Predicting the Winner in Professional Tennis

http://thesport-info.blogspot.com/2012/03/roger-federer-hd-wallpapers-2012.html

Team 5

Neha Bora
Iowa State University

John Lynch
University of Wisconsin - Madison

Dara Zirlin
University of Illinois - Urbana Champaign

Issa Tahir
Auburn University

Melissa Gaddy
North Carolina State University

https://www.thesun.co.uk/living/3013973/theres-a-good-reason-why-you-should-always-take-a-tennis-ball-on-flights/

Objectives

1. Identify important predictors
2. Model to predict the outcome of tennis matches
3. Predict the outcome of the 2017 Wimbledon final

Outline

1. Overview of the dataset and tennis
2. Predictor Selection
3. Classification Models
4. Wimbledon 2017
5. Recommendations

Data Snapshot (Men's Singles)

Response

Predictors

tourney_date	Player_1	Player_2	surface	hands	age_diff	ranks_diff	...	won
20000110	Marc Rosset	John Van Lottum	Hard	TRUE	5.42368241	-20	...	1
20000110	Jan Michael Gambill	Magnus Norman	Hard	TRUE	-1.01026694	43		0
20000110	Juan Carlos Ferrero	Juan Balcells	Hard	TRUE	-4.64887064	-166		0
20000110	Michael Chang	Magnus Gustafsson	Hard	TRUE	-5.13620808	-10		1
20000110	Gaston Gaudio	Sjeng Schalken	Hard	TRUE	-2.25051335	35		1
20000110	Magnus Norman	Marc Rosset	Hard	TRUE	-5.56057495	-32		1

Data for the years 2000-2017

Predictors in our model

Surface (Hard, Clay, Grass, Carpet)

Hands (=TRUE if players have same handedness)

Current Player 1 & 2 Differences:

- Age
- Rank points
- Log(rank points)
- Rank

How players accumulate rank points:

Tournament category	W	F	SF	QF	R16	R32	R64	R128	Q
Grand Slam	2000	1200	720	360	180	90	45	10	25
ATP World Tour Finals	+500 (1500 max)	+400 (1000 max)	(200 for each round robin match win) (600 max)						
Masters 1000	1000	600	360	180	90	45	10 (25)	(10)	25 (16)
500 Series	500	300	180	90	45	(20)			20 (10)
250 Series	250	150	90	45	20	(5)			12 (5)
ATP Challenger Tour Finals	+50	+30	(15 for each round robin match win)						
Challenger 125,000 +H	125	75	45	25	10				5
Challenger 125,000	110	65	40	20	9				5
Challenger 100,000	100	60	35	18	8				5

ATP Rankings (singles), as of 17 July 2017[17]			
#	Player	Points	Move†
1	Andy Murray (GBR)	7,750	—
2	Rafael Nadal (ESP)	7,465	—
3	Roger Federer (SUI)	6,545	▲ 2
4	Novak Djokovic (SRB)	6,325	—
5	Stan Wawrinka (SUI)	6,140	▼ 2
6	Marin Čilić (CRO)	5,075	—

http://www.atpworldtour.com/Rankings/Rankings-FAQ.aspx

Historical predictors in our model

	Ace	1stIn	1stWon	2ndWon	Break point
Server	Serve ✓	1st serve ✓	1st Serve ✓	1st Serve - ✕ 2nd Serve - ✓	Serves ✓
Receiver	Doesn't touch the ball		Loses the point	Loses the point	If wins the point → gains service

Ratio of aces/ total serve points

Ratio of 1st In/ total serve points

Ratio of 1st Won/ total serve points

Ratio of 2nd Won/ total serve points

How would YOU predict the outcome of a tennis match?

https://mariposaxprs.wordpress.com/page/7/

How would YOU predict the outcome of a tennis match?

- Ask a friend
- Pick a favorite
- Compare the rankings

How would YOU predict the outcome of a tennis match?

- Ask a friend
- Pick a favorite
- Compare the rankings

Rank comparison gives 67.3% prediction accuracy!

What else can we try?

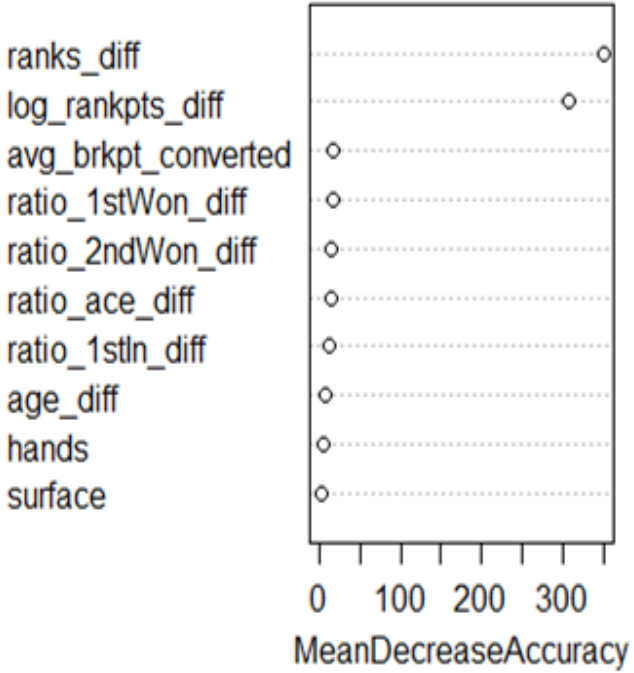
	Methods for predictor importance
1	Random Forest
2	Best Subset Selection

Machine Learning

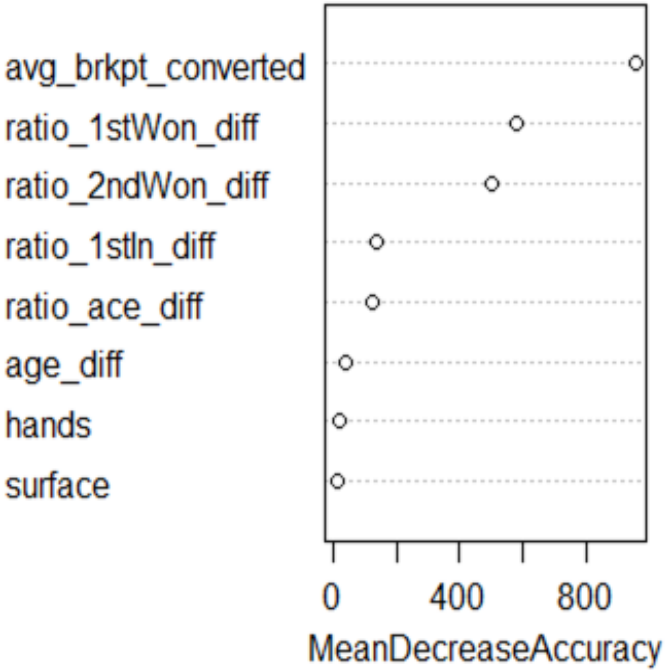
	Classification models
1	Rank Comparison
2	Logistic Regression
3	Decision Trees
4	Support Vector Machines
5	Linear Discriminant Analysis
6	Neural Network

Predictors of Importance for general model

With all the predictors

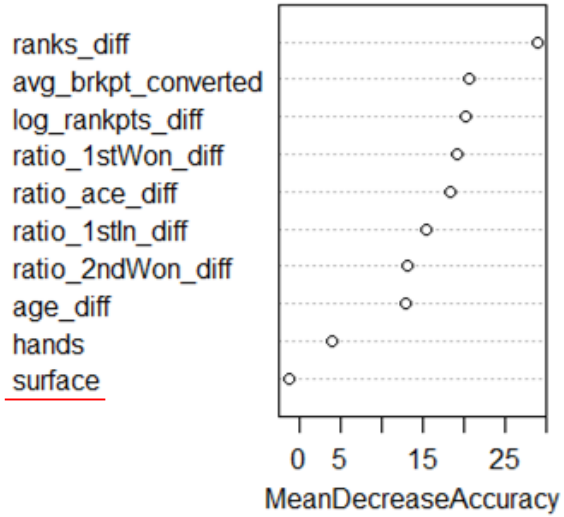


With all the predictors **except ranks**

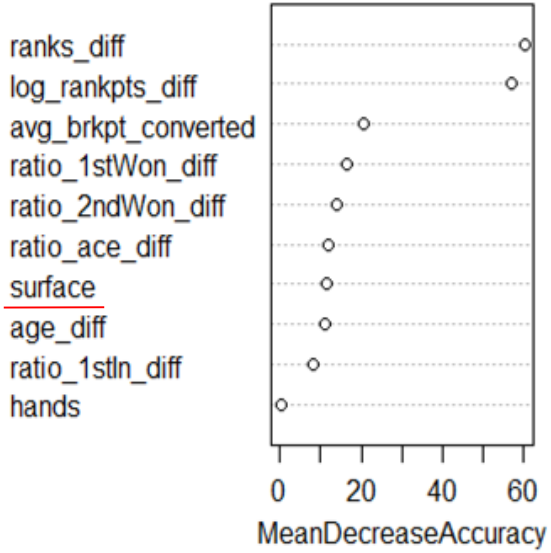


Predictors of importance for top players

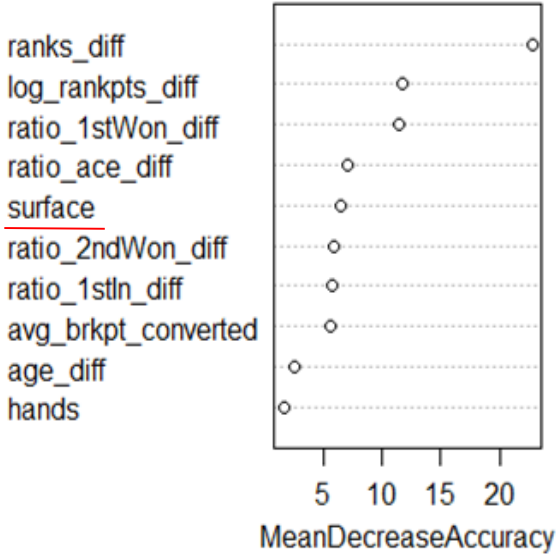
Roger Federer



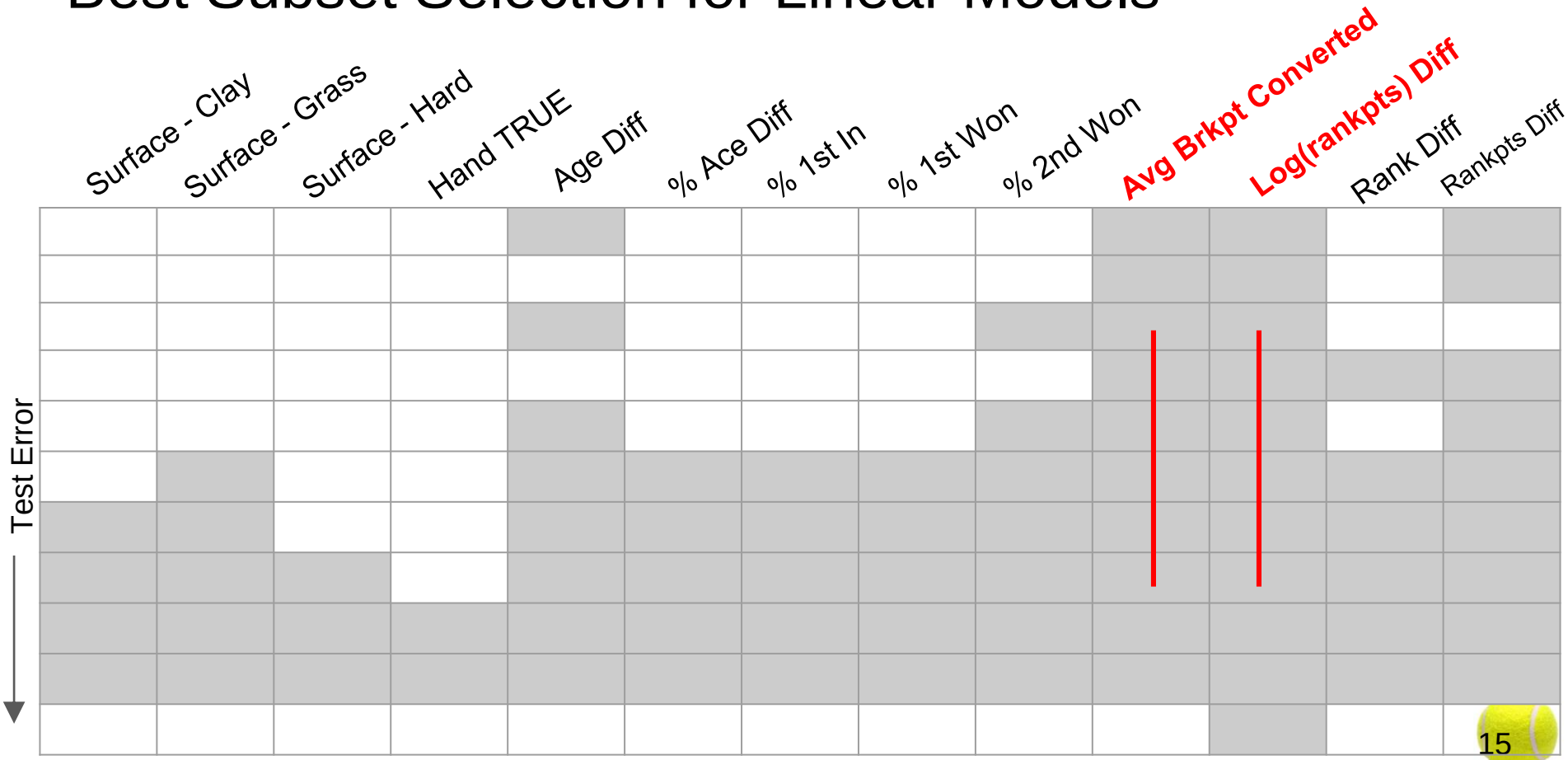
Rafael Nadal



Andy Murray



Best Subset Selection for Linear Models



Logistic Regression Model

Computes a probability that Player 1 wins

Accuracy of classification:

Name	With Rank Predictors	Without Rank Predictors
General Model	68.1%	64.4%
Roger Federer	87.3%	86.6%
Andy Murray	87.8%	82.9%
Rafael Nadal	80.6%	79.9%

Summary of Results

	Methods for predictor importance
1	Best subset regression
2	Random Forest

Machine Learning

Classification models	Highest Accuracy obtained
Rank Comparison	67.3 %
Logistic Regression	68.1%
Linear Discriminant Analysis	67%
Support Vector Machines	66.3%
Neural Network	66%
Tree	58%

2017 Wimbledon Championship Final:

Marin Cilic (6)

vs.

Roger Federer (5)

Model	Predicted Outcome
Tree (all Predictors)	Federer
Tree (no Rank Predictors)	Federer
Support Vector Machines (no Rank Predictors)	Federer
Logistic regression without ranks and surface	Federer (with prob. of 0.73)
Logistic regression with ranks only	Federer (with prob. of 0.503)

https://www.theguardian.com/sport/live/2017/jul/16/wimbledon-2017-roger-federer-v-marin-cilic-in-mens-singles-final-live

Recommendations

Rank is best predictor

Recommendations

Rank is best predictor

When the rank difference is small, we can use other predictors:

- Average breakpoints converted
- 1st and 2nd Won

Recommendations

Rank is best predictor

When the rank difference is small, we can use other predictors:

Average breakpoints converted

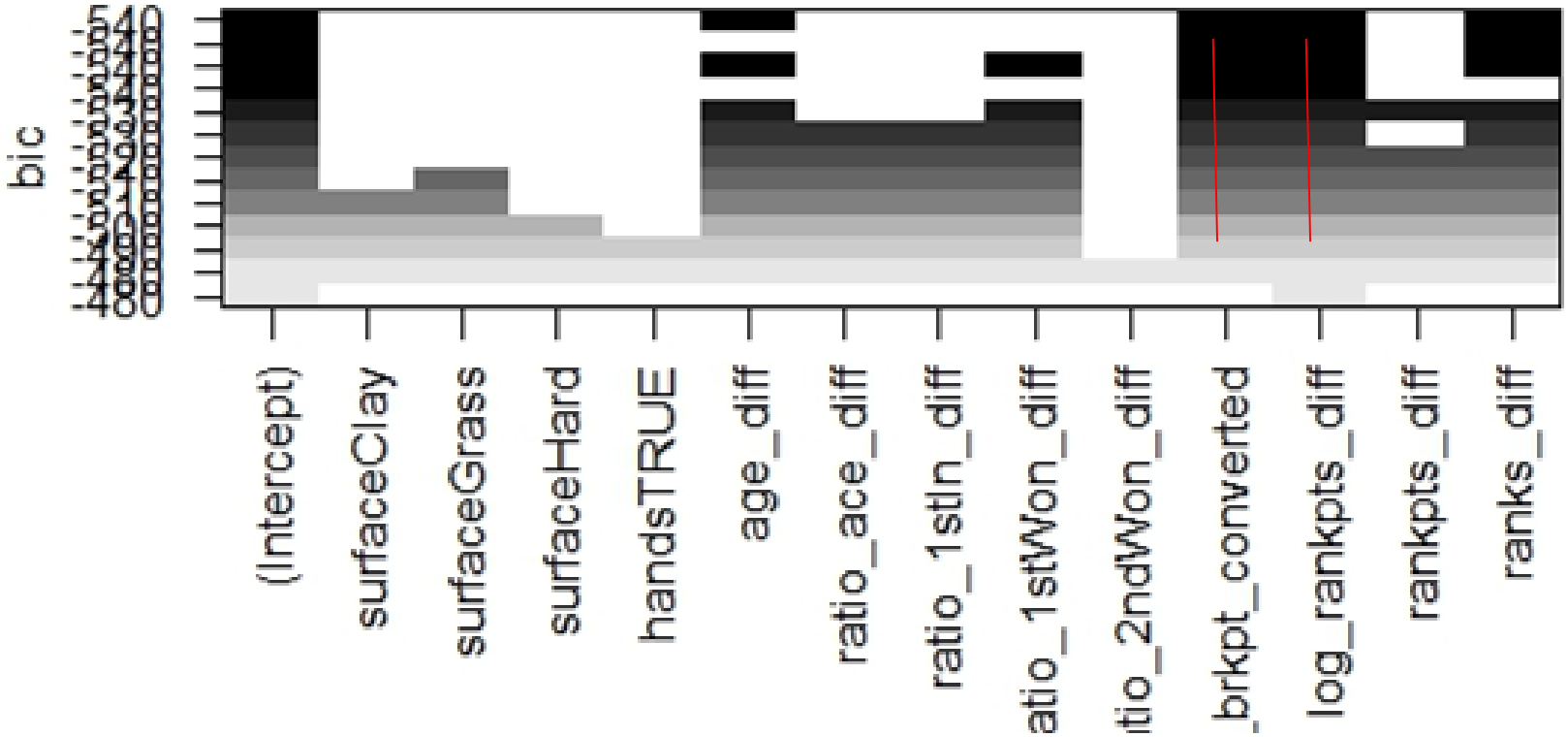
1st and 2nd Won

A small percent increase in accuracy = $$$

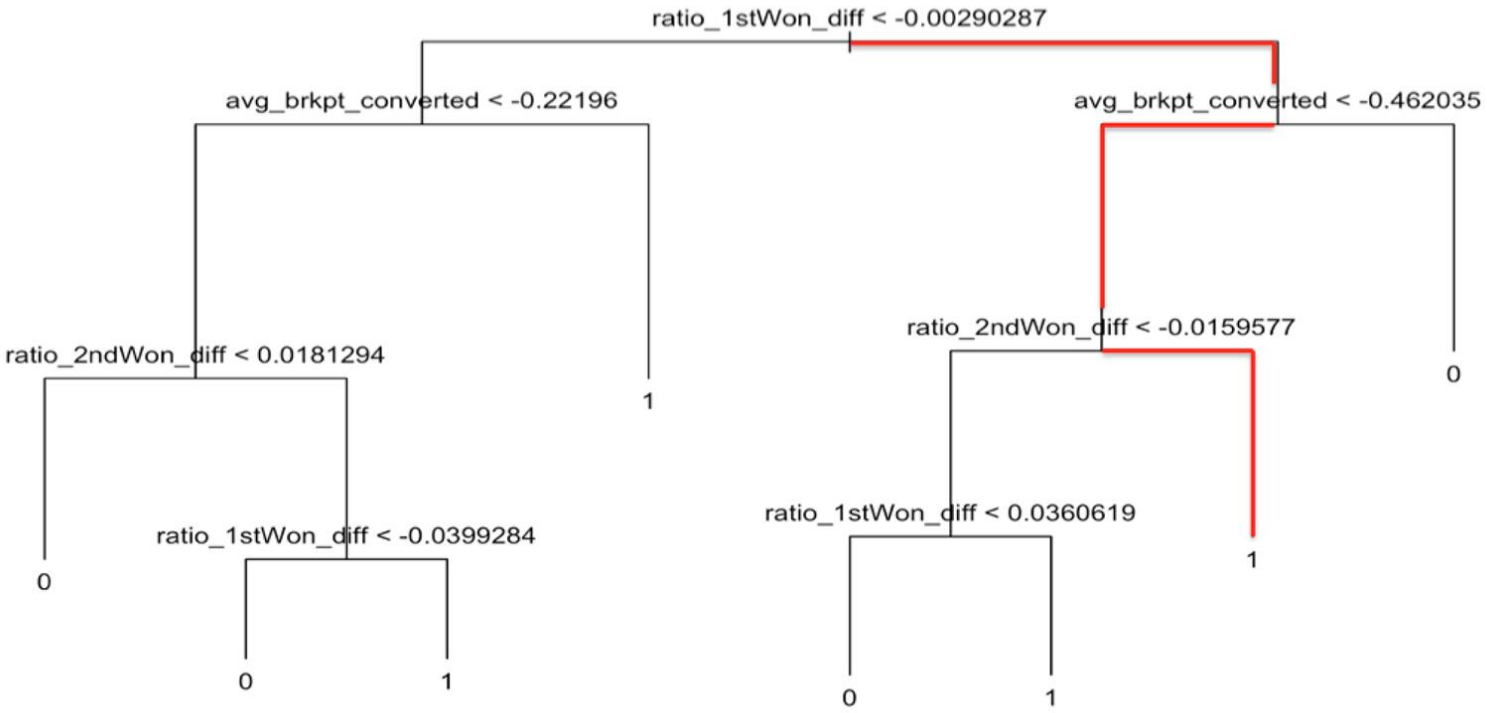
Thank you!

What questions do you have?

Best Subset Selection for Linear Models



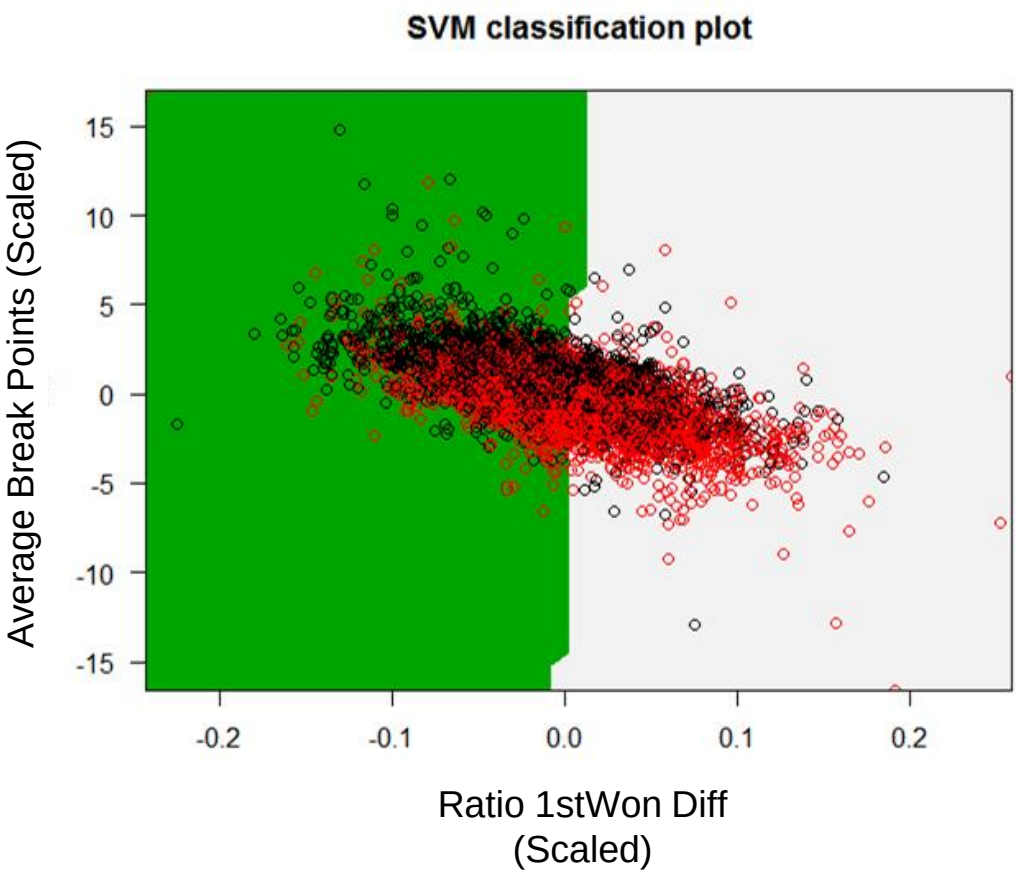
Decision Tree - All Predictors **Except Rank**



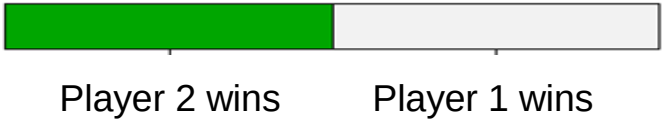
1 = Player 1 wins, 0 = Player 2 wins.

Gives 58% accuracy

Support Vector Machine



SVM Predictions:



Data Set:

Player 1 wins

Player 2 wins

Highest accuracy: 66.3%