Netflix Prize: Problem Statement

- Who Rated What in 2006?
- Training data from pre-2006: user, movie, date
- ~17,000 movies, 480,000 users
  - ~1.17% rated user/movie pairs
- Model performance measured with RMSE

Features: NMF

Decompose the movie-user matrix and minimize squared error between R and R-hat using gradient descent.

	U1	U2	U3	U4
M1	1	1	-	1
M2	1	-	-	1
M3	1	1	-	1
M4	1	-	-	1
M5	-	1	1	1

$$\mathbf{R} \approx \mathbf{P} \times \mathbf{Q}^T = \hat{\mathbf{R}}$$

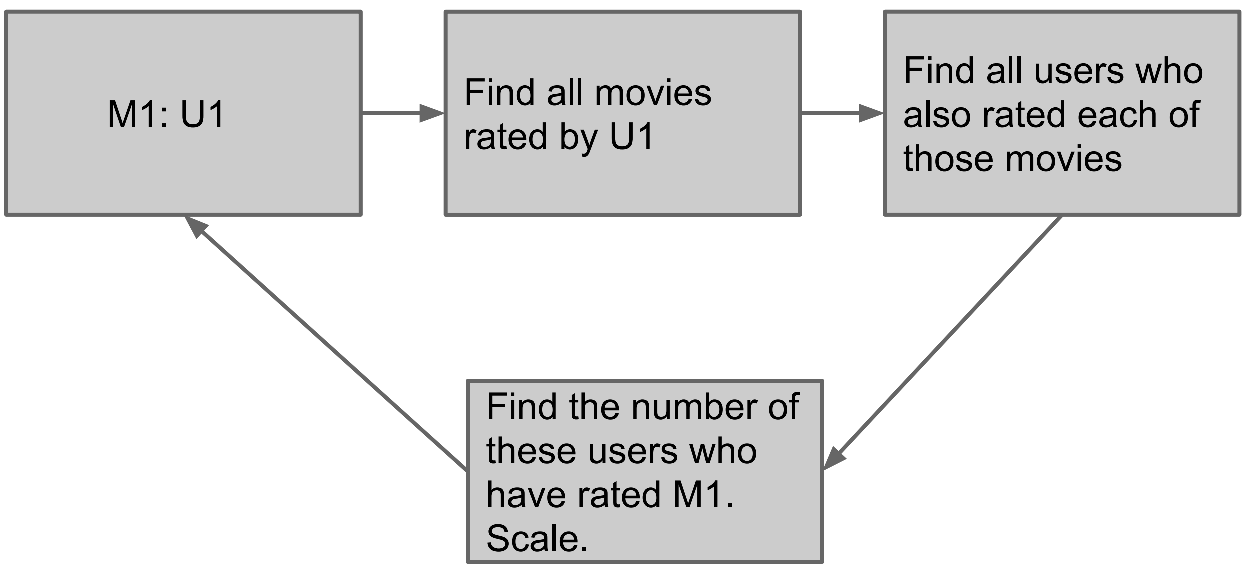
$$\hat{r}_{ij} = p_i^T q_j = \sum_{k=1}^{k} p_{ik} q_{kj}$$

$$e_{ij}^2 = (r_{ij} - \hat{r}_{ij})^2 = (r_{ij} - \sum_{k=1}^{K} p_{ik} q_{kj})^2$$

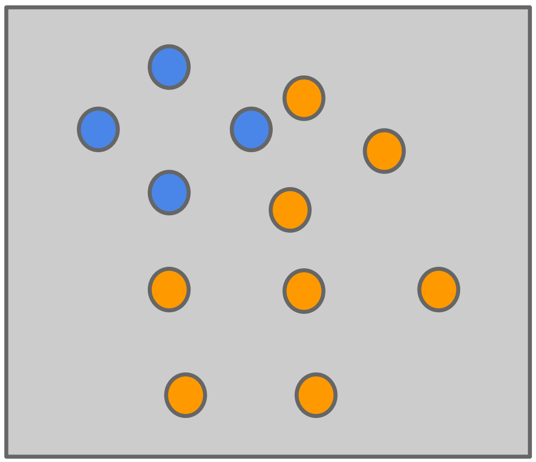
Minimize!

Features: Cosine Similarity

Given a user:movie pair



Features: Global Effects



User and Movie Averages:
For a given movie, the movie average is the number of times the movie has been rated divided by the total number of movies.

i.e. Blue movie average is 4/12

User and Movie Time Effects:
Adds information about the age of the movie and how long the user has been rating movies

Overall Rating Probability + User and Movie Averages + User and Movie Time Effects = Global Effects Feature

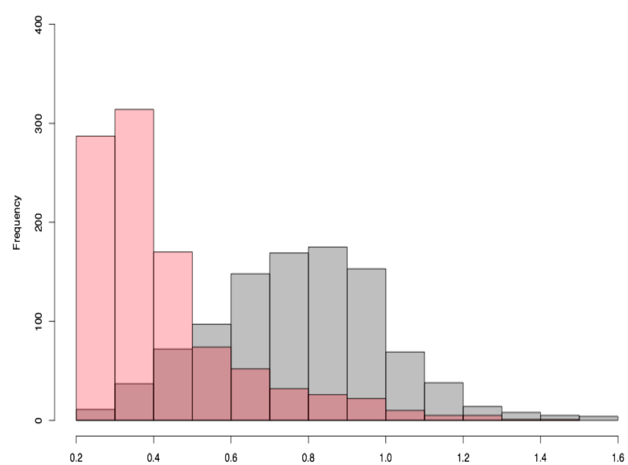
Model Process

Feature Engineering → Transform to "probabilities" → Logistic Regression → Prediction

- NMF
- Cosine similarity
- Global effects

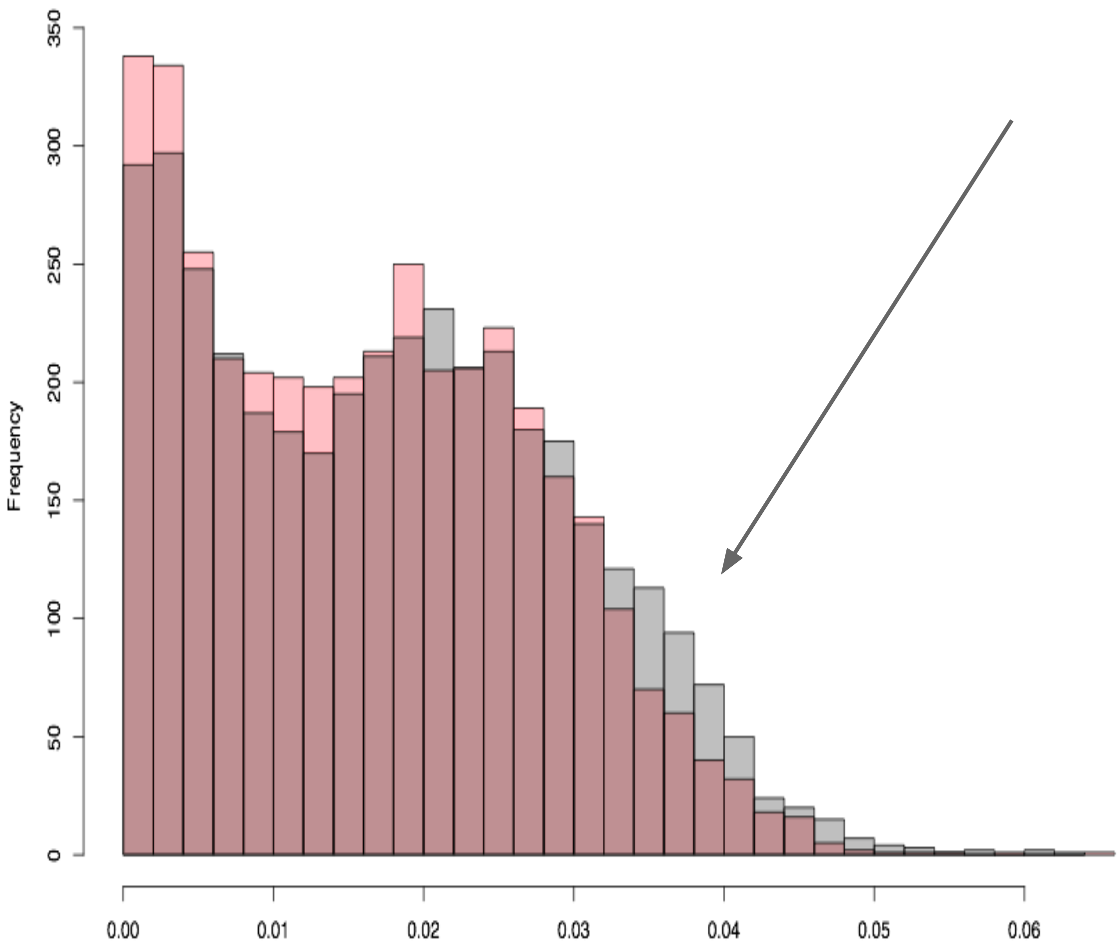
- Feature probabilities

Using the validation set, we assess likelihood of being rated or unrated based on feature value.



Less likely to be rated →

Distribution for Cosine Similarity



We should be able to find differential distributions for non-ratings (pink) versus ratings (grey)

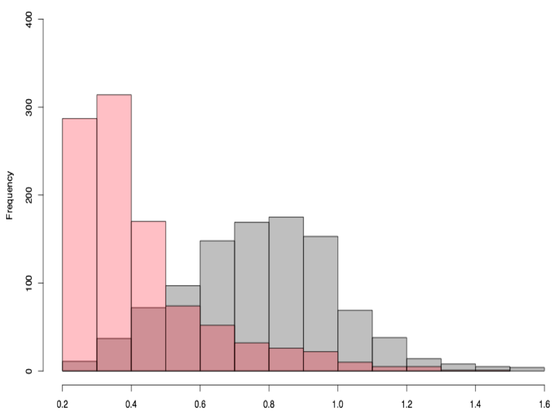
Results

Baseline	Global Effects	NMF	Cosine Sim	RMSE
----				.2684
----	-----			.2755
----		----		.2683
----			----	.2724
----	----	----	----	.2754

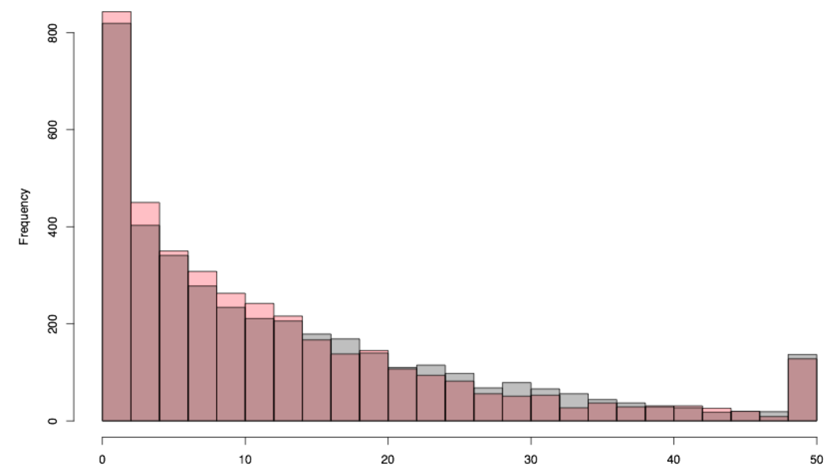
Conclusions

1) Understand how the testing set is created and model validation set accordingly
2) Feature engineering is the bulk of the work, but it doesn’t always pay off— predicting all zeros is a good model here!
3) Work further on using packages and refining models.

Importance of correct validation set



vs



Picked validation set using uniformly selected random pairs.

Feature looks good! :)

Picked validation set weighted by frequency of appearance in the ratings as was done for the test set.

Features looks not so good :(

Conclusions (cont.)

2) Feature engineering is the bulk of the work, but it doesn’t always pay off— predicting all zeros is a good model here!
3) Work further on using packages and refining models, especially using logistic regression as a second predictive model.