Features That Affects Incomes of the Movies in IMDb - Group 11-

Kerem Kurnaz - 26586

Berna Yıldıran - 26431

Taha Can Karaman - 26365

Batuhan Yıldırım - 26478

Outline

- Background
- Problem Definition
- Dataset Description & Utilized Datasets
- Data Exploration
- Correlations & Results & Discussion
- Machine Learning Models
- Conclusion
- Future Work
- Work Division

Introduction

IMDb (International Movie Database)

Movies, television programs, video games, cast, and production crew

⬇

Ratings, genre, cast and crew, storyline, language, release date, runtime, cumulative worldwide gross, trivia, quotes, user reviews, and company credits, personal biographies and details, filmographies, and quotes of the cast and production crew members ...

⬇

- Founded by Col Needham in UK, 1996
- Bought by Jeff Bezos (Amazon) in 1998 for $55 Million

IMDb (International Movie Database)

6.5 million titles & episodes ...

10.4 million personalities ...

WHERE does all this data come from?

83 million registered users !

- contribute information to the website
- can submit new materials
- can edit the existing entries

IMDb (International Movie Database)

Can we TRUST this data created by users ?

Users that have an approved record of submitting data are given instant approval for their corrections and additions to the database.

Changes made in image, name, character name, plot summaries, and titles are screened before publication within 24 - 72 hours.

IMDb (International Movie Database)

Every registered user can rate!

Can we TRUST THE RATINGS that people gave?

Rating Totals

Converted into a Weighted Mean-Rating

Displayed beside each entry title, with online filters employed to deter ballot-stuffing.

Our Project...

Finding the optimal combination of features

To increase the income obtained from the movies.

Analyze different correlations from movie datas

Why are we doing this project?

- to establish realistic connections between features that may seem unrelated with the income obtained from the movie at first.

- The outcome of the project will indicate the strong correlations between features like the duration of the movie, gender ratio of the cast, etc. with the worldwide gross income of the movie.

Problem Definition

The Problem - Income Optimization

Detect the effect of some features of the movies on profits made out of the movie

- Correlations are shaped around “Worldwide Gross Income” of the movies

What are these correlations?

Word Selection in the Title
Genre Choice
Actor/Actress Choice
Gender Ratio of the Cast
Duration
Average Vote Rate

Machine Learning

Random Forest & Decision Tree

- average vote, duration and actor/actress ratio

Dataset Description & Utilized Datasets

IMDb movies extensive dataset

Why did we choose IMDB as a reference ?

IMDb is the most popular movie website and it combines movie plot description, Metascore ratings, critic and user ratings and reviews, release dates, and many more aspects.

- Stores almost every movie and series that has existed (6 million titles !)

Why did we choose this dataset and what does it include?

Dataset consists of 4 different sub-datasets:

1. 85,855 movies with attributes such as movie description, average rating, budget, genre, etc.
2. 85,855 rating details from demographic perspective
3. 297,705 cast members with personal attributes such as birth details, death details, height, spouses, children, etc.
4. 835,513 cast members roles in movies with attributes such as IMDb title id, IMDb name id, order of importance in the movie, role, and characters played.

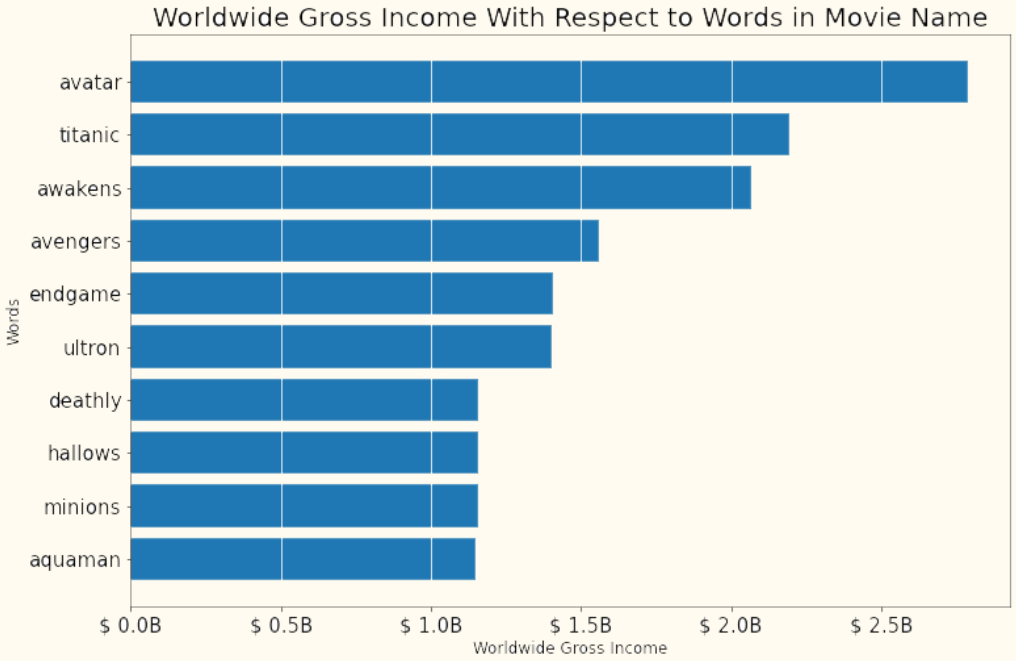
Data Exploration

- The given dataset (Netflix) was not detailed enough
- Tried to merge 2 datasets but did not work
- Discarded the Netflix dataset
- Worked on the new dataset that found

Correlations

- Movie Title Keywords - Worldwide Gross Income
- Movie Genre - Worldwide Gross Income
- Actor/Actress - Worldwide Gross Income
- Gender Ratio of the Cast - Worldwide Gross Income
- Durations - Worldwide Gross Income
- Average Vote Rate - Worldwide Gross Income
- Gender/ Age /Average Vote Rate - Duration

Movie Title Keywords - Worldwide Gross Income



- The mean income value for each word in movie names
- Some words dominated the dataset
- Therefore, excluded from algorithm

Movie Genre - Worldwide Gross Income

How does the choice of genre affect the worldwide gross income?

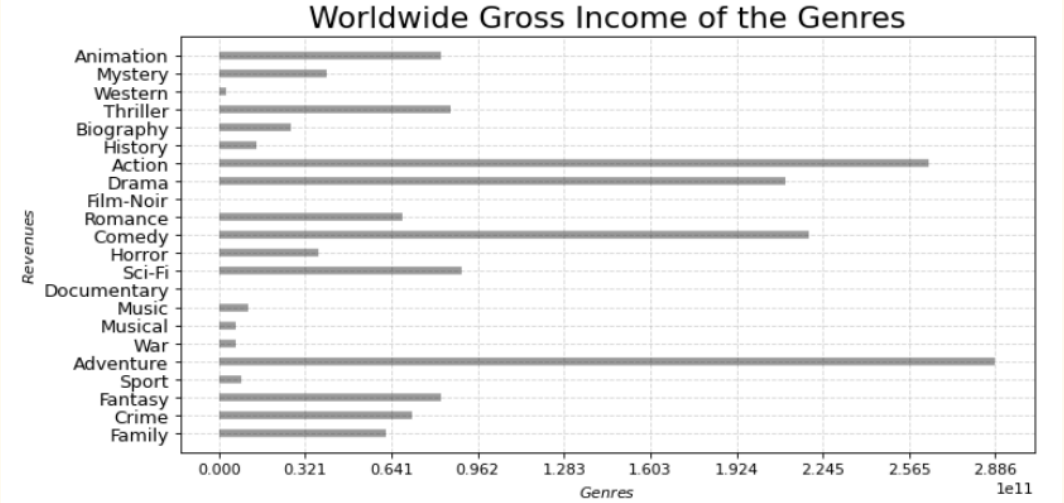
(*These genres are suitable for all ages)

- Adventure*, Action*, Comedy, and Drama
- Discarded Film-noir and Documentary.

(High sample size!)

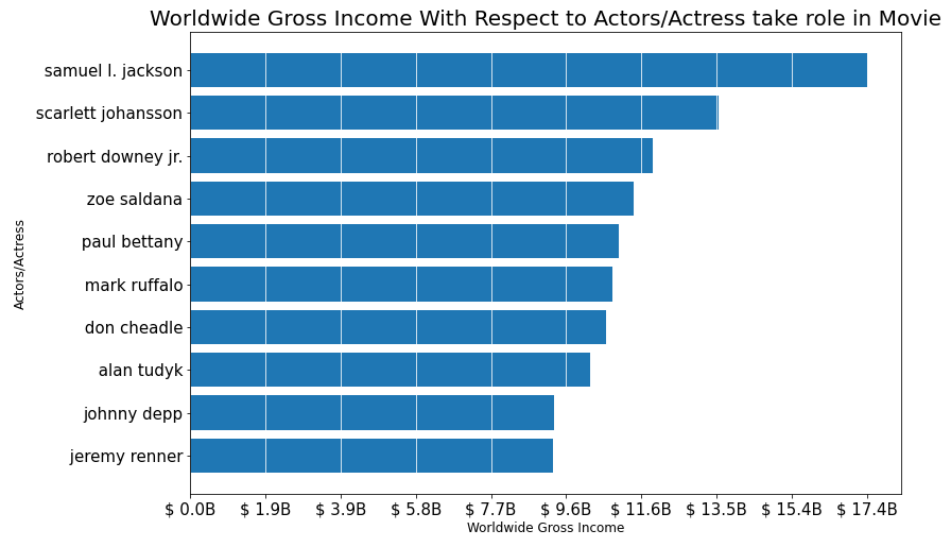
- We expected Drama and Comedy to be higher.
- Genres with low samples?

	genre	total_no_movies_with_genre
15	Documentary	1
16	Horror	1236
17	Sport	265
18	Biography	740
19	Film-Noir	17



Actor/Actress - Worldwide Gross Income

- Dominated actors / Actress took role in same movies
- Outliers has a lot of effect
- Categorical data
- Too many categories
- Inefficient for Machine learning



Gender Ratio of the Cast - Worldwide Gross Income

What is the correlation between gender ratio and the worldwide gross income? **[Ratio=Actress/Actor]**

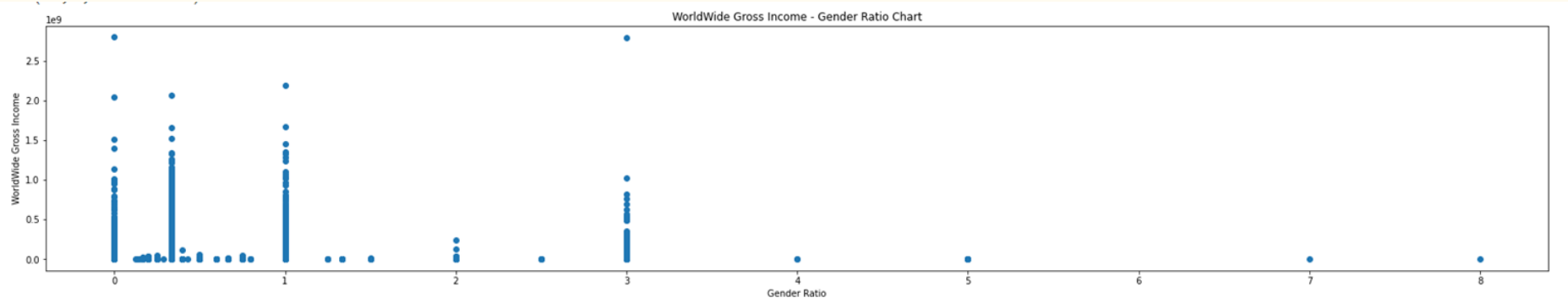
Gender ratio <= 1, higher worldwide income is expected

- Movies that gave more roles to actors rather than actresses achieve higher worldwide gross income

There is a CORRELATION!

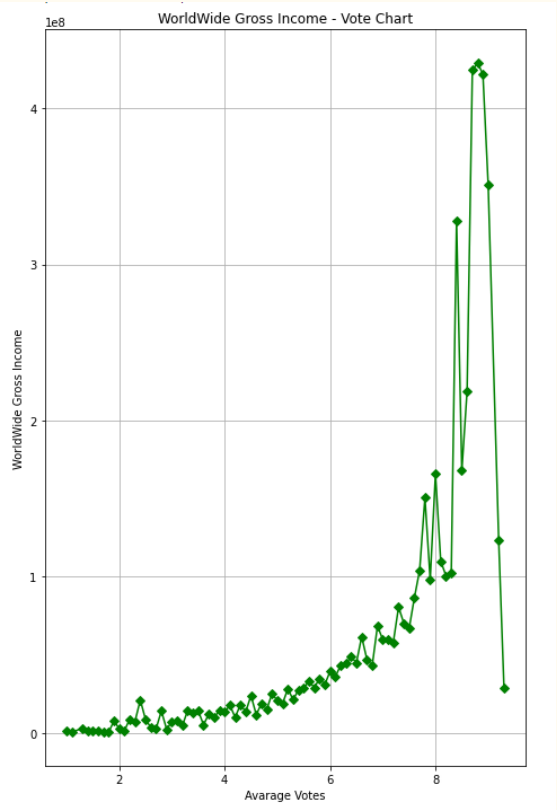
Why is there such a correlation?

The INEQUALITY of the genders and male DOMINATION

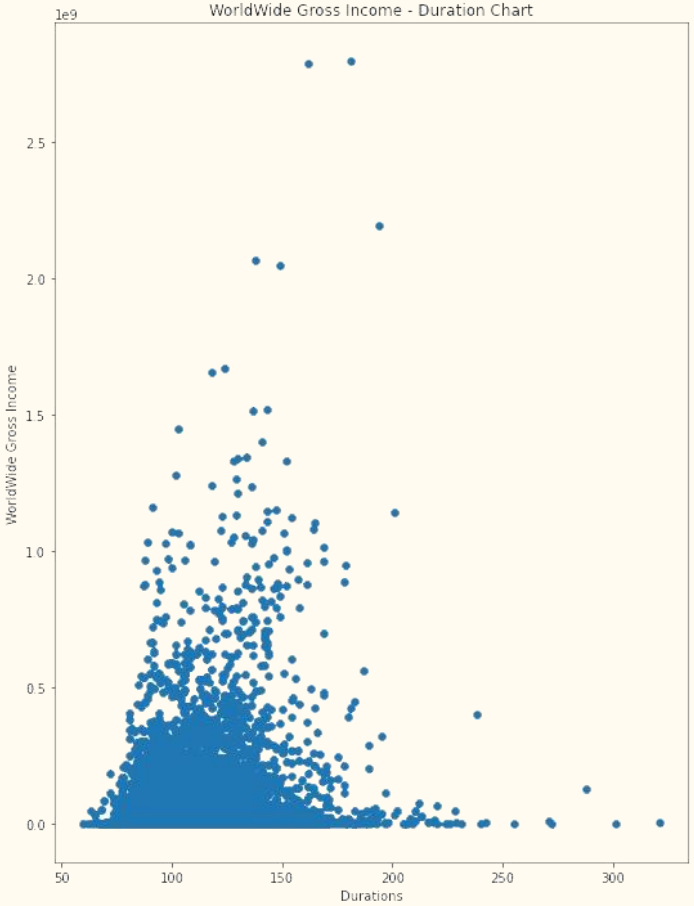


Average Vote Rate - Worldwide Gross Income

- Obvious correlation until 9
- Less effect of outliers
- Negative correlation after 9
- Numeric values
- Suitable for machine learning
- Sometimes unavailable before casting



Durations - Worldwide Gross Income

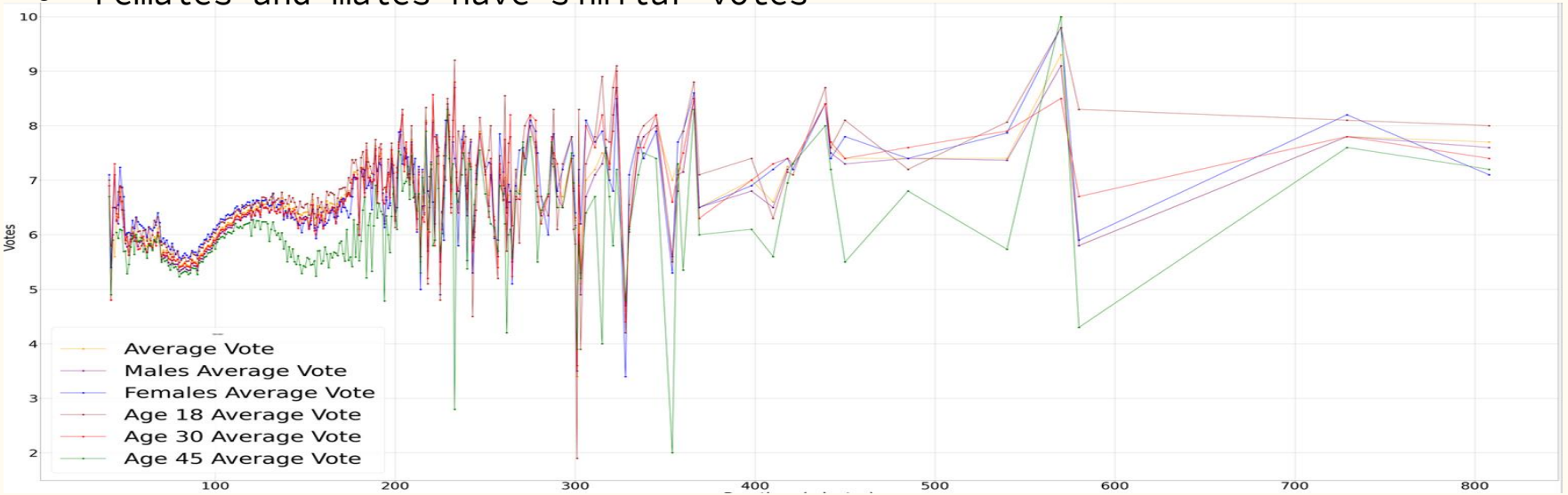


- shows each income value of every movie with respect to duration
- there are some outliers that have a huge impact on mean income
- not a good idea to include the mean value to the algorithm

Duration - Gender ,Age , Votes Correlation

Plotting the graph of the votes with respect to duration for different age and gender groups.

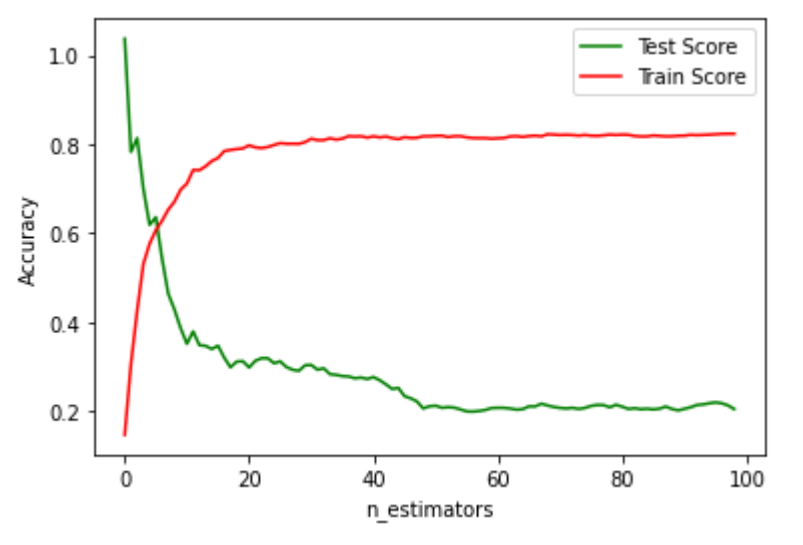
- Old people’s ratings are lower generally
- Young people votes higher, above average
- Females and males have similar votes



Machine Learning Models

Parameter Tuning

- Searching for best n_estimator
- Test score
- Train score
- Difference and overfitting
- Optimum point



Conclusion

Effective Correlations

- Durations
- Average Vote Rates
- Gender ratio of the cast

Ineffective Correlations

- Movie Title Keywords
- Movie Genre
- Actor / Actress

- Machine Learning Models
- Accuracies

Future Work

- Graphical User Interface
- More features for ML model
- ML model improvements
- More visualizations for the model

Can	Kerem	Batuhan	Berna
Movie Title Keywords - Worldwide Gross Income Correlation (Data Process)	Genre - Worldwide Gross Income Correlation (Data Process)	Actor - Worldwide Gross Income Correlation (Data Process)	Finding related datasets that will be used in the project
Machine Learning - Decision Tree Implementation (¼)	Machine Learning - Decision Tree Implementation (¼)	Machine Learning - Decision Tree Implementation (¼)	Gender Ratio of the Cast - Worldwide Gross Income (Data Process)
Helping and trying to resolve the issues in other team members codes	Helping and trying to resolve the issues in other team members codes	Helping and trying to resolve the issues in other team members codes	Helping and trying to resolve the issues in other team members codes
Writing Data Exploration - Data Preprocessing part of the project report	Writing description of the datasets (½)	Writing Data Exploration - Methods used in the project and feature generation part of the project report	Machine Learning - Decision Tree Implementation (1/4)
Creating division of work table among the team members	Creating division of work table among the team members	Creating division of work table among the team members	Writing the Introduction and the problem definition part of the project report
Durations - Worldwide Gross Income Correlation (Data Process)	Gender/ Age /Average Vote Rate - Duration Correlation (Data Process)	Average Vote Rate - Worldwide Gross Income Correlation (Data Process)	Writing description of the datasets (½)
Writing Results & Discussion of 1. Movie Title Keywords - Worldwide Gross Income Correlation 5. Durations - Worldwide Gross Income Correlation	Writing Results & Discussion of 2. Movie Genre - Worldwide Gross Income Correlation 7. Gender/ Age /Average Vote Rate - Duration Correlation	Writing Results & Discussion of 6. Average Vote Rate - Worldwide Gross Income Correlation	Creating division of work table among the team members
Machine Learning - Random Forest Implementation(¼)	Machine Learning - Random Forest Implementation(¼)	Machine Learning - Random Forest Implementation(¼)	General format and language check of the project report
			Writing Results & Discussion of 3. Actor/Actress - Worldwide Gross Income Correlation 4. Gender of the Cast - Worldwide Gross Income Correlation
			Machine Learning - Random Forest Implementation(¼)

OUR PROJECT