Class 17: Regression Trees and Bagging

MGSC 310

Prof. Jonathan Hersh

Class 17 Announcements

1. No Quiz This Week
2. Today! (11/5): Interview with Sumair Shah (Minnesota Twins)
3. Final project one sheet due November 10
4. Pset 5 due Tuesday, November 17 (posted)
5. Midterms will grade by next week

Today: Sumair Shah

R&D / Player Development
Minnesota Twins

Chapman ’18

BA, Business Admin and Finance

Georgia Institute of Tech ‘21

MA, Analytics and Machine Learning

Data Analytics Association Guest Speaker

Annie Wang

JD Candidate, Yale

Formerly:

Director of Data Science, Warren for President

Director of Research and Analytics, Analyst Institute

Civis Analytics, Senior Applied Data Scientist

Tuesday, Nov 10 @ 7pm

Final Project: One-Page Outline Due Nov 10

- **Your final project should take a real-world data set, and estimate a series of predictive models against this dataset.**
- You should identify a business use-case for the prediction, and stimate at least three predictive models we covered in this class against the dataset.

November 10th – Due: students must upload to Canvas a one-page outline of their project. This outline should include:

a) identify a dataset you will use
b) the outcome you are trying to predict, and what variables you will use to predict it
c) motivation to your project -- as in the business or practical management use case of such a prediction
d) three methods you will use to analyze your question of interest
e) the names of the students who will be part of your group (up to four)
f) Rmarkdown document showing summary statistics (mean, std dev, min max) against the baseline dataset.

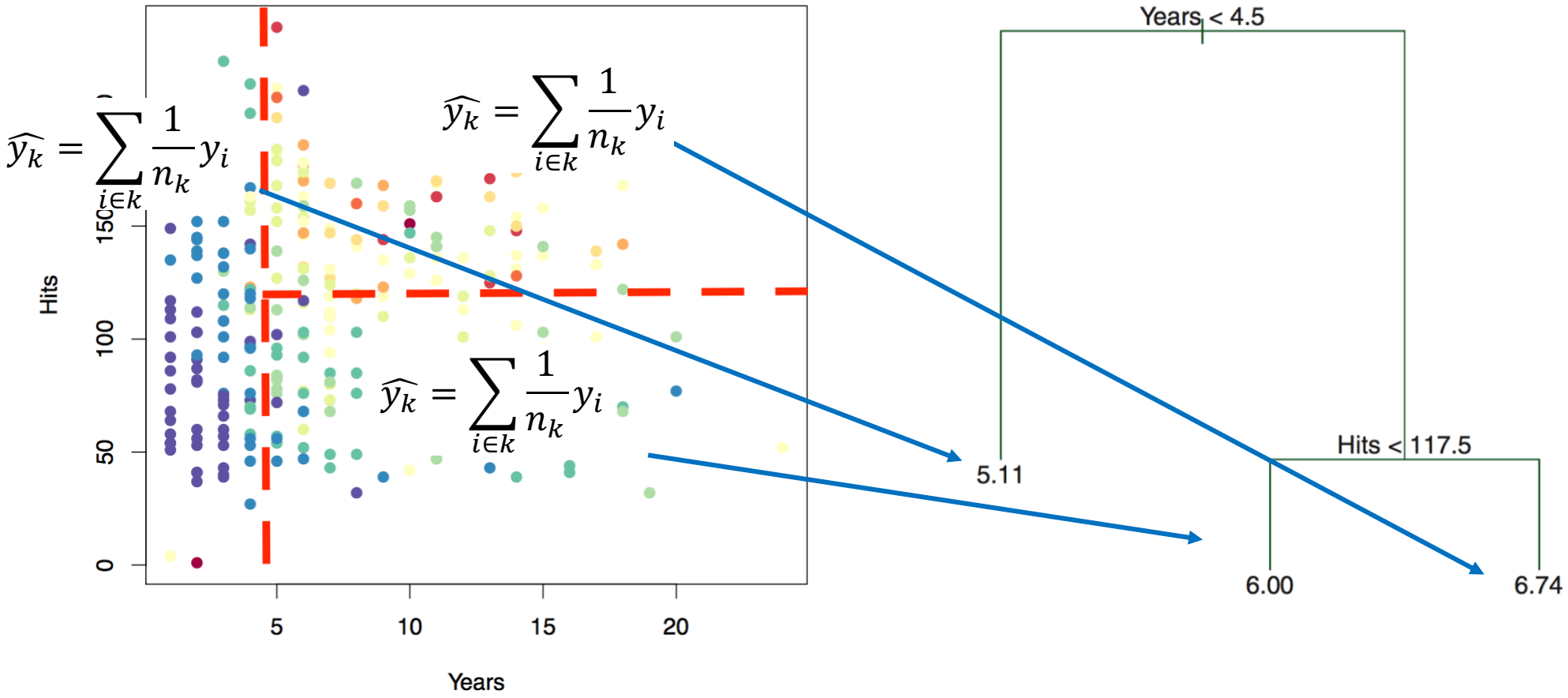
Where to Find Datasets?

- Kaggle: https://www.kaggle.com/datasets
- Kaggle: https://www.kaggle.com/annavictoria/ml-friendly-public-datasets
- FiveThirtyEight https://data.fivethirtyeight.com/
- TidyTuesday: https://github.com/rfordatascience/tidytuesday
- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php

Class 17: Outline

1. Interview Sumair Shah
2. Review Regression Trees
3. Regression Trees Lab
4. Intro to Bagging and Ensemble Methods

Review: Predictions for Trees? “Leaf” Values



- Leaf predictions are average values in each partition, k

Review: Regression Tree Estimation

```
# clean data
titanic_df <- titanic_train %>% as_tibble() %>%
  mutate(Suvived = if_else(Survived == 1,
                           "Survived", "Dead"),
         Survived = as.factor(Survived),
         Sex = as.factor(Sex),
         Pclass = as.factor(Pclass)) %>%
         mutate_if(is.character, as.factor)
```

```
# ctree to estimate model
titanic_tree <- ctree(Survived ~ Sex + Pclass,
                      data = titanic_df)
titanic_tree
```

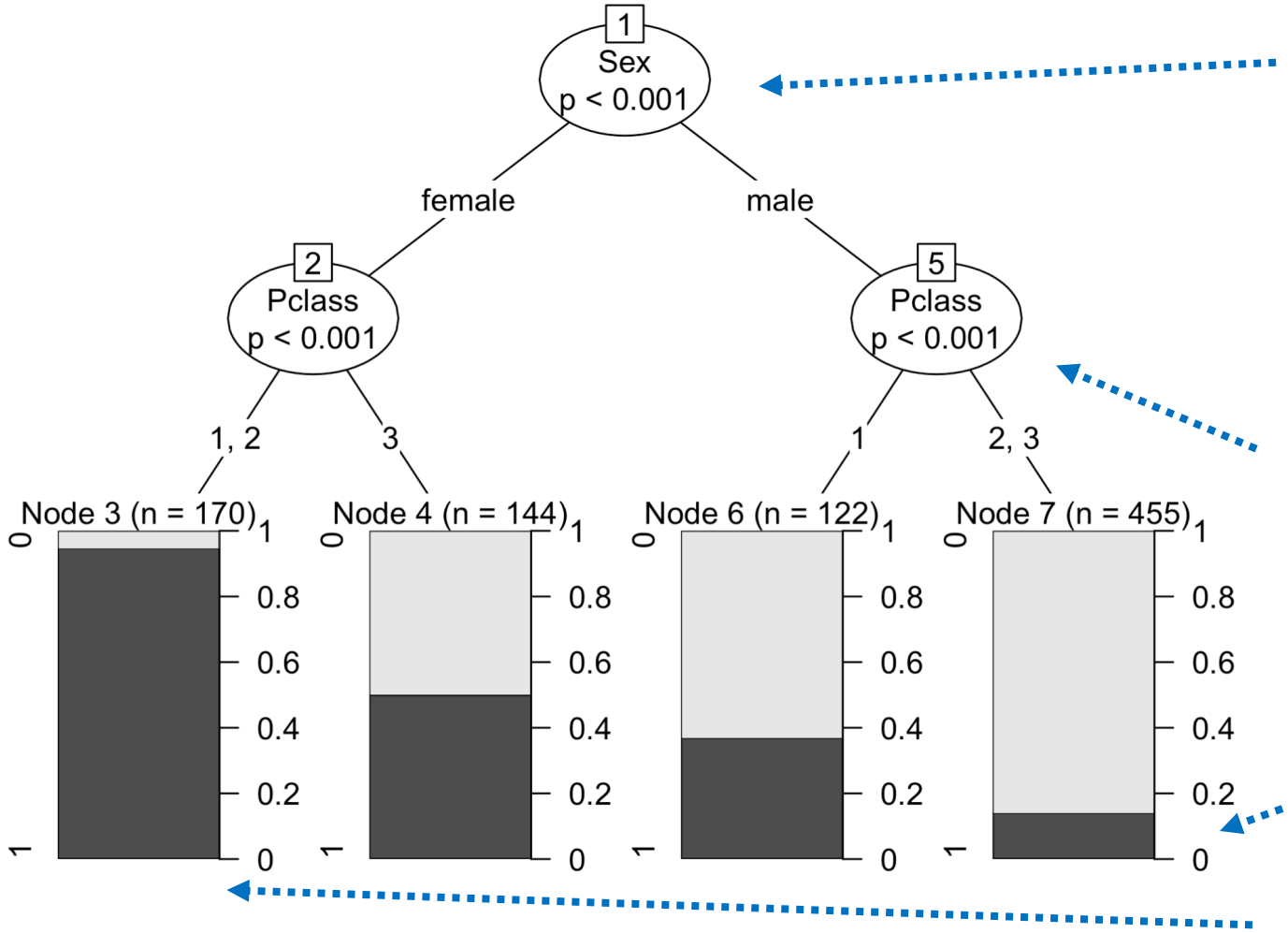
```
> titanic_tree

Model formula:
Survived ~ Sex + Pclass

Fitted party:
[1] root
|   [2] Sex in female
|   |   [3] Pclass in 1, 2: 1 (n = 170, err = 5.3%)
|   |   [4] Pclass in 3: 0 (n = 144, err = 50.0%)
|   [5] Sex in male
|   |   [6] Pclass in 1: 0 (n = 122, err = 36.9%)
|   |   [7] Pclass in 2, 3: 0 (n = 455, err = 14.1%)

Number of inner nodes:    3
Number of terminal nodes: 4
```

- First a bit of data cleaning against the titanic data frame.
- Note we change the characters to factors using mutate_if()

- ctree() function estimates a regression tree
- We need: formula (Y and Xs)
- Data set to estimate (cleaned training)

- Raw model output isn't the prettiest to read (but does show fitted model.)

Review: Plotting Regression Tree



- First split is shown at top with p-value with null hypothesis that split doesn't increase class "separation". Sex of passenger is the most important variable, and p-value shows it improves model fit
- Second split is given by successive node. For males, passenger class is second most important variable
- Survival rates are shown in the leaves summaries at the bottom. There are 455 males with passenger class 2 or 3, and about 15% of them survived.
- For females in classes 1 or 2, nearly all survive. Class 3 females have a survival rate of 50%.

```
#-----------------------------------------------------
# Regression Tree Exercises
#-----------------------------------------------------
# 1. Estimate a regression tree model predicting survival as a function
#    of Sex, Pclass, Age, SibSp and Fare paid using the ctree package.
#    Store this model as titanic_tree_mod2
# 2. Use the print function against the fitted model to view the text
#    Descriptions of the model fit
# 3. Use the plot function on the fitted object to produce the tree plot
#    (you can use the option "gp = gpar(fontsize = 6)")
#    to change the text font size.
# 4. Who has the best chance of survival? Who has the worst?
```

Lab Time!

ENTERTAINMENT

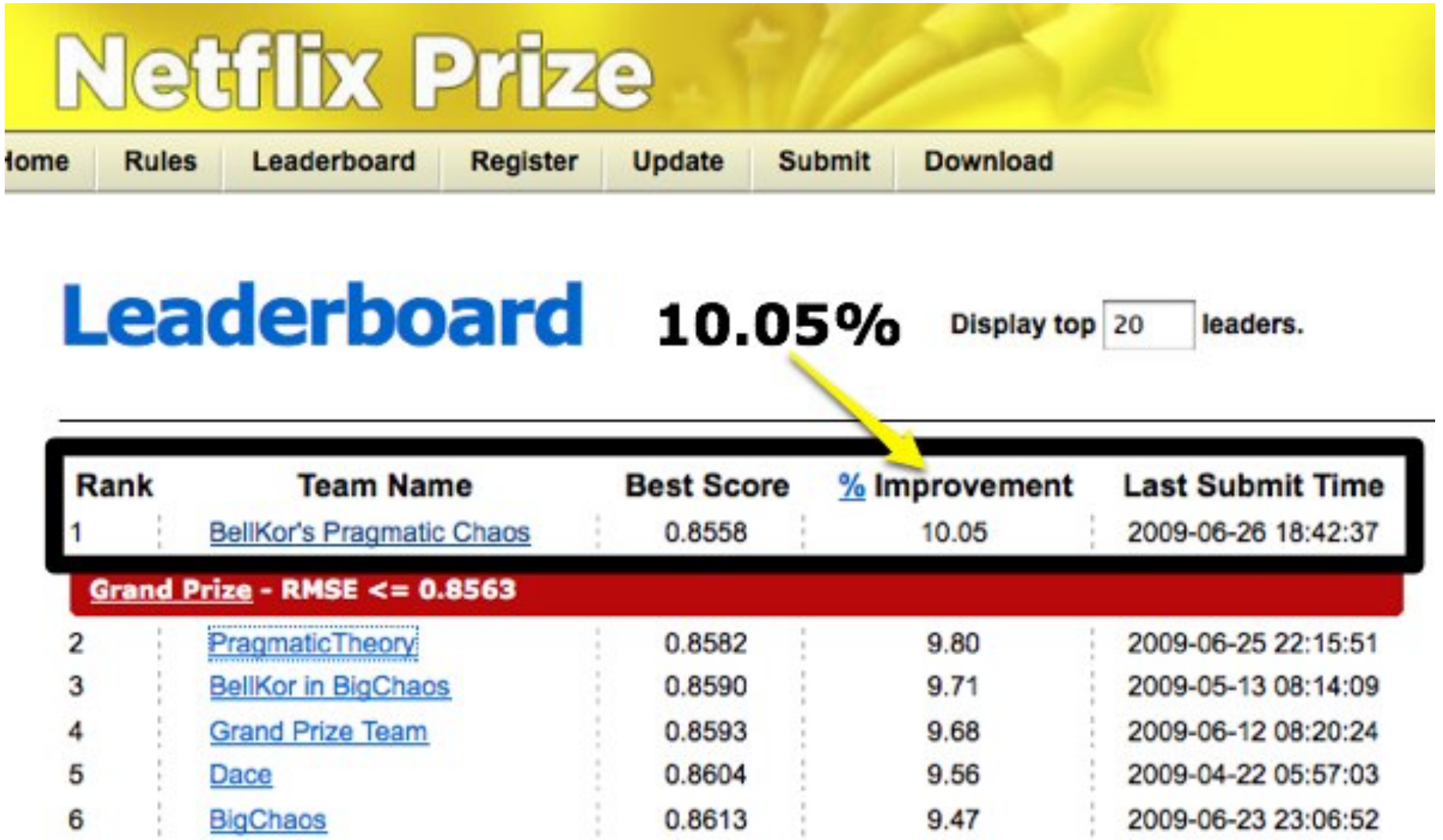
The Netflix Prize: How a $1 Million Contest Changed Binge-Watching Forever

By DAN JACKSON
Published On 07/07/2017
@danielvjackson

In October 2006, Netflix, then a service peddling discs of every movie and TV show under the sun, announced "The Netflix Prize," a competition that lured Mackey and his contemporaries for the computer programmer equivalent of the *Cannonball Run*. The mission: Make the company's recommendation engine 10% more accurate -- or die coding. Word of the competition immediately spread like a virus through comp-sci circles, tech blogs, research communities, and even the mainstream media. ("And if You Liked the Movie, a Netflix Contest May Reward You Handsomely" read the *New York Times* headline.) And while a million dollars created attention, it was the data set -- over 100 million ratings of 17,770 movies from 480,189 customers -- that had number-crunching nuts salivating. There was nothing like it at the time. There hasn't been anything quite like it since.

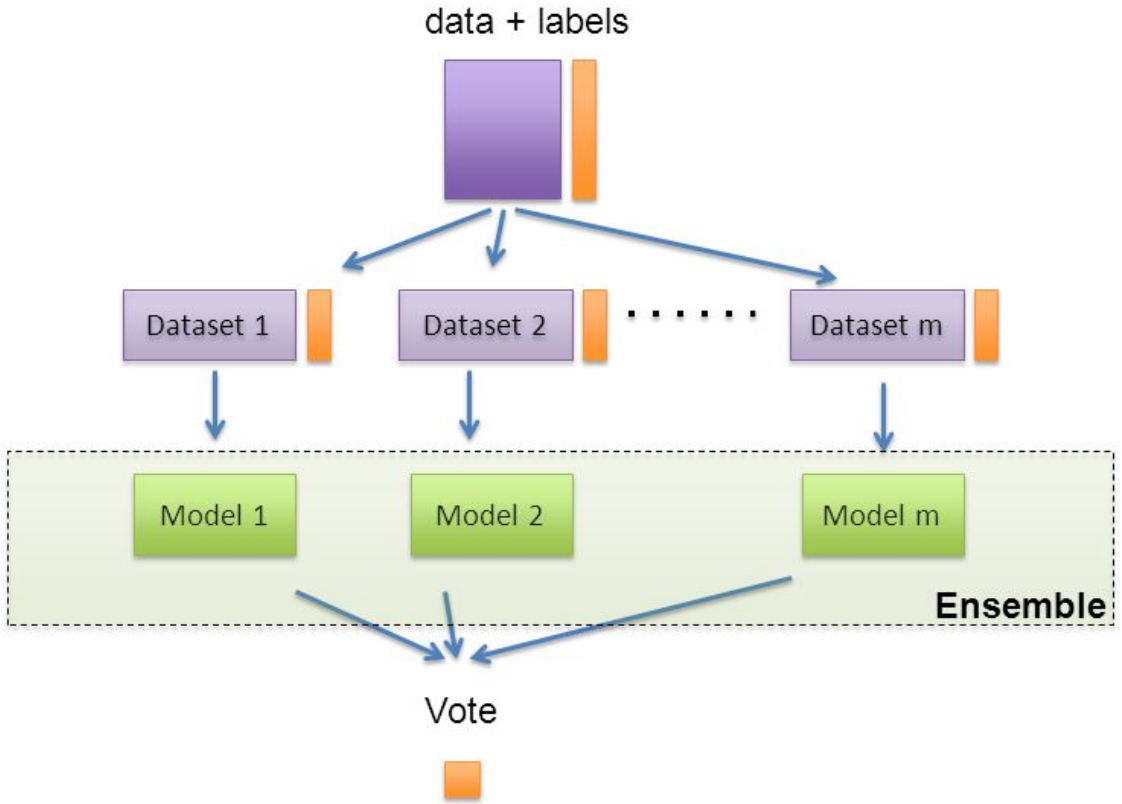
https://www.thrillist.com/entertainment/nation/the-netflix-prize

Netflix prize



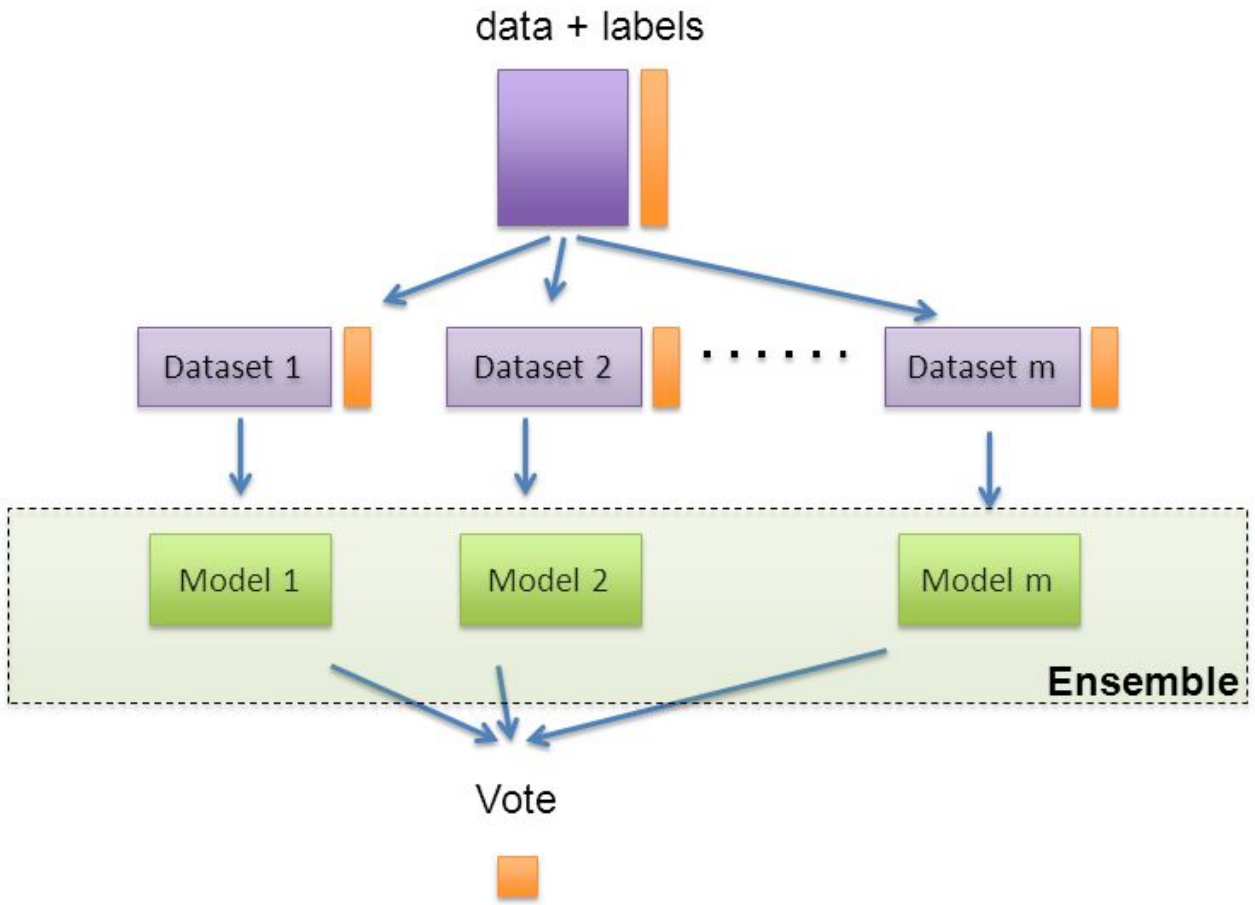
Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	PragmaticTheory	0.8582	9.80	2009-06-25 22:15:51
3	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
4	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
5	Dace	0.8604	9.56	2009-04-22 05:57:03
6	BigChaos	0.8613	9.47	2009-06-23 23:06:52

Netflix prize winners



- Punchline: simple models beat out one very deep model

Netflix prize conclusion: ensemble of simple methods beats one complex method



Bagging

- Bagging is short for bootstrap aggregation
- **It's a general purpose method for reducing variance in any machine learning method**

- With n independent observations, $z_1, z_2, \ldots, z_n$ each with variance σ^2 , the variance of the mean ($\bar{z}$) is given by $\sigma^2/\bar{z}$
- We usually cannot do this because we don't have multiple training datasets

Pruning trees

- How do we know when to stop splitting the data?
- **Trees with many splits can overfit the data**
- Solution is to grow a large tree T_0, then prune it to obtain a smaller sub-tree

Baseball example, unpruned tree

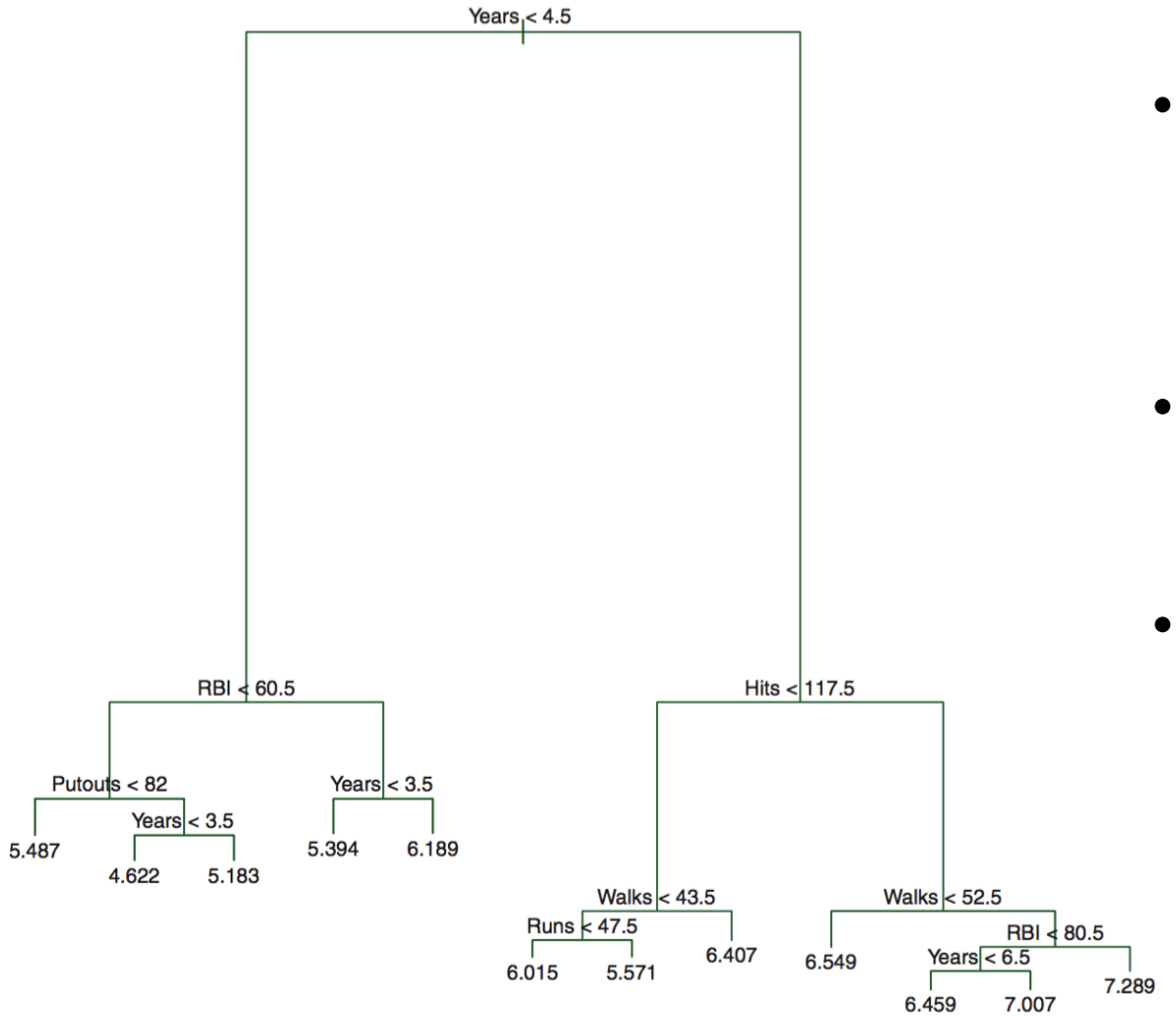
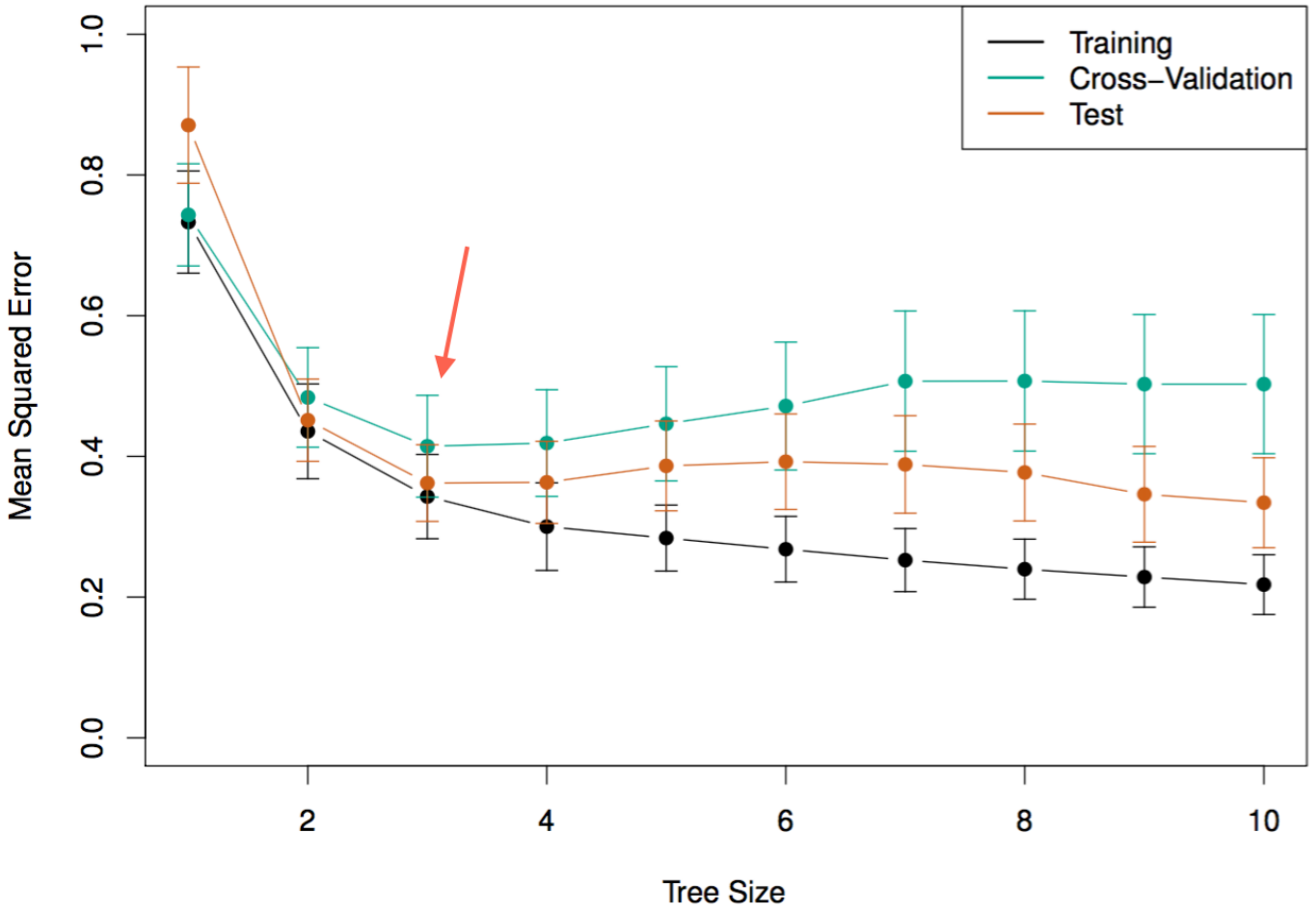


FIGURE 8.4. *Regression tree analysis for the* `Hitters` *data. The unpruned tree that results from top-down greedy splitting on the training data is shown.*

- This tree has too many splits at the bottom and will likely overfit any other data set
- Therefore we must adjust the "depth" of the tree.
- As in prevent the tree from having too many successive splits that only explain the training data and not the test

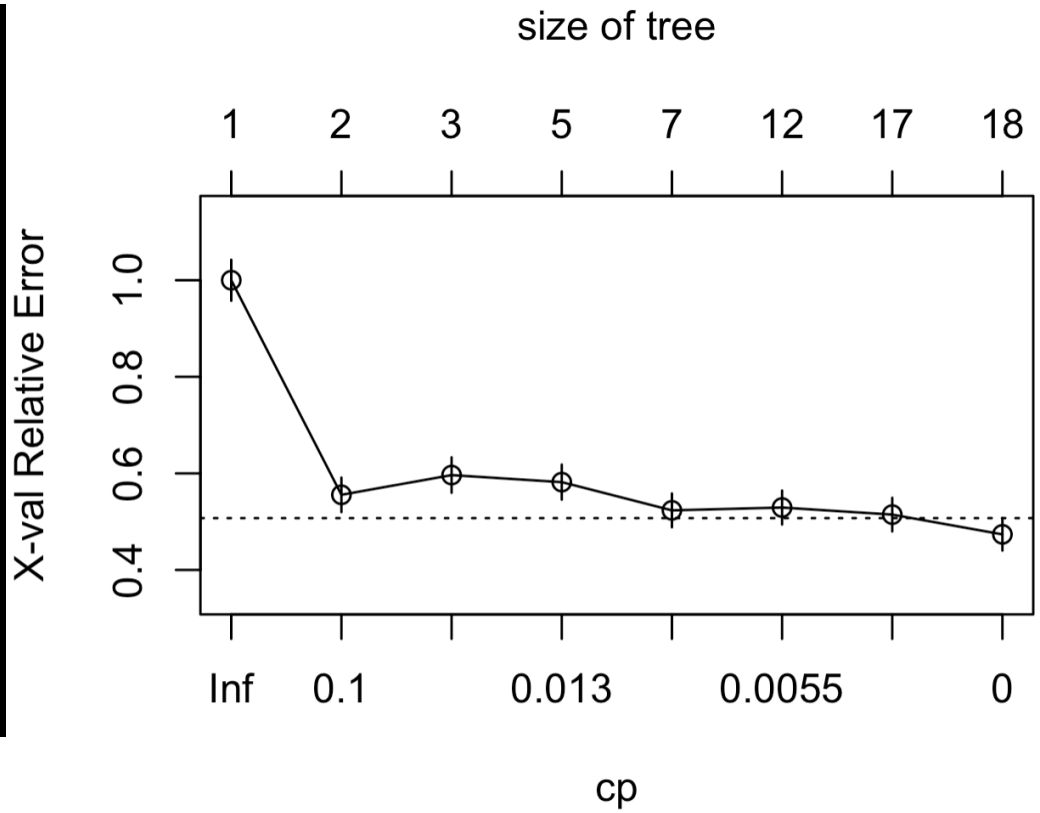
Cross-validate to find alpha



- Complexity in decision trees is controlled by the tree size or depth
- We must cross-validate to select the optimal tree depth, which is here a tree size of depth 3
- Note training error declines with tree depth! (Make sure you know why.)

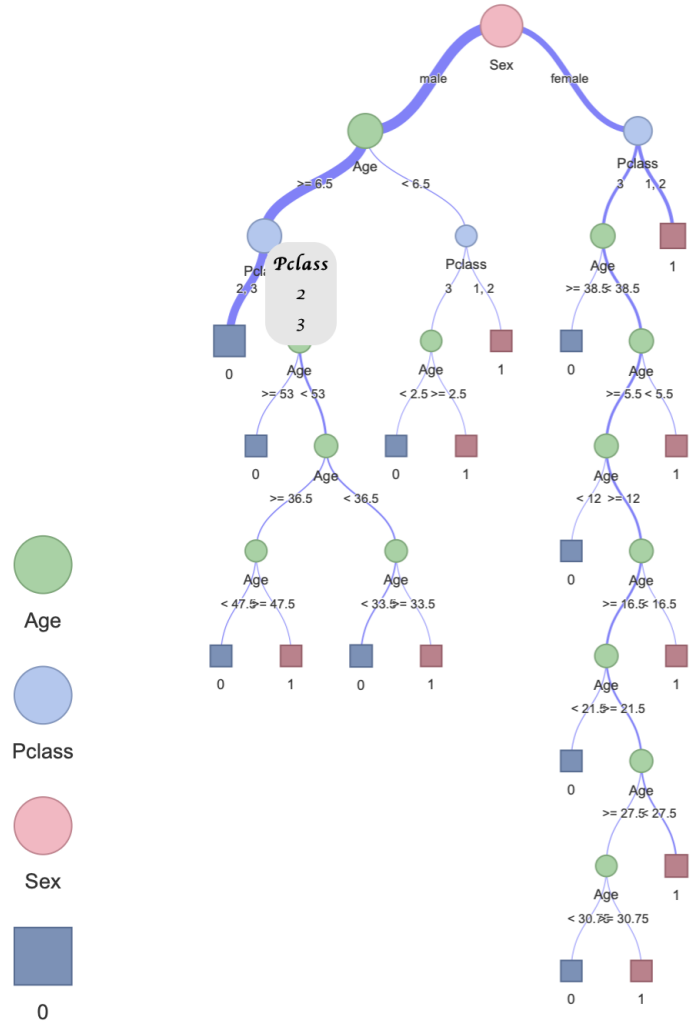
Cross –Validate Our Titanic Data to Determine Optimal Tree Depth

```
# rpart function to select optimal depth of tree
# read the help() file for rpart.control to learn about
#  the different function options
# max depth = 6 ensures the final tree only has this
#  many splits
# min split means minimum observations in a node before
#  a split can be attempted
# cp is the complexity parameter, overall Rsq must
#  increase by cp at each step
titanic_mod_rpart <- rpart(Survived ~ Sex + Pclass + Age,
                           data = titanic_df,
                           method = "class",
                           control = list(cp = 0,
                                          minsplit = 10,
                                          maxdepth = 15))
```



Visualize the Fitted Tree Interactively Using vizTree in VizNetworks Package

```
# fancy, interactive tree visual
install.packages('visNetwork')
install.packages('sparkline')
visNetwork::visTree(titanic_mod_rpart,
                    nodesPopSize = TRUE,
                    edgesFontSize = 18,
                    nodesFontSize = 20,
                    width = "100%",
                    height = "1200px")
```



https://datastorm-open.github.io/visNetwork/

Class 17 Summary

- Regression trees use binary decisions of the X variables to predict the outcome.
- Regression trees can be used for either regression or classification problems.
- We must control the depth and/or complexity of a tree or else it is prone to overfitting.
- Bagging is an ensemble method that combines many models.
- Bagging stands for “bootstrap aggregation” where we combine predictions from models estimated against many bootstrapped datasets.