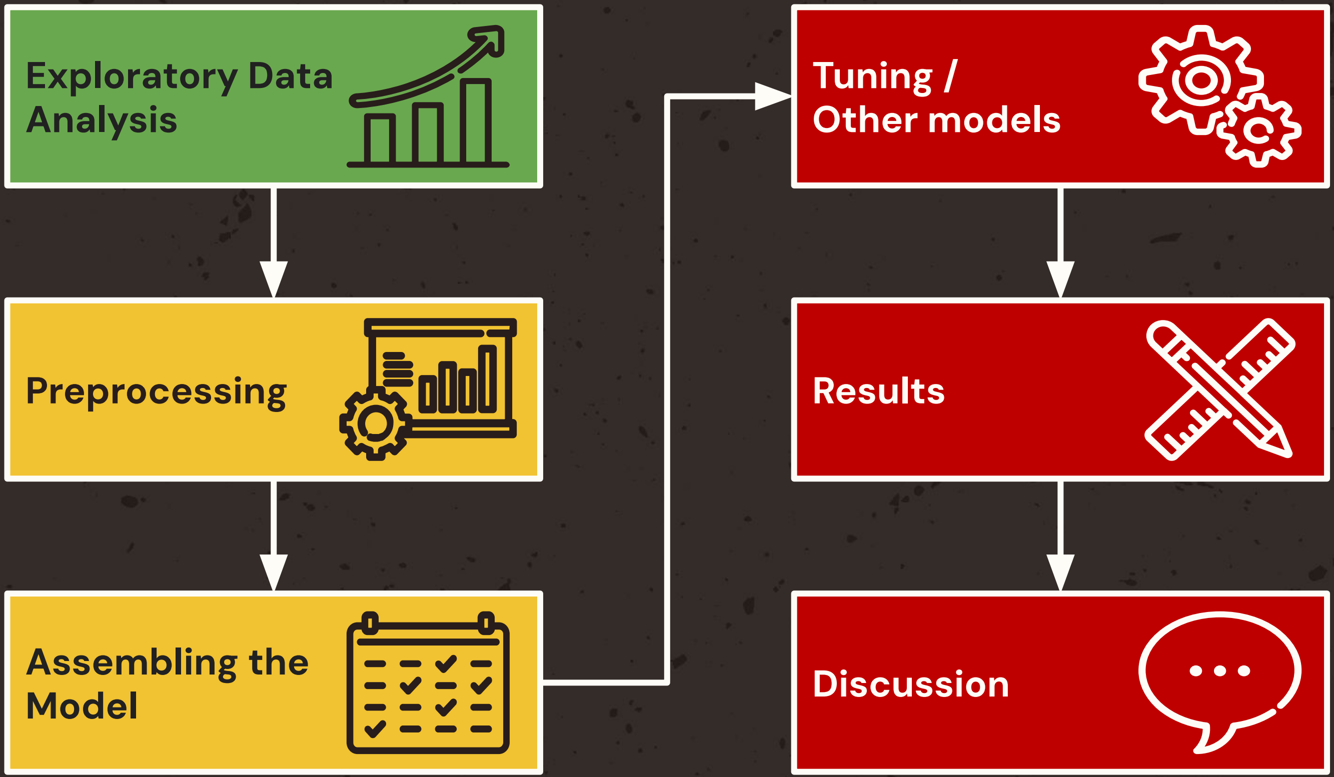
Assembling the Next Blockbuster

Leon, Julia, Mehmet, and Leticia


Exploratory Data Analysis
Preprocessing
Assembling the Model
Tuning / Other models
Results
Discussion

01

Research Question and Data Creation

Topic

Original Research Question: Assemble the next blockbuster film.

Imagined Scenario: A client comes to us with a description of a future movie. It is up to us to indicate them whether or not this movie will become a blockbuster.

Modified Research Question: Can we predict whether or not a movie will be a blockbuster, based on its pre-release characteristics?

Classification

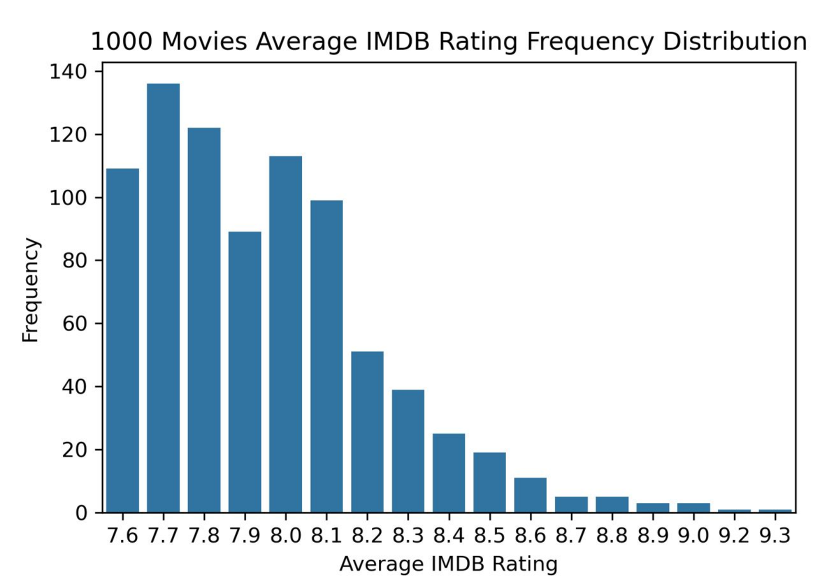
	Low Gross Income	High Gross Income
Low IMDB Rating	Flop	Critically-Disliked Blockbuster
High IMDB Rating	Hidden Gem	***<u>Critically-Acclaimed Blockbuster</u>***

Original Dataset

Top 1000 Rated Movies on IMDb

Why we didn't like it:

- Low Variance -> Could lead to high model bias.
    - IMDb average rating range [7.6, 9.3]
- Not many features (11)
- Low sample size



Webscraping

title.basics (11.5M) + title.ratings (1.5M)

Merge + Filter out non-movies and movies released pre 2020

Pre_scraping (54K)

Scrape from IMDB

Post_scraping (54K)

02

EDA + Preprocessing

Dataset Synopsis

54095 movies (all movies released from 2019 onwards)

16 columns

Features include: runtime, genres, storyline, themes, director, actor, gross income, average IMDB rating

Target variable: movie type (decided based on gross income and IMDB rating).

Dropped rows with null entries in gross income column.
Dropped columns:

- tconst
- titleType
- Unnamed
- originalTitle
- endYear

Unique Values

unique_values()

- {'Action', 'Adult', 'Adventure', 'Animation', 'Biography', 'Comedy', 'Crime', ...}

count_unique_values()

- {'Action': 1529, 'Adventure': 971, 'Fantasy': 454, 'Comedy': 3426, 'Thriller': 1564, ...}

unique_combos()

- ['Action,Adventure,Fantasy', 'Action, Comedy,Thriller', ...']

apply_10_unique()

- First 10 unique values for column primaryTitle: ['Scrambled', 'Enys Men', 'Tayna amuleta', 'Fanon', 'Electric Malady', 'Kick', 'Season of Love', 'The Academy of Magic', 'Petta Rap', 'Everybody Loves Jeanne']

print_amount_unique()

- The number of unique values for column averageRating is 86

Exploratory Data Analysis Overview

- **Goal:** Understand data structure, patterns, and relationships
- Manual inspection helped detect inconsistencies missed
- Guided variable selection and model preparation

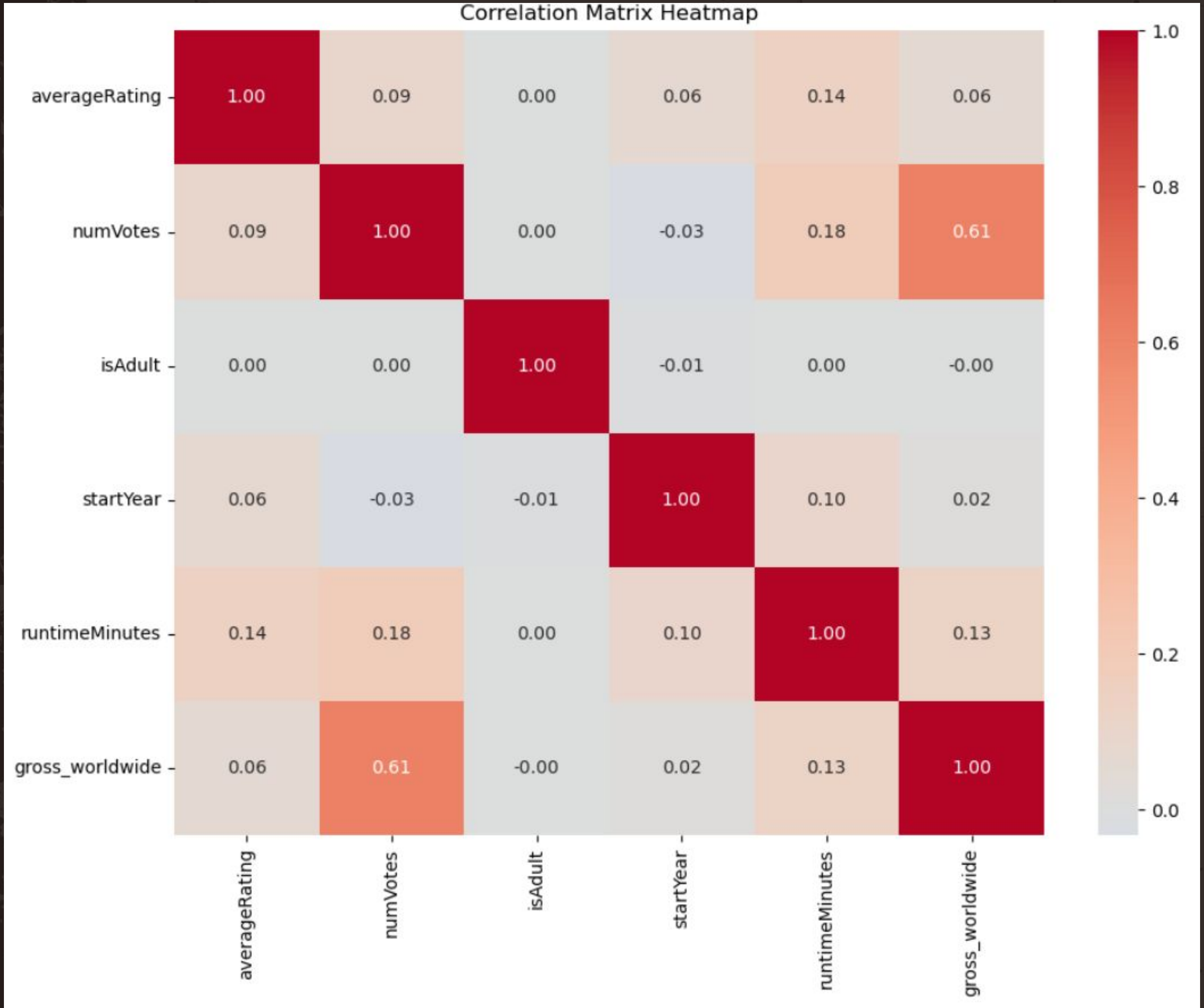
EDA Steps:

Checked dataset shape, size, dimensions, and types

Summary statistics: distributions, central tendencies

Visualizations: histograms, box plots, scatter plots

Correlation matrix and heatmap for variable relationships



averageRating: User ratings for the movie (1-10).

numVotes: Number of users who voted.

isAdult: Indicates whether the film is for adults.

startYear: Year of release (2020-2025)

runtimeMinutes: Duration of the movie in minutes.

gross_worldwide: Worldwide box office gross in USD.

Key Findings:

- gross_worldwide ⇔ numVotes: strong positive correlation (r = 0.61)

Weak correlations:

- runtimeMinutes (r = 0.13)
- averageRating (r = 0.06)
- averageRating weakly correlated with all variables

Other Observations:

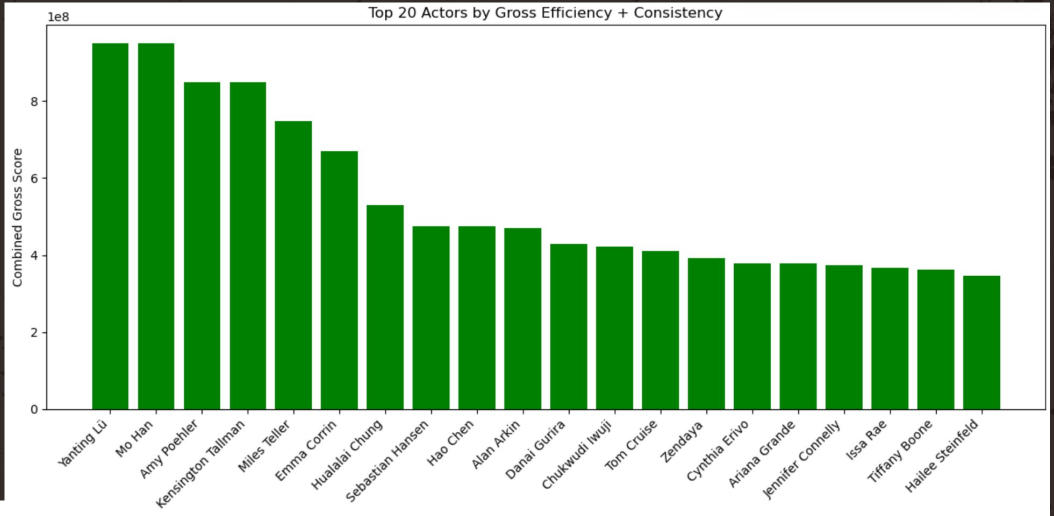
- runtimeMinutes correlates with:
  numVotes (r = 0.18)
  startYear (r = 0.10)

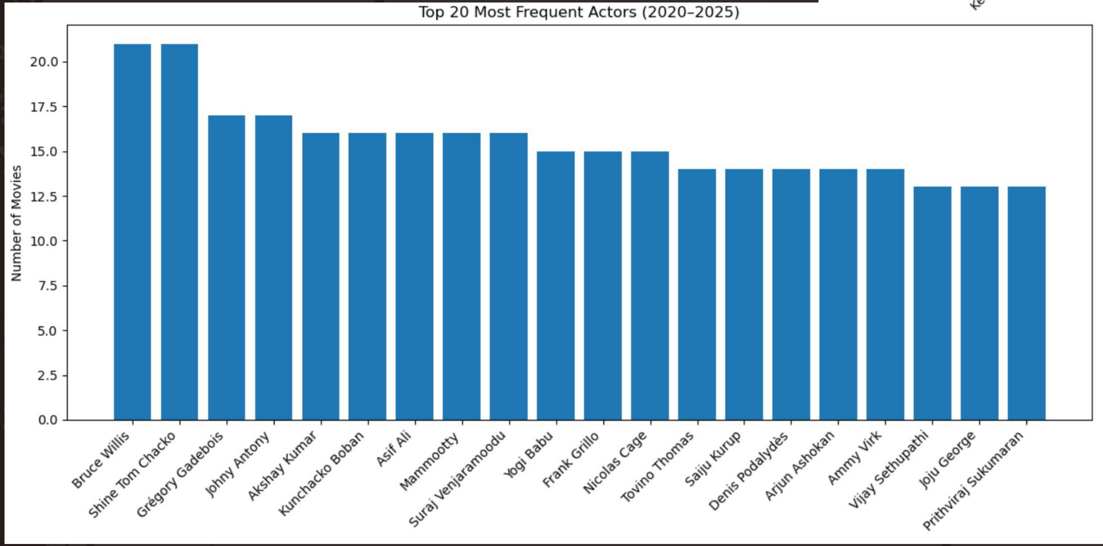
Why is averageRating weakly correlated?

- Ratings are subjective
- Niche/cult films: high ratings, low votes/gross
- Blockbusters: high votes/gross, not necessarily high ratings

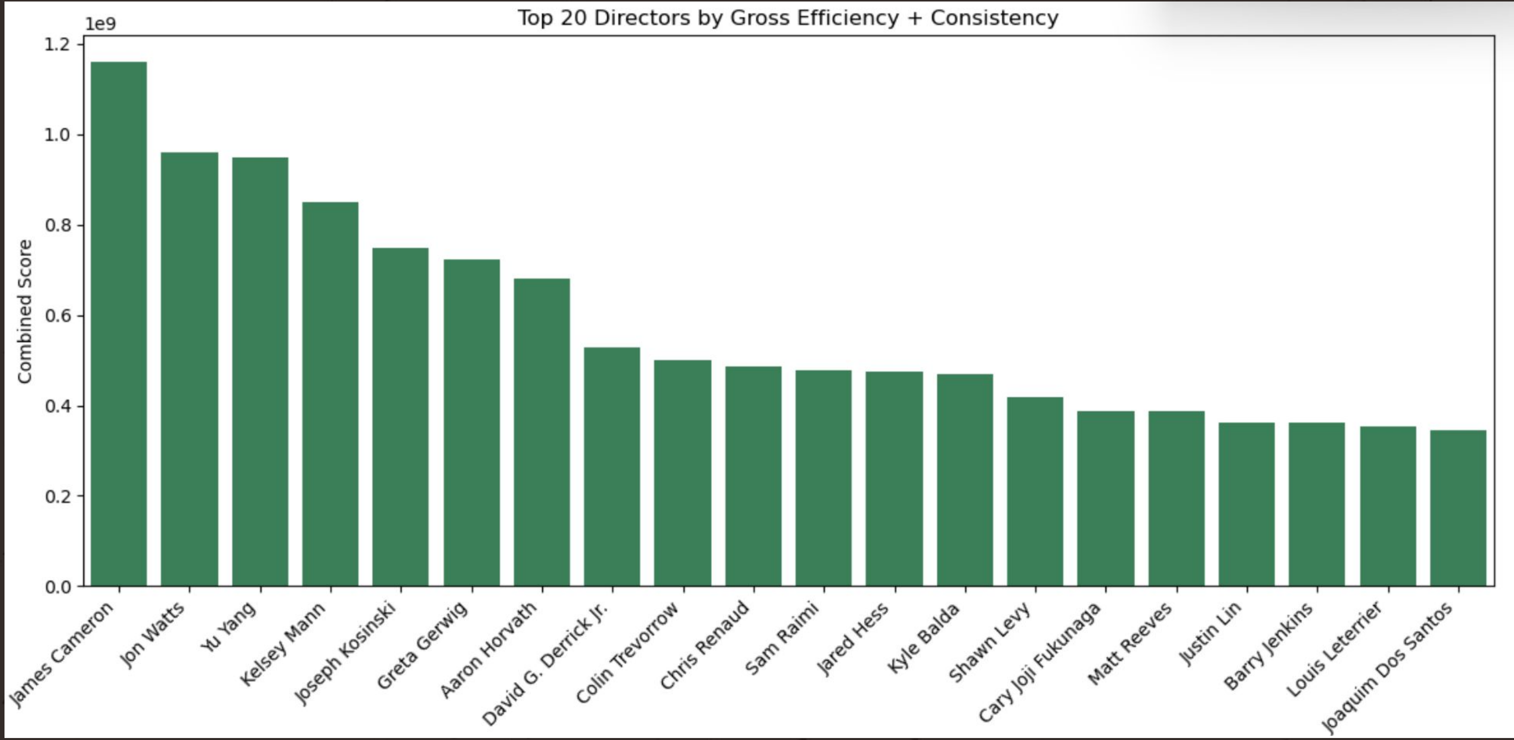
EDA is an iterative process → ongoing refinement

Actors


Top 20 Actors by Gross Efficiency + Consistency
1e8
Combined Gross Score
0
2
4
6
8
Yanting Lü
Mo Han
Amy Poehler
Kensington Tallman
Miles Teller
Emma Corrin
Hualalai Chung
Sebastian Hansen
Hao Chen
Alan Arkin
Danai Gurira
Chukwudi Iwuji
Tom Cruise
Zendaya
Cynthia Erivo
Ariana Grande
Jennifer Connelly
Issa Rae
Tiffany Boone
Hailee Steinfeld


Top 20 Most Frequent Actors (2020–2025)
Number of Movies
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Bruce Willis
Shine Tom Chacko
Grégory Gadebois
Johny Antony
Akshay Kumar
Kunchacko Boban
Asif Ali
Mammootty
Suraj Venjaramoodu
Yogi Babu
Frank Grillo
Nicolas Cage
Tovino Thomas
Saiju Kurup
Denis Podalydès
Arjun Ashokan
Ammy Virk
Vijay Sethupathi
Joju George
Prithviraj Sukumaran

Directors



Top 20 Directors by Gross Efficiency + Consistency
1e9
Combined Score
1.2
1.0
0.8
0.6
0.4
0.2
0.0
James Cameron
Jon Watts
Yu Yang
Kelsey Mann
Joseph Kosinski
Greta Gerwig
Aaron Horvath
David G. Derrick Jr.
Colin Trevorrow
Chris Renaud
Sam Raimi
Jared Hess
Kyle Balda
Shawn Levy
Cary Joji Fukunaga
Matt Reeves
Justin Lin
Barry Jenkins
Louis Leterrier
Joaquim Dos Santos

Target Variable

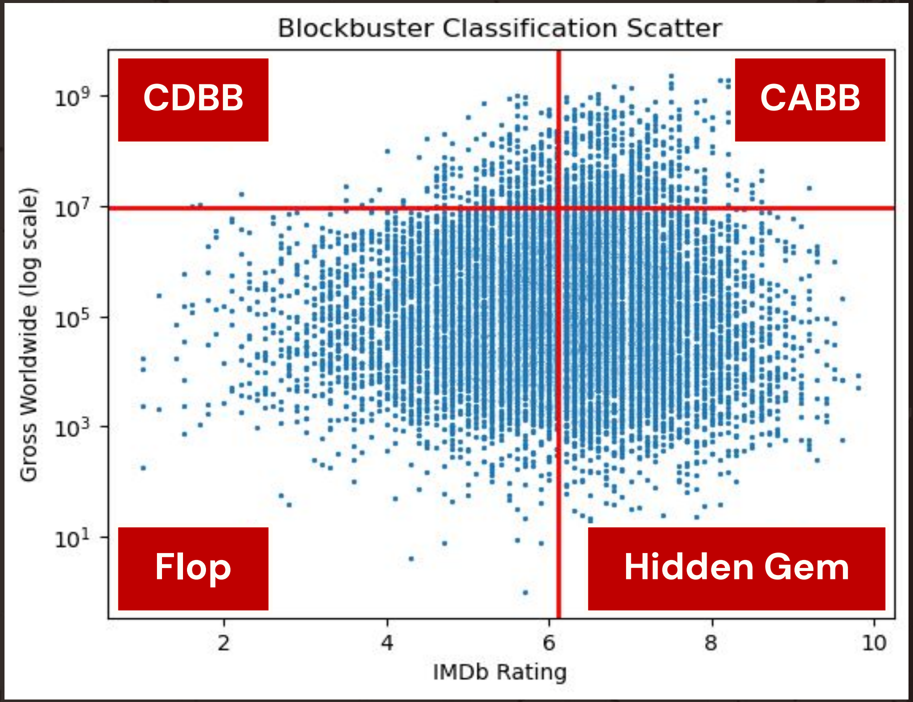
New column called movieType

Gross income boundary: 100 million USD (research online)

IMDB rating boundary: 6.2 (median)

Leads to significant class imbalance:

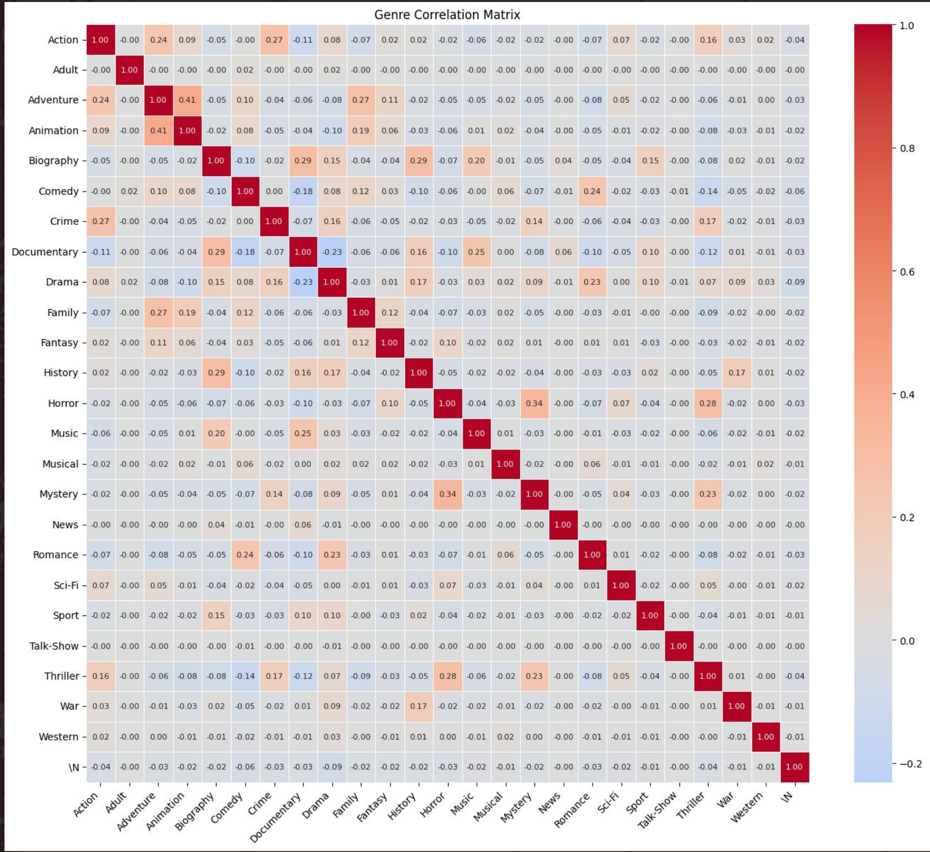
- Flop: 5910 (50.87%)
- Hidden Gem: 5474 (47.12%)
- Critically-Disliked Blockbuster: 73 (0.63%)
- Critically-Acclaimed Blockbuster: 160 (1.38%)

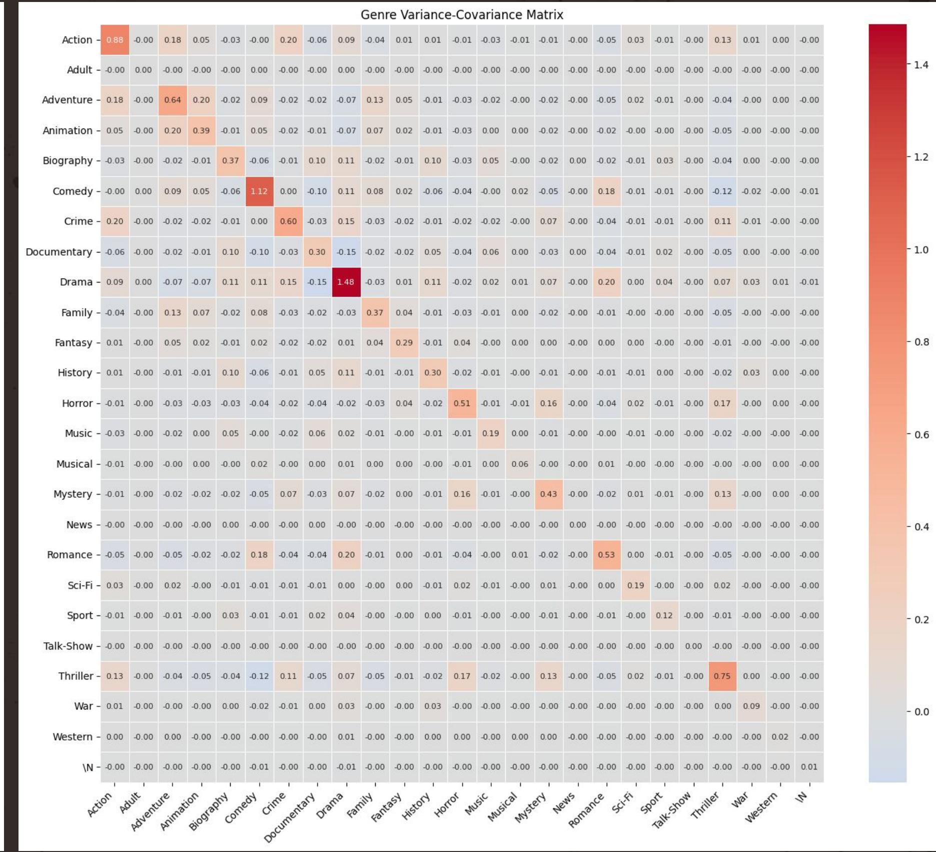


	Low Gross Income	High Gross Income
Low IMDB Rating	Flop	Critically-Disliked Blockbuster
High IMDB Rating	Hidden Gem	***Critically-Acclaimed Blockbuster***

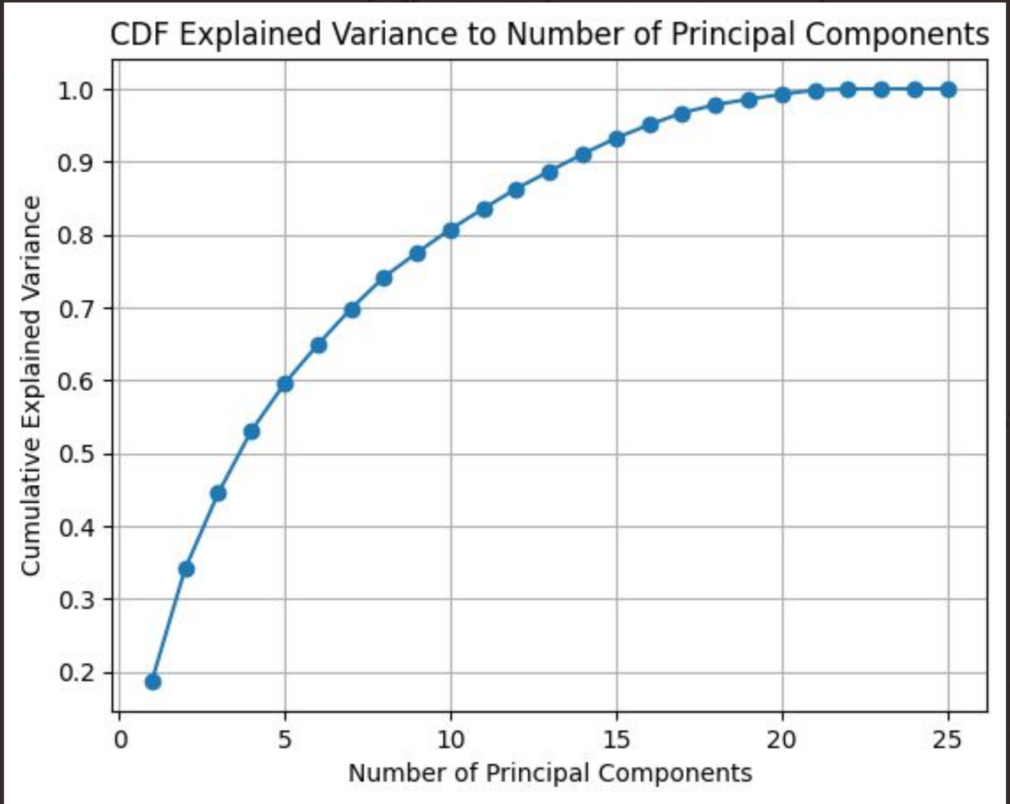
Preprocessing for Model (so far...)

1. Drop numVotes column
2. Encoding (Genres)
   - 25 Categories
   - Usually co occurring and somewhat covarying.
   - Multiple Hot Encoding
3. Feature Making from Natural Language Columns (Storyline, Themes, Primary Title)


Genre Correlation Matrix
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
News
Romance
Sci-Fi
Sport
Talk-Show
Thriller
War
Western
\N
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Genre Variance-Covariance Matrix
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
News
Romance
Sci-Fi
Sport
Talk-Show
Thriller
War
Western
\N
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0

Dimension Reduction of Genres



PC1: 0.1870 (18.70%)
PC2: 0.1535 (15.35%)
PC3: 0.1042 (10.42%)
PC4: 0.0855 (8.55%)
PC5: 0.0650 (6.50%)
PC6: 0.0532 (5.32%)
PC7: 0.0493 (4.93%)
PC8: 0.0437 (4.37%)
PC9: 0.0336 (3.36%)
PC10: 0.0320 (3.20%)
PC11: 0.0283 (2.83%)
PC12: 0.0277 (2.77%)
PC13: 0.0240 (2.40%)
PC14: 0.0234 (2.34%)
PC15: 0.0219 (2.19%)
PC16: 0.0183 (1.83%)
PC17: 0.0164 (1.64%)
PC18: 0.0111 (1.11%)
PC19: 0.0078 (0.78%)
PC20: 0.0064 (0.64%)
PC21: 0.0059 (0.59%)
PC22: 0.0017 (0.17%)
PC23: 0.0001 (0.01%)
PC24: 0.0001 (0.01%)
PC25: 0.0001 (0.01%)

What are the New Dimensions?

- Orthogonal Varimax Rotation to polarize factor loadings

PC 5 – Top 5 genres:

Adventure	0.651382
Family	0.433414
Animation	0.420057
Horror	0.295027
Romance	0.194875

PC 7 – Top 5 genres:

Thriller	0.941778
Horror	0.254712
Crime	0.113421
Comedy	0.099579
Mystery	0.089317

PC 8 – Top 5 genres:

Crime	0.828648
Horror	0.483166
Documentary	0.125824
Action	0.105045
Comedy	0.100730

PC 2 - Top 5 genres:

Comedy	0.904071
Romance	0.224629
Crime	0.206120
Horror	0.173700
Mystery	0.153393

PC 1 - Top 5 genres:

Drama	0.935124
Documentary	0.230906
Comedy	0.192865
Horror	0.152643
History	0.047907

Topic and Sentiment Analysis

- Columns that are not categorical or numeric
- Storyline (str), theme (sex, nudity, family, friendship, dog lover, amusement park, race…)(list of str), Primary Title (str)
- As many themes, primary titles, and storylines as there are movies!
- Multiple Hot Encoding won’t work on this raw data.
- But we can’t dispense with them, they are essentially the film’s entire content.
- Solution: Topic and Sentiment Analysis and then encoding.

Sentiment and Topic Analysis

- Solution: Topic and Sentiment Analysis and then encoding.”

```
theme_sentiment_label
neutral     7534
negative    2234
positive    1849
Name: count, dtype: int64
story_sentiment_label
positive    6368
neutral     3577
negative    1672
Name: count, dtype: int64
title_sentiment_label
neutral     9300
positive    1252
negative    1065
Name: count, dtype: int64
```

```
1    [movie, the, is, and, in, to, of, with, good, ...
click to scroll output; double click to hide, that, was, to, and, o...
3    [kids, animation, and, the, it, movie, to, for...
4    [and, the, in, spanish, of, by, as, with, de, is]
..                                                 ...
66   [love, romantic, comedy, sir, grudge, you, is,...
67   [you, movie, do, we, pun, have, it, other, int...
68   [veterans, ptsd, war, movie, army, not, he, di...
69   [than, itaposs, because, it, yet, tastefully, ...
```

```
                                      Representation  \
0     [relationship, based, on, friendship, word, ti...
1     [nudity, frontal, rear, full, male, female, to...
2                                [, , , , , , , , , ]
3     [cancer, abuse, pregnancy, abusive, opioid, di...
4     [relationships, brother, family, parents, rela...
..                                                 ...
314                              [, , , , , , , , , ]
315   [basement, airbnb, cake, beans, maggot, jerusa...
316                              [, , , , , , , , , ]
317                              [, , , , , , , , , ]
318   [group, friend, screenlife, success, wrong, ev...
```

03

Plans Moving Forward

Brainstorming the Model

XGBoost:

- Classification
- Accurate model
- Handles mixed data types
- Is interpretable
- Is efficient

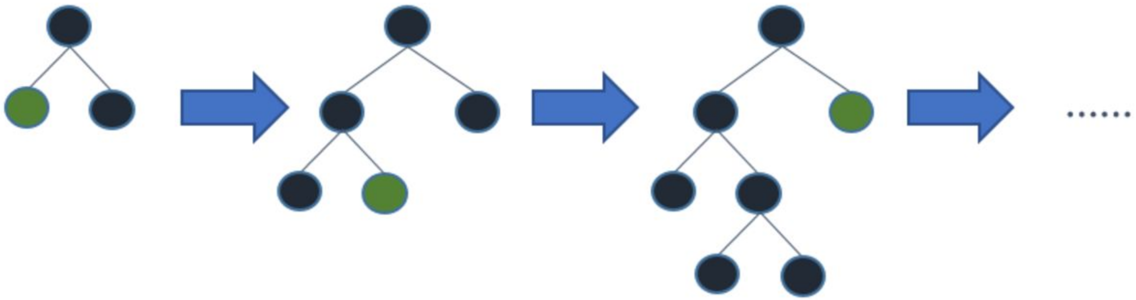
LightGBM:

- Classification
- Accurate model
- Handles mixed data types
- Is interpretable
- Is efficient

LightGBM

Base model → Error → Train next → Repeat → Combine

- Light gradient boosting machine
    - Errors from the previous model is used to train a new one
        - Chooses a leaf with most error to grow it vertically
    - At the end the best model is a combination of all models
- Uses decision trees → **splits the tree leaf wise**



Leaf-wise tree growth

LightGBM

Big advantage:

- Handles mixed data types
  - Integer code for categorical data, focuses on categorical data splitting

Disadvantages:

- Prone to overfitting
  - max_depth to limit tree height
- Slightly less interpretable
  - SHAP values

Thank you!

Any questions?