KAGGLE REPORT

HUANHUAN GE

Abstract. ~~The abstract will be put here, ...~~ In this report,I will talk about my Kaggle Project.In previous studies, I learned the use of Latex and Git and mastered their basic operations. I also learned about Python and data visualization. Now I'm going to use the Kaggle project to demonstrate what I've learned.

Contents

Date: (None).
2020 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Python,Machine Learning, Data Processing,Git,Latex.

1. Introduction

At a high level, what is the problem area you are working in and why is it important? It is important to set the larger context here. Why is the problem of interest and importance to the larger community?

This paragraph narrows down the topic area of the paper.In the first paragraph you have established general context and importance.Here you establish specific context and background.

1.1. **Description.** In a world... where movies made an estimated $41.7 billion in 2018, the film industry is more popular than ever. But what movies make the most money at the box office? How much does a director matter? Or the budget? For some movies, it's "In this paper, we show that . ..You had me at 'Hello.'" . This is the key paragraph in the intro - you summarize, in one paragraph, what are the main contributions of your paper given the context you have established in paragraphs 1 and 2. What is the general approach taken? Why are the specific results significant? This paragraph must be really good.For others, the trailer falls short of expectations and you think "What we have here is a failure to communicate."

1.2. **Target.** In this competition, you're presented with metadata on over 7,000 past films from The Movie Database to try and predict their overall worldwide box office revenue. Data points provided include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries. You can collect other publicly available data to use in your model predictions, but in the spirit of this competition, use only data that would have been available before a movie's release.

You should think about how to structure these one or two paragraph summaries of what your paper is all about. If there are two or three main results,then you might consider itemizing them with bullets or in test.

2. Data Processing

2.1. **Data Description.** In the dataset, it includes 7,398 movies and various metadata from the Movie Database (TMDB). Movies are labeled with id.Data points include cast,crew,plot keywords, budget, posters, release dates, languages, production companies, and countries.Predict the worldwide revenue for 4398 movies.

2.2. **Basic Information of Data.**

- e. g. , First . .. **train.csv** – it contains 3000 rows and 23 columns.
- e. g., Second ...
- e. g., Third . .. If the results fall broadly into two categories,you can bring out that distinction here. For example, **test.csv** – it contains 4398 rows and 22 columns. Compared with the train data, there are fewer "Our results are both theoretical and applied in nature. (two sentences follow, one each on theory and application)revenue"

  Keep this at a high level, you can refer to a future section where specific details and differences will be given. But it is important for the reader to know at a high level, what is new about this work compared to other work in the area.

  column.

- **sample_submission.csv** – it clarifies the data submission format. It just contains 2 columns that is "~~The remainder of this paper is structured as follows...~~id" ~~Give the reader a roadmap for the rest of the paper. Avoid redundant phrasing,~~ and "~~In Section 2, In section 3,... In Section 4, ... " etc.~~

  ~~Test citation [?]. and [?] or ?].~~
  ~~This is for ??, and this is for ??.~~
  ~~Number: 123. 10, 30, 50 and 70, 10 to 30, 10 m, 30 m and 45 m, and 10 %~~
  ~~We have 10 Hz, kg m s^{-1}, the range: 10 Hz to 100 Hz.1/2.~~
  ~~For ??,as shown below:~~

$$a = b \times \sqrt{ab}$$

3. ~~Preliminaries~~

3. ~~Method~~

3. ~~Experiment and Analysis~~

~~Precision Comparison on Event Detection Methods~~ revenue".

2.1. **Data Fields.** The following is basic information of data.

Name	~~OR Event Detection~~ Description	~~AC Event Detection TC Event Detection~~ Attribute
~~precision~~ train.csv	~~0.83~~ Training set(Movies from 1970-2018)	~~0.69 0.46~~ id,belongs_to_collection,budget,genres,homepage, imdb_id,original_language,original_title,overview, popularity,poster_path,production_companies, production_countries,release_date,runtime, spoken_languages,status,tagline,title,Keywords, cast,crew,revenue
~~recall~~ test.csv	~~0.68~~ Test set(Predict revenue)	~~0.48 0.36~~ id,belongs_to_collection,budget,genres,homepage, imdb_id,original_language,original_title,overview, popularity,poster_path,production_companies, production_countries,release_date,runtime, spoken_languages,status,tagline,title,Keywords, cast,crew
~~F-score~~ sample_submission.csv	~~0.747~~ Format of submission	~~0.57 0.4~~ id,revenue

2.2. **Numerical features.**

- There are 4 numerical features in total.
- The minimum of budget is 0.
- There are some missing values in the runtime, and the minimum of runtime is 0.

~~Conclusions~~

	id	budget	popularity	runtime	revenue
count	3000.000000	3.000000e+03	3000.000000	2998.000000	3.000000e+03
mean	1500.500000	2.253133e+07	8.463274	107.856571	6.672585e+07
std	866.169729	3.702609e+07	12.104000	22.086434	1.375323e+08
min	1.000000	0.000000e+00	0.000001	0.000000	1.000000e+00
25%	750.750000	0.000000e+00	4.018053	94.000000	2.379808e+06
50%	1500.500000	8.000000e+06	7.374861	104.000000	1.680707e+07
75%	2250.250000	2.900000e+07	10.890983	118.000000	6.891920e+07
max	3000.000000	3.800000e+08	294.337037	338.000000	1.519558e+09

FIGURE 1. Numerical features

2.3. **Missing Value.** Below is the missing field information.
We're going to remove some characteristic dimensions,such as columns that contain

id	0
belongs_to_collection	2396
budget	0
genres	7
homepage	2054
imdb_id	0
original_language	0
original_title	0
overview	8
popularity	0
poster_path	1
production_companies	156
production_countries	55
release_date	0
runtime	2
spoken_languages	20
status	0
tagline	597
title	0
Keywords	276
cast	13
crew	16
revenue	0
dtype: int64	

FIGURE 2. Missing values analysis

many null-valued features, columns from which Prediction of Revenue does not affect.

~~ACKNOWLEDGEMENT~~

2.4. **Process Genres.** Genres contains all type names and TMDB ids in JSON format. The type information in JSON format needs to be parsed. The following

figure shows the parsing result.

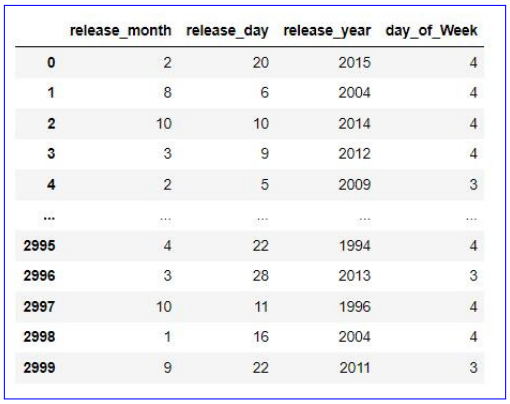
~~The authors would like to thank . . .~~

	genres
0	[Comedy]
1	[Comedy, Drama, Family, Romance]
2	[Drama]
3	[Thriller, Drama]
4	[Action, Thriller]
...	...
2995	[Comedy, Romance]
2996	[Drama, Music]
2997	[Crime, Action, Mystery, Thriller]
2998	[Comedy, Romance]
2999	[Thriller, Action, Mystery]

3000 rows × 1 columns

FIGURE 3. Genres

2.5. **Release date.** The Date data of the Date type needs to be parsed into three dimensions: release year, release month, and release Date. The analysis result is shown in the following figure.



	release_month	release_day	release_year	day_of_Week
0	2	20	2015	4
1	8	6	2004	4
2	10	10	2014	4
3	3	9	2012	4
4	2	5	2009	3
...	...	...	...	...
2995	4	22	1994	4
2996	3	28	2013	3
2997	10	11	1996	4
2998	1	16	2004	4
2999	9	22	2011	3

FIGURE 4. Date

2.6. **Data visualization.** Next, I will make a visual analysis of the impact of budget, popularity, Genres,runtime and date on revenue.

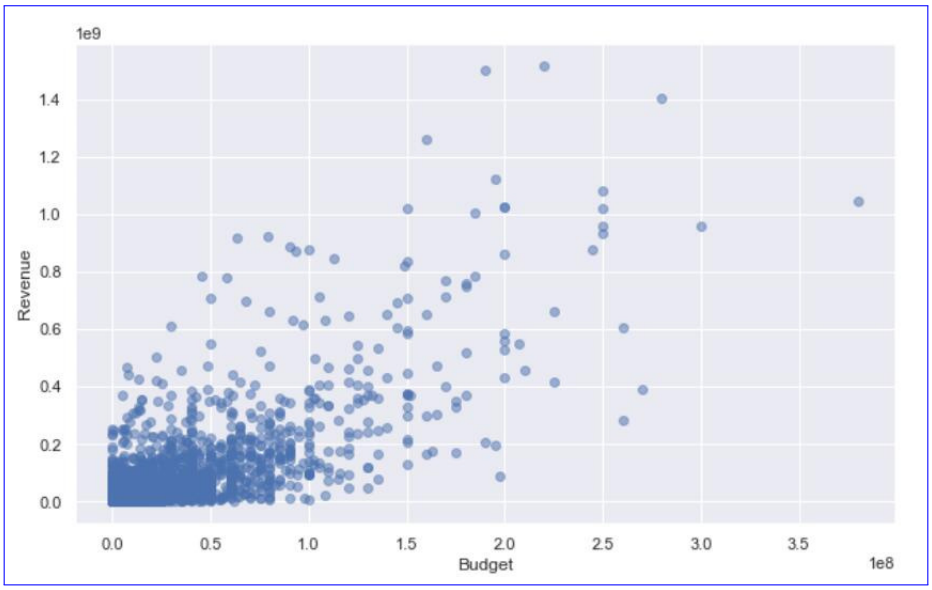


FIGURE 5. budget

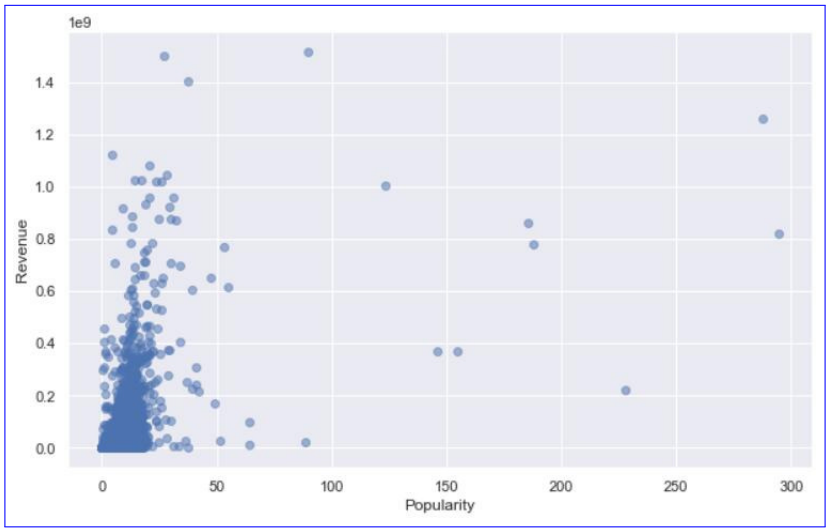


FIGURE 6. Popularity

Through correlation analysis of data visualization, we can find that there should be a positive correlation between budget and revenue, and the correlation between popularity and revenue should not be as strong as budget. In addition, runtime

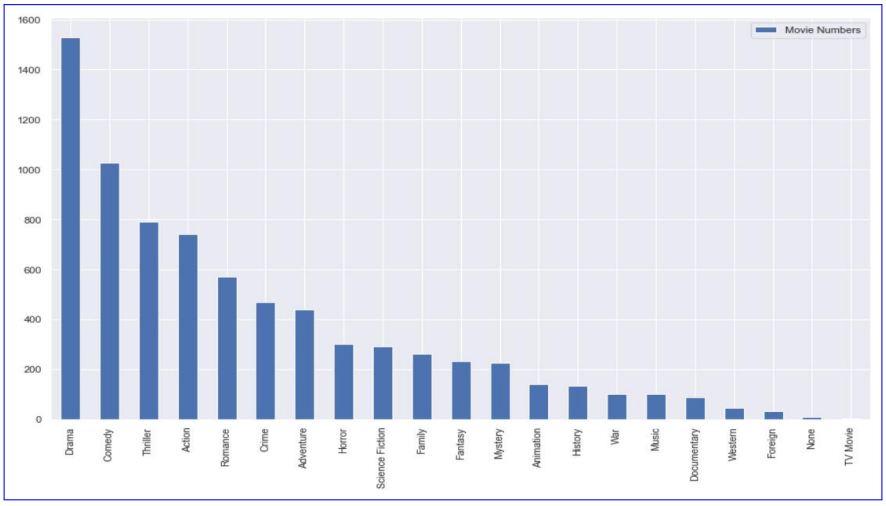


FIGURE 7. Genres

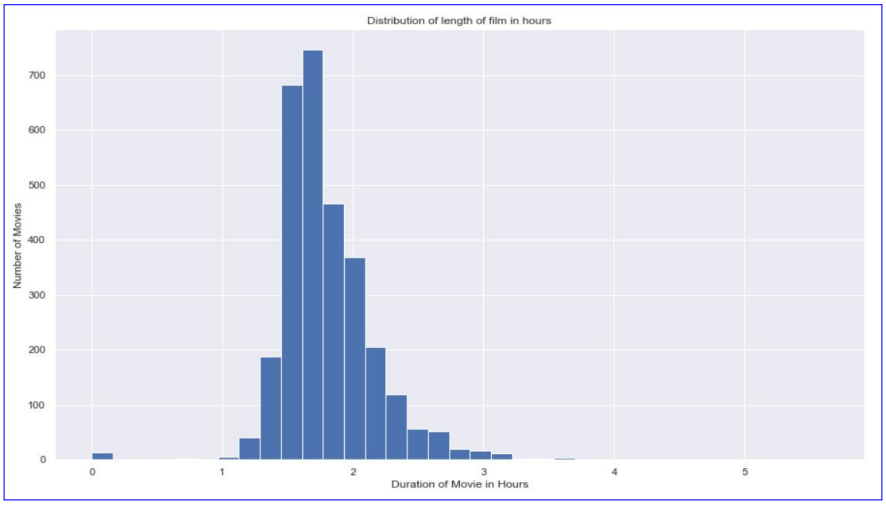


FIGURE 8. Runtime

and release dates should also have an impact on revenue.

3. Feature Selection And Model

3.1. **Release date.** I selected a total of 9 dimensions, including release year, release month, release week, language, number of companies, number of crew, budget, popularity and Runtime.

3.2. **Model and Result.** I used a random forest to train the data.A random forest is a classifier that contains multiple decision trees and whose output class is

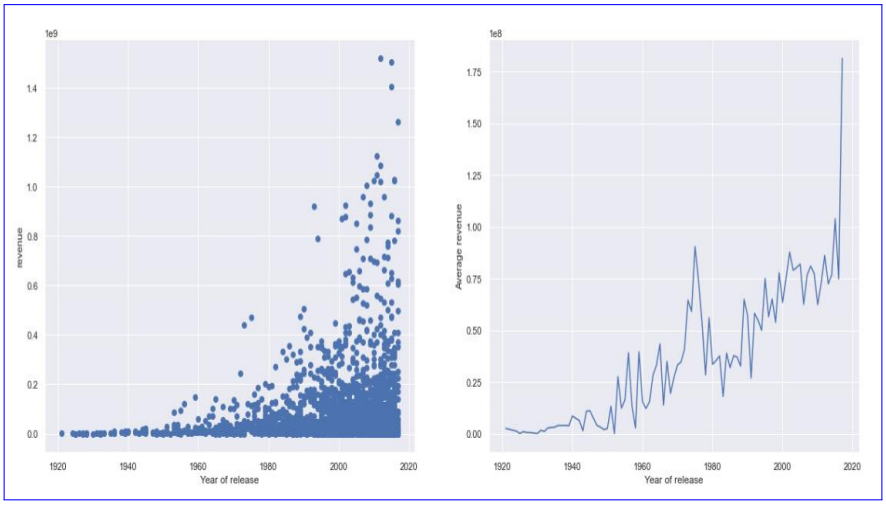


FIGURE 9. Date

determined by the plurality of the classes of the individual tree outputs.
I use the root mean square error to measure the prediction gap.The root mean square error is the square root of the ratio of the deviation between the predicted value and the true value to the number of observations n. In addition,I divided the data set into 0.8 training data set and 0.2 test data set.
Score:1.19

4. CONCLUSION

The effect of the model is general, the main reason is that the selected features are not comprehensive enough, and the learning and thinking of feature selection should be strengthened in the future.

(A. 1) SCHOOL OF COMPUTER SCIENCE,, QINGDAO UNIVERSITY OF TECHNOLOGY, QINGDAO, CHINA

Email address, A. 1: 18852862723@163.com