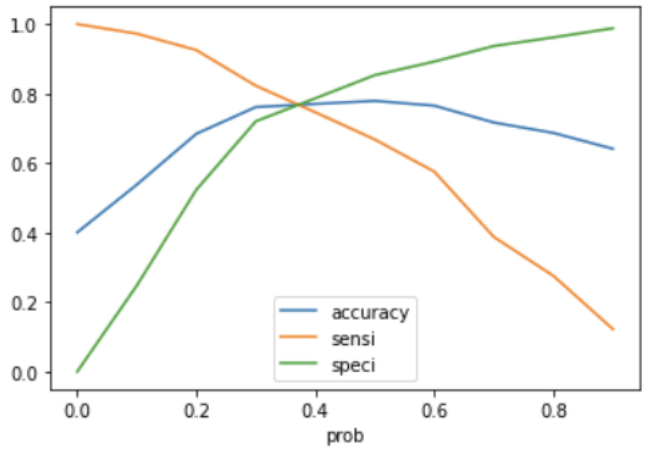
The curve is closer to the left side of the border than to the right side hence our model is having great accuracy.

The area under the curve is 84% of the total area.

Probability Cutoff Point

Probability cutoff point was at around 0.35 as this is where the sensitivity, accuracy and specificity converged.





Conclusion (LR Model)

When compared to the model produced using PCA, our Logistic Regression Model is adequate and accurate, with 78.6 percent Accuracy on Test Set, 73.3 percent Sensitivity, and 82.3 percent Specificity. We may alter these factors by adjusting the cut-off value and so forecast Hot leads in response to circumstances such as the availability of more resources and vice versa.

Conclusion (Recommendation)

X Education Company must prioritize the following critical areas in order to increase overall conversion rates: Enhance user engagement on their website, since this results in more conversions. Increase the frequency of SMS alerts, as this contributes to greater conversion; Increase total visits via advertising, etc., as this contributes to higher conversion; and Improve the Olark Chat service, as this contributes to lower conversion.

This concludes that the model is in a stable state and we can successfully draw business related conclusions from it.