CSE 519 PROJECT FINAL REPORT

Which Movies Endure and Why?

OBJECTIVE:

Our task here is to develop a model that can predict how a movie endures over time based on the information present at the time of its release and also current information available about the movie.

INTRODUCTION

The Internet Movie Database (IMDB) is an online database which stores information related to films, television programs, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, fan and critical reviews, and ratings. Apart from this we have also used data from The Movie Database (TMDB). The TMDB API is a resource for developers to integrate movie, TV show and cast data along with posters or movie fan art. It is a free and community edited database.

BACKGROUND RESEARCH

The American Movie Industry, popularly known as 'Hollywood' is considered as one of the oldest film industries in the world. It produces the largest number of films of any single-language national cinema, with more than 700 English-language films released on average every year. It is also known as the birthplace of various genres of cinema—among them comedy, drama, action, musical, romance, horror, science fiction, and the war epic—having set an example for other national film industries.

Hollywood over the years:

1900 - 1950s: There was a great advancement for film and motion picture technology. Exploration into editing, backdrops, and visual flow motivated aspiring filmmakers to push into new creative territory. With hundreds of movies releasing every year, Hollywood was the rise of an American force. The use of audio tracks in motion pictures created a new viewer dynamic and also initiated Hollywood's leverage in the upcoming World War II in the 1930s. Production saw a rebound due to advances in technology such as special effects, better sound recording quality, and the beginning of color film use in the 1940s.

1960- 2000s: The 1960's saw a great push for social change. In the 1970's, there was a rush of creativity due to changes in restrictions on language, sex, violence, and other strong thematic content. Due to the use of special effects in the 1980s, the budget of film production increased and consequently launched the names of many actors into overblown stardom. The creation of CD-ROM's paved the way for movies on DVD. DVD's featured a much better image quality as well as the capacity for interactive content, and videotapes became obsolete a few years later and this led to a huge spike of movies being watched.

DATA COLLECTION

In order to get our final dataset that we worked on, we combined datasets from the following sources:

- IMDB - https://datasets.imdbws.com/
  All information regarding movies can be downloaded from above IMDB link. The data is refreshed daily.
- TMDB-https://www.themoviedb.org
  The Movie Database (TMDB) provides an API which can be used to pull movie data.
- Numbers.com-https://www.the-numbers.com/movies/production-companies/#production_companies_overview=od3
  The production company's popularity was collected from this link.
- https://www.oscars.org/oscars/awards-databases-0
  https://datahub.io/rufuspollock/oscars-nominees-and-winners#data
  https://www.kaggle.com/theacademy/academy-awards#database.csv
  Information about various nominations and awards that movies and actors got in various categories in a particular year (E.g.: Oscar Awards and Academy Awards)
- IMDBPro- https://www.imdb.com/list/ls009809456/
  The director's popularity was taken from this link
- IMDBPro-https://www.imdb.com/list/ls022928819/
  https://www.imdb.com/list/ls023445281/
  Combining these two sources gives us the cast's popularity.
  IMDBPro uses proprietary algorithms that take into account several measures of popularity for people, titles and companies. The primary measure is who and what people are looking at on IMDb. The rankings are updated on a weekly basis, typically by the end of Monday.
- https://towardsdatascience.com/the-what-and-why-of-inflation-adjustment-5eedb496e080
  Inflation information that we used for adjusting the budget and revenue data.

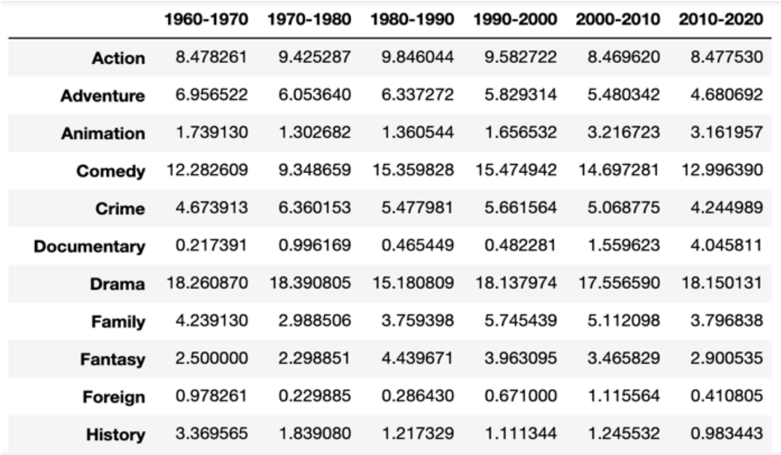
FEATURES

Since our analysis is based on the comparison between the current popularity of a film and the popularity at the time of release, we divided our entire set of features into two parts i.e. features providing information about current popularity and features giving information about popularity at the time of release of the movie.

- **Features measuring popularity at the time of release:**

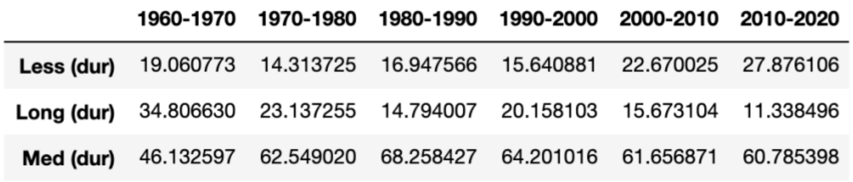
  During our analysis phase in the previous progress report, we found out that these features impact the time of release popularity of a movie the most:

1. **Genre Popularity**: In order to calculate the genre popularity, we divided the time periods into a duration of 10 years and calculated the popularity of a particular genre in that decade. We calculated the percentage of the number of movies of a particular genre released in that particular decade and for a movie having multiple genres we combined these percentages to get a final score.



	1960-1970	1970-1980	1980-1990	1990-2000	2000-2010	2010-2020
Action	8.478261	9.425287	9.846044	9.582722	8.469620	8.477530
Adventure	6.956522	6.053640	6.337272	5.829314	5.480342	4.680692
Animation	1.739130	1.302682	1.360544	1.656532	3.216723	3.161957
Comedy	12.282609	9.348659	15.359828	15.474942	14.697281	12.996390
Crime	4.673913	6.360153	5.477981	5.661564	5.068775	4.244989
Documentary	0.217391	0.996169	0.465449	0.482281	1.559623	4.045811
Drama	18.260870	18.390805	15.180809	18.137974	17.556590	18.150131
Family	4.239130	2.988506	3.759398	5.745439	5.112098	3.796838
Fantasy	2.500000	2.298851	4.439671	3.963095	3.465829	2.900535
Foreign	0.978261	0.229885	0.286430	0.671000	1.115564	0.410805
History	3.369565	1.839080	1.217329	1.111344	1.245532	0.983443

2. **Budget of the movie**: We already have this information with us from the IMDB and TMDB data sources. Initially we thought about using the gross income of a movie at the time of its release or the gross income during the first week of its release, but since we could not find this information, we decided to proceed with budget of a movie. We also considered inflation as a factor and we adjusted the budget according to the Consumer Price Index (CPI) for that year.
3. **Runtime of the movie**: In order to calculate the score of runtime of a movie we employed a similar approach as while calculating the genre score. So, we divided the time period into decades and divided the runtime into 3 categories into Less, Medium and Long. And then we calculated the popularity score in the same way that we calculated the genre popularity.

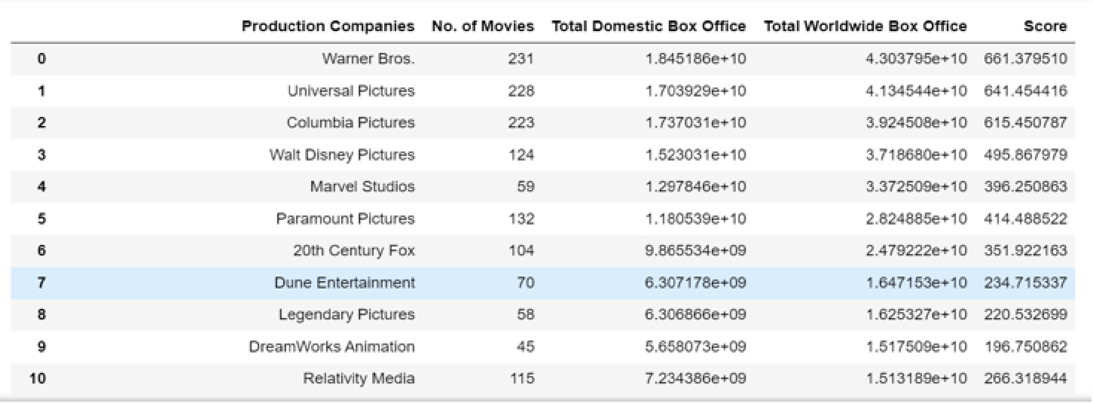


	1960-1970	1970-1980	1980-1990	1990-2000	2000-2010	2010-2020
Less (dur)	19.060773	14.313725	16.947566	15.640881	22.670025	27.876106
Long (dur)	34.806630	23.137255	14.794007	20.158103	15.673104	11.338496
Med (dur)	46.132597	62.549020	68.258427	64.201016	61.656871	60.785398

- **Features measuring current popularity:**

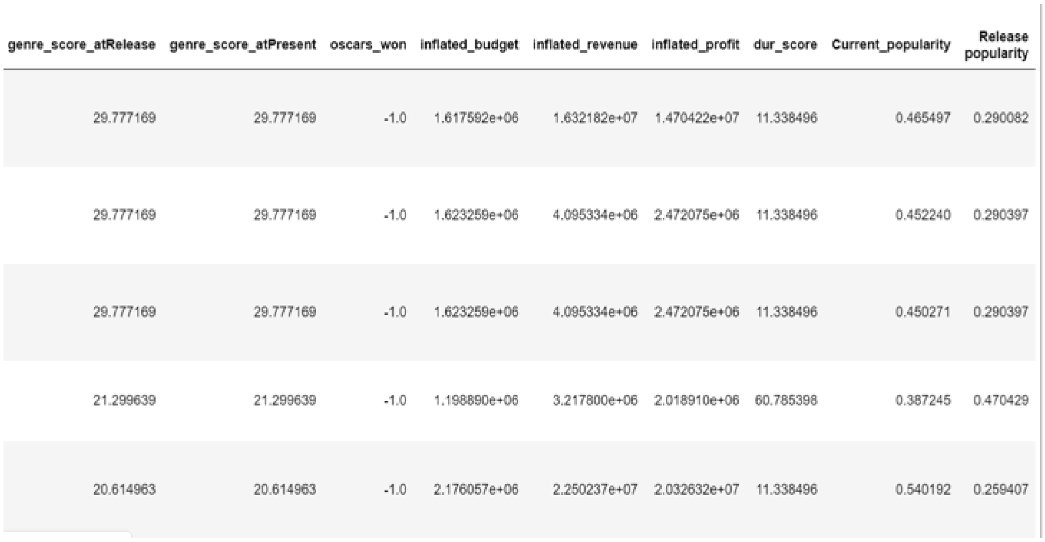
  During our analysis phase in the previous progress report, we found out that these features impact the current popularity of a movie the most:

1. **Director’s popularity:** We got the popularity score for each director from the IMDBPro link and we cleaned the data and converted the string into float value and merged it into our main Dataframe. In order to deal with NaN values after combining, we replaced these values with the 25th quartile.
2. **Cast’s popularity:** We combined the actor’s and actresses’ popularity scores into one Dataframe and performed data cleaning on this and merged it into our main Dataframe.
3. **Production Company’s popularity:** We got the production company information from the IMDBPro link and then we performed data cleaning on it. We had three important columns here- Number of movies, Total Domestic Gross Income and Total Worldwide Gross Income. We came up with a scoring function to combine these features into a single score. We normalized the values of the gross income as the number of movies was in the order of 10^2 and the gross income was in the order of billions. So, we divided it by 10^8 to have all the columns in the same range.



	Production Companies	No. of Movies	Total Domestic Box Office	Total Worldwide Box Office	Score
0	Warner Bros.	231	1.845186e+10	4.303795e+10	661.379510
1	Universal Pictures	228	1.703929e+10	4.134544e+10	641.454416
2	Columbia Pictures	223	1.737031e+10	3.924508e+10	615.450787
3	Walt Disney Pictures	124	1.523031e+10	3.718680e+10	495.867979
4	Marvel Studios	59	1.297846e+10	3.372509e+10	396.250863
5	Paramount Pictures	132	1.180539e+10	2.824885e+10	414.488522
6	20th Century Fox	104	9.865534e+09	2.479222e+10	351.922163
7	Dune Entertainment	70	6.307178e+09	1.647153e+10	234.715337
8	Legendary Pictures	58	6.306866e+09	1.625327e+10	220.532699
9	DreamWorks Animation	45	5.658073e+09	1.517509e+10	196.750862
10	Relativity Media	115	7.234386e+09	1.513189e+10	266.318944

4. **IMDB rating of a movie:** We already had this information from the IMDB and TMDB data source.
5. **Genre popularity (current):** We calculated the current genre popularity by calculating the percentage of movies of a particular genre released in the current decade i.e. 2010-present year. The computation logic is same as before.
6. **Awards won by the movie:** Based on the source that we found we assigned a value of 1 to those movies who won an award and a 0 to those movies who just got nominated but didn’t win. After combining this Dataframe with our main Dataframe we got lots of NaN values for movies that didn’t win and didn’t get nominated and assigned a value of -1 to such movies.
7. **Profit of a movie:** Based on the budget and revenue values present in the source from IMDB and TMDB we first adjusted these values according to inflation and then calculated the inflated profit by subtracting the inflated budget from the inflated revenue.
8. **Vote average:** The average voting score that a movie got. We got this from the IMDB and TMDB data source.



genre_score_atRelease	genre_score_atPresent	oscars_won	inflated_budget	inflated_revenue	inflated_profit	dur_score	Current_popularity	Release popularity
29.777169	29.777169	-1.0	1.617592e+06	1.632182e+07	1.470422e+07	11.338496	0.465497	0.290082
29.777169	29.777169	-1.0	1.623259e+06	4.095334e+06	2.472075e+06	11.338496	0.452240	0.290397
29.777169	29.777169	-1.0	1.623259e+06	4.095334e+06	2.472075e+06	11.338496	0.450271	0.290397
21.299639	21.299639	-1.0	1.198890e+06	3.217800e+06	2.018910e+06	60.785398	0.387245	0.470429
20.614963	20.614963	-1.0	2.176057e+06	2.250237e+07	2.032632e+07	11.338496	0.540192	0.259407

METHODS

- **Calculating Endurance of a film:**

  From all the above scores, we came up with a formula to combine all these scores into a single popularity score.

*Current popularity = (0.5) * vote_average + (0.1) * cast_points + (0.1) * production_points + (0.1) * Director_points + (0.1) * inflated_profit +(0.1) * genre_score_atPresent*

*Release popularity = (0.4) * genre_score_atRelease + (0.3) * dur_score + (0.3) * inflated_budget*

After this we calculated the endurance using the formula below:

*Endurance = ((0.6) * (current popularity- release popularity) + (0.4) * (current popularity + release popularity))/2*

- **Modelling:**

  We considered ‘Endurance’ to be our target variable and our aim was to predict it and compare it with the endurance values we calculated above. We tried out the following regression models on the selected features:

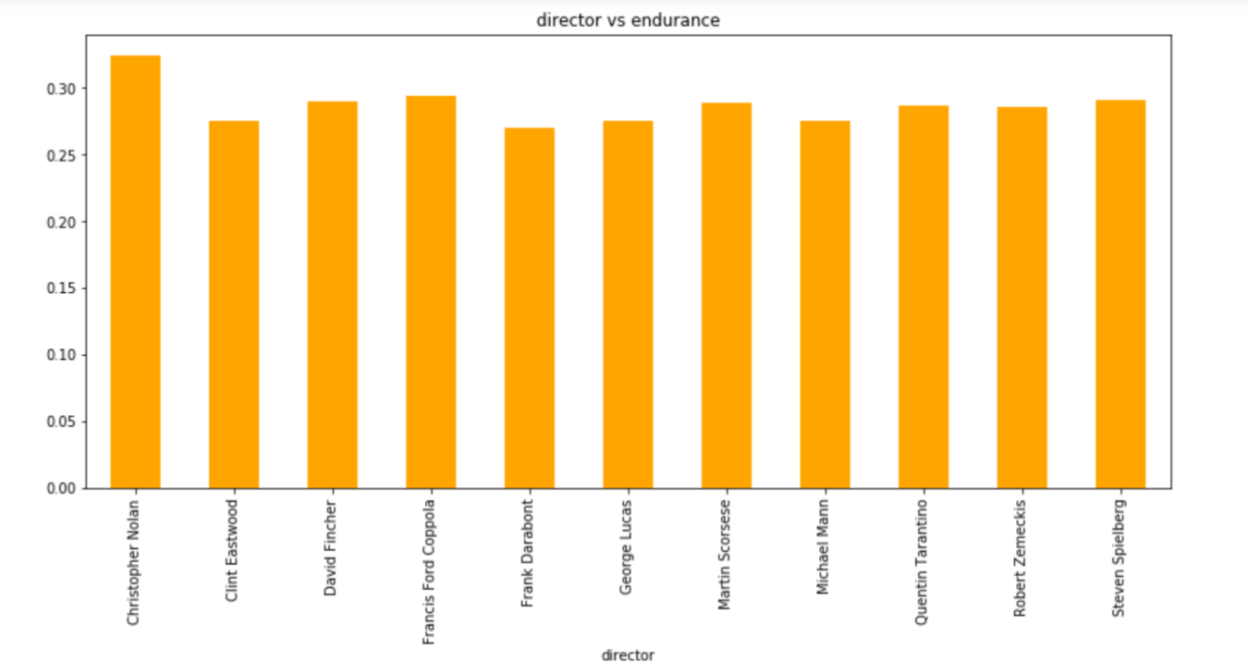
1. Linear Regression: This was our baseline model and the Root Mean Squared Error (RMSE) that we got was 0.00929.
2. Ridge Regression: In order to get better results, we tried out the ridge regression model with alpha values as 0.1, 10 and 100.
3. Lasso Regression: We also tried out Lasso Regression with regularization term values as 0.1, 10 and 100.

Models	Alpha = 0.1	Alpha = 10	Alpha = 100
Ridge	0.01121	0.01073	0.01021
Lasso	0.01112	0.01261	0.03778

RMSE values for Ridge and Lasso Regression

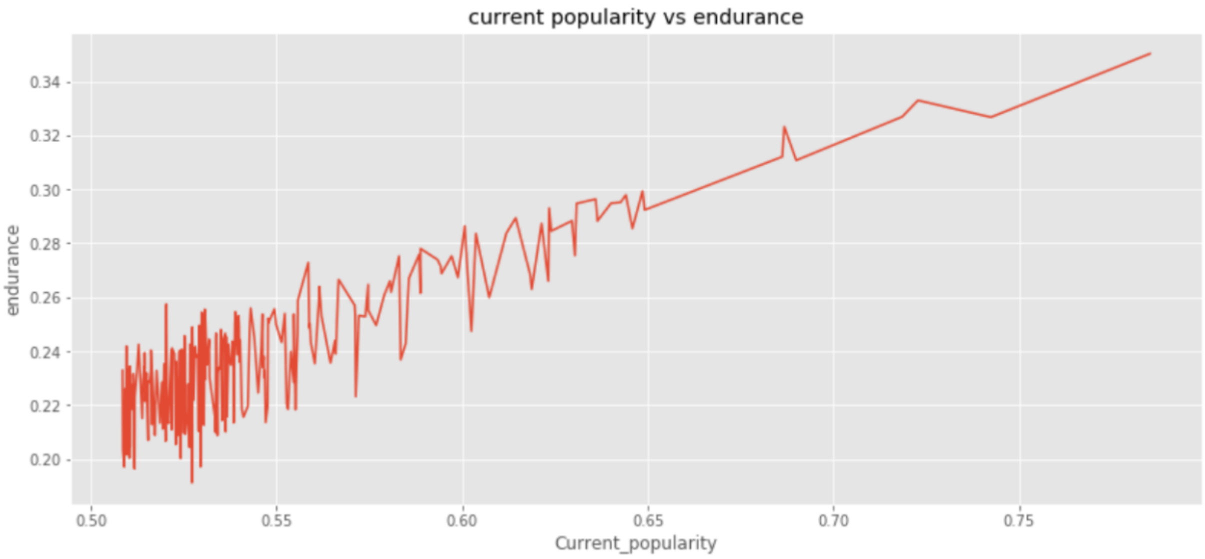
ANALYSIS & RESULTS

Analysis:

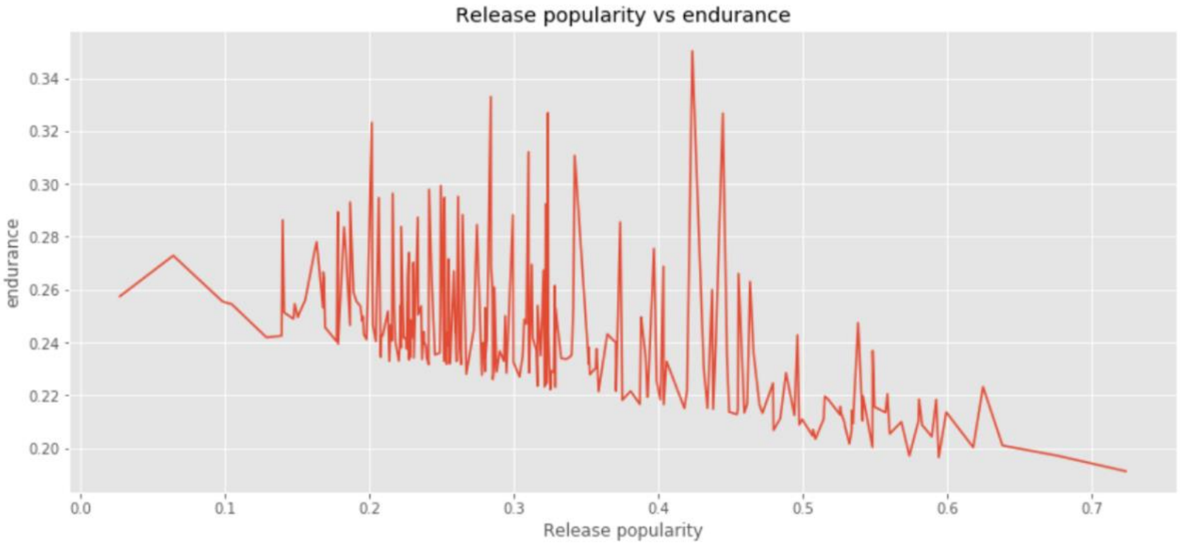


We have drawn a plot between the director’s names and the endurance of the movie directed by them. We compared above names with the most popular directors from the IMDBPro list and 8 out of 10 were matched.

This indicates that the movies directed by the most popular directors are likely to endure over time.



In case of current popularity vs endurance, we can see that an increase in current popularity leads to an increase in endurance which seems logical.



On the other hand, in case of time of release popularity vs endurance, we can see that an increase in time of release popularity leads to an decrease in endurance.

Result: The model that performed the best in our case was Linear Regression with an RMSE of 0.00929.

We tried to answer some questions asked when the topic was introduced.

- Consider the effects of genre: how have different genres performed? Can you demonstrate that war movies and westerns have gone out of style?

To answer this question, we took a mean of the current popularity score of war and western genre movies released in 1960-1970 and in 2010-present. We observed that the latter is lesser than the former which means that war and western movies did not endure well over time.

- How much do Academy Awards (Oscars) make?

  To answer this question, we calculated the endurance of the movies which won Oscar award and the endurance of movies that got nominated but did not win. We found that the endurance of movies that won an award was higher than the endurance of movies that just got nominated but not by a significant amount. Our hypothesis is the movies that won should definitely get a higher endurance value but the movies that just got nominated will also have endured well because they are also good movies.

CONCLUSION

In this report we have successfully developed a metric to determine which movies endure and why. We have also taken into account if a director's popularity or that of an actor/actresses' or that of a production company's affects the popularity of a movie in any way. In addition, we examined the effect of other features such as duration, budget, revenue of a movie on the endurance of movies. Whether a certain genre of movie was likely to endure over a period of time was also taken into consideration.

REFERENCES

[1]https://en.wikipedia.org/wiki/Cinema_of_the_United_States
[2] https://historycooperative.org/the-history-of-the-hollywood-movie-industry/