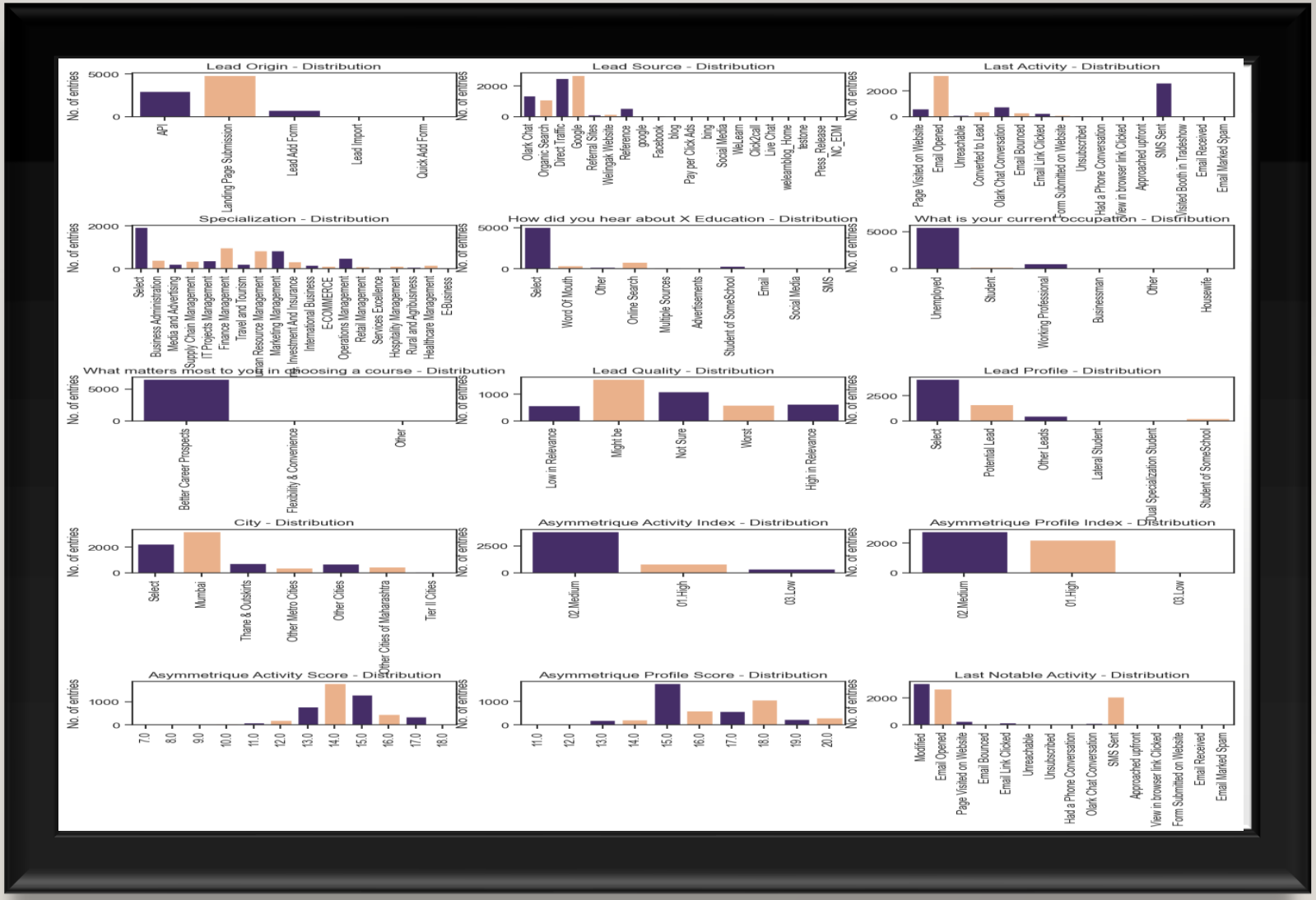
Lead Scoring Case Study

Steps Involved to Build the Model :

- ☐ Cleaning the Data
- ☐ Preparation of Data
- ☐ Exploratory Data Analysis
- ☐ Building the Model
- ☐ Evaluating the Model

Exploratory Data Analysis

UNIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES



UNIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

We can infer from the previous Univariate Analysis of Categorical Variables

1. Some of the fields contain the value "Select“

2. This drop-down value was largely left in its current state, which is equivalent to data not available.

3. For all practical purposes, we can treat it as NaN.

BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

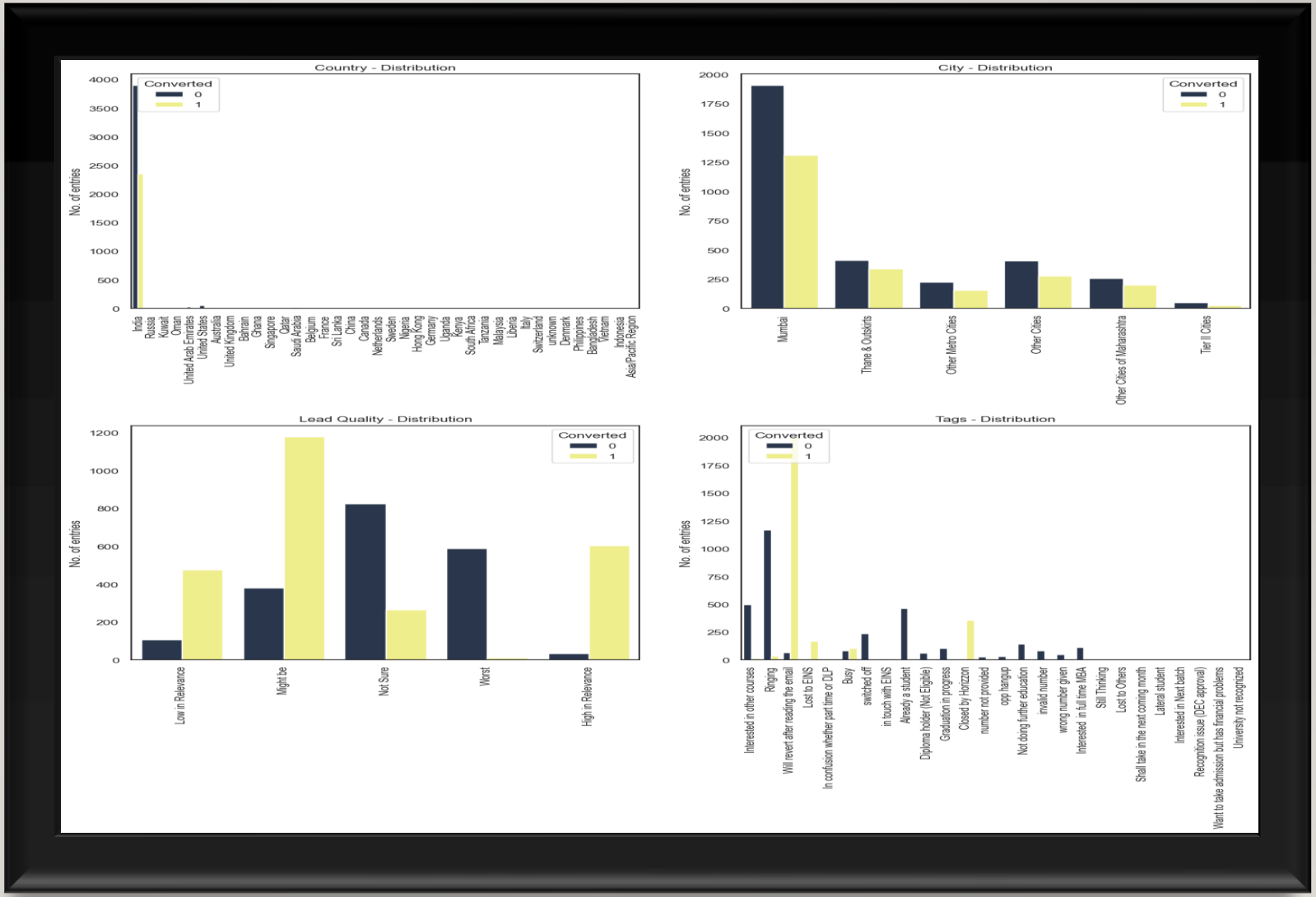
We can infer from the previous Bivariate Analysis of Categorical Variables

1. We plot data for target in terms of the total count

2. Online search and Student of some school show max values in "How did you hear about X education" distribution

3. Only potential lead shows max values in "Lead Profile " distribution.

DISPLAYING THE DISTRIBUTION OF ALL CATEGORY VARIABLES IN HISTOGRAMS



DISPLAYING THE DISTRIBUTION OF ALL CATEGORY VARIABLES IN HISTOGRAMS

Locations include a country and city.

1. Of the 39 country values, 23% are null.

2. Eight different city values, of which 35% are zero

3. It is evident from the count plots above that neither the country nor the city significantly affect conversion rates.

DISPLAYING THE DISTRIBUTION OF ALL CATEGORICAL VARIABLES IN HISTOGRAMS

DISPLAYING THE DISTRIBUTION OF ALL CATEGORICAL VARIABLES IN HISTOGRAMS

1. What's most important to you in a course is: "Better Career Prospects" is the value that most people choose for this column.

2. As 99.99% of the non-null values in the accessible data set have the same value of "Better Career Prospects," this column does not significantly add value.

Displaying the distribution of categorical variables' histograms

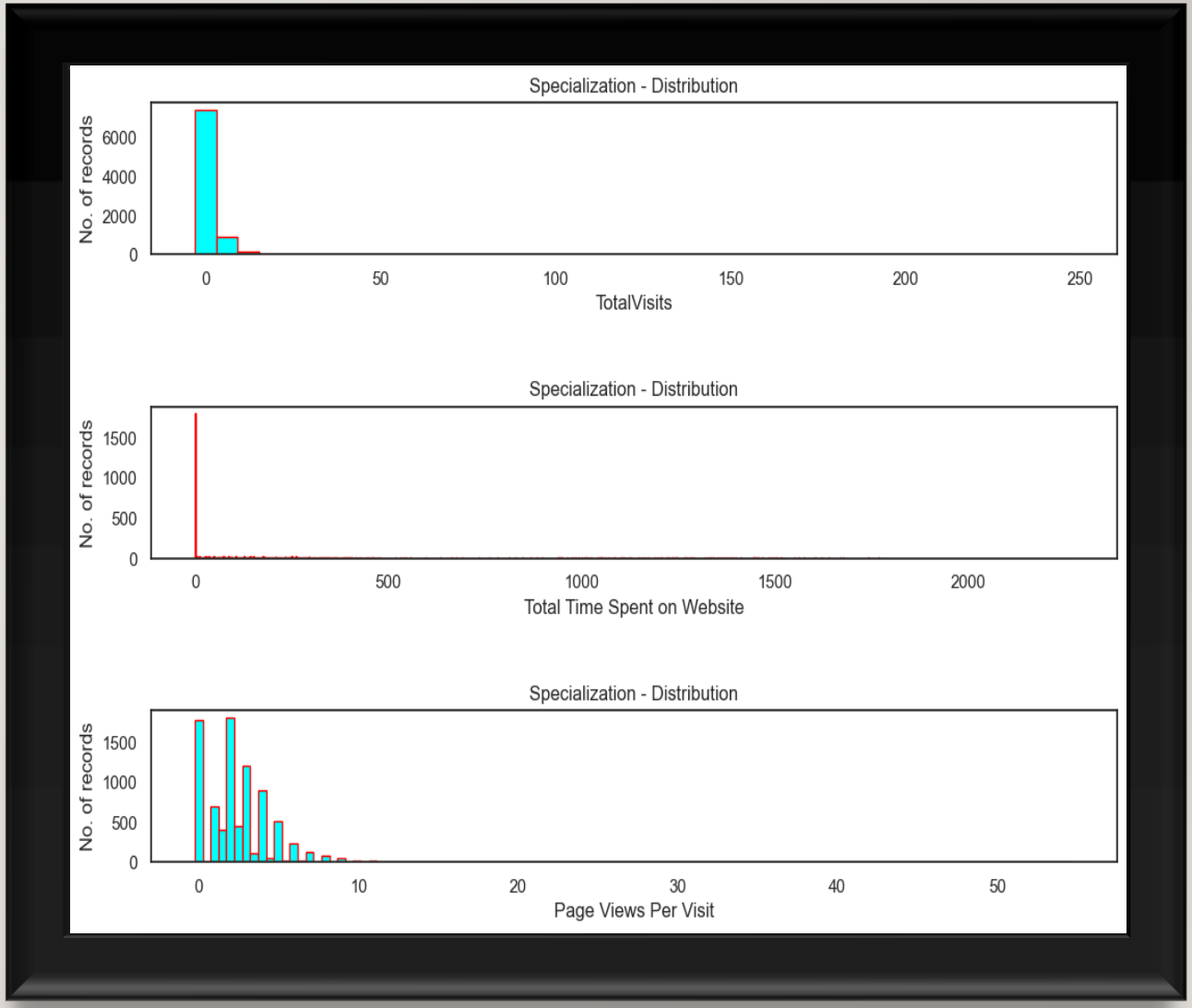
DISPLAYING THE DISTRIBUTION OF CATEGORICAL VARIABLES' HISTOGRAMS

It is evident from the above graph that a small number of specific values for "Tags," "Lead Profile," and "Lead Quality" have a significant impact on the "Converted" result.

Input of these fields will therefore introduce bias. For now, let's refer to them as "Unknown".

The Dummy Variables phase is where they can be handled.

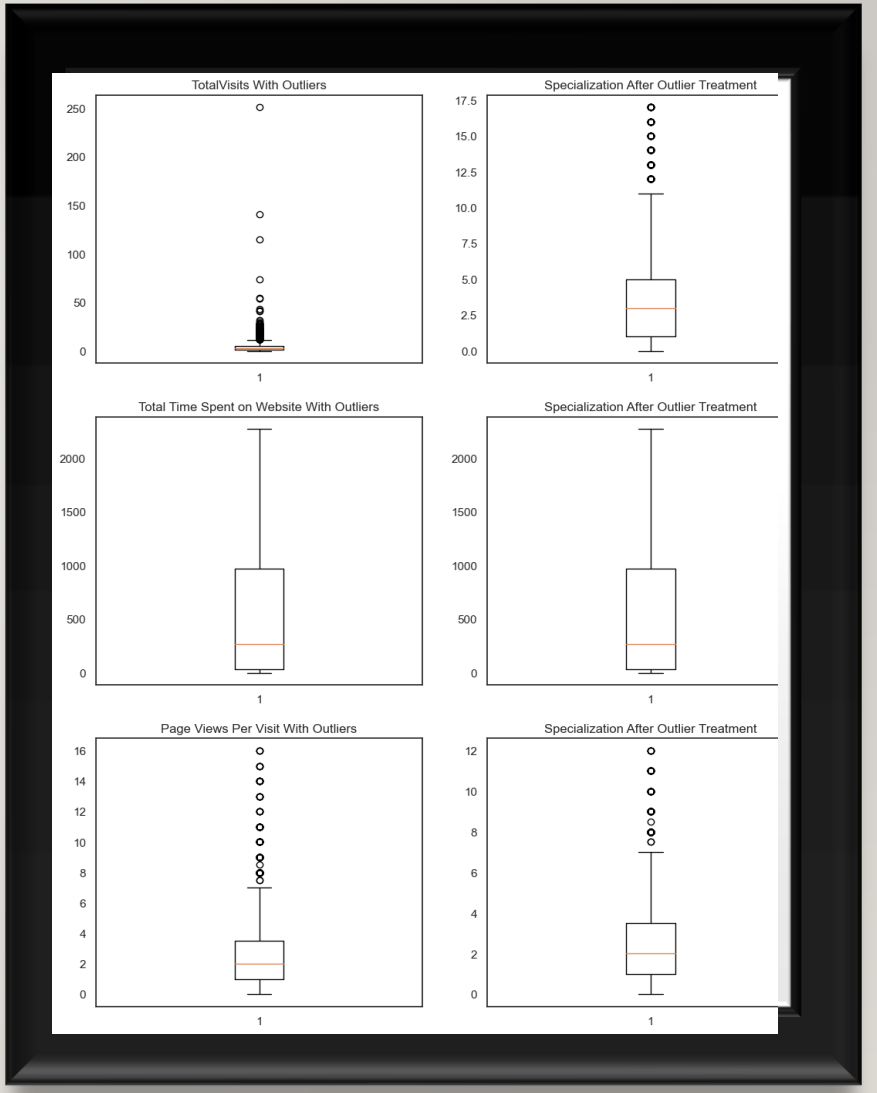
VISUALIZING THE HISTOGRAM OF THE DISTRIBUTION OF ALL NUMERIC VARIABLES



VISUALIZING THE HISTOGRAM OF THE DISTRIBUTION OF ALL NUMERIC VARIABLES

The previous graph indicates possible outliers in the data.

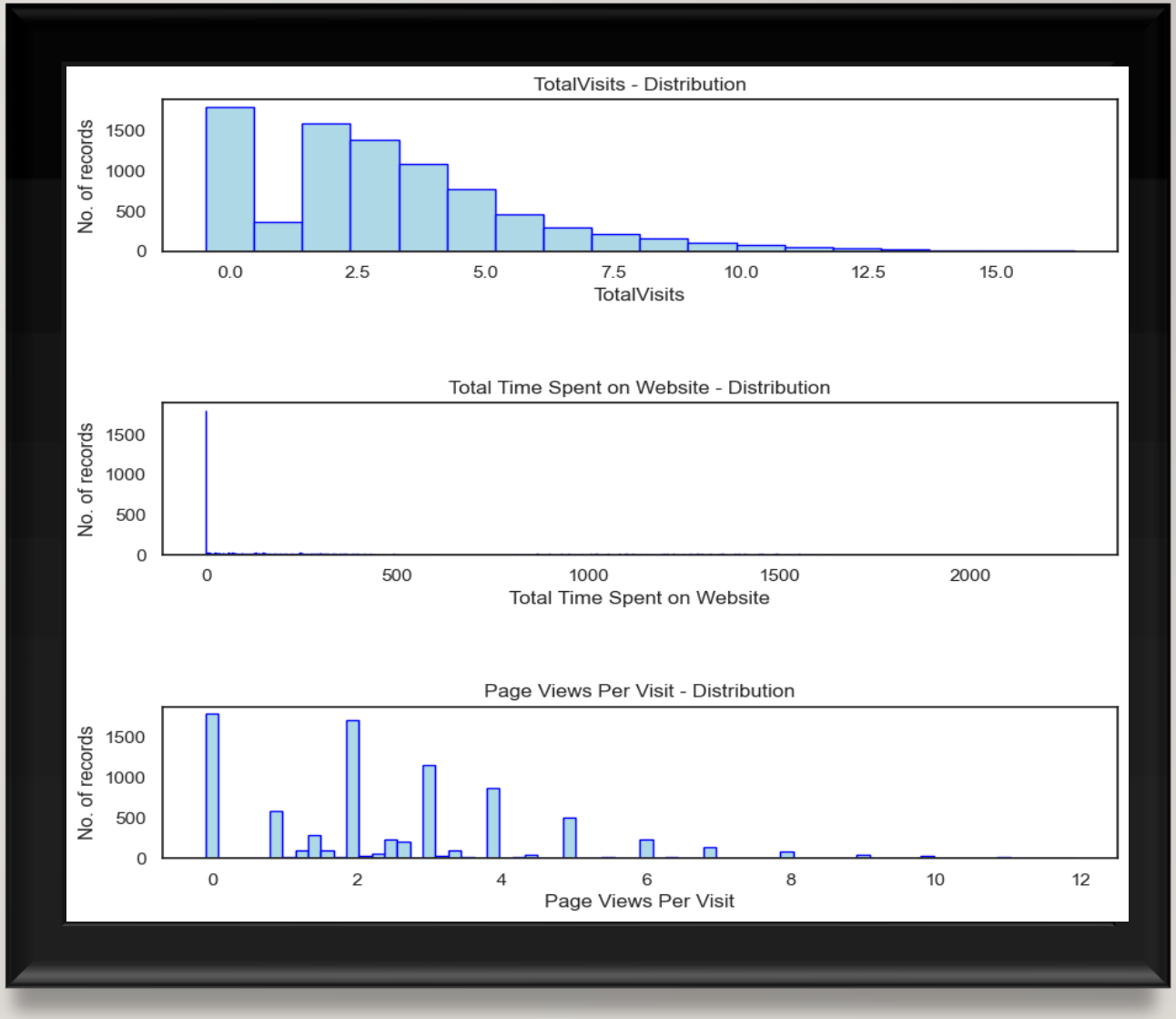
BOX PLOT OF ALL THE VARIABLES BEFORE AND AFTER OUTLIER TREATMENT



Visualizing the histogram of the distribution of all numeric variables after outlier treatment

The box plots are plotted before the outlier treatment.

The Histograms are plotted after the outlier treatment.



Heat Map

- **TotalVisits and Page Views Per Visit have high correlation of 0.72**

TRAIN DATA MATRIX

```
:  # Confusion Matrix
   confusion = metrics.confusion_matrix(y_train_pred_final.Converted,y_train_pred_final.predicted)
   print(confusion)
```

```
[[3446   59]
 [ 586 1840]]
```

TEST DATA MATRIX

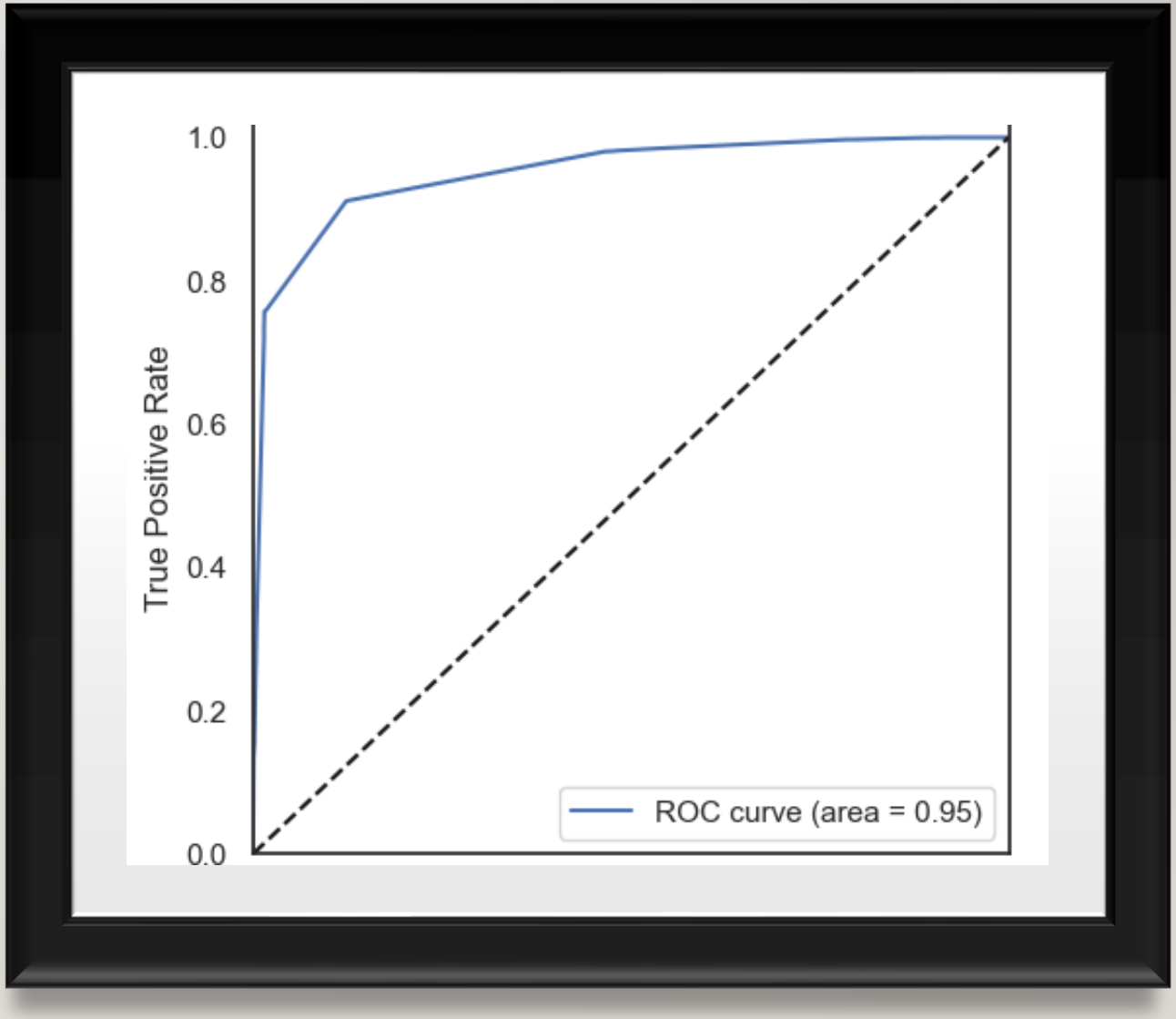
Confusion Matrix for Test Data Creation

```
confusion_test = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.final_predicted )
confusion_test
```

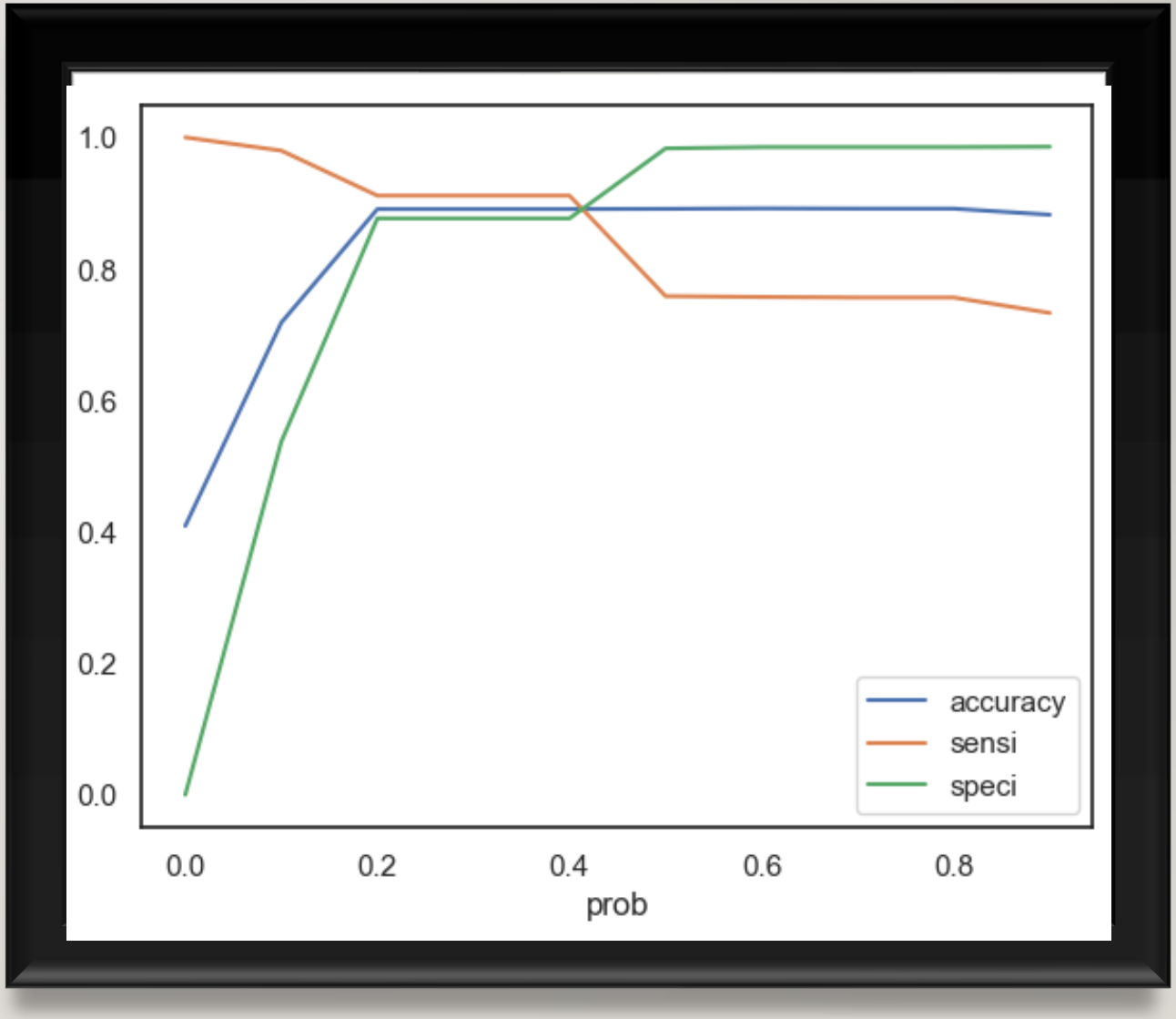
```
array([[1347,  189],
       [  85,  922]], dtype=int64)
```

Roc Curve

- The Area under curve is 0.95



- The different cut-offs are displayed, and 0.4 is the point at which all values converge.



Now that the Logistic Regression model has been developed with all near-ZERO p-values and low VIF (multicollinearity),

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9079	0.079	-24.092	0.000	-2.063	-1.753
Lead Source_Welingak Website	4.3909	1.009	4.351	0.000	2.413	6.369
What is your current occupation_Unemployed	1.8385	0.105	17.524	0.000	1.633	2.044
What is your current occupation_Working Professional	1.9865	0.315	6.297	0.000	1.368	2.605
Tags_Already a student	-4.0378	0.717	-5.635	0.000	-5.442	-2.633
Tags_Closed by Horizzon	5.7010	1.015	5.618	0.000	3.712	7.690
Tags_Diploma holder (Not Eligible)	-3.3832	1.018	-3.323	0.001	-5.379	-1.388
Tags_Interested in full time MBA	-2.8539	0.597	-4.780	0.000	-4.024	-1.684
Tags_Interested in other courses	-3.0180	0.303	-9.946	0.000	-3.613	-2.423
Tags_Lost to EINS	4.3704	0.538	8.127	0.000	3.316	5.424
Tags_Not doing further education	-3.6396	1.014	-3.588	0.000	-5.628	-1.651
Tags_Ringing	-3.3131	0.208	-15.898	0.000	-3.722	-2.905
Tags_Will revert after reading the email	3.6220	0.172	21.000	0.000	3.284	3.960
Tags_switched off	-3.9105	0.587	-6.661	0.000	-5.061	-2.760
Lead Quality_Worst	-3.5416	0.721	-4.909	0.000	-4.956	-2.128

FINAL MODEL

```
# SENSITIVITY
print("Sensitivity of the Test Predictions:",round(100*(TP_test/float(FN_test+TP_test)),2),"%")
```

```
Sensitivity of the Test Predictions: 91.56 %
```

```
# SPECIFICITY
print("Specificity of the Test Predictions:",round(100*(TN_test/float(TN_test+FP_test)),2),"%")
```

```
Specificity of the Test Predictions: 87.7 %
```

```
# ACCURACY SCORE
print("ACCURACY SCORE of the Test Data Predictions:",round(100*((TP_test+TN_test)/(TP_test+TN_test+FP_test+FN_test)),2),"%")
```

```
ACCURACY SCORE of the Test Data Predictions: 89.23 %
```

Precision and Recall

```
print("Precision Score of the Test Data Predictions:",round(100*(precision_score(y_pred_final.Converted, y_pred_final.final_predicted)),2),"%")
```

```
Precision Score of the Test Data Predictions: 82.99 %
```

```
print("Recall Score of the Test Data Predictions:",round(100*(recall_score(y_pred_final.Converted, y_pred_final.final_predicted)),2),"%")
```

```
Recall Score of the Test Data Predictions: 91.56 %
```

```
# Let's check the overall accuracy.
print("Accuracy of the Test Data Predictions:",round(100*(metrics.accuracy_score(y_pred_final.Converted,y_pred_final.final_predicted)),2),"%")
```

```
Accuracy of the Test Data Predictions: 89.23 %
```