# Linear Regression

Machine Learning Algorithm for Thrillers

Presented 7.12.22

Created by: Jenni Hawk

# Business Situtation

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?**

# Project Steps

## ACTION

## TOOLS USED

### WEBSCRAPING

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

Request Module, BeautifulSoup Library

### 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

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

### DETERMINE BASELINE MODEL

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

Pandas, Sklearn

### TRAIN – VALIDATE - TEST

- Utilized cross validation
- Tested two models

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

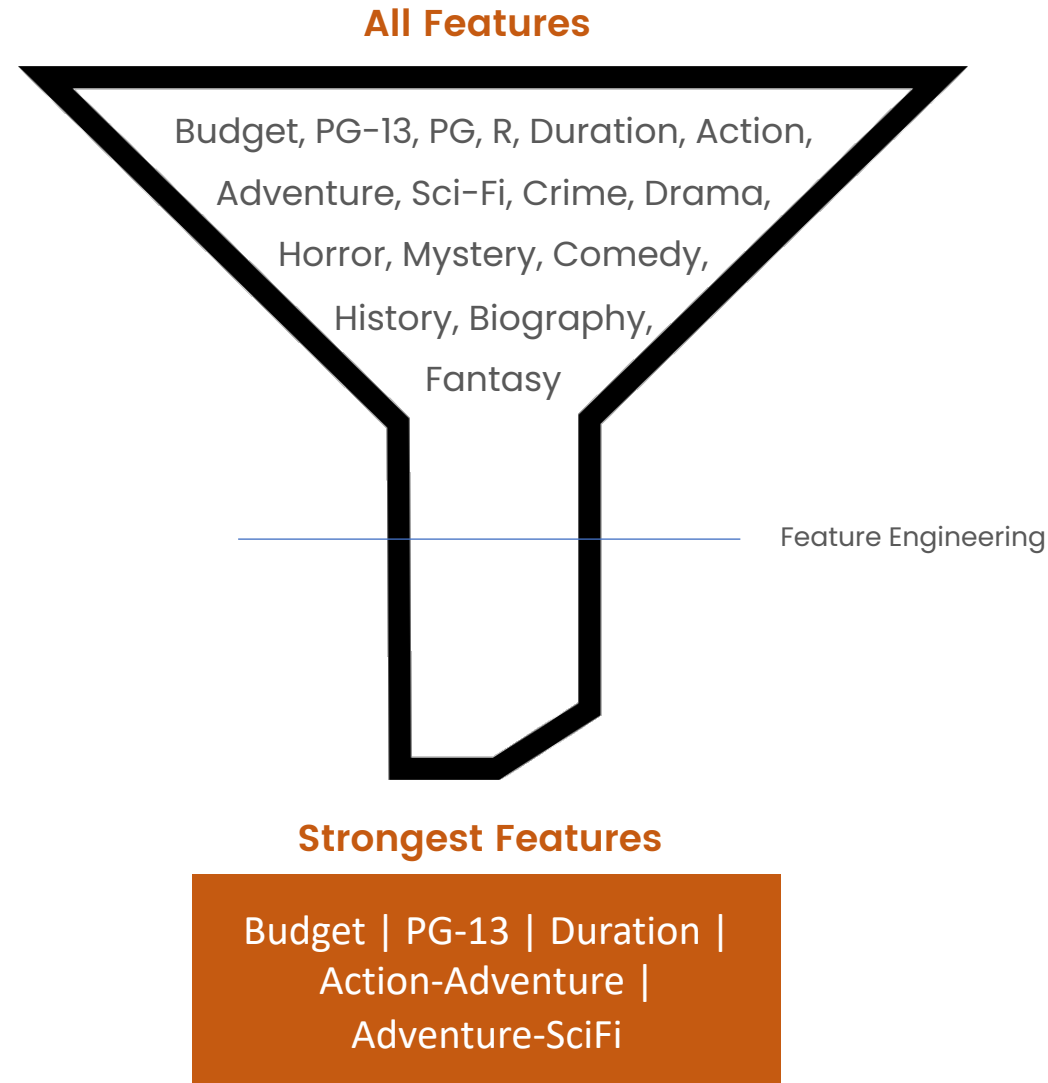Discuss key insights from each phase of the Linear Regression process.

**REGRESSION VIABILITY**

**MODEL + OPTIMIZATIONS**

**QUESTIONS ANSWERED**

# Determine Features for Model Inclusion

**All Features**

Budget, PG-13, PG, R, Duration, Action, Adventure, Sci-Fi, Crime, Drama, Horror, Mystery, Comedy, History, Biography, Fantasy

Feature Engineering

**Strongest Features**

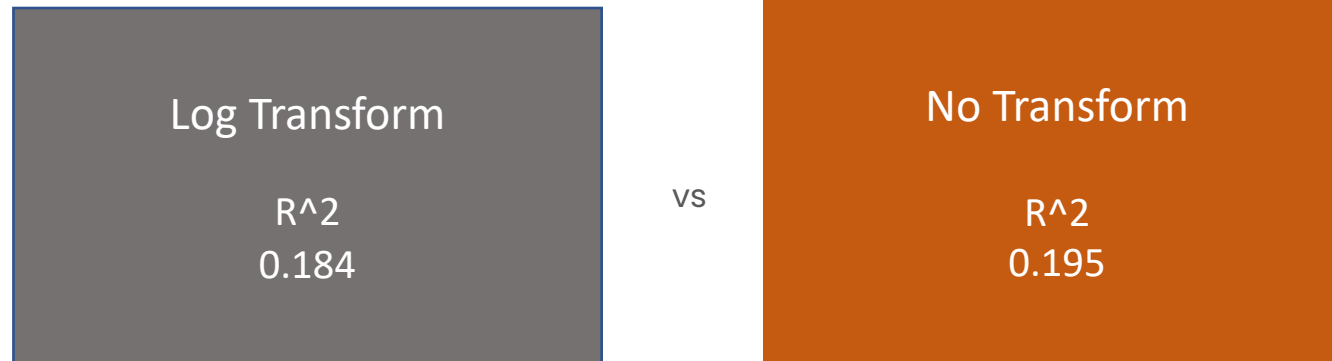Budget | PG-13 | Duration | Action-Adventure | Adventure-SciFi

Methodology:
- Correlation Heatmap
- Features must have strong correlation with US Gross
- Addressed collinearity amongst action + adventure by combining
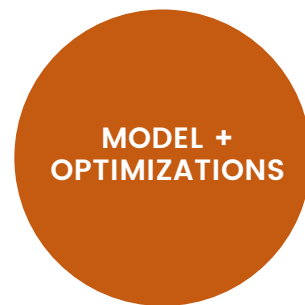
# Tested Log Transform vs No Transform

No Transform performed slightly better than Log Transform

R^2 of model with cross validation

Log Transform

R^2
0.184

vs

No Transform

R^2
0.195

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

- Don't make thrillers too short in duration

Positive Per-Unit Impact on US Gross

| Feature | Coefficient |
|---|---|
| PG-13 | 10,206,365.81 |
| Action_Adv | 8,177,097.38 |
| Adv_Scifi | 2,761,602.71 |

Categorical variables - binary values

Positive Per-Unit Impact on US Gross

| Feature | Coefficient |
|---|---|
| Budget | .51 |

Dollars - Continuous values

Positive Per-Unit Impact on US Gross

| Feature | Coefficient |
|---|---|
| Duration | 934,552 |

Minutes - Integer

- US Gross prediction to increase by $10M with PG-13 ($10Million over any other ratings.)
- US Gross prediction to increase $8MM with Action_Adv
- US Gross prediction to increase $2.7MM with Adv_Scifi

- US Gross prediction to increase by .51

- US Gross prediction to increase by $934,000 per minute when at zero

# Discussion