DCS-52 Data Science(January 2023)

Lead Score Case Study

Prepared by:
1)Divya Dharshini
2)Dharshak Chandra P
3) Digmakumari Patel

Introduction

The assignment study is designed in a way to applied the knowledge of Machine learning, Logistic regression , EDA into the real scenario. This study not only aims at applying learned methods and tricks into the analysis but also be helpful for understanding the E-commerce education sector for its risk analysis, maximising the lead generation for better profitability with help of its data sets.

Problem Statement

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Aim for Lead Scoring Assignment

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

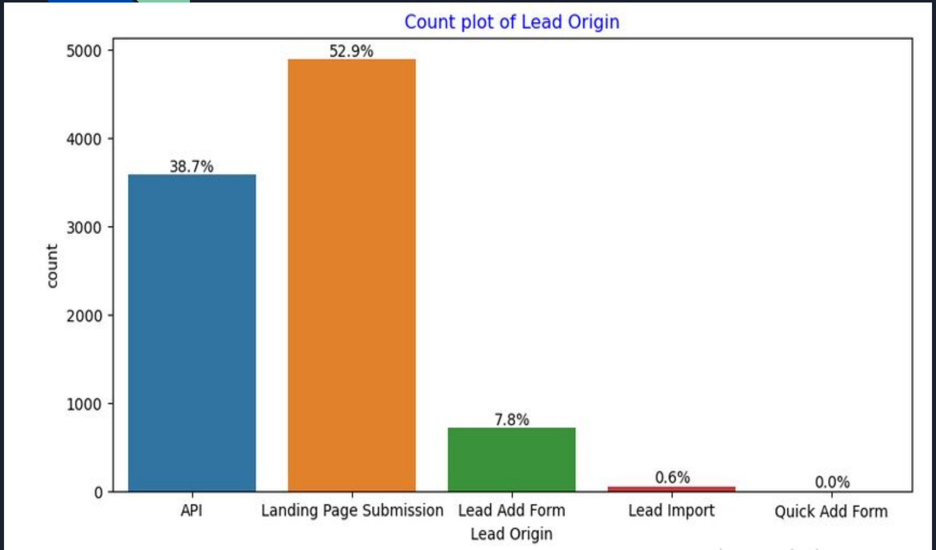
Procedure

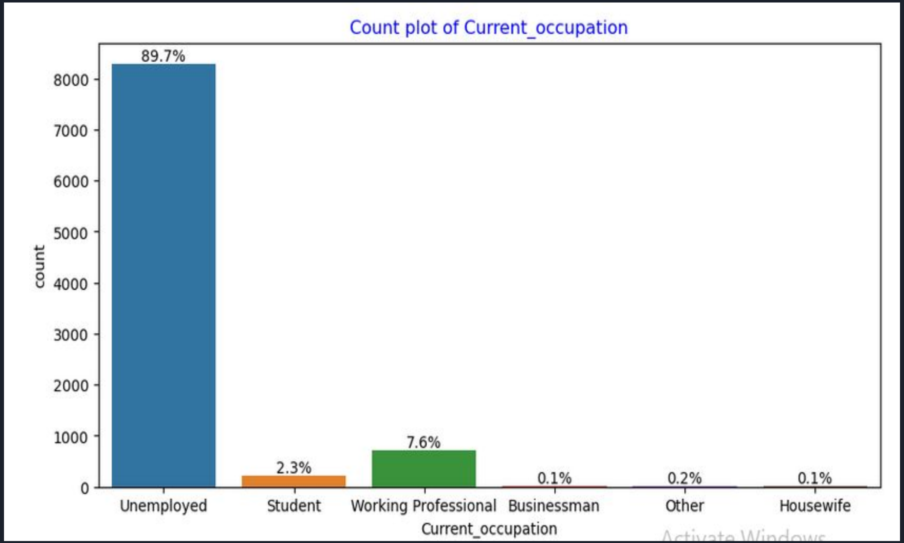
1. Reading Dataset
2. Data Handling
3. Data Cleaning
4. EDA
5. Data Preparation
6. Model Building
7. Drawing insight from the modeled data

Data Handling & Data Cleaning

- Eliminating unnecessary columns based upon the missing values maximality like :
  - Tags
  - Country
  - What matters most to you in choosing a course
  - City
  - Prospect ID
  - Lead Number
  - Last Notable Activity
- Filling missing categorical column with appropriate values:
  - Specialization,
  - Lead Source,
  - Last Activity,
  - What is your current occupation
- Handling Numerical column using value filling of mode and cleaning the data, checking and treating outliers.
- Finally handling “Select” value upon replacing & considering it as a NaN value.
- Eliminating skewed columns due to inaccurate prediction from regression model : Do Not Call, Search,Newspaper Article,X Education Forums,Newspaper,Digital Advertising,Through Recommendations.
- Standardising the value for entries in column
- Mapping binary values

EDA - Univariate Analysis

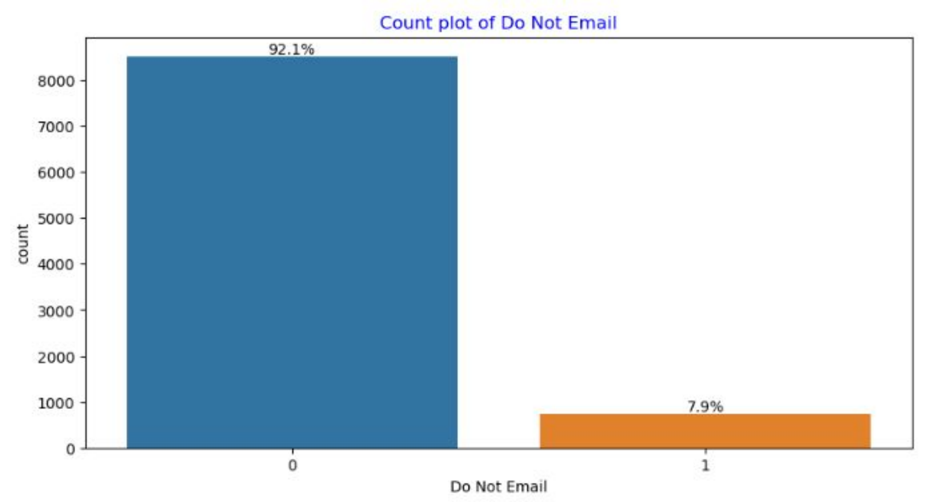


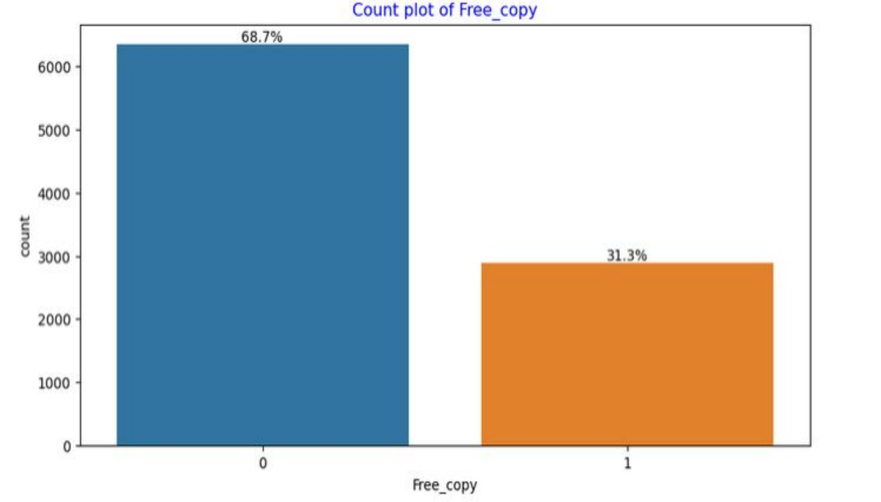


In Lead Origin : "Landing Page Submission" identified 53% customers, "API" identified 39%.

In Current Occupation: Those with nearly 90% being unemployed are likely to be converted to a hot lead

EDA- Univariate Analysis

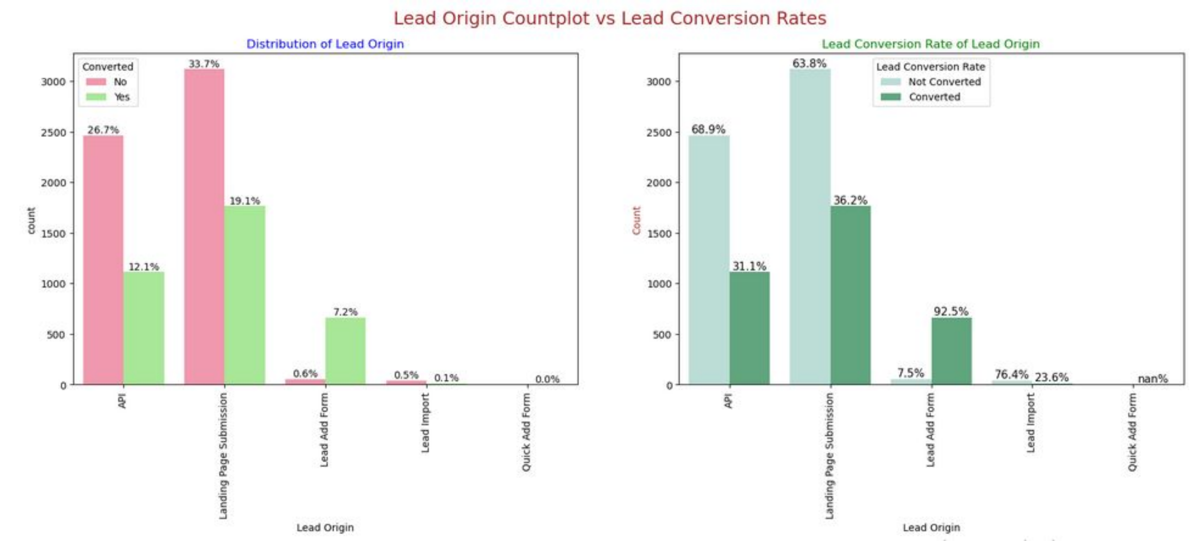




In Do Not Email, 92% of those polled said they do not want to be emailed about the course.

In Free copy, maximum user do not want any copy.

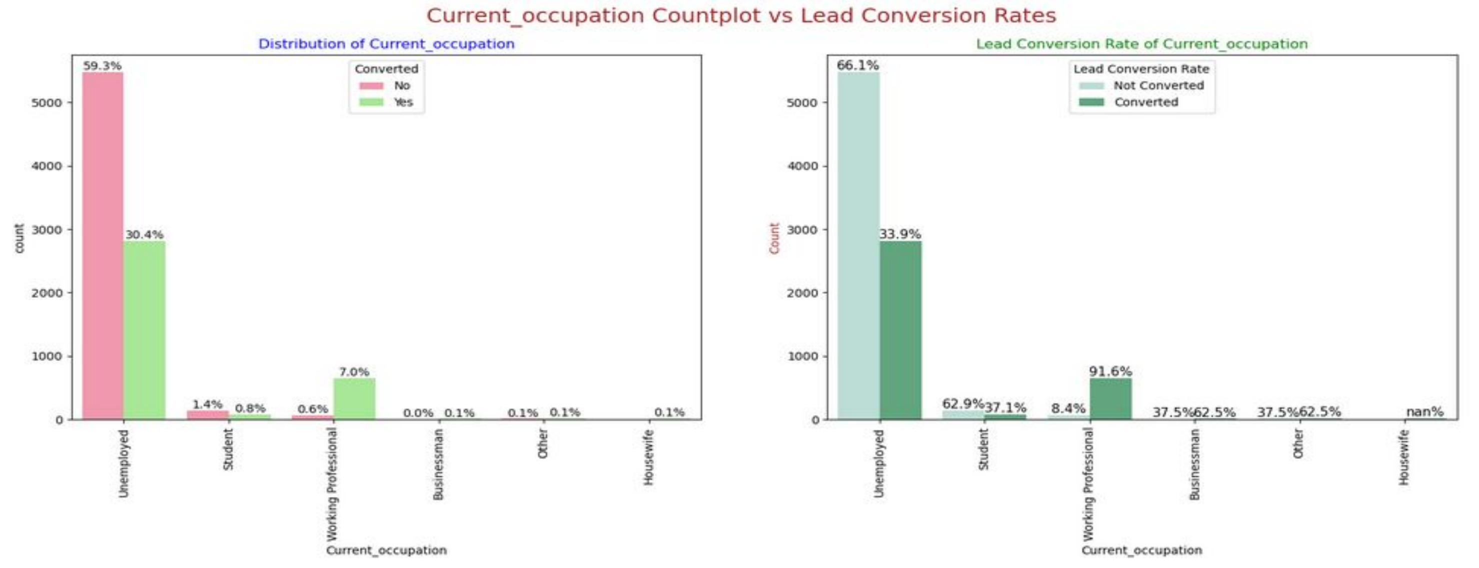
EDA - Bivariate Analysis



Lead Origin vs Lead Conversion Rates

"Landing Page Submission" generated around 52% of all leads, with a **lead conversion rate (LCR) of 36%.** With a lead conversion rate (LCR) of 31%, the **"API"** identified around 39% of customers.

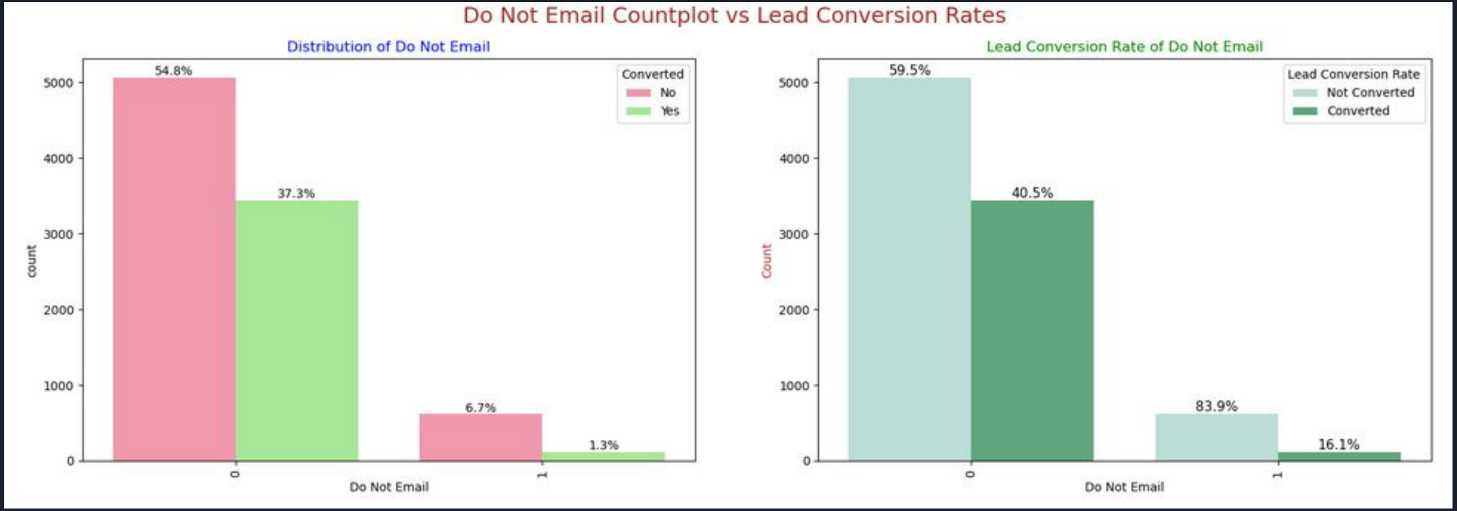
EDA - Bivariate Analysis



Current Occupation Vs Lead Conversion rates

With a **lead conversion rate (LCR) of 34%**, almost 90% of the consumers are unemployed. While Working Professionals account for only 7.6% of total clients, they have a nearly **92% lead conversion rate (LCR).**

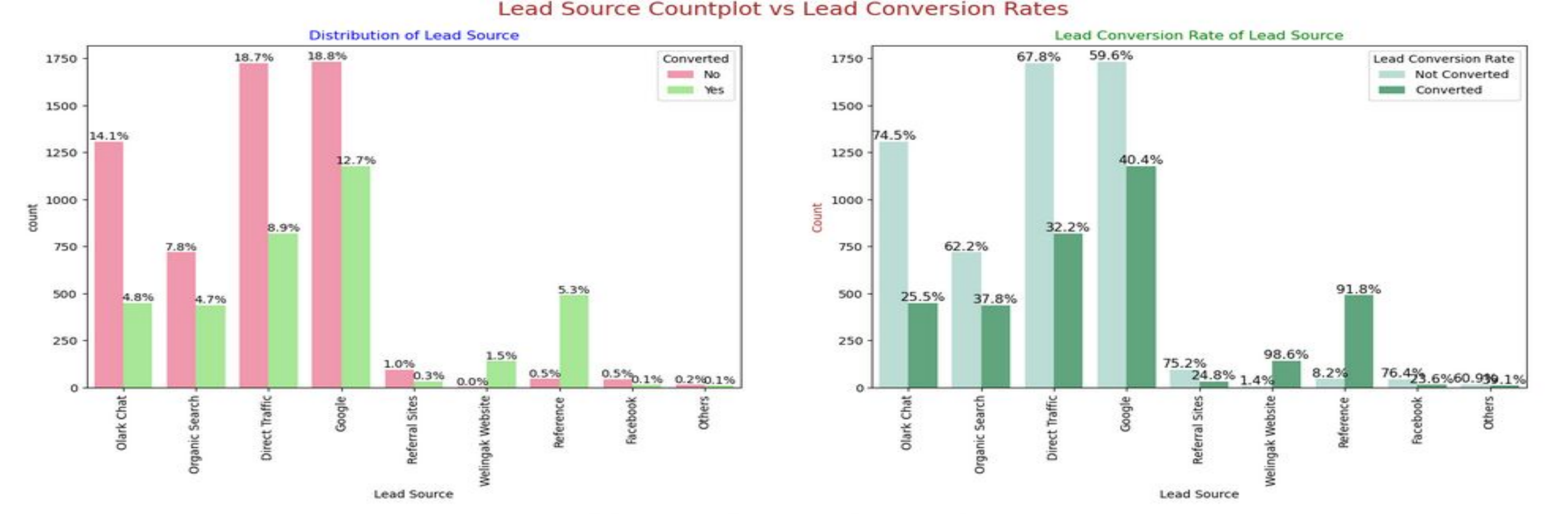
EDA - Bivariate Analysis



Do not email vs Lead Conversion Rates

92% of those polled said they do not want to be emailed about the course.

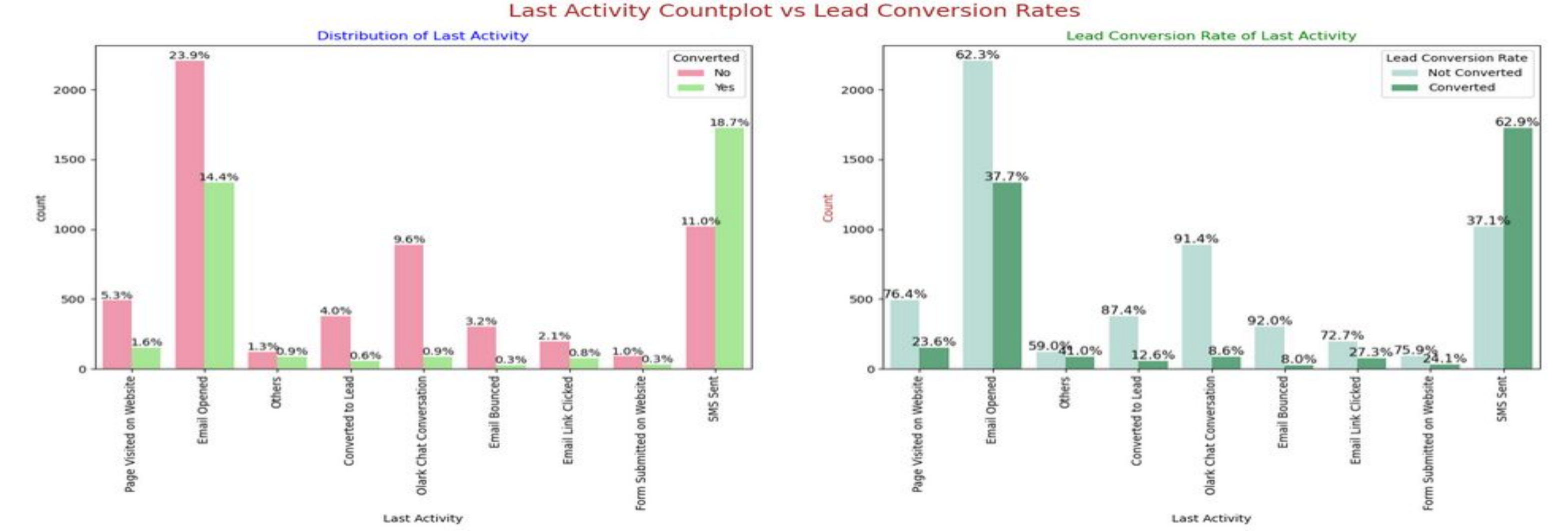
EDA- Bivariate Analysis



Lead source vs Lead conversion

Google has a **LCR of 40%** out of 31% customers, Direct Traffic has a **32% LCR** with 27% customers, which is lower than Google, Organic Search has a **37.8%** LCR but only 12.5% of customers contribute, and Reference has a **LCR of 91%** but only 6% of customers come from this Lead Source.

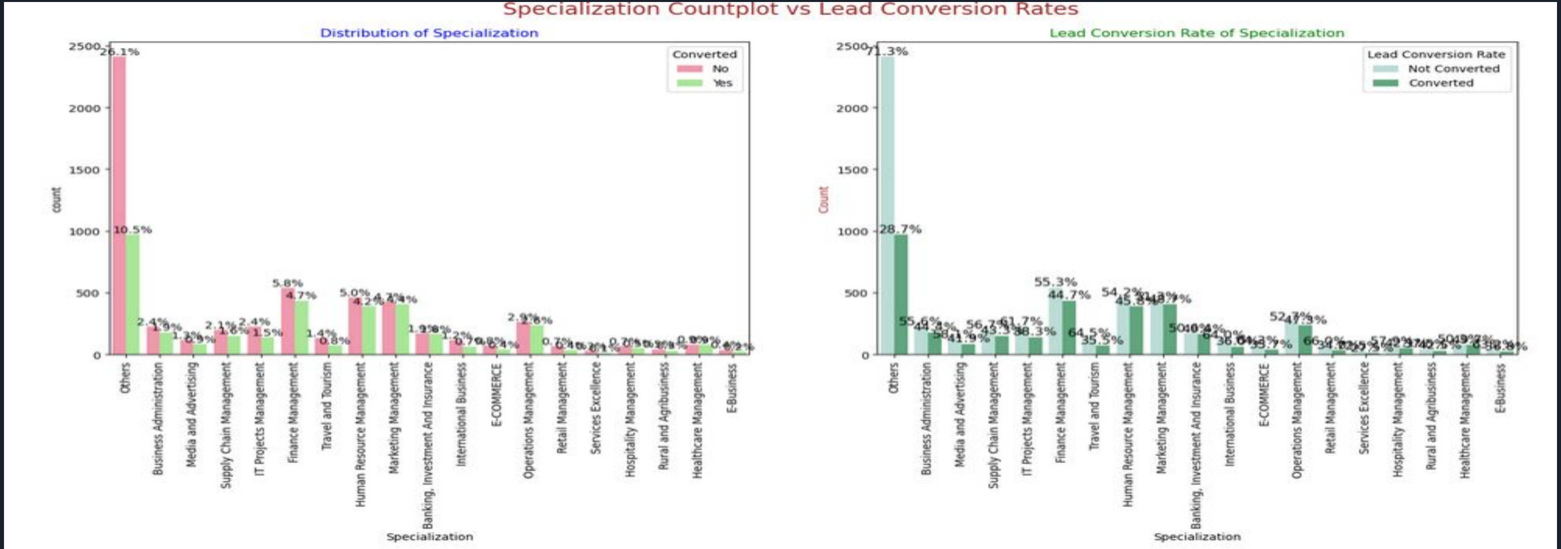
EDA- Bivariate Analysis



Last Activity Vs Conversion Rates

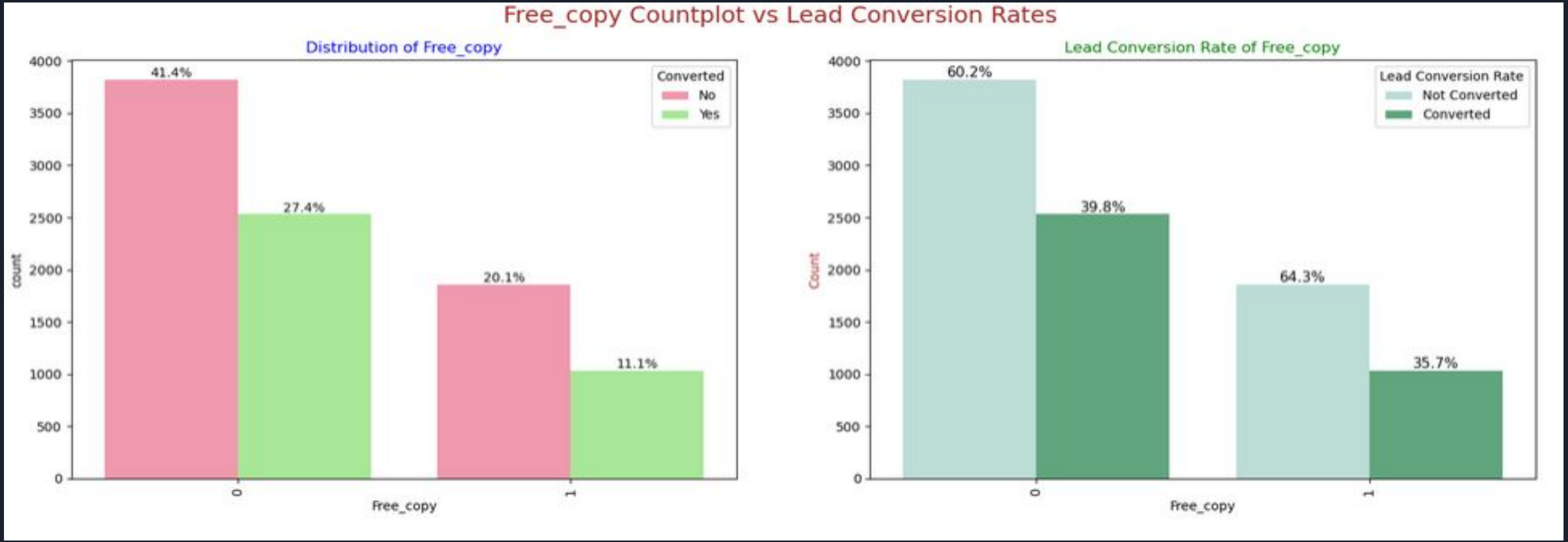
'SMS Sent has a **high lead conversion rate of 63%** with a 30% contribution from previous activities, whereas 'Email Opened' has a 38% contribution from previous activities and a 37% lead conversion rate.

EDA - Bivariate Analysis



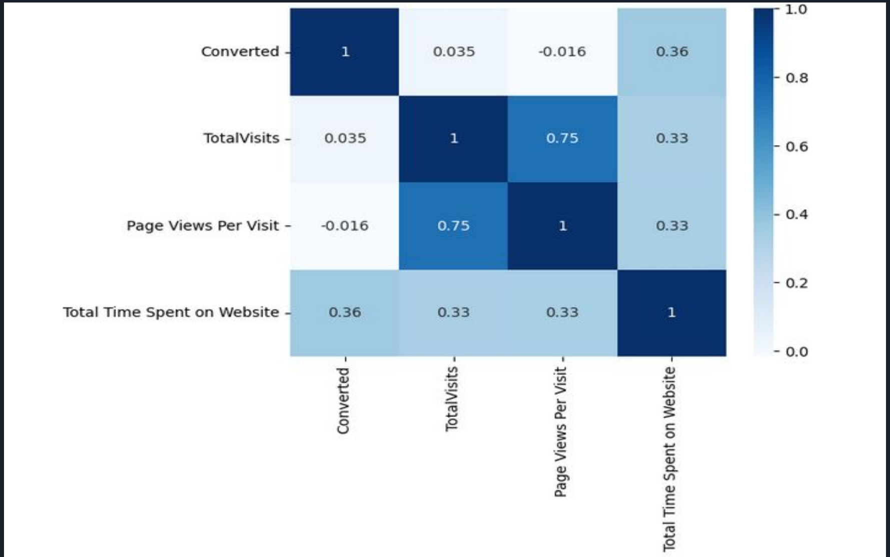
Specilization vs Lead conversion
Marketing Management, Human Resources Management, and Finance Management all make significant contributions.

EDA - Bivariate Analysis



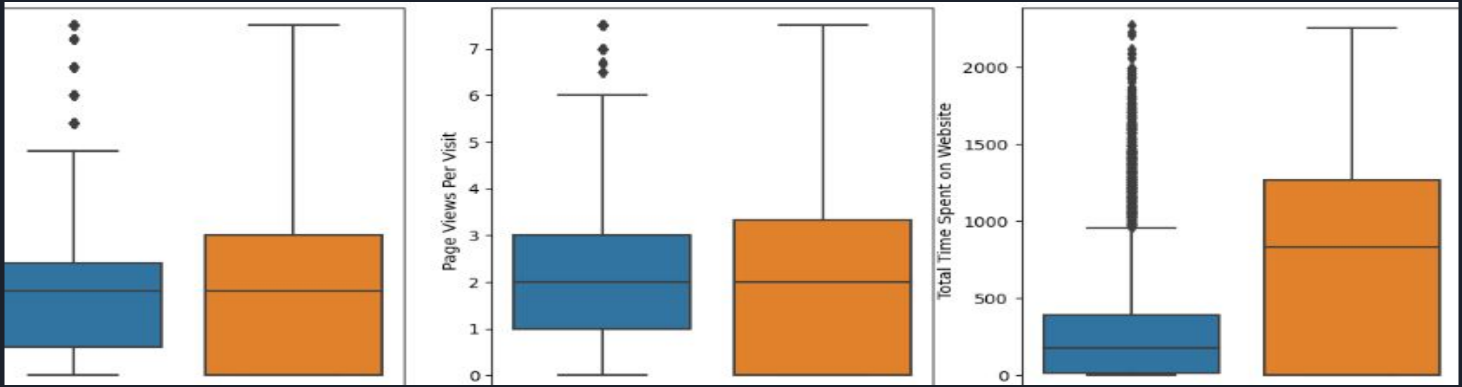
Free copy vs Lead conversion

EDA - Numerical Bivariate analysis



Heat map defines all possible conversion among the data with dark fields.

EDA - Numerical Bivariate Analysis



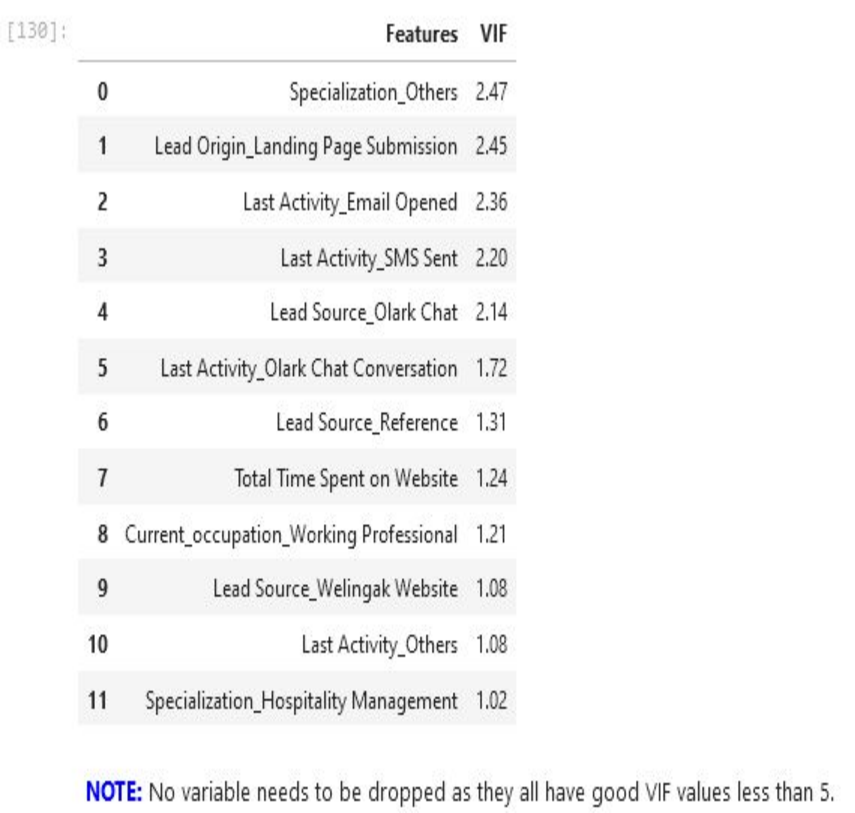
As seen in the boxplot, past leads who spend more time on the website are more likely to convert than those who spend less time.

Data preparation

- Creating dummy variables for building model for columns as
  - Lead Origin
  - Lead Source
  - Last Activity
  - Specialization
  - Current_occupation
- Generating Training and test data for model
- Featuring scaling to determining the conversion rate which is nearly 38.5%

Data model building

1. Build logistic regression model
2. Feature selection : It will be based out on the basis of minimal p values
3. Manual Fine tunning using p values & VIF <5

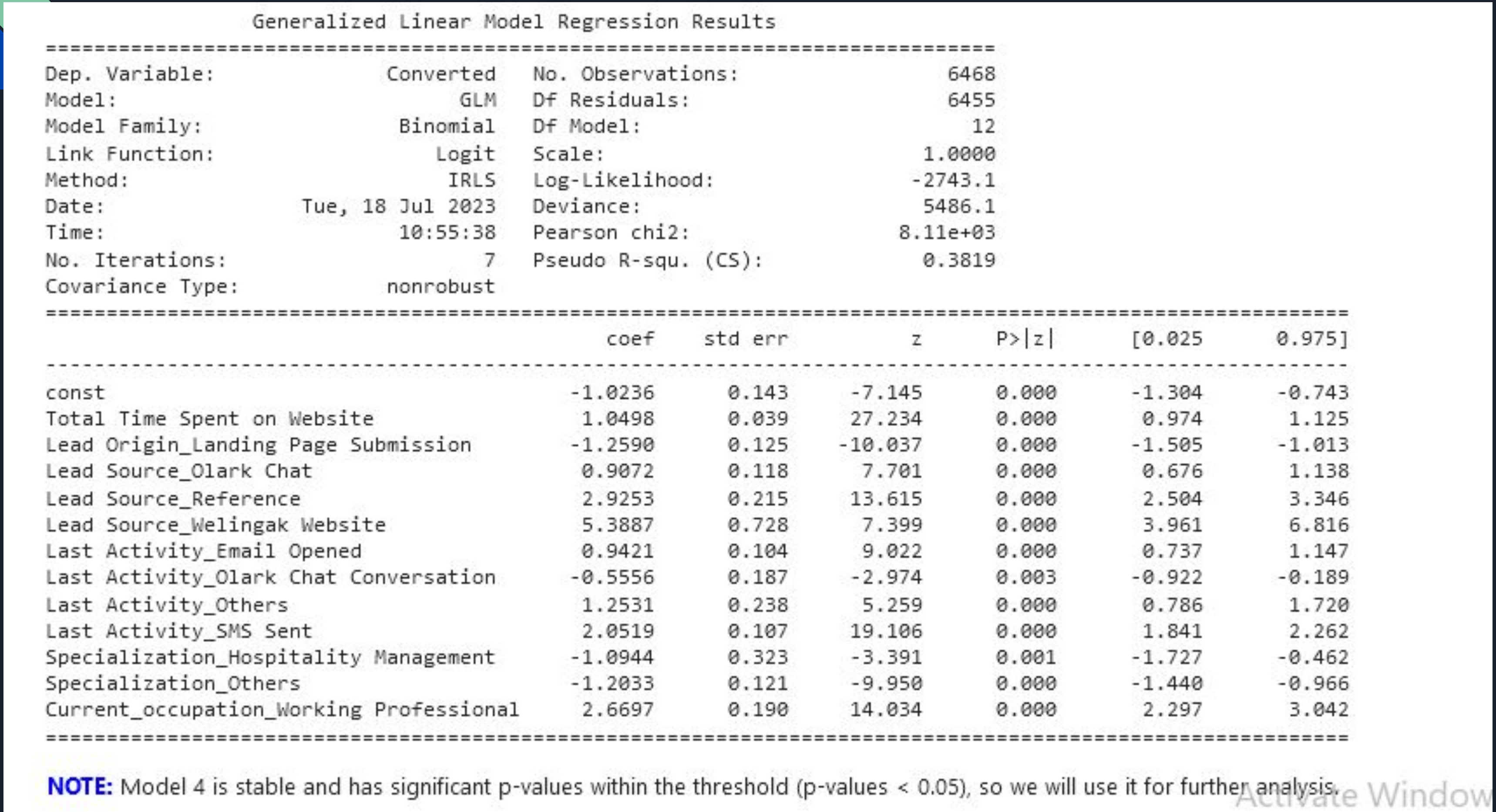


[130]:

	Features	VIF
0	Specialization_Others	2.47
1	Lead Origin_Landing Page Submission	2.45
2	Last Activity_Email Opened	2.36
3	Last Activity_SMS Sent	2.20
4	Lead Source_Olark Chat	2.14
5	Last Activity_Olark Chat Conversation	1.72
6	Lead Source_Reference	1.31
7	Total Time Spent on Website	1.24
8	Current_occupation_Working Professional	1.21
9	Lead Source_Welingak Website	1.08
10	Last Activity_Others	1.08
11	Specialization_Hospitality Management	1.02

NOTE: No variable needs to be dropped as they all have good VIF values less than 5.

Stable model with minimal p value to be used in further analysis



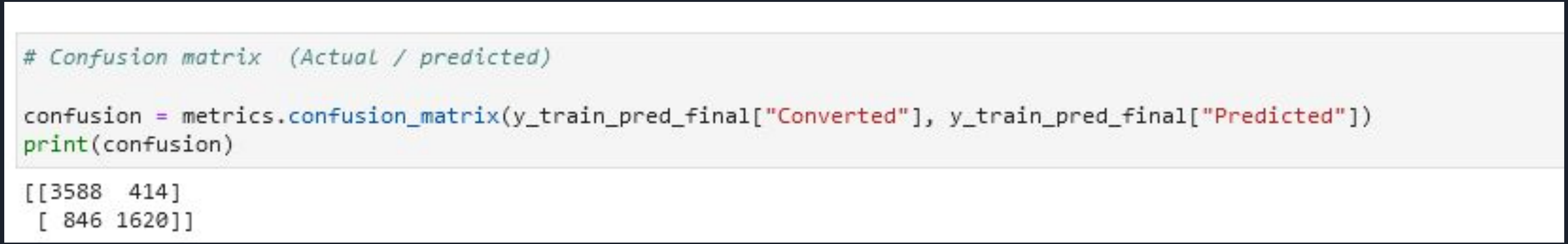
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6455
Model Family:                Binomial   Df Model:                           12
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2743.1
Date:                Tue, 18 Jul 2023   Deviance:                       5486.1
Time:                        10:55:38   Pearson chi2:                 8.11e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.3819
Covariance Type:            nonrobust
============================================================================================================
                                               coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                       -1.0236      0.143     -7.145      0.000      -1.304      -0.743
Total Time Spent on Website                  1.0498      0.039     27.234      0.000       0.974       1.125
Lead Origin_Landing Page Submission         -1.2590      0.125    -10.037      0.000      -1.505      -1.013
Lead Source_Olark Chat                       0.9072      0.118      7.701      0.000       0.676       1.138
Lead Source_Reference                        2.9253      0.215     13.615      0.000       2.504       3.346
Lead Source_Welingak Website                 5.3887      0.728      7.399      0.000       3.961       6.816
Last Activity_Email Opened                   0.9421      0.104      9.022      0.000       0.737       1.147
Last Activity_Olark Chat Conversation       -0.5556      0.187     -2.974      0.003      -0.922      -0.189
Last Activity_Others                         1.2531      0.238      5.259      0.000       0.786       1.720
Last Activity_SMS Sent                       2.0519      0.107     19.106      0.000       1.841       2.262
Specialization_Hospitality Management       -1.0944      0.323     -3.391      0.001      -1.727      -0.462
Specialization_Others                       -1.2033      0.121     -9.950      0.000      -1.440      -0.966
Current_occupation_Working Professional      2.6697      0.190     14.034      0.000       2.297       3.042
============================================================================================================
```

NOTE: Model 4 is stable and has significant p-values within the threshold (p-values < 0.05), so we will use it for further analysis.

Model Evaluation

- Confusion Matrix
- Accuracy
- Sensitivity and Specificity
- Threshold determination using ROC & Finding Optimal cutoff point
- Precision and Recall

Model Evaluation - Confusion matrix, Accuracy



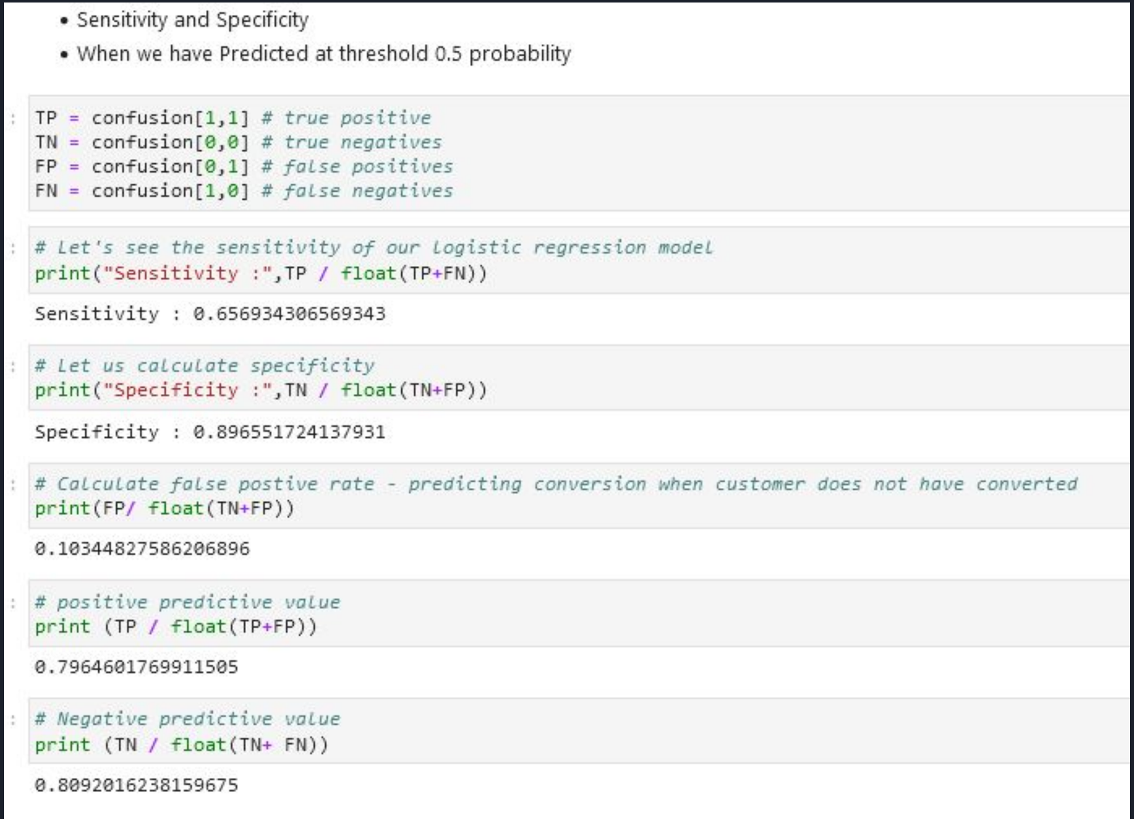
```
# Confusion matrix  (Actual / predicted)

confusion = metrics.confusion_matrix(y_train_pred_final["Converted"], y_train_pred_final["Predicted"])
print(confusion)
```

```
[[3588  414]
 [ 846 1620]]
```

Confusion matrix build using the probability value as 0.5 being calculated earlier.

The accurate conversion that could be made would be around 0.805

Model evaluation: Simplicity & specificity



- Sensitivity and Specificity
- When we have Predicted at threshold 0.5 probability

```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
# Let's see the sensitivity of our logistic regression model
print("Sensitivity :",TP / float(TP+FN))
```

```
Sensitivity : 0.656934306569343
```

```
# Let us calculate specificity
print("Specificity :",TN / float(TN+FP))
```

```
Specificity : 0.896551724137931
```

```
# Calculate false postive rate - predicting conversion when customer does not have converted
print(FP/ float(TN+FP))
```

```
0.10344827586206896
```

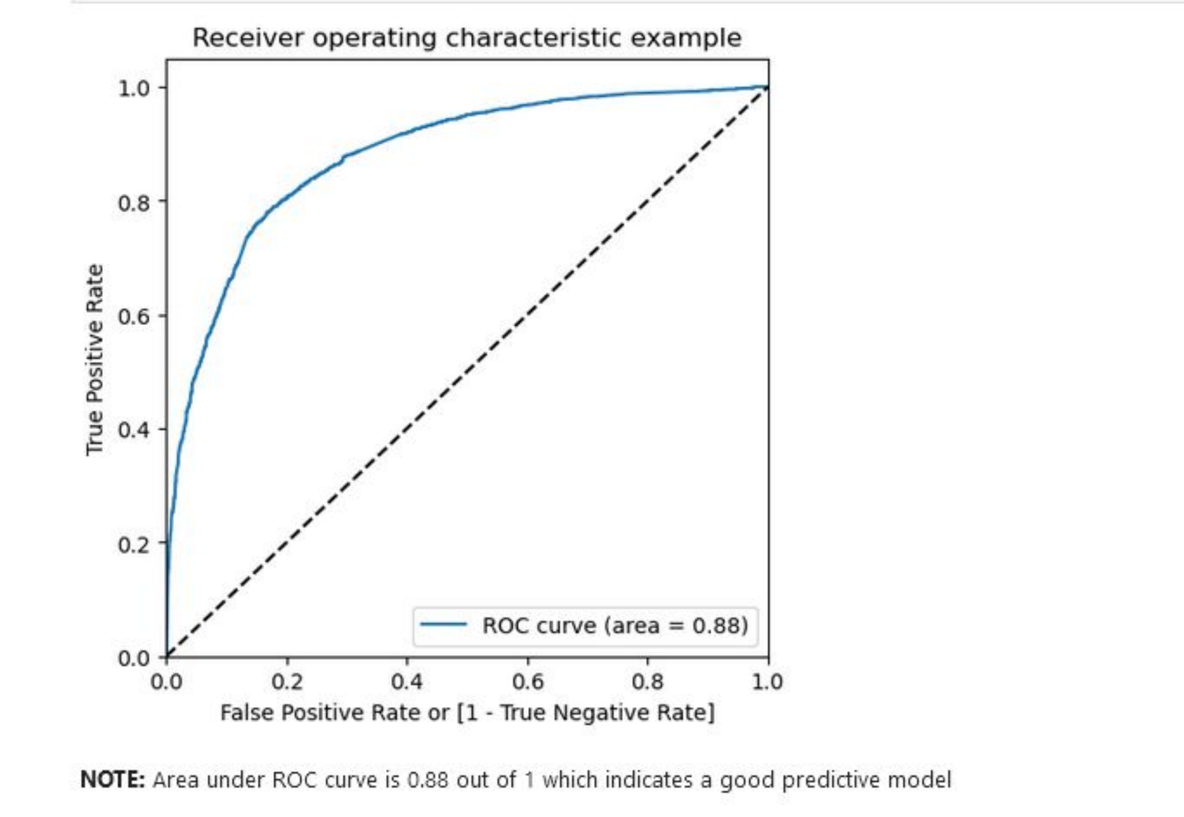
```
# positive predictive value
print (TP / float(TP+FP))
```

```
0.7964601769911505
```

```
# Negative predictive value
print (TN / float(TN+ FN))
```

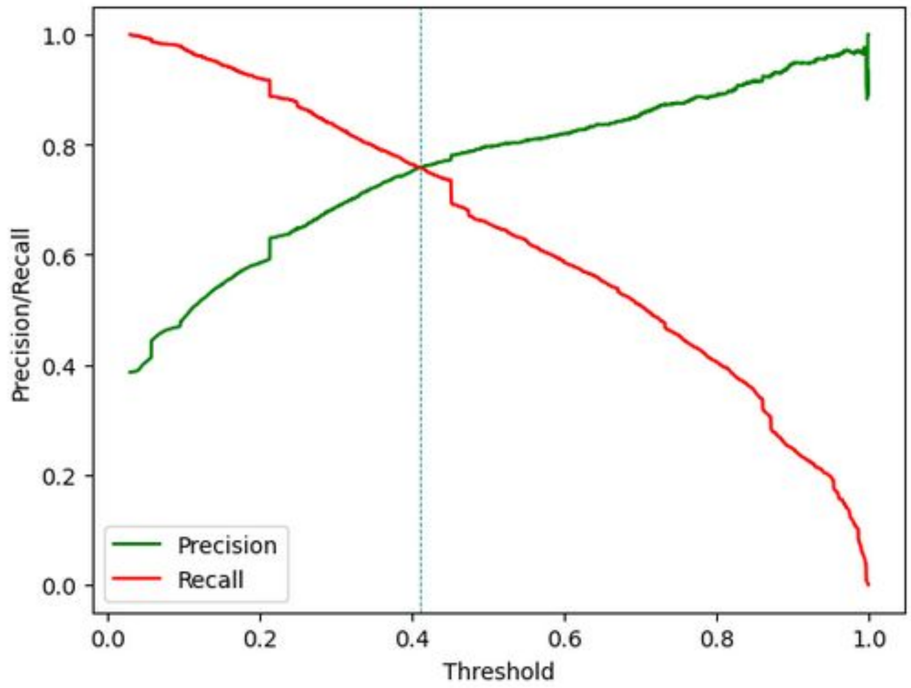
```
0.8092016238159675
```

Model evaluation: Roc Curve for train data



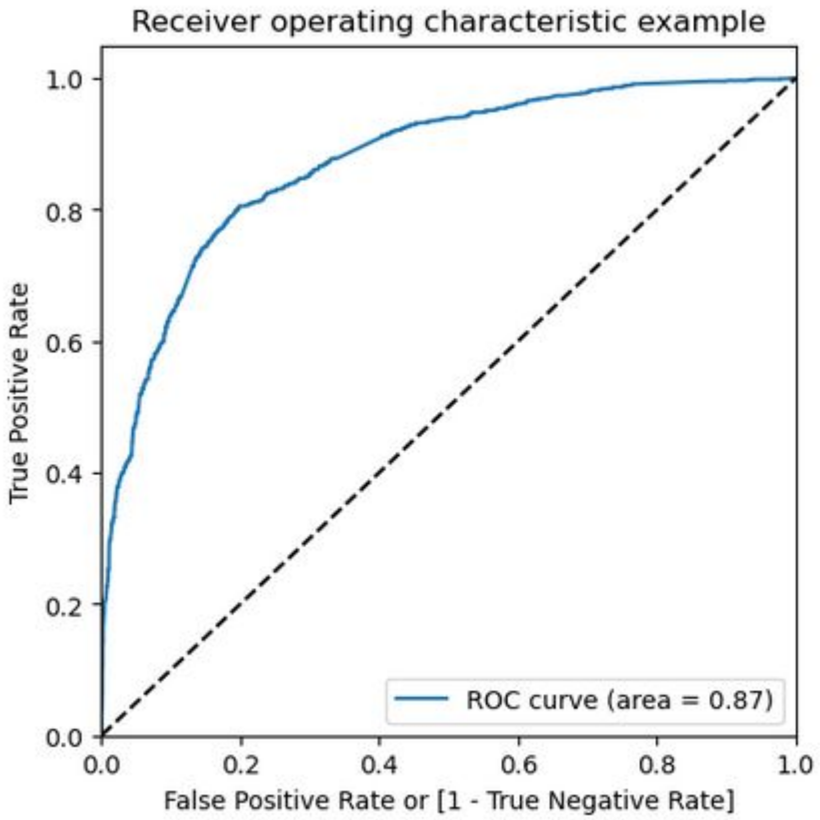
NOTE: Area under ROC curve is 0.88 out of 1 which indicates a good predictive model

Model Evaluation: Precision & Recall



NOTE: The point where the intersection of the curve represents the value at which the model reaches a balance of precision and recall. It can be used to optimize the model's performance based on business requirements. Our probability threshold is approximately 0.41 from the above curve.

Test data model prediction



NOTE: Area under ROC curve is 0.87 out of 1 which indicates a good predictive model

Recommendations

- For focused marketing efforts, prioritize traits with positive coefficients.

- Create methods for attracting high-quality leads from high-performance lead sources.

- Use personalized messages to engage working professionals.

- Optimize communication channels based on the impact of lead engagement.

- More money can be spent on advertising on the Welingak website.

- Incentives/discounts for delivering referrals that convert to leads, to encourage more referrals.

- Working professionals should be aggressively targeted because they have a high conversion rate and are in a better financial position to pay greater fees.