Build a Better Netflix, Win a Million Dollars?

Lester Mackey
August 10, 2014
• Rents & streams movies and TV shows
• 100,000 movie titles
• 26 million customers

Recommends “Movies You’ll ♥”
Recommending Movies You’ll ♥

Rate what you’ve seen to discover suggestions for you

- Mrs. Doubtfire: Haven’t Seen It
- Iron Man 2: Hated it!
- Food, Inc.: Loved it!
Recommending Movies You’ll ♥

Visually-striking Sci-Fi & Fantasy

Feel-good TV Shows

Documentaries
Recommending Movies You’ll ♥

Sci-Fi & Fantasy

Inception

2010  PG-13  148 minutes

Dom Cobb earns a tidy sum infiltrating the dreams of corporate titans to steal their most closely held secrets.

Starring: Leonardo DiCaprio, Joseph Gordon-Levitt

Director: Christopher Nolan

Genre: Sci-Fi & Fantasy

Availability: DVD and Blu-ray

4.4  Our best guess for LESTER

Recommended based on your interest in: Batman Begins, The Matrix and Memento
How This Works

Top Secret Computer Program

Cinematch

Your Predicted Rating:

I don’t ♥ this movie.

[Images of movies and characters, indicating a rating system]

How This Works

Top Secret

Computer Program

Cinematch

Your Predicted Rating:

I don’t ♥ this movie.

[Images of movies and characters, indicating a rating system]
Back at Netflix

How can we improve Cinematch?

Let’s have a contest!

What should the prize be?

How about $1 million?
The Netflix Prize

October 2, 2006
- Contest open to the world
- 100 million movie ratings released to public
- **Goal:** Create computer program to predict ratings
  - **$1 Million** Grand Prize for beating Cinematch accuracy by 10%
  - **$50,000** Progress Prize for the team with the best predictions each year

5,100 teams from 186 countries entered
Dinosaur Planet

David Lin

David Weiss

Lester Mackey

Team Dinosaur Planet
The Ratings

• Training Set
  – What computer programs use to learn customer preferences
  – Each entry:
    – 100,500,000 ratings in total
    – 480,000 customers and 18,000 movies

July 5, 1999
The Ratings: A Closer Look

### Highest Rated Movies

<table>
<thead>
<tr>
<th>Movie</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The Shawshank Redemption</td>
<td></td>
</tr>
<tr>
<td>Lord of the Rings: The Return of the King</td>
<td></td>
</tr>
<tr>
<td>Raiders of the Lost Ark</td>
<td></td>
</tr>
<tr>
<td>Lord of the Rings: The Two Towers</td>
<td></td>
</tr>
<tr>
<td>Finding Nemo</td>
<td></td>
</tr>
<tr>
<td>The Green Mile</td>
<td></td>
</tr>
</tbody>
</table>

### Most Divisive Movies

<table>
<thead>
<tr>
<th>Movie</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fahrenheit 9/11</td>
<td></td>
</tr>
<tr>
<td>Napoleon Dynamite</td>
<td></td>
</tr>
<tr>
<td>Pearl Harbor</td>
<td></td>
</tr>
<tr>
<td>Miss Congeniality</td>
<td></td>
</tr>
<tr>
<td>Lost in Translation</td>
<td></td>
</tr>
<tr>
<td>The Royal Tenenbaums</td>
<td></td>
</tr>
</tbody>
</table>
How the Contest Worked

• Quiz Set & Test Set
  – Used to evaluate accuracy of computer programs
  – Each entry:

• Each team predicts Quiz Set and Test Set ratings once per day
• Netflix displays Quiz score on public Leaderboard

Rating unknown! Sept. 9, 2006
<table>
<thead>
<tr>
<th>Team Name</th>
<th>Best Score</th>
<th>% Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Grand Prize candidates yet</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Grand Prize - RMSE &lt;= 0.8563</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Thought Gang</td>
<td>0.9413</td>
<td>1.06</td>
</tr>
<tr>
<td>Progress Prize 2007 - RMSE &lt;= 0.9419</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wxyzconsulting.com</td>
<td>0.9430</td>
<td>0.88</td>
</tr>
<tr>
<td>Sparkling_Destiny</td>
<td>0.9488</td>
<td>0.27</td>
</tr>
<tr>
<td>Cinematic score on quiz subset - RMSE = 0.9514</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline0</td>
<td>0.9525</td>
<td>-0.12</td>
</tr>
<tr>
<td>CodeMonkey</td>
<td>0.9571</td>
<td>-0.60</td>
</tr>
<tr>
<td>Bjornson</td>
<td>0.9648</td>
<td>-1.41</td>
</tr>
<tr>
<td>jsnell</td>
<td>0.9670</td>
<td>-1.64</td>
</tr>
</tbody>
</table>
How the Contest Worked

• Quiz Set & Test Set
  – Used to evaluate accuracy of computer programs
  – Each entry:

• Each team predicts Quiz Set and Test Set ratings once per day
• Netflix displays Quiz score on public Leaderboard
• Test score is hidden – but best Test score wins!
  – 10% improvement ➞ $1 Million Grand Prize
  – Most improvement in 1 year ➞ $50,000 Progress Prize

Rating unknown!
Sept. 9, 2006
A First Approach: Clustering

- Divide users (or movies) into groups based on similarities
  - Use group information to predict user ratings
    - e.g. The average action-lover gives Indiana Jones a 5
  - Hard clustering: each user belongs to a single cluster
  - Soft clustering: each user fractionally belongs to all clusters
Clustering with Missing Data

• Centroid-based clustering
  – Represent user by incomplete ratings vector, $r_u$
    \[ r_u = (1, 5, ?, ?, 3, ?, 4) \]
  – Represent cluster by centroid vector, $c_k$
    • Typically, $c_k$ is average of user vectors in cluster $k$
  – Minimize (estimated) distance between users and their cluster centers

• Result: \textbf{-0.3\%} improvement over Cinematch
Matching Cinematch

• Incorporate prior information
  – Positive ratings {3,4,5} vs. negative ratings {1,2}
    • Estimate \( E[r| r \geq 3], E[r| r < 3], P(r < 3) \) and combine
  – Ordinal nature of rating data
    • Estimate \( P(r < t) \) for \( t \in \{2, 3, 4, 5\} \) and combine
  – Result: 0.5% improvement over Cinematch
Training on Errors

Recurring theme

• Train one model to predict and hence correct the errors of another model
• Long history in statistics and machine learning
  – Tukey's twicing (1977)
  – Boosting (Schapire, 1990)
  – Gradient boosting (Friedman, 1999)
• e.g., Cluster on errors of clustering predictions
Clustering on Errors

Result: 3.0% improvement over Cinematch
The Three Pillars: Nearest Neighbors

Nearest Neighbor Rule

- Find customer with the most similar ratings
- Use her rating as best guess for new guy’s rating
The Three Pillars: Nearest Neighbors

<table>
<thead>
<tr>
<th></th>
<th>Nearest Neighbor Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Find movie with the most similar ratings</td>
</tr>
<tr>
<td></td>
<td>Use its rating as best guess for new guy’s rating</td>
</tr>
</tbody>
</table>

New guy's rating unknown!
The Three Pillars: Nearest Neighbors

K-Nearest Neighbor Methods

- Classical KNN: 0.5% improvement
- KNN with learned weights: 4.6% improvement
Map of Movie Neighbors
The Three Pillars

• Matrix factorization
  – Alternating least squares
  – Online/Stochastic gradient descent
    • Dec. 2, 2006, Simon Funk (Brandyn Webb)
    • Enormous impact: anyone could beat Cinematch after a few minutes of training

• Typical improvement: 4%
The Three Pillars

- Restricted Boltzmann machines
  - May 2007, Salakhutdinov and Mnih

- Typical improvement: 5%
Milestones

• Spring 2007: Dinosaur Planet enters “Top 10”
• June 2007: DP graduates from college
Model Ensembling

Recurring theme

• Combining the predictions of multiple models to yield improved performance

• Motivation:
  – Diminishing returns from optimizing a single algorithm
    • Best single model improvement: 8.24% (Aron Miller)
    • The Ensemble’s final improvement: 10.09%
  – Different models capture different aspects of the data
    • Global commonalities of MF vs. Local similarities of KNN
  – Variance reduction from uncorrelated inputs
Model Ensembling

• Stacked linear regression (Wolpert, Breiman)
  – Target = held-out ratings, r
  – Covariates = model predictions, P
  – Tikhonov regularization to reduce overfitting

\[
\min_{\beta} \| \mathbf{r} - \mathbf{P} \beta \|^2 + \lambda \| \beta \|^2
\]
Result: 2.0% improvement over best model
Model Ensembling Variations

• Add user, item, and date features as covariates
  – User rating count
  – Date of rating
  – Average inverse user rating count per movie
• Sparse regression: L1 regularizer or nonnegativity constraints
• Regress on pairwise interactions
  – Greedy selection or bagging with random subsets
• Result: 7.96% improvement over Cinematch
The First Progress Prize

• Sept. 3, 2007
  – Dinosaur Planet takes first place (from reigning champion BellKor)
• One hour later
  – BellKor takes back first place

Recurring theme
The First Progress Prize

One day before the deadline...

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>Best Score</th>
<th>% Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BellKor</td>
<td>0.8728</td>
<td>8.26</td>
</tr>
<tr>
<td>2</td>
<td>Gravity</td>
<td>0.8750</td>
<td>8.03</td>
</tr>
<tr>
<td>3</td>
<td>Dinosaur Planet</td>
<td>0.8753</td>
<td>8.00</td>
</tr>
<tr>
<td>4</td>
<td><a href="mailto:ML@UToronto.A">ML@UToronto.A</a></td>
<td>0.8787</td>
<td>7.64</td>
</tr>
<tr>
<td>5</td>
<td>Arek Paterek</td>
<td>0.8789</td>
<td>7.62</td>
</tr>
<tr>
<td>6</td>
<td>basho</td>
<td>0.8805</td>
<td>7.45</td>
</tr>
<tr>
<td>7</td>
<td>NIPS Reject</td>
<td>0.8808</td>
<td>7.42</td>
</tr>
<tr>
<td>8</td>
<td>Ensemble Experts</td>
<td>0.8841</td>
<td>7.07</td>
</tr>
</tbody>
</table>

Grand Prize - RMSE \(\leq 0.8563\)

Progress Prize 2007 - RMSE: 0.9419
The First Progress Prize

• Sept. 3, 2007
  – Dinosaur Planet takes first place (from reigning champion BellKor)

• One hour later
  – BellKor takes back first place

• Sept. 19, 2007
  – Gravity contacts DP about potential collaboration
  – Gabor Takacs, Istvan Pilaszy, Bottyany Nemeth, Domonkos Tikk

• Oct. 1, 2007
  – When Gravity and Dinosaurs Unite overtake BellKor with 8.38%

• 76 seconds later
  – BellKor ties with 8.38%

• Oct. 2, 2007
  – KorBell wins the first $50,000 progress prize with 8.43% improvement
The Power of Teamwork

• First Progress Prize
  –  joins forces with for 8.38% improvement
  –  BellKor improves to 8.43%, wins $50,000
  –  76 seconds later: BellKor ties with 8.38%

• Second Progress Prize
  –  BigChaos joins BellKor for 9.44% and $50,000

• Grand Prize Team (GPT) founded by +
  –  Anyone could join
  –  The more you improve the GPT score, the bigger
  your share of the $1 million Grand Prize
  –  Many joined and brought new techniques with them
Gaussian Missing Data Model (Roberts)

- Assume each vector of user ratings drawn from common multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$
  - Incomplete vector of observed ratings drawn from marginal distribution
- Choose $(\mu, \Sigma)$ to maximize likelihood
  - Expectation-Maximization or gradient ascent
- Predict missing ratings as conditional expectation given observed ratings
- Result: 6.38% improvement
Feature-Weighted Linear Stacking (Sill, Takacs, Mackey, Lin)

• An adaptive approach to stacked linear regression
• Allow model ensembling weights to depend linearly on known features of the user, movie, and date
  – Did the user rate more than 3 movies on this date?
  – Log number of times the movie has been rated
  – Log number of distinct dates on which a user has rated
  – Log of average correlation between movies rated by user and movie to be predicted
• Result: 8.82% $\Rightarrow$ 9.46% improvement for GPT
Remaining teams had 30 days to respond
Over the next 24 days, **30 individuals** from **11 countries** combined forces to challenge BPC
The Last Call

• June 26, 2009
  – Top 3 teams (BellKor, BigChaos, and Pragmatic Theory) combine to pass the Grand Prize threshold
  – Initiates 30 day last call period for $1 million grand prize
• June 30, 2009
  – GPT begins deeper collaboration
  – Message board to share ideas, server to share code and predictions
• July 5, 2009
  – Vandelay Industries! contacts GPT about potential collaboration
• July 7, 2009
  – Opera Solutions joins Vandelay Industries!
• July 20, 2009
  – The Ensemble is born
The Ensemble

- **Grand Prize Team**
  - **Gravity**
    - Gabor Takacs, Istvan Pilaszy, Bottyan Nemeth, Domonkos Tikk
  - **Dinosaur Planet**
    - David Lin, Lester Mackey, David Weiss
  - Joe Sill
  - Ces Bertino
  - Dan Nabutovsky
  - William Roberts
  - Wojtek Kulik
  - Willem Mestrom
  - David Purdy

- **Vandelay Industries!**
  - Greg McAlpin
  - Bill Bame
  - Bo Yang
  - Chris Hefele
  - Jeff Howbert
  - Xiang Liang
  - Larry Ya Luo
  - Aron Miller
  - Steve Pagliarulo
  - Opera Solutions
    - Bruce Deng, Peng Zhou, Priyanka Rastog, Arvind Gangadha, Jacob Spoelstra
  - Craig Carmichael
  - Mike Linacre
  - Edward de Grijs
  - Clive Gifford
  - Feeds2
    - Nicholas Ampazis, George Tsagas

Learn more at http://the-ensemble.com/
The Road to the Grand Prize

• Next to Last Day
  – The Ensemble submits
  – 10.09% improvement on Quiz Set

• Final Day, 6:18pm
  – BellKor’s Pragmatic Chaos responds
  – 10.09% improvement on Quiz Set

• Final Day, 6:38pm
  – The Ensemble makes its final submission
  – 10.10% improvement on Quiz Set

• Final Day, 6:42pm: Contest closes
The Other Road to the Grand Prize

• Next to Last Day
  – The Ensemble submits
  – 10.05% improvement on Test Set

• Final Day, 6:18pm
  – BellKor’s Pragmatic Chaos responds
  – 10.06% improvement on Test Set

• Final Day, 6:38pm
  – The Ensemble makes its final submission
  – 10.06% improvement on Test Set

• Tie breaker: Time of submission
The End

And then there were two...

Teams shown from first appearance in top 20.

Dinosaur Planet

Gravity

WDPU

Newman!

Opera Solutions

Ces

David Purdy
Craig Carmichael
Edward de Grijs
xiangliang
Mike Linacre
Aron Miller
Steve Pagliarulo (ESP)

Dace

Chef Ele

OfADifferentKind

Vandelay

Industries!

clueless

William J.J. Roberts
photek
Green Circle
Expensive Lunch
blednotik

Grand Prize Team

The Ensemble

BellKor

Big Chaos

BellKor in BigChaos

Pragmatic Theory

1st Progress Prize | BellKor Paper Published

2nd Progress Prize | BellKor Paper Published

2007 8.43 2008 9.44 2009 10.05

Copyright AT&T