Predicting a Film's Rating Utilizing Machine Learning

Joel Martinez
Texas State University

Jose Garcia
Texas State University

Kameron Bush
Texas State University

May 6, 2019

Abstract

We present a program that will utilize existing data on films that have been released in order to predict the rating of a film that has yet to be released. This prediction will be given to the film industries that created the film in order to give them a chance to improve the film if needed. The factors being used are IMDB rating, budget, USA gross, USA opening weekend gross, tomatometer, metascore, and worldwide gross. We used using linear regression, k-nearest neighbors, and logistic regression as our model for the project. K-nearest neighbors and logistic regression will utilize k-fold cross validation and will be used as a contender against our linear model. Our methodology to solve this problem included intense research and rigorous testing.

1 Problem Description

When a film industry decides to create a film, the creators have to factor in many attributes that are important to how the public will interpret and perceive the film. Some of these factors are the companys film budget, the film critic ratings, and how much the film makes on its first weekend.
This public interpretation is very important to the film company because how the public perceives the film ultimately determines how many people go and see the film. The amount of people who go see the movie ultimately determines if the film company makes a profit off of the movie, i.e. if it is a hit or a flop.
The fate of the film cannot be determined until after the film is released to the public making it too late to make any changes if it fails. Our program is aimed to fix this problem by analyzing factors of previous films in order to predict the audience rating the new film will receive. This audience rating will give the film company a general idea of how well their film will do before they release the movie. Based on the audience rating presented, the film company will be able to fix their film before release so they can maybe save the film from flopping.
The film industry is a staple in many cultures. It affects almost anyone you can think of. This, along with the fact that we are big fans of films and what it takes to create a work of art worth seeing, is what pushed us to pursue this problem.

2 Survey

To see how much of an impact this project would have with the general population, we asked 50 students their opinion.

Q: Do you believe this project will have an impact on which movies you will watch based on the rating generated?

37 students said it would have an impact.
9 students said it would not impact them.
4 students said they did not know.

3 Plan

March 8th - Understand the Problem in-depth
- Our group has got a now gotten a better understanding of how we can attack this problem. Our attributes will be able to help us use our model to estimate the probability of success (IMBD Rating, Budget, USA Gross, USA Opening Weekend Gross, Tomatometer, Metascore, Worldwide Gross).

March 14th - Develop a data base for the code
- We first created a database where we stored our x attributes and our y audience rating (1.0-10.0) in excel.

March 25th - Complete a base algorithm for Project
- Our model first takes the data from the database (excel), then splits our data to a X and Y dataset. Our model will consist of a train/test split of the X and Y dataset. We used Linear Regression to find a relation between the X attributes and the rating, we then proceeded to use the K-Fold method to optimize our model.

March 29th - Complete Intermediate Project Report
- Done and done.

April 10th - Research ways to make algorithm more accurate
- This step took us some time. We found that doubling the size of our dataset was the most influential change that increased the accuracy of our models.

April 20th - Edit algorithm to be more effective and more precise
- We implemented cross validation in our k-nearest neighbors and logistic regression models in order to increase the precision of our predictor. In the end we saw an increase of about 20% using cross validation but our linear regression model continued to be the more accurate model of the three.

April 29th - Complete Final Project Presentation
- Done.

May 4th - Complete Final Project Report
- This report is proof of completion.

4 Data

The data set contains seven attributes (or features, denoted by X1. . . X8) and one response (or outcomes, denoted by y1). The aim is to use these seven features to predict the outcome.
We specifically choose these seven features because the information is easy to find and open to the public. Some other attributes that we considered for the data was harder to come across on the internet and left many holes in our data. This would have required us to have to clean our data beforehand. With the attributes we are using there is no need to clean the data because we handpicked each movie individually to contain all or most of the attributes.
Data was obtained from the Rotten Tomatoes website which provided us with the Tomatometer (Rotten Tomatoe's score of the movie), and the audience score (the value that we intend to predict.) We also utilized the IMDb website to obtain the gross USA (the amount of money that was made in the USA), the gross opening weekend USA (the amount of money that was made in the first week that the movie was released to the public), worldwide gross (the total amount of money made around the globe), budget (the amount of money the studio spent to create the movie), metascore (the IMDb user's score of the film), and IMDb rating (the score that IMDb critics gave the film.)
All this would come together to predict what an average person watching the film would rate it. Because this score is given by a person and not so calculating machine these scores have an entropic nature. Some may give a low score while others may give a high score. To combat this problem, we have simple stuck to the audience score from Rotten Tomatoes that takes the average of all user's scores.

Specifically:

5 Results

X1 Gross USA

X2 Gross Opening Weekend USA

X3 Worldwide Gross

X4 Budget

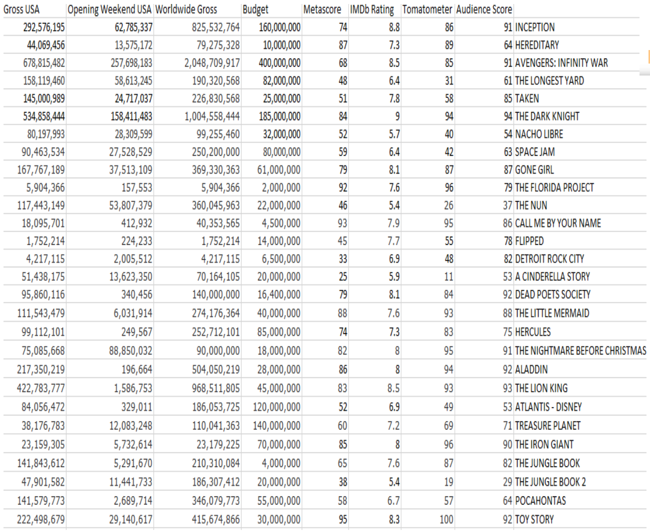
X5 Metascore

X6 IMDb rating

X7 Tomatometer

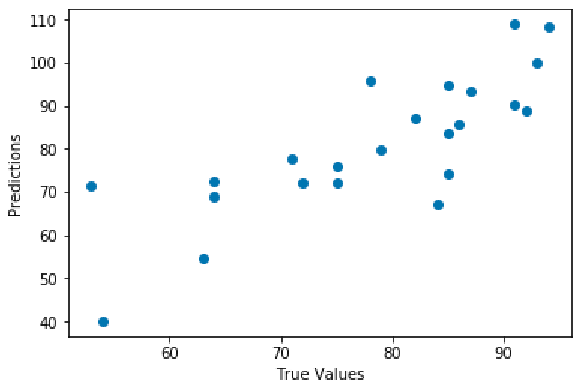
y1 Audience Score

Below is a screenshot of the first few lines in the dataset that we are utilizing for our model. As you can see, most attributes are filled in and few are null. We intend to expand the size of the dataset as we continue to work on the project. Much of the data is gathered by hand and requires a large amount of time to find. Fortunately, the doubling of our dataset was enough to increase our accuracy dramatically so we felt that there was not much benefit in adding much more data to our set.

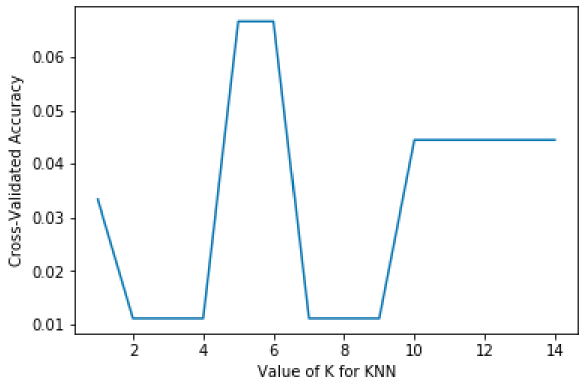


Gross USA	Opening Weekend USA	Worldwide Gross	Budget	Metascore	IMDb Rating	Tomatometer	Audience Score	
292,576,195	62,785,337	825,532,764	160,000,000	74	8.8	86	91	INCEPTION
44,069,456	13,575,172	79,275,328	10,000,000	87	7.3	89	64	HEREDITARY
678,815,482	257,698,183	2,048,709,917	400,000,000	68	8.5	85	91	AVENGERS: INFINITY WAR
158,119,460	58,613,245	190,320,568	82,000,000	48	6.4	31	61	THE LONGEST YARD
145,000,989	24,717,037	226,830,568	25,000,000	51	7.8	58	85	TAKEN
534,858,444	158,411,483	1,004,558,444	185,000,000	84	9	94	94	THE DARK KNIGHT
80,197,993	28,309,599	99,255,460	32,000,000	52	5.7	40	54	NACHO LIBRE
90,463,534	27,528,529	250,200,000	80,000,000	59	6.4	42	63	SPACE JAM
167,767,189	37,513,109	369,330,363	61,000,000	79	8.1	87	87	GONE GIRL
5,904,366	157,553	5,904,366	2,000,000	92	7.6	96	79	THE FLORIDA PROJECT
117,443,149	53,807,379	360,045,963	22,000,000	46	5.4	26	37	THE NUN
18,095,701	412,932	40,353,565	4,500,000	93	7.9	95	86	CALL ME BY YOUR NAME
1,752,214	224,233	1,752,214	14,000,000	45	7.7	55	78	FLIPPED
4,217,115	2,005,512	4,217,115	6,500,000	33	6.9	48	82	DETROIT ROCK CITY
51,438,175	13,623,350	70,164,105	20,000,000	25	5.9	11	53	A CINDERELLA STORY
95,860,116	340,456	140,000,000	16,400,000	79	8.1	84	92	DEAD POETS SOCIETY
111,543,479	6,031,914	274,176,364	40,000,000	88	7.6	93	88	THE LITTLE MERMAID
99,112,101	249,567	252,712,101	85,000,000	74	7.3	83	75	HERCULES
75,085,668	88,850,032	90,000,000	18,000,000	82	8	95	91	THE NIGHTMARE BEFORE CHRISTMAS
217,350,219	196,664	504,050,219	28,000,000	86	8	94	92	ALADDIN
422,783,777	1,586,753	968,511,805	45,000,000	83	8.5	93	93	THE LION KING
84,056,472	329,011	186,053,725	120,000,000	52	6.9	49	53	ATLANTIS - DISNEY
38,176,783	12,083,248	110,041,363	140,000,000	60	7.2	69	71	TREASURE PLANET
23,159,305	5,732,614	23,179,225	70,000,000	85	8	96	90	THE IRON GIANT
141,843,612	5,291,670	210,310,084	4,000,000	65	7.6	87	82	THE JUNGLE BOOK
47,901,582	11,441,733	186,307,412	20,000,000	38	5.4	19	29	THE JUNGLE BOOK 2
141,579,773	2,689,714	346,079,773	55,000,000	58	6.7	57	64	POCAHONTAS
222,498,679	29,140,617	415,674,866	30,000,000	95	8.3	100	92	TOY STORY

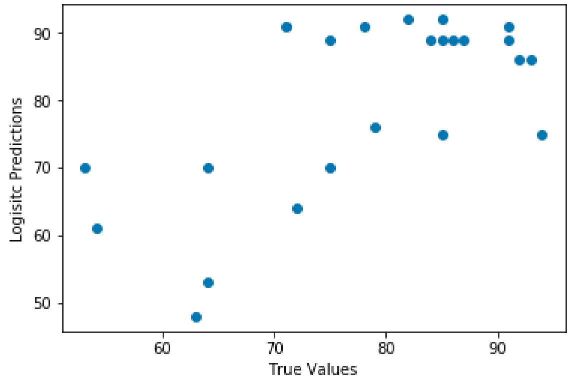
```
Predicting Infinity War: score = ~91, Predicted = 89
```

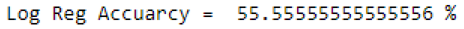


```
Linear Regression Mean Absolute Error =  [7.67876371]
```



```
KNN Accuarcy =  66.66666666666667 %
```





```
Log Reg Accuarcy =  55.55555555555556 %
```

Considering the fact that predicting a score exactly correct is very difficult we decided to take a different approach as to how to calculate the accuracy of our linear regression model. We decided on allowing a small window of +3/-3 to the model's prediction. For example, if our prediction was either 71, 72, 73, 74, 75, 76, or 77 for a film with a score of 74 it would be considered as a accurately predicted audience score. When you take this into consideration, our accuracy increased dramatically.
This is an acceptable method of reading our data because film ratings are all based on biased opinions and many other factors that cannot be considered. The window covers the high variance that comes with predicting this type of data.

6 Code

```
df=pd.DataFrame(df,columns=[’Gross USA’,’Opening Weekend USA’,’Worldwide Gross’,’Budget’,’Metascore’,’IMDb Rating’,’Tomatometer’,’Audience Score’])
```

```
X=df[[’Gross USA’,’Opening Weekend USA’,’Worldwide Gross’,’Budget’,’Metascore’,’IMDb Rating’,’Tomatometer’]]
y=df[’Audience Score’]
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.5)
```

```
regr=linear_model.LinearRegression()
model=regr.fit(X_train, y_train)
predictions=regr.predict(X_test)
```

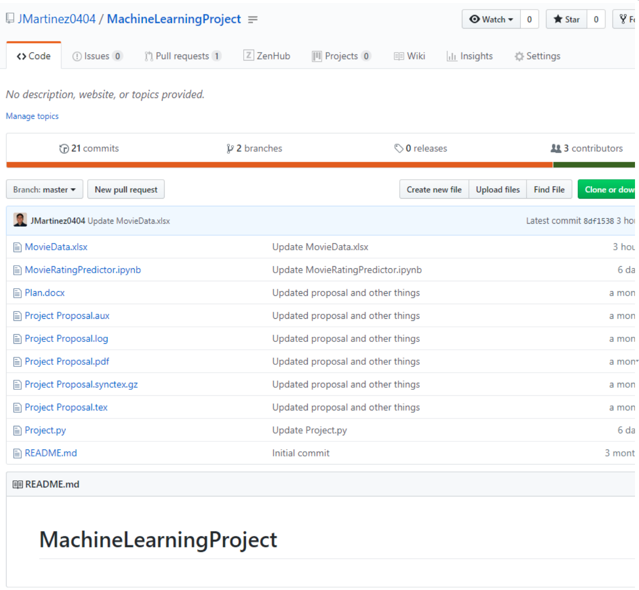
```
print(”Linear Regression Mean Absolute Error = ”, mean_absolute_error(y_test, predictions, multioutput=’raw_values’))
```

```
k_scores=[]
for k in k_range:
knn=KNeighborsClassifier(n_neighbors = k)
scores=cross_val_score(knn, X, y, cv=3, scoring=’accuracy’)
k_scores.append(scores.mean())
```

```
knn=KNeighborsClassifier(n_neighbors=6)
print(”KNN Accuarcy = ”, (cross_val_score(knn, X, y, cv=3, scoring=’accuracy’).mean()) * 1000, ”
```

```
logreg=LogisticRegression()
logmodel=logreg.fit(X_train, y_train)
logpreds=logreg.predict(X_test)
```

7 GitHub



JMartinez0404 / MachineLearningProject

No description, website, or topics provided.

Manage topics

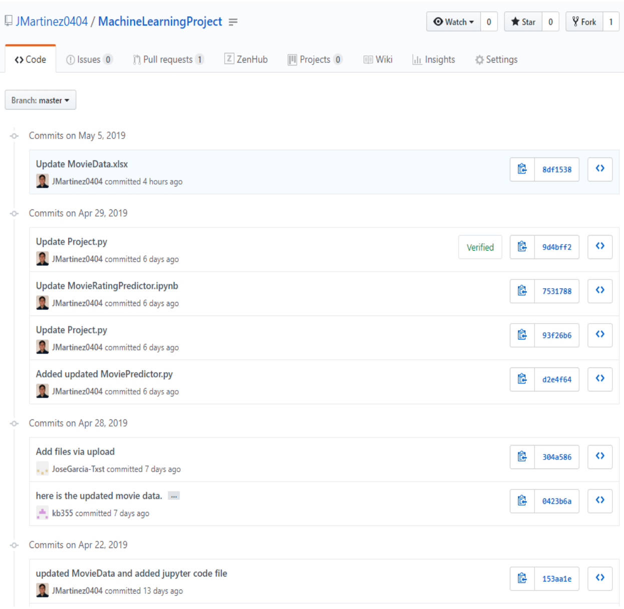
21 commits | 2 branches | 0 releases | 3 contributors

JMartinez0404 Update MovieData.xlsx — Latest commit 8df1538 3 ho

MovieData.xlsx	Update MovieData.xlsx	3 hou
MovieRatingPredictor.ipynb	Update MovieRatingPredictor.ipynb	6 da
Plan.docx	Updated proposal and other things	a mon
Project Proposal.aux	Updated proposal and other things	a mon
Project Proposal.log	Updated proposal and other things	a mon
Project Proposal.pdf	Updated proposal and other things	a month ago
Project Proposal.synctex.gz	Updated proposal and other things	a mon
Project Proposal.tex	Updated proposal and other things	a mon
Project.py	Update Project.py	6 da
README.md	Initial commit	3 mont

README.md

MachineLearningProject



JMartinez0404 / MachineLearningProject

Commits on May 5, 2019

Update MovieData.xlsx — JMartinez0404 committed 4 hours ago — 8df1538

Commits on Apr 29, 2019

Update Project.py — JMartinez0404 committed 6 days ago — Verified — 9d4bff2

Update MovieRatingPredictor.ipynb — JMartinez0404 committed 6 days ago — 7531788

Update Project.py — JMartinez0404 committed 6 days ago — 93f26b6

Added updated MoviePredictor.py — JMartinez0404 committed 6 days ago — d2e4f64

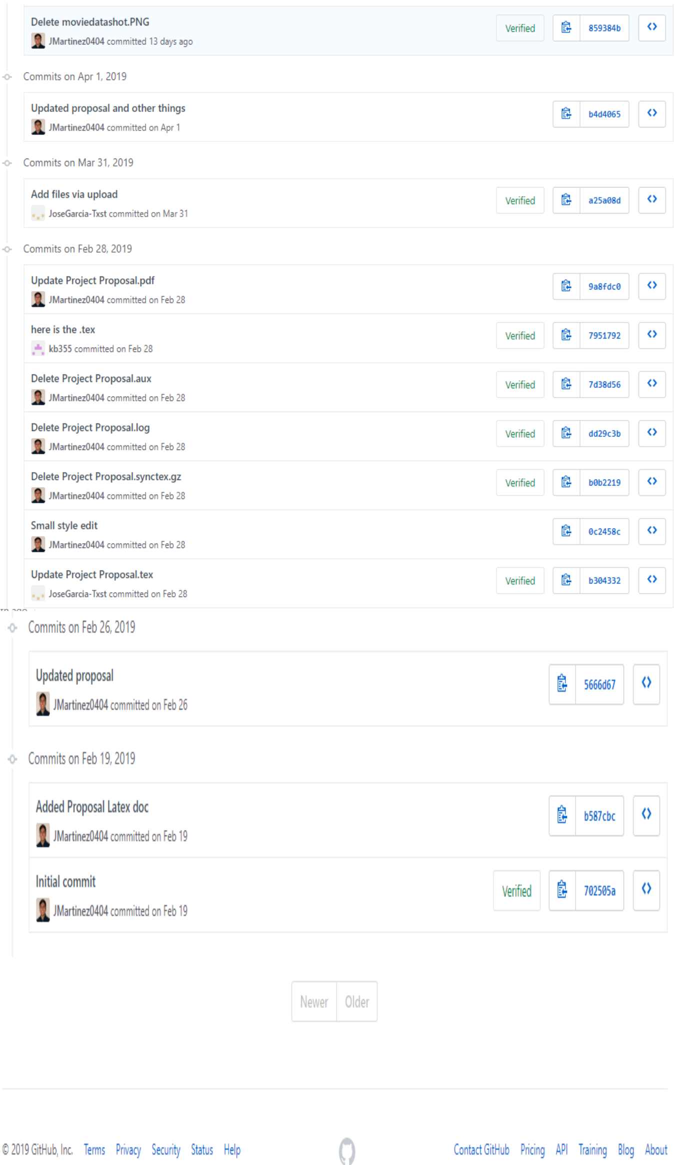
Commits on Apr 28, 2019

Add files via upload — JoseGarcia-Txst committed 7 days ago — 304a586

here is the updated movie data. … — kb355 committed 7 days ago — 0423b6a

Commits on Apr 22, 2019

updated MovieData and added jupyter code file — JMartinez0404 committed 13 days ago — 153aa1e



Delete moviedatashot.PNG — JMartinez0404 committed 13 days ago — Verified — 859384b

Commits on Apr 1, 2019

Updated proposal and other things — JMartinez0404 committed on Apr 1 — b4d4065

Commits on Mar 31, 2019

Add files via upload — JoseGarcia-Txst committed on Mar 31 — Verified — a25a08d

Commits on Feb 28, 2019

Update Project Proposal.pdf — JMartinez0404 committed on Feb 28 — 9a8fdc0

here is the .tex — kb355 committed on Feb 28 — Verified — 7951792

Delete Project Proposal.aux — JMartinez0404 committed on Feb 28 — Verified — 7d38d56

Delete Project Proposal.log — JMartinez0404 committed on Feb 28 — Verified — dd29c3b

Delete Project Proposal.synctex.gz — JMartinez0404 committed on Feb 28 — Verified — b0b2219

Small style edit — JMartinez0404 committed on Feb 28 — 0c2458c

Update Project Proposal.tex — JoseGarcia-Txst committed on Feb 28 — Verified — b304332

Commits on Feb 26, 2019

Updated proposal — JMartinez0404 committed on Feb 26 — 5666d67

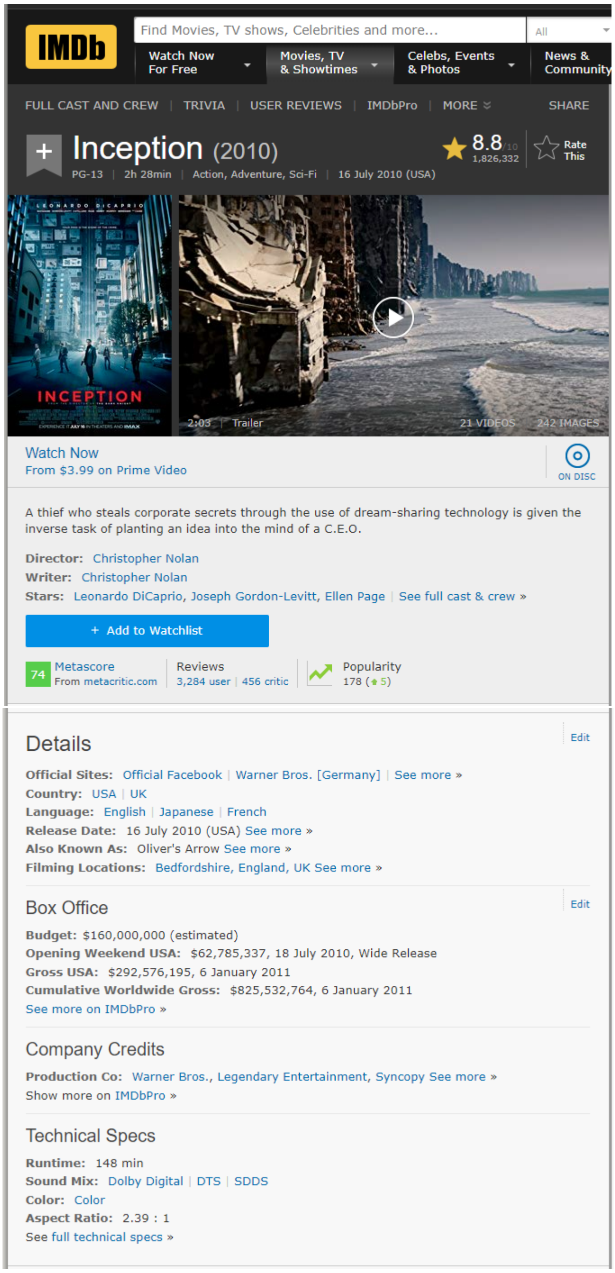
Commits on Feb 19, 2019

Added Proposal Latex doc — JMartinez0404 committed on Feb 19 — b587cbc

Initial commit — JMartinez0404 committed on Feb 19 — Verified — 702505a

Above are screenshots of the code and proposal being managed in our GitHub repo.

8 Films



Inception (2010)

8.8/10 1,826,332 Rate This

PG-13 | 2h 28min | Action, Adventure, Sci-Fi | 16 July 2010 (USA)

Watch Now
From $3.99 on Prime Video
ON DISC

A thief who steals corporate secrets through the use of dream-sharing technology is given the inverse task of planting an idea into the mind of a C.E.O.

Director: Christopher Nolan
Writer: Christopher Nolan
Stars: Leonardo DiCaprio, Joseph Gordon-Levitt, Ellen Page | See full cast & crew »

+ Add to Watchlist

74 Metascore From metacritic.com | Reviews 3,284 user | 456 critic | Popularity 178 (↑5)

Details — Edit

Official Sites: Official Facebook | Warner Bros. [Germany] | See more »
Country: USA | UK
Language: English | Japanese | French
Release Date: 16 July 2010 (USA) See more »
Also Known As: Oliver's Arrow See more »
Filming Locations: Bedfordshire, England, UK See more »

Box Office — Edit

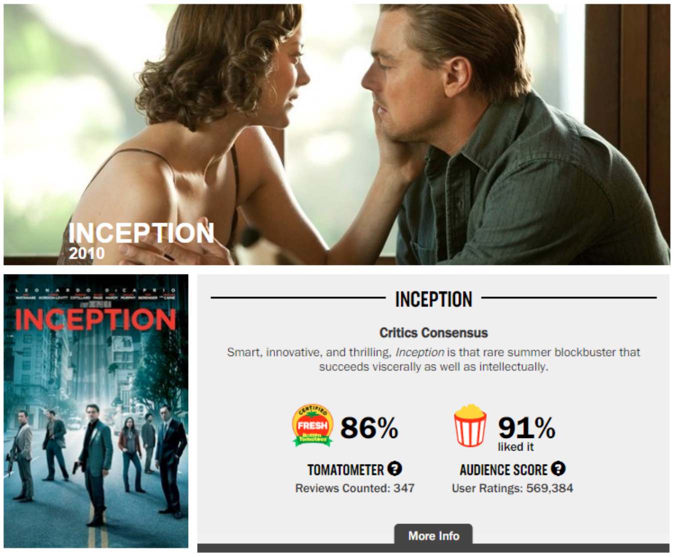
Budget: $160,000,000 (estimated)
Opening Weekend USA: $62,785,337, 18 July 2010, Wide Release
Gross USA: $292,576,195, 6 January 2011
Cumulative Worldwide Gross: $825,532,764, 6 January 2011
See more on IMDbPro »

Company Credits

Production Co: Warner Bros., Legendary Entertainment, Syncopy See more »
Show more on IMDbPro »

Technical Specs

Runtime: 148 min
Sound Mix: Dolby Digital | DTS | SDDS
Color: Color
Aspect Ratio: 2.39 : 1
See full technical specs »



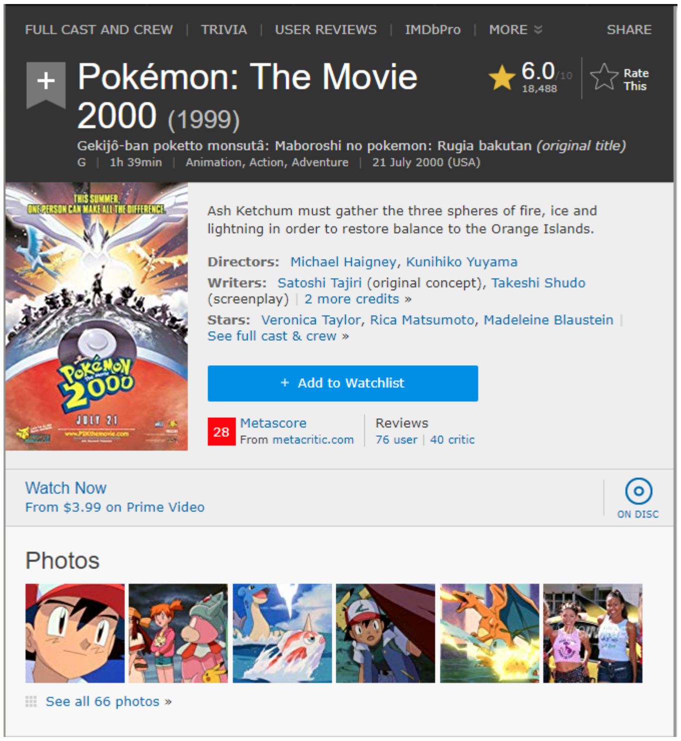
INCEPTION

Critics Consensus

Smart, innovative, and thrilling, *Inception* is that rare summer blockbuster that succeeds viscerally as well as intellectually.

86% TOMATOMETER — Reviews Counted: 347

91% liked it AUDIENCE SCORE — User Ratings: 569,384



FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE | SHARE
THIS SUMMER ONE PERSON CAN MAKE ALL THE DIFFERENCE
Pokémon 2000
JULY 21

Pokémon: The Movie 2000 (1999)

6.0/10 18,488 Rate This

Gekijô-ban poketto monsutâ: Maboroshi no pokemon: Rugia bakutan *(original title)*
G | 1h 39min | Animation, Action, Adventure | 21 July 2000 (USA)

Ash Ketchum must gather the three spheres of fire, ice and lightning in order to restore balance to the Orange Islands.

Directors: Michael Haigney, Kunihiko Yuyama
Writers: Satoshi Tajiri (original concept), Takeshi Shudo (screenplay) | 2 more credits »
Stars: Veronica Taylor, Rica Matsumoto, Madeleine Blaustein | See full cast & crew »

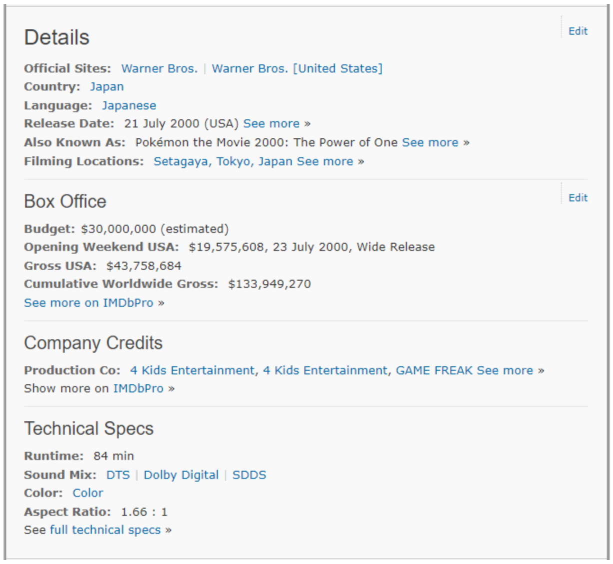
+ Add to Watchlist

28 Metascore From metacritic.com | Reviews 76 user | 40 critic

Watch Now
From $3.99 on Prime Video
ON DISC

Photos

See all 66 photos »



Details

Edit

Official Sites: Warner Bros. | Warner Bros. [United States]
Country: Japan
Language: Japanese
Release Date: 21 July 2000 (USA) See more »
Also Known As: Pokémon the Movie 2000: The Power of One See more »
Filming Locations: Setagaya, Tokyo, Japan See more »

Box Office

Edit

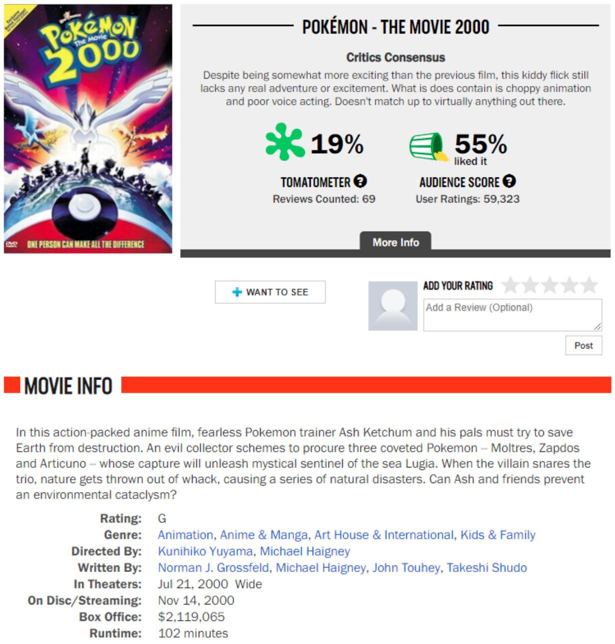
Budget: $30,000,000 (estimated)
Opening Weekend USA: $19,575,608, 23 July 2000, Wide Release
Gross USA: $43,758,684
Cumulative Worldwide Gross: $133,949,270
See more on IMDbPro »

Company Credits

Production Co: 4 Kids Entertainment, 4 Kids Entertainment, GAME FREAK See more »
Show more on IMDbPro »

Technical Specs

Runtime: 84 min
Sound Mix: DTS | Dolby Digital | SDDS
Color: Color
Aspect Ratio: 1.66 : 1
See full technical specs »



POKÉMON - THE MOVIE 2000

Critics Consensus

Despite being somewhat more exciting than the previous film, this kiddy flick still lacks any real adventure or excitement. What is does contain is choppy animation and poor voice acting. Doesn't match up to virtually anything out there.

19% TOMATOMETER
Reviews Counted: 69

55% liked it AUDIENCE SCORE
User Ratings: 59,323

More Info

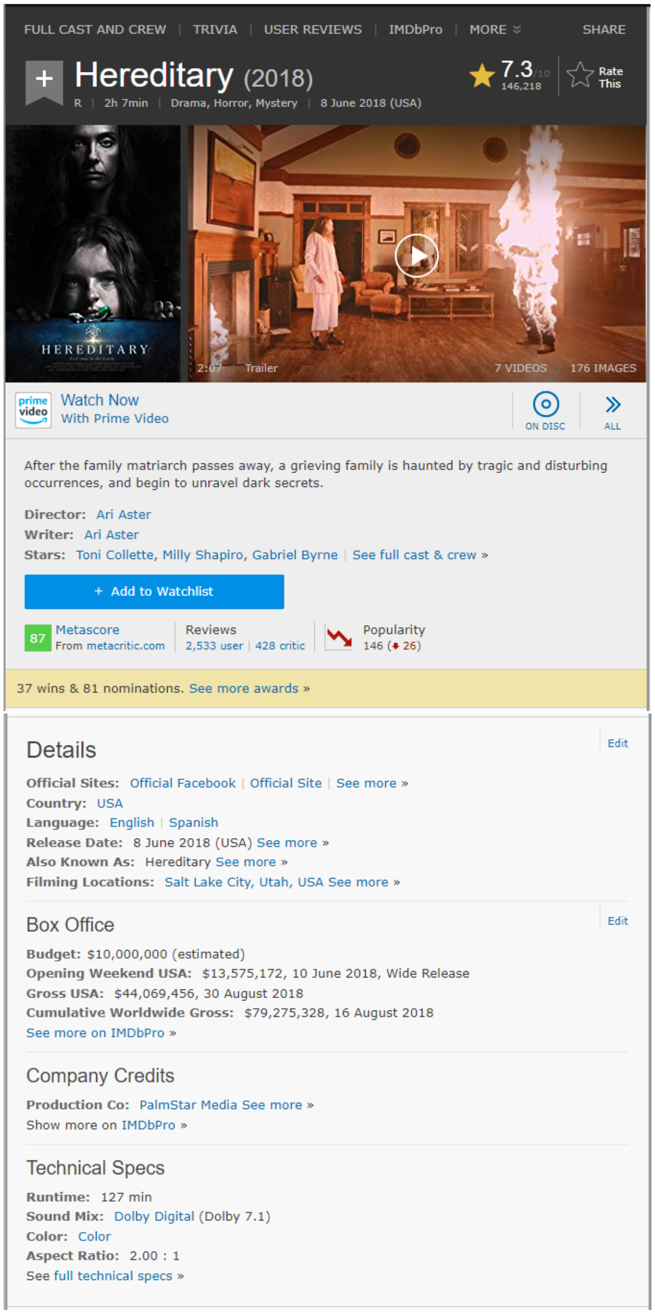
+ WANT TO SEE

ADD YOUR RATING ☆☆☆☆☆
Add a Review (Optional)
Post

MOVIE INFO

In this action-packed anime film, fearless Pokemon trainer Ash Ketchum and his pals must try to save Earth from destruction. An evil collector schemes to procure three coveted Pokemon -- Moltres, Zapdos and Articuno -- whose capture will unleash mystical sentinel of the sea Lugia. When the villain snares the trio, nature gets thrown out of whack, causing a series of natural disasters. Can Ash and friends prevent an environmental cataclysm?

Rating:	G
Genre:	Animation, Anime & Manga, Art House & International, Kids & Family
Directed By:	Kunihiko Yuyama, Michael Haigney
Written By:	Norman J. Grossfeld, Michael Haigney, John Touhey, Takeshi Shudo
In Theaters:	Jul 21, 2000 Wide
On Disc/Streaming:	Nov 14, 2000
Box Office:	$2,119,065
Runtime:	102 minutes


FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE | SHARE
HEREDITARY
2:07 Trailer
7 VIDEOS
176 IMAGES
ON DISC
ALL

Hereditary (2018)

7.3/10 146,218 | Rate This

R | 2h 7min | Drama, Horror, Mystery | 8 June 2018 (USA)

Watch Now
With Prime Video

After the family matriarch passes away, a grieving family is haunted by tragic and disturbing occurrences, and begin to unravel dark secrets.

Director: Ari Aster
Writer: Ari Aster
Stars: Toni Collette, Milly Shapiro, Gabriel Byrne | See full cast & crew »

+ Add to Watchlist

87 Metascore From metacritic.com | Reviews 2,533 user | 428 critic | Popularity 146 (↓ 26)

37 wins & 81 nominations. See more awards »

Details

Edit

Official Sites: Official Facebook | Official Site | See more »
Country: USA
Language: English | Spanish
Release Date: 8 June 2018 (USA) See more »
Also Known As: Hereditary See more »
Filming Locations: Salt Lake City, Utah, USA See more »

Box Office

Edit

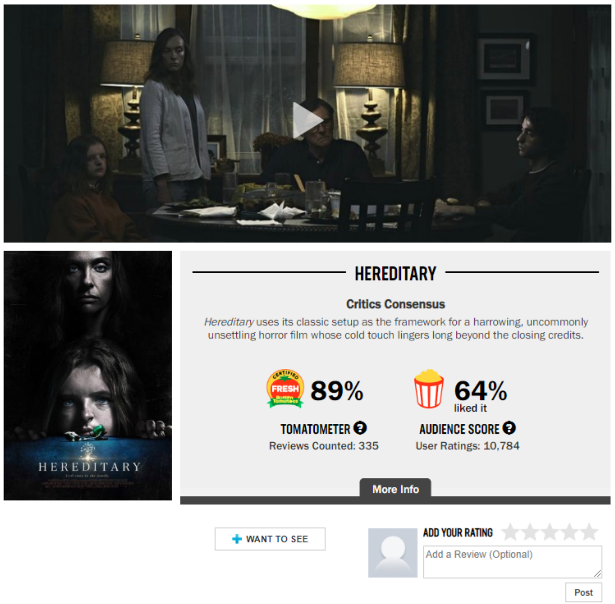
Budget: $10,000,000 (estimated)
Opening Weekend USA: $13,575,172, 10 June 2018, Wide Release
Gross USA: $44,069,456, 30 August 2018
Cumulative Worldwide Gross: $79,275,328, 16 August 2018
See more on IMDbPro »

Company Credits

Production Co: PalmStar Media See more »
Show more on IMDbPro »

Technical Specs

Runtime: 127 min
Sound Mix: Dolby Digital (Dolby 7.1)
Color: Color
Aspect Ratio: 2.00 : 1
See full technical specs »


HEREDITARY

Critics Consensus

Hereditary uses its classic setup as the framework for a harrowing, uncommonly unsettling horror film whose cold touch lingers long beyond the closing credits.

89% TOMATOMETER Reviews Counted: 335

64% liked it AUDIENCE SCORE User Ratings: 10,784

9 Future Work

Obviously, there is much room for improving the accuracy of our models. Different algorithms or combinations of algorithms could increase our accuracy. Maybe, at some point we would be able to rid our evaluation of the +3/-3 score window altogether. More data and a better understanding of bias trends could help us attack this problem more efficiently.

There are other methods of handling our data that we could attempt. For example, there could be a way to combine regression and classification models in order to create different labels for low, mid-low, mid, mid-high, and high scoring films.

Our main focus for continuing this project would be able to increase the accuracy of our model and make the model more user friendly. One could say that a user interface or app could be marketable to a target audience.

10 Conclusion

One important observation that we came across when creating our project is the importance of a quality dataset. They're are a couple of movie database API's on the internet that we could have utilized to create our dataset but in order to ensure a clean set of data creating our own seem the logical step.

Something else we learned from this experience was the importance of understanding the multitude of algorithms available and which cases call for which algorithms. Machine learning is a broad subject as there is much room for improvement.

When it comes to the contributions of the team members they are as follows:

Joel Martinez: worked on program code; worked on report; worked on dataset

Kameron Bush: worked on report; worked on dataset; worked on presentation

Jose Garcia: worked on program code; worked on report; worked on presentation

References

[1] Quader, Nahid & Gani, Md. & Chaki, Dipankar & Ali, Md. *A Machine Learning Approach to Predict Movie Box-Office Success.* Reading, Bangladesh, 2018.

[2] Sharda, Ramesh & Delen, Dursun *Predicting box-office success of motion pictures with neural networks.* Expert Systems with Applications. 30. 243-254. 10.1016/j.eswa.2005.07.018.

[3] B. R. Litman & H. Ahn *Predicting financial success of motion pictures.* In B. R. Litman (Ed.), the motion picture mega-industry. Boston, MA: Allyn & Bacon Publishing, Inc. (1998)

[4] M.H Latif & H. Afzal *Prediction of Movies popularity Using Machine Learning Techniques.* National University of Sciences and technology, H-12, ISB, Pakistan.