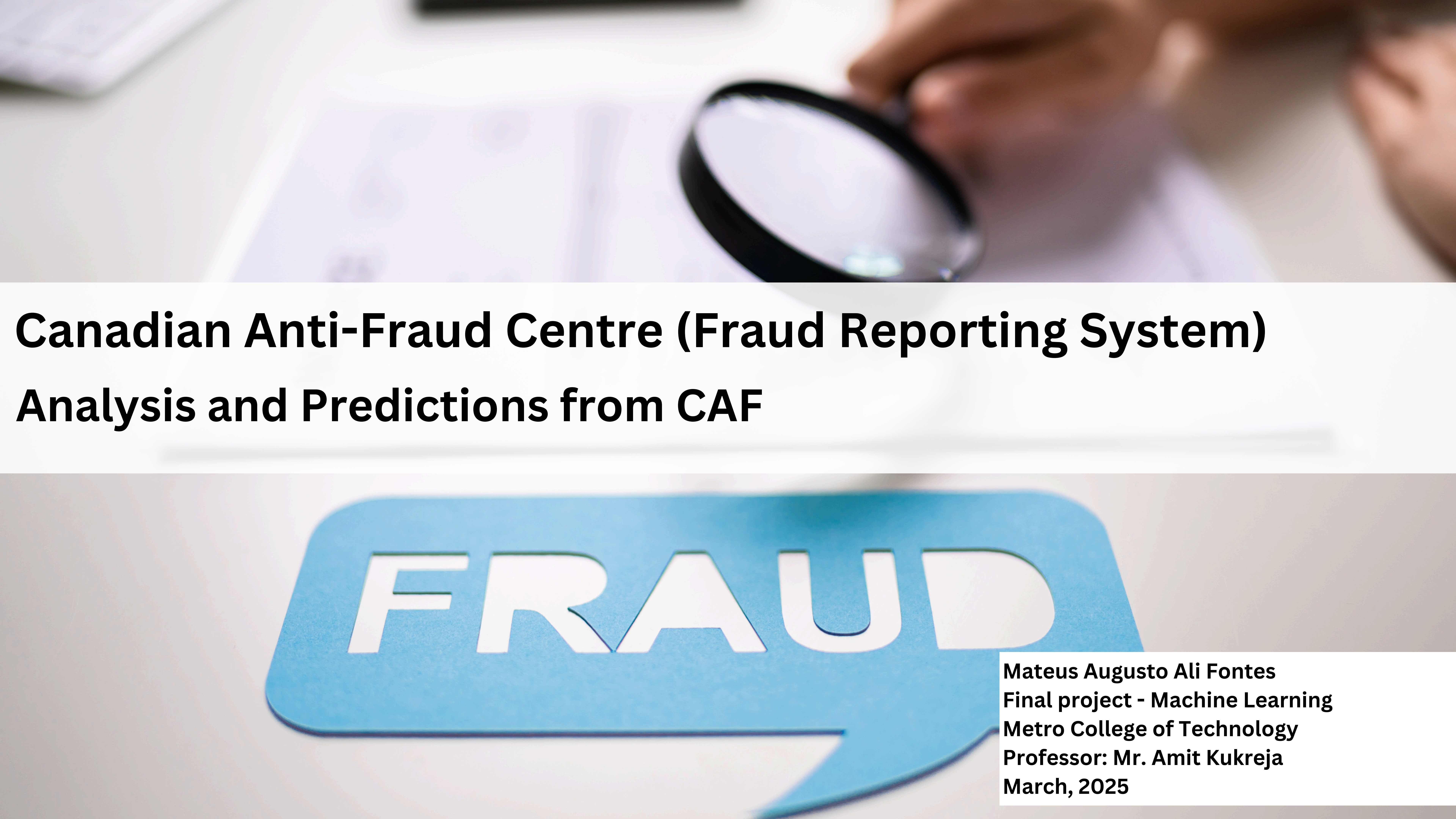


# Canadian Anti-Fraud Centre (Fraud Reporting System)

## Analysis and Predictions from CAF

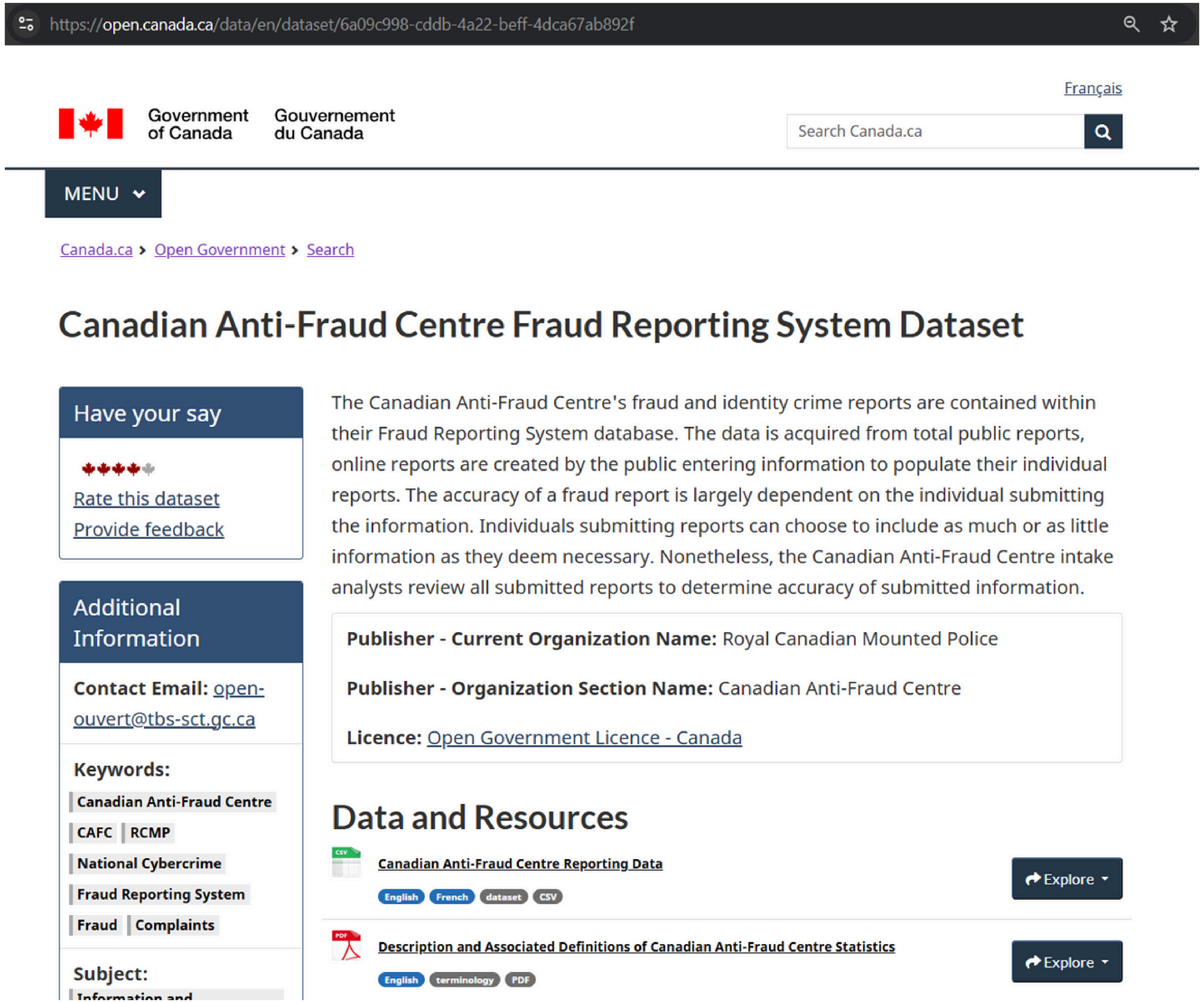
Mateus Augusto Ali Fontes
Final project - Machine Learning
Metro College of Technology
Professor: Mr. Amit Kukreja
March, 2025

# Objective & Overview

- **Content:**

    - **Dataset overview**
    - **Univariate analysis (Numeric variables)**
    - **Univariate analysis (Categoric variables)**
    - **Bivariate analysis**
    - **Performance of your classification models.**
    - **Findings presentation from the analysis and modeling.**

# Dataset overview - CAF



https://open.canada.ca/data/en/dataset/6a09c998-cddb-4a22-beff-4dca67ab892f
Français
Government of Canada
Gouvernement du Canada
Search Canada.ca
MENU

Canada.ca > Open Government > Search

## Canadian Anti-Fraud Centre Fraud Reporting System Dataset

**Have your say**

Rate this dataset
Provide feedback

**Additional Information**

**Contact Email:** open-ouvert@tbs-sct.gc.ca

**Keywords:**
Canadian Anti-Fraud Centre | CAFC | RCMP | National Cybercrime | Fraud Reporting System | Fraud | Complaints

**Subject:**
Information and

The Canadian Anti-Fraud Centre's fraud and identity crime reports are contained within their Fraud Reporting System database. The data is acquired from total public reports, online reports are created by the public entering information to populate their individual reports. The accuracy of a fraud report is largely dependent on the individual submitting the information. Individuals submitting reports can choose to include as much or as little information as they deem necessary. Nonetheless, the Canadian Anti-Fraud Centre intake analysts review all submitted reports to determine accuracy of submitted information.

**Publisher - Current Organization Name:** Royal Canadian Mounted Police

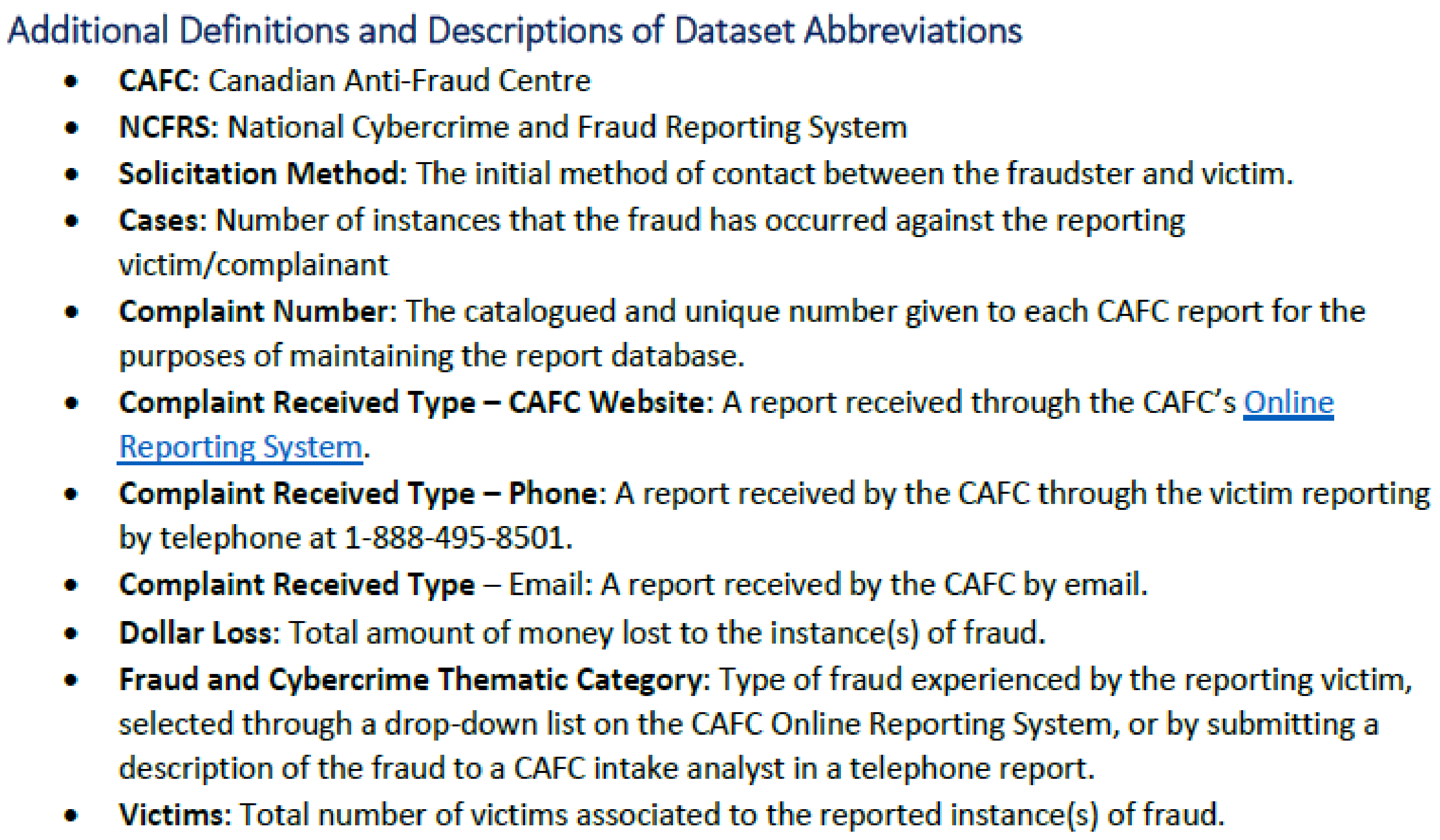
**Publisher - Organization Section Name:** Canadian Anti-Fraud Centre

**Licence:** Open Government Licence - Canada

## Data and Resources

Canadian Anti-Fraud Centre Reporting Data
English | French | dataset | CSV
Explore

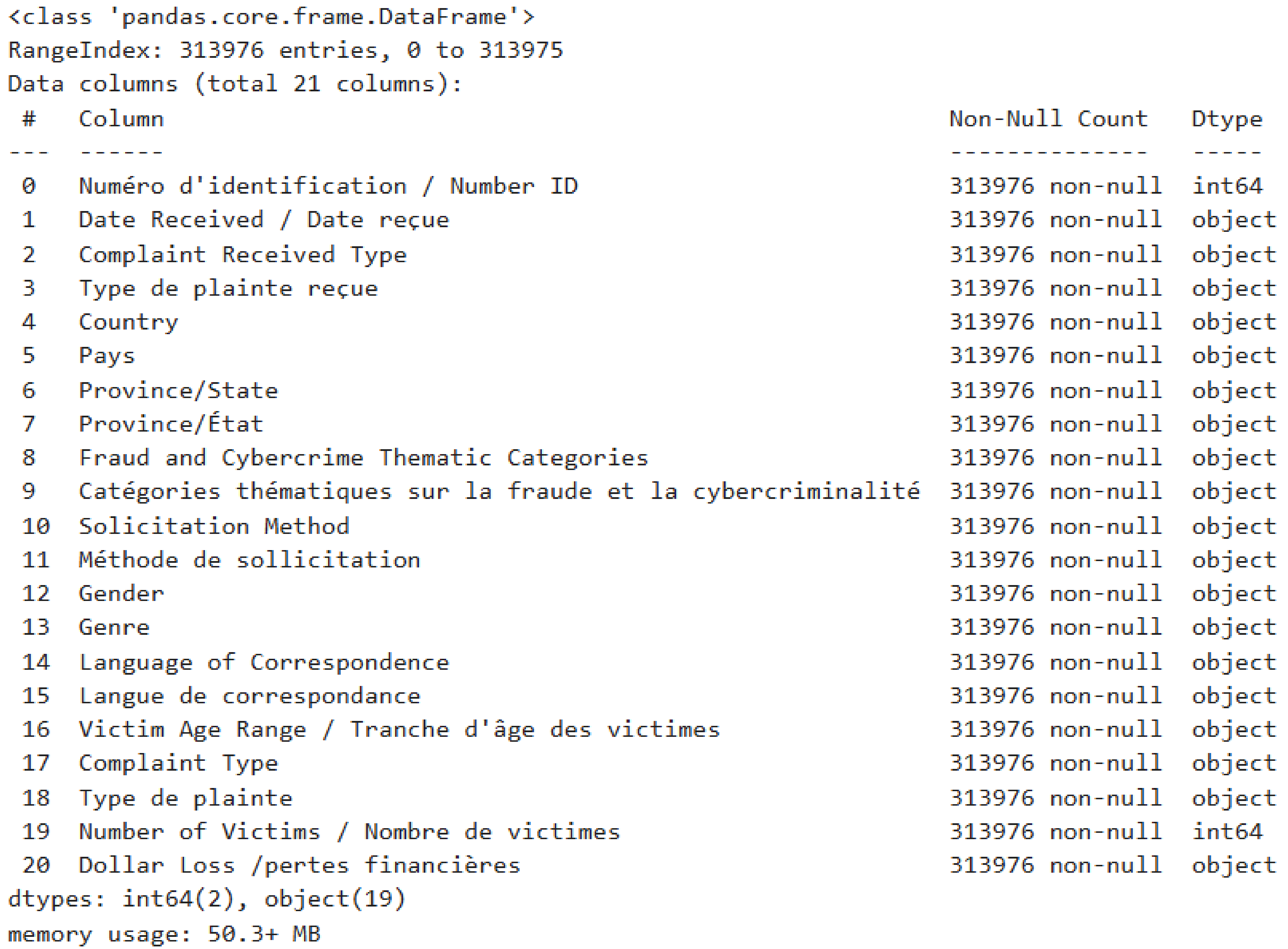
Description and Associated Definitions of Canadian Anti-Fraud Centre Statistics
English | terminology | PDF
Explore



### Additional Definitions and Descriptions of Dataset Abbreviations

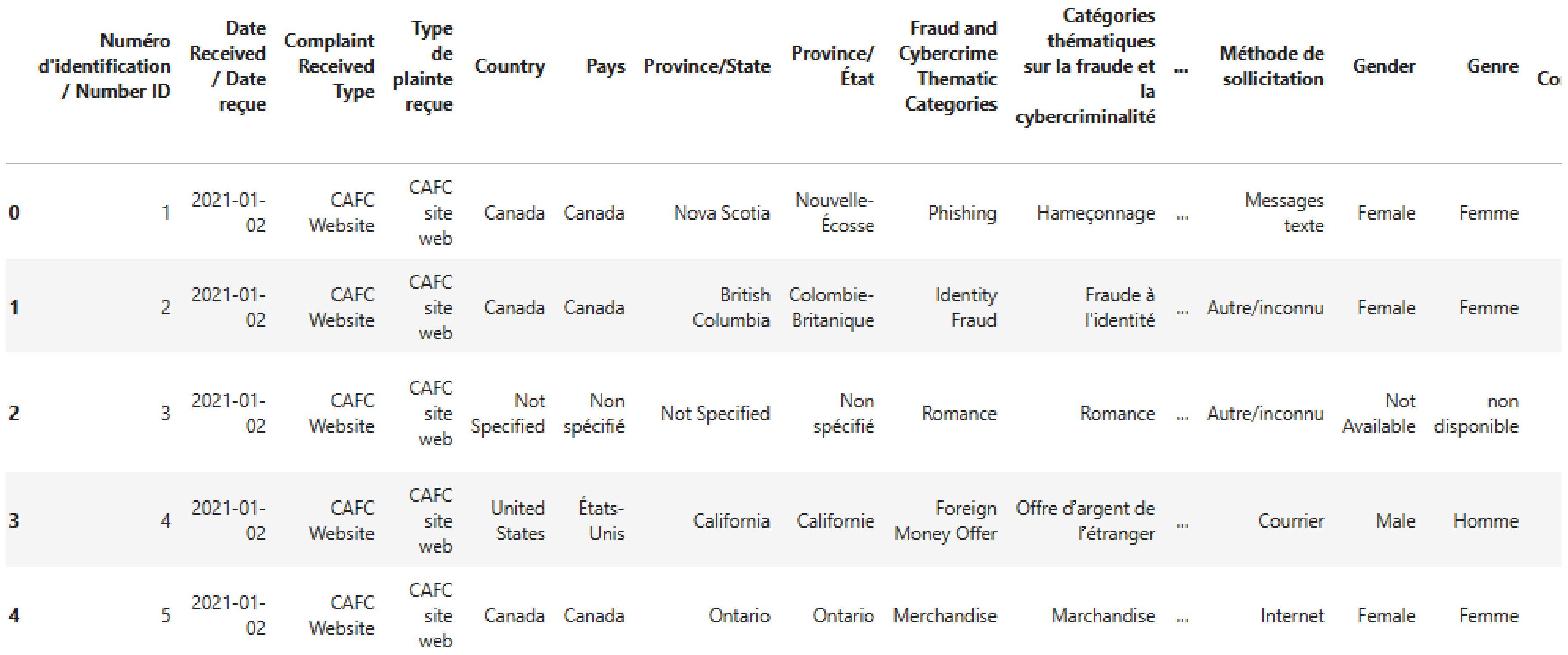
- **CAFC**: Canadian Anti-Fraud Centre
- **NCFRS**: National Cybercrime and Fraud Reporting System
- **Solicitation Method**: The initial method of contact between the fraudster and victim.
- **Cases**: Number of instances that the fraud has occurred against the reporting victim/complainant
- **Complaint Number**: The catalogued and unique number given to each CAFC report for the purposes of maintaining the report database.
- **Complaint Received Type – CAFC Website**: A report received through the CAFC's Online Reporting System.
- **Complaint Received Type – Phone**: A report received by the CAFC through the victim reporting by telephone at 1-888-495-8501.
- **Complaint Received Type** – Email: A report received by the CAFC by email.
- **Dollar Loss**: Total amount of money lost to the instance(s) of fraud.
- **Fraud and Cybercrime Thematic Category**: Type of fraud experienced by the reporting victim, selected through a drop-down list on the CAFC Online Reporting System, or by submitting a description of the fraud to a CAFC intake analyst in a telephone report.
- **Victims**: Total number of victims associated to the reported instance(s) of fraud.

# Dataset overview - CAF



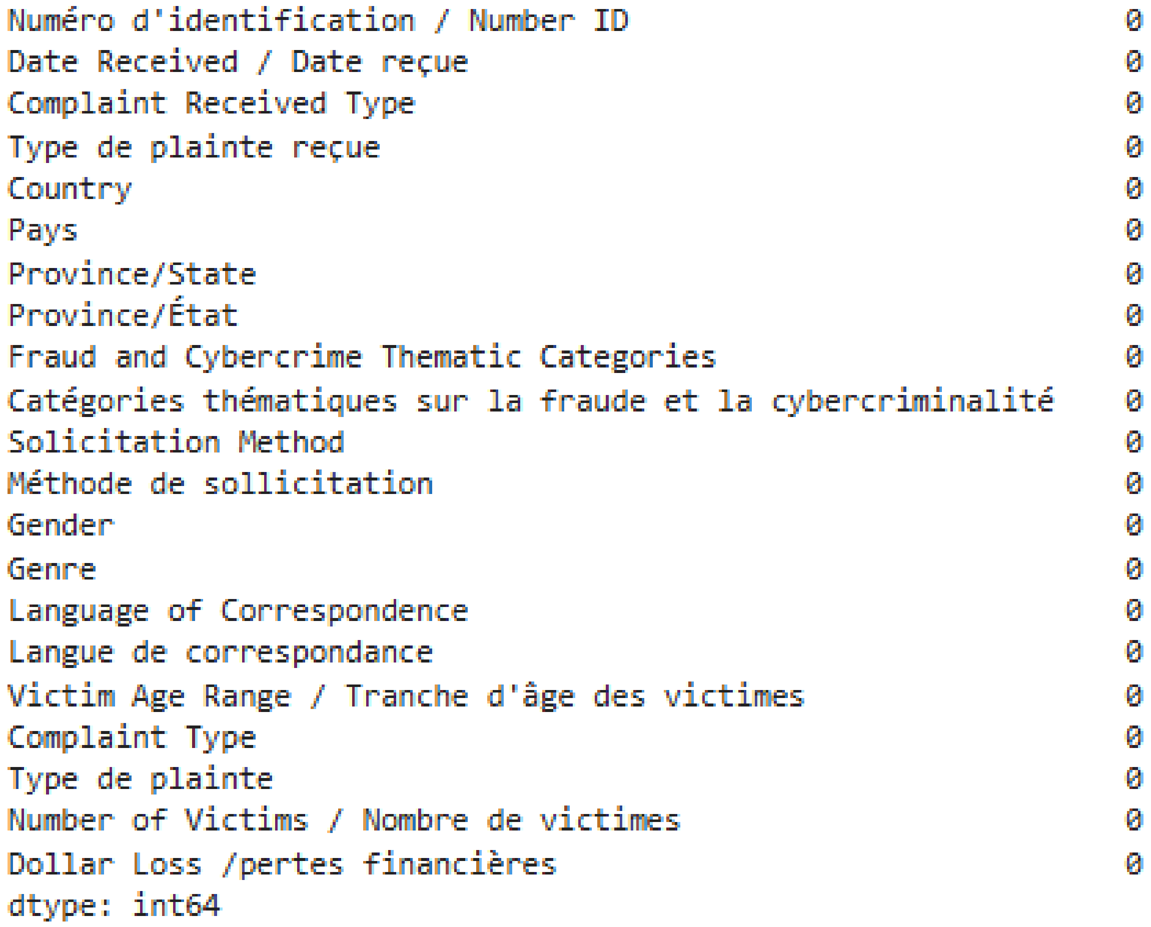
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 313976 entries, 0 to 313975
Data columns (total 21 columns):
 #   Column                                                       Non-Null Count   Dtype
---  ------                                                       --------------   -----
 0   Numéro d'identification / Number ID                          313976 non-null  int64
 1   Date Received / Date reçue                                   313976 non-null  object
 2   Complaint Received Type                                      313976 non-null  object
 3   Type de plainte reçue                                        313976 non-null  object
 4   Country                                                      313976 non-null  object
 5   Pays                                                         313976 non-null  object
 6   Province/State                                               313976 non-null  object
 7   Province/État                                                313976 non-null  object
 8   Fraud and Cybercrime Thematic Categories                     313976 non-null  object
 9   Catégories thématiques sur la fraude et la cybercriminalité  313976 non-null  object
 10  Solicitation Method                                          313976 non-null  object
 11  Méthode de sollicitation                                     313976 non-null  object
 12  Gender                                                       313976 non-null  object
 13  Genre                                                        313976 non-null  object
 14  Language of Correspondence                                   313976 non-null  object
 15  Langue de correspondance                                     313976 non-null  object
 16  Victim Age Range / Tranche d'âge des victimes                313976 non-null  object
 17  Complaint Type                                               313976 non-null  object
 18  Type de plainte                                              313976 non-null  object
 19  Number of Victims / Nombre de victimes                       313976 non-null  int64
 20  Dollar Loss /pertes financières                              313976 non-null  object
dtypes: int64(2), object(19)
memory usage: 50.3+ MB
```

# Dataset overview - CAF

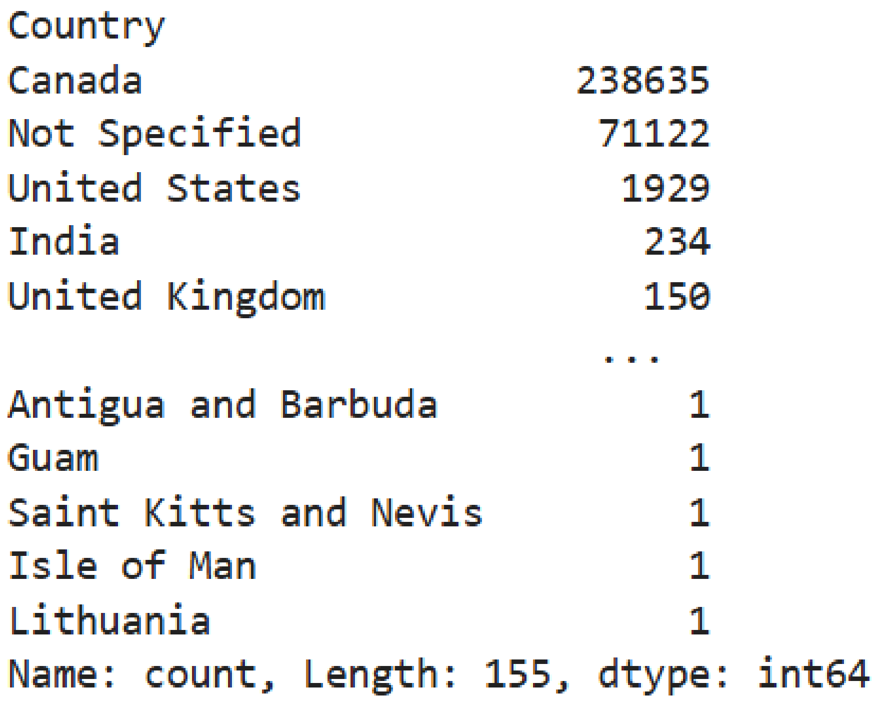


| | Numéro d'identification / Number ID | Date Received / Date reçue | Complaint Received Type | Type de plainte reçue | Country | Pays | Province/State | Province/ État | Fraud and Cybercrime Thematic Categories | Catégories thématiques sur la fraude et la cybercriminalité | ... | Méthode de sollicitation | Gender | Genre | Co |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 2021-01-02 | CAFC Website | CAFC site web | Canada | Canada | Nova Scotia | Nouvelle-Écosse | Phishing | Hameçonnage | ... | Messages texte | Female | Femme | |
| **1** | 2 | 2021-01-02 | CAFC Website | CAFC site web | Canada | Canada | British Columbia | Colombie-Britanique | Identity Fraud | Fraude à l'identité | ... | Autre/inconnu | Female | Femme | |
| **2** | 3 | 2021-01-02 | CAFC Website | CAFC site web | Not Specified | Non spécifié | Not Specified | Non spécifié | Romance | Romance | ... | Autre/inconnu | Not Available | non disponible | |
| **3** | 4 | 2021-01-02 | CAFC Website | CAFC site web | United States | États-Unis | California | Californie | Foreign Money Offer | Offre d'argent de l'étranger | ... | Courrier | Male | Homme | |
| **4** | 5 | 2021-01-02 | CAFC Website | CAFC site web | Canada | Canada | Ontario | Ontario | Merchandise | Marchandise | ... | Internet | Female | Femme | |

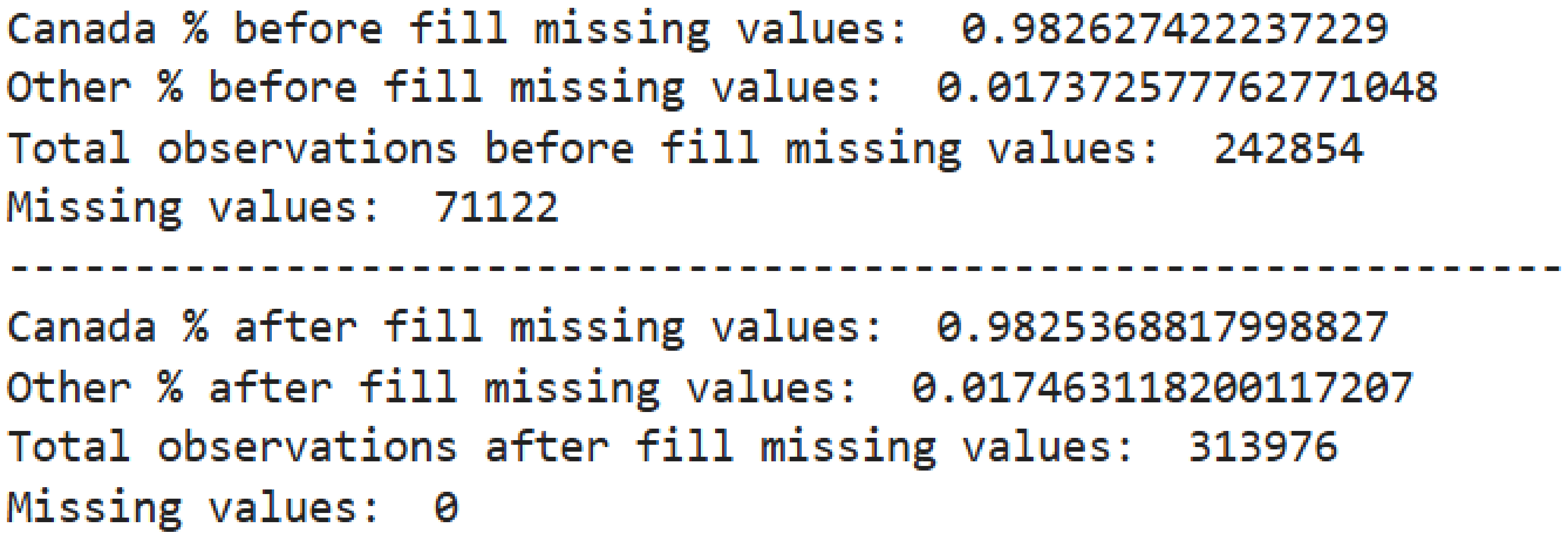
## NO MISSING VALUES



```
Numéro d'identification / Number ID                                0
Date Received / Date reçue                                         0
Complaint Received Type                                            0
Type de plainte reçue                                              0
Country                                                            0
Pays                                                               0
Province/State                                                     0
Province/État                                                      0
Fraud and Cybercrime Thematic Categories                           0
Catégories thématiques sur la fraude et la cybercriminalité        0
Solicitation Method                                                0
Méthode de sollicitation                                           0
Gender                                                             0
Genre                                                              0
Language of Correspondence                                         0
Langue de correspondance                                           0
Victim Age Range / Tranche d'âge des victimes                      0
Complaint Type                                                     0
Type de plainte                                                    0
Number of Victims / Nombre de victimes                             0
Dollar Loss /pertes financières                                    0
dtype: int64
```



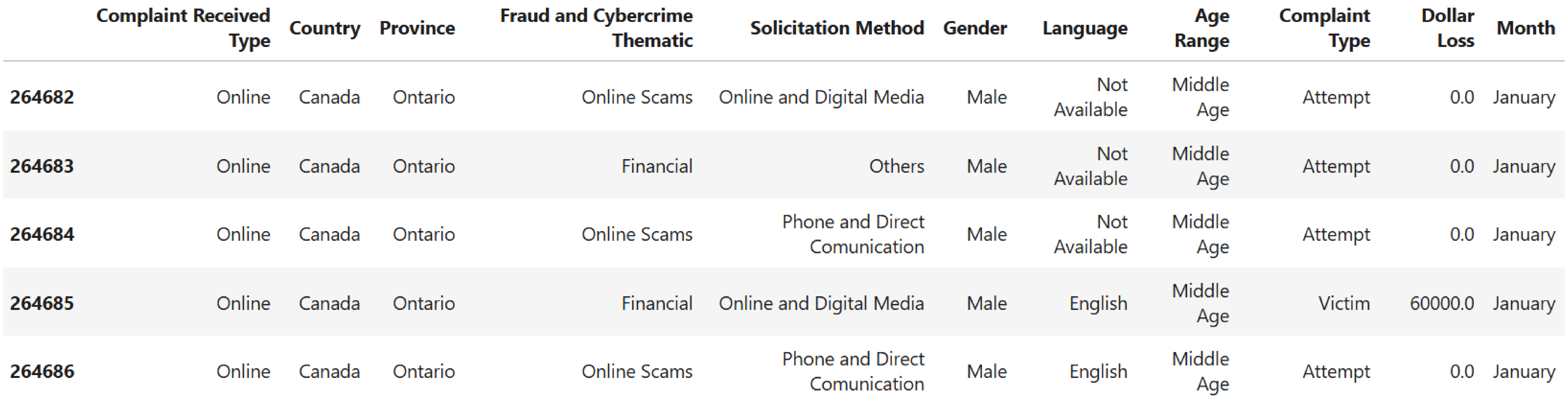
```
Country
Canada                    238635
Not Specified              71122
United States               1929
India                        234
United Kingdom               150
                           ...
Antigua and Barbuda            1
Guam                           1
Saint Kitts and Nevis          1
Isle of Man                    1
Lithuania                      1
Name: count, Length: 155, dtype: int64
```



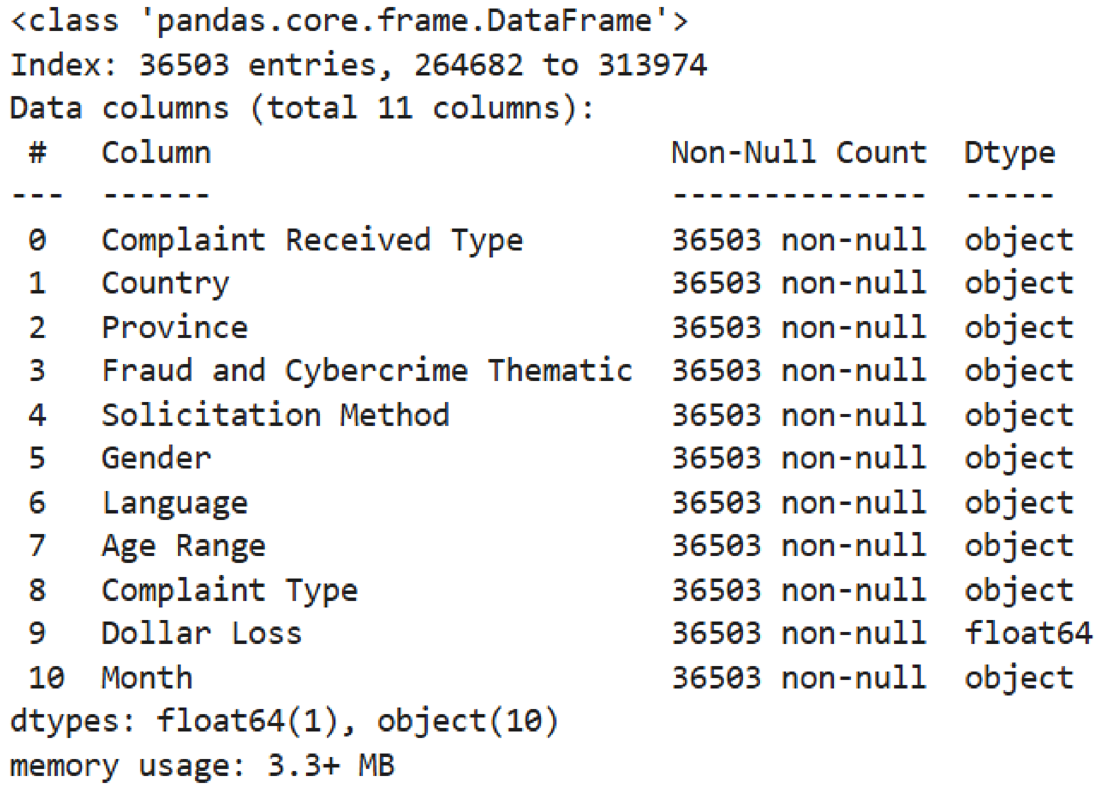
```
Canada % before fill missing values:  0.982627422237229
Other % before fill missing values:  0.017372577762771048
Total observations before fill missing values:  242854
Missing values:  71122
--------------------------------------------------------------
Canada % after fill missing values:  0.9825368817998827
Other % after fill missing values:  0.017463118200117207
Total observations after fill missing values:  313976
Missing values:  0
```

Segmentation column: Country

# Dataset overview - CAF

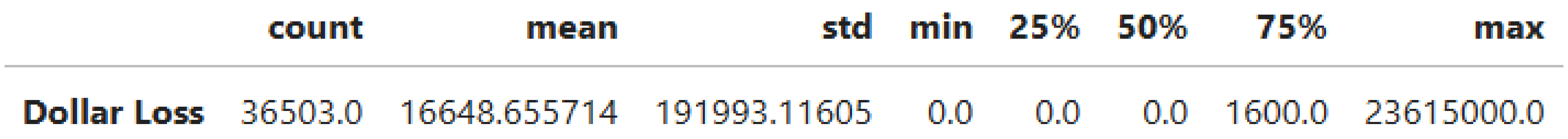


| | Complaint Received Type | Country | Province | Fraud and Cybercrime Thematic | Solicitation Method | Gender | Language | Age Range | Complaint Type | Dollar Loss | Month |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **264682** | Online | Canada | Ontario | Online Scams | Online and Digital Media | Male | Not Available | Middle Age | Attempt | 0.0 | January |
| **264683** | Online | Canada | Ontario | Financial | Others | Male | Not Available | Middle Age | Attempt | 0.0 | January |
| **264684** | Online | Canada | Ontario | Online Scams | Phone and Direct Comunication | Male | Not Available | Middle Age | Attempt | 0.0 | January |
| **264685** | Online | Canada | Ontario | Financial | Online and Digital Media | Male | English | Middle Age | Victim | 60000.0 | January |
| **264686** | Online | Canada | Ontario | Online Scams | Phone and Direct Comunication | Male | English | Middle Age | Attempt | 0.0 | January |

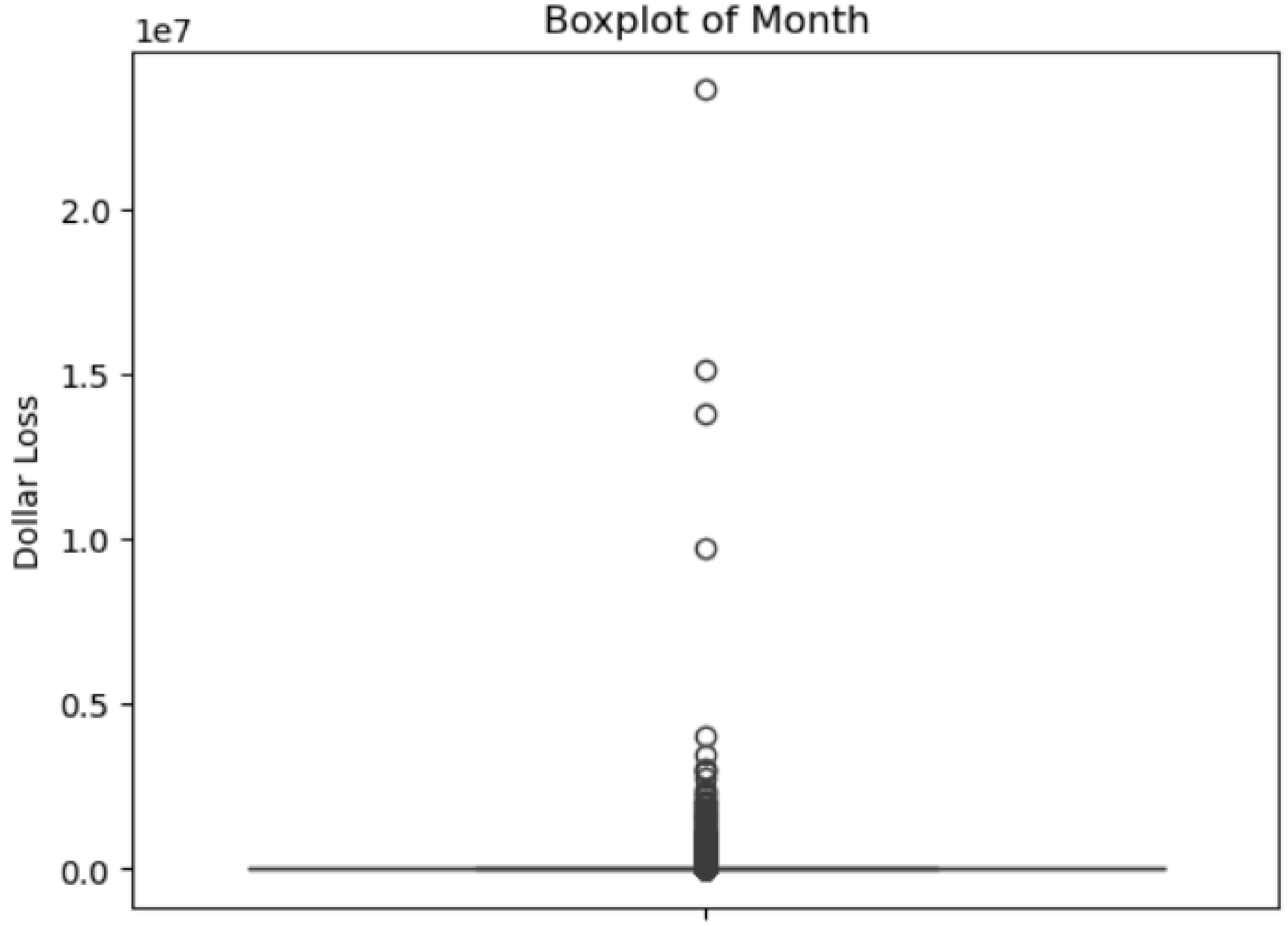


```
<class 'pandas.core.frame.DataFrame'>
Index: 36503 entries, 264682 to 313974
Data columns (total 11 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   Complaint Received Type        36503 non-null  object 
 1   Country                        36503 non-null  object 
 2   Province                       36503 non-null  object 
 3   Fraud and Cybercrime Thematic  36503 non-null  object 
 4   Solicitation Method            36503 non-null  object 
 5   Gender                         36503 non-null  object 
 6   Language                       36503 non-null  object 
 7   Age Range                      36503 non-null  object 
 8   Complaint Type                 36503 non-null  object 
 9   Dollar Loss                    36503 non-null  float64
 10  Month                          36503 non-null  object 
dtypes: float64(1), object(10)
memory usage: 3.3+ MB
```

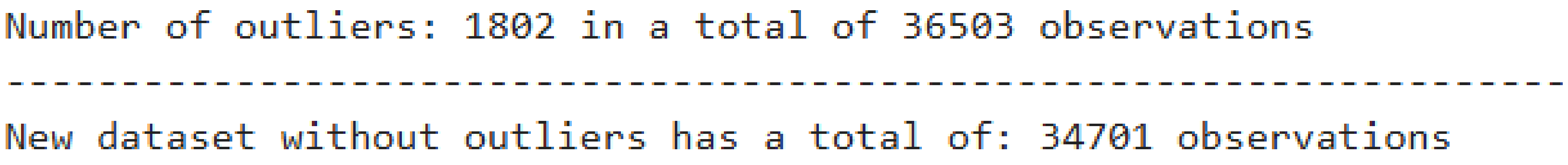
# Univariate Analysis (Numeric variables)



| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Dollar Loss** | 36503.0 | 16648.655714 | 191993.11605 | 0.0 | 0.0 | 0.0 | 1600.0 | 23615000.0 |

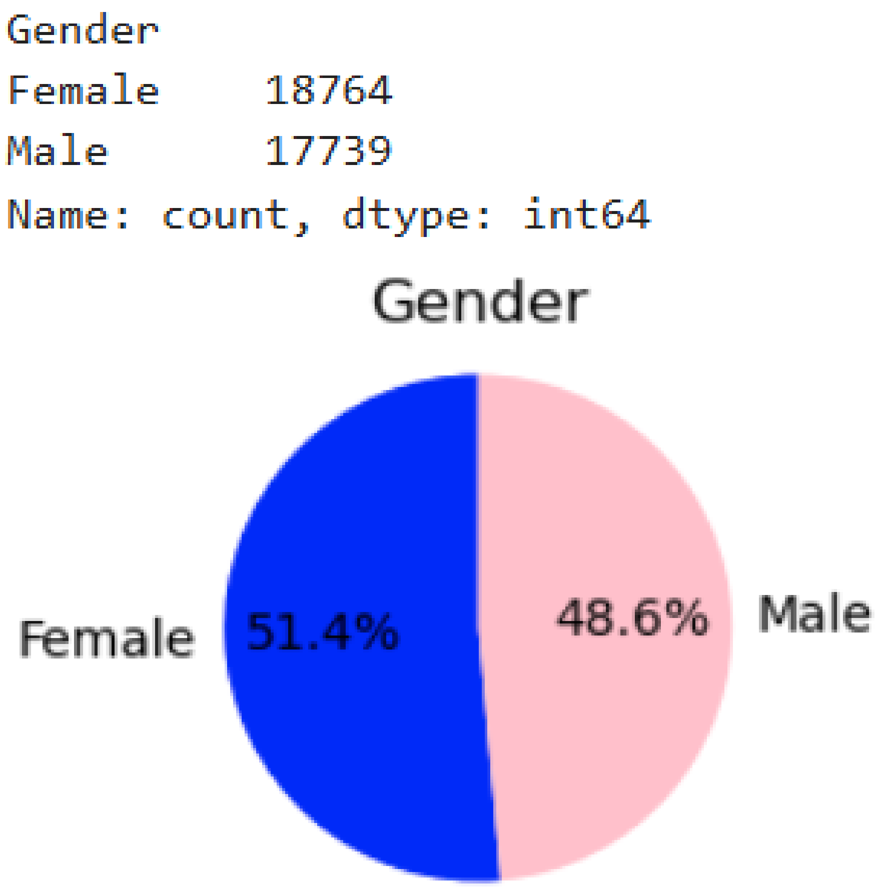


**Dollar Loss: Total amount of money lost to the instance(s) of fraud.**



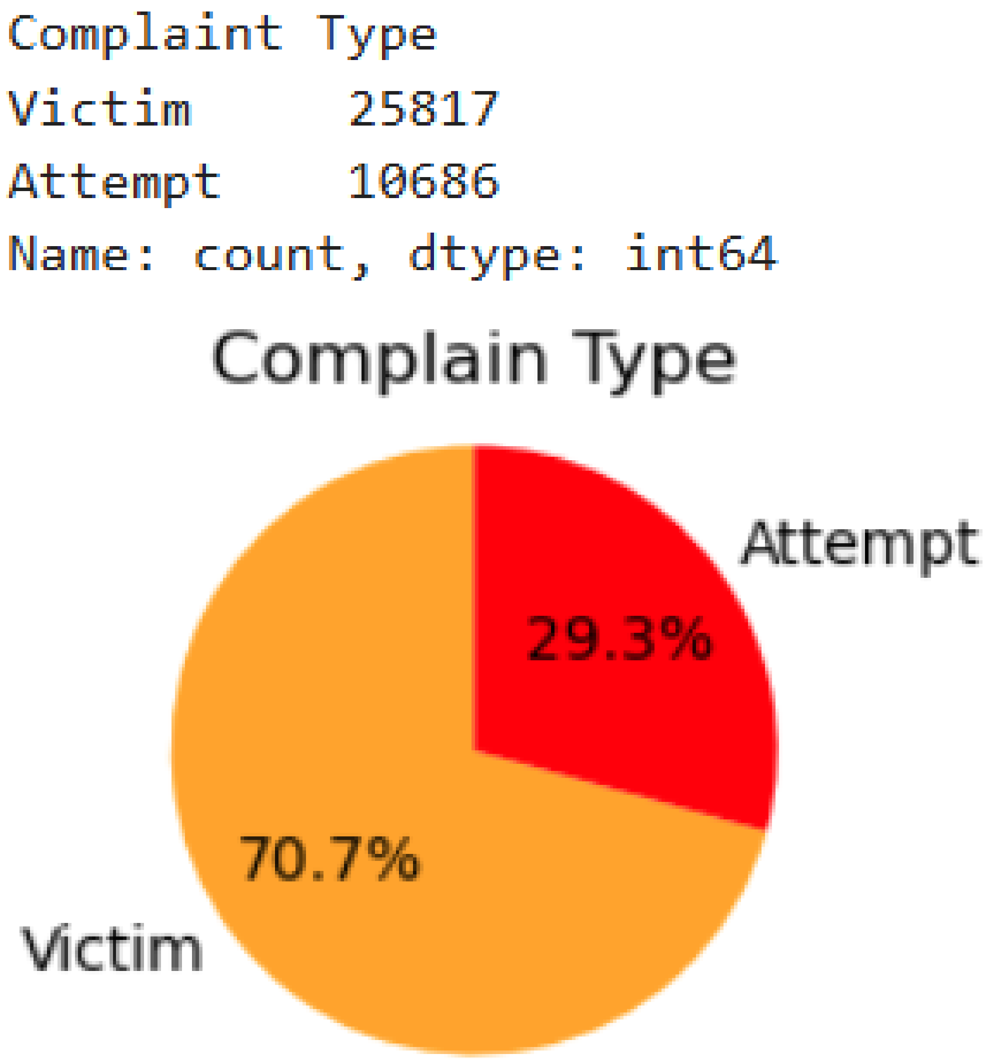
```
Number of outliers: 1802 in a total of 36503 observations
------------------------------------------------------------------
New dataset without outliers has a total of: 34701 observations
```

# Univariate Analysis (Categorical variables)

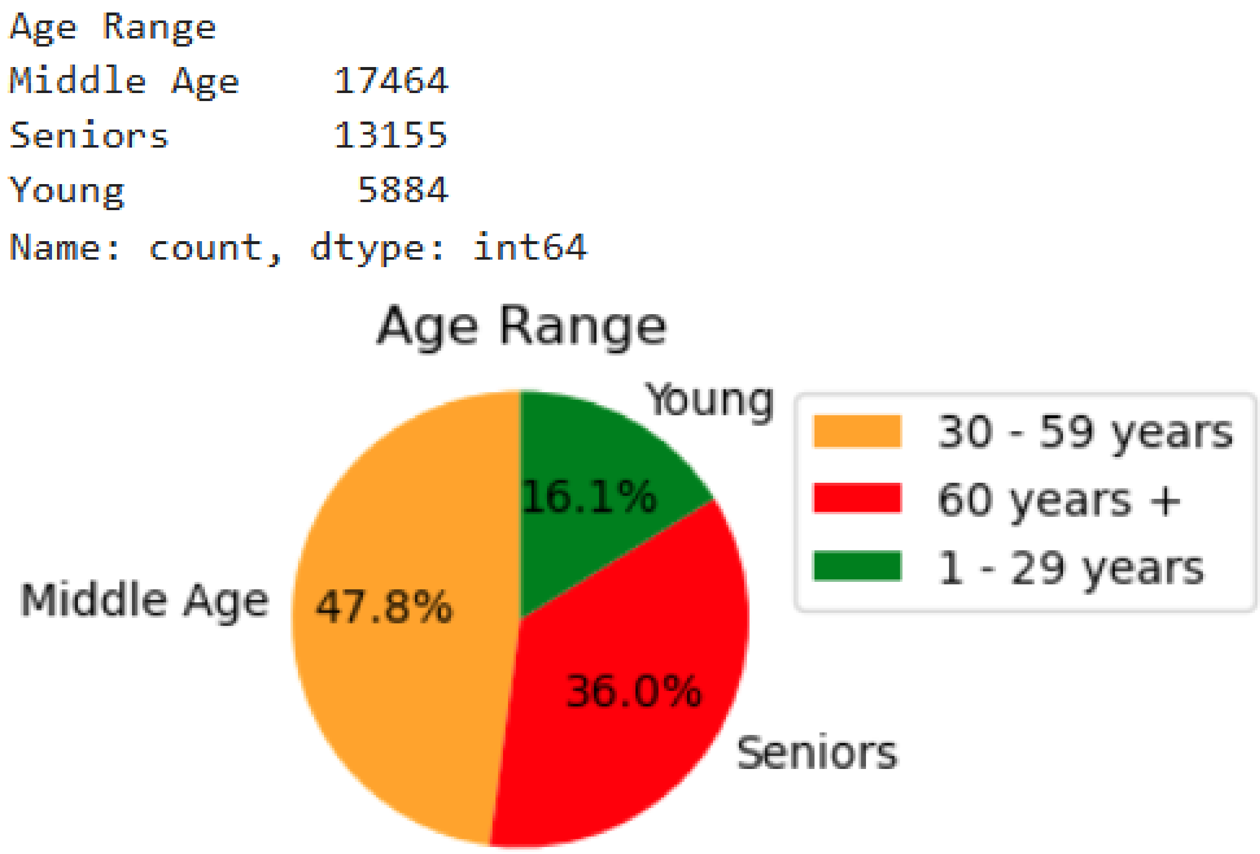


```
Gender
Female     18764
Male       17739
Name: count, dtype: int64
```

**TARGET**

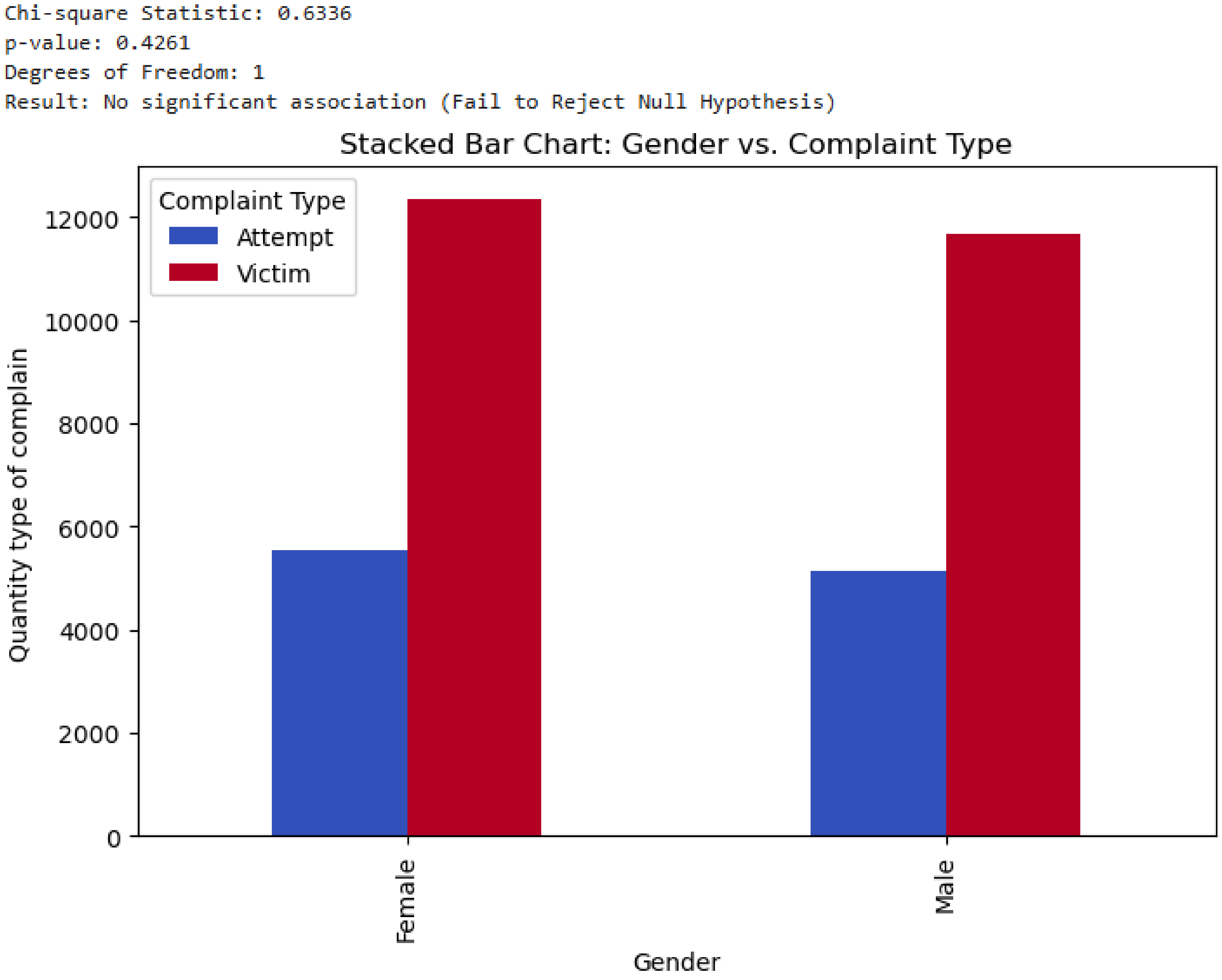


```
Complaint Type
Victim      25817
Attempt     10686
Name: count, dtype: int64
```



```
Age Range
Middle Age     17464
Seniors        13155
Young           5884
Name: count, dtype: int64
```

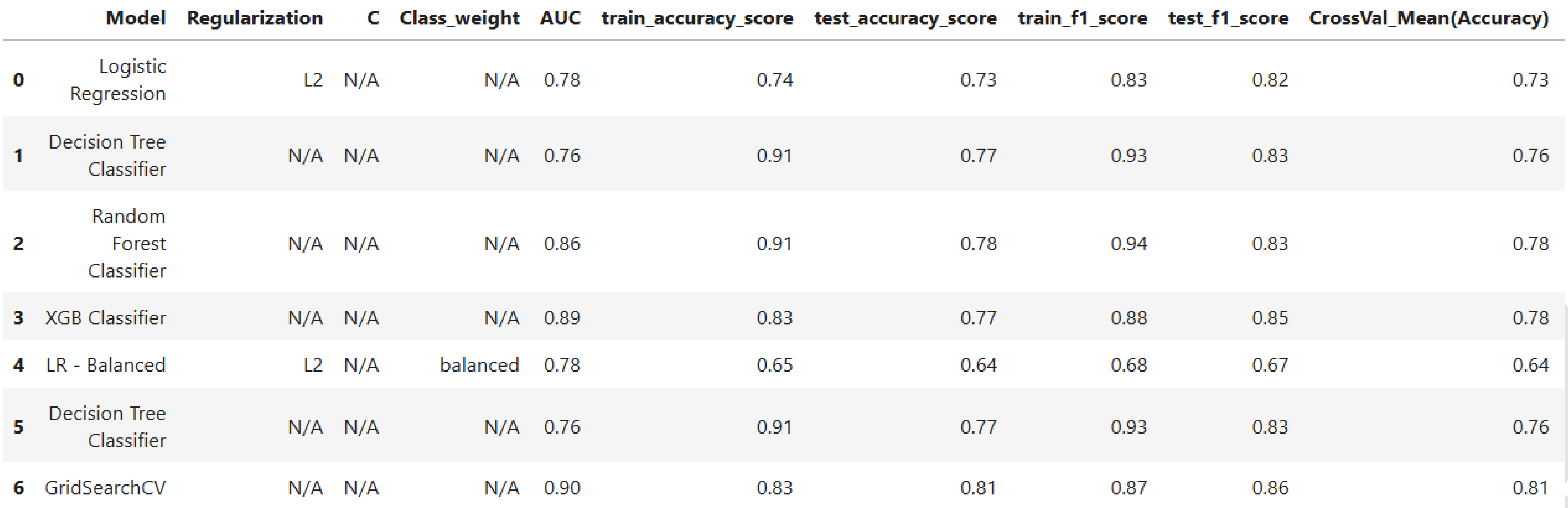
# Bivariate Analysis



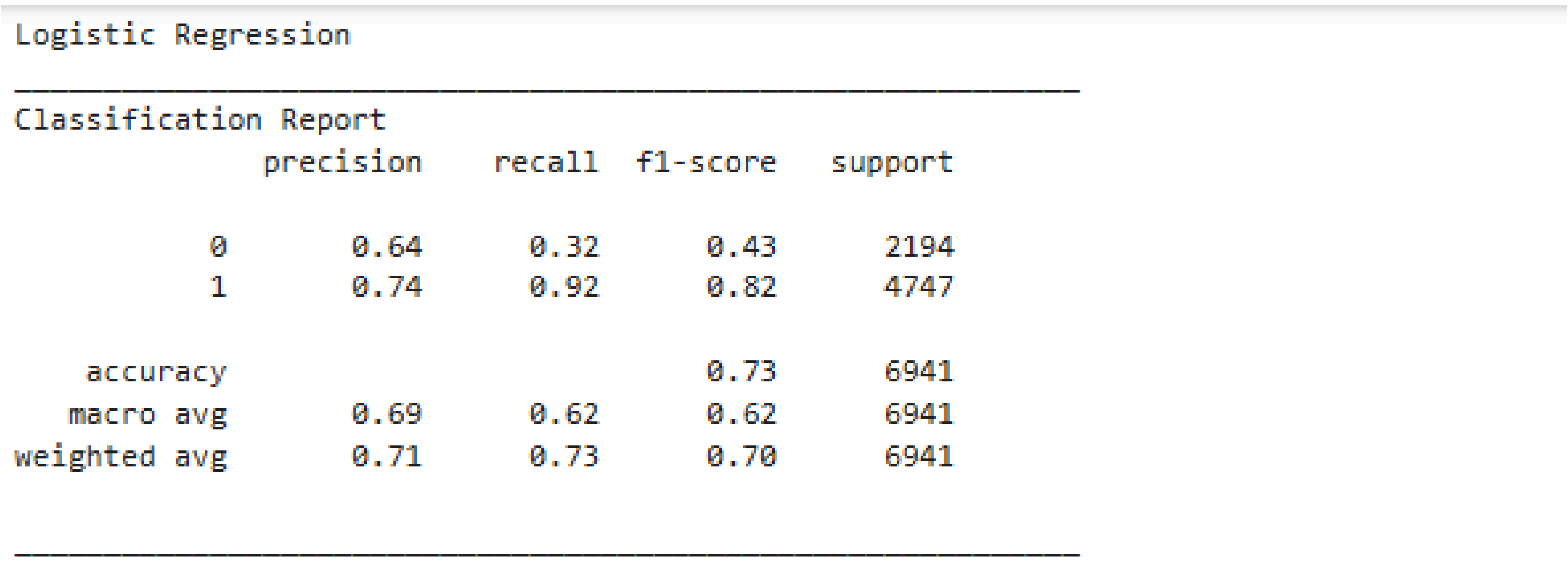
```
Chi-square Statistic: 0.6336
p-value: 0.4261
Degrees of Freedom: 1
Result: No significant association (Fail to Reject Null Hypothesis)
```

# Comparing the performance models.

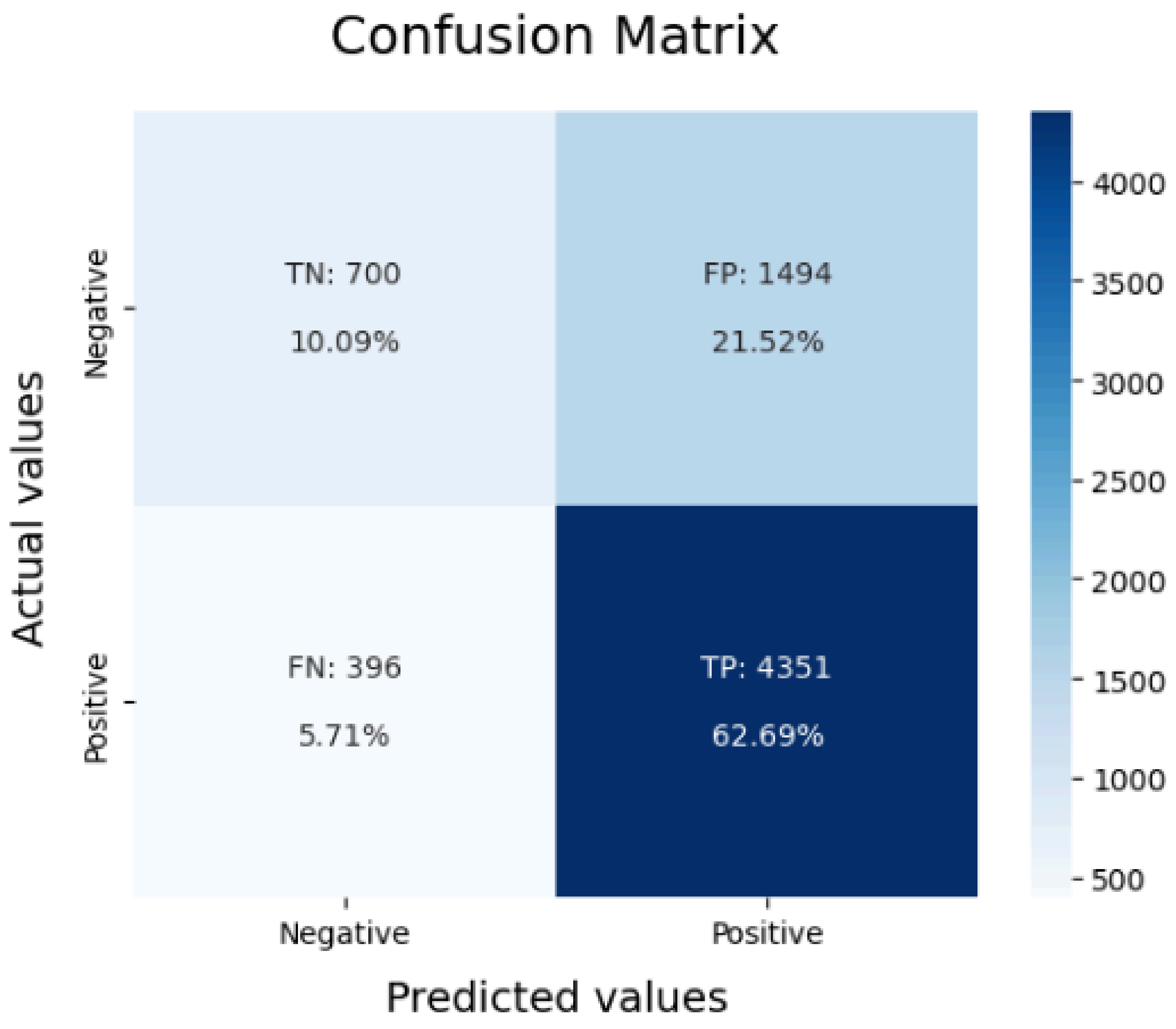
**Since we have an imbalanced dataset, accuracy can give us false assumptions regarding the classifier's performance, it's better to rely on precision and recall, in the same way, a Precision-Recall curve is better to calibrate the probability threshold in an imbalanced class scenario as a ROC curve. Precision and recall can be combined into a single score that seeks to balance both concerns, called the F-score or the F-measure. In this project, I will select the best model based on F-score.**



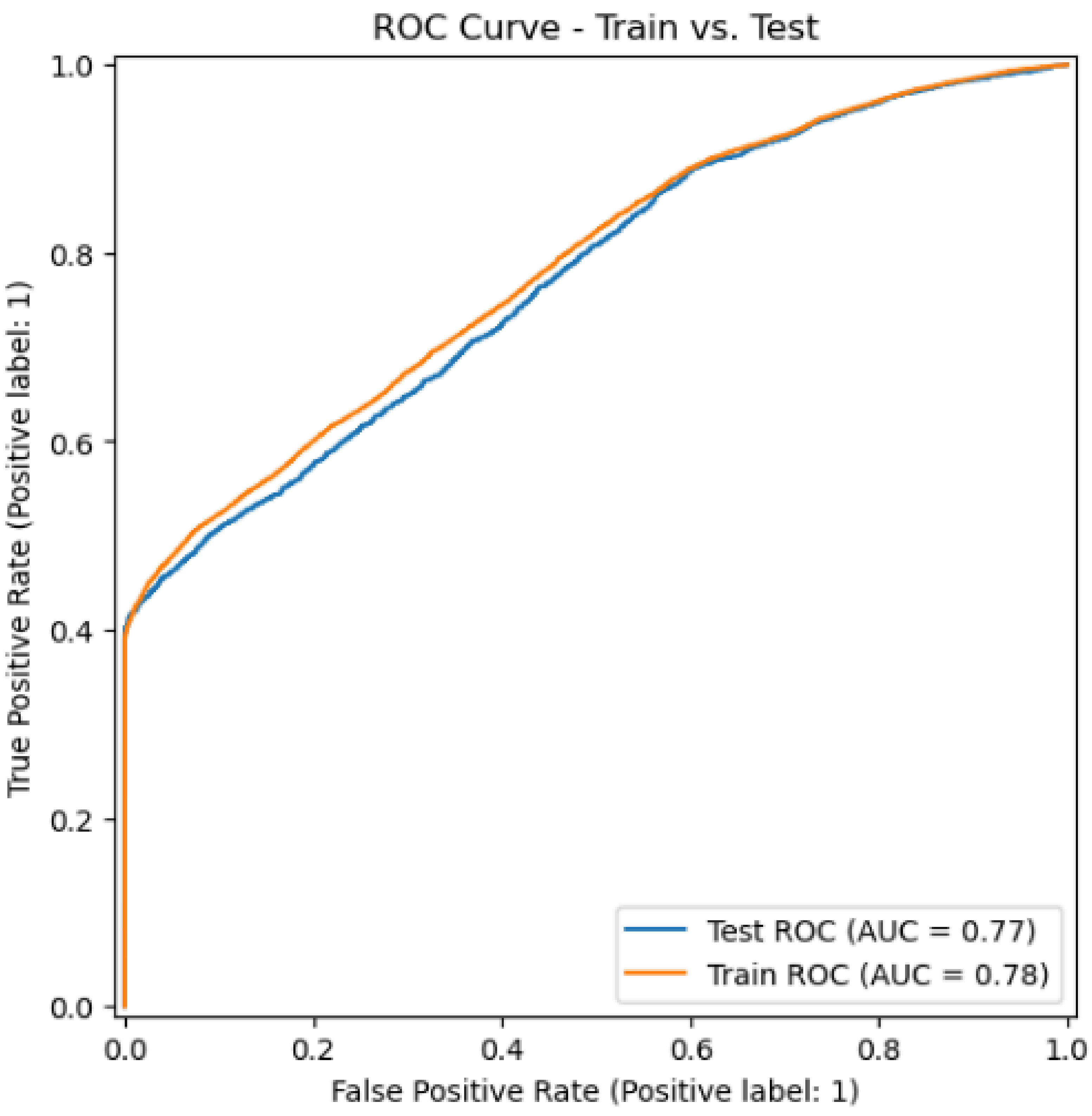
| | Model | Regularization | C | Class_weight | AUC | train_accuracy_score | test_accuracy_score | train_f1_score | test_f1_score | CrossVal_Mean(Accuracy) |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Logistic Regression | L2 | N/A | N/A | 0.78 | 0.74 | 0.73 | 0.83 | 0.82 | 0.73 |
| 1 | Decision Tree Classifier | N/A | N/A | N/A | 0.76 | 0.91 | 0.77 | 0.93 | 0.83 | 0.76 |
| 2 | Random Forest Classifier | N/A | N/A | N/A | 0.86 | 0.91 | 0.78 | 0.94 | 0.83 | 0.78 |
| 3 | XGB Classifier | N/A | N/A | N/A | 0.89 | 0.83 | 0.77 | 0.88 | 0.85 | 0.78 |
| 4 | LR - Balanced | L2 | N/A | balanced | 0.78 | 0.65 | 0.64 | 0.68 | 0.67 | 0.64 |
| 5 | Decision Tree Classifier | N/A | N/A | N/A | 0.76 | 0.91 | 0.77 | 0.93 | 0.83 | 0.76 |
| 6 | GridSearchCV | N/A | N/A | N/A | 0.90 | 0.83 | 0.81 | 0.87 | 0.86 | 0.81 |

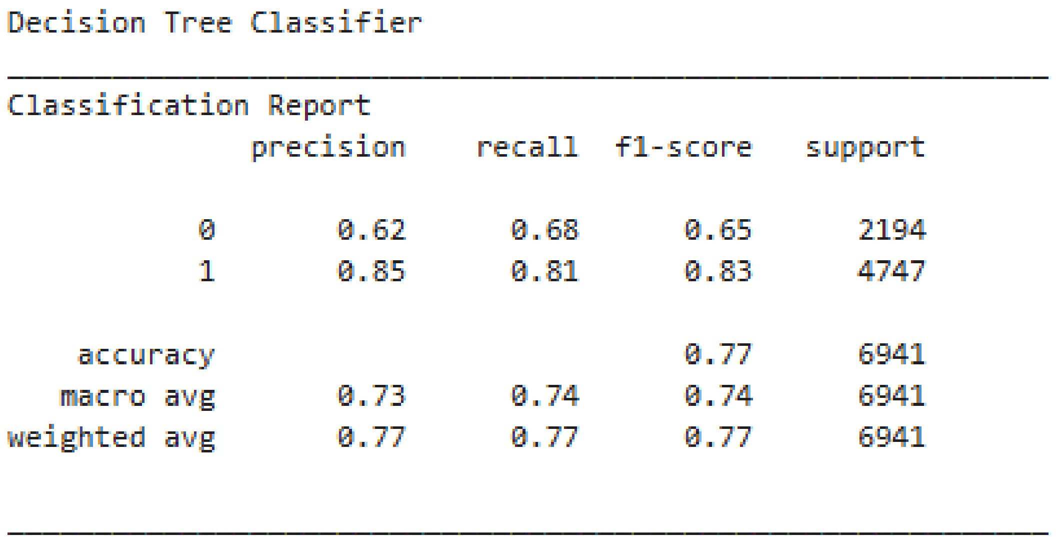


```
Logistic Regression
_________________________________________________________________
Classification Report
              precision    recall  f1-score   support

           0       0.64      0.32      0.43      2194
           1       0.74      0.92      0.82      4747

    accuracy                           0.73      6941
   macro avg       0.69      0.62      0.62      6941
weighted avg       0.71      0.73      0.70      6941

_________________________________________________________________
```

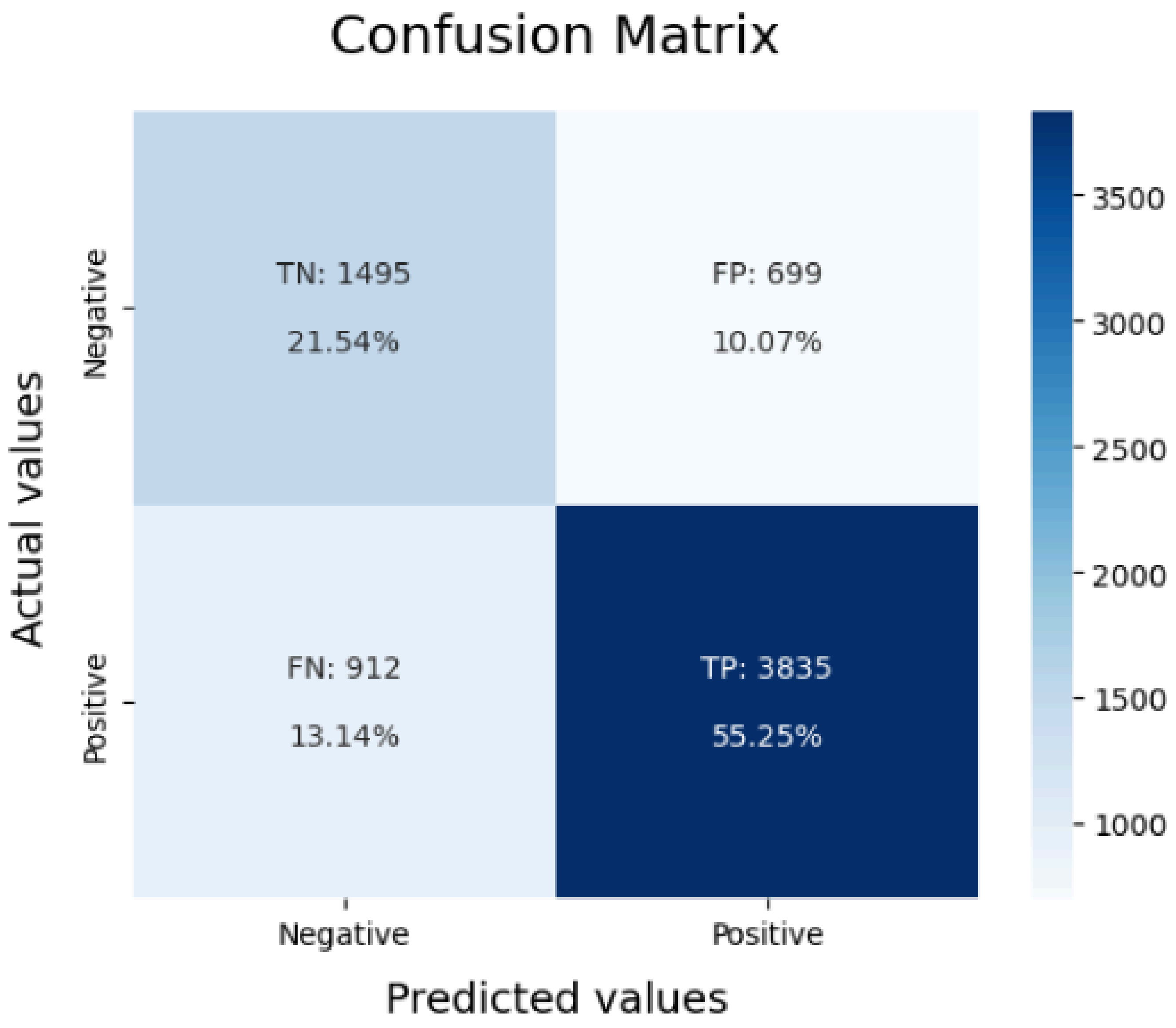


```
================================================================================
```



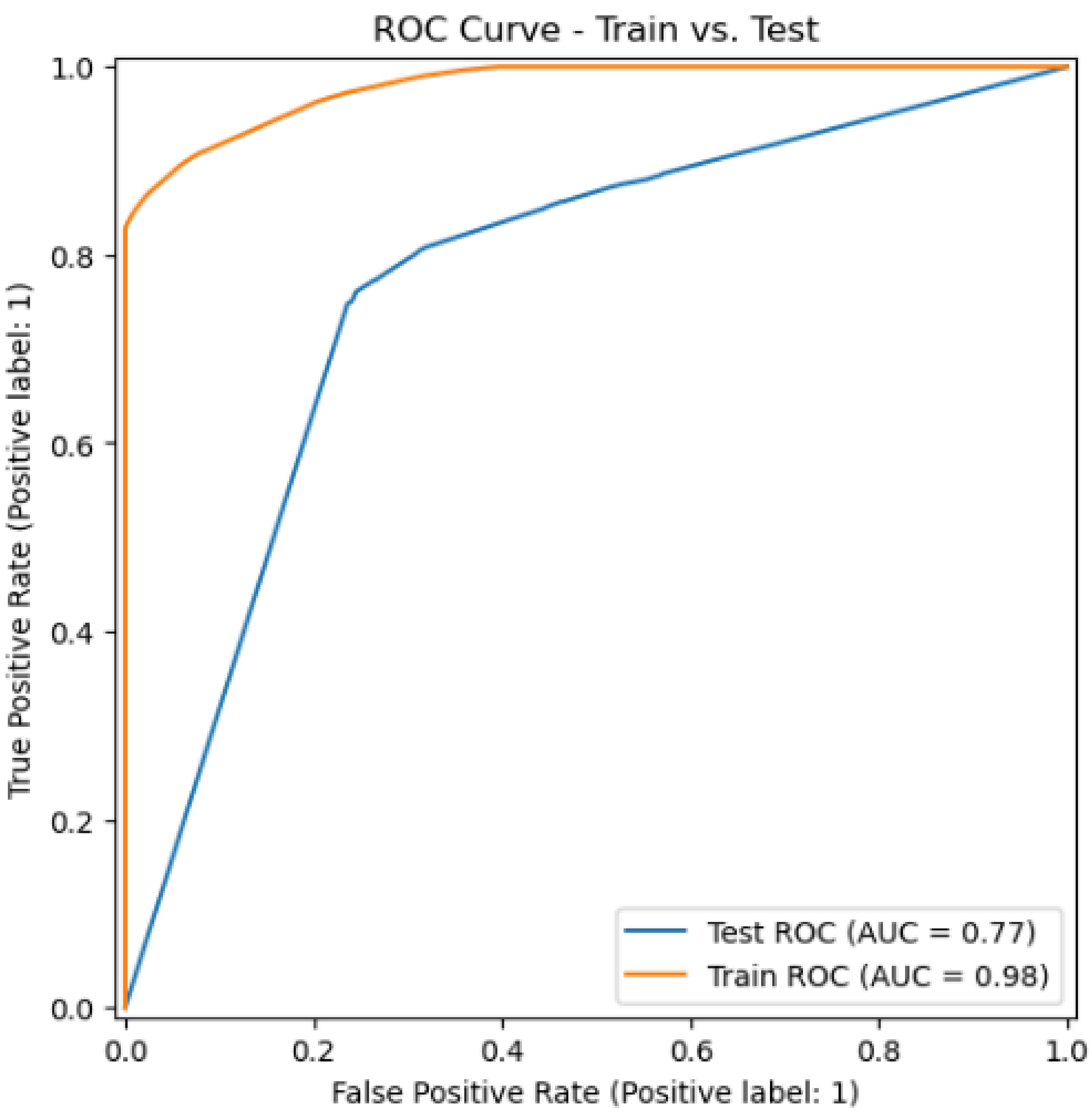


```
Decision Tree Classifier
____________________________________________________________
Classification Report
              precision    recall  f1-score   support

           0       0.62      0.68      0.65      2194
           1       0.85      0.81      0.83      4747

    accuracy                           0.77      6941
   macro avg       0.73      0.74      0.74      6941
weighted avg       0.77      0.77      0.77      6941

____________________________________________________________
```

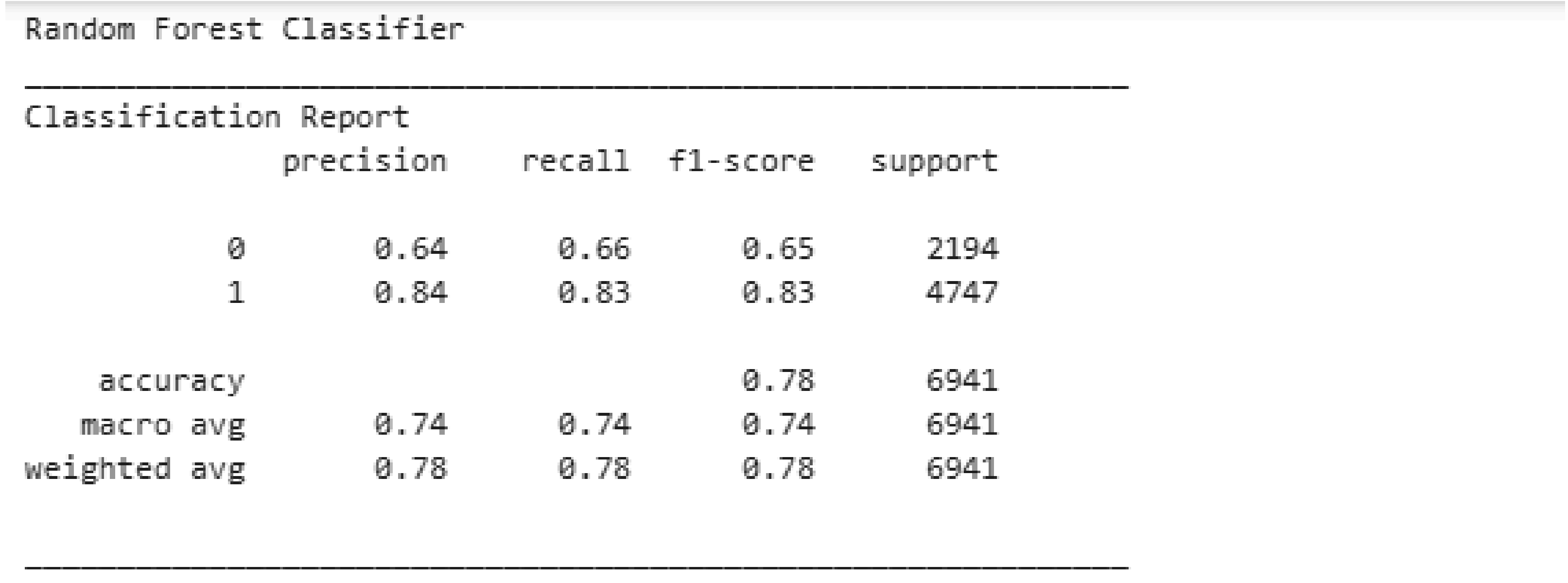


Confusion Matrix

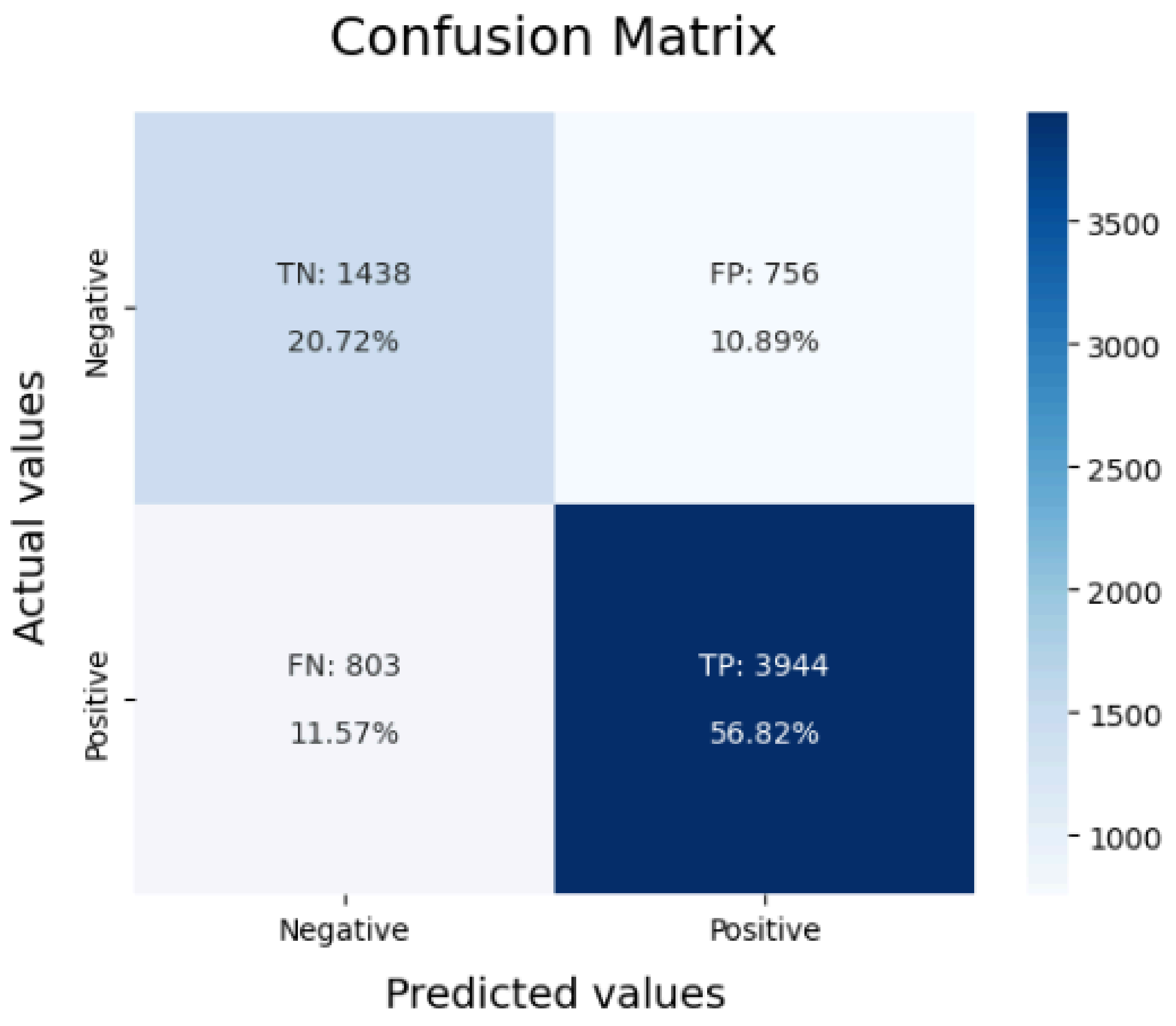
```
===================================================================================
```

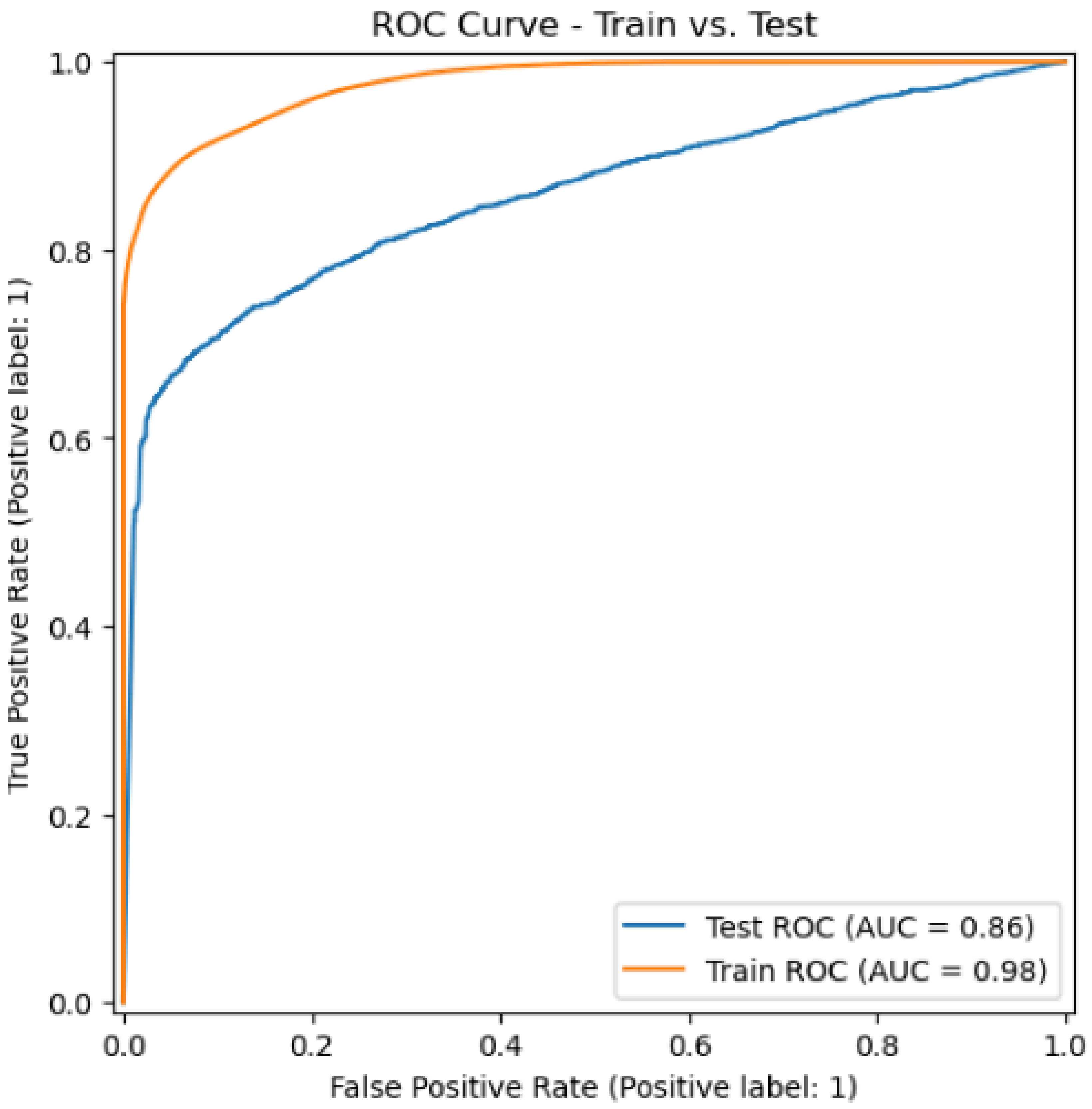


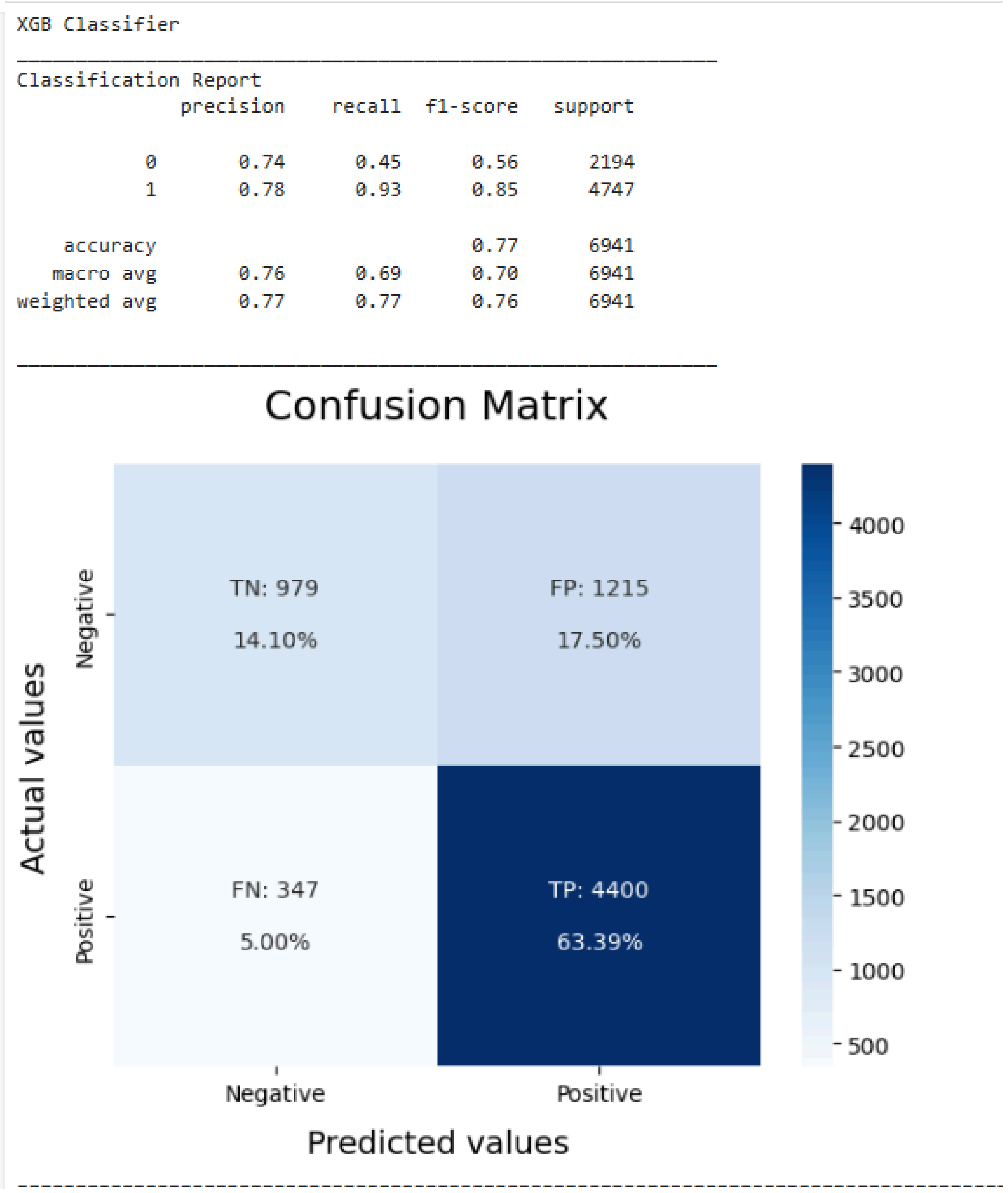
ROC Curve - Train vs. Test



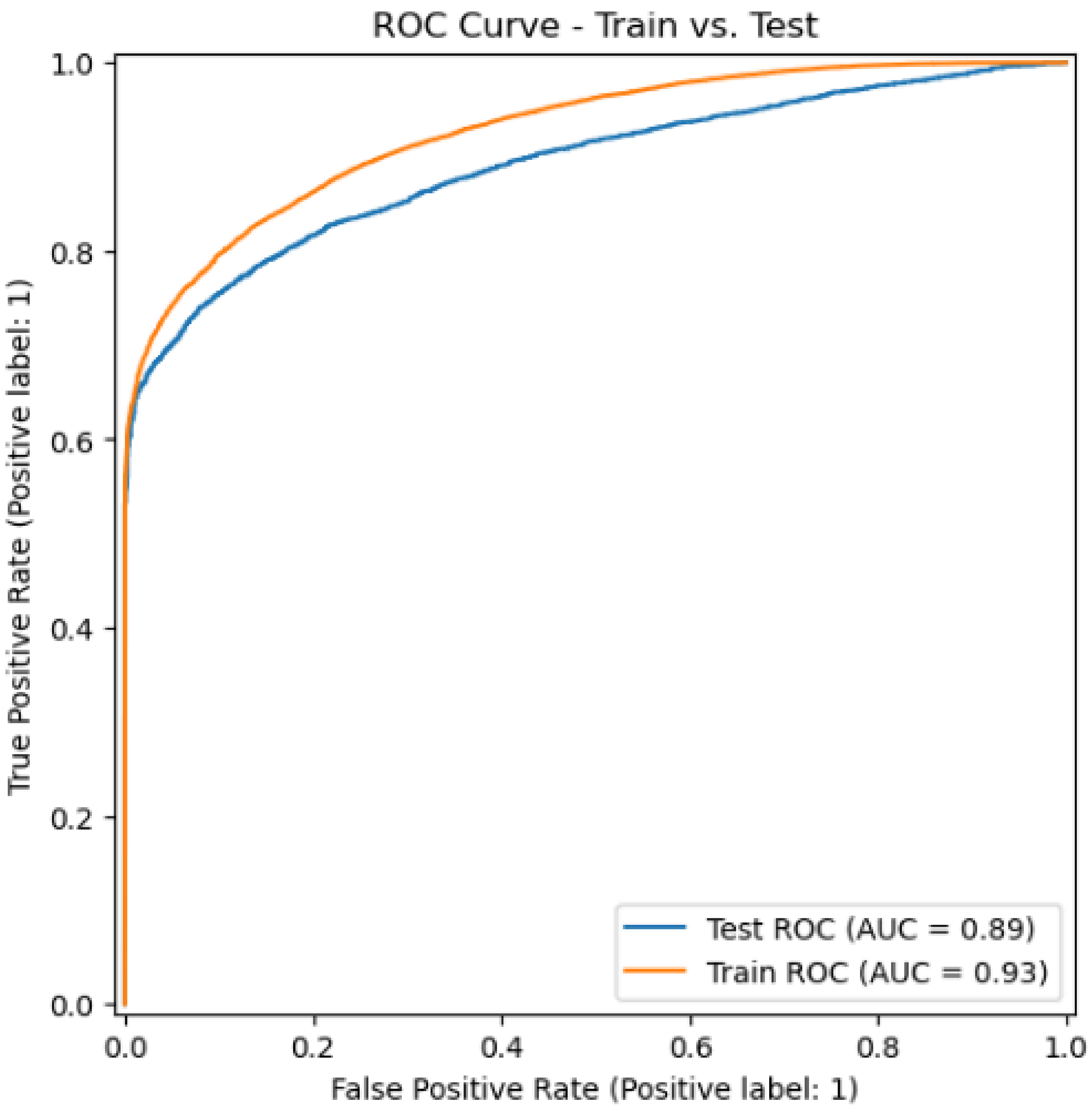
```
Random Forest Classifier
_______________________________________________________________
Classification Report
              precision    recall  f1-score   support

           0       0.64      0.66      0.65      2194
           1       0.84      0.83      0.83      4747

    accuracy                           0.78      6941
   macro avg       0.74      0.74      0.74      6941
weighted avg       0.78      0.78      0.78      6941

_______________________________________________________________
```

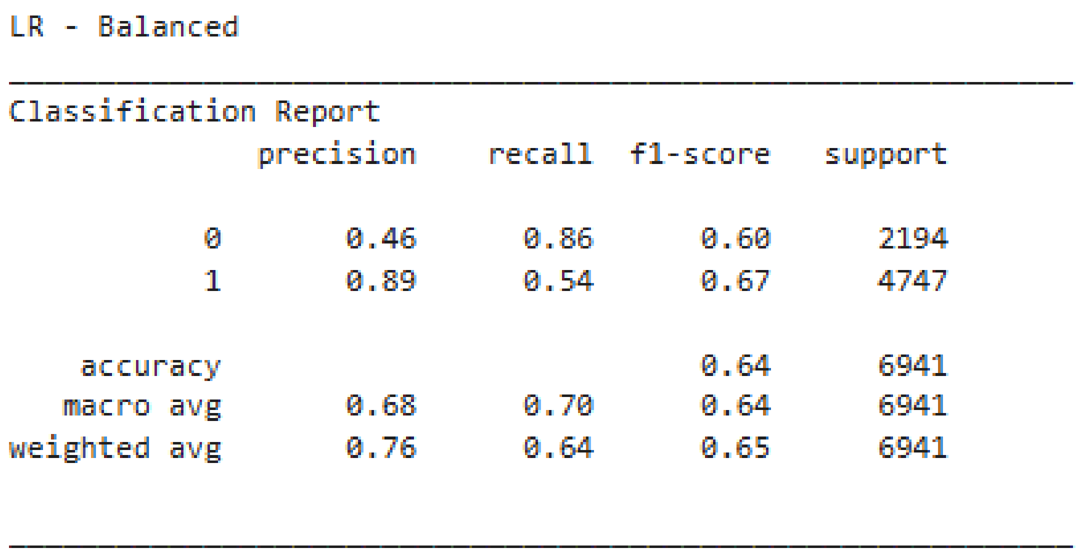




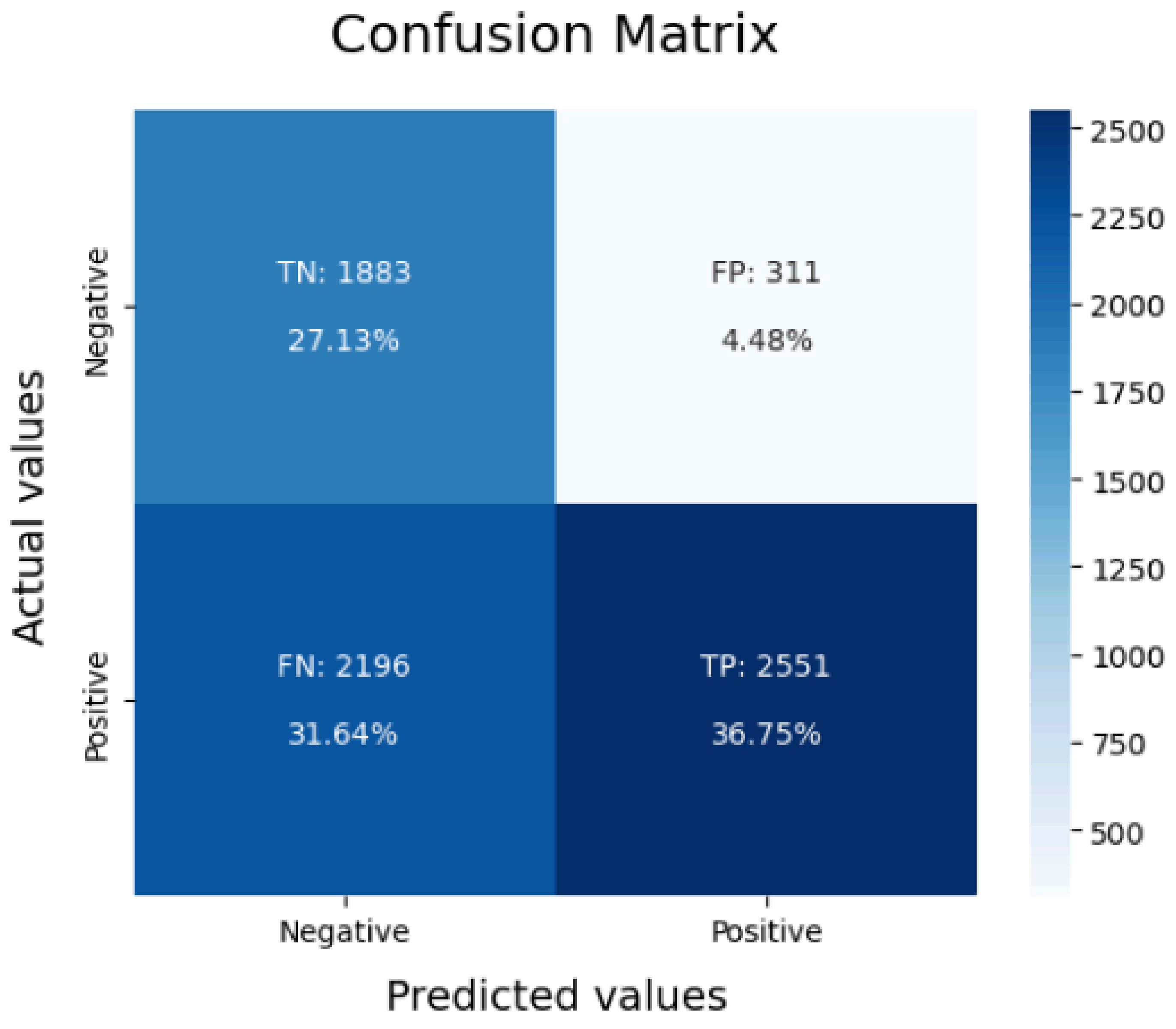


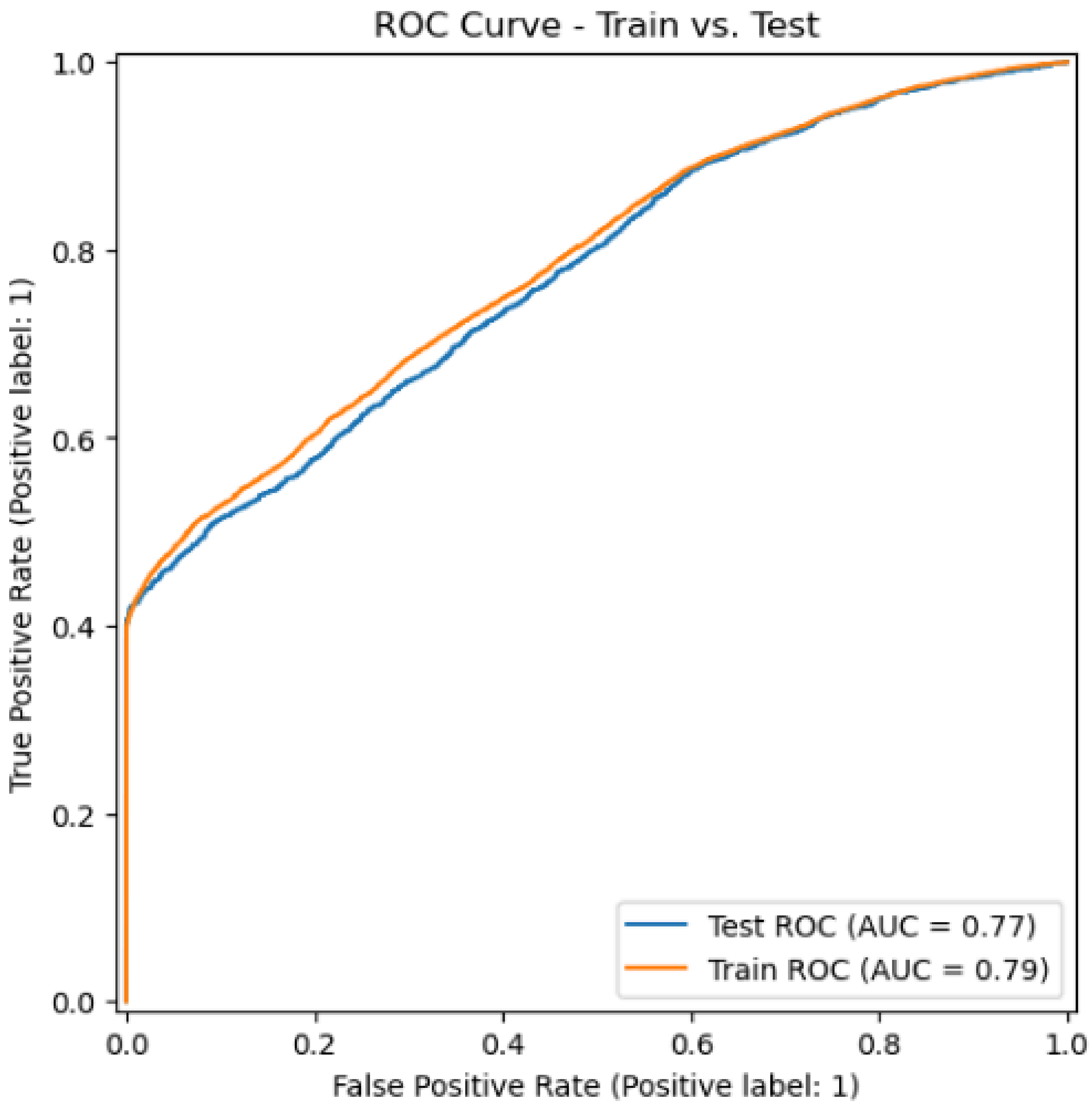
```
XGB Classifier
________________________________________________________________
Classification Report
              precision    recall  f1-score   support

           0       0.74      0.45      0.56      2194
           1       0.78      0.93      0.85      4747

    accuracy                           0.77      6941
   macro avg       0.76      0.69      0.70      6941
weighted avg       0.77      0.77      0.76      6941

________________________________________________________________
```

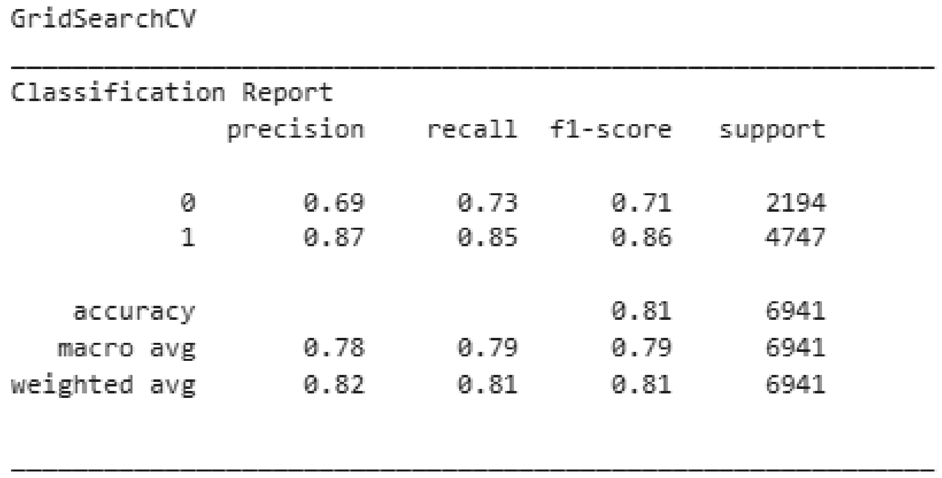




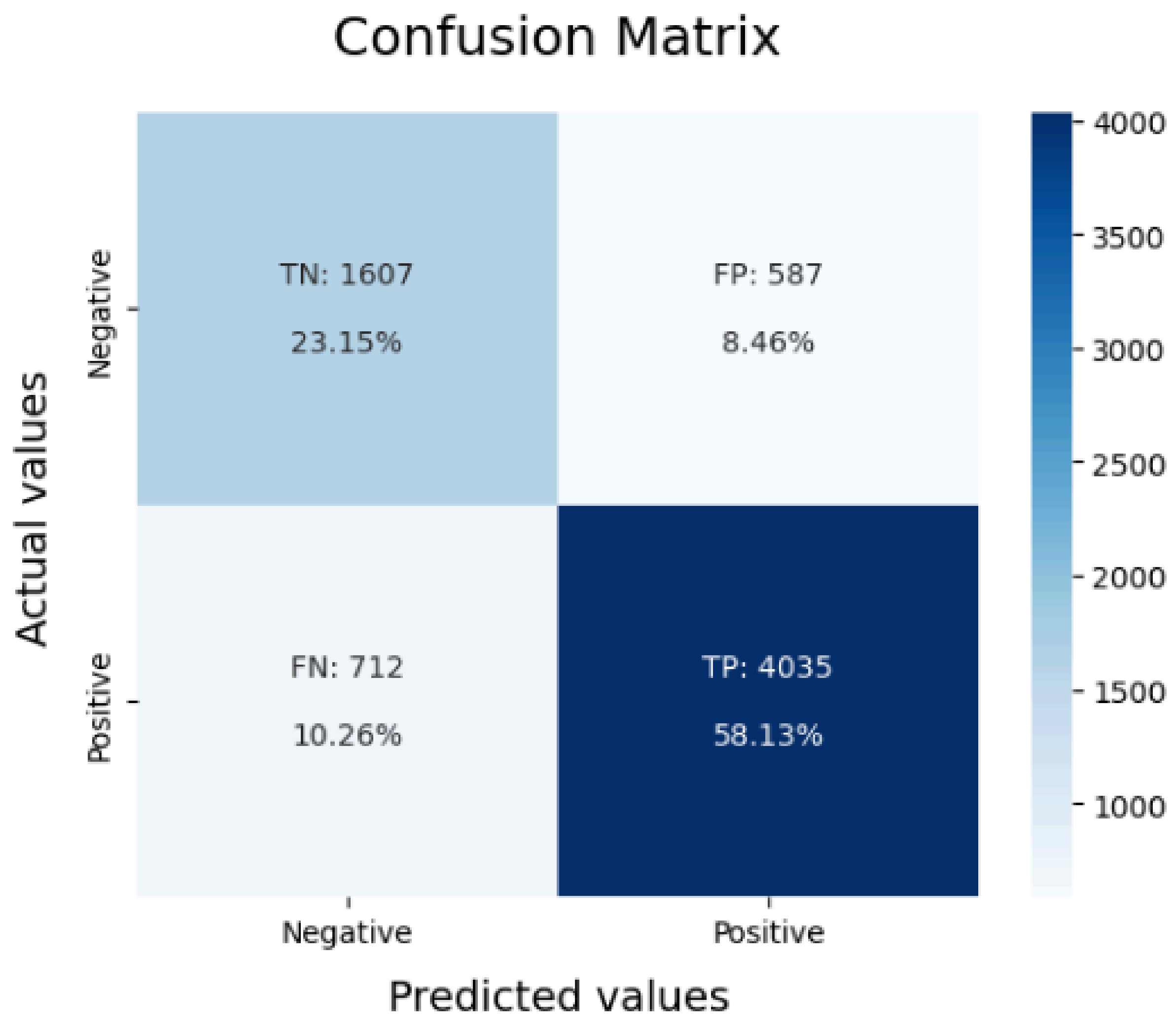
```
LR - Balanced
__________________________________________________________
Classification Report
              precision    recall  f1-score   support

           0       0.46      0.86      0.60      2194
           1       0.89      0.54      0.67      4747

    accuracy                           0.64      6941
   macro avg       0.68      0.70      0.64      6941
weighted avg       0.76      0.64      0.65      6941

__________________________________________________________
```

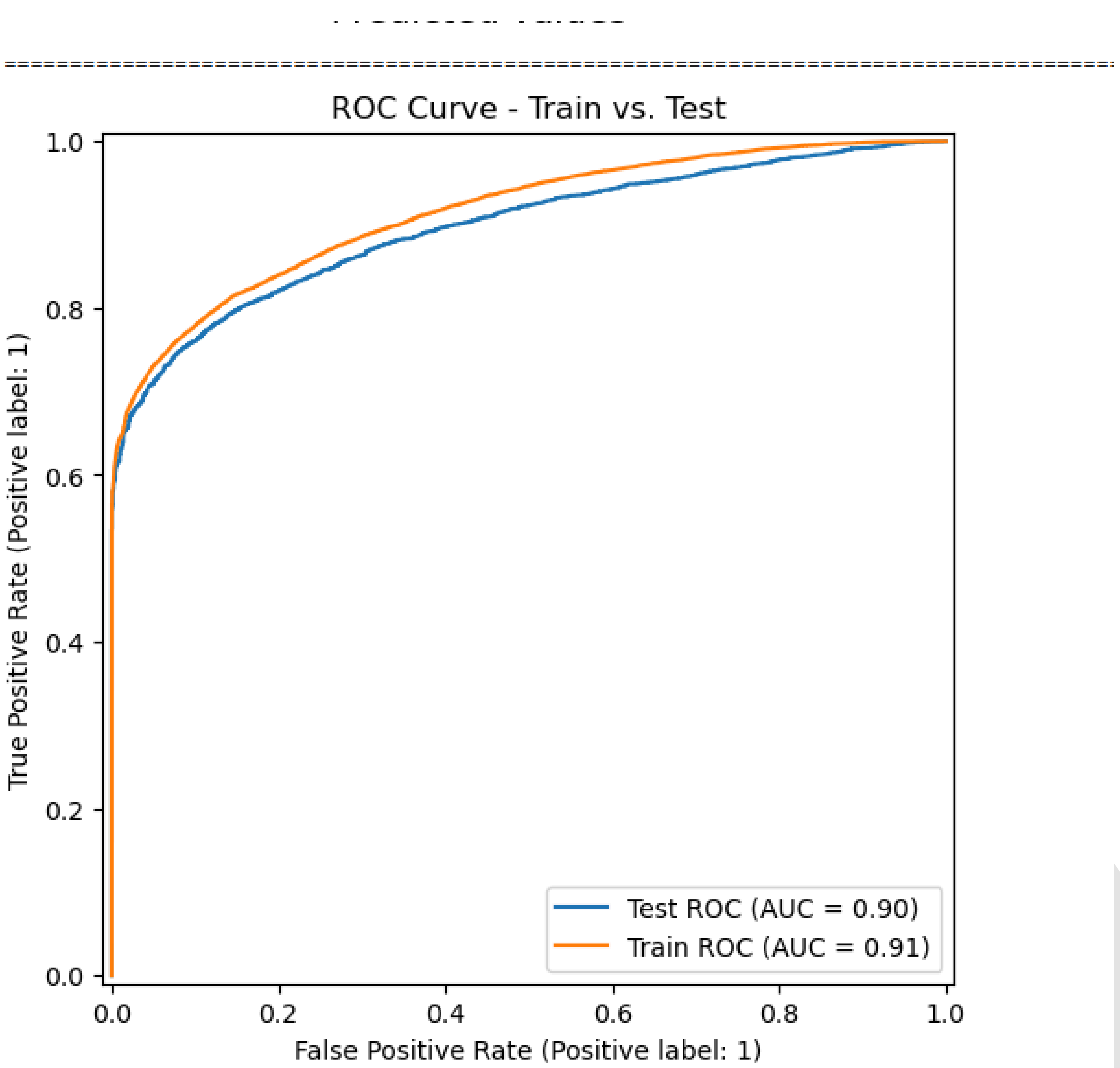






```
GridSearchCV
____________________________________________________________
Classification Report
              precision    recall  f1-score   support

           0       0.69      0.73      0.71      2194
           1       0.87      0.85      0.86      4747

    accuracy                           0.81      6941
   macro avg       0.78      0.79      0.79      6941
weighted avg       0.82      0.81      0.81      6941

____________________________________________________________
```


Confusion Matrix
TN: 1607
FP: 587
FN: 712
TP: 4035
Actual values
Negative
Positive
Predicted values
Negative
Positive


ROC Curve - Train vs. Test
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
Test ROC (AUC = 0.90)
Train ROC (AUC = 0.91)