Movie data prediction

Technical report

Table of content

[1. About the project 1](#_Toc180947310)

[2. Data Collection 1](#_Toc180947311)

[2.1 The Numbers 1](#_Toc180947314)

[2.2 IMDB 1](#_Toc180947315)

[2.3 OMDB 1](#_Toc180947316)

[2.4 MovieLens 1](#_Toc180947317)

[3. Data Wrangling/Preprocessing 1](#_Toc180947318)

[3.1 Linking different datasets 1](#_Toc180947320)

[3.2 Adjustment for inflation 2](#_Toc180947321)

[3.3 Average profits, average production budgets and number of films 2](#_Toc180947322)

[4. Methods used to process data 2](#_Toc180947323)

[4.1 Visualization 2](#_Toc180947325)

[4.2 Vectorization of keywords 2](#_Toc180947326)

[4.3 Clustering of actors, directors and writers 2](#_Toc180947327)

[4.4 Prediction model 2](#_Toc180947328)

[5. Web application 2](#_Toc180947329)

[6. Added value and conclusions 2](#_Toc180947330)

[Appendix 3](#_Toc180947331)

# About the project

The idea was to gather data about movies and try to predict the profitability and success of the movies with features such as actors, directors, writers, release time, production budget, genre etc. We decided to use the US domestic box-office revenue as the measure for the profit of the movie. There were couple reasons for this decision. One was that if we would have used the gross revenue, that would include VFS, DVD sales and TV and streaming profits. That would give benefit for older movies, since they would have had more time to generate gross. Some movies might get shown in theatres multiple times during long periods of time, but we assumed, that those movies and the revenue they would generate from reshowing the movies, would be marginal. We also used only US domestic box office revenue, simply because that was available for us. So, the profit we have computed and used in this project, is not the overall profit. It still gives us a pretty good approximation of the success of the movie.

# Data Collection



## The Numbers

The Numbers (<https://www.the-numbers.com/>) is a database that holds production budget and box-office records for movies. This was the only page we were able to find that had production budgets readily available for our use. We asked for an API for this website, but their API was available only for payment. Luckily this website showed the data in tables with 100 entries at a time, so we were able to gather the data by copying and pasting. The website showed production budgets for 6531 movies, and we copied the information of all those movies. Like stated above, we wanted to use the box office revenue, and not the gross, so we could only use the *release date, movie name* and *production budget* from these tables. Example of the data tables in The Numbers is in the appendix.

## IMDB

IMDB had subsets of their data available for non-commercial use as .tsv files (<https://developer.imdb.com/non-commercial-datasets/>). We used three different datasets from the IMDB database. The datasets and the features of the IMDB data are shown in the appendix.

## OMDB

## MovieLens

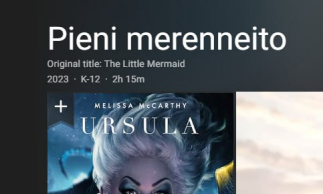
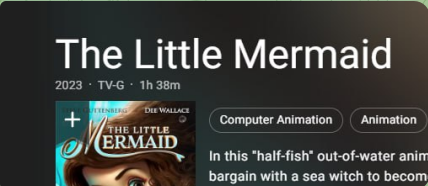
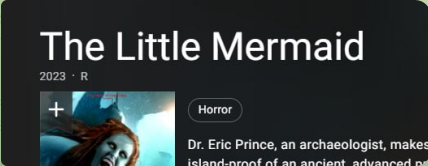
We used…

# Data Wrangling/Preprocessing



## Linking different datasets

The rest of the datasets had the imdb codes for each movie title, so we could use those to link the different datasets together, but the data from The Numbers didn’t include the IMDB codes. For the start, we had to discard 190 titles out of the 6500 titles, because they had duplicated title name and release year, and we didn’t have any means to link these to the right IMDB codes. For example, there were three different movies with the title “The Little Mermaid” released in 2023.



We used the title name to link the production budgets to the IMDB data. We also included the release year to the match, so movies with same titles but from different years wouldn’t mix. We found a match for about half of the titles. The rest of the titles had different spelling. For example *Star Wars: Episode VII - The Force Awakens* in the IMDB data vs *Star Wars Ep. VII: The Force Awakens* in The Numbers data. We used SequenceMatcher from the difflib in Python with threshold 0.8 to match as many more titles as we could. In the end, we were able to match total of 5859 movie production budget to their IMDB codes. Rest of the linking between datasets was done with the IMDB code.

## Adjustment for inflation

To be able to compare the production budgets from different years. We computed inflation adjustments for all production budgets and box office revenues.

## Average profits, average production budgets and number of films

For all actors, directors and writers, we computed the number of the movies they have been involved in and the average profit and average production budget of those movies. We then computed for every movie title the average of the average profits of the actors that were involved in that movie. This was also done for the average production budgets and number of films. This was repeated for the directors and writers as well. These averages were used in the prediction models to evaluate the crew, their experience and their effect on the success of the movie.

# Methods used to process data



## Visualization

For basic EDA we did simple visualizations. Scatterplots and histograms to see how different variables relate to each other and how the variables were distributed. Based on this we picked the most interesting relations between the variables to use in the final model and chose the most interesting findings to display on the website. We mainly used the matplotlib and seaborn libraries to make plots and histograms with colour coding.

## Vectorization of keywords

Already in the beginning of the project we decided that we want to find out what keywords and tags are most related to the successful movies. We ended up using TF-IDF method to try to specify words in different documents, the documents being classes of movies with different success. We decided to use the profit ratio as the metric for success. So, we decided to divide the movies into six different categories based on their profit ratio. Original plan was to use the tags on the movies as well as the synopses, but the tags would have had to be scraped from web. We had hopes on using the tags from MovieLens research dataset but after combining the movies from that dataset with the movies we had the production budgets for, it would have reduced the final size too much. Eventually we decided to just use the synopses.

What we did was cluster the synopses together in each group, lowercased the text, removed the punctuation and stopwords and used the nltklibrary to stem the words. After that we could use the documents for the word vectorization. We used the TfidfVectorizer from the sklearn library for the method. The result was not impressive though, all the six classes of movies ended up having almost the same words as the most characterizing words even between the most profitable and least profitable classes. There were a few distinct words between the classes, so we ended up displaying the 30 most profitable and least profitable words on the website. The significance for the word vectorization in the overall project ended up being much less than we anticipated, however.

## Clustering of actors, directors and writers

We used average production budgets, average profits and number of films of actors, directors and writers to place them in clusters using KMeans method from scikit-learn library. Before the clustering, we used StandardScaler to standardize the features. We computed the inertia of different number of clusters and used that to evaluate the optimal number of clusters.

## Prediction model

We used the TPOTClassifier and TPOTRegressor to find optimal prediction models for our data. Features that were included in the data for the prediction models was inflation adjusted production budget, runtime, release month, genre (each title had 1 to 3 different genres selected), and the average production budgets, average profits and average number of films of the actors, directors and writers. For the prediction models, after removing all rows with missing cells, we had a dataset of 4919 movie titles. We tried the TPOTRegressor to find model to predict the inflation adjusted box office revenue, profit (box office / production budget) and the logarithm of profit. Best regression model was a random forest regressor that predicted the logarithm of the profit. The model had a R2 score of 0.73 and mean squared error of 0.11.

We also used the TPOTClassifier to find best model to predict if the movie would make profit or not (profit<1 or profit>1). The best model was a random forest classifier with accuracy of approximately 0.84. Although the R2 score was ok for the regressor model, we welt that the MSE was too big for any meaningful predictions (notice that it was for the log of profit). We felt that the accuracy of the classifier was decent enough so we decided to use that as our predictive model.

# Web application

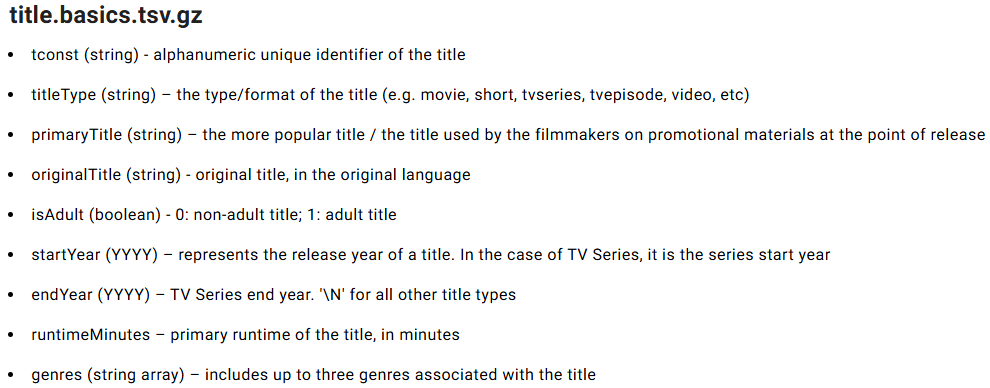
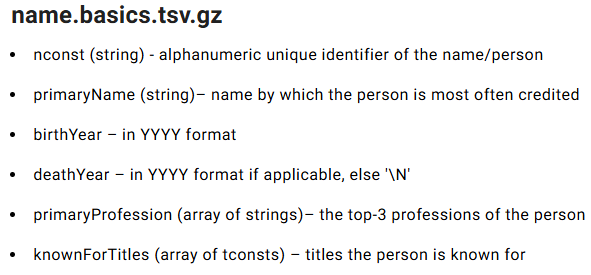
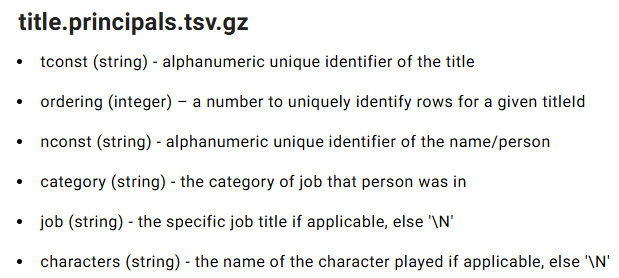
# Added value and conclusions

# Appendix

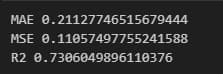
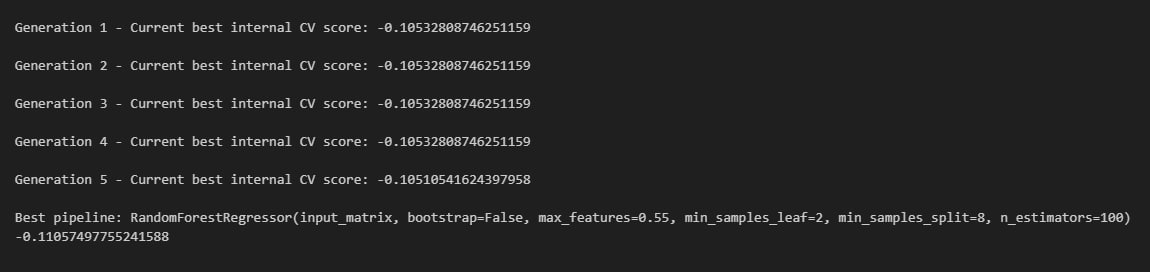
Kuva, joka sisältää kohteen teksti, kuvakaappaus, Fontti, numero

Kuvaus luotu automaattisestiExample of data from The Numbers:

Features in IMDB data:



Best regressor:



Kuva, joka sisältää kohteen teksti, kuvakaappaus, Fontti, algebra

Kuvaus luotu automaattisestiBest classifier: