DS 5110

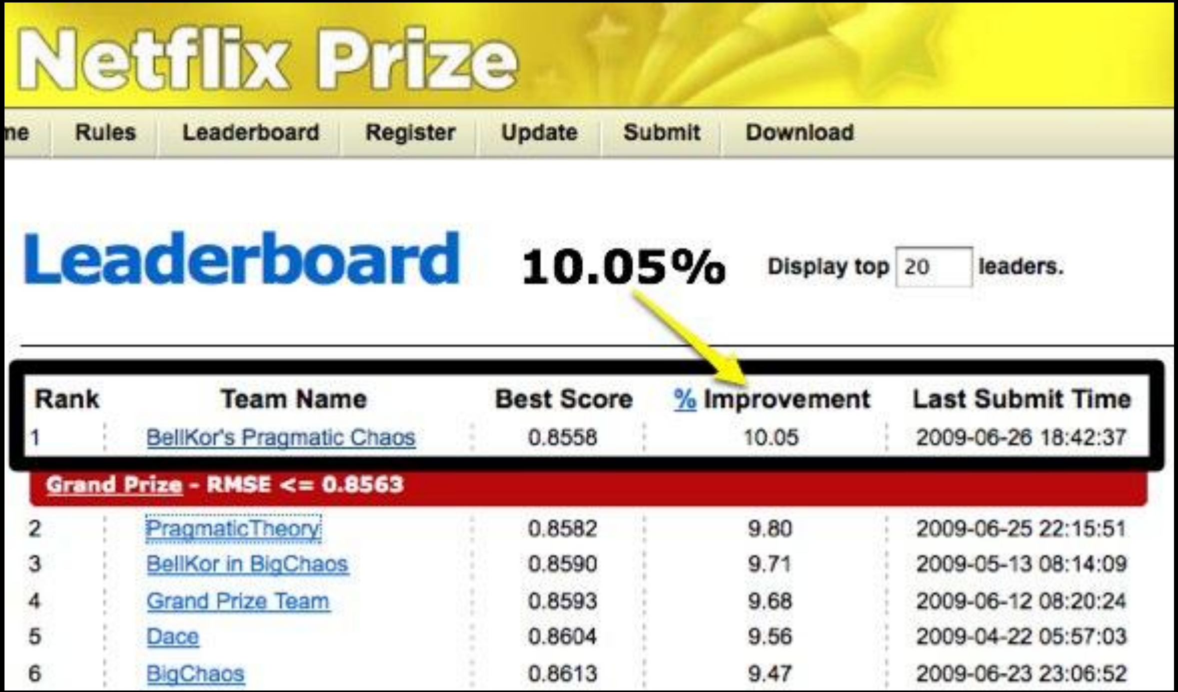
2021 TV-PG

In 2006, Netflix charged participants in the Netflix Prize Challenge with obtaining an RMSE (root mean squared error) value of Netflix's then-best 0.9525 or less on a subset of a data set of 100MM+ ratings—ideally reducing it to 0.8572 or less (a 10% reduction in the accuracy of Netflix's existing system, Cinematch)—, with the prize being awarded to the team that achieved the lowest RMSE on the remaining observations in the "qualifying" subset (Töscher et al. 2009). The objective of this project was to create a model that could replicate or best the top-performing submission to the original Netflix Prize challenge using the same data.

▸ PLAY + MY LIST

Starring:
Lauren Neal *(ln9bv)*
Melanie Sattler *(ms9py)*
Nicholas Thompson *(nat3fa)*
Nima Beheshti *(nb9pp)*
Genre: Data Science
Studio: UVA School of Data Science

PROBING THE NETFLIX PRIZE:
Recommended Algorithms for Recommender Systems

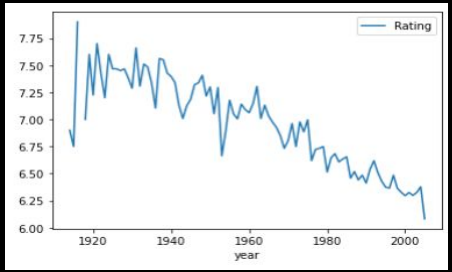


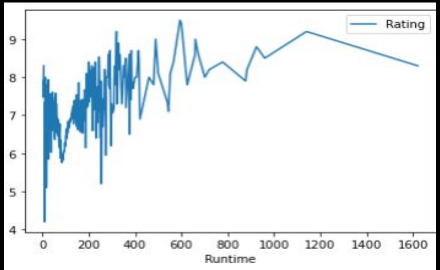
Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	PragmaticTheory	0.8582	9.80	2009-06-25 22:15:51
3	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
4	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
5	Dace	0.8604	9.56	2009-04-22 05:57:03
6	BigChaos	0.8613	9.47	2009-06-23 23:06:52

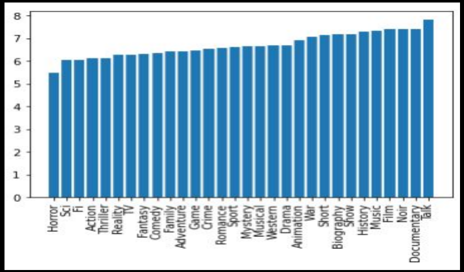
Data Summary

Season 1

```
13368:
2385003,4,2004-07-08
659432,3,2005-03-16
751812,2,2002-12-16
2625420,2,2004-05-25
1650301,1,2005-08-30
2269227,4,2005-10-27
2220672,4,2002-08-19
```







Data Structure and Size

- The image shows 7 reviews for Movie ID 13368.
- Each review has a User ID, Rating, and Date Rated.
- The dataset spans 17,770 titles and 100MM+ reviews.
- This data alone was 2GB.

Supplemental Movie Data

- Expanded upon our current dataset with supplemental data.
- We could tie each Movie ID to its: Year, Title, Runtime, Average Rating, Directors, Writers, Production Companies, and Genres.

Data Analysis

- Performed additional analysis on distribution of ratings
- Average Rating vs Year
- Average Rating vs Runtime
- Average Rating vs Genre

Difficulties with Additional Data

- Genre column had to be expanded due to initial format
- Adding additional features to our current 100,000,000 data points posed a problem
- Total data size expanded beyond allowed limit (15GB)

Data Clean-up

```
1:
1488844,3,2005-09-06
822109,5,2005-05-13
885013,4,2005-10-19
30878,4,2005-12-26
823519,3,2004-05-03
893988,3,2005-11-17
```

```python
df_file1 = df_file1[pd.notnull(df_file1['rating'])]

# Add the movie_id column
df_file1['movie_id'] = movie_np_1.astype(int)
df_file1['user_id'] = df_file1['user_id'].astype(int)
print(df_file1.columns)

print(df_file1.iloc[::5000000,:])

new_cols = df_file1.columns.tolist()
new_cols = new_cols[:1]+new_cols[-1:]+new_cols[1:2]
df_file1 = df_file1[new_cols]

df_file1.to_csv("processed_pt1.txt", encoding='utf-8', index=False)
```

	user_id	rating	movie_id
1	1488844	3.0	1
2	822109	5.0	1
3	885013	4.0	1
4	30878	4.0	1
5	823519	3.0	1

- Data was complete (no missing values, although not every user had viewed/rated every title: ultimately resulting in a **sparse user-item** requiring collaborative filtering / non-negative matrix factorization)
- Had to combine files containing only portions of the reviews together into one complete file
  - 17,000+ movie ids
  - 100MM+ unique ratings from 480,189 users
- Cleaned up with loop: extracted each title's unique ID and placed in a new column (**movie_id**) column with rating and user id, then concatenated the resulting dataframes

Models & Selected Results

Linear Regression

Model	Train/Test Split	ElasticNetParam	MSE	RMSE	R^2
Linear Regression	80/20	0.8	1.177682	1.3869	-5.2653E-9
Lasso Regression	80/20	1.0	1.177682	1.3869	-5.2653E-9
Ridge Regression	80/20	0.0	1.1776	1.3867	7.15E-5

K-Means

k_value	Precision	Recall	F1Score
2	0.8469857478842945	0.5499432240886487	0.666882972998638 9
5	0.07027559450821155	0.6633776091081593	0.12708799098460477
10	0.01692564375741251	0.622320768662232 1	0.03295499021526419
15	0.0179307294912256	0.7181964573268921	0.03498793857498676
20	0.04719882605986290 6	0.662528216704289	0.08811994520650766

Linear Regression

- Started out to get a baseline
- Did not perform well
    - Insignificant R^2
- Research indicated does not perform well with recommender type data

K-Means

- Classify ratings > 3.0 as “recommended” (1) and the remainder as “not recommended” (0)
- Model did not perform well
- Small F-Scores
- Low Precision

ALS Prediction

- Best model for this type of data
- Best RMSE of 0.8526 (rank = 15, alpha = 0.01, 75/25 train/test split) barely bested the Netflix Prize-winning RMSE of 0.8558

ALS

[9]:

	rank	MSE	RMSE
0	5	0.747698	0.864695
1	10	0.727663	0.853032
2	15	0.726944	0.852610
3	20	0.732890	0.856090

Conclusion

Results/Conclusion

- Overall Linear and Classification Models performed poorly for prediction
- ALS model performed the best for prediction
- ALS model expanded to give recommendation based on rating

Future Work

- Expand with more data
  - Budget
  - Box office
  - PR Campaign
  - Social Media exposure
  - News Discussion
- Audience type
  - Children Movie
  - Teen
  - Adult