CS584 – machine learning

fall 2016

Regression Model for

Predicting Movie Box Office Gross

Group Members:

Ying Wu (A20370189)

Yingjuan Wu (A20326320)

Sahand Zeinali (A20318383)

Table of Contents

[Task 2](#_Toc467172373)

[Dataset 2](#_Toc467172374)

[Data source 2](#_Toc467172375)

[Target variable 2](#_Toc467172376)

[Features 2](#_Toc467172377)

[Data size 3](#_Toc467172378)

[Preprocessing 3](#_Toc467172379)

[Visualization 3](#_Toc467172380)

[Target 3](#_Toc467172381)

[Features 4](#_Toc467172382)

[Evaluation 6](#_Toc467172383)

[Performance Measure 6](#_Toc467172384)

[Classifiers 6](#_Toc467172385)

[Evaluation Strategy 7](#_Toc467172386)

[Performance Results 7](#_Toc467172387)

[Top Features 8](#_Toc467172388)

[Discussion 8](#_Toc467172389)

[Interesting/Unexpected Results 8](#_Toc467172390)

[Contributions of Each Group Member 8](#_Toc467172391)

[Conclusion 9](#_Toc467172392)

[References 9](#_Toc467172393)

**Regression Model for Predicting Movie Box Office Gross**

Group Members: Ying Wu, Yingjuan Wu, Sahand Zeinali

# Task

Is there a way to determine the greatness of a movie before it is released? How do we decide if a movie is worth watching instead of relying on instincts or critics? To answer this question, we explored a Kaggle dataset [1] which scrapes more than 5000 movies from IMDB website. Each movie has 28 features including its box office gross. In this project, we are trying to build a regression model which could predict a movie’s box office gross using relevant features available before a movie is released.

# Dataset

This dataset is available online [1]. It contains 28 features for 5043 movies spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors or actresses. Figure 1 shows the data for the first 20 movies in this dataset.

|  |
| --- |
|  |

Fig. 1 Screen shot of the first 20 movies in raw dataset

## Data source

We used this raw dataset on Kaggle directly, which scraped from IMDB website. We didn’t collect additional data or manually label any data since this dataset is large enough with a lot of information to be processed.

## Target variable

Our target variable is the box office gross of movies. We intend to predict this value based on regression method.

## Features

Apart from gross, which we selected as target variable, there are other 27 features in the datasets. These features are: movie\_title, color, num\_critic\_for\_reviews, movie\_facebook\_likes, duration, director\_name, director\_facebook\_likes, actor\_3\_name, actor\_3\_facebook\_likes, actor\_2\_name, actor\_2\_facebook\_likes, actor\_1\_name, actor\_1\_facebook\_likes, genres, num\_voted\_users, cast\_total\_facebook\_likes, facenumber\_in\_poster, plot\_keywords, movie\_imdb\_link, num\_user\_for\_reviews, language, country, content\_rating, budget, title\_year, mdb\_score, aspect\_ratio.

## Data size

There are 5403 movies and each movie has 28 features.

# Preprocessing

To preprocess the dataset, we did the following five things:

1. Remove redundant data. All duplicated movies are deleted.
2. Remove irrelevant features.

Features collected after a movie is released are removed, like aspect\_ratio and num\_user\_for\_reviews.

Features that are hard to be interpreted are removed, including movie\_title, movie\_imdb\_link, plot\_keywords, color, director\_name and actor\_name .

Features that are repetitive others are removed. For example, actor\_facebook\_likes is removed because we have cast\_total\_facebook\_likes, language is removed because country represents the same thing with more classifications.

1. Remove movies which were released before 1960 since they are too old and their features may be inaccurate.
2. Deal with missing values.

For movies that have missing gross or director\_facebook\_likes, they are removed.

For movies that have missing num\_critic\_for\_reviews, duration, facenumber\_in\_poster or budget, their missing value is replaced with the mean of all other movies.

For movies that have missing content-rating, their content-rating is set to “Not rated”.

1. Binarize categorical feature: Genres, Country and Content-rating.

After preprocessing, we have a clean dataset with 3321 rows and 87 columns.

# Visualization

## Target

|  |
| --- |
| system:Users:yingwu:Desktop:Screen Shot 2016-11-22 at 3.04.12 PM.png |

Fig.2 Histogram of gross

## Features

Seven Numeric features as shown in Fig. 3 to Fig. 9:

|  |  |
| --- | --- |
| Fig.3 Histogram of num\_critic\_for\_reviews  Mean is 150.954, Variance is 12831.002 | Fig.4 Histogram of duration  Mean is 107.637, Variance is 403.224 |
| Fig.5 Histogram of director\_facebook\_likes  Mean is 910.898, Variance is 10532433.827 | Fig.6 Histogram of cast\_total\_facebook\_likes  Mean is 10634.358, Variance is 368733196.877 |
| Fig.7 Histogram of facenumber\_in\_poster  Mean is 1.367, Variance is 3.867 | Fig.8 Histogram of budget  Mean is 40389715.08, Variance is 7.411e15 |
| Fig.9 Histogram of imdb\_score  Mean is 6.397, Variance is 1.0732 |  |

We have three Categorical features as shown in Figure 10 to Figure 12:

|  |  |
| --- | --- |
| system:Users:yingwu:Desktop:Screen Shot 2016-11-27 at 7.12.12 PM.png  Fig.10 Bar Plot of Countries | Fig.11 Bar Plot of Content-ratings |
| system:Users:yingwu:Desktop:Screen Shot 2016-11-27 at 7.12.19 PM.png  Fig.12 Bar Plot of Genres |  |

# Evaluation

## Performance Measure

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. The other performance measure used was R^2 Score, because it measures how well future instances are likely to be predicted by the model.

## Classifiers

We tried 5 different models, including Dummy Regressor (baseline), Ordinary Least Squares Regression, Ridge Regression, Bayesian Ridge Regression, and Lasso Regression. These models were used because they are commonly used for solving regression problems. Moreover, these models include different penalty and by comparing the performance of these models on predicting the target, we can identify the best model and parameter setting for our problem. Also, we chose important parameters for each model and compared model performance by setting different values to each parameter. Finally, the setting that yielded highest model performance was chosen for each parameter and listed below.

In the baseline, the parameter selected is mean of target.

The parameter settings (5 different 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, respectively.

For Ridge, the parameter settings (4 different settings) were default, alpha set to 0.5, Fit\_intercept set to False, solver set to ‘lsqr’, respectively.

The parameter settings (4 different settings) used in the Bayesian Ridge were default, 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, since otherwise there would be convergence warnings. The additional parameter settings (8 different 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

We used 10-fold cross validation. The cross validation split the data into 10 even subsets, and in each iteration, it used 1 subset as testing set and 9 subsets as training set. By using cross validation, we used all instances for both training and testing, which helped us get balanced results and more accurate performance measures.

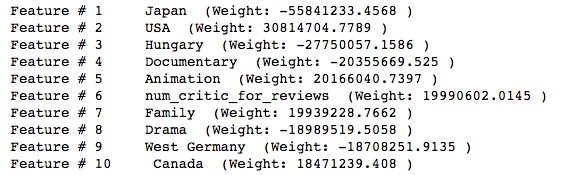
## Performance Results

Table 1. Model Selection Results

|  |  |  |
| --- | --- | --- |
| 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 |
| 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

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



## Discussion

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, Ordinary Least Squares model had the worst performance among all 5 models. 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

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. By contrast, drama (as a genre) is one of the top negative features. This is possibly due to the fact that many non-US movies that have low gross are drama movies, which leads to a low average gross for all drama movies.

Moreover, it is also found 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

|  |
| --- |
| Ying Wu:  Data Exploration:  Project description  Load data, print data shape  Visualize target variable  Evaluation:  Load processed data, Print data shape |
| Performance measures description |
| Report performance of baselines |
| Build linear regression model using different parameter settings and report MSE, R2  Presentation and Report:  Prepare presentation slides  Visualization and conclusion sections of the report |
|  |
|  |
| Yingjuan Wu:  Data Exploration:  Binarize all categorical features, i.e. Genres, Country and Content-rating  Visualize five numerical features  Print Mean and Variance of all numerical features  Evaluation: |
| Continuous features scaling |
| Build ridge regression model with different parameters and report MSE, R2 |
| Build Bayesian ridge regression model with different parameters and report MSE, R2 |
| Report top features and their weights  Presentation and Report:  Add two slides in Presentation.  Complete the first four sections in the report, i.e. Task, Dataset, Preprocessing, and Visualization. |
|  |
|  |
| Sahand Zeinali:  Data Exploration:  Visualize target variable  Visualize five features (numerical and non numerical)  Visulize using bar plots and histograms  Evaluation: |
| Regression model use research |
| Build Lasso regression model with different parameters and report MSE, R2 |

Presentation and Report:

Added slides into the presentation regarding the formatting.

Evaluation sections of the reports were added.

# Conclusion

In this project, we used 5 regression models to predict the movie box office gross and compared their performance. The Ridge model with parameter solver set to ‘lsqr’ achieved the highest performance with MSE of 2.446e+15 and R^2 of 0.47. The MSE is relatively large mainly due to the fact that the range of our target variable is large, from below 100 dollars to hundreds of million dollars. The R^2 score was around 0.5 meaning that the model has moderate performance when predicting the target.

Among all the input variables, country (categorical), genre (categorical), and number of critic reviews (continuous) are most important features for predicting movie gross, because they were the top features of the best model.

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

[1] Kaggle dataset. Available [online] <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

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.