DBMS Hackathon Report

Devansh Agarwal(ES16BTECH11009)

Deep Diwani(EE16BTECH11006)

Observations in Dataset:

- Most the data set(10 million rows) belong to the year 2005
- Half of the consumers have given rating less than 20 times
- Most of the ratings are for a small number of products.
- In year 2005 all the products except 1 has been alloted a rating.
- Rating 4 occurs the most commonly.
- Year 1999 has only around 400 ratings given.

Methods Tried

1. Building similarity matrix between different products using standard methods, but this method seems to be taking more than 10 hrs to build the similarity matrix, here we were only tried to use 1 million points training points for creating the similarity matrix.
2. So we took the top 100 products from the test data based on their count and then tried to build the similarity matrix only for these products (100 x 4661 size) but here we needed to use all the 20 million points because the similarity with many products was coming zero due to their absence in the 1 mil dataset that we were using above. This was also taking a lot of time to create so we were not able to build it.
3. Then we used the mean of rating for each product to predict their ratings.
4. We also used the mode of rating for each product to predict their ratings.
5. We also tried finding probability of each rating to occur and use that distribution to do rating prediction.