# Lead score - Case study

MOCHARLA INDRAKIRAN REDDY

# PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. Company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their details, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, out of 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. 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.

# BUSINESS GOAL & STEPS INVOLVED IN SOLVING LEAD SCORE CASE STUDY

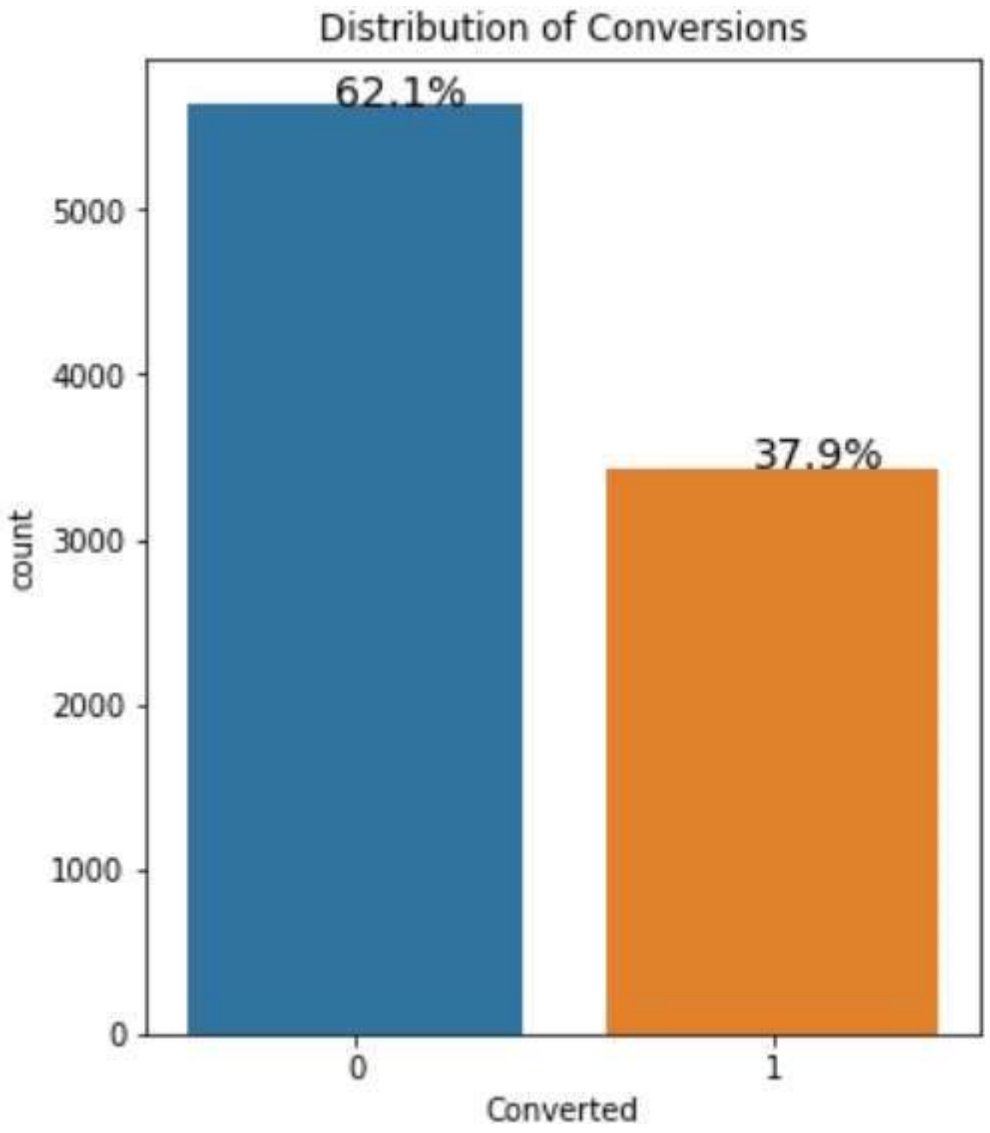
Build a model where we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

**STEPS INVOLVED**

1. Imported required libraries and Understand the data
2. Data cleaning and EDA
3. Data Pre-Processing (Creating Dummy Variables, Train-Test Split, Scaling, Looking at Correlation)
4. Feature Selection using RFE
5. Model Building and Checking P-Value and VIF
6. Plotting ROC Curve and finding optimal probability cutoff
7. Model Evaluation on Test Set
8. Finding the lead scores ( 100*probability of conversion)
9. Performance Comparison for train and test scores
10. Summary

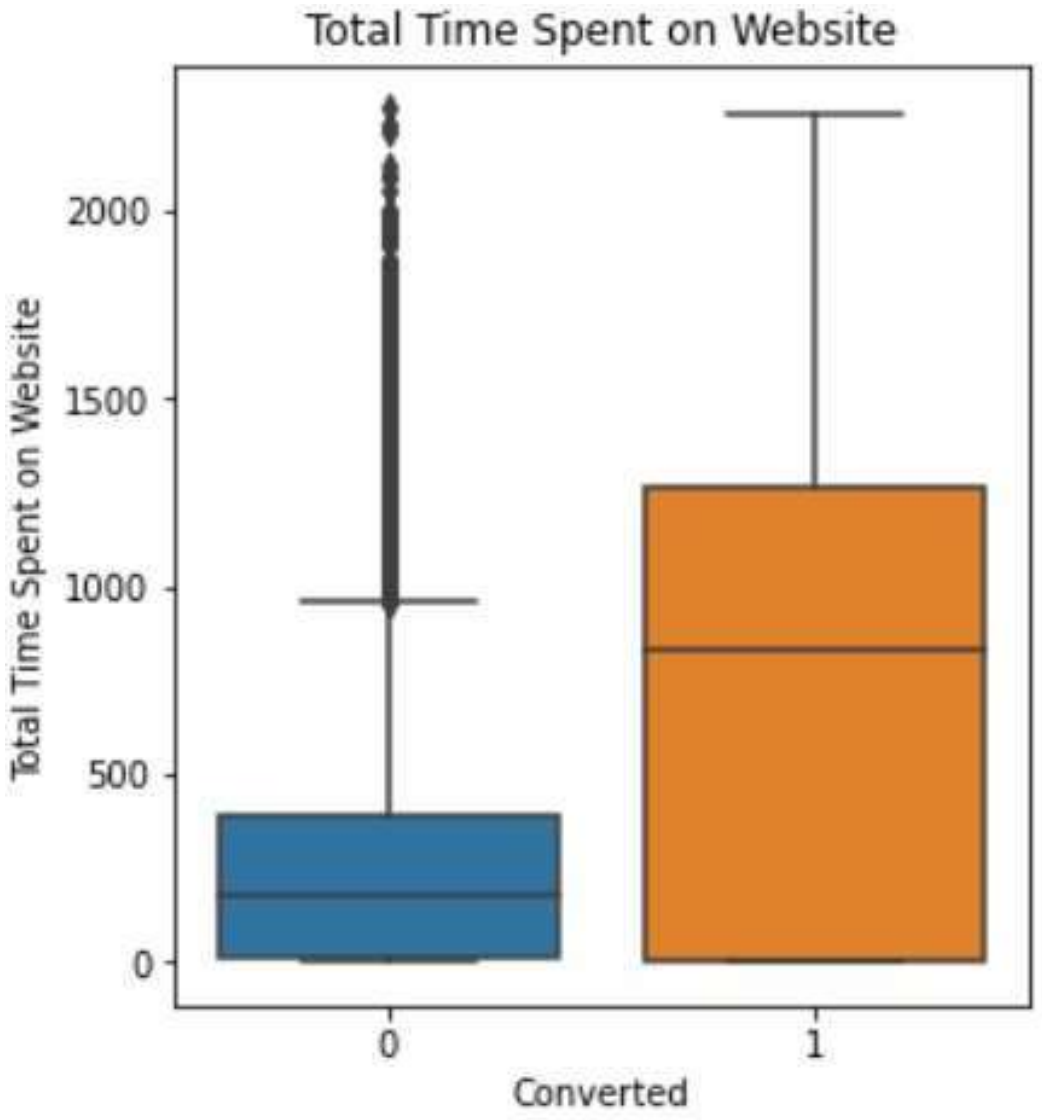
# TARGET VARIABLE IMBALANCE

- Leads converted to paying customers account for 37.9% of the data.
- Although the 50:50 situation is not ideal, we have enough converted consumers to proceed with additional analysis on this data set.

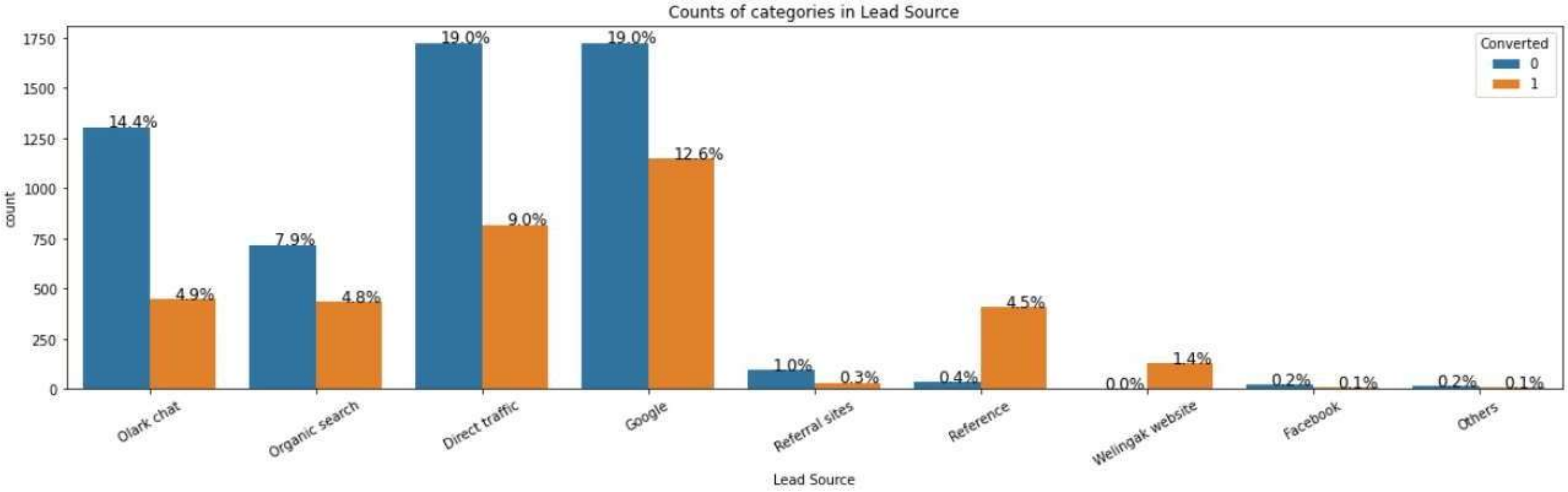


# TOTAL TIME SPENT ON WEBSITE (VS) CONVERTED

- The box plot clearly shows that the average time spent on the website by converted consumers is significantly higher than that of non-converted customers.
- As a result, leads who spend more time on the website have a higher chance of becoming paying clients.

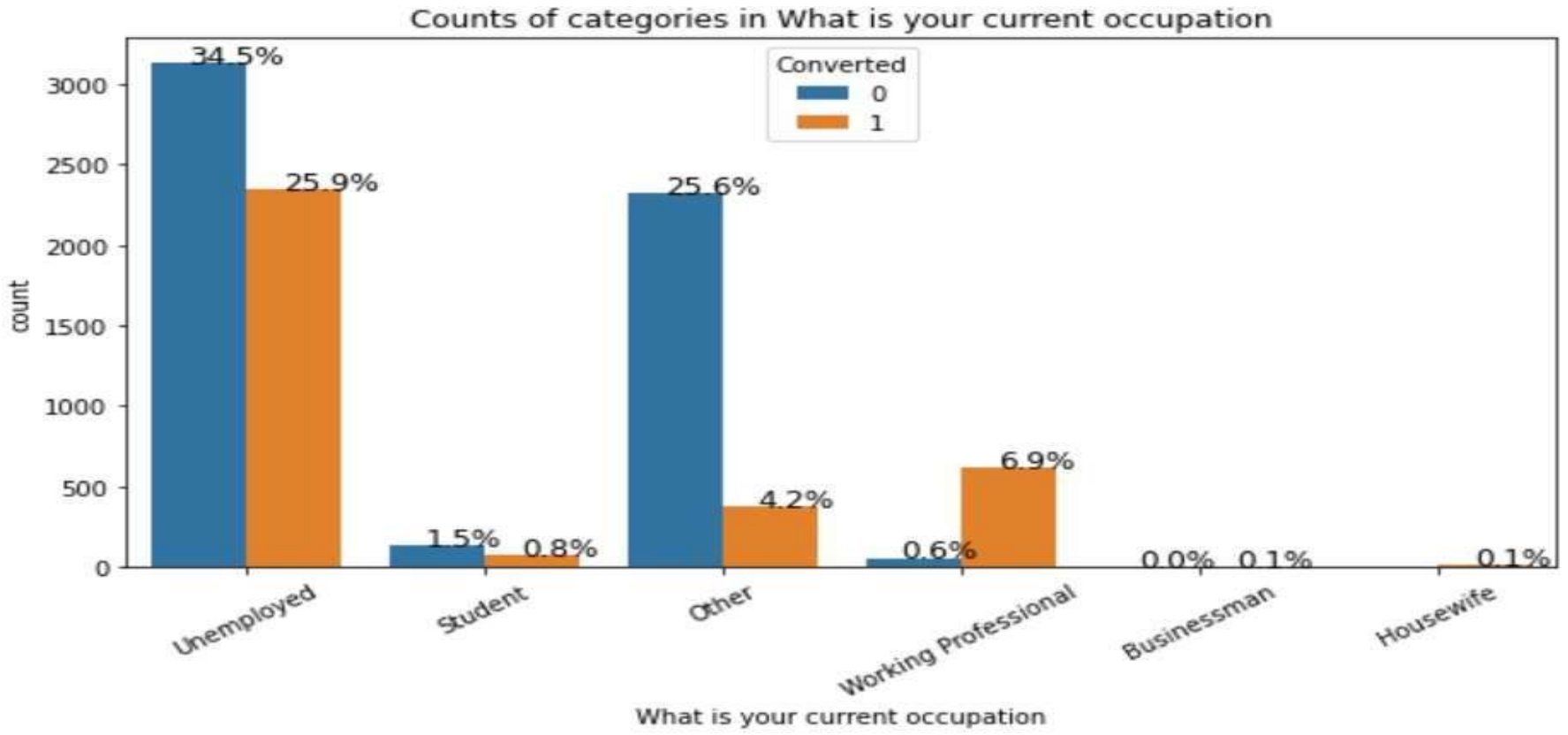


# COUNT PLOT OF DIFFERENT LEAD SOURCE CATEGORIES




- The lead sources "reference" and "welingak website" have a high conversion rate (i.e., more than 90% of the leads with source "reference" and "welingak website" have converted to paying customers).
- As a result, the business team can prioritise leads from these two sources.

# COUNT PLOT OF LEAD OCCUPATION

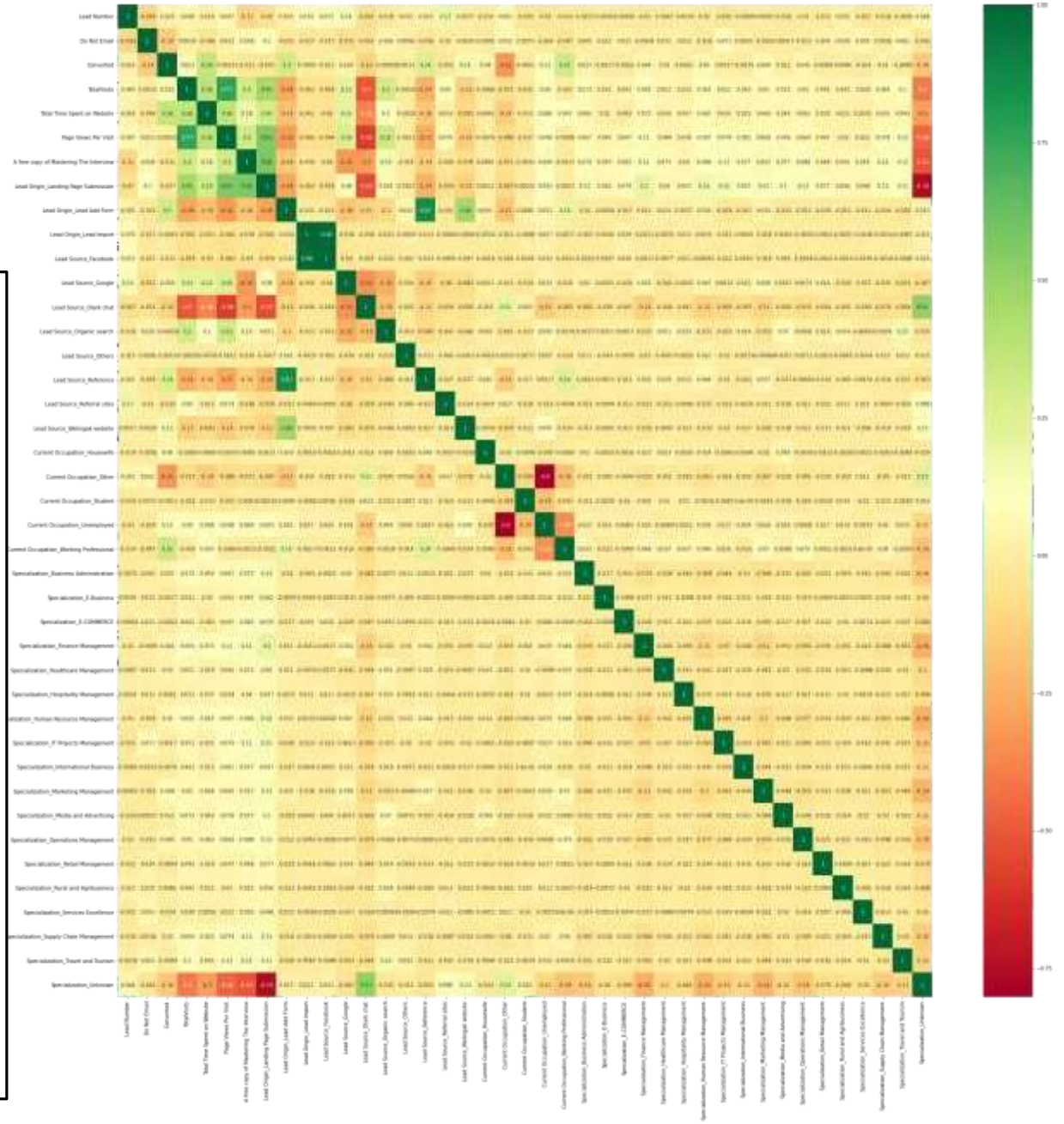



- Among the working professional conversion rate is high.
- To increase conversion rates, the X education team should explore contacting working professionals.

# CORRELATION OF ALL THE VARIABLES

- **Observations:**
- Between target variable and other columns there is no strong correlations
- It was observed that multiple other columns have high correlations (>0.7).
- These columns (mentioned below) have been excluded during model building to reduce multi-collinearity

- Pages per visit (0.77)
- Lead source_facebook (0.98)
- Specialization_unknown (-0.76)
- Lead origin_lead add form (0.87)
- Current occupation_other (-0.8)



# FINAL MODEL SUMMARY

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Converted | No. Observations: | 6351 |
| Model: | GLM | Df Residuals: | 6340 |
| Model Family: | Binomial | Df Model: | 10 |
| Link Function: | logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -2854.9 |
| Date: | Sun, 10 Apr 2022 | Deviance: | 5709.8 |
| Time: | 03:44:29 | Pearson chi2: | 6.40e+03 |
| No. Iterations: | 7 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -2.0718 | 0.080 | -25.771 | 0.000 | -2.229 | -1.914 |
| Do Not Email | -1.4877 | 0.165 | -9.014 | 0.000 | -1.811 | -1.164 |
| Total Time Spent on Website | 1.1434 | 0.039 | 29.270 | 0.000 | 1.067 | 1.220 |
| Lead Origin_Lead Import | 1.1405 | 0.451 | 2.528 | 0.011 | 0.256 | 2.025 |
| Lead Source_Olark chat | 1.0992 | 0.097 | 11.341 | 0.000 | 0.909 | 1.289 |
| Lead Source_Reference | 3.8622 | 0.225 | 17.171 | 0.000 | 3.421 | 4.303 |
| Lead Source_Welingak website | 5.8439 | 0.722 | 8.089 | 0.000 | 4.428 | 7.260 |
| Current Occupation_Student | 1.1485 | 0.214 | 5.369 | 0.000 | 0.729 | 1.568 |
| Current Occupation_Unemployed | 1.3264 | 0.083 | 16.067 | 0.000 | 1.165 | 1.488 |
| Current Occupation_Working Professional | 3.7716 | 0.191 | 19.768 | 0.000 | 3.398 | 4.146 |
| Specialization_Marketing Management | 0.2691 | 0.114 | 2.363 | 0.018 | 0.046 | 0.492 |

# METRICS ON TESTING AND TRAINING DATASETS

| | Train Data | Test Data |
|---|---|---|
| Accuracy | 78.25% | 77.67% |
| Sensitivity | 81.03% | 79.88% |
| Specificity | 76.52% | 76.41% |
| Precision | 68.37% | 65.88% |
| Recall | 81.03% | 79.88% |

- **Observations:**
- The identification of promising clients is a business requirement. Sensitivity (TP / (TP+FN)) should therefore be enhanced.
- On the training and test datasets, the model has a sensitivity/recall of 81.03 percent and 79.88 percent, respectively. This is in line with the CEO of X Education's target lead conversion rate of around 80%.

# CONCLUSION

Conclusion:

•Major signs that a lead will be converted into a hot lead include:

- Website of Lead Source Welingak - A lead generated through the Welingak website is more likely to convert.
- Lead Source Reference - A lead that has been referred by another hot lead has a higher chance of being converted.
- Working Professional Current Occupation - A working professional lead is more likely to convert.

•Major signs that a lead will not be converted into a hot lead include:

- Do Not Email - A lead who refuses to receive emails is less likely to be converted into a potential lead.

•The model's sensitivity score is around 81 percent, which is close to the CEO's target of 80 percent.