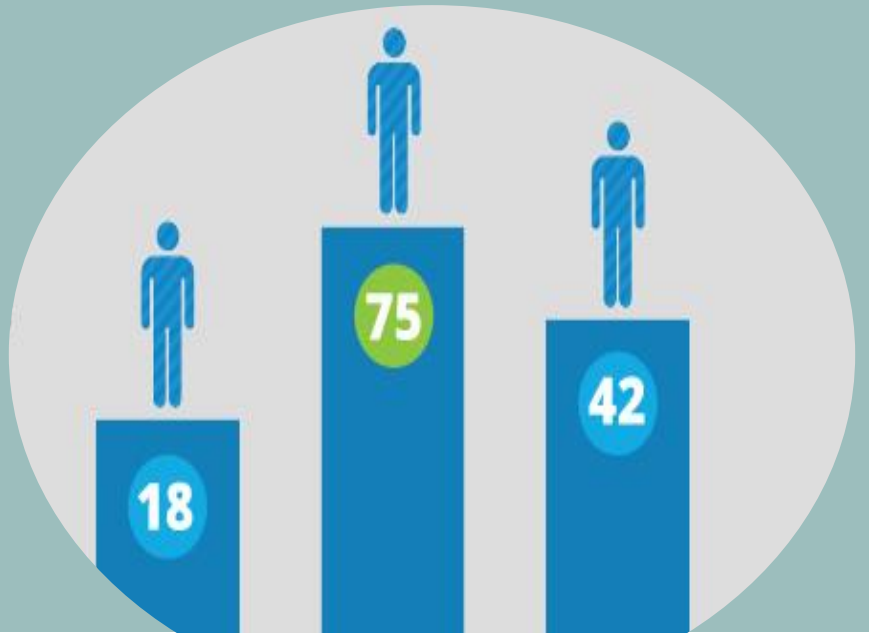


LEAD SCORING CASE STUDY

By Parmila Lama

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. The typical lead conversion rate at X education is around 30%. Help X Education select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Business Objectives:

- Find Hot leads, using a Machine Learning Model.

SOLUTION METHODOLOGY

Step 1: Clean Data

- Categorical values as “Select” was equivalent to null, had converted to null.
- Removed columns with null % of 50 and above.
- Imputed with appropriate values for other variables.
- Removed variables that were insignificant for the case study.

Step 2: EDA

- Performed Univariate Analysis for each variable against target variable “Converted”.
- Removed variables that were not contributing towards classification of the lead conversion.
- Encoded the binary categorical variables to 1/0 for Yes/No values.

SOLUTION METHODOLOGY

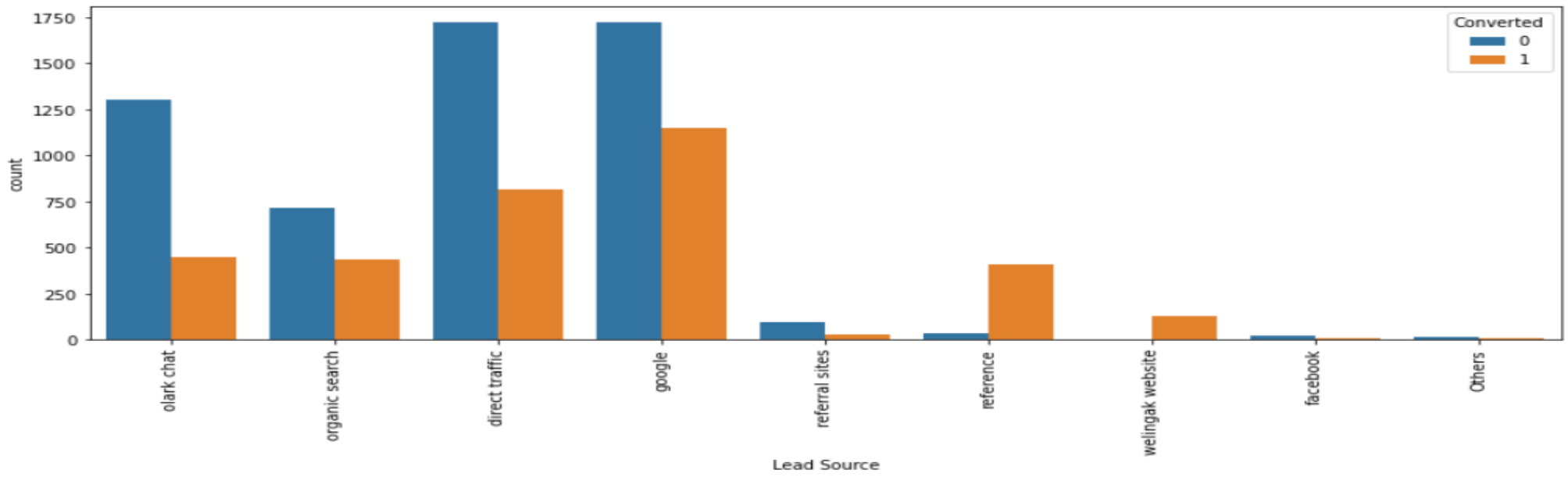
Step 3: Prepare Data

- Created dummy variables using one-hot encoding technique.
- Removed the original variables after creation of dummy variables

Step 4: Model Creation

- Created x, y variables
- Performed a Train-Test Split
- Performed Feature Scaling : Standardize the variables with high Variance
- Train the model
- Feature Selection using techniques: RFE, VIF,P Values
- Asses the models using metrics and ROC.
- Calculated the optimum cutoff probability
- Test the Model
- Generate the Lead scores between 0-100

EDA



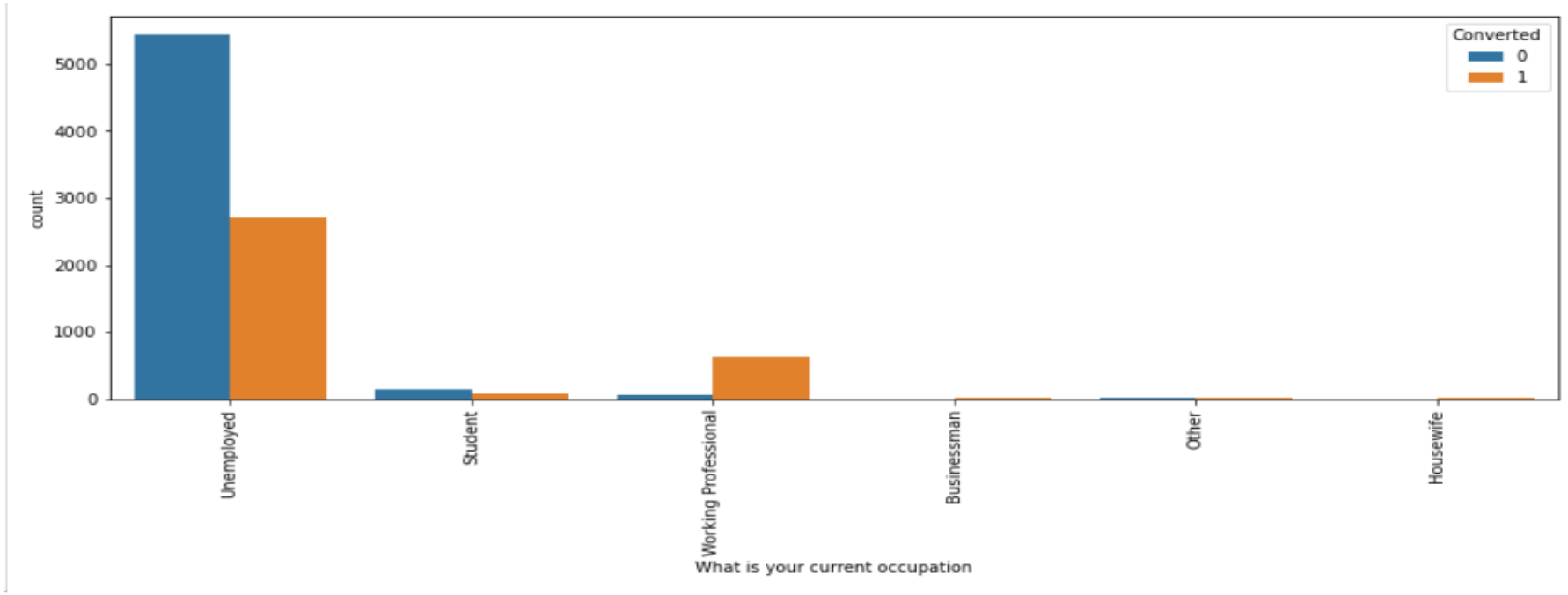
Inference

1. Leads from sources such as "google", direct traffic, organic search and olark chat must focussed on.
2. High Number of Leads coming from Source "Olark chat", but lead conversion is low.

Recommendation

1. Leads from Olark chat should be focussed as many leads from this source.

EDA

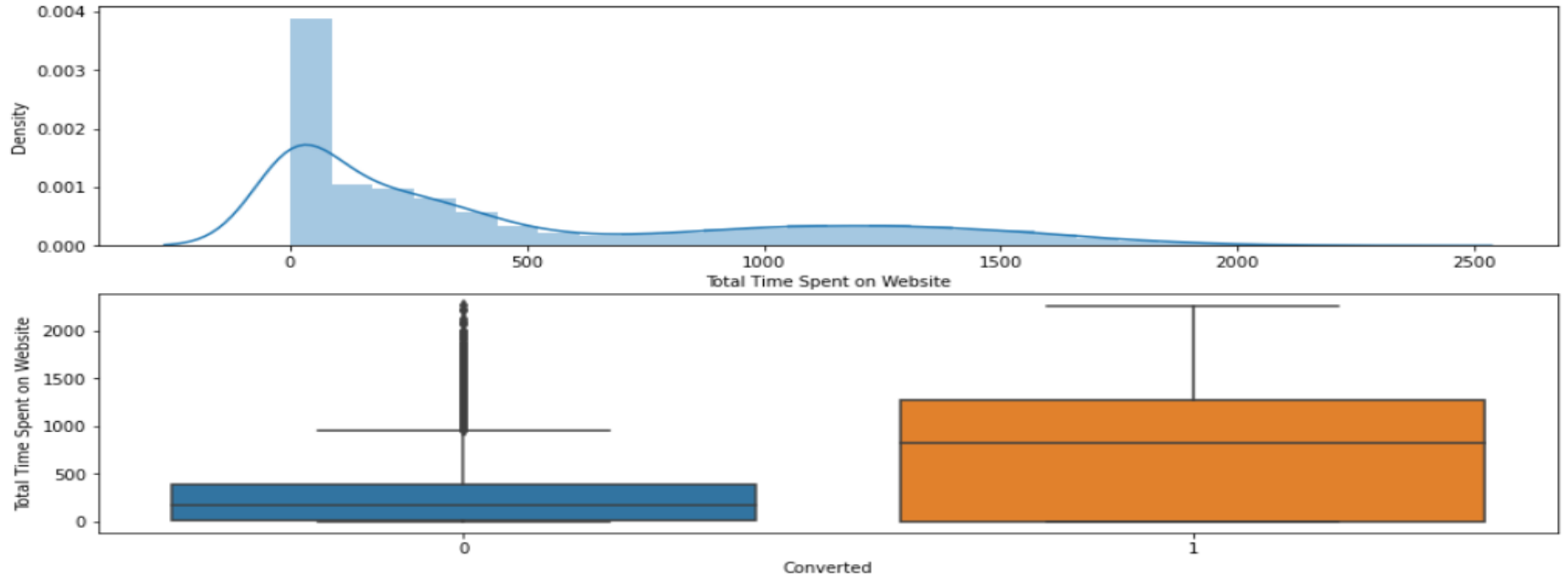


Inference

1. Highest Number of leads are Unemployed.
2. Working professional has high lead converstion rate

EDA

Out[99]: <AxesSubplot:xlabel='Converted', ylabel='Total Time Spent on Website'>



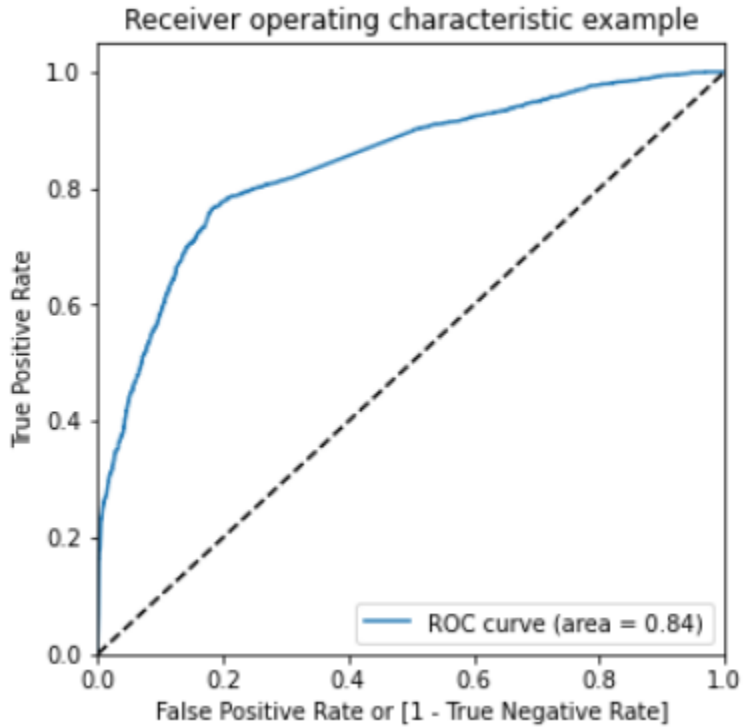
Inference

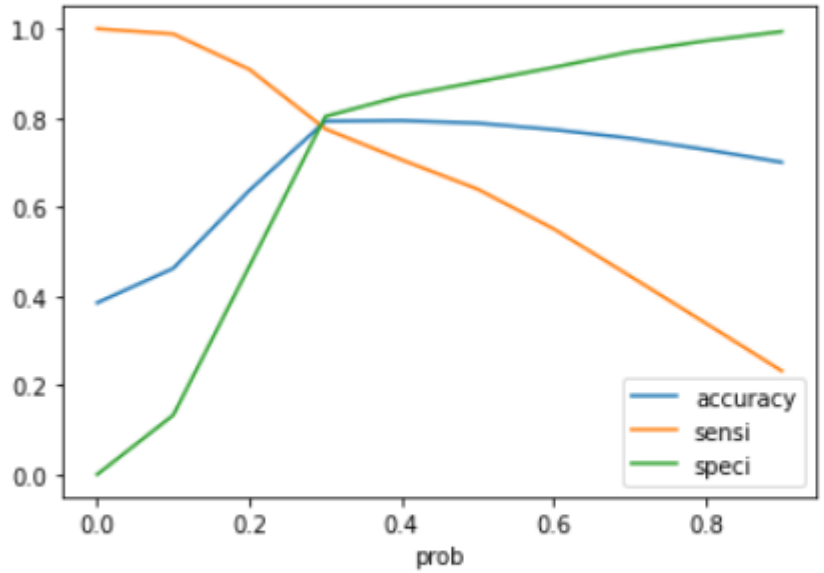
Leads spending more time on the weblise are more likely to be converted.

MODEL BUILDING

- Created x, y variables
- Performed a Train-Test Split
- Performed Feature Scaling
  - Standardize the variables with high Variance
- Train the model
- Feature Selection using techniques:
  - RFE
  - VIF
  - P Values
- Asses the models using metrics and ROC.
- Calculated the optimum cutoff probability
- Test the Model
- Generate the Lead scores between 0-100

ROC CURVE





From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

FINAL MODEL METRICS

Train Set:

```
Overall Accuracy : 79.24%
Sensitivity : 77.51%
Specificity : 80.33%
Precision : 71.17%
Recall : 77.51%
```

Test Set:

```
Overall Accuracy : 79.47%
Sensitivity : 76.84%
Specificity : 80.96%
Precision : 69.72%
Recall : 76.84%
```

CONCLUSION

It was found that the following variables contributed towards most Customers:

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:

  => Google

  => Direct traffic

  => Organic search

  =>Welingak website

- When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.