THE FACTORS OF THE GROSS EARNING IN THE MOVIES INDUSTRY

Jhih-Rou Huang

INTRODUCTION

Motivation

People like to watch movie as an entertainment. There are different kinds of movies and budgets and earning for them are all different.

Object and Goals

Our customer want to invest the film industry but they are new to this field. Hence, our goal is to investigate what are the important factors and sign we can predict whether a movie is profitable or not.

METHODOLOGY

Data

Scrapping data from Imdb website (multiple pages)

Merge two dataset based on the movie name

Algorithms

Delete duplicates and drop rows with null

Feature analysis and linear regression

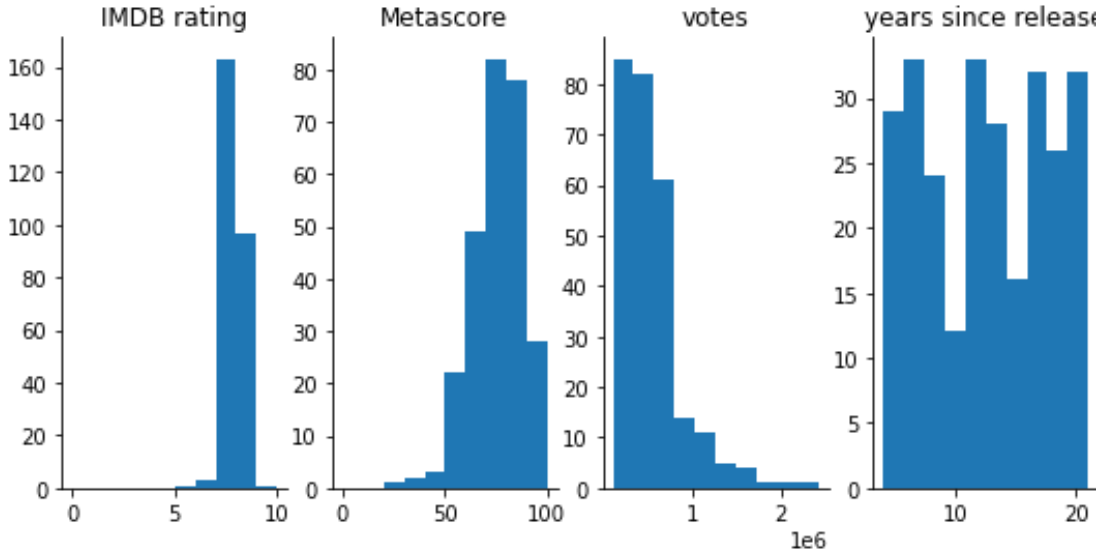
Rescale data

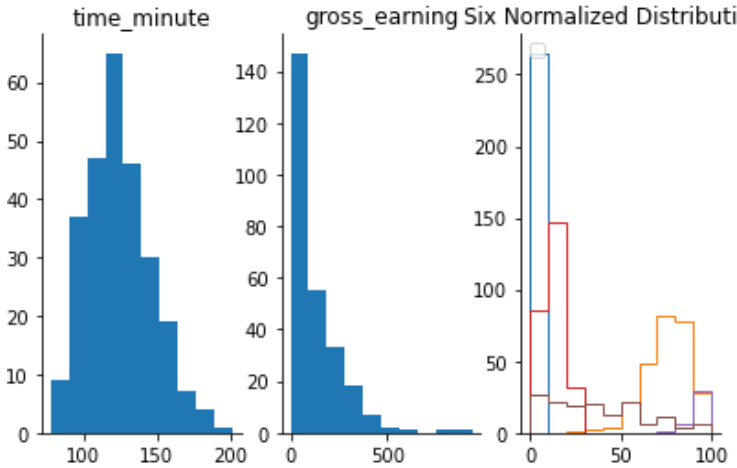
Kfold/ Train/ Test/ Validation

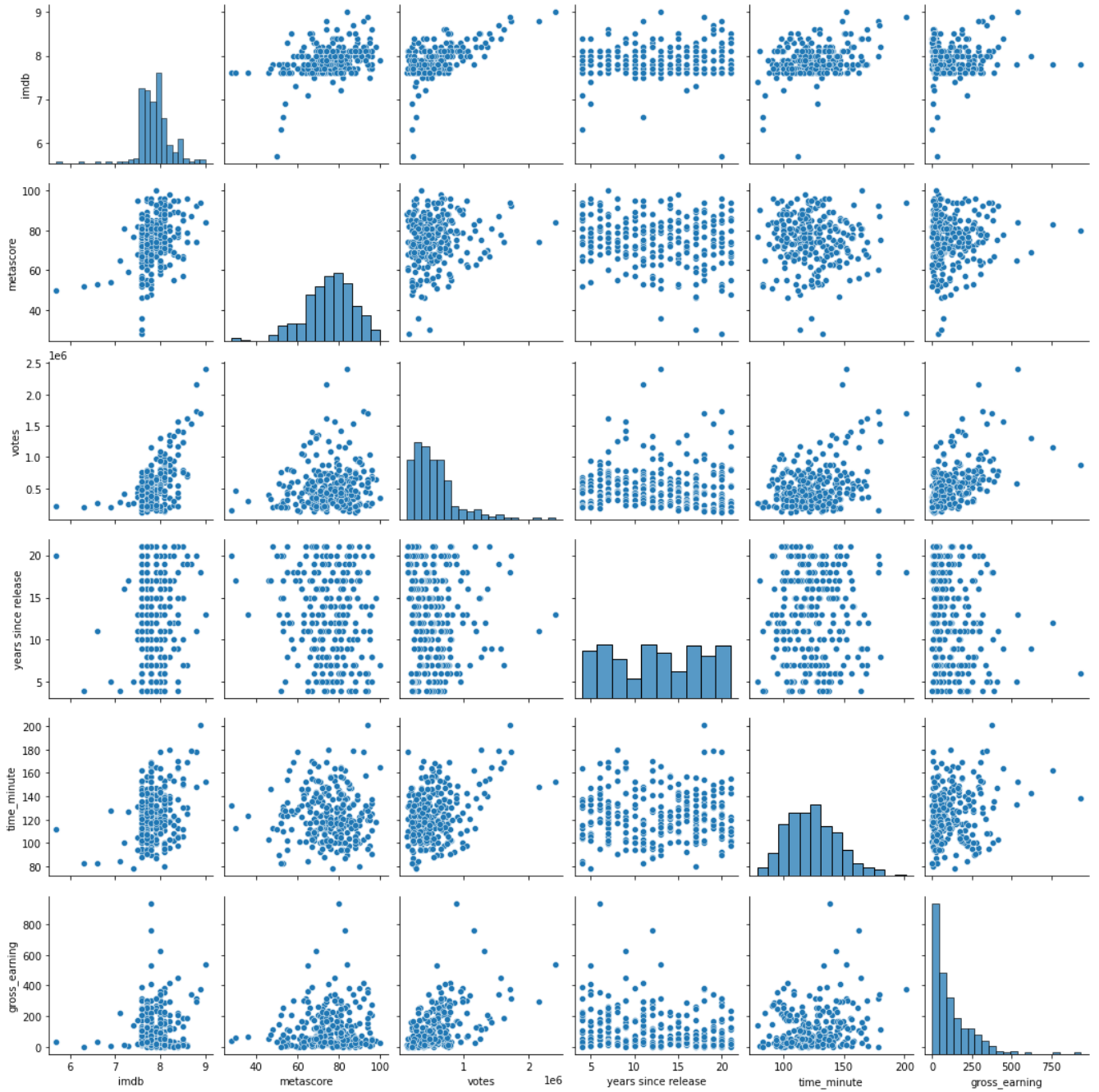
Tools

BeautifulSoup/ Numpy/ Pandas/ Scikit-learn/ Statsmodels/ Matplotlib/ Seaborn

FEATURES


IMDB rating
Metascore
votes
years since release
1e6


time_minute
gross_earning Six Normalized Distributions


imdb
metascore
votes
1e6
years since release
time_minute
gross_earning

FEATURES

```
x=movies[['imdb','metascore','votes','years since release','time_mi
x=sm.add_constant(x)
model = sm.OLS(movies['gross_earning'], x).fit()
model.summary()
```

OLS Regression Results

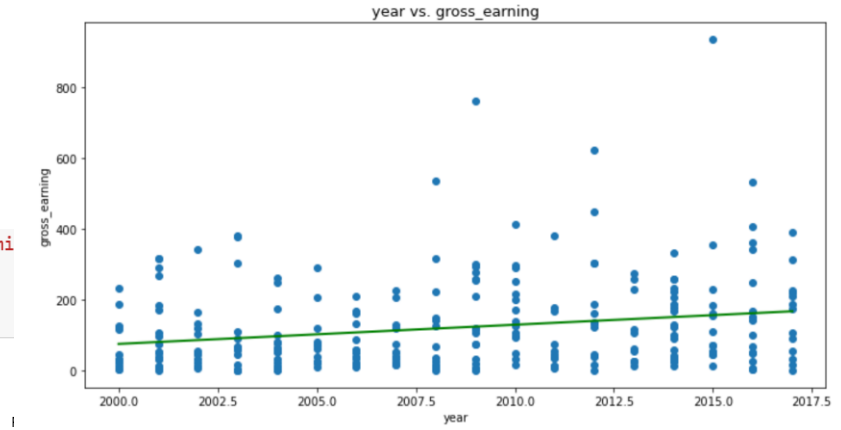
Dep. Variable:	gross_earning	**R-squared:**	0.383
Model:	OLS	**Adj. R-squared:**	0.372
Method:	Least Squares	**F-statistic:**	32.22
Date:	Tue, 14 Sep 2021	**Prob (F-statistic):**	1.72e-25
Time:	17:24:47	**Log-Likelihood:**	-1603.6
No. Observations:	265	**AIC:**	3219.
Df Residuals:	259	**BIC:**	3241.
Df Model:	5		
Covariance Type:	nonrobust		

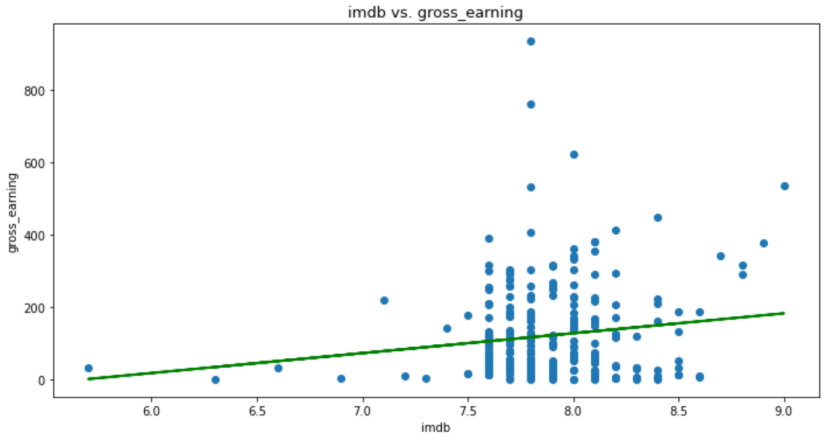
	coef	std err	t	P>|t|	[0.025	0.975]
const	804.1760	174.428	4.610	0.000	460.697	1147.655
imdb	-112.9820	25.024	-4.515	0.000	-162.258	-63.706
metascore	0.7416	0.582	1.275	0.204	-0.404	1.887
votes	0.0003	2.38e-05	10.752	0.000	0.000	0.000
years since release	-4.4981	1.263	-3.560	0.000	-6.986	-2.010
time_minute	0.5500	0.327	1.680	0.094	-0.095	1.195

Omnibus:	123.008	**Durbin-Watson:**	2.124
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	803.343
Skew:	1.747	**Prob(JB):**	3.60e-175
Kurtosis:	10.781	**Cond. No.**	1.78e+07

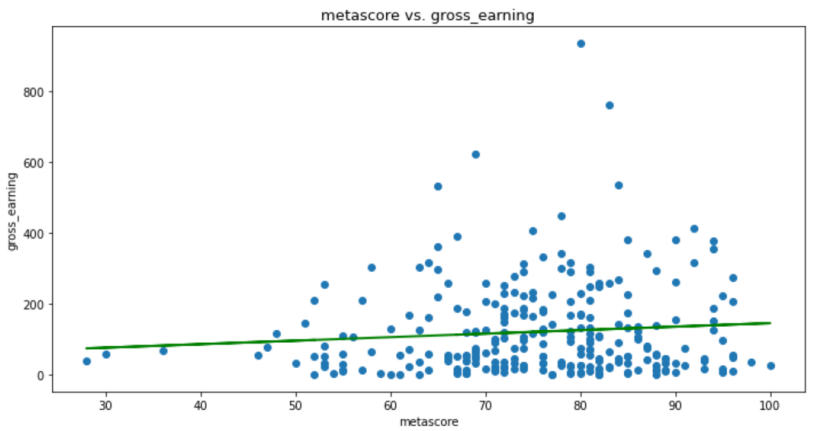
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly sp
[2] The condition number is large, 1.78e+07. This might indicate that there are strong multicollinearity or other numerical problems.

R-squared: 0.0459

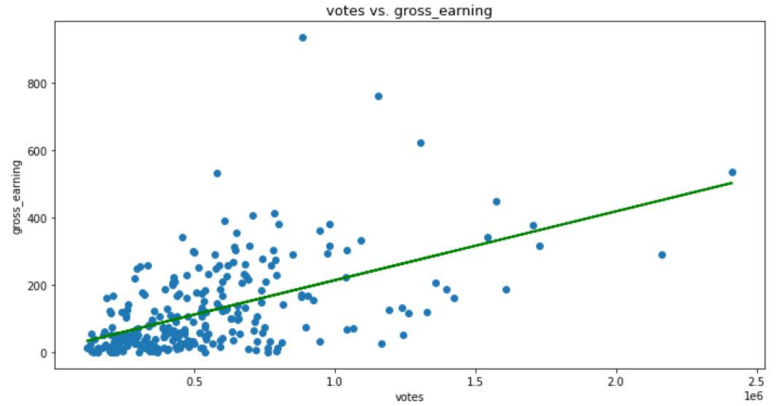




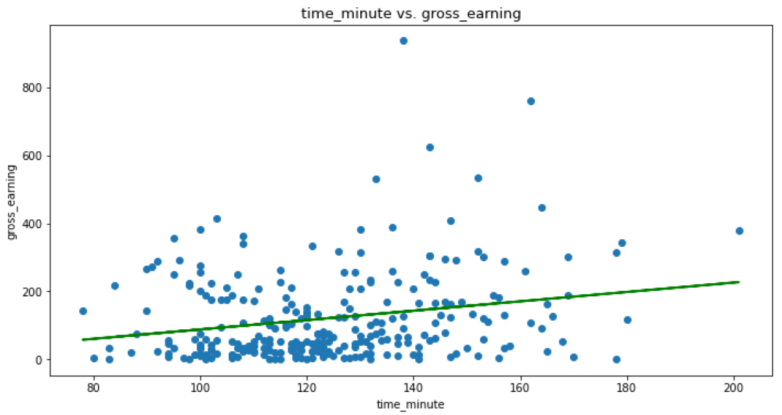
R-squared: 0.0084



R-squared: 0.2901



R-squared: 0.0501



TRAIN / VALIDATION / TEST

Linear Regression R-squared: 0.22

2 Degree polynomial regression val R-squared: 0.29

2 degree polynomial regression is slightly better than the linear one

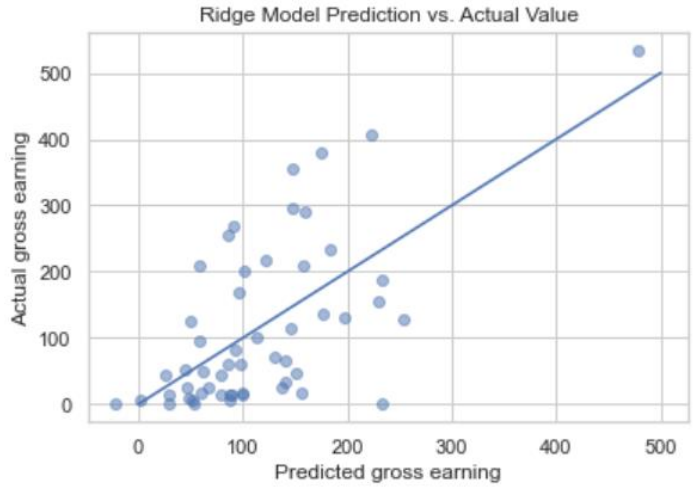
Ridge R-squared: 0.3845

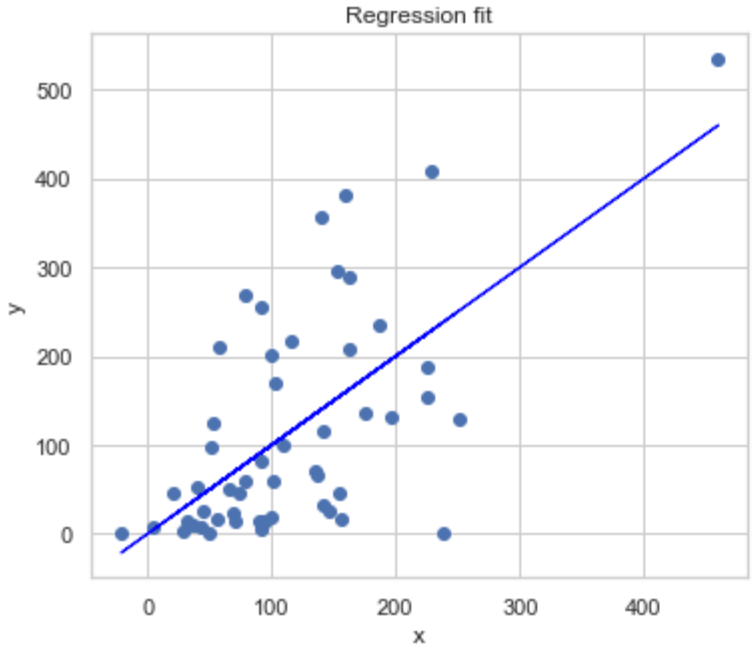
The ridge model is slightly better than the linear regression model.

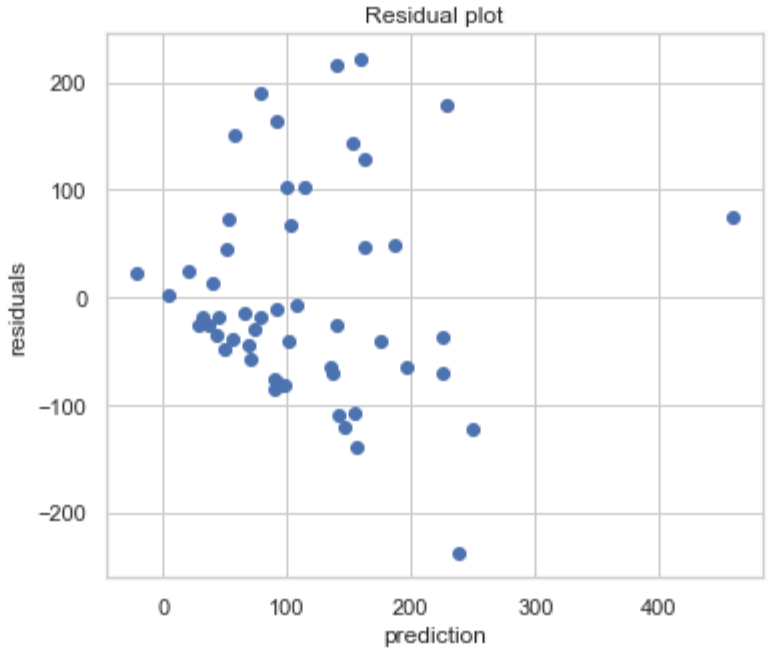
TRAIN / VALIDATION / TEST

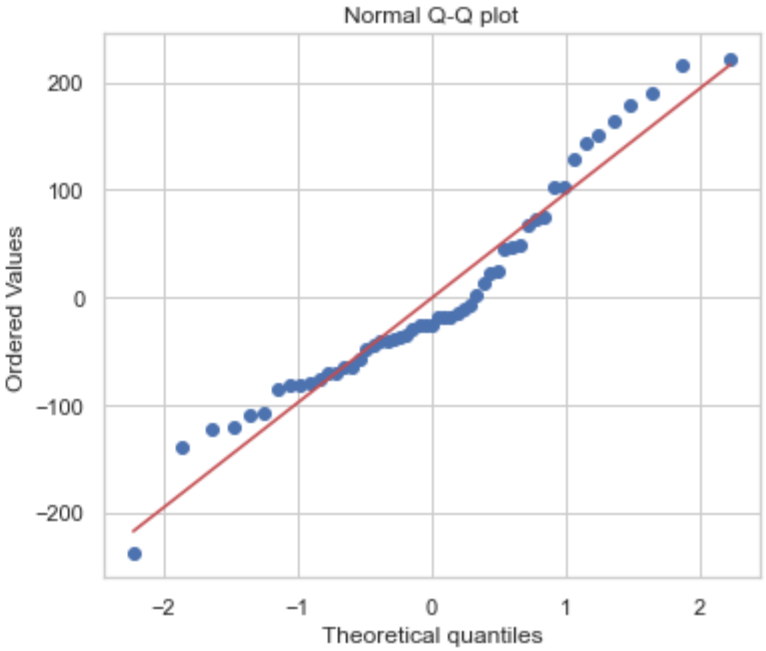
Rescale the values of every features and do the ridge model

```
Ridge R-squared: 0.3984
MAE: 76.7009
```







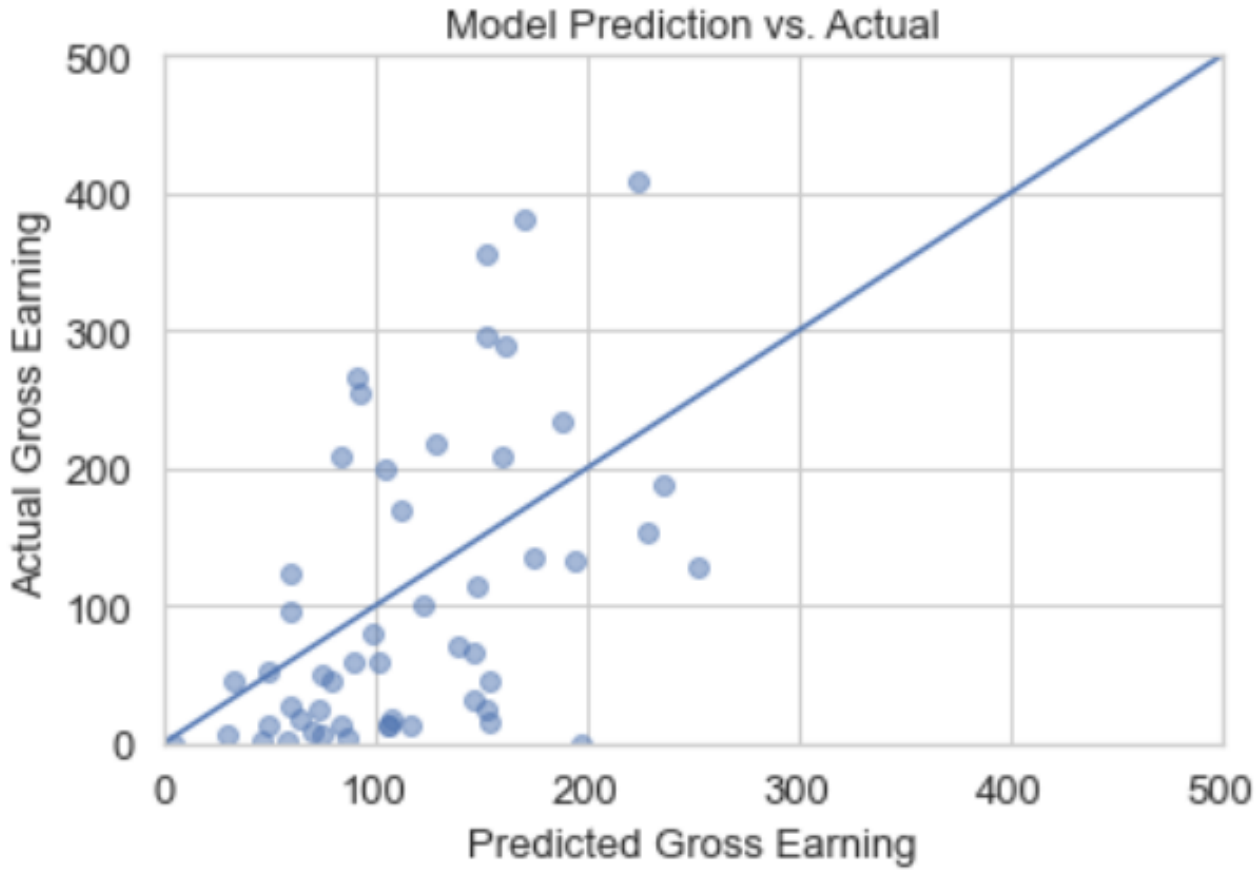


RIDGE

With the best alpha 10

Ridge R-squared: 0.40267

```
Ridge r2: 0.40267
Ridge MAE 79.23316
```



CONCLUSIONS