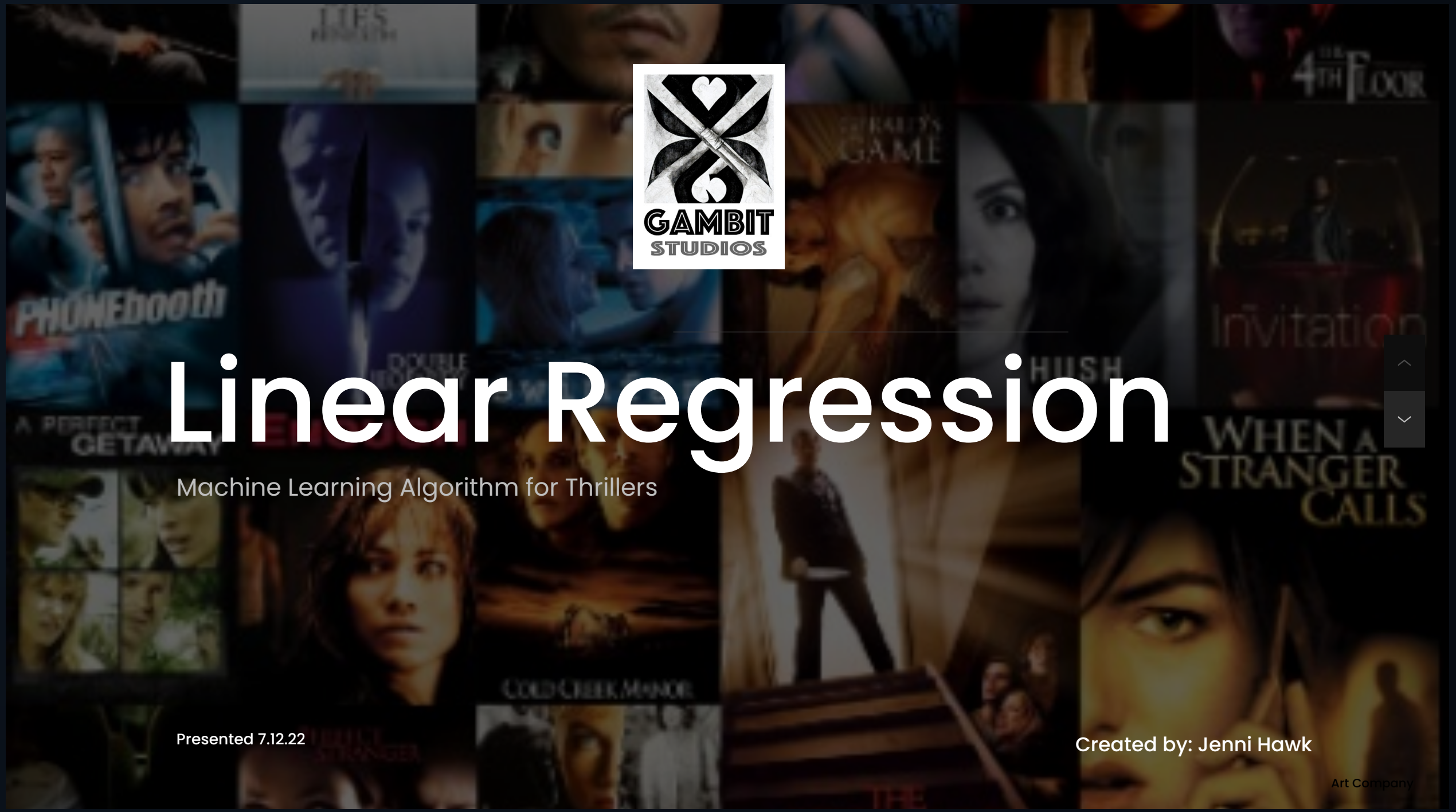


Linear Regression

Machine Learning Algorithm for Thrillers

Presented 7.12.22

Created by: Jenni Hawk

Art Company

Assignment Goals

- Learn the fundamentals of linear regression
- Viability Analysis
- Instantiate model, Train + Test, Score
- Learn coding fundamentals: Statsmodels, Sklearn, etc

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?

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 • Data cleanup, address missing values, etc • Create dummy variables for categorical features • Correlation matrix and regression plots to check linear relationships	Pandas, Seaborn, Statsmodels cpi library (to apply inflation to budget based on year)
Linear Regression Modeling • Fit data to the model • Train, Test, Score • Coefficients: most impactful features	Pandas, Sklearn
EDA & Continued Model Optimization • Predicted vs Actuals • Regularization	Sklearn, Matplotlib, Seaborn

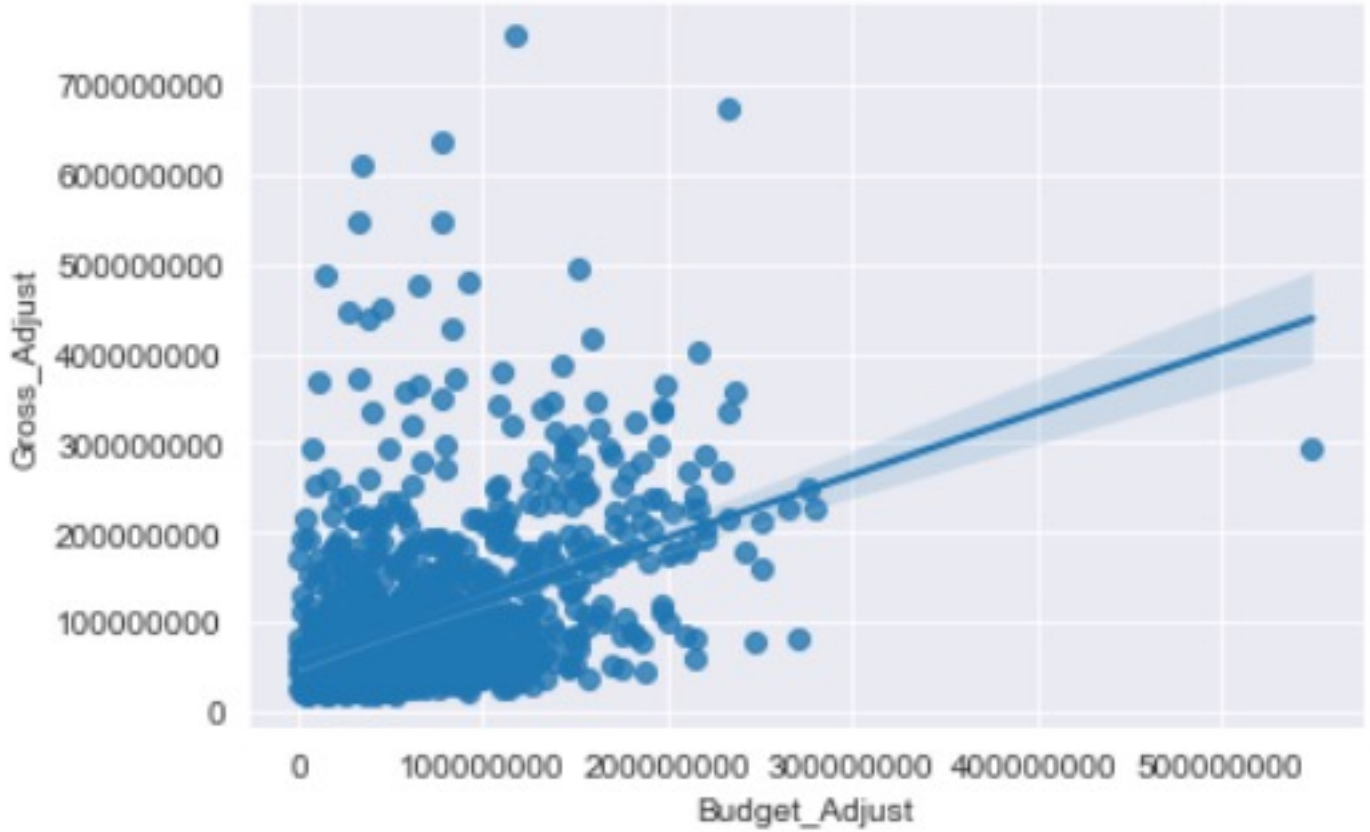
Features Scraped From Thriller List IMDB

IMDB: Thrillers Categorized by Genre

- Features scraped

Linear Regression Modeling

Check Linear Relationships



Feature Target Correlation Analysis

Low correlation doesn't mean it won't contain signal that can be discovered

Positive Correlation with US Gross

Budget	0.72
Duration	0.56
Adventure	0.49
Action	0.42
PG	0.42
PG-13	0.33
Sci-Fi	0.24
Music	0.18

Data: Pearson Correlation / Seaborn Heatmap

Negative Correlation with US Gross

R Rating - 0.47
Horror -0.36
Crime -0.25
Drama -0.21
Mystery -0.21
Romance -0.14
Biography -0.11
Fantasy -0.08
Comedy -0.03
History -0.03

Close to zero -.10 to .10

War -0.10
Sport -0.09
Western -.07
Musical -0.06
Animation – 0.00
Family 0.01

Keep these features in the modeling

Fit the data on the linear regression model

Scores

R^2

Slight overfitting expected

Train R^2 .318

Test R^2 .276 **28% of Variance explained by model**

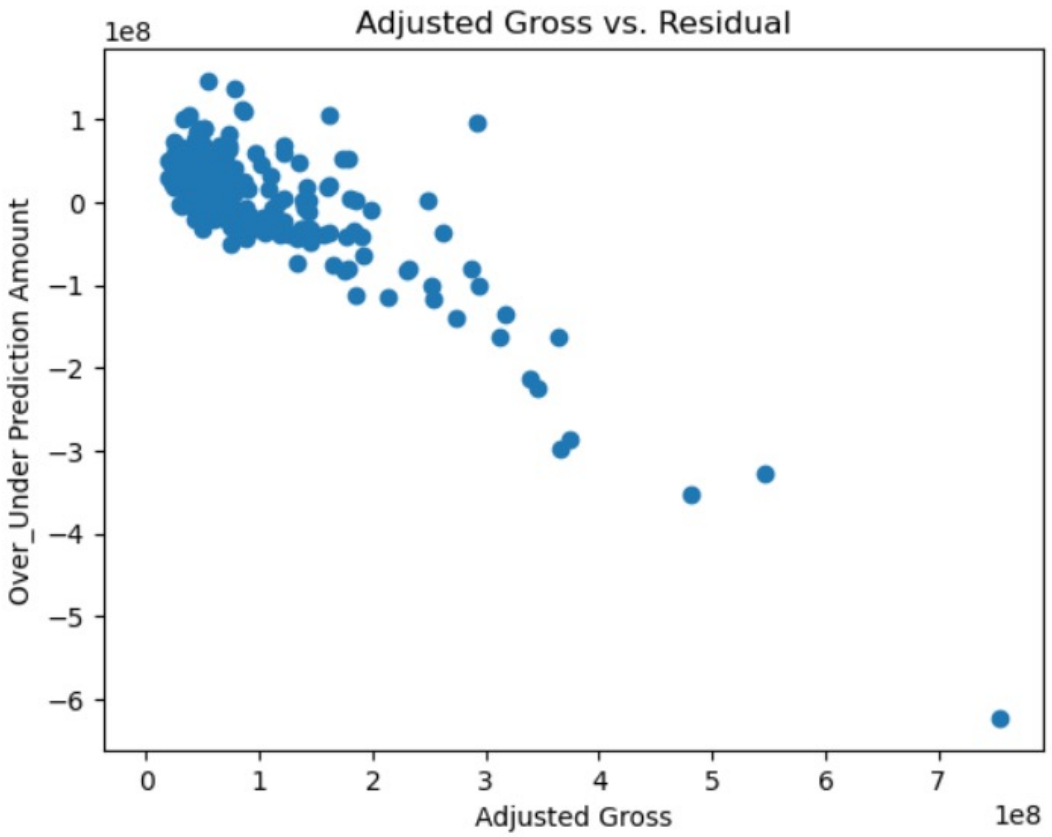
MAE: Mean Absolute Error

How close the prediction is against the real value

$46,825,271

Predicted vs Actuals

Underpredicting – big blockbusters may be the issue



Findings

- Residuals are problematic
  - Heteroskedasticity: unequal variability (scatter)
- Hypothesize that block busters are underpredicting

Future Work

- Address the blockbusters

Regression Coefficients

What the model considers to be the most impactful features and the per-unit impact on US Gross

Positive Impact on Thriller US Gross	
PG	68,017,516.47
Adventure	19,793,585.00
Sci-Fi	19,450,110.29
Comedy	12,575,077.35
Duration	1,647,406.25
Budget	0.39

Negative Impact on Thriller US Gross	
History	-76,092,866.62
Musical	-67,808,047.35
Biography	-57,640,257.49
Animation	-54,468,082.93
R Rating	-40,303,575.50
Drama	-28,094,114.77
PG-13	-27,713,940.97
Romance	-29,425,524.70
Action	-20,658,701.63
Horror	-14,634,199.66
Mystery	-12,610,294.09
Crime	-10,495,540.63
Fantasy	-7,649,883.02

Neither Positive Nor Negative Impact	
War	0.00
Sport	0.00
Western	0.00
Family	0.00
Music	0.00