Box Office Revenue Prediction

Project Proposal Paper

Nhan Nguyen
Computer Science
University of Colorado Boulder
Colorado United States
Nhan.N.Nguyen@colorado.edu

Melissa Pollich
Computer Science
University of Colorado Boulder
Colorado United States
Melissa.Pollich@colorado.edu

Benjamin Breyfogle
Computer Science
University of Colorado Boulder
Colorado United States
Benjamin.Breyfogle@colorado.edu

ABSTRACT

In this paper, we want to see how various elements of a movie contribute to its revenue in box office, which are divided in to external (vote, popularity…) and internal factors (budget, languages…). These factors are quantified by online data sources from Kaggle so some attributes are available partially or unavailable. A lot of cleaning and processing had been done, as well as the removal of some unnecessary attributes before the data can be used to analyze and predict revenue. We use both numerical and categorial attributes of the dataset to establish linear, random forest and decision tree regression. The predicted values are not close to the reported data values as we had hoped to achieve, but they are not too far off from the actual data. Furthermore, the predicted values are consistent between three regression model, which indicate that adding textual data like plot summary, movie name could help improve the result. Knowing about the interactions outside the data like the amount of likes for each trailer, the popularity of actors-directors and better technique like Deep Neural Network could help make the prediction more accurate.

INTRODUCTION

As of 2018, it has been reported that the global box office is worth \$41.7 billion. However, when the home entertainment revenue is included, the global film industry is worth over three times that at \$136 billion. Despite the massive profits, movie industry is one of the riskiest markets for investors due to its uncertainty and unpredictability. Once a movie fails to meet expectation, it can potentially place a stress on the financial status of the movie studio, and lead to the withdrawal of funds from the investors. Therefore, through analyzing the performance of movies on the market, we could find the correlation between attributes of a movie and predict the revenue on release.

CCS CONCEPTS

• **Computing methodologies** → Retrieval model and ranking

KEYWORDS

Revenue, profitability, variables, data, regression

ACM Reference format:

N. Nguyen, M. Pollich and B. Breyfogle. 2018. Box Office Revenue Prediction : Project Proposal Paper. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

RELATED WORK

Galvão, M. and Henriques, R. (2018). Forecasting Movie Box Office Profitability. Journal of Information Systems Engineering & Management, 3(3), 22. https://doi.org/10.20897/jisem/2658

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

WOODSTOCK'18, June, 2018, El Paso, Texas USA

The revenue of a movie can be predicted by constructing a predictive model based on regression, decision trees or neural networks. The paper proposed 3 different variables to use: an interval variable for the value of profitability of individual movies, a categorial variable with multiple classes of profitability values, and a binary variable that shows whether the movie is profitable or not. Using these three variables can increase the accuracy of the prediction and help answer how the market behaves, what appeal to the audiences and the risk associated with movies production.

Overall, our proposed methodology has three components: Data Collection, Data Cleaning, Data Analysis and Prediction. Cleaning the dataset and discarding the irrelevant data from the datasets as well as thorough analysis of the data within these sets. We compare the attributes that are provided across all of the datasets and build a database that encompasses only the attributes that we care about.

1. Data collection: We looked through the three different datasets that all contain different attributes for all the movies that were produced within the last decade or so. Within these selected datasets, some common attributes included: ratings (IMDB), title, actors, actresses, studio, and release date. We explored in more detail which attributes we want to focus on in and pick that dataset for our analysis. The Movies Dataset was chosen since it contains the attributes we are looking for. We chose not to collect data from the IMDB dataset due to multiple reviews expressing the difficulty with the formatting of the data.

2. Data Cleaning: In order to clean the data from the Movies Dataset, we constructed a relational database to make finding duplicates and unnecessary attributes easier. Moreover, with this database we were able to write various queries in order to narrow down relevant information. Many attributes were dropped like collection, homepage, original title… since we don't need them to predict the revenue. Some of the numerical attributes like budget and popularity were formatted in string and contains 0 values so they need to be reformatted to float or removed. Some categorial attributes are also converted to numerical values for consistency in the dataset, though using them as textual data could have preserved interrelations and make the prediction better.

3. Data Analysis and Prediction: After the data is cleaned and organized efficiently, we analyzed it and look for distinguishing patterns. As proven to be effective in the surveyed literature, our prediction model was based on three regression model and had some success in predicting the revenue.

DATA SET

Boxofficemojo Alltime Domestic Data

The data contains 16223 unique values of the lifetime gross, ranking, title, studio and production year of Hollywood movies. They were scrapped from BoxofficeMojo's listing and based on domestic gross.

https://www.kaggle.com/eliasdabbas/boxofficemojo-alltime-domestic-data/version/3

The Movies Dataset

This dataset contains files containing metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

https://www.kaggle.com/rounakbanik/the-movies-dataset/downloads/the-movies-dataset.zip/7

MovieLens Dataset

This dataset contains six files: genome-scores.csv, genome-tags.csv, links.csv, movies.csv, ratings.csv and tags.csv and describes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service. It contains 20000263 ratings and 465564 tag applications across 27278 movies. These datasets were created by 138493 users between January 09, 1995 and March 31, 2015 and was generated on October 17, 2016. Users were selected at random for inclusion. All selected users had rated at least 20 movies. No demographic information is included. Each user is represented by an id, and no other information is provided.

https://grouplens.org/datasets/movielens/20m/

MAIN TECHNIQUES APPLIED

Data cleaning to make a new, better dataset for analysis.

Data processing to remove unwanted attributes.

Mapping of budget to vote count, which are two most correlated attributes.

Training set and testing data set for analysis. This allow us to determine our models' accuracies.

Different factors were evaluated to see their contributions on movie revenue at box office

Most of the attributes that we used in the project are adopted from previous studies that have found especially useful and relevant. We considered not only the most relevant studies with a higher success rate (popularity, actor, director, budget, nominations, genre), but also other attributes that are not as commonly considered (vote average, runtime, spoken language). Some of the attributes that we have found to be used across all studies include:

Budget

Previous studies have shown that the budget of a movie is strongly correlated with the predictability of its popularity and reception. However, while some argue that big budgets represent better quality and positive reviews from both the audience and movie critics, the studies have shown that a big budget does not ensure a high revenue. Despite this finding, this attribute proved valuable in our analysis and design of the prediction model.

Genre

While many authors and critics state through their studies that action movies and thrillers are the most significant and popular among audiences, the evaluation of each genre and its correlation to the film's overall success can ultimately help weed out variables that are not relevant.

Premier Date

The movies premiere date is a variable that influences box office revenue. This is due to the fact that movie attendance, overall, tends to increase significantly on holidays or festive seasons.

TOOLS

Python – Pandas, Dataframe, Numpy, Sklearn, Seaborn

Github

Discord

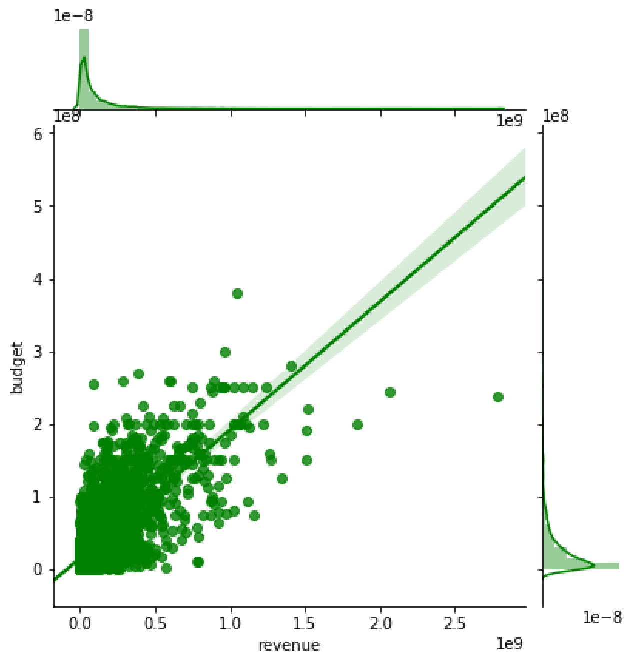
Slack

Jupyter Notebook

VISUALIZATION

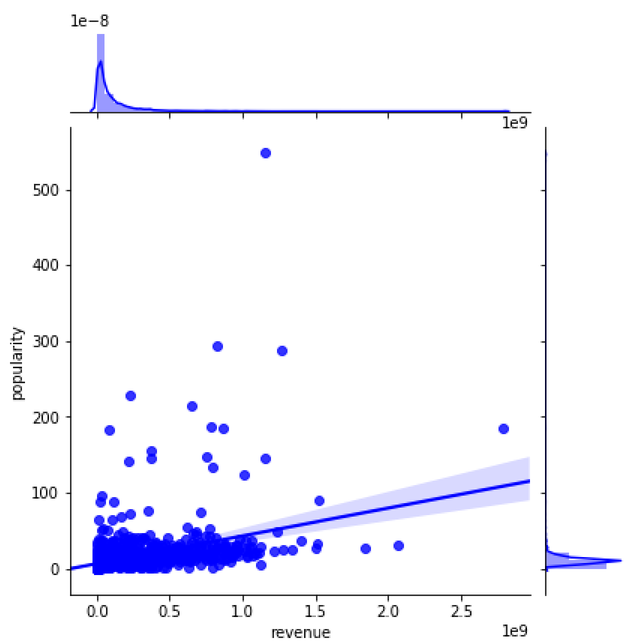
The movies dataset (more specifically, movies_metadata.csv) was chosen because in contains the necessary attributes like genres, budget, vote count, vote average, popularity… and revenue.

	adult	belongs_to_collection	budget	genres	homepage	id	imdb_id	original_language	o
0	False	{'id': 10194, 'name': 'Toy Story Collection', ...	30000000	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...	http://toystory.disney.com/toy-story	862	tt0114709	en	T
1	False	NaN	65000000	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	NaN	8844	tt0113497	en	J
2	False	{'id': 119050, 'name': 'Grumpy Old Men Collect...	0	[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...	NaN	15602	tt0113228	en	G C
3	False	NaN	16000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18,	NaN	31357	tt0114885	en	W E

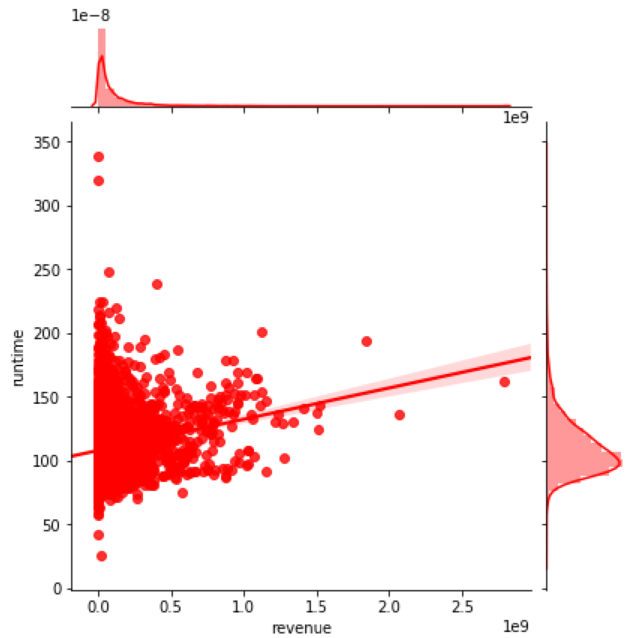


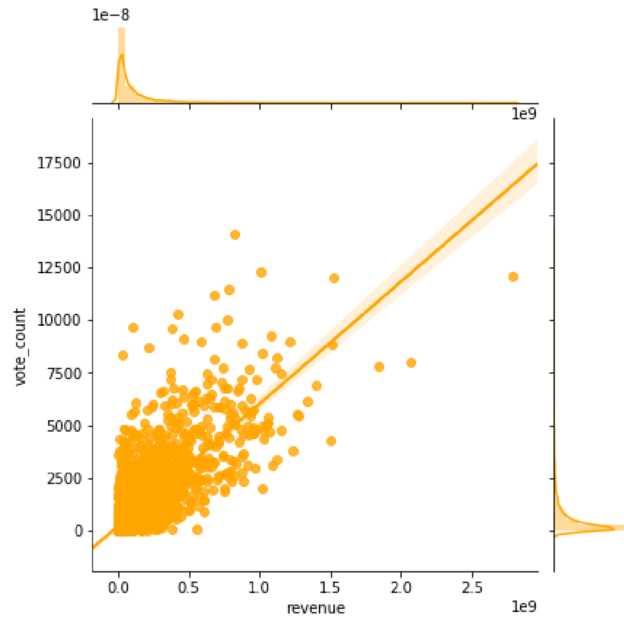
There were a lot of empty data in the original list (NaN value), and the values in budget and popularity was classified as string type instead of int and float type. Also, there was some movies with no budget, or no revenue, so they need to be removed as well. Some attributes like collections, homepage, status… are not required so they should also be dropped from the table. Therefore, data cleaning and processing are required, and the new dataset is as followed:

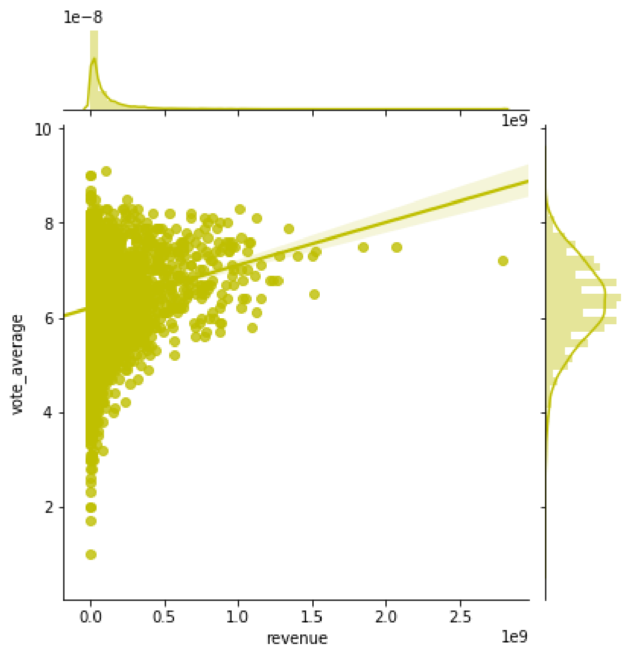
budget	genres	original_language	popularity	poster_path	production_countries	revenue	runtime	vote_average	vote_count	day	weekday	montl
30000000.0	'Animation', 'Comedy', 'Family'	en	21.946943	1	'United States of America'	373554033.0	81.0	7.7	5415.0	30	0	1
65000000.0	'Adventure', 'Fantasy', 'Family'	en	17.015539	1	'United States of America'	262797249.0	104.0	6.9	2413.0	15	4	1
16000000.0	'Comedy', 'Drama', 'Romance'	en	3.859495	1	'United States of America'	81452156.0	127.0	6.1	34.0	22	4	1
60000000.0	'Action', 'Crime', 'Drama', 'Thriller'	en	17.924927	1	'United States of America'	187436818.0	170.0	7.7	1886.0	15	4	1
35000000.0	'Action', 'Adventure', 'Thriller'	en	5.231580	1	'United States of America'	64350171.0	106.0	5.5	174.0	22	4	1



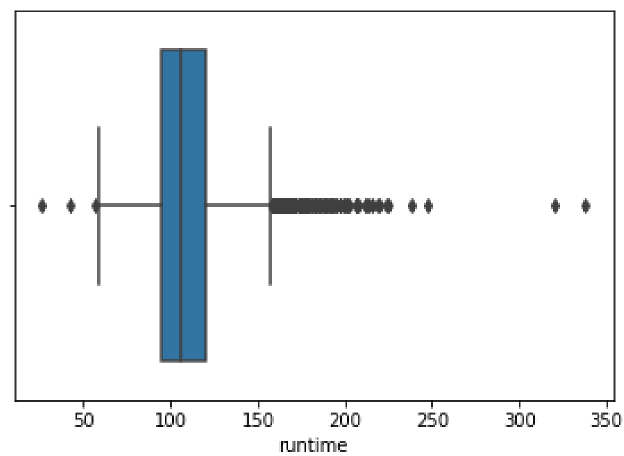
We then look at the relationship between revenue and some of the prominent attributes like budget, popularity, runtime, vote average, vote count. So far, budget and vote count show strong correlation with each other:

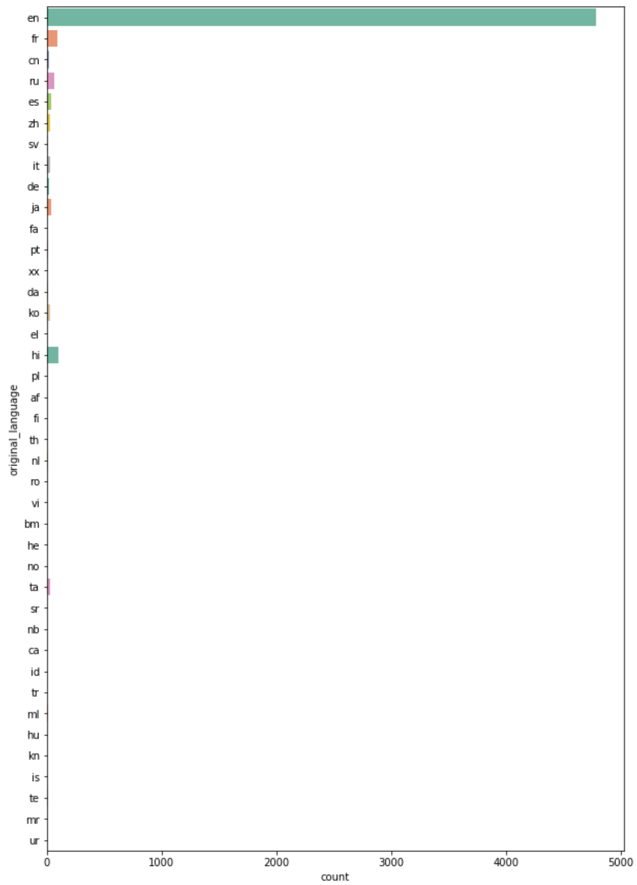


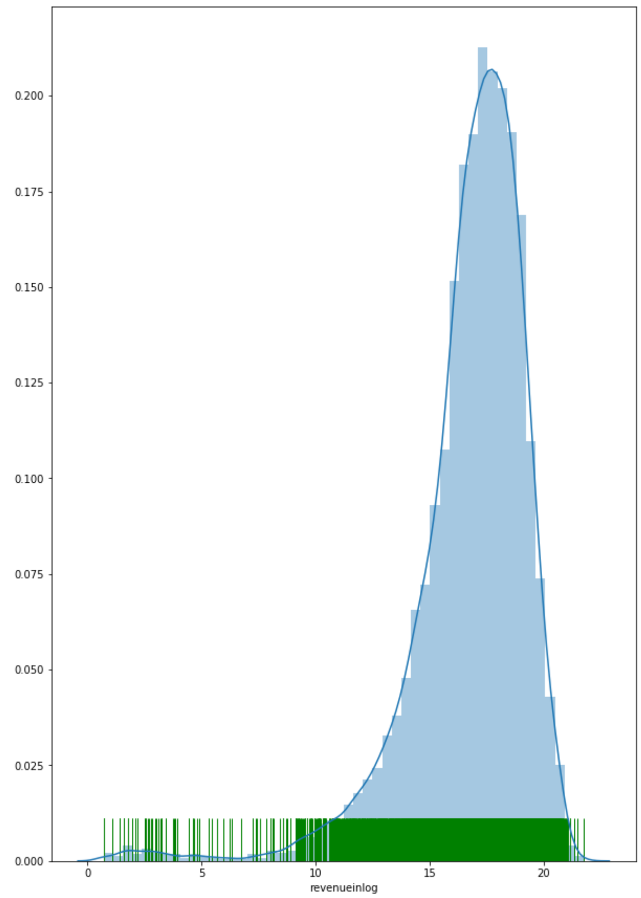


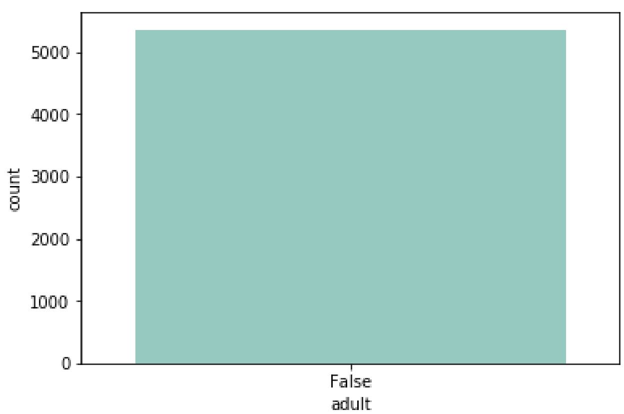


Some of the other attributes were also interesting. For examples, most movies have average runtime, are in English and none of the movies in the list are for adult:

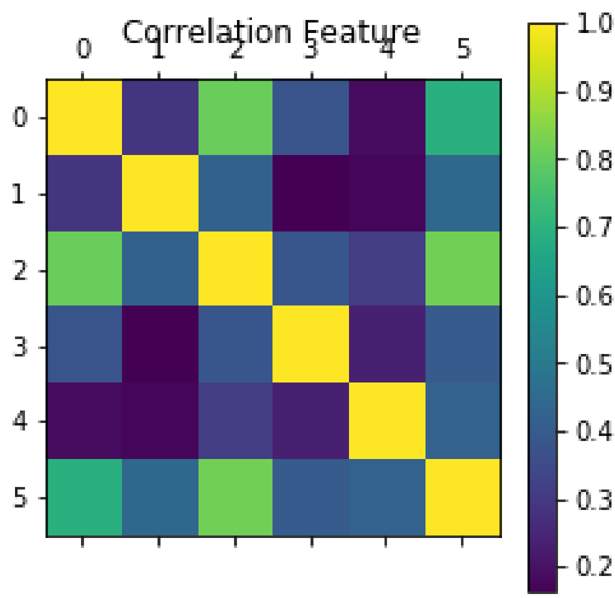




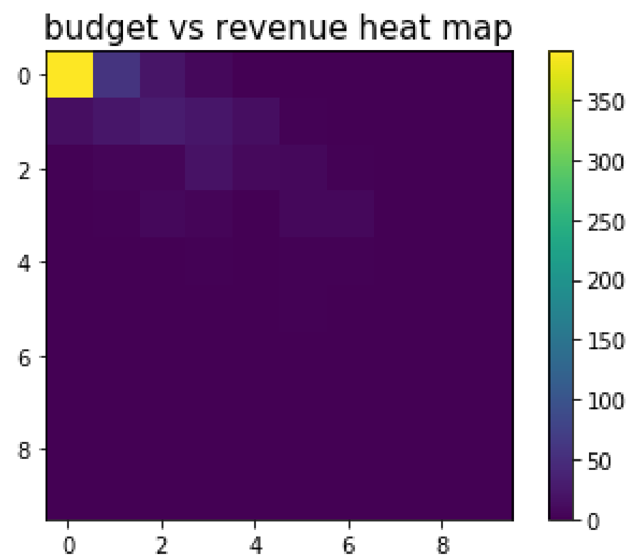


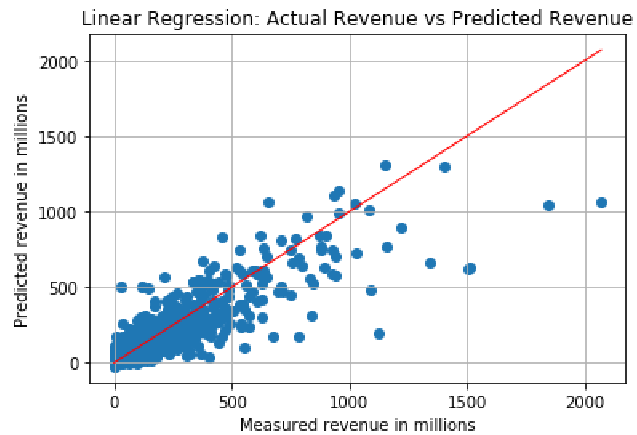


We also have a budget – revenue heat map and its correlation:



The revenue was also changed to log for easier observation.

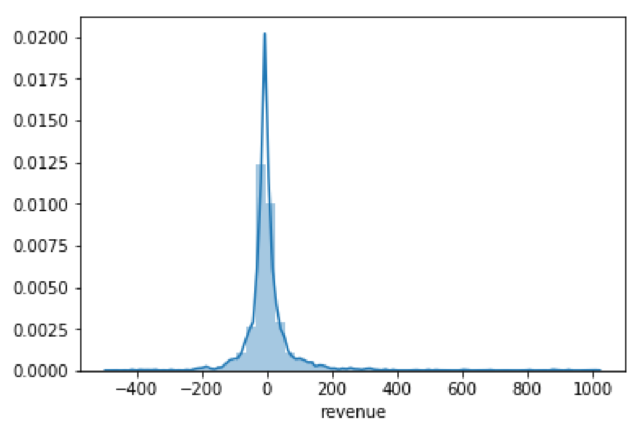




KEY RESULTS

We used train test split function (this does random splitting of the data), and multiple statistical values have been calculated like Mean Absolute Error, Mean Squared Error, Median Absolute Error, Explained Var Score, R^2 Score. Additionally, for Linear Regression, we also calculate Root Mean Squared Error, and Root Mean Squared Logarithmic Error.

Linear regression on all values:



Regression Scores(train_test_split):

Mean Absolute Error: 39.43043039854889

Mean Squared Error: 6817.850459880546

Median Absolute Error: 17.201820528760287

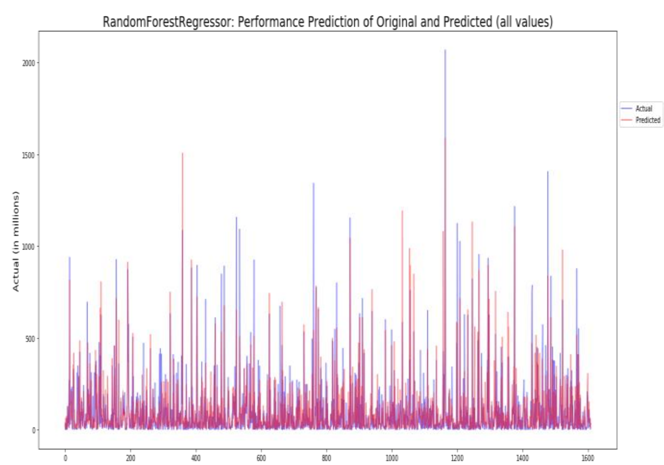
Explained Var Score: 0.7808385785290977

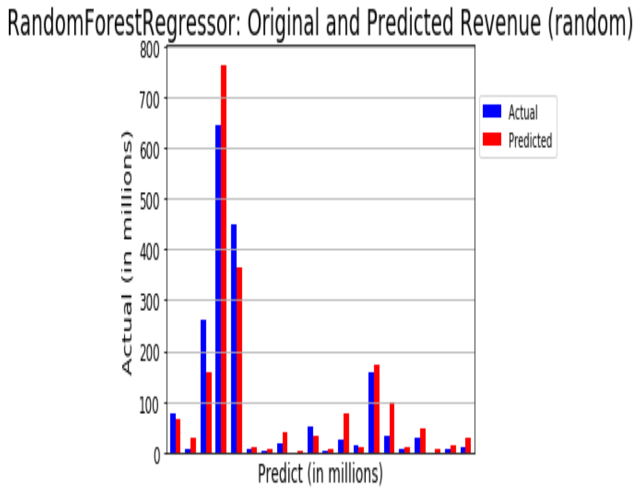
R^2 Score: 0.7804132535210337

Root Mean Squared Error: 82.57027588594183

Root Mean Squared Logarithmic Error: 1.0133662599662634

Random Forest Regression:





Regression Scores(train_test_split):

Mean Absolute Error: 44.32481404481044

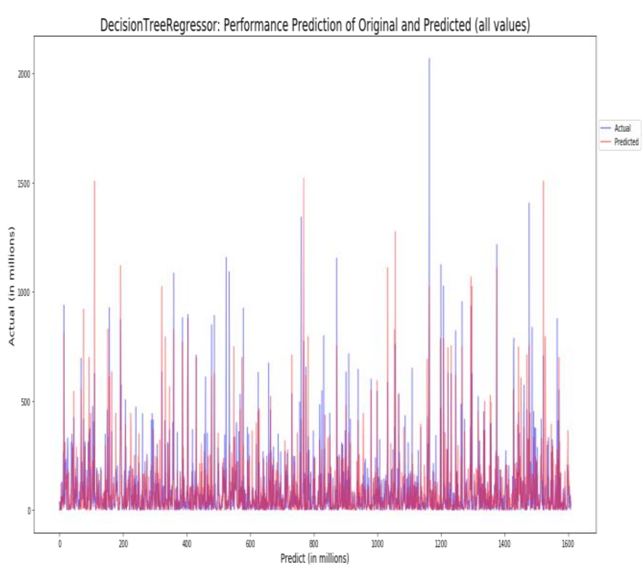
Mean Squared Error: 7290.243562892095

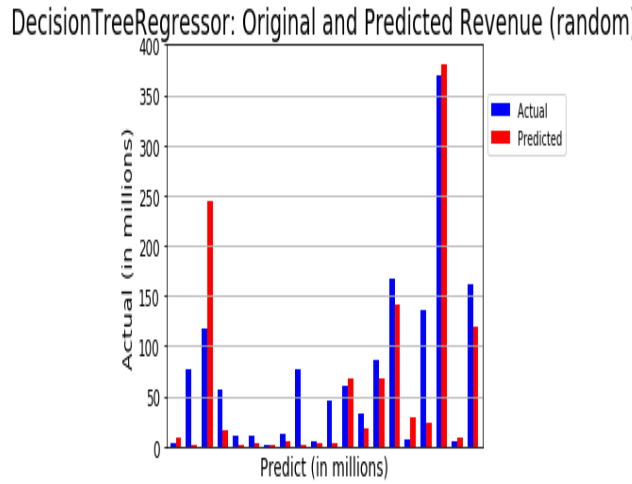
Median Absolute Error: 19.789213599999997

Explained Var Score: 0.7602265722831311

R^2 Score: 0.7600637723525838

Decision Tree Regression:





Regression Scores(train_test_split):

Mean Absolute Error: 59.009372080795515

Mean Squared Error: 14342.770276189405

Median Absolute Error: 22.663409

Explained Var Score: 0.5280692922902059

R^2 Score: 0.5279512729040885

Regression Performance Evaluation for revenue prediction of all 3 models:

Metrics	Linear	Random Forest	Decision Tree
Mean Absolute Error	39.430	44.325	59.009
Mean Squared Error	6817.850	7290.244	14342.770
Median Absolute Error	17.202	19.789	22.663
Explained Var Score	0.781	0.760	0.528

R^2 Score	0.780	0.760	0.528

Other datas (Linear Regression):

Root Mean Squared Error: 82.570

Root Mean Squared Logarithmic Error: 1.013

APPLICATIONS

These are not accurate model, but we can use this as references for revenues of unreleased movies because of some of the most correlated attributes like budget and vote count.

These models could be improved in the future with better techniques like Scaling, Transforming, analyzing of Textual attributes like plot summary, preview…

Including the amount of interactions on social media, list of actors and directors… from other datasets (integration) could help further improve the predictive model.

ACKNOWLEDGMENTS

REFERENCES

[1] Galvão, M. and Henriques, R. (2018). Forecasting Movie Box Office Profitability. Journal of Information Systems Engineering & Management, 3(3), 22. https://doi.org/10.20897/jisem/2658

[2] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. http://dx.doi.org/10.1145/2827872