Title of This Paper*

What is my Subtitle?†

Ben Trovato‡
Institute for Clarity in Documentation
Dublin, Ohio
trovato@corporation.com

Gang Li§
Deakin University
Geelong, VIC 3216, Australia
gang.li@deakin.edu.au

ABSTRACT

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.

KEYWORDS

Python,Machine Learning, Data Processing,Git,Latex

ACM Reference Format:
Ben Trovato and Gang Li. 2022. Title of This Paper: What is my Subtitle?. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, **??** pages. https://doi.org/10.475/123_4

1 INTRODUCTION

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 "You had me at 'Hello.'" 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.

*Produces the permission block, and copyright information
†The full version of the author's guide is available as `acmart.pdf` document
‡Dr. Trovato insisted his name be first.
§Corresponding Author

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

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

- **train.csv** – it contains 3000 rows and 23 columns.
- **test.csv** – it contains 4398 rows and 22 columns. Compared with the train data, there are fewer "revenue" column.
- **sample_submission.csv** – it clarifies the data submission format. It just contains 2 columns that is "id" and "revenue".

2.3 Data Fields

The following is basic information of data.

Name	Description	Attribute
train.csv	Training set(Movies from 1970-2018)	id,belongs_to_collection,budget,ge imdb_id,original_language,origina popularity,poster_path,production production_countries,release_date ken_languages,status,tagline,title,I cast,crew,revenue
test.csv	Test set(Predict revenue)	id,belongs_to_collection,budget,ge imdb_id,original_language,origina popularity,poster_path,production production_countries,release_date ken_languages,status,tagline,title,K
sample_submission.csv	Format of submission	id,revenue

2.4 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.

2.5 Missing Value

Below is the missing field information.
We're going to remove some characteristic dimensions, such as columns that contain many null-valued features, columns from which Prediction of Revenue does not affect.

	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

```
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

	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

	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 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.

2.7 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.

2.8 Data visualization

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

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 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.

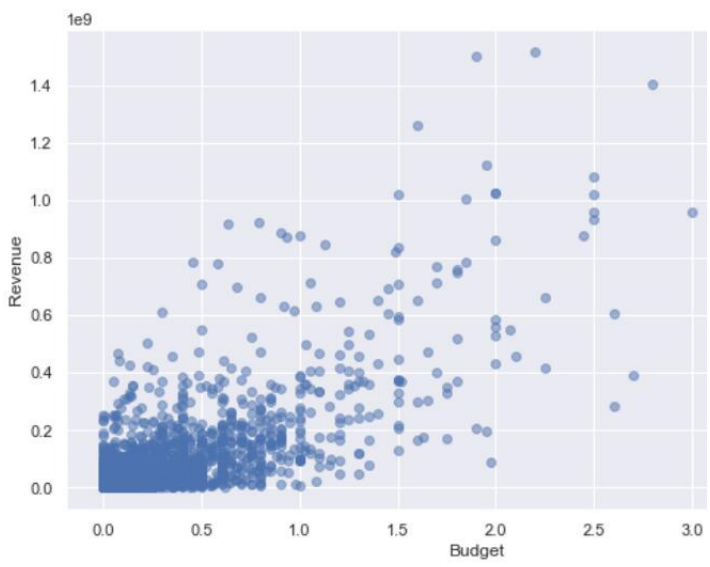

Figure 5: budget

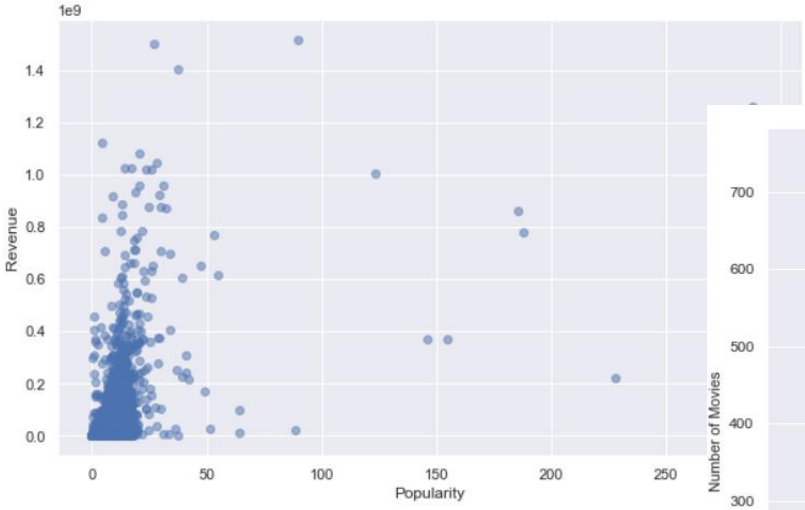

Figure 6: Popularity

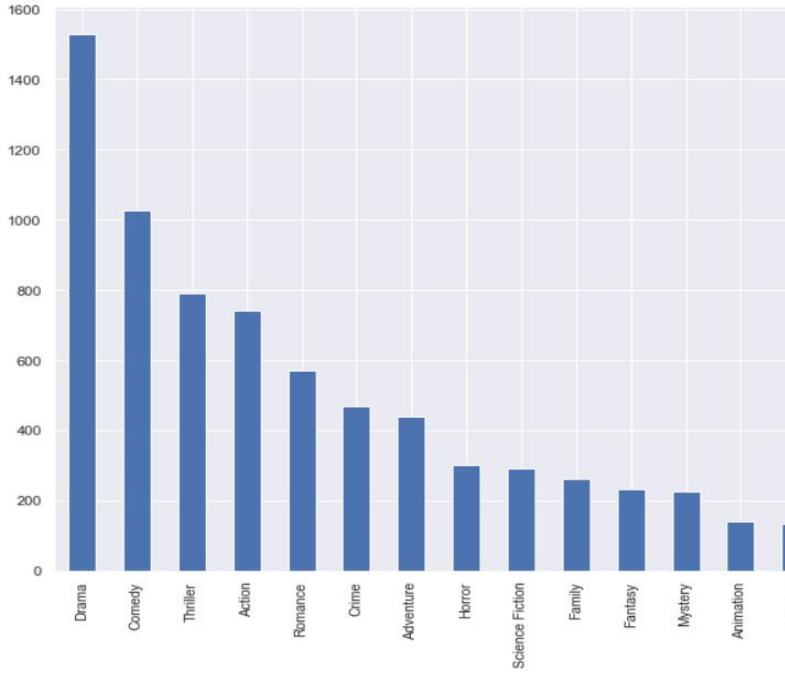

Figure 7: Genres

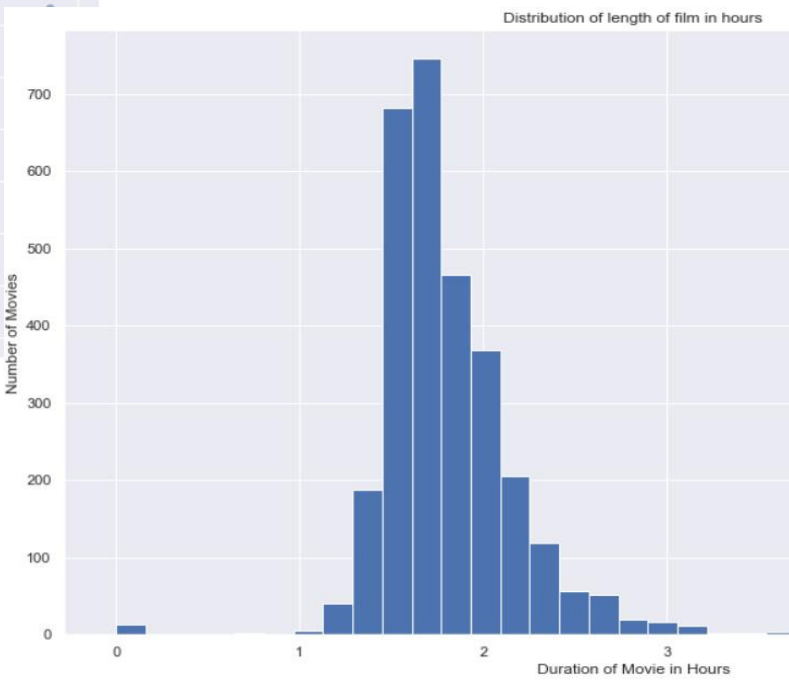

Figure 8: Runtime

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 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

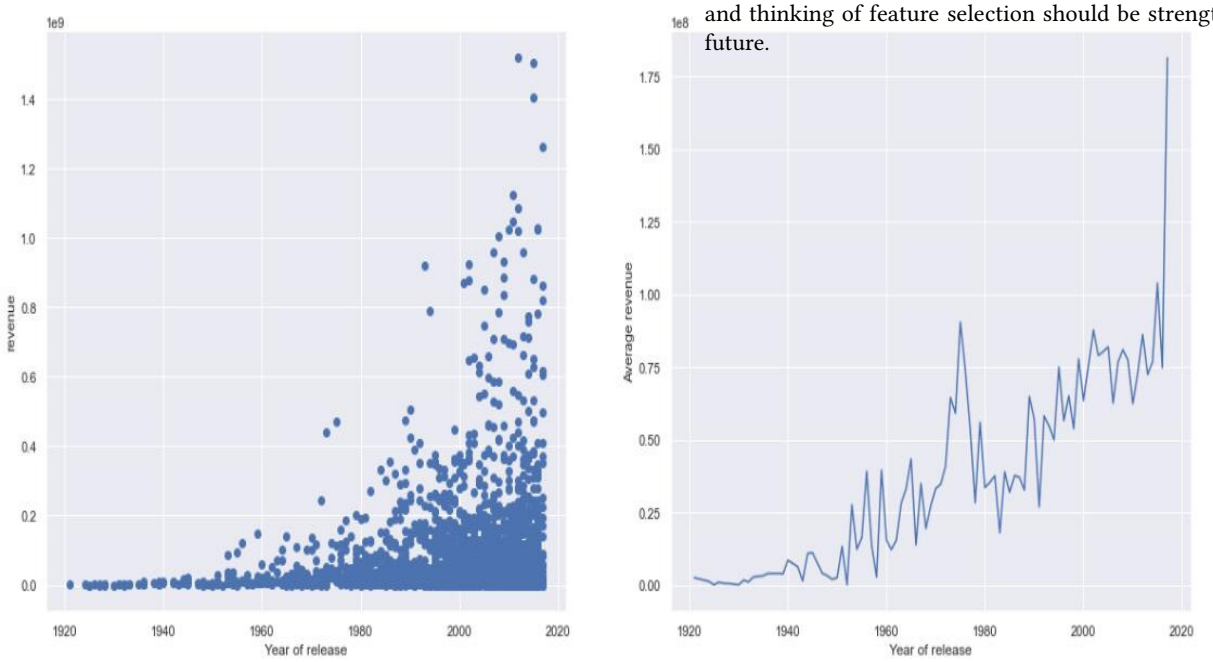

Figure 9: Date

and thinking of feature selection should be strengthened in the future.