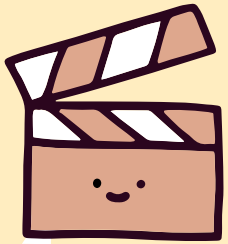

# PREDICTING BOX OFFICE SUCCESS

Marc Muszik
Fei Xu
Rita Zou
Andrew Conaway
Xinyu Cao



# Agenda

01

Data Cleaning

02

Data Preparation

03

Clustering

04

Model Building

05

Summary

# Real World Application

- ❏ Film budgeting processing
- ❏ Casting crew evaluation
- ❏ Potential challenges

- Algorithm should be more dynamic to adapt changes
- Handle Imbalanced Dataset


# Data Cleaning

- Removal of text heavy columns (Poster Link + Overview)
- Fixed data misplacement issue
- Correcting data types
- Dummy Variables for Certificate

# Data Cleaning

## Step 1: Remove all rows with NA values for Gross

| Title | Release Year | Runtime | Genre | Director | Star1 | Star2 | Gross |
|---|---|---|---|---|---|---|---|
| Office Space | 1999 | 89 min | Comedy | Mike Judge | Ron Livingston | Jennifer Aniston | NA |
| Happiness | 1998 | 134 min | Comedy, Drama | Todd Solondz | Jane Adams | Jon Lovitz | 2,807,390 |
| Training Day | 2001 | 122 min | Crime, Drama, Thriller | Antoine Fuqua | Denzel Washington | Ethan Hawke | 76,631,907 |
| Rushmore | 93 min | NA | Comedy, Drama, Romance | Wes Anderson | Jason Schwartzman | Bill Murray | 17,105,219 |
| Abre los ojos | 1997 | 119 min | Drama, Mystery, Sci-Fi | Alejandro Amenbar | Eduardo Noriega | Penelope Cruz | 368,234 |

# Data Cleaning

## Step 2: Fix formatting for Gross

| Title | Release Year | Runtime | Genre | Director | Star1 | Star2 | Gross |
|---|---|---|---|---|---|---|---|
| Happiness | 1998 | 134 min | Comedy, Drama | Todd Solondz | Jane Adams | Jon Lovitz | 2,807,390 |
| Training Day | 2001 | 122 min | Crime, Drama, Thriller | Antoine Fuqua | Denzel Washington | Ethan Hawke | 76,631,907 |
| Rushmore | 93 min | NA | Comedy, Drama, Romance | Wes Anderson | Jason Schwartzman | Bill Murray | 17,105,219 |
| Abre los ojos | 1997 | 119 min | Drama, Mystery, Sci-Fi | Alejandro Amenbar | Eduardo Noriega | Penelope Cruz | 368,234 |

# Data Cleaning

## Step 3: Fix data misplacement issue

| Title | Release Year | Runtime | Genre | Director | Star1 | Star2 | Gross |
|---|---|---|---|---|---|---|---|
| Happiness | 1998 | 134 min | Comedy, Drama | Todd Solondz | Jane Adams | Jon Lovitz | 2807390 |
| Training Day | 2001 | 122 min | Crime, Drama, Thriller | Antoine Fuqua | Denzel Washington | Ethan Hawke | 76631907 |
| Rushmore | 93 min | NA | Comedy, Drama, Romance | Wes Anderson | Jason Schwartzman | Bill Murray | 17105219 |
| Abre los ojos | 1997 | 119 min | Drama, Mystery, Sci-Fi | Alejandro Amenbar | Eduardo Noriega | Penelope Cruz | 368234 |

# Data Cleaning

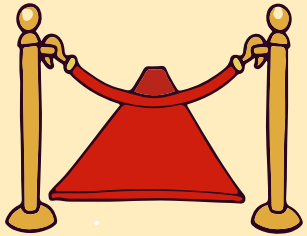
## Step 4: Fix Runtime formatting

| Title | Release Year | Runtime | Genre | Director | Star1 | Star2 | Gross |
|---|---|---|---|---|---|---|---|
| Happiness | 1998 | 134 min | Comedy, Drama | Todd Solondz | Jane Adams | Jon Lovitz | 2807390 |
| Training Day | 2001 | 122 min | Crime, Drama, Thriller | Antoine Fuqua | Denzel Washington | Ethan Hawke | 76631907 |
| Rushmore | 1998 | 93 min | Comedy, Drama, Romance | Wes Anderson | Jason Schwartzman | Bill Murray | 17105219 |
| Abre los ojos | 1997 | 119 min | Drama, Mystery, Sci-Fi | Alejandro Amenbar | Eduardo Noriega | Penelope Cruz | 368234 |

# Data Cleaning

## Step 4: Fix Runtime formatting

| Title | Release Year | Runtime | Genre | Director | Star1 | Star2 | Gross |
|---|---|---|---|---|---|---|---|
| Happiness | 1998 | 134 | Comedy, Drama | Todd Solondz | Jane Adams | Jon Lovitz | 2807390 |
| Training Day | 2001 | 122 | Crime, Drama, Thriller | Antoine Fuqua | Denzel Washington | Ethan Hawke | 76631907 |
| Rushmore | 1998 | 93 | Comedy, Drama, Romance | Wes Anderson | Jason Schwartzman | Bill Murray | 17105219 |
| Abre los ojos | 1997 | 119 | Drama, Mystery, Sci-Fi | Alejandro Amenbar | Eduardo Noriega | Penelope Cruz | 368234 |

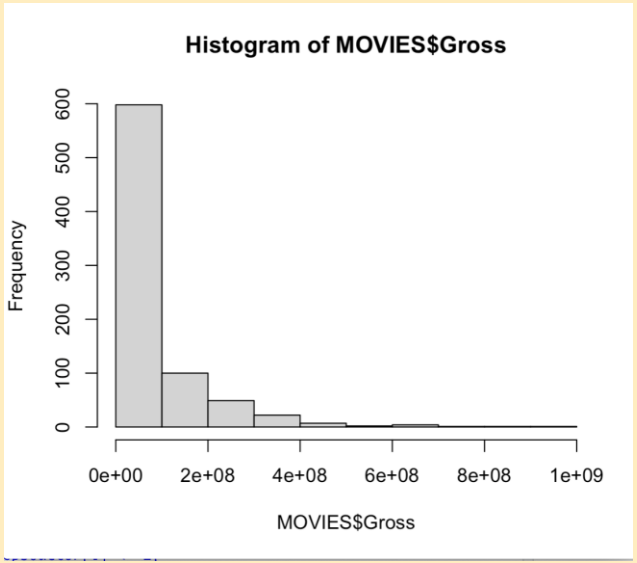


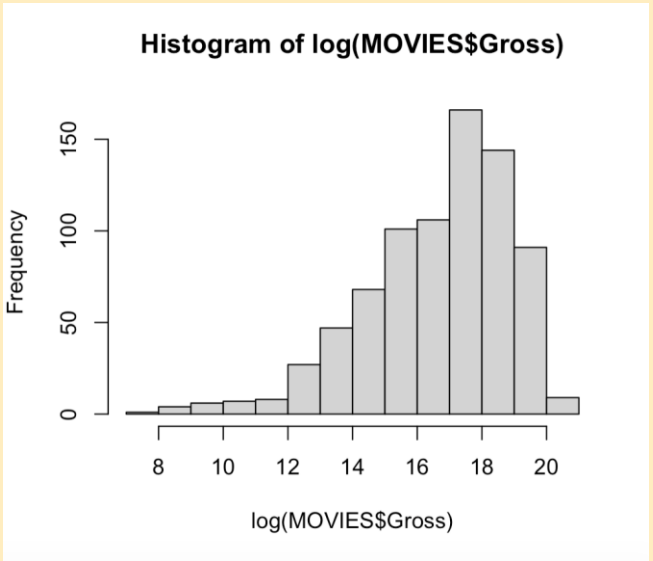
# Data Preparation

- Transform Gross to normalize
- Ensure all NA’s are removed
- Transform factor variables into indicator variables

# Data Preparation

Histogram of Gross variable before and after log transformation



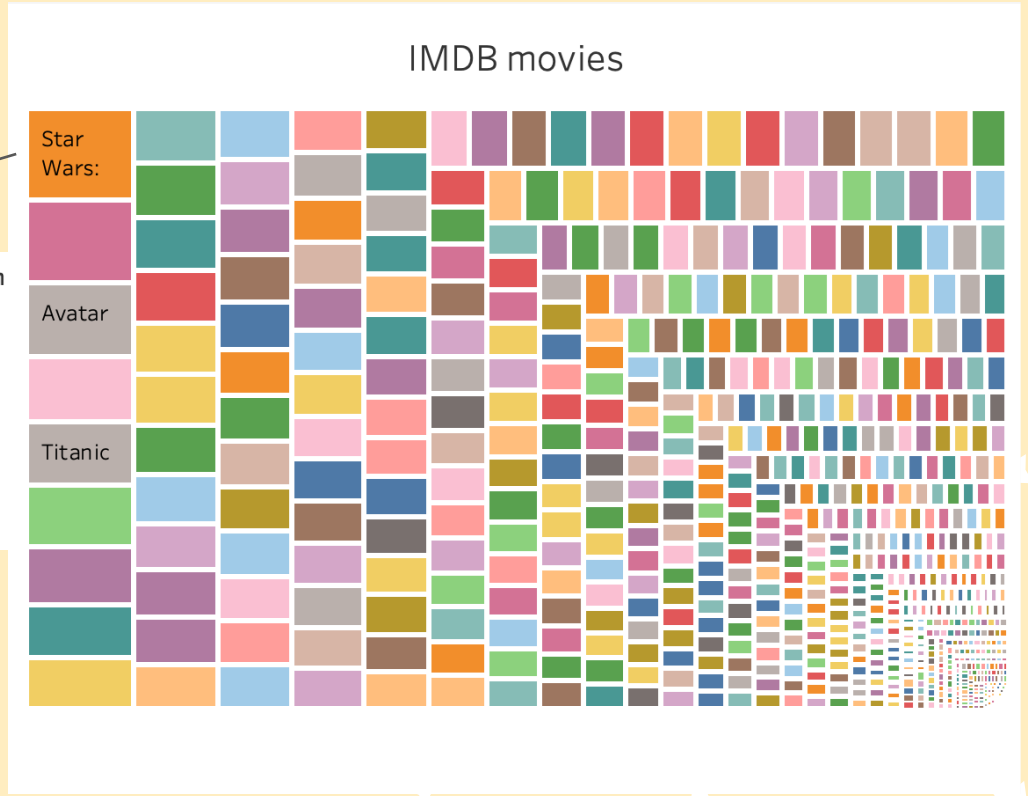



# Encoding Actors/Directors

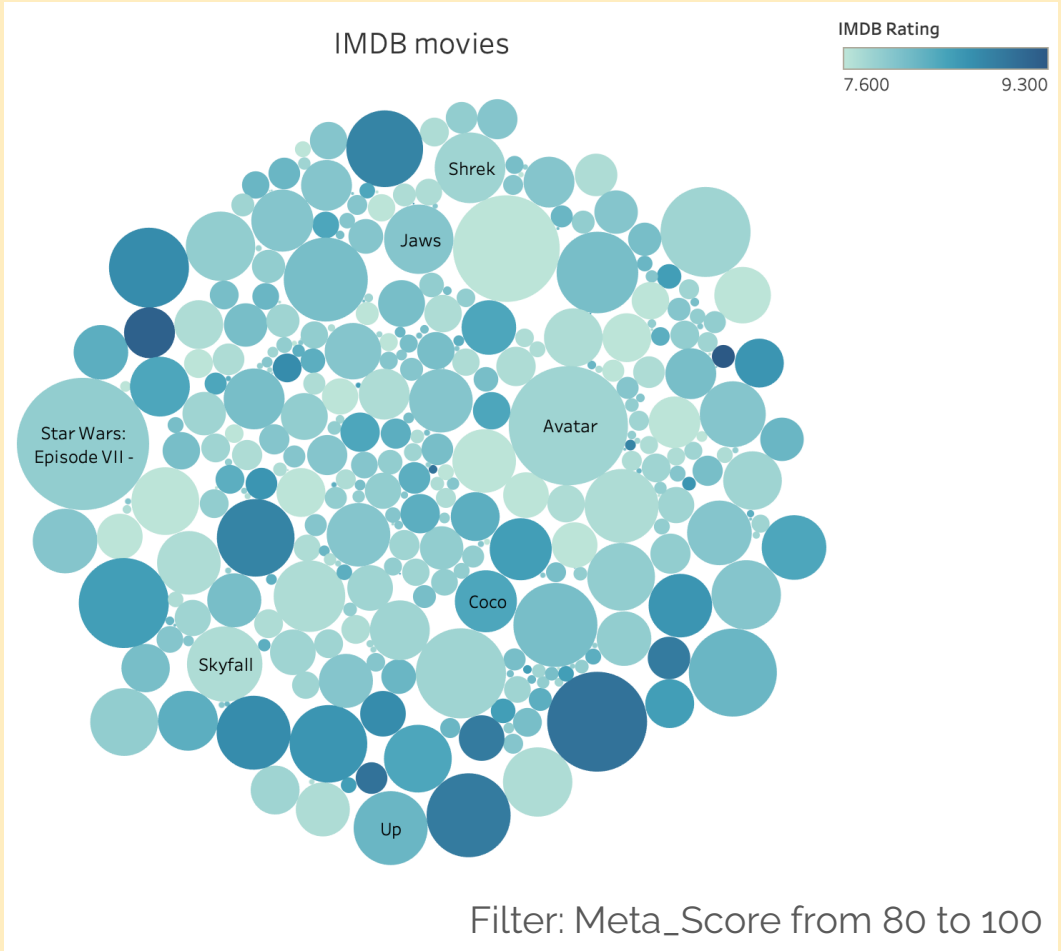
- ❏ Found Top 50 rankings online (ranker.com)
  - ❏ Published April 2021
  - ❏ General public's consensus

# IMDB Visualization

| | |
|---|---|
| Stars: | Daisy Ridley,John Boyega,Oscar Isaac,Domhnall Gleeson |
| Director: | J.J. Abrams |
| Genre: | Action, Adventure, Sci-Fi |
| Released Year: | 2015 |
| Runtime(mins): | 138 |
| Series Title: | Star Wars: Episode VII - The Force Awakens |
| Gross Income($): | 936,662,225 |
| IMDB Rating: | 7.900 |
| Meta score: | 80 |



# Meta_Score vs. IMDB_Rating



# Data Preparation

Director and Stars columns become binary variables, given presence of 1 or more

After

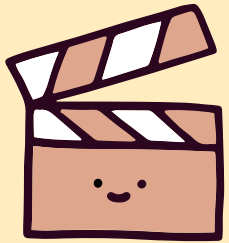
| Title | Release Year | Runtime | Genre | Top 50 Director | Top 50 Actor | Gross |
|---|---|---|---|---|---|---|
| Happiness | 1998 | 134 | Comedy, Drama | 0 | 0 | 14.84777 |
| Training Day | 2001 | 122 | Crime, Drama, Thriller | 0 | 1 | 18.15452 |
| Rushmore | 1998 | 93 | Comedy, Drama, Romance | 1 | 1 | 16.65489 |
| Abre los ojos | 1997 | 119 | Drama, Mystery, Sci-Fi | 0 | 0 | 12.81647 |

# Data Preparation

Also decided to create binary variables for genre

After

| Title | Release Year | Runtime | Action | Comedy | Drama | Top 50 Director | Top 50 Actor | Gross |
|---|---|---|---|---|---|---|---|---|
| Happiness | 1998 | 134 | 0 | 0 | 1 | 0 | 0 | 14.84777 |
| Training Day | 2001 | 122 | 0 | 1 | 0 | 0 | 1 | 18.15452 |
| Rushmore | 1998 | 93 | 0 | 0 | 1 | 1 | 1 | 16.65489 |
| Abre los ojos | 1997 | 119 | 1 | 0 | 1 | 0 | 0 | 12.81647 |

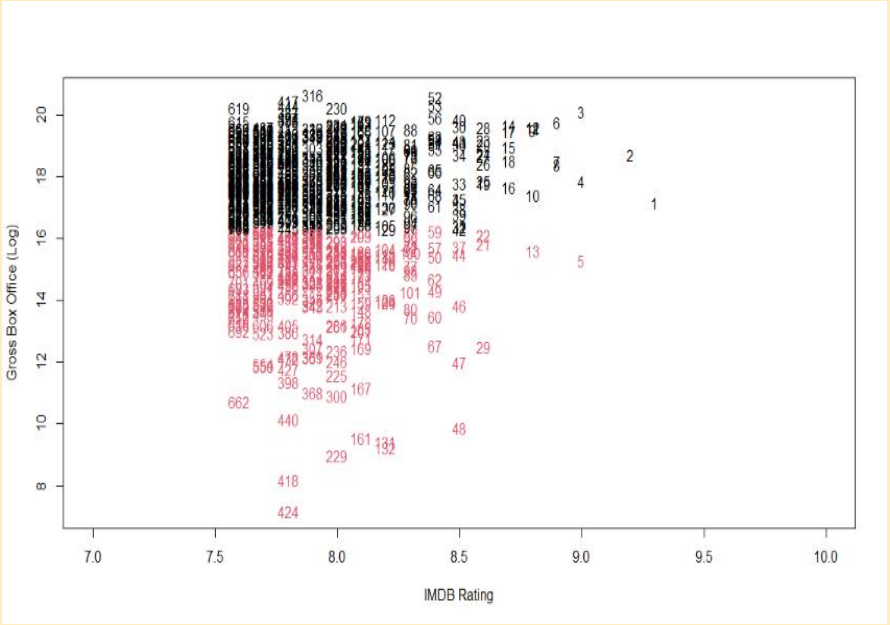


# Clustering

- ❏ IMDB v Gross
- ❏ K-means with all numerical variables
- ❏ DBSCAN

# Imdb v Gross

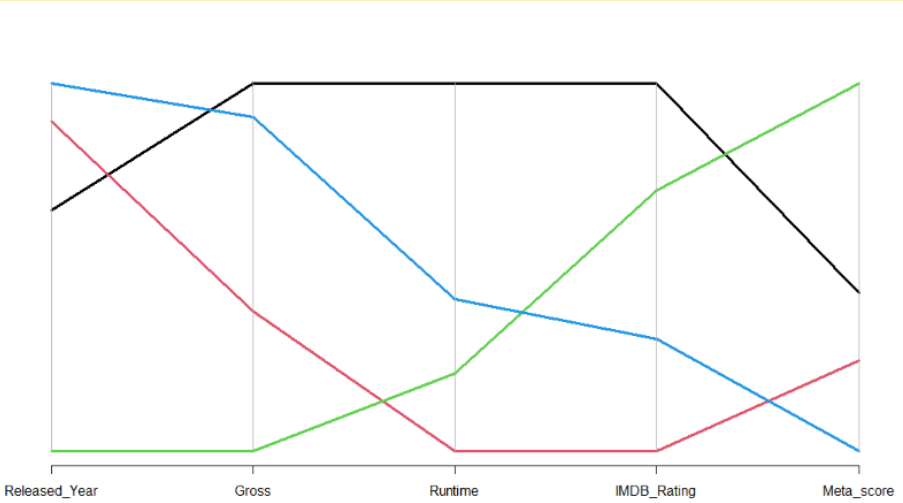
- Used Elbow method to determine centers = 2
- Index 316 - Star Wars Episode VII
- Index 1 - Shawshank Redemption
- From output, found that data was originally input, in order, by IMDb rating



# Clustering with all numerical variables

- Elbow method determines 4 centers. Sizes (89,250,108,267)
- between_SS / total_SS = 63.2% variance explained
- Centers are shown below:

| | Released_Year | Gross | Runtime | IMDB_Rating | Meta_score |
|---|---|---|---|---|---|
| 1 | 1991.045 | 17.36916 | 174.1236 | 8.102247 | 79.00000 |
| 2 | 2000.940 | 16.67104 | 101.0120 | 7.850000 | 76.56400 |
| 3 | 1964.278 | 16.24484 | 116.3241 | 8.028704 | 86.58333 |
| 4 | 2005.150 | 17.26614 | 131.1610 | 7.926592 | 73.28839 |



# DBSCAN

- Reverse-elbow method deemed eps = 10
- Produced 3 clusters(sizes 618, 10 , 5)

  and 81 noise points
- Cluster 2:

```
> head(MOVIES[which(db$cluster == 2),c(1:6,13)])
          Series_Title Released_Year Runtime                        Genre IMDB_Rating Meta_score     Gross
82          Braveheart          1995     178 Biography, Drama, History          8.3         68 18.14097
120               Heat          1995     170    Crime, Drama, Thriller          8.2         76 18.02670
121             Casino          1995     178              Crime, Drama          8.2         73 17.56356
252           Magnolia          1999     188                     Drama          8.0         77 16.92707
267                JFK          1991     189  Drama, History, Thriller          8.0         72 18.06978
269 Dances with Wolves          1990     181 Adventure, Drama, Western          8.0         72 19.03158
```

- Cluster 3:

```
> MOVIES[which(db$cluster == 3),c(1:6,13)]
        Series_Title Released_Year Runtime                  Genre IMDB_Rating Meta_score     Gross
2      The Godfather          1972     175            Crime, Drama          9.2        100 18.72054
129 The Great Escape          1963     172 Adventure, Drama, History          8.2         86 16.30872
294    La dolce vita          1960     174           Comedy, Drama          8.0         95 16.78675
370           Patton          1970     172    Biography, Drama, War          7.9         91 17.93779
483     My Fair Lady          1964     170  Drama, Family, Musical          7.8         95 18.09218
```

# DBSCAN's Outliers

- Cluster 0:
- Outlier qualities: high Runtime's, IMDB_Rating's, Meta_score's, Gross

```
> head(MOVIES[which(db$cluster == 0),c(1:6,13)])
                                                  Series_Title Released_Year Runtime                     Genre IMDB_Rating Meta_score    Gross
4                                     The Godfather: Part II          1974     202               Crime, Drama         9.0          90 17.86381
6                The Lord of the Rings: The Return of the King          2003     201  Action, Adventure, Drama         8.9          94 19.75000
8                                           Schindler's List          1993     195 Biography, Drama, History         8.9          94 18.38918
11 The Lord of the Rings: The Fellowship of the Ring          2001     178  Action, Adventure, Drama         8.8          92 19.56981
13                          Il buono, il brutto, il cattivo          1966     161                   Western         8.8          90 15.62380
14                     The Lord of the Rings: The Two Towers          2002     179  Action, Adventure, Drama         8.7          87 19.65193
```

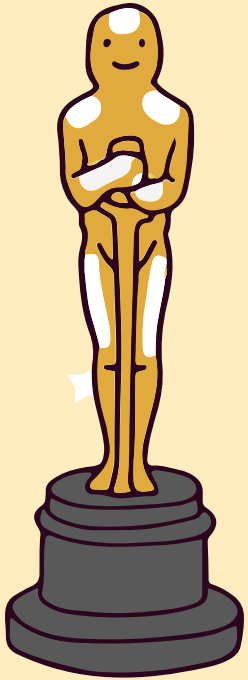
# Discretize Runtime + Gross


## Runtime

| Runtime | Runtime_Lengths |
|---|---|
| Less than 90 min | sub1.5(hours) |
| (90, 120] | 1.5 to 2 |
| (120, 150] | 2 to 2.5 |
| More than 150 min | 2.5 and above |

## Gross

| Gross Amount | Box_Office_Performance |
|---|---|
| less than 5 million(log(5e6)) | flop |
| 5 million to 25 million | underperforming |
| 25 million to 100 million | mod success |
| 100 million to 300 million | successful |
| more than 300 million | blockbuster |



# Box_Office_Performance

- Used as response variable for algorithms
- Important to verify significantly different levels
- Connecting letters report:

```
$Box_Office_Performance
     blockbuster       successful moderate success  underperforming            flop
             "a"              "b"              "c"              "d"             "e"
```

# Test / Train split

- Rule : 80/20 split randomly sampled

80%

## Train Dataset

Size :571

20%

## Test Dataset

Size :143

# Naïve Bayes

- Originally used Released Year, Runtime, IMDb_Rating, Meta_score, top50director, top50actor
- Accuracy of 31.53% but errors with running predictions
- Having both Meta_score & IMDb_Rating problematic, remove weaker predictor
- Pass vector of choices to fL and adjust (optimal is fL=1 and adjust=1)

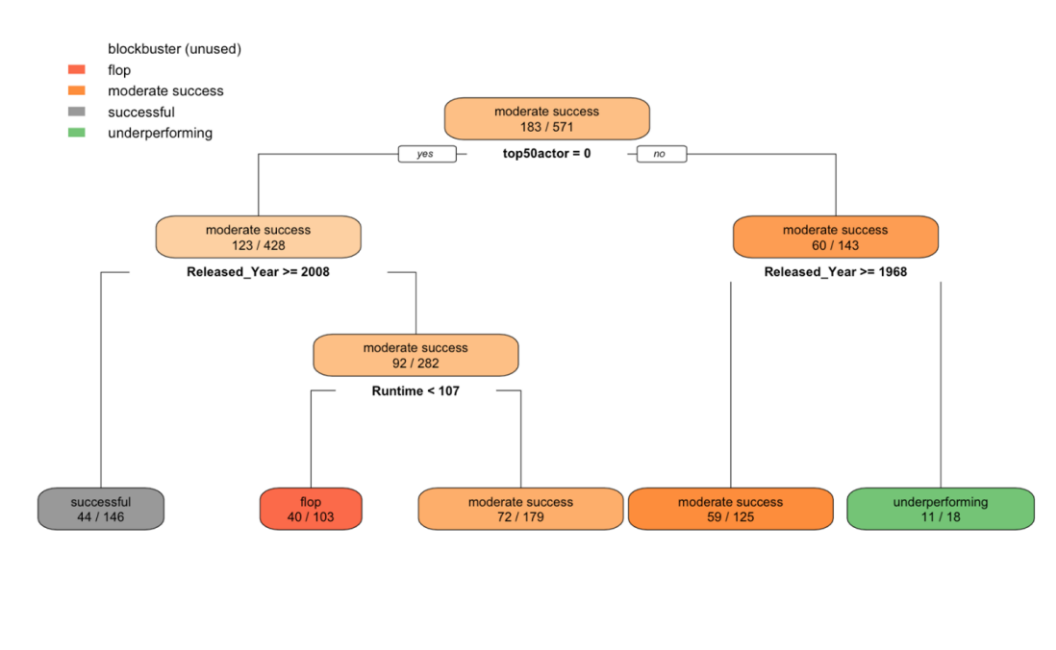
# Naïve Bayes

- Achieve 33.48% accuracy with 5 predictors, compared to predicting all to be in the majority class (30% accuracy)
- Confusion Matrix: Predicting Flop when actual Blockbuster, eliminates opportunity for large profits..predicting blockbuster when actual flop, guarantees losses

| | ybayes.pred | | | | | |
|---|---|---|---|---|---|---|
| ybayestest | blockbuster | flop | moderate success | successful | underperforming | |
| blockbuster | 2 | 4 | 2 | 2 | 0 | |
| flop | 0 | 23 | 9 | 0 | 1 | |
| moderate success | 0 | 18 | 12 | 1 | 1 | |
| successful | 0 | 17 | 12 | 1 | 1 | |
| underperforming | 0 | 25 | 11 | 0 | 1 | |

# DECISION TREE

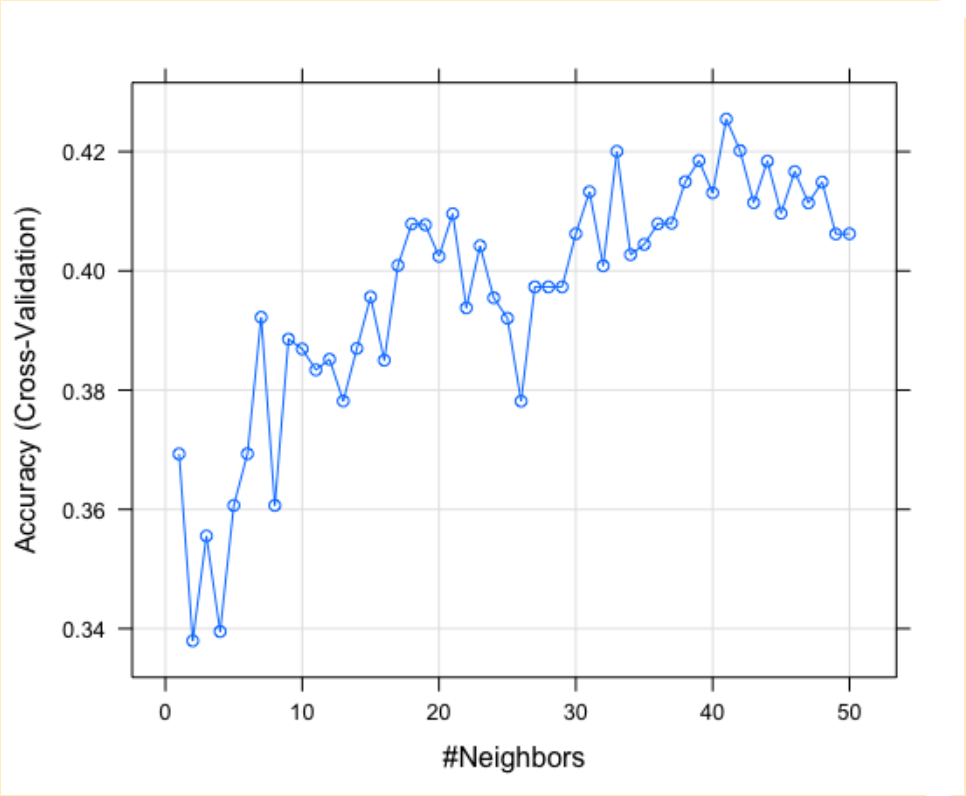
- Used Released Year, Runtime, top50actor, top50director,Meta_score,IMDb_Rating in model
- Did not include Genre because too many levels
- Model only includes Runtime + top50actor + Released Year
- Accuracy of 32.17% on test data 32.05% on train, slightly higher than random guessing
- Complexity parameter = 0.25



# K NEGREST NEIGHBOR

First step: choose the best number of neighbors k

Considering values of k 1-50



# K NEGREST NEIGHBOR

- On Training data:

| Accuracy | Kappa |
|---|---|
| 42.54% | 22.69% |

- On Testing data:

| Accuracy | Kappa |
|---|---|
| 46.85% | 31.41% |

# K NEGREST NEIGHBOR

## ❑ Confusion Matrix:

| | Blockbuster | Success | Moderate Success | Underperforming | Flop |
|---|---|---|---|---|---|
| Blockbuster | 0 | 0 | 0 | 0 | 0 |
| Success | 8 | 19 | 7 | 2 | 4 |
| Moderate Success | 2 | 11 | 20 | 22 | 7 |
| Underperforming | 0 | 1 | 3 | 6 | 2 |
| Flop | 0 | 0 | 2 | 7 | 20 |

# K NEGREST NEIGHBOR

- In practice:
  - Is Euclidean distance the correct measure?
  - Assumes all variables are of equal importance
  - Distance vastly affected by transforming Y

# RANDOM FOREST

- Accuracy on Training set: 0.4344748
- Parameters:
  - Mtry = 2
  - Ntree = 500

| mtry | Accuracy | Kappa |
|---|---|---|
| 2 | 0.4344748 | 0.2056142 |
| 20 | 0.4325983 | 0.2390672 |
| 38 | 0.4134763 | 0.2159663 |

# RANDOM FOREST

- Confusion matrix:

| | Blockbuster | Success | Moderate Success | Underperforming | Flop |
|---|---|---|---|---|---|
| Blockbuster | 0 | 7 | 3 | 0 | 0 |
| Success | 0 | 7 | 24 | 0 | 0 |
| Moderate Success | 0 | 2 | 26 | 0 | 4 |
| Underperforming | 0 | 3 | 19 | 1 | 14 |
| Flop | 0 | 2 | 20 | 0 | 11 |

# SVM Classifier(Linear)

- On Training data:

```
> svmFit$results
  C  Accuracy     Kappa AccuracySD    KappaSD
1 1 0.4325772 0.2398827 0.05786573 0.07969059
```

- On Testing data:

Accuracy: 0.4125874

# SVM Classifier(Linear)

## ❑ Confusion Matrix:

| | Blockbuster | Flop | Moderate Success | Success | Underperforming |
|---|---|---|---|---|---|
| Blockbuster | 4 | 0 | 1 | 5 | 0 |
| Flop | 0 | 18 | 10 | 2 | 3 |
| Moderate Success | 0 | 8 | 20 | 2 | 2 |
| Success | 2 | 1 | 15 | 12 | 1 |
| Underperforming | 0 | 19 | 10 | 3 | 5 |

# SVM Classifier(Radial)

- On Training data:

| C | Accuracy | Kappa |
|---|---|---|
| 0.25 | 0.3273333 | 0.04026568 |
| 0.50 | 0.3431563 | 0.08190693 |
| 1.00 | 0.3444113 | 0.08914005 |

```
> svmFit2$bestTune
         sigma C
3 0.003082603 1
```

- On Testing data:

Accuracy: 0.3286713

# BOOSTED TREE

- On Training data: *Accuracy: 0.4432404 Kappa: 0.2523365*

```
> xgbFit$results
   eta max_depth gamma colsample_bytree min_child_weight subsample nrounds
1  0.3         1     0              0.6                1      0.50      50
4  0.3         1     0              0.6                1      0.75      50
7  0.3         1     0              0.6                1      1.00      50
```

```
> xgbFit$finalModel$tuneValue
   nrounds max_depth eta gamma colsample_bytree min_child_weight subsample
58      50         1 0.4     0              0.6                1      0.75
```

- On Testing data:
  Accuracy: 0.4195804

# ARTIFICIAL NEURAL NETWORK

❑ On Training data:

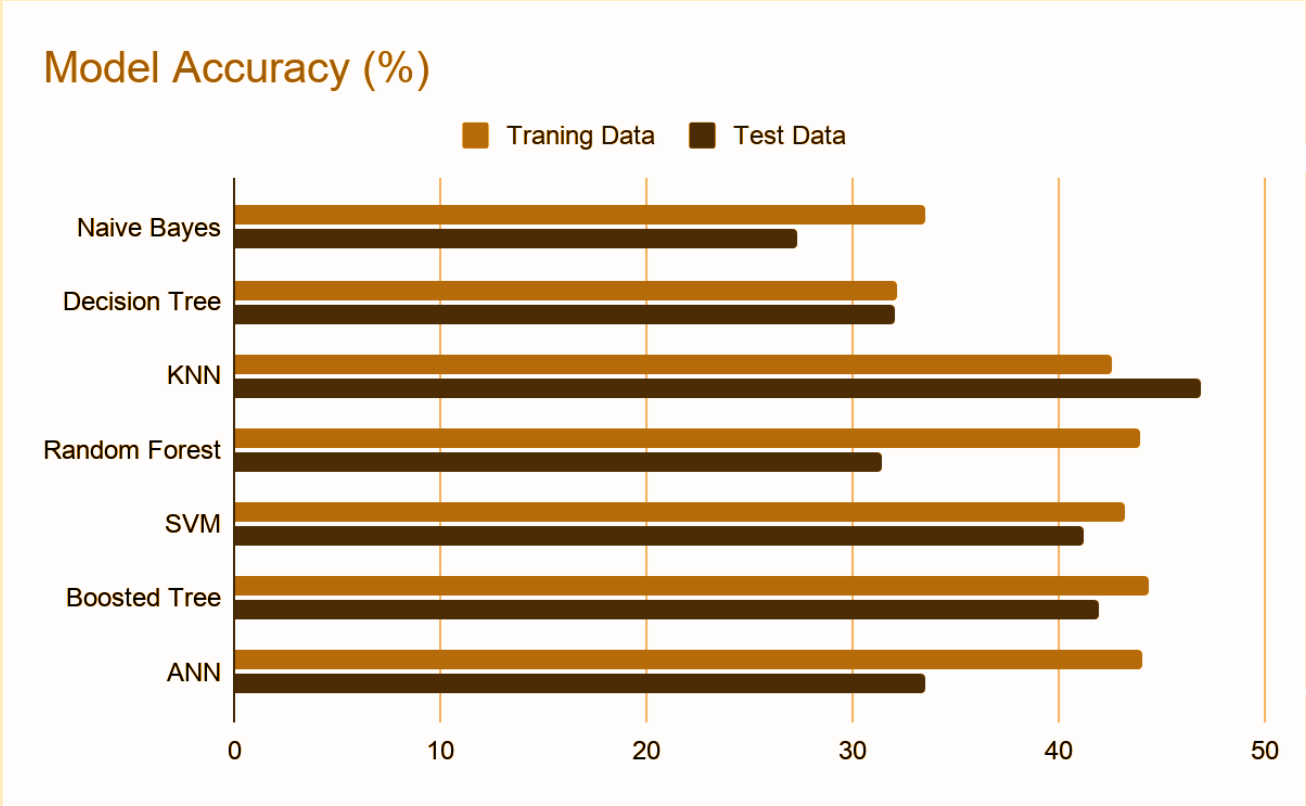
```
> nnetFit$results
   size decay  Accuracy      Kappa  AccuracySD    KappaSD
23    9 1e-03 0.3537292 0.10048852 0.052208275 0.07965439
24    9 1e-02 0.4408813 0.23392421 0.067925710 0.09798944
25    9 1e-01 0.4060669 0.19806593 0.045173581 0.06339580
```

❑ On Testing data:

Accuracy: 0.3356643

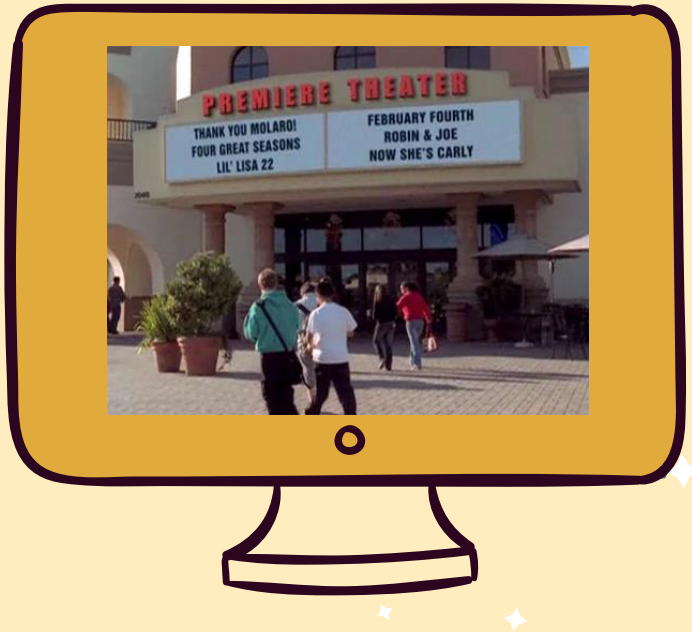
```
> nnetFit$bestTune
   size decay
24    9  0.01
```

# Model Accuracy Comparison



# ALGORITHMS' SUMMARY

- ❏ Boosted Tree has the best performance on training dataset
- ❏ KNN has the best performance on testing dataset
- ❏ Parameter Tuning in Machine learning
- Understanding Bias-Variance Tradeoff
- Control Overfitting
- Faster training performance

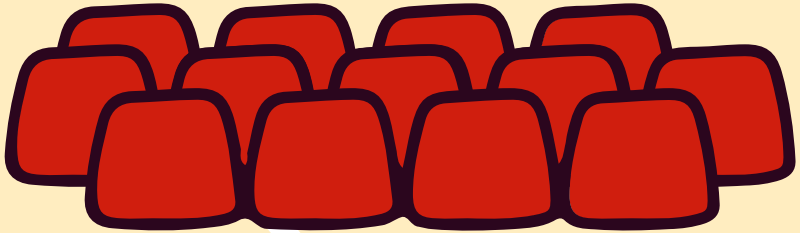


# SHORTCOMINGS WITH DATASET

- Data does not include Budget
- Data includes release year, but not month

# Thanks!

Questions?



CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

# Alternative vectors