



Linear Regression

Machine Learning Algorithm for Thrillers

Presented 7.12.22

Created by: Jenni Hawk

Art Company

Business Situation

A newly emerged production studio plans to make movies in the thriller genre and would like to know which characteristics of thrillers are predictors of US Box Office Gross.

Key Questions:



Does a set of features do a good job in predicting US Gross for thrillers?



Which features are significant predictors of US Gross for thrillers?



GAMBIT
STUDIOS

Project Steps

ACTION



WEBSCRAPING

- Scraped IMDB Thrillers for target and feature data
- 1100 thriller titles, 16 potential predictor variables



EDA & REGRESSION VIABILITY

- Ensure data correct and appears as expected.
- Data cleanup, address missing values, etc
- Correlation matrix, reg plots, R^2 score
- Feature engineering



DETERMINE BASELINE MODEL

- Tested log transform vs no transform
- Tested regularization methods
- Identified features with meaningful coefficients



TRAIN – VALIDATE – TEST

- Utilized cross validation
- Tested two models

TOOLS USED

Request Module, BeautifulSoup Library

Pandas, Seaborn, Statsmodels
cpi library (to apply inflation to budget based on year)

Pandas, Sklearn

Sklearn



Features Scraped From Thriller List IMDB

IMDB: Thrillers Categorized by Genre

Thriller (Sorted by US Box Office Descending)

1-50 of 295,049 titles. | [Next »](#)

View Mode: [Compact](#) | [Detailed](#)

Sort by: [Popularity](#) | [A-Z](#) | [User Rating](#) | [Number of Votes](#) | [US Box Office▼](#) | [Runtime](#) | [Year](#)
| [Release Date](#) | [Date of Your Rating](#) | [Your Rating](#)



1. **The Dark Knight** (2008)

PG-13 | 152 min | Action, Crime, Drama

★ 9.0

☆ [Rate this](#)

84 Metascore

When the menace known as the Joker wreaks havoc and chaos on the people of Gotham, Batman must accept one of the greatest psychological and physical tests of his ability to fight injustice.

Director: Christopher Nolan | **Stars:** Christian Bale, Heath Ledger, Aaron Eckhart, Michael Caine

Votes: 2,580,159 | **Gross:** \$534.86M



● Features scraped

Today

Discuss **key insights** from **each phase** of the Linear Regression process.



REGRESSION
VIABILITY

MODEL +
OPTIMIZATIONS

QUESTIONS
ANSWERED

Today

Discuss **key insights** from **each phase** of the Linear Regression process.

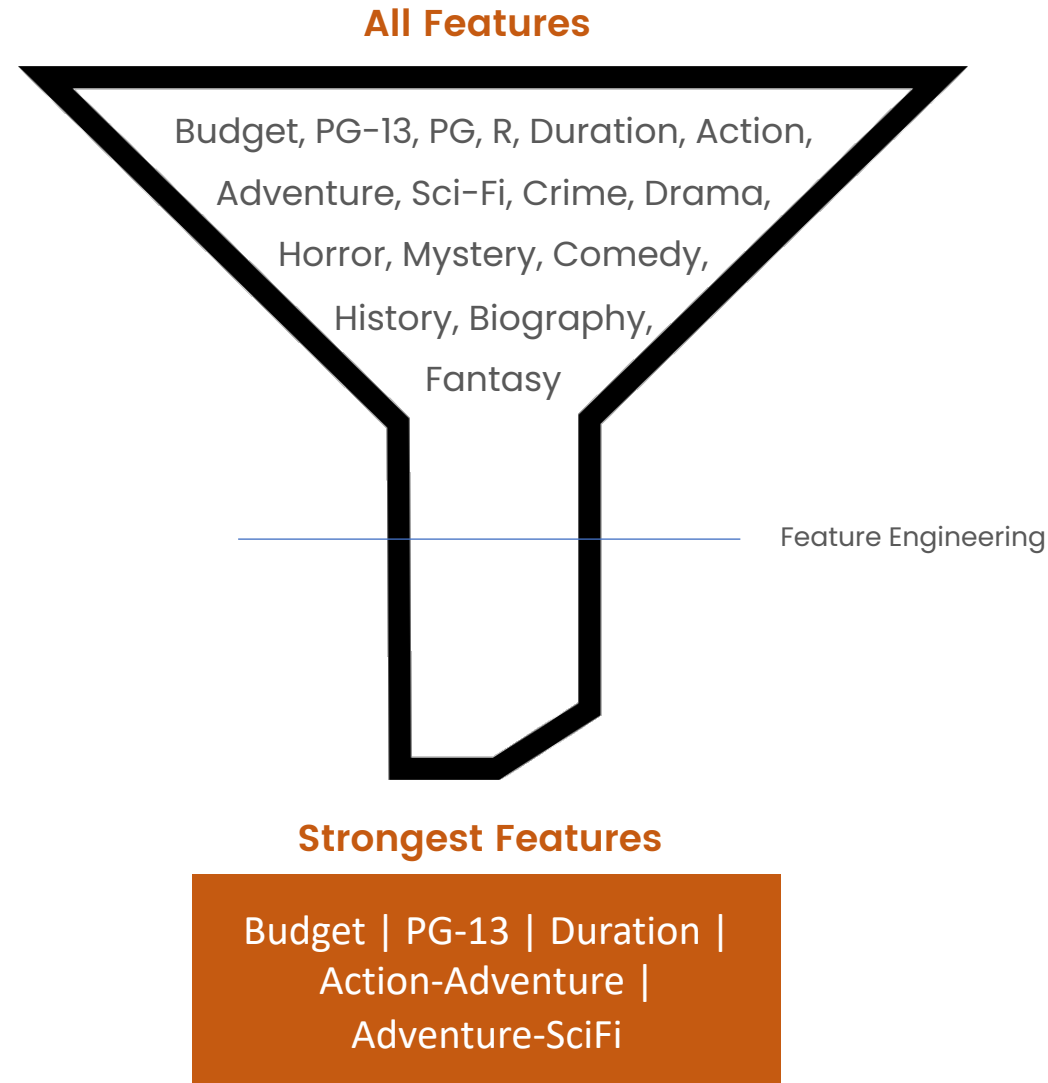


REGRESSION
VIABILITY

MODEL +
OPTIMIZATIONS

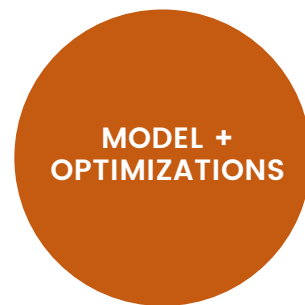
QUESTIONS
ANSWERED

Determine Features for Model Inclusion



Methodology:

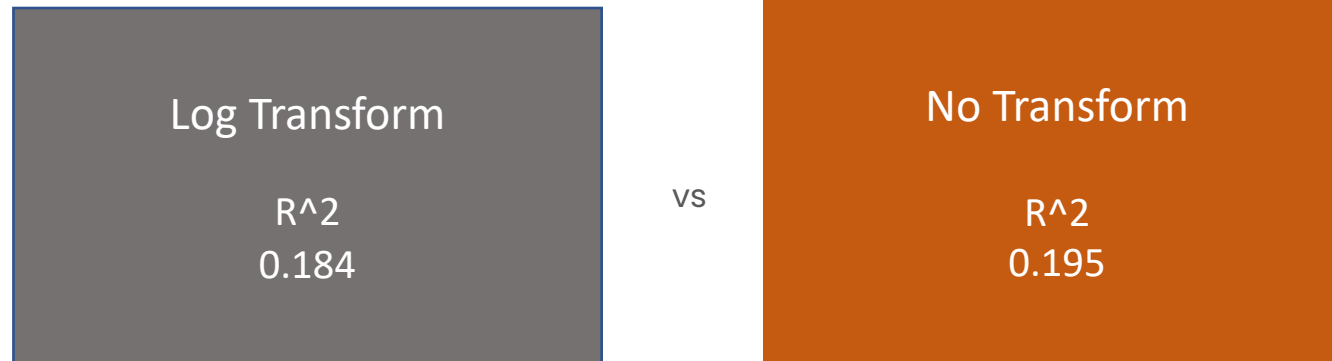
- Correlation Heatmap
- Features must have strong correlation with US Gross
- Addressed collinearity amongst action + adventure by combining



Tested Log Transform vs No Transform

No Transform performed slightly better than Log Transform

R^2 of model with cross validation



Predictor Variables Included:

- Action_Adventure
- Budget
- PG-13
- Duration

Does Lasso increase performance?

Linear Regression	Train R ² Score	Test R ² Score
	0.213	0.178

- While there's not obvious overfitting of the model, Lasso was conducted to determine if score could be made better

Regularization

Lasso	Train R ² Score	Test R ² Score
	0.212	0.179

- Lasso made a very small improvement in prediction equation

Predictor Variables Included

- Action_Adventure
- Adventure_Scifi
- Budget
- PG-13
- Duration

How well is the model predicting?

How close the prediction is against the real value

Linear Regression

MAE
(Mean Absolute Error)

\$52,916,385

- Tells us the average difference between the actual data value and the value predicted by the model
- Establishes baseline metric to be used in further model testing
- Goal is to improve this model by reducing this error

While the model may be far from perfect,
let's see what we've learned...







Does a set of features do a good job in predicting US Gross for Thrillers?

Answer: The current set of features do not do a good job of accurately predicting US Gross.



Which features are significant predictors of US Gross?

Answer: The current set of features do a partial job in predicting US Gross



What else can we say?

Coefficient Analysis Provides Meaning

When reviewing movie projects:

- A **PG-13 rating** should be preferred over R Rating
 - When possible + When it works with larger business strategy
- Lean more into Thrillers that are **Action-Adventure** and **Adventure-SciFi** versus other genres
- **Budget** and **Duration** also indicated positive per-unit impact on US Gross. Further analysis required to provide actionable insight

Positive Per-Unit Impact on US Gross

Feature	Coefficient
PG-13	1.02
Action_Adv	8.17
Adv_Scifi	2.76

Categorical variables – binary values

- US Gross prediction to increase by \$1.02MM with PG-13
- US Gross prediction to increase \$8.17MM with Action_Adv
- US Gross prediction to increase \$2.76 with Adv_Scifi

Positive Per-Unit Impact on US Gross

Feature	Coefficient
Budget	5.09

Dollars – Continuous values

- US Gross prediction to increase by \$5.09MM for every dollar increase in budget given all variables remain the same

Positive Per-Unit Impact on US Gross

Feature	Coefficient
Duration	9.34

Minutes – Integer

- US Gross prediction to increase by \$9.34MM for every minute increase.

Will likely remove from final – practicing interpretation here

Discussion

