Case study: Movie Recommendation Engine

Netflix, Inc. is an American media services provider. The company's primary business is its subscription-based streaming service, which offers online streaming of a library of films and television programs including those produced in-house. From 2006 through 2009, Netflix ran a contest asking the public to submit algorithms to predict user ratings for movies. This algorithm would be useful for Netflix when making recommendations to users. Netflix provided a training data set of about 100 million user ratings and a test data set of about three million user ratings.

They offered a grand prize of one million dollars to the team who could beat Netflix's current algorithm, called Cinematch, by more than 10% measured in terms of root mean squared error. Even if the grand prize was not yet reached, progress prizes of $50,000 per year was awarded for the best result, as long as it was at least a 1% improvement over the previous year. The contest was hugely popular all over the world. By June 2007, over 20,000 teams had registered from over 150 countries.

You can refer to the contest website here: https://www.netflixprize.com/index.html

Variables used in this study

We are using the data obtained from a website called movielens.org, you can find out more using this link https://movielens.org/info/about.

In our movie lens dataset, we have binary vectors for each movie, classifying that movie into genres. In total we have 19 different genres. The movie Toy Story is categorized as an animation, comedy, and children's movie. So the data for Toy Story has a 1 in the spot for these three genres and a 0 everywhere else. The movie Batman Forever is categorized as an action, adventure, comedy, and crime movie. So Batman Forever has a 1 in the spot for these four genres and a 0 everywhere else.

Exercise:

1) Read the data in R using appropriate commands
2) Study the data structure
3) Add headers to the data
4) Remove unwanted variables
5) Plot the data to study clusters
6) Limiting the number of clusters to 10 form a table of mean cluster ratings by each genre
7) Discuss how this model serves its purpose

Conclusion:

The contest went live on October 2, 2006. By October 8, only six days later, a team submitted an algorithm that beat Cinematch. A week later, on October 15, there were three teams already submitting algorithms beating Cinematch.

The 2007 progress prize went to a team called BellKor, with an 8.43% improvement over Cinematch. In 2008, the progress prize again went to team BellKor. But this time, the team included members from the team BigChaos in addition to the original members of BellKor. This was the last progress prize because another 1% improvement would reach the grand prize goal of 10%. On June 26, 2009, the team BellKor's Pragmatic Chaos, composed of members from three different original teams, submitted a 10.05% improvement over Cinematch, signaling the last call for the contest.

Other teams had 30 days to submit algorithms before the contest closed. These 30 days were filled with intense competition and even more progress. 29 days after last call was announced, on July 25, 2009, the team The Ensemble submitted a 10.09% improvement, beating the 10.05% improvement that was submitted by Bellkor's

But by the time Netflix stopped accepting submissions, the next day, Bellkor's Pragmatic Chaos had also submitted a 10.09% improvement, and The Ensemble had submitted a 10.10% improvement.

To really test the algorithms, Netflix tested them on a private test set that the teams had never seen before. This is the true test of predictive ability. On September 18, 2009, Netflix announced that the winning team was Bellkor's Pragmatic Chaos.

They won the competition and the $1 million grand prize!