Using Machine Learning to Predict Gross Income of Documentaries

Presented by Aisulu Omar
aisuluomar123@gmail.com

The goal of the project

This is Tina.

Independent filmmaker.

Documentary - “The Quiet Ones”.

My goal is to predict the gross income of the documentary.

Data and Features

1609 movies after cleaning. From 1982 to the present, scraped from www.boxofficemojo.com.

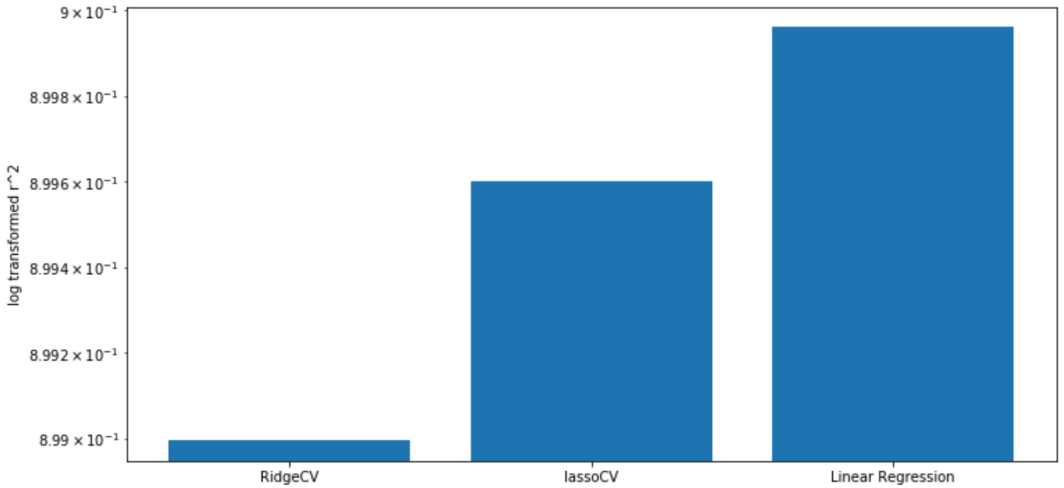
Target: Gross Income

Parameters:

- Total number of theatres the movie will be shown
- Opening gross income
- Opening number of theatres the movies was shown
- Date of release
- Year of release
- Month of release
- Film Studio

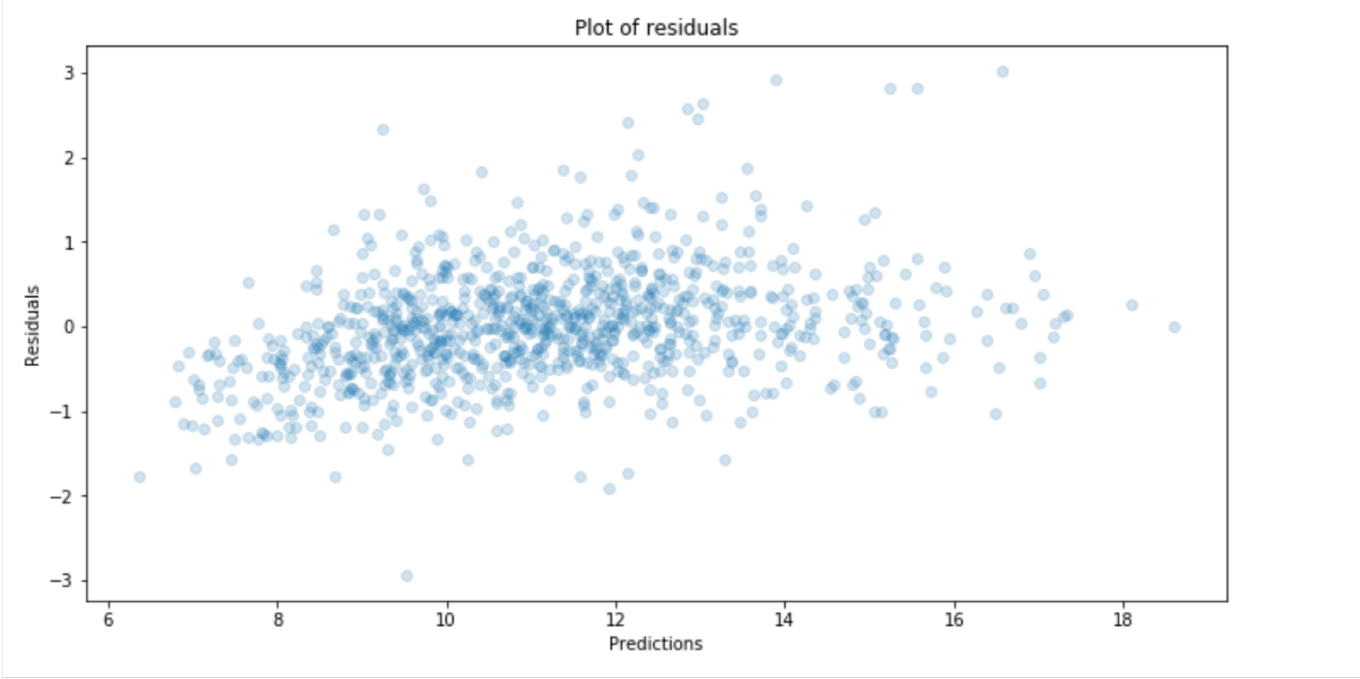
Process

Train (60% of data) -> **Test** (20% of data) -> **Validation** (20% of data) sets.



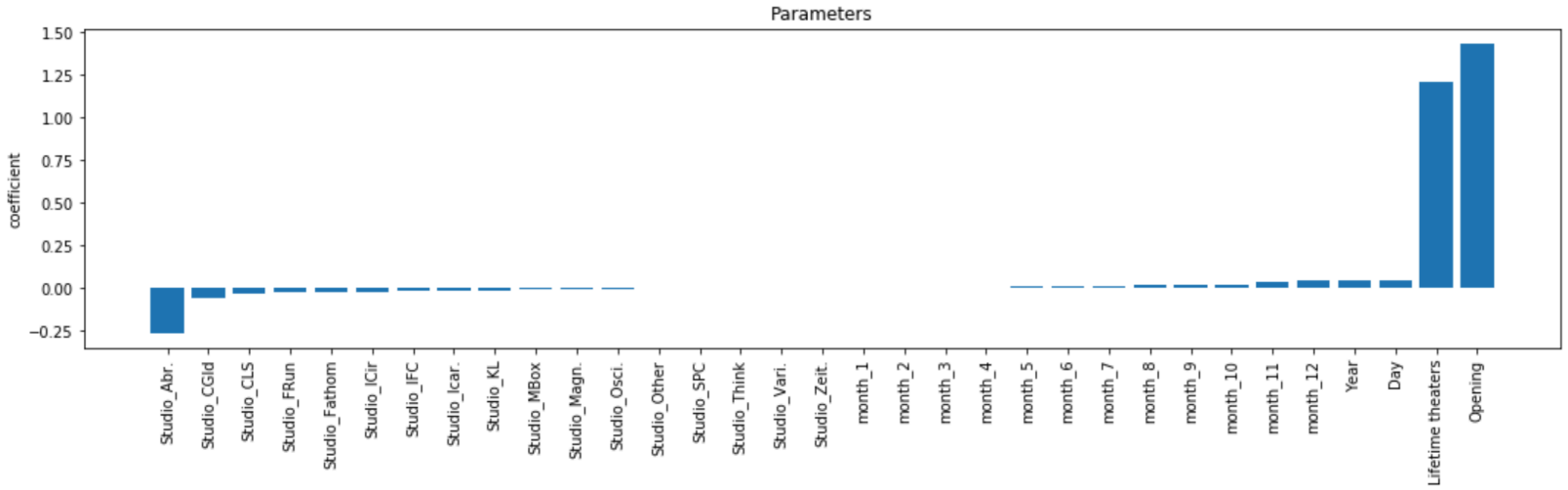
R squared
RidgeCV = 0.898
LassoCV = 0.8996
Linear Regression = 0.8999

Why should you trust my model?



Parameters

Interpreting my parameters and performing feature engineering



The Outcomes

Predicted: $180,120.

Actual outcome: $178,875.

0.9% higher than the actual income.

Released on 28 November 2018
Opening gross: $65,000.00
Number of opening theaters: 3
Number of total theaters the movie will be shown: 13
Studio: Independent

Features I would like to add

- Runtime of a movie
- Marketing channels
- Topic of a documentary

Targets I would like to predict

- Audience score
- Critic score

Thank you.

Linear Regression

Before training my model, I splitted my datset into train (60% of data) -> test (20% of data) -> validation (20% of data) sets.

Results of the Linear Regression vs Validation set:

```
MAE: 0.5197582152647009
MSE: 0.4422856242984291
RMSE: 0.6650455806171702
R^2: 0.899960806950367
```

- **MAE** (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.
- **MSE** (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.
- **RMSE** (Root Mean Squared Error) is the error rate by the square root of MSE.
- **R-squared** (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.

LassoCV

I used LassoCV model to get more insights on the effect of my parameters.

LassoCV train data vs Validation set.

- R^2: 0.899.

Based on the results of LassoCV model coefficients, I removed features with 0 coefficient, and rerun OLS model on the training set with fewer features and got my final results.

RidgeCV

Before fitting my data into RigeCV and LassoCV models, I perfomed scaler transformation on my parameters.

I ran a RidgeCV model and compared results with validation set.

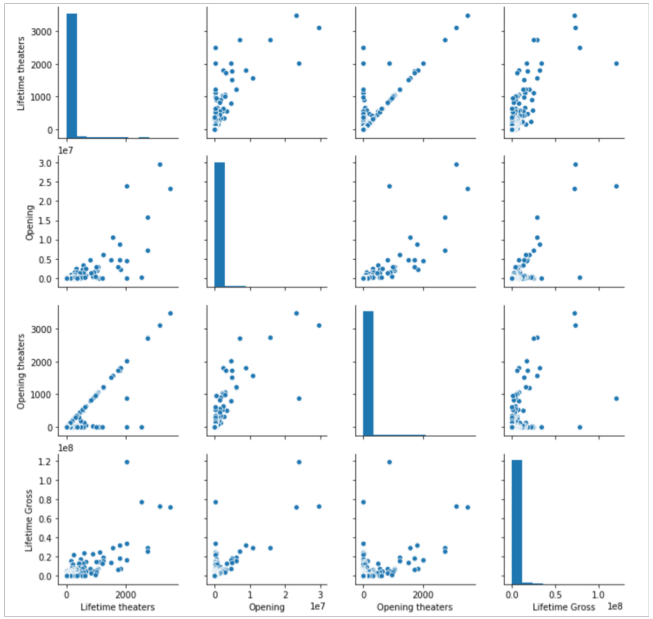
- R^2 slightly decreased: 0.898

I used RidgeCV to increase accuracy of my model, but it did not work.

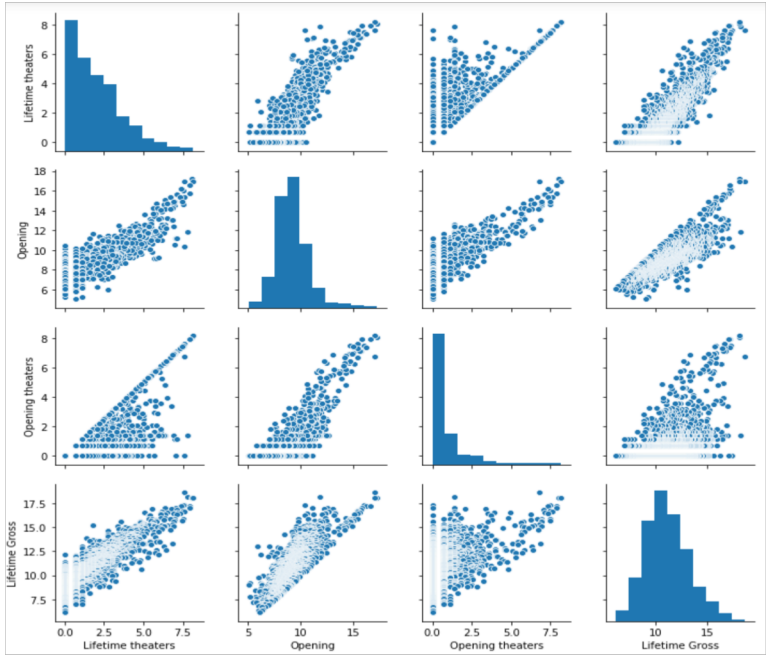
Let’s move to LassoCV.

Exploritory Data Analysis After Cleaning

Looks like my quantitative parameters are not normally distributed.



I needed to perform logarithmic transformation.



First and Final OLS models on training data

90 % variability explained by my model

Dep. Variable:	Lifetime Gross	**R-squared:**	0.902
Model:	OLS	**Adj. R-squared:**	0.899
Method:	Least Squares	**F-statistic:**	260.4
Date:	Wed, 17 Jul 2019	**Prob (F-statistic):**	0.00
Time:	15:55:29	**Log-Likelihood:**	-964.98
No. Observations:	965	**AIC:**	1998.
Df Residuals:	931	**BIC:**	2164.
Df Model:	33		
Covariance Type:	nonrobust		

99% variability explained by my model

Dep. Variable:	Lifetime Gross	**R-squared (uncentered):**	0.996
Model:	OLS	**Adj. R-squared (uncentered):**	0.996
Method:	Least Squares	**F-statistic:**	9193.
Date:	Wed, 17 Jul 2019	**Prob (F-statistic):**	0.00
Time:	16:04:57	**Log-Likelihood:**	-998.19
No. Observations:	965	**AIC:**	2052.
Df Residuals:	937	**BIC:**	2189.
Df Model:	28		
Covariance Type:	nonrobust		

Do you want to know how I got there?