MACHINE LEARNING PROJECT

Nakul Ramanathan (2016168)
Sanchit Malhotra (2016264)

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

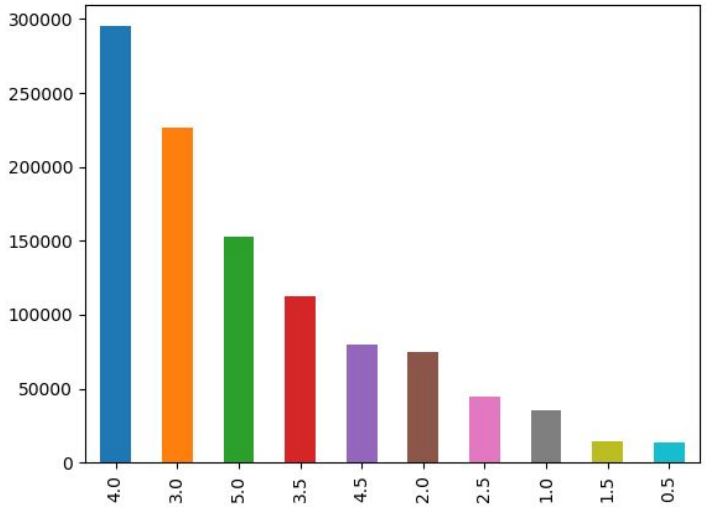
Problem Statement

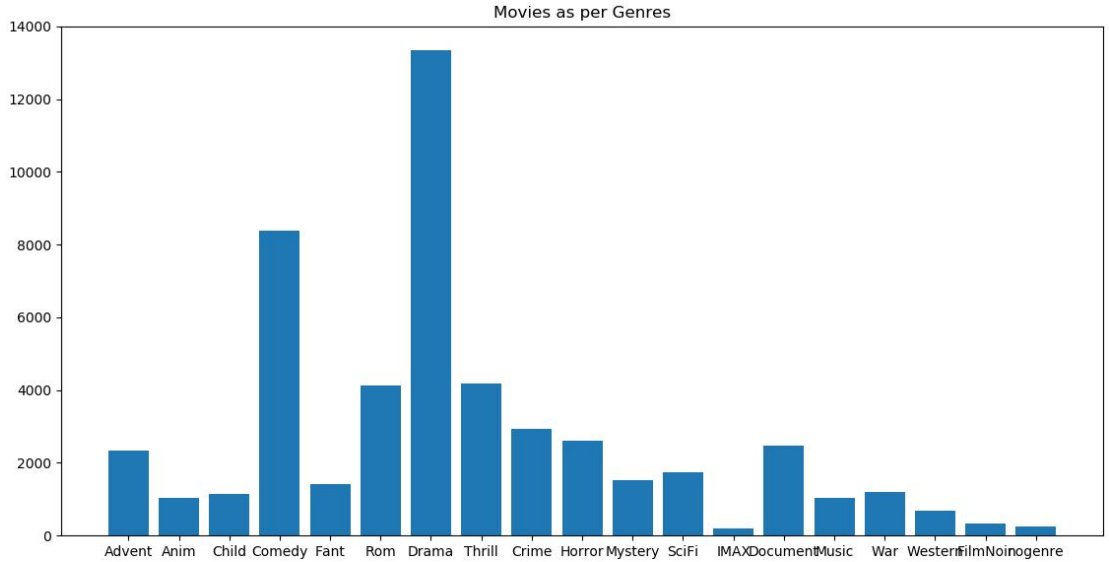
The key to good sales is understanding the customer’s needs. Thus, recommendation systems have lots of use cases in online businesses such as Amazon, Netflix etc..

We aim to make a recommender system which suggests movies using Content Based Filtering and Collaborative Filtering .

Dataset

The dataset used for baseline and the second advanced algorithm can be accessed here.





Approaches Used

Approaches Used :

- User Based Collaborative Filtering (k-NN) {Baseline}
  (using distance in user vector to find the nearest similar user and recommend his high rated movies)
- Matrix Factorization (SVD) {Advanced}
  (using the user-item interaction matrix to determine latent features)
- Content Based Filtering (SVM Classifier) {Advanced}
  (for each movie we fit a separate SVM)

Content Based Filtering (SVM Classifier) {Advanced}

- We treat each movie as a different SVM problem. For each movie there is 1 multi-class SVM.
- The classes of each SVM are -1, 0 and 1
- For each movie we remove the users who have not rated that movie. We assign the rating of the user as the output labels and consider the rest of the ratings of the user as inputs. We then use an SVM classifier to classify.
- We further predict the ratings of the movie based on the ratings of previous movies.

Results

Baseline

	Size of Dataset	
Metric Used	1,00,000	1 million
Root Mean Squared Error	0.9722	0.9253
Mean Absolute Error	0.7676	0.73
Recall	0.58	0.679
Accuracy (Test Set)	0.56	0.601
Precision	0.533	0.59

Advanced (Matrix Factorization)

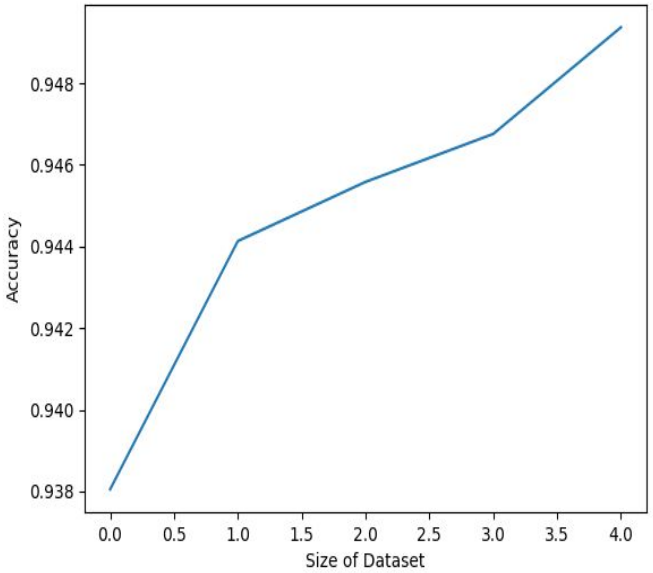
	Size of Dataset	
Metric	100,000	1 million
Root Mean Square Error	0.9398	0.8786
Mean Absolute Error	0.7415	0.6897
Recall	0.564	0.634
Accuracy(Test Set)	0.586	0.622
Precision	0.57	0.61

Advanced (SVM)

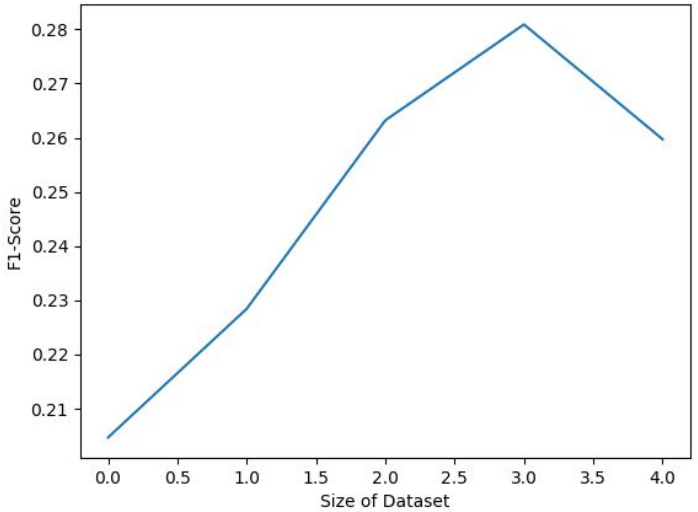
	SVM	
Metrics	RBF Kernel	Linear Kernel
Recall	0.1957	0.3499
Accuracy Values	0.947	0.943
Precision	0.673	0.497

Results

1



2



SVM Linear kernel on increasing size of data set.

Analysis

- From figure 1 it can be observed that accuracy increases as we increase the number of movies in the dataset. But it can also be seen that increase in accuracy is negligible so the model doesn’t really learn anything extra.
- From figure 2 we can see that F1 score or the balance between recall and precision is increasing at a steady rate until 900 movie at which it drops at a sudden rate. This is generally observed when the data contains skewed ratings and is irregular.

Analysis

- The Accuracy of SVM turns out to be very high in comparison to k-NN. But during the training of the models SVM was slow in comparison to k-NN.
  For SVM : While the accuracy of RBF and Linear kernel is almost same, when RBF is used we get overall high precision in comparison to Linear kernel whereas we get high recall in Linear kernel.
- While training the model using 1 million data points, we get more accuracy in comparison to 100k data points. This is because the density of the users increases in the same area in the n dimensional plane.

Analysis

- We tried using TF-IDF (Relevance Feedback) to train a content based type algorithm. The output of this could not be interpreted using mathematical metrics such as accuracy, precision etc. as relevance is a subjective term.
- The convergence of different algorithms was highly affected by the size of the dataset.
- Though SVM has the higher accuracy, but it is computationally expensive.

Contribution

Nakul Ramanathan - k-NN implementation, Midterm Review ppt, TF-IDF

Sanchit Malhotra - SVM implementation, Model Analysis and Validation

Advanced algorithm 1 was implemented equally by both.

Some Help Taken from:

1)https://github.com/alexvlis/movie-recommendation-system

2)Surprise library