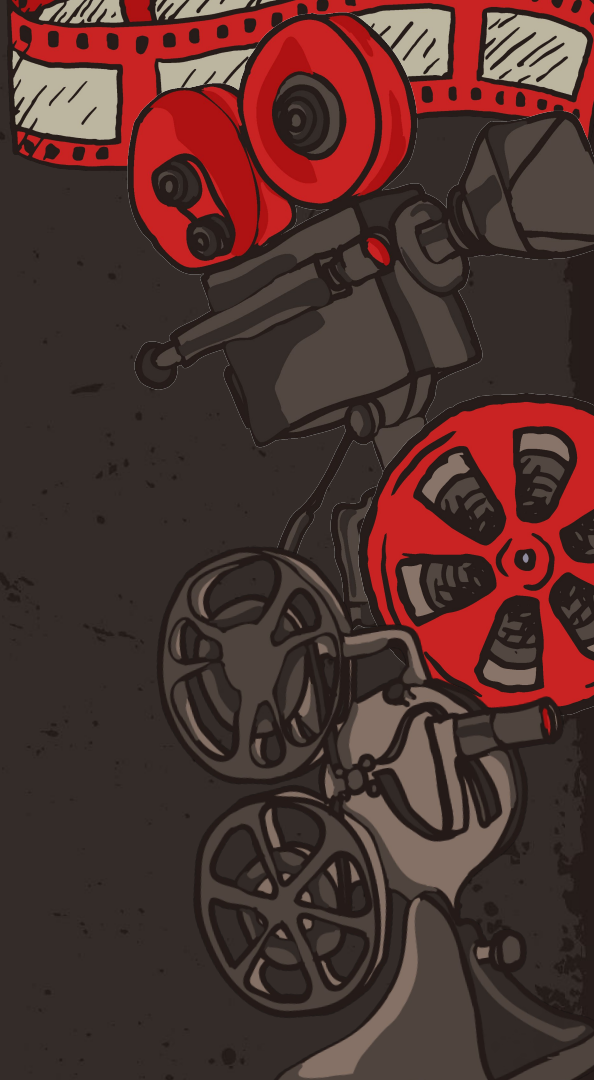
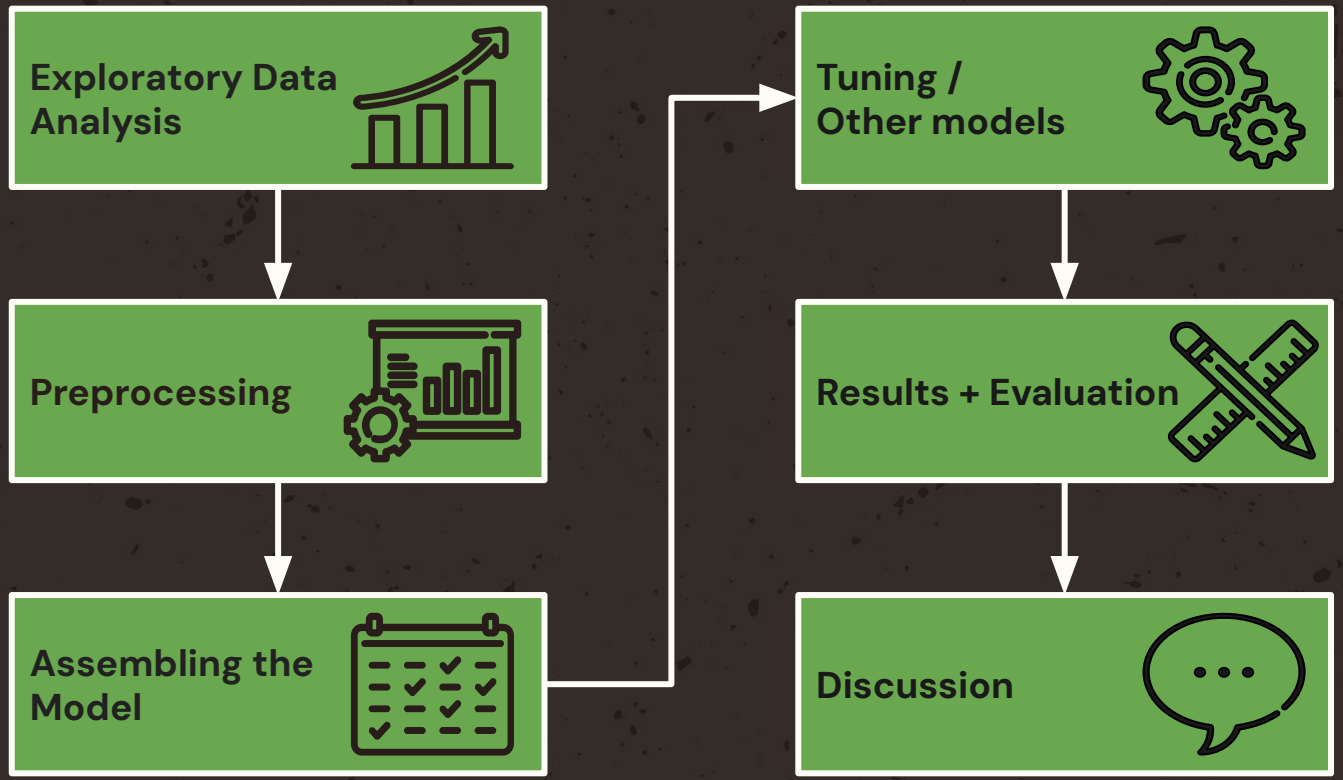
Assembling the Next Blockbuster

Team Parrot 
Leon, Julia, Mehmet, and Leticia

01

Interim Recap




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

Research Question

Original RQ: 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 RQ: 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	***Critically-Acclaimed Blockbuster***

EDA

- Webscraped 54,095 movies (released from 2019 onwards)

- 16 columns
- **Features include:**
  - Numerical: runtime, gross income, average IMDB rating
  - Multilabel: genres, themes, actors
  - Nominal: director

- Unique values
- Frequency Distributions
- Correlation Matrices

02

Preprocessing Data (Old and New)

Themes and Genres

Themes

- Topic and sentiment analysis
  - Turn themes into categories.

Genres

- Multiple hot encoding?
- Dimension reduction?
- Actually Ordinal Encoding
  - LightGBM supports categorical data!

Number of meaningful topics found: 140

	Topic	Count	Name
0	-1	1880	-1_relationship_father_son_nudity
1	0	1511	0_unfortunate_right_wrong_
2	1	190	1_ghost_supernatural_horror_terror
3	2	153	2_gay_interest_kiss_homosexual
4	3	102	3_protagonist_directed_by_girl
..	...	...	...
136	135	10	135_night_fall_friend_job
137	136	10	136_son_mother_demented_grandparents
138	137	10	137_son_army_person_father
139	138	10	138_rape_abduction_revenge_mental
140	139	10	139_india_asia_airforce_tragedy

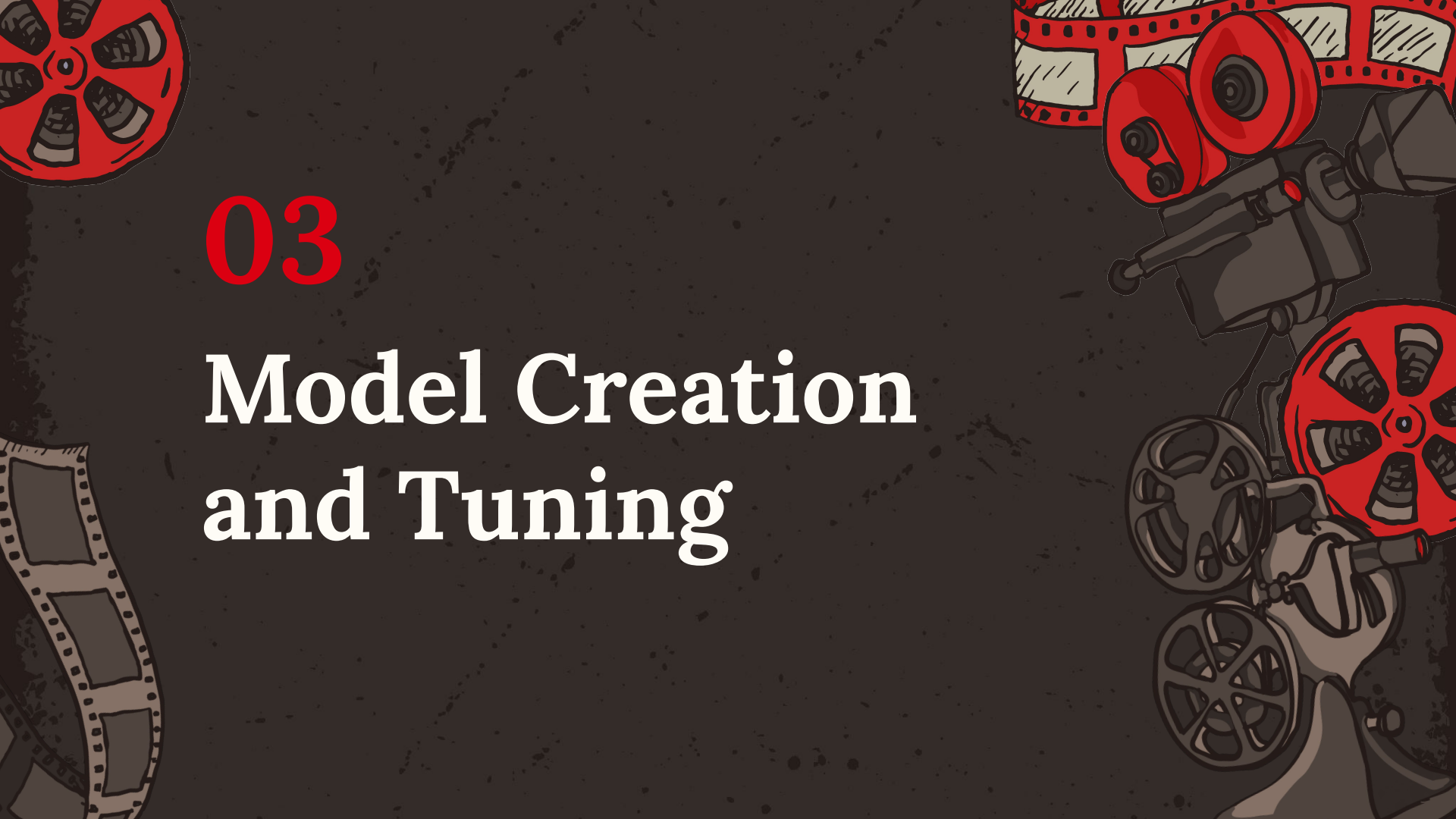
Actor and Director Encoding

Frequency Encoding:

- Actor/Director replaced by their number of appearances

Target Encoding:

- Average gross + average rating
- New actors/directors in the test data receive default value
    - Average gross income of actors/directors that only have one movie in the training data

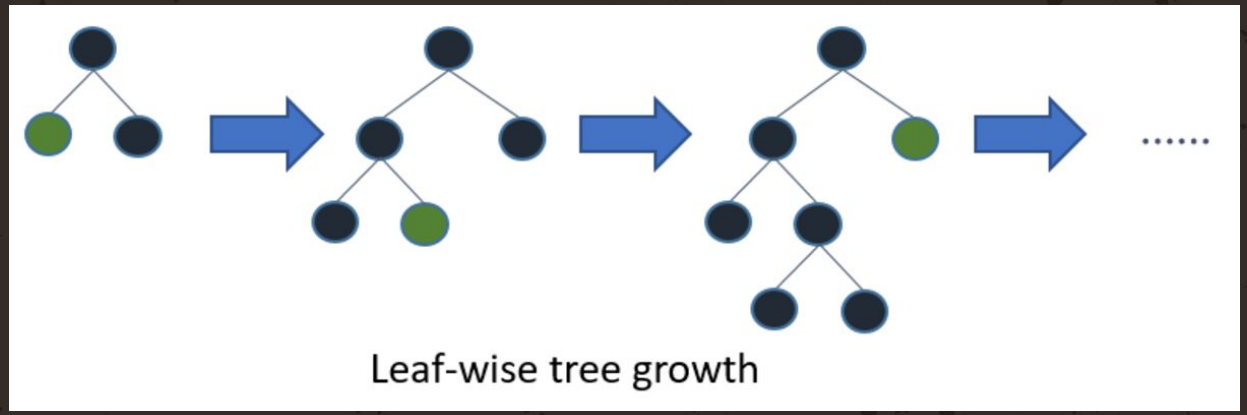


03

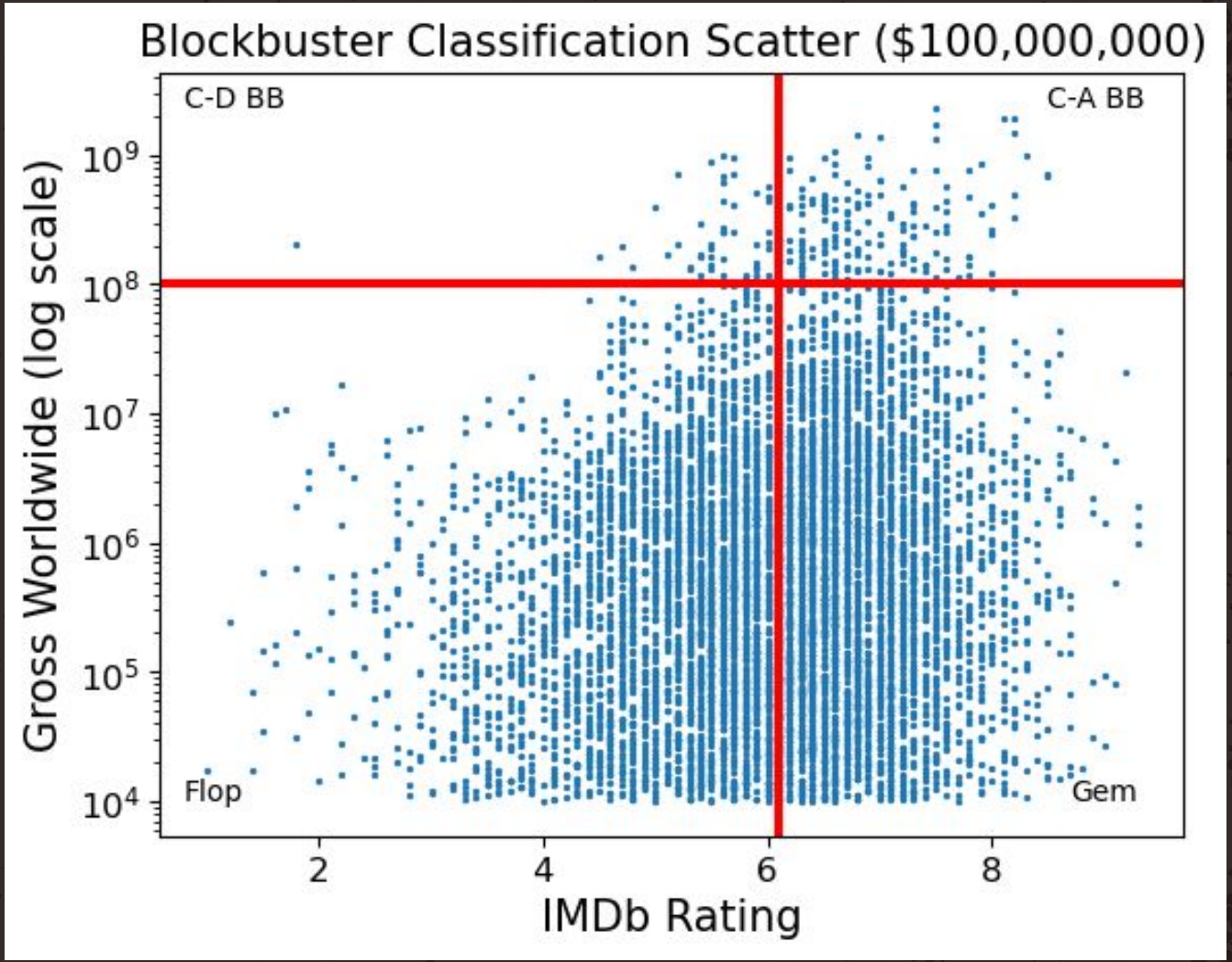
Model Creation and Tuning

LightGBM

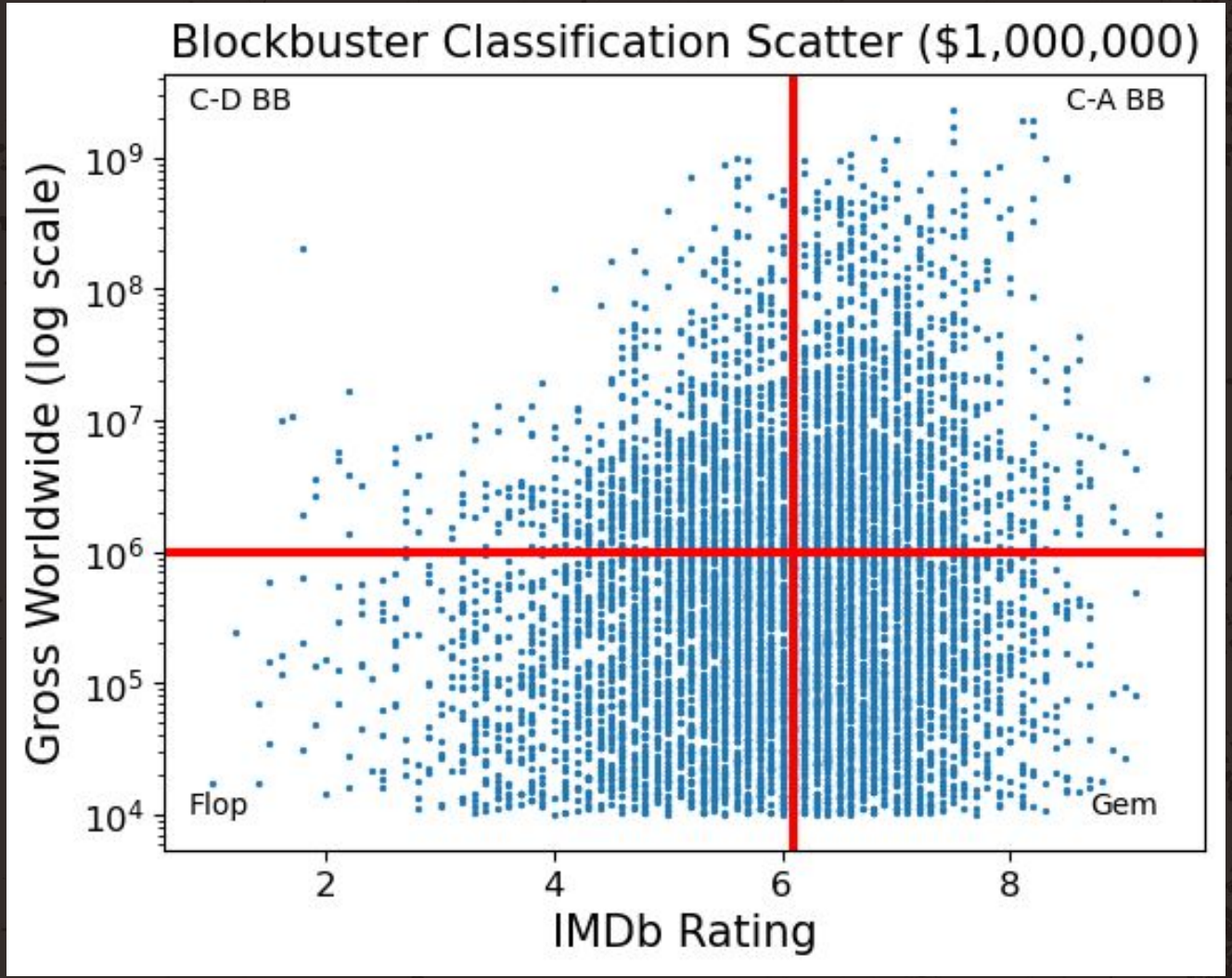
- Light gradient boosting machine with decision trees
  - Errors from the previous model is used to train a new one
- Handles mixed data types
- Low interpretability → fixed with SHAP



Class division of movies for classification with $100 million dollar threshold



Class division of movies for classification with $1 million dollar threshold



Models

Features used: averageRating, isAdult, startYear, runtimeMinutes, theme_sentiment_label, theme_topic_label, Action, Adventure, Animation, ... , director_avg_grossWorldwide, director_avg_averageRating, actor_avg_grossWorldwide, actor_avg_averageRating

Models:

- LightGBM Regression → TV = grossWorldwide
- LightGBM Classification ($1 million threshold) → TV = movieType1M
- LightGBM Classification ($100 million threshold) → TV = movieType100M

Regression

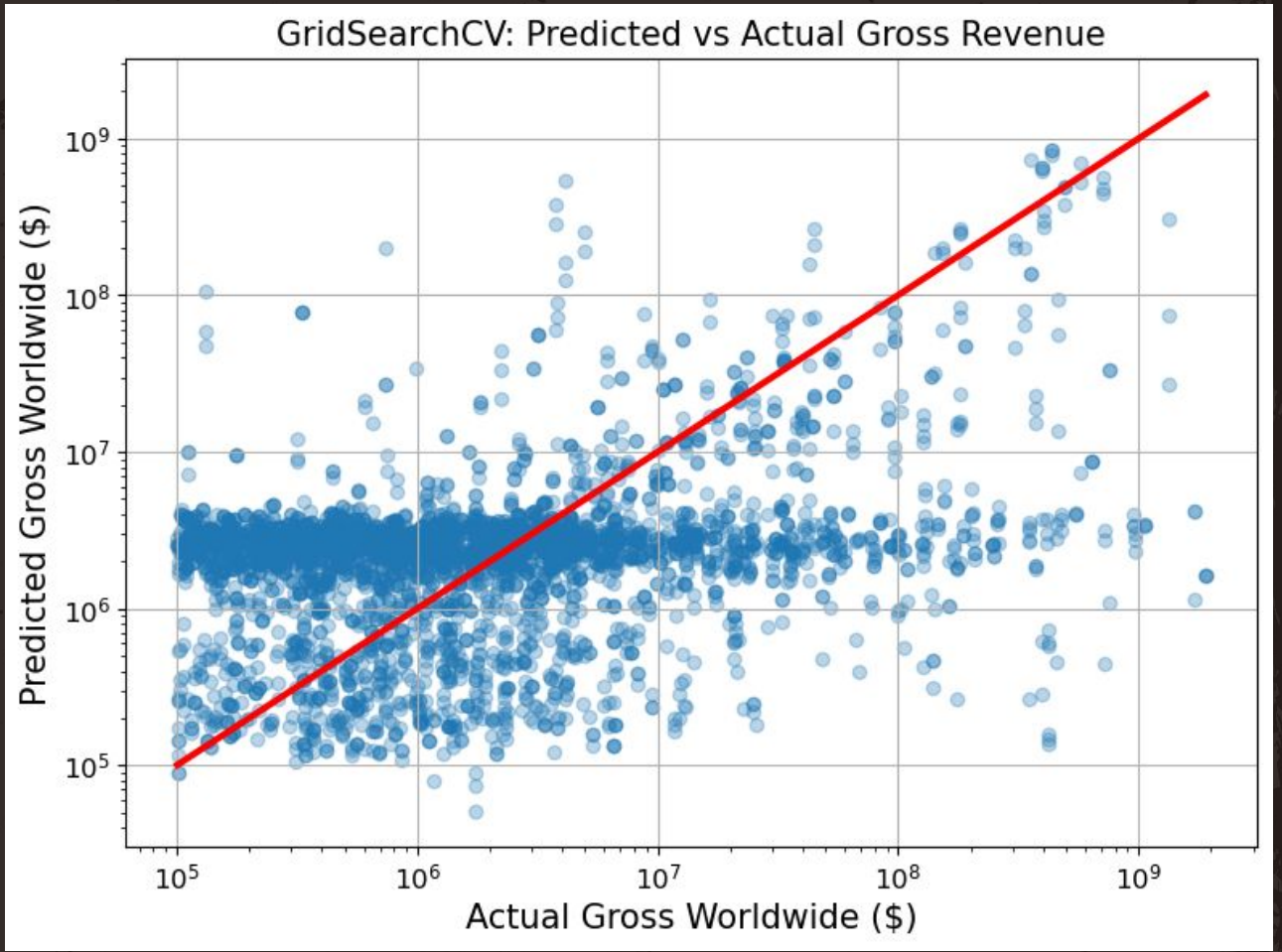
Outputs predicted gross income

Explanation

- Default values for actors and directors

RMSE

- 1.9 in log scale
- 100 million

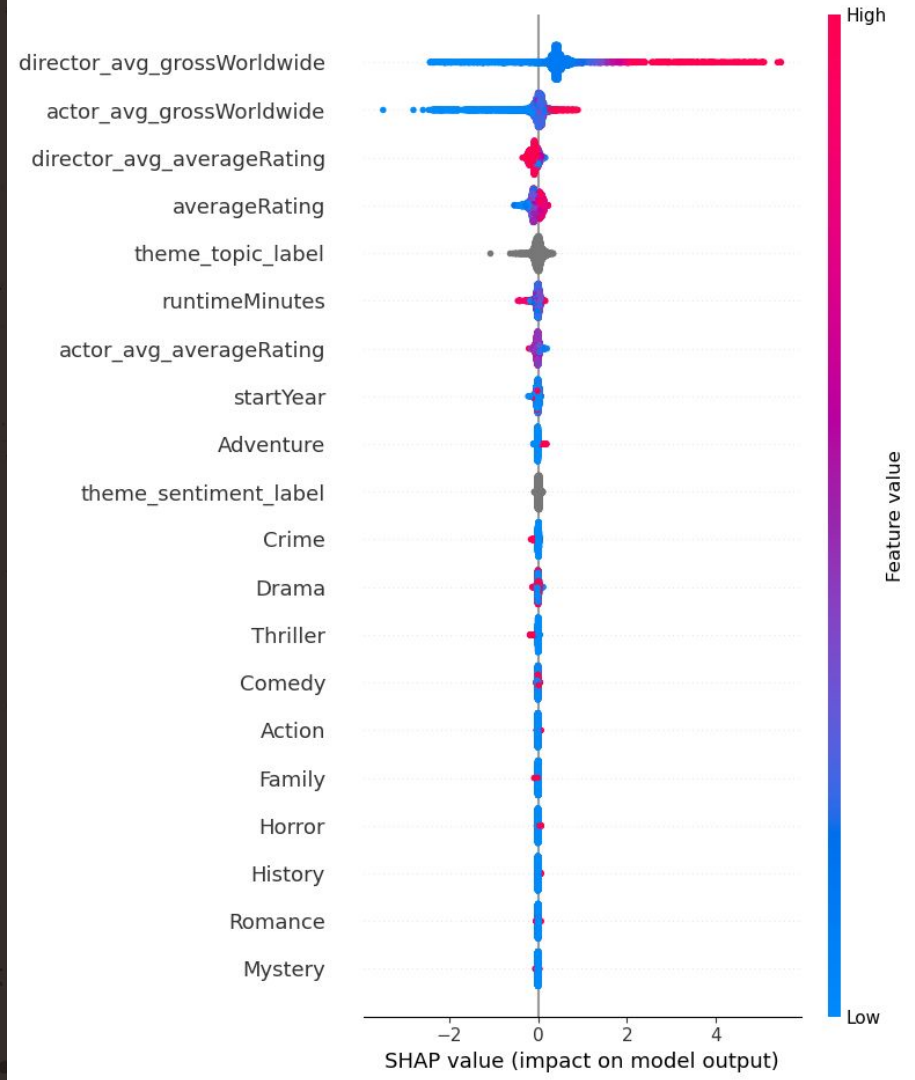


SHAP Values

- Measure how much each feature contributes to the result
- Determine feature importance

Graph explanation

- Pink = positively influences prediction
- Blue = negatively influences prediction
- The longer the line, the more influence



Hyperparameter Tuning

- Hyperparameter optimization
  - RandomSearchCV & GridSearchCV
- num_leaves, max_depth, learning_rate, n_estimators, feature_fraction, bagging_fraction
- What is RMSE and why it helped
  - Lower RMSE = better model 2.70 -> 1.95

Method	What it Does	Pros	Cons
RandomizedSearchCV	Samples random combinations of hyperparameters from defined ranges	Faster for large search spaces	May miss the optimal combination
GridSearchCV	Exhaustively tests all combinations from a predefined grid	More thorough and reliable	Can be slow with many combinations

Class Numbers

Class 0: Critically Acclaimed Blockbuster

Class 1: Critically Disliked Blockbuster

Class 2: Hidden Gem

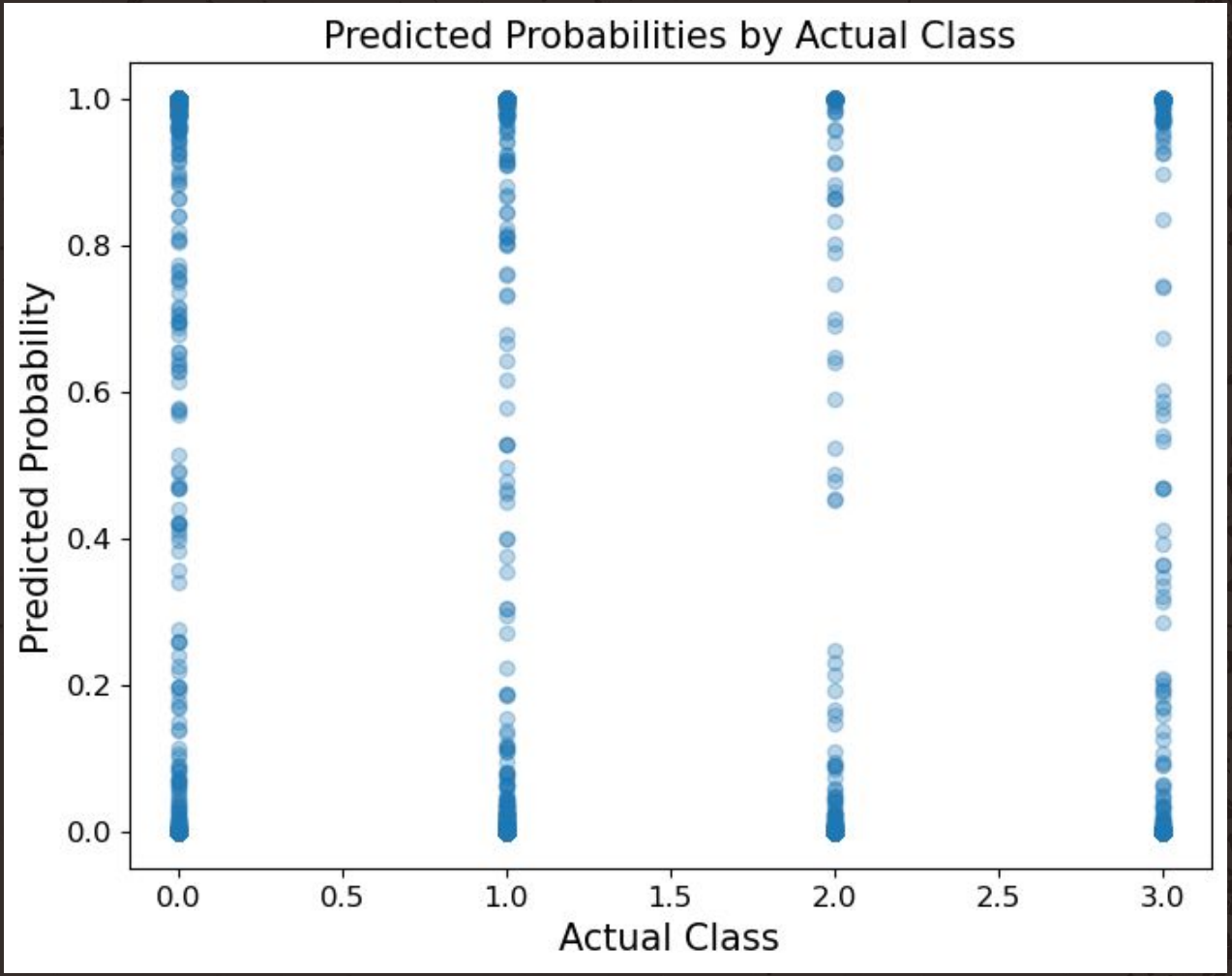
Class 3: Flop

Classification $1M

Outputs class probability

Results

- CAB(0): 0.805
- CDB(1): 0.828
- Flop(2): 0.303
- Hidden gem(3): 0.209

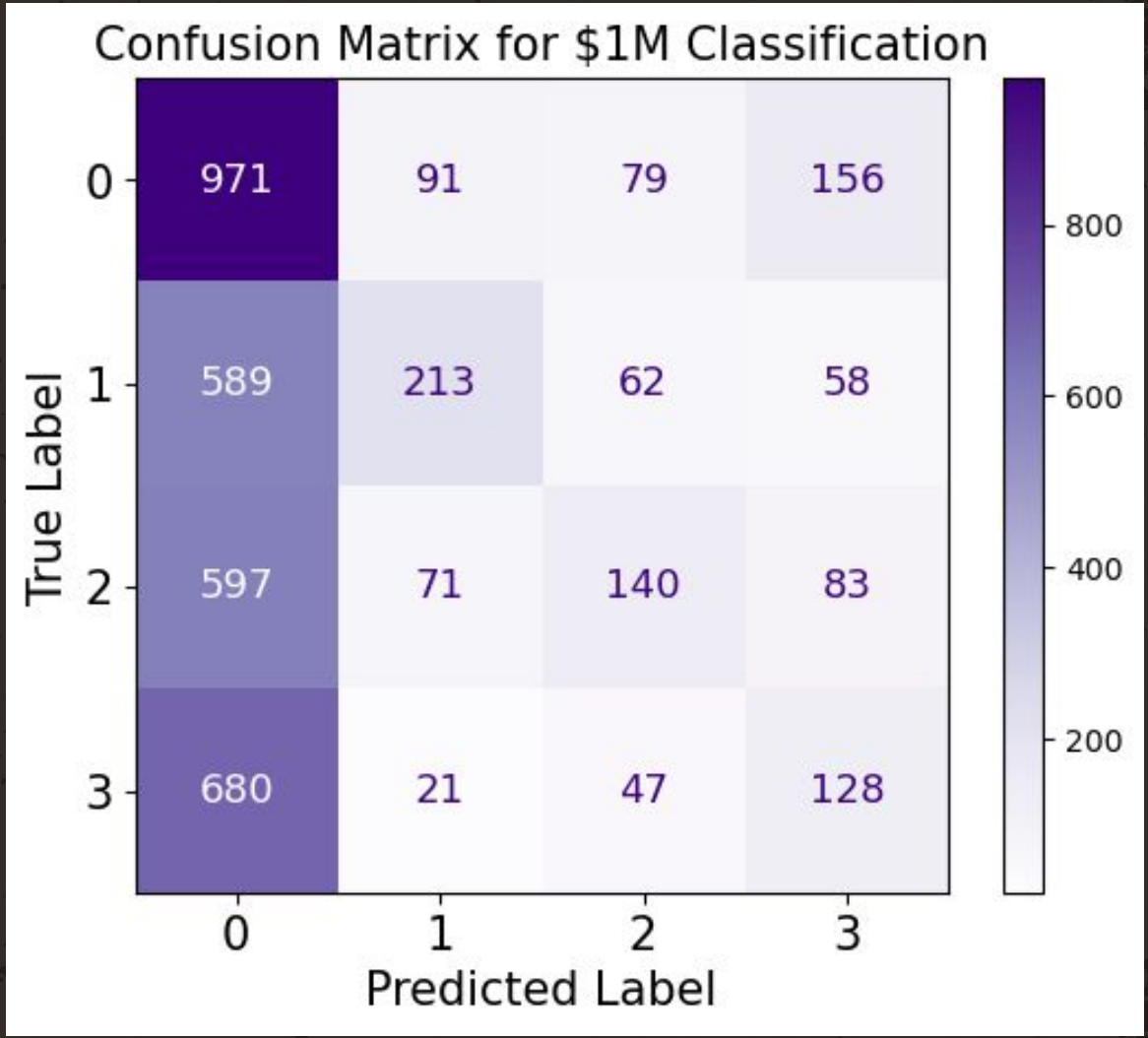


Classification $1M

Outputs class probability

Results

- CAB(0): 0.805
- CDB(1): 0.828
- Flop(2): 0.303
- Hidden gem(3): 0.209



Classification $100M

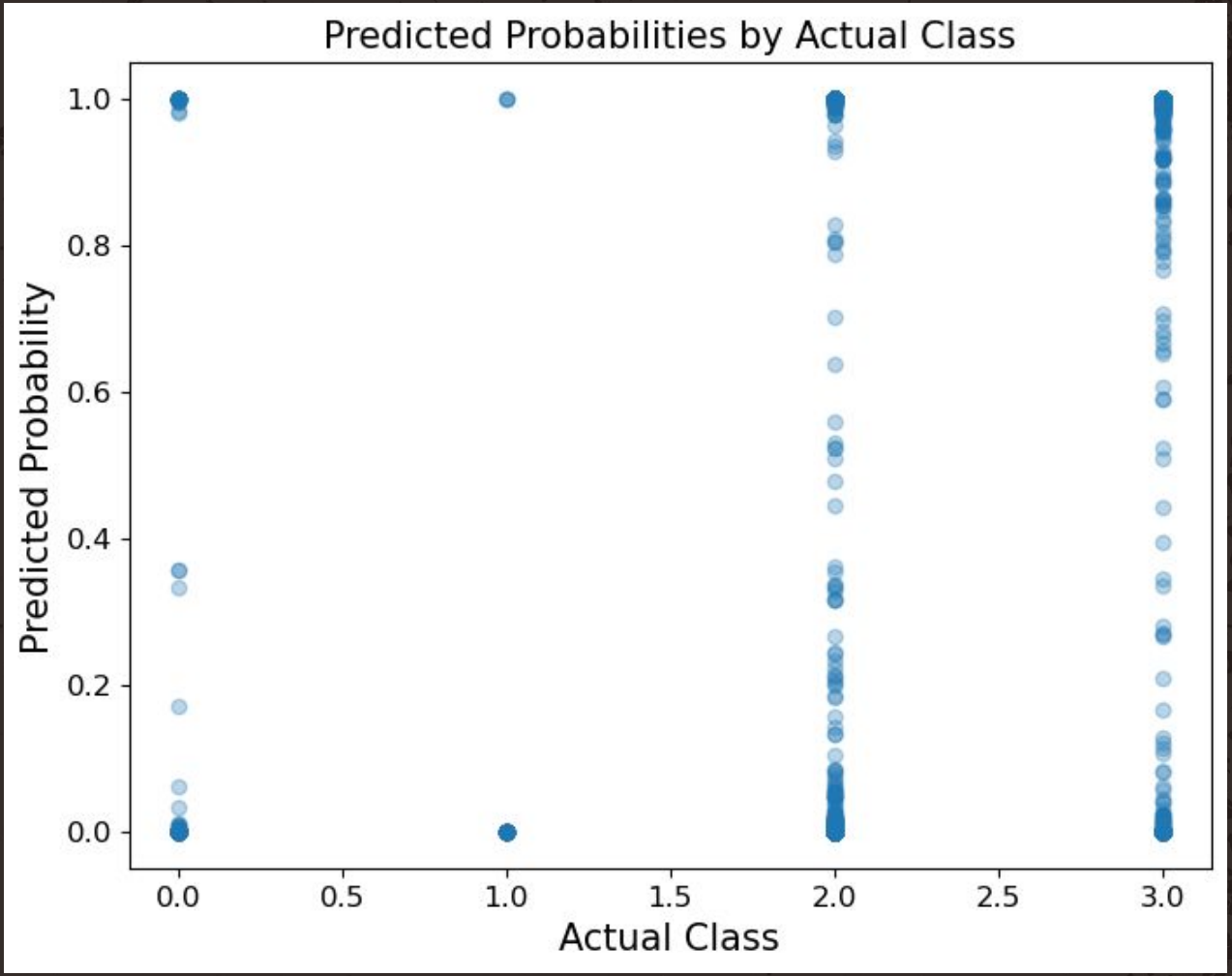
Outputs class probability

Explanation

- Class imbalance

Results

- CAB(0): 0.214
- CDB(1): 0.128
- Flop(2): 0.997
- Hidden gem(3): 0.990



Classification $100M

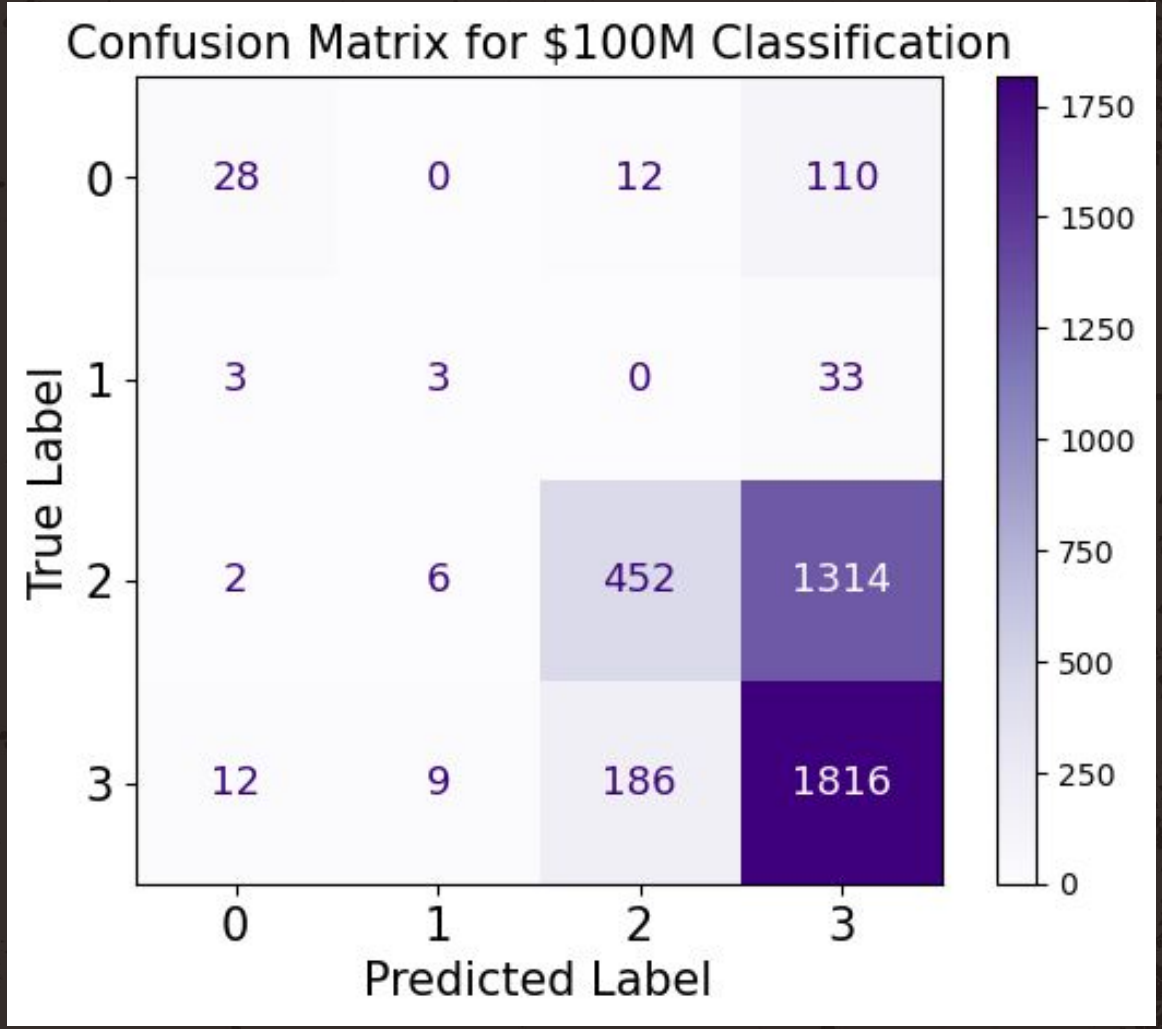
Outputs class probability

Explanation

- Class imbalance

Results

- CAB(0): 0.214
- CDB(1): 0.129
- Flop(2): 0.997
- Hidden gem(3): 0.990



Discussion

- Goal: Predict whether a movie becomes a blockbuster and figure out what makes one
- Commercial success -> budget, marketing, timing, franchise.
    - Not as available on IMDb
- Proxy for budget -> average revenues of actors and directors
- Skewed distribution
- Labeling ambiguity

If We Had More Time...

- Add Variance of Average Box Office and Rating of Actors and Directors (Incorporate risk analysis for client)
- Implement different buzz mechanisms' importance for success (marketing, TV)
- Include Trends (What type of movies get famous now? Has that been changing with time?)
- Use the budget approximation for more real-world accurate results

Thank you!

Any questions?