upGrad

# Lead Score Case Study

By Krishna Katta and Neetima Verma

Executive Program in DS C-39

13-06-2022

upGrad

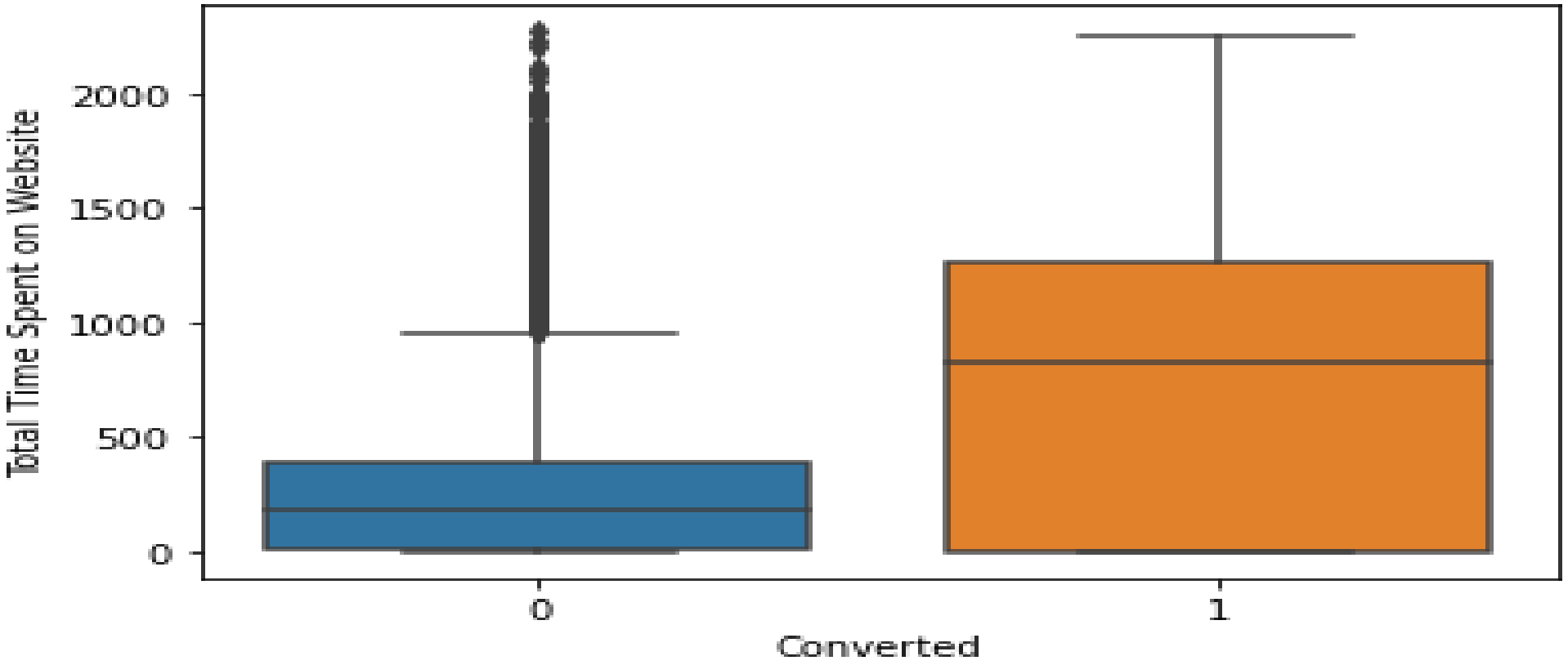
*Problem Statement:*

The Data of customer has been provided by an education company with numerus predictors. The task is to construct a model to predict if a particular customer will enroll in the program or not. The customer who enroll in the program is called Hot Leads. This will help the sales team, their attention to potential Hot Leads to increase the conversion rate.

Analysis based on “Total Time Spent on Website”:

Observation: The following can be observed.

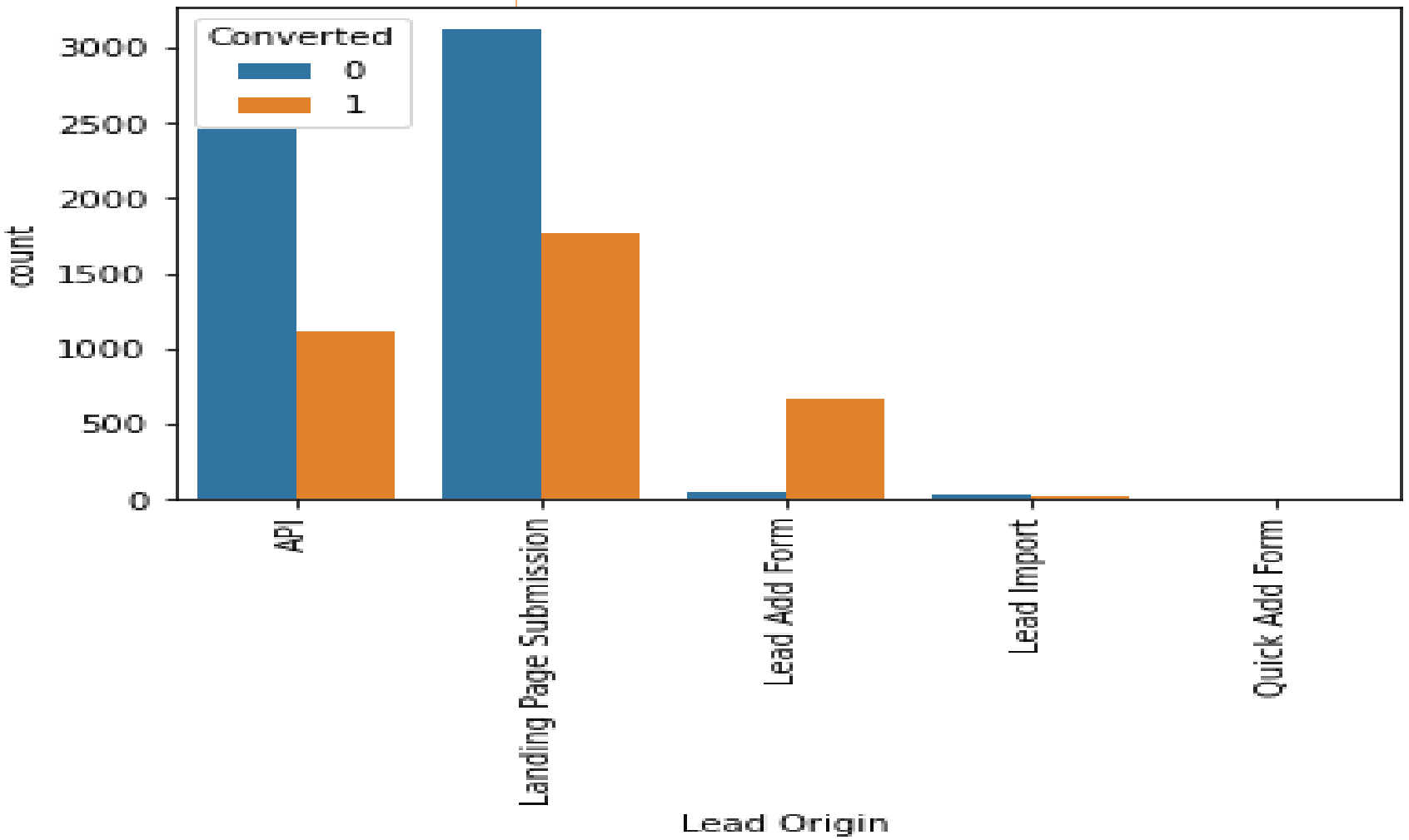
- Clearly customers who have spend more time on the website have higher probability of enrolling with the Academy.



Analysis based on “Lead Origin”:

Observation: The following can be observed.

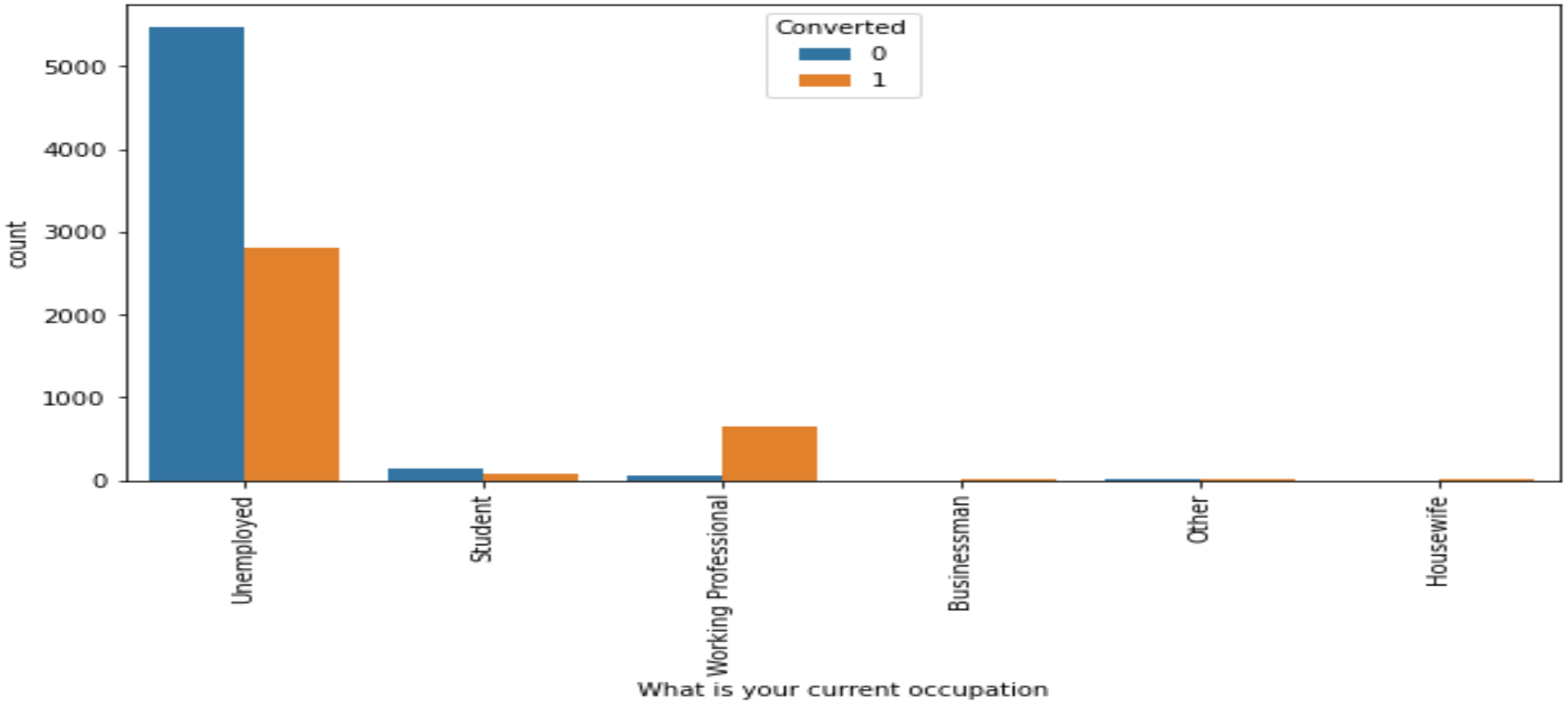
- Lead Add Form has high conversion ratio.



Analysis based on “Occupation”:

Observation: The following can be observed.

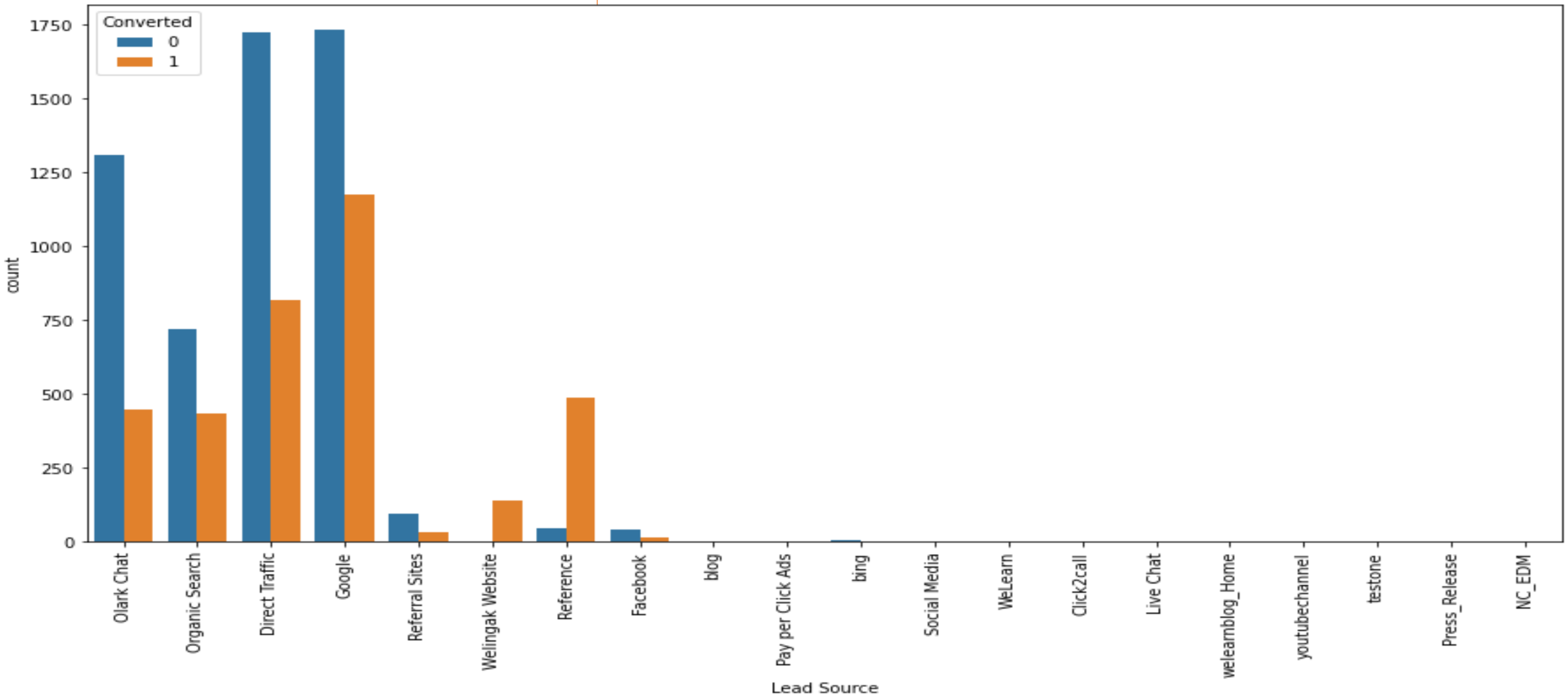
- Though number of converted customers are high in "Unemployed" category, It is the working professionals who have higher conversion rate



Analysis based on “Lead Source”:

Observation: The following can be observed.

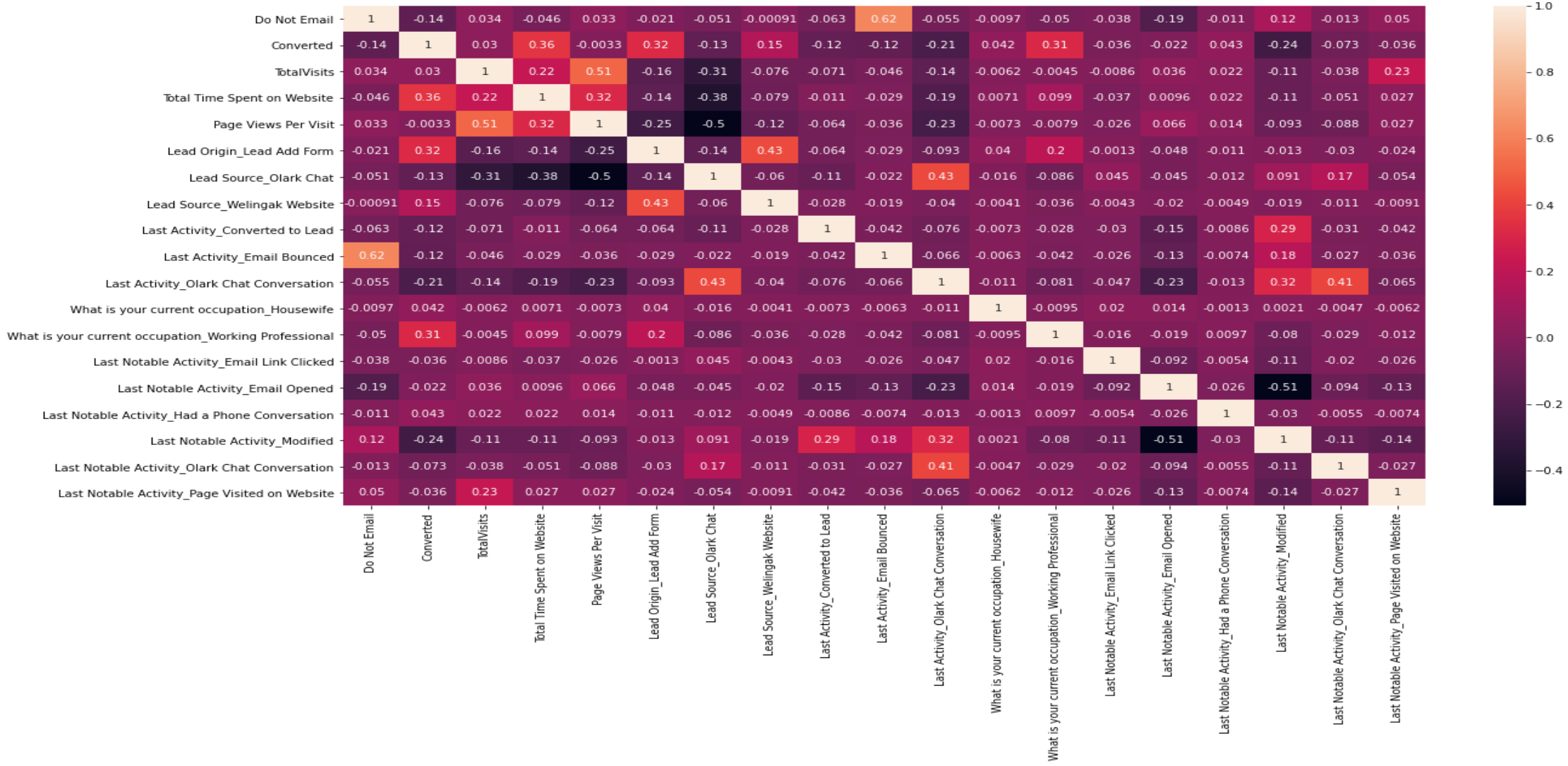
- “Reference” leads and “Welingak Website” categories have high conversion rate.



Analysis based on Correlation on RFEs:

Observation: By observing Heatmap, the following can be observed.

- From the above Heatmap it is clear that "Total time Spent on Website", "Lead Origin_Lead Add Form", "Working professionals" and "Lead Source_Welingak Website" have high correlation with target variable "Converted"



# Data preparation

upGrad

- Replacing “Select” with nan
- Dropping columns which has more that 40% null values.
- Replacing “Yes” and “No” with 1’s and 0’s
- Imputing null values in Categorical variables with Mode.
- Imputing null values in Numerical variables with Mean.
- Dropping variables which had single unique values.
- Created Dummy variables for all the Categorical variables and Dropping the Categorical Variables.

# Test-Train Split

upGrad

- Splitting data into X data which are predicted variables and y as target variable.
- Splitting the data into 70% Training Data and 30% Testing Data.

# Feature Scaling

upGrad

- Used MinMaxScaler to re-scale the following numerical variables.
  - ✓ TotalVisits
  - ✓ Total Time Spent on Website
  - ✓ Page Views Per Visit

# Logistic Regression Model Building

upGrad

- Model Building using GLM

  *logm1 = sm.GLM(y_train,(sm.add_constant(X_train)), family = sm.families.Binomial())*
  *logm1.fit().summary()*

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Converted | **No. Observations:** | 6468 |
| **Model:** | GLM | **Df Residuals:** | 6384 |
| **Model Family:** | Binomial | **Df Model:** | 83 |
| **Link Function:** | logit | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | nan |
| **Date:** | Sun, 20 Jun 2021 | **Deviance:** | nan |
| **Time:** | 12:26:33 | **Pearson chi2:** | 8.48e+18 |
| **No. Iterations:** | 100 | | |
| **Covariance Type:** | nonrobust | | |

# Feature Selection Using RFEs

upGrad

- Recursive feature elimination carried out on the data set.
- GLM Models built on using the chosen variables and metrics like Accuracy, Sensitivity and Specificity are compared for different number of variables in the RFE method.
- 18 variables are chosen based on trial and error method. More than 18 variables does not increase the accuracy of the model by much. Less than 18 Variables reduces the accuracy of the model.

# Variance Inflation Factor (VIF)

upGrad

- VIF for 18 variables are checked for values greater than 5. None found.

| | Features | VIF |
|---|---|---|
| 3 | Page Views Per Visit | 3.00 |
| 15 | Last Notable Activity_Modified | 2.56 |
| 1 | TotalVisits | 2.00 |
| 9 | Last Activity_Olark Chat Conversation | 1.99 |
| 2 | Total Time Spent on Website | 1.87 |
| 0 | Do Not Email | 1.86 |
| 8 | Last Activity_Email Bounced | 1.82 |
| 5 | Lead Source_Olark Chat | 1.69 |
| 13 | Last Notable Activity_Email Opened | 1.67 |
| 4 | Lead Origin_Lead Add Form | 1.43 |
| 16 | Last Notable Activity_Olark Chat Conversation | 1.36 |
| 7 | Last Activity_Converted to Lead | 1.25 |
| 6 | Lead Source_Welingak Website | 1.24 |
| 11 | What is your current occupation_Working Profes... | 1.17 |
| 17 | Last Notable Activity_Page Visited on Website | 1.16 |
| 12 | Last Notable Activity_Email Link Clicked | 1.05 |
| 10 | What is your current occupation_Housewife | 1.01 |
| 14 | Last Notable Activity_Had a Phone Conversation | 1.01 |

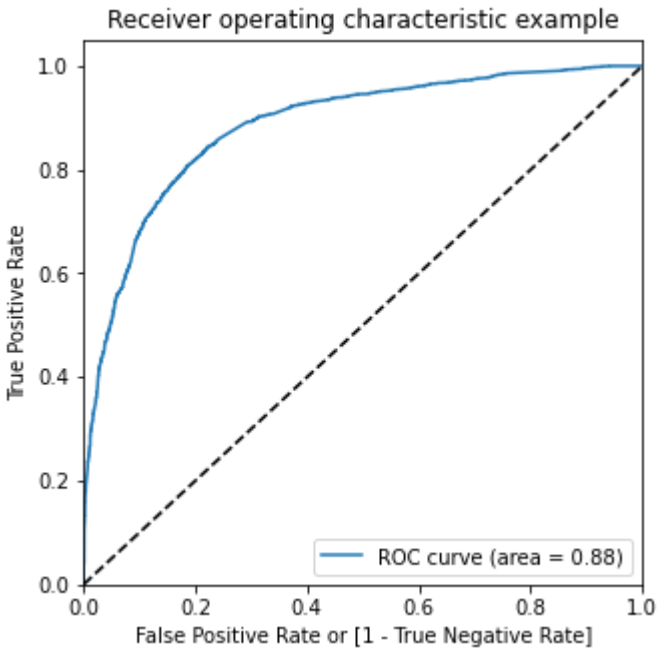
# ROC

Assuming 0.3 as cutoff probability:

Sensitivity: 0.8381

Specificity: 0.7813

Accuracy: 0.8030

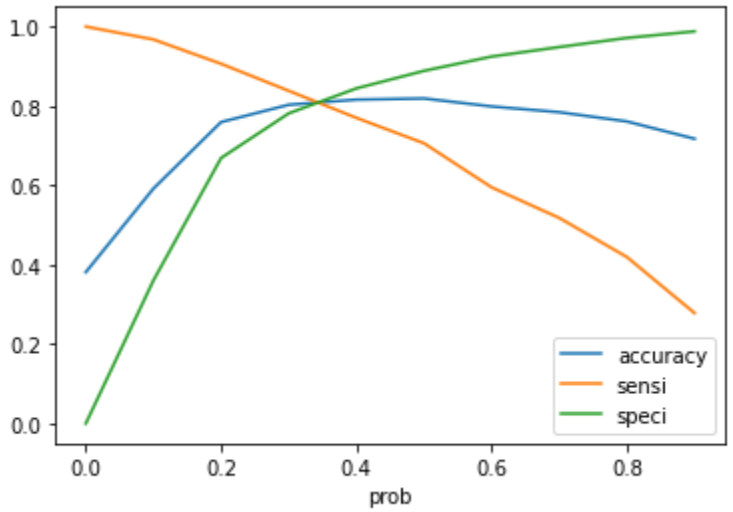



The Shape of the curve confirms that of a good model.

# Finding Optimal Cutoff Probability

upGrad

Plotting Sensitivity Vs Specificity Vs Accuracy for cutoff probability values between 0 and 1




Inference: From the curve above 0.35 is the optimum point to take it as cutoff probability.

# Finding Lead Score

upGrad

Lead score is calculated for each customer in the training data set using the probability value of y.

| | Converted | Converted_Prob | ID | predicted | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | final_predicted | Lead_Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0.255123 | 1871 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 26 |
| **1** | 0 | 0.245922 | 6795 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 |
| **2** | 0 | 0.299014 | 3516 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 30 |
| **3** | 0 | 0.816244 | 8105 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 82 |
| **4** | 0 | 0.194206 | 3934 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 |

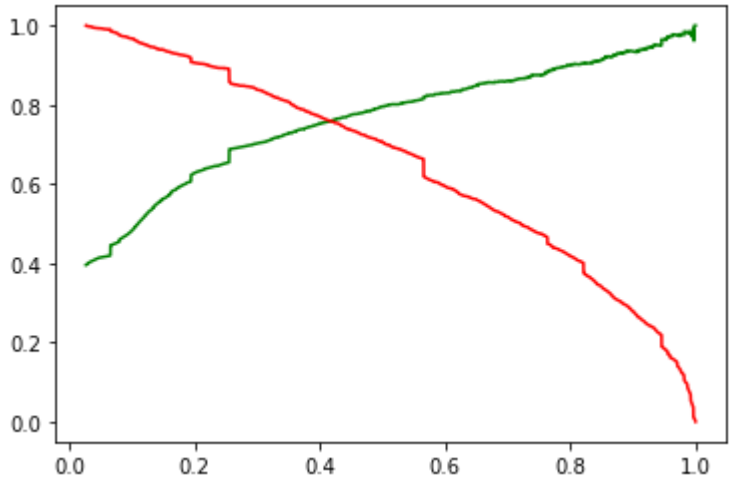
# Precision and Recall

upGrad

Precision Score = 0.7025

Recall Score = 0.8381

Plotting precision score Vs Recall score for various cutoff probabilities between 0 and 1




Inference: From the graph above the intersection point is 0.4 which is close to the value infered from ROC Curve.

# Evaluation on the test set

upGrad

The GLM model built using the training data is used to predict the target variable in the test data.

Accuracy = 0.8170

Sensitivity = 0.8036

Specificity = 0.8258

**The model prediction on the test data is satisfactory based on the above Model evaluation**

# Recommendation

upGrad

- To increase conversion rate, the sales team should focus on customers who are predicted as hot leads by the model.
- To improve the conversion rate the academy should also focus on the variables having high impact on the prediction.