

Name: Zanning Wang

zID: z5224151

Part1:

In the first part, in order to predict the revenue, I preprocess the runtime, production countries, spoken language, cast, crew, homepage and genres before loading them into different model. As we all know, the movie revenue is directly linked to the popularity of actors and directors, so I picked some famous actors and directors after 1990 from authoritative website, cause most of movies in the database are from 1990. The website shows below:

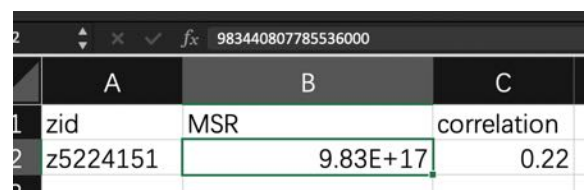
<https://www.imdb.com/list/ls058011111/>

<https://www.studiobinder.com/blog/best-movie-directors-of-all-time/>

In order to distinguish the influence of film companies on profits, we divide them into three types according to whether they are famous or not, so that the model can be classified. I also divide the runtime into four parts, because the movie which run longer are more likely to get more revenue.

By choosing three models logistic regression, linear regression and random forest regression, randomforest model has the best prediction, which correlation is 0.33.

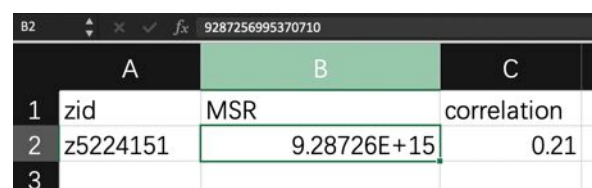
LogisticRegression model



Excel screenshot showing the results for the LogisticRegression model. The formula bar displays 983440807785536000. The table has columns A, B, and C. Row 1 contains headers: zid, MSR, and correlation. Row 2 contains the data for zid z5224151, with MSR value 9.83E+17 and correlation 0.22.

	A	B	C
1	zid	MSR	correlation
2	z5224151	9.83E+17	0.22

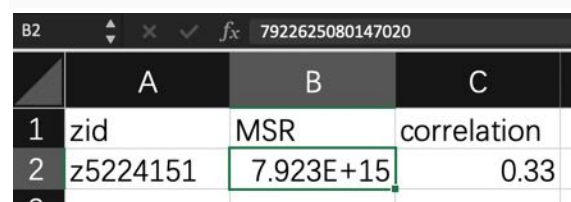
LinearRegression model



Excel screenshot showing the results for the LinearRegression model. The formula bar displays 9287256995370710. The table has columns A, B, and C. Row 1 contains headers: zid, MSR, and correlation. Row 2 contains the data for zid z5224151, with MSR value 9.28726E+15 and correlation 0.21.

	A	B	C
1	zid	MSR	correlation
2	z5224151	9.28726E+15	0.21

RandomForestRegressor model



Excel screenshot showing the results for the RandomForestRegressor model. The formula bar displays 7922625080147020. The table has columns A, B, and C. Row 1 contains headers: zid, MSR, and correlation. Row 2 contains the data for zid z5224151, with MSR value 7.923E+15 and correlation 0.33.

	A	B	C
1	zid	MSR	correlation
2	z5224151	7.923E+15	0.33

Part2:

In order to increase the accuracy for the rating prediction, I try to modify the function:

“preprocess_cast” and “preprocess_crew” , We know that the more movies with famous actors, the more likely to get a higher score. Therefore, I divided the number of famous actors into four categories to improve accuracy.

For this part, I choose three models to predict the rating: KNN, Bagging classifier and the gradient boosting classifier, the gradient boosting classifier and KNN both have the best accuracy which almost 70%

KNN model

A	B	C	D	E
	zid	average_pre	average_rec	accuracy
0	z5224151	0.62	0.53	0.7

GradientBoostingClassifier model

	A	B	C	D	E
1		zid	average_pre	average_rec	accuracy
2		0 z5224151	0.63	0.59	0.7
3					

#BaggingClassifier model

A	B	C	D	E
	zid	average_pre	average_rec	accuracy
0	z5224151	0.62	0.62	0.68