Mining Steam

WEIJIE WU, SEBASTIAN LINDNER

Goal

Using games informations that the user owned to infer the continent they belong to.

Data Summary

52665 random users

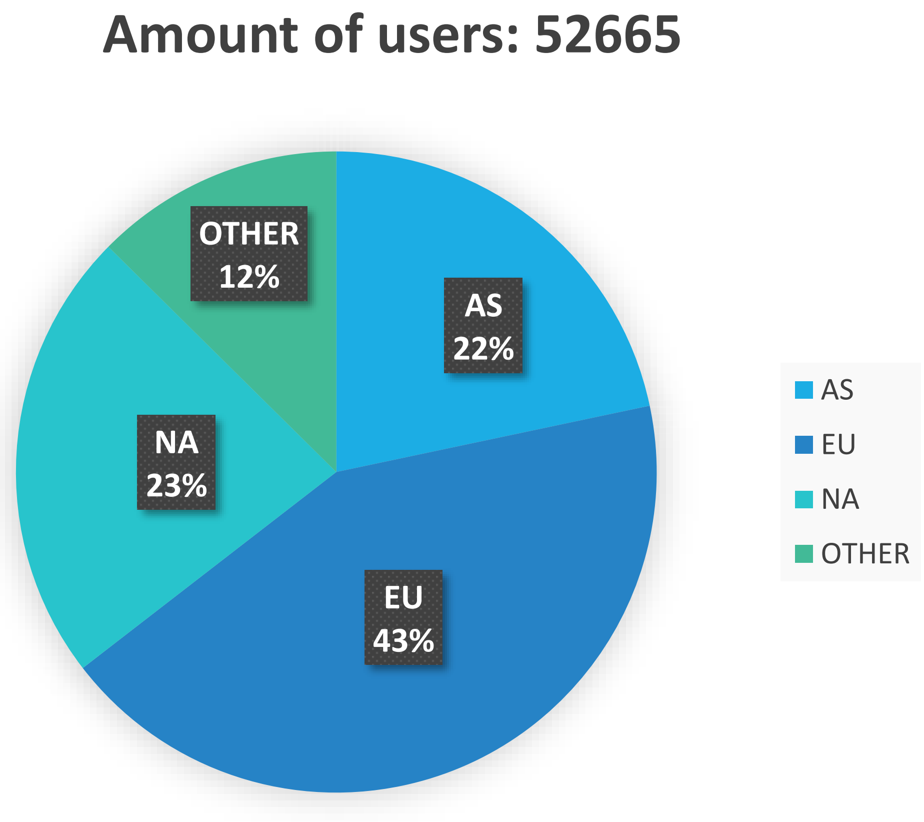
(>= 1 owned games):

AS: 21.68%, 11420

EU: 42.84%, 22564

NA: 22.87%, 12046

OTHER: 12.6%, 6635



Data Summary

In last presentation, we‘ve proved with statistical method to proved that, users‘ region is related to the following factor:

amount of games

playtime of games

which game they own

Data Structure

Combining this factors, we created the vector as following:

[game_amount, game1_playtime, game2_playtime, game3_playtime, ...]

(playtime of the 100 most popular games)

Define An Easy Classifier

Baseline: for those who has more than the average game amount of users from EU, we assume that they’re the users from EU.

Define An Easy Classifier

Result:

	predict->EU	predict->OTHER
real EU	5852 (TP)	16712 (FN)
real OTHER	7361 (FP)	22740 (TN)

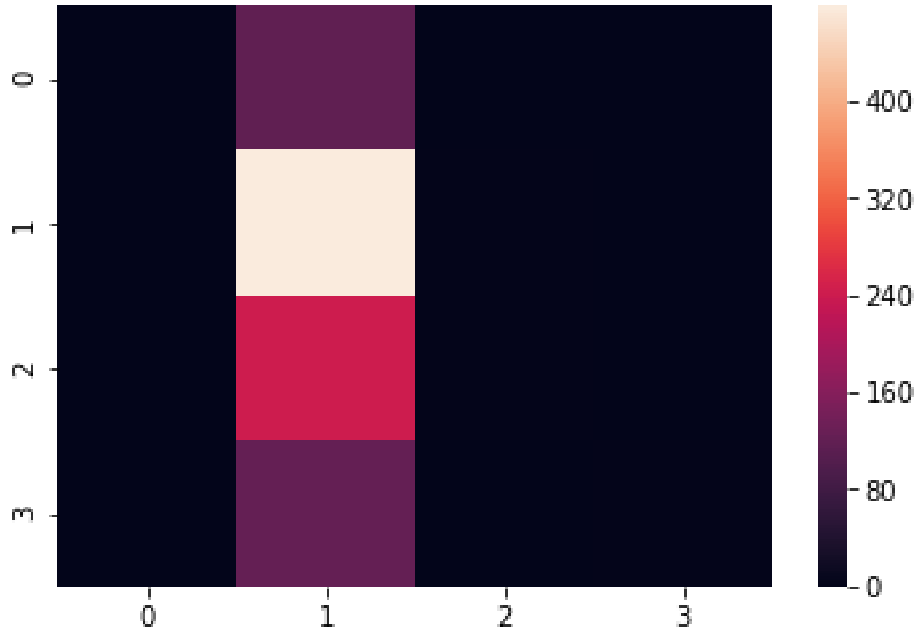
precision	recall	f1
44%	26%	33%

Machine Learning

- SVC, Adaboost
- PCA
- MinMaxScaler, GridSearch
- Adjusting dataset

Classifier-SVC

```
       precision recall f1-score
AS     0.50        0.01     0.02
EU     0.50        0.99     0.66
NA     0.40        0.01     0.02
OT     1.00        0.02     0.03
avg    0.54        0.50     0.34
```



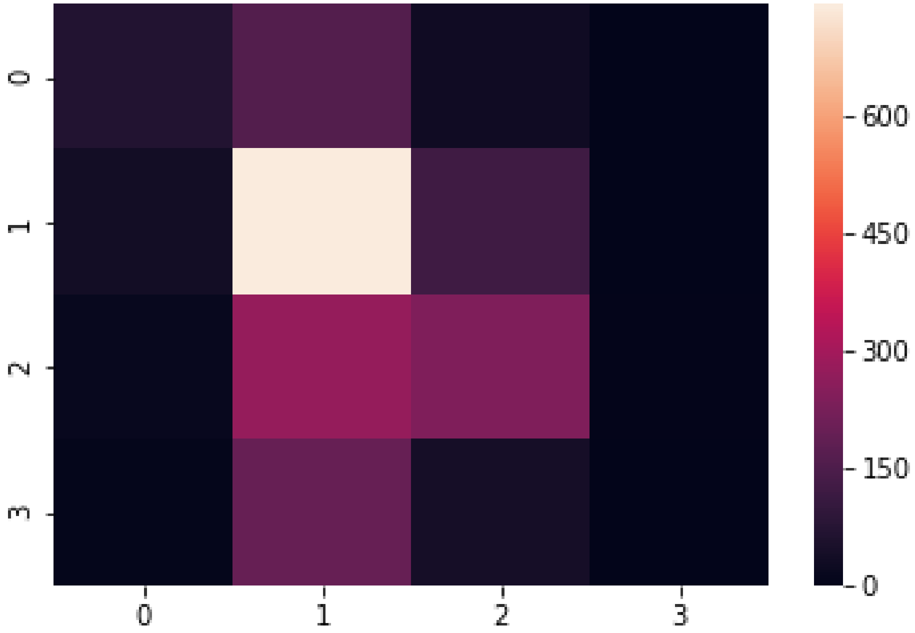
SVC

After using GridSearch and MinMaxScaler(), we can also get a better result from SVC.

```
      precision    recall  f1-score
AS       0.44       0.15       0.23
EU       0.53       0.88       0.66
NA       0.50       0.24       0.32
OT       0.08       0.00       0.01
avg      0.46       0.52       0.44
```

Classifier-AdaBoost

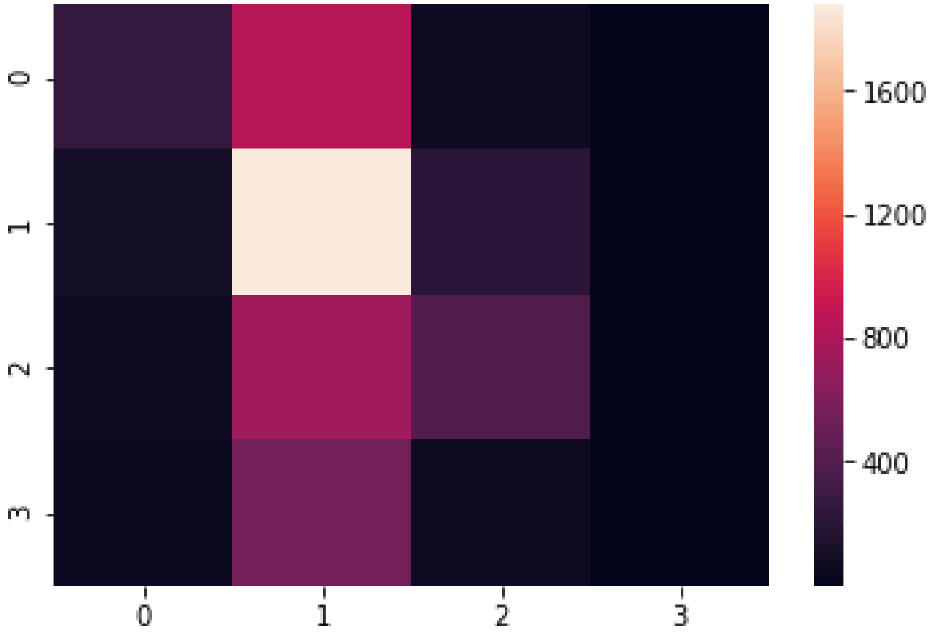
	precision	**recall**	**f1-score**
AS	0.54	0.27	0.36
EU	0.54	0.82	0.65
NA	0.55	0.45	0.49
OT	0.14	0.00	0.01
avg	0.49	0.54	0.49



PCA(Principal component analysis)

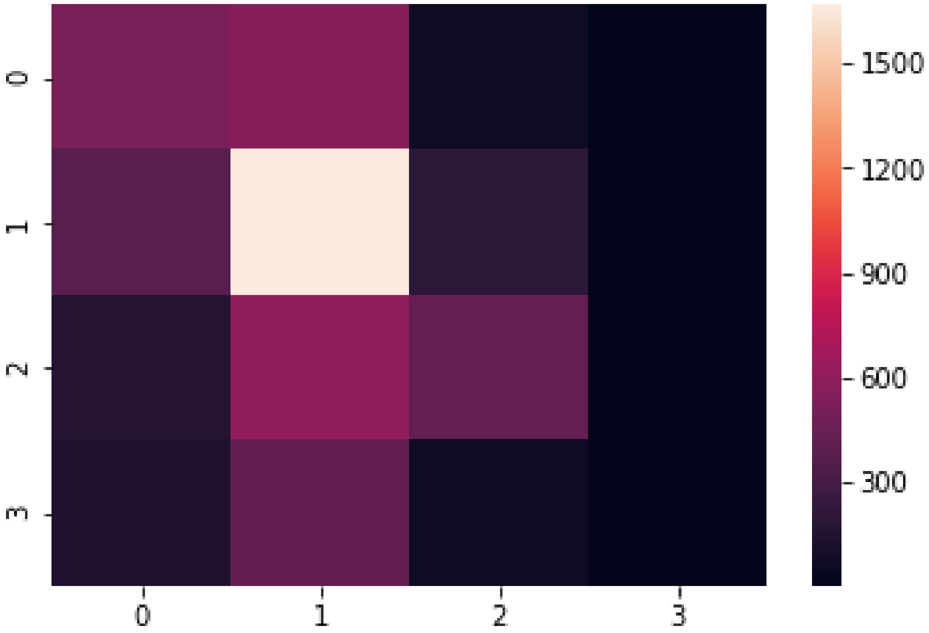
- Reduce noise
- Compress components(101 -> 90, still keep 99.7% information)

PCA(Principal component analysis)



without PCA

Avg-F1: 0.46



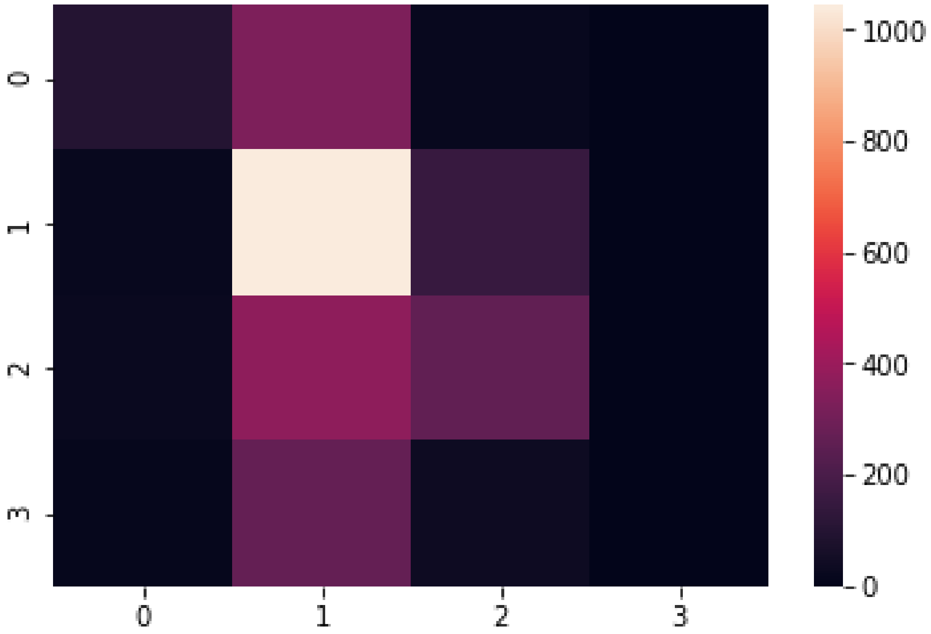
with PCA

Avg-F1: 0.49

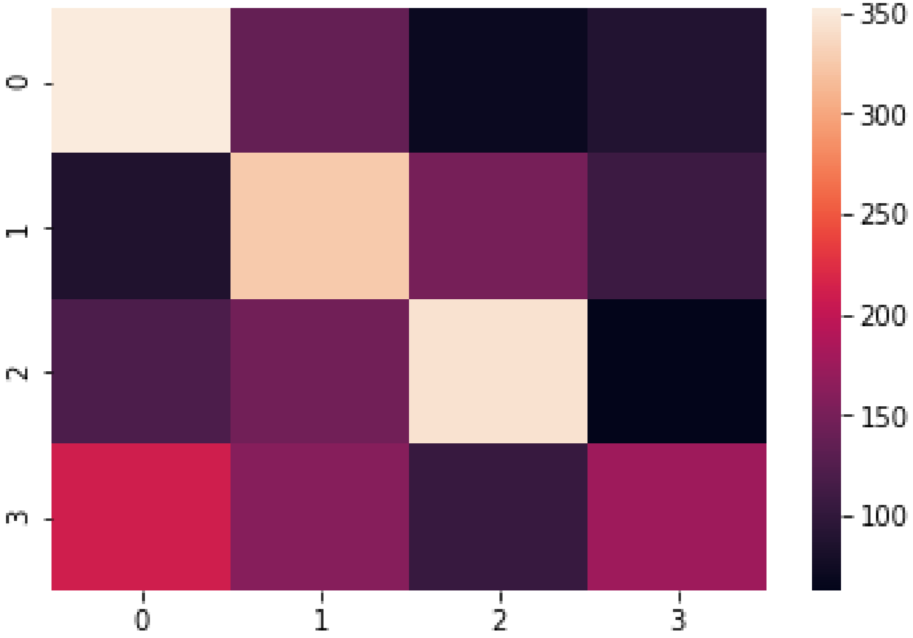
Adjusting Dataset

Almost half of the users are from Europe, so we tried to create a dataset, which the number of users from EU, AS, NA, OTHER are same.

Adjusting Dataset



Adaboost(normal-dataset)
Avg-F1: 0.47



Adaboost(equal-distributed-dataset)
Avg-F1: 0.46

Summary

- The users from OTHER-continent usually don’t have common games
- Only by using games’ information is still not enough to precisely infer the region of the users.