Can six simple measures predict movie ratings?

Data Analytics in Engineering project
Jingyu Chen
Xiuping Yang
Yi Yu

Introduction

- The quality of movies: UNEVEN
- Imagine being a movie investor and you want to invest in great movies, can you predict the rating of a movie, or whether a movie is great or not?

Objectives

- To predict rating (1 to 10)
- To predict classes of movies: Great or Average
- To give helpful suggestions on making great movies

We use IMDb Dataset

- Use R to scrape data from IMDb.com

  IMDb.com is the largest online database, which has tons of information for over 4 million titles, including movies, TV series and others.

What is our scraping?

- ‘Scraping’ is also named ‘Fetching’
- It is a bunch of codes which can recognize the detail information on the website and collect them
- 10000+ movies are randomly collected in several hours
- All of our data is fetched from IMDB instead of simply downloading!

Dataset with budget information

- 3809 objects
- 28 variables

Description of the variables

- Variables we use after data preprocessing:

Rating [1.9, 9.3]
Year [1980, 2017]
Budget [$4500, $12 billion]
Runtime [63, 325](min)

Genres (20 types)
Top 210 Director?
Top 1000 Actor?

- Genres, top directors and top actors: Binary variables

- Lists of top directors / actors are from IMDb.com

We have 20 genres:

- Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy, History, Horror, Music, Musical, Mystery, Romance, Sci-Fi, Sport, Thriller, War, Western
- Within such genre: 1, otherwise: 0

Rating ≥ 8.0 means Great!

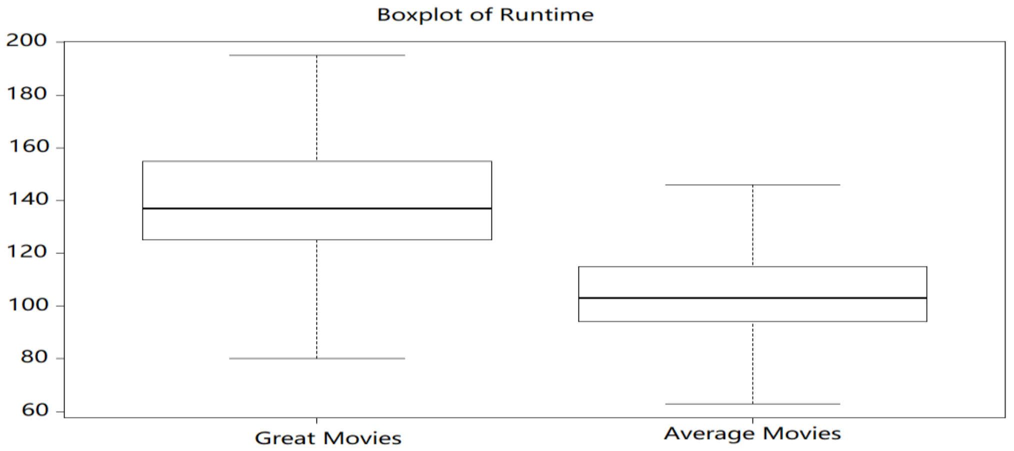
When it comes to classification, we have to draw a line between great movies and average movies.

We decide to choose rating of 8 as the dividing threshold.

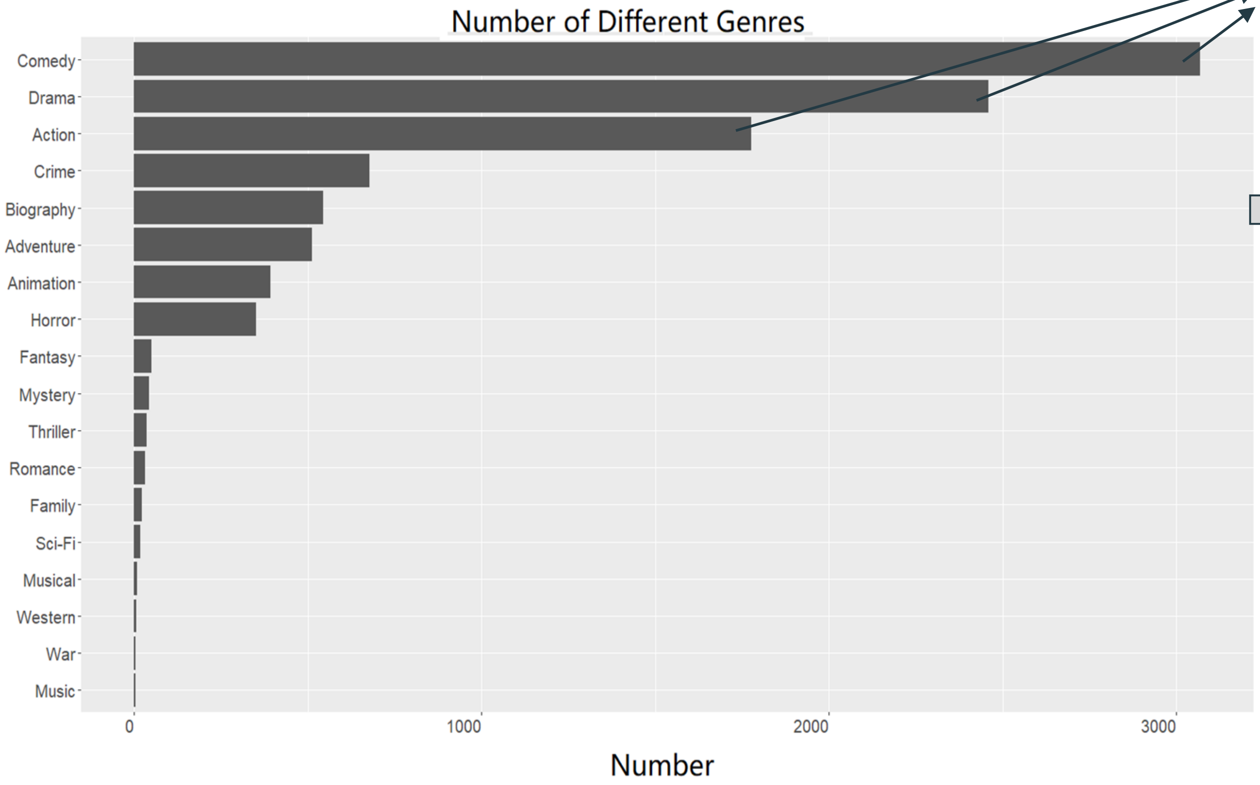
Key takeaways:

- Longer runtime associates with higher rating and great movies
- Genres have a significant influence on the movies
  - Drama, Adventure, Animation ---> Better!
  - Comedy, Action, Romance, Horror ---> Not good
- Top actors and directors involved really matters

Great movies tend to be longer!

	Great Movies	Average Movies
Average runtime	139.82 minutes	106.38 minutes
Median runtime	137 minutes	103 minutes
Percentage of long runtime (> 120 mins) movies	78.18%	17.27%
Box-plot		

Comedy, Drama and Action movies are more



1st Comedy
2nd Drama
3rd Action

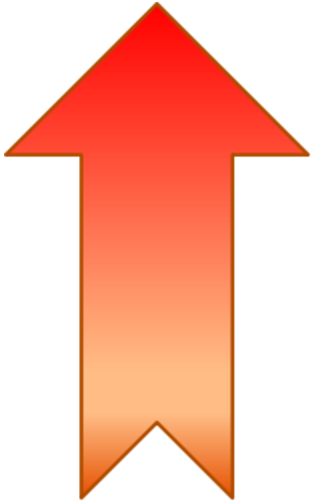
Chi - Square test will tell whether each genre is a significant attribute or not

Chi-Square test says Genres matter

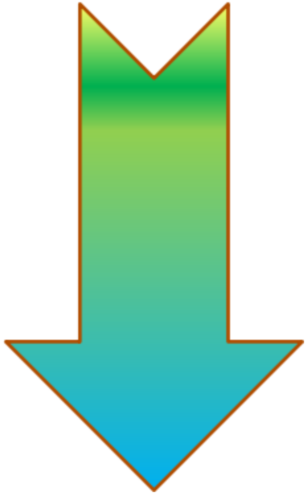
	Great Movies	Average Movies
Action	39%	26%
Adventure	31%	21%
Animation	12%	5%
Comedy	13%	40%
Drama	65%	50%
Horror	2%	11%
Romance	7%	18%

Chi-Square test says Genres matter

These genres have statistically significant impact on chances of being a great movie:



Drama
Adventure
Animation



Comedy
Action
Romance
Horror

Chi-Square says Top Actor & Top Director Matters!

	Great Movies	Average Movies
% with top actors	15%	6%
% with top directors	4%	<1%
Average rating	8.2	6.3

Leonardo DiCaprio:
Inception(8.8),
The wolf of Wall Street(8.2),
Shutter Island(8.1)

Christopher Nolan:
The dark knight(9.0),
Inception(8.8),
Interstellar(8.6)

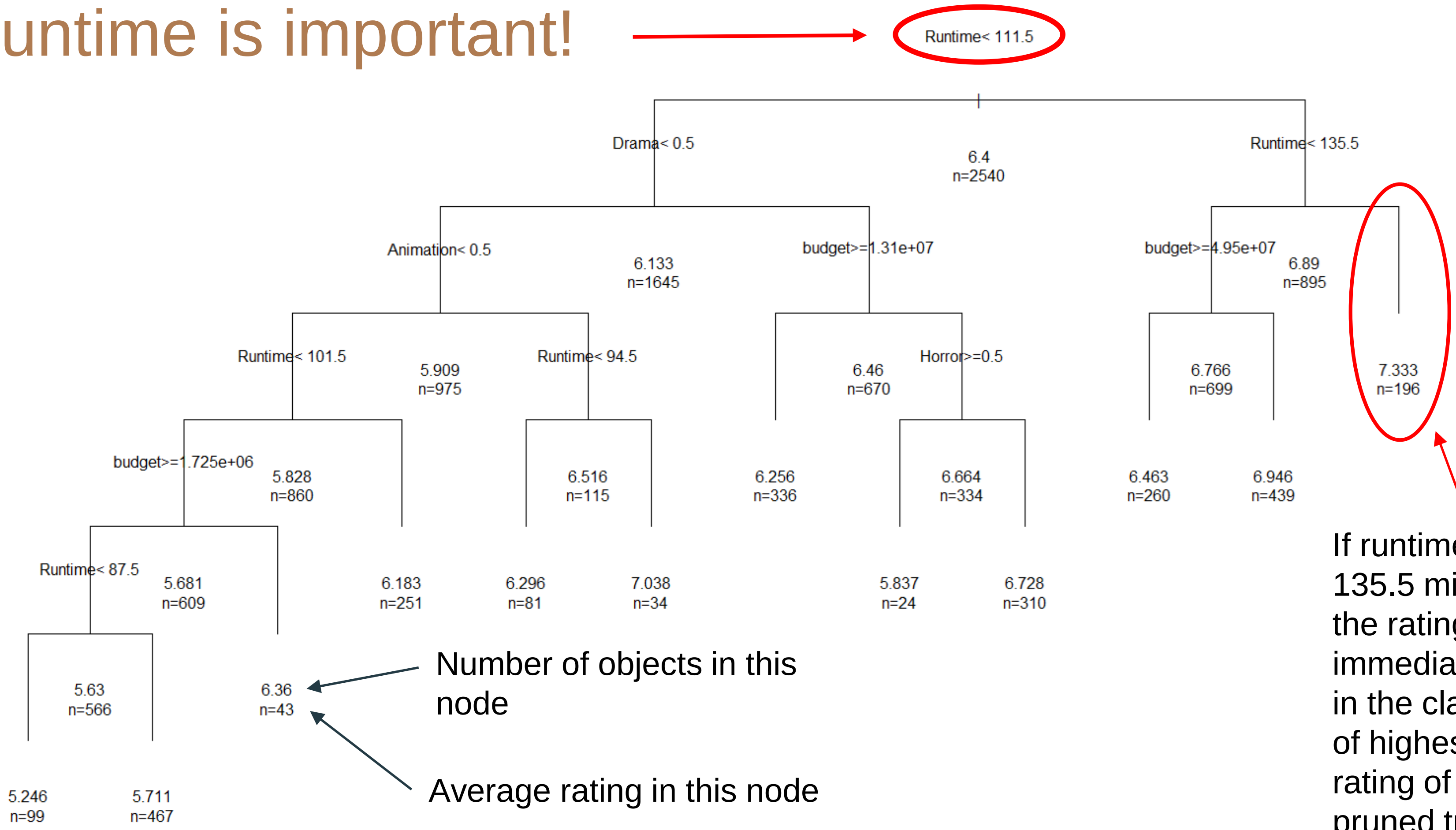
Let’s predict ratings!

- Data is randomly splitted into training (⅔) and testing (⅓).
- Regression tree and random forest are used to train on the training set and predict rating of the testing set.

Conclusions from rating prediction

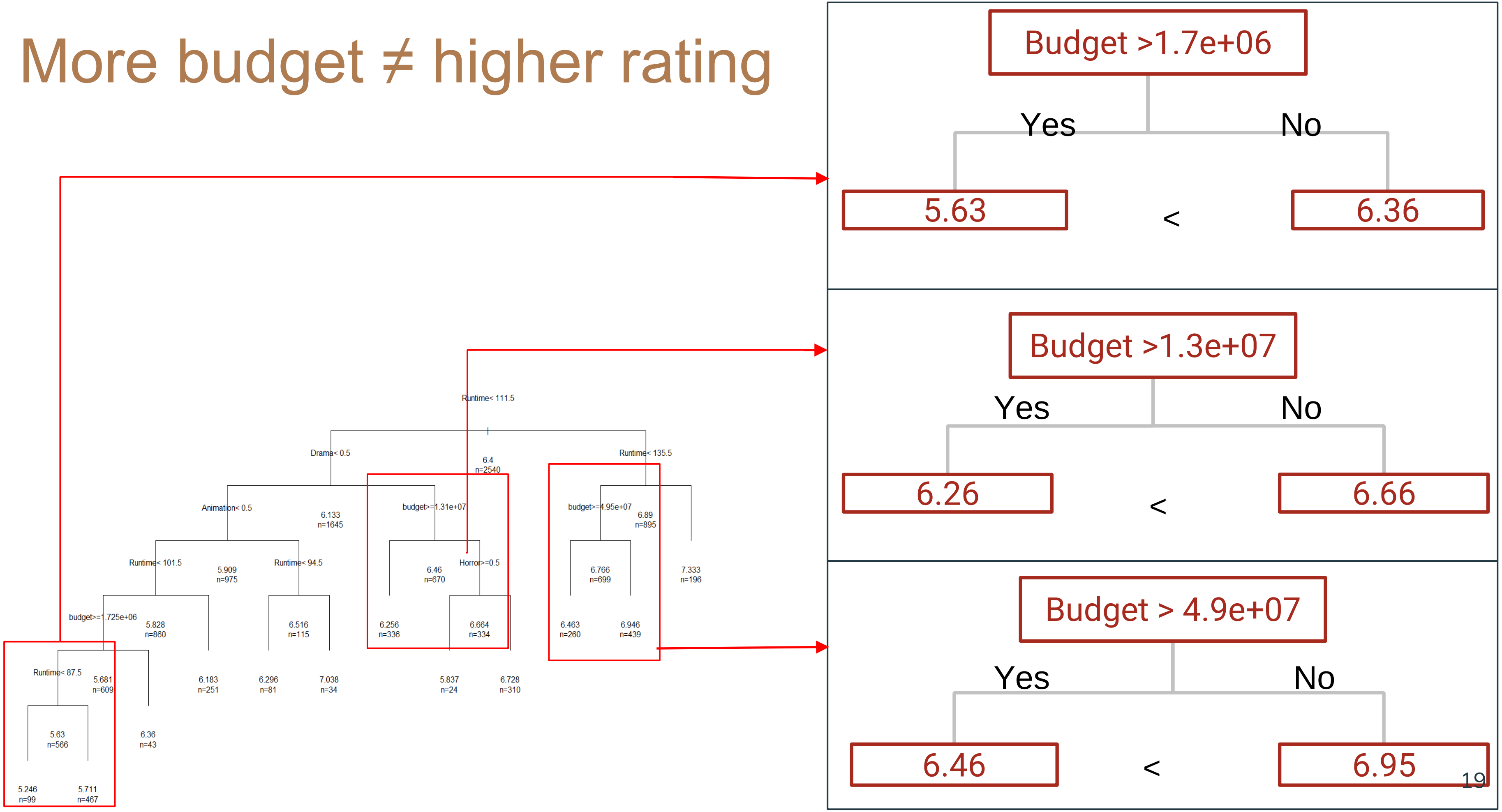
- Longer runtime tends to have higher rating
- Higher budget ≠ higher rating
- Drama and animation tends to have higher rating
- Predicting error is large

Runtime is important!

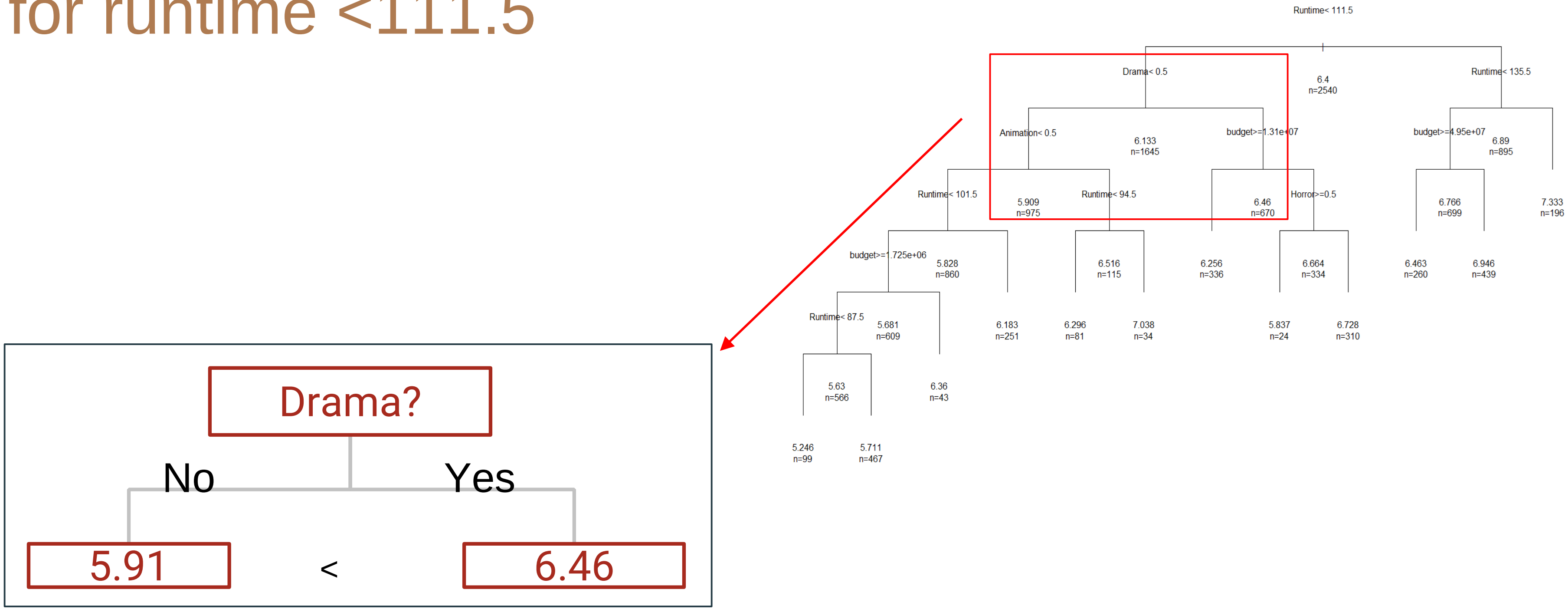


If runtime > 135.5 min, the rating is immediately in the class of highest rating of this pruned tree

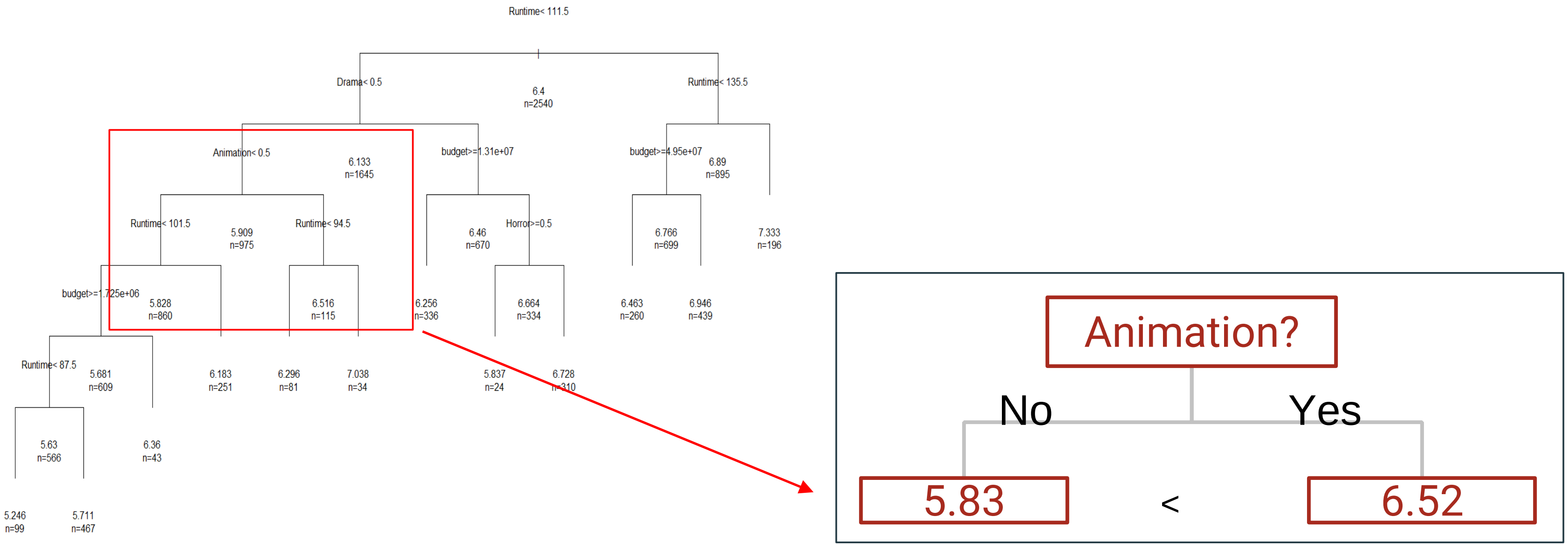
More budget ≠ higher rating



Drama has higher rating for runtime <111.5



Animation has higher ratings



Random forest has better accuracy than regression tree

Measure	Decision Tree	Random Forest
Root mean square error (standard error)	0.91	0.85
Mean absolute error	0.68	0.65

But how good is Random forest?

Predict rating of the test set using the average rating of the training set

Measure	Using average rating of the training set as the predicted rating
Root mean square error (standard error)	1.04
Mean absolute error	0.81

Let's treat this as the baseline.

Random forest improves the baseline by only 20%

Measure	Baseline	Random Forest
Root mean square error (standard error)	1.04	0.85
Mean absolute error	0.81	0.65

Conclusions from rating prediction

- Longer runtime tends to have higher rating
- Higher budget ≠ higher rating
- Drama and animation tends to have higher rating
- Predicting error is large

How about Classification Model?

Classification Models:Dataset

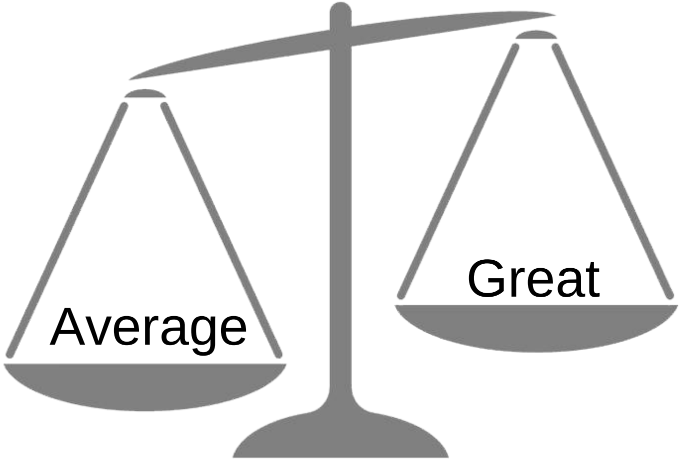
Rating ≥ 8.0: Great movies (4%)

Train set: 70%   Test set: 30%

Stratified Sampling (4% great movies in train set & test set)

The data is Unbalanced!

Only 4% is considered as Great movies

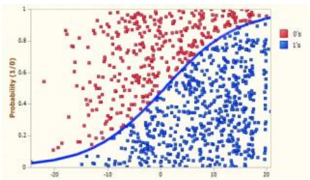


Using statistically significant attributes to build models

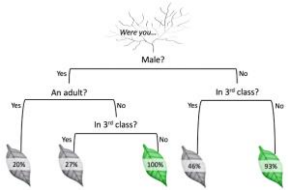
Runtime
Top 210 Directors?
Top 1000 Actors?
Genres:

Drama
Adventure
Animation

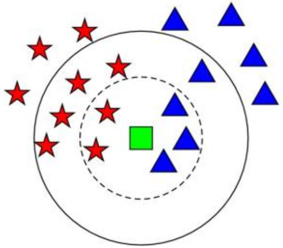
Comedy
Action
Romance
Horror



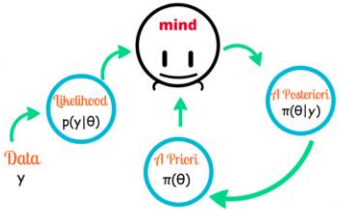
Logistic Regression



Decision Tree

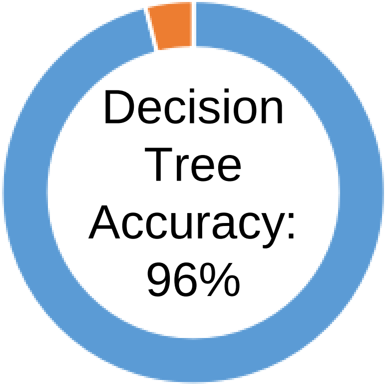


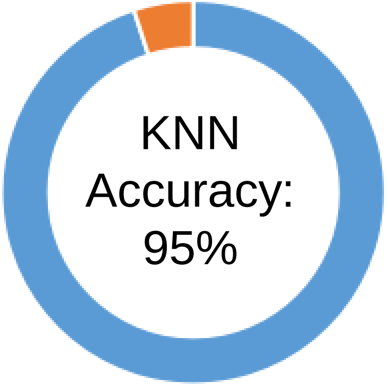
KNN

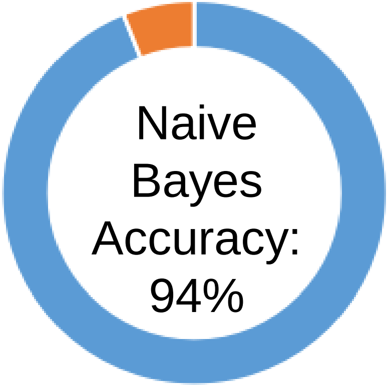


Naive Bayes

Measuring Accuracy







They are pretty much the same

Sensitivity vs. Specificity

Sensitivity:

Percentage of great movies that are correctly identified

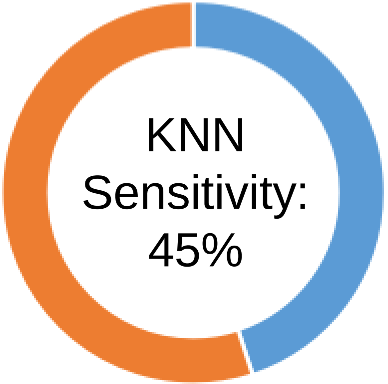
Sensitivity = TP / P

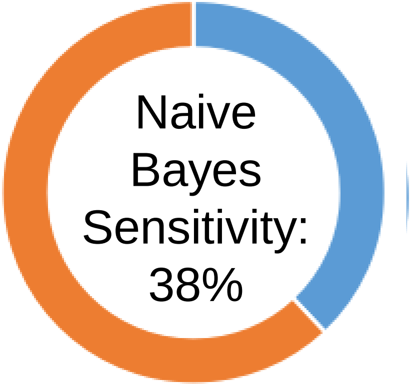
Specificity:

Percentage of average movies that are correctly identified

Sensitivity = TN / N

Logistic Regression has the highest sensitivity





However, the sensitivities are not satisfying
How about another technique?

Using method of Upsampling

Upsampling: increase the frequency of great movies in sample, to balance the ratio of average movies and great movies to about 50-50

Logistic Regression keeps ahead

	Logistic Regression	Decision Tree	KNN	Naive-Bayes
Accuracy	77%	70%	68%	63%
Sensitivity	77%	74%	70%	36%

The accuracies are pretty low

We can’t predict class with only 6 simple measures

But Logistic Regression model is the best classification model since it’s in the first place twice

Details on Logistic Regression model

If the coefficient is positive, then there's greater chance to make the movie a great one by increasing that variable

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.3684	0.4311	-7.813	5.60e-15	***
Year	-1.4905	0.4459	-3.342	0.000831	***
Runtime	8.4365	1.2091	6.978	3.00e-12	***
Action	-1.1449	0.2815	-4.068	4.75e-05	***
Animation	2.0235	0.3635	5.566	2.60e-08	***
Comedy	-1.3474	0.3229	-4.173	3.01e-05	***
Horror	-1.4211	0.6062	-2.344	0.019060	*
Romance	-1.8124	0.5263	-3.444	0.000574	***
TopActor	0.6318	0.3304	1.912	0.055813	.

Suggestions from classification models

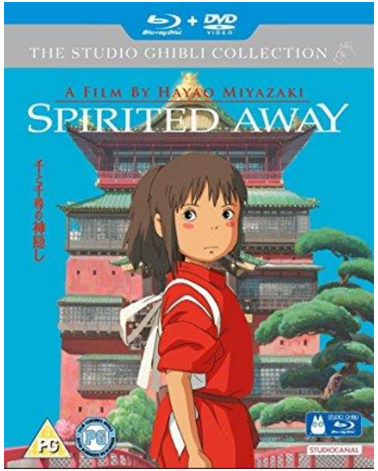
- Make it LONG!
- Make it ANIMATION!
- Hire top ACTORS!

It's helpful to...

Make it LONG!

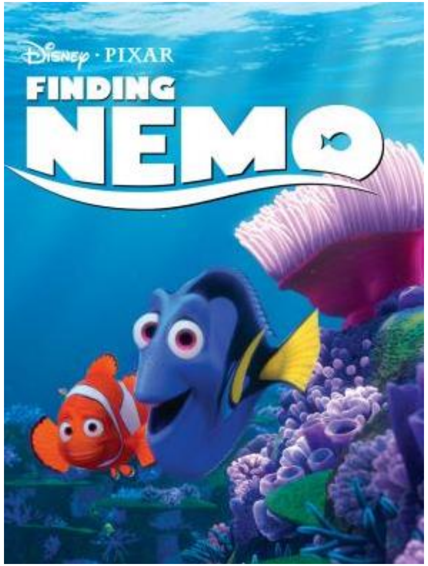
It’s helpful to…

Make it ANIMATION



It’s helpful to...

Have top ACTORS

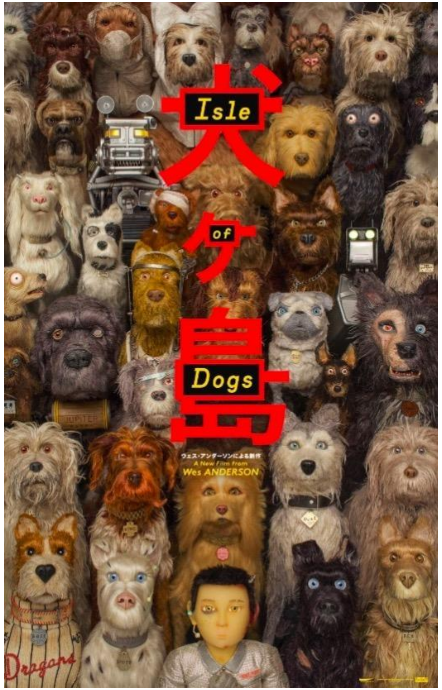


Prediction of 2018’s movies

Prediction:
Average /6.3
Actual: Average /7.7

Prediction:
Average /6.1
Actual: Average /4.2

Prediction of 2018's movies



Prediction:
Great
/7.3
Actual:
Great
/8.2

Prediction:
Average
/6.7
Actual:
Yet to see

If you want to make a great movie...

- Longer runtime associates with higher rating and great movies
- Genres have a significant influence on the movies
  - Drama, Adventure, Animation ---> Better!
  - Comedy, Action, Romance, Horror ---> Not good
- Top actors and directors involved really matters

Further Improvement:

- The results of our models are not so good
  - There are many relevant factors that we can’t use or are hard to quantify
- Try to get more data from other sources
- Try new methods other than the ones we learned from class


tusind tak
谢谢
dakujem vám
ありがとう
ngiyabonga
dziękuję
merci
thank
baie dankie
धन्यवाद
molte grazie
suksema
danke
gracias
takk
you
obrigada
obrigado
dank u
mahalo
gràcies
teşekkür ederim
شكرا
tänan
tack så mycket

Questions?