# 10-301/601: Introduction to Machine Learning Lecture 26 – Random Forests

Henry Chai

8/2/22

# Front Matter

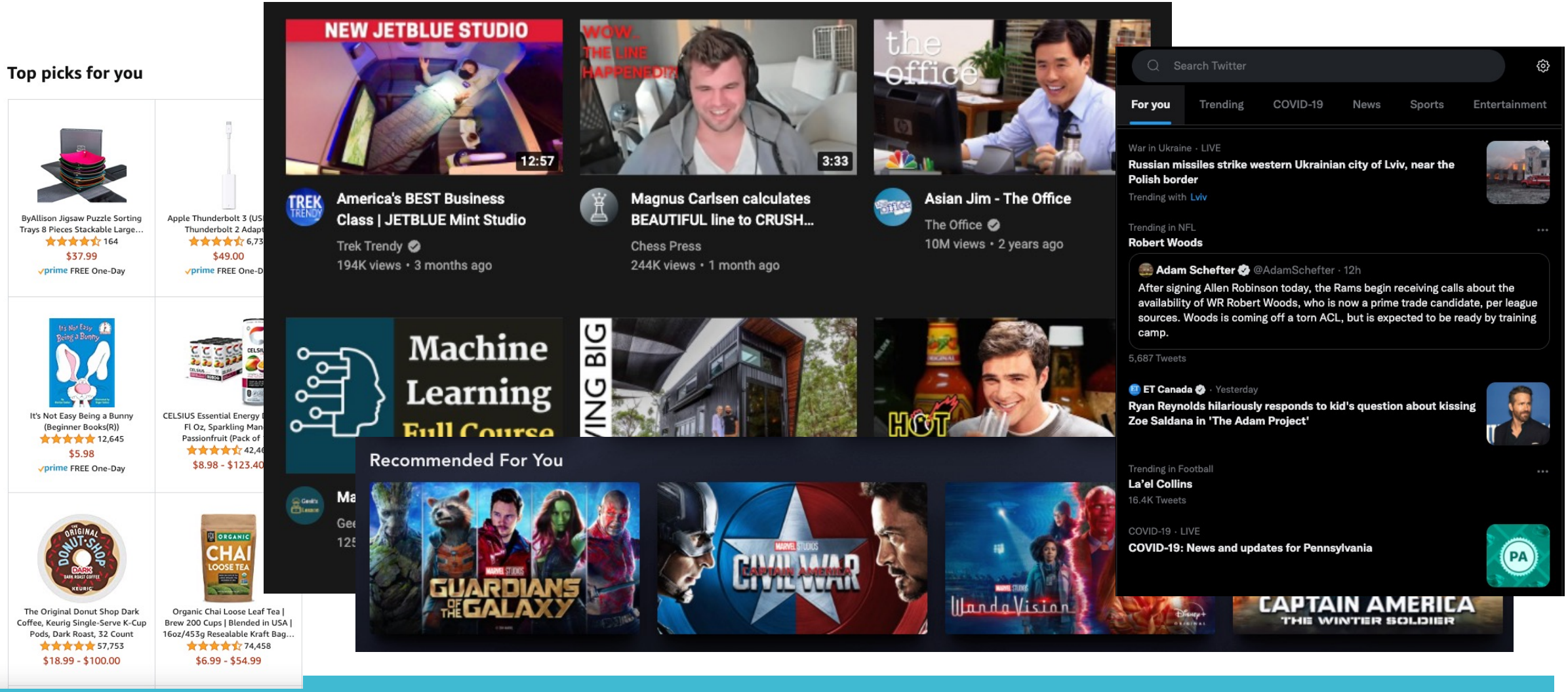
- Announcements
  - HW8 released 7/27, due 8/3 (tomorrow) at 1 PM
    - Please be mindful of your grace day usage (see the course syllabus for the policy)
  - Exam 3 on 8/12, one week from Friday!

# Exam 3 Logistics

- Friday, 8/12 from 4:00 – 5:20 PM in PH 100
- Closed book/notes
  - 1-page cheatsheet allowed, both back and front; can be typeset or handwritten
- Covered material: Lectures 17 – 27
  - Graphical Models: Bayesian Networks & HMMs
  - Reinforcement Learning
  - Unsupervised Learning: Clustering & Dimensionality Reduction
  - Ensemble Methods: Random Forests & Boosting
  - Exam 1 & 2 content may be referenced but will not be the primary focus of any question

# Exam 3 Preparation

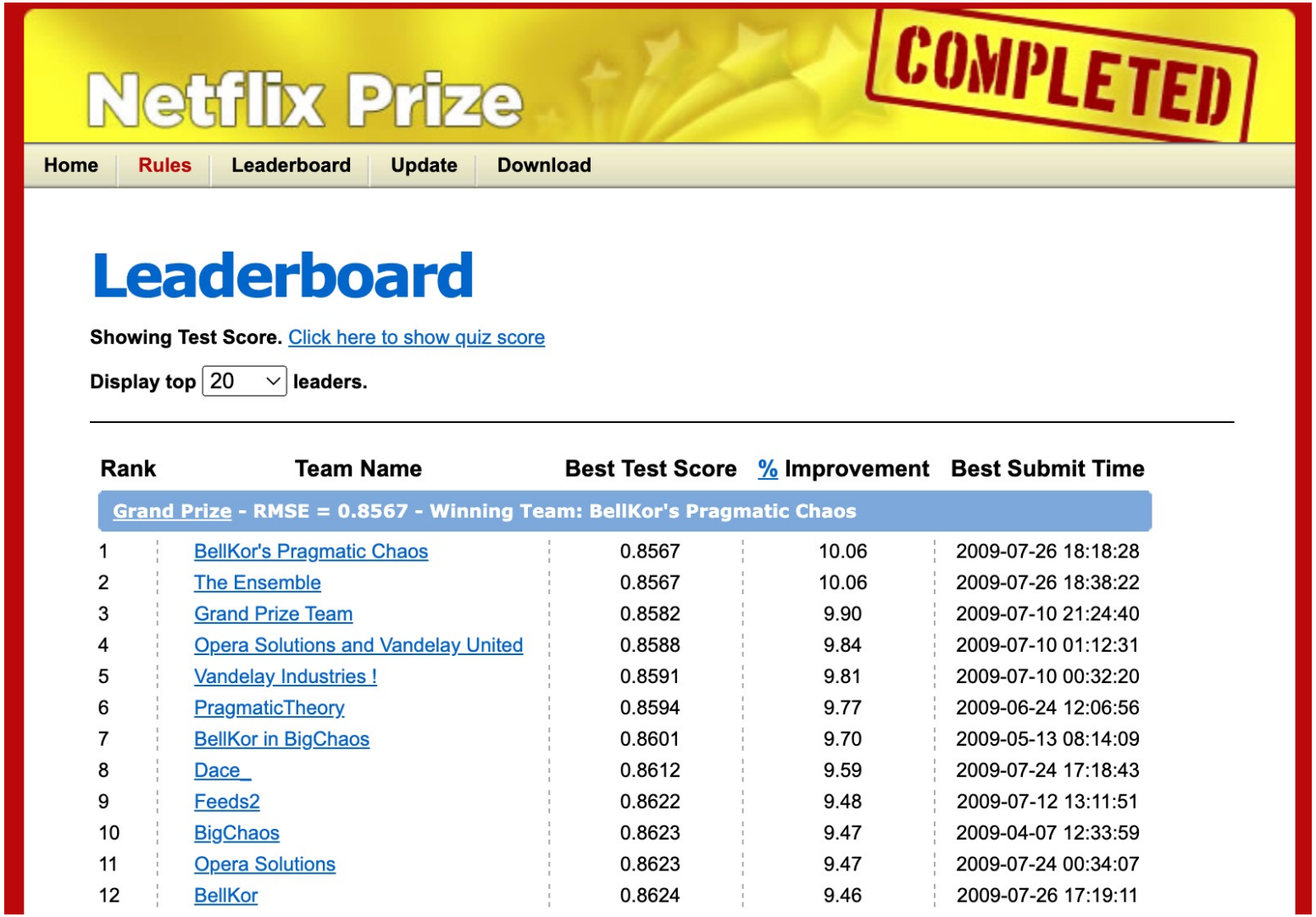
- Review exam practice problems, posted to the course website (under Recitations)
- Attend the exam review recitation (Thursday, 8/11)
- Attend the exam review lecture (Tuesday, 8/9)
- Review this year's homework problems
- Consider whether you understand the "Key Takeaways" for each lecture / section
- Write your cheat sheet




# Personalized Recommendations

# The Netflix Prize

# The Netflix Prize




Source: https://web.archive.org/web/20090926213457/http://www.netflixprize.com/leaderboard

# The Wisdom of Crowds

- In 1906, Francis Galton asked ~800 people at a farmer's fair to guess the weight of a cow, including "experts"
  - Actual weight: 1198 lbs
  - Mean guess: 1197 lbs
  - Mean guess was more accurate than any single guess, even the experts



Source: https://galton.org/essays/1900-1911/galton-1907-vox-populi.pdf

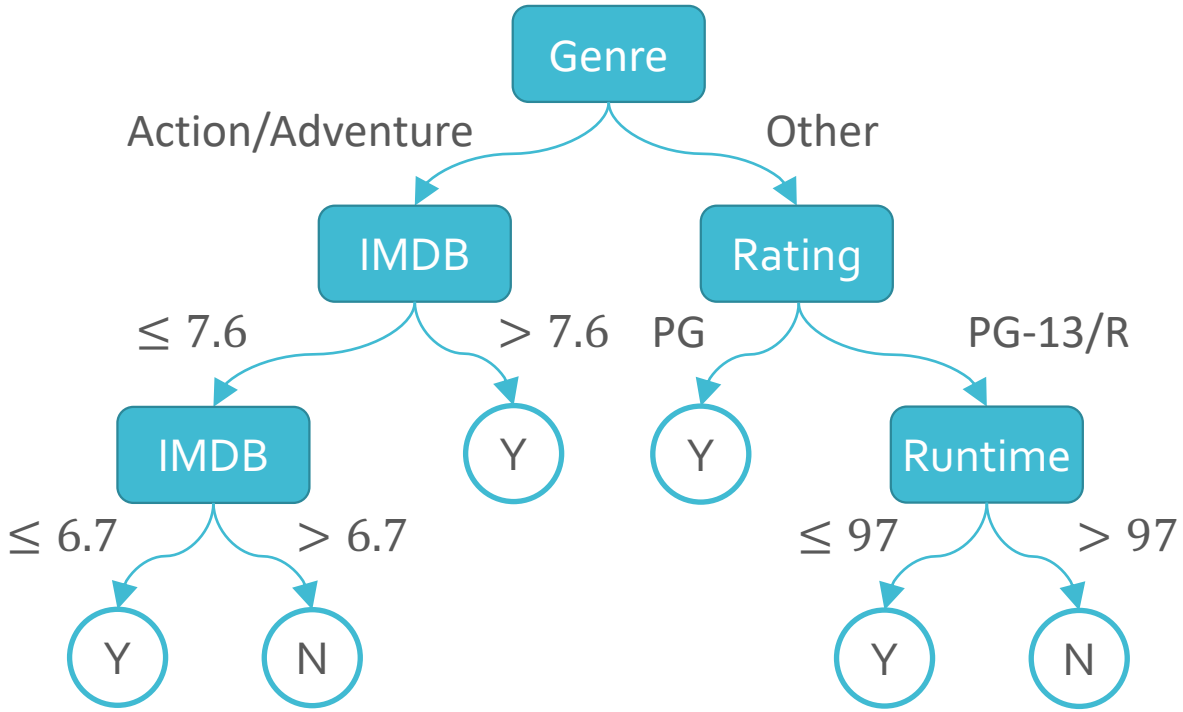
| MovieID | Runtime | Genre | Budget | Year | IMDB | Rating | Liked? |
|---|---|---|---|---|---|---|---|
| 1 | 124 | Action | 18M | 1980 | 8.7 | PG | Y |
| 2 | 105 | Action | 30M | 1984 | 7.8 | PG | Y |
| 3 | 103 | Comedy | 6M | 1986 | 7.8 | PG-13 | N |
| 4 | 98 | Adventure | 16M | 1987 | 8.1 | PG | Y |
| 5 | 128 | Comedy | 16.4M | 1989 | 8.1 | PG | Y |
| 6 | 120 | Comedy | 11M | 1992 | 7.6 | R | N |
| 7 | 120 | Drama | 14.5M | 1996 | 6.7 | PG-13 | N |
| 8 | 136 | Action | 115M | 1999 | 6.5 | PG | Y |
| 9 | 90 | Action | 90M | 2001 | 6.6 | PG-13 | Y |
| 10 | 161 | Adventure | 100M | 2002 | 7.4 | PG | N |
| 11 | 201 | Action | 94M | 2003 | 8.9 | PG-13 | Y |
| 12 | 94 | Comedy | 26M | 2004 | 7.2 | PG-13 | Y |
| 13 | 157 | Biography | 100M | 2007 | 7.8 | R | N |
| 14 | 128 | Action | 110M | 2007 | 7.1 | PG-13 | N |
| 15 | 107 | Drama | 39M | 2009 | 7.1 | PG-13 | N |
| 16 | 158 | Drama | 61M | 2012 | 7.6 | PG-13 | N |
| 17 | 169 | Adventure | 165M | 2014 | 8.6 | PG-13 | Y |
| 18 | 100 | Biography | 9M | 2016 | 6.7 | R | N |
| 19 | 130 | Action | 180M | 2017 | 7.9 | PG-13 | Y |
| 20 | 141 | Action | 275M | 2019 | 6.5 | PG-13 | Y |

# Movie Recommendations



Source: https://www.kaggle.com/datasets/danielgrijalvas/movies

| MovieID | Runtime | Genre | Budget | Year | IMDB | Rating | Liked? |
|---|---|---|---|---|---|---|---|
| 1 | 124 | Action | 18M | 1980 | 8.7 | PG | Y |
| 2 | 105 | Action | 30M | 1984 | 7.8 | PG | Y |
| 3 | 103 | Comedy | 6M | 1986 | 7.8 | PG-13 | N |
| 4 | 98 | Adventure | 16M | 1987 | 8.1 | PG | Y |
| 5 | 128 | Comedy | 16.4M | 1989 | 8.1 | PG | Y |
| 6 | 120 | Comedy | 11M | 1992 | 7.6 | R | N |
| 7 | 120 | Drama | 14.5M | 1996 | 6.7 | PG-13 | N |
| 8 | 136 | Action | 115M | 1999 | 6.5 | PG | Y |
| 9 | 90 | Action | 90M | 2001 | 6.6 | PG-13 | Y |
| 10 | 161 | Adventure | 100M | 2002 | 7.4 | PG | N |
| 11 | 201 | Action | 94M | 2003 | 8.9 | PG-13 | Y |
| 12 | 94 | Comedy | 26M | 2004 | 7.2 | PG-13 | Y |
| 13 | 157 | Biography | 100M | 2007 | 7.8 | R | N |
| 14 | 128 | Action | 110M | 2007 | 7.1 | PG-13 | N |
| 15 | 107 | Drama | 39M | 2009 | 7.1 | PG-13 | N |
| 16 | 158 | Drama | 61M | 2012 | 7.6 | PG-13 | N |
| 17 | 169 | Adventure | 165M | 2014 | 8.6 | PG-13 | Y |
| 18 | 100 | Biography | 9M | 2016 | 6.7 | R | N |
| 19 | 130 | Action | 180M | 2017 | 7.9 | PG-13 | Y |
| 20 | 141 | Action | 275M | 2019 | 6.5 | PG-13 | Y |




# Decision Trees

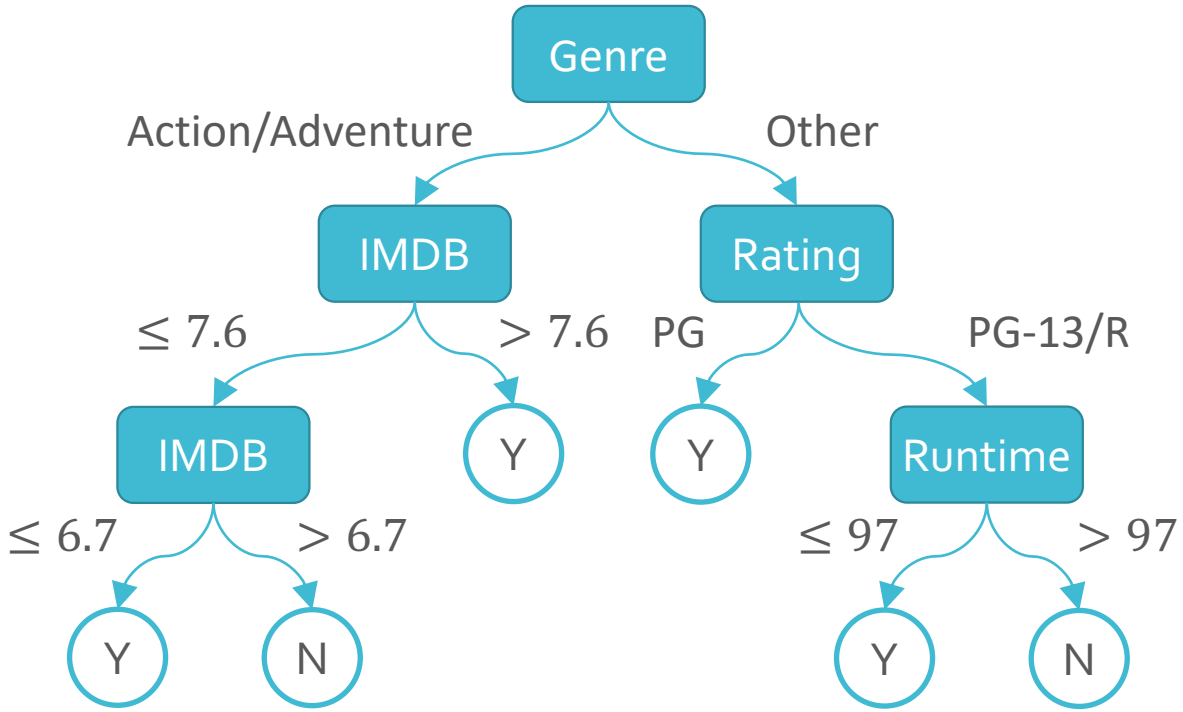


Source: https://www.kaggle.com/datasets/danielgrijalvas/movies

# Recall: Decision Tree Pros & Cons

- Pros
  - Interpretable
  - Efficient (computational cost and storage)
  - Can be used for classification and regression tasks
  - Compatible with categorical and real-valued features
- Cons
  - Learned greedily: each split only considers the immediate impact on the splitting criterion
    - Not guaranteed to find the smallest (fewest number of splits) tree that achieves a training error rate of 0.
  - Prone to overfit
  - High variance

| MovieID | Runtime | Genre | Budget | Year | IMDB | Rating | Liked? |
|---|---|---|---|---|---|---|---|
| 1 | 124 | Action | 18M | 1980 | 8.7 | PG | Y |
| 2 | 105 | Action | 30M | 1984 | 7.8 | PG | Y |
| 3 | 103 | Comedy | 6M | 1986 | 7.8 | PG-13 | N |
| 4 | 98 | Adventure | 16M | 1987 | 8.1 | PG | Y |
| 5 | 128 | Comedy | 16.4M | 1989 | 8.1 | PG | Y |
| 6 | 120 | Comedy | 11M | 1992 | 7.6 | R | N |
| 7 | 120 | Drama | 14.5M | 1996 | 6.7 | PG-13 | N |
| 8 | 136 | Action | 115M | 1999 | 6.5 | PG | Y |
| 9 | 90 | Action | 90M | 2001 | 6.6 | PG-13 | Y |
| 10 | 161 | Adventure | 100M | 2002 | 7.4 | PG | N |
| 11 | 201 | Action | 94M | 2003 | 8.9 | PG-13 | Y |
| 12 | 94 | Comedy | 26M | 2004 | 7.2 | PG-13 | Y |
| 13 | 157 | Biography | 100M | 2007 | 7.8 | R | N |
| 14 | 128 | Action | 110M | 2007 | 7.1 | PG-13 | N |
| 15 | 107 | Drama | 39M | 2009 | 7.1 | PG-13 | N |
| 16 | 158 | Drama | 61M | 2012 | 7.6 | PG-13 | N |
| 17 | 169 | Adventure | 165M | 2014 | 8.6 | PG-13 | Y |
| 18 | 100 | Biography | 9M | 2016 | 6.7 | R | N |
| 19 | 130 | Action | 180M | 2017 | 7.9 | PG-13 | Y |
| 20 | 141 | Action | 275M | 2019 | 6.5 | PG-13 | Y |

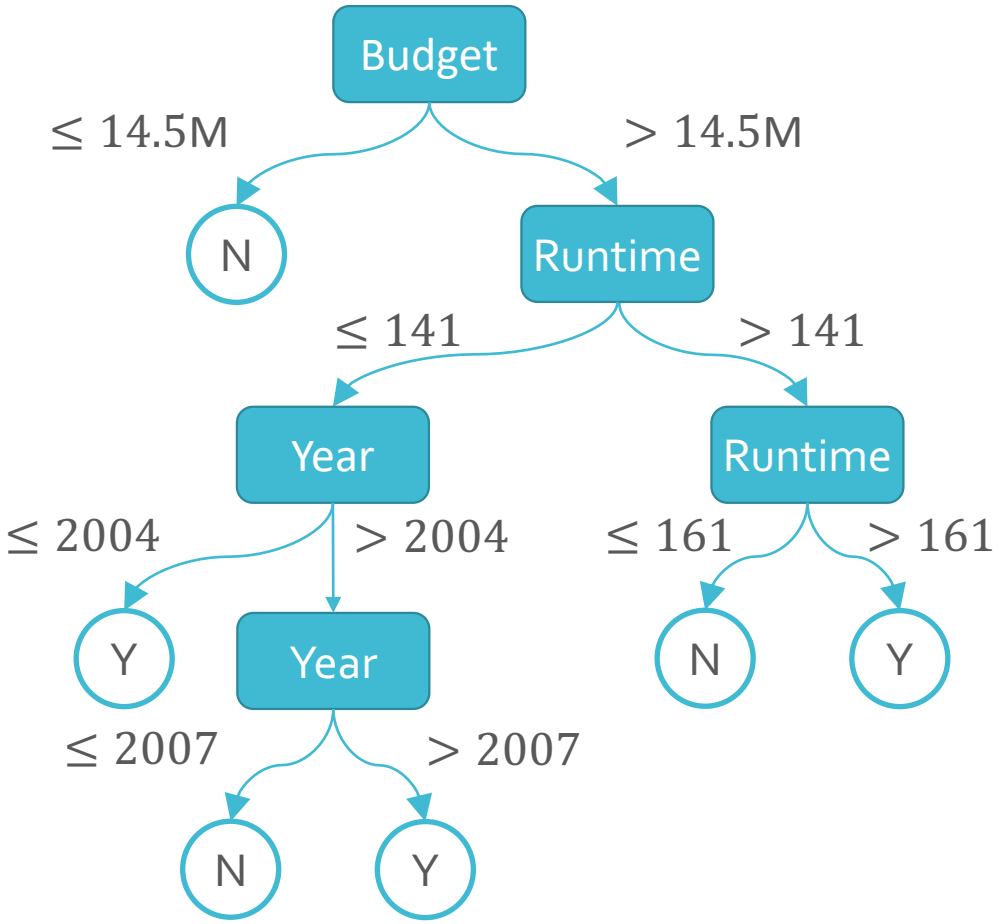



# Decision Trees



Source: https://www.kaggle.com/datasets/danielgrijalvas/movies

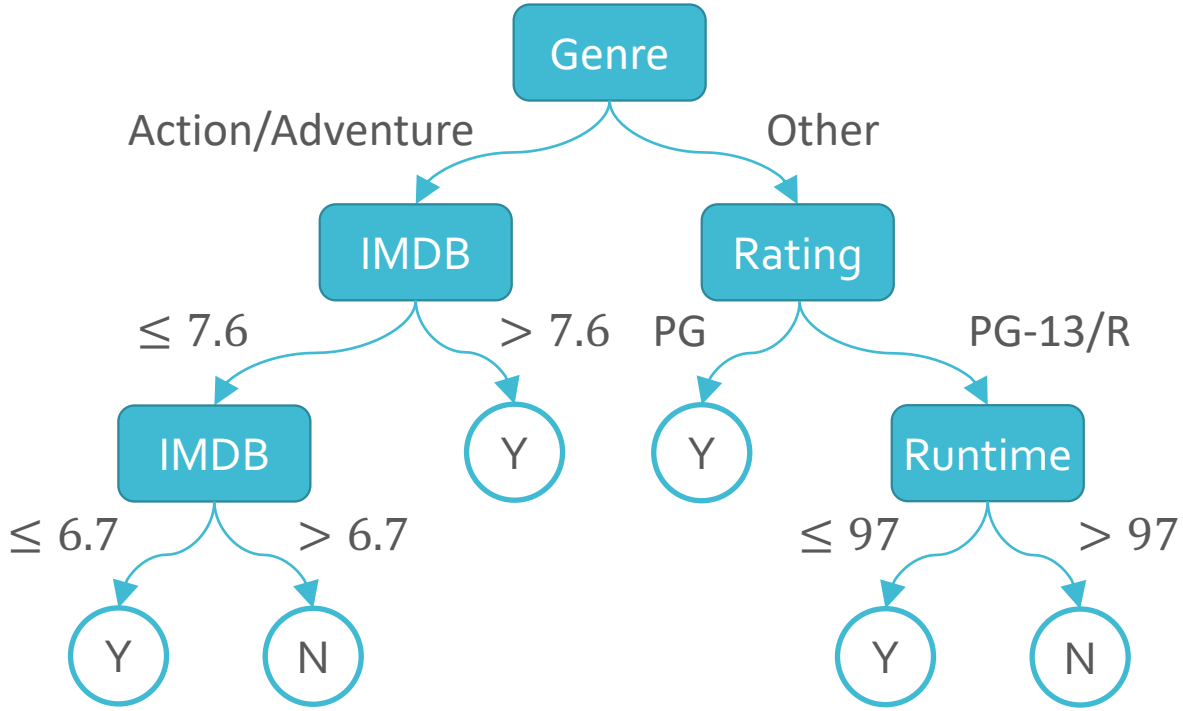
| MovieID | Runtime | Genre | Budget | Year | IMDB | Rating | Liked? |
|---|---|---|---|---|---|---|---|
| 1 | 124 | Action | 18M | 1980 | 8.7 | PG | Y |
| 2 | 105 | Action | 30M | 1984 | 7.8 | PG | Y |
| 3 | 103 | Comedy | 6M | 1986 | 7.8 | PG-13 | N |
| 4 | 98 | Adventure | 16M | 1987 | 8.1 | PG | Y |
| 5 | 128 | Comedy | 16.4M | 1989 | 8.1 | PG | Y |
| 6 | 120 | Comedy | 11M | 1992 | 7.6 | R | N |
| 7 | 120 | Drama | 14.5M | 1996 | 6.7 | PG-13 | N |
| 8 | 136 | Action | 115M | 1999 | 6.5 | PG | Y |
| 9 | 90 | Action | 90M | 2001 | 6.6 | PG-13 | Y |
| 10 | 161 | Adventure | 100M | 2002 | 7.4 | PG | N |
| 11 | 201 | Action | 94M | 2003 | 8.9 | PG-13 | Y |
| 12 | 94 | Comedy | 26M | 2004 | 7.2 | PG-13 | Y |
| 13 | 157 | Biography | 100M | 2007 | 7.8 | R | N |
| 14 | 128 | Action | 110M | 2007 | 7.1 | PG-13 | N |
| 15 | 107 | Drama | 39M | 2009 | 7.1 | PG-13 | N |
| 16 | 158 | Drama | 61M | 2012 | 7.6 | PG-13 | ***Y*** |
| 17 | 169 | Adventure | 165M | 2014 | 8.6 | PG-13 | Y |
| 18 | 100 | Biography | 9M | 2016 | 6.7 | R | N |
| 19 | 130 | Action | 180M | 2017 | 7.9 | PG-13 | Y |
| 20 | 141 | Action | 275M | 2019 | 6.5 | PG-13 | Y |

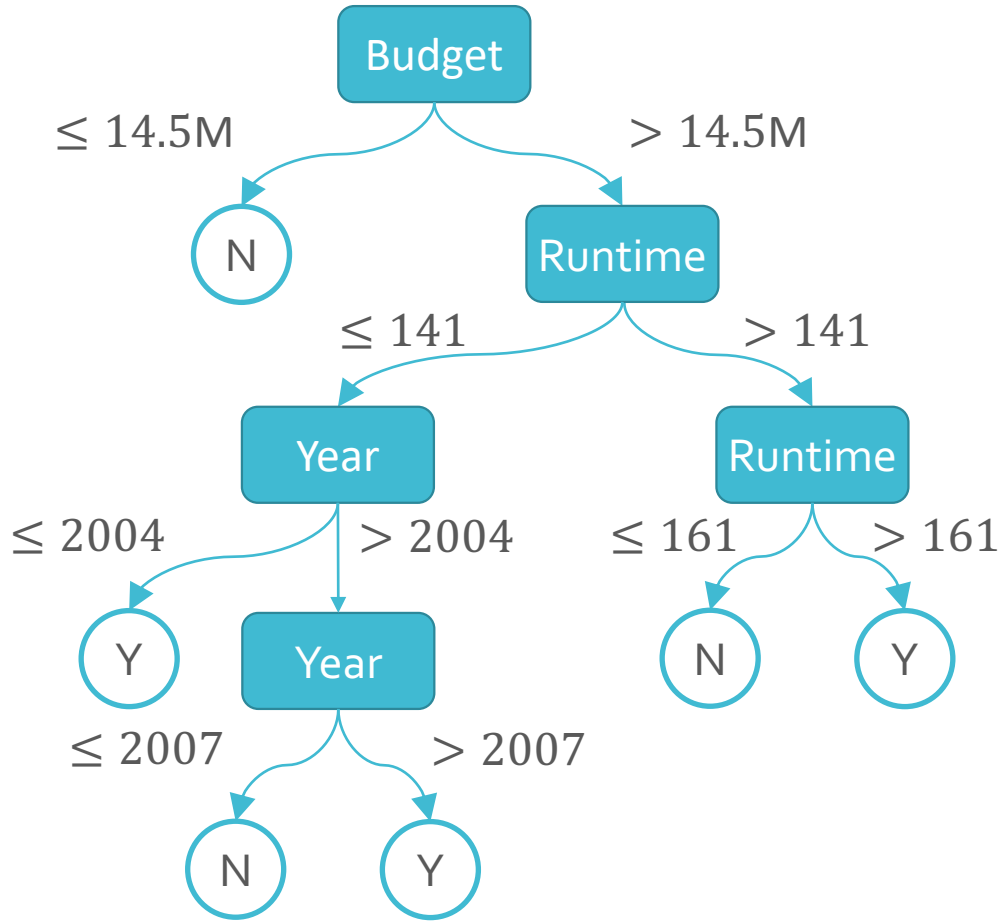



# Decision Trees



Source: https://www.kaggle.com/datasets/danielgrijalvas/movies




# Decision Trees

# Decision Trees: Pros & Cons

- Pros
  - Interpretable
  - Efficient (computational cost and storage)
  - Can be used for classification and regression tasks
  - Compatible with categorical and real-valued features
- Cons
  - Learned greedily: each split only considers the immediate impact on the splitting criterion
    - Not guaranteed to find the smallest (fewest number of splits) tree that achieves a training error rate of 0.
  - Prone to overfit
  - High variance
    - Can be addressed via ensembles → random forests

# Random Forests

- Combines the prediction of many diverse decision trees to reduce their variability
- If $B$ independent random variables $x^{(1)}, x^{(2)}, \ldots, x^{(B)}$ all have variance $\sigma^2$, then the variance of $\frac{1}{B}\sum_{b=1}^{B} x^{(b)}$ is $\frac{\sigma^2}{B}$
- Random forests = bagging + split-feature randomization

  = **b**ootstrap **agg**regat**ing** + split-feature randomization

# Random Forests

- Combines the prediction of many diverse decision trees to reduce their variability
- If $B$ independent random variables $x^{(1)}, x^{(2)}, \ldots, x^{(B)}$ all have variance $\sigma^2$, then the variance of $\frac{1}{B}\sum_{b=1}^{B} x^{(b)}$ is $\frac{\sigma^2}{B}$
- Random forests = bagging + split-feature randomization

  = bootstrap aggregating + split-feature randomization

# Aggregating

- How can we combine multiple decision trees, $\{t_1, t_2, \ldots, t_B\}$, to arrive at a single prediction?
- Regression - average the predictions:

$$\bar{t}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} t_b(\boldsymbol{x})$$

- Classification - plurality (or majority) vote; for binary labels encoded as $\{-1, +1\}$:

$$\bar{t}(\boldsymbol{x}) = \text{sign}\left(\frac{1}{B}\sum_{b=1}^{B} t_b(\boldsymbol{x})\right)$$

# Random Forests

- Combines the prediction of many diverse decision trees to reduce their variability
- If $B$ independent random variables $x^{(1)}, x^{(2)}, \ldots, x^{(B)}$ all have variance $\sigma^2$, then the variance of $\frac{1}{B}\sum_{b=1}^{B} x^{(b)}$ is $\frac{\sigma^2}{B}$
- Random forests = bagging + split-feature randomization

  = bootstrap aggregating + split-feature randomization

# Bootstrapping

- Insight: one way of generating different decision trees is by changing the training data set
- Issue: often, we only have one fixed set of training data
- Idea: resample the data multiple times ***with replacement***

| MovieID | ⋯ |
|---|---|
| 1 | ⋯ |
| 2 | ⋯ |
| 3 | ⋯ |
| ⋮ | ⋮ |
| 19 | ⋯ |
| 20 | ⋯ |

Training data

| MovieID | ⋯ |
|---|---|
| 1 | ⋯ |
| 1 | ⋯ |
| 1 | ⋯ |
| ⋮ | ⋮ |
| 14 | ⋯ |
| 19 | ⋯ |

Bootstrapped Sample 1

| MovieID | ⋯ |
|---|---|
| 4 | ⋯ |
| 4 | ⋯ |
| 5 | ⋯ |
| ⋮ | ⋮ |
| 16 | ⋯ |
| 16 | ⋯ |

Bootstrapped Sample 2

⋯

⋯

# Bootstrapping

- Idea: resample the data multiple times ***with replacement***
  - Each bootstrapped sample has the same number of data points as the original data set
  - Duplicated points cause different decision trees to focus on different parts of the input space

| MovieID | ⋯ |
|---|---|
| 1 | ⋯ |
| 2 | ⋯ |
| 3 | ⋯ |
| ⋮ | ⋮ |
| 19 | ⋯ |
| 20 | ⋯ |

Training data

| MovieID | ⋯ |
|---|---|
| 1 | ⋯ |
| 1 | ⋯ |
| 1 | ⋯ |
| ⋮ | ⋮ |
| 14 | ⋯ |
| 19 | ⋯ |

Bootstrapped Sample 1

| MovieID | ⋯ |
|---|---|
| 4 | ⋯ |
| 4 | ⋯ |
| 5 | ⋯ |
| ⋮ | ⋮ |
| 16 | ⋯ |
| 16 | ⋯ |

Bootstrapped Sample 2

⋯

⋯

# Split-feature Randomization

- Issue: decision trees trained on bootstrapped samples still behave similarly
- Idea: in addition to sampling the data points (i.e., the rows), also sample the features (i.e., the columns)
- Each time a split is being considered, limit the possible features to a randomly sampled subset

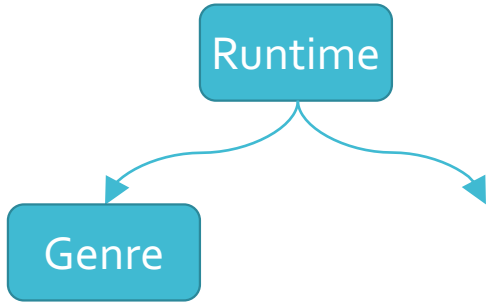
| Runtime | Genre | Budget | Year | IMDB | Rating |
|---|---|---|---|---|---|

# Split-feature Randomization

- Issue: decision trees trained on bootstrapped samples still behave similarly
- Idea: in addition to sampling the data points (i.e., the rows), also sample the features (i.e., the columns)
- Each time a split is being considered, limit the possible features to a randomly sampled subset

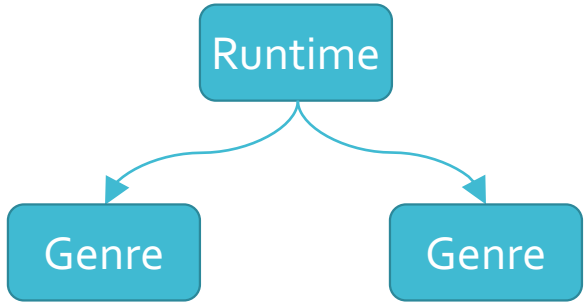

# Split-feature Randomization

- Issue: decision trees trained on bootstrapped samples still behave similarly
- Idea: in addition to sampling the data points (i.e., the rows), also sample the features (i.e., the columns)
- Each time a split is being considered, limit the possible features to a randomly sampled subset

# Split-feature Randomization

- Issue: decision trees trained on bootstrapped samples still behave similarly
- Idea: in addition to sampling the data points (i.e., the rows), also sample the features (i.e., the columns)
- Each time a split is being considered, limit the possible features to a randomly sampled subset

# Random Forests

- Input: $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}, B, \rho$
- For $b = 1, 2, \ldots, B$
  - Create a dataset, $\mathcal{D}_b$, by sampling $N$ points from the original training data $\mathcal{D}$ ***with replacement***
  - Learn a decision tree, $t_b$, using $\mathcal{D}_b$ and the ID3 algorithm ***with split-feature randomization***, sampling $\rho$ features for each split
- Output: $\bar{t} = f(t_1, \ldots, t_B)$, the aggregated hypothesis

# How can we set $B$ and $\rho$?

- Input: $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}, B, \rho$
- For $b = 1, 2, \ldots, B$
  - Create a dataset, $\mathcal{D}_b$, by sampling $N$ points from the original training data $\mathcal{D}$ ***with replacement***
  - Learn a decision tree, $t_b$, using $\mathcal{D}_b$ and the ID3 algorithm ***with split-feature randomization***, sampling $\rho$ features for each split
- Output: $\bar{t} = f(t_1, \ldots, t_B)$, the aggregated hypothesis

# Recall: Validation Sets


- Suppose we want to compare multiple hyperparameter settings $\theta_1, \ldots, \theta_K$
- For $k = 1, 2, \ldots, K$
  - Train a model on $D_{train}$ using $\theta_k$
- Evaluate each model on $D_{val}$ and find the best hyperparameter setting, $\theta_{k^*}$
- Compute the error of a model trained with $\theta_{k^*}$ on $D_{test}$

# Out-of-bag Error

- For each training point, $\boldsymbol{x}^{(n)}$, there are some decision trees which $\boldsymbol{x}^{(n)}$ was not used to train (roughly $B/e$ trees or 37%)
  - Let these be $t^{(-n)} = \left\{t_1^{(-n)}, t_2^{(-n)}, \ldots, t_{N_{-n}}^{(-n)}\right\}$
- Compute an aggregated prediction for each $\boldsymbol{x}^{(n)}$ using the trees in $t^{(-n)}$, $\bar{t}^{(-n)}\left(\boldsymbol{x}^{(n)}\right)$
- Compute the out-of-bag (OOB) error, e.g., for regression

$$E_{OOB} = \frac{1}{N}\sum_{n=1}^{N}\left(\bar{t}^{(-n)}\left(\boldsymbol{x}^{(n)}\right) - y^{(n)}\right)^2$$

# Out-of-bag Error

- For each training point, $\boldsymbol{x}^{(n)}$, there are some decision trees which $\boldsymbol{x}^{(n)}$ was not used to train (roughly $B/e$ trees or 37%)
  - Let these be $t^{(-n)} = \left\{t_1^{(-n)}, t_2^{(-n)}, \ldots, t_{N_{-n}}^{(-n)}\right\}$
- Compute an aggregated prediction for each $\boldsymbol{x}^{(n)}$ using the trees in $t^{(-n)}$, $\bar{t}^{(-n)}\left(\boldsymbol{x}^{(n)}\right)$
- Compute the out-of-bag (OOB) error, e.g., for classification

$$E_{OOB} = \frac{1}{N}\sum_{n=1}^{N}\left[\!\left[\bar{t}^{(-n)}\left(\boldsymbol{x}^{(n)}\right) \neq y^{(n)}\right]\!\right]$$

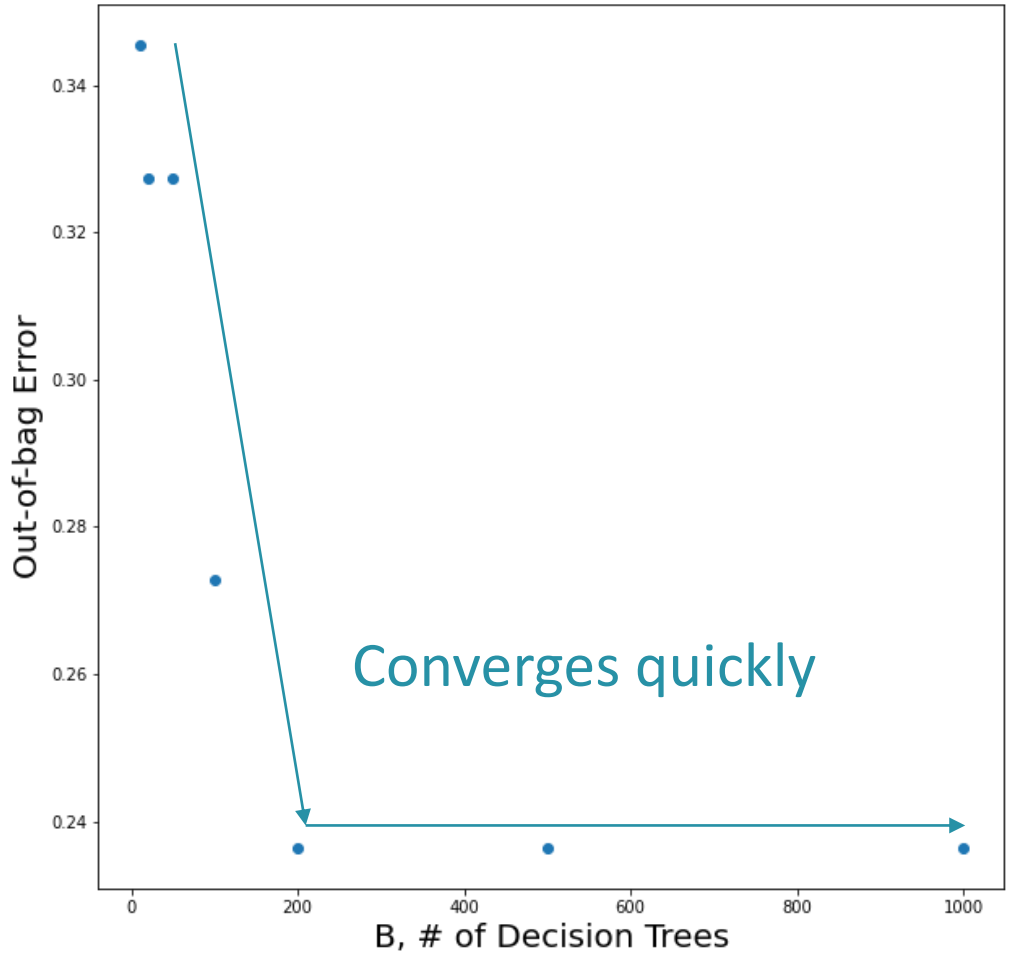
- $E_{OOB}$ can be used for hyperparameter optimization!

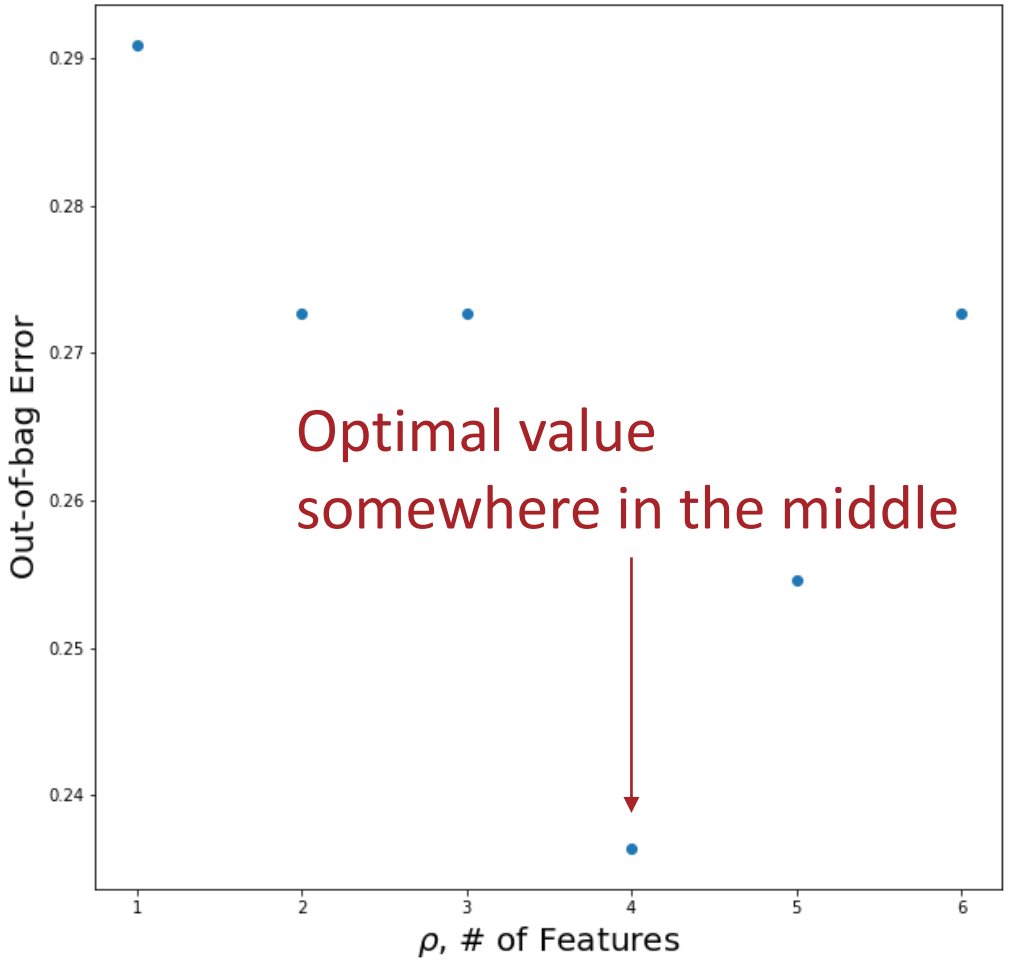
# Out-of-bag Error


- Suppose we want to compare different numbers of trees in our random forest $B_1, \ldots, B_K$
- For $k = 1, 2, \ldots, K$
  - Train a random forest on $D_{train}$ with $B_k$ trees
- Compute $E_{OOB}$ for each random forest and find the best number of trees, $B_{k^*}$
- Evaluate the random forest with $B_{k^*}$ trees on $D_{test}$

# Out-of-bag Error


- Suppose we want to compare different numbers of trees in our random forest $B_1, \ldots, B_K$
- For $k = 1, 2, \ldots, K$
  - Train a random forest on $D_{train}$ with $B_k$ trees
- Compute $E_{OOB}$ for each random forest and find the best number of trees, $B_{k^*}$
- Evaluate the random forest with $B_{k^*}$ trees on $D_{test}$



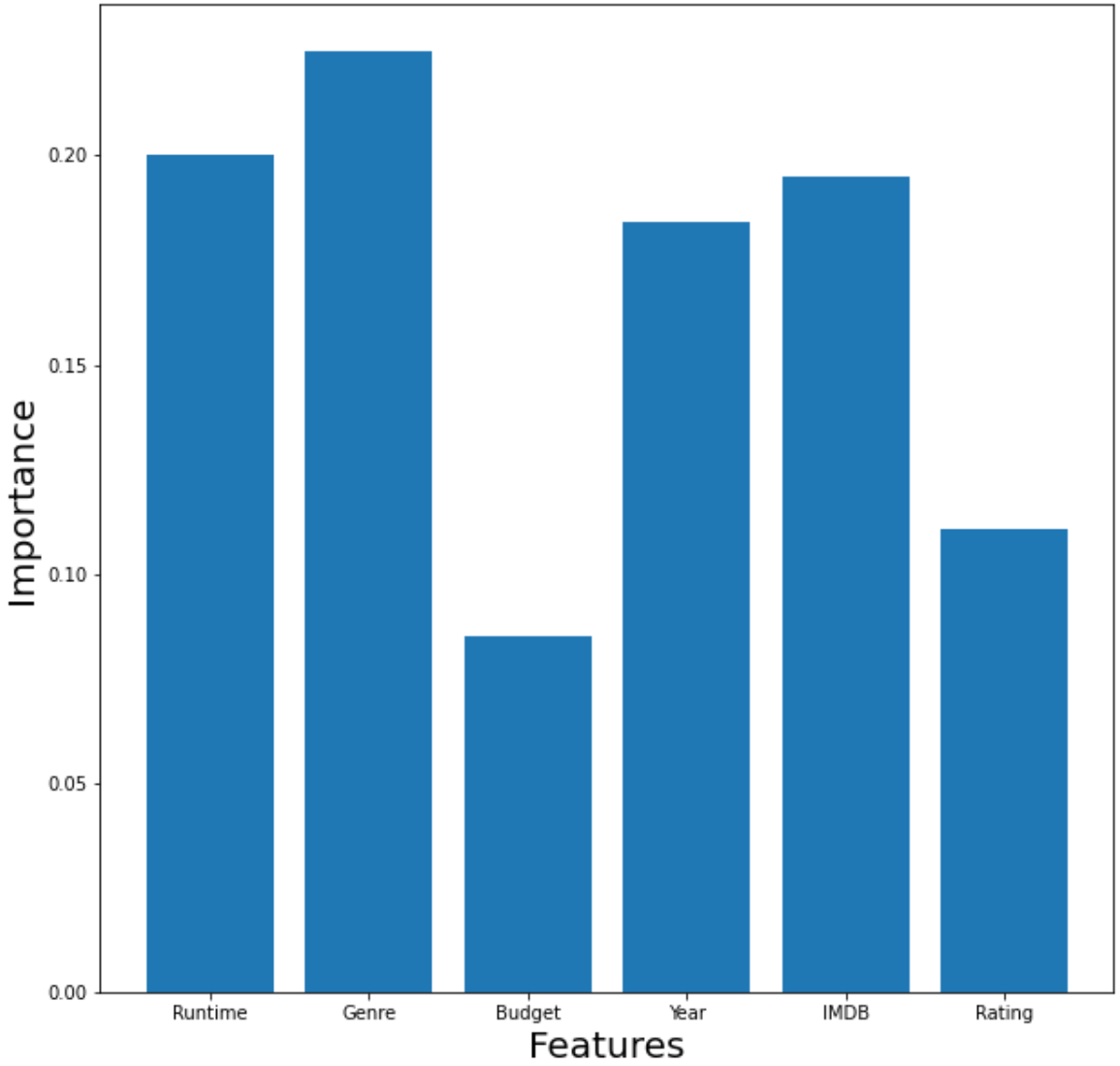




# Setting Hyperparameters

# Feature Importance

- Some of the interpretability of decision trees gets lost when switching to random forests
- Random forests allow for the computation of “feature importance”, a way of ranking features based on how useful they are at predicting the target
- Initialize each feature’s importance to zero
- Each time a feature is chosen to be split on, add the reduction in Gini impurity (weighted by the number of data points in the split) to its importance

# Feature Importance



# Key Takeaways

- Ensemble methods employ a “wisdom of crowds” philosophy
  - Can reduce the variance of high variance methods
- Random forests = bagging + split-feature randomization
  - Aggregate multiple decision trees together
  - Bootstrapping and split-feature randomization increase diversity in the decision trees
  - Use out-of-bag errors for hyperparameter optimization
  - Use feature importance to identify useful attributes