Homework 2
Data Mining Technology for Business and Society

Theodoros Sofianos (1867968)
Hassan Ismail (1735885)

Recommendation Systems Evaluation

Part 1.1

The results of applying all recommendation system algorithms provided by surprise library on the given dataset (taking 5folds):

1. **KNNBasic**

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9468	0.9517	0.9505	0.9530	0.9560	0.9516	0.0030
Fit time	0.76	1.00	1.25	1.23	0.98	1.04	0.18
Test time	10.84	12.48	12.12	9.83	8.08	10.67	1.60

2. **KNNWithMeans**

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9287	0.9347	0.9312	0.9331	0.9344	0.9324	0.0022
Fit time	0.86	1.09	1.33	1.35	1.11	1.15	0.18
Test time	12.41	14.03	13.38	10.84	8.81	11.89	1.88

3. **KNNWithZScore**

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9272	0.9338	0.9307	0.9322	0.9330	0.9314	0.0023
Fit time	0.96	1.24	1.59	1.60	1.21	1.32	0.24
Test time	13.61	16.14	15.73	12.84	10.45	13.76	2.07

4. KNNBaseline

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9051	0.9101	0.9065	0.9100	0.9107	0.9085	0.0022
Fit time	0.89	1.19	1.62	1.59	1.25	1.31	0.27
Test time	16.00	18.51	17.85	14.73	11.49	15.71	2.50

5. SVD

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9054	0.9075	0.9048	0.9102	0.9115	0.9079	0.0026
Fit time	17.98	21.73	22.35	19.34	15.18	19.32	2.60
Test time	1.09	1.12	0.81	0.62	0.56	0.84	0.23

6. SVDpp

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.8915	0.8941	0.8915	0.8972	0.8920	0.8933	0.0022
Fit time	1699.09	1709.61	1695.63	1708.35	907.65	1544.07	318.25
Test time	37.84	34.76	36.97	29.77	19.41	31.75	6.78

7. NMF

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9323	0.9394	0.9322	0.9398	0.9409	0.9369	0.0038
Fit time	18.21	21.95	22.03	18.69	15.24	19.22	2.55
Test time	1.02	0.80	0.66	0.48	0.46	0.69	0.21

8. SlopeOne

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9199	0.9253	0.9209	0.9236	0.9253	0.9230	0.0022
Fit time	3.61	4.79	5.57	8.01	6.12	5.62	1.46
Test time	27.61	29.61	29.57	22.17	15.38	24.87	5.46

9. CoClustering

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9322	0.9365	0.9257	0.9387	0.9424	0.9351	0.0058
Fit time	2.49	2.50	2.49	2.52	2.39	2.48	0.05
Test time	0.43	0.40	0.40	0.41	0.39	0.41	0.02

The algorithms which have a better RMSE are:

- “SVDpp” with a RMSE equal to 0. 8951.
- “SVD” with a RMSE equal to 0. 9084.
- “KNNBaseline” with a RMSE equal to 0.9085.

To use all cpu-cores we used the parameter ***n_jobs= 4*** in the cross_validate command. Where 4 is the number of our cpu core (we got it by command *cpu_count()*), we were able to use n_jobs = -1 as well.

Part 1.2

Performing hyper parameter tuning:

1. SVD:

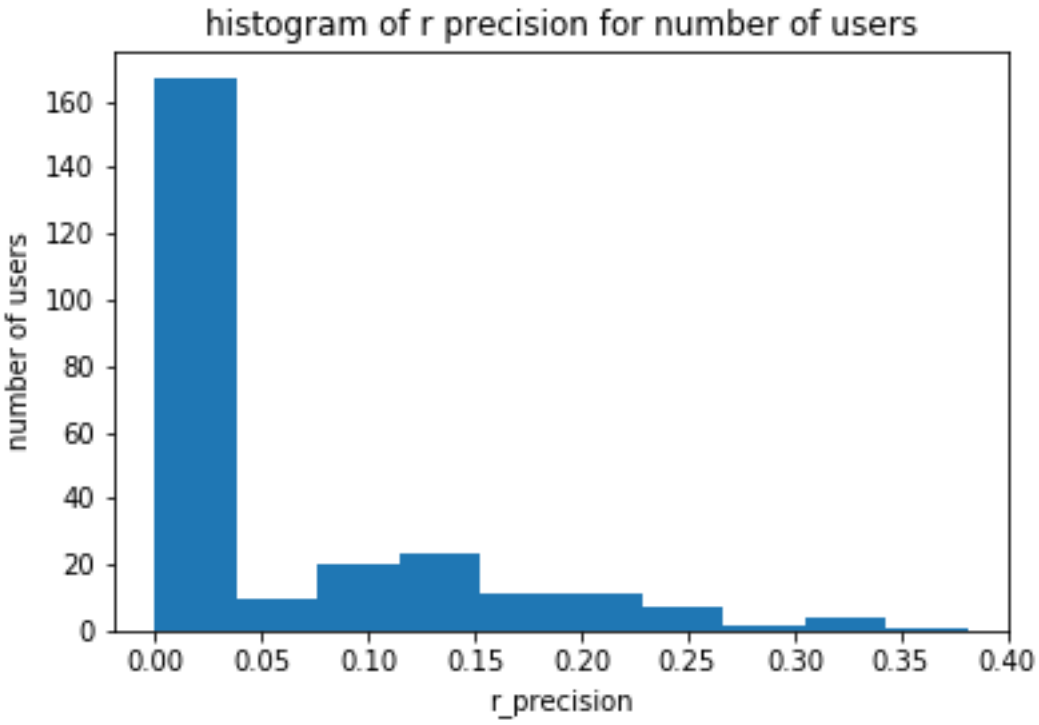
The grid search creates combinations from the below values of the parameters and apply the SVD algorithm and check the results:

- n_factors (number of factors): [50, 200, 700]
- n_epochs (number of iterations): [50, 70, 150]
- lr_all (learning rate for all the parameters): [0.003, 0.008, 0.01]
- reg_all (regularization term for all parameters): [0.0, 0.06, 0.08, 0.1]

And as result, the optimal configuration was:

- n_factors = 200
- n_epochs = 70
- lr_all = 0.008
- reg_all = 0.1

Part 2.1



Avg.r_precision=0.05350647946283566

```
Running time(from start till end of file) = 15.933274746 minutes
```

Part 2.2

Since its inefficient to store all the tuples of (item-probability) for all items for each and every user, we thought about taking into account only the items that have a correlation with other items above a threshold, which in our case is 0.000330402869469384 and we calculated it by taking (from part 2.1)the median probability of all the recommended items JUST for user1(1683). (the median of all recommendations, not the top-x of them). The idea was either to change this threshold for every user or to find the average median for all users, but for running time purposes we stick to the threshold of user1683, which seems to take enough items into account, even for other users.

```
Avg r_precision=0.21572458514843526

Running time(from start till end of file)= 18.762144681 minutes
```