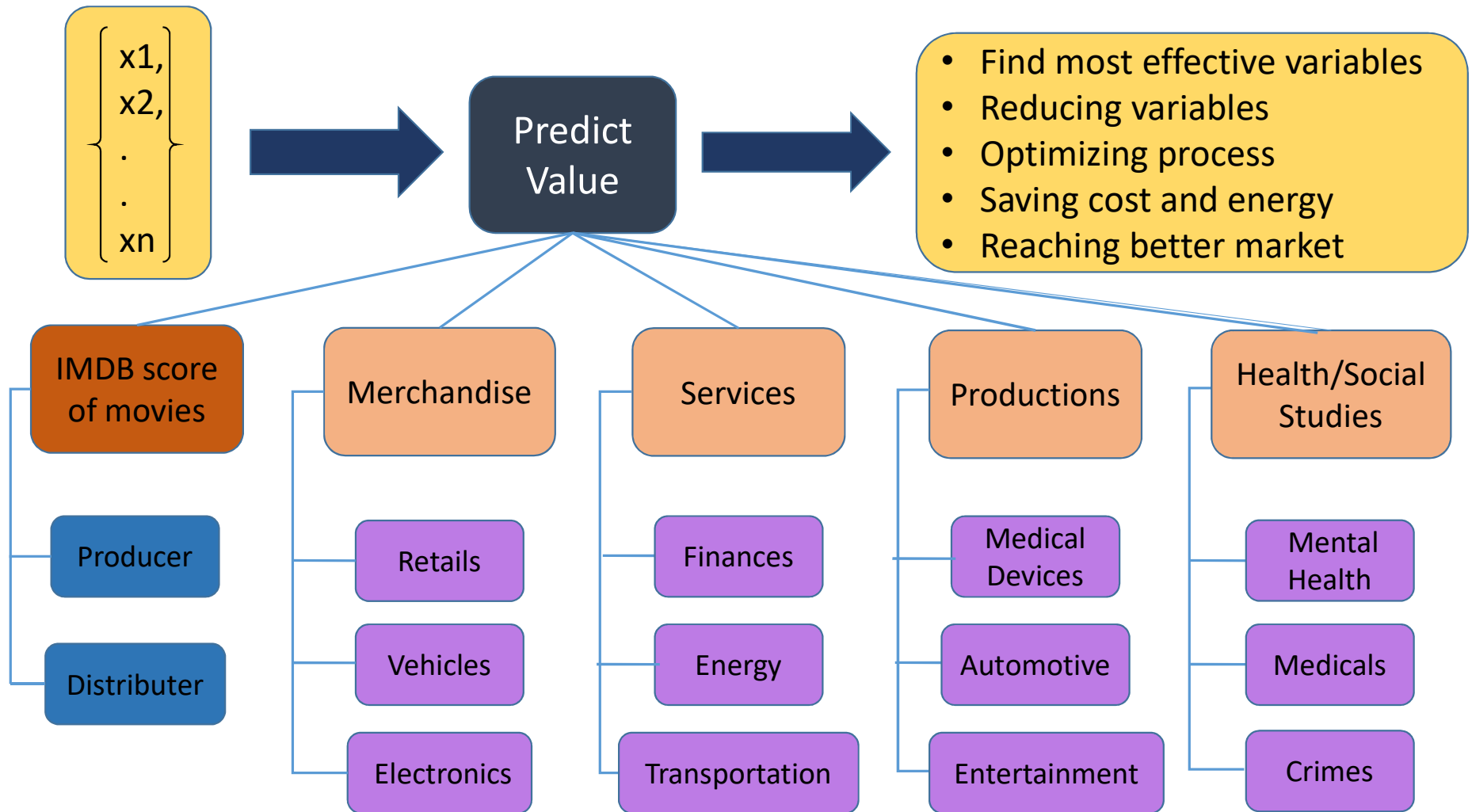


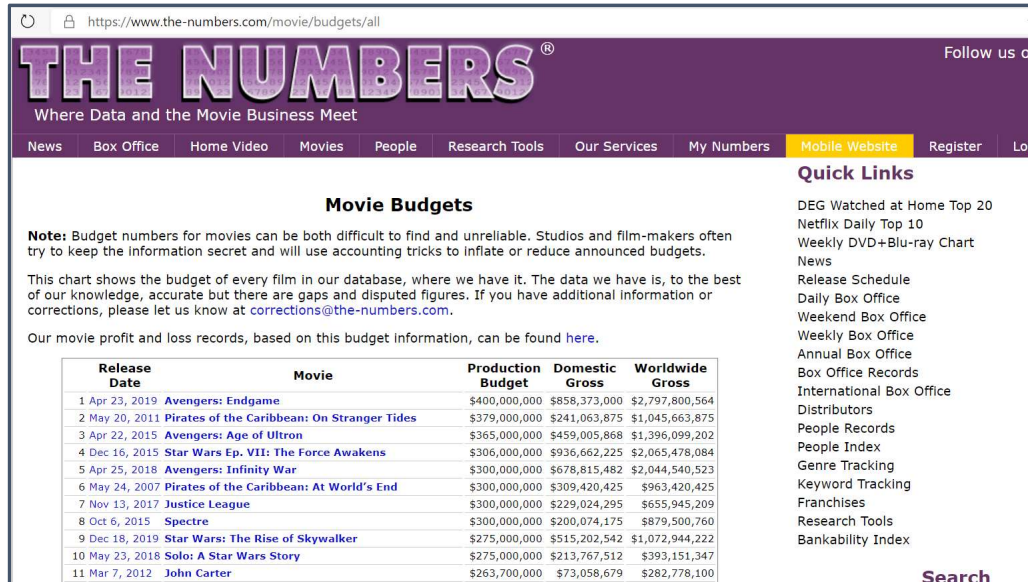
Prediction of IMDB Score of movies



Prediction of IMDB Score of movies

- Extraction of data of more than 5000 movies from different websites:
 - <https://www.the-numbers.com/movie/budgets/all>
 - Movie title, release date, production budget, domestic gross and worldwide gross
 - Link of each movie from: <https://www.the-numbers.com/movie>
 - MPAA, duration, genre, keywords, language, country and company
 - More details from: <https://www.imdb.com>
 - IMDB score, director, crews, color, aspect ratio, number of votes and critics
- Data cleaning and selecting meaningful data
 - Nulls, dtypes, duplications, unity, outliers, one-hot-encoding
- Data analysis
 - Distribution and deviation of different features
 - Correlations between different features
- Prediction methods (ML)
 - Regression (Linear and Non-Linear)
 - Random Forest
 - Neural Network
- Metrics

Data Extraction



The screenshot shows the 'THE NUMBERS' website with the title 'Where Data and the Movie Business Meet'. The navigation bar includes links for News, Box Office, Home Video, Movies, People, Research Tools, Our Services, My Numbers, Mobile Website, Register, and Log. The 'Movie Budgets' section contains a note about budget reliability and a table of movie budgets. The table has columns for Release Date, Movie, Production Budget, Domestic Gross, and Worldwide Gross. The first 11 rows of the table are visible, showing movies like Avengers: Endgame, Pirates of the Caribbean: On Stranger Tides, and Star Wars: The Force Awakens. A 'Quick Links' sidebar on the right lists various data sources and tools. A search bar is located at the bottom right of the table.

Release Date	Movie	Production Budget	Domestic Gross	Worldwide Gross
1 Apr 23, 2019	Avengers: Endgame	\$400,000,000	\$858,373,000	\$2,797,800,564
2 May 20, 2011	Pirates of the Caribbean: On Stranger Tides	\$379,000,000	\$241,063,875	\$1,045,663,875
3 Apr 22, 2015	Avengers: Age of Ultron	\$365,000,000	\$459,005,868	\$1,396,099,202
4 Dec 16, 2015	Star Wars Ep. VII: The Force Awakens	\$306,000,000	\$936,662,225	\$2,065,478,084
5 Apr 25, 2018	Avengers: Infinity War	\$300,000,000	\$678,815,482	\$2,044,540,523
6 May 24, 2007	Pirates of the Caribbean: At World's End	\$300,000,000	\$309,420,425	\$963,420,425
7 Nov 13, 2017	Justice League	\$300,000,000	\$229,024,295	\$655,945,209
8 Oct 6, 2015	Spectre	\$300,000,000	\$200,074,175	\$879,500,760
9 Dec 18, 2019	Star Wars: The Rise of Skywalker	\$275,000,000	\$515,202,542	\$1,072,944,222
10 May 23, 2018	Solo: A Star Wars Story	\$275,000,000	\$213,767,512	\$393,151,347
11 Mar 7, 2012	John Carter	\$263,700,000	\$73,058,679	\$282,778,100

- Selecting correct table
- Extract movie links from titles
- Generate address of other pages

96	Nov 14, 2016	Fantastic Beasts and Where to Find Them	\$180,000,000	\$234,037,575	\$811,724,385
97	Jun 27, 2008	WALL-E	\$180,000,000	\$223,808,164	\$532,508,025
98	Aug 10, 2007	Rush Hour 3	\$180,000,000	\$140,125,968	\$256,585,882
99	Jun 27, 2016	The Legend of Tarzan	\$180,000,000	\$126,643,061	\$348,902,025
100	Nov 23, 2011	Hugo	\$180,000,000	\$73,864,507	\$180,047,784

1-100

101-200

201-300

301-400

401-500

501-1,000 »

Data Extraction

Unnamed: 0	ReleaseDate	Movie	ProductionBudget	DomesticGross	WorldwideGross	movie_link
97	98 Aug 10, 2007	Rush Hour 3	\$180,000,000	\$140,125,968	\$256,585,882	https://www.the-numbers.com/movie/Rush-Hour-3#...
98	99 Jun 27, 2016	The Legend of Tarzan	\$180,000,000	\$126,643,061	\$348,902,025	https://www.the-numbers.com/movie/Legend-of-Ta...
99	100 Nov 23, 2011	Hugo	\$180,000,000	\$73,864,507	\$180,047,784	https://www.the-numbers.com/movie/Hugo#tab=sum...
0	101 Jul 20, 2017	Valerian and the City of a Thousand Planets	\$180,000,000	\$40,479,370	\$215,098,356	https://www.the-numbers.com/movie/Valerian-and...
1	102 Feb 6, 2015	Jupiter Ascending	\$179,000,000	\$47,482,519	\$181,982,519	https://www.the-numbers.com/movie/Jupiter-Asce...
2	103 Jul 16, 2018	Mission: Impossible—Fallout	\$178,000,000	\$220,159,104	\$787,176,729	https://www.the-numbers.com/movie/Mission-Impo...

- Repeated movie name
- Fix Indices
- Deal with problematic characters

Avengers: Endgame (2019)

Theatrical Performance

Domestic Box Office	\$858,373,000	Details
International Box Office	\$1,939,427,564	Details
Worldwide Box Office	\$2,797,800,564	

Home Market Performance

Est. Domestic DVD Sales	\$23,661,532	Details
Est. Domestic Blu-ray Sales	\$83,234,207	Details
Total Est. Domestic Video Sales	\$106,895,739	

[Further financial details...](#)

Quick Links

- DEG Watched at Home Top 20
- Netflix Daily Top 10
- Weekly DVD+Blu-ray Chart
- News
- Release Schedule
- Daily Box Office
- Weekend Box Office
- Weekly Box Office
- Annual Box Office
- Box Office Records
- International Box Office
- Distributors
- People Records
- People Index
- Genre Tracking
- Keyword Tracking
- Franchises
- Research Tools
- Bankability Index

Synopsis

The grave course of events set in motion by Thanos that wiped out half the universe and fractured the Avengers ranks compels the remaining Avengers to take one final stand in Marvel Studios' grand conclusion to twenty-two films, "Avengers: Endgame."

Metrics

Opening Weekend: \$357,115,007 (41.6% of total gross)
 Legs: 2.40 (domestic box office/biggest weekend)
 Domestic Share: 30.7% (domestic box office/worldwide)

- Go to the link of each movie
- Select the correct table
- Transpose and define columns
- Add columns to DataFrame
- Handle lack of data

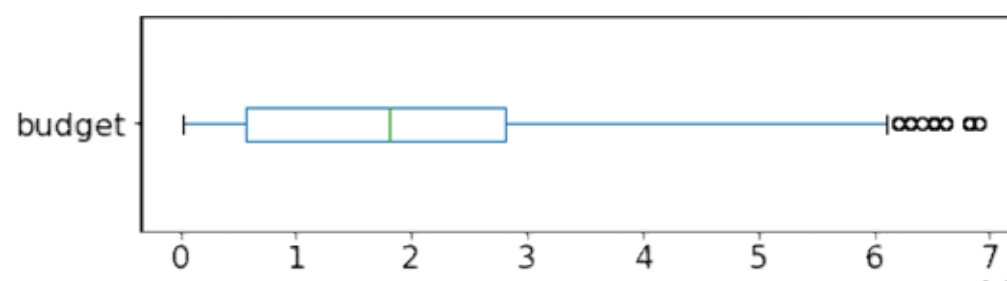
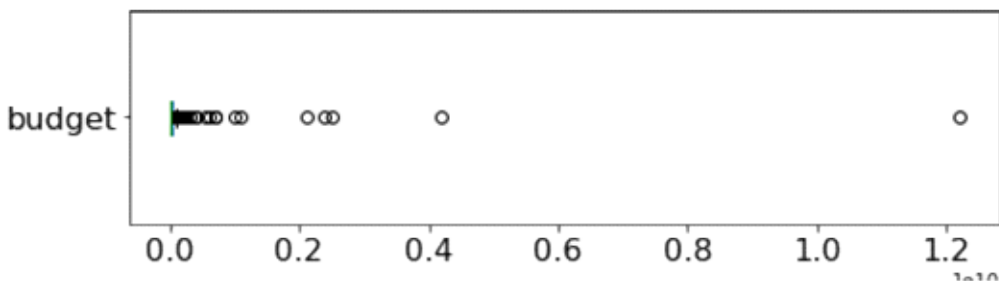
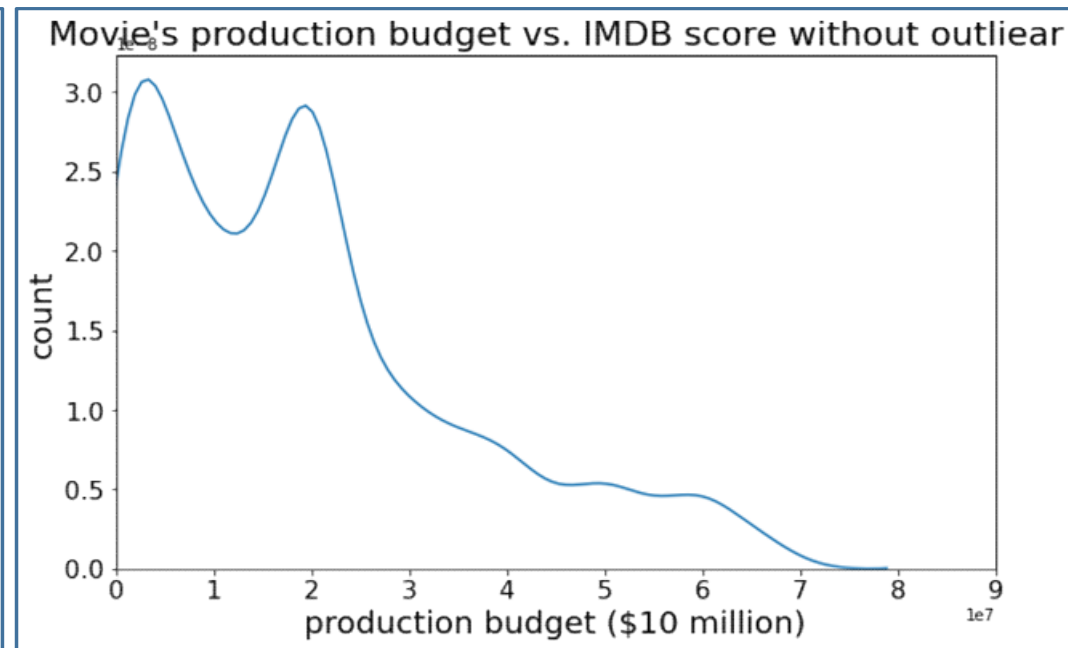
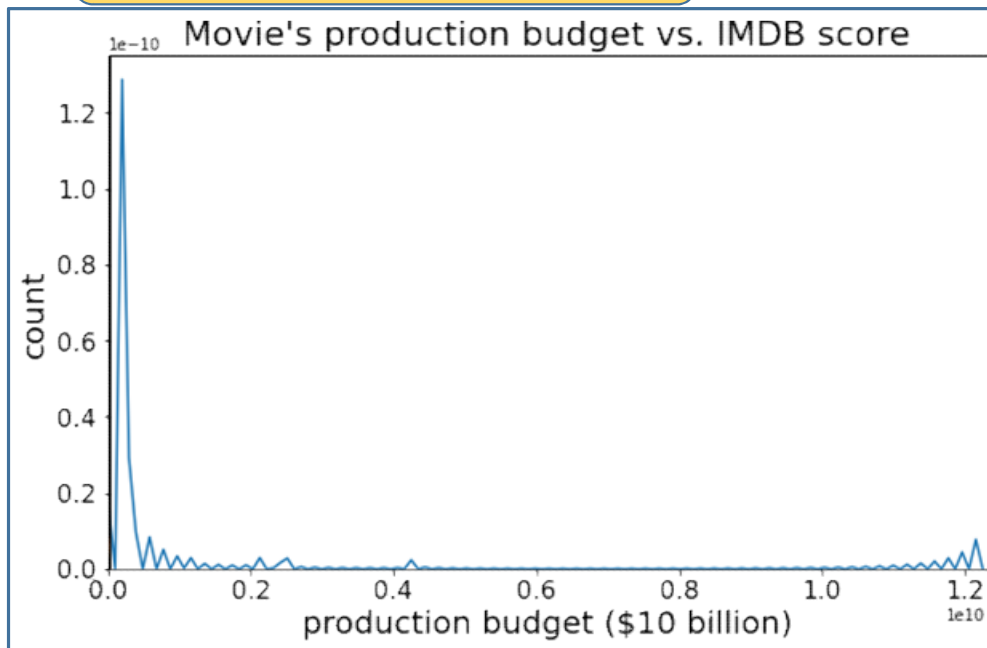
- Go to each movie's IMDB page and extract more data

Data Cleaning

- Repeated movies
- Problematic characters
- Nulls for each feature
 - Budget, gross
 - Color
 - Language, country, company
- Data types
- Redundancy
 - Genre, MPAA rating
- Unity
- One-Hot-Encoding
- Normalizing
- Distributions, STDs and Outlier

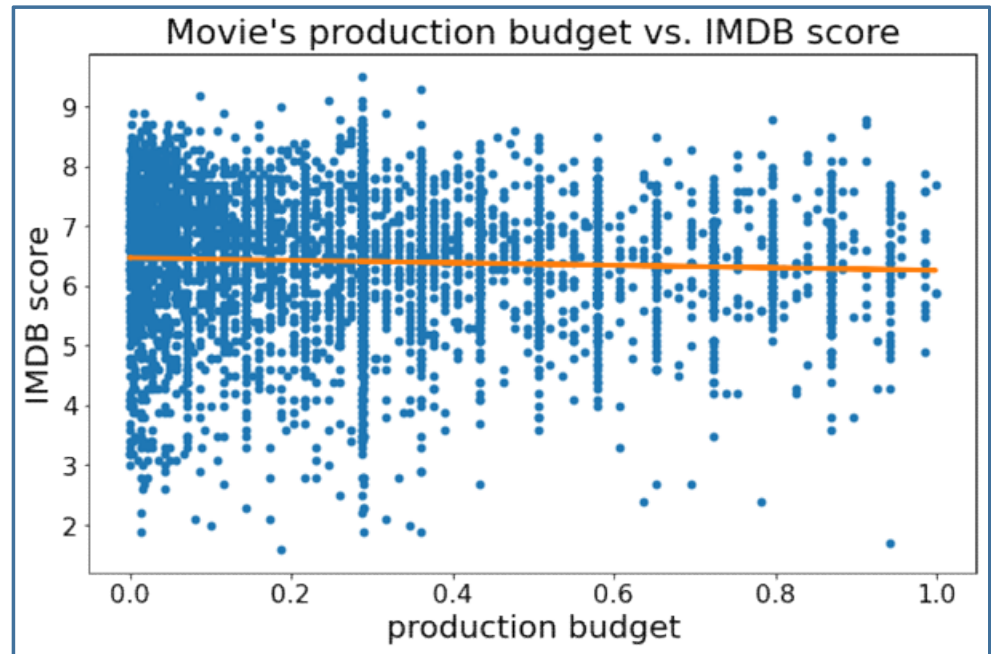
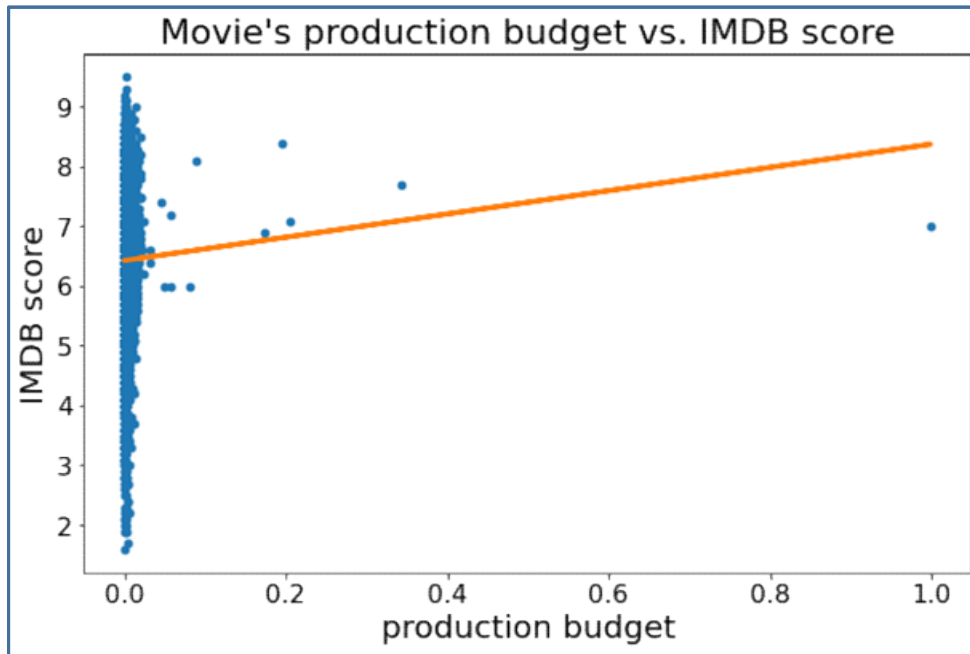
Distributions and Outliers

- Budget Distribution:



Data Analysis

- Budget vs IMDB score:



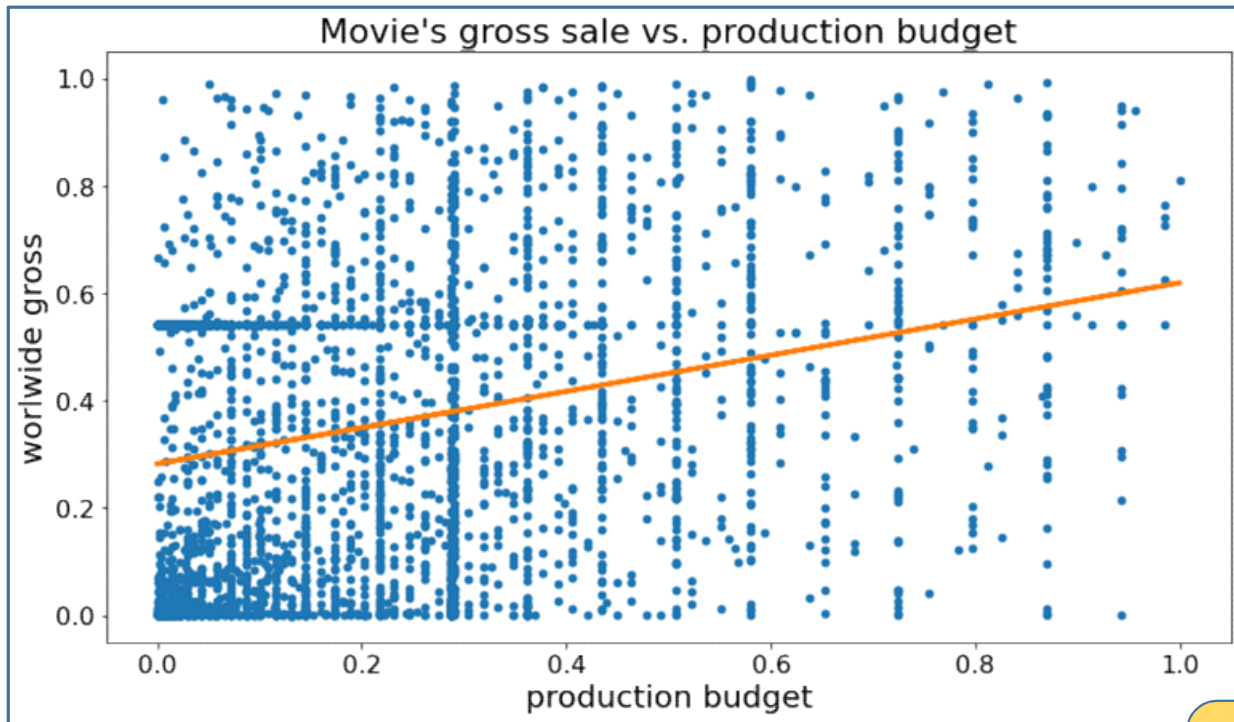
R^2 : .002

Pearson Corr.: -.044

MSE: 1.306

Data Analysis

- Budget vs Gross:



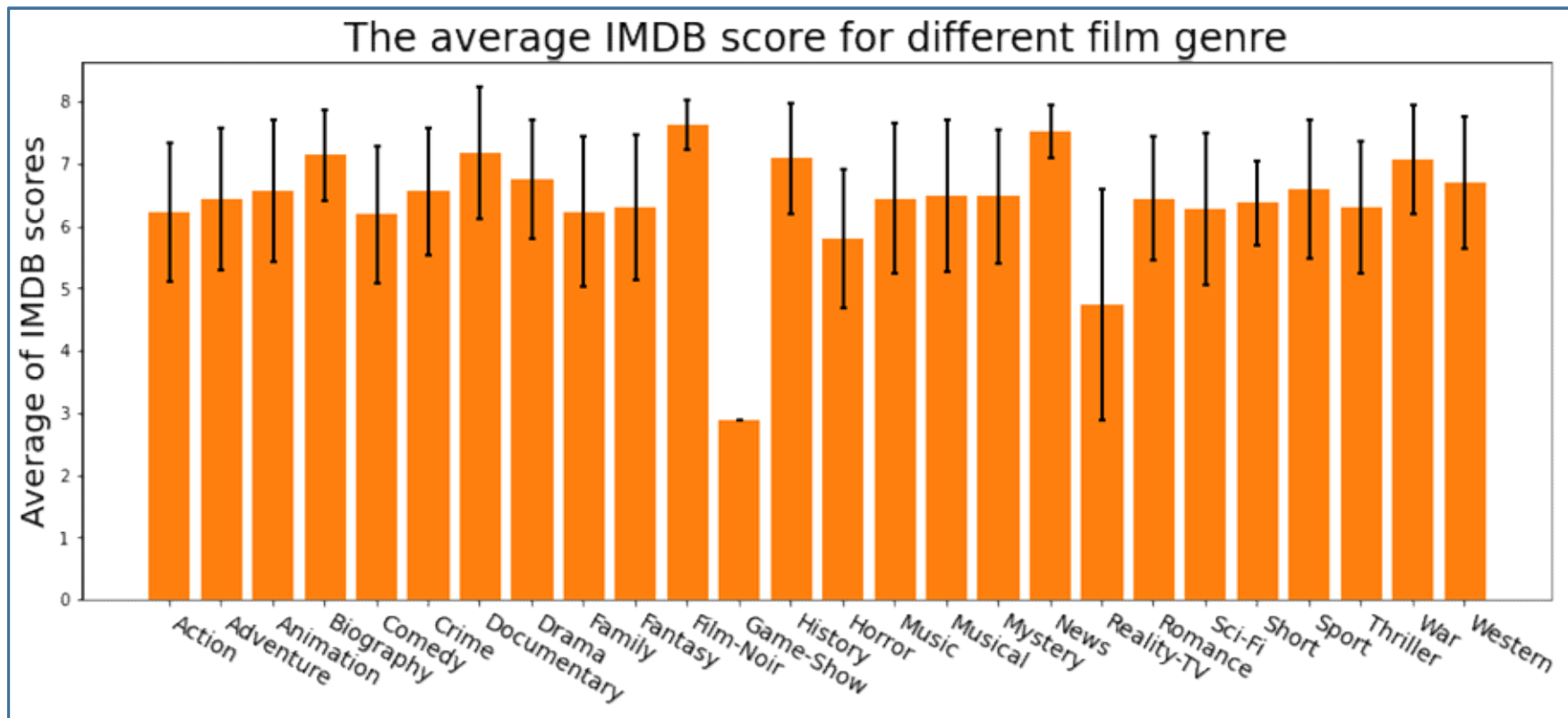
R^2 : .067

Pearson Corr.: .259

MSE: 37.295

Data Analysis

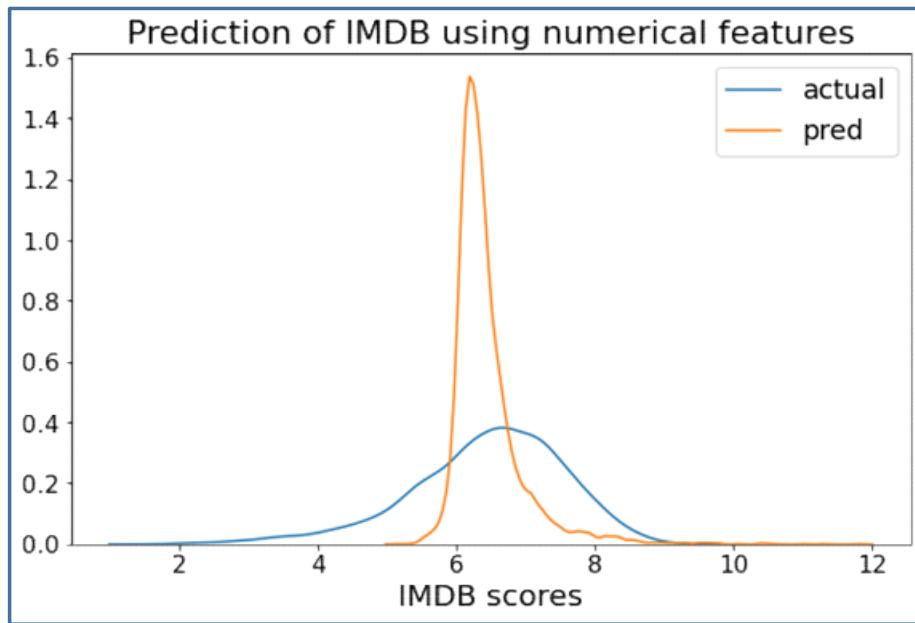
- Genres vs IMDB score:



Prediction

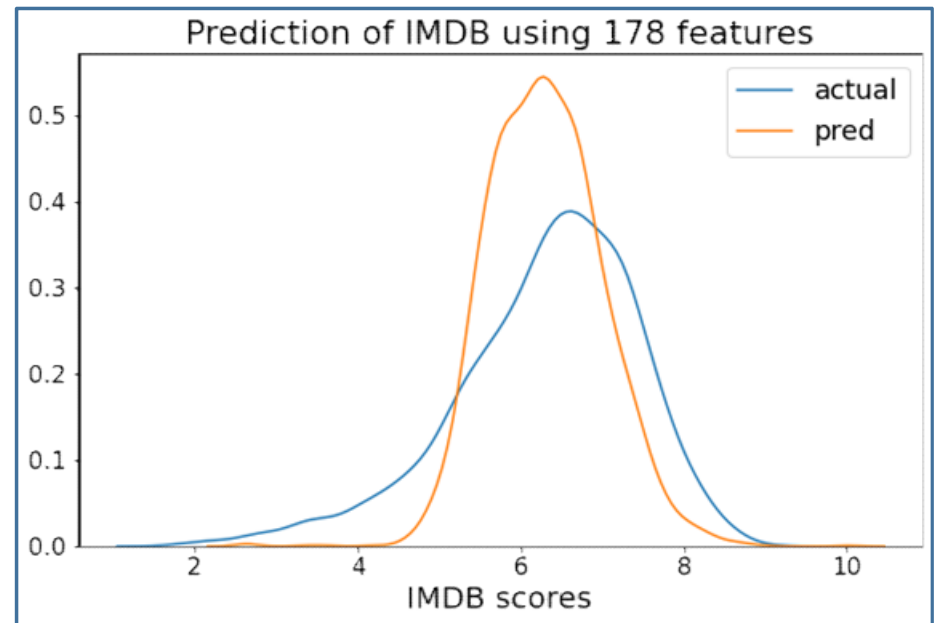
- Linear Regression:

Numerical features: 7



R^2 : .215
Pearson Corr.: .464
MSE: .997

Numerical + categorical features: 178



R^2 : .349
Pearson Corr.: .591
MSE: .827

Data Analysis

- Linear Regression:
 - Year
 - Director
 - Actors
- Non-Linear Regression
- Random Forest
- Neural Network