Predicting Movie Success Before Release: A Machine Learning Approach to IMDb Ratings

1st Ahmet Aksünger
Technische Hochschule Augsburg
Augsburg, Germany
ahmet.aksuenger@tha.de - 2214186

***Abstract*—Predicting a movie's success before its release is a really challenging but valuable task. In this study, we do exploratory data analysis (EDA) and apply different machine learning (ML) techniques to predict IMDb scores by categorizing them into score ranges. There are many factors affecting the IMDb scores including budget, cast, director, genre, etc., which can be collected for interpretation before the movie is released. However, there are also some data which cannot be known before release, such as viewer count, audience reception, revenue, popularity, and many more. Hence, we have a limited amount of features to interpret, which makes this study even more challenging. Therefore, we are applying detailed feature engineering methods to enhance our limited features to get better results.**

In this study, many machine learning models are tested including XGBoost, CatBoost, Random Forest, Bagging etc. Also an AutoML framework (AutoGluon) is tested to automate model selection and optimization.

The results show that AutoML based models perform better than manually tuned models, with an accuracy of %76.45 after extensive future engineering.

***Index Terms*—imdb, prediction, movie, machine learning, classification, automl**

I. Introduction

The movie field improves day by day, introducing more and more movies each year as well as new actors/actresses and directors. However, are the qualities of the movies also improving? Are movies with some certain genders have the possibility to be more successful? How about movies with certain keywords? People are investing more money as the time passes, they try new technologies, they hire experienced and well-known actors/actresses, they hire the most popular directors but is having more budget on a movie enough to make it successful? Does it help with the IMDb? How about hiring the most popular actors and actresses such as Leonardo DiCaprio, Brad Pitt, Scarlett Johannson? Would the runtime of my movie affect its IMDb rating? Lastly, the most important question: is it possible to predict the success of a movie before it is released?

This study aims to apply an extensive data analysis on movies to answer all of the questions above. The data analysis helps us to build the background for our main purpose which is to develop a predictive model for IMDb scores. The Movies Dataset is used to achieve all of these. The dataset initially includes 45466 movies. However, after some filtering we are only using the 4546 movies to build and evaluate our predictive model. The dataset includes many features about the movies which can be categorized to 3 main subjects: the meta data about the movies, the credits and the keywords. As expected, not all of these features are helpful for us as we are trying to predict the IMDb score range of a movie before it is released. Hence, features such as revenue, vote count and popularity are not in our interest.

To identify the best predictive model, we experimented with several algorithms including XGBoost, CatBoost, Random Forest, AutoGluon. We aim to achieve the highest possible classification accuracy and provide a meaningful prediction for movie success. This research contributes both to the field of machine learning and film making. It demonstrates the effectiveness of AutoML frameworks and it helps the filmmakers with the conclusions of the detailed data analysis also with the predictive model.

II. Related Work

In literature, there are some related works about predicting the movie success. Although the word "succes"s for movies may be defined differently in each study, the actions taken to get to the conslusion are similar. In 2019, a study focusing on prediction movie success was published [1]. They used data mining to gather data about the movies which could be used later to build a model. They scraped information from IMDb, Rotten Tomato and YouTube. There are many features they are using to develop a predictive model, some of the features they have are directors, budget, YouTube trailer views and so on. However, they also used the total gross and critic reviews to build their model, which means the study was aimed to predict a movie success after some time it is released. As total gross and critic reviews would not be possible to be known before the release of a movie. They defined the success by categorizing the IMDb score as Poor, Average, Good and Very Good.

Another study in the field was about comparing the machine learning approaches for movie success prediction [5]. In this study they used the IMDb Dataset consisting of 5000 movies. They also used the total earning of a movie (gross) as a feature. Additionally they used the profit percentage of a movie to support their model more. Considering these two features mentioned, the aim in this study is also to predict a movie success after release. Differently, they categorized the score range to two categories: Hit and Flop. They trained different

machine learning algorithms and compared the results. They achieved a classification accuracy score of %85.4 using the Gaussian Naive Bayes (GNB).

Success is not always defined as the IMDb score of a movie. The cited study aims to predict the financial success of a movie [7]. They categorized the financial performance of a movie to 9 classes. They then used Neural Networks to train a classification model which achieved an accuracy score of %75.2.

Some studies priotirized the power of social media [6] [2]. This study published in 2023 uses the data about the movie from the Facebook to predict both IMDb score and movie success [8]. It scrapes some useful information from Facebook (FB) such as Director FB likes, Actor FB likes, Actress FB likes etc. It then tests several different machine learning algorithms to build the model with the highest accuracy possible.

III. Methodology

The main workflow of the study is shown in Figure 1.

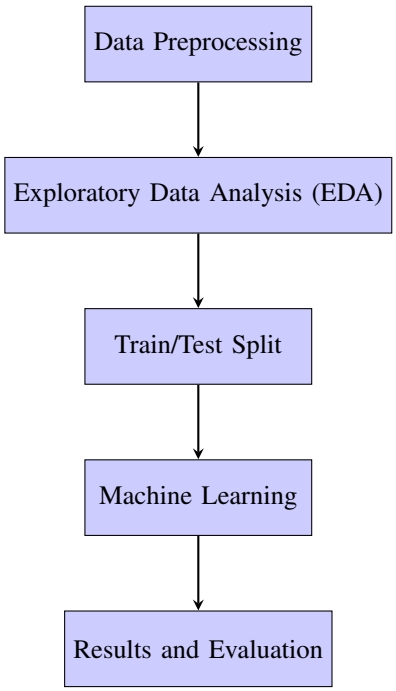

Fig. 1. Workflow of the Study

A. Data Preprocessing

As mentioned in the Introduction part, we are using The Movies dataset which consists of 45466 movies. Some of the movies in the dataset have null values for some important features we are planning to use. These movies have been removed from the dataset. Another important task would be to have finalized and fair IMDb scores which requires a certain amount of users votes. For that case, movies that have less than 80 votes have been removed.

After removing the unnecessary features and revising some of the features, the dataset feature information overall is shown in Table I.

Column Name	Description
`title`	Official title of the movie
`top_1_genre`	Most relevant genre of the movie
`top_1_company`	Leading production company
`top_1_country`	Country where the movie was produced
`vote_average`	IMDb score
`budget`	Production budget in dollars
`runtime`	Duration of the movie in minutes
`original_language`	Original language of the movie
`release_year`	Year the movie was released
`release_month`	Month the movie was released
`top_1_actor`	Lead actor in the movie
`top_1_director`	Director of the movie
`top_1_keyword`	Most important keyword

TABLE I
Description of Movie Columns

After the data preprocessing process, the size of the dataset drops to 4546 movies.

As the aim of this study is to predict the success of a movie by IMDb score, a categorization was applied to the `vote_average` column as shown in Table II.

Interval	Category
$0 \leq x < 4$	Flop
$4 \leq x < 6$	Average
$6 \leq x < 8$	Hit
$8 \leq x < 10$	Super Hit

TABLE II
Movie Rating Categories

B. Exploratory Data Analysis (EDA)

A detailed exploratory data analysis has been made to get to know the dataset and features better. The information gathered in this step is used to train a high-accuracy classification model. The statistics gathered here may be useful for the movie field and future movie related researches. The EDA is focused on interpreting a feature combined with the IMDb score.

1) IMDb Score Analysis: The distribution of IMDb score follows the normal distribution as shown in Figure 2. However, the curve is skewed slightly left, meaning that movies with really low IMDb scores (categorized as Flop) are more than movies with really high IMDb scores (categorized as Super Hit). Overall average IMDb score is 6.38.

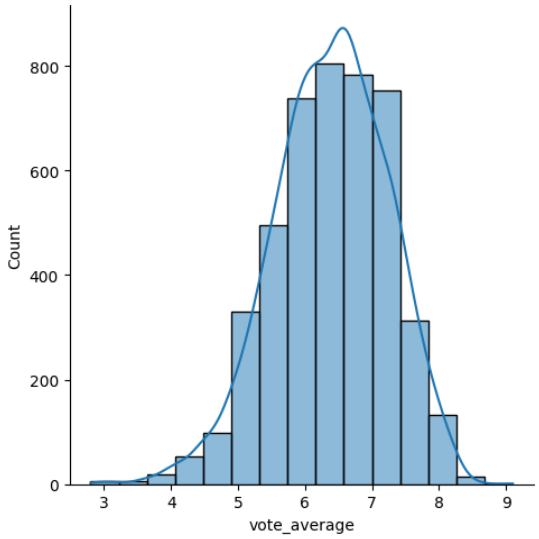

Fig. 2. Distribution of IMDb score

2) Runtime Analysis: Runtime is one of the first feature that comes to mind when talking about movies. In our dataset, average runtime is 109 minutes. As shown in Figure 3, we observe a correlation coefficient of 0.32 between runtime and IMDb score. This suggests us a weak positive correlation. Therefore, in general longer movies tend to have slightly higher IMDb score. It is worth to note that this relation is not strong.

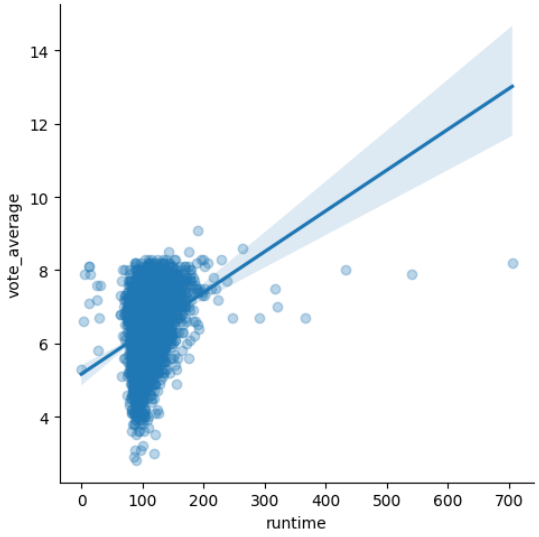

Fig. 3. Correlation between runtime and IMDb score

3) Budget Analysis: At first glance, you might think that budget has a lot of influence on the IMDb score of a movie. However, budget analysis shows that there is no direct relation between an IMDb score and budget. The correlation coefficient is observed as -0.08, which proves that there is no linear relation between these two features. Investing more money on a movie does not guarantee a better IMDb scores.

We also observe that the distribution of the budget does not match the normal distribution, shown in Figure 4. Most of the movies are low to mid budget movies.

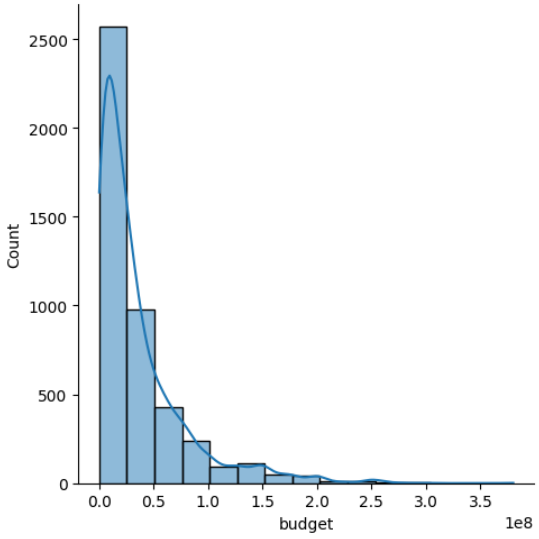

Fig. 4. Distribution of budget

4) Time Analysis: The distribution of the years of release of the movies is shown in Figure 5. Most of the movies are released recently, supporting the theory of movie field growing rapidly.

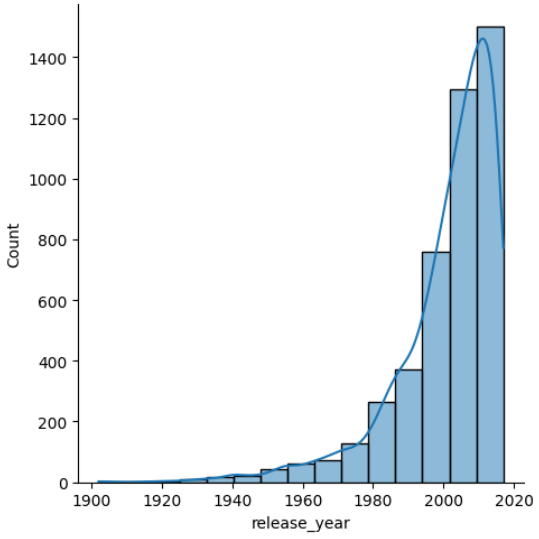

Fig. 5. Distribution of release years

Meanwhile, the correlation analysis of the release years shows us that there is a weak negative correlation, with a coefficient of -0.3. This states us that IMDb scores are slightly decreasing over time. The graph can be found in Figure 6.

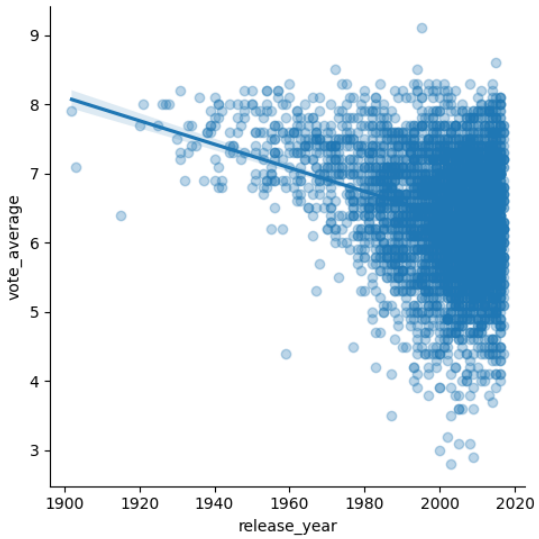

Fig. 6. Correlation between release year and IMDb score

When movies are categorized to 20th and 21st centuries, a significant difference is observed. After applying the t-test, it is proven that 20th Century movies are significantly better. As shown in Table III, 21st century movies have approximately 0.4 lower average IMDb score than 20th century movies.

Century	**Avg**	**Count**
20th Century	6.615696	1631
21st Century	6.252762	2915

TABLE III
AVERAGE RATINGS AND MOVIE COUNT BY CENTURY

5) Genre Analysis: The bar chart in Figure 7 presents the average IMDb scores across different genres. The highest-rated genres are Western and Documentary, both over 7. This indicates us that these types of movies tend to receive higher appreciation from the audience. Drama (6.78), History (6.73) and War (6.68) also have relatively high ratings, followed closely by Animation (6.61) and Crime (6.60).

On the other hand, Horror (5.93) and TV Movie (5.50) have the lowest average IMDb ratings. These genres are not appreciated much by the audience.

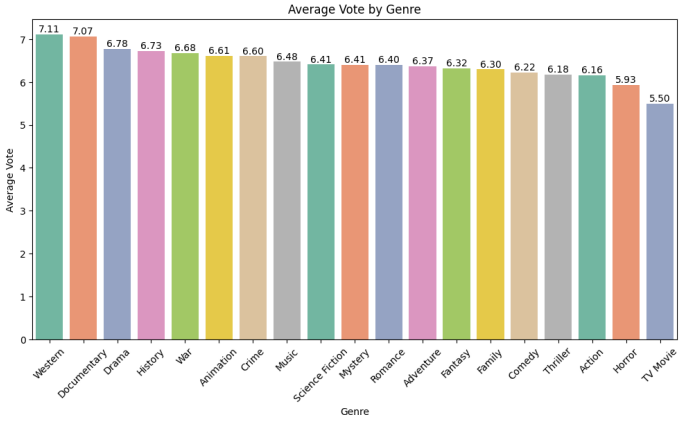

Fig. 7. Average IMDb Scores by Genre

6) Actor Analysis: The analysis of actors' IMDb scores is based on the number of movies they have as a lead actor. We observed that most of the actors appeared in limited amount of movies. The average number of movies for an actor is 2.66.

Since the distribution of movie counts is not normal, a 90th percentile threshold (6 movies) was applied. This helps us to compare the IMDb ratings of actors in a more meaningful and stabil way.

Table IV shows the top 10 actors with the highest average IMDb scores, after applying the threshold mentioned above.

TABLE IV
TOP 10 HIGHEST-RATED ACTORS BY IMDB SCORE

top_1_actor	movie_count	movie_score
Charlie Chaplin	6	8.02
James Stewart	6	7.87
Cary Grant	6	7.48
Edward Norton	6	7.33
Daniel Radcliffe	10	7.31
Paul Newman	7	7.30
Aamir Khan	7	7.26
Daniel Day-Lewis	8	7.19
Leonardo DiCaprio	15	7.19
Al Pacino	16	7.16

Table V shows the top 10 lowest rated actors.

TABLE V
BOTTOM 10 LOWEST-RATED ACTORS BY IMDB SCORE

top_1_actor	movie_count	movie_score
Jesse Bradford	8	4.90
Christopher Lambert	6	5.22
Steven Seagal	11	5.35
Jean-Claude Van Damme	21	5.40
Vince Vaughn	8	5.49
Jessica Alba	6	5.53
Tim Allen	8	5.58
Eddie Murphy	25	5.58
Jennifer Lopez	10	5.60
Renée Zellweger	9	5.63

7) Director Analysis: Similar to the actor analyis, we took the 90th percentile (5 movies) as a threshold. This focuses on

directors who directed multiple movies and are consistently getting high IMDb ratings.

Table VI shows the top 10 highest rated directors. As Charlie Chaplin both directed and played in his movies, he is again in the lead with an impressive IMDb rating of 8.02. This is followed by Sergio Leone (7.98) and Hayao Miyazaki.

TABLE VI
TOP 10 HIGHEST-RATED DIRECTORS BY IMDB SCORE

top_1_director	movie_count	movie_score
Charlie Chaplin	6	8.02
Sergio Leone	5	7.98
Hayao Miyazaki	7	7.89
Billy Wilder	10	7.77
Christopher Nolan	10	7.72
Quentin Tarantino	9	7.68
Stanley Kubrick	12	7.62
David Lynch	9	7.42
John Ford	5	7.42
Jean-Luc Godard	5	7.42

Table VII showcases the top 10 lowest rated directors.

TABLE VII
BOTTOM 10 LOWEST-RATED DIRECTORS BY IMDB SCORE

top_1_director	movie_count	movie_score
Jason Friedberg	6	3.88
Brian Levant	6	5.15
Jonathan Frakes	11	5.15
Andrzej Bartkowiak	5	5.20
Eli Roth	5	5.34
Brian Robbins	5	5.38
Rob Zombie	5	5.40
Raja Gosnell	8	5.44
John Moore	6	5.45
David R. Ellis	5	5.48

8) Keyword Analysis: In the Data Preprocessing part, the most important keyword was extracted for each movie. In this section, we are evaluating the impacts of keywords. There are 1282 unique keywords for 4546 movies. Almost each movie has its own unique keyword. In order to find the most frequent and impactful keywords, we again applied the 90th percentile technique, which lead us to a threshold of 8 movies.

As we observe from Table VIII, the keyword "individual" has the best IMDb score with a value of 7.45. This keyword might include the movies which focus mainly on character development. This seems to be well appreciated by the audience. The second most high rated keyword is "Germany" (7.27), which seems to be coming from historical war related movies. The third most impactful keyword is "Child Abuse", showing that audience likes movies that touch the important and dramatic social themes.

TABLE VIII
TOP 10 HIGHEST-RATED KEYWORDS BY IMDB SCORE

top_1_keyword	movie_count	movie_score
Individual	21	7.45
Germany	11	7.27
Child Abuse	10	7.18
Transporter	7	7.11
Life and Death	9	7.07
Rebel	8	7.05
Shyness	7	7.00
Civil War	8	6.99
San Francisco	16	6.98
Judge	8	6.96

C. Feature Engineering

Most of the features mentioned in the Exploratory Data Analysis (EDA) are used in building the predictive model. However, they are not enough to train a high accuracy classification. Also, some of the features mentioned above are non-numeric features, which should be converted to numeric values.

1) Categorical Features: Genre, company, country, language, actor, director and keyword are not numerical values. They cannot be used in a model directly. Hence, they need to be presented with a numerical value using different techniques. After trying different methods with different features, the best results obtained were from one-hot encoding and target encoding.

a) One-Hot Encoding: A method that creates binary columns for each category, representing the presence of that categorical value with 1 and else with 0.

b) Target Encoding: A method that replaces each category with the mean target value. For example, for each director, replacing its name with its average IMDb rating.

Target encoding gave the best results for most of the categorical features. Genre, company, country, actor and director features were encoded using target encoding. Using one-hot encoding with the language column worked better, as there are not many unique language categories.

As there are too many unique keywords (1282), both target encoding and one-hot encoding did not increase the accuracy in our model. So a different feature engineering approach had to be applied for the keyword feature.

2) Keyword Sentiment Analysis: We still needed a way to use the keyword information, as it could increase the accuracy of our model if used wisely. For that, we used the advantages of sentiment analysis.

In order to assign a sentiment score for each keyword, we used an open-source rule-based model called VADER [4]. VADER is a rule-based model that takes a word or group of words and generates a sentiment score taking a continues value between -1 and 1. The closer the result is to -1, the more negative the word is. Table IX shows an example of outputs.

Keyword	Sentiment Score
Jealousy	-0.3182
Terrorist	-0.6908
Paris	0.0000
Holiday	0.4019
Diving	0.0772
Kiss	0.4215

TABLE IX
SENTIMENT SCORES OF KEYWORDS

3) Interaction Terms: To improve the accuracy score of our predictive model, we created new features by combining existing ones.

- `runtime_budget`: This feature is a multiplication of a movie's budget with its runtime. It increased the accuracy. The reason why it helps our model is not known yet. It might be about the relation of higher-budget movies with longer runtimes impacting the IMDb ratings.
- `director_year`: This feature is a multiplication of the directors average IMDb rating and the release year of the movie. Any further analysis about how this is helpful has not been made.

4) Additional Terms:

- `actor_career_length`: The actor career length is computed as the difference between the release year of a movie, and the release year of the first movie of the actor. This feature simply represents how experienced an actor is. It has helped our model to give better results.
- `is_franchise`: This feature is an indicator whether a movie is a franchise movie of an initial movie. A binary column was added, where 1 means the movie specified is franchise and 0 represents it is not. In order to determine whether a movie is franchise or not, we applied the following formula.

$$\text{is\_franchise}_i = \begin{cases} 1, & \text{if title contains "2", "3", or "Part"} \\ 0, & \text{otherwise} \end{cases}$$

- `actor_avg_gross`: In order to gain more information from the actors experiences and historical statistics, a feature that represents the total earning of an actor from its previous movies was added.
- `director_avg_gross`: The same logic was applied for directors.

D. Model Training

After selection all features, we now have a dataset with 45 columns. 30 of these columns are the one-hot encoded Language columns. In order to train a model, train and test sets are required. Training set is the set which our model is trained from. Test set is the set where we evaluate the metrics of the model such as accuracy score and F1 score.

Table X shows how the data was splitted to training and test sets.

Dataset	Count	Percentage
Training Set	3,636	80%
Test Set	910	20%

TABLE X
TRAINING AND TEST DATASET SPLIT

During the study, different models were tested to find the model that gives the highest accuracy output. Models tested are:

- **CatBoost** - CatBoost is an algorithm for gradient boosting on decision trees. It was developed by Yandex and it works well with the categorical features.
- **XGBoost** - Same as CatBoost, XGBoost is also working with gradient boosted decision trees. It handles the tabular data well.
- **Random Forest** - A commonly-used machine learning algorithm that combines the output of multiple decision trees.
- **Bagging Classifier** - An ensembling classifier that focuses to fix the overfitting problems in classification models.
- **Extra Trees Classifier** - It is really similar to Random Forest. However, it takes some extra actions in the feature selection part to add additional randomness.
- **AutoGluon** - A new open source AutoML framework that automates model selection, feature selection and hyperparameter tuning [3].

Each model was trained several times with the help of cross-validation techniques. 5 crosses were used for this. The trained models were tested on test sets to evaluate their performance. The primary metric compared was the "accuracy".

IV. RESULTS AND EVALUATION

After training all the models mentioned in Model Training part, we computed their performances on the test set. We took the "accuracy" as the most important metric. Each model was trained and evaluated 5 times with different splits of the training and test datasets, with the help of k-fold cross-validation technique. This ensures that the evaluated accuracy is not getting affected by a bad dataset split by luck.

A. Model Performances

Table XI showcases the accuracy scores of the mentioned models.

Model	Accuracy (%)
XGBoost	75.2
CatBoost	74.1
Random Forest	74.3
Bagging Classifier	74.3
Extra Trees Classifier	72.5
AutoGluon	**76.4**

TABLE XI
MODEL ACCURACY COMPARISON ON TEST SET

As observed, **AutoGluon achieved the highest accuracy (%76.4)**. This shows us how much more effective automated models are rather than manually tuned models. The second highest performed model is XGBoost, with an accuracy of %75.2 off by %1.2 which could be considered as a big gap, since the evaluation metric is accuracy.

B. Feature Importance

To understand what impacts the most on a movie's IMDb rating, a feature importance analysis was made. Figure 8 shows the the top features affecting movies success.

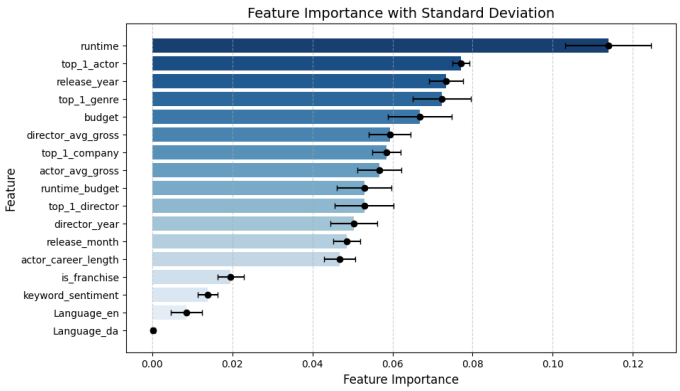

Fig. 8. Feature Importance with Standard Deviation

The most powerful factor when predicting a movies IMDb success is runtime. This is followed by top_1_actor and **release_year** with a big gap between the first element. This indicates how important the runtime is for our model's decisions.

The genres are the 4th most important features. Although there was no direct correlation between the IMDb rating and the budget, we still can see that the predictive model found a way to gather information out of it.

Additionally added features in the feature engineering part overall have a moderate impact. Looking at the **director_avg_gross**, **actor_avg_gross**, **runtime_budget** and **director_year**; we can see that they have a moderate impact on our model's decision-making process. However, **actor_career_length** and **is_franchise** did not contribute much.

The least important features are the categorical language columns. This is expected as most of the movies are in english, so it is not providing any useful information.

V. Conclusion

This study focused on the goal of predicting a movies success **before its release**. We performed different data science techniques to conclude a meaningful result.

It also showed the power of Auto Machine Learning (AutoML) frameworks achieving an accuracy of %76.4 outperforming other manually tuned models. The study also provided many different statistics about the movie field including the factors that influence a movie success the most. It also gives idea on how helpful feature engineering is, and how it can contribute to your model. Even though the impact of the feature engineered features are moderate to low, it still increased the accuracies in a way.

Overall, it is possible to predict a movies success before its release up to a certain point. There are still limitations. Specifically, about the audience amount, such as popularity and vote count.

A. Future Work

Although vote count and popularity are mainly features after the release of a movie, the popularity might be computed before release too. Data mining techniques might be used to extract data from social media platforms such as X, YouTube, Instagram and Reddit. The view count of the trailer of the movie on YouTube might contribute to the model.

The current keyword sentiment analysis is really limited. This could be enhanced, and different techniques to use the keywords column might be found.

Deep Learning models such as Neural Networks (NN) can be used to analyze the text data like overview, keywords and movie titles.

References

[1] Partha Chakraborty, Md Zahidur, and Saifur Rahman. Movie success prediction using historical and current data mining. *International Journal of Computer Applications*, 178(47):1–5, 2019.

[2] Beyza Çizmeci and Şule Gündüz Ögüdücü. Predicting imdb ratings of pre-release movies with factorization machines using social media. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pages 173–178. IEEE, 2018.

[3] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

[4] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[5] M Lakshmi, K Aditya Shastry, Aditya Sandilya, Rahul Shekhar, et al. A comparative analysis of machine learning approaches for movie success prediction. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 684–689. IEEE, 2020.

[6] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten De Rijke. Predicting imdb movie ratings using social media. In *Advances in Information Retrieval: 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings 34*, pages 503–507. Springer, 2012.

[7] Ramesh Sharda and Dursun Delen. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2):243–254, 2006.

[8] Irum Sindhu. and Faryal Shamsi. Prediction of imdb movie score movie success by using the facebook. In *2023 International Multi-disciplinary Conference in Emerging Research Trends (IMCERT)*, volume I, pages 1–5, 2023.