Movie Box-office Prediction Based on Early Critic Vote

Amin Khoeini

In the era of social media, Community-driven platforms such as Twitter and Facebook gave a voice to every user of a cultural commodity. In the 1980's, people bought the Cahiers du Cinéma magazine to see if a film is acclaimed by critics and is worth watching. Today, the new generation uses websites such as IMDB and Rotten Tomatoes to see what ratings the movie earned. Each user of these websites can rate a movie and if they decide to initiate a twitter storm about it to hype or destroy its reputation, they have all the means to do so. In this era, the fate of a movie's success is in the hands of the users of those websites.

Squid Game is a South Korean TV show that struggled to find a producer for almost 10 years. Finally, it was picked up by Netflix in 2018, and during the first week of its release, was wildly unpopular. Then, there was a huge presence of the show on Twitter and other social media platforms and by the end of the second week, Squid Game was number one show on Netflix and ended up being the most viewed TV show on Netflix. Although a movie critic might call the show below average or even call it worthless, it currently has an 8/10 vote on IMDB with more than 380,000 votes. This is just one example about how ratings on websites impact the success of movies and TV shows in our era. This intrigues my interest to dig deeper to see how much the internet and social media impacts the box office of movies today.

The goal of this project is to predict the box-office of movies based on ratings from popular websites such as IMDB, Rotten Tomatoes and Metacritics. IMDB voting is only done by users of the website who are regular movie viewers. Metacritic votes are done solely by critics and only normal users can see the rating. Rotten Tomatoes on the other hand has two rating systems; one for critics which is called Tomatometer and one for normal users which is called an audience score. The first step was to create a model using a combination of audience votes and critic votes to potentially measure how accurate this model can be to predict the box-office of a movie. However, after initial analysis and for the purpose of this project, the final model presented only uses early critic votes to predict the box-office of a movie.

1. Data

The data set for this project consists of 3312 rating sets of movies. Datasets were gathered from 4 different sources:

- **Box-Office Mojo:** which provides data about the box-office of movies. Data was manually scraped from the website using the Beautiful Soup library.
- **IMDB:** which can be found on Kaggle and consists of average votes, male votes, female votes, US votes and non-US votes accompanied by how many people voted for the movie.
- **Rotten Tomatoes:** which is also found on Kaggle. This set has both a value for critic votes and audience votes.
- **Metacritic:** which was scraped manually from the website and consists of the average of critics and audience vote, with vote counts.

1.1 Data Extraction

While IMDB and Rotten Tomatoes were gathered from the Kaggle website, the data from Metacritic and Box-office Mojo websites were manually extracted. Python's BeautifulSoup was the package that was used for this extraction. Box-office Mojo website has rather a simple html layout. Just by going to the worldwide tab in the mail page, more than 250 box-office data are available for each year. First a list of the year from 2010 to 2019 created, and used that list as an iterator, all the data for each year extracted by BeautifulSoup's "select" method.

For the Metacritic the html is more complicated and is different based on the movie rating. Also each movie has its own page and there is no page that has all the ratings, like what is shown in Mojo. So the first step was gathering the name of the all movies on the website for each year from 2010 to 2019. Doing this, a list of all the available films on the website were created with the corresponding link of the movie on the website. Then by using this list as an iterator, BeautifulSoup can access the html data of each film easily. The real challenge was finding the html class code for the rating of each movie, because the website chose a different class code for the rating based on its number. If the movie is labeled as a positive rating, the site refers to it as a ""metascore_w header_size movie positive", and it's different for the negative, mixed, and perfect positive. Same concept followed for the number of the vote. So 4 conditional commands had to be created for each iteration to find the vote and vote number for each movie, while it was not clear how each movie was labeled beforehand.

1.2 Data Wrangling

IMDB and Rotten Tomatoes dataset, both have a column called imdb_id and used as a unique identifier to merge these two without any issue. To merge these dataset to Metacritic and Mojo, movie title is the only option as a common column. Merging two dataset based on movie title could be a challenging task while titles can type differently in each site, some has a forigen name and website might use the original, some has a sequel using a number at the end, and some has a long title that the website might choose to shorten it. For solving that issue, the RecordLinkage package was used in this project. Pairs for different dataset created based on Year, Title, Alternative Title and Director using a 85% threshold, and these pairs used to merge datasets more accurately.

The final dataset had lots of unrelated columns, after merge, that needed to be dropped from the dataset before the modeling process. Mojo website also provides the two columns for domestic and international box-office of the movie. While these two columns had lots of NaN values, they can not be used for any regional prediction and those also drop from the dataset.

The voting data also has a different format for each website. While IMDB and Metacritics used 0-100 range, Rotten Tomatoes used 0-10, so some changes had to be made to make these data all unified. All the data types also had to change to numeric for the purpose of the project.

2. Method

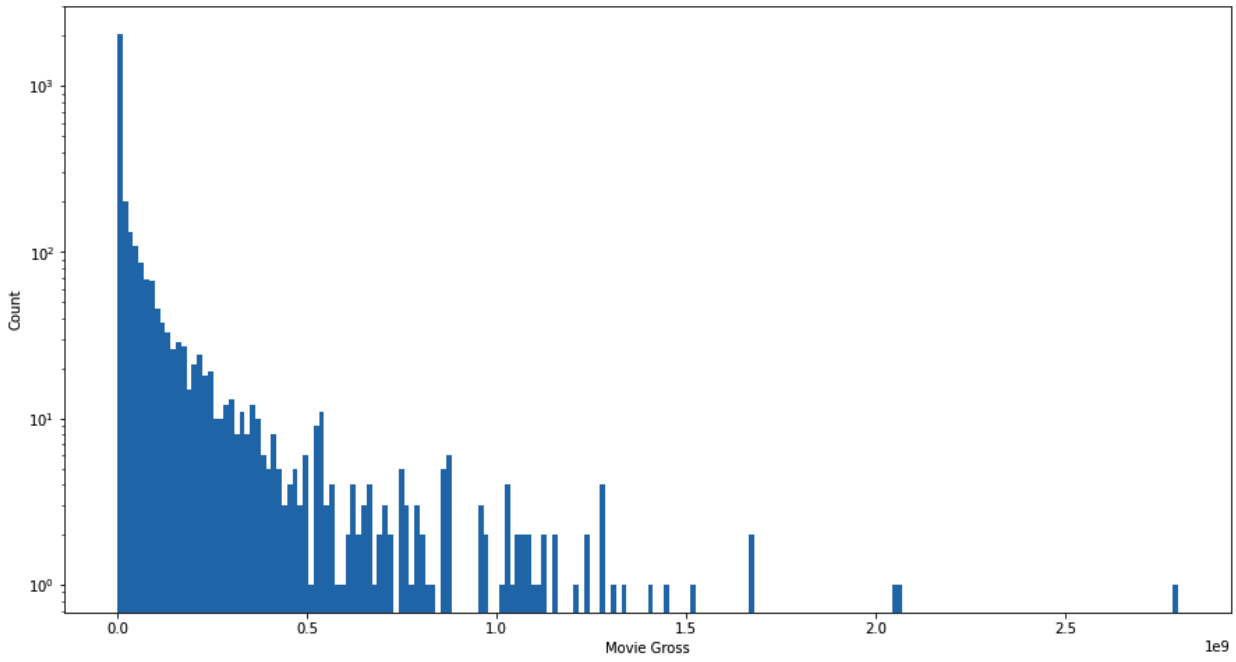
Predicting a box-office number is a supervised regression task. Features in the dataset consist of vote averages and the number of critics/audience who voted. These features will be used as a predictor. Movie box-office will be the dependent variable.

In the first step, early critics' votes and audience votes were chosen to see how efficiently they can be used for the prediction purpose. Then by using Gridesearch and performing hyperparameters tuning, different regression models' performance scores were gathered. This helped to pick the best model based on mean absolute error score.

Finally, and for the purpose of this project, the final model created only uses the critics early votes data to see if it can provide an accurate prediction for the movie box-office before the movie's release.

3. EDA

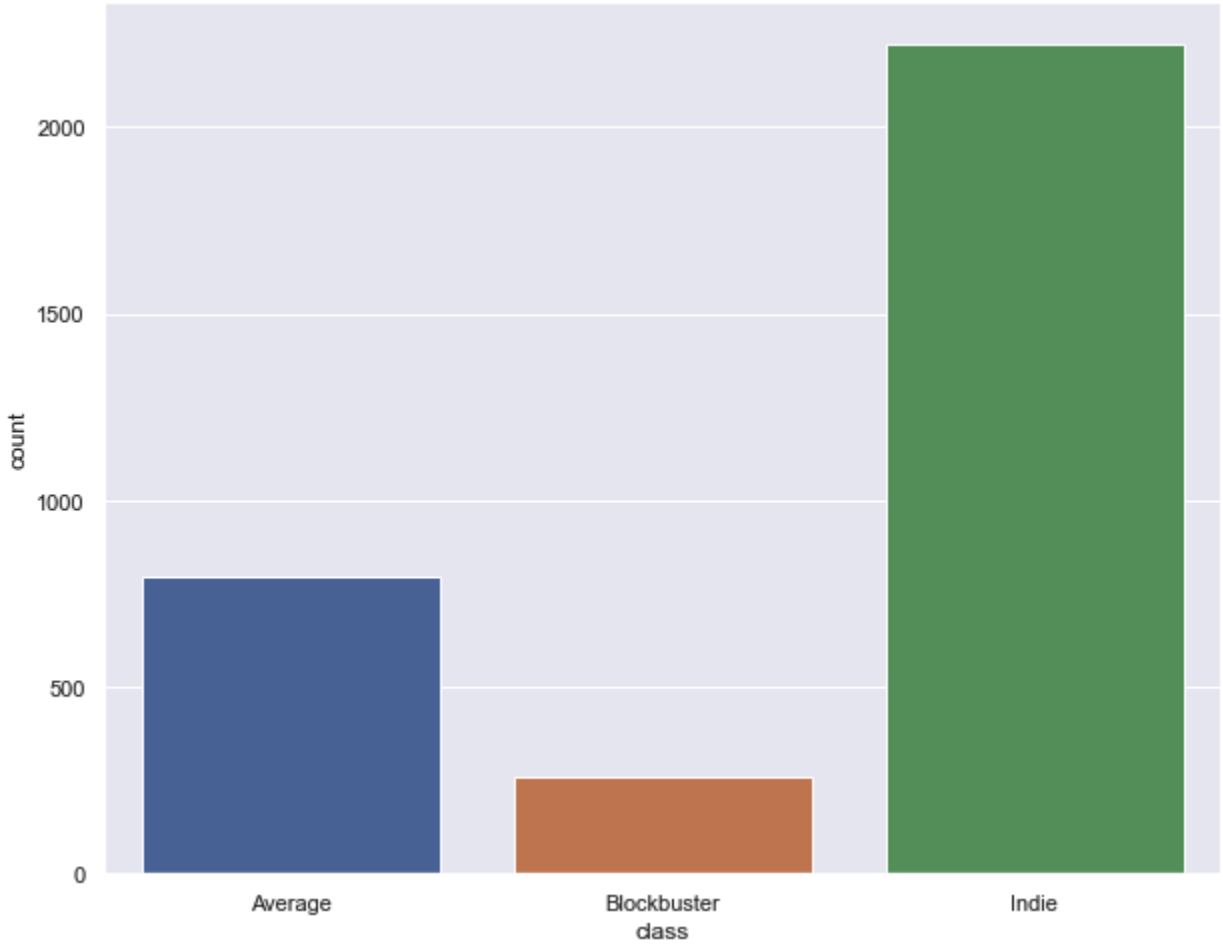
3.1 Movie box-office



The histogram shows the nature of the box-office as a long range value, between 10,000 to 3,000,000,000 and visibility is not a continuous value. This makes the prediction task, which is based on the regression model, very hard and inaccurate. The dependent variable in the regression model needs to be a continuous number in order to train the model for all possible scenarios to have an accurate prediction. Knowing this, the model will not create a very accurate box-office prediction.

Due to the dependent variable's natural value, movie box-office has been categorized in three different classes: Blockbuster, Average and Indie. This way, in addition to obtaining a prediction for box-office, the regression model can predict which class of movies will end up in that model.

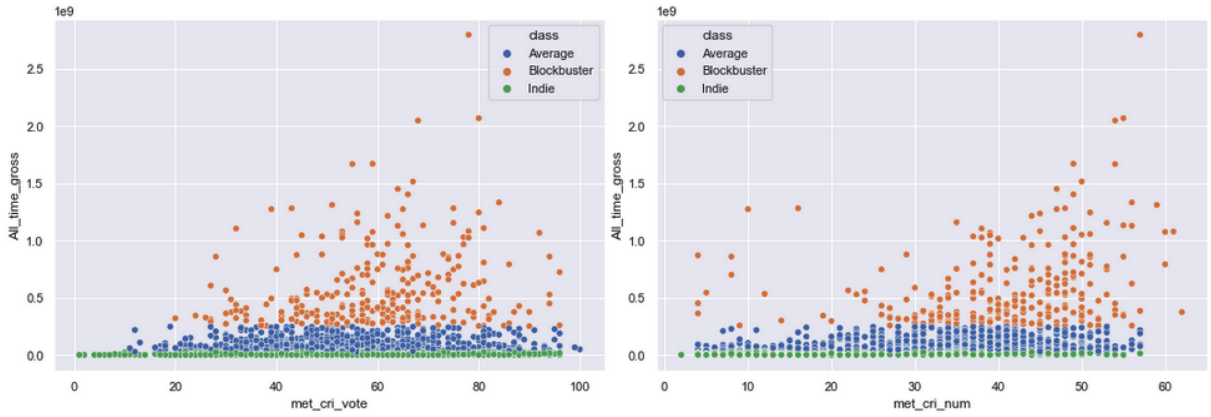
Criteria: Blockbuster = > $250 million
Average = between $25 million and $250 million
Indie = < $25 million



3.2 Predictor values: Votes Data

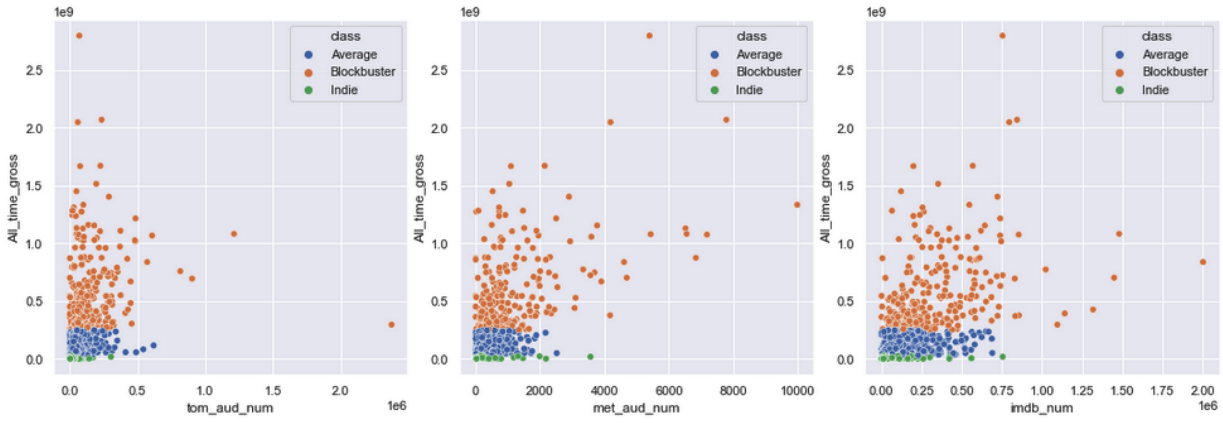
As the only predictor values, the dataset consists of vote averages and number of votes used. It contains 18 columns gathered from three different websites; Metacritic, Rotten Tomatoes and IMDB. Initial analysis can show the importance of each feature in the regression model and

can also help to pick the best feature for it. It can also help visualize the difference between critics' votes and audience votes.



For Metacritic, it is clear that the vote number shows a more linear correlation with the box-office data. All three different classes also have a balanced distribution both in vote average and vote number.

The Rotten Tomatoes critics' vote scatter plot has even more linearity in comparison to Metacritic and as a result, these columns will have the most importance in the prediction model.



Just by taking a glance at the graph containing audience number, this has more of a correlation with the box-office. The audience vote number after the release of the movie and as

the movie's popularity increases among the audience, they tend to vote higher. Therefore, it can be assumed here that the model which only consists of the critics' votes will perform worse than the model that used the combination of the audience and critics vote.

4. Algorithms & Machine Learning

Python scikit-learn was used for training my recommendation model. Four different algorithms were tested on the full votes dataset, and concluded that the Random Forest Regressor algorithm performed the best. It should be noted that this algorithm, although it is the most accurate, is also the most computationally expensive, and this should be taken into account if this were to go into production.

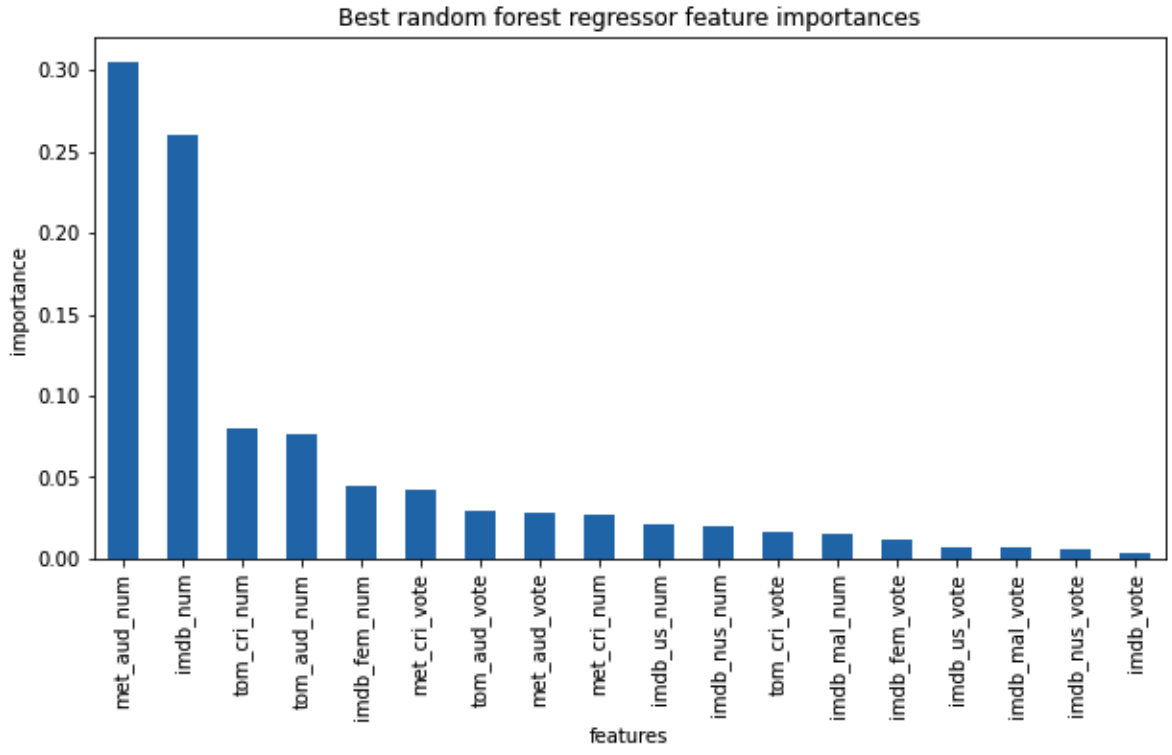
Algorithms	MAE(in million dollars)	R^2	fit_time
Linear Regression	50.61	0.49	0.0037
Random Forest Regression	40.82	0.62	1.24
Lasso Regresion	50.74	0.51	0.65
ElasticNet REgression	50.01	0.63	0.05

MAE was chosen as the accuracy metric over RMSE because it is more robust against the outlier. Box-office data, which is the dependent value, has many blockbuster movies with values over one billion dollars as the outlier and MAE performed better as a metric score in this case. The smaller the MAE, the more accurate the prediction. The Random Forest Regressor clearly

performed better in comparison to others. Using the GridSearch, the below hyperparameters were chosen for best model performance.

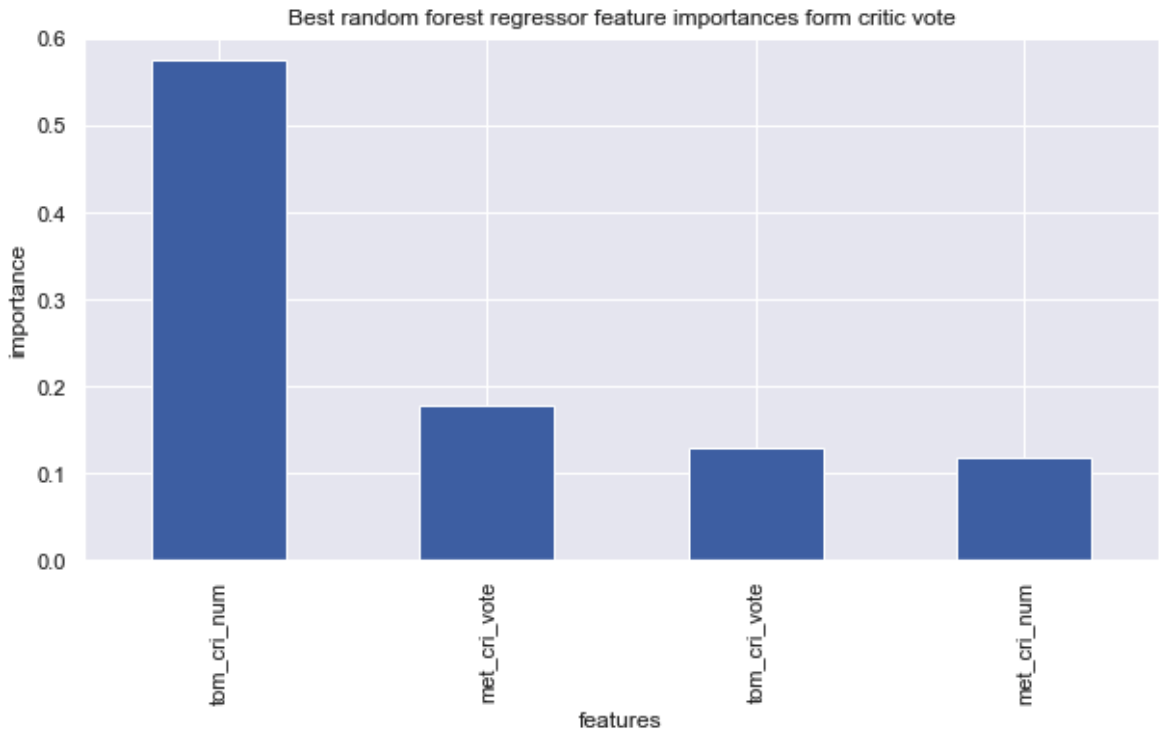
```
{'randomforestregressor__max_depth': 60,
 'randomforestregressor__min_samples_leaf': 2,
 'randomforestregressor__min_samples_split': 6,
 'randomforestregressor__n_estimators': 50,
 'standardscaler': None}
```

By looking at the featured importances, we can see that the hypothesis about the audience being the most important factor was true. The Rotten Tomatoes critics' vote number has the most weight among the critics' vote information.



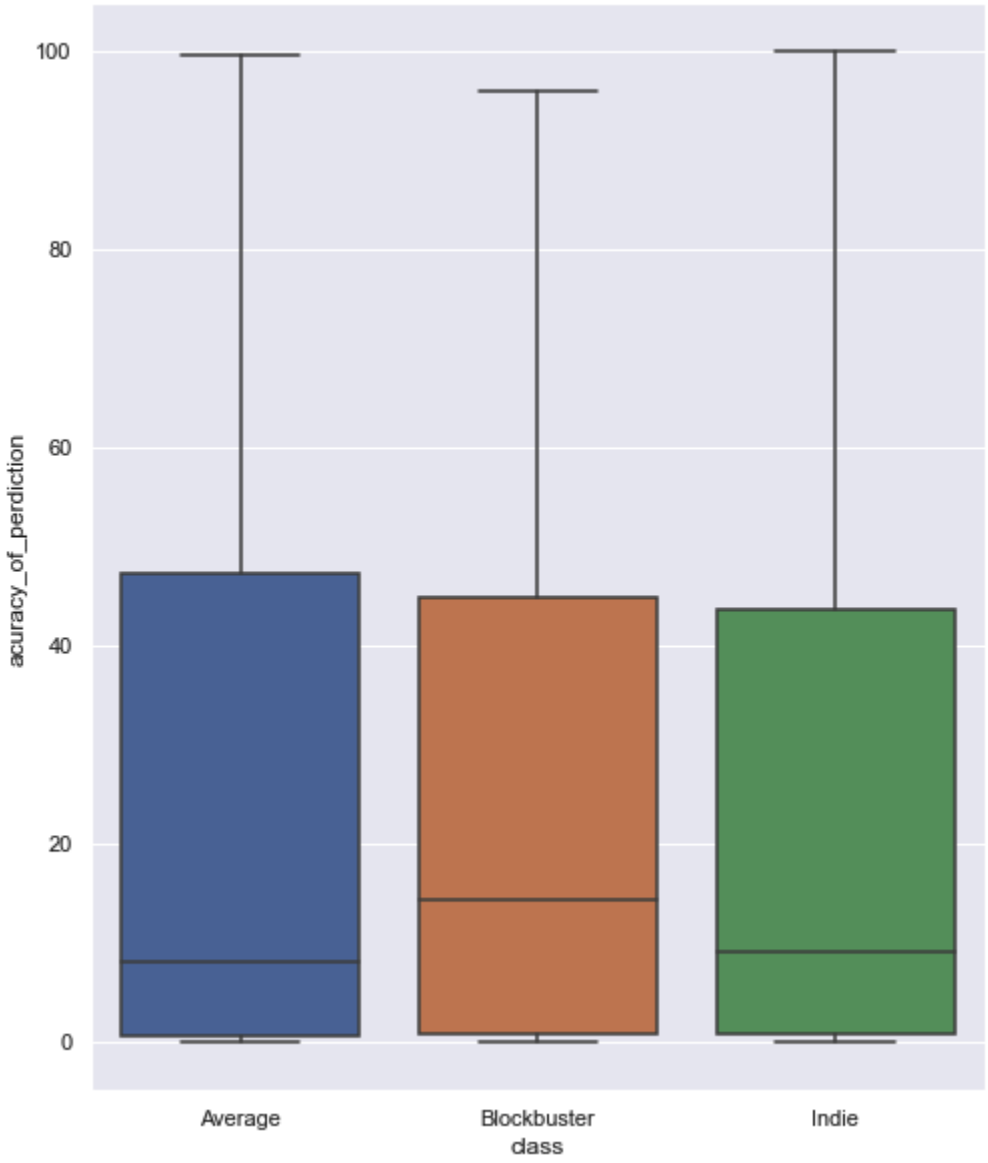
5. Modeling

The Random Forest Regressor was chosen with the hyperparameters that were calculated in the previous step as the final model. For the goal of the project, this model was trained only on critics' early vote information. Rotten Tomatoes vote number still plays the most important role in this model and has more than 60% weight of the model performance. Each movie that received more votes on this website from critics, regardless of what the actual vote was, will perform better in box-office according to this model.

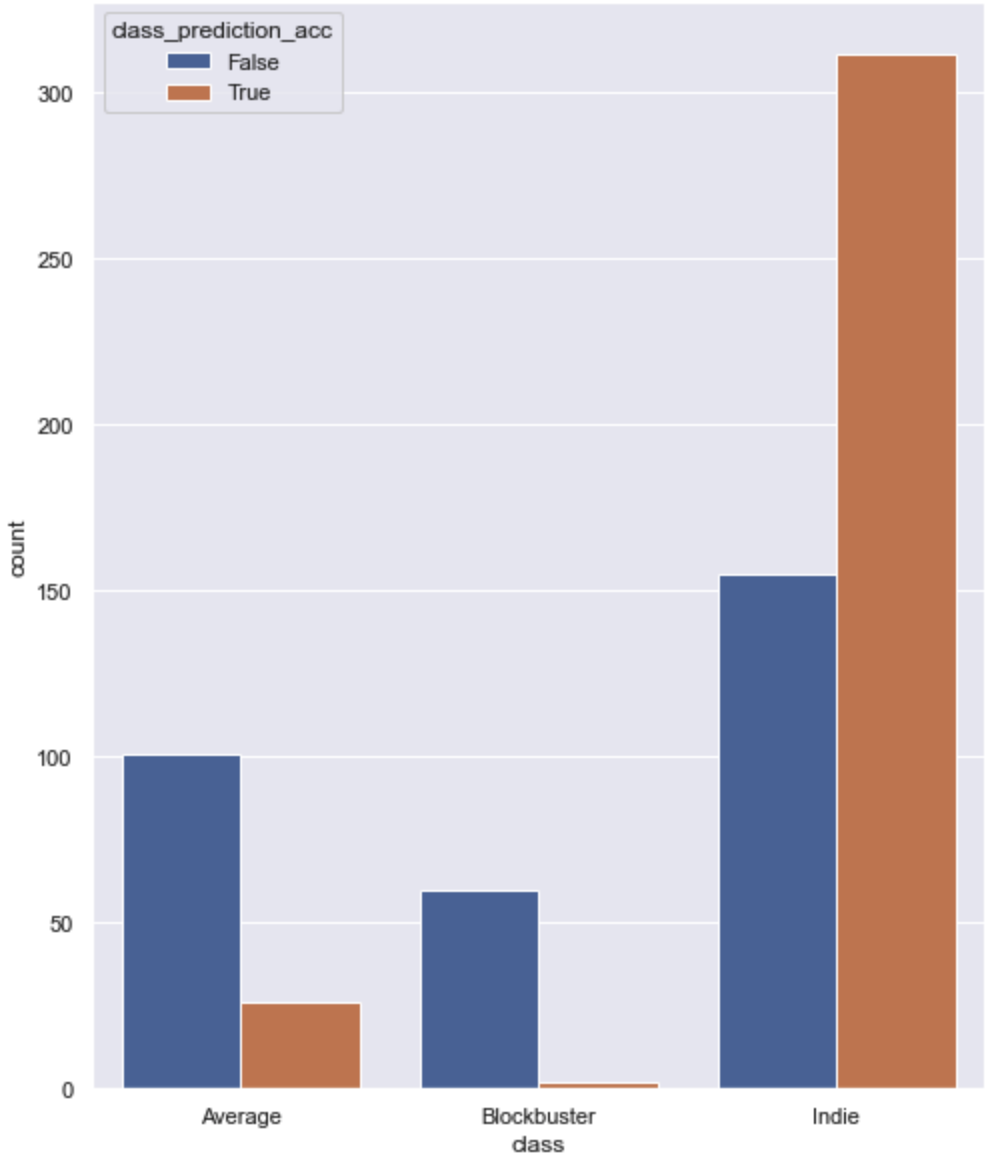


To see how this model performed, a numerical value was calculated by using the absolute error of the prediction and the actual box-office of each film and called it prediction accuracy. This value ranges from 5% to 45% and is almost the same for different classes of the film. So by

using the early vote of the critics, the model can predict the box-office of the film by 5 to 45 percent accuracy. This number is low mostly because of the nature of the box-office as a large range of discontinuous values.



This model can predict which class the movie will end up in. Especially for the Indie class, by looking at the critics early vote, it is safe to assume that the movie is going to end up in this class, or an average/Blockbuster movie.



6. Obtaining a prediction for 2021 movies

For further investigation and to see how this model works with the more recent data, three movies were picked, one from each class, to see how the early critics' vote can predict the box-office of that movie and if the model will be able to pick the right class for the movie or not.

6.1 Dune

Using the early critics' votes of this movie, my model predicts that Dune will collect $966 million at the box-office. This movie in reality ends up with only $400 million. Because the critics love this film and vote more for it, we saw higher than what the film actually made in box-office. While the MAE is large, the model predicts correctly that Dune will be a Blockbuster movie.

6.2 House of Gucci

Despite the fact that this movie is directed by Ridley Scott and has very well-known and popular stars such as Al Pacino and Lady GaGa, it only gathered $100 million in box-office. People simply didn't like it, or because of the pandemic restrictions people still hesitated to go back to movie theaters. But critics voted high for this film, and this is why my model predicted a $450 million box-office for this film. Not only is it not close to the real number, it also predicts the wrong class for the film, only because critics like it more than the audience.

6.3 Bergman Island

This is the movie that was adored by critics in festivals and on average got more than 90% positive votes. But outside of the festivals, it did not get enough attention and that is why it does not have a high number of votes even among the critics. Therefore, the model predicted that this movie would sell $2 million at the box-office and in reality, Bergman Island ended up with only $700,000. So the model correctly predicts the class and also guesses a very close number for the real box-office.

7. Future Improvements

- Predicting the box-office for a movie is a hard task. While early critics' vote showed to be a decent factor in this prediction, it is not enough for having a very accurate prediction. If more information such as how many times people google the name of the film in the first week of the movie release, or how many times people tweet about the film or even know the advertising budget of the movie can be gathered, a more accurate model for box-office prediction can be generated.

- Besides these three websites, there are others that critics submit their vote to. Although these are the most popular ones, more data would greatly help the prediction.

- The data gathered was from the years of 2010 through 2019 and only has 3000 rows. If I am able to gather data for all years available on these websites and be able to merge them with the current model, it would have much better performance.