CS584 – machine learning

fall 2016

Regression Model for Predicting Movie Box Office Gross

Group Members:

Ying Wu

Yingjuan Wu

Sahand Zeinali

Table of Contents

[Task 2](#_Toc467172373)

[Dataset 2](#_Toc467172374)

[Data source 2](#_Toc467172375)

[Target variable 2](#_Toc467172376)

[Features 2](#_Toc467172377)

[Data size 2](#_Toc467172378)

[Preprocessing 2](#_Toc467172379)

[Visualization 2](#_Toc467172380)

[Target 2](#_Toc467172381)

[Features 2](#_Toc467172382)

[Evaluation 2](#_Toc467172383)

[Performance Measure 2](#_Toc467172384)

[Classifiers 3](#_Toc467172385)

[Evaluation Strategy 3](#_Toc467172386)

[Performance Results 3](#_Toc467172387)

[Top Features 3](#_Toc467172388)

[Discussion 3](#_Toc467172389)

[Interesting/Unexpected Results 3](#_Toc467172390)

[Contributions of Each Group Member 3](#_Toc467172391)

[Conclusion 3](#_Toc467172392)

[References 3](#_Toc467172393)

Regression Model for Predicting Movie Box Office Gross

Group Members: Ying Wu, Yingjuan Wu and Sahand Zeinali

# Task

<Describe the task here. What is the problem? What is the classification/regression problem? Why is it interesting?>

Currently There are some predict a movie’s worldwide box-office gross using relevant features. This is a useful feature to obtain since we can use this model to predict a movie’s commercial success during pre-production.

# Dataset

<Describe the dataset here.>

## Data source

<Where did you get the dataset from? Did you collect any additional data? Did you manually label any data?>

The data was collected from Keggle and no additional was collected nor was it manually labeled.

## Target variable

<What is the target variable?>

The selected target variable is gross for the project.

## Features

<What are the input features and how many are there?>

There are ten input features selected and the selected input features are critic reviews, duration of movie, face number in poster, genres, budget, country, content-rating, imdb score, number of director facebook likes, number of cast total facebook likes.

## Data size

<How many instances do you have in the data?>

The data includes over 5000 instances.

# Preprocessing

<Describe any preprocessing you have done for this dataset.>

# Visualization

## Target

<Provide statistics about the target variable; if classification, provide counts for each class; if regression, print mean and variance.>

<Provide a histogram of the target variable.>

## Features

<Visualize the features. For categorical variables, provide bar plots, for numerical features, provide histograms and statistics. If there are too many features, visualize and provide statistics for only 10 of them that you choose. If your data is text, print the list of most frequent words. If your data is raw (like images), describe how your model is handling it.>

# Evaluation

## Performance Measure

<Which performance measure you chose and why?>

There were two performance measures used. The performance measures used for the project were Mean Squared Error(MSE) and R^2 Score. MSE was used since it helps measure the average of the squares of the errors. It is a risk function corresponding to the expected value of the squared error loss. 'neg\_mean\_squared\_error' has been used as a scoring parameter in the cross-validation function to calculate the MSE scores. R^2 Score is used because it measures how well future instances are likely to be predicted by the model. 'r2\_score' is used as scoring parameter in the cross-validation function to calculate the R^2 scores. The cross validation performed is 10-fold.

## Classifiers

<Which classifiers and parameter settings did you try and why?>

In the baseline, the parameter selected is mean of target. The classifiers selected are Ordinary Least Squared, Ridge, Bayesian Ridge and Lasso. The parameter settings of ordinary least squares other than the default were Fit\_intercept set to False, N\_jobs set to 1, Normalize set to True and Copy\_X set to False. The parameter settings used in the Bayesian Ridge (except for the default) are when alpha\_1 is set to 1.e1 and alpha\_2 is set to 1.e2, lambda\_1 is set to 1.e^3 and lambda\_2 is set to 1.e4. additionality, when Fit\_intercept is set to False and compute\_score is set to True. In the Lasso model, Tol is set to 1 in all the parameter settings including the default. The additional parameter settings are alpha set to 0.5, normalize set to True and fit\_intercept set to False, precomputer is set to True, positive is set to True, warm\_start is set to True, copy\_X is set to False.

## Evaluation Strategy

<Did you do train-test split or cross-validation and why?>

Yes, cross validation was performed with 10 folds. The reason that cross validation was preffered was

## Performance Results

<Report your results, including baselines, using a table similar to the one on slide 7 of the project presentation template file.>

Table 1: Baseline and Ordinary Least Squares Model

|  |  |  |
| --- | --- | --- |
| **Model** | **Parameters** | **Performance** |
| Baseline | Mean of target | MSE=4.127e+15, R^2=-1.523 |
| Ordinary Least Squares | Default | MSE=2.073e+40, R^2=0.472 |
|  | Fit\_intercept=False | MSE=5.662e+37  R^2=-0.075 |
|  | N\_jobs=-1 | MSE=2.261e+40  R^2=0.472 |
|  | Normalize=True | MSE=1.853e+73  R^2=0.471 |
|  | Copy\_X=False | MSE=1.534e+40  R^2=0.472 |

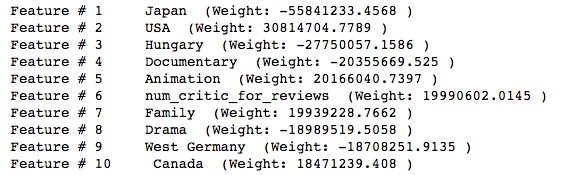
Table 2: Ridge, Bayesian Ridge and Lasso Models

|  |  |  |
| --- | --- | --- |
| **Model** | **Parameters** | **Performance** |
| Ridge | Default | MSE=2.466e+15, R^2=0.471 |
|  | Alpha=0.5 | MSE=2.497e+15, R^2=0.472 |
|  | Fit\_intercept=False | MSE=2.466e+15  R^2=0.471 |
|  | Solver=‘lsqr’ | MSE=2.446e+15  R^2=0.470 |
| Bayesian Ridge | Default | MSE=4.127e+15  R^2=9.300e-13 |
|  | alpha\_1=1.e1, alpha\_2=1.e2, lambda\_1=1.e3, lambda\_2=1.e4 | MSE=4.127e+15  R^2=9.351e-12 |
|  | Fit\_intercept=False | MSE=5.874e+15, R^2=-0.547 |
|  | Compute\_score=True | MSE=4.127e+15, R^2=9.300e-13 |
| Lasso | Tol=1 | MSE=2.533e+15  R^2=0.458 |
|  | Tol=1,alpha=0.5 | MSE=2.533e+15  R^2=0.458 |
|  | Tol=1,normalize=True | MSE=2.533e+15  R^2=0.458 |
|  | Tol=1,fit\_intercept=False | MSE=2.753e+15  R^2=0.391 |
|  | Tol=1,precomputer=True | MSE=2.533e+15  R^2=0.458 |
|  | Tol=1,positive=True | MSE=2.495e+15, R^2=0.441 |
|  | Tol=1,warm\_start=True | MSE=2.533e+15  R^2=0.458 |
|  | Tol=1,copy\_X=False | MSE=2.533e+15  R^2=0.178 |

## Top Features

<Present the top features with respect to your model.>

Based on the Ridge regression model that has been used, the top features are ranked in this order:



## Discussion

<Briefly discuss your results. Did the best classifier perform as well as you expected? If things did not work out as well, why do you think they did not work? Did one classifier perform much better (or worse) than others? And so on.>

Classifiers had different results but the Ridge model consistently had better results all around. Following Ridge, Lasso had more consistent results (both MSE AND R^2 Score). Thirdly, Bayesian Ridge had consistent MSE results but not R^2 Scores. Lastly it was Ordinary Least Squares. OLS lacked both consistency and positive results. The results for Ridge in general were positive since MSE was lower and R^2 was higher than expected.

# Interesting/Unexpected Results

<Discuss a few interesting/unexpected cases. See slide 9 of the project presentation template file.>

There were some unexpected results developed in the project. First interesting result was that amongst the top 10 features, there are 3 non-US country features (Japan, Hungary, West Germany) that have strong negative correlation with the target – movie gross. By looking closely into the data, it was found that the average gross for these three countries are all below 100,000 while the average gross for all movies are 45,608,461. Thus, the gross of movies produced in these three countries are significantly below the world-wide average, which explains why the three features are among the top negative features. Another interesting finding was that animation and family-theme movies achieve highest world-wide box office gross amongst all the genres. This is useful information for movie production companies since it suggests what themes will be more popular and commercially successful. Another interesting result is that if numerical features are not scaled, the performance of models are similar to scaled case. But the top six important features will be numerical features since our target, i.e. gross, is a very large number. And lastly it was observed how some changes could increase and decrease MSE and R^2 Score at the same time.

# Contributions of Each Group Member

<If you are working in a group, please discuss in detail what each member did for this project.>

The project has been developed in a group of three by Ying Wu, Yingjuan Wu and Sahand Zeinali. The proposal for the project was submitted by Ying Wu and in the second phase (Data exploration), Yingjuan and Sahand joined the team.

# Conclusion

<Provide concluding remarks.>

The results of the project caused in interesting results and added a good understanding to what the expectations are from a regression model. The progress made in the project followed with a good set of results since the best performing model was consistent and had better results than the expected. Some of the top features from the ridge model were unexpected but mostly, they were consistent with the expectation.

# References

<Provide references if you have any.>

Notes (erase these notes before you save and submit):

1. Once you are done editing, this file, update Table of Contents; you can simply click on Table of Contents and click on the button Update Table once it appears. Then, save this file as a pdf file.
2. Do not copy text from any website/paper. All the wording has to be your own.
3. Delete instructions (the text that appear inside <> above) and these notes before you save and submit.
4. Anywhere it says “discuss”, one or two paragraphs should be sufficient. Do not write anything unnecessary, but also do not omit any necessary details.