LetterBoxd Film Analysis

By: Clayton Brittan

Primary Goal and Motivation

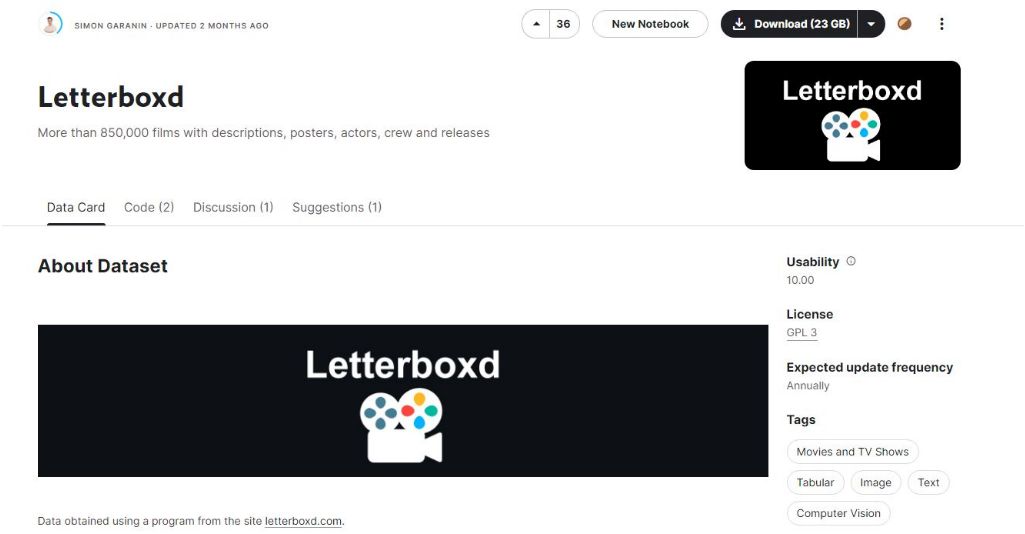
- Understanding evolving trends in the film industry
- Recognizing factors influencing film success
- Predictive potential for future film success
- Importance for filmmakers and viewers alike

Questions to Answer

- Key factors influencing audience reception and ratings in films?
- Influence of factors like release year, duration, genre themes, etc., on a film's ratings?
- Discernible trends or patterns in audience preferences over time and their correlation with industry changes?
- Accuracy of predicting film success or ratings based on attributes like release year, duration, and genre themes?
- Insights gained from analyzing crowd-sourced film reviews and ratings on platforms like Letterboxd
- How these insights can inform filmmakers and viewers about evolving industry trends?

Dataset

Source: https://www.kaggle.com/datasets/gsimonx37/letterboxd/data



Data Processing and Preparation

- Analyze structure of the Data
- Combine Individual Data Points per Movie
- Remove Unwanted Data Points
- Combine Individual CSV Files
- Check For Missing or Erroneous Data
- Encode Non-Numeric Variables

id movie identifier (primary key)	name the name of the film	date year of release of the film	tagline the slogan of the film	description description of the film	minute movie dur minutes)
1.00m 1.90m	759487 unique values	1874 2031	[null] 85% animation short 0% Other (131670) 15%	[null] 17% Mexican feature film 0% Other (740307) 83%	1
1000001	Barbie	2023	She's everything. He's just Ken.	Barbie and Ken are having the time of their lives in the colorful and seemingly perfect world of Bar...	114
1000002	Parasite	2019	Act like you own the place.	All unemployed, Ki-taek's family takes peculiar interest in the wealthy and glamorous Parks for thei...	133
1000003	Everything Everywhere All at Once	2022	The universe is so much bigger than you realize.	An aging Chinese immigrant is swept up in an insane adventure, where she alone can save what's impor...	140
1000004	Fight Club	1999	Mischief. Mayhem. Soap.	A ticking-time-bomb insomniac and a slippery soap salesman channel primal male aggression into a sho...	139

Tools Used

- Python
- Pandas
- Numpy
- SK-Learn
- PyPlot
- Seaborn
- CSV
- MatplotLib

```python
import pandas as pd
import numpy as np
import csv

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import MultiLabelBinarizer
from sklearn.metrics import mean_squared_error

from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor

import matplotlib.pyplot as plt
import seaborn as sns
```

Classification Applied

- K-Nearest Neighbors Classification:

```
Mean Squared Error: 0.19265167636861547
R-squared: 0.33444441948388737
```

```
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

- Random Forest Classification:

```
Mean Squared Error: 0.007615754843475398
R-squared: 0.973733676771311
```

```
model = RandomForestRegressor()
model.fit(X_train, y_train)
```

Testing the Model

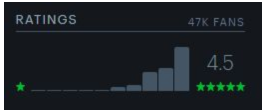
Testing the model using Unseen Data (Dune 2)

```
new_data = {
    'date': [2024.0],
    'minute': [167.0],
    'Action': [1.0],
    'Science Fiction': [1.0]
}

new_data_df = pd.DataFrame(new_data)

# Fill zeros for all other theme columns
theme_columns = [col for col in filtered_df.columns if col not in ['date', 'minute', 'Action', 'Science Fiction']]
for column in theme_columns:
    new_data_df[column] = 0.0

new_data_df = new_data_df[X.columns]
```

K-Nearest Neighbors: [2.76184456]

Random Forest: [2.82]

https://m.media-amazon.com/images/M/MV5BN2QyZGU4ZDctOWMzMy00NTc5LThlOGQtODhmNDI1NmY5YzAwXkEyXkFqcGdeQXVyMDM2NDM2MQ@@._V1_FMjpg_UX1000_.jpg

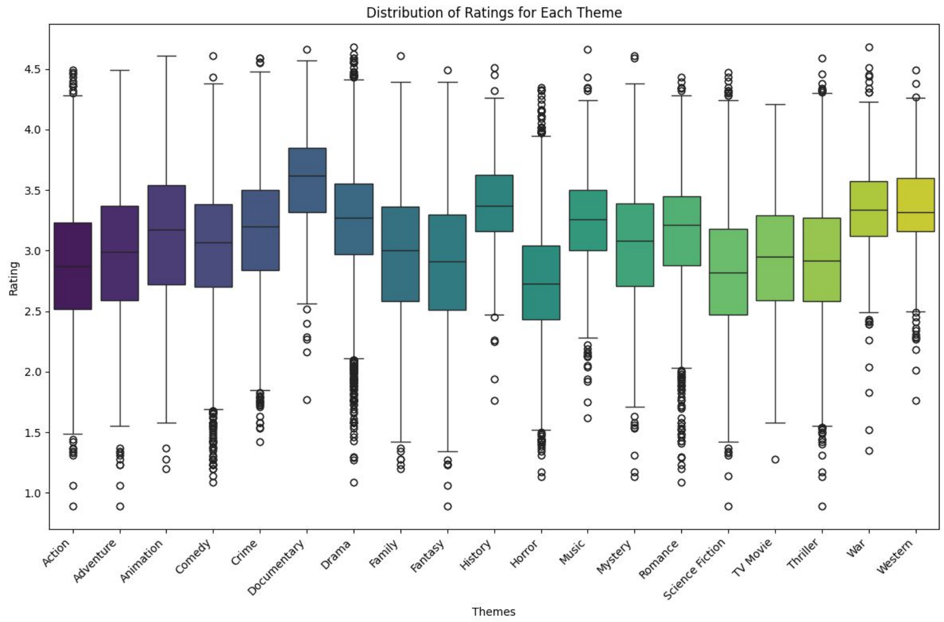


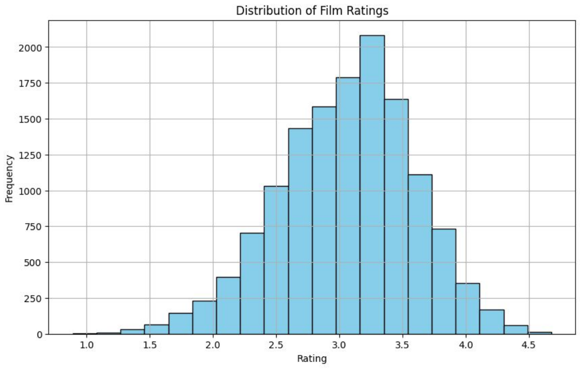
What We Can Learn From The Model

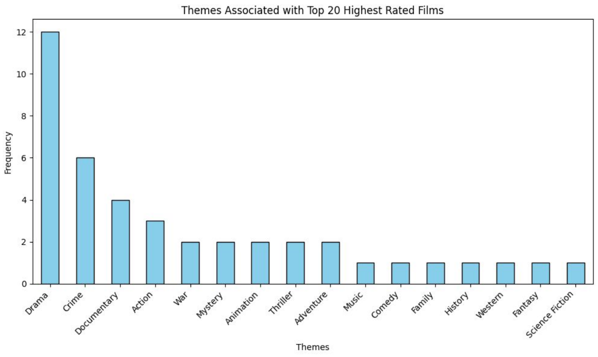
Most Important Features

```
Feature: date, Importance: 0.2920995759567058
Feature: minute, Importance: 0.28041295964962637
Feature: Action, Importance: 0.022061717644103532
Feature: Adventure, Importance: 0.016622086520230325
Feature: Animation, Importance: 0.017016204973485265
Feature: Comedy, Importance: 0.02280449500587551
Feature: Crime, Importance: 0.019811566512077418
Feature: Documentary, Importance: 0.05476953774718346
Feature: Drama, Importance: 0.08775456452275479
Feature: Family, Importance: 0.011421064795180493
Feature: Fantasy, Importance: 0.016524496813703292
Feature: History, Importance: 0.005490644826045301
Feature: Horror, Importance: 0.04600198399619078
Feature: Music, Importance: 0.01024557471289118
Feature: Mystery, Importance: 0.014706751708994954
Feature: Romance, Importance: 0.017302486797070846
Feature: Science Fiction, Importance: 0.024242919197987985
Feature: TV Movie, Importance: 0.00655563572662662
Feature: Thriller, Importance: 0.02428795347620403
Feature: War, Importance: 0.005209405092866976
Feature: Western, Importance: 0.0046583743241950205
```

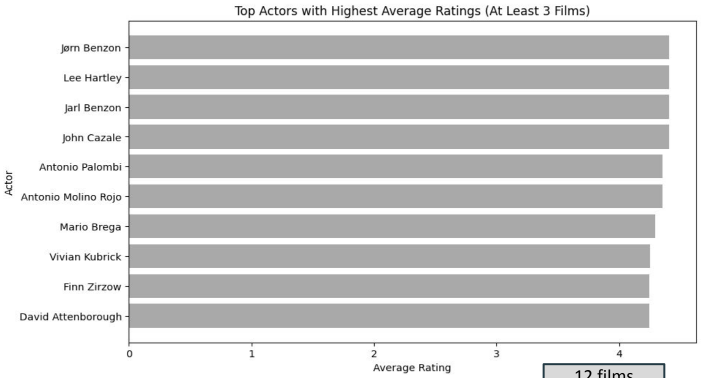
Using Model To Fuel Visual Analysis

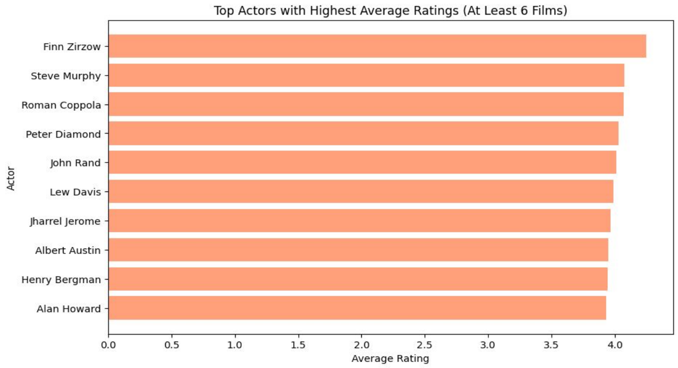


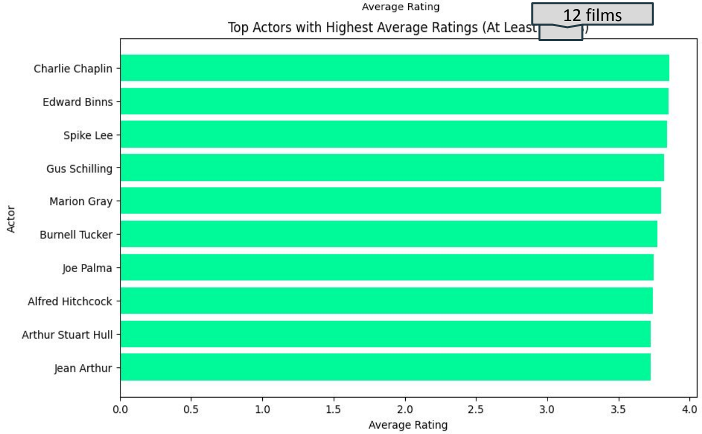


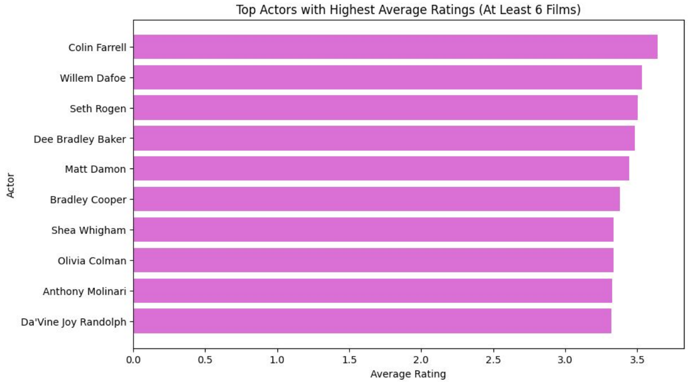


What the Model Can't Capture



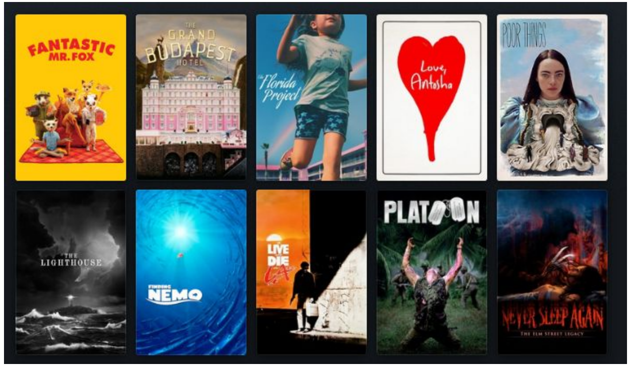






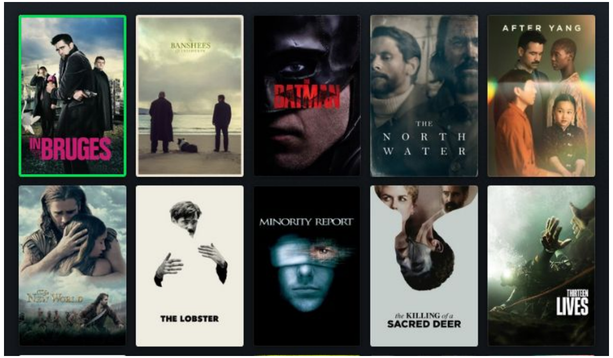
Finn Zirzow:

Willem Dafoe:



Spike Lee:

Colin Farrell:



Knowledge Gained

- Comprehensive feature selection is crucial for effective predictive modeling in film score prediction.
- Visual analysis plays a pivotal role in revealing nuanced patterns and trends in film datasets.
- Film scores tend to exhibit intrinsic skewness in distribution, typically ranging between 2.5 and 3.5.
- Genre dynamics significantly influence film reception and evaluation, with genres like drama and documentaries performing better than horror productions.
- The relationship between fiscal investment, production value, and film acceptance suggests potential insights if budget data were included in analyses.
- Gaps in data, such as budgetary considerations, highlight the need for more comprehensive datasets to understand the dynamics of movie ratings.
- The convergence of empirical insights and methodological requirements creates opportunities for future inquiry in predictive modeling and film evaluation.

Applications

- Production companies: Strategic resource allocation for optimized returns and risk reduction.
- Marketing agencies: Tailored advertising campaigns based on audience preferences for improved engagement.
- Content platforms: Curated catalogs reflecting diverse audience tastes, minimizing underperforming titles.
- Improved Models: Using the limitations discovered in this model we can gather more data and build better models