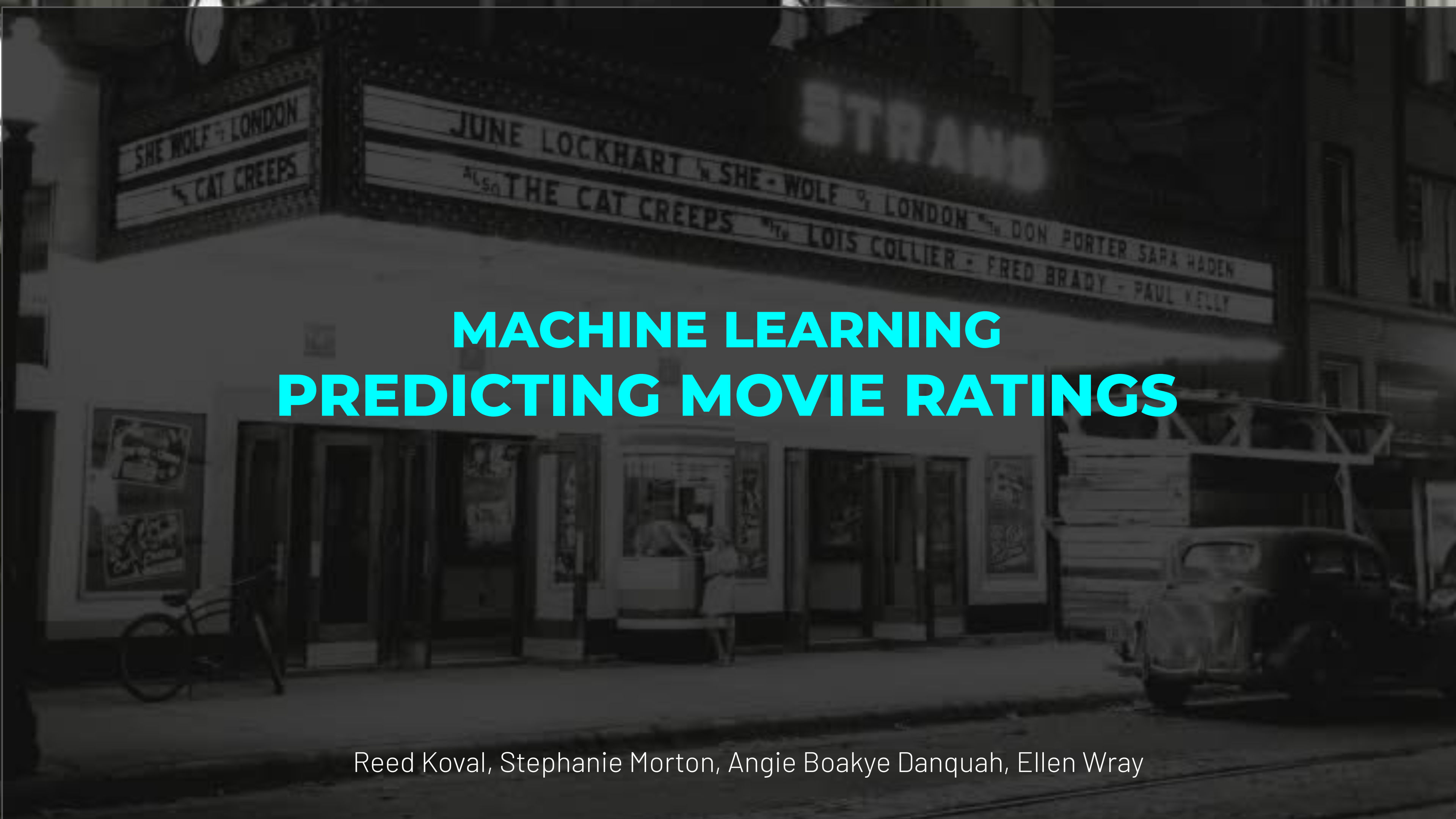

SHE WOLF OF LONDON
CAT CREEPS
JUNE LOCKHART IN SHE-WOLF OF LONDON WITH DON PORTER SARA HADEN
ALSO THE CAT CREEPS WITH LOIS COLLIER - FRED BRADY - PAUL KELLY
STRAND

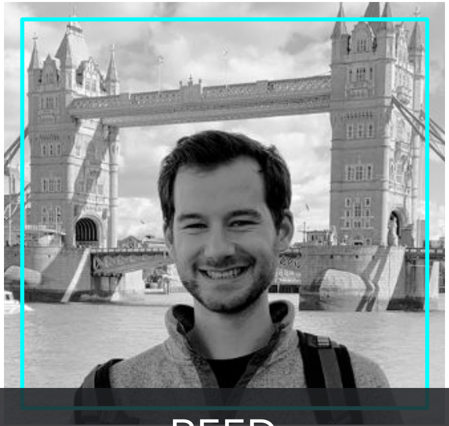
MACHINE LEARNING
PREDICTING MOVIE RATINGS

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

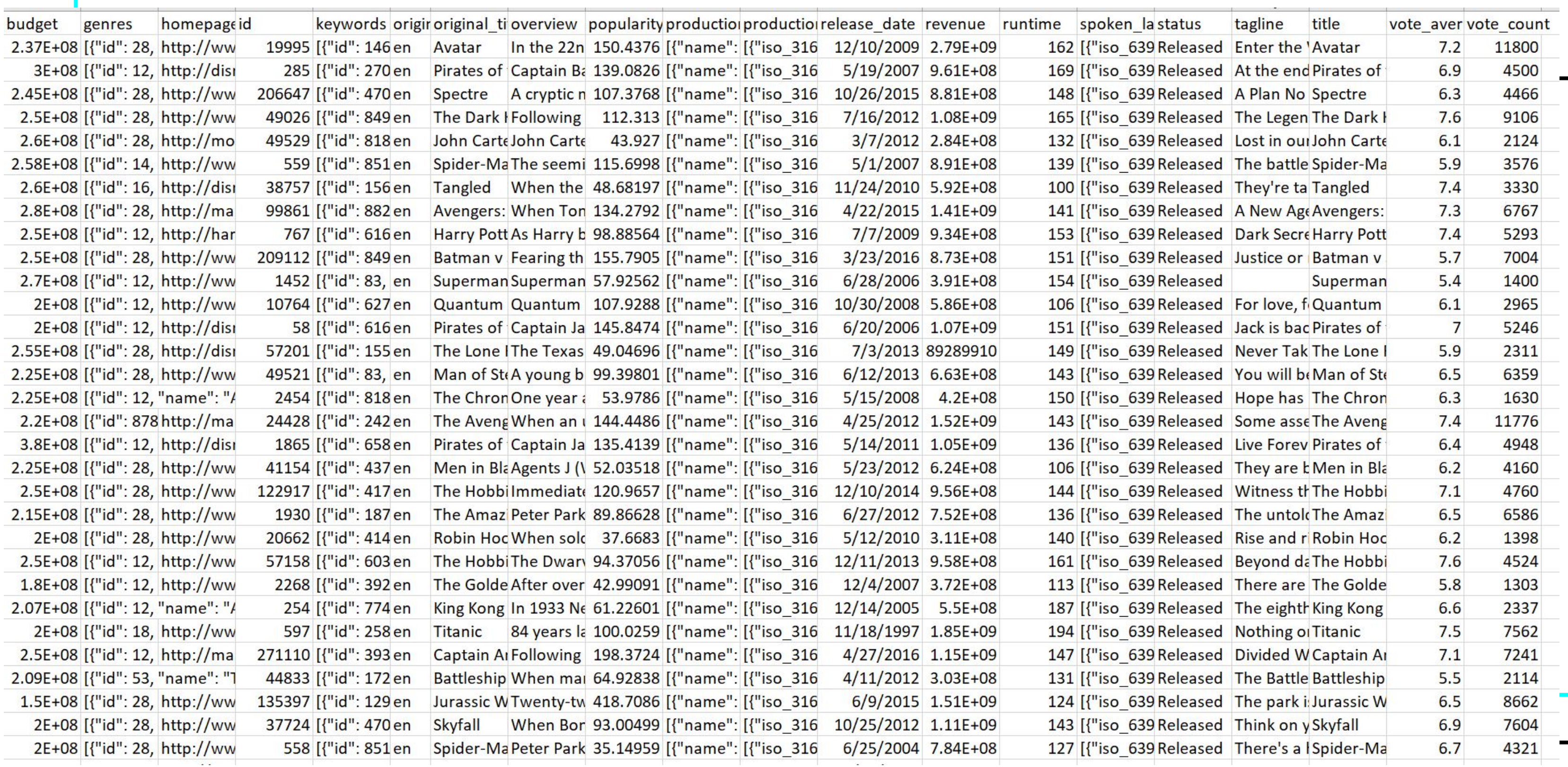
4th Year at UVA
Statistics & Economics

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

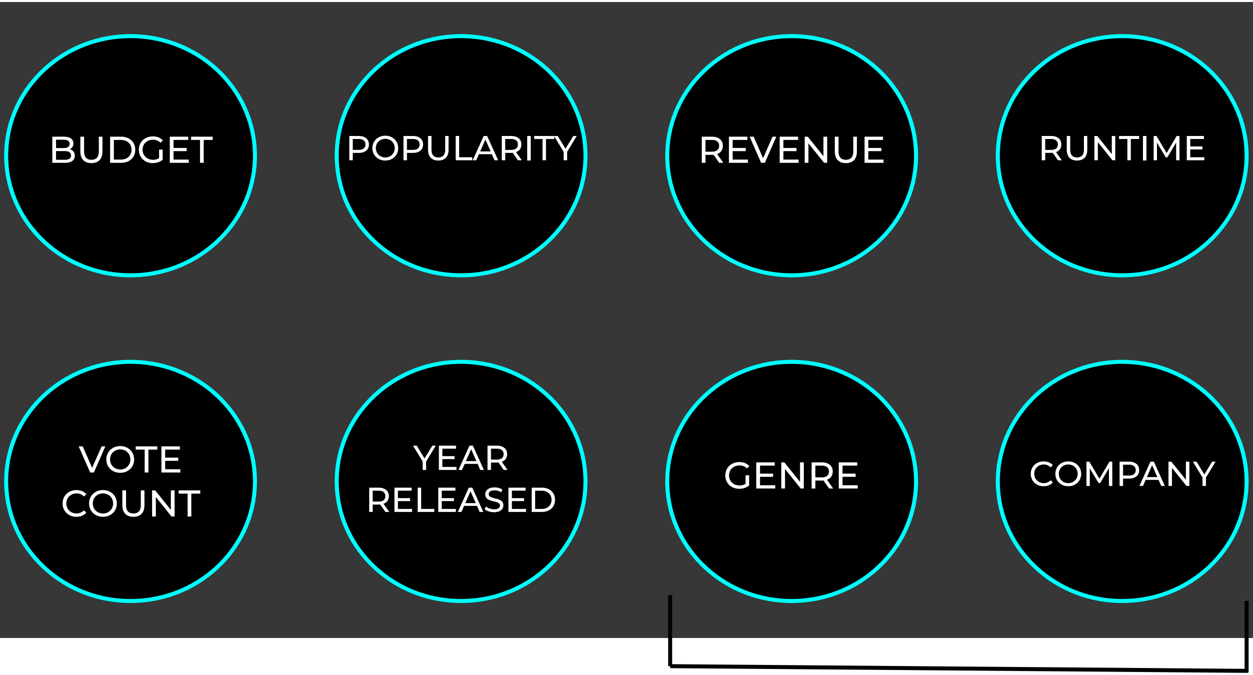
Our Data Set



budget	genres	homepage	id	keywords	origir	original_ti	overview	popularity	productioi	productioi	release_date	revenue	runtime	spoken_la	status	tagline	title	vote_aver	vote_count
2.37E+08	[{"id": 28,	http://ww	19995	[{"id": 146	en	Avatar	In the 22n	150.4376	[{"name":	[{"iso_316	12/10/2009	2.79E+09	162	[{"iso_639	Released	Enter the '	Avatar	7.2	11800
3E+08	[{"id": 12,	http://disı	285	[{"id": 270	en	Pirates of	Captain Ba	139.0826	[{"name":	[{"iso_316	5/19/2007	9.61E+08	169	[{"iso_639	Released	At the end	Pirates of '	6.9	4500
2.45E+08	[{"id": 28,	http://ww	206647	[{"id": 470	en	Spectre	A cryptic n	107.3768	[{"name":	[{"iso_316	10/26/2015	8.81E+08	148	[{"iso_639	Released	A Plan No	Spectre	6.3	4466
2.5E+08	[{"id": 28,	http://ww	49026	[{"id": 849	en	The Dark I	Following	112.313	[{"name":	[{"iso_316	7/16/2012	1.08E+09	165	[{"iso_639	Released	The Legen	The Dark I	7.6	9106
2.6E+08	[{"id": 28,	http://mo	49529	[{"id": 818	en	John Carte	John Carte	43.927	[{"name":	[{"iso_316	3/7/2012	2.84E+08	132	[{"iso_639	Released	Lost in oul	John Carte	6.1	2124
2.58E+08	[{"id": 14,	http://ww	559	[{"id": 851	en	Spider-Ma	The seemi	115.6998	[{"name":	[{"iso_316	5/1/2007	8.91E+08	139	[{"iso_639	Released	The battle	Spider-Ma	5.9	3576
2.6E+08	[{"id": 16,	http://disı	38757	[{"id": 156	en	Tangled	When the	48.68197	[{"name":	[{"iso_316	11/24/2010	5.92E+08	100	[{"iso_639	Released	They're ta	Tangled	7.4	3330
2.8E+08	[{"id": 28,	http://ma	99861	[{"id": 882	en	Avengers:	When Ton	134.2792	[{"name":	[{"iso_316	4/22/2015	1.41E+09	141	[{"iso_639	Released	A New Age	Avengers:	7.3	6767
2.5E+08	[{"id": 12,	http://har	767	[{"id": 616	en	Harry Pott	As Harry b	98.88564	[{"name":	[{"iso_316	7/7/2009	9.34E+08	153	[{"iso_639	Released	Dark Secre	Harry Pott	7.4	5293
2.5E+08	[{"id": 28,	http://ww	209112	[{"id": 849	en	Batman v	Fearing th	155.7905	[{"name":	[{"iso_316	3/23/2016	8.73E+08	151	[{"iso_639	Released	Justice or	Batman v	5.7	7004
2.7E+08	[{"id": 12,	http://ww	1452	[{"id": 83,	en	Superman	Superman	57.92562	[{"name":	[{"iso_316	6/28/2006	3.91E+08	154	[{"iso_639	Released		Superman	5.4	1400
2E+08	[{"id": 12,	http://ww	10764	[{"id": 627	en	Quantum	Quantum	107.9288	[{"name":	[{"iso_316	10/30/2008	5.86E+08	106	[{"iso_639	Released	For love, f	Quantum	6.1	2965
2E+08	[{"id": 12,	http://disı	58	[{"id": 616	en	Pirates of	Captain Ja	145.8474	[{"name":	[{"iso_316	6/20/2006	1.07E+09	151	[{"iso_639	Released	Jack is bac	Pirates of '	7	5246
2.55E+08	[{"id": 28,	http://disı	57201	[{"id": 155	en	The Lone I	The Texas	49.04696	[{"name":	[{"iso_316	7/3/2013	89289910	149	[{"iso_639	Released	Never Tak	The Lone I	5.9	2311
2.25E+08	[{"id": 28,	http://ww	49521	[{"id": 83,	en	Man of Ste	A young b	99.39801	[{"name":	[{"iso_316	6/12/2013	6.63E+08	143	[{"iso_639	Released	You will be	Man of Ste	6.5	6359
2.25E+08	[{"id": 12,	"name": "/	2454	[{"id": 818	en	The Chron	One year a	53.9786	[{"name":	[{"iso_316	5/15/2008	4.2E+08	150	[{"iso_639	Released	Hope has	The Chron	6.3	1630
2.2E+08	[{"id": 878	http://ma	24428	[{"id": 242	en	The Aveng	When an u	144.4486	[{"name":	[{"iso_316	4/25/2012	1.52E+09	143	[{"iso_639	Released	Some asse	The Aveng	7.4	11776
3.8E+08	[{"id": 12,	http://disı	1865	[{"id": 658	en	Pirates of	Captain Ja	135.4139	[{"name":	[{"iso_316	5/14/2011	1.05E+09	136	[{"iso_639	Released	Live Forev	Pirates of '	6.4	4948
2.25E+08	[{"id": 28,	http://ww	41154	[{"id": 437	en	Men in Bla	Agents J (\	52.03518	[{"name":	[{"iso_316	5/23/2012	6.24E+08	106	[{"iso_639	Released	They are b	Men in Bla	6.2	4160
2.5E+08	[{"id": 28,	http://ww	122917	[{"id": 417	en	The Hobbi	Immediate	120.9657	[{"name":	[{"iso_316	12/10/2014	9.56E+08	144	[{"iso_639	Released	Witness th	The Hobbi	7.1	4760
2.15E+08	[{"id": 28,	http://ww	1930	[{"id": 187	en	The Amaz	Peter Park	89.86628	[{"name":	[{"iso_316	6/27/2012	7.52E+08	136	[{"iso_639	Released	The untolc	The Amaz	6.5	6586
2E+08	[{"id": 28,	http://ww	20662	[{"id": 414	en	Robin Hoc	When sold	37.6683	[{"name":	[{"iso_316	5/12/2010	3.11E+08	140	[{"iso_639	Released	Rise and r	Robin Hoc	6.2	1398
2.5E+08	[{"id": 12,	http://ww	57158	[{"id": 603	en	The Hobbi	The Dwarv	94.37056	[{"name":	[{"iso_316	12/11/2013	9.58E+08	161	[{"iso_639	Released	Beyond da	The Hobbi	7.6	4524
1.8E+08	[{"id": 12,	http://ww	2268	[{"id": 392	en	The Golde	After over	42.99091	[{"name":	[{"iso_316	12/4/2007	3.72E+08	113	[{"iso_639	Released	There are	The Golde	5.8	1303
2.07E+08	[{"id": 12,	"name": "/	254	[{"id": 774	en	King Kong	In 1933 Ne	61.22601	[{"name":	[{"iso_316	12/14/2005	5.5E+08	187	[{"iso_639	Released	The eighth	King Kong	6.6	2337
2E+08	[{"id": 18,	http://ww	597	[{"id": 258	en	Titanic	84 years la	100.0259	[{"name":	[{"iso_316	11/18/1997	1.85E+09	194	[{"iso_639	Released	Nothing oı	Titanic	7.5	7562
2.5E+08	[{"id": 12,	http://ma	271110	[{"id": 393	en	Captain Ar	Following	198.3724	[{"name":	[{"iso_316	4/27/2016	1.15E+09	147	[{"iso_639	Released	Divided W	Captain Ar	7.1	7241
2.09E+08	[{"id": 53,	"name": "T	44833	[{"id": 172	en	Battleship	When mai	64.92838	[{"name":	[{"iso_316	4/11/2012	3.03E+08	131	[{"iso_639	Released	The Battle	Battleship	5.5	2114
1.5E+08	[{"id": 28,	http://ww	135397	[{"id": 129	en	Jurassic W	Twenty-tw	418.7086	[{"name":	[{"iso_316	6/9/2015	1.51E+09	124	[{"iso_639	Released	The park i	Jurassic W	6.5	8662
2E+08	[{"id": 28,	http://ww	37724	[{"id": 470	en	Skyfall	When Bor	93.00499	[{"name":	[{"iso_316	10/25/2012	1.11E+09	143	[{"iso_639	Released	Think on y	Skyfall	6.9	7604
2E+08	[{"id": 28,	http://ww	558	[{"id": 851	en	Spider-Ma	Peter Park	35.14959	[{"name":	[{"iso_316	6/25/2004	7.84E+08	127	[{"iso_639	Released	There's a I	Spider-Ma	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

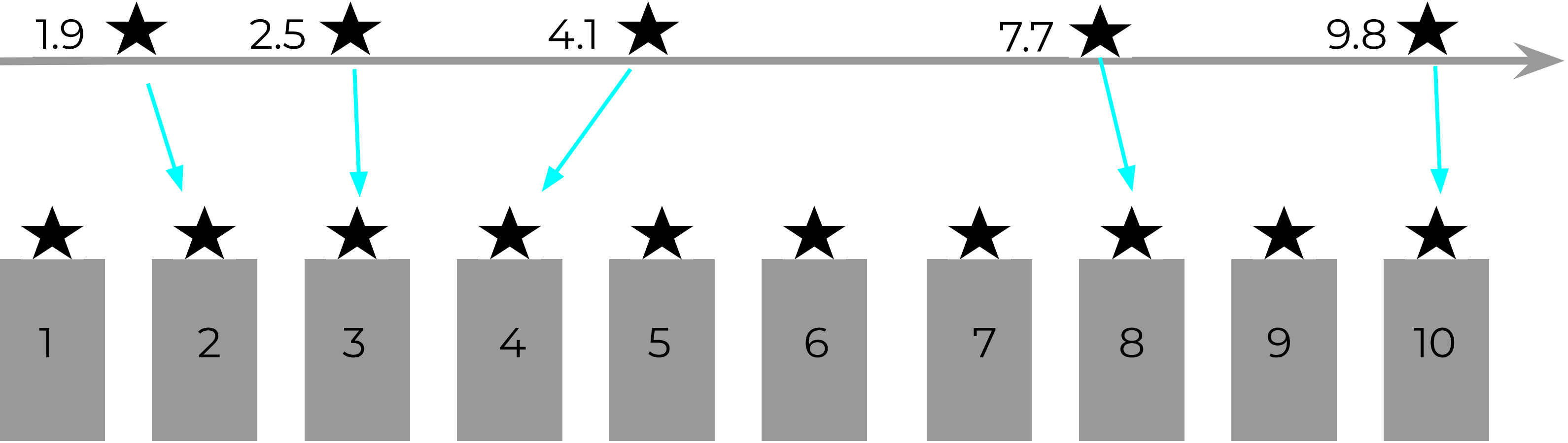
GENRE

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

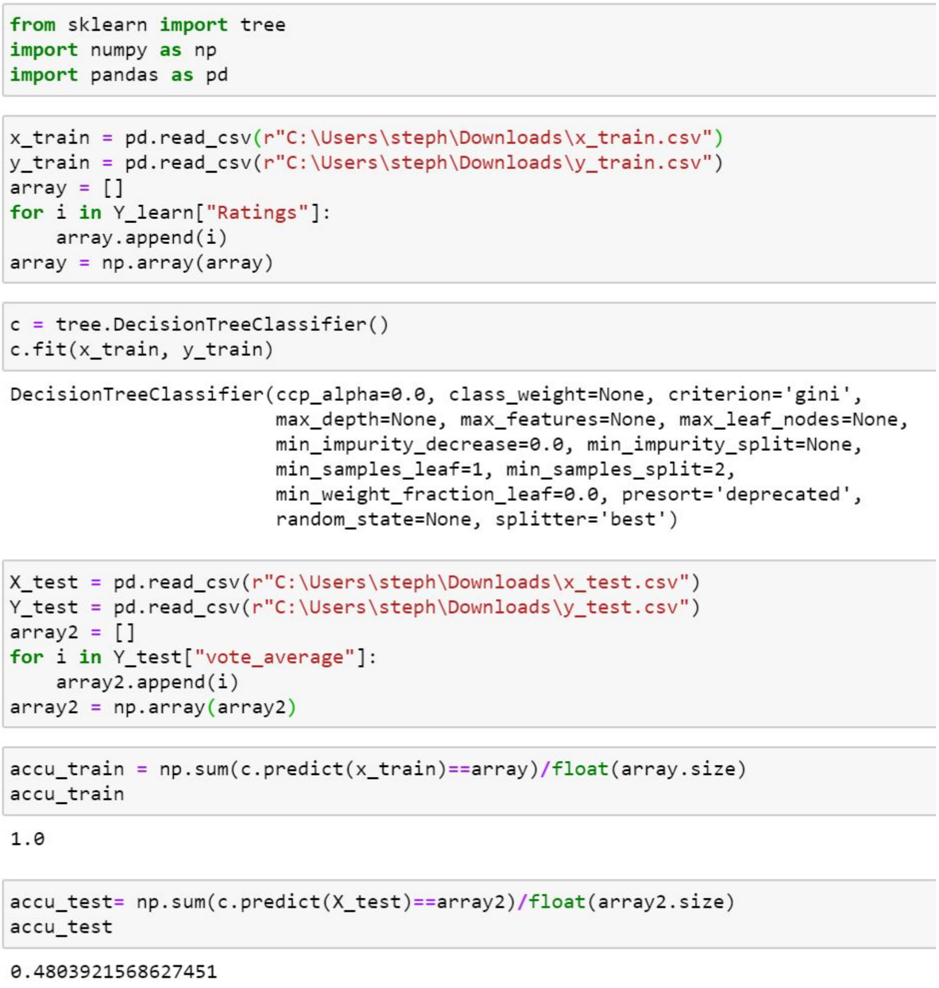
Problem 2

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
```

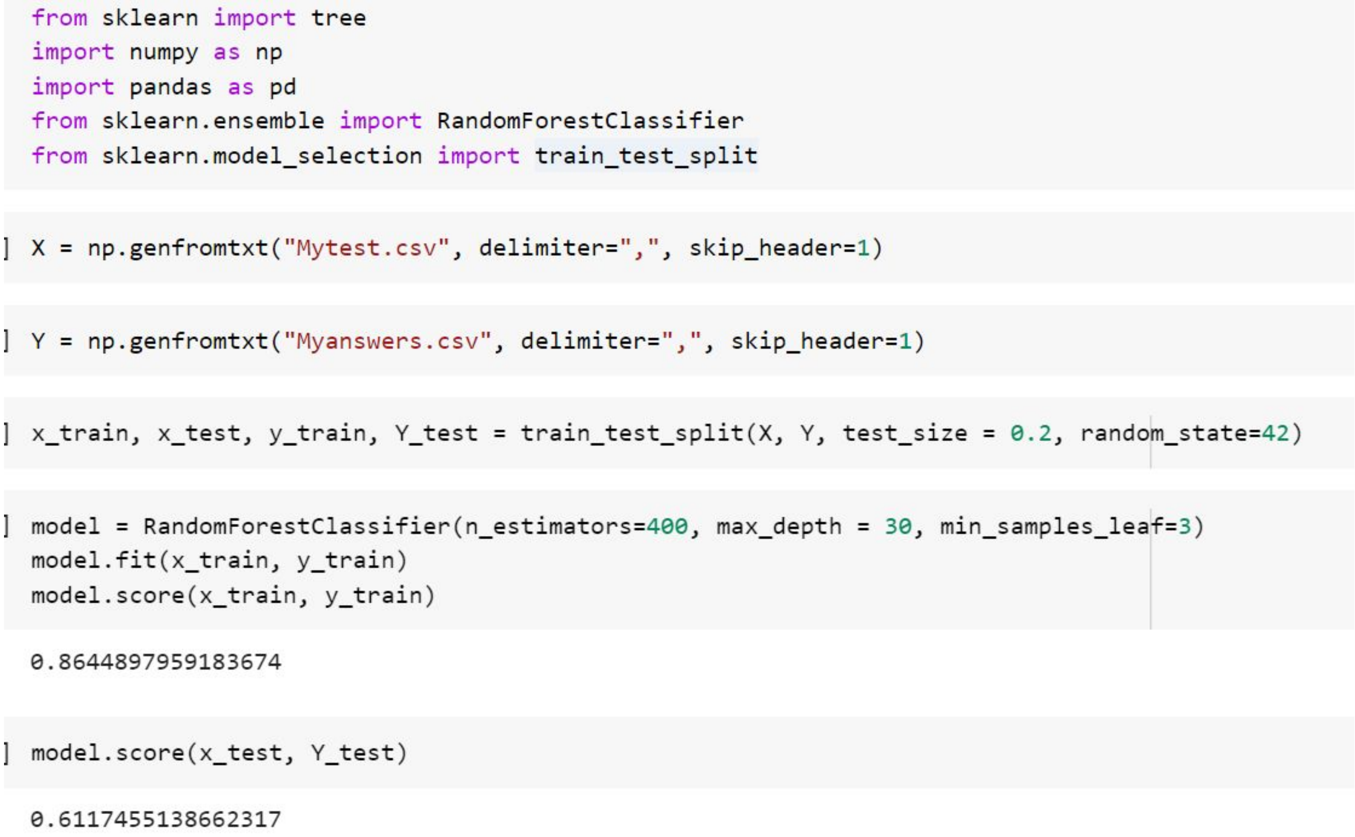
```
1.0
```

```
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)
```

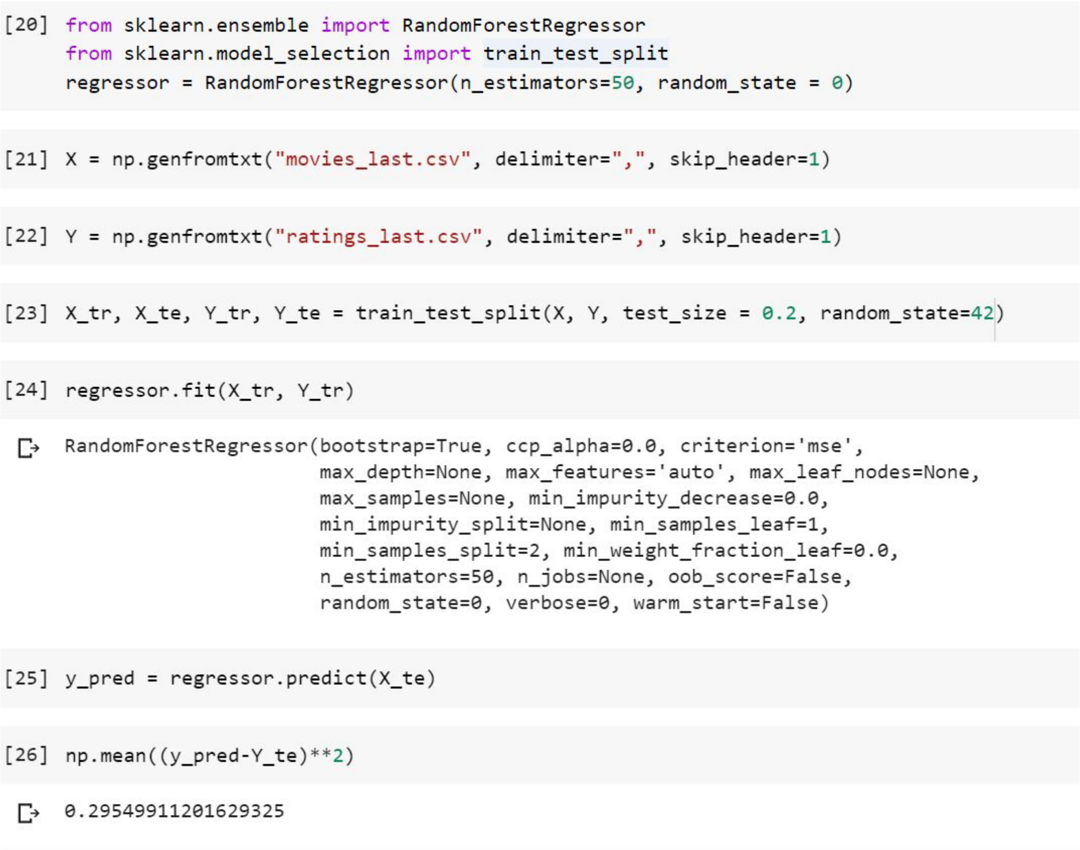
```
0.8644897959183674
```

```
] 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.