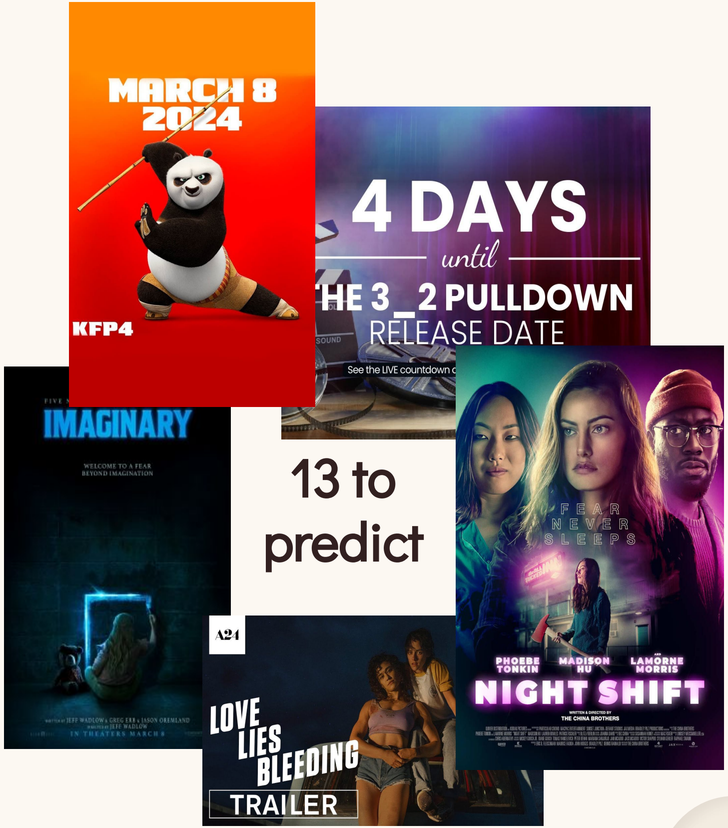
# Project 5: Predicting Movie Box Office Gross

Ayato Hisanaga
Geethu Devarajan
Ivy Chen
Madhusmita Oke
Shan Ming Gao
Vera Hu

# 01 Topic Introduction

# Overview

- **Project Goal:** Use  ML to predict the box office revenue of movies set to premiere in the final week of the quarter in the United States, specifically between  March 4th and 10th.
- **Potential  Business Audience:**  Key decision-makers in the entertainment industry, such as movie studios, distributors, and cinema chains.
- **Practical  Explanation and Value:**
  - Assist stakeholders in making informed decisions
  - The ability to accurately predict box office grosses can significantly impact business outcomes.  It can guide decisions on advertising budgets, release strategies, and production investments, leading to increased revenue and profitability
- **Potential Additional  Learnings from the Analysis:** Uncover trends in movie performance, audience preferences, and the impact of various features on box office revenue. These insights could inform long-term business strategies and improve decision-making processes.




13 to predict

# 02 Methodology

# Data Sources & Collection

Time range: 2010-2024    Region: USA

1 **IMDB:** IMDb provides a Python package **Cinemagoer** for retrieving the data of the IMDb movie data about movies, people and companies.

2 **TMDB:** TMDB provides an API on data from TMDB database, including IMDb ID that allows us to join data from different sources.

3 **Web Scraping :** Web Scraping of Wikipedia pages was done to collect data about awards won and nominations.

4 **RottenTomatoes & Metacritic:** Used for critic review information collection

# Features Collected

Number of features: 13    Number of rows: 1418

| Feature | Description | Data Type |
|---|---|---|
| imdb_id | IMDb movie ID | int |
| original_title | Original title of the movie | string |
| release_date | Release date of the movie | datetime |
| budget | Movie budget in USD | float |
| runtime | Runtime of the movie in minutes | int |
| popularity | TMDB metric for lifetime user engagement | float |
| vote_average | TMDB movie rating | float |
| vote_count | # of user votes received by a movie on TMDB | int |

# Features Collected

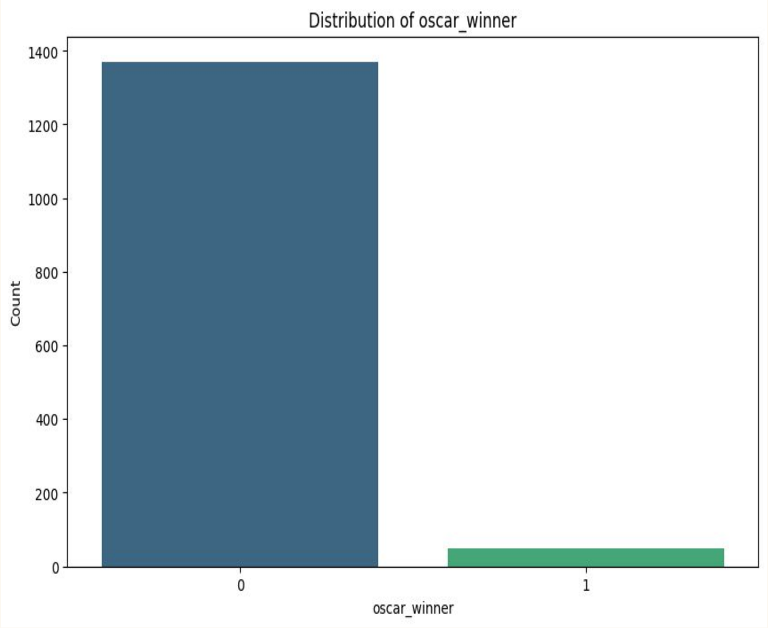
Number of features: 13    Number of rows: 1418

| Feature | Description | Data Type |
|---|---|---|
| ratings | IMDb movie rating | float |
| genre_list | A list of genres the movie belongs to | string |
| oscar_winner | Binary values, 1 meaning the movie won at least one oscar | int |
| oscar_noms | Binary values, 1 meaning the movie got nominated for oscar | int |
| box_office | Global movie box office in USD | float |
| earn_class | Based on Return on Investment (%) (loser, earner, super-earner). Super_earner at 90th percentile. | string |

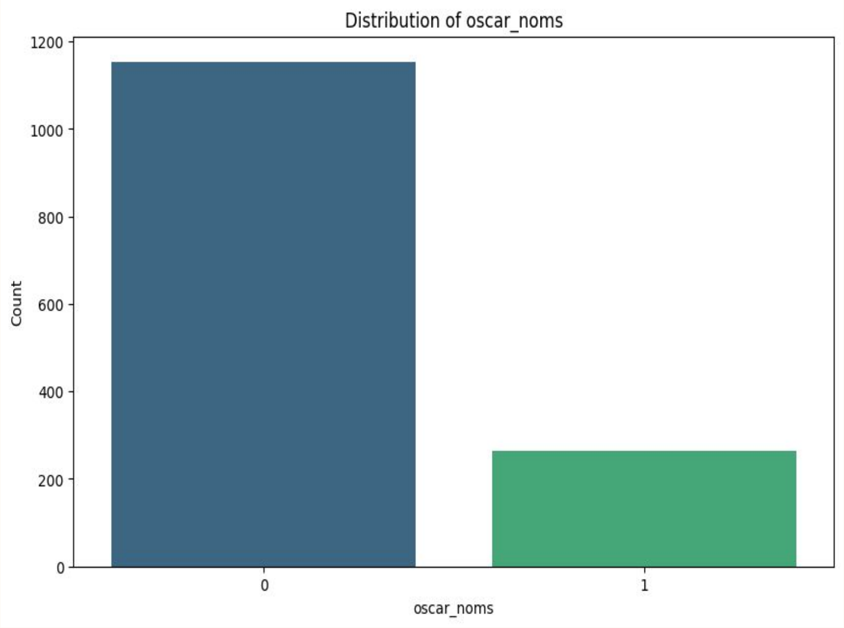
# Data Cleaning

- **"box_office"**: Stripped "$" in the original data and transformed it into numeric values
- **"budget":** Filled 180 null values with median value per genre
- **"genre_list":** Transformed "genre_list" into binary values, with each genre as a single column
- **"release_date":** Broke down "release_date" into year, month and day of week.
- **Post-release data in test set**: 'popularity', 'vote_average', 'vote_count' and 'ratings' will be available after release, so we estimated these values from genre medians and means.

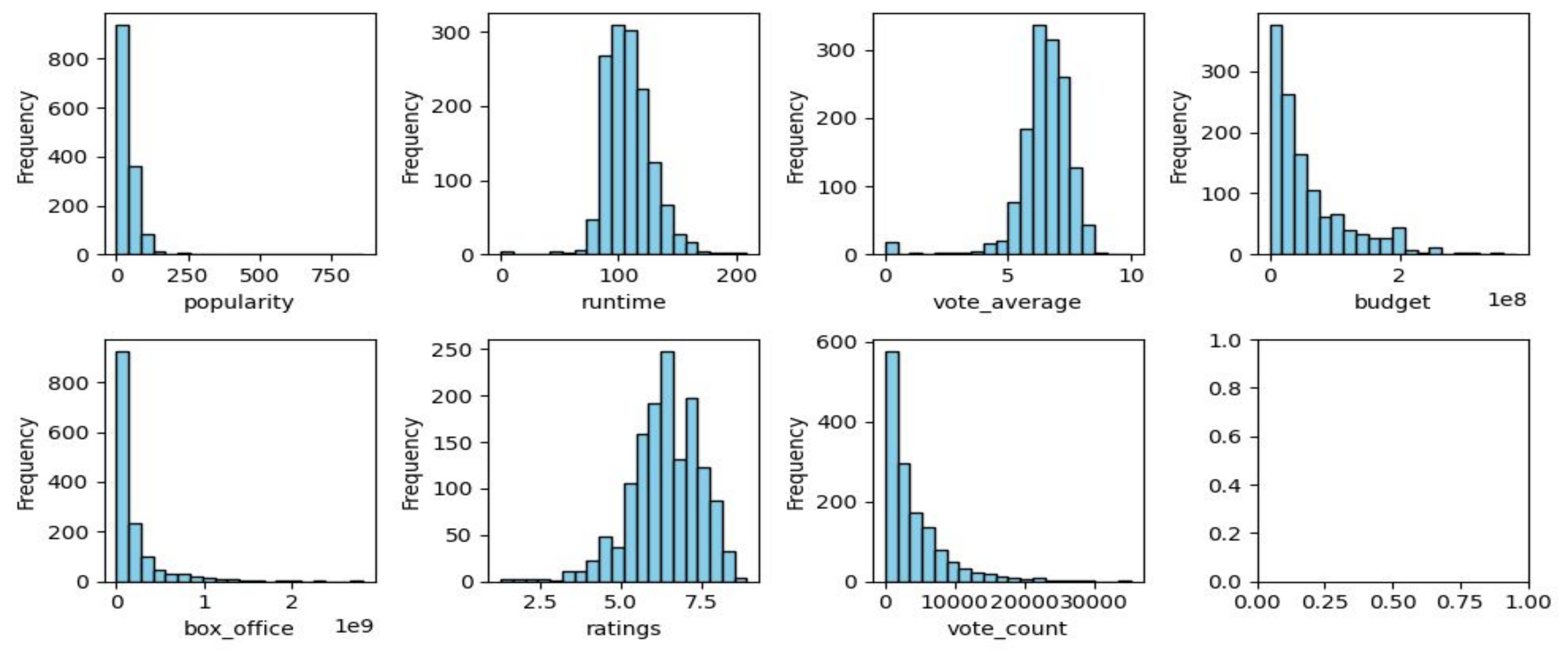
# Oscar Winning Movies



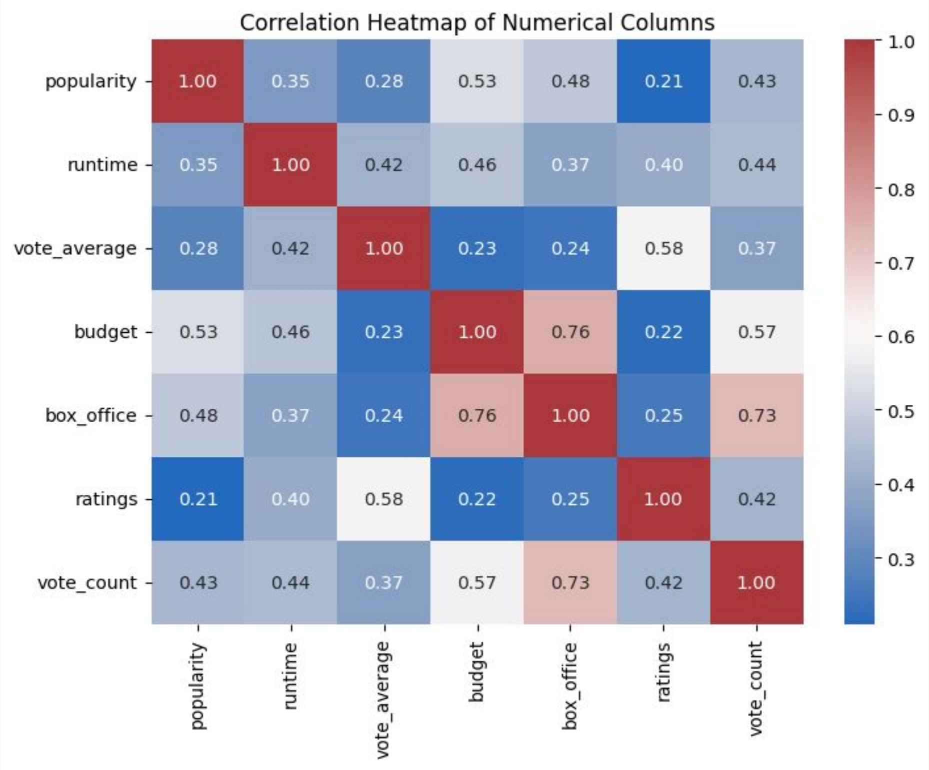

# Oscar Nominated Movies



# Data Exploration – Basic Distribution



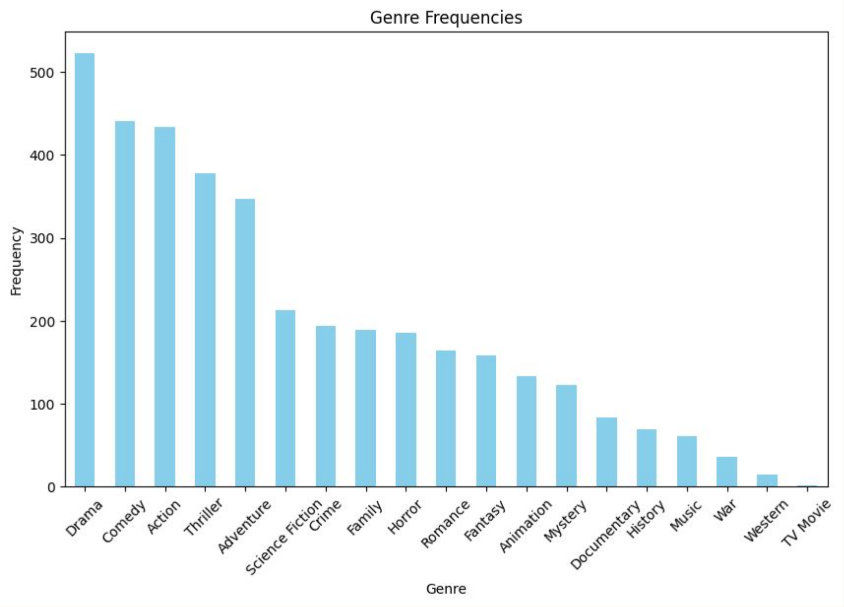
# Data Exploration – Correlation Matrix



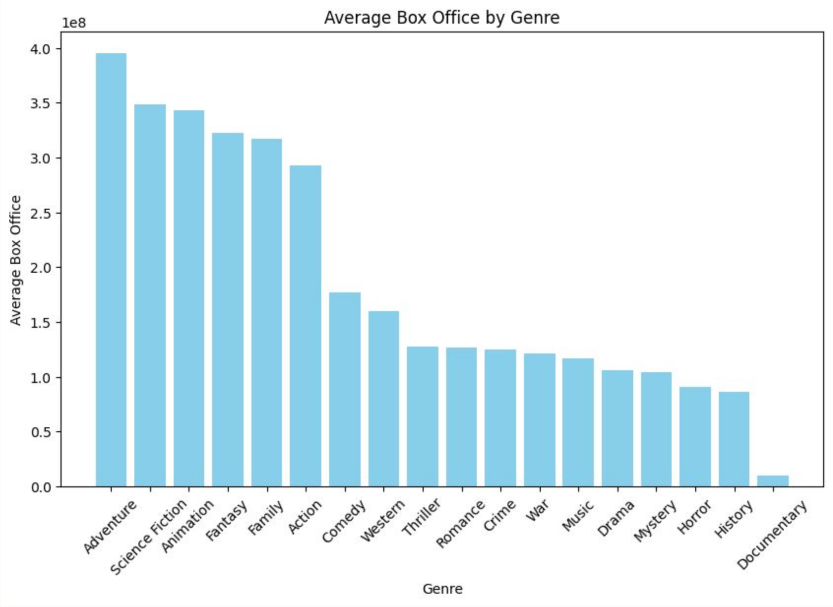

**Features mostly related to box office:**
budget, vote_count

**Features least related to box office:**
vote_average, ratings

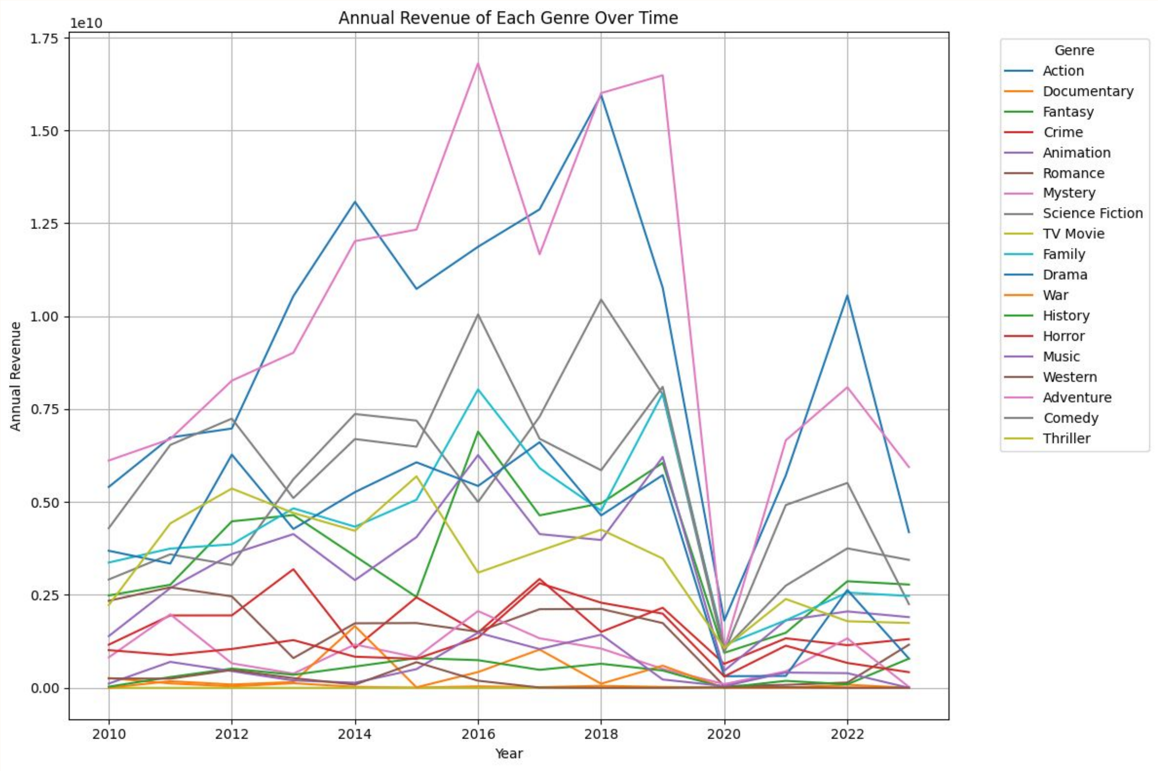
# Distribution of Genres



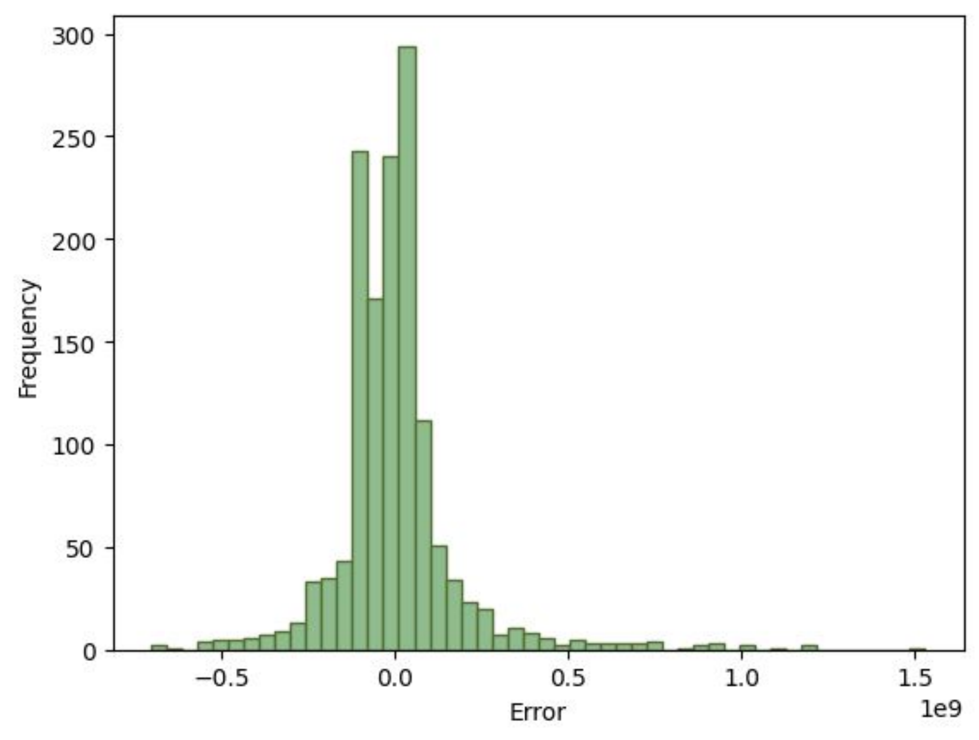

# Average Box Office by Genre



# Annual Revenue by Genre



# Data Exploration – Baseline Model




Simple Linear Regression model with "budget" as predictor variable

- **R Square:** 0.577

# 03 Compare Models

# Techniques and Models

## Continuous: it's our main problem

- Linear Regression
- kNN
- XGBoost Regressor

## Categorical

- Decision Tree and Random Forest

# Linear Regression model

| Movie | Predicted Gross Revenue |
|---|---|
| Love Lies Bleeding | $534,591,335 |
| Kung Fu Panda 4 | $369,112,691 |
| Imaginary | $91,714,640 |
| Cabrini | $100,119,156 |
| Accidental Texan | $21,756,308 |
| The 3_2 Pulldown | $70,817,789 |
| American Dreamer | $66,927,417 |
| Night Shift | $80,041,821 |
| 5lbs of Pressure | $107,867,255 |
| The Ballad of Davy Crockett | $235,208,981 |
| Space: The Longest Goodbye | $3,654,450 |
| The Piper | $86,765,660 |
| Glitter & Doom | $65,733,534 |

- Standardize features to ensure consistency between training and prediction
- Training model: **R-squared of 0.747**
- Predicted gross revenue for **13 upcoming movies**, with estimates ranging from **$3M to $534M**
- Emphasizes the importance of financial and social metrics in driving box office success

# kNN Regressor

- Effectively group movies that share similar characteristics and performance trends

- Capture complex, non-linear relationships between features and the target variable (box office)

- Performed feature scaling to prevent features with larger magnitudes from dominating the distance calculations

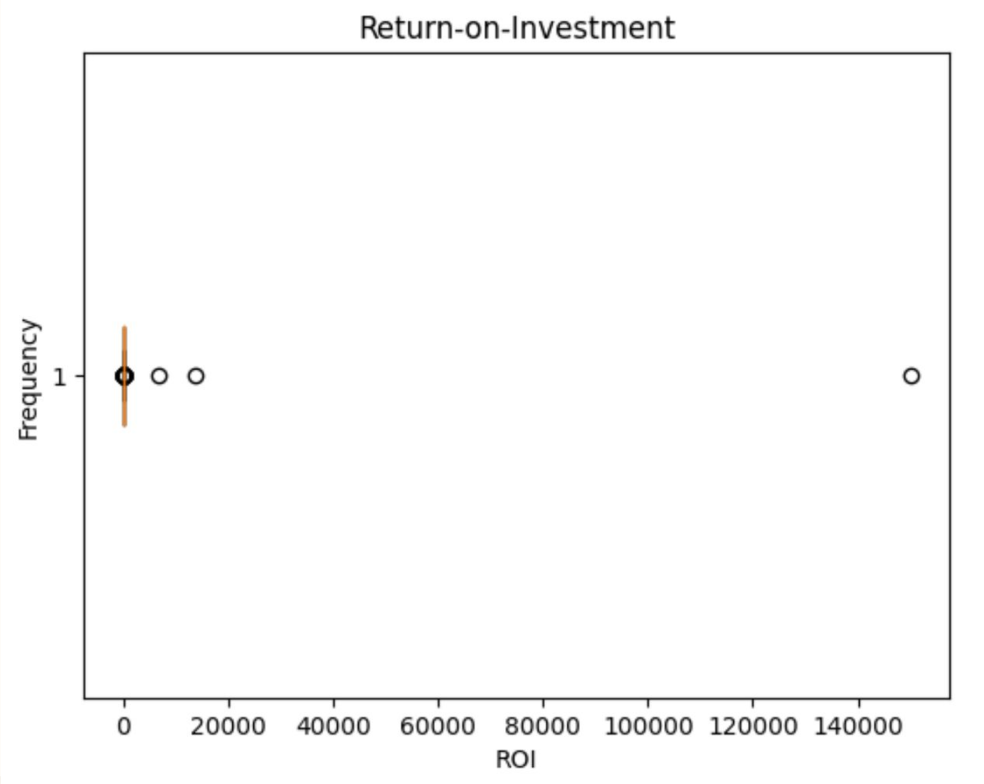
- Chosen value of **k = 5, Average  R-squared = 0.5758**

```
k=1, R2: 0.2868545970598171
k=3, R2: 0.5239415174160968
k=5, R2: 0.5701285685449033
k=7, R2: 0.5619884016971377
k=9, R2: 0.5359938525752554
k=11, R2: 0.5073837436348648
k=13, R2: 0.47563854839042996
k=15, R2: 0.46323213319148737
k=17, R2: 0.4633540784463628
k=19, R2: 0.4636887071541944
k=21, R2: 0.47033048382539455
k=23, R2: 0.4681098695188427
k=25, R2: 0.47587621012965586
k=27, R2: 0.4713002592952441
k=29, R2: 0.4675601991013547
```

# XGBoost Regressor

- XGBoost (Extreme Gradient Boosting) Regressor works well with non-linear relationships within data
- Gradient Boosted Decision Tree Model
- Underlying technique - boosting to minimize residual errors at each step
- Regularization Parameters - gamma (penalize further partitions of tree node), lambda (penalize attaching higher weights to features), alpha (penalize non-zero coefficients)
- Features used : movie genres, time specific & budget
- **R-squared: 0.7712260451971712**

# Categorical: Return on Investment (ROI)

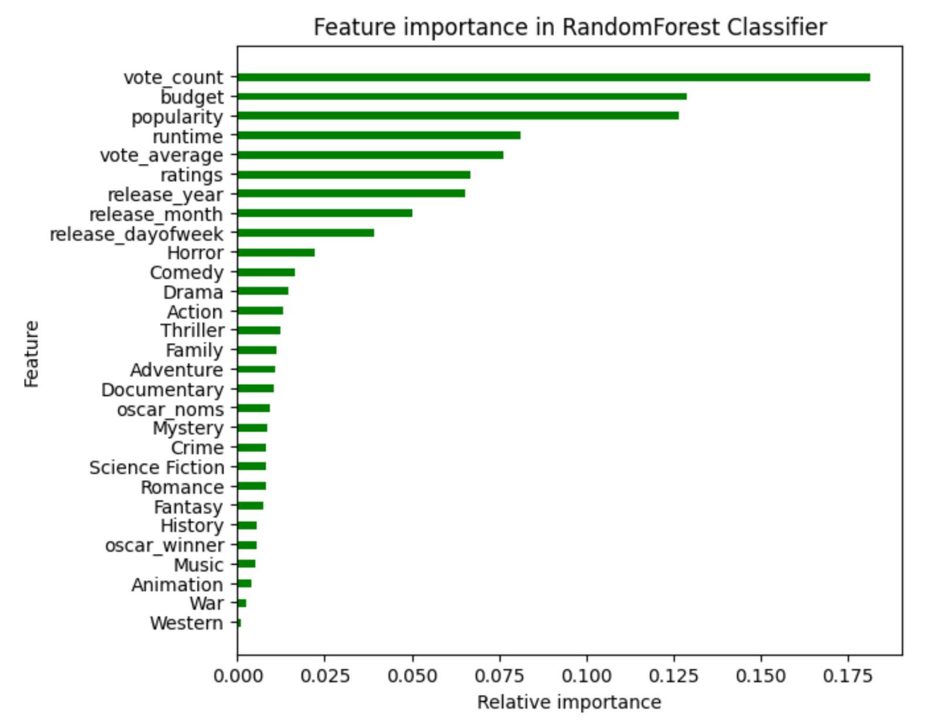



**ROI = (Gross-budget)/(budget)**

**mean = 123.75**
**median = 1.44**
(typical ROI is 144% of budget)

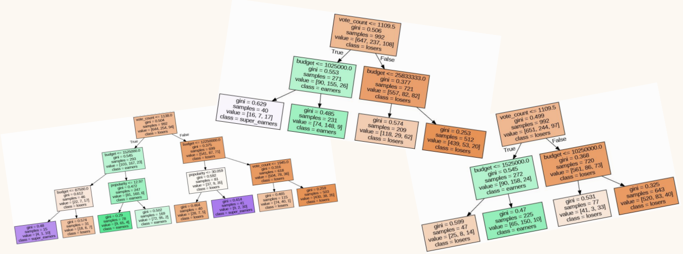
**Highly skewed!**
Luckily, Decision Tree and Random Forest more robustly tackle skews and outliers.

# Decision Tree and Random Forest




**Recall**

**3-classes of ROI:** losers, earners, super-earners (90th percentile)



**Decision Tree**

DT accuracy ~ 0.75 (max_depth 3)

Important features = vote_count, ratings, popularity, budget

**Random Forest**

RF accuracy ~ 0.75 (10,000 trees)

| movie | earn_class |
|---|---|
| Love Lies Bleeding | earner |
| Kung Fu Panda 4 | earner |
| Imaginary | earner |
| Cabrini | earner |
| Accidental Texan | earner |
| The 3_2 Pulldown | earner |
| American Dreamer | earner |
| Night Shift | earner |
| 5lbs of Pressure | earner |
| The Ballad of Davy Crockett | earner |
| Space: The Longest Goodbye | loser |
| The Piper | earner |
| Glitter & Doom | loser |

# RF Prediction:

## 2 losers
## 0 super-earners

# 04 Takeaway

# Engagement and $$$ matter! Here's what to expect.

Eg. Vote_count, budget, popularity, runtime, and vote average

## Impactful Features

## Model of Choice:

XGBoost Regressor

Why this model's predictive value?

- R-squared
- Mean-squared-error
- Data over time

Surprisingly, similar performance in linear regression.

# Predictions for New Movies

| Movie | Gross_Predicted_Revenue |
|---:|---:|
| Love Lies Bleeding | $105,777,760 |
| Kung Fu Panda 4 | $250,660,060 |
| Imaginary | $100,146,830 |
| Cabrini | $26,659,300 |
| Accidental Texan | $42,818,496 |
| The 3_2 Pulldown | $72,286,784 |
| American Dreamer | $79,604,320 |
| Night Shift | $62,907,676 |
| 5lbs of Pressure | $51,473,488 |
| The Ballad of Davy Crockett | $194,965,580 |
| Space: The Longest Goodbye | $16,639,981 |
| The Piper | $95,083,088 |
| Glitter & Doom | $36,235,776 |


**Limitations/Future Directions**

- Data Accuracy and Completeness
- Genre dependent characterization for Nulls
- Time Series Analysis




13 new movies



# Thanks!

Time to see how they fare in theaters!