08_MovieMining_Part2

Discovering Trends in Film Runtime

Caleb Stillman
Computer Science Post Bacc.
University of Colorado Boulder
Boulder Colorado United States
cast9172@colorado.edu

Graydon Sinclair
Computer Science Post Bacc.
University of Colorado Boulder
Boulder Colorado United States
grsi7978@colorado.edu

Stephen Tynan
Computer Science Post Bacc.
University of Colorado Boulder
Boulder Colorado United States
stty1424@colorado.edu

Problem Statement / Motivation

Group 8's project goal is the mining of data from a repository with information on over 950,000 films. The mining will have a focus primarily on the runtime feature of the data. It will look at both its impact on other features and the way other features impact it.

Some questions that will be answered via data examination include whether it is possible to predict the runtime of a film based on its genre, country of origin, and/or date of release. Another goal is to identify the ideal film length associated with high ratings.

The intent of determining if these relationships exist, and to what extent, is that this information could be applied within the film industry. This application would theoretically help produce movies with a greater likelihood of positive public reception, resulting in an increased box office.

Literature Survey

A survey of existing literature suggests that some work has been done to examine the impact of runtime on other facets of films. However, the work has been minimal.

A piece in Cognitive Research Journal [1] speaks to runtime tangentially as it related to trends in the pacing of films throughout history. However, the author examines a limited 210 films and only includes films within an arbitrary runtime cutoff at the two-and-a-half-hour mark.

Another article from Film and Digital Media [2] does focus on runtime but is a shallow piece. The intent of this project is to mine further information than is presented in this strictly topical article.

IMDBPro [3] published an article that focuses on how the success of a film is measured, which is related to this project, but it does not speak to runtime at all. This may lead one to believe that runtime is an overlooked aspect of success.

Proposed Work

This project will begin with EDA, including concurrent data cleaning and preprocessing.

As part of the EDA process outliers in the data will be identified and handled. For instance, a cursory glance at the data shows that there are movies over the 7-hour mark in the data set. Both whether these are actual movie lengths or bad data, and how the outlier is handled needs to be determined. Some of the outliers may be determined from area expertise instead of raw data. Such as the fact that the data set was last updated in July 2024, meaning that any films with release dates past June 2024 likely have bad or limited ratings data as only select viewers may have seen it. This kind of outlier will also need to be determined and handled.

Not only will outliers need to be either removed, transformed, or replaced depending on the type of feature but the overall data characteristics will be recorded and documented. During EDA visualization will be leveraged to help determine if any patterns arise. Visualizing the data and documenting characteristics will help understand the distributions and relationships between features and may influence which features are kept for which parts of the project.

In an earlier iteration of this project, it was discussed that other data sets may be brought in to get additional information on box office, budget, and ratings for films. Since the initial planning it has been decided to abstain from these additions. The primary data set from Letterboxd includes viewer ratings which can be indexed as a success metric. It was determined that there is already a large sum of literature on the analysis of the box office and budgets of films [3] and additional research would be less impactful than that pertaining to runtime. As there are no other data sets being leveraged there is less data preprocessing and integration than initially planned.

Some integration is still required however, as the Letterboxd data set is split into ten separate sub-sets in CSV format. A primary key is shared across all these sub-sets allowing for merging during data frame creation. Other steps in the preprocessing will include transforming categorical data into numerical, using one-hot encoding, so that the information can be more easily parsed. Depending on which evaluation techniques are being leveraged there may also be instances of classification occurring, e.g. the assignment of ratings to categories such as great, good, okay, poor, and bad.

The evaluation stage comes after the EDA, cleaning, and preprocessing, wherein primary features and correlations will be discovered. Mean Squared Error (MSE), regression analysis, Mean Absolute Error (MAE), precision, recall, F1 scores, and lift are just some of the measurements that will be used to determine the strength of correlation and general relationships between the features mined. While precision and recall do not require binary metrics lift and F1 scores are better suited for them. With this in mind some additional avenues for model building may be better suited for final analysis. Models may be built with specific limitations on features in mind and/or arbitrary rules may be tested to see which yields the most interesting data.

Data Set

The data set is sourced from the website / app Letterboxd and hosted on Kaggle [4]. Letterboxd is a social media film site that aggregates film data and viewer interactions such as ratings, likes, and views with said films. The data set has been downloaded locally in case of an unforeseen event where it is no longer available via the web.

The data set is quite large. It includes 10 sub-sets of data and has information on over 950,000 films. Each of these films can have many entries across the sub-sets of data. For instance, one film may have 3 genres (rows) of data in the genre sub-set alone. The sub-sets include information on genres, country of origin, actors, crew, studios, languages, poster urls, release type and date information, themes, and the movies themselves. The sub-sets vary in size. As mentioned above each film can have many rows in each sub-set other than the movies sub-set. No sub-set has more than seven columns (including the primary key).

The data set is updated annually. It was last updated July 2024, meaning the runtime data included is very up to date for current research. However, when looking at some features, such as ratings, it may be important to note that this feature may not have stabilized for a movie that came out close to the data set update. This is another candidate for domain knowledge-based outliers for certain features.

It is important to note that most of the data in these data sets is categorical. Therefore, if regression is to be done on any of these subsets or the final monolithic set then some encoding will be required.

Evaluation Methods

For evaluation accuracy measurements will be used when looking at the success of classification prediction. Lift will be measured to look at the performance and importance of the associations found. MSE will be used for regression analysis such as the prediction of runtimes. Precision and recall will be used to measure the success of model predictions. F1 scores will be leveraged for tasks such as genre classification. Matplotlib and Seaborn will also be used to create visual plots to aid in evaluation, such as providing a comparison between predicted and known values.

Tools

Python, and its associated libraries will be the primary tools for this project. The following associated libraries will be used for mining the data; Pandas for producing and manipulating data frames, NumPy to aid in calculations, Seaborn and Matplotlib to produce plots and other visualizations, Scikit-learn to help build regression models. Specifically, Scikit-learn will also provide additional tools and functionality for manipulating the data frames, such as splitting variables into testing and training data sets. There was initial discussion around the use of Orange, but due to the size of the data set it was determined to not be ideal for this project.

The size of the data set limit has limited the tools available for use. Due to its size the only library and option for cleaning and merging the data was pandas. An initial merge of all the data lead to a data frame with millions of entries. This is too large for viewing. Excel for instance has a ~ 1 million row limit. However, after additional cleaning and encoding the size of the data set was drastically reduced providing an opportunity to review and use additional tools again.

Milestones

1.1 Identification

Expected completion date: 10/01/24

1. Identification of viable datasets
2. Identification of the primary and secondary features.

2.1 EDA – Cleaning

Expected completion date: 10/15/24

Cleaning of data:

Removal of NaN and bad data

Identification and potential removal of outliers (by both domain expertise and statistical determination)

2.2 Data Preprocessing

Expected completion date: 10/31/24

Data has been combined into a singular set containing all pertinent data.

Transformation of data – e.g. one-hot encoding

Trend analysis and precursory evaluations

3.1 Analysis

Expected completion date: 11/30/24

Complete research and analysis on the data. Apply evaluation methods and add visualizations.

Build a predictive model and compare model to test data.

Apply methodologies to the following big questions (more may arise during EDA):

1. What is impact of runtime on rating? / What is the ideal runtime?
2. What is impact of genre on runtime?
3. What is impact of country of origin on runtime?
4. What is the impact of runtime on rating by country of origin?
5. What is the impact of runtime on rating by genre?

4.1 Finalize Results

Expected completion date: 12/6/24

1. Finalize evaluation
2. Synthesize results – determine if there are new questions to be asked
3. Produce final presentation.

Milestones Completed :

1. Identification
2. EDA - Cleaning
3. Data Preprocessing

Milestones To Do :

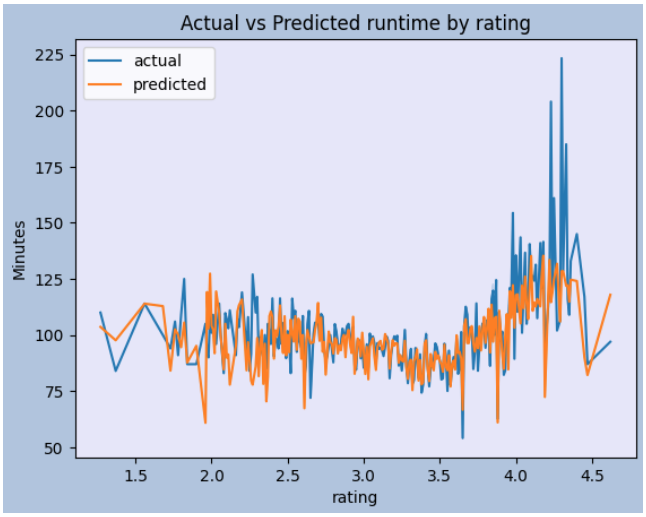
1. Analysis
2. Finalize Results

Results So Far

The EDA and cleaning of the data set uncovered some unforeseen issues. As regression is going to be used as a key feature for the project it was determined that we must encode the categorical data. This is also important as most libraries that are going to be leveraged, such as sklearn, do not generally handle categorical data. To complete this vital step one-hot encoding was used. This lead to more complications as the data set became so large if all features were one-hot encoded that it was unmanageable. Some features were one-hot encoded as is, such as genre, ensuring that all potential feature values were kept. However, features such as studios had unmanageable amounts of potential values. Studios alone had 160,882 unique values, so direct one-hot encoding was not an option. With studios, and other features such as release country and languages, some purging of data was required. To complete this the top n numbers of values were kept, where n marked a value where there was a precipitous drop in associated films or there was already a healthy amount of data points above the value at n. For instance, countries that released more than 10,000 films were left in the data set, and those with less were removed, leaving a total of 27 features to be encoded for this category. This process allowed for a large number of films to remain in the data. After all the cleaning and encoding the final data set had 17313 entries with 247 features.

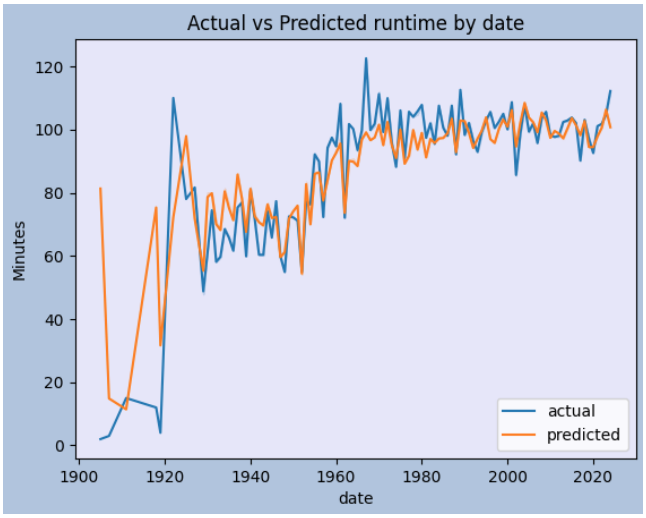
As mentioned in the Proposed Work section above we have leveraged visualizations during our preliminary EDA to determine whether any patterns arose in the data, both those we expect to find or those we might find surprising. To this end, we fitted and trained a linear regression model on the “minutes” (of runtime) feature. This was used to plot a few graphs: Actual vs. Predicted Runtime by Rating, Actual vs. Predicted Runtime by Date, Actual vs. Predicted Linear, and Rating vs. Minutes. These will each be addressed in turn below.

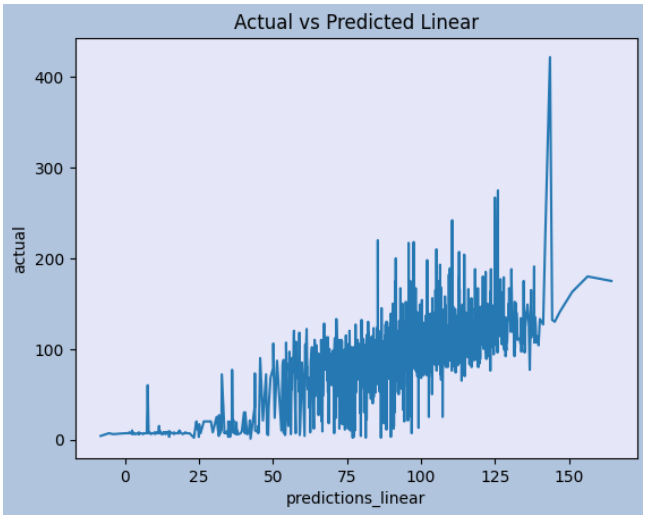
The Actual vs. Predicted Runtime by Rating graph served well as a test of our preliminary linear regression model.



The results of the graph show an encouraging level of correlation between the actual and the predicted runtimes by rating, particularly among middle-range values, indicating that one of our most important questions is indeed worth further investigation. That said, the graph revealed that our model could use tuning to provide better predictions around the lower and upper ranges of ratings.

Another feature we graphed with runtime was release date. Our interest in a preliminary graph of these features together at this early stage was primarily to discern how trends in runtime have developed over time.



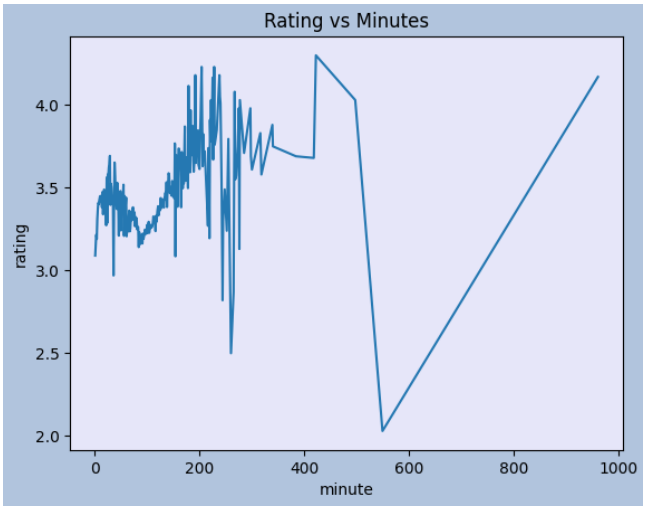


The results of this graph did largely confirm our expectations, showing that early films had a much shorter average runtime. There was a large spike in average length of films in the early 1920's, stabilizing at a much higher range of runtime (60-80 minutes, up from <20 minutes) beginning around the release of the first non-silent film, The Jazz Singer, in 1927. Once sound was introduced, runtimes stayed in roughly the 60-80 minute duration range for about the next 20 years. Starting in the 50's, film length began to creep up over the course of the next decade, stabilizing again in the late 60's/early 70's to an average length range between 90 and 110 minutes, a trend in film length which has continued through the present day. Here again, we tested our linear regression model and saw that our model predictions follow actual values to an encouraging degree (the exception being in the earliest two decades of film history).

In speaking about our model, our next plot was of actual runtime values vs. our predicted linear regression values, to determine the initial efficacy of our linear regression model.

One thing of note in this graph is that one can see that there are not many films included in the dataset with runtimes between 0 and 25 minutes, and there are a few included in our dataset with runtimes between 25-40 minutes, but after this point is where the data really spikes. This is as expected, as the dataset is meant to capture feature films. The Academy of Motion Picture Arts and Sciences, the American Film Institute, and the British Film Institute all consider feature films to be those with runtimes greater than 40 minutes [5]. We are able to see that the data is most tightly clustered around the 45-138 minute range, which is also to be expected, as we could see on the previous graph as well. That said, this graph also shows that we need to work on fine-tuning the performance of our linear regression model. We can also conclude that we may want to look into other metrics such as RMSE to supplement the findings of our linear regression model.

Lastly, we created a graph of the pure rating vs. minutes of runtime, without also graphing the results of our model.



The most striking revelation of this plot is that we have further cleaning on outliers to complete, as there is not much data worth noting past the 350 minute mark, but we have data in the 400-1000 minute range which is probably contributing to the skewing of our linear regression model.

All told, our results thus far have been largely encouraging that we're on the right track, seeing data that backs up conclusions we would expect about known runtime trends, and that we have at least a good start on a prediction model for answering some of our proposed questions.

REFERENCES

[1] James E. Cutting, 2016. *The evolution of pace in popular movies*, Cognitive Research Journal DOI: https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-016-0029-0

[2] acmk19. 2017. *Time Isn't Money: Does Runtime Affect a Film's Box Office?*, Film and Digital Media, DOI: https://filmanddigitalmedia.wordpress.com/2017/10/19/time-isnt-money-does-runtime-affect-a-films-box-office/

[3] *How is the success of films and TV shows measured?*, IMDBPro DOI: https://pro.imdb.com/content/article/entertainment-industry-resources/featured-articles/how-is-the-success-of-films-and-tv-shows-measured/GLFTC8ZLBBUSNTM3

[4] Letterboxd Kaggle Data. DOI: https://www.kaggle.com/datasets/gsimonx37/letterboxd/data

[5] Feature Film Wikipedia Data. DOI: https://en.wikipedia.org/wiki/Feature_film