Kaggle Project Report

Huanhuan Ge
Qingdao University of Technology, China

Introduction

This is a prediction problem. The datasets includes 7,398 movies and various metadata from the Movie Database (TMDB).We need to predict the worldwide revenue for 4398 movies.The datasets' attributes are shown below.

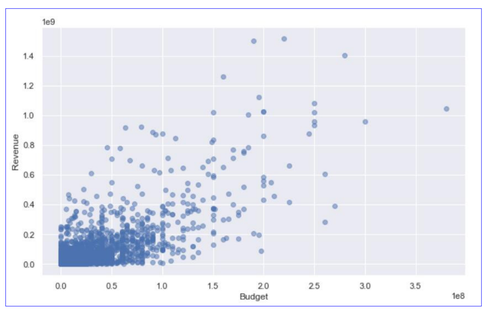
Name	Attirbute
sales_train.csv	id,belongs_to_collection,budget,genres,homepage, imdb_id,original_language,original_title,overview, popularity,poster_path,production_companies, production_countries,release_date,runtime, spoken_languages,status,tagline,title,Keywords, cast,crew,revenue
test.csv	id,belongs_to_collection,budget,genres,homepage, 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
sample_submission.csv	id,revenue

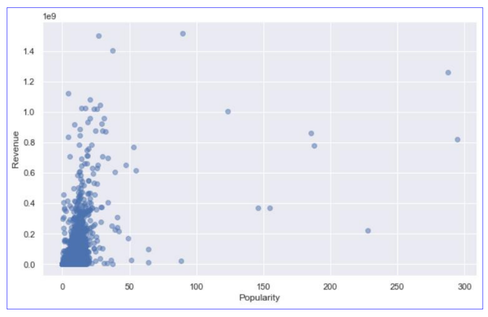
Data processing

- Basic Information of Data
- Numerical features
- Remove missing value and NaN value
- Filter outliers Data
- Process Genres
- Process Date

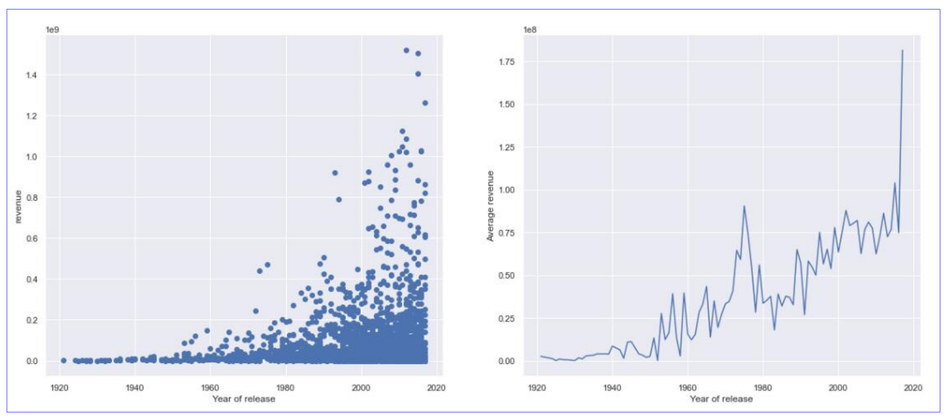
Data Visualization

Graph of budget, popularity and revenue.





Year And Revenue.



Feature Selection

- release_data: release_year, release_day, release_month
- original_language,number_companies,crew_size
- budget,popularity, runtime

	release_year	release_day	release_month	original_language	budget	popularity	runtime	number_companies	crew_size
0	2015	20	2	7	14000000	6.575393	33	3	72
1	2004	6	8	7	40000000	8.248895	53	1	9
2	2014	10	10	7	3300000	64.299990	45	3	64
3	2012	9	3	13	1200000	3.174936	62	0	3
4	2009	5	2	18	0	1.148070	58	0	2
...	...	...	...	...	...	...	...	...	...
2995	1994	22	4	7	0	9.853270	42	2	17
2996	2013	28	3	29	0	3.727996	42	2	15
2997	1996	11	10	7	65000000	14.482345	60	3	10
2998	2004	16	1	7	42000000	15.725542	30	2	89
2999	2011	22	9	7	35000000	10.512109	46	6	48

TULIP Team for Universal Learning and Intelligent Processing

Modeling and Result

- Model:~~Lightgbm~~ Random Forest
  - ~~Score:0.93740(RMSE)~~ 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.
- RMSE
  - ~~Rank: 3027/8738~~ 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.
- Train set: 0.8 , Test set: 0.2
- Score: 1.19

Conclusion

~~Compare to midterm presentation,RMSE decreased from 1.04885 to 0.93740. The reason mainly is more features and different modeling. In the figure~~ 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 ~~importance and monthly sales, some historical feature play an important role~~selection should be strengthened in the future.
~~And lightgbm models and xgboost models result have a marginal difference, but processing speed of lightgbm model is faster.~~