



Predicting Piracy Rates of Movies using ML Algorithms

Alisha Baral

Background

- Piracy has significant economic impact on the entertainment industry. It is the unauthorized use or reproduction of copyrighted works, and it has become increasingly prevalent with the rise of the internet and digital media.
- ML algorithms are a powerful tool that can be used to predict piracy rates of movies by analyzing various factors that can influence piracy.

About this data

- The movie dataset used for this projects was taken from Kaggle where it was gathered from a pirated website that has a user base of around 2M visitors per month.
- It contains information about movies, including details such as movie tittle, genre, director, release date, runtime, storyline, language, writer, and user ratings.

Research question

- 1) What factors influence the number of downloads (Also meaning mostly pirated) a movie receives? How much of an impact does the industry, language, or release date have on the number of downloads?
- 2) Does the IMDb rate of any movie affect the download number of the movie? What is the highest pirated movie together and from each industry?
- 3) Which industry is mostly affected by piracy?

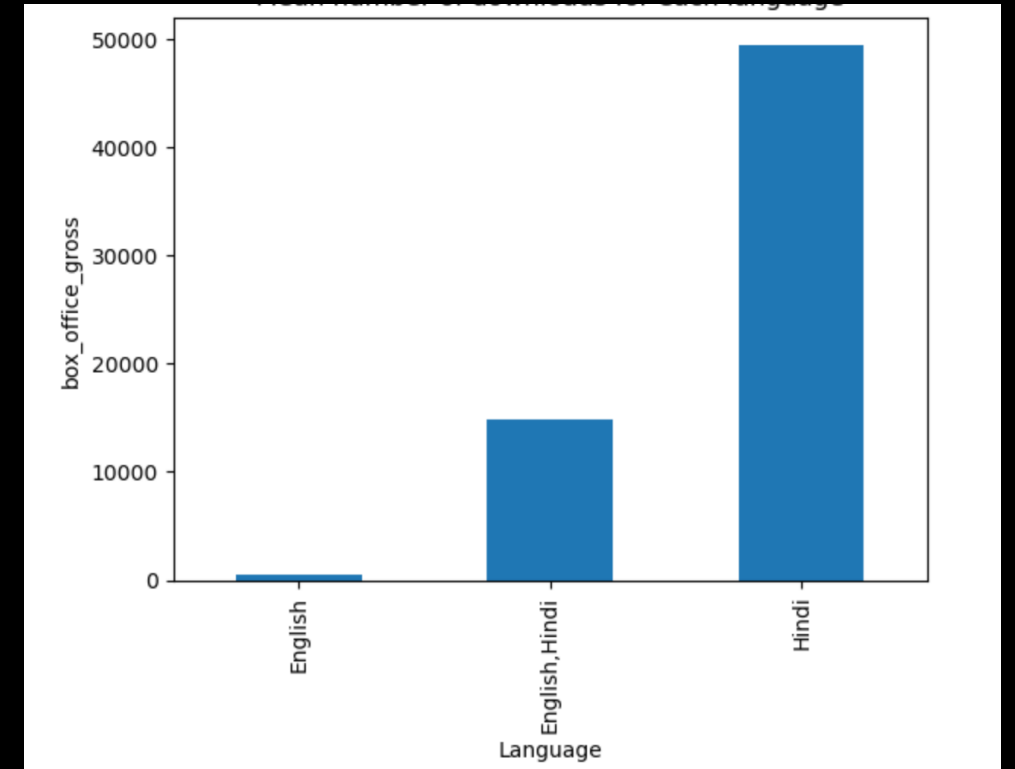
Mdb-rating	appropriate_for	director	downloads	id	industry	language	posted_date	release_date	run_time	storyline	title	views	writer
4.8	R	John Swab	304	372092	Hollywood / English	English	20 Feb, 2023	Jan 28 2023	105	Doc\r\nfacilitates a fragile truce between th...	Little Dixie	2,794	John Swab
6.4	TV-PG	Paul Ziller	73	372091	Hollywood / English	English	20 Feb, 2023	Feb 05 2023	84	Caterer\r\nGoldy Berry reunites with detectiv...	Grilling Season: A Curious Caterer Mystery	1,002	John Christian Plummer
5.2	F	Ben Wheatley	1,42	13381	Hollywood / English	English,Hindi	20 Apr, 2021	Jun 18 2021	1h 47min	As the world searches for a cure to a disastro...	In the Earth	14,419	Ben Wheatley
8.1	NaN	Venky Atluri	1,549	372090	Tollywood	Hindi	20 Feb, 2023	Feb 17 2023	139	The life of a young man and his struggles agai...	Vaathi	4,878	Venky Atluri
4.6	NaN	Shaji Kailas	657	372089	Tollywood	Hindi	20 Feb, 2023	Jan 26 2023	122	A man named Kalidas gets stranded due to the p...	Alone	2,438	Rajesh Jayaraman
...

How does
the data
look?

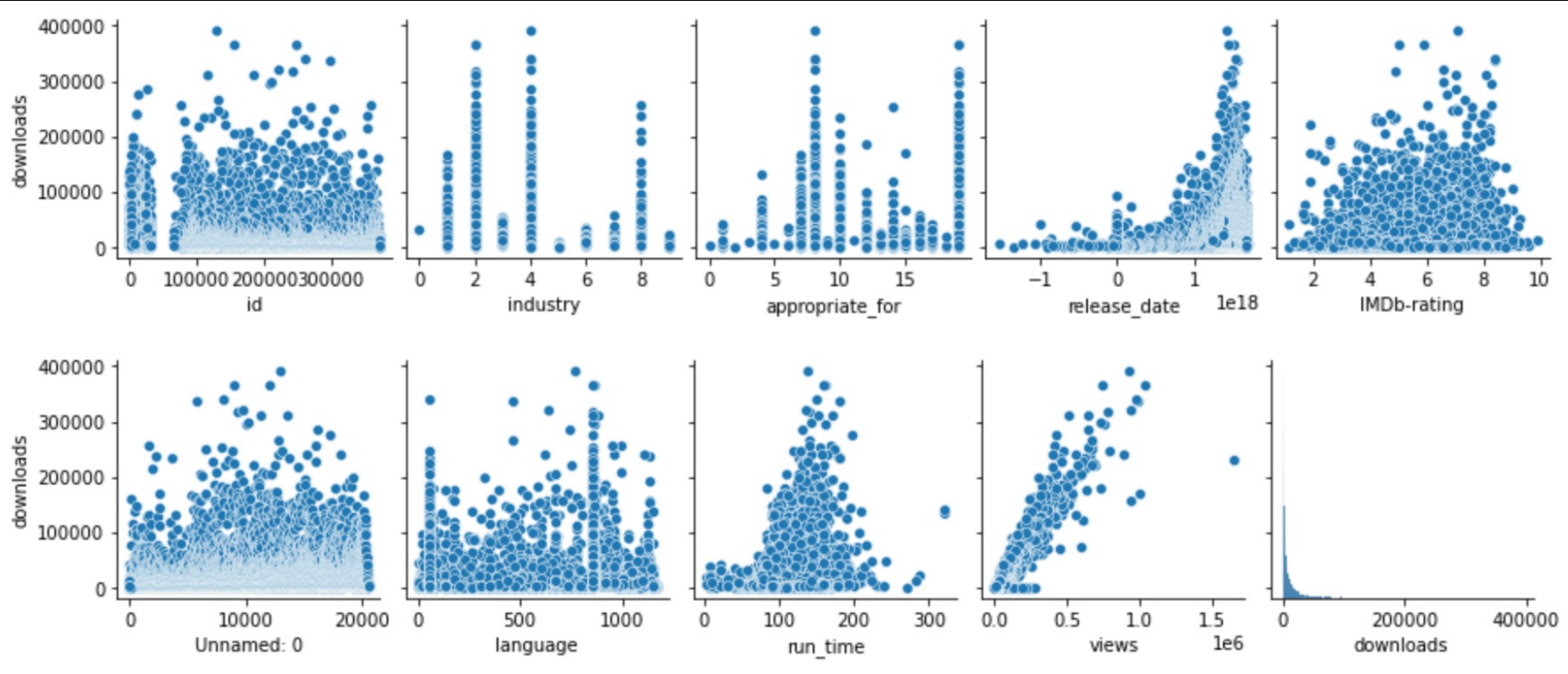


Data preprocessing and EDA

Date ▾	Day	Day #	Top 10 Gross ▾	%± YD ▾	%± LW ▾	Releases ▾	#1 Release	Gross ▾
Dec 31 New Year's Eve	Saturday	365	\$27,962,493	+1.3%	-40.3%	31	Avatar: The Way of Water	\$18,053,159
Dec 30	Friday	364	\$37,900,613	+21.5%	+29.7%	32	Avatar: The Way of Water	\$24,836,835
Dec 29	Thursday	363	\$31,204,428	-2.1%	+55.2%	28	Avatar: The Way of Water	\$20,117,061
Dec 28	Wednesday	362	\$31,866,473	-14.7%	+49.5%	28	Avatar: The Way of Water	\$20,582,014
Dec 27	Tuesday	361	\$37,343,124	-20.5%	+70.3%	28	Avatar: The Way of Water	\$24,128,503
Dec 26	Monday	360	\$46,963,277	+8%	+146.5%	29	Avatar: The Way of Water	\$32,270,430
Dec 25 Christmas Day	Sunday	359	\$43,485,546	-23%	-54.6%	28	Avatar: The Way of Water	\$29,179,791
Dec 24	Saturday	358	\$21,255,461	-27.2%	-58.7%	28	Avatar: The Way of Water	\$14,869,288
Dec 23	Friday	357	\$29,211,733	+45.3%	-49.5%	29	Avatar: The Way of Water	\$19,289,141
Dec 22	Thursday	356	\$20,101,091	-5.7%	+780.5%	30	Avatar: The Way of Water	\$14,632,040
Dec 21	Wednesday	355	\$21,315,270	-2.8%	+606.1%	31	Avatar: The Way of Water	\$14,403,438
Dec 20	Tuesday	354	\$21,925,733	+15.1%	+457.5%	31	Avatar: The Way of Water	\$18,288,904

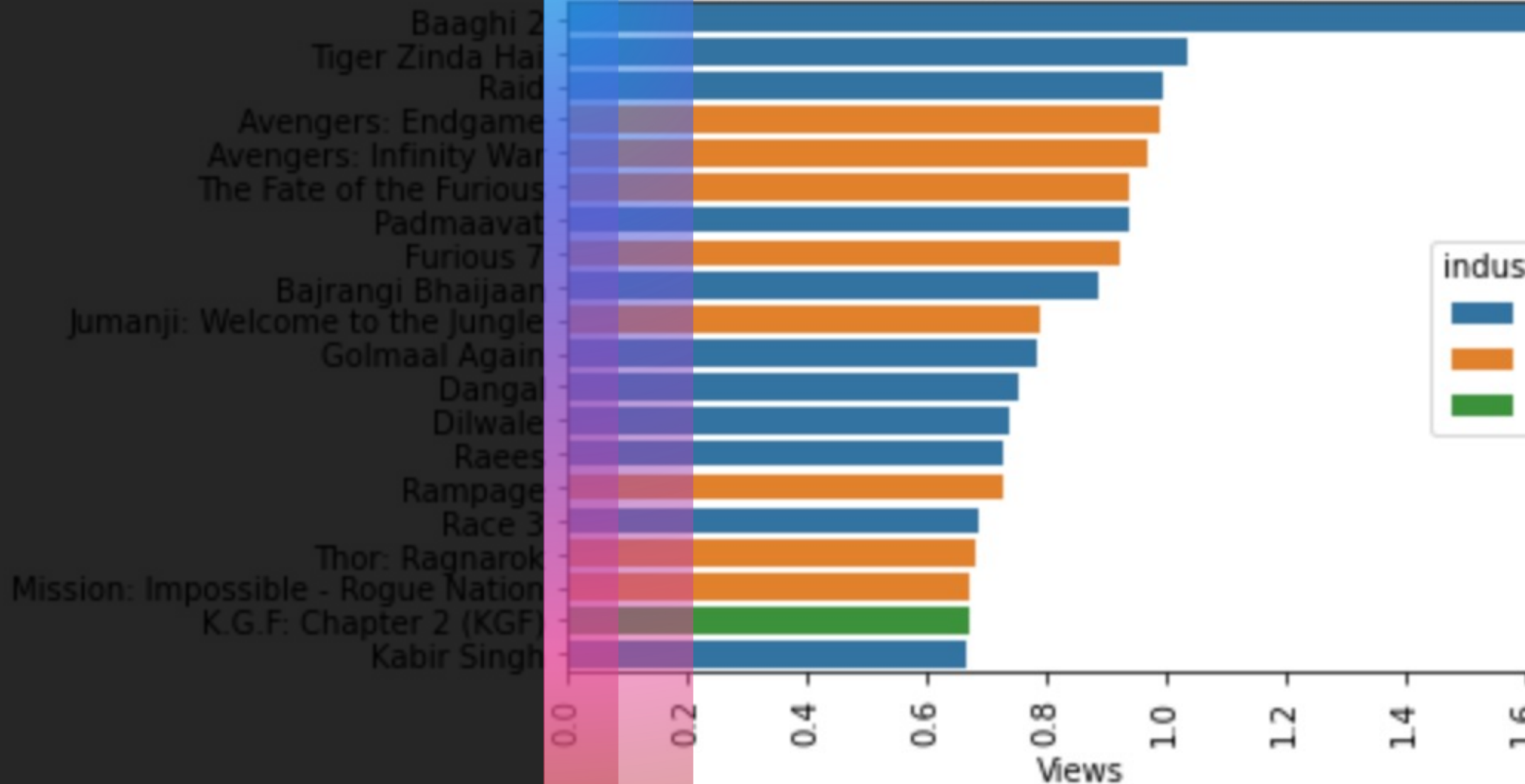


Limitations of Integrating Box Office Data

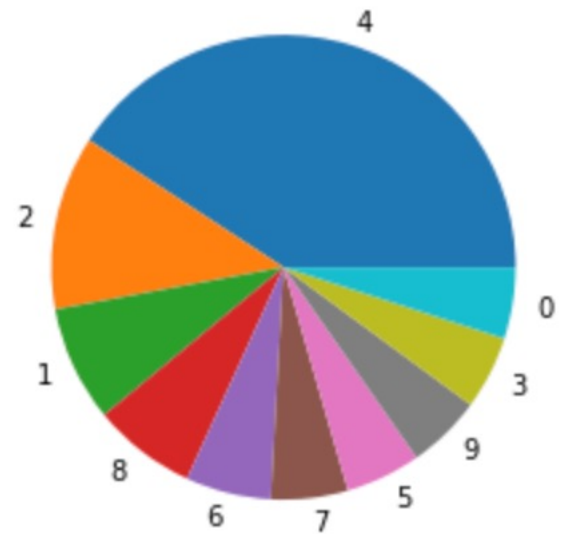


Highest viewed pirated Movies by Industry

Movie name

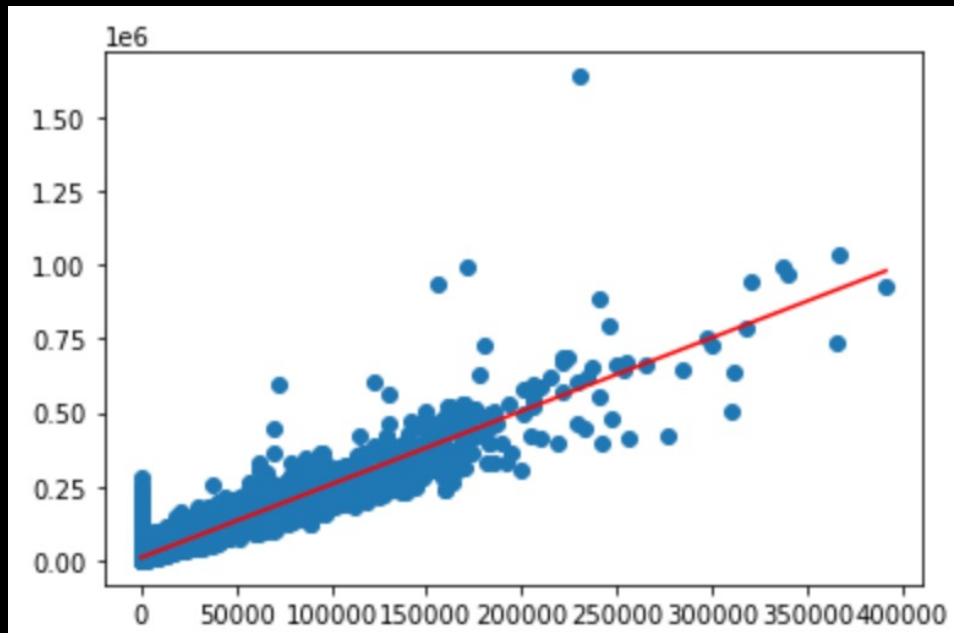


Industry affected by Piracy (Normalized Values)





ML



Mean Absolute Error: 299537.80400399014
Mean Squared Error: 117709157823.732
Root Mean Squared Error: 343087.68241330376

