****

**Predicting Movie rating and Box Office Gross by PCA and LR model**

Project by: Yunfeng Liu, Yash Naik, Erdong Xia

Instructor: Prof A. Gandomi

**Contents**

Abstract 3

Background Introduction 3

Data Introduction 4

Data Processing 4

Methods Introduction 6

Dimension Reduction 7

Regression Analysis 9

Conclusion 15

Future Research 16

References 16

Appendix 17

**Abstract**

The purpose of this project is to create two prediction models, among which the rating model will be able to predict the movie rating on IMDB website while the box office gross model will be able to predict the gross revenue the movie can generate in American market. The influencing factors for the two models in consideration are duration of the movie, the budget of the movie, the year of the movie release, the actors and directors involved and movie genres. Before multivariate regression analysis, principal component analysis is conducted as dimension reduction method and cross validation method is conducted for testing and model selection.

Key word: Movie Rating

Box Office Gross

Multivariate Regression Analysis

Principal Component Analysis

Cross Validation

**Background Introduction**

With the prosperity of entertainment industry, movie has always been one of the most popular topics among people in leisure time and there are mainly two aspects for a definition of good movies. On the one hand, it is a high rating on movie websites with millions of reviewers ranging from professional commenter to common audience. On the other hand, it is a great success in box office in terms of earning money. Both aspects contribute to an excellent movie and this project is mainly aimed at the following 2 questions:

1. What are the influencing factors of movie rating score and box office gross?

2. How to predict the rating score and box office gross of a movie with the influencing factors?

In this project, the research will be processed in the following steps.

1. Data collection and processing

2. Principal component analysis for dimension reduction

3. Multivariate regression for prediction model in which stepwise will be the variable selection method and cross validation will be the model selection method

Conclusion and future research will be at the end of the paper with the SAS code for analysis attached below.

**Database Introduction**

The project data source is from [www.kaggle.com](http://www.kaggle.com) containing over 5,000 movies data in mainland America during over past 20 years with 16 variables as follows:

Variable Label: descriptions (variable types).

1. Movie title: the title of the movies (strings).

2. Movie year: the year of the movies published (integrals from 1990 to 2016).

3. Director name: the name of director (strings).

4. Director Facebook likes: the Facebook likes of the director (integrals).

5. Actor 1 name: the name of actor 1 (strings).

6. Actor 1 Facebook likes: the Facebook likes of actor 1 (integrals).

7. Actor 2 name: the name of actor 2 (strings).

8. Actor 2 Facebook likes: the Facebook likes of actor 2 (integrals).

9. Actor 3 name: the name of actor 3 (strings).

10. Actor 3 Facebook likes: the Facebook likes of actor 3 (integrals).

11. Budget: the total budgets to produce the movies in dollars (integrals).

12. Gross: the total revenue of box office in mainland America in dollars (integrals).

13. Reviewer number: the number of reviewers under the movie on IMDB website (integrals).

14. IMDB score: the score of the movie on IMDB website (decimal 1 from 0.0 to 10.0).

15. Movie genres: the genres of the movie belonging to. A movie may belong to 1 to 5 different genres. The total number of genres is 17, which are action, adventure, family, animation, comedy, drama, biography, mystery, horror, musical, crime, documentary, fantasy, sci-fi, romance, western, thriller (strings).

16. Movie duration: the duration of movie in minutes (integrals).

**Data Processing**

**1. Delete null and missing values from raw data (5500 rows-3116 rows).**

There are missing values in the raw dataset due to various kinds of reasons such as unknown information for Facebook likes for directors and actors because some of them did not have Facebook accounts, corrupted statistics of budgets and gross due to long time period, etc.

**2. Select movies with sufficient reviewer numbers (3116 rows-1558 rows).**

The score on IMDB website may go into extremes due to deficiency of reviewers, which may hinder the preciseness of the prediction model. As a result, movies below the medium reviewer number are deleted and the movies remained all have a reviewer number greater than 128.

**3. Compute the total actor Facebook likes by the three actors Facebook likes.**

Three actors are treated equally in contributing to the actor level. The actor Facebook likes = Actor 1 Facebook likes + Actor 2 Facebook likes + Actor 3 Facebook likes.

**4. Transform directors and actors Facebook likes into score level from 1 to 5.**

The exact number of Facebook likes of director and actors are redundant for analysis and may cause serious overfitting problem. As a result, a transformation work is conducted with the level of their Facebook likes among their fellows. The exact corresponding table is in Table 1:

**Table 1**

|  |  |
| --- | --- |
| Score | Level |
| 5 | Top 5 % |
| 4 | 85%-95% |
| 3 | 70%-85% |
| 2 | 50%-70% |
| 1 | Last 50% |

**5. Transform string variable of genres into 17 virtual variables representing 17 specific categories.**

To analyze the influence of movie genres on score and gross, it is essential to transform string variable to virtual variable. The virtual variable equals to 1 if the movie belongs to the genre it represents and equals to 0 if the movie does not belong to the genre it represents.

**6. For cross validation, divide the population into four testing datasets evenly and set training dataset according to different testing dataset.**

The logic of cross validation is shown in Table 2, in which the green cells are testing datasets and yellow cells are training datasets. The models from each training dataset will be validated in the corresponding testing datasets and by examining the difference of the predicting performance between these four parts, the overfitting problem will get exposed which is beneficial to the model selection process.

**Table 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train 1** | Test 1 | Test 2 | Test 3 | Test 4 |
| **Train 2** | Test 1 | Test 2 | Test 3 | Test 4 |
| **Train 3** | Test 1 | Test 2 | Test 3 | Test 4 |
| **Train 4** | Test 1 | Test 2 | Test 3 | Test 4 |

**Methods Introduction**

**1. Multivariate Linear Regression Model.**

Method Theory:

Multivariate linear regression model is a generalization of linear regression by considering more than one independent variable. The general model for multivariate linear regression is

{\displaystyle Y\_{i}=\beta \_{0}+\beta \_{1}X\_{i1}+\beta \_{2}X\_{i2}+\ldots +\beta \_{p}X\_{ip}+\epsilon \_{i}.}

In the above formula, Y is the dependent variable, are independent variables, are the coefficient estimates of s and is the intercept.

Project Example:

Dependent Variable: 1. IMDB score

2. Gross

Independent Variables: 1. Director Score

2. Actors Score

3. Duration

4. Budget

5. Movie Year

6-22. 17 Virtual Variables for Movie Genres

**2. Principal Component Analysis (PCA).**

Method Theory:

Principal Component Analysis is an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables.

Project Example:

Among the independent variables, the first five variables can be supposed to be uncorrelated due to correlation coefficient matrix, but the 6-22 variables which are Movie Genres Virtual Variables are highly collinearly and contain too many columns describing the same movie attribute of genre, which is fatal for model fitting.

PCA can transform these variables into uncorrelated factors and appropriate number of principal components can be selected among these factors to realize dimension reduction.

**Dimension Reduction**

**1. PCA Preparation**

Select 17 virtual variables representing movie genres as analysis objectives, transform them into uncorrelated factors and reduce the dimension by PCA

**2. PCA Processing**

Table 3 is the eigenvalues of the correlation matrix which indicates that PCA process transforms the original 17 variables into 17 uncorrelated factors with descending eigenvalue and ascending cumulative variance explained proportion.

**Table 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 2.72633669 | 0.48833478 | 0.1604 | 0.1604 |
| **2** | 2.23800192 | 0.82505004 | 0.1316 | 0.292 |
| **3** | 1.41295188 | 0.1264659 | 0.0831 | 0.3751 |
| **4** | 1.28648597 | 0.07129985 | 0.0757 | 0.4508 |
| **5** | 1.21518613 | 0.16262036 | 0.0715 | 0.5223 |
| **6** | 1.05256577 | 0.06203863 | 0.0619 | 0.5842 |
| **7** | 0.99052714 | 0.05436379 | 0.0583 | 0.6425 |
| **8** | 0.93616336 | 0.04746732 | 0.0551 | 0.6975 |
| **9** | 0.88869603 | 0.01982364 | 0.0523 | 0.7498 |
| **10** | 0.8688724 | 0.21195669 | 0.0511 | 0.8009 |
| **11** | 0.6569157 | 0.05279898 | 0.0386 | 0.8396 |
| **12** | 0.60411672 | 0.01705175 | 0.0355 | 0.8751 |
| **13** | 0.58706498 | 0.14446462 | 0.0345 | 0.9096 |
| **14** | 0.44260035 | 0.04430786 | 0.026 | 0.9357 |
| **15** | 0.39829249 | 0.04440597 | 0.0234 | 0.9591 |
| **16** | 0.35388652 | 0.01255059 | 0.0208 | 0.9799 |
| **17** | 0.34133594 |  | 0.0201 | 1 |

**Table 4**

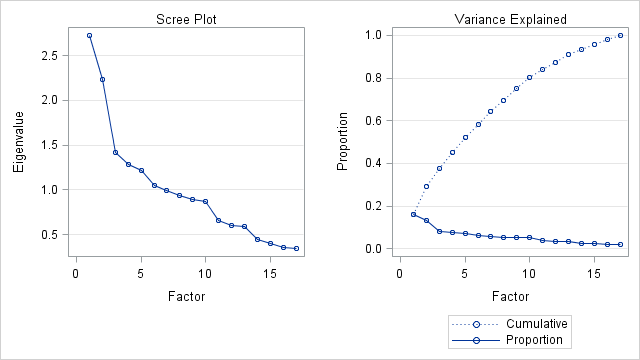
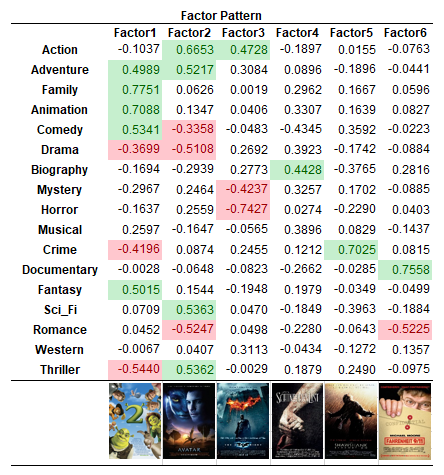


Table 4 is the Scree Plot and Variance Explained Plot corresponding to it. From scree plot, first 6 factors are selected as principal components having eigenvalues greater than 1.0, which form a steep tendency on left and flat tendency on right. The cumulative variance explained by the 6 factors is about 60%, which is enough for the total information in movie genres.

**3. PCA Results Analysis**

Table 5 is the Factor Pattern Table which indicates the coefficients of each genre contributing to each factor.

**Table 5**

****

From Table 5, the positive indicators of each factor are labeled as green cells while negative indicators of each factor are labeled as red cells. The realistic indications of the 6 principal components are concluded as follows:

PC1: Family animation movies with comedy, fantasy and adventure while without thriller, crime or drama.

PC2: Action movies with sci-fi, thriller and adventure while without drama or comedy.

PC3: Action movies without horror or mystery.

PC4: Biography movies.

PC5: Crime movies.

PC6: Documentary movies without romance.

**4. PCA to LR**

1. Compute the coefficients of each movie under the 6 principal components from factor pattern table.

2. Replace the original 17 virtual variables representing movie genres by the 6 principal components.

3. The LR models are updated as follows:

Dependent Variable: 1. IMDB score

2. Gross

Independent Variables: 1. Director Score

2. Actors Score

3. Duration

4. Budget

5. Movie Year

6-11. 6 Principal Components for Movie Genres named as PC1 to PC6

**Regression Analysis**

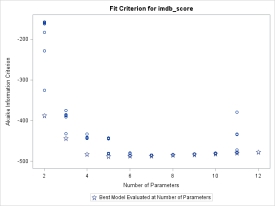
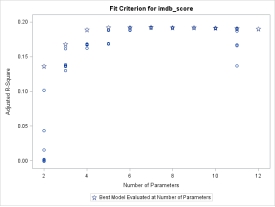
**1. Variable selection**

After PCA, 11 variables are remained, and variable selection is performed along with regression analysis. For the variable selection method, stepwise is employed to select the independent variables which are highly correlated with either rating or gross.

1.1 Rating model

For rating model, the stepwise results are showed in Table 6

**Table 6**



As the figures showed,4 to 6 parameters which includes intercept are the three combinations of variables which have least Akaike information criterion (AIC) and the highest adjust R square. And the best of them is showed in Table 7.

**Table 7**

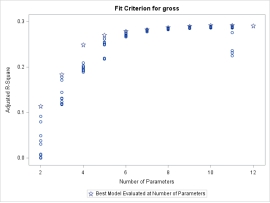
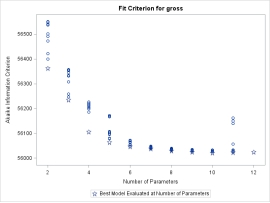
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Parameter Estimates of Rating Model | | | | | | |
| Variable | DF | Parameter | Standard | t Value | Pr > |t| | Standardized |
| Estimate | Error | Estimate |
| Intercept | 1 | 41.94717 | 5.43902 | 7.71 | <.0001 | 0 |
| director\_star | 1 | 0.13974 | 0.02038 | 6.86 | <.0001 | 0.17509 |
| actors\_star | 1 | 0.05001 | 0.01854 | 2.7 | 0.0071 | 0.06428 |
| duration | 1 | 0.01275 | 0.00126 | 10.12 | <.0001 | 0.25958 |
| title\_year | 1 | -0.01848 | 0.0027 | -6.84 | <.0001 | -0.16053 |

Thus, from the results of stepwise, the best combination including four variables: director star, actors star, duration and title year, along with the second-best combination that have five variables: director star, actors star, duration, title year and budget, and the third best combination that have three variables: director star, duration and title year, are chosen for cross-validation.

1.2 Box office gross model

For box office gross model, the stepwise results are showed in Table 8.

**Table 8**



In the case of gross revenue, 9 to 11 parameters which include intercept are the best three combinations due to the AIC value and adjust R square. The best model with 9 independent variables in is shown in Table 9.

**Table 9**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Parameter Estimates of Box Office Gross Model | | | | | | |
| Variable | DF | Parameter | Standard | t Value | Pr > |t| | Standardized |
| Estimate | Error | Estimate |
| Intercept | 1 | 905658009 | 409360130 | 2.21 | 0.0271 | 0 |
| director\_star | 1 | 3955034 | 1534556 | 2.58 | 0.01 | 0.06184 |
| actors\_star | 1 | 8184642 | 1400512 | 5.84 | <.0001 | 0.13128 |
| budget | 1 | 0.22634 | 0.02054 | 11.02 | <.0001 | 0.24388 |
| duration | 1 | 227612 | 95540 | 2.38 | 0.0173 | 0.05783 |
| PC1 | 1 | 19734251 | 1639592 | 12.04 | <.0001 | 0.25931 |
| PC2 | 1 | 18745529 | 1636019 | 11.46 | <.0001 | 0.24632 |
| PC3 | 1 | -5333250 | 1629191 | -3.27 | 0.0011 | -0.07008 |
| PC5 | 1 | -5107989 | 1625136 | -3.14 | 0.0017 | -0.06712 |
| title\_year | 1 | -447572 | 203386 | -2.2 | 0.0279 | -0.04853 |

Thus, from the table, the best combination including nine variables: director star, actors star, budget, duration, PC 1, PC 2, PC 3, PC 5 and title year, along with the second best combination that have eight variables: director star, actors star, budget, duration, PC 1 PC 2 PC 3 and PC 5, and the third best combination that have ten variables: director star, actors star, budget, duration, PC 1, PC 2, PC 3, PC 5, title year and PC 4, are chose for cross-validation.

**2. Model Selection**

To eliminate overfitting as much as we can, cross-validation are used to select the best model based on the results of prediction errors. With the testing – training groups and the combinations of variables selected by stepwise which are renamed by model 1, model 2 and model 3, each group is trained and tested for each model, then their mean square errors (MSE) are computed and compared for evaluation. The model which has the least average MSE will be considered as the best model.

2.1 Rating model

For rating model, the cross-validation results are showed in Table 10.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 10**  MSE of rating model | | | | |
|  | test1 | test2 | test3 | test4 |
| model1 | 0.3168 | 0.3487 | 0.3872 | 0.4128 |
| model2 | 0.3171 | 0.3627 | 0.3872 | 0.4128 |
| model3 | 0.3195 | 0.3492 | 0.389 | 0.4148 |

From the table, it is obviously that model 1 have the least MSE. Thus, model 1 which has four independent variables is selected to be the rating model for regression analysis.

2.2 Box office gross model

For gross model, the cross-validation results are showed in Table 11.

**Table 11**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MSE of gross model | | | | |  |
|  | test1 | test2 | test3 | test4 | avg |
| model1 | 2.278 × 1015 | 4.017 × 1015 | 1.926 × 1015 | 1.635 × 1015 | 2.464 × 1015 |
| model2 | 2.269 × 1015 | 3.960 × 1015 | 1.928 × 1015 | 1.650 × 1015 | 2.452 × 1015 |
| model3 | 2.277 × 1015 | 4.011 × 1015 | 1.925 × 1015 | 1.636 × 1015 | 2.462 × 1015 |

After calculating the average MSE of each model, we find that model 2 is the best one which has the least average MSE. Thus, model 2 with its nine independent variables is selected to be the box office gross model for regression analysis.

**3. Regression results**

After variable selection and model selection, regression analysis is performed with whole samples and the models we have selected.

3.1 Rating model

Table 12 shows the parameter estimate results of rating model.

**Table 12**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Parameter Estimates of Rating Model | | | | | | |
| Variable | DF | Parameter | Standard | t Value | Pr > |t| | Standardized |
| Estimate | Error | Estimate |
| Intercept | 1 | 41.94717 | 5.43902 | 7.71 | <.0001 | 0 |
| duration | 1 | 0.01275 | 0.00126 | 10.12 | <.0001 | 0.25958 |
| director\_star | 1 | 0.13974 | 0.02038 | 6.86 | <.0001 | 0.17509 |
| actors\_star | 1 | 0.05001 | 0.01854 | 2.7 | 0.0071 | 0.06428 |
| title\_year | 1 | -0.01848 | 0.0027 | -6.84 | <.0001 | -0.16053 |

The variance analysis result is shown in Table 13.

**Table 13**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Analysis of Variance of Rating Model | | | | | |
| Source | DF | Sum of | Mean | F Value | Pr > F |
| Squares | Square |
| Model | 4 | 272.31162 | 68.07791 | 93.41 | <.0001 |
| Error | 1553 | 1131.87985 | 0.72883 |  |  |
| Corrected Total | 1557 | 1404.19148 |  |  |  |
| Root MSE | 0.85372 | R-Square | 0.1939 |  |  |
| Dependent Mean | 6.65815 | Adj R-Sq | 0.1919 |  |  |
| Coeff Var | 12.82215 |  |  |  |  |

From the result of standardized estimate coefficient of each variable, it showed that duration is the most important variable to rating score predication of a new movie. And, both director star and title year are important, though the former one has a positive coefficient and the later one has a negative coefficient. However, title year is not a parameter we can easily change when a movie is about to come out. Thus, to make a movie harvest a high rating score from IMDB reviewers, a long duration with a famous director is the best choice.

3.2 Box office gross model

Table 14 shows the parameter estimate results of gross model.

**Table 14**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Parameter Estimates of Box Office Gross Model | | | | | | |
| Variable | DF | Parameter | Standard | t Value | Pr > |t| | Standardized |
| Estimate | Error | Estimate |
| Intercept | 1 | 5071218 | 9685173 | 0.52 | 0.6006 | 0 |
| PC1 | 1 | 19734643 | 1641624 | 12.02 | <.0001 | 0.25932 |
| PC2 | 1 | 18797712 | 1637875 | 11.48 | <.0001 | 0.24701 |
| PC3 | 1 | -5250693 | 1630778 | -3.22 | 0.0013 | -0.069 |
| PC5 | 1 | -5160318 | 1626976 | -3.17 | 0.0015 | -0.06781 |
| duration | 1 | 254446 | 94876 | 2.68 | 0.0074 | 0.06465 |
| director\_star | 1 | 4300439 | 1528400 | 2.81 | 0.005 | 0.06724 |
| actors\_star | 1 | 7635889 | 1379842 | 5.53 | <.0001 | 0.12248 |
| budget | 1 | 0.22329 | 0.02052 | 10.88 | <.0001 | 0.24059 |

The variance analysis result is showed in the Table 15

**Table 15**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Analysis of Variance of Box Office Gross Model | | | | | |
| Source | DF | Sum of | Mean | F Value | Pr > F |
| Squares | Square |
| Model | 8 | 2.64 × 1018 | 3.30 × 1017 | 80.3 | <.0001 |
| Error | 1549 | 6.37 × 1018 | 4.11 × 1015 |  |  |
| Corrected Total | 1557 | 9.02 × 1018 |  |  |  |
| Root MSE | 64147174 | R-Square | 0.2931 |  |  |
| Dependent Mean | 68575445 | Adj R-Sq | 0.2895 |  |  |
| Coeff Var | 93.54248 |  |  |  |  |

From the results of standardized coefficient of each independent variable, it indicates that PC 1 is the most influential variable of box office gross. Besides, both PC 2 and budget are also important to gross predication. In addition, the negative coefficient of PC 3 and PC 5 should also be attached attention. Plus, different from rating model, actors star’s coefficient is higher than director star’s. Thus, from the coefficients of these independent variables, it can be concluded that to make a movie earn more money, a big budget with the genre of PC 1 or PC 2 is what the movie maker should consider first, also famous actors are more functional than famous directors to collect money in American market.

**Conclusion**

1. From movie rating model, a long movie shot by a famous director is likely to earn a high rating score on reviewer websites and the genres of the movie are irrelevant

2. From movie gross model, a big-budget movie with the genres of   family-animation or action-sci-fi is likely to earn a successful box office. Plus, in terms of earning money, actors are more important than director.

**Future Research**

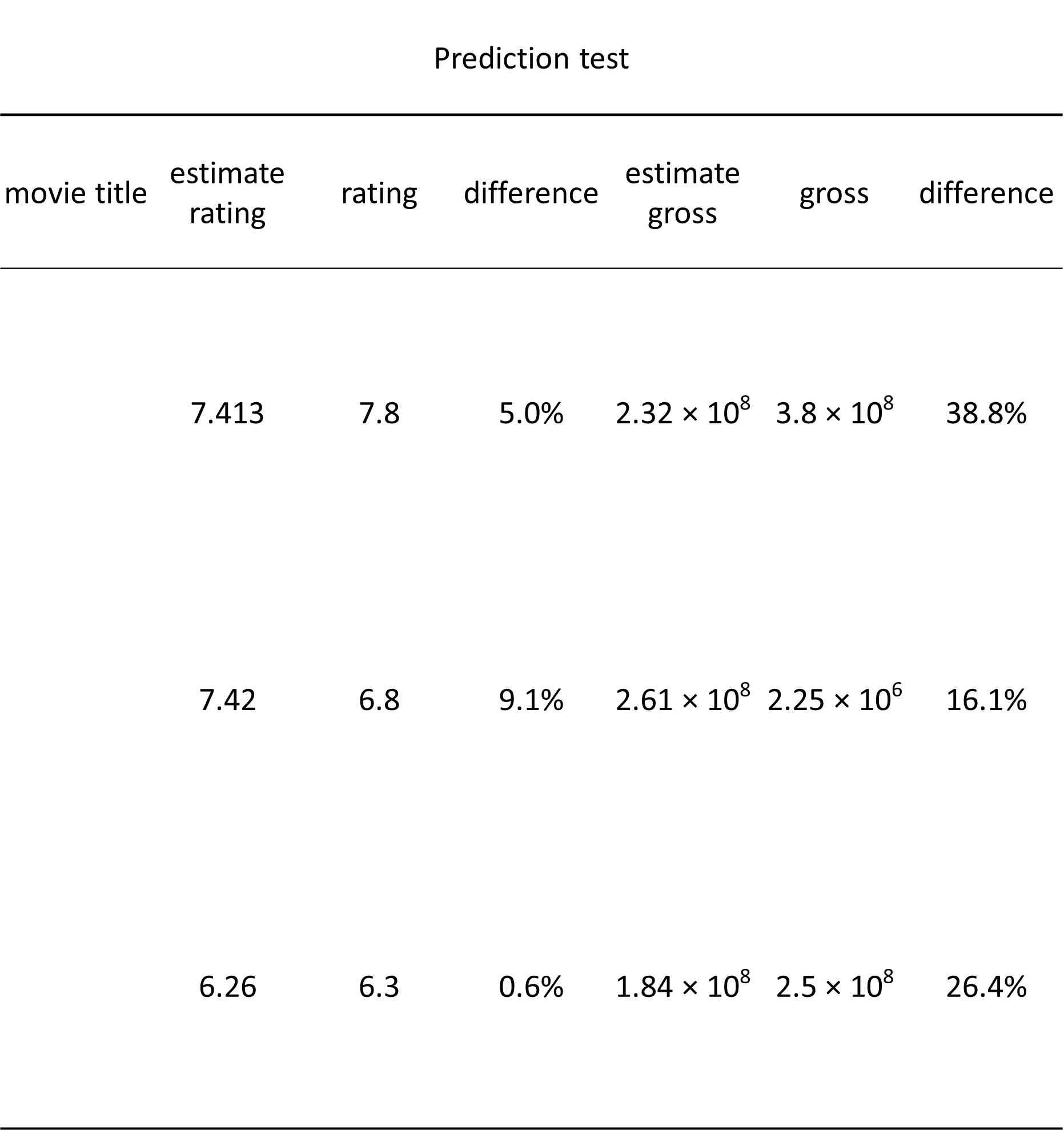
According to the project conclusion, a future testing research is conducted among three new movies released in first half year of 2017 and the results and difference is shown in Table 16.

**Table 16**

A picture containing book, text

Description generated with very high confidenceA close up of a mans face

Description generated with very high confidenceA close up of a womans face

Description generated with very high confidence

**References**

1. www.Kaggle.com for dataset retrieval.

2. Practical Multivariate Analysis, Fifth Edition book by A. Afifi, S. May, V. Clark.

3. SAS Help. www.sas.com/documentation

4. www.imdb.com for collecting movie ratings in future research.

5. www.facebook.com for Social media data.

**Appendix**

SAS code for the project

\*data import as movie;

PROC IMPORT OUT= WORK.movie

DATAFILE= "C:\Users\A08399\Desktop\movie.txt"

DBMS=TAB REPLACE;

GETNAMES=YES;

DATAROW=2;

RUN;

\*descriptive statistic analysis for each variable;

proc univariate data=movie;

var num\_critic\_for\_reviews duration director\_facebook\_likes actor\_1\_facebook\_likes actor\_2\_facebook\_likes actor\_3\_facebook\_likes;

run;

\*delete movie with few review numbers;

data movie1;

set movie;

if num\_critic\_for\_reviews>=128;

three\_actors\_facebook\_likes=actor\_1\_facebook\_likes+actor\_2\_facebook\_likes+actor\_3\_facebook\_likes;

run;

\*transform facebook likes of directors and actors to star ranging from 1 to 5;

proc univariate data=movie1;

var director\_facebook\_likes;

output pctlpts=50 70 85 95 pctlpre=pwid;

run;

proc univariate data=movie1;

var three\_actors\_facebook\_likes;

output pctlpts=50 70 85 95 pctlpre=pwid;

run;

data movie2;

set movie1;

if director\_facebook\_likes<=157 then director\_star=1;

else if director\_facebook\_likes>157 and director\_facebook\_likes<=350 then director\_star=2;

else if director\_facebook\_likes>350 and director\_facebook\_likes<=670 then director\_star=3;

else if director\_facebook\_likes>670 and director\_facebook\_likes<=14000 then director\_star=4;

else director\_star=5;

run;

data movie3;

set movie2;

if three\_actors\_facebook\_likes<=6101 then actors\_star=1;

else if three\_actors\_facebook\_likes>6101 and three\_actors\_facebook\_likes<=15974 then actors\_star=2;

else if three\_actors\_facebook\_likes>15974 and three\_actors\_facebook\_likes<=26722 then actors\_star=3;

else if three\_actors\_facebook\_likes>26722 and three\_actors\_facebook\_likes<=44000 then actors\_star=4;

else actors\_star=5;

run;

\*principal components analysis for genres of movies;

proc factor data=movie3 preplot plot scree nfactors=6 rotate=varimax out=movie4;

title "Factor Analysis for Catagory";

var Action Adventure Family Animation Comedy Drama

Biography Mystery Horror Musical Crime Documentary

Fantasy Sci\_Fi Romance Western Thriller;

run;

\*divide population into 4 subsets for cross validation;

proc surveyselect data=movie4 method =srs n=390 out=test1 seed = 25070419;

run ;

proc sql ;

create table movie5 as

select \*

from movie4

where movie4.movie\_title not in(select test1.movie\_title from test1);

quit;

proc surveyselect data=movie5 method =srs n=390 out=test2 seed = 25070419;

run ;

proc sql ;

create table movie6 as

select \*

from movie5

where movie5.movie\_title not in(select test2.movie\_title from test2);

quit;

proc surveyselect data=movie6 method =srs n=389 out=test3 seed = 25070419;

run ;

proc sql ;

create table test4 as

select \*

from movie6

where movie6.movie\_title not in(select test3.movie\_title from test3);

quit;

data train1;

set test2 test3 test4;

data train2;

set test1 test3 test4;

data train3;

set test1 test2 test3;

data train4;

set test1 test2 test3;

run;

\*conduct regression model for rating for each training set;

proc reg data=train1 outest=regout1;

title "regression model for movie score (test1)";

model imdb\_score=duration director\_star title\_year actors\_star budget/vif stb;

run;

quit;

proc score data=test1 score=regout1 out=testout1

type=parms;

var duration director\_star title\_year actors\_star budget;

run;

data testout1;

set testout1;

deviation=model1-imdb\_score;

deviation\_square=deviation\*deviation;

run;

proc sql;

select (sum(deviation\_square)/390)\*\*1/2 as mse1

from testout1;

quit;

proc reg data=train2 outest=regout2;

title "regression model for movie score (test2)";

model imdb\_score=duration director\_star title\_year actors\_star budget/vif stb;

run;

quit;

proc score data=test2 score=regout2 out=testout2

type=parms;

var duration director\_star title\_year actors\_star budget;

run;

data testout2;

set testout2;

deviation=model1-imdb\_score;

deviation\_square=deviation\*deviation;

run;

proc sql;

select (sum(deviation\_square)/390)\*\*1/2 as mse2

from testout2;

quit;

proc reg data=train3 outest=regout3;

title "regression model for movie score (test3)";

model imdb\_score=duration director\_star title\_year actors\_star budget/vif stb;

run;

quit;

proc score data=test3 score=regout3 out=testout3

type=parms;

var duration director\_star title\_year actors\_star budget;

run;

data testout3;

set testout3;

deviation=model1-imdb\_score;

deviation\_square=deviation\*deviation;

run;

proc sql;

select (sum(deviation\_square)/389)\*\*1/2 as mse3

from testout3;

quit;

proc reg data=train4 outest=regout4;

title "regression model for movie score (test4)";

model imdb\_score=duration director\_star title\_year actors\_star budget/vif stb;

run;

quit;

proc score data=test4 score=regout4 out=testout4

type=parms;

var duration director\_star title\_year actors\_star budget;

run;

data testout4;

set testout4;

deviation=model1-imdb\_score;

deviation\_square=deviation\*deviation;

run;

proc sql;

select (sum(deviation\_square)/389)\*\*1/2 as mse4

from testout4;

quit;

proc reg data=movie4;

title "regression model for movie score (all)";

model gross=factor1 factor2 factor3 factor4 factor5 factor6 duration director\_star actors\_star budget title\_year/selection=stepwise vif stb;

run;

quit;

proc score data=movie4 score=regout out=testout

type=parms;

var factor1 factor2 factor3 factor4 factor5 factor6 duration director\_star actors\_star budget title\_year;

run;

\*conduct regression model for gross for each training set;

proc reg data=train1 outest=regout5;

title "regression model for movie gross (test1)";

model gross=factor1 factor2 factor3 factor5 duration director\_star actors\_star budget/ vif stb;

run;

quit;

proc score data=test1 score=regout5 out=testout5

type=parms;

var factor1 factor2 factor3 factor5 duration director\_star actors\_star budget;

run;

data testout5;

set testout5;

deviation=model1-gross;

deviation\_square=deviation\*deviation;

run;

proc sql;

select (sum(deviation\_square)/390)\*\*1/2 as mse1

from testout5;

quit;

proc reg data=train2 outest=regout6;

title "regression model for movie gross (test2)";

model gross=factor1 factor2 factor3 factor5 duration director\_star actors\_star budget/vif stb;

run;

quit;

proc score data=test2 score=regout6 out=testout6

type=parms;

var factor1 factor2 factor3 factor5 duration director\_star actors\_star budget;

run;

data testout6;

set testout6;

deviation=model1-gross;

deviation\_square=deviation\*deviation;

run;

proc sql;

select (sum(deviation\_square)/390)\*\*1/2 as mse2

from testout6;

quit;

proc reg data=train3 outest=regout7;

title "regression model for movie gross (test3)";

model gross=factor1 factor2 factor3 factor5 duration director\_star actors\_star budget/vif stb;

run;

quit;

proc score data=test3 score=regout7 out=testout7

type=parms;

var factor1 factor2 factor3 factor5 duration director\_star actors\_star budget;

run;

data testout7;

set testout7;

deviation=model1-gross;

deviation\_square=deviation\*deviation;

run;

proc sql;

select (sum(deviation\_square)/389)\*\*1/2 as mse3

from testout7;

quit;

proc reg data=train4 outest=regout8;

title "regression model for movie gross (test4)";

model gross=factor1 factor2 factor3 factor5 duration director\_star actors\_star budget/ vif stb;

run;

quit;

proc score data=test4 score=regout8 out=testout8

type=parms;

var factor1 factor2 factor3 factor5 duration director\_star actors\_star budget;

run;

data testout8;

set testout8;

deviation=model1-gross;

deviation\_square=deviation\*deviation;

run;

proc sql;

select (sum(deviation\_square)/389)\*\*1/2 as mse4

from testout8;

quit;

proc reg data=movie4;

title "regression model for movie gross (all)";

model gross=factor1 factor2 factor3 factor4 factor5 factor6 duration director\_star actors\_star budget title\_year/selection=stepwise vif stb;

run;

quit;

proc score data=movie4 score=regout out=testout

type=parms;

var factor1 factor2 factor3 factor4 factor5 factor6 duration director\_star actors\_star budget title\_year;

run;