Predicting the Covering of NCAA Football Spreads

Matt Hinger
EECS 349
Northwestern University
Evanston, IL 60201, USA

Abstract

The goal of my research was to look at the relationship between different aspects of college football game and the covering of the spread. In the world of college football, it is hard to predict the outcomes of games due to the random nature of sports. Anything can happen at any time. I wanted to determine if there was any way to predict whether or not the spread of certain game would be covered, or if college football were truly random. I approached the project by looking at data from both the 2009 and 2010 football seasons and in four phases. Each phase used the same set of machine learning techniques but the data sets used to both train and test were tweaked each time to incorporate different attributes or leave out certain data elements. After running the models on the different data sets, the results were somewhat surprising. I was able to achieve multiple models in the high 40's percent predicted correctly area as well as four models achieving over 50% predicted correctly

Introduction

The thought going into this research was that it would be very interesting if you would be able to develop a model capable of accurately predicting if the spread on a college football game would be covered. My main motivation for this project is the fact that I find college football so fascinating and I am an avid fan of the sport. I wanted to look at the raw data from a large amount of games and determine what aspects of the game, if any, played a role in or influenced the outcome of games. Those aspects ranged from rankings, to the week of the season, to general statistics of each team. It would be interesting to note if certain features of the game had more influence in the outcome of the event than is evident to the outside observer, or if college football were truly random. Therefore the goal of the project is to find out what features play the most important role in the prediction of games covering the spread.

Methodology

I first began my research by gathering a large amount of data on games. I used the websites found in the Resources section as my main tools for gathering data. For my training data set, I only used games in which both teams were FBS teams. This led to my data set being 679 games from the 2009 season. My testing data consisted of the 682 games from the 2010 season in which both teams were FBS. I chose to use games with only FBS teams because the Las Vegas spreads often left out games with non-FBS opponents and also the fact that those games tend to be blowouts which could have influenced or even biased my data set. The features that I used were as follows: Week, Avg. Rush Attempts Offense, Avg. Rush Yards Offense, Avg. Pass Attempts Offense, Avg. Pass Completions Offense, Avg. Pass Yards Offense, Avg. Interceptions Offense, Avg. Fumbles Offense, Avg. Rush Attempts Defense, Avg. Rush Yards Defense, Avg. Pass Attempts Defense, Avg. Pass Completions Defense, Avg. Pass Yards Defense, Avg. Interceptions Defense, Avg. Fumbles Defense, Avg. Points Scored (these statistics were used for both the home and away teams), Neutral Site, Las Vegas Line. In later

phases of my research I added the Anderson & Hester computer rankings (available starting on week 6 games). I chose these computer rankings as opposed to the coaches' poll because it gave me rankings for all 120 FBS teams. I chose it over other computer rankings because the rankings did not reward teams for running up scores (that is to say margin of victory is not considered), the rankings don't prejudice teams (they look at actual accomplishments), and these rankings compute the most accurate strength of schedule ratings. I chose to use cumulative averages for the statistical elements, that is to say for week-1 games there were no statistics and for week-3 games the statistics were the average of the previous two weeks. I felt that this gave a better indicator of how teams were performing and could help decrease the effect of blowout wins on my results.

I focused my analysis on the following learning algorithms: J48 Decision Trees, Random Forest, Bayes Net, Rotation Forest and ZeroR (used a metric for performance). After phases 1 and 2, I ended up dropping Bayes Net and Rotation Forest my model set because they were only returning results that were equal with ZeroR at best. Overall my different J48 models performed the best on average for every phase.

The design of the different phases of my project was done with the intention of determining which features had the greatest impact on the results that were being generated. For all phases I used the same metrics for each of the model types. The only thing that changed from phase to phase was the data sets and features used. The models I used were as follows: ZeroR, J48 –C 0.25 –M 2, J48 –C 0.15 –M 3, J48 –C 0.2 –M 2, Random Forest –I 10 –K 0 –S 1, Random Forest –I 25 –K 10 –S 1, Bayes Net, Rotation Forest. Phase 1 was looking at the entire data set for training and testing. Therefore it included all weeks and all features (with the exception of computer rankings). Phase 2 was looking at all features (with the exception of computer rankings) and weeks 2-14. I left out week 1 games because the statistics for those games were 0, and therefore it seemed to be hurting my performance since it was just "guessing" for those games. Phase 3 included all features plus computer rankings and weeks 2-14. Phase 4 was looking at all features including computer rankings and weeks 6-14. This was done to see if rankings played an important role, since they're only available week 6 to the end of the season.

Results

After completing all phases, I achieved some surprising results. Bayes nets and Rotation Forest achieved ZeroR performance at best for all four phases. The best performing models were the J48 decision trees. The random forest models' performance varied depending on the phase. Phase 3 yielded the best overall results when I used the J48 –C 0.15 –M 3 model. I was able to achieve 51.4774% predicted correctly, which I consider to be extremely good because of the fact that the spreads are supposed to yield 50/50 results. The ordering of the phases by their best result is as follows: Phase 3, Phase 2, Phase 1, and Phase 4. Using this information, I was able to come to the conclusion that week-1 games, since they have no values for the statistical attributes, tended to be a random guess by the machine learning algorithms. By eliminating week-1 games from the data set, I saw an increase of roughly 1.5% predicted correctly between the best models of Phase 1 and Phase 2. Using the knowledge that eliminating week-1 games improved performance, I decided to add computer rankings to the mix. This further improved the best performance by roughly .15% predicted correctly when compared to the best model from Phase 2. Phase 4 did not perform nearly as well as anticipated, possibly due to the elimination of a large chunk of the training data by not including weeks 1-5. By looking at the decision tree of the best model, I was able to see that the week of the game was the first attribute split on. It split

on less than or equal to week 12 and greater than week 12, which could be due to the fact that later on in the season teams tend to be playing more seasoned and better teams as their schedule gets tougher with rivalry games and conference championships. I also found it surprising that whether or not a game was played on a neutral site only came into play very deep in the decision tree and only on the greater than week 12 branch.

For a table of results for each phase see the Appendix.

Future Work

Due to the fact that I achieved promising results on multiple models from various attempts, the logical next steps would be to continue tinkering with the features that are used. Inclusion of aspects of the game such as injuries, star players, weather, recruiting, coach record vs. the team, amount of junior/senior starters, etc. I am also going to look at the 2010 bowl games and run my best performing models on the games and see how they perform.

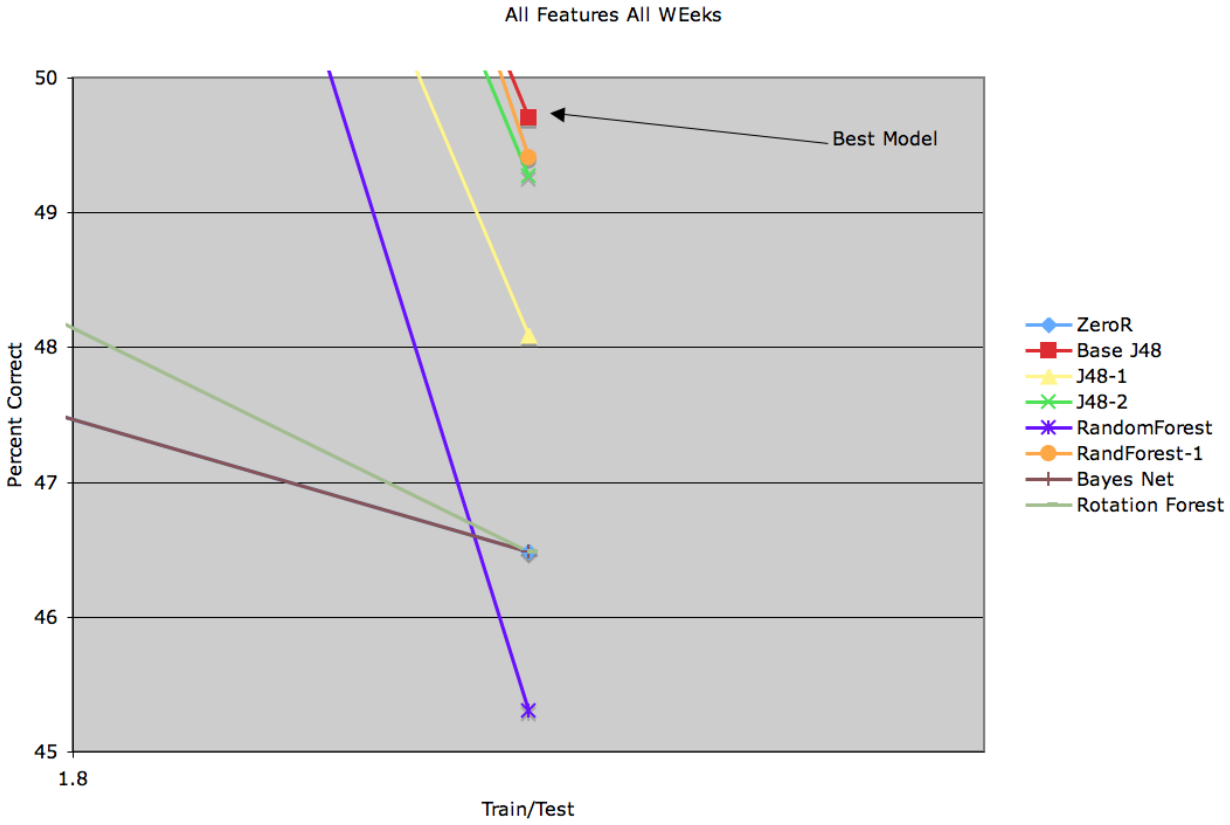
Resources

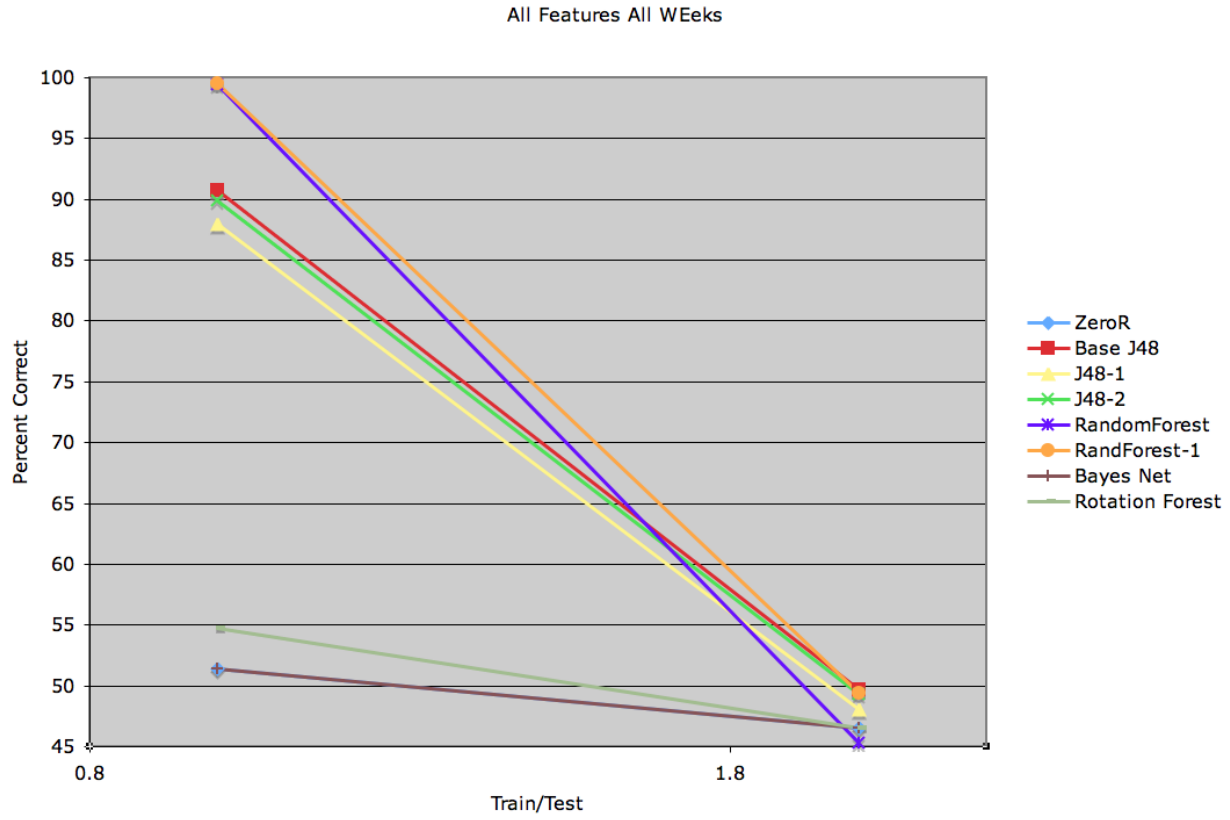
Anderson & Hester Computer Rankings: http://andersonsports.com/football/ACF_frnk.html
ESPN: http://espn.com
The Sunshine Forecast: http://www.repole.com/sun4cast/data.html

Appendix

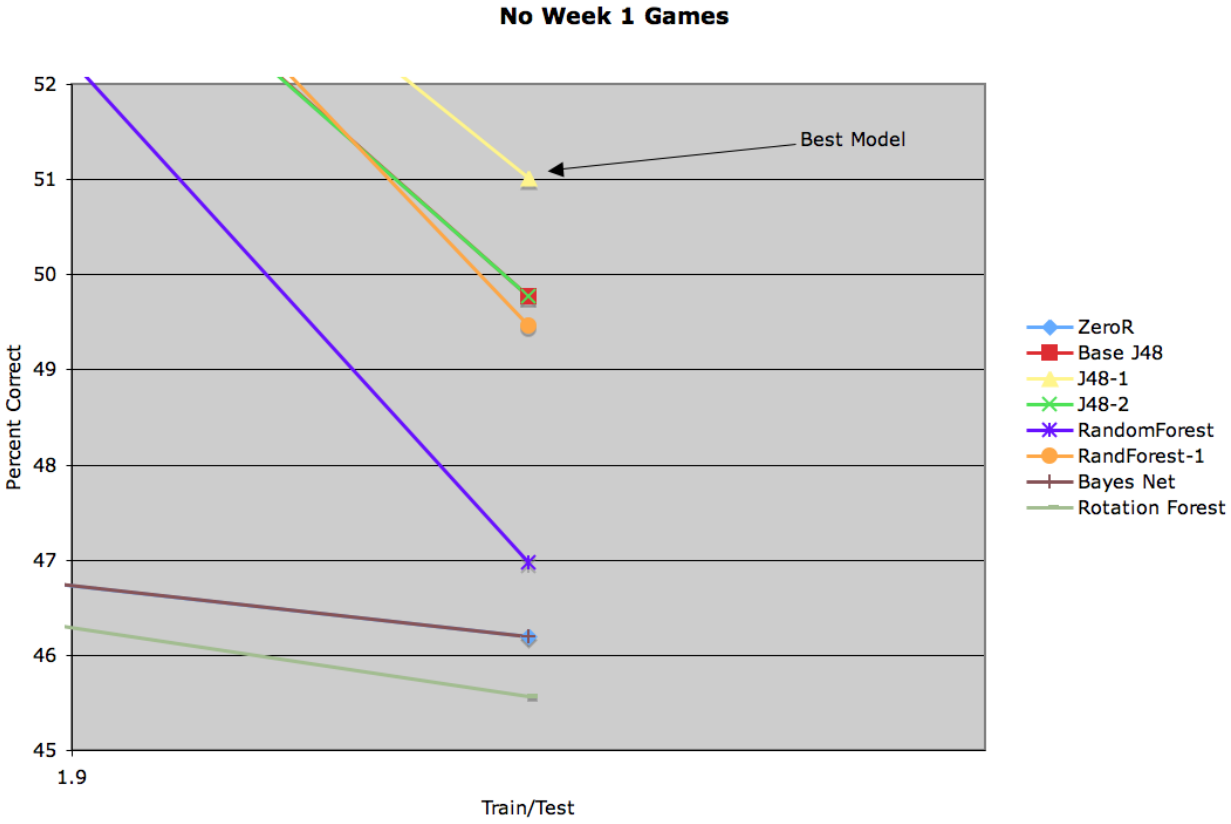
Percent Correct for Various Models on Test Data		
Phase 1 - All Base Features, All Weeks		
J48 -C 0.25 -M 2		49.7067
Random Forest -I 25 -K 10 -S 1		49.4135
J48 -C 0.2 -M 2		49.2669
J48 -C 0.15 -M 3		48.0938
ZeroR		46.4809
Bayes Net		46.4809
Rotation Forest		46.4809
Random Forest -I 10 -K 0 -S 1		45.3079
Phase 2 - All Base Features, No Week 1 Games		
J48 -C 0.15 -M 3		51.0109
J48 -C 0.25 -M 2		49.7667
J48 -C 0.2 -M 2		49.7667
Random Forest -I 25 -K 10 -S 1		49.4557
Random Forest -I 10 -K 0 -S 1		46.9673
ZeroR		46.1897
Bayes Net		46.1897
Rotation Forest		45.5677
Phase 2 - All Base Features, No Week 1 Games		
J48 -C 0.15 -M 3		51.4774
J48 -C 0.25 -M 2		51.1664
J48 -C 0.2 -M 2		51.1664
Random Forest -I 10 -K 0 -S 1		48.0560
Random Forest -I 25 -K 10 -S 1		47.2784
ZeroR		46.1897
Phase 2 - All Base Features, No Week 1 Games		
J48 -C 0.15 -M 3		48.5459
J48 -C 0.2 -M 2		48.5459
J48 -C 0.25 -M 2		48.0984
Random Forest -I 25 -K 10 -S 1		47.4273
ZeroR		46.7562
Random Forest -I 10 -K 0 -S 1		44.0716

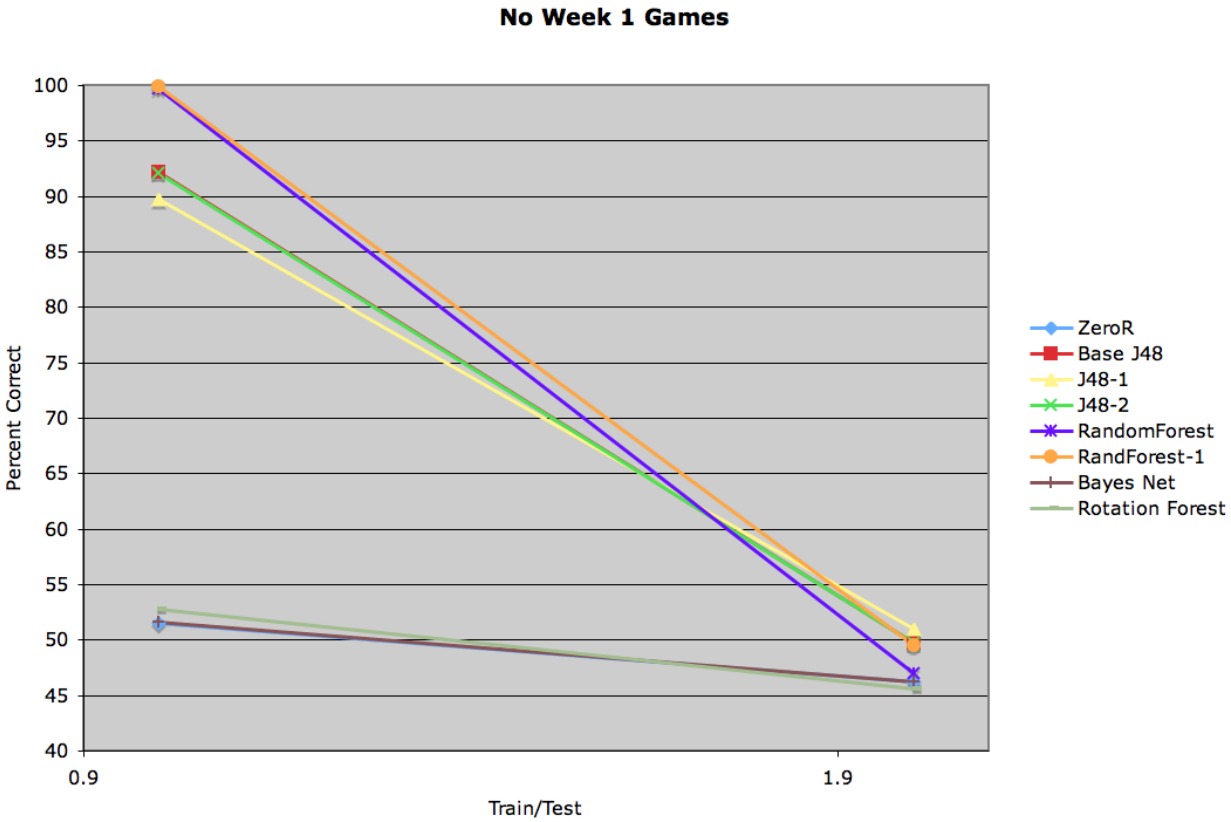
PHASE 1:


All Features All WEeks
Best Model
Percent Correct
50
49
48
47
46
45
1.8
Train/Test
ZeroR
Base J48
J48-1
J48-2
RandomForest
RandForest-1
Bayes Net
Rotation Forest

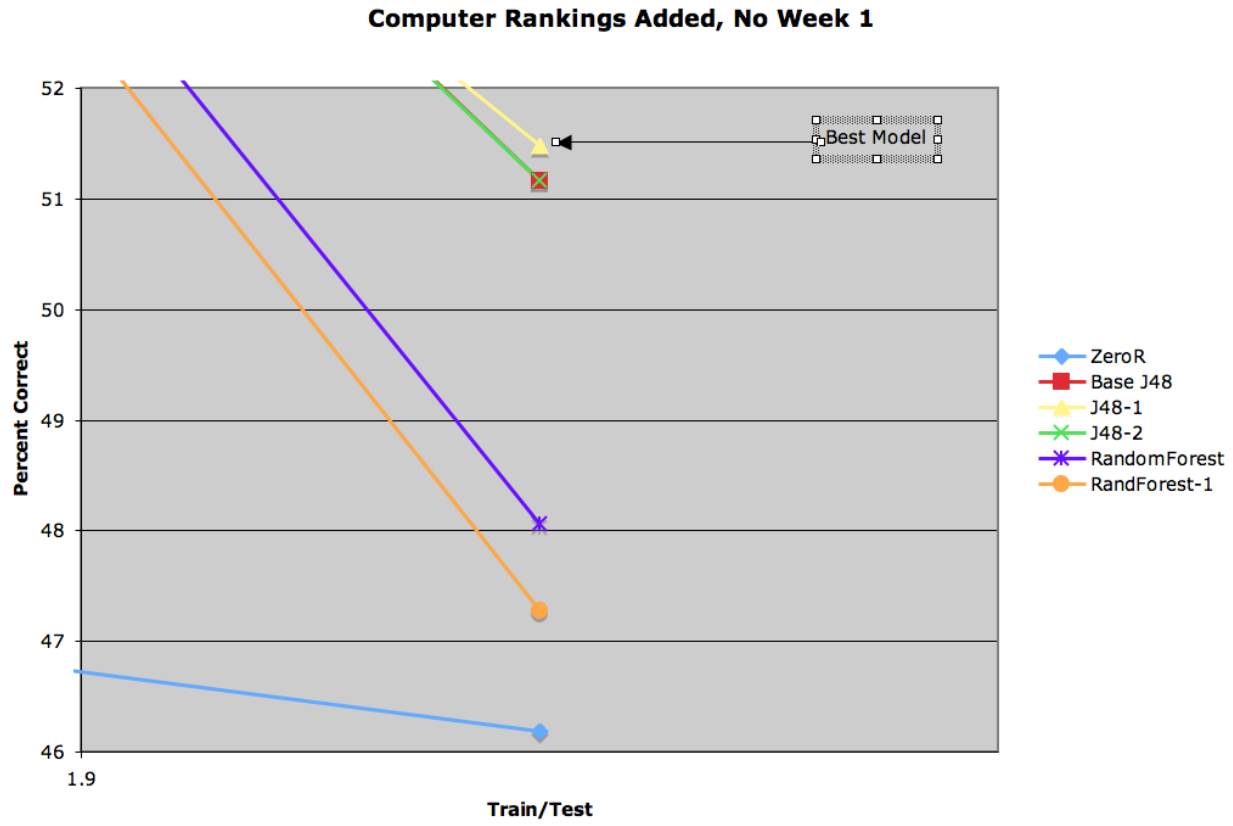

All Features All WEeks
Percent Correct
100
95
90
85
80
75
70
65
60
55
50
45
0.8
1.8
Train/Test
ZeroR
Base J48
J48-1
J48-2
RandomForest
RandForest-1
Bayes Net
Rotation Forest

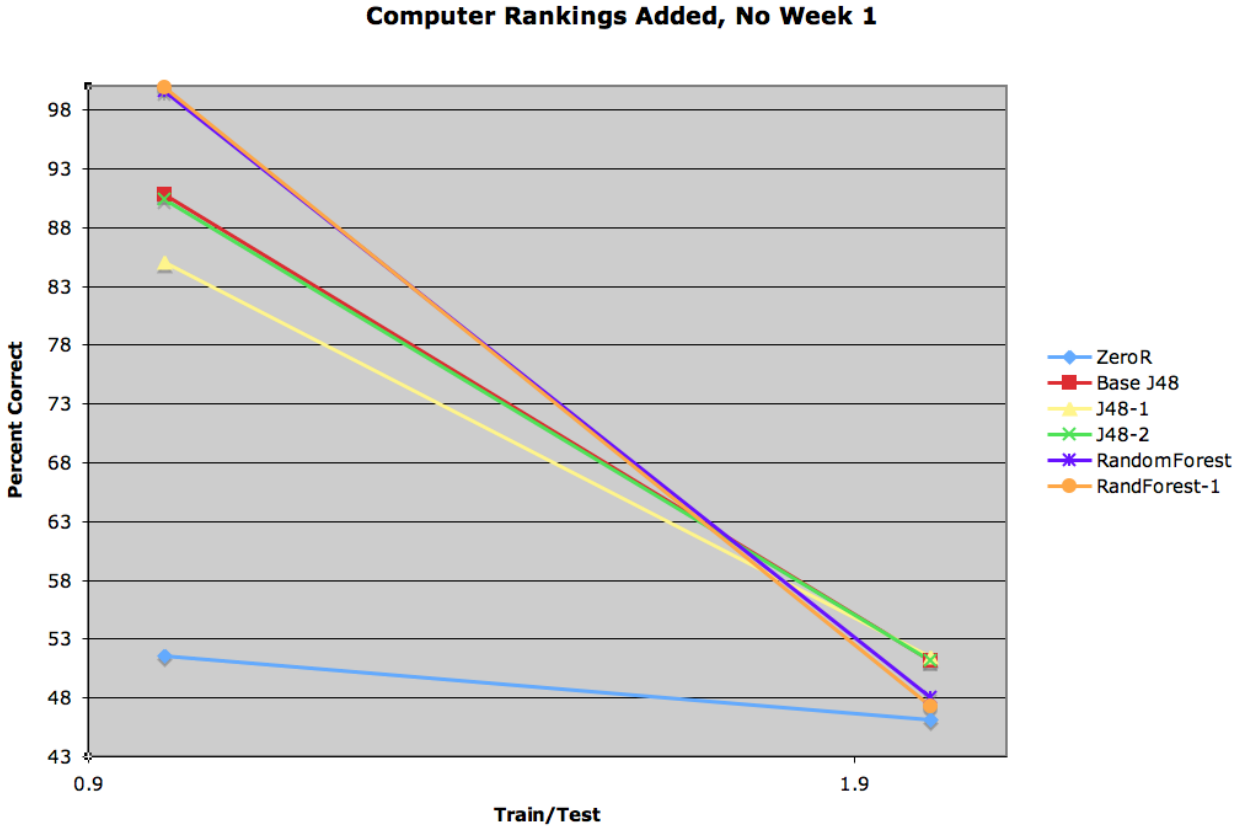
PHASE 2:


No Week 1 Games
Best Model
Percent Correct
52
51
50
49
48
47
46
45
1.9
Train/Test
ZeroR
Base J48
J48-1
J48-2
RandomForest
RandForest-1
Bayes Net
Rotation Forest

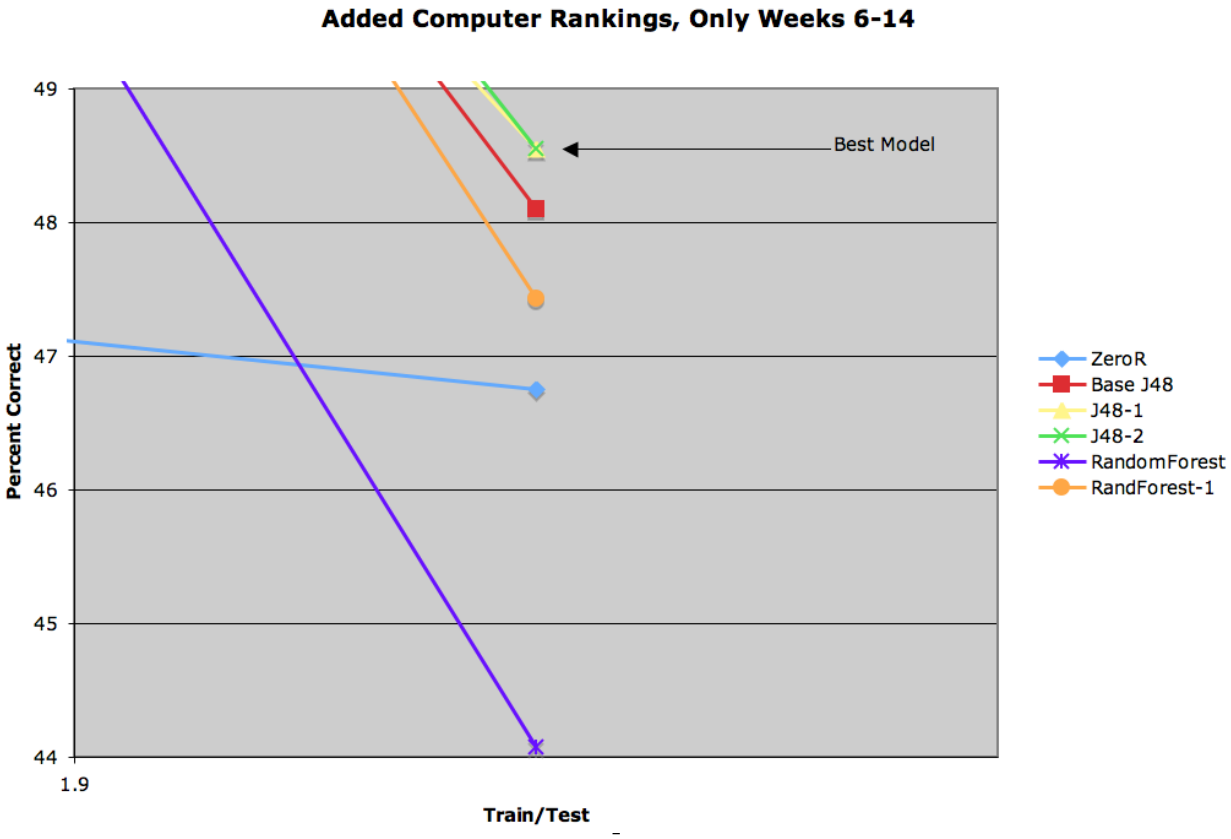

No Week 1 Games
Percent Correct
100
95
90
85
80
75
70
65
60
55
50
45
40
0.9
1.9
Train/Test
ZeroR
Base J48
J48-1
J48-2
RandomForest
RandForest-1
Bayes Net
Rotation Forest

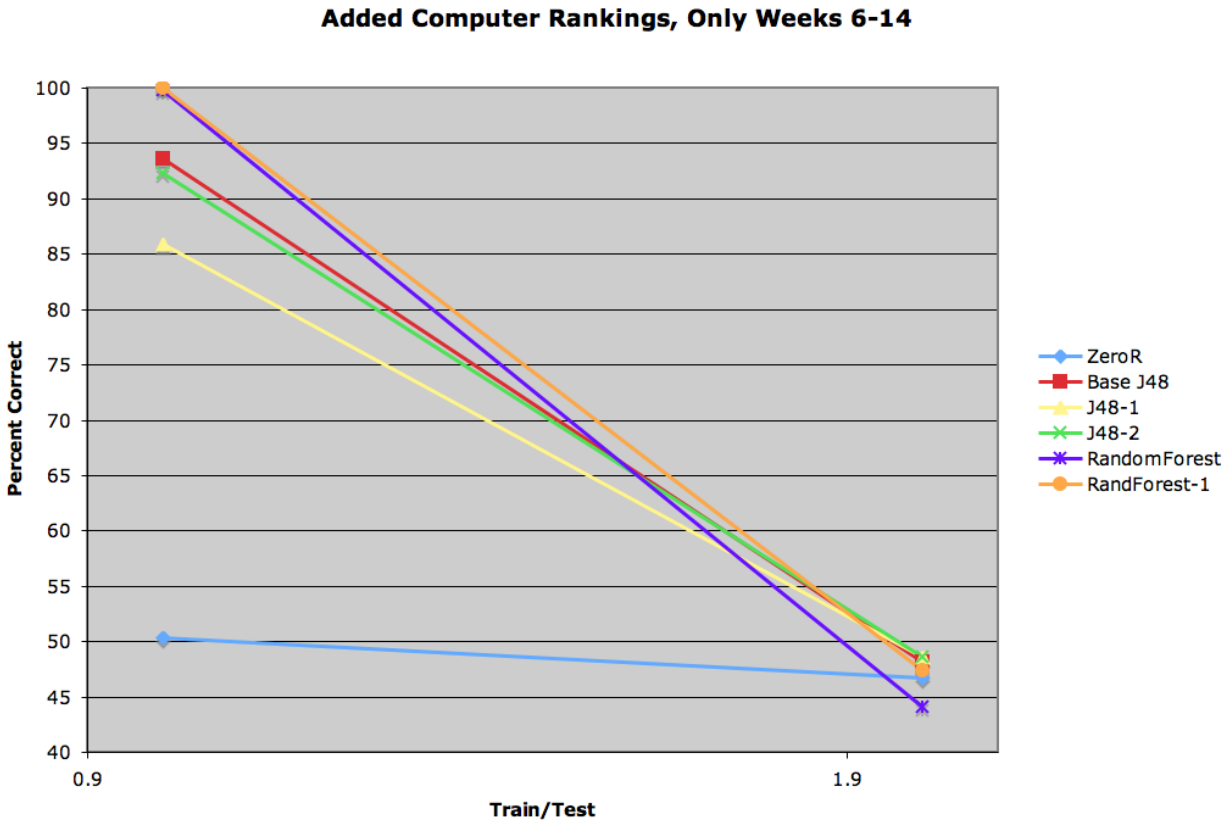
PHASE 3:


Computer Rankings Added, No Week 1
Best Model
Percent Correct
52
51
50
49
48
47
46
1.9
Train/Test
ZeroR
Base J48
J48-1
J48-2
RandomForest
RandForest-1


Computer Rankings Added, No Week 1
Percent Correct
98
93
88
83
78
73
68
63
58
53
48
43
0.9
1.9
Train/Test
ZeroR
Base J48
J48-1
J48-2
RandomForest
RandForest-1

PHASE 4:


Added Computer Rankings, Only Weeks 6-14
Best Model
Percent Correct
49
48
47
46
45
44
1.9
Train/Test
ZeroR
Base J48
J48-1
J48-2
RandomForest
RandForest-1


Added Computer Rankings, Only Weeks 6-14
Percent Correct
100
95
90
85
80
75
70
65
60
55
50
45
40
0.9
1.9
Train/Test
ZeroR
Base J48
J48-1
J48-2
RandomForest
RandForest-1