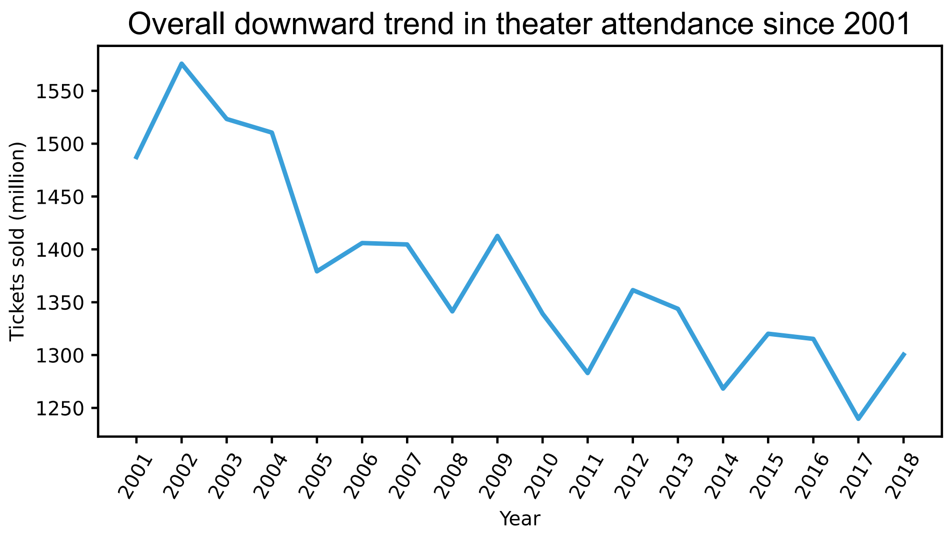
Predicting movie sales

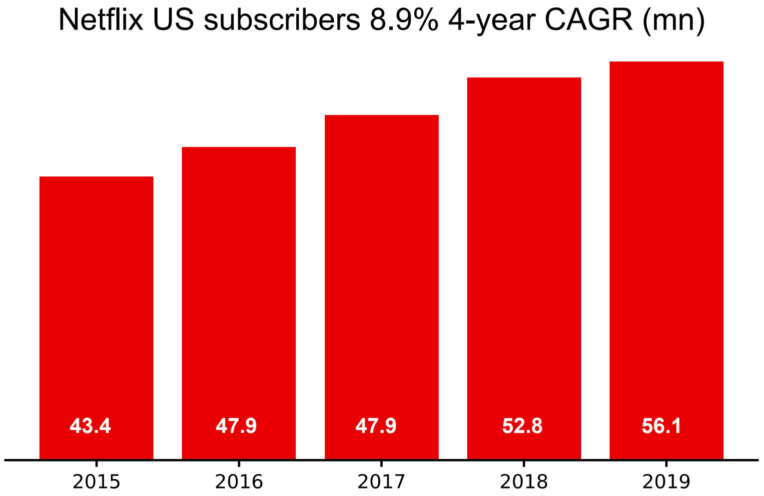
22 January 2021
Tawney Kirkland

Agenda

- Introduction
- Methodology
- Findings
- Insights
- Lessons

The movie industry is facing stiff competition from streaming services





Source: Statista

Against this backdrop, the project aims to enable the movie industry to effectively respond to the changing environment

Project objective:
Develop a model that can **predict domestic gross revenues** prior to movie release

Assumptions

Focus on highest grossing movies regardless of MPAA ratings

Period 2000 to 2020

Interpretability is important

Linear regression modeling was used to support this objective, bolstered by insights from industry experts

Tools

Web scraping:
- Beautiful Soup

Manipulation and analysis:
- Python
- Pandas
- Numpy
- Sklearn
- LassoCV

Visualization:
- Matplotlib
- Seaborn

Methods

Desktop research: Understanding industry trends

Stakeholder interviews: Semi-structured interviews with two industry stakeholders (production design and animation) to gain deeper insights on the industry and identify key features to investigate

Exploratory data analysis and linear regression: Using secondary data scraped from **BoxOffice Mojo** and **Numbers**, conducted exploratory and linear regression to predict domestic gross sales prior to movie release

A list of features is provided in the Appendix

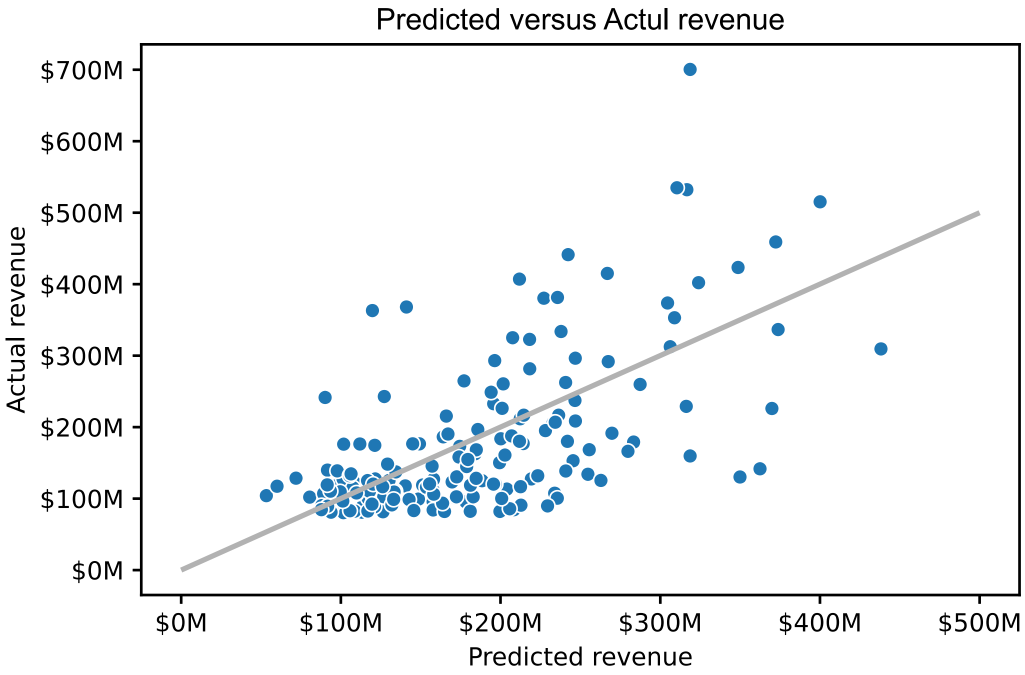
Using LassoCV enabled identification of the 11 most important features

	Budget	Budget * Sequel	Top 50 director	Runtime	Sequel	Year	PG-13	PG	Budget * Horror	Horror	R
Domestic sales	0.56	0.44	0.35	0.32	0.23	0.16	0.15	0.03	-0.05	-0.12	-0.20

Correlations for the selected features

A list of removed features and the full correlation matrix are provided in the Appendix

Approximately 39% of revenue variance can be explained by the model



R^2: 0.393

Predicting smaller revenue movies but underpredicted blockbusters

On average, the predicted value deviates from the actual value by ~ $ 82.5 million

Refer to the appendix for a plot of the predicted values versus the residuals

These findings provide important insights for the movie industry

Insights

High budget, high revenue - save the blockbusters for theater release

Low budget sequels do not fare as well at the box office as higher budget sequels - send to streaming or reconsider production

Lower budget horrors have high returns - worth studios investing in as it is a comparatively low entry cost with a high payoff (to a point)

Predicted domestic revenue =

0.593 * budget +
0.414 * budget * sequel +
-0.273 * budget * Horror +
36095819 * top50_director +
- 31964253 * sequel +
17975425 * Horror +
928330 * rating_PG +
-12239725 * rating_PG-13 +
-19517330 * rating_R +
641161 * year +
736997 * runtime +
-1267362725

Lessons and next steps

It is difficult to predict variable phenomena

Important to be very discerning around the parameters of the problem (e.g. budget / revenue outliers)

Regularization, your friend is

Next steps:

Remove outliers and remodel

Fine tune for month

Investigate and remove features (e.g. runtime)

Credits

BoxOffice Mojo	Top Lifetime Grosses, https://www.boxofficemojo.com/chart/top_lifetime_gross/?ref_=bo_cso_ac
The Numbers	Movie Budgets, https://www.the-numbers.com/movie/budgets/all
Statista	North America Box Office, https://www.statista.com/statistics/187076/tickets-sold-at-the-north-american-box-office-since-2001/

Appendix

The approach resulted in a sample of 723 movies from 2000 to 2020

Base features

- Budget
- Year
- Month
- Genre
- Runtime (minutes)
- MPAA rating
- Domestic distributor
- Crew
- Cast
- Sequel (yes / no)

Engineered features

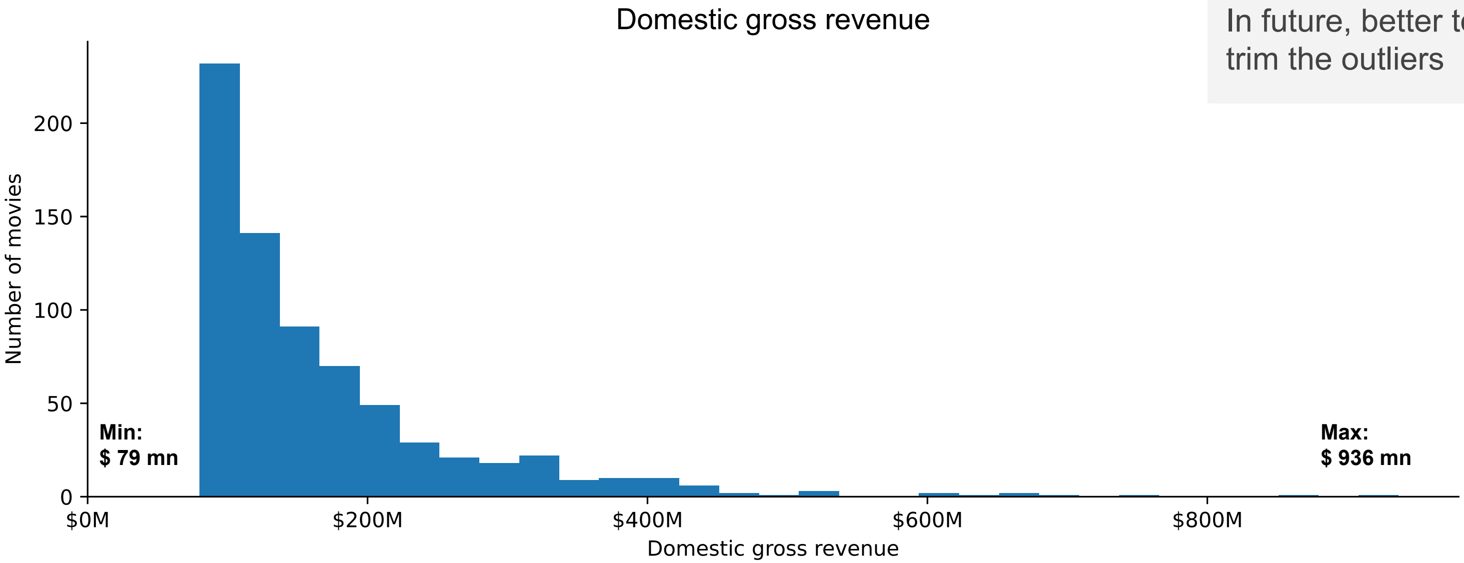
- Top 50 directors [1]
- Bankable actors (count) [2]

Key interactions

- Budget * Sequel
- Budget * Horror

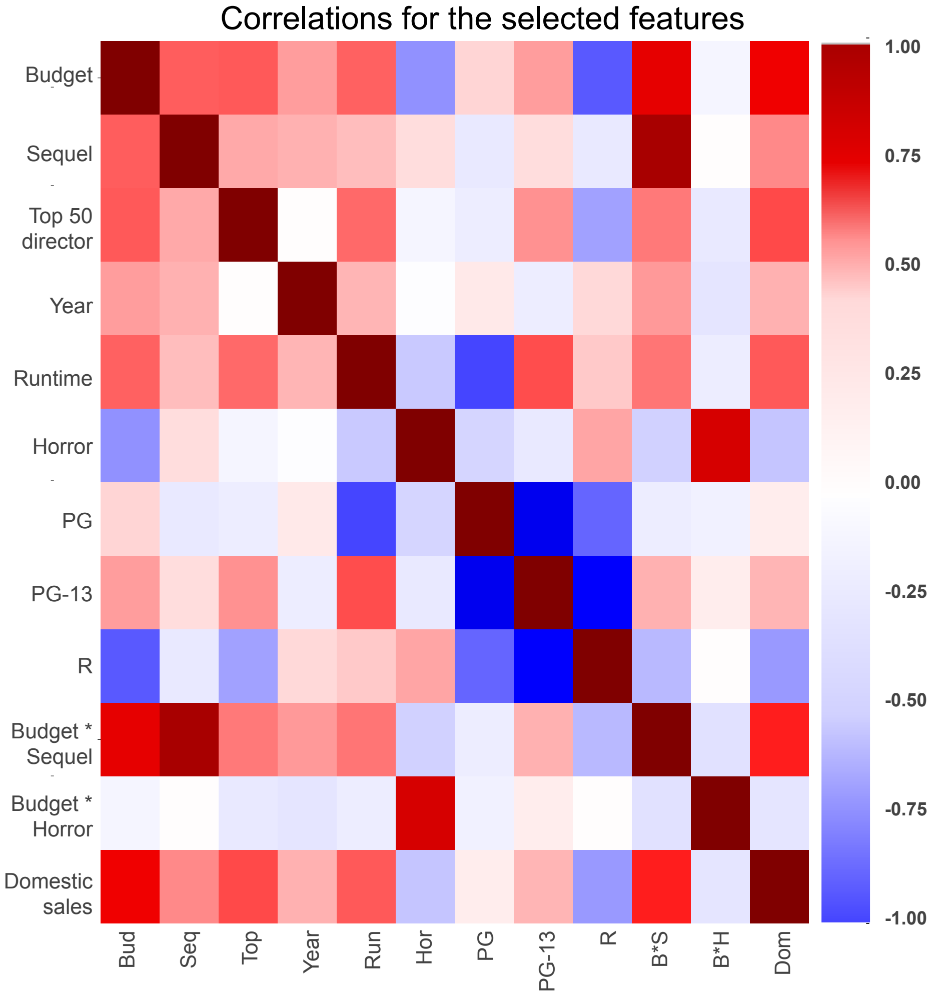
1. Top 50 directors defined in terms of the 50 directors (15% of total number of directors) contributing ~45% of total revenues
2. Based on The Numbers Index of highest value-adding actors

Heavily skewed target variable made prediction difficult



In future, better to trim the outliers

Using LassoCV enabled identification of the 11 most important features



A list of removed features is provided in the Appendix

A number of features had little to no effect in the model and were left on the cutting room floor

Removed features

- Month
- Domestic distributor
- Bankable actors

Genres including:
- Action
- Adventure
- Drama
- etc

The model both under and over predicted revenues for blockbusters