# Report

PMDL Assignment 2

Ivan Inchin, i.inchin@innopolis.university

## Introduction

The project consists of a solution to the Movie Recommendation system problem. Using MovieLens 100K Dataset I've introduced 2 Machine Learning models SVD(*Singular value decomposition)* and KNN *(k Nearest Neighbor).* Models represent collaborative filtering approach for movie recommendations. For model evaluation I used RMSE (*Root-mean-square deviation)* metric and HitRate metric. As a result I've got the movie recommendation system that is available to recommend movies based on user's data and does it better than random recommendations.

## Data analysis

MovieLens 100K Dataset consists of 6 files:

- **u.data** - Ratings of movies by users
- **u.genre** - Binding of genre to int representations
- **u.info** - General information about dataset
- **u.item** - Information about movies
- **u.user** - Information about user
- **u.occupation** - List of users' occupations

u.data, u.user, u.item files were transformed into pandas datasets, get column names and merged into one big dataset with fields:

> 'uid', 'iid', 'rating', 'timestamp', 'movie_name', 'date', 'url', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', '15', '16', '17','18', '19', '20', '21', '22', '23', 'age', 'gender', 'occupation','zip'

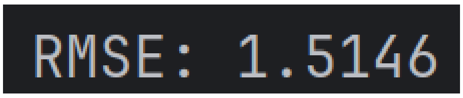
Numbers in naming are one-hotted genres names and mapping to **u.genre** int representations is saved with *i - 5* shift, where i is the column name.

## Model Implementation

For creating Movie Recommendation System I've used a Collaborative Filtering approach and implemented two models SVG and KNN. Their implementations can be found in model_initialisation.ipynb notebook. In the same notebook functions for calculating RMSE and HitRate can be found. Both models were created as an experiment to see which one will show better results.

## Model Advantages and Disadvantages

KNN and SVD models showed better results than just random generation of recommendation. Results of random recommendations in terms of RMSE can be seen in  Fig 1.



```
RMSE: 1.5146
```

Fig 1. RMSE for random recommendation

Both models showed 1,6x performance in terms of RMSE score.

The disadvantage of the models is that they predict movies according to similarity, but rate them by movie rating, that means that the similar movie with bad rating won’t be in the recommendations.
Also KNN model try to suggest movies according to their rating and may skip the fact about genre similarity.

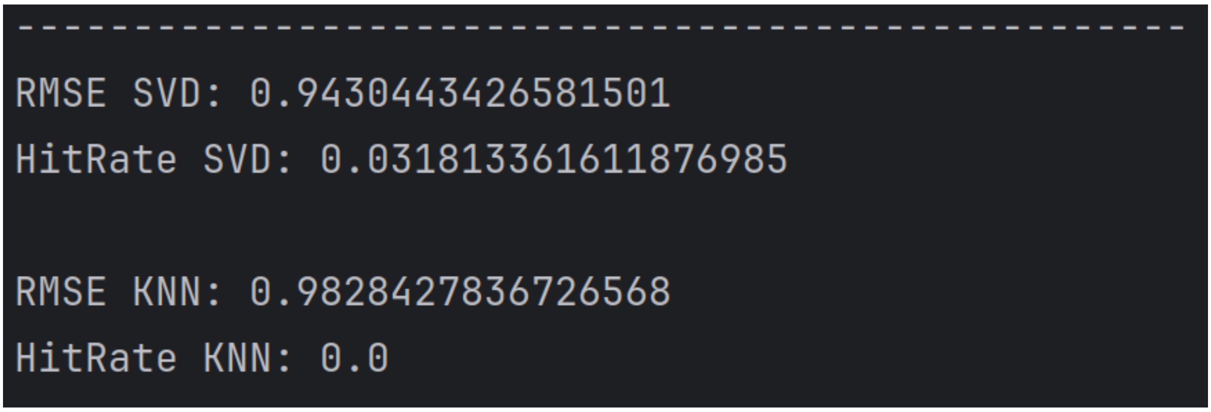
## Training Process

KNN is an unsupervised learning method, so to predict the best suitable movies, the whole dataset is marked during the training process.

SVD is an unsupervised learning method too and its training consists of matrix decomposition of the whole dataset.

## Evaluation

For evaluation I used two metrics: RMSE and HitRate. RMSE shows how far is the current predicted movie from the initial one. HitRate counts how many predicted movies for the users are in his top 10 out of his total movies. evaluate.py is used for evaluating KNN and SVD models. For evaluation in these metrics, SVD shows better results than KNN model. In Fig 2. you can see the results of evaluation.



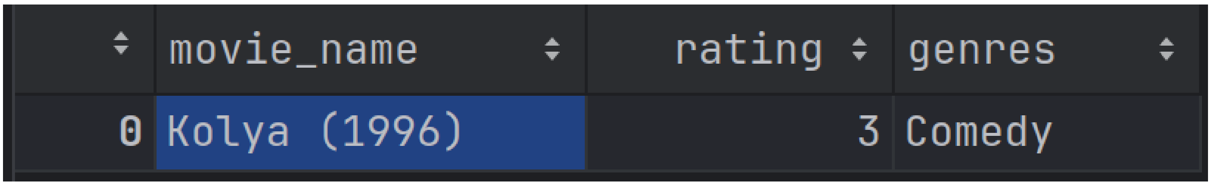
```
--------------------------------------------------

RMSE SVD: 0.9430443426581501
HitRate SVD: 0.031813361611876985


RMSE KNN: 0.9828427836726568
HitRate KNN: 0.0
```

Fig 2. Results of evaluations of SVG and KNN models

# Results

As a result, the SVD model showed the best result in the both RMSE and HitRate metric by 0.039 and 0.032 points. Both models did better than random recommendations by 0,5716 and 0,5322 points accordingly.

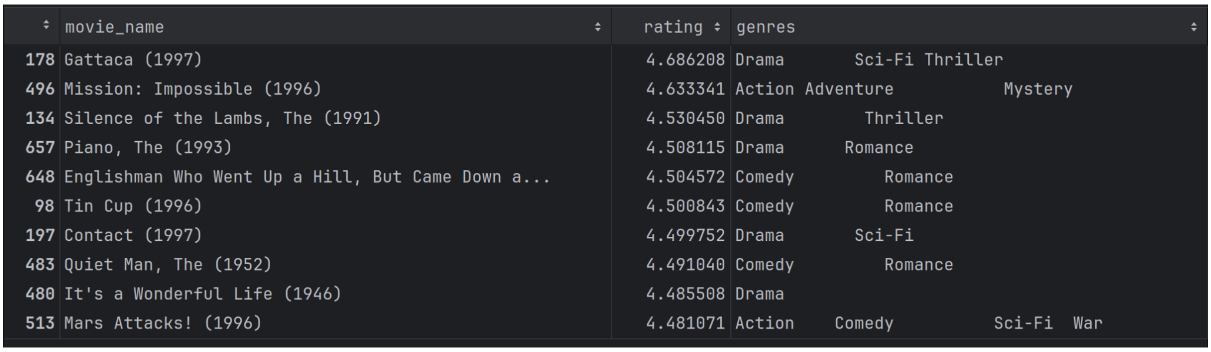
As an example of recommendations, let’s see the recommendation for a user with *id 196* for the film with *id* 242 and *name Kolya (1996).* Initial info about the Kolya movie is shown in Fig 3.



| | movie_name | rating | genres |
|---|---|---|---|
| 0 | Kolya (1996) | 3 | Comedy |

Fig 3. Basic info about movie

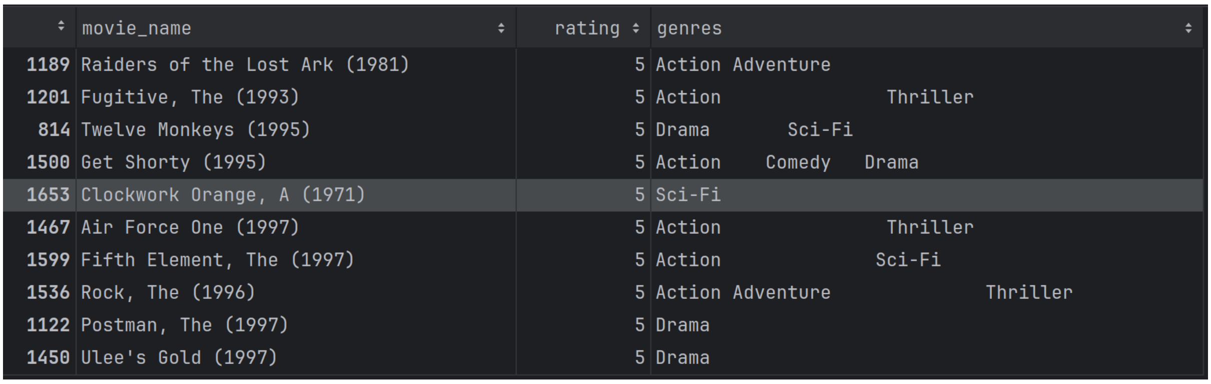
In Fig 4. We can see that SVD model gave us movies with average rating 4.5  and even tried to match Comedy genre for recommendations.



| | movie_name | rating | genres |
|---|---|---|---|
| 178 | Gattaca (1997) | 4.686208 | Drama Sci-Fi Thriller |
| 496 | Mission: Impossible (1996) | 4.633341 | Action Adventure Mystery |
| 134 | Silence of the Lambs, The (1991) | 4.530450 | Drama Thriller |
| 657 | Piano, The (1993) | 4.508115 | Drama Romance |
| 648 | Englishman Who Went Up a Hill, But Came Down a... | 4.504572 | Comedy Romance |
| 98 | Tin Cup (1996) | 4.500843 | Comedy Romance |
| 197 | Contact (1997) | 4.499752 | Drama Sci-Fi |
| 483 | Quiet Man, The (1952) | 4.491040 | Comedy Romance |
| 480 | It's a Wonderful Life (1946) | 4.485508 | Drama |
| 513 | Mars Attacks! (1996) | 4.481071 | Action Comedy Sci-Fi War |

Fig 4. Recommendations by SVD

Fig 5. shows that KNN just gave for that particular example movies with the maximum rating. Only one film has genre Comedy *(Get Shorty),* and there are three films with genre Drama. and three with Thiller.



| | movie_name | rating | genres |
|---|---|---|---|
| 1189 | Raiders of the Lost Ark (1981) | 5 | Action Adventure |
| 1201 | Fugitive, The (1993) | 5 | Action Thriller |
| 814 | Twelve Monkeys (1995) | 5 | Drama Sci-Fi |
| 1500 | Get Shorty (1995) | 5 | Action Comedy Drama |
| 1653 | Clockwork Orange, A (1971) | 5 | Sci-Fi |
| 1467 | Air Force One (1997) | 5 | Action Thriller |
| 1599 | Fifth Element, The (1997) | 5 | Action Sci-Fi |
| 1536 | Rock, The (1996) | 5 | Action Adventure Thriller |
| 1122 | Postman, The (1997) | 5 | Drama |
| 1450 | Ulee's Gold (1997) | 5 | Drama |

Fig 5. Recommendations by KNN

I can conclude that KNN shows worse results than SVD according to RMSE and HitRate metrics and an example for a particular user and film showed it.

Overall SVD-based recommendation system can be used to suggest next movies to watch.