ML 512: Course Project (Choice 1: explore methods)

Venkata Subba Narasa Bharath, Meadam
venkatasubban.meadam@stonybrook.edu
SBU ID: 112672986

I. Abstract/Task

Any given old movie will have had a certain popularity at its time of release, perhaps measured by release specific attributes like budget, box office gross, video-tape, sales etc.. The current popularity of the film might be in-line or in contradiction to its popularity at its time of release. In this work, we develop a model to predict the current popularity of a film to identify whether a movie is still relevant (popular) or lost its shine over time. This works presents a comprehensive study of the endurance of a film's popularity over time.

II. Dataset

For this work, we have used the TMDB 5000 Movies dataset available on Kaggle[4] .

Dataset Info:

a. Total features: 21
b. Samples: 10866

how are these
features encoded?
concepts are a bit vague

III. Introduction

Major factors influencing a movies popularity

Perhaps the most intriguing part of this project is figuring out the factors that influence a movie popularity the most. The main contributing factors for a movie's popularity are (1) development, (2) pre-production, (3) production, (4) post-production, (5) marketing and distribution, and (6) exhibition.

Furthermore, after a careful study of these points, we showed that all the factors are related to the performance of the production house. And, in order to judge how popular a movie could be based on the feature during its release, we gauged the performance of the production house for that movie. We argued that one obvious metric to measure this numerically, was net profit; and to measure it we subtracted the movie budget from the world-wide business gross of the movie.

This time, we wanted to figure out more factors that are subtle in nature, but influence the popularity of a feature film greatly. For instance, there is a term called Heirloom Movies. Keith Simanton, senior film editor for the movie website IMDb.com, says that most of the people whose hearts were touched at a young age are now passing the movie love onto their children — or in the case of filmmakers, onto a whole new audience[2].

Movies like Stand By Me or Ferris Bueller's Day Off is still popular after 30 years after their original release date, and some of them are getting re-released in major theatres across United States. "I refer to these movies as heirloom movies, movies people saw as teenagers then and now want to show their kids," says Simanton. "Films like Children of A Lesser God and Out of Africa, darn good movies which won Oscars, were not seen by a certain impressionable age group. So 1 they are not along on this emotional wagon train." [2] Also, great dialogue and good songs aid to a movie popularity. These factors do not directly affect the box-office gross, and are technically not release-specific-attributes; but nonetheless, they remain as major impacting factors on the endurance of a movie's popularity.

IV. Data cleaning and pre-processing

Our data cleaning and pre-processing effort largely had following steps, as described below.

Data cleaning steps:

1) Removal of short duration movies with no budget, revenue or award or nomination details
2) Removal of entries with no duration
3) Removal of outliers and NaN values

Data Pre-processing steps:

1) **Normalize popularity in a scale of (1-10):** From the data we extracted, we found out that the TMDB popularity (a numeric value with two digits after the decimal point) ranges from mid-40s to below 1. We decided to normalize this down to a scale of 1-10, to compare with our endurance metric.
2) **Inflation adjustment**
Inflation adjustment or deflation is the process of removing the effect of price inflation from data. We have adjusted inflation with the following formula:

$$\boxed{Adjusted\ Value = \frac{Actual\ Value}{Index\ Value} * 100}$$

V. Feature Engineering

From the data we have already obtained there are few parameters that need attention or can be modified to be of best use for our model. We have discussed them below:

Production company score:

Production company can be one big factor which can influence the popularity of a movie. From THE NUMBERS[9] website, we have obtained production company domestic gross, worldwide gross and the number of movies made by each production company over these years. In order to provide a score for the production company, we have calculated production points for each production house using the formula

$$\boxed{\begin{aligned} Production\ Company\ Score = \\ \alpha{*}worldwide\ gross \\ +\beta{*}domestic\ gross \\ +\gamma{*}number\ of\ movies\ made \end{aligned}}$$

where α,β γ are co-coefficients of worldwide gross ,domestic-gross and number of movies made respectively.

Director and Cast rank:

Director and cast plays an important role in bringing the audience to the movie and the movie might stay with the audience due to this factor. So, we have come up with scores for the director and the top three cast of the movie.

From the IMDB website, we have the list of directors and list of cast along with their points correspondingly. We have mapped this data with our original dataset using the title of the film and the year of release.

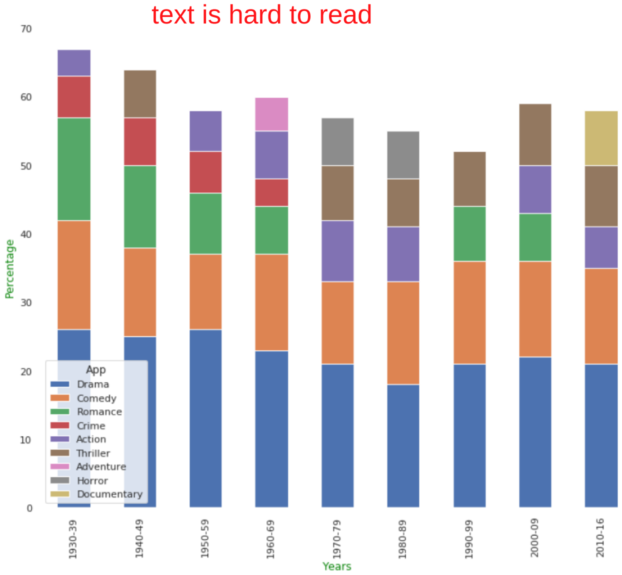
- In order to determine a score for the director, we have directly used the points that were provided by the IMDB dataset.
- In case of actors, we have taken the top three actors and we have take the summation of their corresponding points from the IMDB dataset.

$$\boxed{\begin{aligned} cast\ score = \alpha * actor1+ \\ \beta * actor2+ \\ \gamma * actor3 \end{aligned}}$$

where α , β γ are cast-scores of lead actor1, lead actor2 and lead actor 3 respectively.

Genre Score:

text is hard to read



We show the most popular movies in each decade from 1930-2010.

From the above graph we can observe that genre can shift over the decades and could be one prominent factor which can really effect the endurance of a movie. So, we have defined a score for genre of a movie as genre_score as the percentage of the number of films made on a genre over a decade

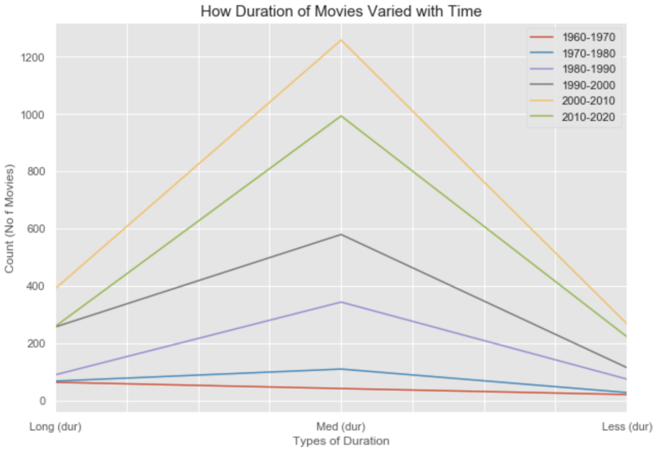
$$\boxed{genre\ score = \frac{movies\ Of\ a\ Genre}{Total\ Movies\ per\ decade} * 100}$$

We have calculated the genre score for movies separated by decade. The intuition behind this idea is to give a good genre score for movies which are a popular choice in that era.

Duration score Score:

Duration of a movie has varied over the decades. We are curious to know if it makes any impact on the endurance of a movie. We have segmented the movies into three categories based on the duration.

$Small : Duration < 90mins$
$Medium : 90 <= Duration < 120mins$
$Long : Duration >= 120mins$



Hence, We have defined duration score as the percentage of the number of films made on a genre over a decade:

$$\boxed{duration\ score = \frac{Movies\ Per\ Segment}{Total Movies} * 100}$$

PREDICTION MODELS

We have defined our prediction models are on the basis of regression. As a target variable, we have added endurance metric (as one of our column in the dataset. After that, we divided our dataset into a raining to test respectively using train-test-split. We calculated Root Mean Square Error value for each prediction model to compare their accuracy. Then, we decided to evaluate our models by finding the top movies.

Models:
Baseline Models:

1) **Linear Regression Model:**
We ran linear regression model on the dataset and got a root mean square error of 0.1032.

2) **Lasso Regression Model:**
We ran lasso regression model on the dataset and got a root mean square error of 0.1070.

3) **Ridge Regression Model:** what is ridge parameter?
We ran ridge regression model on the dataset and got a root mean square error of 0.1032.

great to always have baselines

Advanced Models:
In the advanced models, we have ran our predictions on LightGBM, XGBoost and SVM. We decided to use these advanced models to run our predictions as they provide several parameters along with advanced kernels. SVM's are meant to perform similar to neural netowrks model. For our use case we decided to choose Support Vector Machine(SVM) instead of neural networks mainly because of the fact that Artificial Neural Networks(ANN) often tends to overfit. Also, Artificial Neural Networks usually converges on the local minima than the global minima. All the three advanced models chosen by us come with different kernel options. We especially wanted to use tree based model such as LightGBM and XGBoost, so that we can go in further depth of their decision trees.

Given that, a decision tree seems like it would have been a more reasonable baseline

1) **LightGBM Regression Model:**
We ran LightGBM regression model on the datset with the following parameters: learning rate = 0.05, boosting type = 'gbdt', objective = 'binary', metric = 'binary logloss', sub feature = 0.5,num leaves =

Top 10 Endured Movies
Titanic
Avatar
The Dark Knight Rises
The Dark Knight
Star Wars
The Warrior's Way
Star Wars: The Force Awakens
Inception
Avengers: Age of Ultron
The Exorcist

It would be better to accumulate all these results in a table somewhere so we can compare them

TABLE I: Current popularity movie attributes

10, min data = 50, max depth = 100. We observed a root mean square error of 0.06126.

2) **XGBoost Regression Model:**
We ran xgboost regression model on the datset with the following parameters, objective ='reg:linear', colsample bytree = 0.3, learning rate = 0.05, max depth = 100, alpha = 10, n estimators = 100. We observed a root mean square error of 0.06626.

VI. Be a a Consultant - Why do I care?

- **The Robustness of the Metric:** The endurance metric proved to be a very robust measure for popularity endurance indeed. We were able to predict some of the all time most popular movies.(All of them are in the IMDB top-250).

- **Myth Busted - Top rated movies are not the most popular movies:** We found out that top rated, or critically acclaimed movies are necessarily not the most popular ones. Rather, popular movies, the ones which touch the heart of the viewers, are often from lighter genres like drama, fantasy etc. Of course these movies have amongst them quite a few critically acclaimed titles, but the correlation between current popularity, endurance of popularity and ratings continue to tell a different story.

These are really interesting observations. It would be good to show them numerically somehow, via a plot or a table.

- **Historical Pattern of Popularity:** We found that with time, the movies got more complex in their story-lines. The genres of the movies intertwined more, and newer genres like thriller, horror, animation (due to advancement of technology) etc became more popular. This shows how the audience's taste of movies have grown matured over years, and the general trend of newly popular movie genres.

VII. Be a Critic – How can I Improve?

– **Absence of revenue details other than box-office gross:** We have argued that from mid-1970s, audience have chosen various other modes of watching a movie over going to the theatre. These alternative modes are home videos (mid-1970s), DVDs (1997), Television, etc. In fact, the domestic box office represented 80% of studio revenues in 1980, but by 1992 it decreased to no more than 25%. However, there is no way we can get the complete video-tape (or DVD) sales data for a given movie. Hope of getting the exact television airing details, and revenue earned are even more bleak. This data, if available would have been invaluable to deduce a better metric and prediction score. ✓

– **Limitation about our models:** Our advanced models such as XGBRegressor, LightGBMRegressor, SVMRegressor are dependent on the values of parameters such as Max Depth, Learning Rate, or the kind of kernels used to train these models. We chose RMSE (root mean square error) as our metric to decide the best performing model. We used a greedy algorithm approach, and kept on trying different combinations of parameters in a random way. ✓

Subsequently, we chose the models with the values nearest to the optimal set of values and performed further calculations from there. This could have been a better choice if we would have used more definite algorithms to choose the best parameters for our models.

- **Limitation about feature engineering:** The analyses of movies are highly dependent on financial data. However, due to scarcity of real records, we had to impute financial data by a regression model (as explained in the data pre-processing section) in quite a few places. Even though slightly, we guess this has impacted our prediction score. ✓

References

[1] Pangarker, N.A., Smit, Eon. *The determinants of box office performance in the film industry revisited*. South African Journal of Business Management,2013

[2] https://www.usatoday.com/story/life/movies/2016/07/17/why-movies-endure-top-gun-aliens-ferris-bueller-stand-by-me-30-anniversary/86780790/. "[Why movies endure]".

[3] https://www.kaggle.com/tmdb/tmdb-movie-metadata

[4] http://boxofficemojo.com/

[5] https://developers.themoviedb.org/3/getting-started/introduction

[6] https://www.imdb.com/list/ls009809456/

[7] https://www.vanityfair.com/hollywood/2014/09/shawshank-redemption-anniversary-story

[8] https://www.the-numbers.com/movies/production-companies/

I think this is a great first pass of a project. You capture a very interesting problem, and your feature engineering / subject motivation are very solid.

The machine learning aspects can be improved a bit. It is good that you establish some regression baselines, but since 2 of your models are tree-based and one is a classification model, there could be more baselines relevant to these, like a simple decision tree. This could start an exploration as to why different models performed differently (which was also hard to see because the results are not shown next to each other).

The introduction and discussion sections definitely show interestingness, but it is weak because you write the conclusions, but don't back them with experiments or plots.

If you improve on these areas, and address some of the issues you mentioned in your "critic" section, I could see this as a submission to a data mining conference, like KDD or WSDM. The interestingness is definitely there!

min crit: 5
writing: 1.5
interestingness: 1.5
overall 8/10