NETFLIX RATING PREDICTION SYSTEM

Shupeng Niu
Kartikeya Pharasi
Alok Kucheria

Background

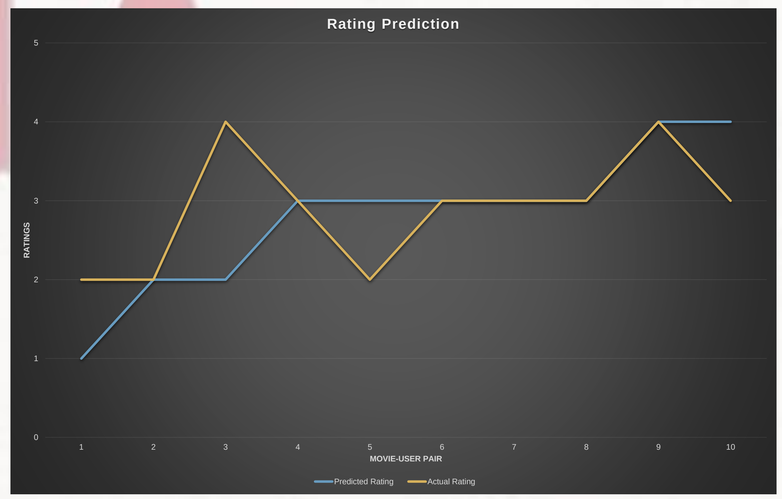
- The Netflix Prize was a contest Netflix sponsored, which sought to substantially improve the accuracy of predictions about how much someone would enjoy a particular movie and rate it based on their previous movie preferences.

Objective

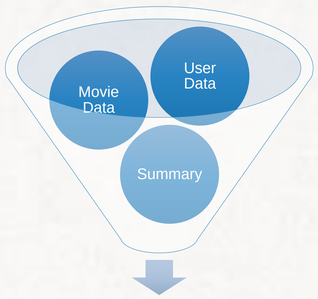
- Our objective is to make a rating prediction for a movie by a user based on similar movies rated by similar users. This can be achieved by using clustering for similar users and movies and applying prediction techniques on the newly formed clusters.

Results

- According to our customized neighboring algorithm we have tried to predict the rating based on clusters created using k-means.
- For Cinematch, the Netflix algorithm, the RMSE is 0.95. The winning team had an RMSE of 0.85.
- Using the subset of the entire dataset, we applied the Leave one out Cross Validation technique. For multiple runs of 10 random movie-user pairs, the classifier gave RMSE varying from 0.84 to 1.45, mostly close to 0.93.
- As we increase the size of the training set and refine cluster sizes a much better prediction is derived.



Methodology

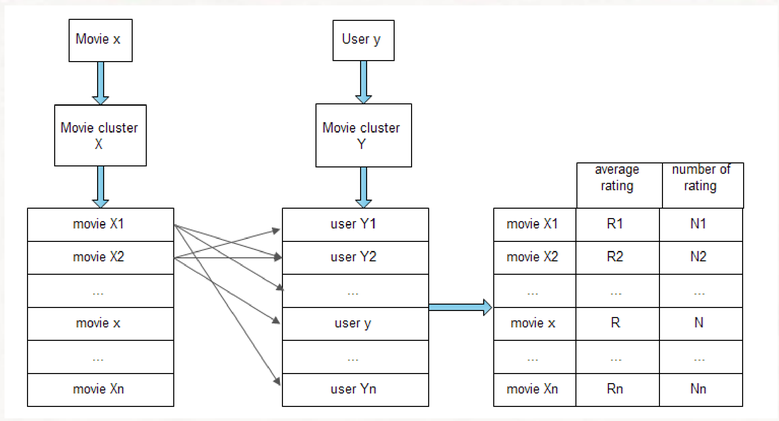


Data Extraction

Parameter Selection
- Selection of parameters from the above results.

K-Means Clustering
- Our data is well suited for k-means clustering. We create user clusters and movie clusters.

Prediction Algorithm
- Based on the cluster averages and weightage.



- **Predicted Rating** = $\frac{R1*N1 + R2*N2 + \ldots + Rn*Nn}{N1+N2+\cdots+Nn} + \alpha$

  $= \frac{\sum_1^n RN}{\sum_1^n N} + \alpha$

- **Where α is the correction factor calculated as**

  $\alpha = \frac{(r1-R1)N1+(r2-R2)N2+\cdots(rn-Rn)Nn}{N1+N2+\cdots Nn}$

  $= \frac{\sum_1^n (rn-Rn)Nn}{\sum_1^n Nn}$

- Here, *rn* is only for the movies which have been rated by our user for whom we need a movie rating prediction.

Conclusion

For every movie and user, we have numerous ratings using which we create groups of similar users and movies. We use the k-means method as it best suits the type of data we have. We use a variation of the KNN method to make a prediction taking into consideration all relevant factors.

Based on the above data, its analysis and the results obtained, we can say that our Netflix Prediction System will work with respectable levels of accuracy. We can further improve this method by adding parameters in the prediction algorithm as we learn more about user behavior.