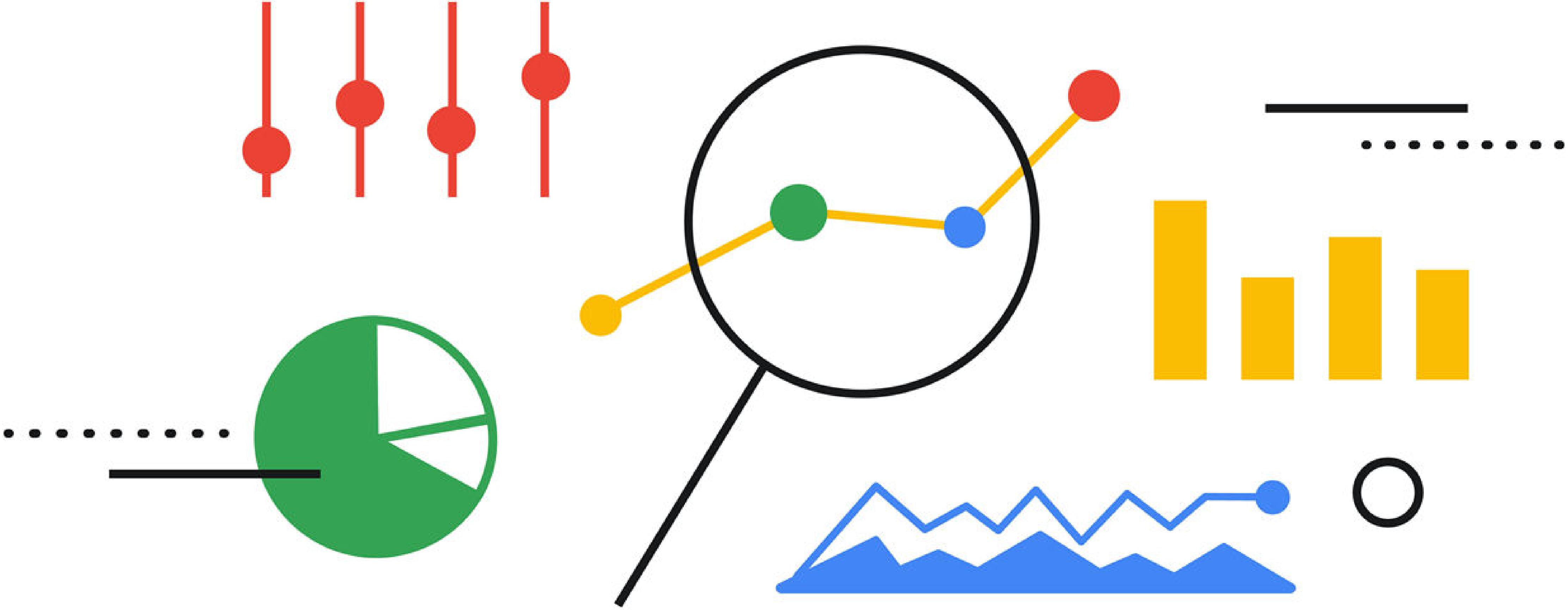
# Lead Scoring Case Study



**Presented By:** **Bishal Dasgupta** (+91 9674271330)

**Shripal Purohit** (+91 9872444857)

# Problem Statement

- In this Case study there is an Edtech company called – X Education, which provide online courses to working professionals. They are doing advertisements and customers are getting to know their product. The customer sees the ads on websites, search engines like Google, social media like Youtube and from there they land on the company's webpage. Once the customer visits the homepage, he/she may fill a form and hence their contact information is saved and the customer becomes a lead.
- Now the sales team calls/mails the leads to explain about their product to help them better understand the product and sell their product. Once the lead buys the product he/she is termed as 'converted'.

# Problem Statement - 2

- The problem that X Education is facing is that they have a low lead conversion rate that means they are getting a lot of leads and they are spending their recourses in reaching to all these leads and very small percentage of them are actually buying their products.
- The company wants to increase their lead conversion rate from 30% to 80%.
- In this case study we have to help X company identify the ‘hot leads’ and build a logistic regression machine learning model which calculates a lead score for the leads.
- With this X company can focus their resources on the hot leads which are most probable to be converted, thereby increasing their lead conversion rate.

# Assumptions

- ‘Prospect ID’ and ‘Lead Number’ being unique identifiers were irrelevant for our model building purpose. Hence, we have dropped the ‘Prospect ID’ but kept ‘Lead Number’ for our convinience.
- There were few variables which hand single outcomes, hence we have dropped those variables as well like ‘Magazine’ , ‘Newspaper’, etc.

# Approach

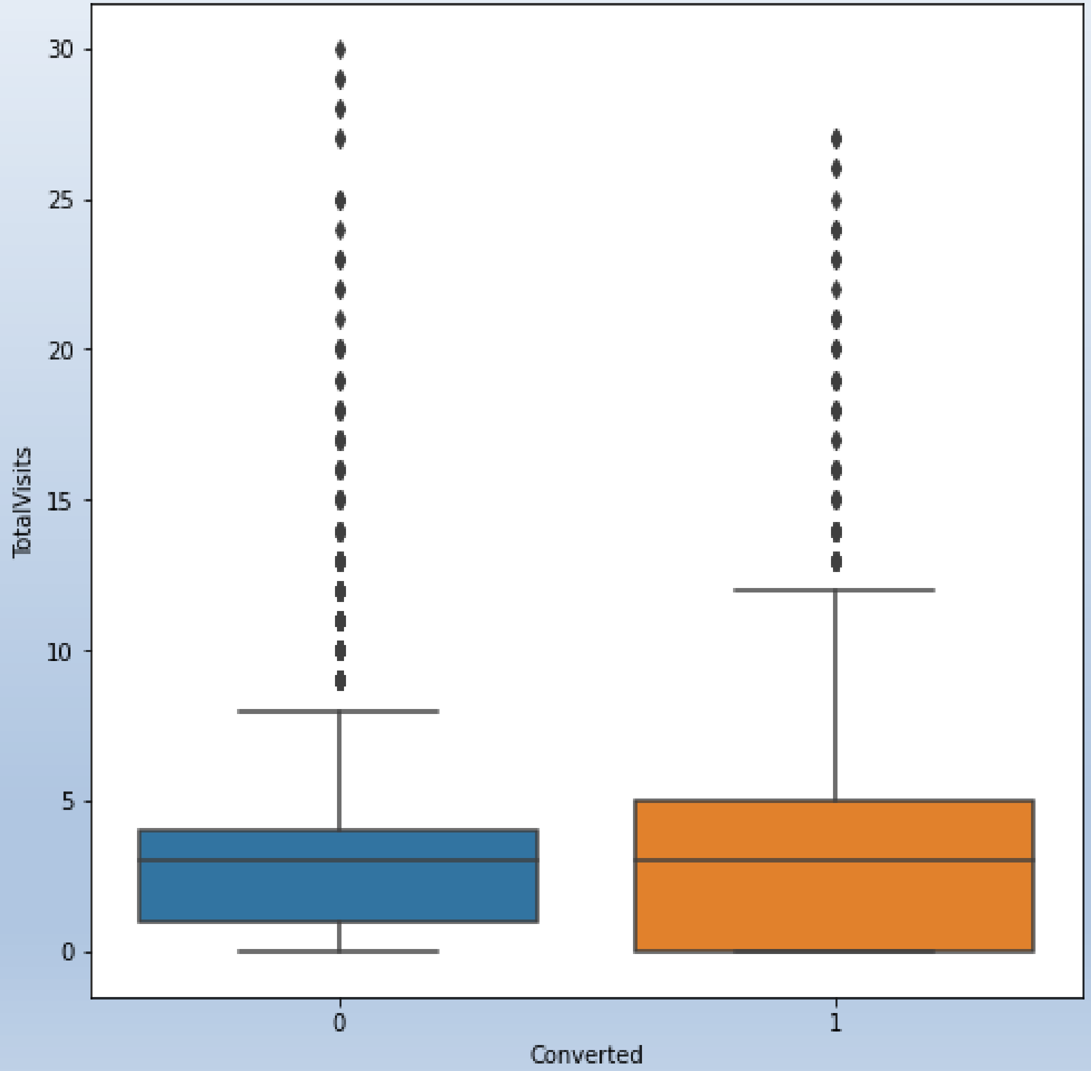
- We have first explored the data –
  - Dimensions, variables and their types
  - Missing values
  - Invalid data values like ‘Select’ in some variables.
- Data Cleaning
  - Handling missing values
  - Handling outliers
  - Removing ‘Select’
- Data Preprocessing
  - Creating dummy variables for categorical variables
  - Scaling continuous variables.
  - Train Test Split

# Approach - 2

- Exploratory Data Analysis
  - Univariate Analysis
  - Bivariate Analysis
  - Multivariate Analysis
- Model Building
  - Stats model as well as linear regression model
  - Hybrid model of automated and manual variable selection
- Prediction and Evaluation
  - Accuracy, Confusion matrix, Sensitivity, Specificity

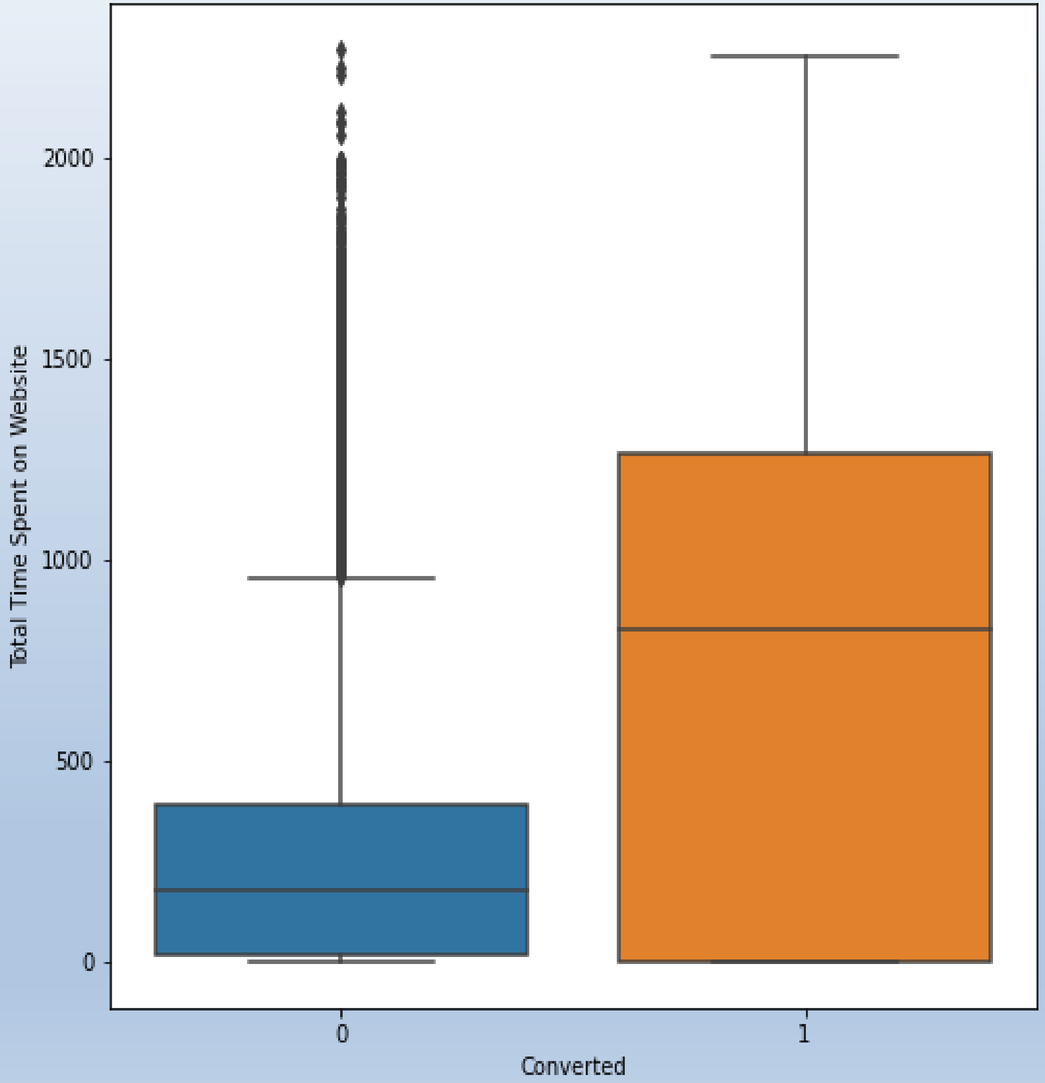
# Insights

- People who have visited the website more frequently tends to have higher lead conversion rate.

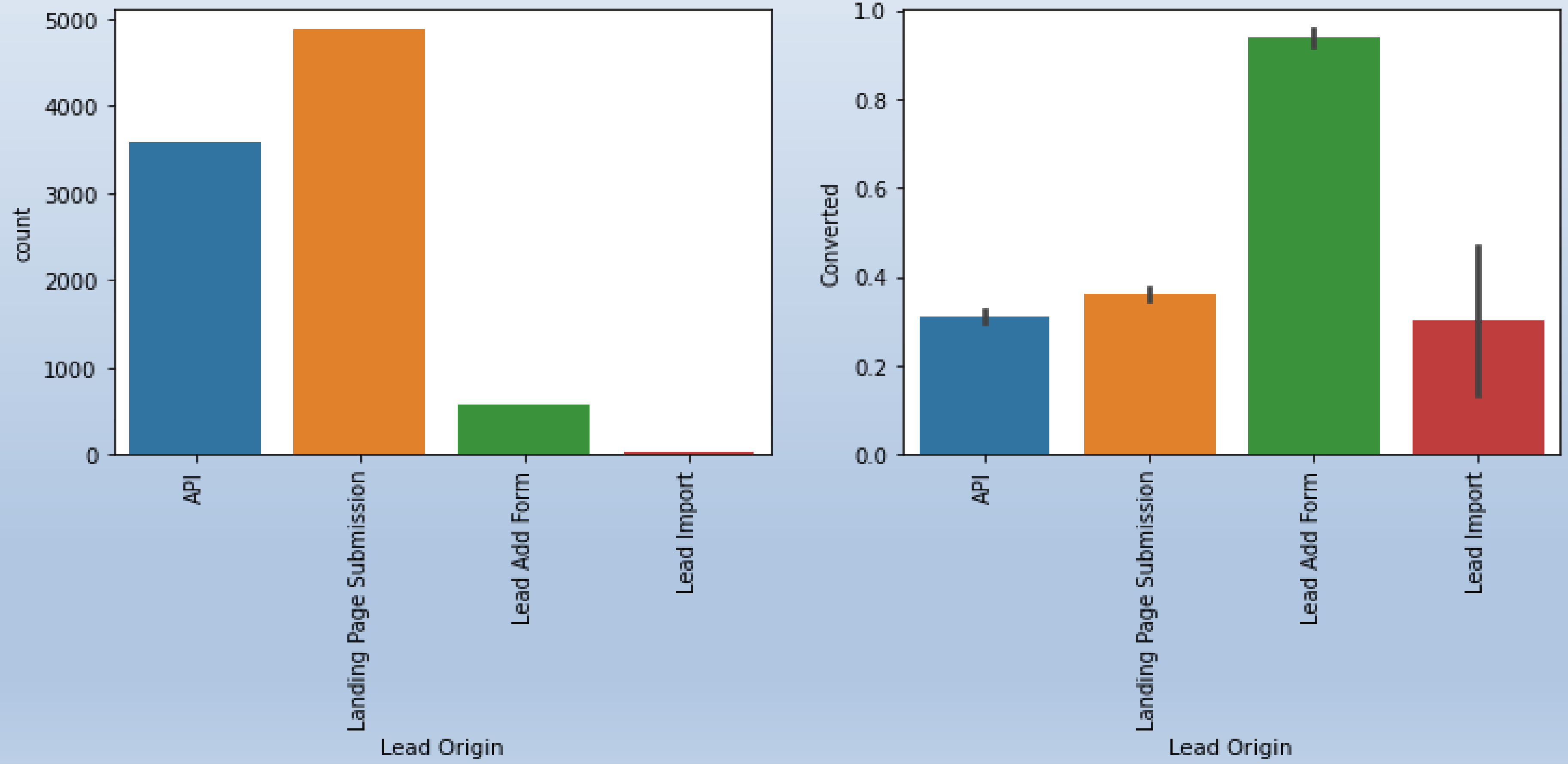


# Insights

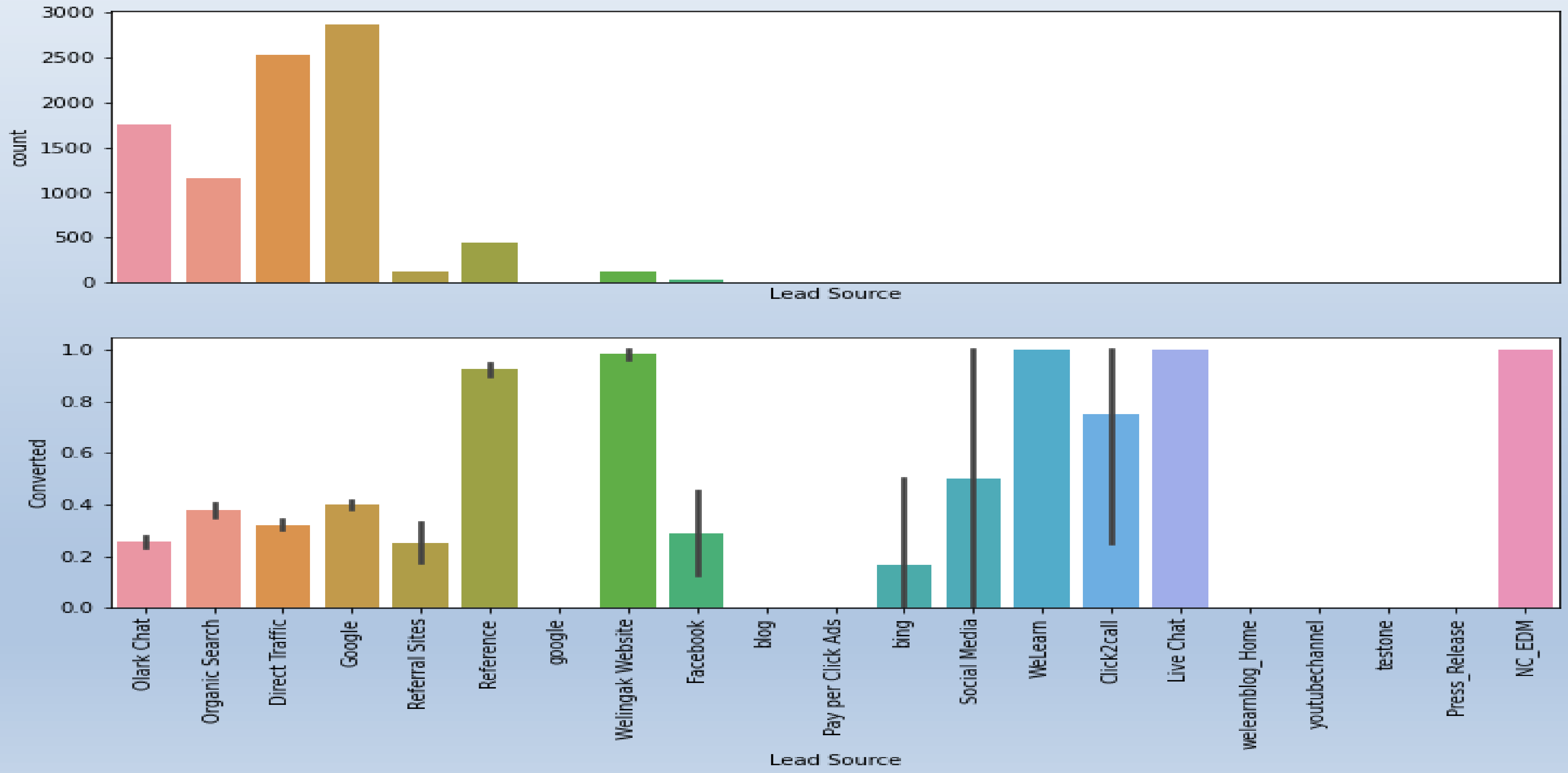
- People who have spent more time on the website tends to have higher lead conversion rate.



# Lead origin from Lead Add form shows maximum lead conversion.

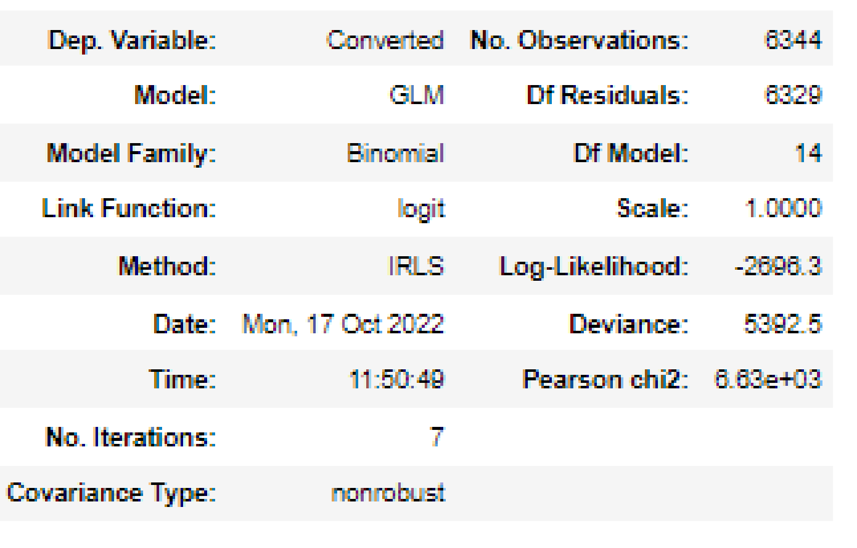


# Lead source ‘Welingak Website’ and ‘Reference’ has significant count and higher lead conversion.

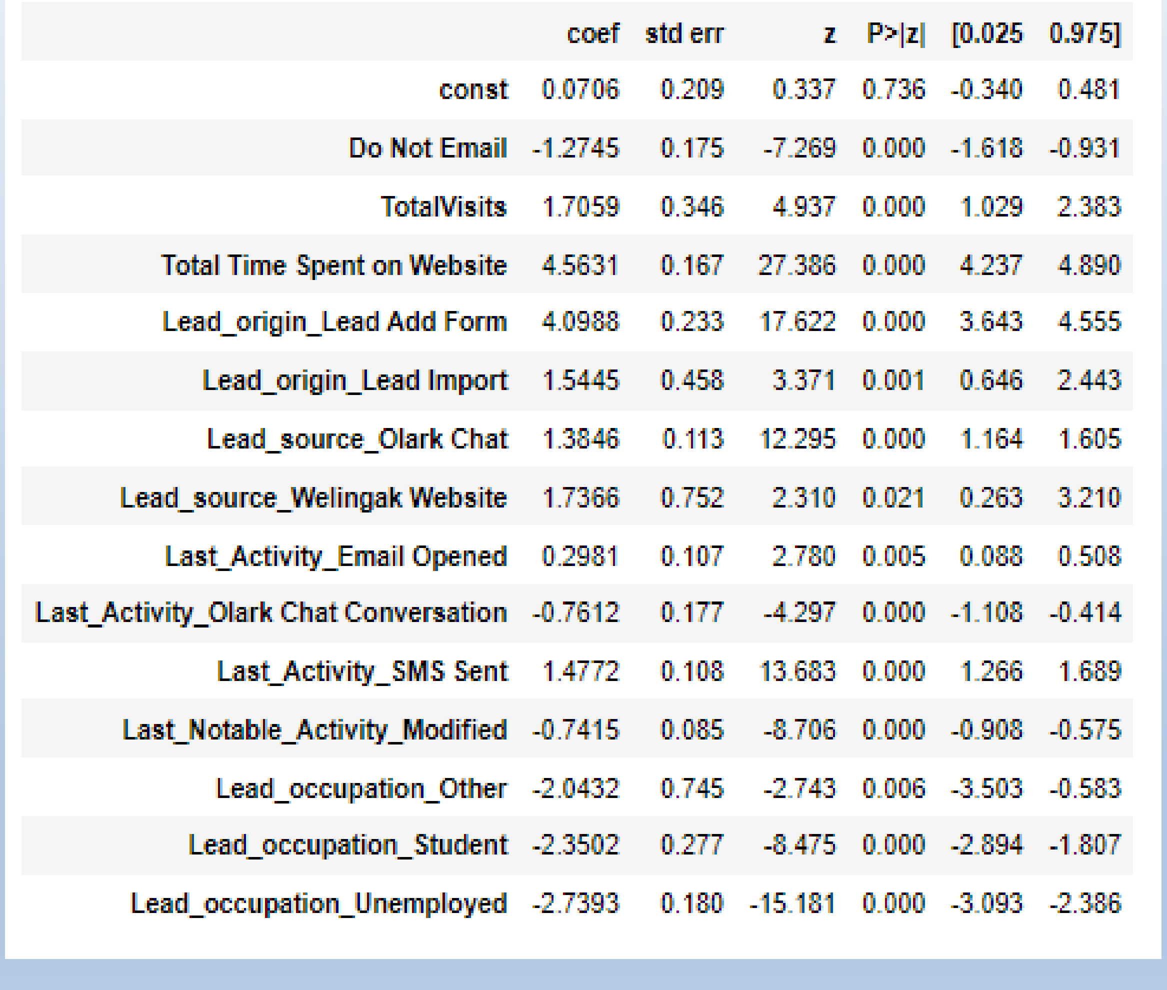


# Results

- Model Performance



| | | | |
|---|---|---|---|
| Dep. Variable: | Converted | No. Observations: | 6344 |
| Model: | GLM | Df Residuals: | 6329 |
| Model Family: | Binomial | Df Model: | 14 |
| Link Function: | logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -2696.3 |
| Date: | Mon, 17 Oct 2022 | Deviance: | 5392.5 |
| Time: | 11:50:49 | Pearson chi2: | 6.63e+03 |
| No. Iterations: | 7 | | |
| Covariance Type: | nonrobust | | |

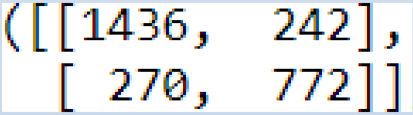


| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0706 | 0.209 | 0.337 | 0.736 | -0.340 | 0.481 |
| Do Not Email | -1.2745 | 0.175 | -7.269 | 0.000 | -1.618 | -0.931 |
| TotalVisits | 1.7059 | 0.346 | 4.937 | 0.000 | 1.029 | 2.383 |
| Total Time Spent on Website | 4.5631 | 0.167 | 27.386 | 0.000 | 4.237 | 4.890 |
| Lead_origin_Lead Add Form | 4.0988 | 0.233 | 17.622 | 0.000 | 3.643 | 4.555 |
| Lead_origin_Lead Import | 1.5445 | 0.458 | 3.371 | 0.001 | 0.646 | 2.443 |
| Lead_source_Olark Chat | 1.3846 | 0.113 | 12.295 | 0.000 | 1.164 | 1.605 |
| Lead_source_Welingak Website | 1.7366 | 0.752 | 2.310 | 0.021 | 0.263 | 3.210 |
| Last_Activity_Email Opened | 0.2981 | 0.107 | 2.780 | 0.005 | 0.088 | 0.508 |
| Last_Activity_Olark Chat Conversation | -0.7612 | 0.177 | -4.297 | 0.000 | -1.108 | -0.414 |
| Last_Activity_SMS Sent | 1.4772 | 0.108 | 13.683 | 0.000 | 1.266 | 1.689 |
| Last_Notable_Activity_Modified | -0.7415 | 0.085 | -8.706 | 0.000 | -0.908 | -0.575 |
| Lead_occupation_Other | -2.0432 | 0.745 | -2.743 | 0.006 | -3.503 | -0.583 |
| Lead_occupation_Student | -2.3502 | 0.277 | -8.475 | 0.000 | -2.894 | -1.807 |
| Lead_occupation_Unemployed | -2.7393 | 0.180 | -15.181 | 0.000 | -3.093 | -2.386 |

All variables are significant.

# Model Evaluation

- Accuracy of our model is 81.17%, which looks fine.
- Confusion matrix:



```
([[1436,  242],
  [ 270,  772]]
```

- Specificity : 85.57%
- Sensitivity: 74.08%
- Precision: 76.13%
- Recall: 74.00%

# Recommendations

- Since data was missing for many variables, we should focus getting more information.
- Sales team should more resources on people who are visiting the website more frequently and spending more time on the website.
- Try to increase the reach for lead to fill up forms.
- Sales team should prioritise leads coming from lead source ‘Welingak Website’ and ‘Reference’.

# Thank You