



1. Business Problem

1.1 Problem Description

Netflix is all about connecting people to the movies they love. To help customers find those movies, they developed world-class movie recommendation system: CinematchSM. Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. Netflix use those predictions to make personal movie recommendations based on each customer's unique tastes. And while **Cinematch** is doing pretty well, it can always be made better.

Now there are a lot of interesting alternative approaches to how Cinematch works that netflix haven't tried. Some are described in the literature, some aren't. We're curious whether any of these can beat Cinematch by making better predictions. Because, frankly, if there is a much better approach it could make a big difference to our customers and our business.

Credits: <https://www.netflixprize.com/rules.html>

1.2 Problem Statement

Netflix provided a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set. (Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.)

1.3 Sources

- <https://www.netflixprize.com/rules.html>
- <https://www.kaggle.com/netflix-inc/netflix-prize-data>
- Netflix blog: <https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429> (very nice blog)
- surprise library: <http://surpriselib.com/> (we use many models from this library)
- surprise library doc: http://surprise.readthedocs.io/en/stable/getting_started.html (we use many models from this library)
- installing surprise: <https://github.com/NicolasHug/Surprise#installation>
- Research paper: <http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/a1-koren.pdf> (most of our work was inspired by this paper)
- SVD Decomposition : <https://www.youtube.com/watch?v=P5mlg91as1c>

1.4 Real world/Business Objectives and constraints

Objectives:

1. Predict the rating that a user would give to a movie that he has not yet rated.
2. Minimize the difference between predicted and actual rating (RMSE and MAPE)

Constraints:

1. Some form of interpretability.

2. Machine Learning Problem

2.1 Data

2.1.1 Data Overview

Get the data from : <https://www.kaggle.com/netflix-inc/netflix-prize-data/data>

Data files :

- combined_data_1.txt
- combined_data_2.txt
- combined_data_3.txt
- combined_data_4.txt
- movie_titles.csv

The first line of each file [combined_data_1.txt, combined_data_2.txt, combined_data_3.txt, combined_data_4.txt] contains the movie id followed by a colon. Each subsequent line in the file corresponds to a rating from a customer and its date in the following format:

CustomerID,Rating,Date

MovieIDs range from 1 to 17770 sequentially.

CustomerIDs range from 1 to 2649429, with gaps. There are 480189 users.

Ratings are on a five star (integral) scale from 1 to 5.

Dates have the format YYYY-MM-DD.

2.1.2 Example Data point

1:

1488844,3,2005-09-06

822109,5,2005-05-13

885013,4,2005-10-19

30878,4,2005-12-26

823519,3,2004-05-03

893988,3,2005-11-17

124105,4,2004-08-05
1248029,3,2004-04-22
1842128,4,2004-05-09
2238063,3,2005-05-11
1503895,4,2005-05-19
2207774,5,2005-06-06
2590061,3,2004-08-12
2442,3,2004-04-14
543865,4,2004-05-28
1209119,4,2004-03-23
804919,4,2004-06-10
1086807,3,2004-12-28
1711859,4,2005-05-08
372233,5,2005-11-23
1080361,3,2005-03-28
1245640,3,2005-12-19
558634,4,2004-12-14
2165002,4,2004-04-06
1181550,3,2004-02-01
1227322,4,2004-02-06
427928,4,2004-02-26
814701,5,2005-09-29
808731,4,2005-10-31
662870,5,2005-08-24
337541,5,2005-03-23
786312,3,2004-11-16
1133214,4,2004-03-07
1537427,4,2004-03-29
1209954,5,2005-05-09
2381599,3,2005-09-12
525356,2,2004-07-11
1910569,4,2004-04-12
2263586,4,2004-08-20
2421815,2,2004-02-26
1009622,1,2005-01-19
1481961,2,2005-05-24
401047,4,2005-06-03
2179073,3,2004-08-29
1434636,3,2004-05-01
93986,5,2005-10-06
1308744,5,2005-10-29
2647871,4,2005-12-30
1905581,5,2005-08-16
2508819,3,2004-05-18
1578279,1,2005-05-19
1159695,4,2005-02-15
2588432,3,2005-03-31
2423091,3,2005-09-12
470232,4,2004-04-08

```
2148699,2,2004-06-05
1342007,3,2004-07-16
466135,4,2004-07-13
2472440,3,2005-08-13
1283744,3,2004-04-17
1927580,4,2004-11-08
716874,5,2005-05-06
4326,4,2005-10-29
```

2.2 Mapping the real world problem to a Machine Learning Problem

2.2.1 Type of Machine Learning Problem

For a given movie and user we need to predict the rating would be given by him/her to the movie.

The given problem is a Recommendation problem

It can also be seen as a Regression problem

2.2.2 Performance metric

- Mean Absolute Percentage Error: https://en.wikipedia.org/wiki/Mean_absolute_percentage_error
- Root Mean Square Error: https://en.wikipedia.org/wiki/Root-mean-square_deviation

2.2.3 Machine Learning Objective and Constraints

1. Minimize RMSE.
2. Try to provide some interpretability.

```
1 # this is just to know how much time will it take to run this entire ipython notebook
2 from datetime import datetime
3 # globalstart = datetime.now()
4 import pandas as pd
5 import numpy as np
6 import matplotlib
7 matplotlib.use('nbagg')
8
9 import matplotlib.pyplot as plt
10 plt.rcParams.update({'figure.max_open_warning': 0})
11
12 import seaborn as sns
13 sns.set_style('whitegrid')
14 import os
15 from scipy import sparse
16 from scipy.sparse import csr_matrix
17
18 from sklearn.decomposition import TruncatedSVD
19 from sklearn.metrics.pairwise import cosine_similarity
20 import random
```

Double-click (or enter) to edit

3. Exploratory Data Analysis

3.1 Preprocessing

3.1.1 Converting / Merging whole data to required format: u_i, m_j, r_ij

```

1 start = datetime.now()
2 if not os.path.isfile('data.csv'):
3     # Create a file 'data.csv' before reading it
4     # Read all the files in netflix and store them in one big file('data.csv')
5     # We re reading from each of the four files and appendig each rating to a global f
6     data = open('data.csv', mode='w')
7
8     row = list()
9     files=['data_folder/combined_data_1.txt','data_folder/combined_data_2.txt',
10           'data_folder/combined_data_3.txt', 'data_folder/combined_data_4.txt']
11     for file in files:
12         print("Reading ratings from {}".format(file))
13         with open(file) as f:
14             for line in f:
15                 del row[:] # you don't have to do this.
16                 line = line.strip()
17                 if line.endswith(':'):
18                     # All below are ratings for this movie, until another movie appear
19                     movie_id = line.replace(':', '')
20                 else:
21                     row = [x for x in line.split(',')]
22                     row.insert(0, movie_id)
23                     data.write(','.join(row))
24                     data.write('\n')
25         print("Done.\n")
26     data.close()
27 print('Time taken :', datetime.now() - start)

```



Reading ratings from data_folder/combined_data_1.txt...
Done.

Reading ratings from data_folder/combined_data_2.txt...
Done.

Reading ratings from data_folder/combined_data_3.txt...
Done.

Reading ratings from data_folder/combined_data_4.txt...
Done.

Time taken : 0:05:03.705966

Double-click (or enter) to edit

```

1 print("creating the dataframe from data.csv file..")
2 df = pd.read_csv('data.csv', sep=',',
3                 names=['movie', 'user', 'rating', 'date'])
4 df.date = pd.to_datetime(df.date)
5 print('Done.\n')
6
7 # we are arranging the ratings according to time.
8 print('Sorting the dataframe by date..')

```

```

9 df.sort_values(by='date', inplace=True)
10 print('Done..')

```

creating the dataframe from data.csv file..
Done.

Sorting the dataframe by date..
Done..

```
1 df.head()
```

| | movie | user | rating | date |
|-----------------|-------|--------|--------|------------|
| 56431994 | 10341 | 510180 | 4 | 1999-11-11 |
| 9056171 | 1798 | 510180 | 5 | 1999-11-11 |
| 58698779 | 10774 | 510180 | 3 | 1999-11-11 |
| 48101611 | 8651 | 510180 | 2 | 1999-11-11 |
| 81893208 | 14660 | 510180 | 2 | 1999-11-11 |

```
1 df.describe()['rating']
```

| | |
|------------------------------|--------------|
| count | 1.004805e+08 |
| mean | 3.604290e+00 |
| std | 1.085219e+00 |
| min | 1.000000e+00 |
| 25% | 3.000000e+00 |
| 50% | 4.000000e+00 |
| 75% | 4.000000e+00 |
| max | 5.000000e+00 |
| Name: rating, dtype: float64 | |

3.1.2 Checking for NaN values

```

1 # just to make sure that all Nan containing rows are deleted..
2 print("No of Nan values in our dataframe : ", sum(df.isnull().any()))

```

No of Nan values in our dataframe : 0

Double-click (or enter) to edit

3.1.3 Removing Duplicates

```

1 dup_bool = df.duplicated(['movie','user','rating'])
2 dups = sum(dup_bool) # by considering all columns..( including timestamp)
3 print("There are {} duplicate rating entries in the data..".format(dups))

```

There are 0 duplicate rating entries in the data..

Double-click (or enter) to edit

3.1.4 Basic Statistics (#Ratings, #Users, and #Movies)

```

1 print("Total data ")
2 print("-"*50)
3 print("\nTotal no of ratings :",df.shape[0])
4 print("Total No of Users      :", len(np.unique(df.user)))
5 print("Total No of movies     :", len(np.unique(df.movie)))

```



Total data

```

-----

Total no of ratings : 100480507
Total No of Users   : 480189
Total No of movies  : 17770

```

3.2 Splitting data into Train and Test(80:20)

```

1 if not os.path.isfile('train.csv'):
2     # create the dataframe and store it in the disk for offline purposes..
3     df.iloc[:int(df.shape[0]*0.80)].to_csv("train.csv", index=False)
4
5 if not os.path.isfile('test.csv'):
6     # create the dataframe and store it in the disk for offline purposes..
7     df.iloc[int(df.shape[0]*0.80):].to_csv("test.csv", index=False)
8
9 train_df = pd.read_csv("train.csv", parse_dates=['date'])
10 test_df = pd.read_csv("test.csv")

```

3.2.1 Basic Statistics in Train data (#Ratings, #Users, and #Movies)

```

1 # movies = train_df.movie.value_counts()
2 # users = train_df.user.value_counts()
3 print("Training data ")
4 print("-"*50)
5 print("\nTotal no of ratings :",train_df.shape[0])
6 print("Total No of Users      :", len(np.unique(train_df.user)))
7 print("Total No of movies     :", len(np.unique(train_df.movie)))

```



Training data

```

-----

Total no of ratings : 80384405
Total No of Users   : 405041
Total No of movies  : 17424

```

3.2.2 Basic Statistics in Test data (#Ratings, #Users, and #Movies)

```

1 print("Test data ")
2 print("-"*50)
3 print("\nTotal no of ratings :",test_df.shape[0])
4 print("Total No of Users      :", len(np.unique(test_df.user)))
5 print("Total No of movies     :", len(np.unique(test_df.movie)))

```



Test data

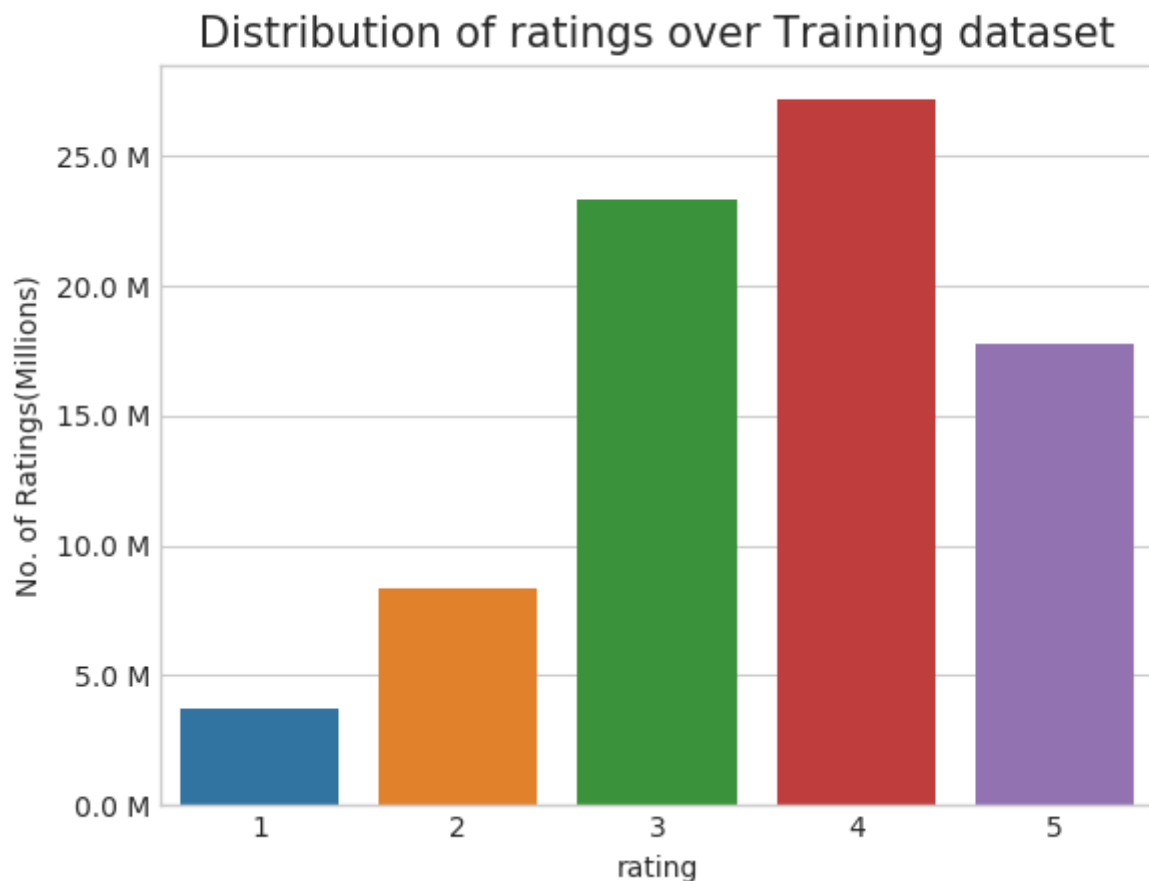
3.3 Exploratory Data Analysis on Train data

Double-click (or enter) to edit

```
1 # method to make y-axis more readable
2 def human(num, units = 'M'):
3     units = units.lower()
4     num = float(num)
5     if units == 'k':
6         return str(num/10**3) + " K"
7     elif units == 'm':
8         return str(num/10**6) + " M"
9     elif units == 'b':
10        return str(num/10**9) + " B"
```

3.3.1 Distribution of ratings

```
1 fig, ax = plt.subplots()
2 plt.title('Distribution of ratings over Training dataset', fontsize=15)
3 sns.countplot(train_df.rating)
4 ax.set_yticklabels([human(item, 'M') for item in ax.get_yticks()])
5 ax.set_ylabel('No. of Ratings(Millions)')
6
7 plt.show()
```



Add new column (week day) to the data set for analysis.

```

1 # It is used to skip the warning 'SettingWithCopyWarning'..
2 pd.options.mode.chained_assignment = None # default='warn'
3
4 train_df['day_of_week'] = train_df.date.dt.weekday_name
5
6 train_df.tail()

```



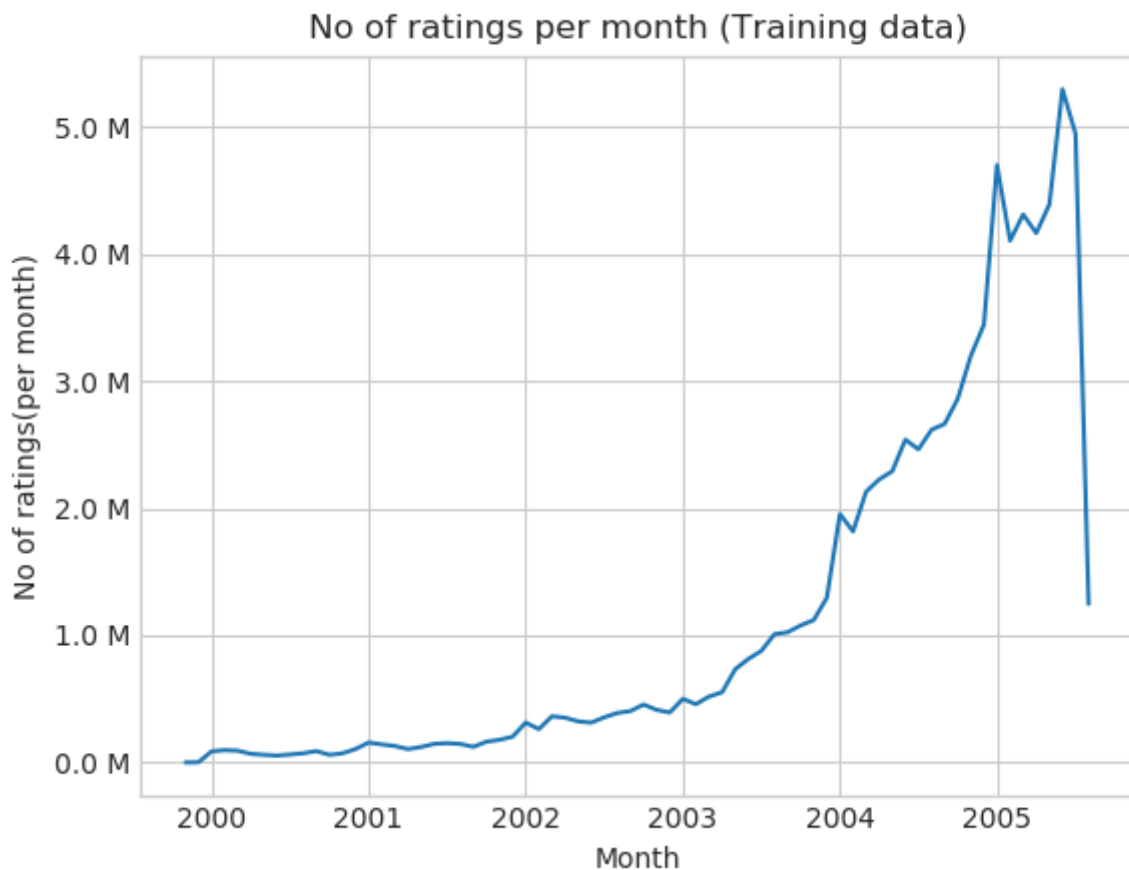
| | movie | user | rating | date | day_of_week |
|-----------------|-------|---------|--------|------------|-------------|
| 80384400 | 12074 | 2033618 | 4 | 2005-08-08 | Monday |
| 80384401 | 862 | 1797061 | 3 | 2005-08-08 | Monday |
| 80384402 | 10986 | 1498715 | 5 | 2005-08-08 | Monday |
| 80384403 | 14861 | 500016 | 4 | 2005-08-08 | Monday |
| 80384404 | 5926 | 1044015 | 5 | 2005-08-08 | Monday |

3.3.2 Number of Ratings per a month

```

1 ax = train_df.resample('m', on='date')['rating'].count().plot()
2 ax.set_title('No of ratings per month (Training data)')
3 plt.xlabel('Month')
4 plt.ylabel('No of ratings(per month)')
5 ax.set_yticklabels([human(item, 'M') for item in ax.get_yticks()])
6 plt.show()

```



Double-click (or enter) to edit

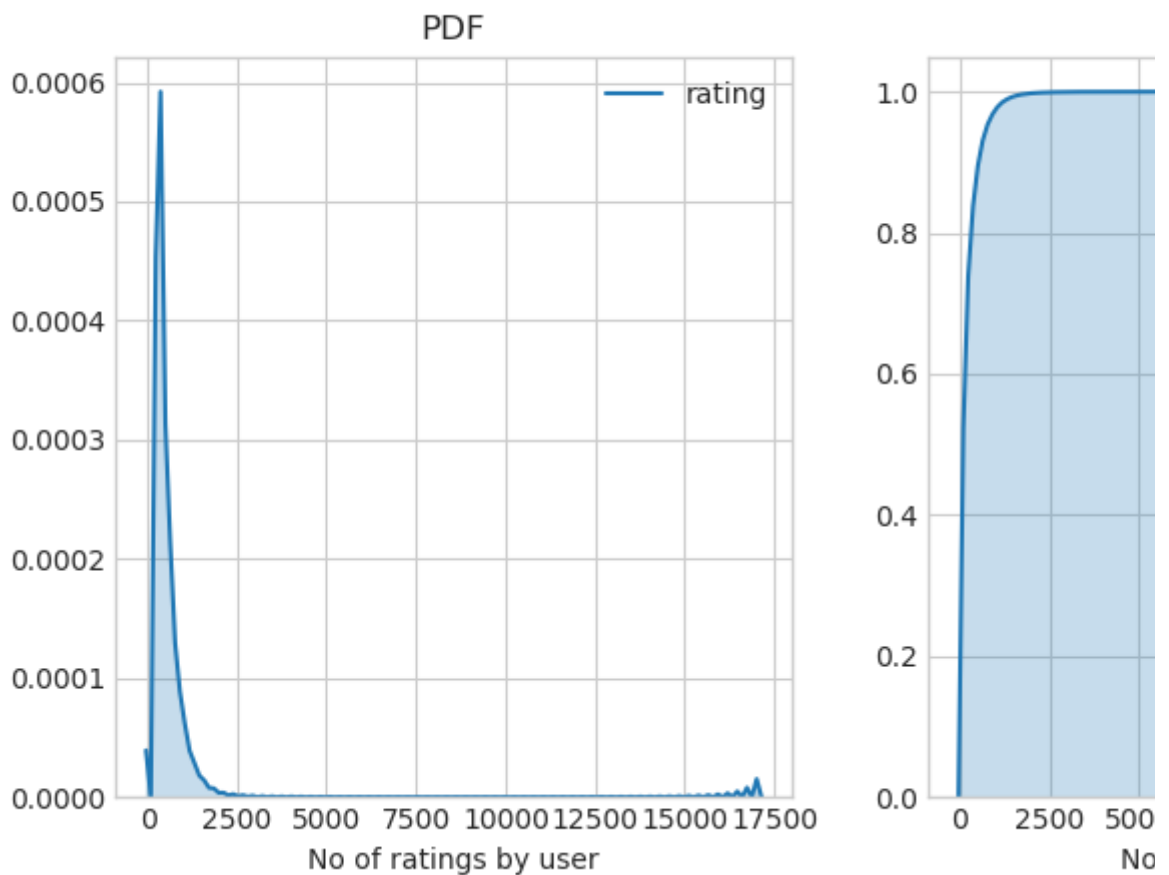
3.3.3 Analysis on the Ratings given by user

```
1 no_of Rated_movies_per_user = train_df.groupby(by='user')['rating'].count().sort_value
2
3 no_of Rated_movies_per_user.head()
```



```
user
305344      17112
2439493      15896
387418       15402
1639792       9767
1461435       9447
Name: rating, dtype: int64
```

```
1 fig = plt.figure(figsize=plt.figaspect(.5))
2
3 ax1 = plt.subplot(121)
4 sns.kdeplot(no_of Rated_movies_per_user, shade=True, ax=ax1)
5 plt.xlabel('No of ratings by user')
6 plt.title("PDF")
7
8 ax2 = plt.subplot(122)
9 sns.kdeplot(no_of Rated_movies_per_user, shade=True, cumulative=True, ax=ax2)
10 plt.xlabel('No of ratings by user')
11 plt.title('CDF')
12
13 plt.show()
```



```
1 no_of_rated_movies_per_user.describe()
```

```
count    405041.000000
mean      198.459921
std       290.793238
min        1.000000
25%       34.000000
50%       89.000000
75%      245.000000
max     17112.000000
Name: rating, dtype: float64
```

There, is something interesting going on with the quantiles..

```
1 quantiles = no_of_rated_movies_per_user.quantile(np.arange(0,1.01,0.01), interpolation
```

```
1 plt.title("Quantiles and their Values")
2 quantiles.plot()
3 # quantiles with 0.05 difference
4 plt.scatter(x=quantiles.index[::5], y=quantiles.values[::5], c='orange', label="quanti
5 # quantiles with 0.25 difference
6 plt.scatter(x=quantiles.index[::25], y=quantiles.values[::25], c='m', label = "quantil
7 plt.ylabel('No of ratings by user')
8 plt.xlabel('Value at the quantile')
9 plt.legend(loc='best')
10
11 # annotate the 25th, 50th, 75th and 100th percentile values....
12 for x,y in zip(quantiles.index[::25], quantiles[::25]):
13     plt.annotate(s="({} , {})".format(x,y), xy=(x,y), xytext=(x-0.05, y+500)
14                 ,fontweight='bold')
15
16
17 plt.show()
```



```
1 quantiles[::5]
```

```
0.00      1
0.05      7
0.10     15
0.15     21
0.20     27
0.25     34
0.30     41
0.35     50
0.40     60
0.45     73
0.50     89
0.55    109
0.60    133
0.65    163
0.70    199
0.75    245
0.80    307
0.85    392
0.90    520
0.95    749
1.00   17112
Name: rating, dtype: int64
```

Value at the quantile

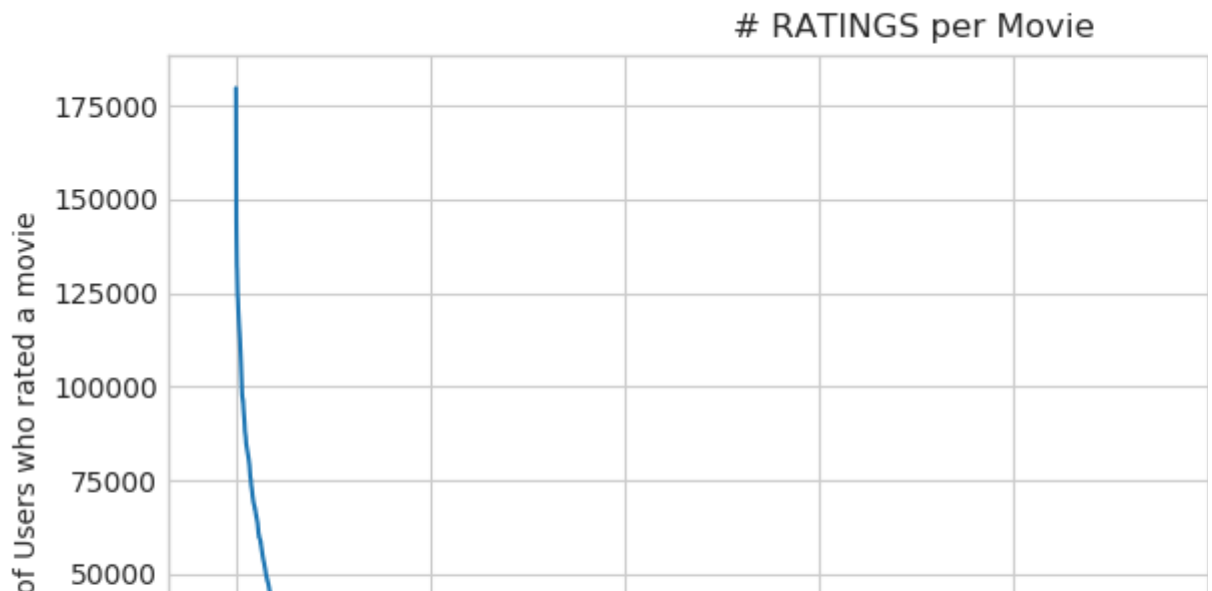
how many ratings at the last 5% of all ratings??

```
1 print('\n No of ratings at last 5 percentile : {}'.format(sum(no_of_rated_movies_per
```

```
No of ratings at last 5 percentile : 20305
```

3.3.4 Analysis of ratings of a movie given by a user

```
1 no_of_ratings_per_movie = train_df.groupby(by='movie')['rating'].count().sort_values(a
2
3 fig = plt.figure(figsize=plt.figaspect(.5))
4 ax = plt.gca()
5 plt.plot(no_of_ratings_per_movie.values)
6 plt.title('# RATINGS per Movie')
7 plt.xlabel('Movie')
8 plt.ylabel('No of Users who rated a movie')
9 ax.set_xticklabels([])
10
11 plt.show()
```



- It is very skewed.. just like number of ratings given per user.
- There are some movies (which are very popular) which are rated by huge number of users.
- But most of the movies(like 90%) got some hundreds of ratings.

3.3.5 Number of ratings on each day of the week

```
1 fig, ax = plt.subplots()
2 sns.countplot(x='day_of_week', data=train_df, ax=ax)
3 plt.title('No of ratings on each day...')
4 plt.ylabel('Total no of ratings')
5 plt.xlabel('')
6 ax.set_yticklabels([human(item, 'M') for item in ax.get_yticks()])
7 plt.show()
```



No of ratings on each day...

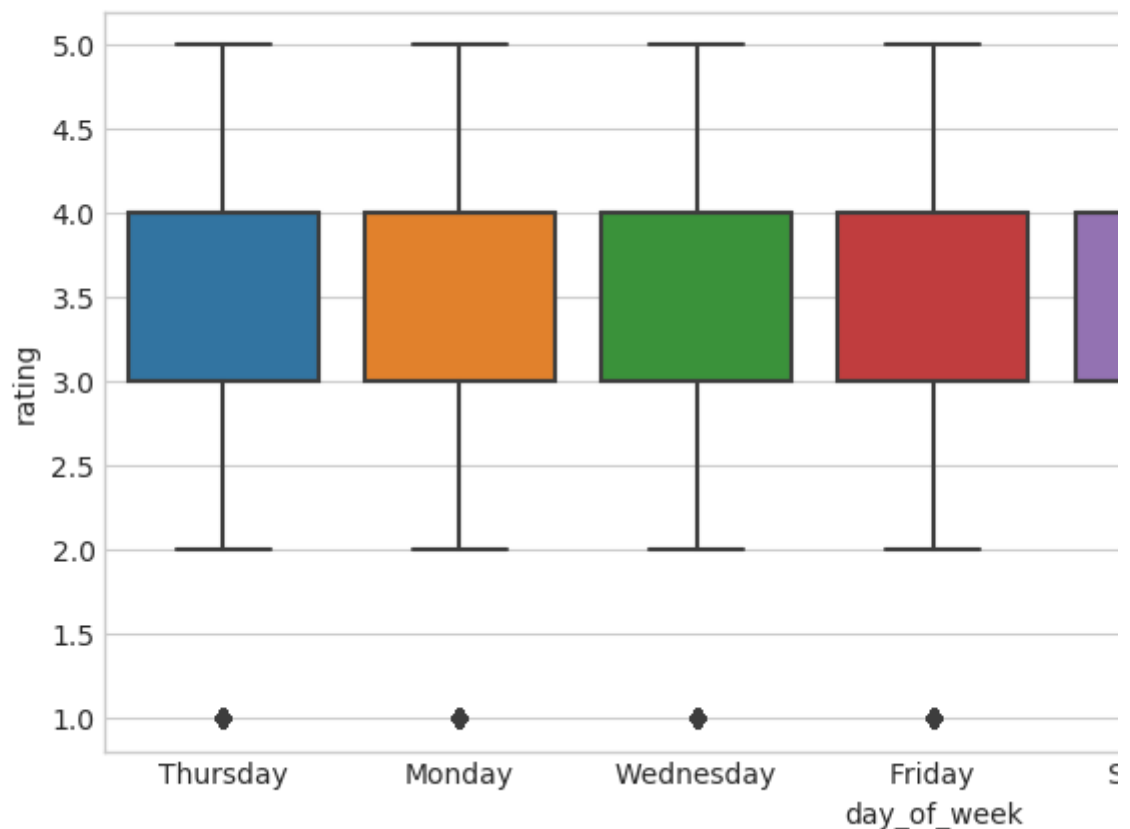
14.0 M



```

1 start = datetime.now()
2 fig = plt.figure(figsize=plt.figaspect(.45))
3 sns.boxplot(y='rating', x='day_of_week', data=train_df)
4 plt.show()
5 print(datetime.now() - start)

```



0:01:10.003761

```

1 avg_week_df = train_df.groupby(by=['day_of_week'])['rating'].mean()
2 print(" AVerage ratings")
3 print("-"*30)
4 print(avg_week_df)
5 print("\n")

```



Average ratings

Double-click (or enter) to edit

Monday 3.577250

3.3.6 Creating sparse matrix from data frame

Tuesday 3.574438



3.3.6.1 Creating sparse matrix from train data frame

```
1 start = datetime.now()
2 if os.path.isfile('train_sparse_matrix.npz'):
3     print("It is present in your pwd, getting it from disk....")
4     # just get it from the disk instead of computing it
5     train_sparse_matrix = sparse.load_npz('train_sparse_matrix.npz')
6     print("DONE..")
7 else:
8     print("We are creating sparse_matrix from the dataframe..")
9     # create sparse_matrix and store it for after usage.
10    # csr_matrix(data_values, (row_index, col_index), shape_of_matrix)
11    # It should be in such a way that, MATRIX[row, col] = data
12    train_sparse_matrix = sparse.csr_matrix((train_df.rating.values, (train_df.user.va
13                                           train_df.movie.values)),)
14
15    print('Done. It\'s shape is : (user, movie) : ',train_sparse_matrix.shape)
16    print('Saving it into disk for furthur usage..')
17    # save it into disk
18    sparse.save_npz("train_sparse_matrix.npz", train_sparse_matrix)
19    print('Done..\n')
20
21 print(datetime.now() - start)
```



We are creating sparse_matrix from the dataframe..
 Done. It's shape is : (user, movie) : (2649430, 17771)
 Saving it into disk for furthur usage..
 Done..

0:01:13.804969

The Sparsity of Train Sparse Matrix

```
1 us,mv = train_sparse_matrix.shape
2 elem = train_sparse_matrix.count_nonzero()
3
4 print("Sparsity Of Train matrix : {} % ".format( (1-(elem/(us*mv))) * 100) )
```



Sparsity Of Train matrix : 99.8292709259195 %

3.3.6.2 Creating sparse matrix from test data frame

```
1 start = datetime.now()
2 if os.path.isfile('test_sparse_matrix.npz'):
3     print("It is present in your pwd, getting it from disk....")
4     # just get it from the disk instead of computing it
5     test_sparse_matrix = sparse.load_npz('test_sparse_matrix.npz')
6     print("DONE..")
7 else:
```

```

8 print("We are creating sparse_matrix from the dataframe..")
9 # create sparse_matrix and store it for after usage.
10 # csr_matrix(data_values, (row_index, col_index), shape_of_matrix)
11 # It should be in such a way that, MATRIX[row, col] = data
12 test_sparse_matrix = sparse.csr_matrix((test_df.rating.values, (test_df.user.value
13                                     test_df.movie.values)))
14
15 print('Done. It\'s shape is : (user, movie) : ',test_sparse_matrix.shape)
16 print('Saving it into disk for furthur usage..')
17 # save it into disk
18 sparse.save_npz("test_sparse_matrix.npz", test_sparse_matrix)
19 print('Done..\n')
20
21 print(datetime.now() - start)

```



We are creating sparse_matrix from the dataframe..
 Done. It's shape is : (user, movie) : (2649430, 17771)
 Saving it into disk for furthur usage..
 Done..

0:00:18.566120

The Sparsity of Test data Matrix

```

1 us,mv = test_sparse_matrix.shape
2 elem = test_sparse_matrix.count_nonzero()
3
4 print("Sparsity Of Test matrix : {} % ".format( (1-(elem/(us*mv))) * 100) )

```



Sparsity Of Test matrix : 99.95731772988694 %

3.3.7 Finding Global average of all movie ratings, Average rating per user, and Average rating per movie

```

1 # get the user averages in dictionary (key: user_id/movie_id, value: avg rating)
2
3 def get_average_ratings(sparse_matrix, of_users):
4
5     # average ratings of user/axes
6     ax = 1 if of_users else 0 # 1 - User axes, 0 - Movie axes
7
8     # ".A1" is for converting Column_Matrix to 1-D numpy array
9     sum_of_ratings = sparse_matrix.sum(axis=ax).A1
10    # Boolean matrix of ratings ( whether a user rated that movie or not)
11    isRated = sparse_matrix!=0
12    # no of ratings that each user OR movie..
13    no_of_ratings = isRated.sum(axis=ax).A1
14
15    # max_user and max_movie ids in sparse matrix
16    u,m = sparse_matrix.shape
17    # creae a dictionary of users and their average ratigns..
18    average_ratings = { i : sum_of_ratings[i]/no_of_ratings[i]
19                      for i in range(u if of_users else m)
20                      if no_of_ratings[i] !=0}
21
22    # return that dictionary of average ratings
23    return average_ratings

```

3.3.7.1 finding global average of all movie ratings

```

1 train_averages = dict()
2 # get the global average of ratings in our train set.


```



```

3 train_global_average = train_sparse_matrix.sum()/train_sparse_matrix.count_nonzero()
4 train_averages['global'] = train_global_average
5 train_averages

```


 {'global': 3.582890686321557}

3.3.7.2 finding average rating per user

```

1 train_averages['user'] = get_average_ratings(train_sparse_matrix, of_users=True)
2 print('\nAverage rating of user 10 :',train_averages['user'][10])

```


 Average rating of user 10 : 3.3781094527363185

3.3.7.3 finding average rating per movie

```

1 train_averages['movie'] = get_average_ratings(train_sparse_matrix, of_users=False)
2 print('\n Average rating of movie 15 :',train_averages['movie'][15])

```

 Average rating of movie 15 : 3.3038461538461537

Double-click (or enter) to edit

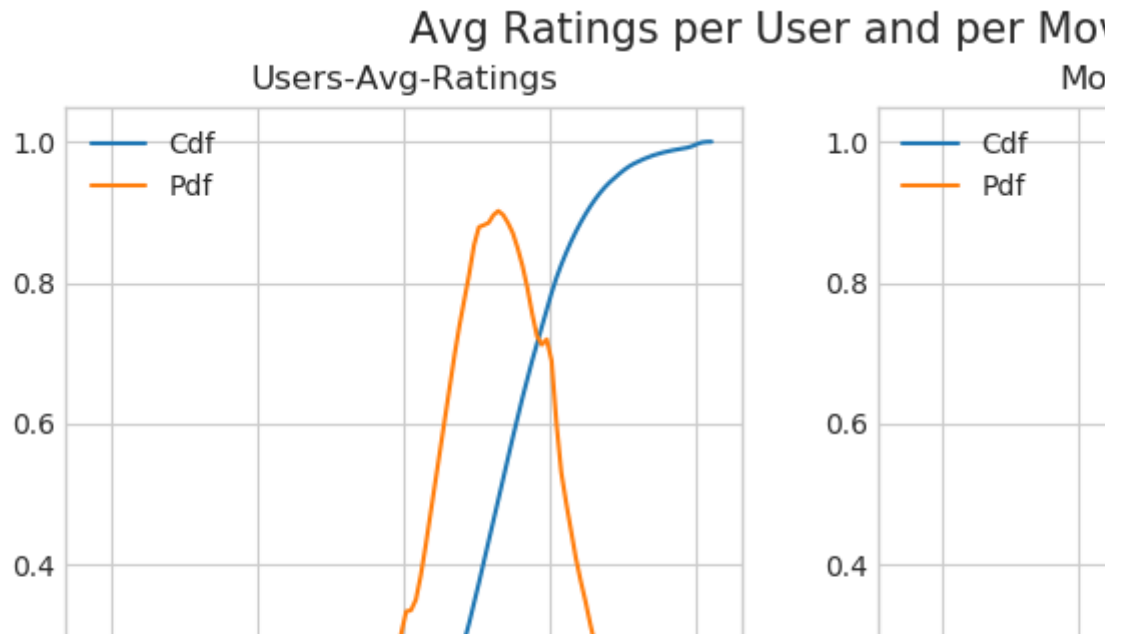
3.3.7.4 PDF's & CDF's of Avg.Ratings of Users & Movies (In Train Data)

```

1 start = datetime.now()
2 # draw pdfs for average rating per user and average
3 fig, (ax1, ax2) = plt.subplots(nrows=1, ncols=2, figsize=plt.figaspect(.5))
4 fig.suptitle('Avg Ratings per User and per Movie', fontsize=15)
5
6 ax1.set_title('Users-Avg-Ratings')
7 # get the list of average user ratings from the averages dictionary..
8 user_averages = [rat for rat in train_averages['user'].values()]
9 sns.distplot(user_averages, ax=ax1, hist=False,
10              kde_kws=dict(cumulative=True), label='Cdf')
11 sns.distplot(user_averages, ax=ax1, hist=False, label='Pdf')
12
13 ax2.set_title('Movies-Avg-Rating')
14 # get the list of movie average ratings from the dictionary..
15 movie_averages = [rat for rat in train_averages['movie'].values()]
16 sns.distplot(movie_averages, ax=ax2, hist=False,
17              kde_kws=dict(cumulative=True), label='Cdf')
18 sns.distplot(movie_averages, ax=ax2, hist=False, label='Pdf')
19
20 plt.show()
21 print(datetime.now() - start)

```





3.3.8 Cold Start problem

3.3.8.1 Cold Start problem with Users

```

1 total_users = len(np.unique(df.user))
2 users_train = len(train_averages['user'])
3 new_users = total_users - users_train
4
5 print('\nTotal number of Users :', total_users)
6 print('\nNumber of Users in Train data :', users_train)
7 print("\nNo of Users that didn't appear in train data: {}({} %) \n ".format(new_users,
8                                     np.round((new_

```



Total number of Users : 480189

Number of Users in Train data : 405041

No of Users that didn't appear in train data: 75148(15.65 %)

We might have to handle **new users (75148)** who didn't appear in train data.

3.3.8.2 Cold Start problem with Movies

```

1 total_movies = len(np.unique(df.movie))
2 movies_train = len(train_averages['movie'])
3 new_movies = total_movies - movies_train
4
5 print('\nTotal number of Movies :', total_movies)
6 print('\nNumber of Users in Train data :', movies_train)
7 print("\nNo of Movies that didn't appear in train data: {}({} %) \n ".format(new_movie
8                                     np.round((new_

```



Total number of Movies : 17770

Number of Users in Train data : 17424

We might have to handle **346 movies** (small comparatively) in test data

Double-click (or enter) to edit

3.4 Computing Similarity matrices

3.4.1 Computing User-User Similarity matrix

1. Calculating User User Similarity_Matrix is **not very easy**(unless you have huge Computing Power and lots of time) because of number of. usersbeing lare.

- You can try if you want to. Your system could crash or the program stops with **Memory Error**

3.4.1.1 Trying with all dimensions (17k dimensions per user)

```

1 from sklearn.metrics.pairwise import cosine_similarity
2
3
4 def compute_user_similarity(sparse_matrix, compute_for_few=False, top = 100, verbose=F
5     draw_time_taken=True):
6     no_of_users, _ = sparse_matrix.shape
7     # get the indices of non zero rows(users) from our sparse matrix
8     row_ind, col_ind = sparse_matrix.nonzero()
9     row_ind = sorted(set(row_ind)) # we don't have to
10    time_taken = list() # time taken for finding similar users for an user..
11
12    # we create rows, cols, and data lists.. which can be used to create sparse matri
13    rows, cols, data = list(), list(), list()
14    if verbose: print("Computing top",top,"similarities for each user..")
15
16    start = datetime.now()
17    temp = 0
18
19    for row in row_ind[:top] if compute_for_few else row_ind:
20        temp = temp+1
21        prev = datetime.now()
22
23        # get the similarity row for this user with all other users
24        sim = cosine_similarity(sparse_matrix.getrow(row), sparse_matrix).ravel()
25        # We will get only the top 'top' most similar users and ignore rest of them.
26        top_sim_ind = sim.argsort()[-top:]
27        top_sim_val = sim[top_sim_ind]
28
29        # add them to our rows, cols and data
30        rows.extend([row]*top)
31        cols.extend(top_sim_ind)
32        data.extend(top_sim_val)
33        time_taken.append(datetime.now().timestamp() - prev.timestamp())
34        if verbose:
35            if temp%verb_for_n_rows == 0:
36                print("computing done for {} users [ time elapsed : {} ]"
37                    .format(temp, datetime.now()-start))
38

```

```

39
40 # lets create sparse matrix out of these and return it
41 if verbose: print('Creating Sparse matrix from the computed similarities')
42 #return rows, cols, data
43
44 if draw_time_taken:
45     plt.plot(time_taken, label = 'time taken for each user')
46     plt.plot(np.cumsum(time_taken), label='Total time')
47     plt.legend(loc='best')
48     plt.xlabel('User')
49     plt.ylabel('Time (seconds)')
50     plt.show()
51
52 return sparse.csr_matrix((data, (rows, cols)), shape=(no_of_users, no_of_users)),

```

```

1 start = datetime.now()
2 u_u_sim_sparse, _ = compute_user_similarity(train_sparse_matrix, compute_for_few=True,
3                                           verbose=True)
4 print("-"*100)
5 print("Time taken :",datetime.now()-start)

```

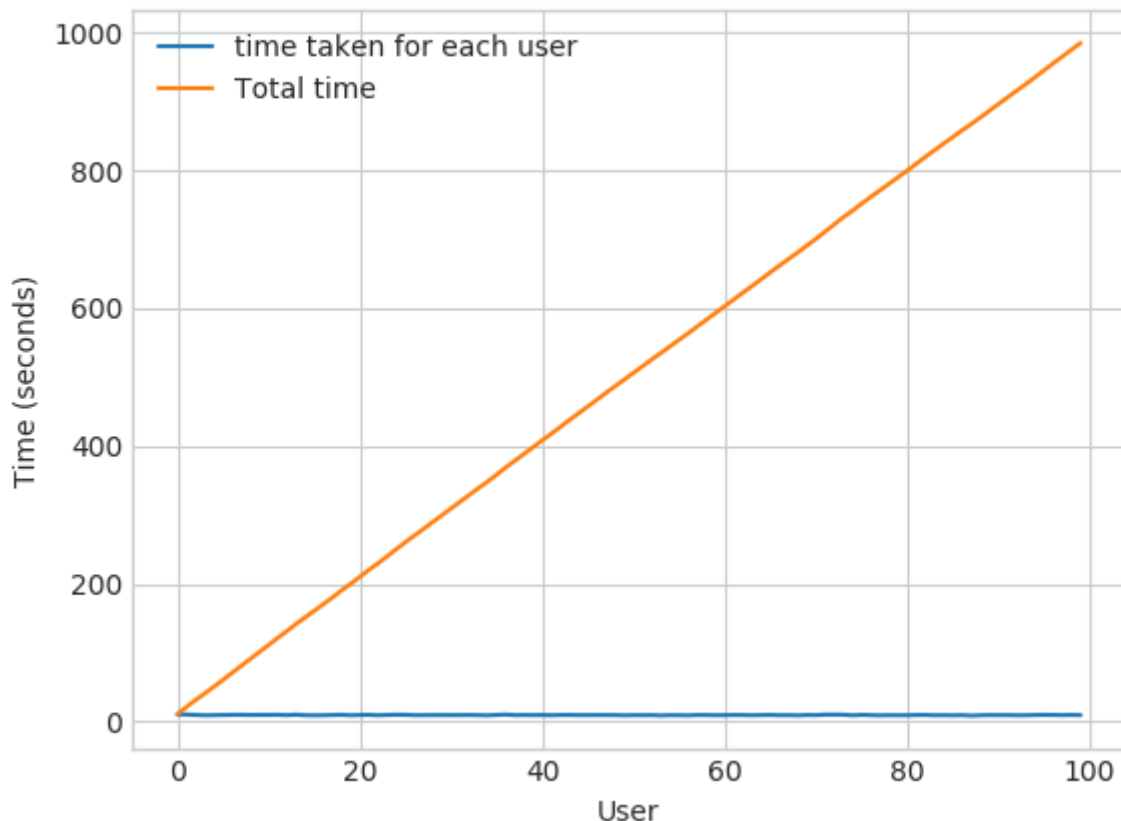


Computing top 100 similarities for each user..

```

computing done for 20 users [ time elapsed : 0:03:20.300488 ]
computing done for 40 users [ time elapsed : 0:06:38.518391 ]
computing done for 60 users [ time elapsed : 0:09:53.143126 ]
computing done for 80 users [ time elapsed : 0:13:10.080447 ]
computing done for 100 users [ time elapsed : 0:16:24.711032 ]
Creating Sparse matrix from the computed similarities

```



Time taken : 0:16:33.618931

3.4.1.2 Trying with reduced dimensions (Using TruncatedSVD for dimensionality reduction of user vector)

Double-click (or enter) to edit

- We have **405,041 users** in our training set and computing similarities between them..(**17K dimensional vector**..) is time consuming..
- From above plot, It took roughly **8.88 sec** for computing similar users for **one user**
- We have **405,041 users** with us in training set.
- - $405041 \times 8.88 = 3596764.08 \text{ sec} = 59946.068 \text{ min} = 999.101133333 \text{ hours} = 41.62921388 \text{ days}$
 - Even if we run on 4 cores parallelly (a typical system now a days), It will still take almost **10 and 1/2 days**.

IDEA: Instead, we will try to reduce the dimensions using SVD, so that **it might** speed up the process...

```
1 from datetime import datetime
2 from sklearn.decomposition import TruncatedSVD
3
4 start = datetime.now()
5
6 # initialize the algorithm with some parameters..
7 # All of them are default except n_components. n_iter is for Randomized SVD solver.
8 netflix_svd = TruncatedSVD(n_components=500, algorithm='randomized', random_state=15)
9 trunc_svd = netflix_svd.fit_transform(train_sparse_matrix)
10
11 print(datetime.now()-start)
```



0:29:07.069783

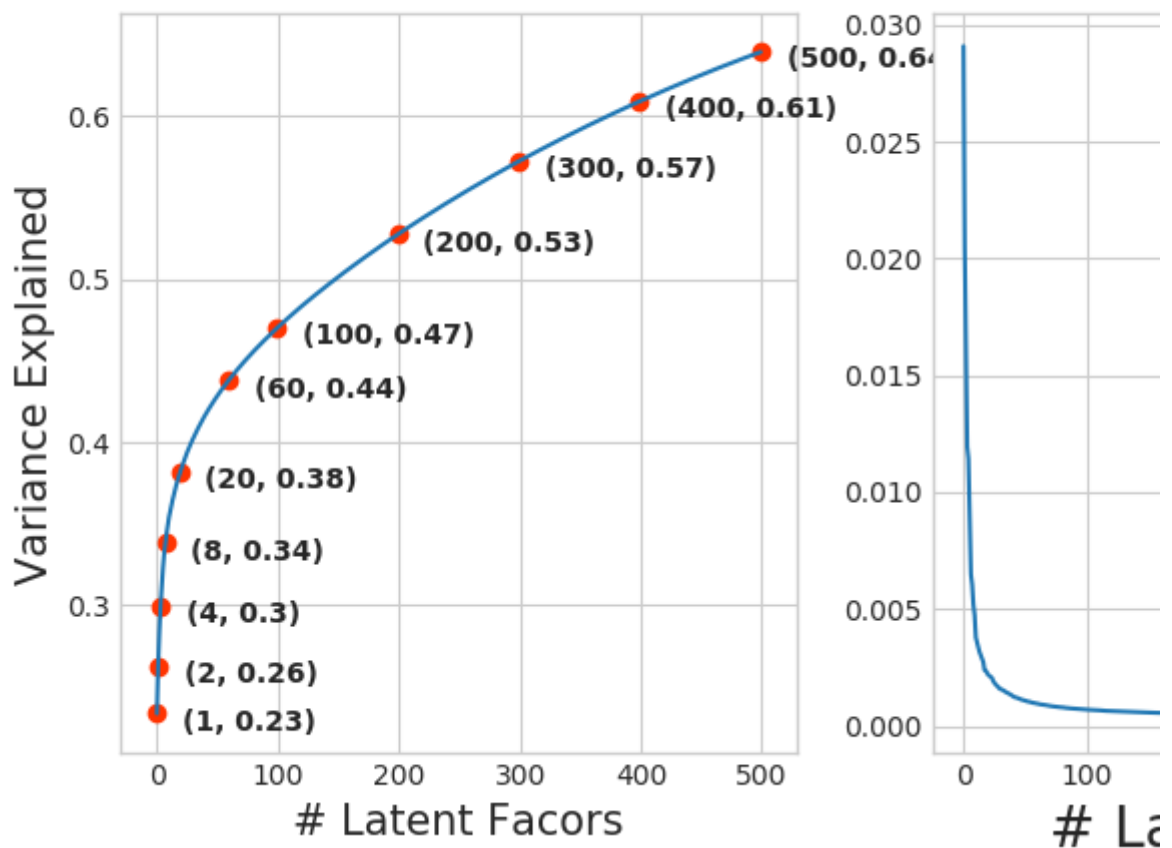
Here,

- $\Sigma \leftarrow (\text{netflix_svd.singular_values_})$
- $V^T \leftarrow (\text{netflix_svd.components_})$
- U is not returned. instead **Projection_of_X** onto the new vectorspace is returned.
- It uses **randomized svd** internally, which returns **All 3 of them separately**. Use that instead..

```
1 expl_var = np.cumsum(netflix_svd.explained_variance_ratio_)
```

```
1 fig, (ax1, ax2) = plt.subplots(nrows=1, ncols=2, figsize=plt.figaspect(.5))
2
3 ax1.set_ylabel("Variance Explained", fontsize=15)
4 ax1.set_xlabel("# Latent Factors", fontsize=15)
5 ax1.plot(expl_var)
6 # annotate some (latent factors, expl_var) to make it clear
7 ind = [1, 2, 4, 8, 20, 60, 100, 200, 300, 400, 500]
8 ax1.scatter(x=[i-1 for i in ind], y=expl_var[[i-1 for i in ind]], c='#ff3300')
9 for i in ind:
10     ax1.annotate(s="{}, {}".format(i, np.round(expl_var[i-1], 2)), xy=(i-1, expl_var[i-1]),
11                  xytext=(i+20, expl_var[i-1] - 0.01), fontweight='bold')
12
13 change_in_expl_var = [expl_var[i+1] - expl_var[i] for i in range(len(expl_var)-1)]
14 ax2.plot(change_in_expl_var)
15
16
17
18 ax2.set_ylabel("Gain in Var_Expl with One Additional LF", fontsize=10)
19 ax2.yaxis.set_label_position("right")
20 ax2.set_xlabel("# Latent Factors", fontsize=20)
21
```

```
22 plt.show()
```



```
1 for i in ind:
2     print("{} {}".format(i, np.round(expl_var[i-1], 2)))
```


```
(1, 0.23)
(2, 0.26)
(4, 0.3)
(8, 0.34)
(20, 0.38)
(60, 0.44)
(100, 0.47)
(200, 0.53)
(300, 0.57)
(400, 0.61)
(500, 0.64)
```

I think 500 dimensions is good enough


- By just taking **(20 to 30)** latent factors, explained variance that we could get is **20 %**.
- To take it to **60%**, we have to take **almost 400 latent factors**. It is not fare.
- It basically is the **gain of variance explained**, if we **add one additional latent factor to it**.
- By adding one by one latent factor too it, the **_gain in explained variance** with that addition is decreasing. (Obviously, because they are sorted that way).
- **LHS Graph:**
 - **x** — (No of latent factors),

- **y** -- (The variance explained by taking x latent factors)
- **More decrease in the line (RHS graph) :**
 - We are getting more explained variance than before.
- **Less decrease in that line (RHS graph) :**
 - We are not getting benefitted from adding latent factor further. This is what is shown in the plots.
- **RHS Graph:**
 - **x** -- (No of latent factors),
 - **y** -- (Gain n Expl_Var by taking one additional latent factor)

```
1 # Let's project our Original U_M matrix into into 500 Dimensional space...
2 start = datetime.now()
3 trunc_matrix = train_sparse_matrix.dot(netflix_svd.components_.T)
4 print(datetime.now()- start)
```

 0:00:45.670265

```
1 type(trunc_matrix), trunc_matrix.shape
```

 (numpy.ndarray, (2649430, 500))

- Let's convert this to actual sparse matrix and store it for future purposes

```
1 if not os.path.isfile('trunc_sparse_matrix.npz'):
2     # create that sparse matrix
3     trunc_sparse_matrix = sparse.csr_matrix(trunc_matrix)
4     # Save this truncated sparse matrix for later usage..
5     sparse.save_npz('trunc_sparse_matrix', trunc_sparse_matrix)
6 else:
7     trunc_sparse_matrix = sparse.load_npz('trunc_sparse_matrix.npz')
```

```
1 trunc_sparse_matrix.shape
```

 (2649430, 500)

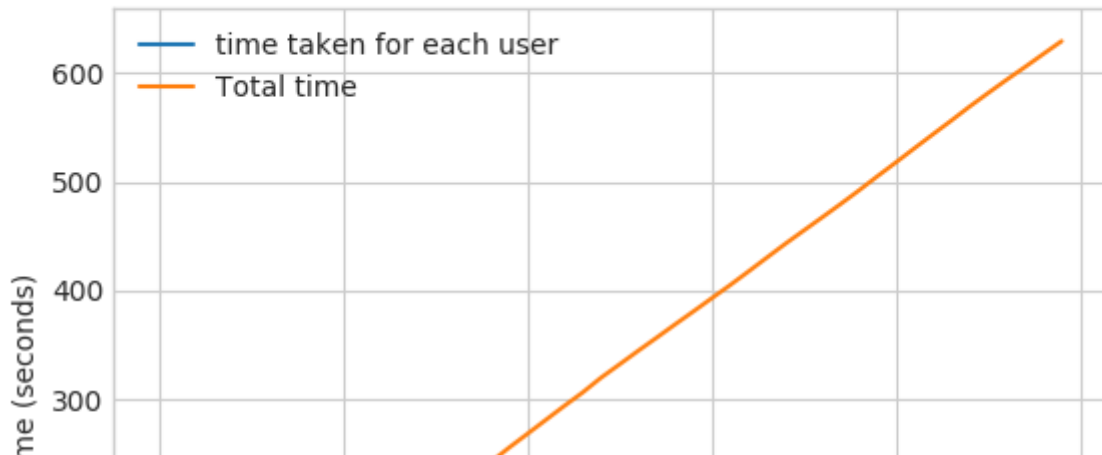
```
1 start = datetime.now()
2 trunc_u_u_sim_matrix, _ = compute_user_similarity(trunc_sparse_matrix, compute_for_few
3                                                     verb_for_n_rows=10)
4 print("-"*50)
5 print("time:",datetime.now()-start)
```



```

Computing top 50 similarities for each user..
computing done for 10 users [ time elapsed : 0:02:09.746324 ]
computing done for 20 users [ time elapsed : 0:04:16.017768 ]
computing done for 30 users [ time elapsed : 0:06:20.861163 ]
computing done for 40 users [ time elapsed : 0:08:24.933316 ]
computing done for 50 users [ time elapsed : 0:10:28.861485 ]
Creating Sparse matrix from the computed similarities

```



: This is taking more time for each user than Original one.

- from above plot, It took almost **12.18** for computing similar users for **one user**
- We have **405041 users** with us in training set.
- $405041 \times 12.18 = 4933399.38 \text{ sec} = 82223.323 \text{ min} = 1370.388716667 \text{ hours}$
 - Even we run on 4 cores parallelly (a typical system now a days), It will still take almost **(14 - 15) days**.

• Why did this happen...??

- Just think about it. It's not that difficult.

----- (sparse & dense.....get it ??) -----

Is there any other way to compute user user similarity..??

-An alternative is to compute similar users for a particular user, whenever required (**ie., Run time**)

- We maintain a binary Vector for users, which tells us whether we already computed or not..
- *****If not***** :
 - Compute top (let's just say, 1000) most similar users for this given user, and add this to our datastructure, so that we can just access it(similar users) without recomputing it again.
- *****If It is already Computed*****:
 - Just get it directly from our datastructure, which has that information.

- In production time, We might have to recompute similarities, if it is computed a long time ago. Because user preferences changes over time. If we could maintain some kind of Timer, which when expires, we have to update it (recompute it).

-
- ***Which datastructure to use:***
 - It is purely implementation dependant.
 - One simple method is to maintain a ****Dictionary Of Dictionaries****.
 -
 - ****key** : **** _userid_**
 - **__value__**: **_Again a dictionary_**
 - **__key__** : **_Similar User_**
 - **__value__**: **_Similarity Value_**

3.4.2 Computing Movie-Movie Similarity matrix

```

1 start = datetime.now()
2 if not os.path.isfile('m_m_sim_sparse.npz'):
3     print("It seems you don't have that file. Computing movie_movie similarity...")
4     start = datetime.now()
5     m_m_sim_sparse = cosine_similarity(X=train_sparse_matrix.T, dense_output=False)
6     print("Done..")
7     # store this sparse matrix in disk before using it. For future purposes.
8     print("Saving it to disk without the need of re-computing it again.. ")
9     sparse.save_npz("m_m_sim_sparse.npz", m_m_sim_sparse)
10    print("Done..")
11 else:
12    print("It is there, We will get it.")
13    m_m_sim_sparse = sparse.load_npz("m_m_sim_sparse.npz")
14    print("Done ...")
15
16 print("It's a ",m_m_sim_sparse.shape," dimensional matrix")
17
18 print(datetime.now() - start)

```

It seems you don't have that file. Computing movie_movie similarity...
 Done..
 Saving it to disk without the need of re-computing it again..
 Done..
 It's a (17771, 17771) dimensional matrix
 0:10:02.736054

```
1 m_m_sim_sparse.shape
```

(17771, 17771)

- Even though we have similarity measure of each movie, with all other movies, We generally don't care much about least similar movies.
- Most of the times, only top_xxx similar items matters. It may be 10 or 100.
- We take only those top similar movie ratings and store them in a saperate dictionary.

```
1 movie_ids = np.unique(m_m_sim_sparse.nonzero()[1])
```

```

1 start = datetime.now()
2 similar_movies = dict()
3 for movie in movie_ids:
4     # get the top similar movies and store them in the dictionary
5     sim_movies = m_m_sim_sparse[movie].toarray().ravel().argsort()[::-1][1:]
6     similar_movies[movie] = sim_movies[:100]
7 print(datetime.now() - start)
8
9 # just testing similar movies for movie_15
10 similar_movies[15]

```



0:00:33.411700

```

array([ 8279,  8013, 16528,  5927, 13105, 12049,  4424, 10193, 17590,
        4549,  3755,   590, 14059, 15144, 15054,  9584,  9071,  6349,
       16402,  3973,  1720,  5370, 16309,  9376,  6116,  4706,  2818,
         778, 15331,  1416, 12979, 17139, 17710,  5452,  2534,   164,
       15188,  8323,  2450, 16331,  9566, 15301, 13213, 14308, 15984,
       10597,  6426,  5500,  7068,  7328,  5720,  9802,   376, 13013,
        8003, 10199,  3338, 15390,  9688, 16455, 11730,  4513,   598,
       12762,  2187,   509,  5865,  9166, 17115, 16334,  1942,  7282,
       17584,  4376,  8988,  8873,  5921,  2716, 14679, 11947, 11981,
        4649,   565, 12954, 10788, 10220, 10963,  9427,  1690,  5107,
        7859,  5969,  1510,  2429,   847,  7845,  6410, 13931,  9840,
       3706])

```

Double-click (or enter) to edit

3.4.3 Finding most similar movies using similarity matrix

Does Similarity really works as the way we expected...?

Let's pick some random movie and check for its similar movies....

```

1 # First Let's load the movie details into soe dataframe..
2 # movie details are in 'netflix/movie_titles.csv'
3
4 movie_titles = pd.read_csv("data_folder/movie_titles.csv", sep=',', header = None,
5                             names=['movie_id', 'year_of_release', 'title'], verbose=True,
6                             index_col = 'movie_id', encoding = "ISO-8859-1")
7
8 movie_titles.head()

```



Tokenization took: 4.50 ms

Type conversion took: 165.72 ms

Parser memory cleanup took: 0.01 ms

| | year_of_release | title |
|----------|-----------------|------------------------------|
| movie_id | | |
| 1 | 2003.0 | Dinosaur Planet |
| 2 | 2004.0 | Isle of Man TT 2004 Review |
| 3 | 1997.0 | Character |
| 4 | 1994.0 | Paula Abdul's Get Up & Dance |
| 5 | 2004.0 | The Rise and Fall of ECW |

Similar Movies for 'Vampire Journals'

```

1 mv_id = 67
2
3 print("\nMovie ----->",movie_titles.loc[mv_id].values[1])
4
5 print("\nIt has {} Ratings from users.".format(train_sparse_matrix[:,mv_id].getnnz()))
6
7 print("\nWe have {} movies which are similarto this  and we will get only top most..".

```



Movie -----> Vampire Journals

It has 270 Ratings from users.

We have 17284 movies which are similarto this and we will get only top most..

```

1 similarities = m_m_sim_sparse[mv_id].toarray().ravel()
2
3 similar_indices = similarities.argsort()[::-1][1:]
4
5 similarities[similar_indices]
6
7 sim_indices = similarities.argsort()[::-1][1:] # It will sort and reverse the array an
8                                             # and return its indices(movie_ids)

```

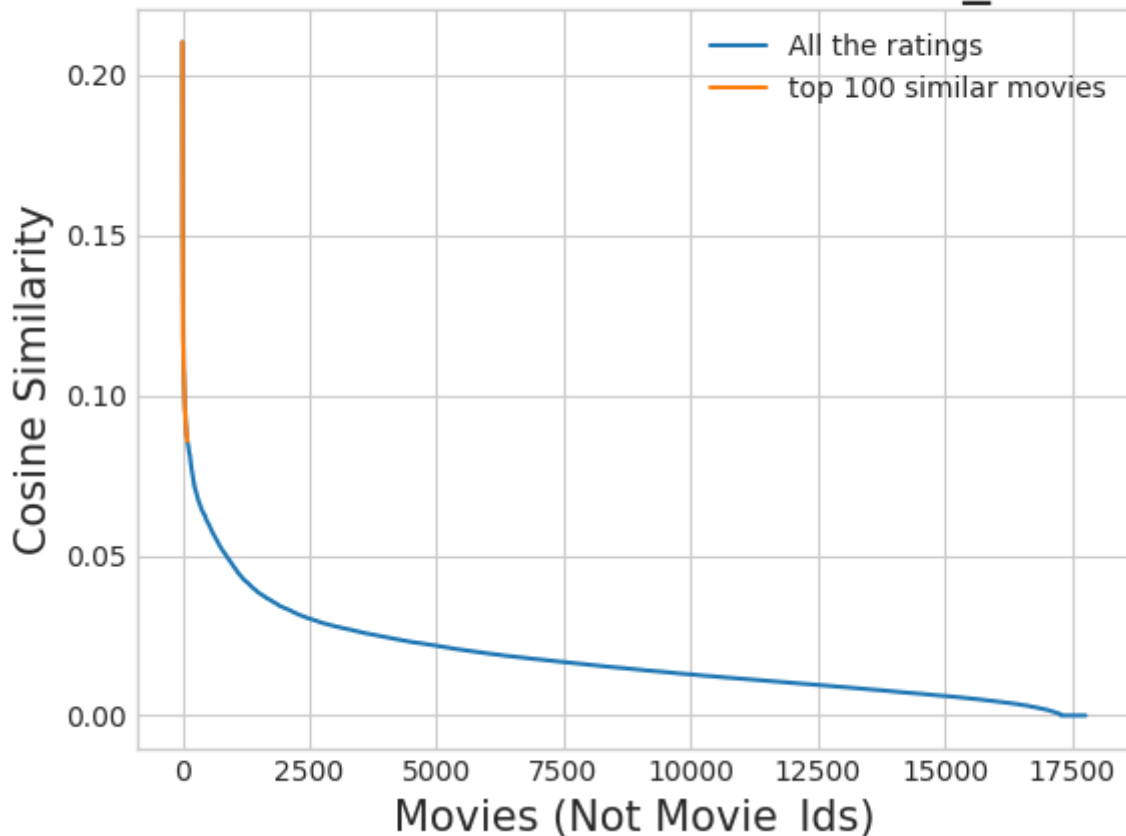
```

1 plt.plot(similarities[sim_indices], label='All the ratings')
2 plt.plot(similarities[sim_indices[:100]], label='top 100 similar movies')
3 plt.title("Similar Movies of {}(movie_id)".format(mv_id), fontsize=20)
4 plt.xlabel("Movies (Not Movie_Ids)", fontsize=15)
5 plt.ylabel("Cosine Similarity",fontsize=15)
6 plt.legend()
7 plt.show()

```



Similar Movies of 67(movie_id)



Double-click (or enter) to edit

Double-click (or enter) to edit

Top 10 similar movies

```
1 movie_titles.loc[sim_indices[:10]]
```



| | year_of_release | title |
|----------|-----------------|--------------------------|
| movie_id | | |
| 323 | 1999.0 | Modern Vampires |
| 4044 | 1998.0 | Subspecies 4: Bloodstorm |
| 1688 | 1993.0 | To Sleep With a Vampire |
| 13962 | 2001.0 | Dracula: The Dark Prince |
| 12053 | 1993.0 | Dracula Rising |
| 16279 | 2002.0 | Vampires: Los Muertos |
| 4667 | 1996.0 | Vampirella |
| 1900 | 1997.0 | Club Vampire |
| 13873 | 2001.0 | The Breed |
| 15867 | 2003.0 | Dracula II: Ascension |

Double-click (or enter) to edit

Similarly, we can **find similar users** and compare how similar they are.

Double-click (or enter) to edit

Double-click (or enter) to edit

4. Machine Learning Models



```
1 def get_sample_sparse_matrix(sparse_matrix, no_users, no_movies, path, verbose = True)
2     """
3     It will get it from the 'path' if it is present or It will create
4     and store the sampled sparse matrix in the path specified.
5     """
6
7     # get (row, col) and (rating) tuple from sparse_matrix...
8     row_ind, col_ind, ratings = sparse.find(sparse_matrix)
9     users = np.unique(row_ind)
```

```

10 movies = np.unique(col_ind)
11
12 print("Original Matrix : (users, movies) -- ({} {})".format(len(users), len(movies))
13 print("Original Matrix : Ratings -- {}{}\n".format(len(ratings)))
14
15 # It just to make sure to get same sample everytime we run this program..
16 # and pick without replacement....
17 np.random.seed(15)
18 sample_users = np.random.choice(users, no_users, replace=False)
19 sample_movies = np.random.choice(movies, no_movies, replace=False)
20 # get the boolean mask or these sampled_items in originl row/col_inds..
21 mask = np.logical_and( np.isin(row_ind, sample_users),
22                        np.isin(col_ind, sample_movies) )
23
24 sample_sparse_matrix = sparse.csr_matrix((ratings[mask], (row_ind[mask], col_ind[m
25                                         shape=(max(sample_users)+1, max(sample_mo
26
27 if verbose:
28     print("Sampled Matrix : (users, movies) -- ({} {})".format(len(sample_users),
29     print("Sampled Matrix : Ratings --", format(ratings[mask].shape[0]))
30
31 print('Saving it into disk for furthur usage..')
32 # save it into disk
33 sparse.save_npz(path, sample_sparse_matrix)
34 if verbose:
35     print('Done..\n')
36
37 return sample_sparse_matrix

```

4.1 Sampling Data

4.1.1 Build sample train data from the train data

```

1 start = datetime.now()
2 path = "sample/small/sample_train_sparse_matrix.npz"
3 if os.path.isfile(path):
4     print("It is present in your pwd, getting it from disk....")
5     # just get it from the disk instead of computing it
6     sample_train_sparse_matrix = sparse.load_npz(path)
7     print("DONE..")
8 else:
9     # get 10k users and 1k movies from available data
10    sample_train_sparse_matrix = get_sample_sparse_matrix(train_sparse_matrix, no_user
11                                                         path = path)
12
13 print(datetime.now() - start)

```



It is present in your pwd, getting it from disk....
 DONE..
 0:00:00.035179

4.1.2 Build sample test data from the test data

```

1 start = datetime.now()
2
3 path = "sample/small/sample_test_sparse_matrix.npz"
4 if os.path.isfile(path):
5     print("It is present in your pwd, getting it from disk....")
6     # just get it from the disk instead of computing it
7     sample_test_sparse_matrix = sparse.load_npz(path)
8     print("DONE..")
9 else:
10    # get 5k users and 500 movies from available data

```

```

11 sample_test_sparse_matrix = get_sample_sparse_matrix(test_sparse_matrix, no_users=
12 path = "sample/small/sample_test_spar
13 print(datetime.now() - start)

```



It is present in your pwd, getting it from disk....
 DONE..
 0:00:00.028740

Double-click (or enter) to edit

4.2 Finding Global Average of all movie ratings, Average rating per User, and Average rating per Movie (from sampled train)

```

1 sample_train_averages = dict()

```

4.2.1 Finding Global Average of all movie ratings

```

1 # get the global average of ratings in our train set.
2 global_average = sample_train_sparse_matrix.sum()/sample_train_sparse_matrix.count_non
3 sample_train_averages['global'] = global_average
4 sample_train_averages

```



{'global': 3.581679377504138}

4.2.2 Finding Average rating per User

```

1 sample_train_averages['user'] = get_average_ratings(sample_train_sparse_matrix, of_use
2 print('\nAverage rating of user 1515220 :',sample_train_averages['user'][1515220])

```



Average rating of user 1515220 : 3.9655172413793105

4.2.3 Finding Average rating per Movie

```

1 sample_train_averages['movie'] = get_average_ratings(sample_train_sparse_matrix, of_u
2 print('\n AVerage rating of movie 15153 :',sample_train_averages['movie'][15153])

```



AVerage rating of movie 15153 : 2.6458333333333335

Double-click (or enter) to edit

4.3 Featurizing data

```

1 print('\n No of ratings in Our Sampled train matrix is : {}'.format(sample_train_spa
2 print('\n No of ratings in Our Sampled test matrix is : {}'.format(sample_test_spar

```



No of ratings in Our Sampled train matrix is : 129286

No of ratings in Our Sampled test matrix is : 7333

4.3.1 Featurizing data for regression problem

4.3.1.1 Featurizing train data

```

1 # get users, movies and ratings from our samples train sparse matrix
2 sample_train_users, sample_train_movies, sample_train_ratings = sparse.find(sample_train_sparse_matrix)

1 #####
2 # It took me almost 10 hours to prepare this train dataset.##
3 #####
4 start = datetime.now()
5 if os.path.isfile('sample/small/reg_train.csv'):
6     print("File already exists you don't have to prepare again..." )
7 else:
8     print('preparing {} tuples for the dataset.\n'.format(len(sample_train_ratings)))
9     with open('sample/small/reg_train.csv', mode='w') as reg_data_file:
10         count = 0
11         for (user, movie, rating) in zip(sample_train_users, sample_train_movies, sample_train_ratings):
12             st = datetime.now()
13             # print(user, movie)
14             #----- Ratings of "movie" by similar users of "user" -----
15             # compute the similar Users of the "user"
16             user_sim = cosine_similarity(sample_train_sparse_matrix[user], sample_train_sparse_matrix[sample_train_users]).T
17             top_sim_users = user_sim.argsort()[::-1][1:] # we are ignoring 'The User'
18             # get the ratings of most similar users for this movie
19             top_ratings = sample_train_sparse_matrix[top_sim_users, movie].toarray().r
20             # we will make it's length "5" by adding movie averages to .
21             top_sim_users_ratings = list(top_ratings[top_ratings != 0][:5])
22             top_sim_users_ratings.extend([sample_train_averages['movie'][movie]]*(5 - len(top_sim_users_ratings)))
23             # print(top_sim_users_ratings, end=" ")
24
25
26             #----- Ratings by "user" to similar movies of "movie" -----
27             # compute the similar movies of the "movie"
28             movie_sim = cosine_similarity(sample_train_sparse_matrix[:,movie].T, sample_train_sparse_matrix[sample_train_movies]).T
29             top_sim_movies = movie_sim.argsort()[::-1][1:] # we are ignoring 'The User'
30             # get the ratings of most similar movie rated by this user..
31             top_ratings = sample_train_sparse_matrix[user, top_sim_movies].toarray().r
32             # we will make it's length "5" by adding user averages to.
33             top_sim_movies_ratings = list(top_ratings[top_ratings != 0][:5])
34             top_sim_movies_ratings.extend([sample_train_averages['user'][user]]*(5 - len(top_sim_movies_ratings)))
35             # print(top_sim_movies_ratings, end=" : -- ")
36
37             #-----prepare the row to be stores in a file-----#
38             row = list()
39             row.append(user)
40             row.append(movie)
41             # Now add the other features to this data...
42             row.append(sample_train_averages['global']) # first feature
43             # next 5 features are similar_users "movie" ratings
44             row.extend(top_sim_users_ratings)
45             # next 5 features are "user" ratings for similar_movies
46             row.extend(top_sim_movies_ratings)
47             # Avg_user rating
48             row.append(sample_train_averages['user'][user])
49             # Avg_movie rating
50             row.append(sample_train_averages['movie'][movie])
51
52             # finalley, The actual Rating of this user-movie pair...
53             row.append(rating)
54             count = count + 1
55

```

```

56         # add rows to the file opened..
57         reg_data_file.write(','.join(map(str, row)))
58         reg_data_file.write('\n')
59         if (count)%10000 == 0:
60             # print(','.join(map(str, row)))
61             print("Done for {} rows----- {}".format(count, datetime.now() - start))
62
63
64 print(datetime.now() - start)

```



preparing 129286 tuples for the dataset..

```

Done for 10000 rows----- 0:53:13.974716
Done for 20000 rows----- 1:47:58.228942
Done for 30000 rows----- 2:42:46.963119
Done for 40000 rows----- 3:36:44.807894
Done for 50000 rows----- 4:28:55.311500
Done for 60000 rows----- 5:24:18.493104
Done for 70000 rows----- 6:17:39.669922
Done for 80000 rows----- 7:11:23.970879
Done for 90000 rows----- 8:05:33.787770
Done for 100000 rows----- 9:00:25.463562
Done for 110000 rows----- 9:51:28.530010
Done for 120000 rows----- 10:42:05.382141
11:30:13.699183

```

Reading from the file to make a Train_dataframe

```

1 reg_train = pd.read_csv('sample/small/reg_train.csv', names = ['user', 'movie', 'GAvg'
2 reg_train.head()

```



| | user | movie | GAvg | sur1 | sur2 | sur3 | sur4 | sur5 | smr1 | smr2 | smr3 | smr4 | smi |
|---|--------|-------|----------|------|------|------|------|------|------|------|------|------|-----|
| 0 | 53406 | 33 | 3.581679 | 4.0 | 5.0 | 5.0 | 4.0 | 1.0 | 5.0 | 2.0 | 5.0 | 3.0 | 1 |
| 1 | 99540 | 33 | 3.581679 | 5.0 | 5.0 | 5.0 | 4.0 | 5.0 | 3.0 | 4.0 | 4.0 | 3.0 | 5 |
| 2 | 99865 | 33 | 3.581679 | 5.0 | 5.0 | 4.0 | 5.0 | 3.0 | 5.0 | 4.0 | 4.0 | 5.0 | 4 |
| 3 | 101620 | 33 | 3.581679 | 2.0 | 3.0 | 5.0 | 5.0 | 4.0 | 4.0 | 3.0 | 3.0 | 4.0 | 5 |
| 4 | 112974 | 33 | 3.581679 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 3.0 | 5.0 | 5.0 | 5.0 | 3 |

- **GAvg** : Average rating of all the ratings
- **Similar users rating of this movie:**
 - sur1, sur2, sur3, sur4, sur5 (top 5 similar users who rated that movie..)
- **Similar movies rated by this user:**
 - smr1, smr2, smr3, smr4, smr5 (top 5 similar movies rated by this movie..)
- **UAvg** : User's Average rating
- **MAvg** : Average rating of this movie
- **rating** : Rating of this movie by this user.

Double-click (or enter) to edit

4.3.1.2 Featurizing test data

```
1 # get users, movies and ratings from the Sampled Test
2 sample_test_users, sample_test_movies, sample_test_ratings = sparse.find(sample_test_s
```

```
1 sample_train_averages['global']
```



3.581679377504138

```
1 start = datetime.now()
2
3 if os.path.isfile('sample/small/reg_test.csv'):
4     print("It is already created...")
5 else:
6
7     print('preparing {} tuples for the dataset.\n'.format(len(sample_test_ratings)))
8     with open('sample/small/reg_test.csv', mode='w') as reg_data_file:
9         count = 0
10        for (user, movie, rating) in zip(sample_test_users, sample_test_movies, sample_test_ratings):
11            st = datetime.now()
12
13            #----- Ratings of "movie" by similar users of "user" -----
14            #print(user, movie)
15            try:
16                # compute the similar Users of the "user"
17                user_sim = cosine_similarity(sample_train_sparse_matrix[user], sample_train_sparse_matrix)
18                top_sim_users = user_sim.argsort()[::-1][1:] # we are ignoring 'The User'
19                # get the ratings of most similar users for this movie
20                top_ratings = sample_train_sparse_matrix[top_sim_users, movie].toarray()
21                # we will make it's length "5" by adding movie averages to .
22                top_sim_users_ratings = list(top_ratings[top_ratings != 0][:5])
23                top_sim_users_ratings.extend([sample_train_averages['movie'][movie]]*(5-len(top_sim_users_ratings)))
24                # print(top_sim_users_ratings, end="--")
25
26            except (IndexError, KeyError):
27                # It is a new User or new Movie or there are no ratings for given user
28                ##### Cold Start Problem #####
29                top_sim_users_ratings.extend([sample_train_averages['global']]*(5 - len(top_sim_users_ratings)))
30                #print(top_sim_users_ratings)
31            except:
32                print(user, movie)
33                # we just want KeyErrors to be resolved. Not every Exception...
34                raise
35
36
37
38            #----- Ratings by "user" to similar movies of "movie" -----
39            try:
40                # compute the similar movies of the "movie"
41                movie_sim = cosine_similarity(sample_train_sparse_matrix[:,movie].T, sample_train_sparse_matrix[:,movie].T)
42                top_sim_movies = movie_sim.argsort()[::-1][1:] # we are ignoring 'The Movie'
43                # get the ratings of most similar movie rated by this user..
44                top_ratings = sample_train_sparse_matrix[user, top_sim_movies].toarray()
45                # we will make it's length "5" by adding user averages to.
46                top_sim_movies_ratings = list(top_ratings[top_ratings != 0][:5])
47                top_sim_movies_ratings.extend([sample_train_averages['user'][user]]*(5-len(top_sim_movies_ratings)))
48                #print(top_sim_movies_ratings)
49            except (IndexError, KeyError):
50                #print(top_sim_movies_ratings, end=" : -- ")
51                top_sim_movies_ratings.extend([sample_train_averages['global']]*(5-len(top_sim_movies_ratings)))
52                #print(top_sim_movies_ratings)
53            except:
54                raise
55
56            #-----prepare the row to be stores in a file-----#
57            row = list()
58            # add usser and movie name first
```

```

59     row.append(user)
60     row.append(movie)
61     row.append(sample_train_averages['global']) # first feature
62     #print(row)
63     # next 5 features are similar_users "movie" ratings
64     row.extend(top_sim_users_ratings)
65     #print(row)
66     # next 5 features are "user" ratings for similar_movies
67     row.extend(top_sim_movies_ratings)
68     #print(row)
69     # Avg_user rating
70     try:
71         row.append(sample_train_averages['user'][user])
72     except KeyError:
73         row.append(sample_train_averages['global'])
74     except:
75         raise
76     #print(row)
77     # Avg_movie rating
78     try:
79         row.append(sample_train_averages['movie'][movie])
80     except KeyError:
81         row.append(sample_train_averages['global'])
82     except:
83         raise
84     #print(row)
85     # finalley, The actual Rating of this user-movie pair...
86     row.append(rating)
87     #print(row)
88     count = count + 1
89
90     # add rows to the file opened..
91     reg_data_file.write(','.join(map(str, row)))
92     #print(','.join(map(str, row)))
93     reg_data_file.write('\n')
94     if (count)%1000 == 0:
95         #print(','.join(map(str, row)))
96         print("Done for {} rows----- {}".format(count, datetime.now() - start))
97     print("",datetime.now() - start)

```



preparing 7333 tuples for the dataset..

```

Done for 1000 rows----- 0:04:29.293783
Done for 2000 rows----- 0:08:57.208002
Done for 3000 rows----- 0:13:30.333223
Done for 4000 rows----- 0:18:04.050813
Done for 5000 rows----- 0:22:38.671673
Done for 6000 rows----- 0:27:09.697009
Done for 7000 rows----- 0:31:41.933568
0:33:12.529731

```

Reading from the file to make a test dataframe

```

1 reg_test_df = pd.read_csv('sample/small/reg_test.csv', names = ['user', 'movie', 'GAvg',
2                               'smr1', 'smr2', 'smr3', 'smr',
3                               'UAvg', 'MAvg', 'rating'], h
4 reg_test_df.head(4)

```



| | user | movie | GAvg | sur1 | sur2 | sur3 | sur4 | sur5 | smr |
|---|--|-------|------|------|------|------|------|------|-----|
| • | GAvg : Average rating of all the ratings | | | | | | | | |
| • | Similar users rating of this movie: | | | | | | | | |
| | ◦ sur1, sur2, sur3, sur4, sur5 (top 5 simiular users who rated that movie..) | | | | | | | | |
| • | Similar movies rated by this user: | | | | | | | | |
| | ◦ smr1, smr2, smr3, smr4, smr5 (top 5 simiular movies rated by this movie..) | | | | | | | | |
| • | UAvg : User AVerage rating | | | | | | | | |
| • | MAvg : Average rating of this movie | | | | | | | | |
| • | rating : Rating of this movie by this user. | | | | | | | | |

Double-click (or enter) to edit

4.3.2 Transforming data for Surprise models

```
1 from surprise import Reader, Dataset
```

4.3.2.1 Transforming train data


- We can't give raw data (movie, user, rating) to train the model in Surprise library.
- They have a saperate format for TRAIN and TEST data, which will be useful for training the models like SVD, KNNBaseLineOnly....etc.,in Surprise.
- We can form the trainset from a file, or from a Pandas DataFrame.
http://surprise.readthedocs.io/en/stable/getting_started.html#load-dom-dataframe-py

```
1 # It is to specify how to read the dataframe.
2 # for our dataframe, we don't have to specify anything extra..
3 reader = Reader(rating_scale=(1,5))
4
5 # create the traindata from the dataframe...
6 train_data = Dataset.load_from_df(reg_train[['user', 'movie', 'rating']], reader)
7
8 # build the trainset from traindata.. It is of dataset format from surprise library..
9 trainset = train_data.build_full_trainset()
```

4.3.2.2 Transforming test data

- Testset is just a list of (user, movie, rating) tuples. (Order in the tuple is impotant)

```
1 testset = list(zip(reg_test_df.user.values, reg_test_df.movie.values, reg_test_df.rati
2 testset[:3]
```

 [(808635, 71, 5), (941866, 71, 4), (1737912, 71, 3)]

4.4 Applying Machine Learning models


Double-click (or enter) to edit

- Global dictionary that stores rmse and mape for all the models....
 - It stores the metrics in a dictionary of dictionaries

keys : model names(string)

value: dict(**key** : metric, **value** : value)

```
1 models_evaluation_train = dict()
2 models_evaluation_test = dict()
3
4 models_evaluation_train, models_evaluation_test
```

 ({} , {})

Double-click (or enter) to edit

Utility functions for running regression models

```
1 # to get rmse and mape given actual and predicted ratings..
2 def get_error_metrics(y_true, y_pred):
3     rmse = np.sqrt(np.mean([ (y_true[i] - y_pred[i])**2 for i in range(len(y_pred)) ]))
4     mape = np.mean(np.abs( (y_true - y_pred)/y_true )) * 100
5     return rmse, mape
6
7 #####
8 #####
9 def run_xgboost(algo, x_train, y_train, x_test, y_test, verbose=True):
10     """
11     It will return train_results and test_results
12     """
13
14     # dictionaries for storing train and test results
15     train_results = dict()
16     test_results = dict()
17
18
19     # fit the model
20     print('Training the model..')
21     start = datetime.now()
22     algo.fit(x_train, y_train, eval_metric = 'rmse')
23     print('Done. Time taken : {}'.format(datetime.now()-start))
24     print('Done \n')
25
26     # from the trained model, get the predictions....
27     print('Evaluating the model with TRAIN data...')
28     start = datetime.now()
29     y_train_pred = algo.predict(x_train)
30     # get the rmse and mape of train data...
31     rmse_train, mape_train = get_error_metrics(y_train.values, y_train_pred)
32
33     # store the results in train_results dictionary..
34     train_results = {'rmse': rmse_train,
35                     'mape' : mape_train,
36                     'predictions' : y_train_pred}
37
38     #####
39     # get the test data predictions and compute rmse and mape
40     print('Evaluating Test data')
41     y_test_pred = algo.predict(x_test)
42     rmse_test, mape_test = get_error_metrics(y_true=y_test.values, y_pred=y_test_pred)
```

```

43 # store them in our test results dictionary.
44 test_results = {'rmse': rmse_test,
45                 'mape' : mape_test,
46                 'predictions':y_test_pred}
47 if verbose:
48     print('\nTEST DATA')
49     print('-'*30)
50     print('RMSE : ', rmse_test)
51     print('MAPE : ', mape_test)
52
53 # return these train and test results...
54 return train_results, test_results
55

```

Utility functions for Surprise modes

```

1 # it is just to makesure that all of our algorithms should produce same results
2 # everytime they run...
3
4 my_seed = 15
5 random.seed(my_seed)
6 np.random.seed(my_seed)
7
8 #####
9 # get (actual_list , predicted_list) ratings given list
10 # of predictions (prediction is a class in Surprise).
11 #####
12 def get_ratings(predictions):
13     actual = np.array([pred.r_ui for pred in predictions])
14     pred = np.array([pred.est for pred in predictions])
15
16     return actual, pred
17
18 #####
19 # get 'rmse' and 'mape' , given list of prediction objects
20 #####
21 def get_errors(predictions, print_them=False):
22
23     actual, pred = get_ratings(predictions)
24     rmse = np.sqrt(np.mean((pred - actual)**2))
25     mape = np.mean(np.abs(pred - actual)/actual)
26
27     return rmse, mape*100
28
29 #####
30 # It will return predicted ratings, rmse and mape of both train and test data #
31 #####
32 def run_surprise(algo, trainset, testset, verbose=True):
33     '''
34         return train_dict, test_dict
35
36         It returns two dictionaries, one for train and the other is for test
37         Each of them have 3 key-value pairs, which specify 'rmse', 'mape', and 'p
38     '''
39     start = datetime.now()
40     # dictionaries that stores metrics for train and test..
41     train = dict()
42     test = dict()
43
44     # train the algorithm with the trainset
45     st = datetime.now()
46     print('Training the model...')
47     algo.fit(trainset)
48     print('Done. time taken : {} \n'.format(datetime.now()-st))
49
50     # ----- Evaluating train data-----#
51     st = datetime.now()
52     print('Evaluating the model with train data..')
53     # get the train predictions (list of prediction class inside Surprise)
54     train_preds = algo.test(trainset.build_testset())

```

```

55 # get predicted ratings from the train predictions..
56 train_actual_ratings, train_pred_ratings = get_ratings(train_preds)
57 # get 'rmse' and 'mape' from the train predictions.
58 train_rmse, train_mape = get_errors(train_preds)
59 print('time taken : {}'.format(datetime.now()-st))
60
61 if verbose:
62     print('-'*15)
63     print('Train Data')
64     print('-'*15)
65     print("RMSE : {}\nMAPE : {}".format(train_rmse, train_mape))
66
67 #store them in the train dictionary
68 if verbose:
69     print('adding train results in the dictionary..')
70 train['rmse'] = train_rmse
71 train['mape'] = train_mape
72 train['predictions'] = train_pred_ratings
73
74 #----- Evaluating Test data-----#
75 st = datetime.now()
76 print('\nEvaluating for test data...')
77 # get the predictions( list of prediction classes) of test data
78 test_preds = algo.test(testset)
79 # get the predicted ratings from the list of predictions
80 test_actual_ratings, test_pred_ratings = get_ratings(test_preds)
81 # get error metrics from the predicted and actual ratings
82 test_rmse, test_mape = get_errors(test_preds)
83 print('time taken : {}'.format(datetime.now()-st))
84
85 if verbose:
86     print('-'*15)
87     print('Test Data')
88     print('-'*15)
89     print("RMSE : {}\nMAPE : {}".format(test_rmse, test_mape))
90 # store them in test dictionary
91 if verbose:
92     print('storing the test results in test dictionary...')
93 test['rmse'] = test_rmse
94 test['mape'] = test_mape
95 test['predictions'] = test_pred_ratings
96
97 print('\n'+ '-'*45)
98 print('Total time taken to run this algorithm :', datetime.now() - start)
99
100 # return two dictionaries train and test
101 return train, test

```

Double-click (or enter) to edit

4.4.1 XGBoost with initial 13 features

```
1 import xgboost as xgb
```

```

1 # prepare Train data
2 x_train = reg_train.drop(['user', 'movie', 'rating'], axis=1)
3 y_train = reg_train['rating']
4
5 # Prepare Test data
6 x_test = reg_test_df.drop(['user', 'movie', 'rating'], axis=1)
7 y_test = reg_test_df['rating']
8
9 # initialize Our first XGBoost model...
10 first_xgb = xgb.XGBRegressor(silent=False, n_jobs=13, random_state=15, n_estimators=10)
11 train_results, test_results = run_xgboost(first_xgb, x_train, y_train, x_test, y_test)
12
13 # store the results in models_evaluations dictionaries
14 models_evaluation_train['first_algo'] = train_results

```

```

15 models_evaluation_test['first_algo'] = test_results
16
17 xgb.plot_importance(first_xgb)
18 plt.show()

```



Training the model..

Done. Time taken : 0:00:01.795787

Done

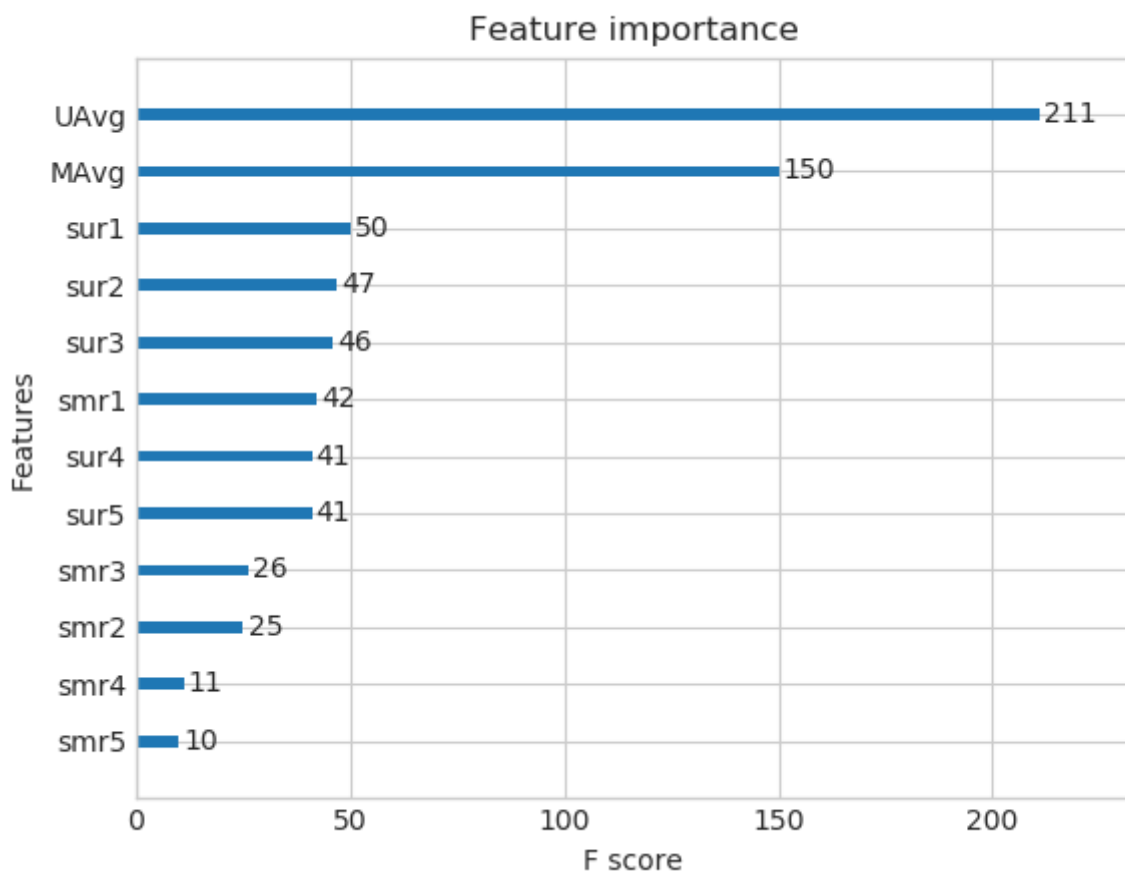
Evaluating the model with TRAIN data...

Evaluating Test data

TEST DATA

RMSE : 1.0761851474385373

MAPE : 34.504887593204884



Double-click (or enter) to edit

4.4.2 Surprise BaselineModel

```
1 from surprise import BaselineOnly
```

Predicted_rating : (baseline prediction)

-

http://surprise.readthedocs.io/en/stable/basic_algorithms.html#surprise.prediction_algorithms.baseline_only.BaselineOnly

$$\hat{r}_{ui} = b_{ui} = \mu + b_u + b_i$$

- μ : Average of all trainings in training data.
- b_u : User bias
- b_i : Item bias (movie biases)

Optimization function (Least Squares Problem)

-

http://surprise.readthedocs.io/en/stable/prediction_algorithms.html#baseline-s-estimates-configuration

$$\sum_{r_{ui} \in R_{train}} (r_{ui} - (\mu + b_u + b_i))^2 + \lambda (b_u^2 + b_i^2) . \text{ [mimimize } b_u, b_i]$$

```

1
2 # options are to specify.., how to compute those user and item biases
3 bsl_options = {'method': 'sgd',
4               'learning_rate': .001
5               }
6 bsl_algo = BaselineOnly(bsl_options=bsl_options)
7 # run this algorithm.., It will return the train and test results..
8 bsl_train_results, bsl_test_results = run_surprise(my_bsl_algo, trainset, testset, ver
9
10
11 # Just store these error metrics in our models_evaluation datastructure
12 models_evaluation_train['bsl_algo'] = bsl_train_results
13 models_evaluation_test['bsl_algo'] = bsl_test_results

```



Training the model...

Estimating biases using sgd...

Done. time taken : 0:00:00.822391

Evaluating the model with train data..

time taken : 0:00:01.116752

Train Data

RMSE : 0.9347153928678286

MAPE : 29.389572652358183

adding train results in the dictionary..

Evaluating for test data...

time taken : 0:00:00.074418

Test Data

RMSE : 1.0730330260516174

MAPE : 35.04995544572911

storing the test results in test dictionary...

Total time taken to run this algorithm : 0:00:02.014073

Double-click (or enter) to edit

4.4.3 XGBoost with initial 13 features + Surprise Baseline predictor

Updating Train Data

```
1 # add our baseline_predicted value as our feature..
2 reg_train['bslpr'] = models_evaluation_train['bsl_algo']['predictions']
3 reg_train.head(2)
```



| | user | movie | GAvg | sur1 | sur2 | sur3 | sur4 | sur5 | smr1 | smr2 | smr3 | smr4 | smr5 |
|---|-------|-------|----------|------|------|------|------|------|------|------|------|------|------|
| 0 | 53406 | 33 | 3.581679 | 4.0 | 5.0 | 5.0 | 4.0 | 1.0 | 5.0 | 2.0 | 5.0 | 3.0 | 1.0 |
| 1 | 99540 | 33 | 3.581679 | 5.0 | 5.0 | 5.0 | 4.0 | 5.0 | 3.0 | 4.0 | 4.0 | 3.0 | 5.0 |

Updating Test Data

```
1 # add that baseline predicted ratings with Surprise to the test data as well
2 reg_test_df['bslpr'] = models_evaluation_test['bsl_algo']['predictions']
3
4 reg_test_df.head(2)
```



| | user | movie | GAvg | sur1 | sur2 | sur3 | sur4 | sur5 | smr1 |
|---|--------|-------|----------|----------|----------|----------|----------|----------|----------|
| 0 | 808635 | 71 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 |
| 1 | 941866 | 71 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 |

```
1 # prepare train data
2 x_train = reg_train.drop(['user', 'movie', 'rating'], axis=1)
3 y_train = reg_train['rating']
4
5 # Prepare Test data
6 x_test = reg_test_df.drop(['user', 'movie', 'rating'], axis=1)
7 y_test = reg_test_df['rating']
8
9 # initialize Our first XGBoost model...
10 xgb_bsl = xgb.XGBRegressor(silent=False, n_jobs=13, random_state=15, n_estimators=100)
11 train_results, test_results = run_xgboost(xgb_bsl, x_train, y_train, x_test, y_test)
12
13 # store the results in models_evaluations dictionaries
14 models_evaluation_train['xgb_bsl'] = train_results
15 models_evaluation_test['xgb_bsl'] = test_results
16
17 xgb.plot_importance(xgb_bsl)
18 plt.show()
19
```



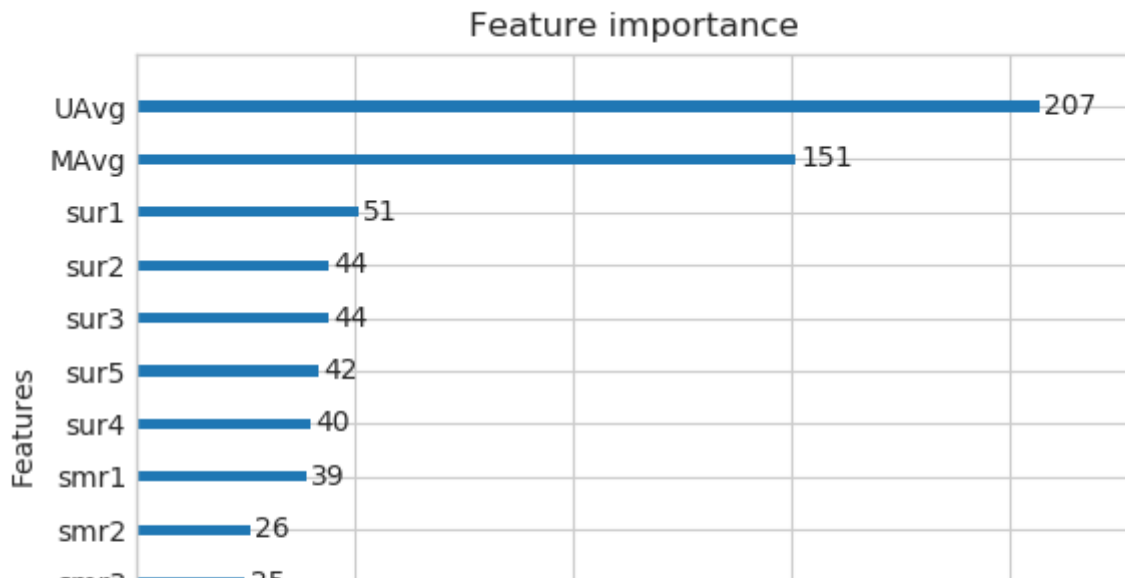
Training the model..
 Done. Time taken : 0:00:02.388635

Done

Evaluating the model with TRAIN data...
 Evaluating Test data

TEST DATA

 RMSE : 1.0763419061709816
 MAPE : 34.491235560745295



Double-click (or enter) to edit

Double-click (or enter) to edit

4.4.4 Surprise KNNBaseline predictor

```
1 from surprise import KNNBaseline
```

- KNN BASELINE
 - http://surprise.readthedocs.io/en/stable/knn_inspired.html#surprise.prediction_algorithm.s.knns.KNNBaseline
- PEARSON_BASELINE SIMILARITY
 - http://surprise.readthedocs.io/en/stable/similarities.html#surprise.similarities.pearson_baseline
- SHRINKAGE
 - 2.2 Neighborhood Models in <http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/a1-koren.pdf>

- **predicted Rating** : (_ based on User-User similarity _)

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{v \in N_i^k(u)} \text{sim}(u, v) \cdot (r_{vi} - b_{vi})}{\sum_{v \in N_i^k(u)} \text{sim}(u, v)}$$

- **b_{ui}** - Baseline prediction of (user, movie) rating
- **$N_i^k(u)$** - Set of **K similar** users (neighbours) of **user (u)** who rated **movie(i)**
- **$\text{sim}(u, v)$** - **Similarity** between users **u** and **v**
 - Generally, it will be cosine similarity or Pearson correlation coefficient.
 - But we use **shrunk Pearson-baseline correlation coefficient**, which is based on the pearsonBaseline similarity (we take base line predictions instead of mean rating of user/item)

Double-click (or enter) to edit

- **Predicted rating** (based on Item Item similarity):

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{j \in N_u^k(i)} \text{sim}(i, j) \cdot (r_{uj} - b_{uj})}{\sum_{j \in N_u^k(i)} \text{sim}(i, j)}$$

- **_Notations follows same as above (user user based predicted rating) _**

4.4.4.1 Surprise KNNBaseline with user user similarities

```

1 # we specify , how to compute similarities and what to consider with sim_options to ou
2 sim_options = {'user_based' : True,
3               'name': 'pearson_baseline',
4               'shrinkage': 100,
5               'min_support': 2
6               }
7 # we keep other parameters like regularization parameter and learning_rate as default
8 bsl_options = {'method': 'sgd'}
9
10 knn_bsl_u = KNNBaseline(k=40, sim_options = sim_options, bsl_options = bsl_options)
11 knn_bsl_u_train_results, knn_bsl_u_test_results = run_surprise(knn_bsl_u, trainset, te
12
13 # Just store these error metrics in our models_evaluation datastructure
14 models_evaluation_train['knn_bsl_u'] = knn_bsl_u_train_results
15 models_evaluation_test['knn_bsl_u'] = knn_bsl_u_test_results
16

```



```

Training the model...
Estimating biases using sgd...
Computing the pearson_baseline similarity matrix...
Done computing similarity matrix.
Done. time taken : 0:00:30.173847

```

```

Evaluating the model with train data..
time taken : 0:01:35.970614

```

```

-----
Train Data
-----

```

```

RMSE : 0.33642097416508826

```

```

MAPE : 9.145093375416348

```

```

adding train results in the dictionary..

```

4.4.4.2 Surprise KNNBaseline with movie movie similarities

```

1 # we specify , how to compute similarities and what to consider with sim_options to ou
2
3 # 'user_based' : Fals => this considers the similarities of movies instead of users
4
5 sim_options = {'user_based' : False,
6               'name': 'pearson_baseline',
7               'shrinkage': 100,
8               'min_support': 2
9               }
10 # we keep other parameters like regularization parameter and learning_rate as default
11 bsl_options = {'method': 'sgd'}
12
13
14 knn_bsl_m = KNNBaseline(k=40, sim_options = sim_options, bsl_options = bsl_options)
15
16 knn_bsl_m_train_results, knn_bsl_m_test_results = run_surprise(knn_bsl_m, trainset, te
17
18 # Just store these error metrics in our models_evaluation datastructure
19 models_evaluation_train['knn_bsl_m'] = knn_bsl_m_train_results
20 models_evaluation_test['knn_bsl_m'] = knn_bsl_m_test_results
21

```



```

Training the model...
Estimating biases using sgd...
Computing the pearson_baseline similarity matrix...
Done computing similarity matrix.
Done. time taken : 0:00:01.093096

```

```

Evaluating the model with train data..
time taken : 0:00:07.964272

```

Double-click (or enter) to edit

4.4.5 XGBoost with initial 13 features + Surprise Baseline predictor + KNNBaseline predictor

adding train results in the dictionary..

- First we will run XGBoost with predictions from both KNN's (that uses User_User and Item_Item similarities along with our previous features.
- Then we will run XGBoost with just predictions form both knn models and preditions from our baseline model.

```
RMSF = 1 072758832653683
```

Preparing Train data

```

1 # add the predicted values from both knns to this dataframe
2 reg_train['knn_bsl_u'] = models_evaluation_train['knn_bsl_u']['predictions']
3 reg_train['knn_bsl_m'] = models_evaluation_train['knn_bsl_m']['predictions']
4
5 reg_train.head(2)

```

| | user | movie | GAvg | sur1 | sur2 | sur3 | sur4 | sur5 | smr1 | smr2 | smr3 | smr4 | smr5 |
|---|-------|-------|----------|------|------|------|------|------|------|------|------|------|------|
| 0 | 53406 | 33 | 3.581679 | 4.0 | 5.0 | 5.0 | 4.0 | 1.0 | 5.0 | 2.0 | 5.0 | 3.0 | 1.0 |
| 1 | 99540 | 33 | 3.581679 | 5.0 | 5.0 | 5.0 | 4.0 | 5.0 | 3.0 | 4.0 | 4.0 | 3.0 | 5.0 |

Preparing Test data

```

1 reg_test_df['knn_bsl_u'] = models_evaluation_test['knn_bsl_u']['predictions']
2 reg_test_df['knn_bsl_m'] = models_evaluation_test['knn_bsl_m']['predictions']
3
4 reg_test_df.head(2)

```

| | user | movie | GAvg | sur1 | sur2 | sur3 | sur4 | sur5 | smr1 |
|---|--------|-------|----------|----------|----------|----------|----------|----------|----------|
| 0 | 808635 | 71 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 |
| 1 | 941866 | 71 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 |

```

1 # prepare the train data....
2 x_train = reg_train.drop(['user', 'movie', 'rating'], axis=1)
3 y_train = reg_train['rating']
4
5 # prepare the train data....
6 x_test = reg_test_df.drop(['user', 'movie', 'rating'], axis=1)
7 y_test = reg_test_df['rating']

```

```

8
9 # declare the model
10 xgb_knn_bsl = xgb.XGBRegressor(n_jobs=10, random_state=15)
11 train_results, test_results = run_xgboost(xgb_knn_bsl, x_train, y_train, x_test, y_test)
12
13 # store the results in models_evaluations dictionaries
14 models_evaluation_train['xgb_knn_bsl'] = train_results
15 models_evaluation_test['xgb_knn_bsl'] = test_results
16
17
18 xgb.plot_importance(xgb_knn_bsl)
19 plt.show()

```



Training the model..

Done. Time taken : 0:00:02.092387

Done

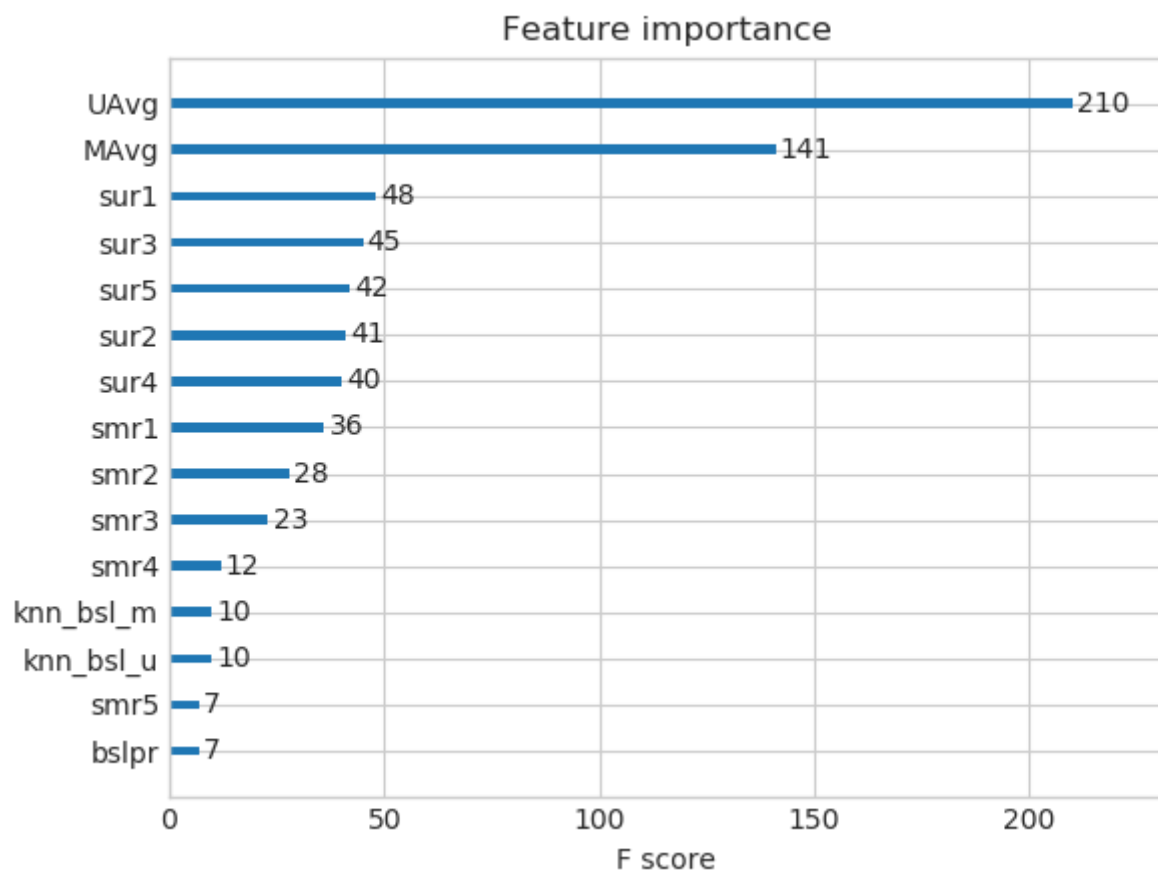
Evaluating the model with TRAIN data...

Evaluating Test data

TEST DATA

RMSE : 1.0763602465199797

MAPE : 34.48862808016984



4.4.6 Matrix Factorization Techniques

4.4.6.1 SVD Matrix Factorization User Movie interactions

```
1 from surprise import SVD
```

http://surprise.readthedocs.io/en/stable/matrix_factorization.html#surprise.prediction_algorithms.matrix_factorization.SVD

- **Predicted Rating :**

-
- $\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$
 - q_i - Representation of item(movie) in latent factor space
 - p_u - Representation of user in new latent factor space

- A BASIC MATRIX FACTORIZATION MODEL in [https://datajobs.com/data-science-repo/Recommender-Systems-\[Netflix\].pdf](https://datajobs.com/data-science-repo/Recommender-Systems-[Netflix].pdf)

- **Optimization problem with user item interactions and regularization (to avoid overfitting)**

-
- $\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda (b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2)$

```
1 # initialize the model
2 svd = SVD(n_factors=100, biased=True, random_state=15, verbose=True)
3 svd_train_results, svd_test_results = run_surprise(svd, trainset, testset, verbose=True)
4
5 # Just store these error metrics in our models_evaluation datastructure
6 models_evaluation_train['svd'] = svd_train_results
7 models_evaluation_test['svd'] = svd_test_results
```



Training the model...

Processing epoch 0
 Processing epoch 1
 Processing epoch 2
 Processing epoch 3
 Processing epoch 4
 Processing epoch 5
 Processing epoch 6
 Processing epoch 7
 Processing epoch 8
 Processing epoch 9
 Processing epoch 10
 Processing epoch 11
 Processing epoch 12
 Processing epoch 13
 Processing epoch 14
 Processing epoch 15

Double-click (or enter) to edit

Processing epoch 16

4.4.6.2 SVD Matrix Factorization with implicit feedback from user (user rated movies)

```
1 from surprise import SVDpp
```

- ----> 2.5 Implicit Feedback in <http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/a1-koren.pdf>

- **Predicted Rating :**

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T \left(p_u + |I_u|^{-\frac{1}{2}} \sum_{j \in I_u} y_j \right)$$

- I_u --- the set of all items rated by user u
- y_j --- Our new set of item factors that capture implicit ratings.

- **Optimization problem with user item interactions and regularization (to avoid overfitting)**

$$\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda \left(b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2 + ||y_j||^2 \right)$$

```
1 # initialize the model
2 svdpp = SVDpp(n_factors=50, random_state=15, verbose=True)
3 svdpp_train_results, svdpp_test_results = run_surprise(svdpp, trainset, testset, verbo
4
5 # Just store these error metrics in our models_evaluation datastructure
6 models_evaluation_train['svdpp'] = svdpp_train_results
7 models_evaluation_test['svdpp'] = svdpp_test_results
8
```




```

Training the model...
processing epoch 0
processing epoch 1
processing epoch 2
processing epoch 3
processing epoch 4
processing epoch 5
processing epoch 6
processing epoch 7
processing epoch 8
processing epoch 9
processing epoch 10
processing epoch 11
processing epoch 12
processing epoch 13
processing epoch 14
processing epoch 15
processing epoch 16
processing epoch 17
processing epoch 18
processing epoch 19
Done. time taken : 0:01:56.765007

Evaluating the model with train data..
time taken : 0:00:06.387920
-----

```

Train Data

RMSE : 0.6032438403305899

MAPE : 17.49285063490268

adding train results in the dictionary..

```

Evaluating for test data...
time taken : 0:00:00.071642
-----

```

Test Data

RMSE : 1.0728491944183447

MAPE : 35.03817913919887

Double-click (or enter) to edit

Double-click (or enter) to edit

4.4.7 XgBoost with 13 features + Surprise Baseline + Surprise KNNbaseline + MF Techniques

Preparing Train data

```

1 # add the predicted values from both knns to this dataframe
2 reg_train['svd'] = models_evaluation_train['svd']['predictions']
3 reg_train['svdpp'] = models_evaluation_train['svdpp']['predictions']
4
5 reg_train.head(2)

```



| | user | movie | GAvg | sur1 | sur2 | sur3 | sur4 | sur5 | smr1 | smr2 | ... | smr4 | smr5 |
|---|-------|-------|----------|------|------|------|------|------|------|------|-----|------|------|
| 0 | 53406 | 33 | 3.581679 | 4.0 | 5.0 | 5.0 | 4.0 | 1.0 | 5.0 | 2.0 | ... | 3.0 | 1.0 |
| 1 | 99540 | 33 | 3.581679 | 5.0 | 5.0 | 5.0 | 4.0 | 5.0 | 3.0 | 4.0 | ... | 3.0 | 5.0 |

2 rows × 21 columns

Preparing Test data

```

1 reg_test_df['svd'] = models_evaluation_test['svd']['predictions']
2 reg_test_df['svdpp'] = models_evaluation_test['svdpp']['predictions']
3
4 reg_test_df.head(2)

```



| | user | movie | GAvg | sur1 | sur2 | sur3 | sur4 | sur5 | smr1 |
|---|--------|-------|----------|----------|----------|----------|----------|----------|----------|
| 0 | 808635 | 71 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 |
| 1 | 941866 | 71 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 | 3.581679 |

2 rows × 21 columns

Double-click (or enter) to edit

```

1 # prepare x_train and y_train
2 x_train = reg_train.drop(['user', 'movie', 'rating'], axis=1)
3 y_train = reg_train['rating']
4
5 # prepare test data
6 x_test = reg_test_df.drop(['user', 'movie', 'rating'], axis=1)
7 y_test = reg_test_df['rating']
8
9
10
11 xgb_final = xgb.XGBRegressor(n_jobs=10, random_state=15)
12 train_results, test_results = run_xgboost(xgb_final, x_train, y_train, x_test, y_test)
13
14 # store the results in models_evaluations dictionaries
15 models_evaluation_train['xgb_final'] = train_results
16 models_evaluation_test['xgb_final'] = test_results
17
18
19 xgb.plot_importance(xgb_final)
20 plt.show()

```



Training the model..

Done. Time taken : 0:00:04.203252

Done

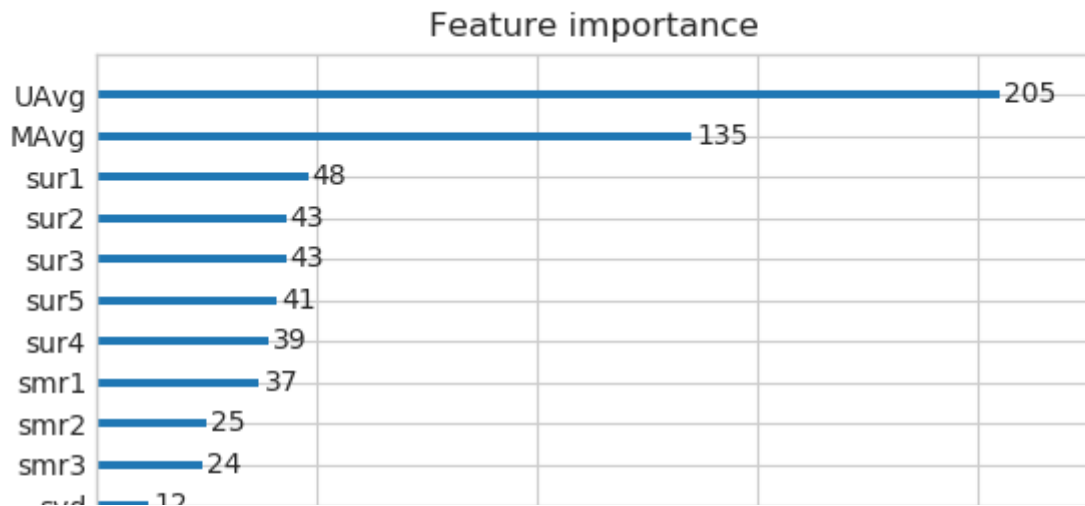
Evaluating the model with TRAIN data...

Evaluating Test data

TEST DATA

RMSE : 1.0763580984894978

MAPE : 34.487391651053336



4.4.8 XgBoost with Surprise Baseline + Surprise KNNbaseline + MF Techniques

```

1 # prepare train data
2 x_train = reg_train[['knn_bsl_u', 'knn_bsl_m', 'svd', 'svdpp']]
3 y_train = reg_train['rating']
4
5 # test data
6 x_test = reg_test_df[['knn_bsl_u', 'knn_bsl_m', 'svd', 'svdpp']]
7 y_test = reg_test_df['rating']
8
9
10 xgb_all_models = xgb.XGBRegressor(n_jobs=10, random_state=15)
11 train_results, test_results = run_xgboost(xgb_all_models, x_train, y_train, x_test, y_
12
13 # store the results in models_evaluations dictionaries
14 models_evaluation_train['xgb_all_models'] = train_results
15 models_evaluation_test['xgb_all_models'] = test_results
16
17 xgb.plot_importance(xgb_all_models)
18 plt.show()

```



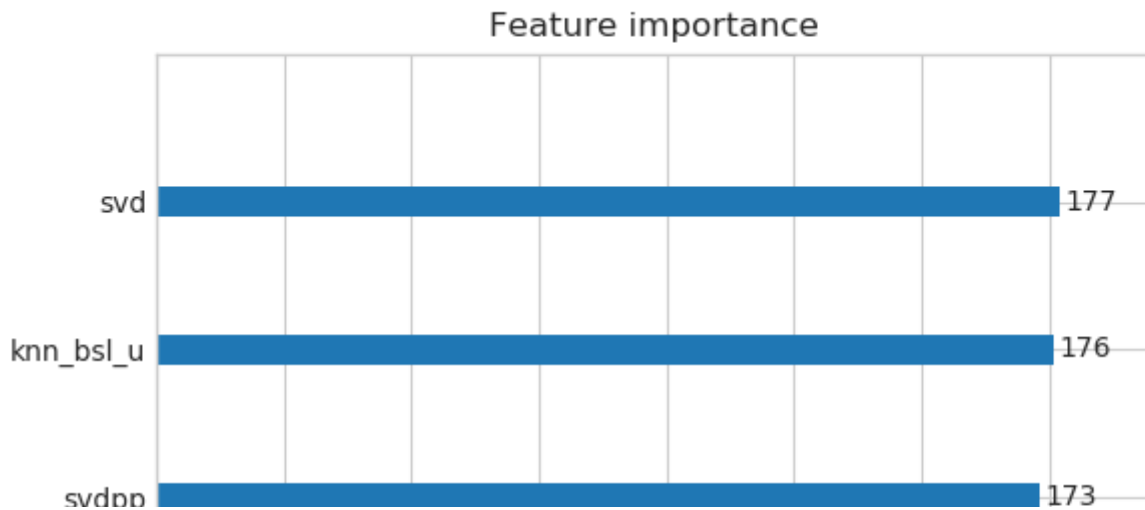
Training the model..
 Done. Time taken : 0:00:01.292225

Done

Evaluating the model with TRAIN data...
 Evaluating Test data

TEST DATA

 RMSE : 1.075480663561971
 MAPE : 35.01826709436013



Double-click (or enter) to edit

Double-click (or enter) to edit

4.5 Comparision between all models

```
1 # Saving our TEST_RESULTS into a dataframe so that you don't have to run it again
2 pd.DataFrame(models_evaluation_test).to_csv('sample/small/small_sample_results.csv')
3 models = pd.read_csv('sample/small/small_sample_results.csv', index_col=0)
4 models.loc['rmse'].sort_values()
```

```
svd                1.0726046873826458
knn_bsl_u          1.0726493739667242
knn_bsl_m          1.072758832653683
svdpp              1.0728491944183447
bsl_algo           1.0730330260516174
xgb_knn_bsl_mu     1.0753229281412784
xgb_all_models     1.075480663561971
first_algo         1.0761851474385373
xgb_bsl            1.0763419061709816
xgb_final          1.0763580984894978
xgb_knn_bsl        1.0763602465199797
Name: rmse, dtype: object
```

Double-click (or enter) to edit

```

1 print("-"*100)
2 print("Total time taken to run this entire notebook ( with saved files) is :",datetime

```



Total time taken to run this entire notebook (with saved files) is : 0:42:08.30276

5. Assignment

1.Instead of using 10K users and 1K movies to train the above models, use 25K users and 3K movies (or more) to train all of the above models. Report the RMSE and MAPE on the test data using larger amount of data and provide a comparison between various models as shown above.

NOTE: Please be patient as some of the code snippets make take many hours to complete execution.

2.Tune hyperparameters of all the Xgboost models above to improve the RMSE.

```

1  %%javascript
2  // Converts integer to roman numeral
3  // https://github.com/kmahelona/ipython_notebook_goodies
4  // https://kmahelona.github.io/ipython_notebook_goodies/ipython_notebook_toc.js
5  function romanize(num) {
6      var lookup = {M:1000,CM:900,D:500,CD:400,C:100,XC:90,L:50,XL:40,X:10,IX:9,V:5,IV:4
7      roman = '',
8      i;
9      for ( i in lookup ) {
10         while ( num >= lookup[i] ) {
11             roman += i;
12             num -= lookup[i];
13         }
14     }
15     return roman;
16 }
17
18 // Builds a <ul> Table of Contents from all <headers> in DOM
19 function createTOC(){
20     var toc = "";
21     var level = 0;
22     var levels = {};
23     $('#toc').html('');
24
25     $(':header').each(function(i){
26         if (this.id=='tocheading'){return;}
27
28         var titleText = this.innerHTML;
29         var openLevel = this.tagName[1];
30
31         if (levels[openLevel]){
32             levels[openLevel] += 1;
33         } else{
34             levels[openLevel] = 1;
35         }
36
37         if (openLevel > level) {
38             toc += (new Array(openLevel - level + 1)).join('<ul class="toc">');
39         } else if (openLevel < level) {
40             toc += (new Array(level - openLevel + 1)).join('</ul>');
41             for (i=level;i>openLevel;i--){levels[i]=0;}
42         }
43
44         level = parseInt(openLevel);
45
46         if (this.id==''){this.id = this.innerHTML.replace(/ /g,"-")}
47         var anchor = this.id;
48
49         toc += '<li><a style="text-decoration:none", href="#" + encodeURIComponent(anchor
50
51     });
52 }

```

```
53  
54  
55     if (level) {  
56         toc += (new Array(level + 1)).join("</ul>");  
57     }  
58  
59  
60     $('#toc').append(toc);  
61  
62 };  
63  
64 // Executes the createToc function  
65 setTimeout(function(){createTOC();},100);  
66  
67 // Rebuild to TOC every minute  
68 setInterval(function(){createTOC();},60000);
```

