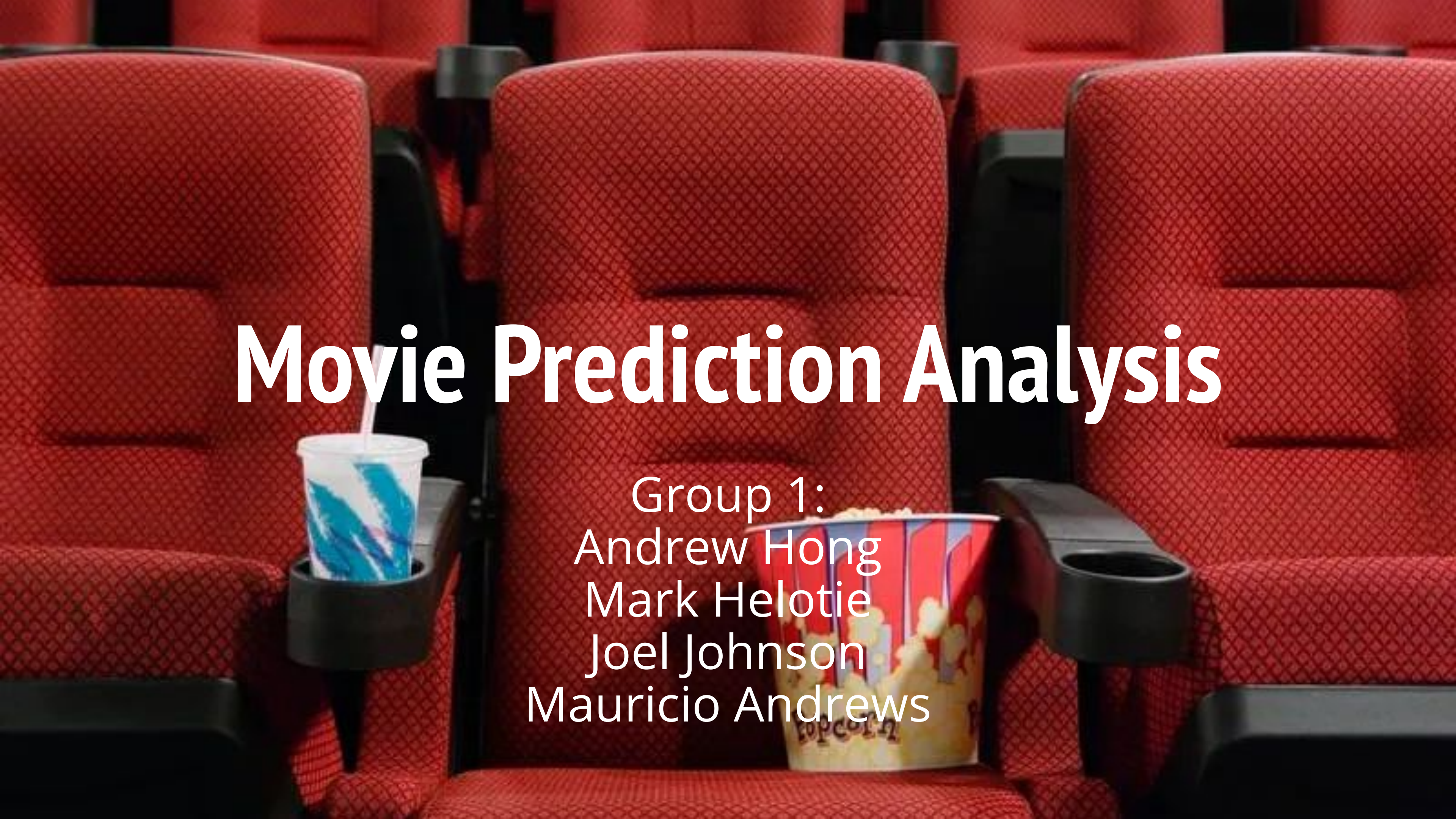


Movie Prediction Analysis

Group 1:
Andrew Hong
Mark Helotie
Joel Johnson
Mauricio Andrews

Introduction:

With a large number of movies available for viewers to watch, it can be hard to predict whether or not an upcoming movie can be a worthwhile investment. In order to do so, we want to figure out the factors that are most important in determining movie success and create a machine learning model using this data that can predict a movie's success.

Objectives

- Using the dataset, create a criteria that would determine if a movie is successful or not.
- Visualize the correlation between a movie's success and various factors.
- Use this to create a machine learning model that is able to predict a movie's success based off the data.
- Determine which factors are most important in a movie's success.

Datasets

All of our datasets were sourced from Kaggle.

Our initial work was done with a dataset from 2017 called TMDB 5000 .

...and then we discovered a much more robust dataset here (from 2022).

In an attempt to get better clarity and representation on our output,
we performed many comparisons between the two datasets.

Success Criteria

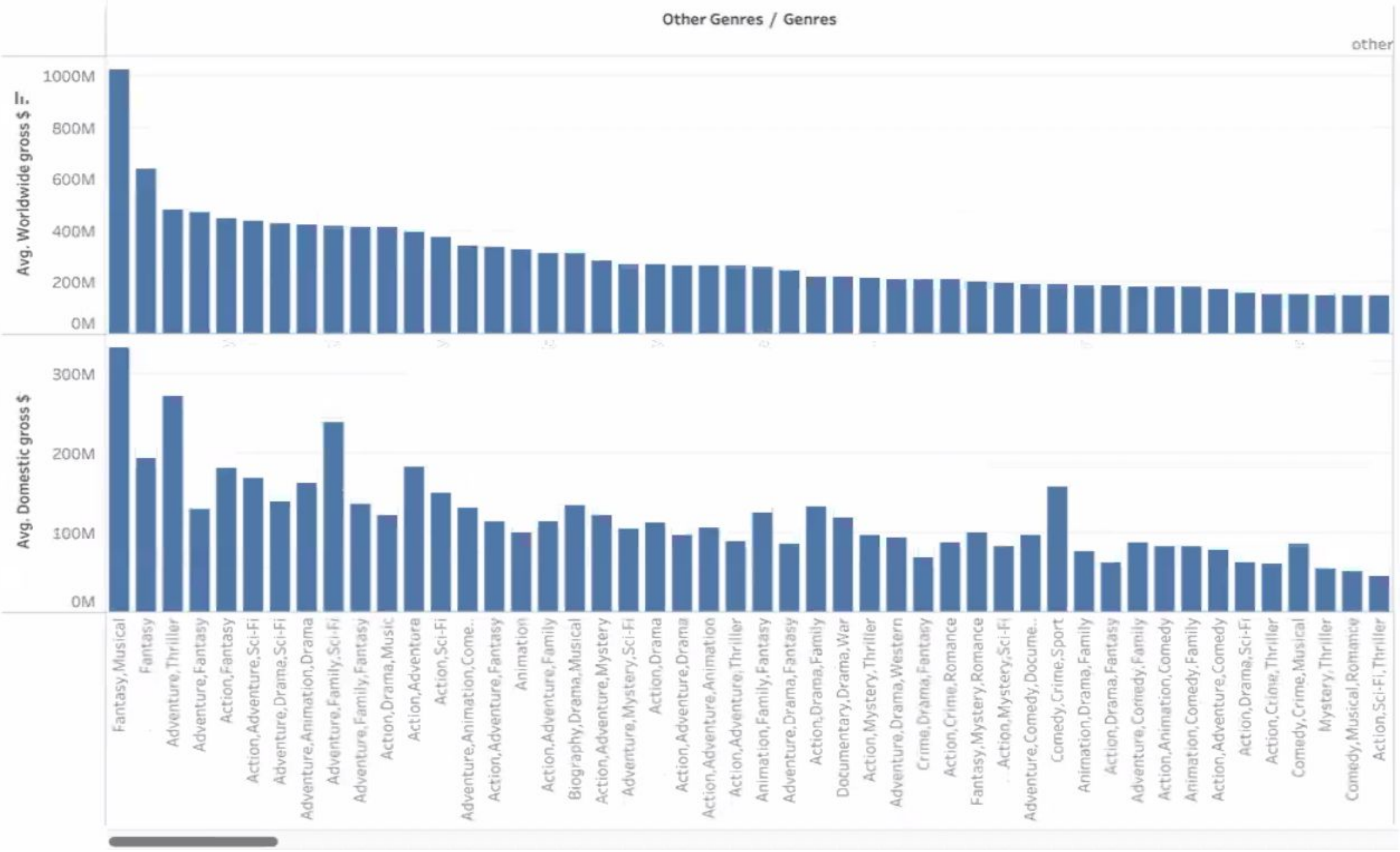
For our machine learning model, we defined a model as “successful” based on the following criteria:

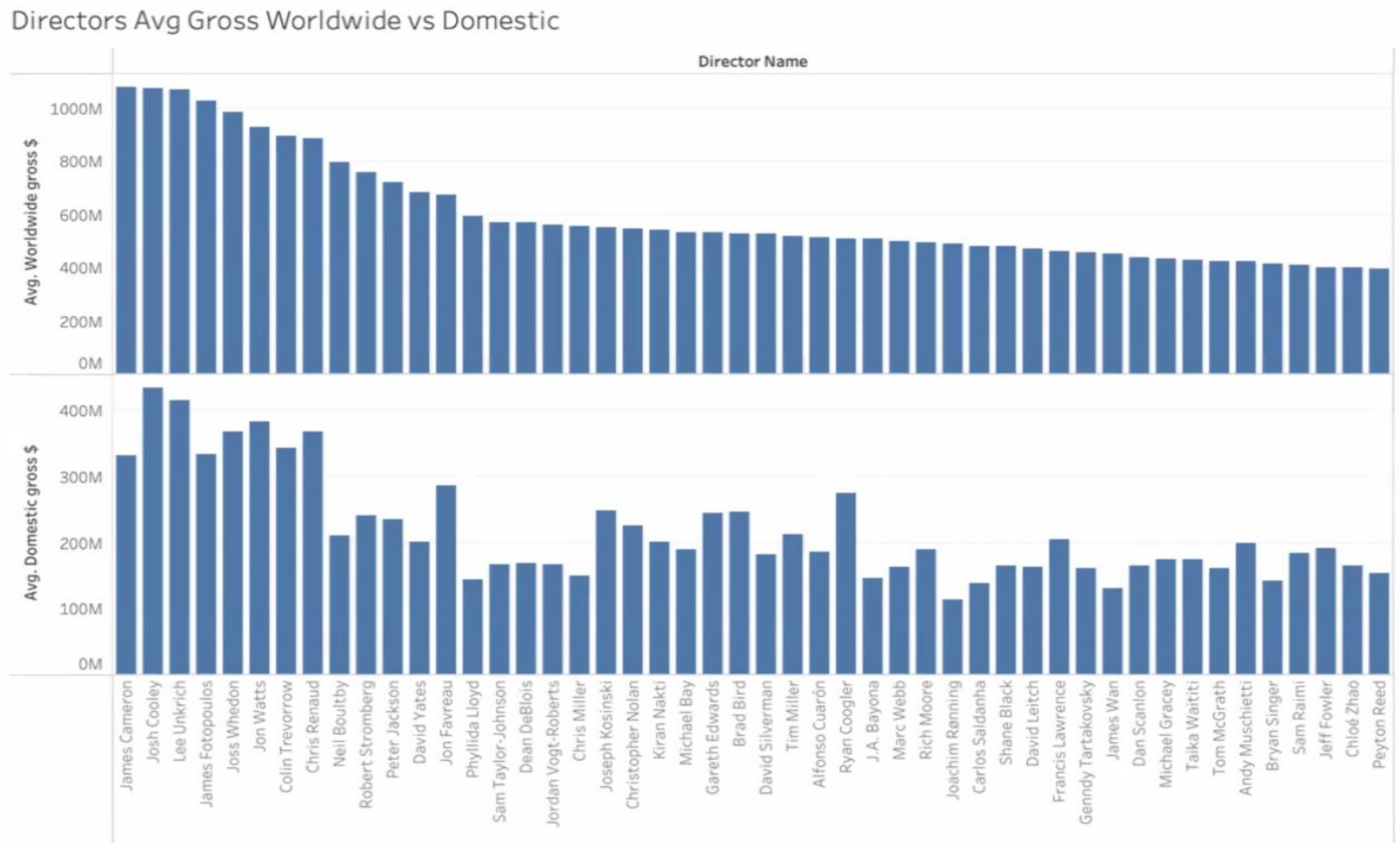
- The movie’s revenue is greater than twice its’ budget.

AND

- The average rating of the movie is greater than the average rating of the dataset overall.

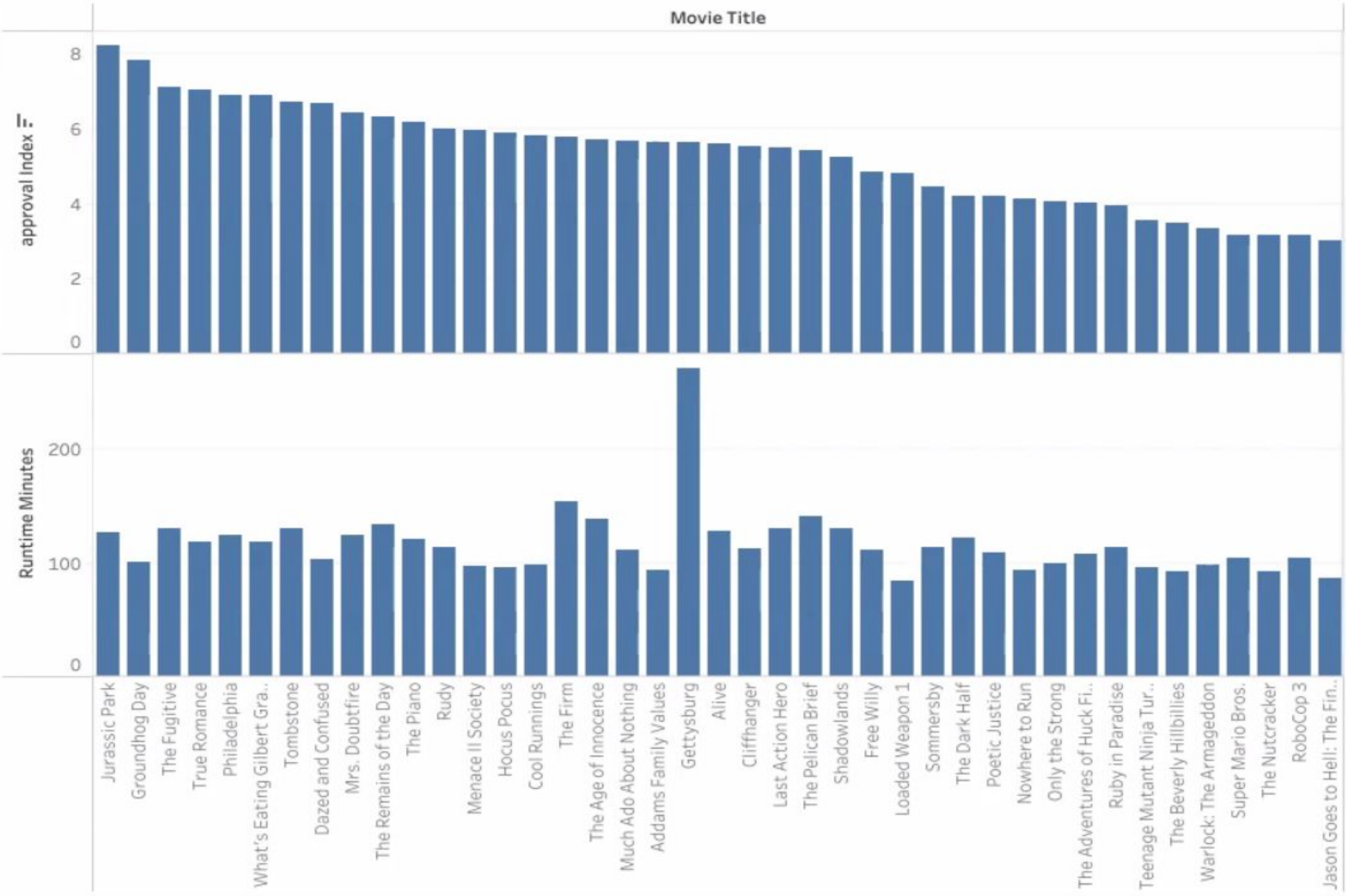
Worldwide Gross





Directors Avg Gross Worldwide vs Domestic
Director Name
Avg. Worldwide gross $
1000M
800M
600M
400M
200M
0M
Avg. Domestic gross $
400M
300M
200M
100M
0M
James Cameron
Josh Cooley
Lee Unkrich
James Fotopoulos
Joss Whedon
Jon Watts
Colin Trevorrow
Chris Renaud
Neil Boultby
Robert Stromberg
Peter Jackson
David Yates
Jon Favreau
Phyllida Lloyd
Sam Taylor-Johnson
Dean DeBlois
Jordan Vogt-Roberts
Chris Miller
Joseph Kosinski
Christopher Nolan
Kiran Nakti
Michael Bay
Gareth Edwards
Brad Bird
David Silverman
Tim Miller
Alfonso Cuarón
Ryan Coogler
J.A. Bayona
Marc Webb
Rich Moore
Joachim Rønning
Carlos Saldanha
Shane Black
David Leitch
Francis Lawrence
Genndy Tartakovsky
James Wan
Dan Scanlon
Michael Gracey
Taika Waititi
Tom McGrath
Andy Muschietti
Bryan Singer
Sam Raimi
Jeff Fowler
Chloé Zhao
Peyton Reed

Movie Popularity and Runtime by Year



Machine Learning Model

After defining our success criteria, we added an additional column into our dataset to showcase each movie's success.

We proceeded to fit our data into the random forest classifier. We used this due to the large number of variables present in our model, which random forest is well suited for.

	budget	popularity	revenue	vote_average	vote_count	success_score
0	237000000	150.437577	2787965087	7.2	11800	1
1	300000000	139.082615	961000000	6.9	4500	1
2	245000000	107.376788	880674609	6.3	4466	1
3	250000000	112.312950	1084939099	7.6	9106	1
4	260000000	43.926995	284139100	6.1	2124	0
...	...	...	...	...	...	...
4798	220000	14.269792	2040920	6.6	238	1
4799	9000	0.642552	0	5.9	5	0
4800	0	1.444476	0	7.0	6	0
4801	0	0.857008	0	5.7	7	0
4802	0	1.929883	0	6.3	16	0

Machine Learning Model - Results

Confusion Matrix and Classification Report

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	833	17
Actual 1	5	346

Accuracy Score: 0.9816819317235637

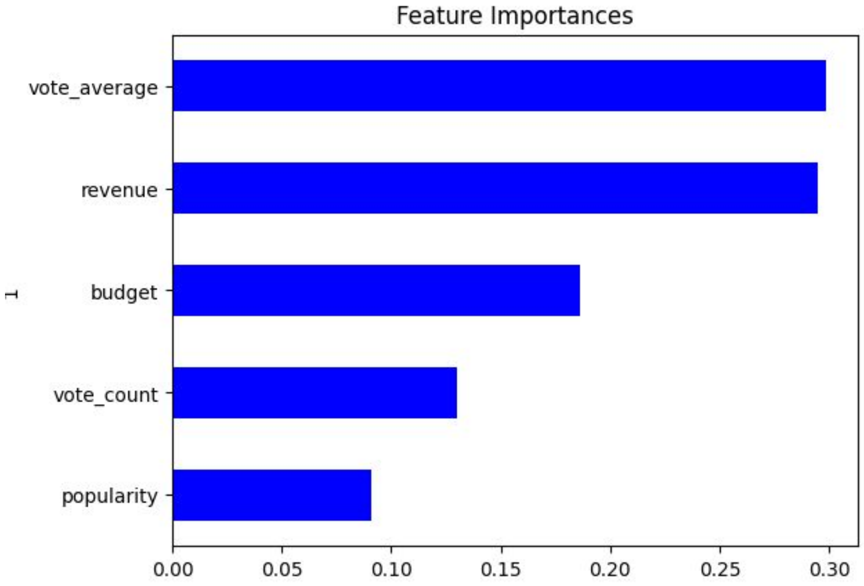
Classification Report

	precision	recall	f1-score	support
0	0.99	0.98	0.99	850
1	0.95	0.99	0.97	351
accuracy			0.98	1201
macro avg	0.97	0.98	0.98	1201
weighted avg	0.98	0.98	0.98	1201

Feature Importance

We can also use **sklearn** to determine the importance of each feature in determining a movie's success. Here are the results:

```
[(0.2983306480889208, 'vote_average'),
 (0.2946290526684361 4, 'revenue'),
 (0.18635902106610636, 'budget'),
 (0.12987841167775563, 'vote_count'),
 (0.09080286649878107, 'popularity')]
```



Machine Learning Model (Alternate Dataset)

We also used a dataset known as the Ultimate Film Statistics and compared our results to our original model.

Comparison:

	runtime_minutes	movie_averageRating	movie_numerOfVotes	approval_Index	production_budget	domestic_gross	worldwide_gross	success_score
0	192.0	7.8	277543.0	7.061101	460000000	667830256	2265935552	1
1	181.0	8.4	1143642.0	8.489533	400000000	858373000	2794731755	1
2	137.0	6.6	533763.0	6.272064	379000000	241071802	1045713802	1
3	141.0	7.3	870573.0	7.214013	365000000	459005868	1395316979	1
4	149.0	8.4	1091968.0	8.460958	300000000	678815482	2048359754	1
...	...	...	...	...	...	...	...	...
4375	100.0	7.2	110078.0	6.017902	65000	11529368	22233808	1
4376	98.0	6.6	7986.0	4.231464	50000	10426506	10426506	1
4377	93.0	4.9	1593.0	2.526405	50000	2335352	2335352	0
4378	98.0	6.2	14595.0	4.242085	50000	391674	424149	0
4379	111.0	6.2	163.0	2.191765	50000	8374	8374	0

Machine Learning Model (Alternate Dataset) - Results

Confusion Matrix and Classification Report

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	715	21
Actual 1	11	348

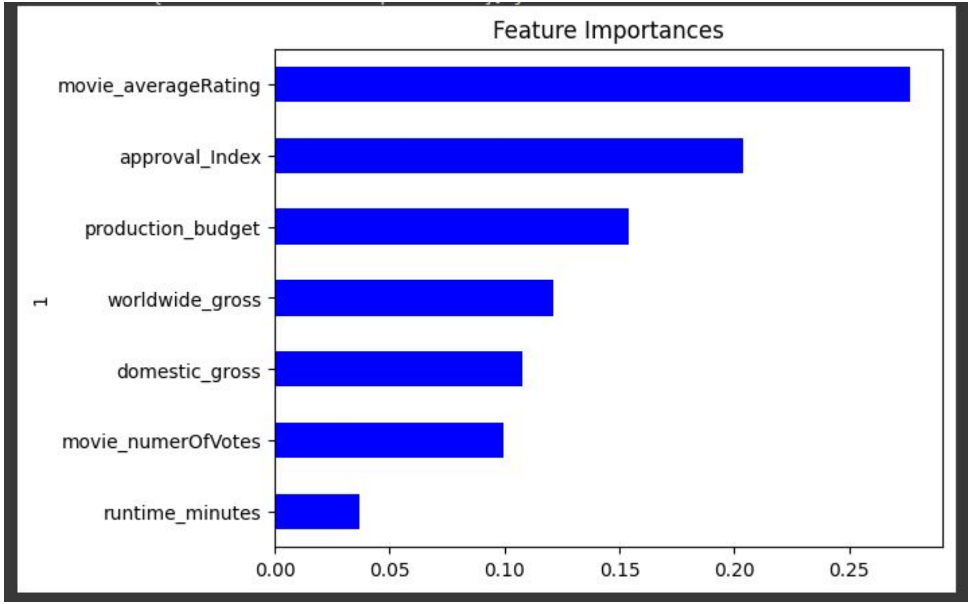
Accuracy Score: 0.9707762557077626

Classification Report

	precision	recall	f1-score	support
0	0.98	0.97	0.98	736
1	0.94	0.97	0.96	359
accuracy			0.97	1095
macro avg	0.96	0.97	0.97	1095
weighted avg	0.97	0.97	0.97	1095

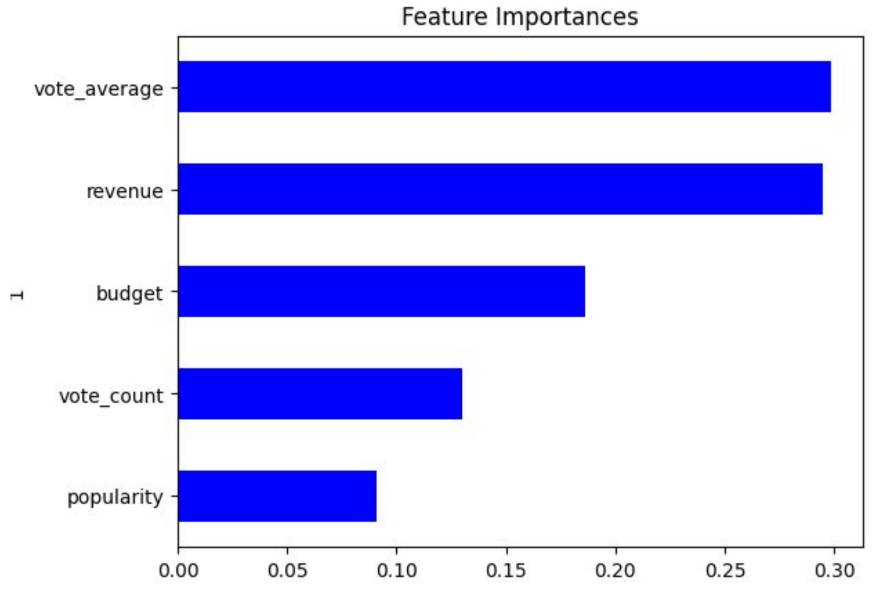
Feature Importance (Alternate Dataset)

```
[(0.2768135760549307, 'movie_averageRating'),
 (0.2037114119853299, 'approval_Index'),
 (0.15399358759855838, 'production_budget'),
 (0.12147781814059178, 'worldwide_gross'),
 (0.10772034904874249, 'domestic_gross'),
 (0.0994849925457004, 'movie_numerOfVotes'),
 (0.036798264626146236, 'runtime_minutes')]
```

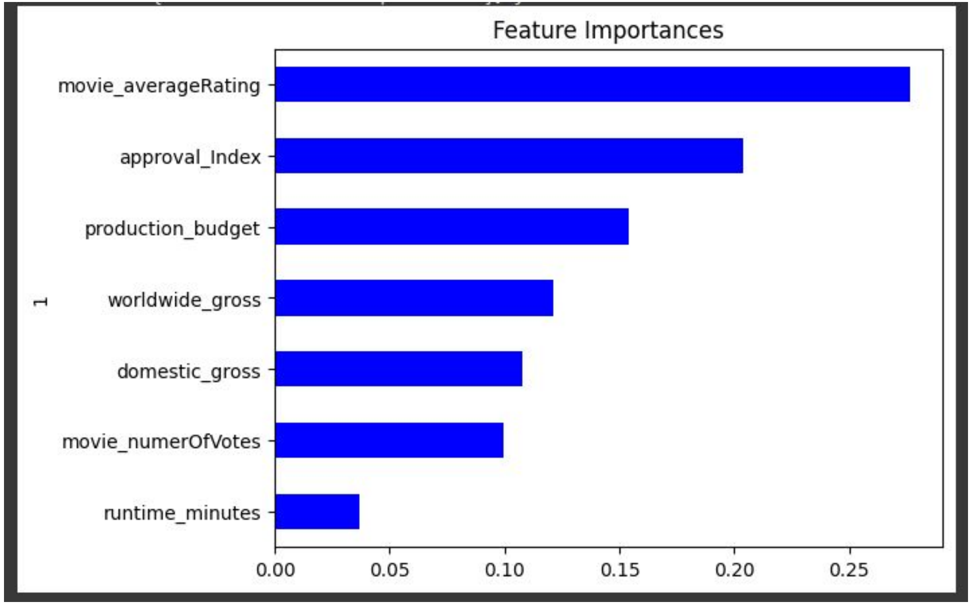


Feature Importance Comparison

TMDB_5000_dataset



Ultimate Film Statistics



Importance of Genres

In this section, we are comparing the impact of the genre of the movie to its success.

We are using the Ultimate Film

Statistics dataset for this one.

Due to there being a large number of genre types, we first renamed all genres that had less than 50 movies in the dataset to “other”

```
Other                           2397
Comedy,Drama,Romance             196
Drama                            160
Adventure,Animation,Comedy       157
Comedy,Drama                     148
Comedy                           147
Comedy,Romance                   135
Action,Adventure,Sci-Fi          119
Drama,Romance                    115
Action,Crime,Drama               109
Action,Adventure,Comedy           97
Action,Comedy,Crime               79
Action,Adventure,Fantasy          76
Crime,Drama,Mystery               75
Action,Crime,Thriller             73
Horror,Mystery,Thriller           63
Crime,Drama,Thriller              62
Biography,Drama,History           61
Action,Adventure,Drama            59
Action,Adventure,Thriller         52
Name: genres, dtype: int64
```

Importance of Genres

Confusion Matrix and Classification Report

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	730	6
Actual 1	349	10

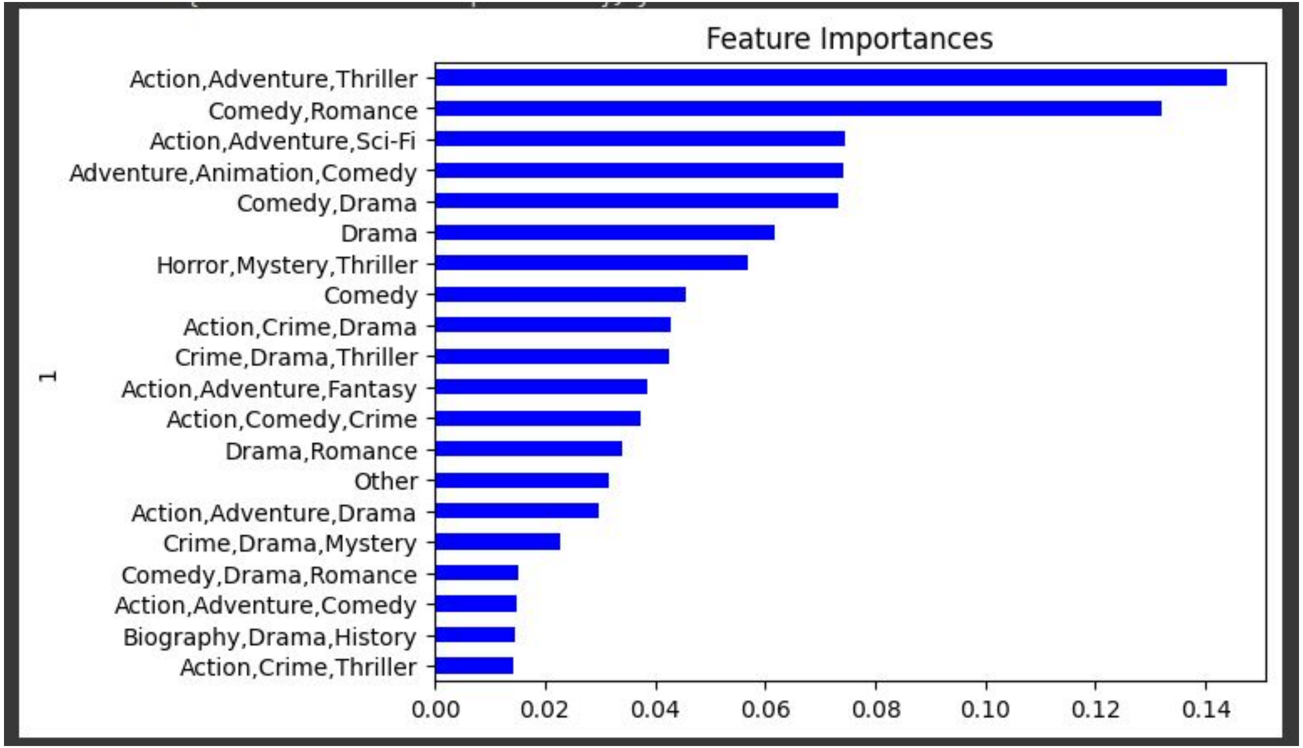
Accuracy Score: 0.6757990867579908

Classification Report

	precision	recall	f1-score	support
0	0.68	0.99	0.80	736
1	0.62	0.03	0.05	359
accuracy			0.68	1095
macro avg	0.65	0.51	0.43	1095
weighted avg	0.66	0.68	0.56	1095

This model is unreliable. In particular, it predicts many movies to be unsuccessful when in reality they actually were.

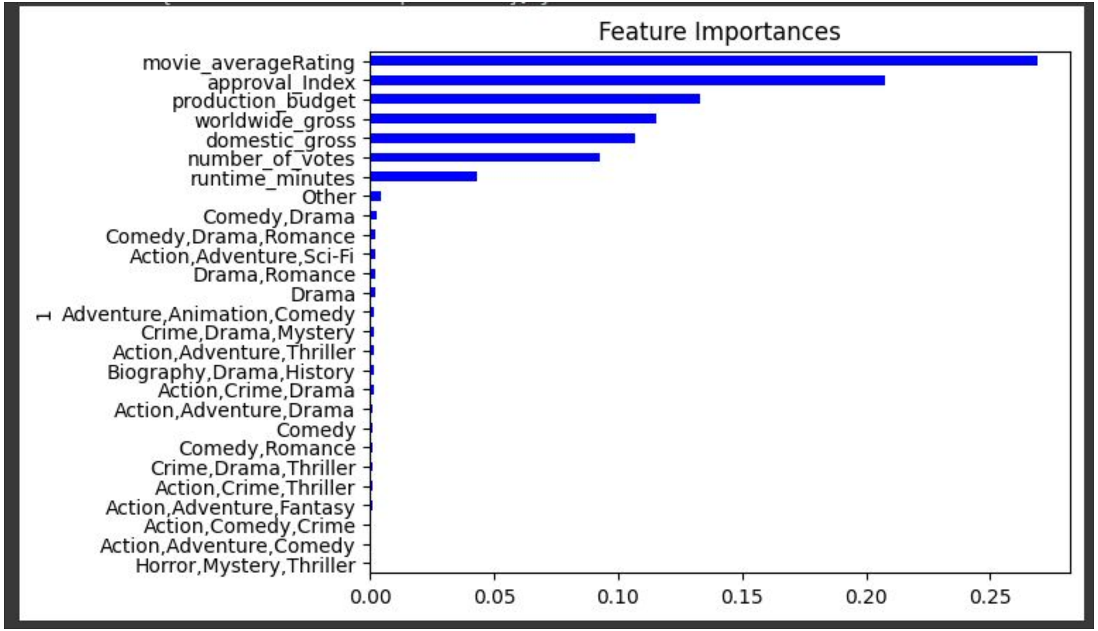
Importance of Genres - Feature Importance



Take this with a grain of salt for the reason listed on the previous slide.

Importance of Genres - Feature Importance

In order to maintain high accuracy, we need to keep in the features from before. Here's our new feature importances.



Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	712	24
Actual 1	16	343

Accuracy Score: 0.9634703196347032

Classification Report

	precision	recall	f1-score	support
0	0.98	0.97	0.97	736
1	0.93	0.96	0.94	359
accuracy			0.96	1095
macro avg	0.96	0.96	0.96	1095
weighted avg	0.96	0.96	0.96	1095

Importance of Directors

The last thing we did was compare the importance of directors. Once again, we used the Ultimate Film Statistics Dataset for this.

First, we filtered out directors by renaming directors with 10 or less movies to “other directors”.

Other_director	4037
Steven Spielberg	25
Clint Eastwood	24
Ridley Scott	21
Woody Allen	20
Martin Scorsese	19
Steven Soderbergh	18
Spike Lee	16
Ron Howard	16
Tim Burton	15
Robert Zemeckis	14
Joel Schumacher	13
Brian De Palma	13
Oliver Stone	13
Renny Harlin	13
Barry Levinson	13
Michael Bay	12
Tony Scott	12
Paul W.S. Anderson	11
Antoine Fuqua	11
Richard Donner	11
M. Night Shyamalan	11
Kevin Smith	11
Roland Emmerich	11

Importance of Directors - Feature Importances

As before, our model is unreliable due to its inaccuracy, particularly in rating movies as unsuccessful when they're actually successful.

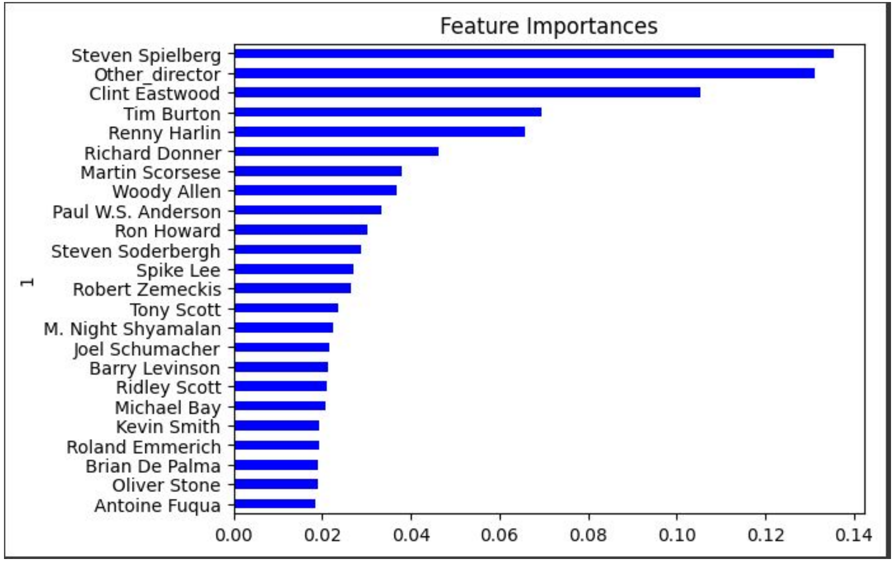
Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	728	8
Actual 1	346	13

Accuracy Score: 0.6767123287671233

Classification Report

	precision	recall	f1-score	support
0	0.68	0.99	0.80	736
1	0.62	0.04	0.07	359
accuracy			0.68	1095
macro avg	0.65	0.51	0.44	1095
weighted avg	0.66	0.68	0.56	1095



Importance of Directors - Feature Importances

To fix this, we had to keep the other features. We can see that directors have little impact on a movie's success, (although Ridley Scott does seem to be the most influential director).

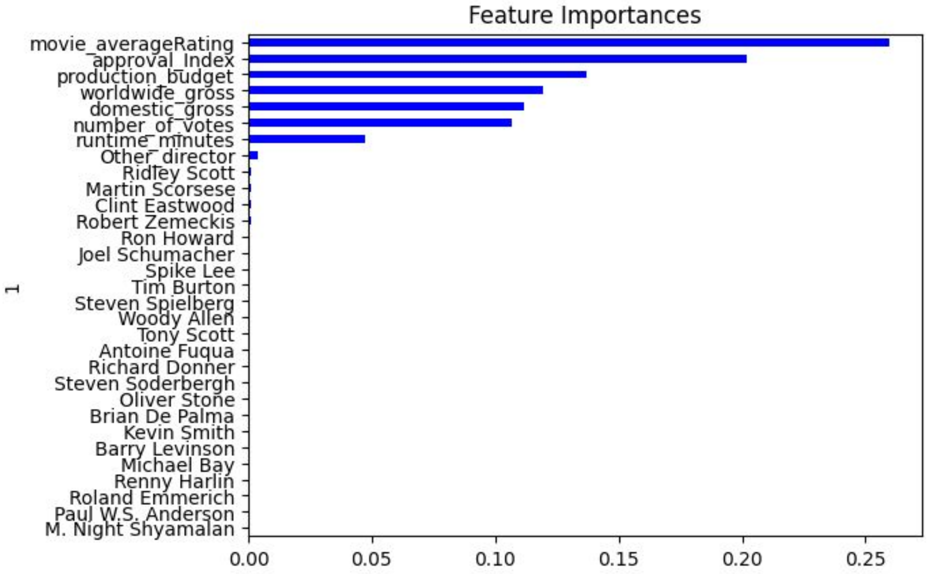
Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	715	21
Actual 1	14	345

Accuracy Score: 0.9680365296803652

Classification Report

	precision	recall	f1-score	support
0	0.98	0.97	0.98	736
1	0.94	0.96	0.95	359
accuracy			0.97	1095
macro avg	0.96	0.97	0.96	1095
weighted avg	0.97	0.97	0.97	1095



Conclusion:

- For both datasets, the movie’s rating contributed the most to the success of a movie. Interestingly, the rating count contributed the least to the movie’s success.
- Directors and genres had very little impact on a movie’s success, so if you want to invest in a movie, these don’t matter as much.
- However, if you do want to use these factors, it seems that comedy/drama movies have the most importance in determining a movie’s success.