Predicting Overtime in the

NBA

KC Kern

15.071 Analytics Edge

May 12, 2014

MITSloan MANAGEMENT

Overview

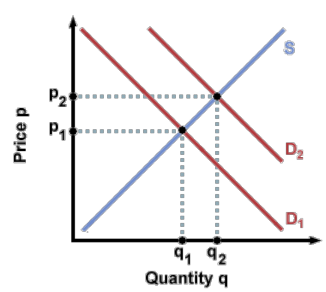
Overtime occurs when the total points accumulated by each time equal each other at the end of the fourth quarter of the game. Overtime puts 5 extra game time minutes on the clock, and the game continues. Over the past 25 years, games have run into overtime 6.1% of the time. The scope of this project is to use statistical, mathematical, and machine learning methods to increase the ability to predict overtime, based on available data.

Business Value

Nielsen is a company dedicated to valuating TV timeslots for the purpose of determining the market value of advertising timeslots. They claim the following:

> There's a big reason why Nielsen is synonymous with television audience measurement. We invented it. Since day one, we've offered the media industry the expertise it needs to make the best marketing decisions possible. Today, our expansive and representative television measurement services capture video viewing across all screens: television, computers and mobile devices. National and local TV ratings help media companies and brands decide how to spend the nearly $70 billion on TV advertising in the U.S. alone. Ratings are only one of the audience measurement services we provide.[1]

Marketing decisions based on the size and quality of a television audience are of paramount importance. NBA games garner large audiences, but the question of weather or not the game will go into overtime adds a level of uncertainty—and for advertisers a level a risk—to the valuation of the timeslot.



When a game goes into overtime, the television audience stays tuned in longer, dramatically increasing the reach (and value) of the advertisements displayed. The demand curve with respect to the timeline shifts outward, producing a new equilibrium level with a higher price point. This increase in price—and the probability of it rising—can be estimated using statistical methods, and factored into valuation models in the same way insurance premium generators compute their probabilistic inputs.

---

[1] http://www.nielsen.com/us/en/nielsen-solutions/nielsen-measurement/nielsen-tv-measurement.html

Hypothesis

Intuitively, better-matched teams should have a higher probability of overtime than unevenly matched teams. Measures of evenly matched teams include pre-game league standings, as well as point deficits at any time during the game.

With a baseline of 6.1%, we would expect that teams with similar records would have a probability of over 6.1%. But how much more? 7%? 10%? 15%? Likewise, a game that should be a blowout (based on team abilities) should have a much lower probability of overtime: far below 6.1%, approaching zero. Probability will change as the game moves forward in time.

This experiment serves to (1) confirm the hypothesis of over/under baseline probabilities based on comparative team standings, and (2) determine the magnitude of the difference in probability that available data offers.

Data Collection

Data was collected through prolonged brute-force HTML scraping of basketballreference.com. Scripts were created to parse the HTML, collect the values, and store them in a systematic manner.

NBA Games Played on May 8, 2010

May 7, 2010 · **May 8, 2010** · May 9, 2010 · May 8 2010 Find Games

2 Games Click **Final** for box score, if available

	1	2	3	4	T
Orlando Magic					105 Final
Atlanta Hawks					75
Magic	28	24	27	26	105
Hawks	18	15	22	20	75

Box Score | Play-By-Play | Shot Charts

PTS	2 tied	22
TRB	Dwight Howard-ORL	16
AST	Jameer Nelson-ORL	4

	1	2	3	4	T
Los Angeles Lakers					111 Final
Utah Jazz					110
Lakers	17	33	32	29	111
Jazz	22	32	26	30	110

Box Score | Play-By-Play | Shot Charts

PTS	Kobe Bryant-LAL	35
TRB	Pau Gasol-LAL	17
AST	Deron Williams-UTA	9

League Standings Standings through May 8, 2010

Eastern Conference	W	L	W-L%	GB	PS/G	PA/G
Atlantic Division						
Boston Celtics*	50	32	.610	—	99.2	95.6
Toronto Raptors	40	42	.488	10	104.1	105.9
New York Knicks	29	53	.354	21	102.1	105.9
Philadelphia 76ers	27	55	.329	23	97.7	101.6
New Jersey Nets	12	70	.146	38	92.4	101.5
Central Division						
Cleveland Cavaliers*	61	21	.744	—	102.1	95.6
Milwaukee Bucks*	46	36	.561	15	97.7	96.0
Chicago Bulls*	41	41	.500	20	97.5	99.1
Indiana Pacers	32	50	.390	29	100.8	103.8
Detroit Pistons	27	55	.329	34	94.0	99.1
Southeast Division						
Orlando Magic*	59	23	.720	—	102.8	95.3
Atlanta Hawks*	53	29	.646	6	101.7	97.0
Miami Heat*	47	35	.573	12	96.5	94.2
Charlotte Bobcats*	44	38	.537	15	95.3	93.8
Washington Wizards	26	56	.317	33	96.2	101.0

Western Conference	W	L	W-L%	GB	PS/G	PA/G
Northwest Division						
Denver Nuggets*	53	29	.646	—	106.5	102.4
Utah Jazz*	53	29	.646	—	104.2	98.9
Oklahoma City Thunder*	50	32	.610	3	101.5	98.0
Portland Trail Blazers*	50	32	.610	3	98.1	94.8
Minnesota Timberwolves	15	67	.183	38	98.2	107.8
Pacific Division						
Los Angeles Lakers*	57	25	.695	—	101.7	97.0
Phoenix Suns*	54	28	.659	3	110.2	105.3
Los Angeles Clippers	29	53	.354	28	95.7	102.1
Golden State Warriors	26	56	.317	31	108.8	112.4
Sacramento Kings	25	57	.305	32	100.0	104.4
Southwest Division						
Dallas Mavericks*	55	27	.671	—	102.0	99.3
San Antonio Spurs*	50	32	.610	5	101.4	96.3
Houston Rockets	42	40	.512	13	102.4	102.7
Memphis Grizzlies	40	42	.488	15	102.5	104.0
New Orleans Hornets	37	45	.451	18	100.2	102.7

Data for over 30,000 games were collected, along with league standings at that point in time.

Upon collection, team differential statistics (Δ deltas) were derived from individual team scores. Upon calculation of all derivative values, the following key data points became available:

Percentage Record Δ	Games Behind Δ	Playoffs (Binary)	Day of Year	Day of Week	Points Score per Game Δ	Season Wins Δ
Score Δ at Q1	Points Left Δ at Q1	Score Δ at Q2	Score Δ at Q3	Points Left Δ at Q3	Points Allowed per Game Δ	Season Losses Δ

In addition, the independent binary vairable of overtime was collected. At a glance, there is some variability based on certain categorical variables, but is generally reasonably constant.

Game Conditions	Probability of Overtime
Overall	**6.10%**
Regular Season	**6.00%**
Playoffs	**7.00%**
1990-1999	**5.80%**
2000-2009	**6.30%**
2010-Present	**6.20%**

The objective is then to construct a model that outperforms the baseline model of a flat 6.1% using the statistical data associated with the game.

Model Building and Preliminary Results

Three methods were used to construct predictive models: (1) Logistic regression, (2) Classification and Regression Trees (CART), and Random Forest Machine learning. For each method, four (4) models were constructed to accommodate each point in time, and the new variables available at that point in time.

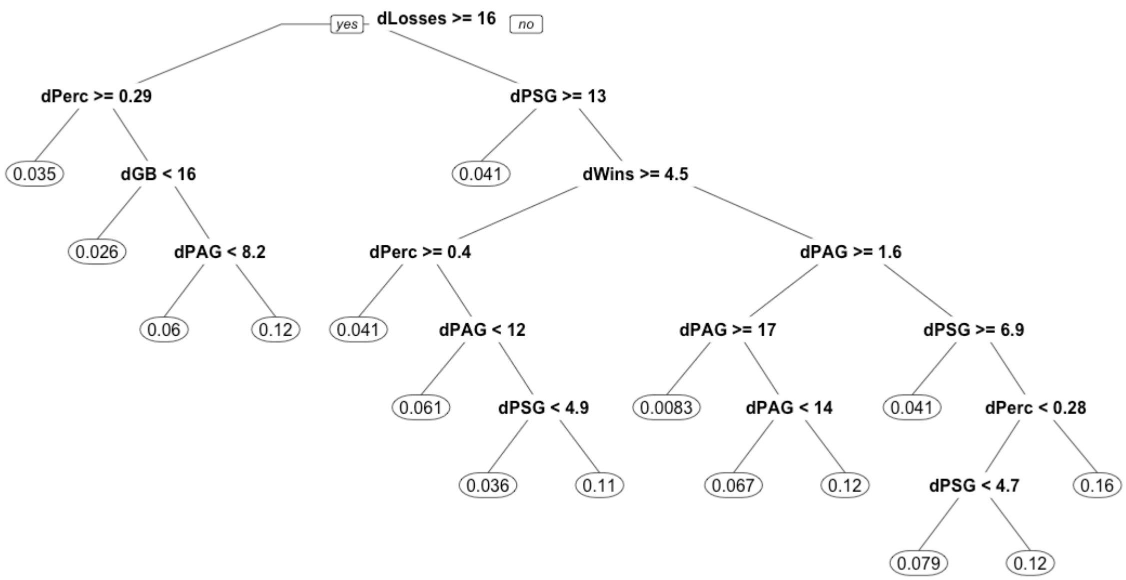
Logistic Regression

Prior to constructing a model to use, a rigorous test was performed to determine the variables with the greatest predictive value at any point in time (pre-game, after the 1st quarter, at the half, and after the 3rd quarter.) The pre-game variables found to be most valuable (based on standard deviation and test-train error composite scores) were: Win/Loss Percent Δ, Conference Games Behind Δ, and Playoff binary. During game time, the points Δ and points left Δ variables were added. With this model in place, the following variables are revealed to have these p values when run against the entire dataset:

Variable	Pre-Game	Q1	Q2	Q3
Percentage Record Δ	0.0135	0.0132	0.0316	0.281
Games Behind Δ	**0**	**0.00013**	**0.0001**	**0.003**
Playoffs (Binary)	0.2348	0.2292	0.2822	0.2855
Score Δ at Q1	X	**0**	0.4153	0.102
Points Left Δ at Q1	X	0.0358	0.3535	0.7719
Score Δ at Q2	X	X	**0**	0.334
Points Left Δ at Q2	X	X	0.0054	0.0246
Score Δ at Q3	X	X	X	**0**
Points Left Δ at Q3	X	X	X	0.5847

Classification and Regression Trees (CART)

Cart modeling was problematic at first due to its eagerness to always want to classify a game as "not overtime," given the probability. However, given an adjusted complexity parameter and all available variables, it became more insightful. A simplified example of the type of pre-game model it constructed is shown in this tree:

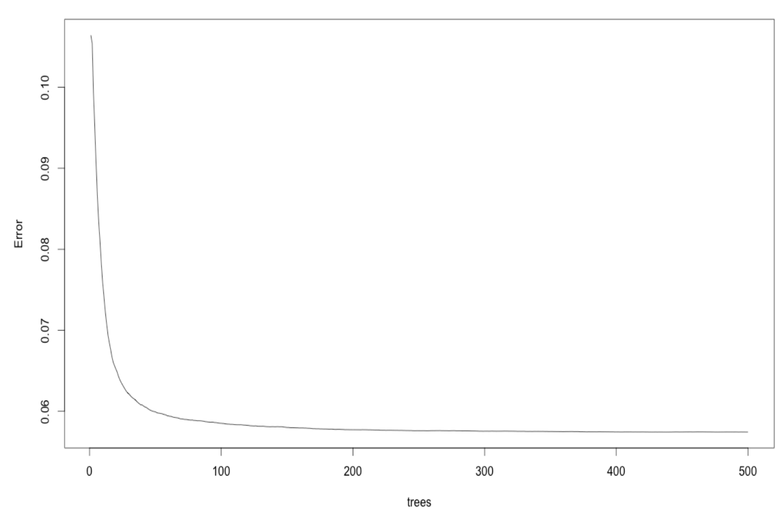


The actual model used a much lower complexity parameter, and had many more branches to follow.

Random Forest Machine Learning

The final method used is Random Forest machine learning, which input all variables, and produced a predicative model by creating a multitude of decision trees.

500 trees are created in the forest, and in the 3rd quarter model, Number of trees: 4 are variables tried at each split, with a mean of squared residuals of 0.057.



Comparative Findings

Each model was created on training data subset, and tested on an independent test set. A separate test set model was created AND run on the test set to provide a baseline. The difference between the baseline and the predicted value provided the test/train error value as an accuracy measure.

The four models across the three methods yielded the following results:

Model	T	Min	Median	Mean	Max	σ	Test/Train Error
Regression	Pre	1.7%	6.3%	6.1%	7.9%	0.9%	0.8%
Regression	Q1	-0.5%	6.3%	6.1%	9.6%	1.2%	1.1%
Regression	Q2	-7.3%	6.5%	6.1%	12.0%	2.4%	2.4%
Regression	Q3	-15.8%	6.9%	6.1%	12.8%	3.8%	3.8%
CART	Pre	0.0%	4.7%	6.1%	17.4%	3.9%	4.0%
CART	Q1	0.0%	4.9%	6.1%	17.0%	4.1%	4.3%
CART	Q2	0.0%	4.7%	6.1%	22.5%	4.8%	4.6%
CART	Q3	0.0%	3.8%	6.1%	22.9%	5.7%	5.9%
RandomForest	Pre	0.0%	5.6%	6.7%	78.3%	5.1%	5.1%
RandomForest	Q1	0.0%	6.0%	6.8%	74.6%	4.5%	4.4%
RandomForest	Q2	0.0%	6.3%	7.1%	51.7%	4.8%	4.5%
RandomForest	Q3	0.0%	6.2%	7.1%	50.7%	6.0%	5.9%

Generally, predictive power (articulated by the standard deviation: the spread of predictions) increases as time moves forward. The standard deviation does appear to be inversely correlated to the accuracy measure, however. After the 3rd quarter one model may predict as high as a 50% chance of overtime, but will have a greater chance of error.

Regression modeling is the most conservative, with the smallest spread of predictions, the smallest increase overtime, and also with the greatest accuracy.

Both tree-modeling methods perform similarly, with CART being slightly more accurate. See appendix 1 and 2 for comparative probability distributions over time.

Case in Point

On November 2, 2009, the Memphis Grizzlies played the Sacramento Kings. It was early in the season, and they were a well matched team. It was a close game.

	MEM	SAC	P(OT)
Wins	1	1	7.02%
Losses	3	3	
Percentage	25%	25%	
Games Behind	2	2.5	
Avg. Points Scored/Game	107	100.5	
Avg. Points Allowed/Game	115.8	107	
1st Quarter Score	**26**	**33**	**7.27%**
Halftime Score	**56**	**54**	**34.08%**
3rd Quarter Score	**79**	**80**	**44.38%**

By the end of the 3rd, the score was 79-80, and the Random Forest model was predicting a 44% chance of overtime—far more likely than the baseline 6.1%. With less than two seconds left, the Kings tied the game at 110 points, sending it into overtime.

Ad buyers can't say they weren't warned.

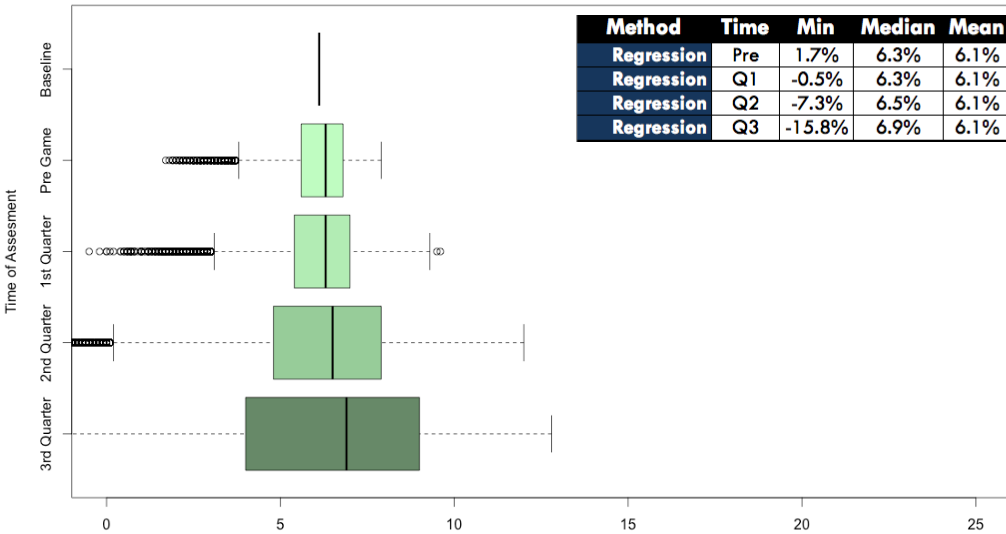
Conclusion and Recommendations

Using a mix of these models, TV networks can continuously be predicting the probability of overtime, and constantly reevaluating the acceptable bid price for a post-game time slot.

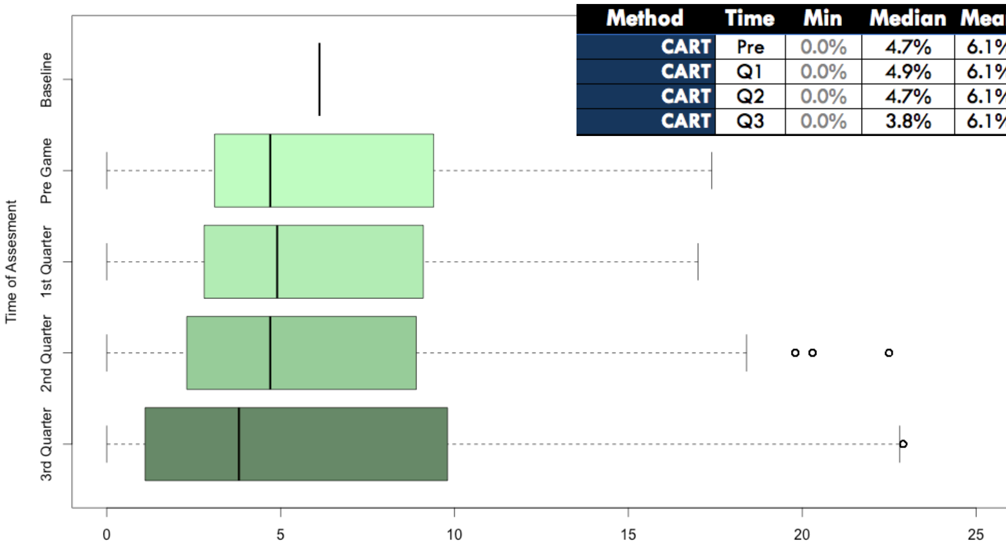
Using a balance of the standard deviation and accuracy measures (depending on risk appetite or other market factors) networks can create a model that capitalizes on the economic value produced by overtime before the time to make the decision has already passed.

Appendix 1: Probability Distributions Over Time, by Method

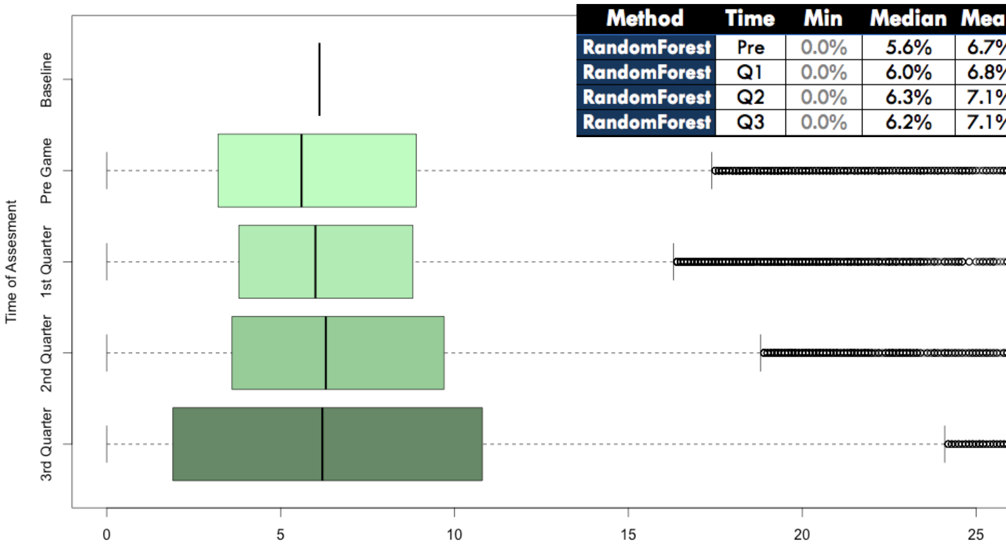
Method	Time	Min	Median	Mean	Max	SdtDv	Test/Train Error
Regression	Pre	1.7%	6.3%	6.1%	7.9%	0.9%	0.8%
Regression	Q1	-0.5%	6.3%	6.1%	9.6%	1.2%	1.1%
Regression	Q2	-7.3%	6.5%	6.1%	12.0%	2.4%	2.4%
Regression	Q3	-15.8%	6.9%	6.1%	12.8%	3.8%	3.8%



Method	Time	Min	Median	Mean	Max	SdtDv	Test/Train Error
CART	Pre	0.0%	4.7%	6.1%	17.4%	3.9%	4.0%
CART	Q1	0.0%	4.9%	6.1%	17.0%	4.1%	4.3%
CART	Q2	0.0%	4.7%	6.1%	22.5%	4.8%	4.6%
CART	Q3	0.0%	3.8%	6.1%	22.9%	5.7%	5.9%

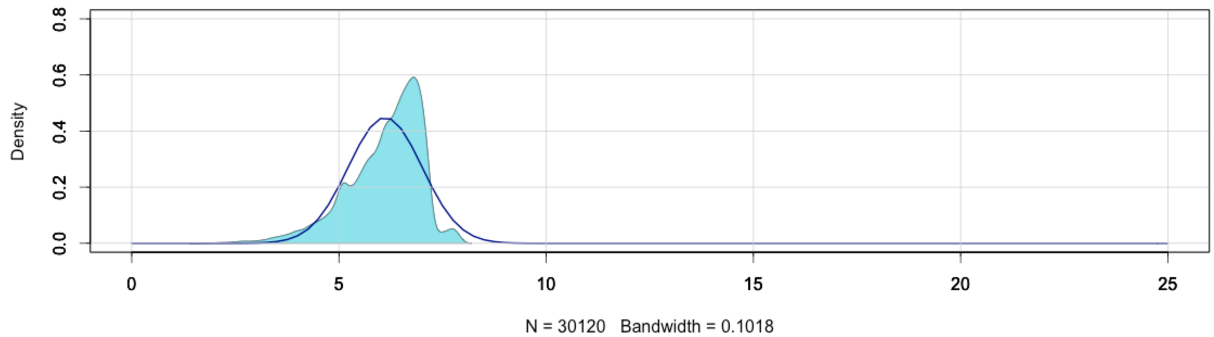


Method	Time	Min	Median	Mean	Max	SdtDv	Test/Train Error
RandomForest	Pre	0.0%	5.6%	6.7%	78.3%	5.1%	5.1%
RandomForest	Q1	0.0%	6.0%	6.8%	74.6%	4.5%	4.4%
RandomForest	Q2	0.0%	6.3%	7.1%	51.7%	4.8%	4.5%
RandomForest	Q3	0.0%	6.2%	7.1%	50.7%	6.0%	5.9%

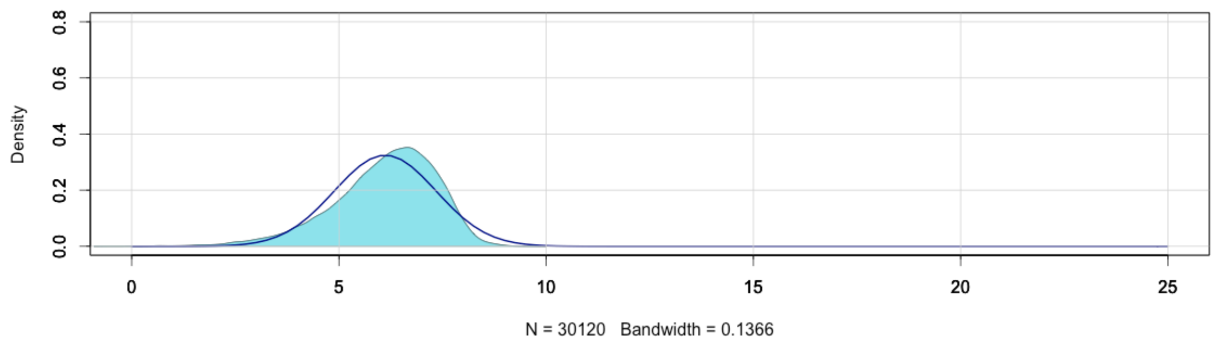


Appendix 2: Probability Distributions for Logistic Regression

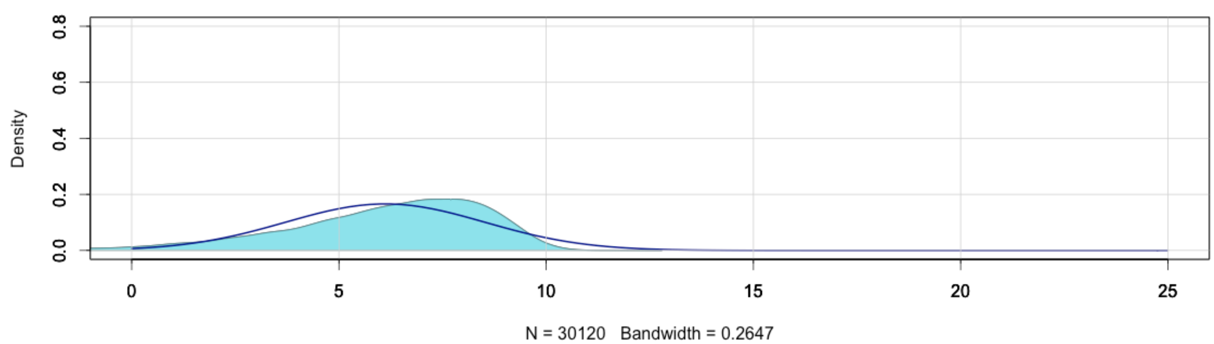
Pre-Game



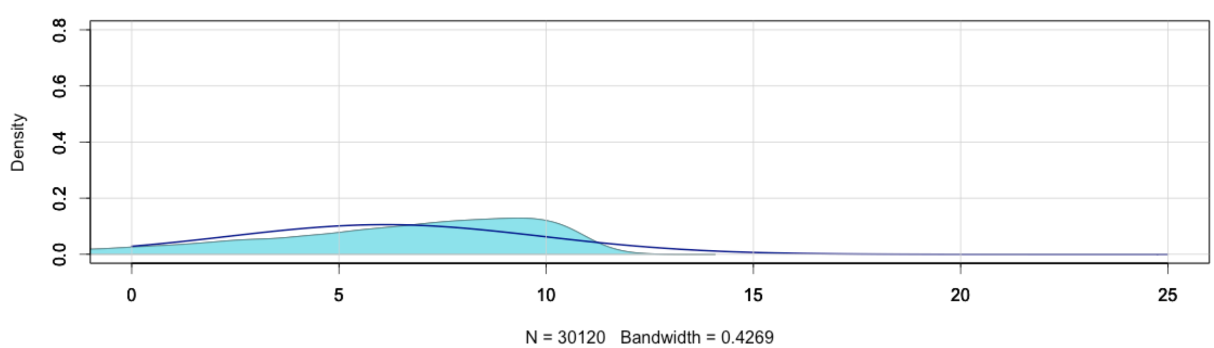
Q1



Q2

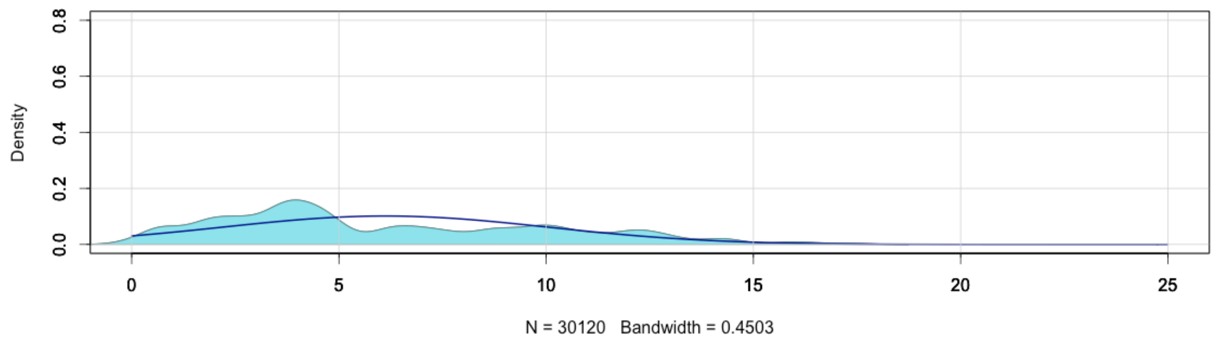


Q3

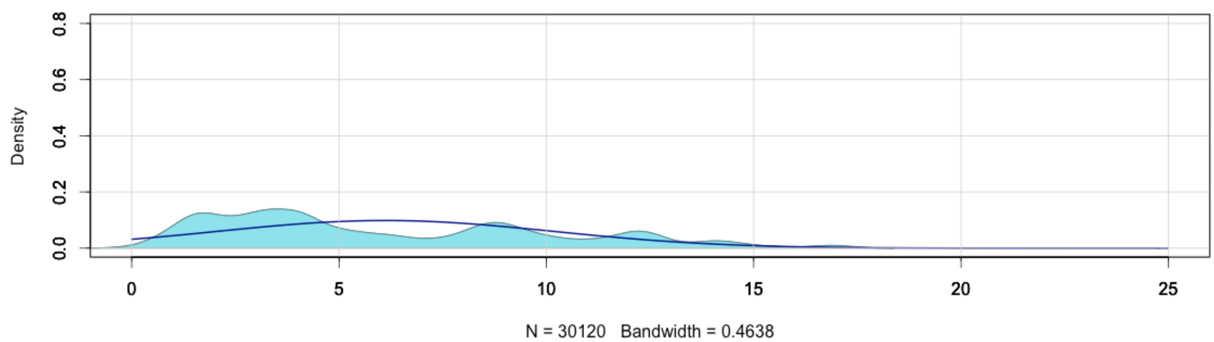


Appendix 3: Probability Distributions for CART

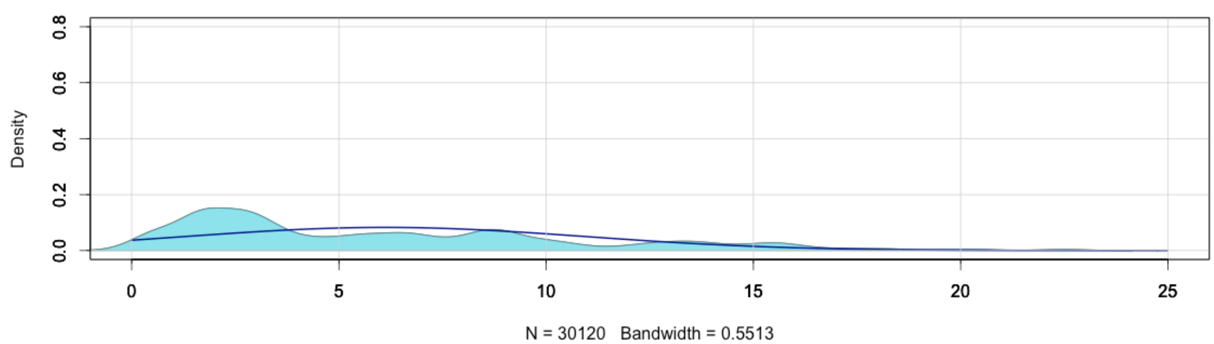
Pre-Game



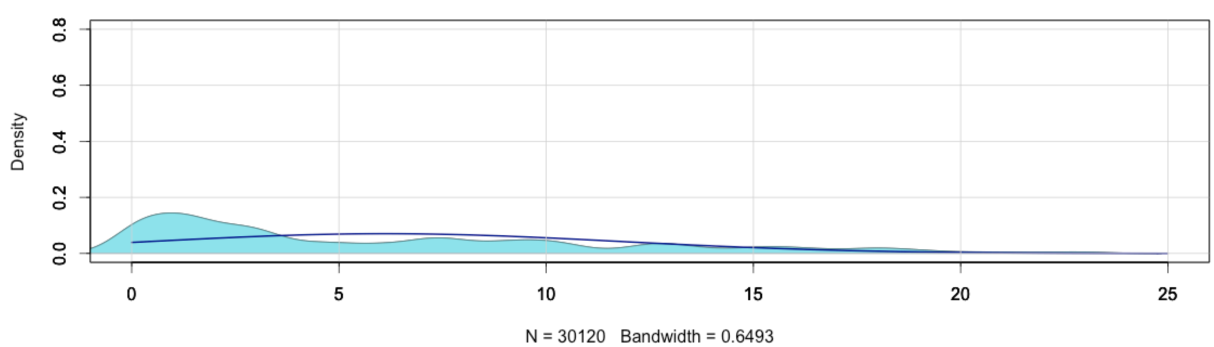
Q1



Q2

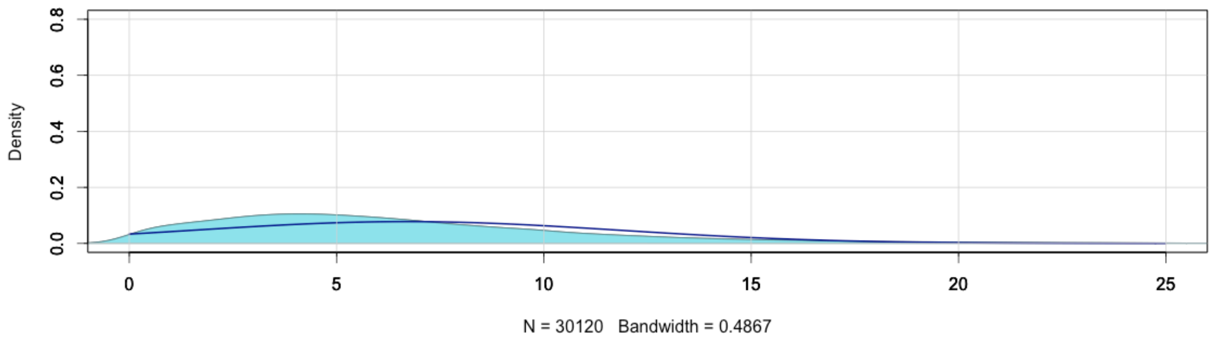


Q3

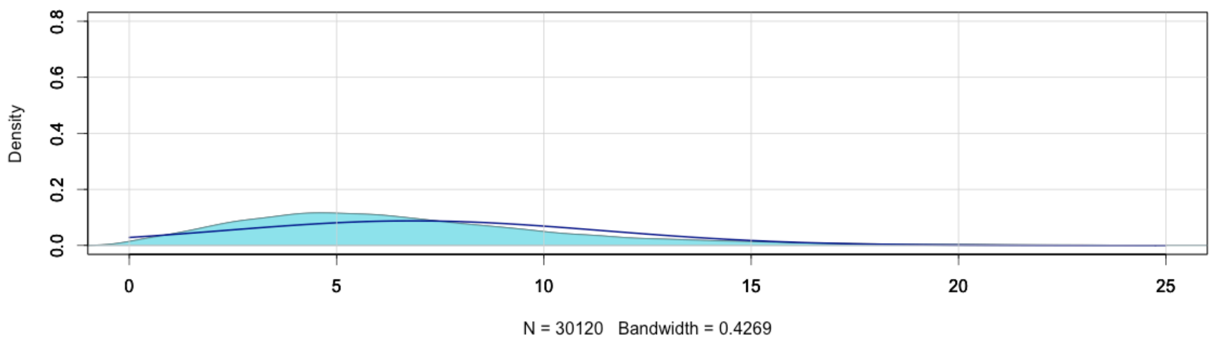


Appendix 4: Probability Distributions for Random Forest

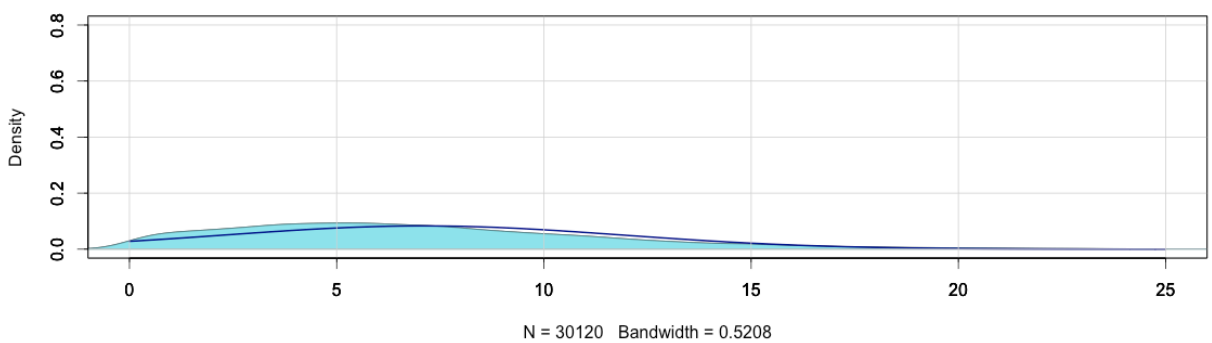
Pre-Game



Q1



Q2



Q3