Project Report

Recasting Movies Using IMDB Data

Gaurav Mishra , Piyush Khemka , Pulkit Sharma

110828660 , 110828688 , 110900867

{gamishra,pkhemka,pusharma}@cs.stonybrook.edu

1 Problem Statement

Given an input movie, can we suggest a suitable replacement for the lead actor male and lead actor female, if the movie were to be remade in the year specified by the user.

Example questions answered by the model:

1. Given the movie - 'Inception', who would be the best actor to play the lead role, if the movie was to be made in 1950s?
2. Given the movie - 'Titanic', who would be the best actor to play Kate Winslet's role, if the movie was to be made in 2016?
3. Who would play Arnold Schwarzenegger's role in Terminator if the movie was to be made in 1940s or if it was made today?

2 Approach

We extracted relevant data from IMDB pertaining to lead actors male & lead actors female and represented each of them as a feature vector. We did this for all actors and represented them on a multi-dimensional cartesian space. We then computed the distance between each of them and sorted them based on various distance. Actors who were nearer to each other were considered more similar. To recommend a replacement for a given query in the form of (Movie, Release Year, Remake year), we applied age fitness & genre fitness on the list of similar actors & yielded only those actors who passed both the fitness tests.

3 Baseline Model

For the initial baseline model, we had extracted data from IMDB and represented each actor as a feature vector of Age, Height, Average Rating of their movies on IMDB and Average number of votes for each movie.

We had tested it on a few examples and found some problems in our baseline such as - unknown actors being suggested as replacements for famous A list stars, couple of non-actors making the list (due to error in data cleaning). The feedback on our progress

report was to collect a more interesting feature set for actors & test them on a large scale dataset.

For our advanced model, we have incorporated all the feedbacks and generated a more comprehensive feature vector set for each actor & tested them robustly for all remakes found in the IMDB dataset.

4 Advanced Model

4.1 Final Feature Set

1. Age - Age of the actor as of today. (2016 - Birth year)
2. Height - Height of the actor as listed on IMDB and all converted to cms scale.
3. Ratings - Average rating of their movies on IMDB.
4. Votes - Average number of votes received on their movies on IMDB. This feature also acts as a proxy for popularity. Popular actors (or movies) receive more views & hence more votes.
5. Genres - Each of the 34 distinct genres listed out by IMDB were considered as a separate feature. This field contains a numeric value which represents the total sum of all movies done by an actor in that particular genre. So, a value of 5 in Drama implies that the actor in question has acted in 5 drama movies in his/her career span.
6. Salary - Average of all the salaries as listed on IMDB. The salaries were adjusted for inflation with the aid of CPI index before averaging.
7. Budget - Average budget of all the movies acted in by the actor. The budget of the movie was also adjusted for inflation with the aid of CPI index.
8. Movie Gross - Average gross of all the movies of an actor. The gross was also adjusted for inflation with CPI index.
9. Awards (Academy awards & Golden Globes) - We assigned points to actors on the basis of number of academy awards and Golden Globes won & award nominations.The formula is listed as follows:
   100 * (number of Academy Awards won) + 75 * (number of Golden Globes won) + 30 * (number of Academy Awards nominations) + 20 * (number of Golden Globes nominations)
   The weights of 100,75,30 & 20 were determined heuristically.

4.2 Normalizing & Cleaning data

1. Salary, Movie Gross & Budget data was found to be very sparse since IMDB collects this data for only top A-list stars. The median values for these fields were found to be 0 with a very high standard deviation. Therefore, we decided not to fill the 0 values with the mean or median and kept them as 0 so as to introduce a large distance between unknown actors and big movie stars.

2. All the features considered were normalized by taking their z-score in order to bring them down to the same scale.

We considered many alternate sources of data but decided against using them. We had scraped Facebook likes & count of Twitter followers but found that most old actors did not have this data or had very skewed data as compared to newer actors. We also considered using Rotten Tomatoes metacritic rating but did not use it because ratings from IMDB was already accounting for the quality of the movie and using additional such data would have skewed our feature set heavily into that direction.

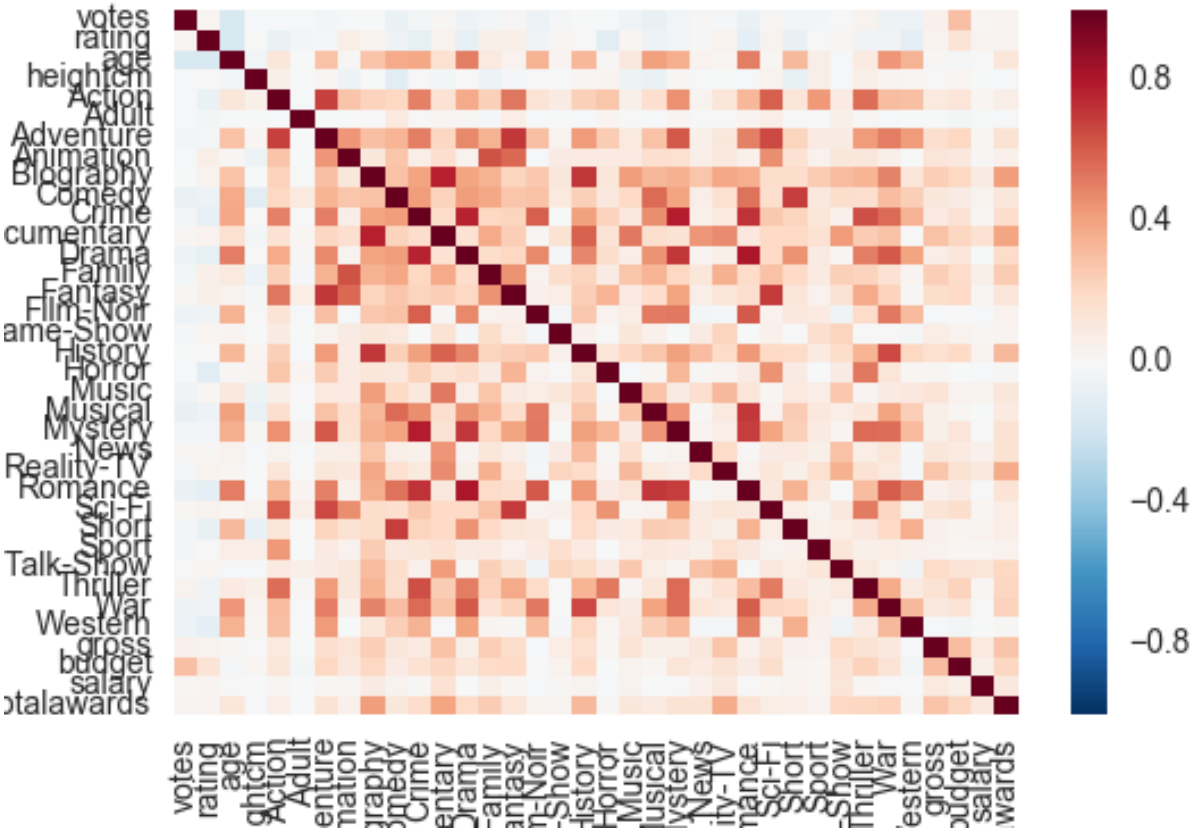

Figure 1: Correlation for male actor

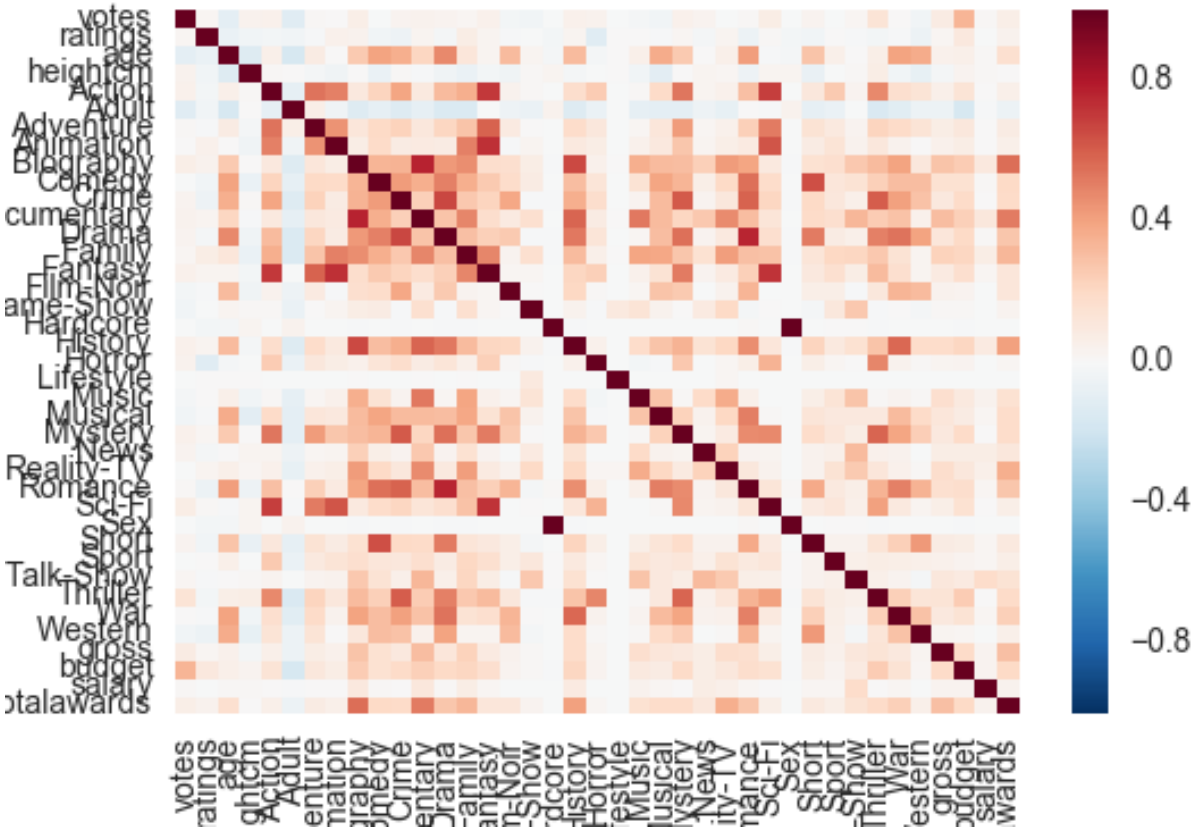

Figure 2: Correlation for female actor

4.3 Methodology

After constructing the feature set for each actor, we applied a distance function on it to find the distance between every pair of actor and sorted them to find actors who had the least distance between them. We used Euclidean, Manhattan and Cosine distance function and compared the results obtained from each of them. They have been outlined in more details in the Results section.

4.3.1 Age Fit

After finding the distance between actors, we applied age fitness on it to remove any actor who may not be a suitable for replacement due to age.

Our model takes into account the age of the actor when they played the role originally. Then it adjusts the age of all other actors according to the given input year i.e. the model will evaluate their age for the year in which the movie is being remade and eliminates all those actors who aren't in the same age bucket as the original actor when they starred in the original movie. This age fitness ensures that child actors are only replaced with other child actors & not adults and similarly roles which require older actors, don't end up suggesting young teenage actors as replacements.

For Example- If we remake the movie 'Inception' in 1950s, then the replacement lead

actor suggested by our model will be in a similar age bracket in 1950s as Di Caprio was when he played the lead role in Inception in 2012.

4.3.2 Genre Fit

After eliminating actors on the basis of age, we also eliminate actors who don't act in the same kind of movies as the original actors. Since our problem requires us to find suitable replacements in movies, it is imperative to suggest actors who seem fit for the type of movie being remade and not just replace them with actors who were found to be most similar to each other in terms of popularity, movie rating etc. For instance, replacing Christian Bale with Adam Sandler as Batman would be a bad result since Christian Bale acts in serious drama movies & Adam Sandler acts in Comedy movies primarily.

To account for genre fitness we are taking the cosine similarity of genres between pair of actors and suggesting only those actors as replacements whose cosine similarity is greater than 0.6.

5 Evaluation Of Model

5.1 Testing Dataset

IMDB tags connections between movies with keywords like - alternate language version of, edited from, edited into, featured in, features, followed by, follows, referenced in, references, remade as, remake of, similar to, spin off, spin off from etc.

To test our model we considered movies which have a **'remade as'** link between them. An interesting finding was that certain movies which we intuitively believe are remakes were connected in by other kind of link.

For example Mad Max: Fury Road (2014) is thought to be a remake of Mad Max (1979) but IMDB tags them as -
Mad Max (1979) was **'followed by'** Mad Max: Fury Road (2014). This is probably because lots of features of movie like nature of characters , the relative order of importance of characters , etc have changed in the 2014 version of movie.

Another interesting observation was that there were more remakes in the pre 1990 era while the seemingly and popular remakes after 1990 are actually **'a version of'**, (case with a lot superhero movies made again and again), or **'followed by'**(story continues from the previous version) .

We collected all the pairs of movies which were connected by **'remade as'** relationship and filtered this list based on following criteria -

- Both the movies in the pair should be released in USA.
- Rating information should exists for both movies in any pair.
- Runtime of both movies should be greater than 60 minutes.
- Movie must have had a theatrical release i.e. (movies which went straight to DVD or TV movies weren't considered).

5.2 Testing Methodology

Quantative: For any given remake we ran our model for the actor who appeared in the original movie and let our model predict possible replacements for him. We let our model come up with multiple (5% of the dataset size) replacements for the original actor. If the set of replacement actors contained the actor who actually appeared in the remake we call that a 'hit' for our model. We counted for what percentage of remakes our model was 'hit'. We did this test for both actors as well as actresses.

Qualitative: Intuitively how good the top 5 replacements for any actor are for any given movie.

6 Results

6.1 Quantative Results

Following are the results from the 3 models which we tested -

Distance Function	Male Dataset Accuracy	Female Dataset Accuracy
`Cosine`	44.62%	61.97%
`Manhattan`	41, 39%	49.47%
`Euclidean`	43.54%	51.04%

Table 1: Accuracy of the model from difference distance functions used

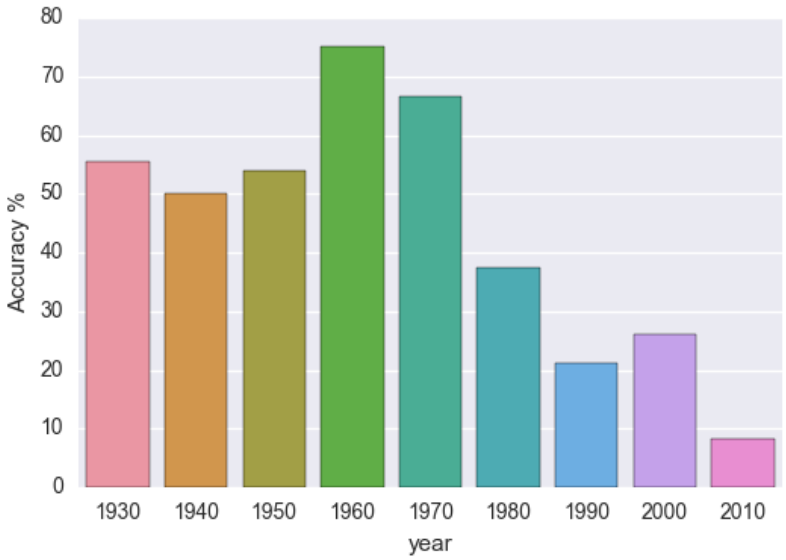

Figure 3: Accuracy for male actors for various year buckets

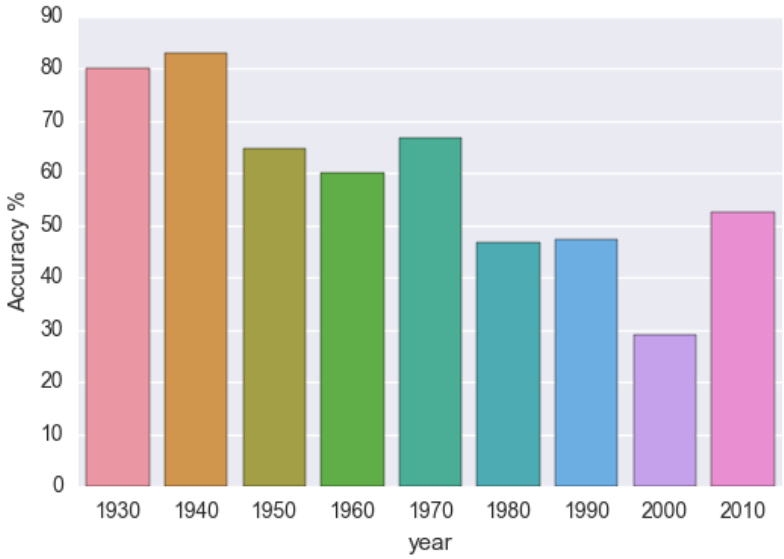

Figure 4: Accuracy for female actors for various year buckets

6.2 Qualitative results:

6.2.1 Comparison with Baseline Model

Remake Inception(2010) in 1960

Advanced Model	Baseline Model
Marlon Brando	Ian McKellen
Paul Newman	Joe Spinell
Robert Duvall	John Williams
James Garner	Newell Alexander
Richard Burton	Michael Caine

Table 2: Top 5 replacements for Leonardo DiCaprio

Analysis Of Inception

As compared to our baseline model the results have improved. No unknown actors are observed also because of a better genre fit results like Marlon Brando are a better fit for DiCaprio than the ones observed in the baseline model.

Analysis of Its a Wonderful life

The quality of results have improved drastically when recasting for older movies. All the replacements in our advanced model are well known actors unlike the baseline model. Additionally, a stronger age fit ensures that the actors are of suitable age bracket and can be good potential replacements.

Its a Wonderful life(1946) to be made in 2016

Advanced Model	Baseline Model
Matt Damon	Johnnie Brannon
Adrien Brody	Chris Muto
Matthew McConaughey	Jeff Pope
James Franco	Edward Norton
Will Smith	Wes Bentley

Table 3: Top 5 Replacements of James Stewart

Remake Terminator (1984) in 2010

Advanced Model	Baseline Model
Keanu Reeves	Brendan Frasier
Matt Damon	John Cusack
Ewan McGregor	Jim Carrey
Will Smith	Kevin Hefferman
James Franco	Will Smith

Table 4: Top 5 replacements for Arnold Schwarzenegger

Analysis Of Terminator

Genre fit ensures that Actors are replaced according to the genre of movies they have worked in. Also, a more robust feature vector ensures that replacement actors are of same repute which wasn't the case in our baseline model.

6.2.2 Some more results:

Testing replacements for an older male actor
Remake Lord of the Rings (2001) in 2011
Top 5 replacements for Ian Mckellen

1. Bill Nighy
2. Liam Neeson
3. John Goodman
4. Bob Hoskins
5. Jeff Daniels

Testing for lead male actor - James Bond
Remake Die Another Day(2000) in 1980
Top 5 replacements for Pierce Brosnan -

1. Roger Moore
2. Sean Connery

3. Harrison Ford
4. Robert Redford
5. Clint Eastwood

Testing for lead female actor in 90s
Remake Mr. & Mrs Smith (2005) in 1990
Top 5 replacements for Angelina Jolie.

1. Nicole Kidman
2. Julia Roberts
3. Cate Blanchett
4. Jodie Foster
5. Sharon Stone

Testing for lead female actor today
Remake Mr. & Mrs Smith (2005) in 2016
Top 5 replacements for Angelina Jolie.

1. Keira Knightley
2. Natalie Portman
3. Jessica Chastan
4. Anne Hathaway
5. Kirsten Dunst

Testing lead female actor in 60s
Recast Kate Winslet in Titanic from 1998 to 1960

1. Maggie Smith
2. Glenda Jackson
3. Joanne Woodward
4. Grace Kelly
5. Judi Dench

Recast Older Actress
Replacing Meryl Streep in The Devil wears Prada (2006) to 2016

1. Jessica Lange
2. Sissy Spacek
3. Holly Hunter
4. Jodie Foster
5. Emma Thompson

Child Actor test
Remake Home Alone 1990 in 2001
Replace Macaulay Culkin

1. Taylor Lautner
2. Daniel Radcliffe
3. Dylan O Brien
4. Dev Patel
5. Logan Lerman

7 Conclusion

We built a comprehensive model to recommend suitable replacements for actors with other similar actors and tested it both quantitatively (on remakes dataset) and qualitatively (based on intuition). During the course of the project we found that a significant amount of our time and effort was spent in gathering and cleaning data. To gather data we had to write complex SQL queries to extract relevant data and apply data science techniques learnt in order to bring them to the same scale, adjust for inflation etc.

We also discovered that while features like rating, votes, salary, height played a huge role in finding similar actors, it was actually the age and genre fitness test which ensured good results in suggesting replacements. Without the two fitness tests, our model wasn?t yielding such good results.

We also found that our results were much better for female actors as compared to male actors. After a thorough investigation, we can conclude that it is due to the reason that heroines typically have a shorter career span as compared to their male counterparts and are more likely to be typecast in a particular genre of movie or role. Therefore, their feature sets aren't as varied as their male counterparts and thus the task of finding similar actresses becomes easier than finding similar actors.