Is this a Good or Bad Movie/Show ?

Table of Contents

Problem Statement

- Tommy is a production executive at IMDb.
- His bosses have tasked him to figure out what makes a tv-show or movie a smash hit.
- However, there are so many variables as to what makes a movie successful, that Tommy is overwhelmed.
- Tommy has thus engaged us, a team of data scientists to help him develop a machine learning model. This model should be able to predict the potential success of a given production based on certain features such as actors present in the production, genre, etc.

Problem Statement

- To develop this model, we will be scraping data from IMDb.
- Exploratory data analysis will then be conducted on the data set.
- This is to identify previous successful productions based on the features of the dataset which includes actor names, genre, content rating, viewership ratings, and more. This analysis will inform the development of a machine learning model that can be used to generalize to new productions, ultimately maximizing the chances of success for future productions.

- In summary, the goal is to create a robust and reliable machine learning model that can assist production executives in making informed decisions about which movies or TV shows to produce, based on data-driven insights into the factors that have contributed to success in the past.

Data Collection – Web scraping

- The data was scraped from IMDB using Beautiful Soup, a python package for parsing HTML documents.
- The scraped data resulted in 700,000+ rows broken down as follows:
- 14 movie genres x 50 titles per page x 1001 pages

Generating y-labels

- The y-labels for the dataset were generated based on the IMDb ratings the movies got (i.e. from a scale of 1-10 stars)

- All genres except horror had a median rating of 6.3, and therefore the threshold for the y-labels was set at 6.3
  - If a movie scored above 6.3 stars, they were labelled as good movies.
  - If a movie scored below 6.3 stars, they were labelled as bad movies.

- For the horror genre, its median rating was 5, and so the threshold for this particular genre was set at 5 instead.

Data Cleaning

- Null value cleaning
  - For null values in the dataset, it was mainly due to two reasons.
    - Movies or TV shows which were not released yet, but were already listed in IMDb, for example movies slated for release in late 2023 or 2024 onwards.
    - Movies which were released many decades ago, and ratings / reviews for it have been lost.
    - Therefore any rows with missing values were dropped.

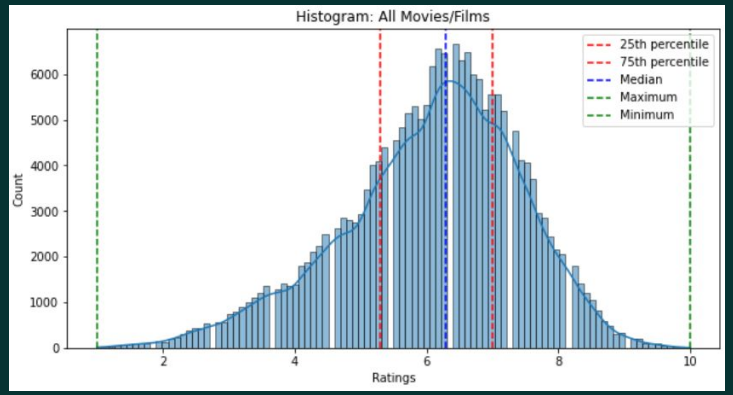
- Duplicates
  - There were genres with multiple labels (e.g. Action-Horror instead of just Action or Horror).
    - In this case, we will assign the main genre (i.e. Action will be assigned for Action-Horror).

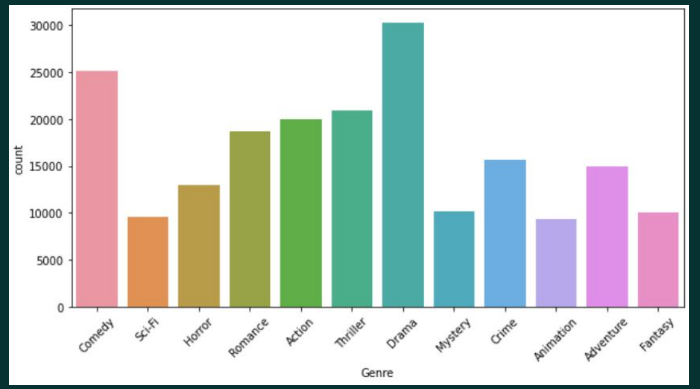
- Text Preprocessing
  - Stopwords removal, and stemming was done on the dataset as well.

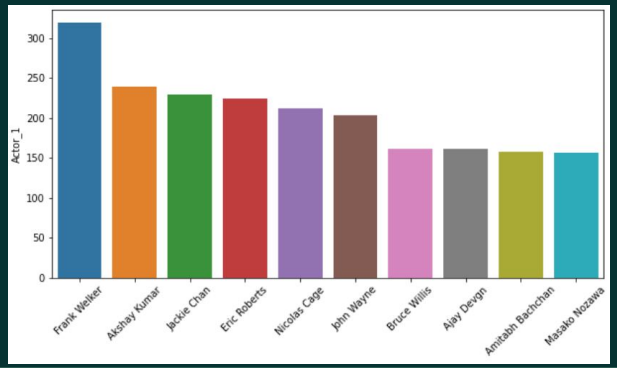
EDA (Feature Engineering)

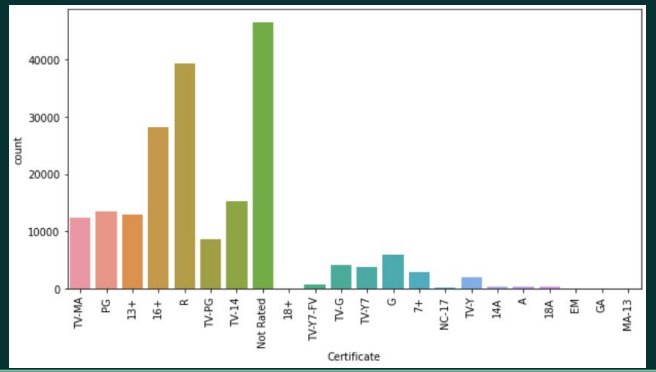
- Certificate Rating. For eg . T to Teens, E to Everyone G to Generic
- Actors. Creating more columns to fit all actors. For eg. Actor1 , Actor 2, Actor 3. Actor 1 will have more weightage.

All Movies/Films

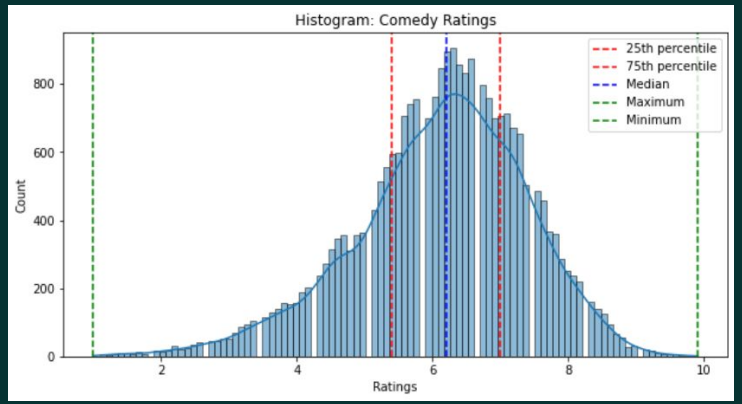


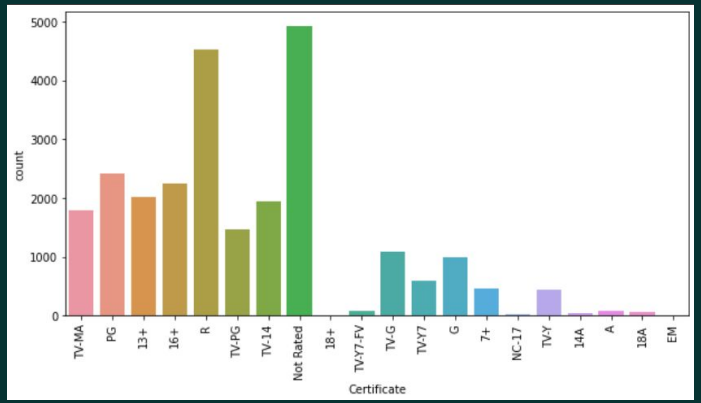


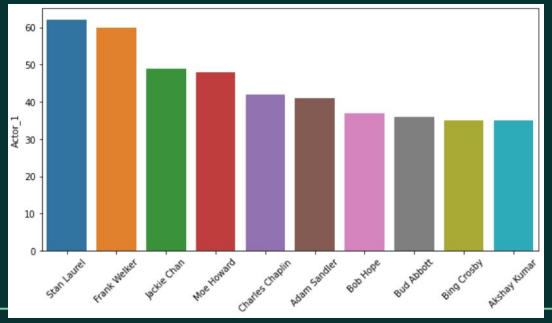




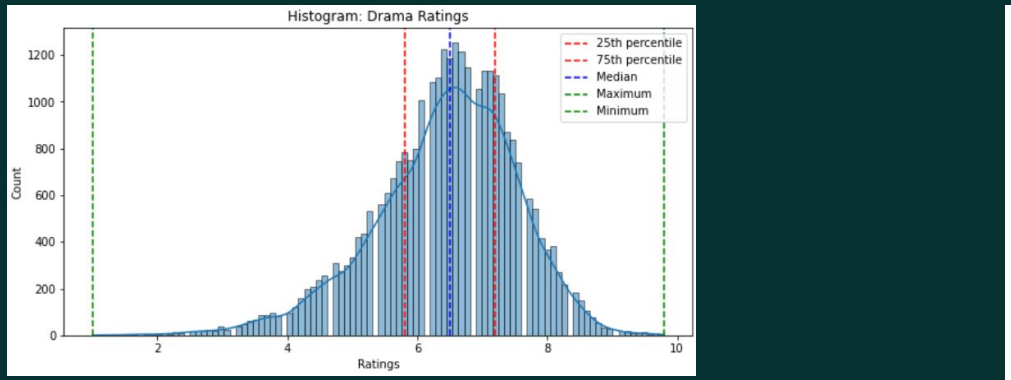
Comedy

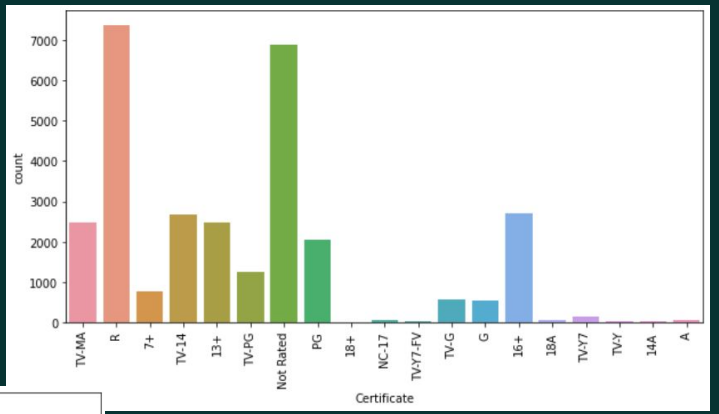

Histogram: Comedy Ratings
25th percentile
75th percentile
Median
Maximum
Minimum
Count
Ratings
0
200
400
600
800
2
4
6
8
10

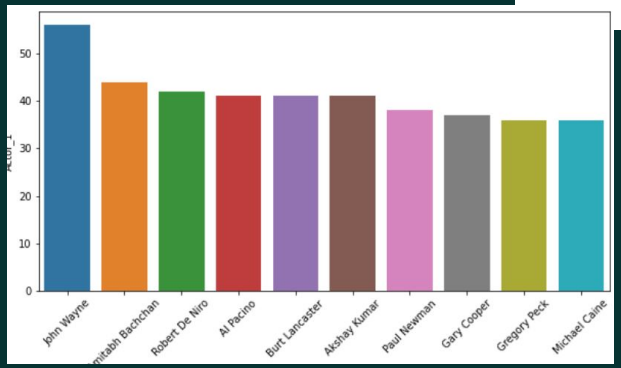

count
Certificate
0
1000
2000
3000
4000
5000
TV-MA
PG
13+
16+
R
TV-PG
TV-14
Not Rated
18+
TV-Y7-FV
TV-G
TV-Y7
G
7+
NC-17
TV-Y
14A
A
18A
EM


Actor_1
0
10
20
30
40
50
60
Stan Laurel
Frank Welker
Jackie Chan
Moe Howard
Charles Chaplin
Adam Sandler
Bob Hope
Bud Abbott
Bing Crosby
Akshay Kumar

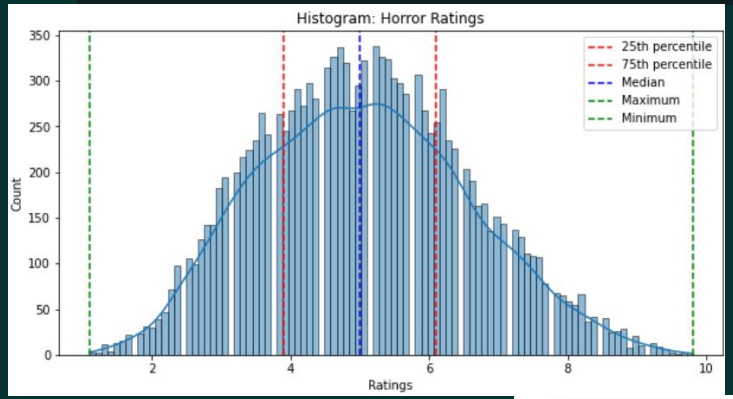
Drama

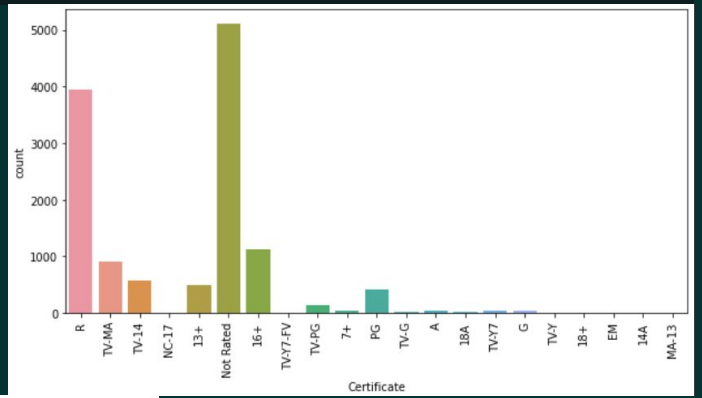

Histogram: Drama Ratings
25th percentile
75th percentile
Median
Maximum
Minimum
Count
Ratings

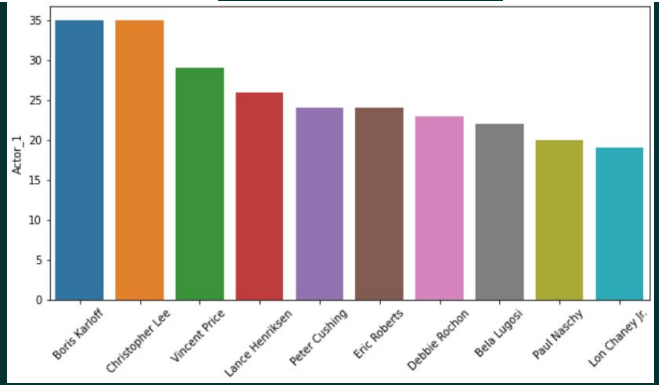

count
Certificate
TV-MA
R
7+
TV-14
13+
TV-PG
Not Rated
PG
18+
NC-17
TV-Y7-FV
TV-G
G
16+
18A
TV-Y7
TV-Y
14A
A


John Wayne
Amitabh Bachchan
Robert De Niro
Al Pacino
Burt Lancaster
Akshay Kumar
Paul Newman
Gary Cooper
Gregory Peck
Michael Caine

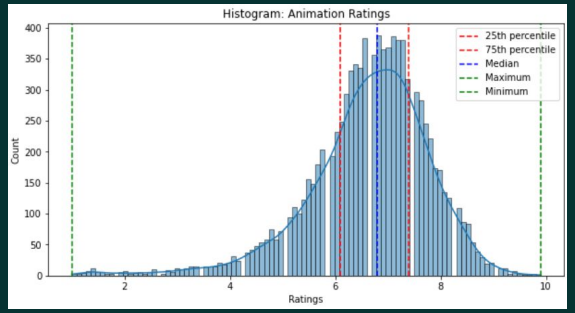
Horror

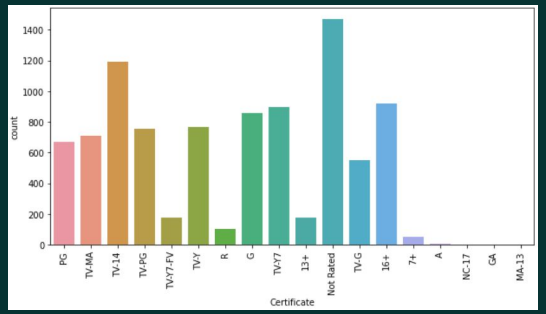

Histogram: Horror Ratings
25th percentile
75th percentile
Median
Maximum
Minimum
Count
Ratings

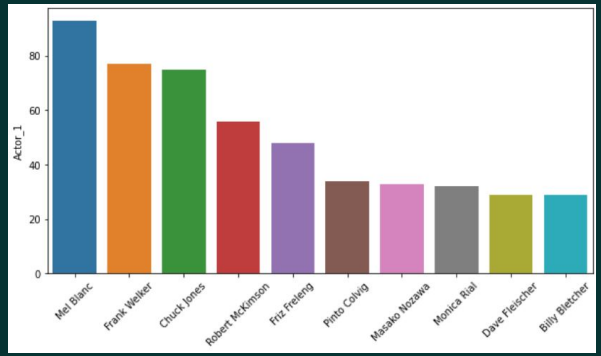

count
R
TV-MA
TV-14
NC-17
13+
Not Rated
16+
TV-Y7-FV
TV-PG
7+
PG
TV-G
A
18A
TV-Y7
G
TV-Y
18+
EM
14A
MA-13
Certificate


Actor_1
Boris Karloff
Christopher Lee
Vincent Price
Lance Henriksen
Peter Cushing
Eric Roberts
Debbie Rochon
Bela Lugosi
Paul Naschy
Lon Chaney Jr.

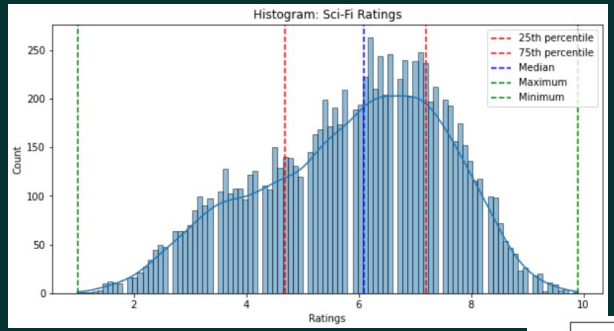
Animation

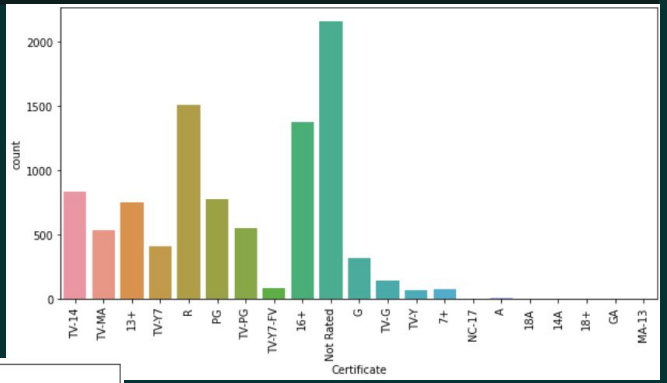

Histogram: Animation Ratings
25th percentile
75th percentile
Median
Maximum
Minimum
Count
Ratings

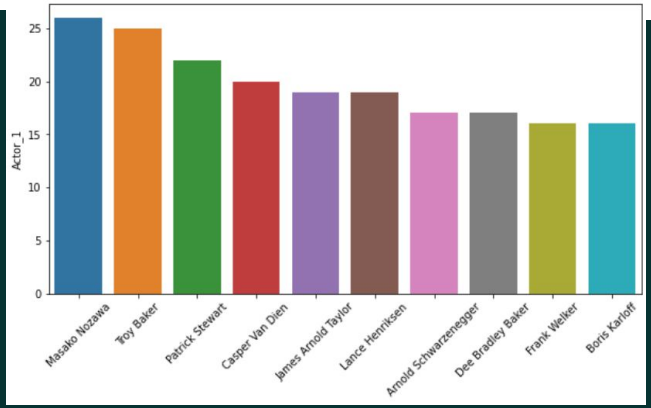

count
Certificate
PG
TV-MA
TV-14
TV-PG
TV-Y7-FV
TV-Y
R
G
TV-Y7
13+
Not Rated
TV-G
16+
7+
A
NC-17
GA
MA-13


Actor_1
Mel Blanc
Frank Welker
Chuck Jones
Robert McKimson
Friz Freleng
Pinto Colvig
Masako Nozawa
Monica Rial
Dave Fleischer
Billy Bletcher

Sci-Fi


Histogram: Sci-Fi Ratings
25th percentile
75th percentile
Median
Maximum
Minimum
Count
Ratings


count
Certificate
TV-14
TV-MA
13+
TV-Y7
R
PG
TV-PG
TV-Y7-FV
16+
Not Rated
G
TV-G
TV-Y
7+
NC-17
A
18A
14A
18+
GA
MA-13


Actor_1
Masako Nozawa
Troy Baker
Patrick Stewart
Casper Van Dien
James Arnold Taylor
Lance Henriksen
Arnold Schwarzenegger
Dee Bradley Baker
Frank Welker
Boris Karloff

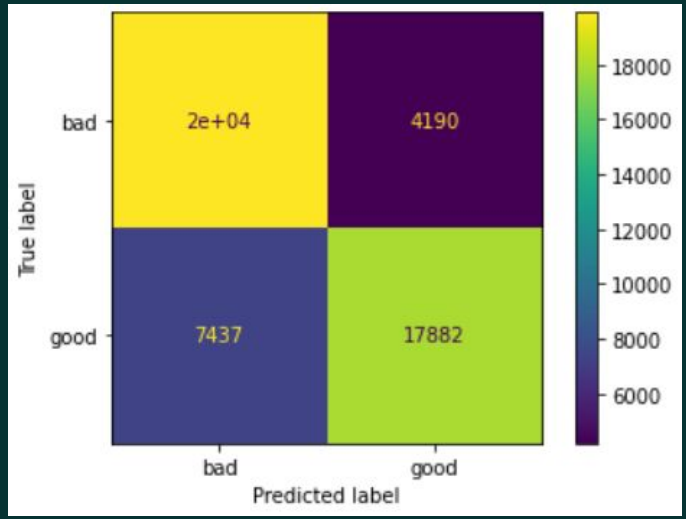
Baseline Model

- The processed dataset was first count vectorized and fit to a Multinomial Naive Bayes model as the baseline.
-

Multinomial Naive Bayes	Accuracy
Train	0.825
Test	0.764 ⬇

- Observations:
  - 0.061 difference between training and testing of data
  - Slightly overfitting of data
- Accuracy is the chosen metric as y true is quite balanced:
  - 1 : 0.511327
  - 2 : 0.488673

Confusion Matrix



Improved Models: Random Forest

- As the baseline model have some overfitting, Random Forest model is explored.

Random Forest		Accuracy	Remarks
Default Setting	Train	0.579	RandomForestClassifier(max_depth = 10, random_state = 42)
	Test	0.574	
Gridsearch	Train	0.579	Best parameters: 'rfc__max_depth': 10, 'rfc__n_estimators': 100, 'rfc__random_state': 42
	Test	0.574	

Improved Models: Logistic Regression

- Logistic Regression model is explored and the results based on F1 score.

Logistic Regression		Accuracy	Remarks
Default Setting	Train	0.931	LogisticRegression(max_iter=500)
	Test	0.829 ⬇	
Gridsearch	Train	0.904	Best parameters: 'lr__C': 0.5, 'lr__penalty': 'l2'
	Test	0.811 ⬇	

- Logistic Regression model is explored and the results based on F1 score
- This model shows signs of over fitting

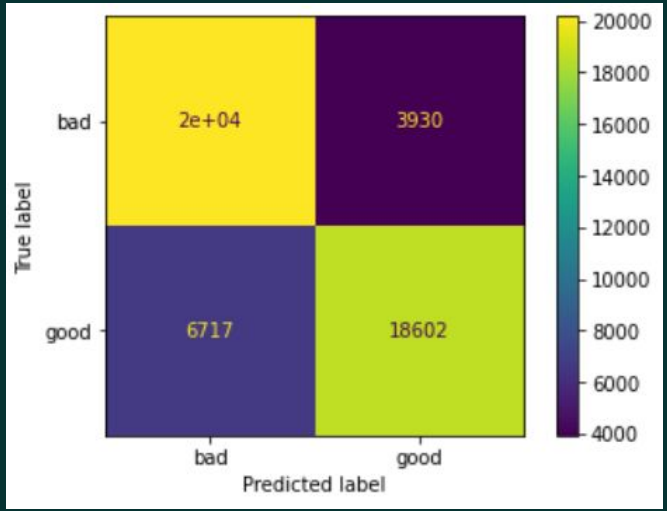
Selected Model: Multinomial Naive Bayes (Stemming)

- Based on the models evaluated, Naive Bayes with stemming is selected as the best model for predicting this dataset.

Naive Bayes		Accuracy	Remarks
Gridsearch	Train	0.855	-Baseline with Stemming 'nb__alpha': 0.1
	Test	0.784	

- Improvement of 0.02 on the test accuracy vs Base model.

Confusion Matrix



Predicting Hold Out Dataset

- Determining Hold out dataset
  - Movies/Tv-shows currently on going
  - Movies/Tv-shows not aired yet

	Genre	Title	Year_produced	Certificate	Ratings	Description	Actor_1	Actor_2	Actor_3	Actor_4	Predicted_values
0	Comedy	Barbie	(2023)	NaN	NaN	To live in Barbie Land is to be a perfect bein...	Margot Robbie	Ariana Greenblatt	Ryan Gosling	Helen Mirren	Good
1	Comedy	Guardians of the Galaxy Vol. 3	(2023)	13+	NaN	Still reeling from the loss of Gamora, Peter Q...	Chris Pratt	Zoe Saldana	Dave Bautista	Vin Diesel	Good
2	Comedy	Lilo & Stitch	NaN	NaN	NaN	Live-action remake of Disney's animated classi...	Sydney Agudong	Billy Magnussen	Tia Carrere	Courtney B. Vance	Good
3	Comedy	Wicked	(2024)	NaN	NaN	The story of how a green-skinned woman framed ...	Michelle Yeoh	Cynthia Erivo	Jeff Goldblum	Ariana Grande	Good
4	Comedy	White Men Can't Jump	(2023)	R	NaN	A remake of the 1992 film about a pair of bask...	Sinqua Walls	Jack Harlow	Lance Reddick	Teyana Taylor	Good

Recommendations / Future Works

Addressing Current Challenges

- Further improve model performance by looking into unigrams/bigrams
- Compare performance of TfidfVectorizer to CountVectorizer
- Include movies that have more than 1 kind of label. E.g Action and Horror and Romance
- To include actor's current career status (active, retired, deceased)
- To consider potential issue of bias in movies e.g horor

Future works

- Can be modified to predicting nominating types of actors to be casted
- Model can be modified to predict genres based on inputs such as: certificate ratings, title, description and actors

Conclusion

Data Visualisation

- Tommy can now identify and shortlist top actors for each genres
- Propose a Drama/Comedy genre as it is the most common type → well received genre
- Proposed to direct a R/16+ type of show as it is the most common → least niche certificate rating

Model Prediction

- Random Forest model has proven to be a model that can be used for solving this problem.
- Our team has made a model which is able to successfully predict whether a movie/TV-show is good/bad , about 70% of the time.
- Tommy is now able to predict if the movie proposal based on the following:
    - Genre
    - Title
    - Description
    - Actors (Top 4)


THANK YOU