

A black and white photograph of the Strand Theatre in London. The building's facade is covered in movie posters and signs. A large sign at the top reads "STRAND". Below it, a marquee lists "JUNE LOCKHART & SHE-WOLF OF LONDON" and "ALSO THE CAT CREEPS". Another sign on the left corner says "SHE-WOLF OF LONDON" and "CAT CREEPS". A vintage car is parked on the street in front of the theatre, and a bicycle is leaning against the entrance on the left.

MACHINE LEARNING PREDICTING MOVIE RATINGS

Reed Koval, Stephanie Morton, Angie Boakye Danquah, Ellen Wray

Our Team



STEPHANIE

3rd Year at UVA
CS & Economics



REED

3rd Year at UVA
Statistics & Economics



ANGIE

4th Year at UVA
Statistics & Economics



ELLEN

4th Year at UVA
Statistics & Economics

GOAL

The **goal** of our project is to use machine learning to generate predictions for the average rating left by reviewers after movies have been released in box office. We used a **data set** from Kaggle* that has information on the genre, title, year produced, production company, budget, revenue, runtime, language, key words, and popularity of movies dating back since the 1920s. The numerical variables of these will be **inputs** to train our computer. Each movie also has an average review calculated by the reviews left by movie goers. This numerical variable will be the **output** and the variable we are attempting to predict using **machine learning**.

*https://www.kaggle.com/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv

Data Set

budget	genres	homepage_id	keywords	original_title	overview	popularity	production_code	production_release_date	revenue	runtime	spoken_language	status	tagline	title	vote_average	vote_count
2.37E+08	["id": 28, "name": "Action"]	http://www.fox.com	19995	["id": 146, "name": "Avatar"]	In the 22nd century, a marine is sent to a distant planet to help the natives fight against a vicious alien species.	150.4376	["name": "Avatar", "iso_3166_2": "US"]	12/10/2009	2.79E+09	162	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Enter the Avatar	Avatar	7.2	11800
3E+08	["id": 12, "name": "Adventure"]	http://disney.com	285	["id": 270, "name": "Pirates of the Caribbean"]	Captain Jack Sparrow leads a band of pirates to fight against a powerful sea monster.	139.0826	["name": "Pirates of the Caribbean", "iso_3166_2": "US"]	5/19/2007	9.61E+08	169	["iso_639_1": "en", "iso_639_2": "Released"]	Released	At the end of the world	Pirates of the Caribbean	6.9	4500
2.45E+08	["id": 28, "name": "Action"]	http://www.fox.com	206647	["id": 470, "name": "Spectre"]	A cryptic new mission for James Bond as he battles a powerful enemy.	107.3768	["name": "Spectre", "iso_3166_2": "US"]	10/26/2015	8.81E+08	148	["iso_639_1": "en", "iso_639_2": "Released"]	Released	A Plan No Spectre	Spectre	6.3	4466
2.5E+08	["id": 28, "name": "Action"]	http://www.fox.com	49026	["id": 849, "name": "The Dark Knight Rises"]	Following the events of The Dark Knight, Batman returns to fight against a powerful enemy.	112.313	["name": "The Dark Knight Rises", "iso_3166_2": "US"]	7/16/2012	1.08E+09	165	["iso_639_1": "en", "iso_639_2": "Released"]	Released	The Legend The Dark Knight	The Dark Knight Rises	7.6	9106
2.6E+08	["id": 28, "name": "Action"]	http://movie.com	49529	["id": 818, "name": "John Carter"]	John Carter is transported to a distant planet where he fights against a powerful enemy.	43.927	["name": "John Carter", "iso_3166_2": "US"]	3/7/2012	2.84E+08	132	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Lost in our John Carter	John Carter	6.1	2124
2.58E+08	["id": 14, "name": "Fantasy"]	http://www.fox.com	559	["id": 851, "name": "Spider-Man"]	The seemingly ordinary Peter Parker becomes the superhero Spider-Man.	115.6998	["name": "Spider-Man", "iso_3166_2": "US"]	5/1/2007	8.91E+08	139	["iso_639_1": "en", "iso_639_2": "Released"]	Released	The battle Spider-Man	Spider-Man	5.9	3576
2.6E+08	["id": 16, "name": "Drama"]	http://disney.com	38757	["id": 156, "name": "Tangled"]	When a princess is kidnapped by an evil witch, a young man goes on a journey to rescue her.	48.68197	["name": "Tangled", "iso_3166_2": "US"]	11/24/2010	5.92E+08	100	["iso_639_1": "en", "iso_639_2": "Released"]	Released	They're ta Tangled	Tangled	7.4	3330
2.8E+08	["id": 28, "name": "Action"]	http://marvel.com	99861	["id": 882, "name": "Avengers: When Time Collides"]	When Tony Stark and his allies team up with Iron Man, they fight against a powerful enemy.	134.2792	["name": "Avengers: When Time Collides", "iso_3166_2": "US"]	4/22/2015	1.41E+09	141	["iso_639_1": "en", "iso_639_2": "Released"]	Released	A New Age Avengers:	Avengers: When Time Collides	7.3	6767
2.5E+08	["id": 12, "name": "Adventure"]	http://harrypotter.com	767	["id": 616, "name": "Harry Potter and the Half-Blood Prince"]	Harry Potter and his friends face a powerful enemy in the seventh year of his magical education.	98.88564	["name": "Harry Potter and the Half-Blood Prince", "iso_3166_2": "US"]	7/7/2009	9.34E+08	153	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Dark Secrets Harry Potter	Harry Potter and the Half-Blood Prince	7.4	5293
2.5E+08	["id": 28, "name": "Action"]	http://www.fox.com	209112	["id": 849, "name": "Batman v Superman: Dawn of Justice"]	Batman and Superman team up to fight against a powerful enemy.	155.7905	["name": "Batman v Superman: Dawn of Justice", "iso_3166_2": "US"]	3/23/2016	8.73E+08	151	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Justice or Batman v Superman	Batman v Superman: Dawn of Justice	5.7	7004
2.7E+08	["id": 12, "name": "Adventure"]	http://www.fox.com	1452	["id": 83, "name": "Superman"]	Superman returns to Earth to fight against a powerful enemy.	57.92562	["name": "Superman", "iso_3166_2": "US"]	6/28/2006	3.91E+08	154	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Superman	Superman	5.4	1400
2E+08	["id": 12, "name": "Adventure"]	http://www.fox.com	10764	["id": 627, "name": "Quantum Leap"]	Quantum Leap is a science fiction series about a man who travels through time.	107.9288	["name": "Quantum Leap", "iso_3166_2": "US"]	10/30/2008	5.86E+08	106	["iso_639_1": "en", "iso_639_2": "Released"]	Released	For love, fr Quantum Leap	Quantum Leap	6.1	2965
2E+08	["id": 12, "name": "Adventure"]	http://disney.com	58	["id": 616, "name": "Pirates of the Caribbean"]	Captain Jack Sparrow leads a band of pirates to fight against a powerful sea monster.	145.8474	["name": "Pirates of the Caribbean", "iso_3166_2": "US"]	6/20/2006	1.07E+09	151	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Jack is bac Pirates of the Caribbean	Pirates of the Caribbean	7	5246
2.55E+08	["id": 28, "name": "Action"]	http://disney.com	57201	["id": 155, "name": "The Lone Ranger"]	The Lone Ranger is a Western film about a man who fights against a powerful enemy.	49.04696	["name": "The Lone Ranger", "iso_3166_2": "US"]	7/3/2013	89289910	149	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Never Tak The Lone Ranger	The Lone Ranger	5.9	2311
2.25E+08	["id": 28, "name": "Action"]	http://www.fox.com	49521	["id": 83, "name": "Man of Steel"]	A young boy discovers he is the son of a powerful alien and must learn to control his powers.	99.39801	["name": "Man of Steel", "iso_3166_2": "US"]	6/12/2013	6.63E+08	143	["iso_639_1": "en", "iso_639_2": "Released"]	Released	You will be Man of Steel	Man of Steel	6.5	6359
2.25E+08	["id": 12, "name": "Drama"]	"name": "The Chronicle"	2454	["id": 818, "name": "The Chronicle"]	One year after the events of The Dark Knight, Batman returns to fight against a powerful enemy.	53.9786	["name": "The Chronicle", "iso_3166_2": "US"]	5/15/2008	4.2E+08	150	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Hope has The Chronicle	The Chronicle	6.3	1630
2.2E+08	["id": 878, "name": "The Avengers"]	http://marvel.com	24428	["id": 242, "name": "The Avengers"]	When an alien invasion threatens Earth, a team of superheroes must unite to save the world.	144.4486	["name": "The Avengers", "iso_3166_2": "US"]	4/25/2012	1.52E+09	143	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Some asse The Avengers	The Avengers	7.4	11776
3.8E+08	["id": 12, "name": "Adventure"]	http://disney.com	1865	["id": 658, "name": "Pirates of the Caribbean"]	Captain Jack Sparrow leads a band of pirates to fight against a powerful sea monster.	135.4139	["name": "Pirates of the Caribbean", "iso_3166_2": "US"]	5/14/2011	1.05E+09	136	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Live Forev Pirates of the Caribbean	Pirates of the Caribbean	6.4	4948
2.25E+08	["id": 28, "name": "Action"]	http://www.fox.com	41154	["id": 437, "name": "Men in Black: International"]	Agents K and J team up with a new partner to fight against a powerful enemy.	52.03518	["name": "Men in Black: International", "iso_3166_2": "US"]	5/23/2012	6.24E+08	106	["iso_639_1": "en", "iso_639_2": "Released"]	Released	They are t Men in Black	Men in Black: International	6.2	4160
2.5E+08	["id": 28, "name": "Action"]	http://www.fox.com	122917	["id": 417, "name": "The Hobbit"]	Immediately after the events of The Lord of the Rings: The Fellowship of the Ring, Bilbo Baggins is forced to go on a journey.	120.9657	["name": "The Hobbit", "iso_3166_2": "US"]	12/10/2014	9.56E+08	144	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Witness th The Hobbit	The Hobbit	7.1	4760
2.15E+08	["id": 28, "name": "Action"]	http://www.fox.com	1930	["id": 187, "name": "The Amazon"]	Peter Parker is bitten by a radioactive spider and gains the ability to become the superhero Spider-Man.	89.86628	["name": "The Amazon", "iso_3166_2": "US"]	6/27/2012	7.52E+08	136	["iso_639_1": "en", "iso_639_2": "Released"]	Released	The untok The Amazon	The Amazon	6.5	6586
2E+08	["id": 28, "name": "Action"]	http://www.fox.com	20662	["id": 414, "name": "Robin Hood"]	Robin Hood is a film about a man who fights against a powerful enemy.	37.6683	["name": "Robin Hood", "iso_3166_2": "US"]	5/12/2010	3.11E+08	140	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Rise and r Robin Hood	Robin Hood	6.2	1398
2.5E+08	["id": 12, "name": "Adventure"]	http://www.fox.com	57158	["id": 603, "name": "The Hobbit"]	The Dwarves are forced to go on a journey to fight against a powerful enemy.	94.37056	["name": "The Hobbit", "iso_3166_2": "US"]	12/11/2013	9.58E+08	161	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Beyond d: The Hobbit	The Hobbit	7.6	4524
1.8E+08	["id": 12, "name": "Adventure"]	http://www.fox.com	2268	["id": 392, "name": "The Goldie"]	After over a year of being a prisoner, Goldie is released and must learn to live with his new life.	42.99091	["name": "The Goldie", "iso_3166_2": "US"]	12/4/2007	3.72E+08	113	["iso_639_1": "en", "iso_639_2": "Released"]	Released	There are The Goldie	The Goldie	5.8	1303
2.07E+08	["id": 12, "name": "Drama"]	"name": "The King Kong"	254	["id": 774, "name": "King Kong In 1933"]	Needing a way to get the giant ape out of the city, the military decides to use a new weapon.	61.22601	["name": "King Kong In 1933", "iso_3166_2": "US"]	12/14/2005	5.5E+08	187	["iso_639_1": "en", "iso_639_2": "Released"]	Released	The eight: King Kong	King Kong	6.6	2337
2E+08	["id": 18, "name": "The Titanic"]	http://www.fox.com	597	["id": 258, "name": "Titanic"]	84 years later, the ship is discovered and the world is amazed by the discovery.	100.0259	["name": "Titanic", "iso_3166_2": "US"]	11/18/1997	1.85E+09	194	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Nothing o Titanic	Titanic	7.5	7562
2.5E+08	["id": 12, "name": "Adventure"]	http://marvel.com	271110	["id": 393, "name": "Captain America"]	Following the events of The Avengers, Captain America is sent to a distant planet to fight against a powerful enemy.	198.3724	["name": "Captain America", "iso_3166_2": "US"]	4/27/2016	1.15E+09	147	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Divided W Captain America	Captain America	7.1	7241
2.09E+08	["id": 53, "name": "The Battleship"]	"name": "The Battleship"	44833	["id": 172, "name": "Battleship"]	When a massive alien ship is discovered, the Navy decides to use it as a weapon.	64.92838	["name": "Battleship", "iso_3166_2": "US"]	4/11/2012	3.03E+08	131	["iso_639_1": "en", "iso_639_2": "Released"]	Released	The Battle Battleship	Battleship	5.5	2114
1.5E+08	["id": 28, "name": "Action"]	http://www.fox.com	135397	["id": 129, "name": "Jurassic World"]	Twenty-two years after the events of Jurassic Park, a new dinosaur park is built.	418.7086	["name": "Jurassic World", "iso_3166_2": "US"]	6/9/2015	1.51E+09	124	["iso_639_1": "en", "iso_639_2": "Released"]	Released	The park i Jurassic World	Jurassic World	6.5	8662
2E+08	["id": 28, "name": "Action"]	http://www.fox.com	37724	["id": 470, "name": "Skyfall"]	When Bond is forced to go on a journey, he must learn to live with his new life.	93.00499	["name": "Skyfall", "iso_3166_2": "US"]	10/25/2012	1.11E+09	143	["iso_639_1": "en", "iso_639_2": "Released"]	Released	Think on y Skyfall	Skyfall	6.9	7604
2E+08	["id": 28, "name": "Action"]	http://www.fox.com	558	["id": 851, "name": "Spider-Man"]	Peter Parker is bitten by a radioactive spider and gains the ability to become the superhero Spider-Man.	35.14959	["name": "Spider-Man", "iso_3166_2": "US"]	6/25/2004	7.84E+08	127	["iso_639_1": "en", "iso_639_2": "Released"]	Released	There's a Spider-Man	Spider-Man	6.7	4321

Our Variables

We chose to use the numerical variables from the data set for our machine learning and two categorical variables

BUDGET

POPULARITY

REVENUE

RUNTIME

VOTE
COUNT

YEAR
RELEASED

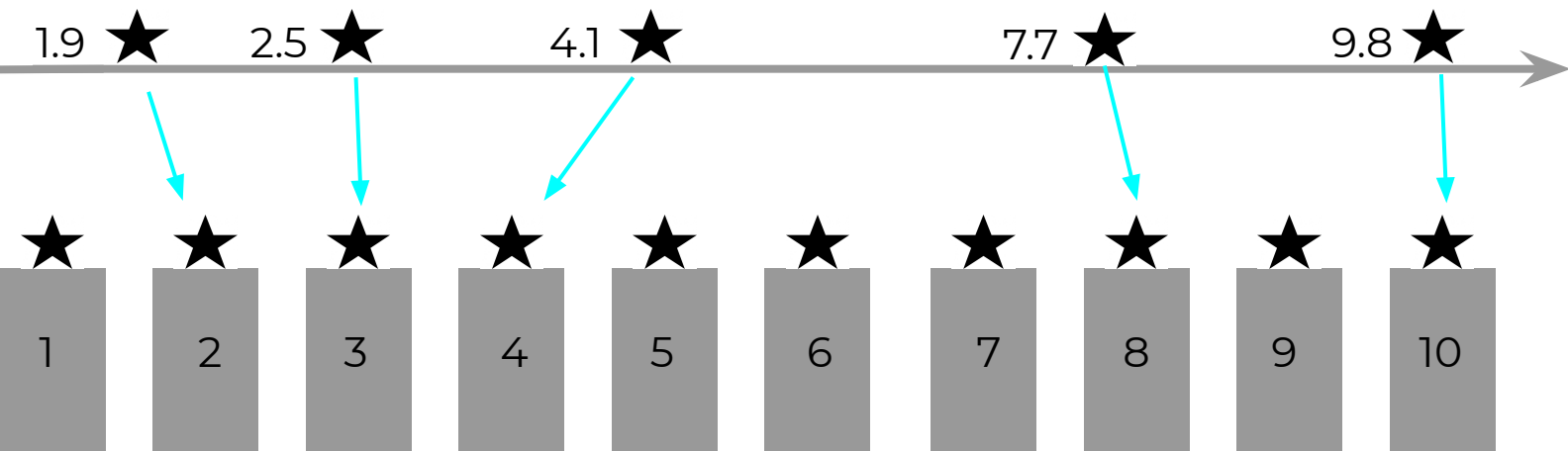
GENRE

COMPANY

These variables are encoded
because they are categorical

Our Measurement: Average Rating

Output variable was a continuous average rating for linear machine learning.
We created another variable rounding each float to an int. This created 10 categorical variables better fit for categorical machine learning.



Cleaning

Each row represents a different movie. Columns in our final data set include values for:

ID	An Identification number for each movie
Title	The final title of the movie
Budget	Given in USD
Popularity	A variable defined and formulated by IMDB
Revenue	Given in USD
Runtime	Length of the movie in minutes
Year	Year of Release
Genres	17 Columns with binary values denoting whether a movie belongs to a given genre
Production Companies	15 Columns with binary values denoting whether the top 15 production companies were involved in the production of the movie

Problems Faced

Problem 1

A lot of our rows had odd instances of \$0 budgets, or were not even released movies yet. We cleaned these rows and got rid of anything that had yet to be released or had \$0 budget or revenues

Problem 2

We wanted to use two categorical variables: genre and production company. However, a machine cannot understand “horror” or “Disney”, so we had to encode these categorical variables to be usable for machine learning.

Problem 3

We wanted to use the date of production data, but it was difficult to see patterns day-to-day. So, we created another column that returned each year the movie was released instead of the day/month/year so we could see trends over years.

Model Decision Tree

```
from sklearn import tree
import numpy as np
import pandas as pd
```

```
x_train = pd.read_csv(r"C:\Users\steph\Downloads\x_train.csv")
y_train = pd.read_csv(r"C:\Users\steph\Downloads\y_train.csv")
array = []
for i in Y_learn["Ratings"]:
    array.append(i)
array = np.array(array)
```

```
c = tree.DecisionTreeClassifier()
c.fit(x_train, y_train)
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                        max_depth=None, max_features=None, max_leaf_nodes=None,
                        min_impurity_decrease=0.0, min_impurity_split=None,
                        min_samples_leaf=1, min_samples_split=2,
                        min_weight_fraction_leaf=0.0, presort='deprecated',
                        random_state=None, splitter='best')
```

```
X_test = pd.read_csv(r"C:\Users\steph\Downloads\x_test.csv")
Y_test = pd.read_csv(r"C:\Users\steph\Downloads\y_test.csv")
array2 = []
for i in Y_test["vote_average"]:
    array2.append(i)
array2 = np.array(array2)
```

```
accu_train = np.sum(c.predict(x_train)==array)/float(array.size)
accu_train
```

1.0

```
accu_test= np.sum(c.predict(X_test)==array2)/float(array2.size)
accu_test
```

0.4803921568627451

Model

Random Forest Classifier

```
from sklearn import tree
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
```

```
] X = np.genfromtxt("Mytest.csv", delimiter=",", skip_header=1)
```

```
] Y = np.genfromtxt("Myanswers.csv", delimiter=",", skip_header=1)
```

```
] x_train, x_test, y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state=42)
```

```
] model = RandomForestClassifier(n_estimators=400, max_depth = 30, min_samples_leaf=3)
model.fit(x_train, y_train)
model.score(x_train, y_train)
```

```
0.8644897959183674
```

```
] model.score(x_test, Y_test)
```

```
0.6117455138662317
```

Final Model

Random Forest Regressor

```
[20] from sklearn.ensemble import RandomForestRegressor
      from sklearn.model_selection import train_test_split
      regressor = RandomForestRegressor(n_estimators=50, random_state = 0)
```

```
[21] X = np.genfromtxt("movies_last.csv", delimiter=",", skip_header=1)
```

```
[22] Y = np.genfromtxt("ratings_last.csv", delimiter=",", skip_header=1)
```

```
[23] X_tr, X_te, Y_tr, Y_te = train_test_split(X, Y, test_size = 0.2, random_state=42)
```

```
[24] regressor.fit(X_tr, Y_tr)
```

```
↳ RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                        max_depth=None, max_features='auto', max_leaf_nodes=None,
                        max_samples=None, min_impurity_decrease=0.0,
                        min_impurity_split=None, min_samples_leaf=1,
                        min_samples_split=2, min_weight_fraction_leaf=0.0,
                        n_estimators=50, n_jobs=None, oob_score=False,
                        random_state=0, verbose=0, warm_start=False)
```

```
[25] y_pred = regressor.predict(X_te)
```

```
[26] np.mean((y_pred-Y_te)**2)
```

```
↳ 0.29549911201629325
```

Our Improvements

ONE HOT ENCODE

Instead of doing it by hand, we would try and choose a defining genre for each movie and one hot encode to create an easier way to identify genre for each movie.

FEWER CATEGORIES

We have 10 “bins” for our average rating, but we could use 3 bins for “low”, “mid”, “high” ratings. This would give better accuracy in our machine learning output.

LINEAR REGRESSION

We would try using our original output variable (in its continuous form) to perform linear regression machine learning. This may produce even better results.