GAMBIT STUDIOS

Linear Regression

Machine Learning Algorithm for Thrillers

Presented 7.12.22

Created by: Jenni Hawk

Art Company

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?

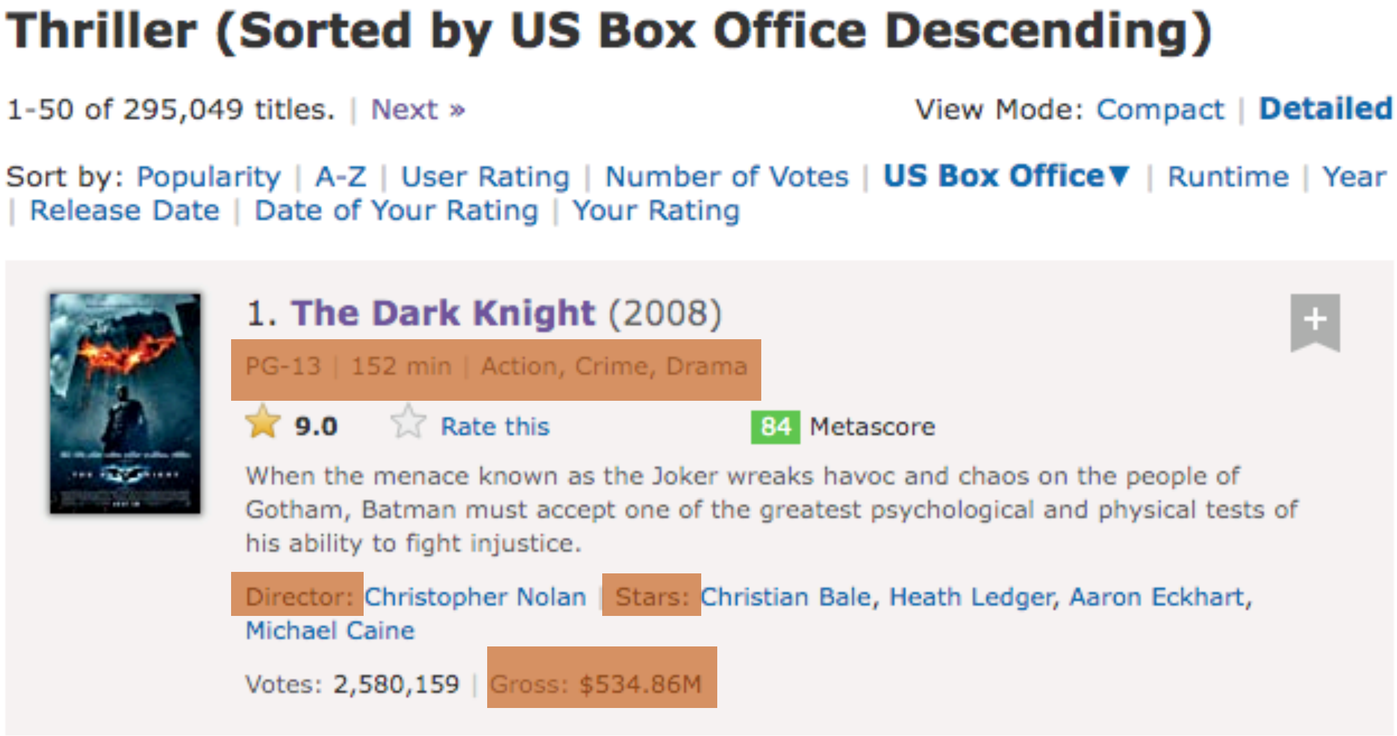
GAMBIT
STUDIOS

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 • Filtered to small set of features that had strongest correlation with US Box Office Gross • Tested log transform vs no transform	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 Linear process.

REGRESSION VIABILITY

BASELINE

MODEL EVALUATION

QUESTIONS ANSWERED

Today

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

REGRESSION VIABILITY

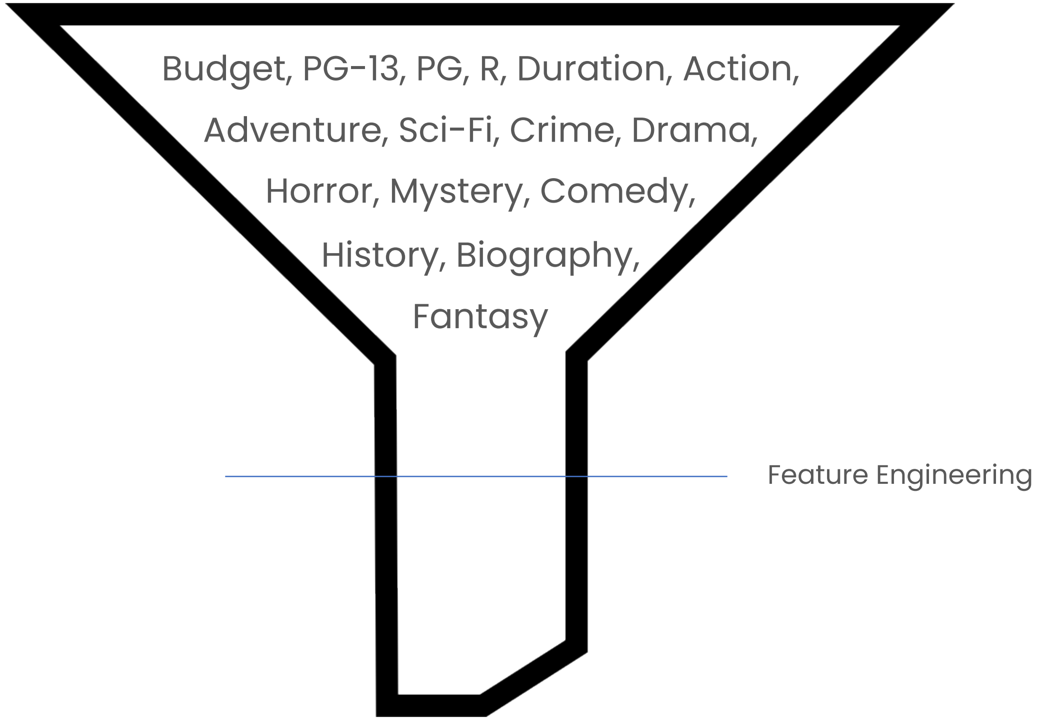
BASELINE

MODEL EVALUATION

QUESTIONS ANSWERED

Determine Features for Baseline

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

REGRESSION VIABILITY

BASELINE

MODEL EVALUATION

QUESTIONS ANSWERED

Getting to a Baseline

No Transform performed slightly better than Log Transform

Log Transform

R^2
0.184

vs

No Transform

R^2
0.195

R^2 of model with cross validation

REGRESSION VIABILITY

BASELINE

MODEL EVALUATION

FUTURE WORK

Model Performance

How close the prediction is against the real value

Test Scores	Findings
Mean Absolute Error MAE = 53.45	• Establishes baseline. metric to be used in further model testing • Goal is to improve this model by reducing this error
R^2 = 0.195	• 19% of the model predictions are correct To improve: • Remove outlier data (ie: impact of blockbusters, movies from 70s/80s) • Find/Test additional features beyond current dataset

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

REGRESSION VIABILITY

BASELINE

MODEL EVALUATION

QUESTIONS ANSWERED

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?

Correlation Status 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 need further analysis to provide actionable insight

Strong Correlation	No Correlation	Strong Neg Corr
Budget | PG-13 | Duration | Action-Adventure | Adventure-SciFi	Comedy | History | Biography | Fantasy | Romance	R | Crime | Drama | Horror

LIES BENEATH
THE 4TH FLOOR
PHONEbooth
DOUBLE JEOPARDY
GERALD'S GAME
HUSH
Invitation
A PERFECT GETAWAY
WHEN A STRANGER CALLS
COLD CREEK MANOR
PERFECT STRANGER
THE

Discussion