Citibike Classification

Alexander Sigrist

Found 2013

- Bike Sharing System
- Fleet of 40,000
- Launched e-bikes in 2019

- Identify a model that can effectively detect usertype

Subscribers vs Customers

- What’s the difference?
- Why is it useful?

Data Overview

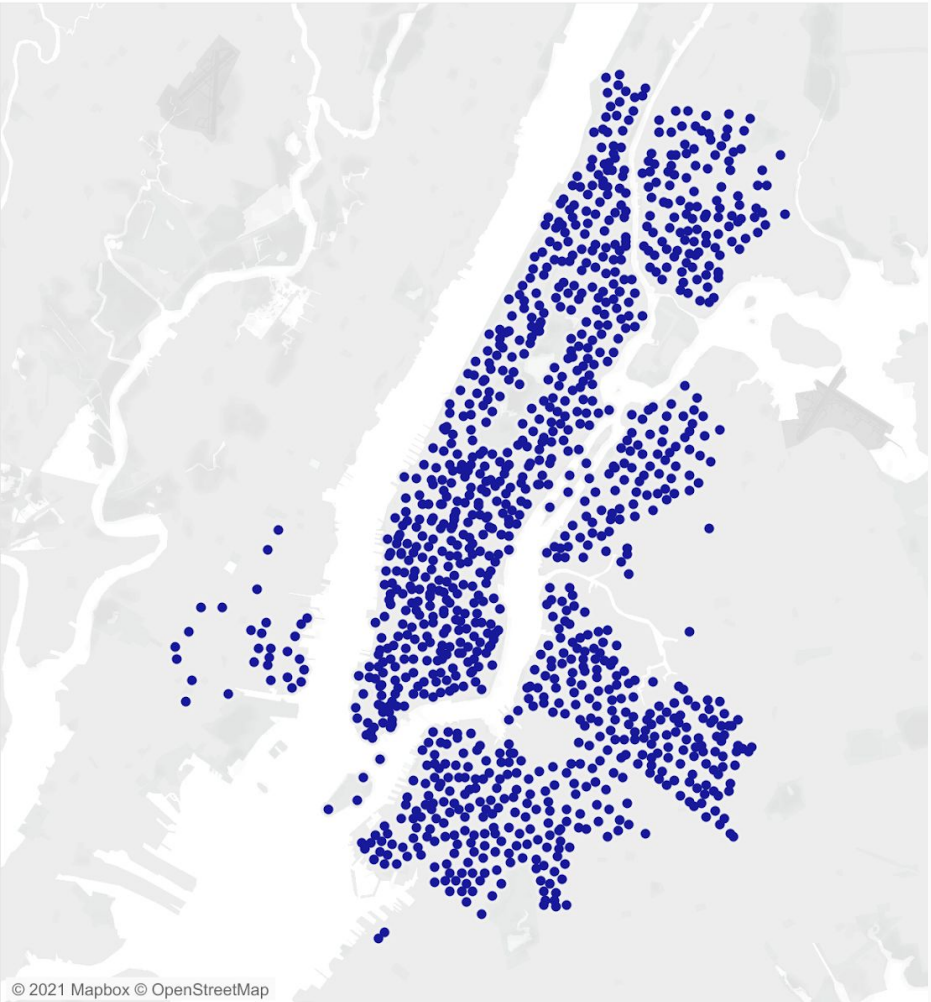
- Citibike Data - Summer 2020
  - Target - Usertype

- Features Anomalies
  - Outliers
  - Class Imbalance

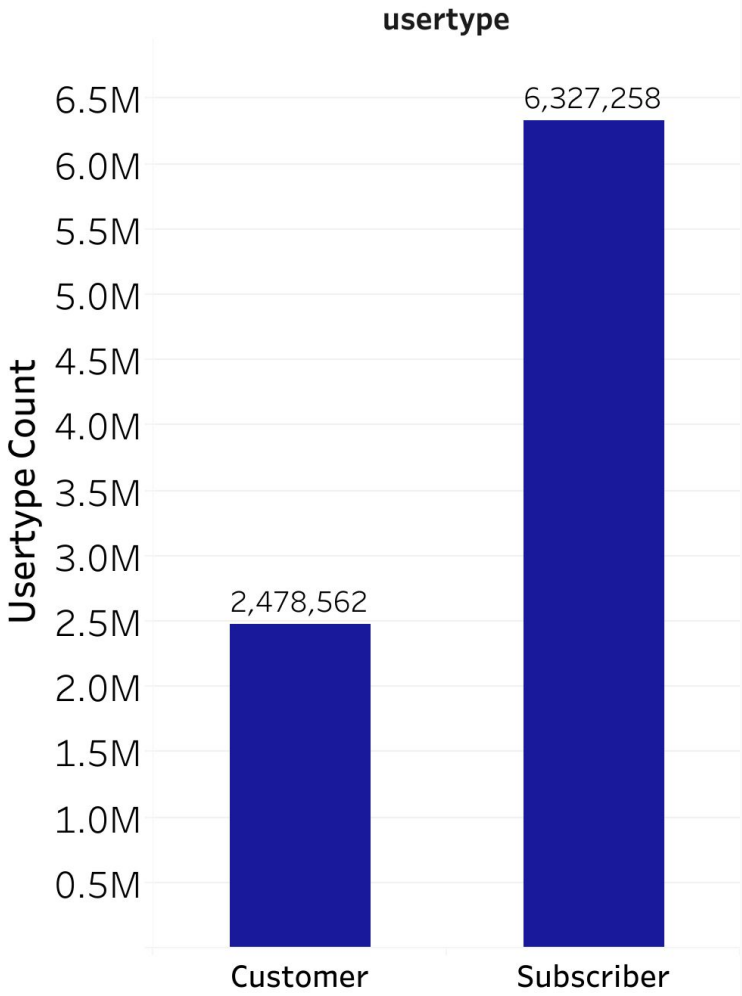
Features

- Trip duration
- Gender
- Month
- Day of the week
- Birth Year
- Day
- Starttime
- Stoptime
- End station latitude & longitude
- Start station latitude & longitude

Citibike Locations 09/2020



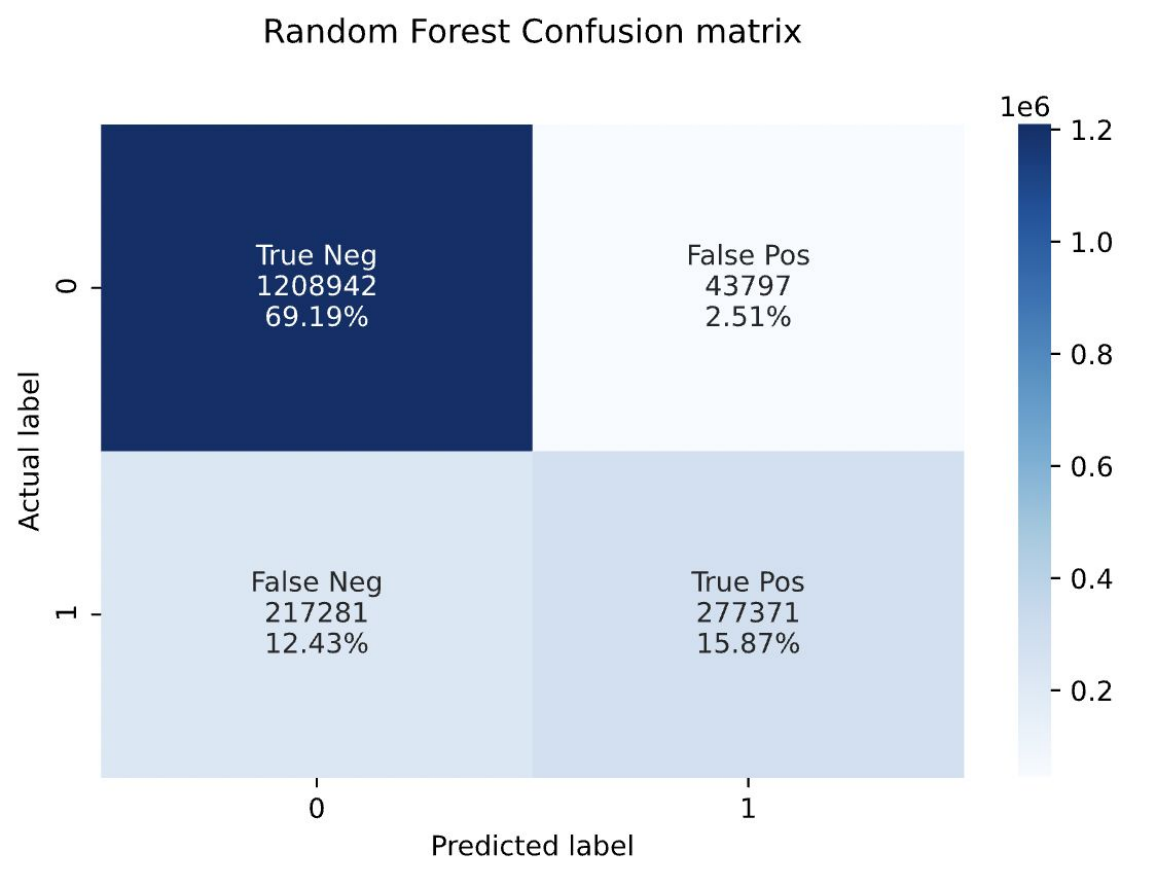
Class Imbalance of 1 : 3 for Usertype



Models & Metrics

Models Tested	Logistic Regression	Naive Bayes Gaussian	Random Forest	XGBoost*
F1 Score:	.523	.573	**.68**	.662

Models & Metrics



0 = Subscriber

1 = Customer

Models & Metrics

Models Tested	Logistic Regression	Naive Bayes Gaussian	Random Forest	XGBoost*
F1 Score:	.523	.573	.68	.662
Beta = 0.25				

Models & Metrics

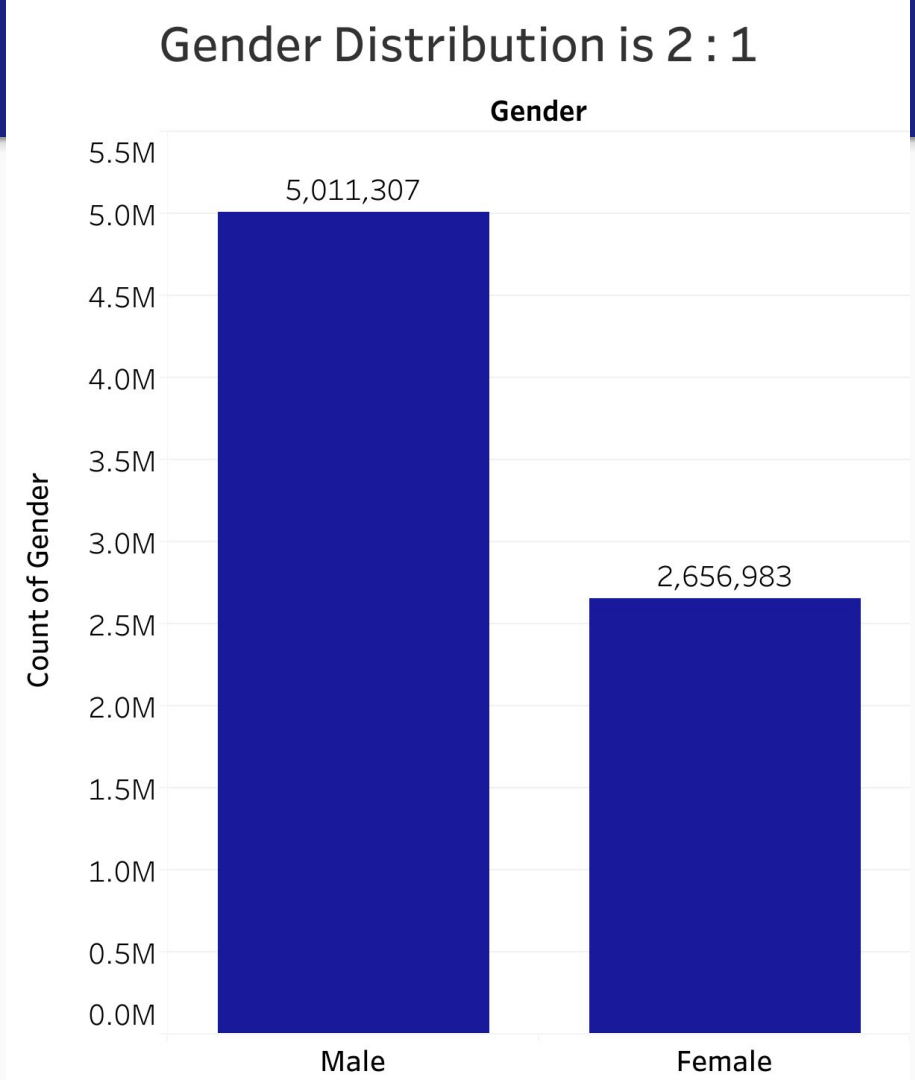
Models Tested	Logistic Regression	Naive Bayes Gaussian	Random Forest	XGBoost*
F1 Score:	.523	.573	.68	.662
Beta = 0.25				
Fbeta Score:	.757	.811	**.837**	.823

Next Steps

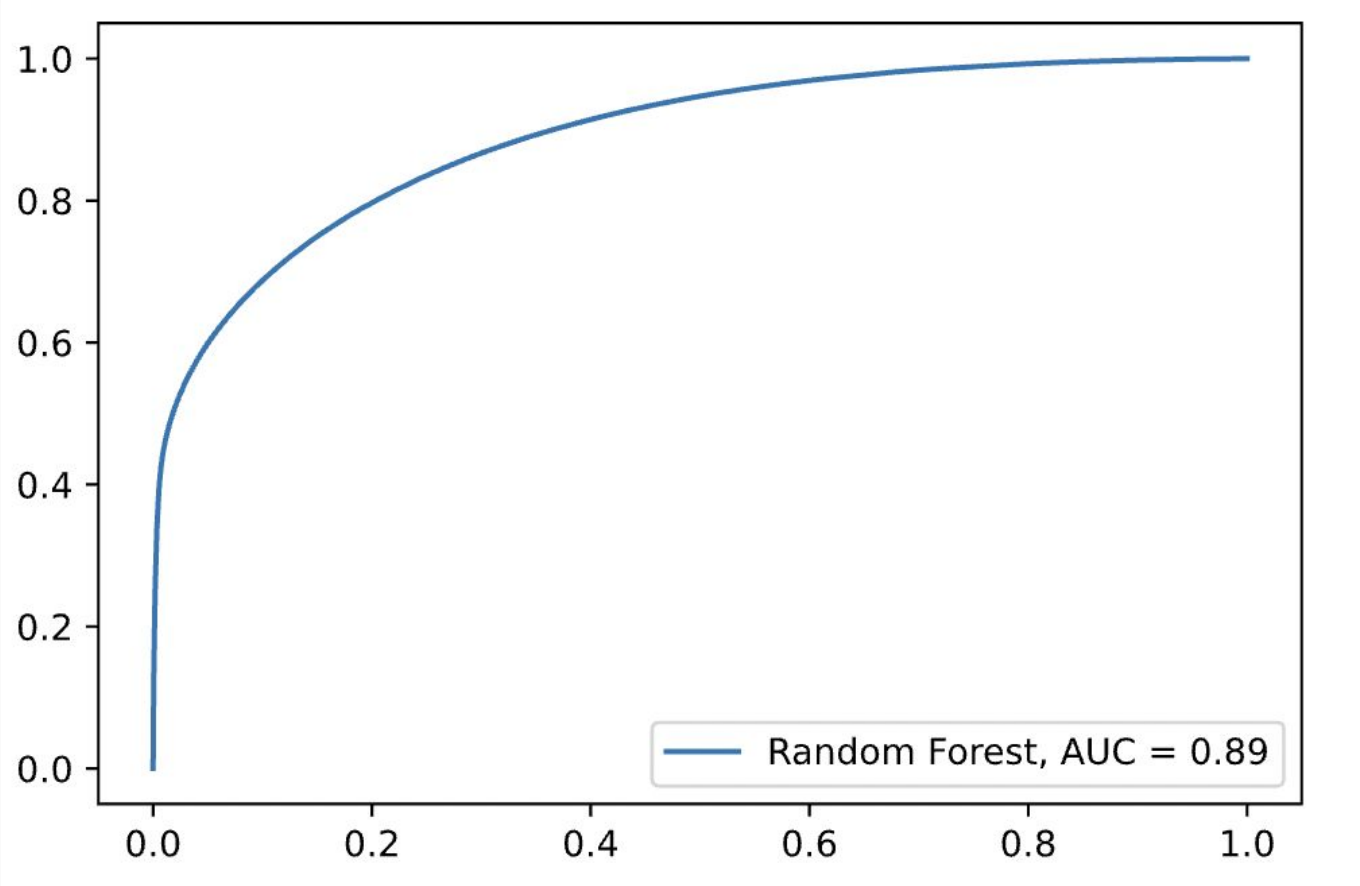
- Examine XGBoost - tune hyperparameters

- Undersample dataset

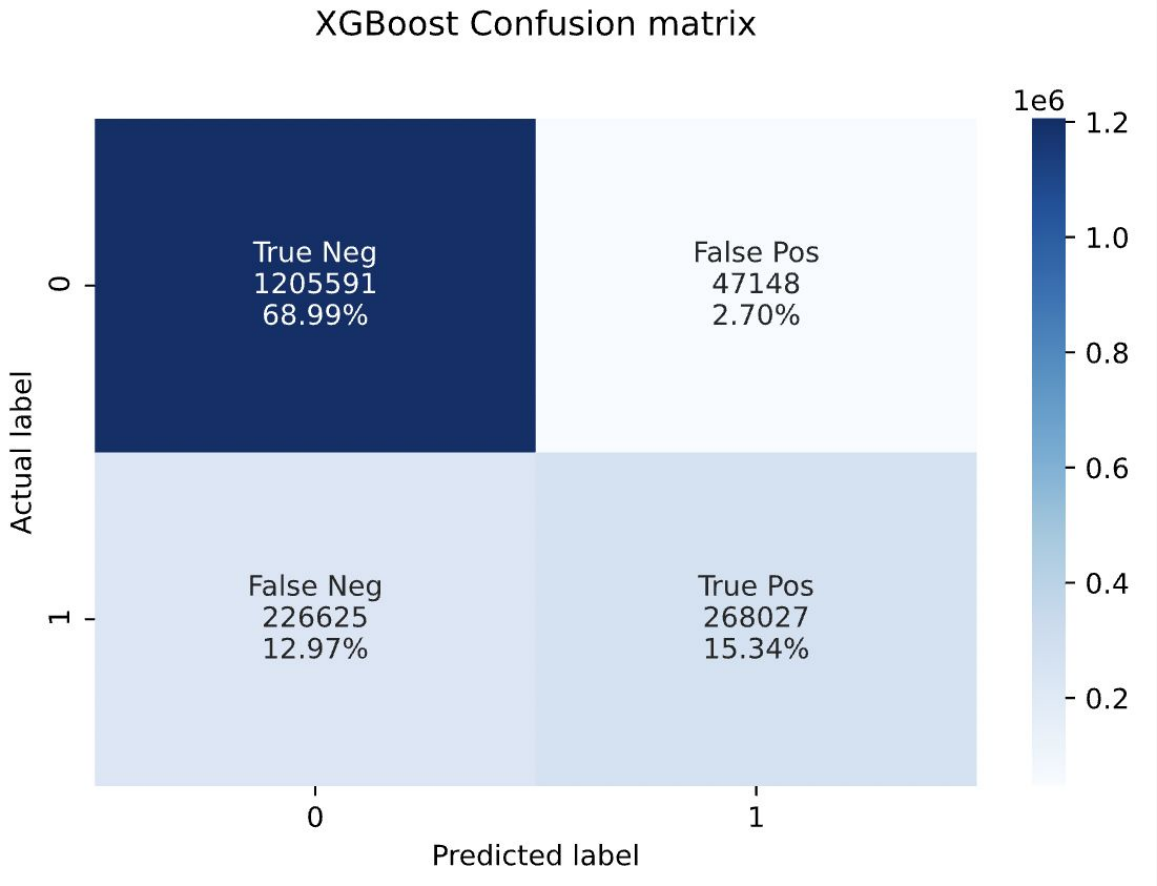
Appendix


Gender Distribution is 2 : 1
Gender
Count of Gender
5.5M
5.0M
4.5M
4.0M
3.5M
3.0M
2.5M
2.0M
1.5M
1.0M
0.5M
0.0M
5,011,307
2,656,983
Male
Female

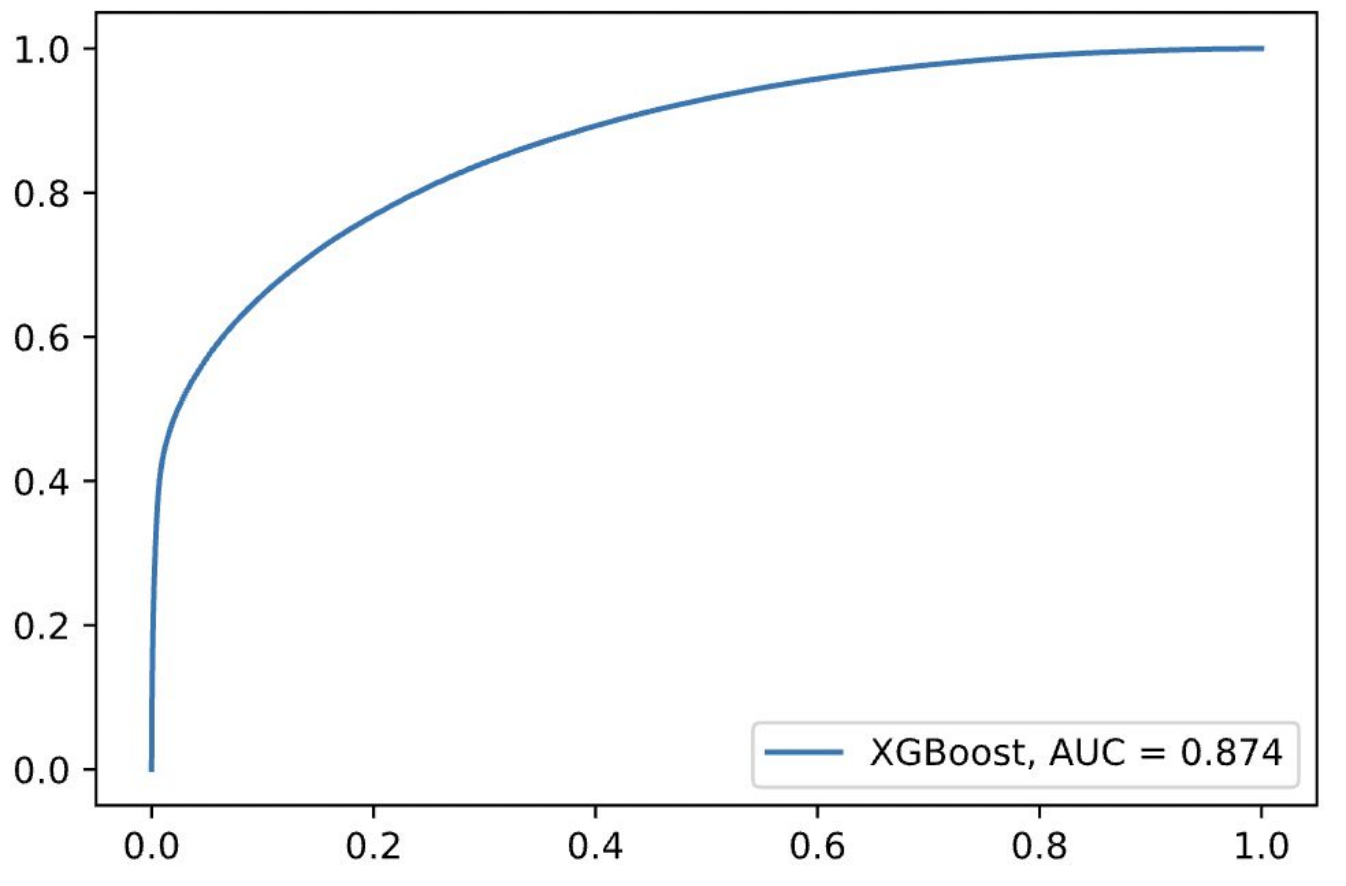
Appendix



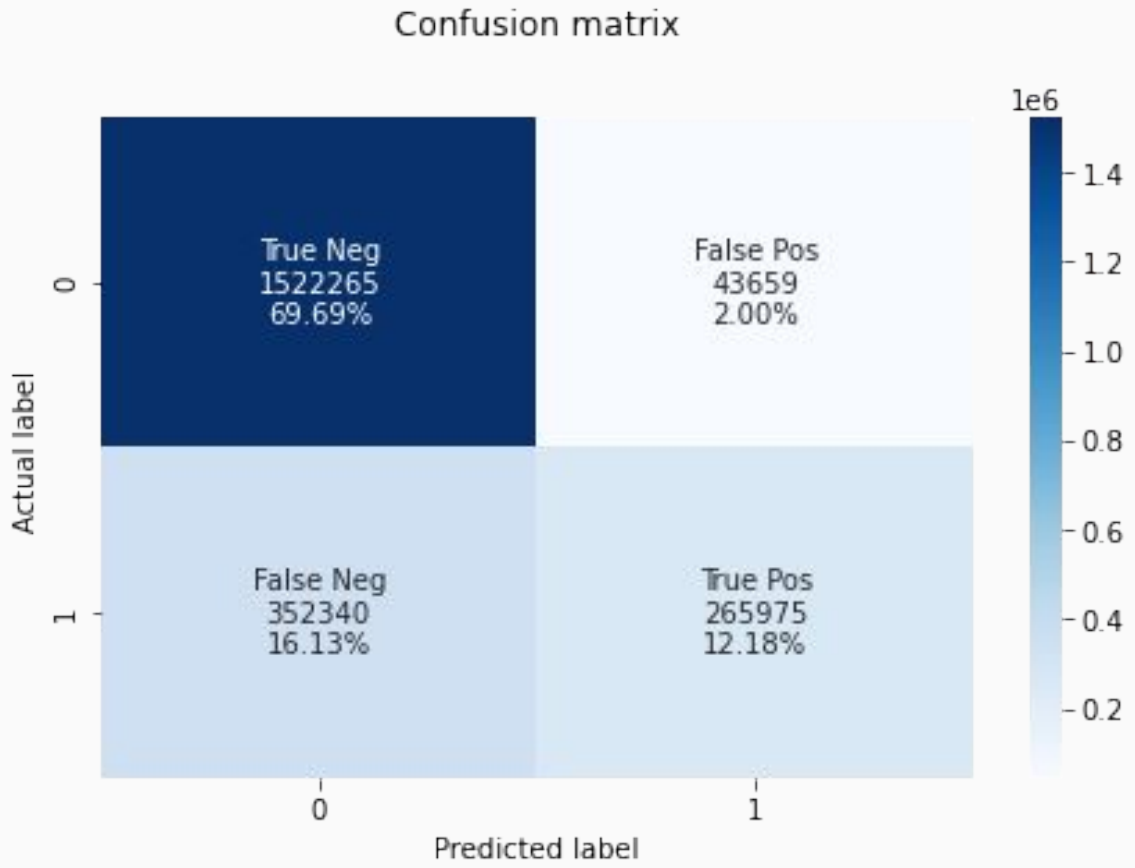
Appendix


XGBoost Confusion matrix
True Neg
1205591
68.99%
False Pos
47148
2.70%
False Neg
226625
12.97%
True Pos
268027
15.34%
Actual label
0
1
0
1
Predicted label
1e6
1.2
1.0
0.8
0.6
0.4
0.2

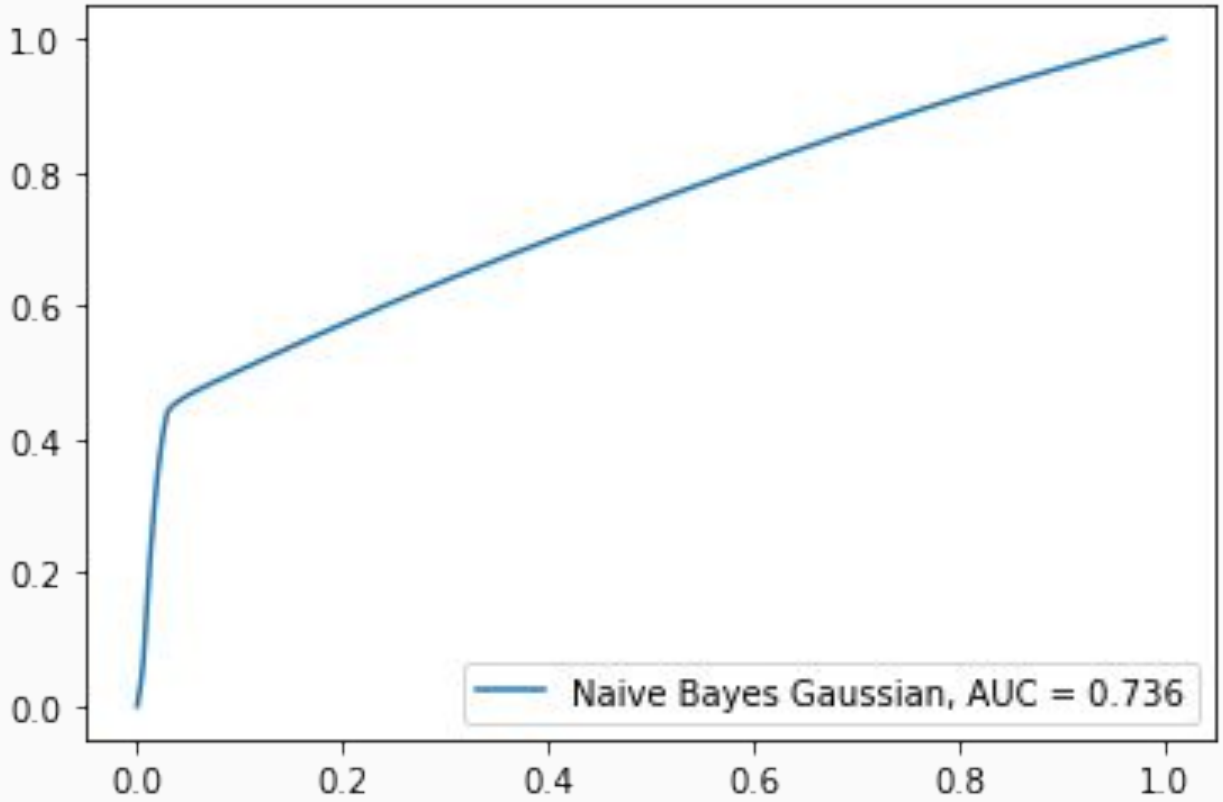
Appendix



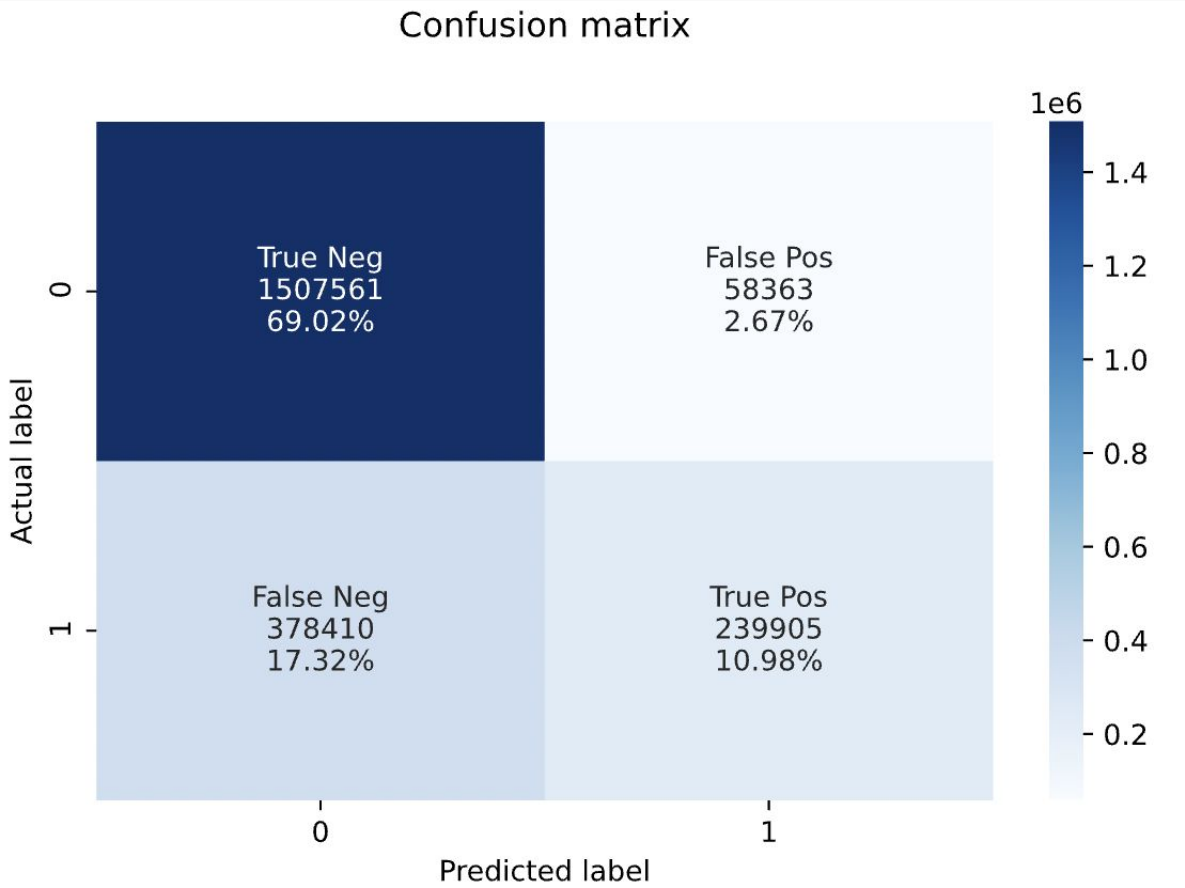
Appendix


Confusion matrix
True Neg
1522265
69.69%
False Pos
43659
2.00%
False Neg
352340
16.13%
True Pos
265975
12.18%
Actual label
0
1
Predicted label
0
1
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2

Appendix



Appendix


Confusion matrix
True Neg
1507561
69.02%
False Pos
58363
2.67%
False Neg
378410
17.32%
True Pos
239905
10.98%
Actual label
0
1
Predicted label
0
1
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2

Appendix