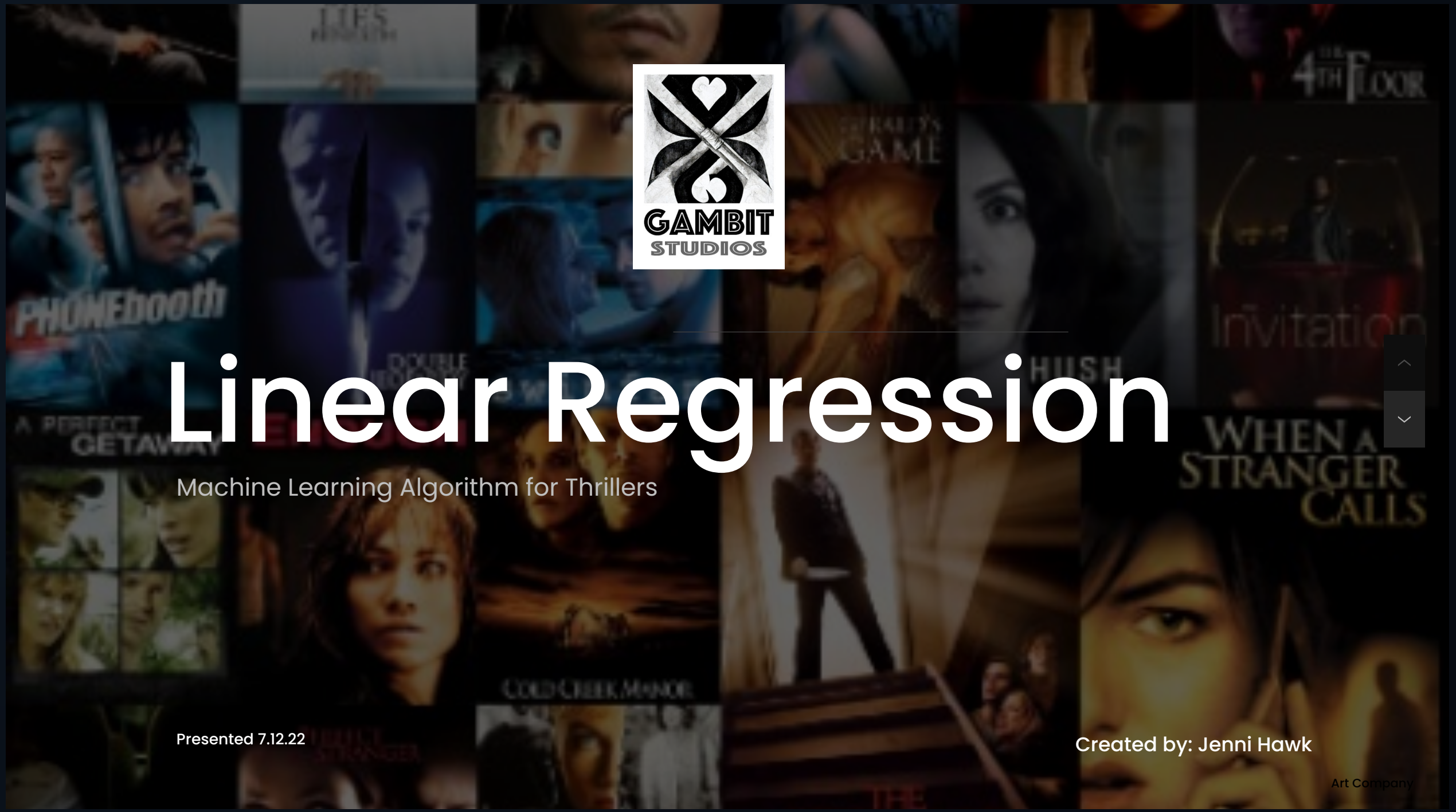


Linear Regression

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?

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

- Features scraped

Today

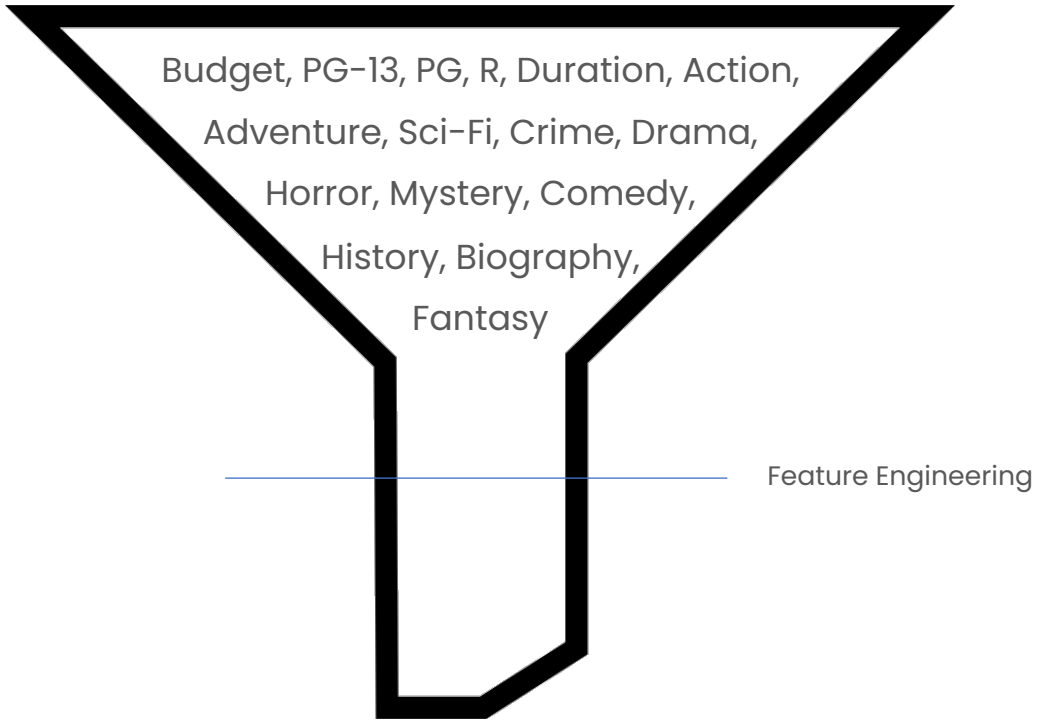
Discuss key insights from each phase of the Linear Regression process.

Today

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

Determine Features for Baseline

All Features



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	vs	No Transform
R^2 0.184		R^2 0.195

R^2 of model with cross validation

REGRESSION VIABILITY

BASELINE

MODEL EVALUATION

FUTURE WORK

Train vs Test – Baseline Model

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

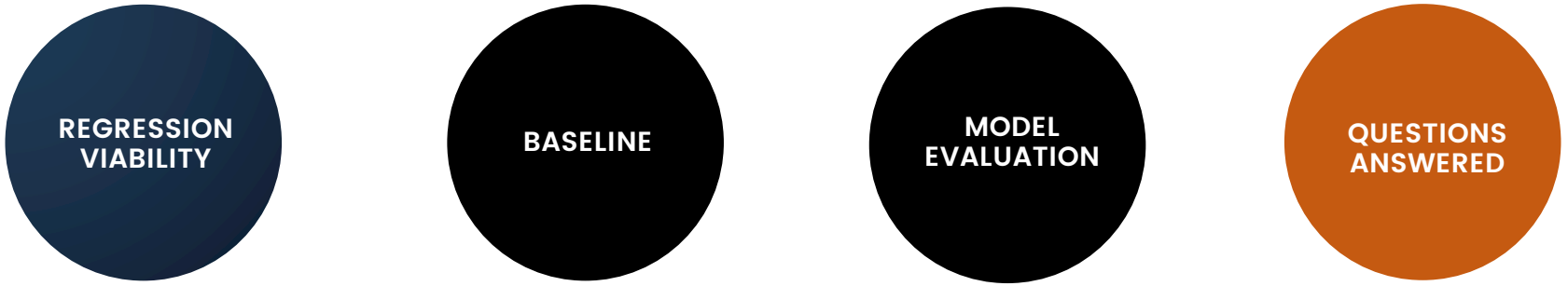
Model Performance

How close the prediction is against the real value

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

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?

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 need further analysis 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
Budget	5.09
Duration	9.34

Discussion

Art Company