



CAPSTONE PROJECT FINAL REPORT

Box-office, Oscars Prediction with
Machine Learning in R

Fall 2020
student id: chun1

Raymond Chun





ABSTRACT

The film industry was once a simple artistic field that showed moving pictures with a story. Nowadays, films are a big business that impacts popular culture in many profound ways. Modern Hollywood amazes and entertains us with endless movies at the theater. They use cutting edge technology, dazzling filming techniques, and beautiful storytelling to reach us. This project aims to predict the box-office performance of the mainstream movies from large production studios. This project uses several regression models and calculates the error of each one. This project will demonstrate the connection between numerical and categorical variables with the box-office ticket sales. Also, a second data set was used to predict Oscar nominated films. Classification models with accuracy results were used in order to demonstrate this.



PROJECT GOALS

Goals:

- Predict box-office performance for a certain criteria of films since 1980
- Predict the ratio of box-office over budget for a filtered subset of films since 1993
- Predict Oscar nominations for movies based off of two data sets since 2009
- Study and examine certain ways to improve movie data and prediction



P R E - P R O C E S S I N G

A lot of effort and time went into studying and understanding the data before actually implementing the models.

1. Correlations: I studied the numerical variables to see which ones were mostly correlated with box-office performance.
2. Lasso: Since Lasso has a built in variable selection process, that was also considered carefully.
3. Time Period: The 1980's are generally considered the beginning of the modern age of cinema. Technology, advertising, genres were starting to solidify into certain standards during this time. However, one could also consider each decade as era too. Films are speeding up in terms of development.
4. Skewness: The box-office performance is highly right skewed. The log and ratio were used to scale this.
5. Categorical Variables: I separated and encoded the genres, directors, and writers variables. These categories were separated using R's dplyr and tidyverse libraries.
6. Missing Values: Any missing values were imputed with the Recipes package in R. Many of the missing values were due to metascore and the usa_box_office columns.

MODEL RESULTS

Section 1: Predicting Box-office performance

LINEAR REGRESSION

```
predictionsLR= predict(linearRegression,  
movies_test)  
RMSE(predictionsLR, movies_test$box_office)  
R2(predictionsLR, movies_test$box_office)
```

RMSE: 0.6091574

R2: 0.7006303

LASSO

```
predictionsL <- predict(lasso, movies_test)  
RMSE(predictionsL,  
movies_test$box_office)  
R2(predictionsL, movies_test$box_office)
```

RMSE: 0.6098021

R2: 0.7023074

RIDGE

```
predictionsRidge1 <-  
predict(ridge,movies_test)  
RMSE(predictionsRidge1,  
movies_test$box_office)  
R2(predictionsRidge1,  
movies_test$box_office)
```

RMSE: 0.6105359

R2: 0.7052497

ENET

```
predictionsElasticNet1 <- predict(enet,  
movies_test)  
RMSE(predictionsElasticNet1,  
movies_test$box_office)  
R2(predictionsElasticNet1,  
movies_test$box_office)
```

RMSE: 0.6120553

R2: 0.7055012

MODEL RESULTS

CONT'D

Section 2: Predicting the ratio of Box-Office over Budget Ratio

LINEAR REGRESSION

```
predictionsLR=  
predict(linearRegression,  
movies_test)  
RMSE(predictionsLR,  
movies_test$ratio)  
R2(predictionsLR,  
movies_test$ratio)
```

RMSE: 0.9644851
R2: 0.2484607

LASSO

```
predictionsL <- predict(lasso,  
movies_test)  
RMSE(predictionsL,  
movies_test$ratio)  
R2(predictionsL,  
movies_test$ratio)
```

RMSE: 0.9623387
R2: 0.2479994

RIDGE

```
predictionsRidge <-  
predict(ridge,movies_test)  
RMSE(predictionsRidge,  
movies_test$ratio)  
R2(predictionsRidge,  
movies_test$ratio)
```

RMSE: 0.9629903
R2: 0.2444957

ENET

```
predictionsElasticNet <-  
predict(enet, movies_test)  
RMSE(predictionsElasticNe  
t, movies_test$ratio)  
R2(predictionsElasticNet,  
movies_test$ratio)
```

RMSE: 0.9621042
R2: 0.2466875

MODEL RESULTS

CONT'D

Section 3: Predicting Oscar Nominations

LASSO

```
prediction(lasso_predictions_prob$`1`, movies_oscars_test$Oscar)
performance(pred_lasso, measure = "auc")@y.values
perf <- performance(pred_lasso,
measure = "tpr", x.measure = "fpr")
```

Area Under the Curve: 0.8469001
Accuracy: 0.895

RIDGE

```
pred_ridge =
prediction(ridge_predictions_prob$`1`, movies_oscars_test$Oscar)
performance(pred_ridge, measure = "auc")@y.values
perfR <- performance(pred_ridge,
measure = "tpr", x.measure = "fpr")
```

Area Under the Curve: 0.8007106
Accuracy: 0.8922

ENET

```
pred_enet =
prediction(enet_predictions_prob$`1`,
movies_oscars_test$Oscar)
performance(pred_enet, measure = "auc")@y.values
perfE <- performance(pred_enet,
measure = "tpr", x.measure = "fpr")
```

Area Under the Curve: 0.820497
Accuracy: 0.9031

LOGISTIC

```
predict.logistic <-
predict(model.logistic,
movies_oscars_test, type="response")
predict.logistic.label =
factor(ifelse(predict.logistic > .5, "Yes", "No"))
actual.label =
movies_oscars_test$Oscar
table(actual.label,
predict.logistic.label)
```

Area Under the Curve: 0.7415
Accuracy: 0.8867

CONCLUSIONS

- Based off the error and accuracy calculations, we can reasonably conclude that the models for predicting box-office and oscar winners **had acceptable-to-good performances**. While there is a lot of noisy info in the dataset, I was able to pre-process carefully and create strong filters for prediction.
- If studios invest a lot into a movie, they tend to have lots of sales. However, the profit ratio isn't always strong. It costs a lot to make movies and marketing can cost even more.
- The main strength of the IMDB dataset seems to be the critic and fan reviews they collect. These show a lot of connections with how well a film performs. Naturally, you can see the films with better reviews typically perform better at the box-office.
- Directors matter quite a bit. When encoded and split, the directors had a large effect on the error results. The Oscars reward serious, well-developed films so it makes sense they reward serious and respected directors.
- Many films with the top ratios were different than expected. Films like 'The Gallows' or 'Fireproof' greatly outperformed their budget but had little impact on culture. Horror films do particularly well at this.
- Movie Data is robust, but could use improvements. I believe films should aim to collect more advanced metrics, such as a pace of a film, scene changes, dialogue speed, camera techniques, etc. There could be more ways to measure the style of a film. This data set is definitely strong, but could use more useful variables.
- Fans and Critics are important. Executives should invest more in movie screeners and previews. This will definitely give you an idea with how a film can perform. Streaming can help deliver that in our current era.



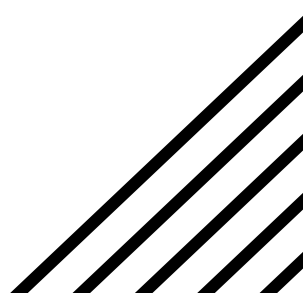
CLOSING REMARKS

I really enjoyed this movie data project and I have no regrets on my choice. It was a great challenge to pre-process all this data before modelling it. I learned so much from just trial and error, but also studying this domain of movies. It was terrific.

In terms of technical skills, I learned a tremendous amount about separating categorical fields and rows. In the past, I usually worked with numerical variables so this was an interesting lesson in learning how to deal with categorical ones. Imputing missing values was useful too.

For analysis and story telling, I definitely learned how budget and results go hand in hand. Investing a lot of money is easy, but doing it mindfully with strong results is hard. I also feel like success is about perception. Even films that do not win major awards can impact culture in positive ways. Fans don't always show up with money, they show up with passion.

Thanks again to Professor Chan for the zoom calls and advice. The suggestions for the project were very useful and intuitive. I appreciate all the time and conversations.





REFERENCES

IMDB data set, Oscars Award data set:

<https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset>

<https://www.kaggle.com/unanimad/the-oscar-award>

History of Film references:

https://en.wikipedia.org/wiki/History_of_film

<https://open.lib.umn.edu/mediaandculture/chapter/8-2-the-history-of-movies/>

Movie Rating Sites:

<https://www.imdb.com/>

<https://www.rottentomatoes.com/>

<https://www.metacritic.com/>

