Lead Scoring Case Study

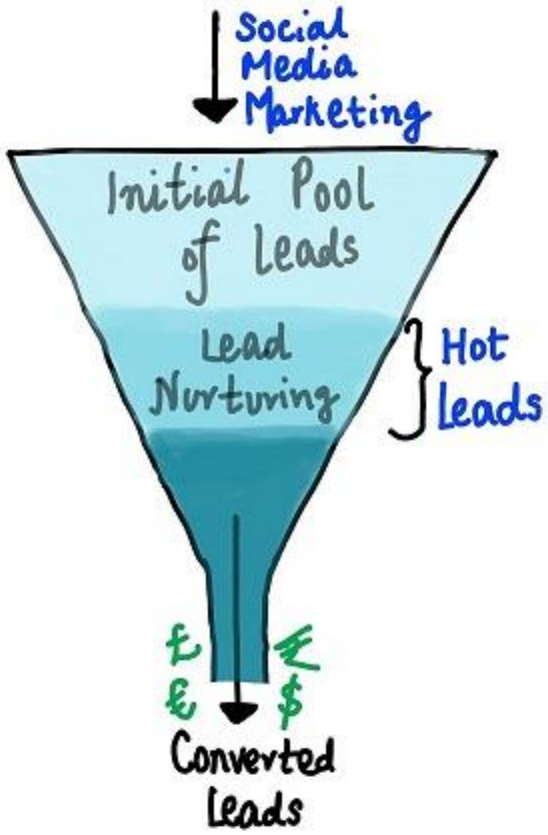
Team Members

Artik Gupta

Pabitra Kumar Sahoo

Abhishek Singh

Problem Statement



An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The 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 email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 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.

Objectives

- In order to aid the company in identifying the most promising leads, commonly referred to as 'Hot Leads,' with a lead conversion rate of approximately 80%.
- To construct a model where each lead is assigned a lead score, ensuring that customers with higher scores have an increased likelihood of conversion, while those with lower scores are associated with a lower chance of conversion.
- Assist the sales team in shifting their attention towards potential leads, preventing them from engaging in unproductive phone calls.

Approach

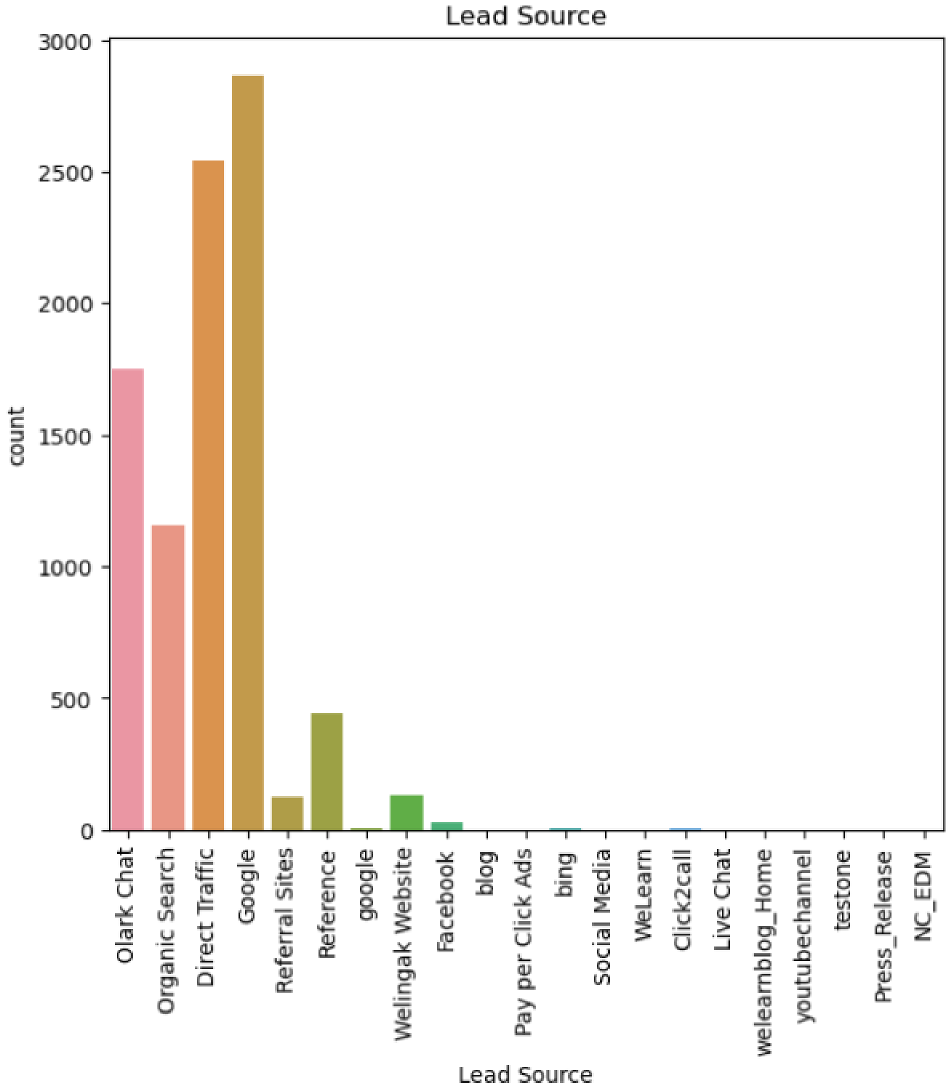
- Data Cleaning
- EDA
- Data Preparation
- Model Building
- Model Evaluation
- Predictions on test data
- Recommendations

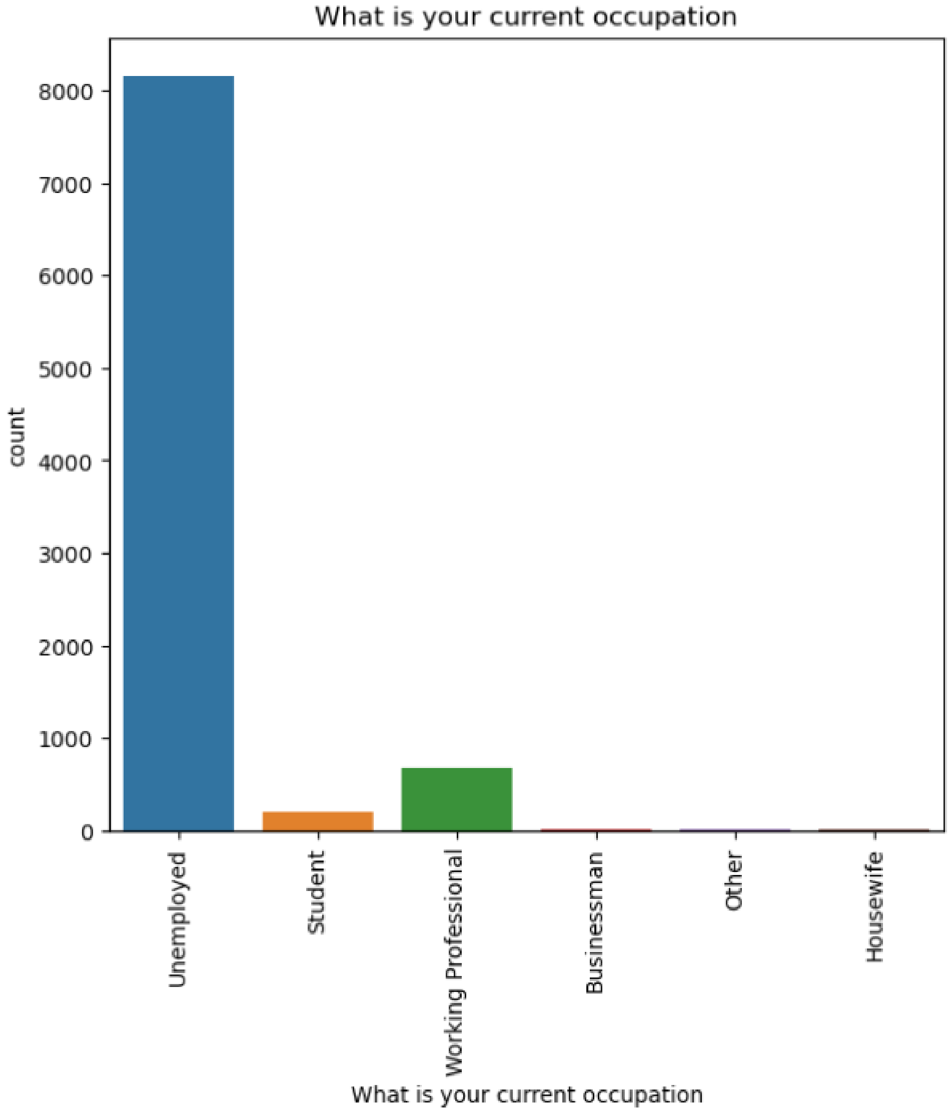
Data Cleaning

- The "Select" level indicates null values for certain categorical variables, signifying that customers did not make any selection from the provided list of options.
- Columns containing more than 40% null values were removed.
- Null values in categorical columns were addressed by considering their frequency and specific considerations during the handling process.
- Remove columns that do not contribute meaningful insights or value to the study objective.
- Imputation was used for some categorical variables.
- Columns that are not useful for modeling or have only one category of response were excluded.
- Values with low frequency were consolidated into the category labeled as "Others."

EDA – Univariate Analysis

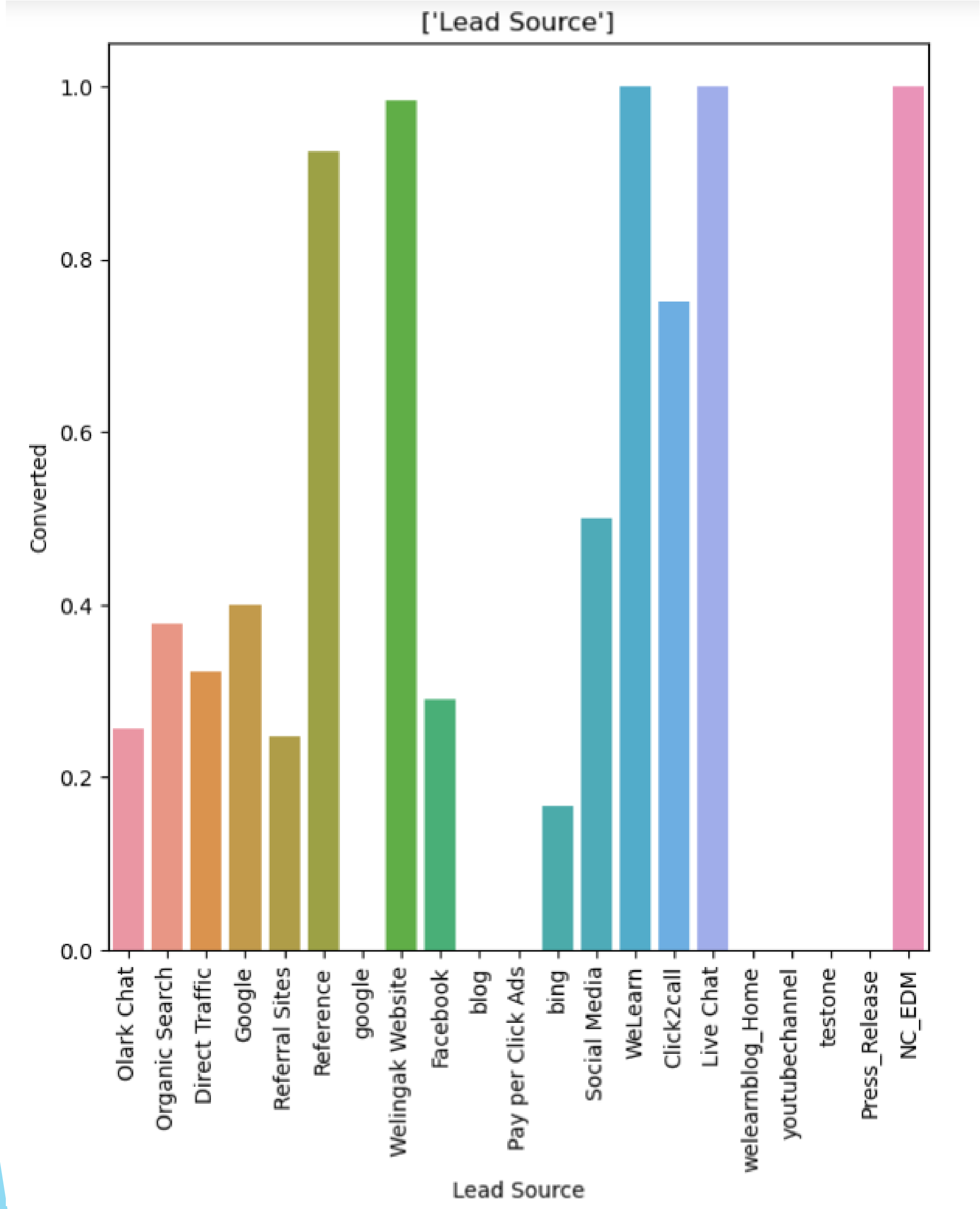
- Checking the count of categories:
  - Most of the leads are from traffic and Olark chat sources.
  - Most of the leads are unemployed.
  - Large proportion of leads are from mumbai city.

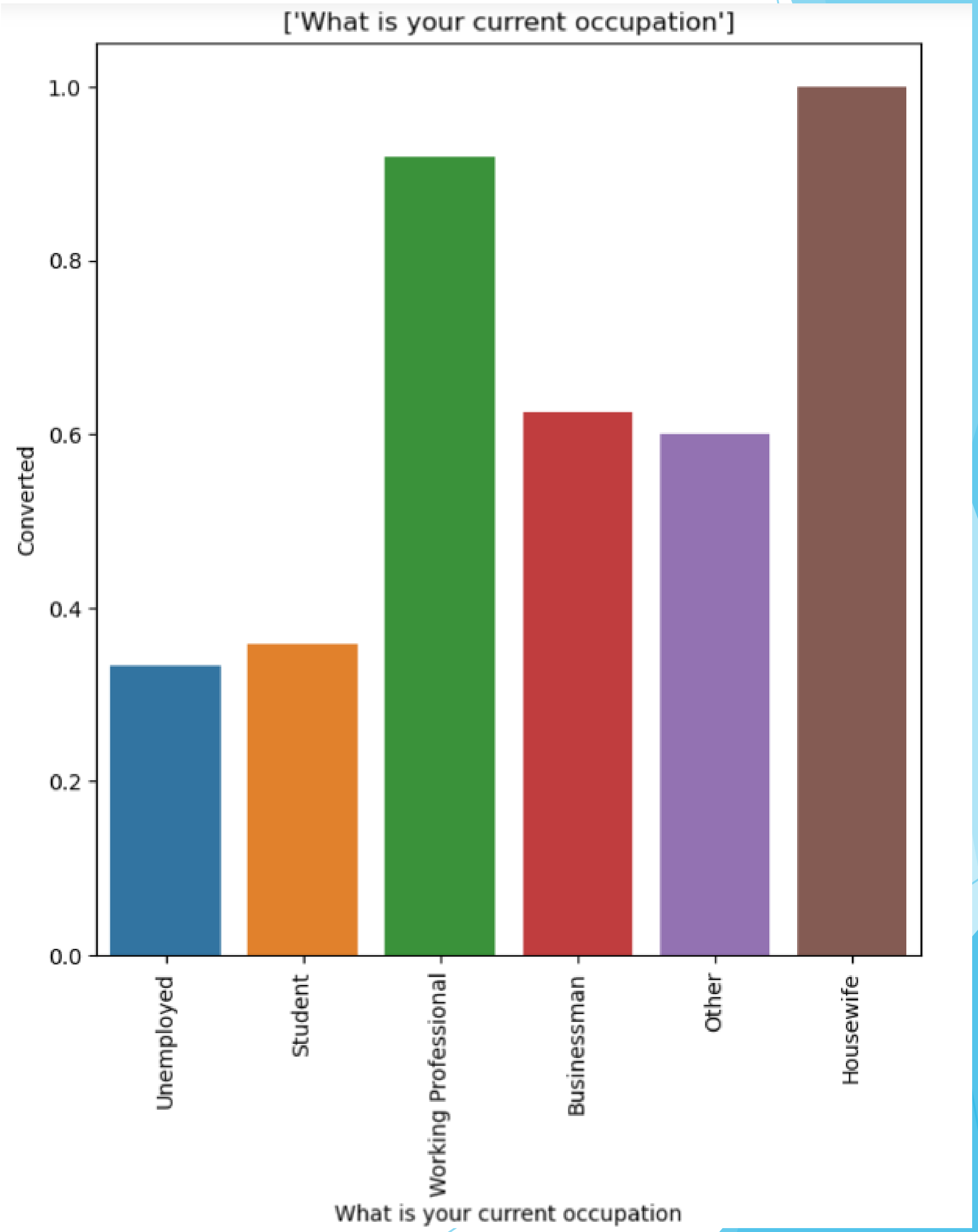

Lead Source
count
3000
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

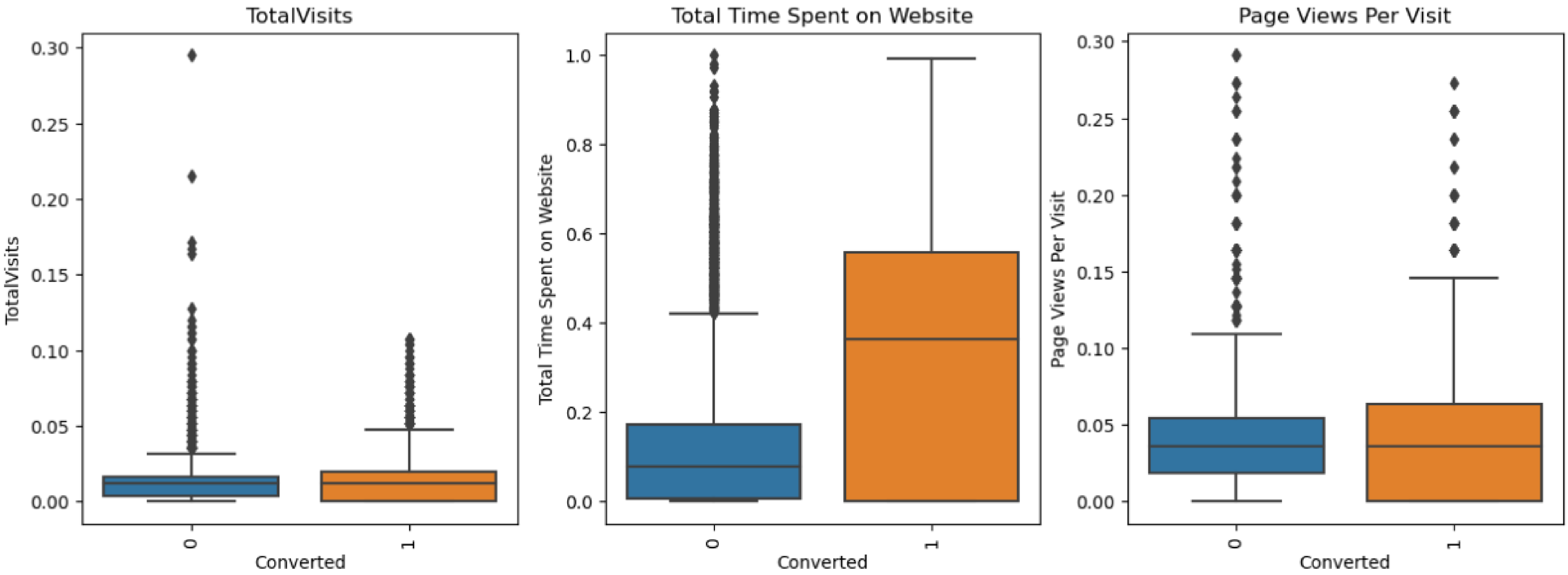

What is your current occupation
count
8000
7000
6000
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

EDA – Bivariate Analysis

- Relevant lead quality have high conversion ratio.
- Leads from online websites have high conversion ratio.


['Lead Source']
Converted
1.0
0.8
0.6
0.4
0.2
0.0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source


['What is your current occupation']
Converted
1.0
0.8
0.6
0.4
0.2
0.0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


TotalVisits
0.30
0.25
0.20
0.15
0.10
0.05
0.00
TotalVisits
0
1
Converted
Total Time Spent on Website
1.0
0.8
0.6
0.4
0.2
0.0
Total Time Spent on Website
0
1
Converted
Page Views Per Visit
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Page Views Per Visit
0
1
Converted

Data Preparation

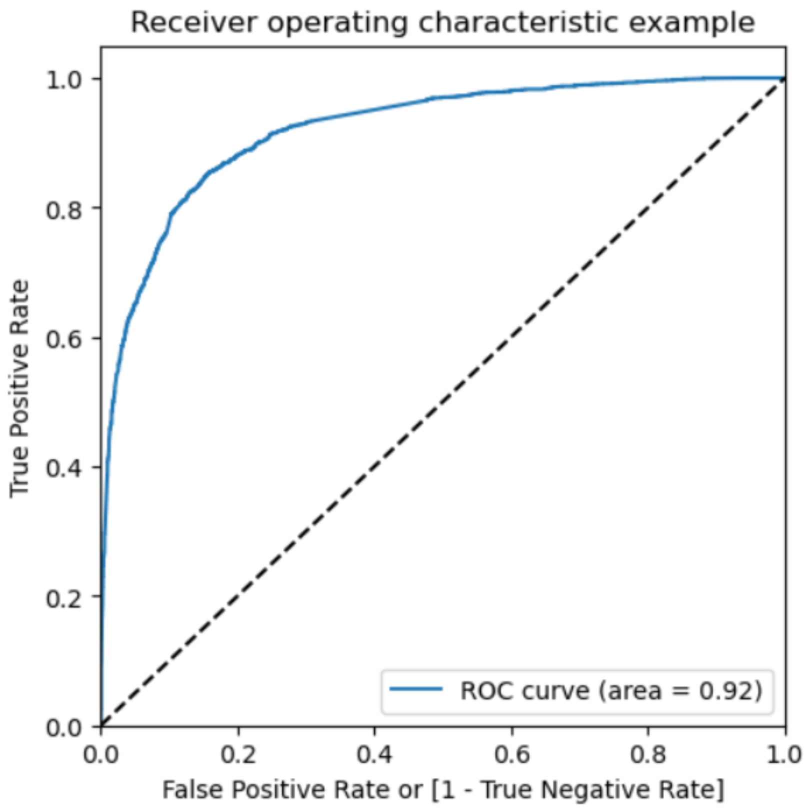
- In preceding steps, binary-level categorical columns were already encoded as 1s and 0s.
- Generated dummy features for categorical variables.
- Splitting Train and Test Sets
  - 80:20 % ratio was chosen for the split.
- Feature scaling
  - The MinMax scaler was employed to scale the features.
- Checking the correlations
  - Predictor variables that exhibited high correlation with each other were excluded.

Model Building

- To increase model performance and decrease computation time we need to reduce the dimensions and large number of features in dataset.
- Performing Recursive Feature Elimination (RFE) is crucial to select only the essential columns.
- A manual feature reduction process was employed to construct models, involving the elimination of variables with a p-value greater than 0.05.
- Model 3 looks stable after three iteration with:
  - P-values within the threshold of significance (p-values < 0.05).
  - No evidence of multicollinearity was found, as indicated by Variance Inflation Factors (VIFs) less than 5.

Model Evaluation

- ROC Curve
  - The Area under the ROC curve is 0.92 out of 1, indicating a strong predictive model.
  - The curve closely aligns with the top-left corner of the plot, reflecting a model with a high true positive rate and a low false positive rate across all threshold values.



Predicting test data

- Train data
  - Confusion Matrix

```
[[4426,   54],
 [1568, 1203]]
```

  - Sensitivity: 0.43
  - Specificity: 0.98
  - Precision: 0.95

- Test data
  - Confusion Matrix

```
[[1033,  118],
 [153,  509]]
```

  - Sensitivity: 0.76
  - Specificity: 0.89
  - Precision: 0.81

Recommendations

- Introduce incentives or discounts for providing references that result in lead conversions, encouraging more referrals.
- Tailor messaging to effectively engage working professionals.
- Allocate a higher budget for advertising on the Welingak Website.
- Aggressively target working professionals, given their high conversion rates and potentially better financial situations to afford higher fees.
- Emphasize features with positive coefficients to enhance targeted marketing strategies.
- Formulate strategies to attract high-quality leads from top-performing lead sources.
- Optimize communication channels by assessing the impact of lead engagement.