PREDICTION OF MOVIE SUCCESS

Contents

1. Introduction

1.1 Background

Movies and TV Shows are one of the most influential and engaging forms of entertainment for almost everyone, however their financial success usually is very difficult to predict [18]. Production Studios invest significant resources into movies production and marketing [1], but the outcome (*box office*) often depends on numerous factors, such as genre, cast, director [25], timing and other factors, like competition and marketing campaign [1]. While popular platforms like IMDB provide ratings and reviews to measure audience reception [4], I believe that worldwide box office revenue remains the “gold standard” for quantifying a real movie's success [1]. Surprisingly, available datasets with detailed budget / revenue numbers are scarce, so there a gap in machine learning-based prediction efforts.

1.2 Motivation

This project related to the area that influence not only me personally, but everyone I know, as well as includes a desire to explore the practical applications of machine learning techniques. Having completed theoretical coursework and hands-on projects in data science and machine learning, I wanted to apply this knowledge in a domain I am passionate about – Movies and TV shows. Most publicly available datasets, such as those available on Kaggle, mostly focus on predicting audience-driven metrics like popularity or IMDB score ratings [5] [7]. However, I chose to focus on predicting worldwide box office revenue (*for simplicity, limited by movies only*) as it reflects the real commercial success of movies and provides an underexplored area for research.

1.3 Objective

The primary objective of this project is to develop a several machine learning techniques that could predict worldwide box office revenue, based on a wide range of features, including budget, actors participated, director, writer, movie genre(s), movie runtime, release year and others.

The objectives include:

- Collecting and preprocessing a comprehensive dataset by merging information from multiple publicly available sources, including TMDB API (*The Movie Database*) [9], IMDB available datasets [5] [6] [7] as well as box office platforms like “IMDB Pro” [8], “Box Office Mojo” [12] and “The Numbers” [11];
- Normalizing financial data (*budgets / box office revenue*) using historical Consumer Price Index (CPI) data to account for inflation [10];
- Exploring, analyzing feature importance and their impact on the result and applying machine learning techniques such as Decision Trees [3], Random Forests [25],

Gradient Boosting (*LightGBM*, *XGBoost*, *CatBoost*, …) and other regressors to train and evaluate performance of different predictive models;

- Comparing the performance of various models to identify the most effective approach for predicting box office revenue.

1.4 Why This Project?

This project offers a unique opportunity to deepen my understanding of traditional machine learning methods by applying them to a real-world problem. By collecting and curating my own dataset, I gain flexibility in feature engineering and problem exploration, avoiding the constraints of already available datasets.

2. Dataset Preparation

2.1 Data Sources

The dataset was created by merging information from multiple sources, including available datasets on Kaggle, TMDB API [9], IMDB publicly available datasets; Box Office Mojo website [12] and The Numbers website [11] for missing budget/revenue numbers. Financial data, such as budgets and revenues, was normalized using Consumer Price Index (CPI) data [10] to account for inflation changes. By combining all of those sources, the final dataset [16] captures a comprehensive picture of main features which are influencing movie financial success.

2.2 Data Collection

Data was collected programmatically using the TMDB API [9] to retrieve movie metadata (*as it's free for educational purposes and not applied any rate limiting for API calls*). However, due to missing any useful API endpoints, budgets / box office data from "IMDB Pro" website [8], "Box Office Mojo" website [12] and "The Numbers" website [11] was scraped using Beautiful Soup python library [13] (*due to applied rate limiting restrictions, those process performed over few weeks to collect ~40k movie details*). Finally, the datasets were joined based on unique identifiers such as '*imdb_id*', '*tmdb_id*' or movie title and release year, ensuring consistency and avoiding duplicates [16].

2.3 Data Cleaning

Missing financial values in box office 'revenue' were handled by dropping movies with no financial information (*as it's our prediction target and mandatory for us*). Text fields like '*overview*', '*tagline*', '*genres*', '*cast*', '*director*', '*writers*', '*producers*' and '*production companies*' were preprocessed by converting them into lowercase and removing special characters (*to get a better match from different datasets*). Financial columns (*like 'budget' and 'revenue'*) were adjusted for inflation using CPI data [10] (*targeting to current, 2024 year as baseline*), ensuring budgets and revenues across decades were comparable. Also, '*imdb_id*' column was replaced by averages for missing values.

2.4 Feature Engineering

All numerical features were scaled / normalized. In additional to that, budget numbers were log-processed to remove outliers. As there could be more than one genre, genres were one-hot encoded. All person-related fields (*like actors, producers, directors, …*) were updated to importance score, including appearance order and weighted contributions to past successful (or unsuccessful) movies. In addition, '*overview*' and '*tagline*' text fields were transformed to sentiment score (*using NLP technics*) – sentiment scores were extracted from movie overviews and taglines using a pre-trained sentiment analysis model [17] (*note: at feature selection, including those sentiment scores improved result by reducing RSME score by ~3.2%, so those features remained in final dataset*).

2.5 Final Dataset

The final dataset consisted of ~40k movies with 29 features, spanning from 1920 to 2024. Features included numerical variables like runtime, year, normalized budget and revenue, as well as categorical variables like genres, sentiment score and person-related fields scores. [16]

3. Data Insights /Analysis

This section explores the dataset's key insights, focusing on the metrics calculated for personalities (*like actors or producers*) and their contribution to movie success. Using a newly designed weighted scoring system, which was used for assessing personality past performances based on the following factors: profitability, box office performance, and IMDB rating. Also this section will cover the feature importance and reducing feature dimensions application (*via principal components, or PCM*). These insights will provide a foundation for understanding the features influencing movie revenue more.

3.1 Scoring Methodology for Person Involvement

To evaluate the importance of a person (*actor / director / producer / writer / …*) in predicting a movie's success, score for each person was calculated, based on their involvement in previous movies. The score integrates metrics like profitability, box office performance and IMDB ratings, with higher weights assigned to recent performances and leading roles (*order appearance in the list*). The used metrics calculated as:

$$\boldsymbol{Profitability} = Normalized\ by\ CLI\ Revenue - \ Normalized\ by\ CLI\ Budget$$

$$\boldsymbol{Norm\ Profitability} = \left|\frac{Current\ \boldsymbol{Profitability} - Average\ Profit}{Standard\ Deviation\ of\ Profit}\right|$$

$$\boldsymbol{Box\ Office\ Performance} = \frac{Normalized\ by\ CLI\ Revenue}{Normalized\ by\ CLI\ Budget}$$

$$\boldsymbol{Time\ Weight} = \frac{1}{Current\ Year - Release\ Year + 1}$$

$$\boldsymbol{Position\ Weight} = \frac{Total\ Number\ Personas - Position}{Total\ Number\ Personas}$$

$$\boldsymbol{Final\ Person\ Score} = \boldsymbol{Time\ Weight} * \boldsymbol{Position\ Weight} * (0.3 * \boldsymbol{Avg\ IMDB} + 0.4 * \boldsymbol{Norm\ Profitability} + 0.3 * \boldsymbol{Box\ Office\ Performance})$$

Based on the formulas and available final Dataset, we calculated the most impactful personalities for movie financial success, as example, Actors data (*top 3 more impactful actors*) would be looking like [16]:

1. Samuel L. Jackson:
Final Person Score: **1289.16**
Actor Weighted Value: **44.63**
Total Profitability: **$28,827,185,506.81**
Normalized Profitability: **$71.36**
Box Office Avg Performance: **3.29**
Appearances: **131**
Avg IMDb Rating: **6.40**
Top 5 Movies: "Jackie Brown", "True Romance", "Snakes on a Plane", "Jurassic Park", "Kill Bill: Vol. 2"

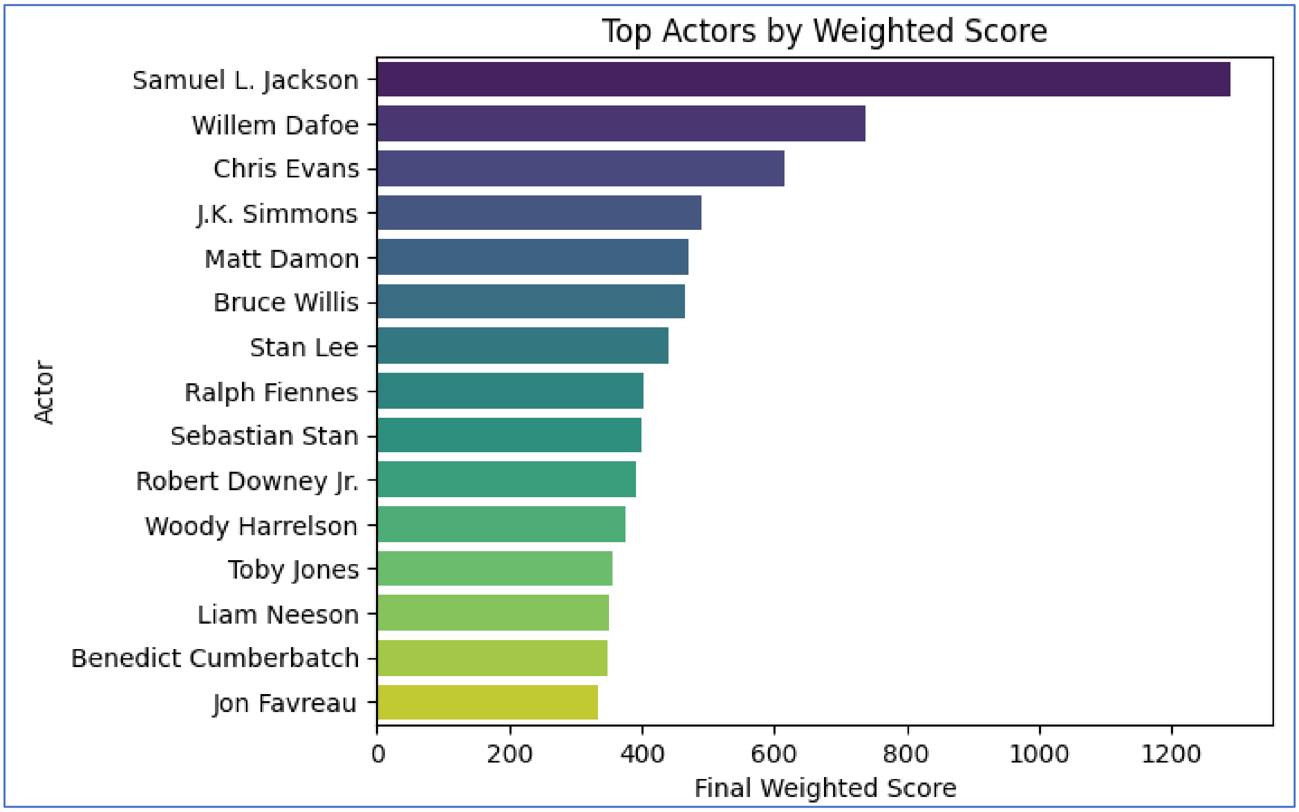
2. Willem Dafoe:
Final Person Score: **738.07**
Actor Weighted Value: **50.40**
Total Profitability: **$14,450,895,321.16**
Normalized Profitability: **$35.66**
Box Office Avg Performance: **2.82**
Appearances: **98**
Avg IMDb Rating: **6.69**
Top 5 Movies: "Inside Man", "The English Patient", "The Life Aquatic with Steve Zissou", "Wild at Heart", "Basquiat"

3. Chris Evans:
Final Person Score: **614.69**
Actor Weighted Value: **34.87**
Total Profitability: **$17,300,433,686.80**
Normalized Profitability: **$42.73**
Box Office Avg Performance: **3.56**
Appearances: **43**
Avg IMDb Rating: **7.02**
Top 5 Movies: "Street Kings", "Sunshine", "Captain America: The First Avenger", "Fantastic Four: Rise of the Silver Surfer", "London"

In the section below provided detailed information in tables for each person type (*top 15 values for each*) [16].

Top Actors and Metrics

The top Actors based on final score shown in the graph below:

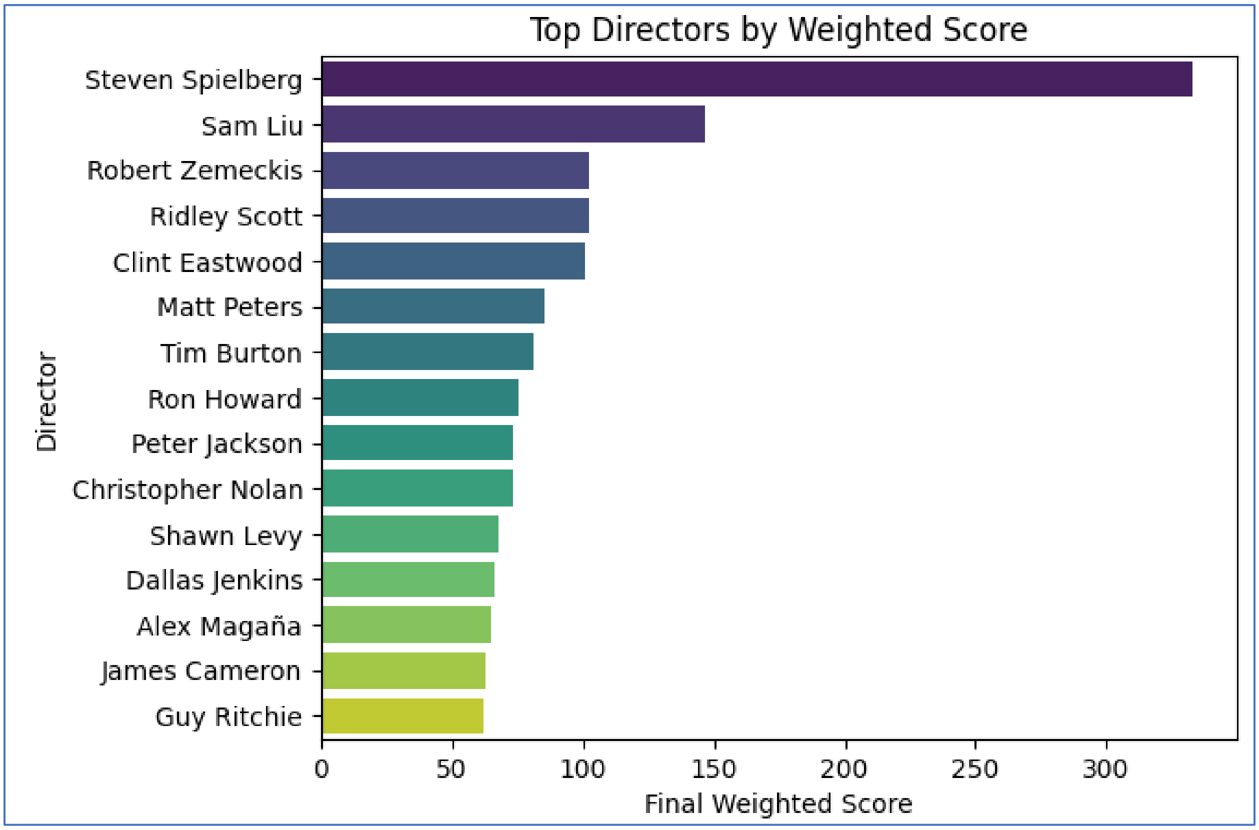


Same top 15 Actors displayed in the table with all metrics calculated:

Rank	Name	Final Score	Weighted Value	Total Profitability	Normalized Profitability	Box Office Avg Performance	Appearances	Avg IMDb Rating
1	Samuel L. Jackson	1289.16	44.63	$28.83 B	$71.36	3.29	131	6.4
2	Willem Dafoe	738.07	50.4	$14.45 B	$35.66	2.82	98	6.69
3	Chris Evans	614.69	34.87	$17.3 B	$42.73	3.56	43	7.02
4	J.K. Simmons	489.8	40.79	$11.79 B	$29.05	2.94	74	6.68
5	Matt Damon	471.22	38.78	$11.94 B	$29.43	2.75	76	6.7
6	Bruce Willis	465.95	27.08	$17.59 B	$43.45	2.86	120	4.69
7	Stan Lee	442.16	13.63	$32.14 B	$79.58	4.32	44	7.02
8	Ralph Fiennes	402.6	26.23	$15.02 B	$37.07	3.71	50	6.93
9	Sebastian Stan	400.18	33.97	$11.41 B	$28.1	4.09	32	6.88
10	Robert Downey Jr.	390.79	23.44	$16.39 B	$40.48	3.12	72	6.95
11	Woody Harrelson	374.92	36.14	$10.11 B	$24.87	2.81	78	6.85
12	Toby Jones	356.27	27.93	$12.56 B	$30.96	3.14	63	6.56
13	Liam Neeson	351.77	29.25	$11.99 B	$29.54	2.44	90	6.19
14	Benedict Cumberbatch	349.08	20.78	$16.5 B	$40.74	3.94	41	6.78
15	Jon Favreau	334.55	21.47	$15.22 B	$37.58	3.63	43	7.05

Top Directors and Metrics

The top Directors based on final score shown in the graph below:

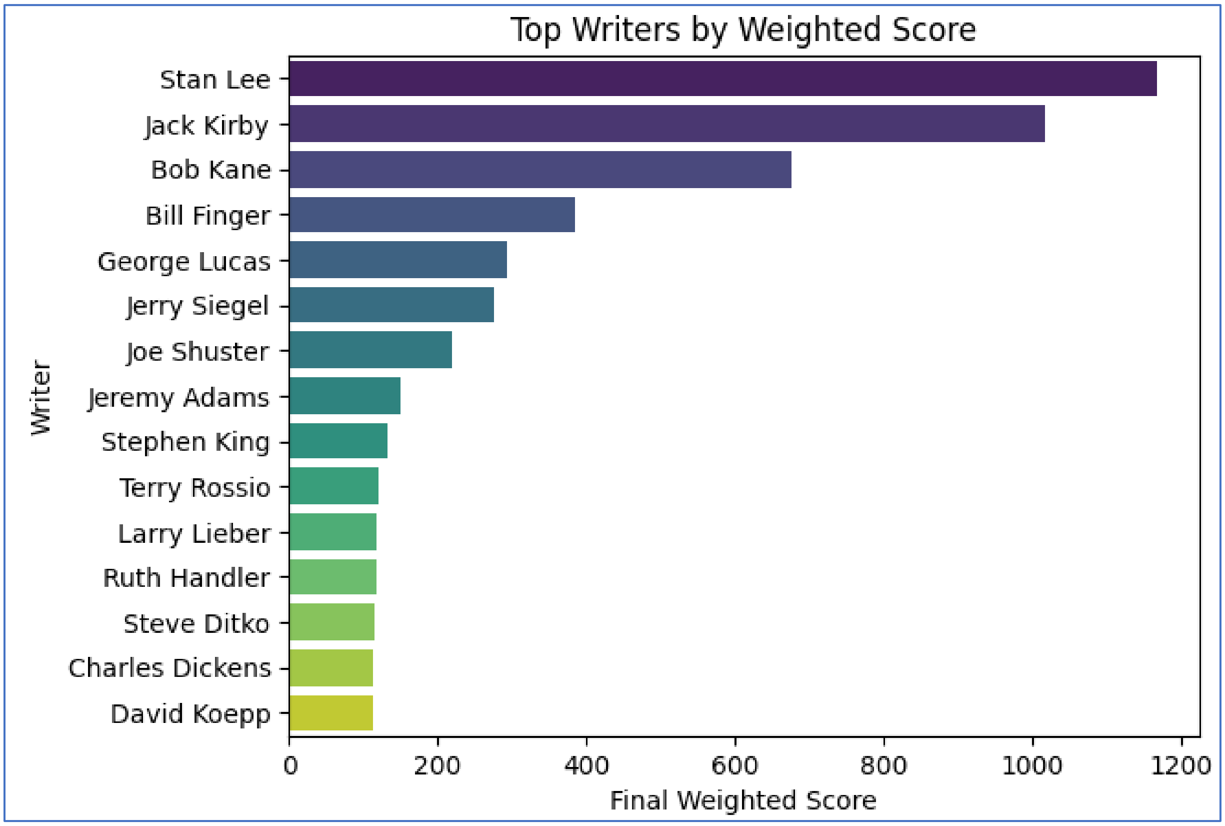


Same top 15 Directors displayed in the table with all metrics calculated:

Rank	Name	Final Score	Weighted Value	Total Profitability	Normalized Profitability	Box Office Avg Performance	Appearances	Avg IMDb Rating
1	Steven Spielberg	333.06	15.48	$21.86 B	$51.64	6.68	36	7.29
2	Sam Liu	146.78	17.01	$8.7 B	$20.41	4.6	21	6.52
3	Robert Zemeckis	102.45	15.78	$6.59 B	$15.38	3.61	22	6.35
4	Ridley Scott	102.15	21.88	$4.69 B	$10.89	2.38	28	6.61
5	Clint Eastwood	101.03	21.97	$4.45 B	$10.32	3.07	40	6.98
6	Matt Peters	85.65	12.16	$6.94 B	$16.23	6.19	9	6.35
7	Tim Burton	80.98	14.16	$5.66 B	$13.19	3.31	20	6.82
8	Ron Howard	75.16	13.09	$5.66 B	$13.17	2.85	27	7.05
9	Peter Jackson	73.7	8.87	$7.91 B	$18.53	4.26	14	8.16
10	Christopher Nolan	73.17	10.93	$6.24 B	$14.55	4.05	12	8.14
11	Shawn Levy	67.65	16.1	$3.99 B	$9.21	3.47	14	7.03
12	Dallas Jenkins	66.37	24.09	$2.16 B	$4.88	5.14	7	7.56
13	Alex Magaña	65.01	31.74	$1.38 B	$3.03	9.43	9	6.41
14	James Cameron	62.48	5.35	$11.62 B	$27.33	5.28	12	7.29
15	Guy Ritchie	62.46	22.27	$2.57 B	$5.84	3.13	15	6.95

Top Writers and Metrics

The top Writers based on final score shown in the graph below:

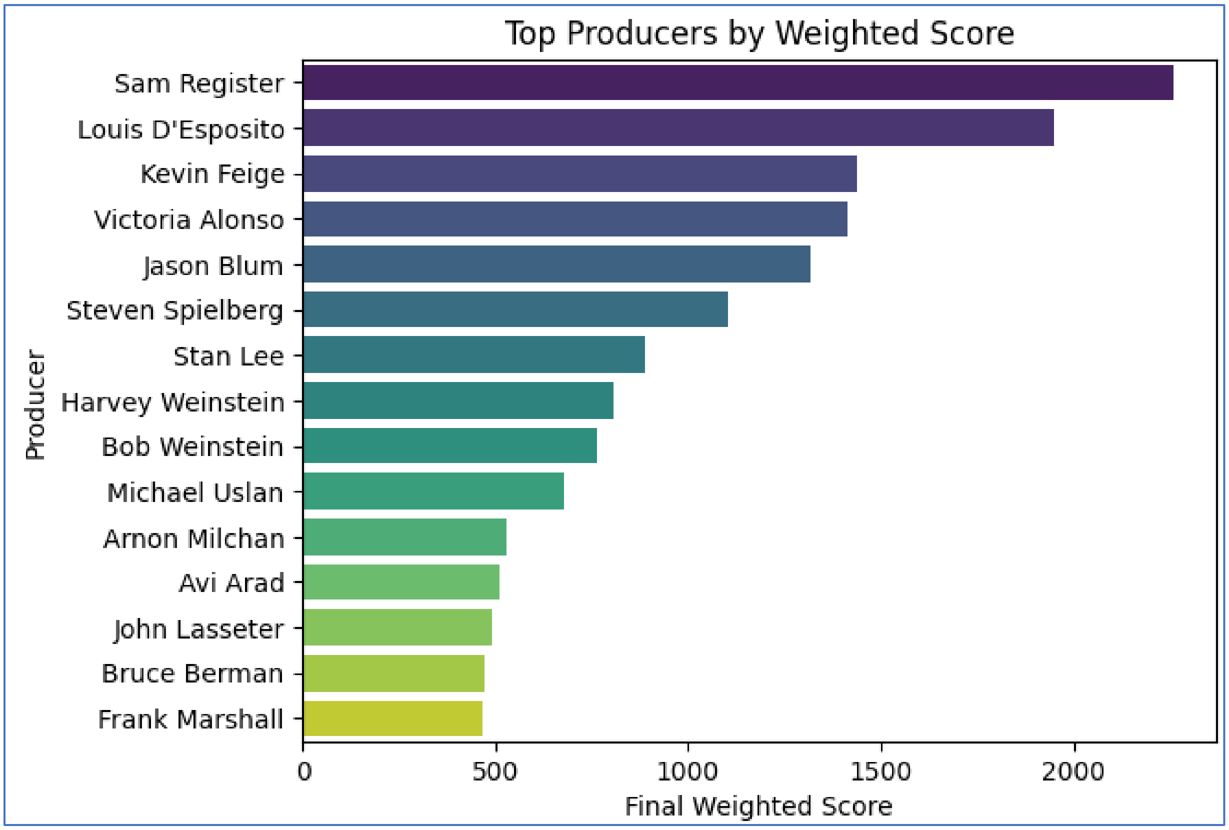


Same top 15 Writers displayed in the table with all metrics calculated:

Rank	Name	Final Score	Weighted Value	Total Profitability	Normalized Profitability	Box Office Avg Performance	Appearances	Avg IMDb Rating
1	Stan Lee	1167.71	31.57	$38.61 B	$91.39	4.1	47	6.53
2	Jack Kirby	1018.32	31.76	$33.31 B	$78.82	3.99	50	6.94
3	Bob Kane	675.63	30.51	$22.96 B	$54.26	4.12	50	6.57
4	Bill Finger	384.16	28.1	$14.03 B	$33.04	4.41	28	6.53
5	George Lucas	293.09	11.58	$26.06 B	$61.61	6.13	28	6.81
6	Jerry Siegel	275.51	21.61	$13.1 B	$30.83	3.75	30	6.59
7	Joe Shuster	219.65	17.93	$12.57 B	$29.58	3.73	29	6.61
8	Jeremy Adams	149.11	16.74	$8.85 B	$20.76	6.69	13	6.4
9	Stephen King	133.55	17.19	$8.09 B	$18.94	3.36	59	5.9
10	Terry Rossio	120.83	11.37	$10.96 B	$25.77	3.29	23	6.37
11	Larry Lieber	118.58	8.36	$14.43 B	$34.0	4.31	17	7.04
12	Ruth Handler	117.65	9.46	$12.92 B	$30.42	4.31	26	5.89
13	Steve Ditko	115.42	8.07	$14.8 B	$34.89	4.6	17	6.07
14	Charles Dickens	113.96	6.87	$17.06 B	$40.25	4.83	24	6.54
15	David Koepp	113.5	12.23	$9.49 B	$22.28	3.96	29	6.35

Top Producers and Metrics

The top Producers based on final score shown in the graph below:

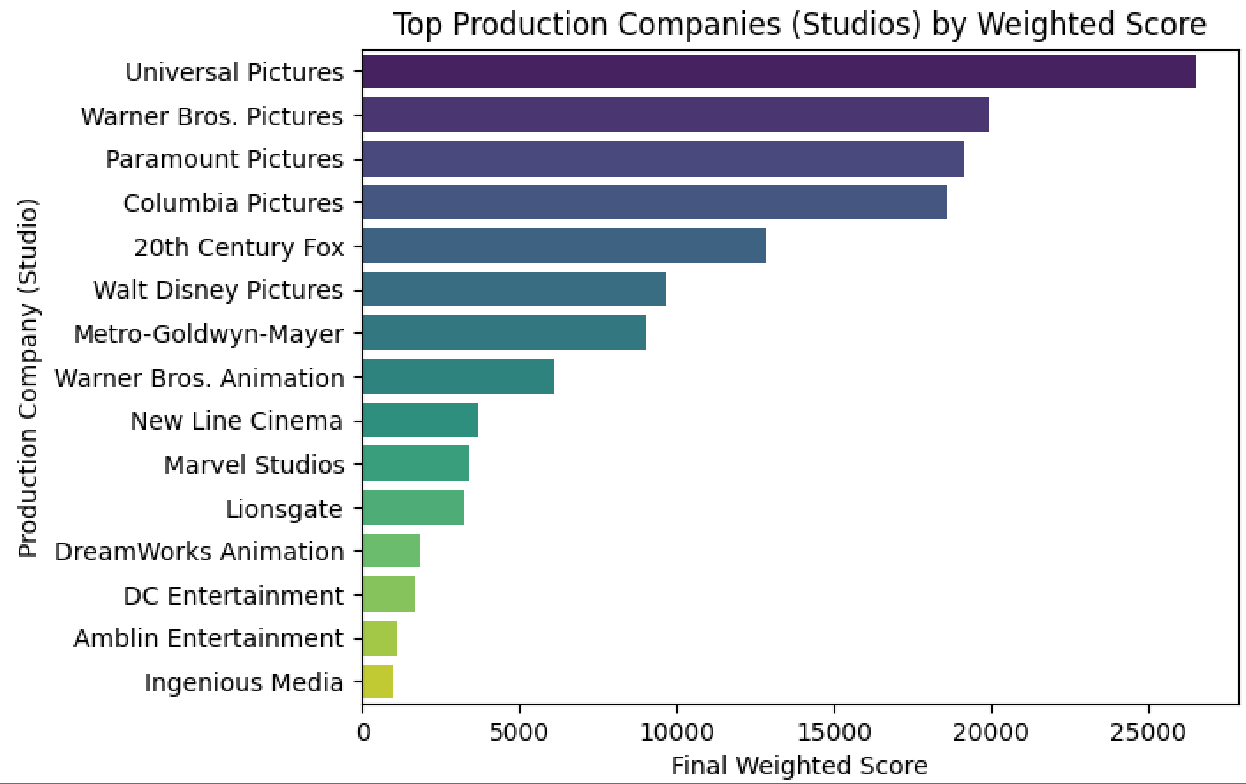


Same top 15 Producers displayed in the table with all metrics calculated:

Rank	Name	Final Score	Weighted Value	Total Profitability	Normalized Profitability	Box Office Avg Performance	Appearances	Avg IMDb Rating
1	Sam Register	2258.69	60.48	$38.92 B	$92.15	5.41	78	6.34
2	Louis D'Esposito	1951.85	54.59	$37.09 B	$87.79	4.56	49	7.12
3	Kevin Feige	1439.03	35.05	$42.72 B	$101.16	4.17	64	7.07
4	Victoria Alonso	1413.13	42.82	$34.19 B	$80.91	4.72	42	7.08
5	Jason Blum	1316.99	73.3	$18.8 B	$44.37	4.96	115	5.52
6	Steven Spielberg	1103.67	38.87	$29.51 B	$69.8	3.69	88	6.81
7	Stan Lee	889.79	24.1	$38.45 B	$91.01	3.98	63	6.88
8	Harvey Weinstein	806.91	47.18	$17.84 B	$42.1	2.76	228	6.34
9	Bob Weinstein	762.61	43.38	$18.35 B	$43.3	2.79	225	6.3
10	Michael Uslan	676.56	33.55	$20.9 B	$49.36	3.88	51	6.56
11	Arnon Milchan	528.54	36.42	$15.14 B	$35.68	2.56	140	6.3
12	Avi Arad	512.62	27.09	$19.69 B	$46.49	3.67	42	6.28
13	John Lasseter	492.03	17.86	$28.46 B	$67.3	4.2	49	7.21
14	Bruce Berman	470.52	30.7	$15.97 B	$37.67	2.48	97	6.4
15	Frank Marshall	466.65	24.54	$19.68 B	$46.46	3.61	69	6.67

Top Production Companies (Studios) and Metrics

The top Production Companies (Studios) based on final score shown in the graph below:

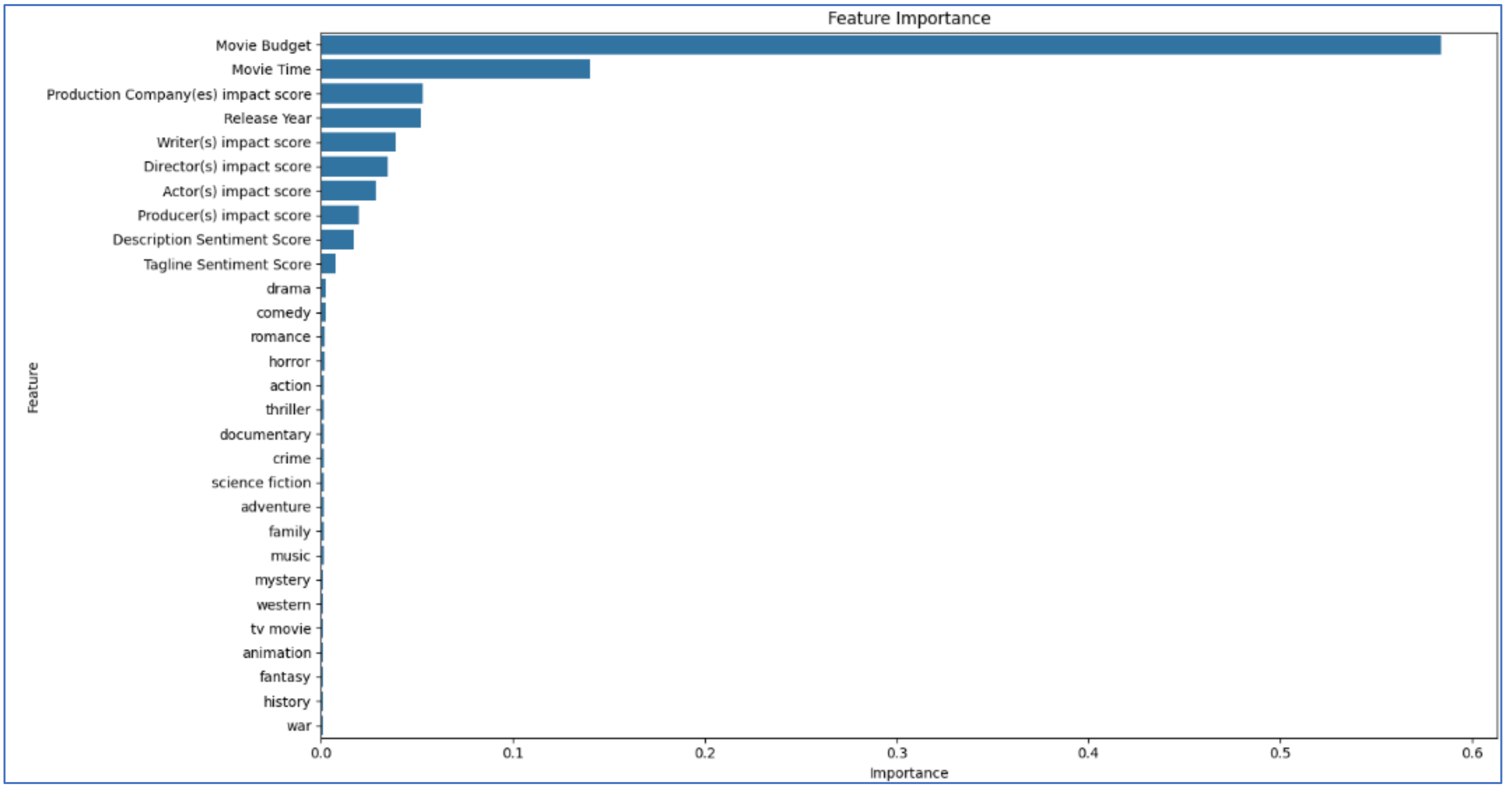


Same top 15 Production Companies (Studios) displayed in the table with all metrics calculated:

Rank	Name	Final Score	Weighted Value	Total Profitability	Normalized Profitability	Box Office Avg Performance	Appearances	Avg IMDb Rating
1	Universal Pictures	26520.51	260.6	$107.04 B	$253.85	2.79	822	6.2
2	Warner Bros. Pictures	19915.36	210.89	$99.3 B	$235.46	2.36	890	6.4
3	Paramount Pictures	19118.65	206.37	$97.39 B	$230.94	2.74	694	6.35
4	Columbia Pictures	18567.81	201.04	$97.11 B	$230.27	2.73	663	6.28
5	20th Century Fox	12847.24	97.8	$138.14 B	$327.66	3.45	639	6.25
6	Walt Disney Pictures	9630.02	144.83	$69.79 B	$165.41	2.96	254	6.5
7	Metro-Goldwyn-Mayer	9064.54	110.35	$86.33 B	$204.67	2.81	633	6.35
8	Warner Bros. Animation	6107.42	95.36	$67.08 B	$158.98	5.1	124	6.32
9	New Line Cinema	3691.72	106.75	$36.3 B	$85.92	2.94	309	6.1
10	Marvel Studios	3403.41	84.54	$41.87 B	$99.14	4.34	61	7.08
11	Lionsgate	3277.56	139.83	$24.71 B	$58.41	2.8	271	5.64
12	DreamWorks Animation	1858.42	64.28	$30.08 B	$71.16	3.77	69	6.69
13	DC Entertainment	1656.76	50.87	$33.89 B	$80.2	5.11	69	6.44
14	Amblin Entertainment	1106.06	39.79	$28.89 B	$68.33	4.12	90	6.67
15	Ingenious Media	984.89	68.17	$15.19 B	$35.81	2.51	122	5.9

3.2 Feature Importance

Based on final Dataset, and using the Random Forest model, the importance of each feature was calculated, based on how often and significantly it was used in the decision splits across all trees in the ensemble. The results, visualized in the accompanying bar chart below, indicate that the most influential feature is a ‘*budget*’, following by ‘*runtime*’ and ‘*Studio*’. While genres fields were less important. What’s interesting – sentiment scores, computed from Description (‘*overview*’) and Tagline were more important than specific flag for any genre.



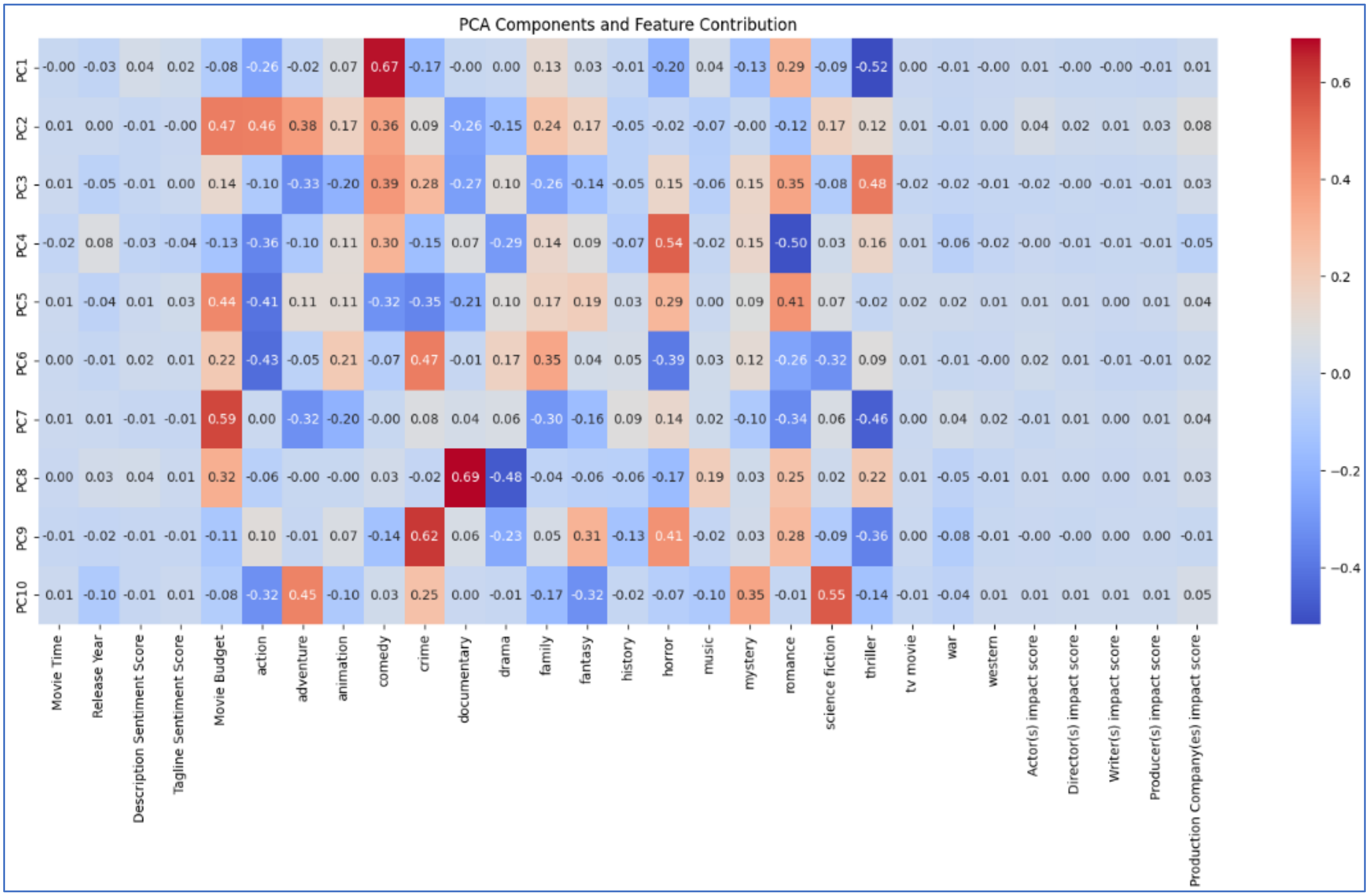
3.3 Principal Components Application

In addition to final Dataset direct usage of all prepared features, the PCM dimensionality reduction technique [21] was also applied with top 10 Principal Components. The goal of applying PCA in this study was to reduce the complexity of the feature set and improve the model's performance.

PCA was applied to the feature set, reducing the dimensionality to top 10 principal components [24]. However, the results showed that the model's performance declined:

- RMSE increased to 2.6359.
- R^2 score dropped to 0.6574.
- Mean Absolute Error (MAE) increased to 1.8485.

The decline in performance indicates that the dimensionality reduction removed some valuable information from our Dataset, which was required for accurately predicting movie financial success. While PCA is effective for simplifying datasets with high multicollinearity or redundant features, in this case, the removed variance likely contained critical predictive information about the target variable [21]. Based on the results, using the original prepared final Dataset without PCA application provided better prediction results. This highlights the importance of retaining detailed features, such as categorical and numerical variables, in this final Dataset. PCA Components and Feature contribution [21] to them displayed below:



4. Regression Models Training and Performance

4.1 Training Process and Models Evaluated

The following regression models were evaluated:

- *Linear Models*: Ridge, Lasso, and Elastic Net;
- *Nonlinear Models*: K-Nearest Neighbors (KNN) [26], Decision Trees [22], Random Forest [22], Gradient Boosting [32], XGBoost [31], LightGBM [27], and CatBoost [29];
- *Ensemble Models*: Combining multiple algorithms (*Random Forest, Gradient Boosting and Linear Regression*);
- *Dimensionality Reduction*: Using PCA to reduce features to the top 10 principal components.

The process of model performance evaluation includes the following steps:

a) Loading final Dataset and filtering all data with no revenue data (*~540k -> ~40k*);
b) Replacing person fields by lists (*to avoid any Test data involvement, we should apply any updates, like calculating actors scores or 'fit', only based on Train data available and after that apply that data or 'transform' on Test as well Train*);
c) Replacing 'overview' and 'tagline' by sentiment scores (*using precalculated sentiment.polarity from TextBlob*) [17];
d) Drop all unused features;
e) Split Dataset to Train and Test (*80% train and 20% test*);
f) "*Fit*" scalers, person score calculation on Train subset;
g) "*Transform*" both Train as well as Test subsets (*avoiding influence from "non-seen" Test*);
h) Train and check prediction for different models, using different regressors.
i) Calculate better hyperparameters using GridSearch library;
j) Evaluate performance of the models.

4.2 Regression Models Performance and Results

The models were compared using the appropriate for regression metrics (*as well as suggested in comments for Final Report proposal*), such as:

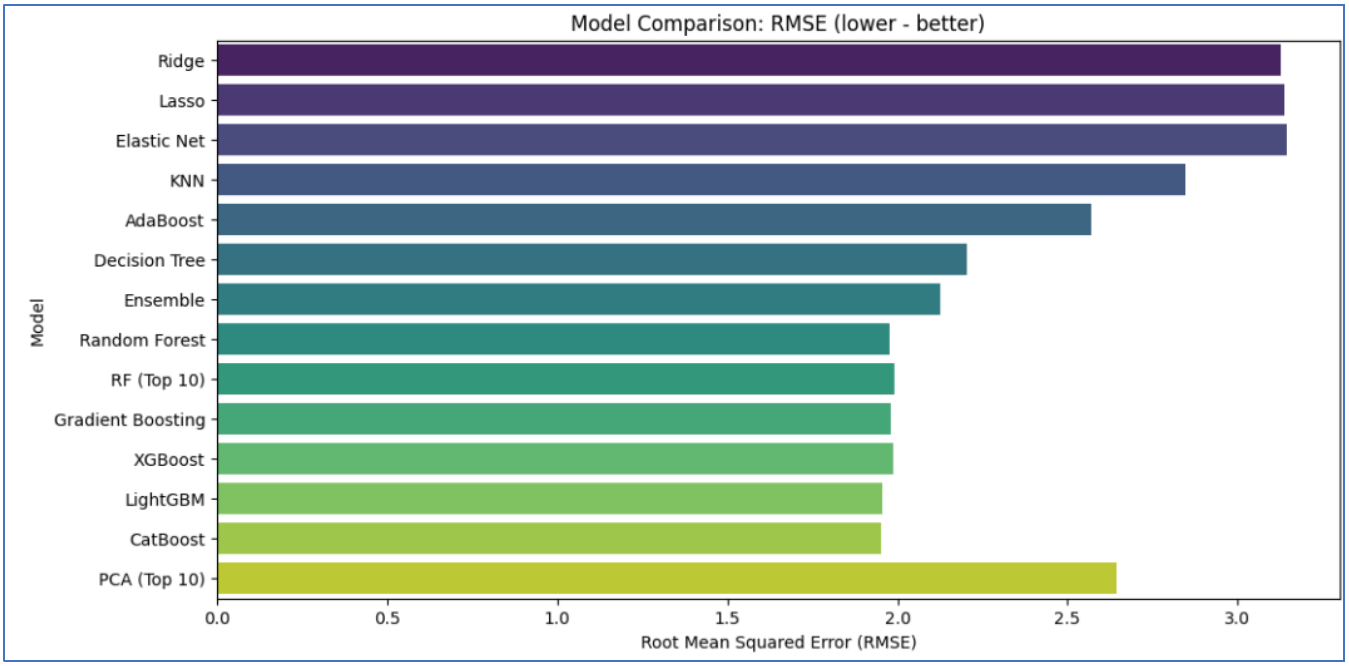
- **RMSE**: Measures the average magnitude of error (*lower is better*).
- **R^2 Score**: Explains the proportion of variance in the dependent variable captured by the model (*higher is better*).
- **MAE**: Measures the average absolute error (*lower is better*).

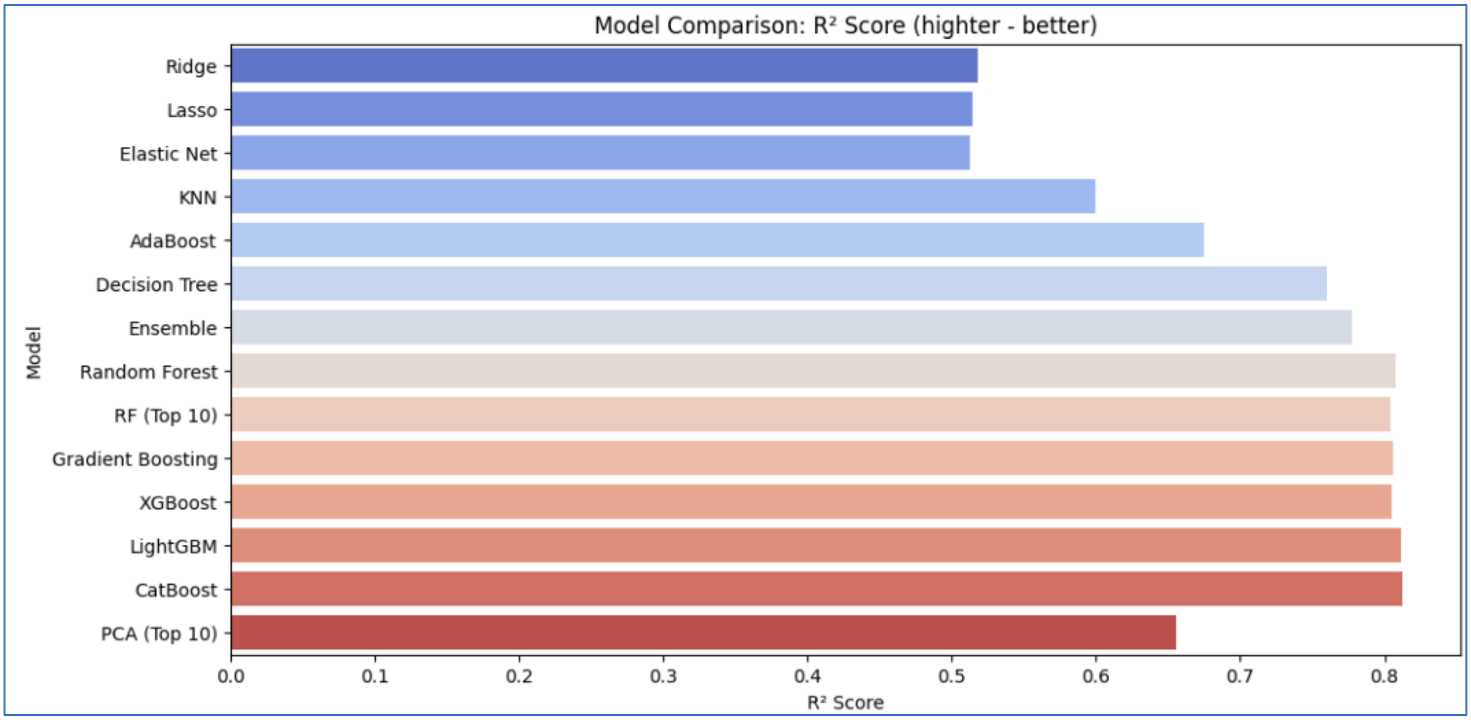
The different model performance displayed in the table below (*with **the best** and **the worst***):

Model	RMSE	R^2 Score	MAE
Ridge Regression	3.125747	0.518158	2.337167
Lasso Regression	3.137679	0.514473	2.309979
Elastic Net	3.143643	0.512625	2.308342
KNN	2.848055	0.599969	2.074419
Ada Boost	2.568481	0.674651	2.087669
Decision Tree	2.204506	0.760327	1.529299
Ensemble (RF + GB + LR)	2.126172	0.777057	1.548371
Random Forest (All features)	1.975257	0.807583	1.380623
Random Forest (Feature selection: 10)	1.992670	0.804175	1.395387
Gradient Boosting	1.982148	0.806238	1.402819
XGBoost	1.988081	0.805076	1.401522
Light GBM	1.956874	0.811148	1.375892
CatBoost	**1.951922**	**0.812102**	**1.370273**
PCA - top 10 components	2.643682	0.655321	1.854603

From the table above, we could see that best performing model was **CatBoost** [30] with hyperparameters: *iterations=300*, *learning_rate=0.09*, *depth=9*. This model outperformed other models by effectively capturing complex relationships in the data without overfitting.

Also **LightGBM** [28]**, Random Forest**, **XGBoost** [31] and **Gradient Boosting** [32] showed very good results (*very close to CatBoost*), while linear models (*Ridge, Lasso, and Elastic Net*) shows worse result, indicating that linear relationships were insufficient to capture the complexity of the dataset. Interestingly, KNN also performed not as good as expected (*maybe because of high dimensionality of data*). Same information available below in graphs.





The evaluation shows that advanced ensemble models, particularly **CatBoost** and **LightGBM**, are the most effective for predicting movie success. These models benefit from their ability to capture complex feature interactions and handle diverse feature types. Future improvements could explore hyperparameter tuning and additional feature engineering to further enhance performance.

5. Conclusions and Future Work

This project aimed to get the complete as possible Dataset as well as predict movie revenue (*box office*) success using machine learning, with focus on regression of revenue numbers as the primary metric [15]. By creating a comprehensive, enriched dataset and employing advanced machine learning techniques, it was possible to demonstrate that ensemble models, particularly **CatBoost** and **LightGBM**, are highly effective for this task. Feature engineering, such as person (*actor, producer, writer and director*) scoring and sentiment analysis, even further enhanced our model performance.

Despite these achievements, the project still faced challenges, including missing complete data for older movies and the missing some extra details, like marketing budgets, which are known to influence box office performance. Addressing these limitations in future analysis (*some of them available on websites [8] [12]*), might further improve predictive accuracy.

Looking ahead, this work can be extended by incorporating additional features such as social media trends [20], marketing budgets, and regional preferences [14]. Moreover, exploring hybrid models combining machine learning and deep learning techniques [19] [23] could significantly improve final performance of predicting movie success. This project not only provides actionable insights for the entertainment industry but also serves as a foundation for future research in movie analytics.

All calculations and final dataset available in GitHub [16].

6. References

[1] Troy Segal (2022) – "What You Need to Know About Investing in Movies". Published online at www.investopedia.com. Retrieved from: 'https://www.investopedia.com/financial-edge/0512/how-to-invest-in-movies.aspx' [Online Resource]

[2] Vr, Nithin & Pranav, M & Babu, PB & Lijiya, A.. (2014). Predicting Movie Success Based on IMDB Data. International Journal of Business Intelligents. 003. 34-36. 10.20894/IJBI.105.003.002.004.

[3] Pradeep, Kavya & TintuRosmin, C & Durom, Sherly & Anisha, G. (2020). Decision Tree Algorithms for Accurate Prediction of Movie Rating. 853-858. 10.1109/ICCMC48092.2020.ICCMC-000158.

[4] IMDB (2024) – "Most Popular Movies". Published online at www.imdb.com. Retrieved from: 'https://www.imdb.com/chart/moviemeter/' [Online Resource]

[5] Samruddhi Mhatre (2020) – "IMDB Movies Analysis". [Data set]. Published online at kaggle.com. Retrieved from: 'https://www.kaggle.com/datasets/samruddhim/imdb-movies-analysis/data' [Online Resource]

[6] Alan Vourc'h (2024) – "The Ultimate 1Million Movies Dataset (TMDB + IMDb)". [Data set]. Published online at kaggle.com. Retrieved from:

‘https://www.kaggle.com/datasets/alanvourch/tmdb-movies-daily-updates’ [Online Resource]

[7] beridzeg45 (2024) – “IMDB Movies - Up to May 2024”. [Data set]. Published online at kaggle.com. Retrieved from: ‘https://www.kaggle.com/datasets/beridzeg45/all-movies-on-imdb’ [Online Resource]

[8] IMDB PRO (2024) – “Box Office Movies”. Published online at www.pro.imdb.com. Retrieved from: ‘ https://pro.imdb.com/box_office’ [Online Resource]

[9] The Movie Database (2024) - "Movie details API". Published online at developer.themoviedb.org. Retrieved from: ‘https://developer.themoviedb.org/reference/movie-details' [Online Resource]

[10] Federal Reserve Bank of Minneapolis (2024) – “Consumer Price Index, 1913-”. Published online at www.minneapolisfed.org. Retrieved from: ’https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator/consumer-price-index-1913-’ [Online Resource]

[11] The Numbers (2024) – “Box Office Records”. Published online at www.the-numbers.com. Retrieved from: ’https://www.the-numbers.com/box-office-records/’ [Online Resource]

[12] Box Office Mojo (2024) – “Worldwide Box Office”. Published online at www.boxofficemojo.com. Retrieved from: ‘https://www.boxofficemojo.com/year/world/?ref_=bo_nb_hm_tab’ [Online Resource]

[13] Leonard Richardson (2024) – “Beautiful Soup Documentation”. Published online at www.crummy.com. Retrieved from: ‘https://www.crummy.com/software/BeautifulSoup/’ [Online Resource]

[14] Nahid Quader and Md. Osman Gani -” A Machine Learning Approach to Predict Movie Box-office,” Information Technology (ICCIT), December 2017.

[15] Vikranth Udandarao, Pratyush Gupta (2024) – “Movie Revenue Prediction using Machine Learning Models”. 10.48550/arXiv.2405.11651

[16] Dmitry Elsakov (2024) – “Movie-Success-ML-“. Published online at github.com. Retrieved from: ‘https://github.com/delsakov/Movie-Success-ML-‘ [Online Resource]

[17] Sloria (2024) – “TextBlob: Sentiment Analysis”. Published online at textblob.readthedocs.io. Retrieved from: ‘https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis‘ [Online Resource]

[18] Paul, Chiranjib & Das, Prabir. (2022). Predicting movie revenue before committing significant investments. Journal of Media Economics. 34. 1-28. 10.1080/08997764.2022.2066108.

[19] D.A., Olubukola & O.M., Stephen & A.K., Funmilayo & Omotunde, Ayokunle & A., Oyebola & Oduroye, Ayorinde & Ajayi, Wumi & Yaw, Mensah. (2021). –“Movie Success Prediction Using Data Mining. British Journal of Computer, Networking and Information Technology”. 4. 22-30. 10.52589/BJCNIT-CQOCIREC.

[20] Beyza Çizmeci and Sule Gündüz Öğüdücü, ”Predicting IMDb Rating of Pre-release Movies with Factorization Machines Using Social Media,” IEEE 3rd International Conference on Computer Science and Engineering, 2018.

[21] Seung Jun Choi (2024) – "Lab 5.1: PCA" Published online at utexas.instructure.com. Retrieved from: 'https://utexas.instructure.com/courses/1404604/files/79040549?wrap=1' [Online Resource]

[22] Seung Jun Choi (2024) – "Lab Sessions 3" Published online at utexas.instructure.com. Retrieved from: 'https://utexas.instructure.com/courses/1404604/pages/lab-sessions-3?module_item_id=14044652' [Online Resource]

[23] R. Sharda and D. Delen (2006) – "Predicting box-office success of motion pictures with neural networks," Expert Systems with Applications, vol. 30 , no. 2, p. 243–254, 2006.

[24] scikit-learn (2024) – "Dimensionality Reduction with Neighborhood Components Analysis". Published online at scikit-learn.org. Retrieved from: 'https://scikit-learn.org/stable/auto_examples/neighbors/plot_nca_dim_reduction.html#sphx-glr-auto-examples-neighbors-plot-nca-dim-reduction-py' [Online Resource]

[25] N. Darapaneni *et al*., "Movie Success Prediction Using ML," *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, USA, 2020, pp. 0869-0874, doi: 10.1109/UEMCON51285.2020.9298145.

[26] Elizabeth Antony, Nimmy Francis (2022) – "Movie Box Office Success Prediction using Machine Learning". Proceedings of the National Conference on Emerging Computer Applications (NCECA)-2022 Vol.4, Issue.1, p. 621-624

[27] LightGBM (2024) – "Training". Published online at lightgbm.readthedocs.io. Retrieved from: 'https://lightgbm.readthedocs.io/en/stable/Python-Intro.html#training' [Online Resource]

[28] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157.

[29] CatBoost (2024) – "Training". Published online at catboost.ai. Retrieved from: 'https://catboost.ai/docs/en/features/training' [Online Resource]

[30] Anna Veronika Dorogush, Vasily Ershov, Andrey Gulin (2017) – "CatBoost: gradient boosting with categorical features support". Published online at learningsys.org. Retrieved from: 'https://learningsys.org/nips17/assets/papers/paper_11.pdf' [Online Resource]

[31] XGBoost (2024) – "Training". Published online at xgboost.readthedocs.io. Retrieved from: 'https://xgboost.readthedocs.io/en/stable/python/python_intro.html#training' [Online Resource]

[32] scikit-learn (2024) – "GradientBoostingRegressor". Published online at scikit-learn.org. Retrieved from: 'https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html' [Online Resource]