Estimating Win Probabilities for College Football Teams Ranked in the AP Poll

Ryan Morgan
December 13th, 2017

Introduction

- Since 1936, the Associated Press has weekly released rankings of the top college football teams in the country in the AP Poll

- The team viewed to be the "best" is ranked 1, the team viewed to be the second best is ranked 2, etc.

- Note: A "higher" ranking means the ranking is numerically higher. This means the "better" teams have lower rankings

- We wanted to see what the relationship was between rankings in the AP Poll and estimated win probabilities

Example *Sports-Reference* Schedule Table

G	Date	Time	Day	School		Opponent	Conf		Pts	Opp	W	L	Streak	TV	Notes
1	Sep 3, 2016	8:00 PM	Sat	(1) Alabama	N	(20) USC	Pac-12	W	52	6	1	0	W 1	ABC	
2	Sep 10, 2016	3:30 PM	Sat	(1) Alabama		Western Kentucky	CUSA	W	38	10	2	0	W 2	ESPN2	
3	Sep 17, 2016	3:30 PM	Sat	(1) Alabama	@	(19) Ole Miss	SEC	W	48	43	3	0	W 3	CBS	
4	Sep 24, 2016	12:00 PM	Sat	(1) Alabama		Kent State	MAC	W	48	0	4	0	W 4		
5	Oct 1, 2016	7:00 PM	Sat	(1) Alabama		Kentucky	SEC	W	34	6	5	0	W 5		
6	Oct 8, 2016	7:00 PM	Sat	(1) Alabama	@	(16) Arkansas	SEC	W	49	30	6	0	W 6		
7	Oct 15, 2016	3:30 PM	Sat	(1) Alabama	@	(9) Tennessee	SEC	W	49	10	7	0	W 7		
8	Oct 22, 2016	3:30 PM	Sat	(1) Alabama		(6) Texas A&M	SEC	W	33	14	8	0	W 8		
9	Nov 5, 2016	8:00 PM	Sat	(1) Alabama	@	(15) LSU	SEC	W	10	0	9	0	W 9		
10	Nov 12, 2016	12:00 PM	Sat	(1) Alabama		Mississippi State	SEC	W	51	3	10	0	W 10		
11	Nov 19, 2016	7:00 PM	Sat	(1) Alabama		Chattanooga	Non-Major	W	31	3	11	0	W 11		
12	Nov 26, 2016	3:30 PM	Sat	(1) Alabama		(16) Auburn	SEC	W	30	12	12	0	W 12		
13	Dec 3, 2016	4:00 PM	Sat	(1) Alabama	N	(15) Florida	SEC	W	54	16	13	0	W 13		SEC Championship Game
14	Dec 31, 2016	3:00 PM	Sat	(1) Alabama	N	(4) Washington	Pac-12	W	24	7	14	0	W 14		Peach Bowl
15	Jan 9, 2017	8:00 PM	Mon	(1) Alabama	N	(3) Clemson	ACC	L	31	35	14	1	L 1	ESPN	College Football Championship

Constructing the Data Set

- Schedules were scraped for every team and every season available on the *Sports-Reference* website
- The AP Poll has been ranking the top 25 since 1989, so only games from 1989 through 2016 were considered
  - From 1968 to 1988, the top 20 teams were ranked
- Only games between two ranked teams were considered
- Each game should only be recorded once in the data set

Constructing the Data Set

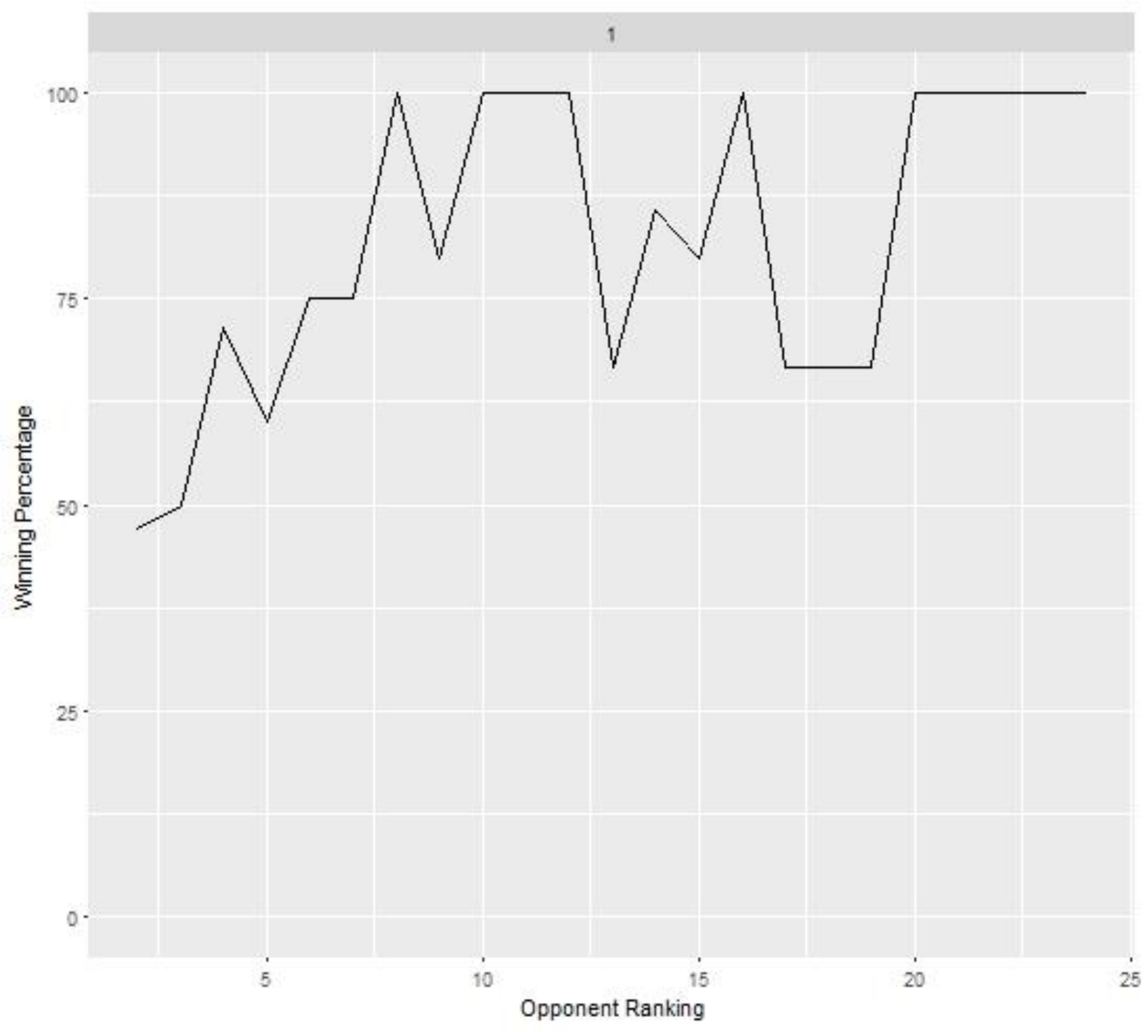
- Each game should only be recorded once in the data set
  - Example: The 2016 game between Alabama and Clemson was originally in the data set twice; once from Alabama's point of view (A loss where the Opponent was Clemson) and once from Clemson's point of view (A win where the Opponent was Alabama)
- Games were filtered down to only include games where the Opponent's name came before the Team's name alphabetically
  - Example: "Alabama" comes before "Clemson" alphabetically, so the game was only included from Clemson's point of view (where the Opponent was Alabama)
- The "Team" and "Opponent" designation for each game is arbitrary

Constructing the Data Set

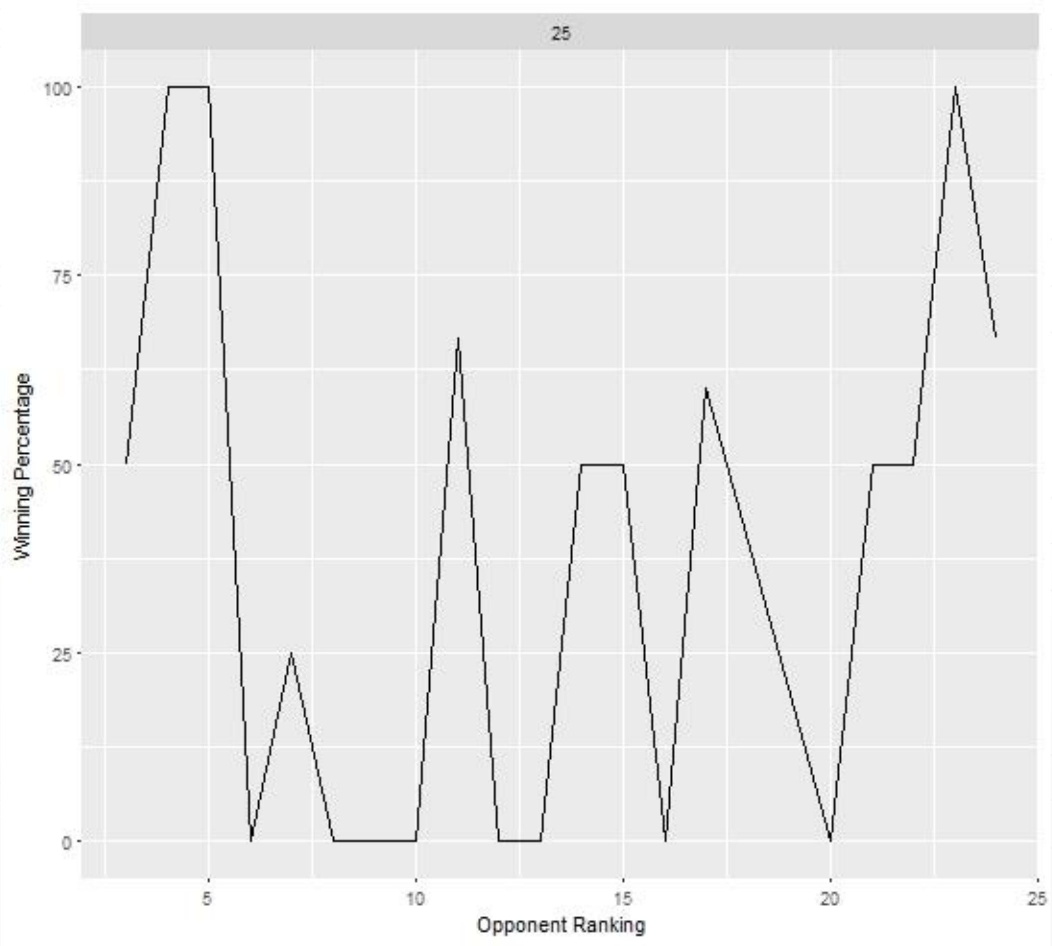
- Since 1989, there have been 1494 games between two teams ranked in the AP Poll

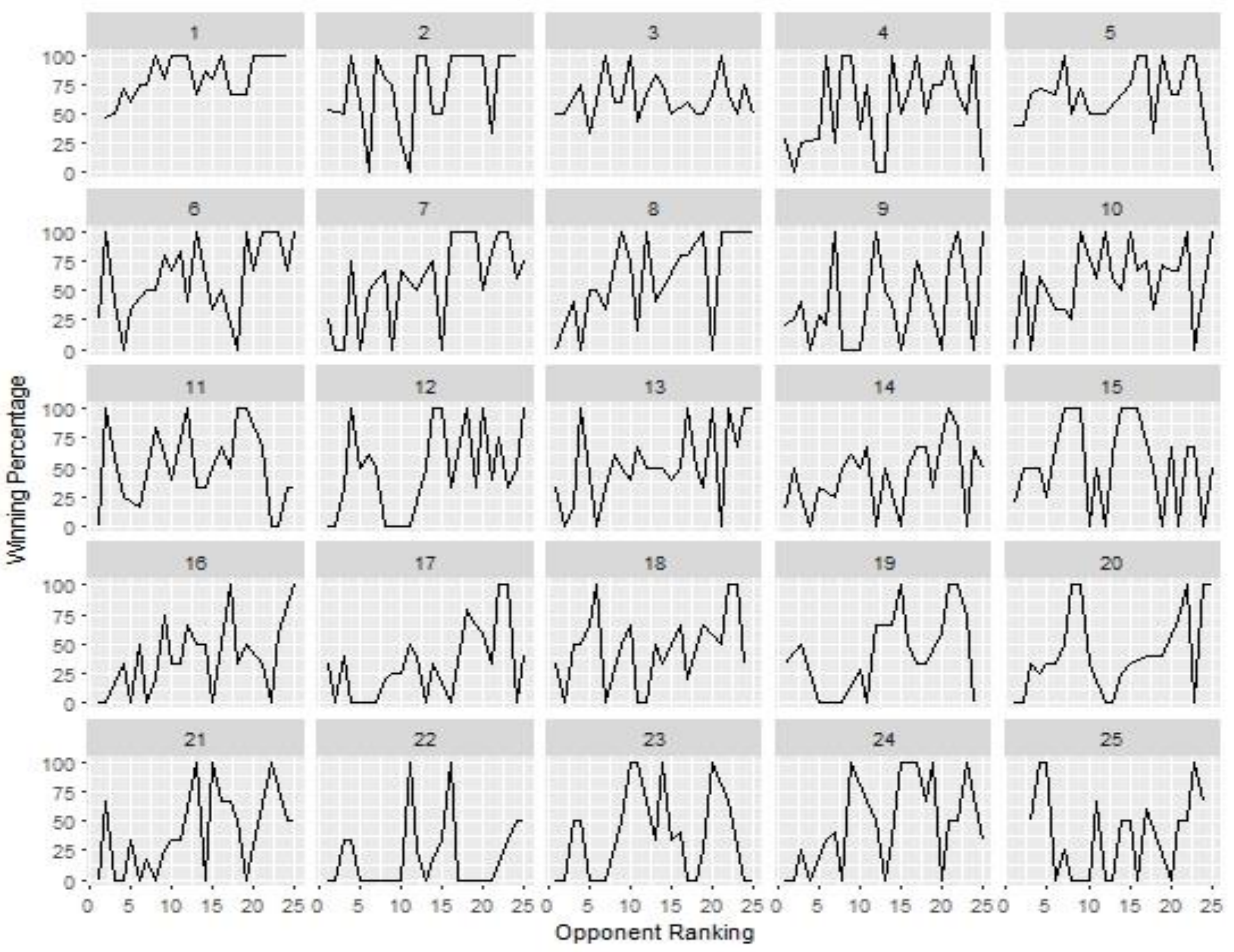
- Our data set has information on the location, the result, the ranks of the teams, the points scored by both teams, the game number, whether a bowl game, and whether a conference game for each of these games

Winning Percentage of Teams Ranked 1 in the AP Poll



Winning Percentage of Teams Ranked 25 in the AP Poll





1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Winning Percentage
Opponent Ranking

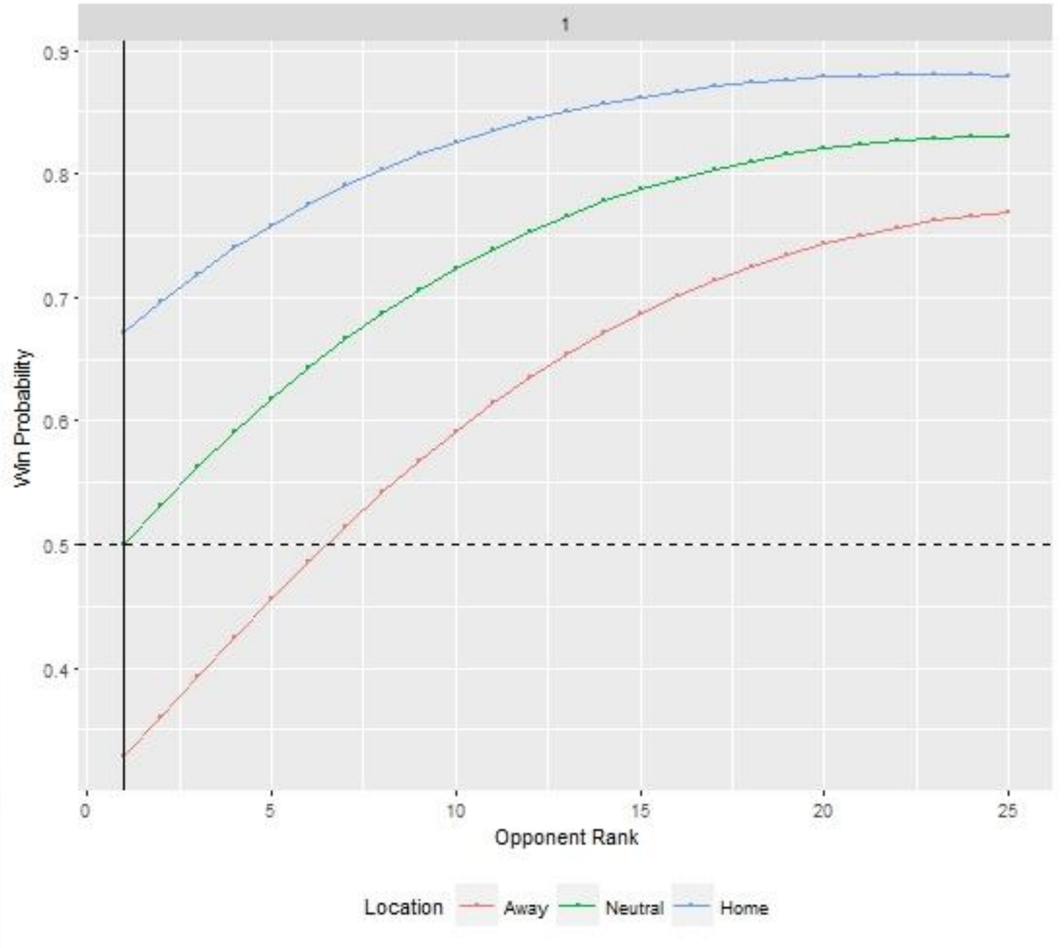
The Goal

- We wanted to explore how win probabilities are associated with rankings in the AP Poll

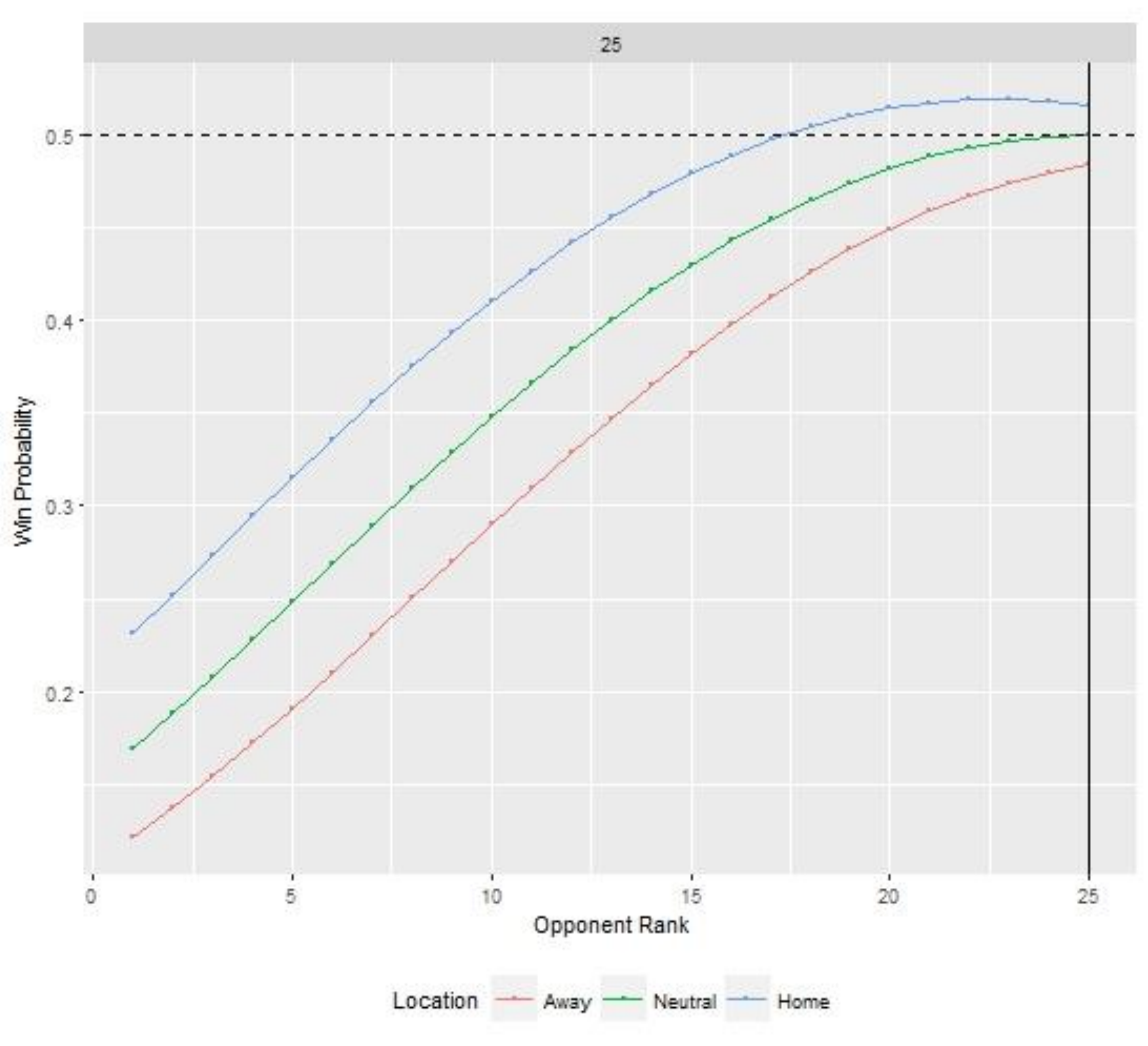
- Want to use results from “similar” games
  - Example, a 1 versus 25 matchup never occurred in our data set. It would seem logical that this would be “similar” to a 1 versus a 24 (which has happened 5 times) or a 3 versus a 25 (which has happened twice)

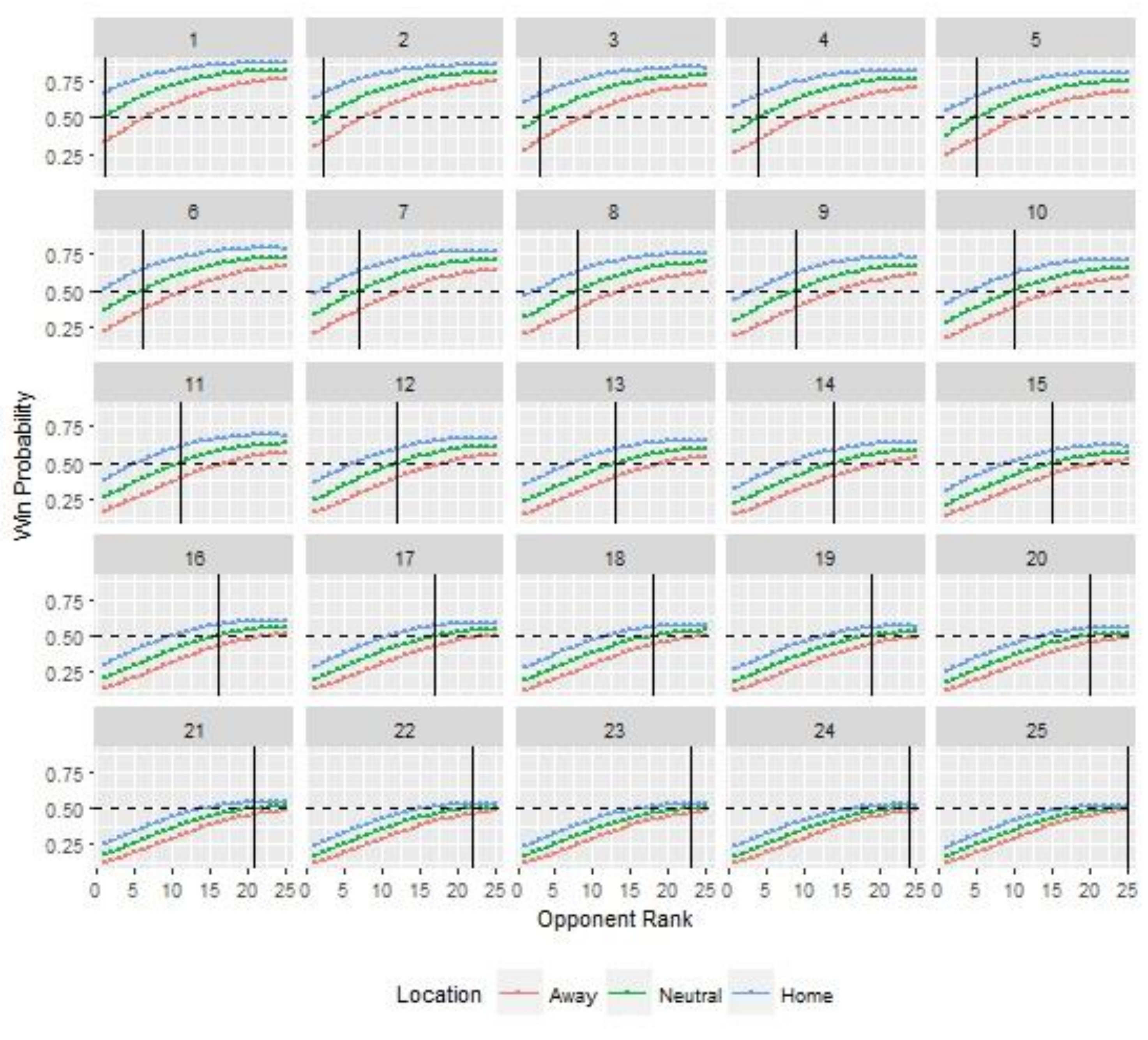
- We estimated win probabilities using 3 different approaches
  - GLMs with a logit link, GLMs with a probit link, Random Forests
- We also predicted point differentials using multiple linear regression models

Example of Win Probability Estimates



Example of Win Probability Estimates




Win Probability
Opponent Rank
Location
Away
Neutral
Home

Generalized Linear Model with a Logit Link

- Allows to construct a model with responses of wins (1's) and losses (0's)
- Accounts for the "Symmetry" problem
- Accounts for the "Even Matchup" problem

The “Symmetry” problem

- In every game, one of the two teams must win
- The win probability for the “Team” should be 1 minus the win probability for the “Opponent”
- If we want to predict the probability a Team will win, we should get 1 minus that probability if we predicted the probability the Opponent will win

The “Even Matchup” problem

- If two teams were the same (as far as our explanatory variables are concerned), neither team should be favored to win

- If the Team has the same ranking as the Opponent and the game is taking place at a neutral site, the team should have a win probability of .5

GLM Logit construction

$$Y_i = \begin{cases} 0 & \text{Team lost game } i \\ 1 & \text{Team won game } i \end{cases}$$

$$X_{1,i} : \text{Location of Game } i, -1 \text{ for away, } 0 \text{ for neutral, and } 1 \text{ for home}$$

$$X_{2,i} : \text{Difference in Team and Opponent Rank (Team minus Opponent)}$$

$$X_{3,i} : \text{Average of Team and Opponent Rank}$$

$$P(Y_i = 1 | X_{1,i}, X_{2,i}, X_{3,i}) = \pi(X_{1,i}, X_{2,i}, X_{3,i})$$

$$Y_i | X_{1,i}, X_{2,i}, X_{3,i} \sim \text{Bernoulli}(\pi(X_{1,i}, X_{2,i}, X_{3,i}))$$

$$\log\left(\frac{\pi(X_{1,i},X_{2,i},X_{3,i})}{1-\pi(X_{1,i},X_{2,i},X_{3,i})}\right) = \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \beta_3 \cdot X_{1,i} \cdot X_{3,i} + \beta_4 \cdot X_{2,i} \cdot X_{3,i} + \beta_5 \cdot X_{1,i} \cdot X_{2,i}$$

$$\pi(X_{1,i},X_{2,i},X_{3,i}) = \frac{1}{1+\exp(-1\cdot[\beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \beta_3 \cdot X_{1,i} \cdot X_{3,i} + \beta_4 \cdot X_{2,i} \cdot X_{3,i} + \beta_5 \cdot X_{1,i} \cdot X_{2,i}])}$$

Variables Considered

Variable Name	Variable Description
X_1	**numLocation**: The location of the game from the Team Perspective. -1: Away, 0: Neutral, 1: Home
X_2	**DiffRanks:** The difference in Team Rank and Opponent Rank (Team minus Opponent)
X_3	**AvgRank:** The average of the Team Rank and the Opponent Rank

Variable
X_1
X_2
$X_1 \cdot X_3$
$X_2 \cdot X_3$
$X_1 \cdot X_2$

Symmetry

- Suppose a Team ranked 3 is playing at home versus an Opponent ranked 5.

- $X_1 = 1,\ X_2 = 3 - 5 = -2,\ X_3 = \frac{3+5}{2} = 4$

- The estimated probability the team wins is

$$\pi(X_1 = 1, X_2 = -2, X_3 = 4) = \frac{1}{1 + \exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}\ .$$

$$1 - \pi(X_1 = 1, X_2 = -2, X_3 = 4) = 1 - \frac{1}{1 + \exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}$$

$$= \frac{\exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}{1 + \exp(-1 \cdot [\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}$$

$$= \frac{\exp([\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])^{-1}}{1 + \exp([\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])^{-1}}$$

$$= \frac{1}{1 + \exp([\beta_1 \cdot 1 + \beta_2 \cdot -2 + \beta_3 \cdot 4 + \beta_4 \cdot -8 + \beta_5 \cdot -2])}$$

$$= \frac{1}{1 + \exp(-1 \cdot [\beta_1 \cdot -1 + \beta_2 \cdot 2 + \beta_3 \cdot -4 + \beta_4 \cdot 8 + \beta_5 \cdot 2])}$$

$$= \pi(X_1 = -1, X_2 = 2, X_3 = 4)$$

Symmetry

- So the estimated probability the Team *loses* (which is the same as the probability the Opponent wins) is

$$1 - \pi(X_1 = 1, X_2 = -2, X_3 = 4) = \pi(X_1 = -1, X_2 = 2, X_3 = 4).$$

- Which give the estimated win probability from the Opponent's perspective (where $X_1 = -1,\ X_2 = 5 - 3 = 2,\ X_3 = \frac{5+3}{2} = 4$

Even Matchup

- If the Team and Opponent have the same ranking and the game is at a neutral site, then $X_1 = 0,\ X_2 = Team\ Rank\ - Opponent\ Rank = 0$
- This accounts for our "Even Matchup" problem, because

$$\pi(X_1 = 0, X_2 = 0, X_3) = \frac{1}{1 + \exp(-1 \cdot [\beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0 + \beta_4 \cdot 0 + \beta_5 \cdot 0])}$$

$$= \frac{1}{1 + \exp(0)}$$

$$= \frac{1}{2}\ .$$

From the 5 variables, we have 8 “variable sets”

Variable Set	Explanatory Variables Used	Parameters Estimated
1	$X_1, X_2, X_1 \cdot X_3, X_2 \cdot X_3, X_1 \cdot X_2$	$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$
2	$X_1, X_2, X_1 \cdot X_3, X_2 \cdot X_3$	$\beta_1, \beta_2, \beta_3, \beta_4$
3	$X_1, X_2, X_1 \cdot X_3, X_1 \cdot X_2$	$\beta_1, \beta_2, \beta_3, \beta_5$
4	$X_1, X_2, X_1 \cdot X_3$	$\beta_1, \beta_2, \beta_3$
5	$X_1, X_2, X_2 \cdot X_3, X_1 \cdot X_2$	$\beta_1, \beta_2, \beta_4, \beta_5$
6	$X_1, X_2, X_2 \cdot X_3$	$\beta_1, \beta_2, \beta_4$
7	$X_1, X_2, X_1 \cdot X_2$	$\beta_1, \beta_2, \beta_5$
8	X_1, X_2	β_1, β_2

A GLM with a logit link was constructed for each variable set

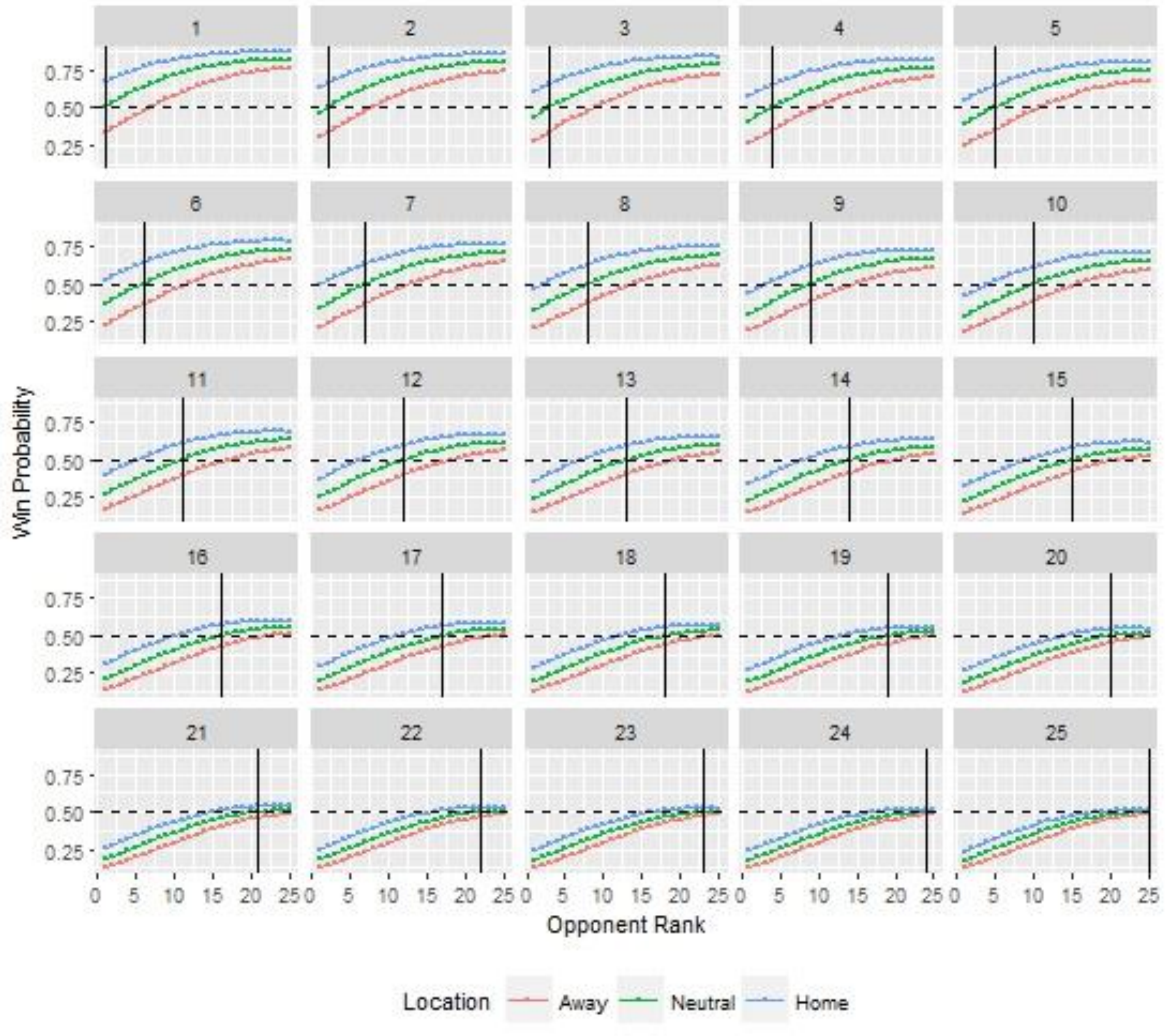
- We can construct a model using each variable set with a GLM logit link
- We will construct separate models for each season from 1989 to 2016
  - To construct the win probability estimates for 1989, we will use the games from 1990 through 2016 to construct the model
  - To construct the win probability estimates for 1990, we will use the games from 1989 and 1991 through 2016 to construct the model

Estimated win probabilities for a Team ranked 1 playing at home against an Opponent ranked 25 using a GLM with a Logit link

	1989	1990	1991	2014	2015	2016
Set 1	0.882	0.889	0.883	0.879	0.880	0.879
Set 2	0.877	0.882	0.876	0.878	0.880	0.874
Set 3	0.885	0.891	0.887	0.882	0.882	0.881
Set 4	0.879	0.883	0.878	0.879	0.881	0.875
Set 5	0.881	0.887	0.882	0.877	0.878	0.878
Set 6	0.877	0.882	0.876	0.877	0.880	0.874
Set 7	0.885	0.891	0.887	0.882	0.882	0.881
Set 8	0.880	0.884	0.879	0.880	0.882	0.877

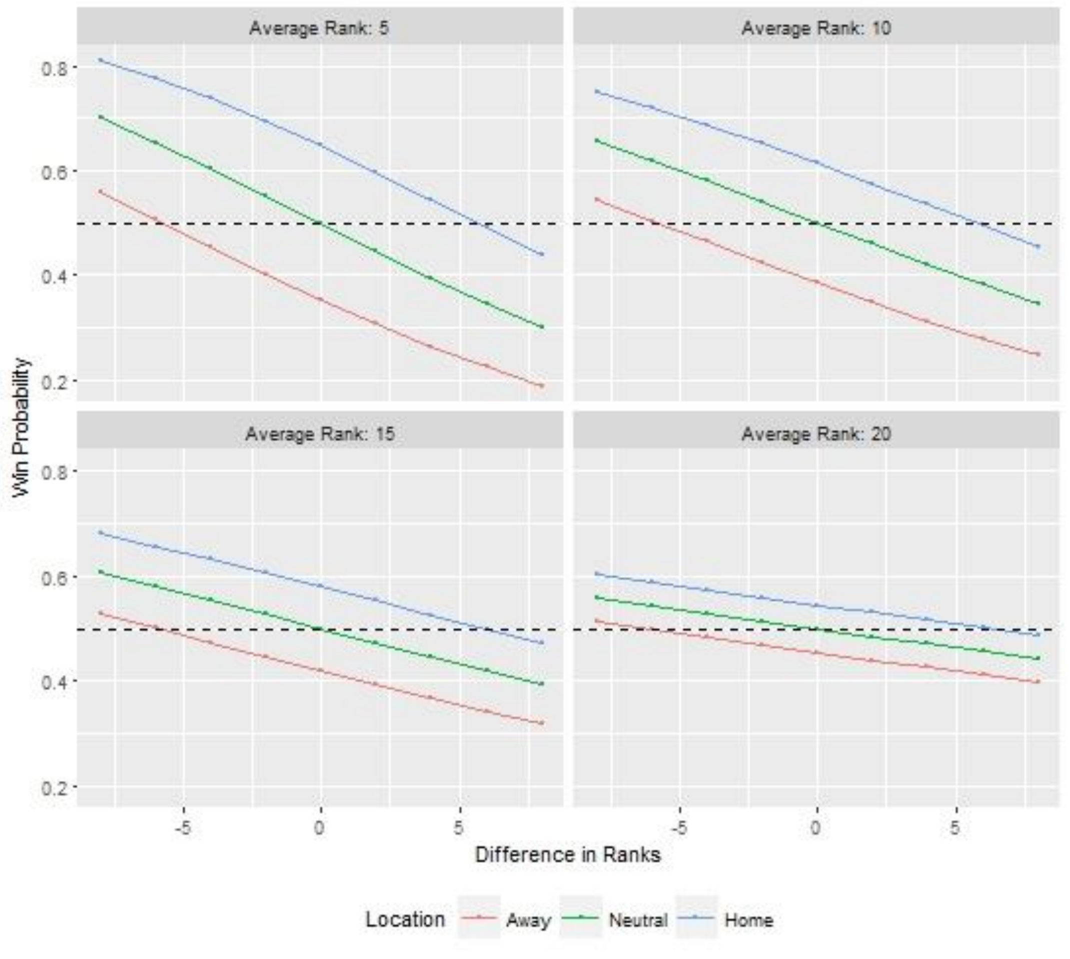
Estimated win probabilities for a Team ranked 24 playing on the road against an Opponent ranked 25 using a GLM with a Logit link

	1989	1990	1991	2014	2015	2016
Set 1	0.468	0.483	0.486	0.494	0.496	0.487
Set 2	0.469	0.484	0.487	0.494	0.496	0.488
Set 3	0.480	0.497	0.496	0.506	0.510	0.498
Set 4	0.480	0.497	0.497	0.507	0.510	0.498
Set 5	0.402	0.403	0.404	0.410	0.401	0.406
Set 6	0.402	0.403	0.405	0.410	0.401	0.406
Set 7	0.417	0.421	0.420	0.426	0.420	0.420
Set 8	0.418	0.422	0.421	0.426	0.420	0.420

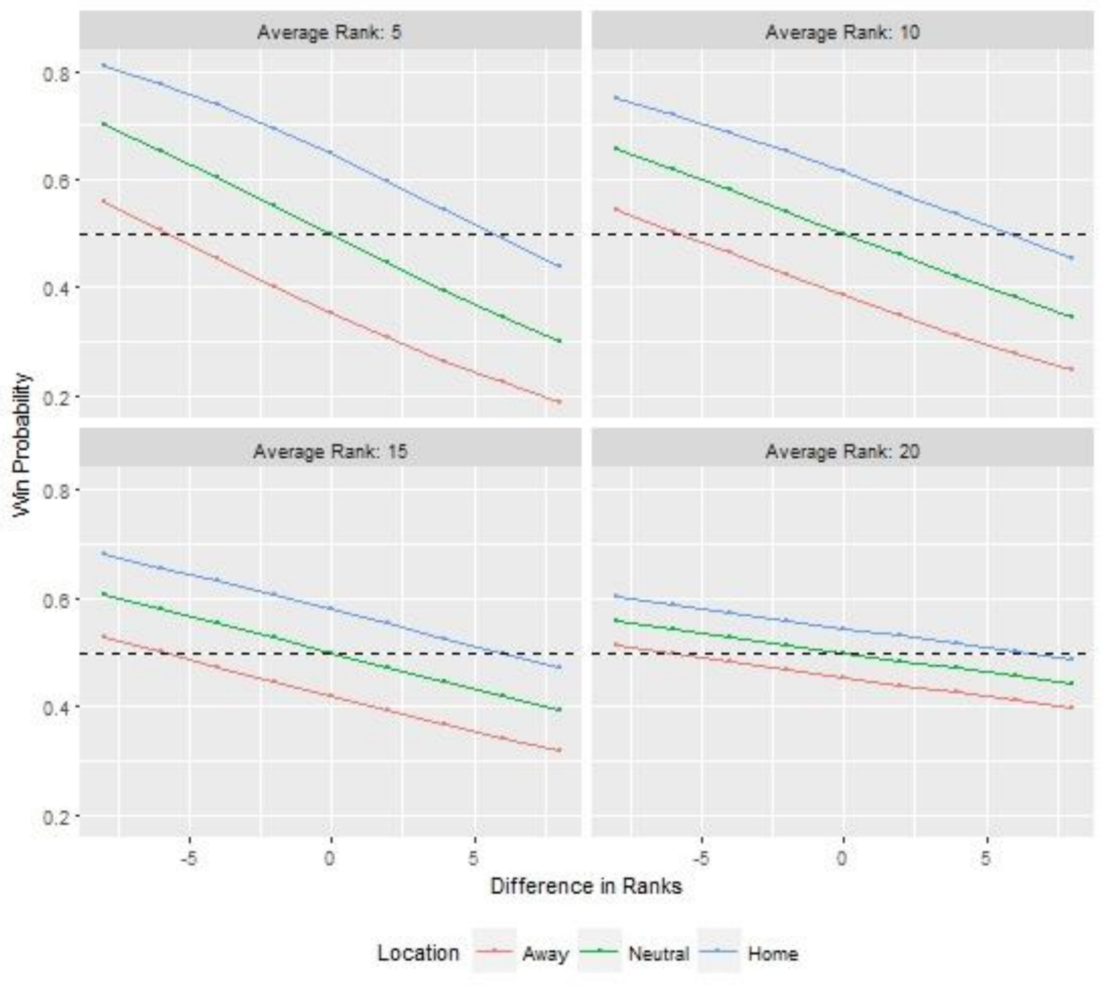

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Win Probability
0.25
0.50
0.75
Opponent Rank
0
5
10
15
20
25
Location
Away
Neutral
Home

- As difference in ranks increases (Team ranking either gets higher or Opponent ranking gets lower), estimated win probability decreases

- Difference in ranks has a bigger impact (steeper drops in win probability) for lower average ranks

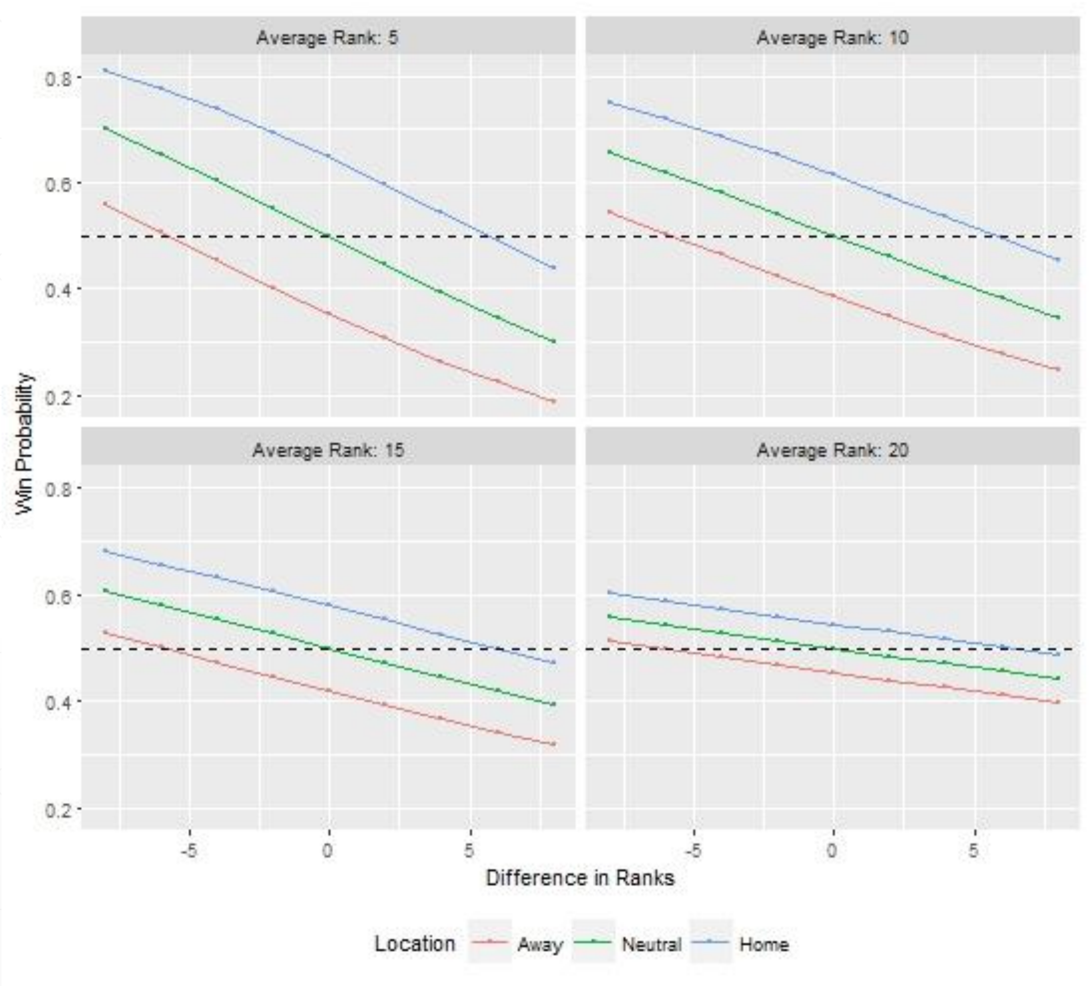


- As average rank increases, estimated win probabilities get closer to .5
  - Games between teams who both have high ranks (ex: 19 and 21) are closer to “toss ups” than games between teams who both have low ranks (ex: 4 and 6).



- Win probability is at its highest for Home games and its lowest for Away games

- Home field advantage is greater for lower ranked teams
  - As average rank increases, the difference in win probability due to location is smaller



GLM with a probit link

- We also used a GLM with a probit link on each variable set
- Model construction is the same, except instead of a logit link, a probit link can be used

$$Y_i | X_{1,i}, X_{2,i}, X_{3,i} \sim \text{Bernoulli}(\pi(X_{1,i}, X_{2,i}, X_{3,i}))$$

$$\Phi^{-1}[\pi(X_{1,i}, X_{2,i}, X_{3,i})] = \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \beta_3 \cdot X_{1,i} \cdot X_{3,i} + \beta_4 \cdot X_{2,i} \cdot X_{3,i} + \beta_5 \cdot X_{1,i} \cdot X_{2,i}$$

$$\pi(X_{1,i}, X_{2,i}, X_{3,i}) = \Phi(\beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \beta_3 \cdot X_{1,i} \cdot X_{3,i} + \beta_4 \cdot X_{2,i} \cdot X_{3,i} + \beta_5 \cdot X_{1,i} \cdot X_{2,i})$$

Symmetry

- Because of the symmetry of the normal distribution, we still guarantee that 1 minus the probability a team wins is equal to the probability the opponent wins

$$1 - \pi(X_1 = 1, X_2 = -2, X_3 = 4) = \pi(X_1 = -1, X_2 = 2, X_3 = 4).$$

Even Matchup

- If the Team and Opponent have the same ranking and the game is at a neutral site, then $X_1 = 0,\ X_2 = Team\ Rank\ - Opponent\ Rank = 0$
- This accounts for our "Even Matchup" problem, because

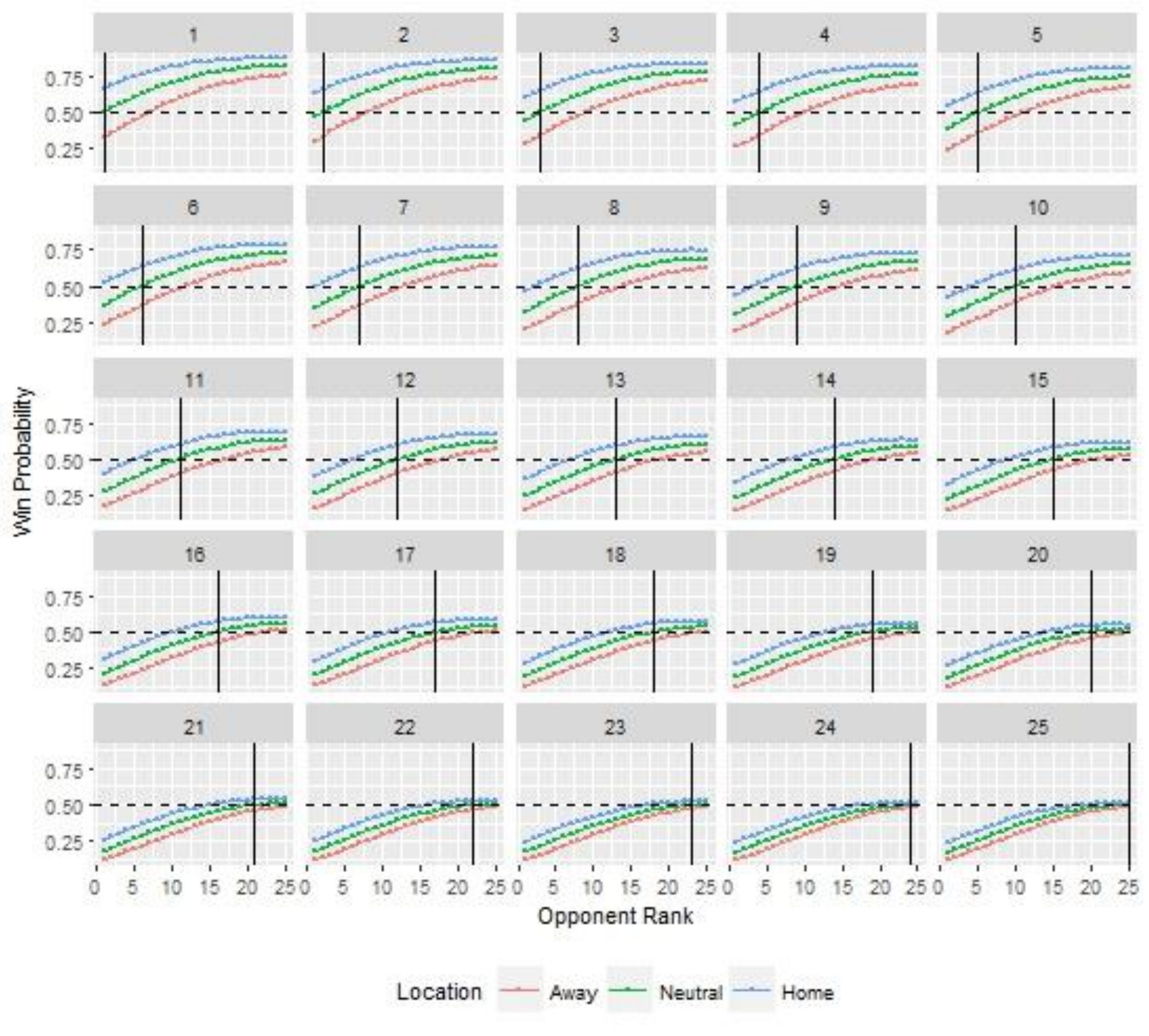
$$\pi(X_1 = 0, X_2 = 0, X_3) = \Phi(\beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0 + \beta_4 \cdot 0 + \beta_5 \cdot 0)$$

$$= \Phi(0)$$

$$= 0.50\ .$$

Estimated win probabilities for a Team ranked 1 playing at home against an Opponent ranked 25 using a GLM with a Probit link

	1989	1990	1991	2014	2015	2016
Set 1	0.892	0.899	0.893	0.889	0.889	0.888
Set 2	0.886	0.891	0.884	0.886	0.888	0.883
Set 3	0.896	0.902	0.897	0.893	0.893	0.891
Set 4	0.888	0.893	0.887	0.889	0.891	0.885
Set 5	0.892	0.898	0.893	0.887	0.887	0.888
Set 6	0.886	0.891	0.885	0.886	0.889	0.883
Set 7	0.896	0.902	0.897	0.892	0.892	0.891
Set 8	0.889	0.894	0.888	0.889	0.892	0.886


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
0.75
0.50
0.25
Win Probability
0 5 10 15 20 25
Opponent Rank
Location
Away
Neutral
Home

Random Forests

- A random forest would allow for more complicated interactions between our explanatory variables
- We used the same 8 variable sets to construct random forests to estimate win probabilities
- The *randomForest* function in the R package *randomForest* was used to construct our random forests
- Forests were constructed with 1500 trees

Determining Tuning Parameters

- The *randomForest* function requires two tuning parameters
  - *mtry: How many predictor variables are considered at each “spilt” in the tree*
  - *nodesize: No splits are attempted for nodes this size and smaller*

- We want to construct a random forest using each variable set for each season
  - 8 variable sets on 28 seasons, resulting in 224 separate forests
- Instead of choosing a “one size fits all” *mtry* and *nodesize*, we found separate tuning parameters for each of the 224 forests

Choosing an *mtry* and *nodesize*

- Possible *mtry* values: 1, 2
- Possible *nodesize* values: 120, 130, 140, … , 250

- A random forest with each combination of *mtry* and *nodesize* values was constructed for each of the 224 separate forests
  - 28 combinations of tuning parameters tested for each of the 224 forests
- The “Out of Bag” (OOB) predictions from each random forest were then used to select which of the tuning parameters would be used for each forest

- We found an OOB negative log likelihood loss, using the formula

$$-1 \cdot \sum_{i=1}^{n} [Y_i \cdot \log(OOB_i) + (1 - Y_i) \cdot \log(1 - OOB_i)] \ .$$

- $Y_i = Result\ of\ Game\ i\ (0{:}\ Loss, 1{:}\ Win)$
- $OOB_i = OOB\ estimated\ win\ probability\ for\ Game\ i$

- The tuning parameters that resulted in the lowest OOB negative log likelihood loss were chosen for each random forest

Example: Random Forest for 2016 season using Variable set 2

- One of the 224 random forests

- Using an *mtry* of 2 and a *nodesize* of 230 resulted in the lowest OOB Loss for random forests for the 2016 season using variable set 2

- Therefore, the forest built using an *mtry* of 2 and a *nodesize* of 230 was used for this random forest

- This process was repeated for each season and variable set combination

- Selected *mtry* values were always 2

- Selected *nodesize* values varied between 140 and 240

Random Forest Win Probabilities

- To account for the “Symmetry” problem, the average of a team’s win probability and 1 minus an opponent’s win probability was found

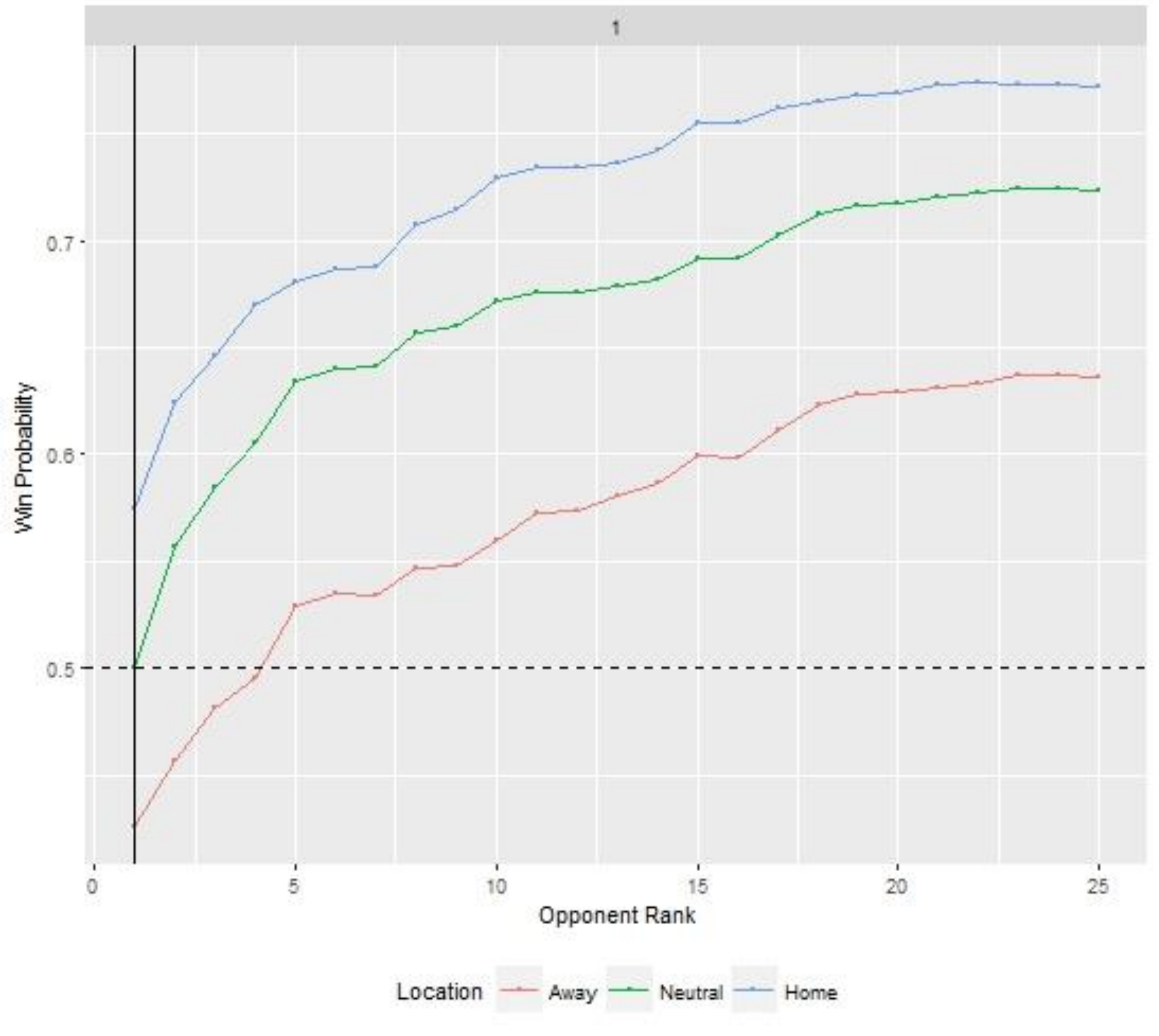
- Example: Suppose we want to estimate the win probability for a Team ranked 25 playing at home against an Opponent ranked 5

Location	Team Rank	Opponent Rank	RF Probability
Home	25	5	.4
Away	5	25	.7

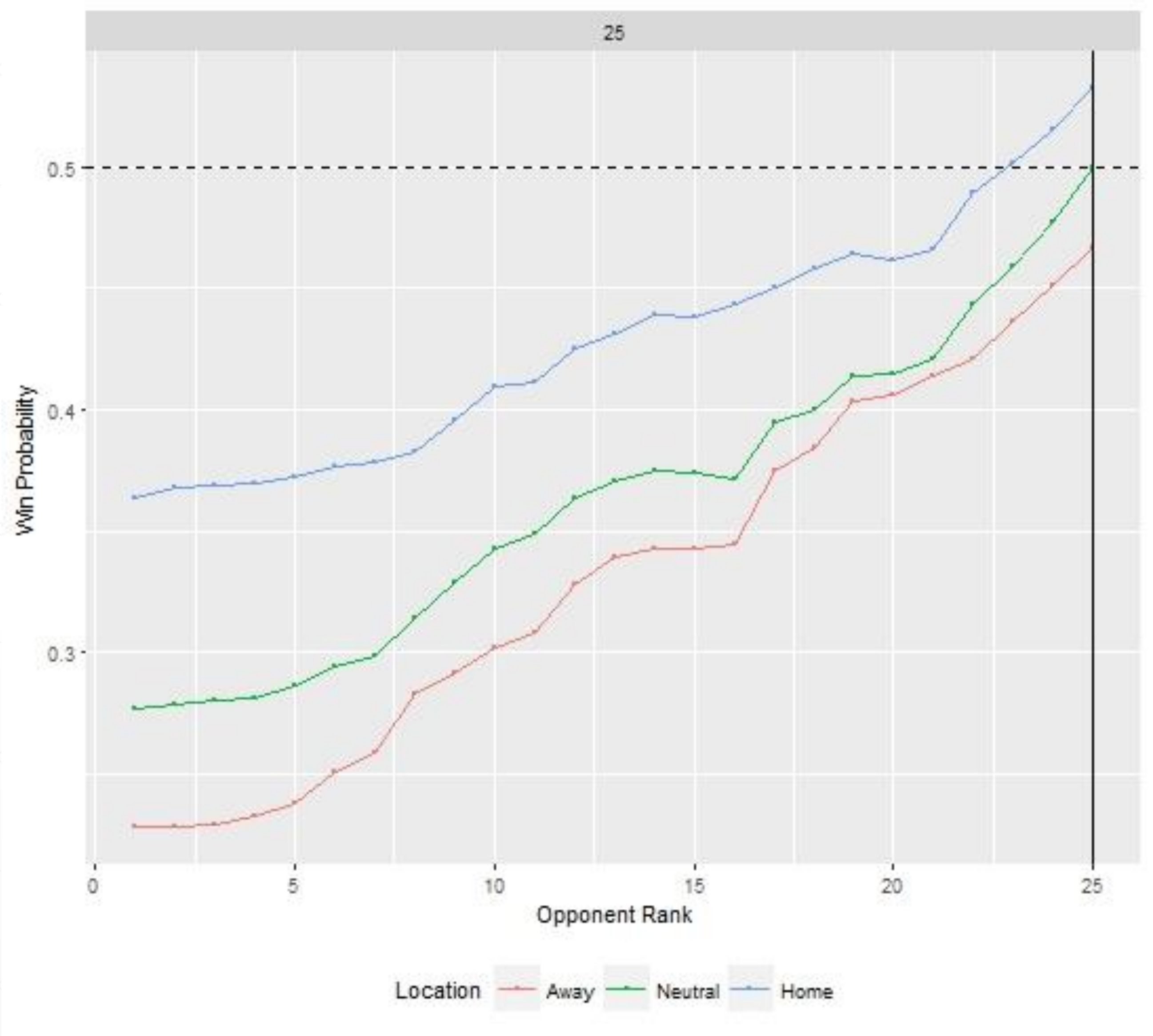
- The predicted win probability would then be calculated to be $\frac{.4+(1-.7)}{2} = .35$

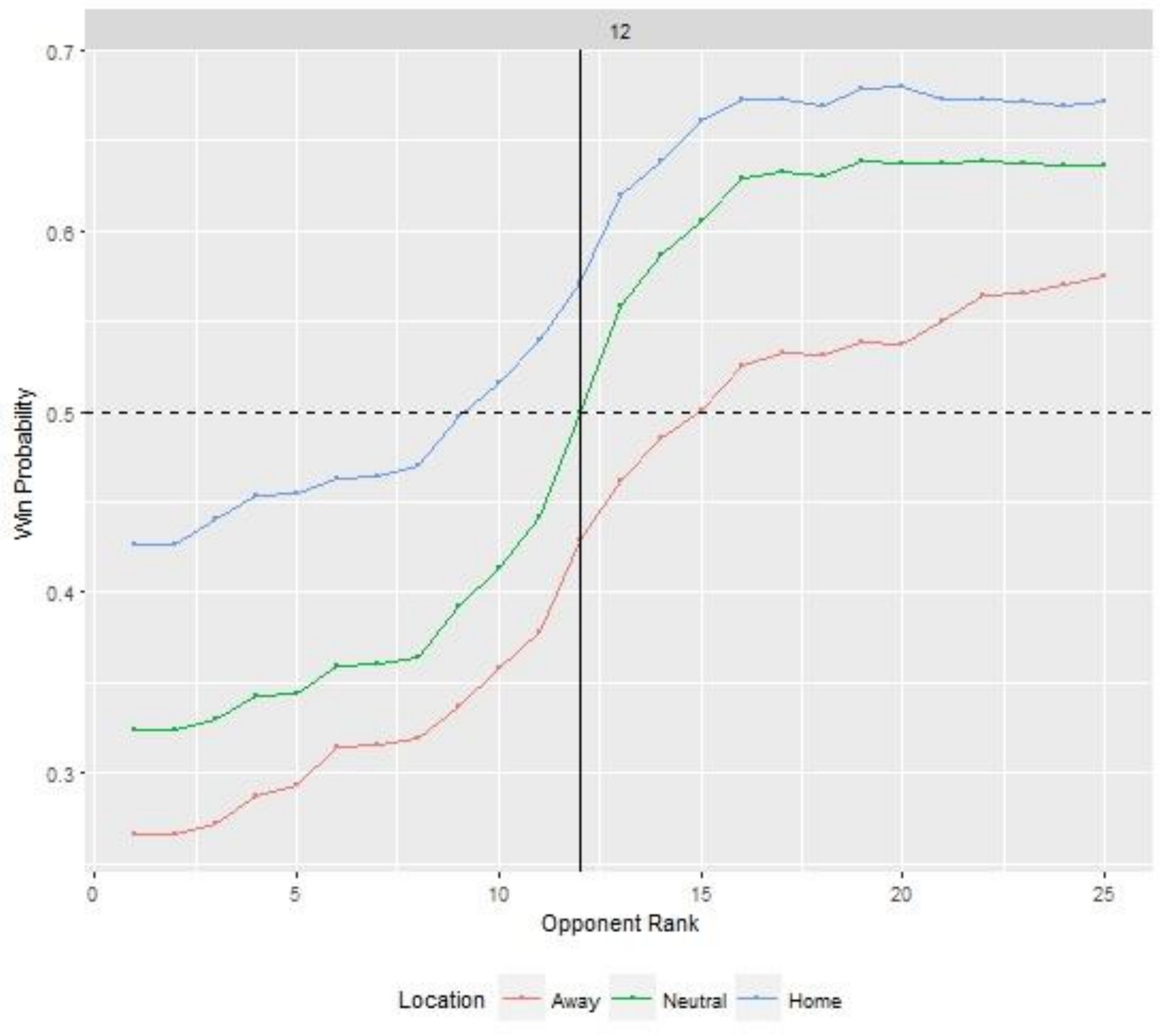
Estimated win probabilities for a Team ranked 25 playing at home against an Opponent ranked 5

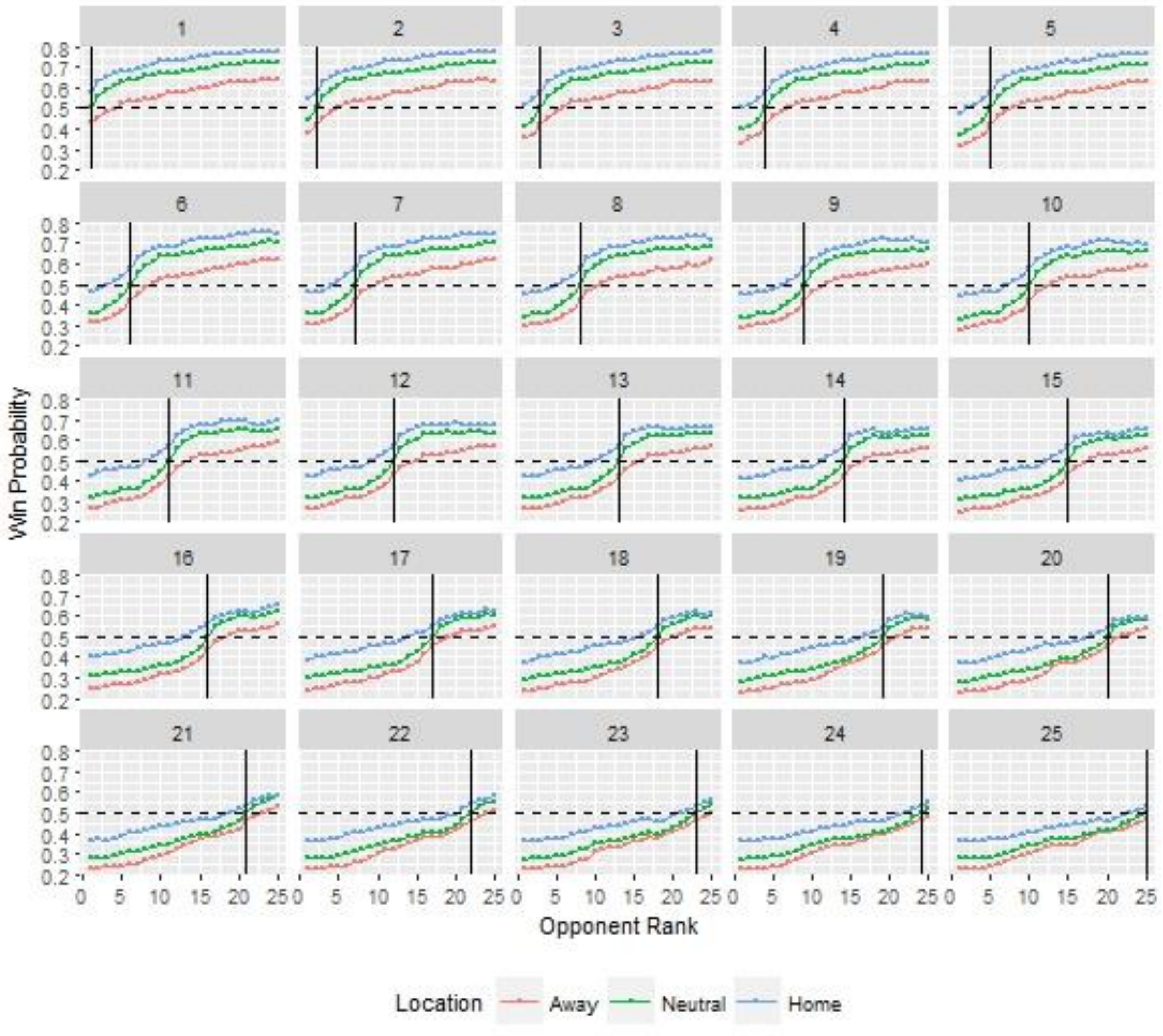
Variable Set	GLM Logit	GLM Probit	Random Forest
Set 1	0.309	0.309	0.373
Set 2	0.316	0.317	0.373
Set 3	0.263	0.262	0.372
Set 4	0.271	0.271	0.375
Set 5	0.324	0.325	0.370
Set 6	0.330	0.332	0.375
Set 7	0.277	0.277	0.350
Set 8	0.284	0.285	0.350



1
Win Probability
0.7
0.6
0.5
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home


25
Win Probability
0.5
0.4
0.3
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home


12
0.7
0.6
0.5
0.4
0.3
Win Probability
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Win Probability
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home

Assessing the Methods: AIC

- For the GLMs, we compared the AIC values for each method to see which was the "best"

- Variable set 2 resulted in the lowest AIC for each season (for both link functions)
  - Location, Difference in Ranks, interaction between Location and Average Rank, interaction between Difference in Ranks and Average Rank

- The probit link models had lower AIC values than the logit link models

- Using AIC, our best GLM is the probit link using variable set 2

AICs for GLMs with a Logit Link

Variable Set	2012 AICs	2013 AICs	2014 AICs	2015 AICs	2016 AICs
Set 1	1826.751	1826.959	1815.593	1815.728	1831.611
Set 2	1824.752	1825.119	1813.600	1813.728	1829.679
Set 3	1838.469	1837.537	1823.824	1827.215	1838.071
Set 4	1836.493	1835.776	1821.867	1825.222	1836.173
Set 5	1831.511	1829.631	1819.368	1820.887	1834.970
Set 6	1829.511	1827.753	1817.368	1818.900	1833.014
Set 7	1842.071	1839.398	1826.841	1831.321	1840.755
Set 8	1840.079	1837.587	1824.855	1829.321	1838.824

AICs for Variable Set 2

Link Used	2012 AICs	2013 AICs	2014 AICs	2015 AICs	2016 AICs
Logit Link	1824.752	1825.119	1813.600	1813.728	1829.679
Probit Link	1824.520	1825.067	1813.519	1813.604	1829.508

Assessing the Methods: Negative log likelihood loss

- Compare our predicted win probabilities to what actually happened each season
- A “perfect” prediction model would give all wins an estimated win probability of 1 and all losses an estimated win probability of 0
- High win probabilities for games that were won and low win probabilities for games that were lost are ideal

- We calculated a negative log likelihood loss for each of the 24 methods (GLM Logit, GLM Probit, Random Forest all used on 8 variable sets) to see which methods were the best at assigning win probabilities.

$$-1 \cdot \sum_{i=1}^{n} [Y_i \cdot \log(\hat{\pi}_i) + (1 - Y_i) \cdot \log(1 - \hat{\pi}_i)]$$

- $Y_i = Result\ of\ Game\ i\ (0{:}\ Loss, 1{:}\ Win)$
- $\widehat{\pi}_i = Estimated\ win\ probability\ for\ Game\ i$

Comparing negative log likelihood losses

Variable Set	Probit Link Losses	Logit Link Losses	Random Forest Losses
Set 1	948.8	948.8	958.3
Set 2	947.5	947.5	958.5
Set 3	953.1	953.2	958.8
Set 4	951.8	951.9	958.7
Set 5	950.7	950.8	958.1
Set 6	949.3	949.4	958.0
Set 7	954.8	954.8	958.7
Set 8	953.4	953.4	959.0

MLR on Point Differential

- In addition to using GLMs and Random Forests to estimate win probabilities, we used multiple linear regression (MLR) models to predict point differentials.

- Doesn't estimate a win probability, instead predicts how many points a team will win (or lose) by

- Can be modified to predict a winner, in that positive point differentials predict a win and negative point differentials predict a loss
- We used the same 8 variable sets

Model Description

$$Y_i = \text{Points scored in game } i - \text{Opponent points scored in game } i$$

$$X_{1,i} : \text{Location of game i, -1 for away, 0 for neutral, and 1 for home}$$

$$X_{2,i} : \text{Difference in Team and Opponent Rank (Team minus Opponent)}$$

$$X_{3,i} : \text{Average of Team and Opponent Rank}$$

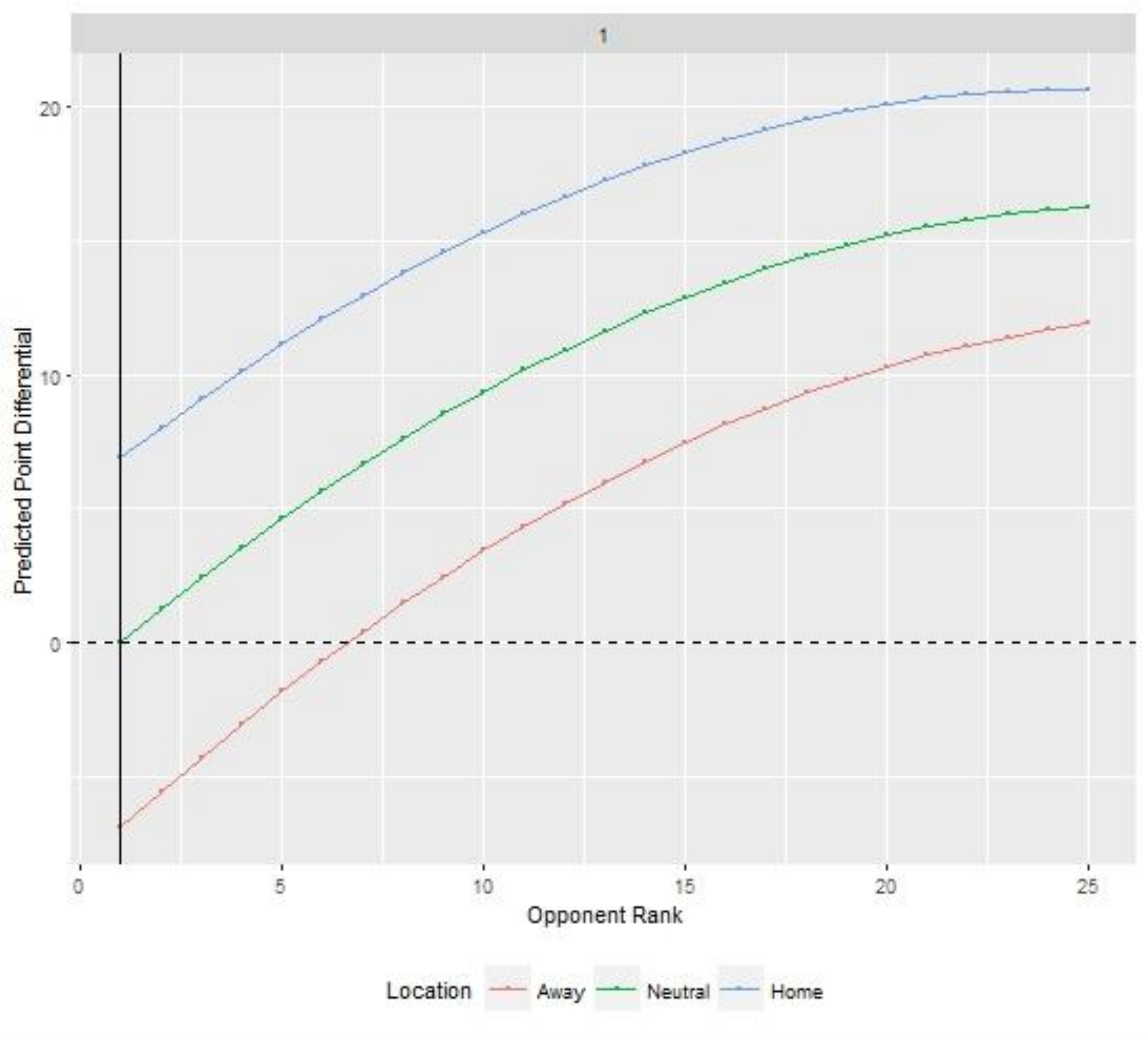
$$Y_i = \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \beta_3 \cdot X_{1,i} \cdot X_{3,i} + \beta_4 \cdot X_{2,i} \cdot X_{3,i} + \beta_5 \cdot X_{1,i} \cdot X_{2,i} + \epsilon_i$$

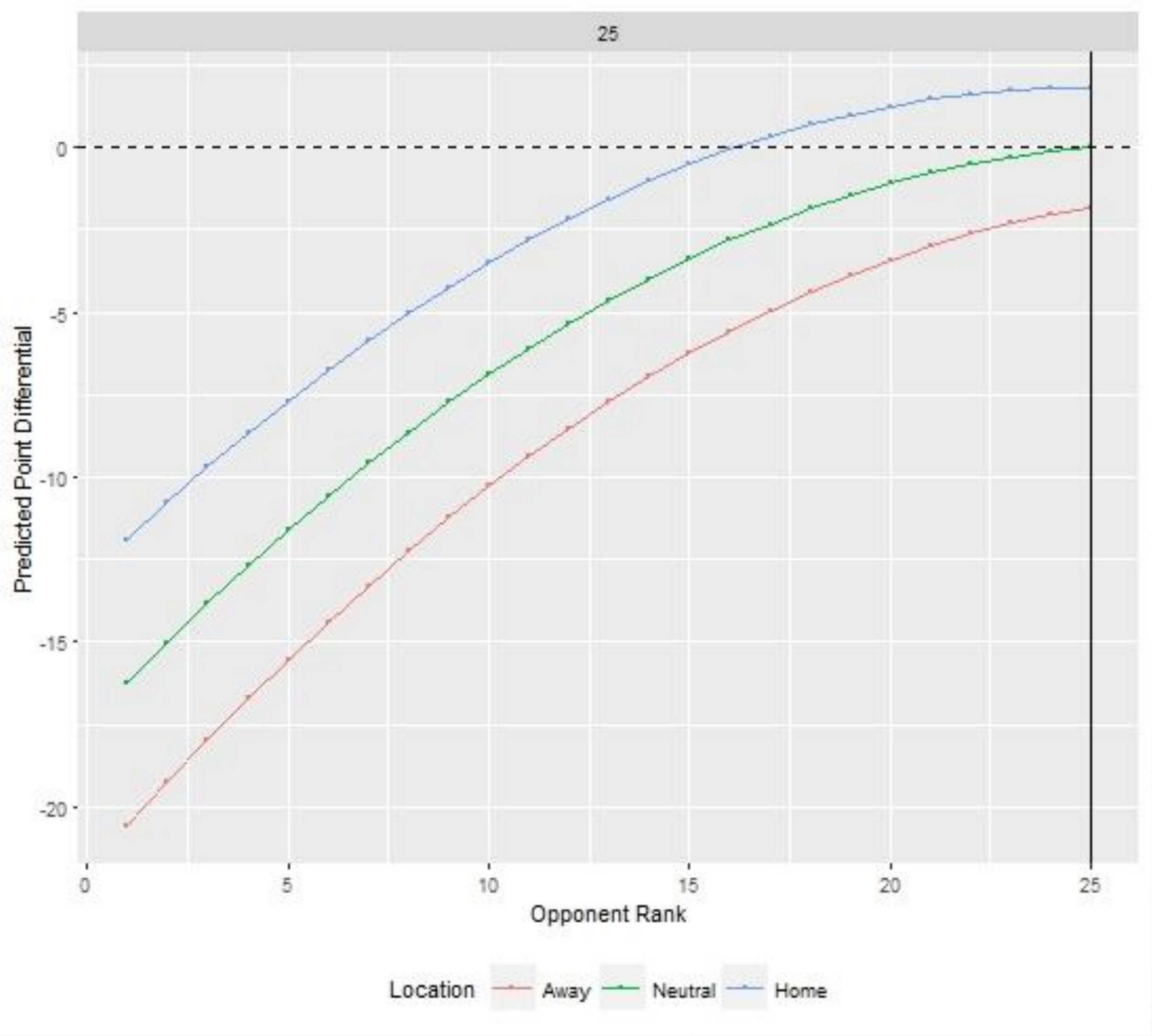
$$\epsilon_i \sim N(0, \sigma_e^2) \ .$$

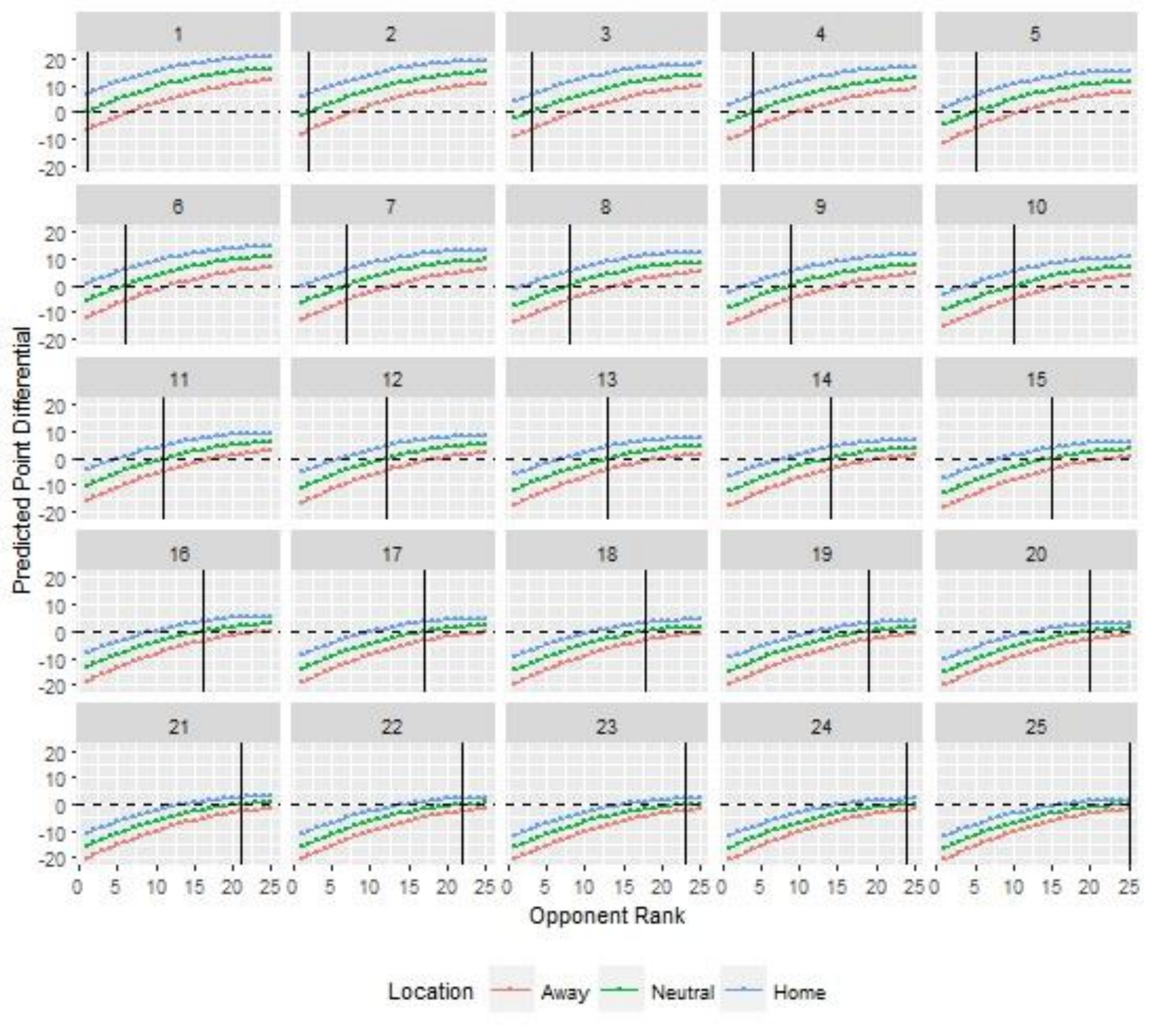
- Still accounts for “symmetry” problem
- Still accounts for “even matchup” problem

Team ranked 25 playing at home against an Opponent ranked 5

Variable Set	GLM Logit	GLM Probit	Random Forest	MLR
Set 1	0.309	0.309	0.373	-7.57
Set 2	0.316	0.317	0.373	-7.68
Set 3	0.263	0.262	0.372	-9.84
Set 4	0.271	0.271	0.375	-9.90
Set 5	0.324	0.325	0.370	-6.97
Set 6	0.330	0.332	0.375	-7.13
Set 7	0.277	0.277	0.350	-9.24
Set 8	0.284	0.285	0.350	-9.35


1
Predicted Point Differential
20
10
0
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home


25
Predicted Point Differential
0
-5
-10
-15
-20
0
5
10
15
20
25
Opponent Rank
Location
Away
Neutral
Home


Predicted Point Differential
Opponent Rank
Location
Away
Neutral
Home

Assessing MLRs: MSE

- To assess the MLR models, we calculated mean square errors for each of the variable sets used

- For Games *i* through *n:*

$$\frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2;$$

- $\hat{Y}_i: Estimated\ Point\ Differential\ for\ Game\ i$

- $Y_i: Point\ Differential\ for\ Game\ i$

Variable Set	Average of MSEs
Set 1	6.30
Set 2	6.17
Set 3	6.22
Set 4	6.09
Set 5	6.18
Set 6	6.05
Set 7	6.11
Set 8	5.99

“Best” methods, according to our assessments

- Using AIC, the best GLM uses a probit link on variable set 2
  - Location, Difference in Ranks, interaction between Location and Average Rank, interaction between Difference in Ranks and Average Rank

- Using negative log likelihood loss, the best win probability estimator is the GLMs that use variable set 2 (probit and logit link losses were nearly identical)
  - Location, Difference in Ranks, interaction between Location and Average Rank, interaction between Difference in Ranks and Average Rank

- Using MSE, the best model for predicting point differential uses variable set 8
  - Location, Difference in Ranks

Summary of the 32 Methods

- We have 4 different approaches to estimating win probability or predicting point differential
  - GLM using Logit Link
  - GLM using Probit Link
  - Random Forests
  - MLR (predicting point differential)

- We used each approach on 8 different variable sets
  - Each set included Location and Difference in Ranks
  - 8 sets were made by including or not including interactions between Location and Average Rank, Difference in Ranks and Average Rank, Location and Difference in Ranks

Predicting Winners

- Fans mostly care about who wins the game
  - Doesn’t make a difference if the win probability was .55 or .95, a win is a win

- We can compare our 32 methods and see which methods were the best at predicting winners for game from the 1989 through 2016 season
  - A win probability above .5 indicates the team is predicted to win
  - A positive point differential indicates the team is predicted to win
  - Games being predicted weren’t used in constructing the model making those predictions
    - Games from the 2016 season weren’t used to make predictions on games in the 2016 season

Example games from the 2016 Season

- Game 1: On September 3rd, 2016, North Carolina (ranked 22) lost 24 to 33 against Georgia (ranked 18) at a neutral site.

- Game 2: On September 24th, 2016, Texas A&M (ranked 10) won 45 to 24 against Arkansas (ranked 17) at a neutral site.

- Game 3: On October 15th, 2016, Wisconsin (ranked 8) lost 23 to 30 against Ohio State (ranked 2) at home.

Example Games

	Game 1	Game 2	Game 3
Team Rank	22	10	8
Opponent Rank	18	17	2
Location	Neutral	Neutral	Home
Result	Loss	Win	Loss
Point Differential	-9	21	-7

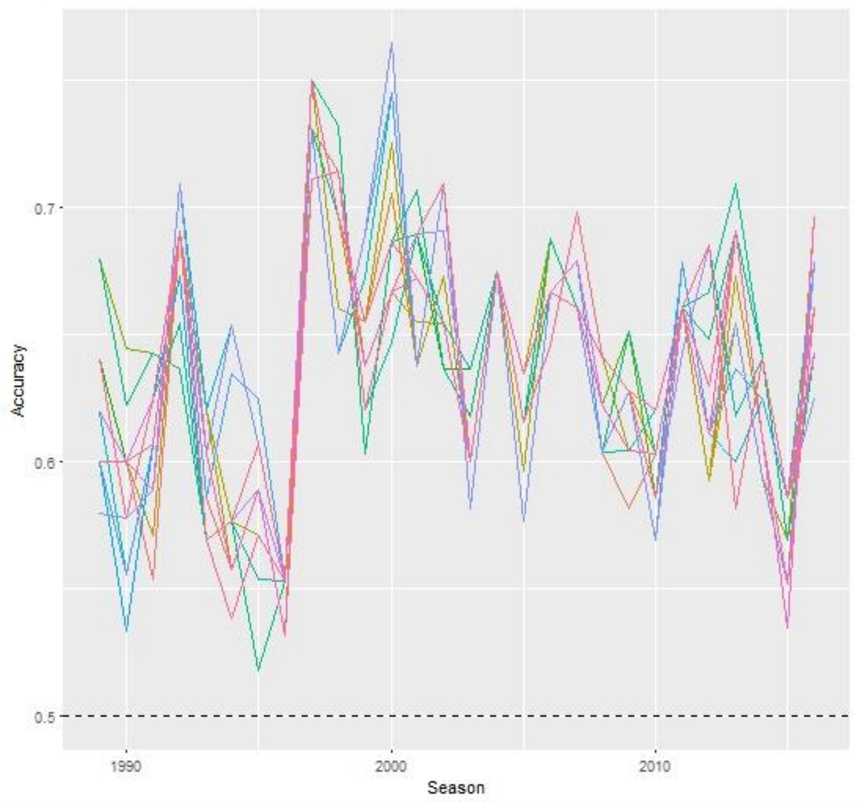
- All 32 methods correctly predicted a loss in Game 1
- All 32 methods correctly predicted a win in Game 2
- 24 methods correctly predicted a loss in Game 3, while the remaining 8 incorrectly predicted a win

Accuracies in the 2016 Season

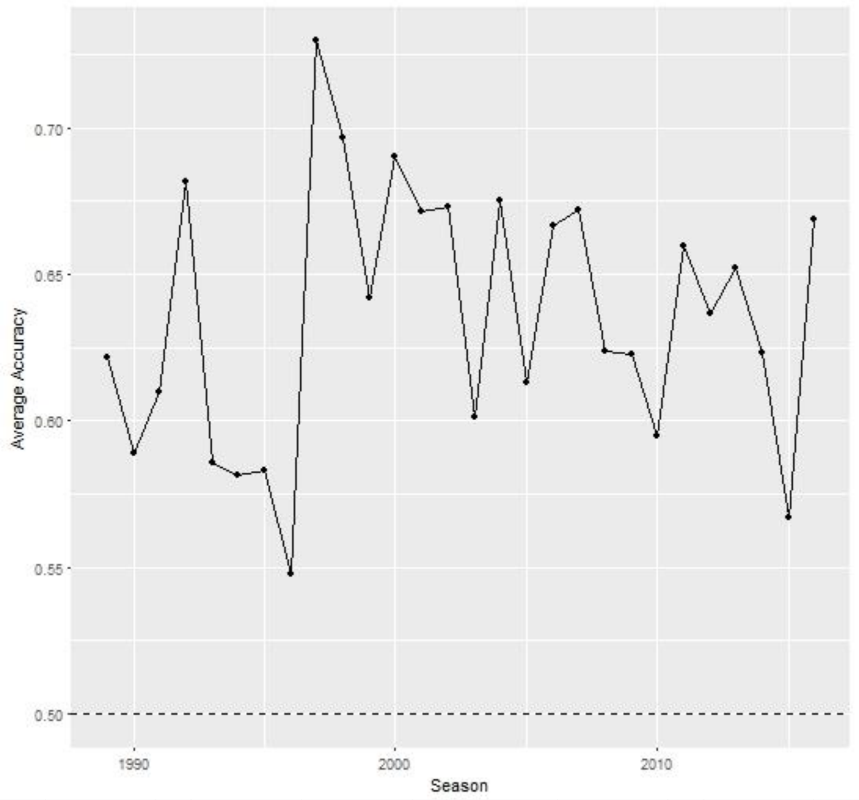
Variable Set	GLM- Logit	GLM- Probit	Random Forests	MLR
Set 1	***.6964***	.6429	.6786	***.6964***
Set 2	***.6964***	.6429	.6786	***.6964***
Set 3	.6786	.6786	.6786	.6607
Set 4	.6429	.6429	.6429	.6429
Set 5	***.6964***	.6429	.6786	***.6964***
Set 6	***.6964***	.6429	.6786	***.6964***
Set 7	.6786	.6786	.6786	.6607
Set 8	.6429	.6429	***.6250***	.6607

Ranged from 35 Correct Games to 39 Correct Games out of 56

Accuracies by Season



Average Accuracy by Season



Total Accuracy

Variable Set	**GLM- Logit**	**GLM- Probit**	**Random Forests**	**MLR**
Set 1	.6292	.6386	.6339	.6352
Set 2	***.6285***	.6386	.6345	.6352
Set 3	.6305	***.6392***	.6345	.6345
Set 4	.6345	.6372	.6352	.6372
Set 5	.6332	.6379	.6352	.6319
Set 6	.6325	.6379	.6359	.6319
Set 7	.6312	***.6392***	.6339	.6345
Set 8	.6359	.6359	.6345	.6379

Ranges from 939 to 955 Correct Predictions out of 1494 games

Our Methods beat picking the lower ranked team

- The lower ranked team won 938 of the 1494 games (62.78% of the time)

- Our lowest accuracy (GLM Logit- Variable set 2) correctly predicted the winner in 939 of the 1494 games
  - 1 game better is still better!
- Our highest accuracy (GLM Probit- Variable sets 3 and 7) correctly predicted the winner in 955 of the 1494
  - 17 games better

Comparing Accuracies

- Two methods tied for most accurate (63.92%); the GLM probits that used variable set 7 and variable set 3
  - Variable set 7: Location, Difference in Ranks, interaction between Location and Difference in Ranks
  - Variable set 3: Location, Difference in Ranks, interaction between Location and Average Rank, interaction between Location and Difference in Ranks

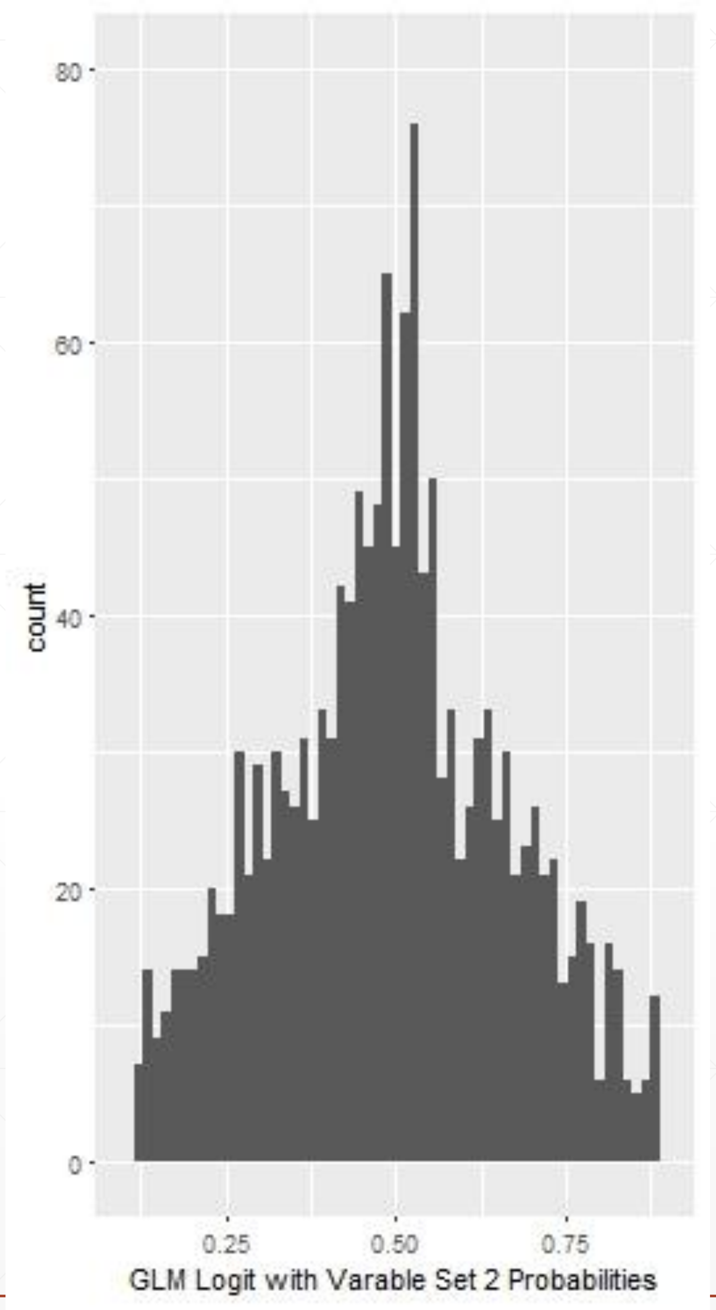
- The least accurate (62.85%) method was the GLM logit that used variable set 2
  - Location, Difference in Ranks, interaction between Location and Average Rank, interaction between Difference in Ranks and Average Rank

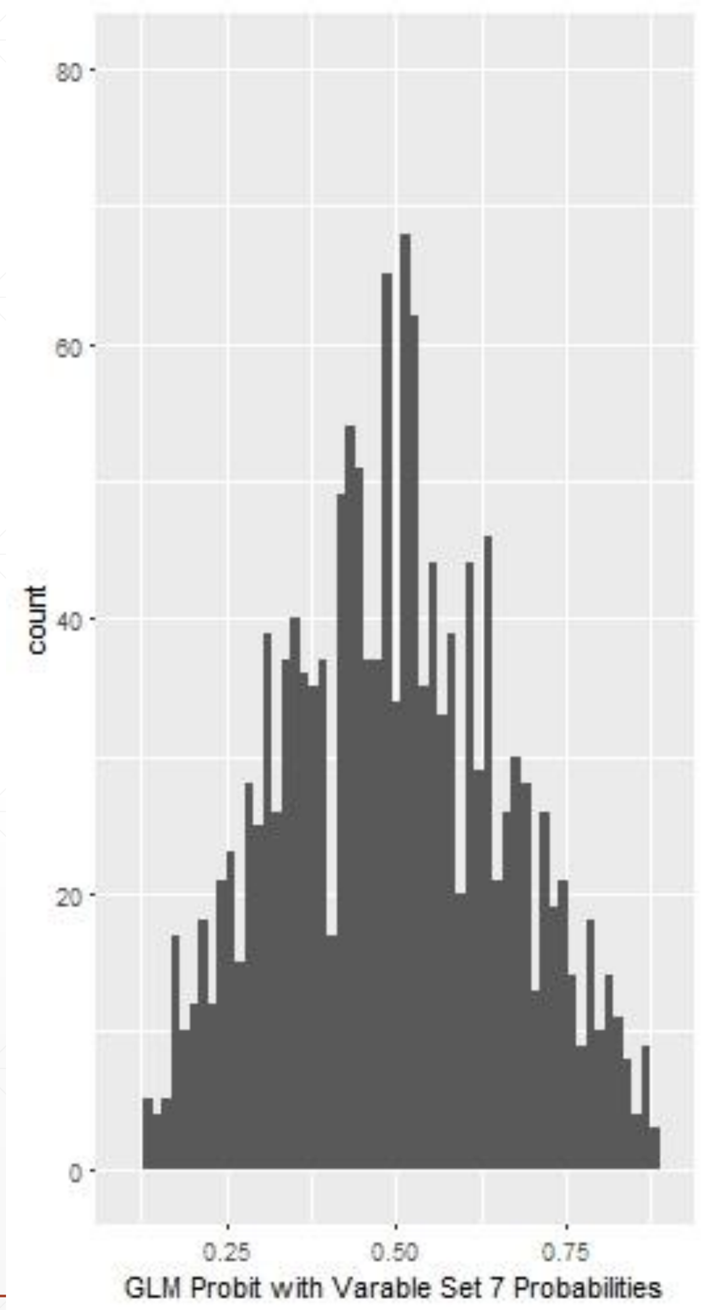
Comparing Accuracies

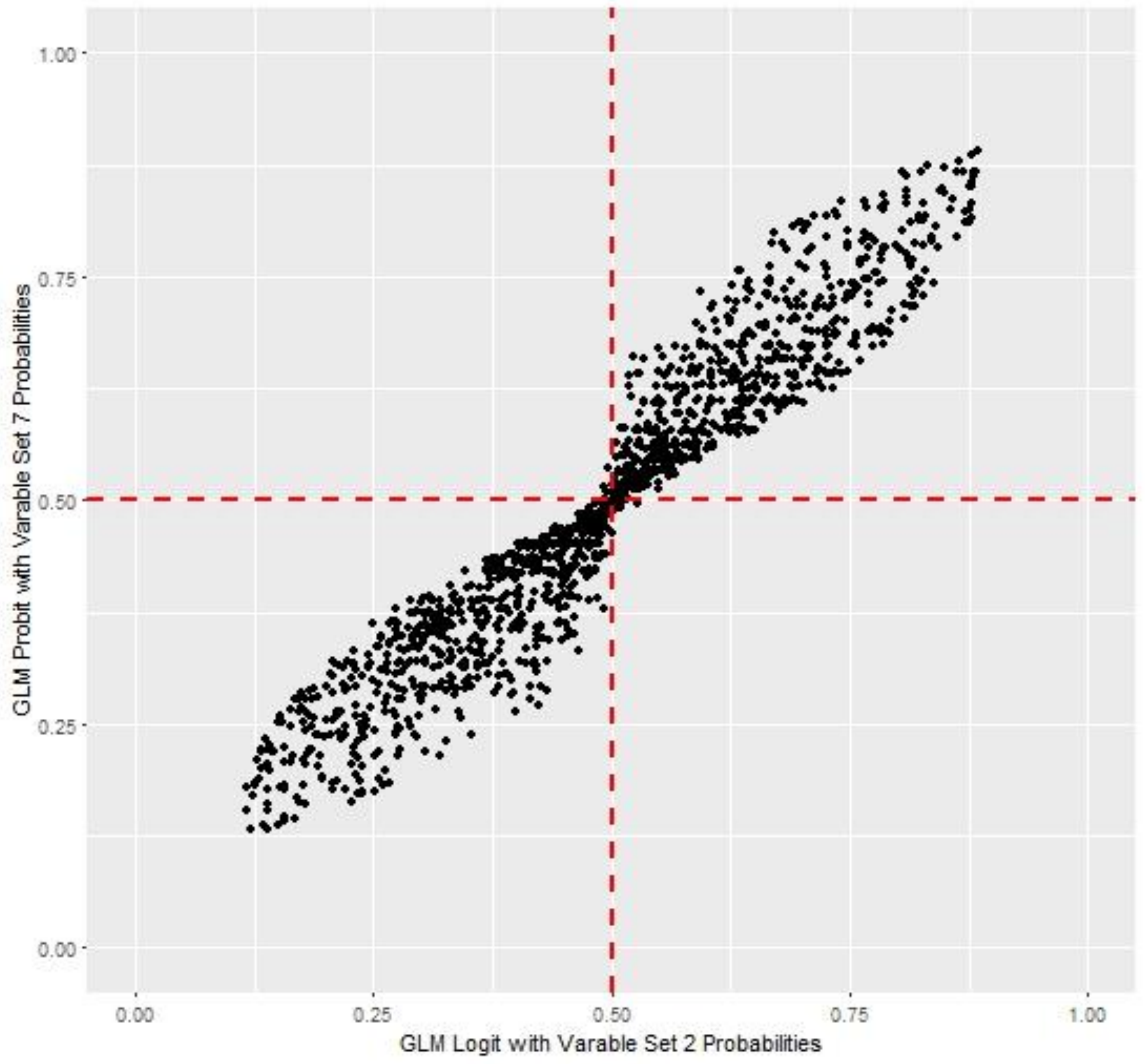
- The difference between our most and least accurate methods was a difference in 1.07 percentage points
  - A 16 game difference out of the 1494 games

- The method that had the lowest negative loglikelihood loss (GLM logit, set 2) ended up having the *worst* accuracy

- The method that had the highest negative loglikelihood loss (GLM probit, set 7) tied for the *best* accuracy


80
60
40
20
0
count
0.25
0.50
0.75
GLM Logit with Varable Set 2 Probabilities


80
60
40
20
0
count
0.25
0.50
0.75
GLM Probit with Varable Set 7 Probabilities



GLM Probit with Varable Set 7 Probabilities
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
GLM Logit with Varable Set 2 Probabilities

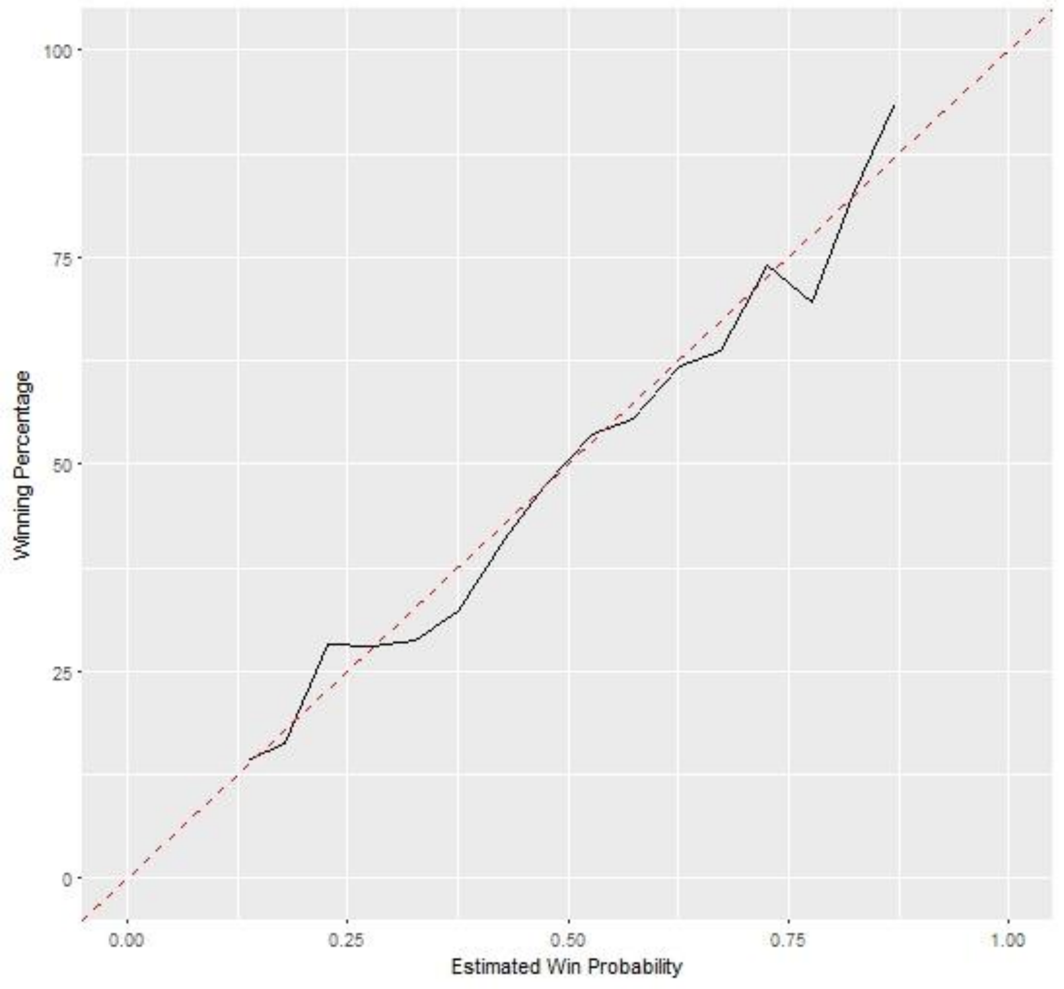
Accuracy at assigning win probabilities

- Fans may mostly care about who wins, but there is a big difference between claiming a team has a 51% chance to win and a 99% chance to win

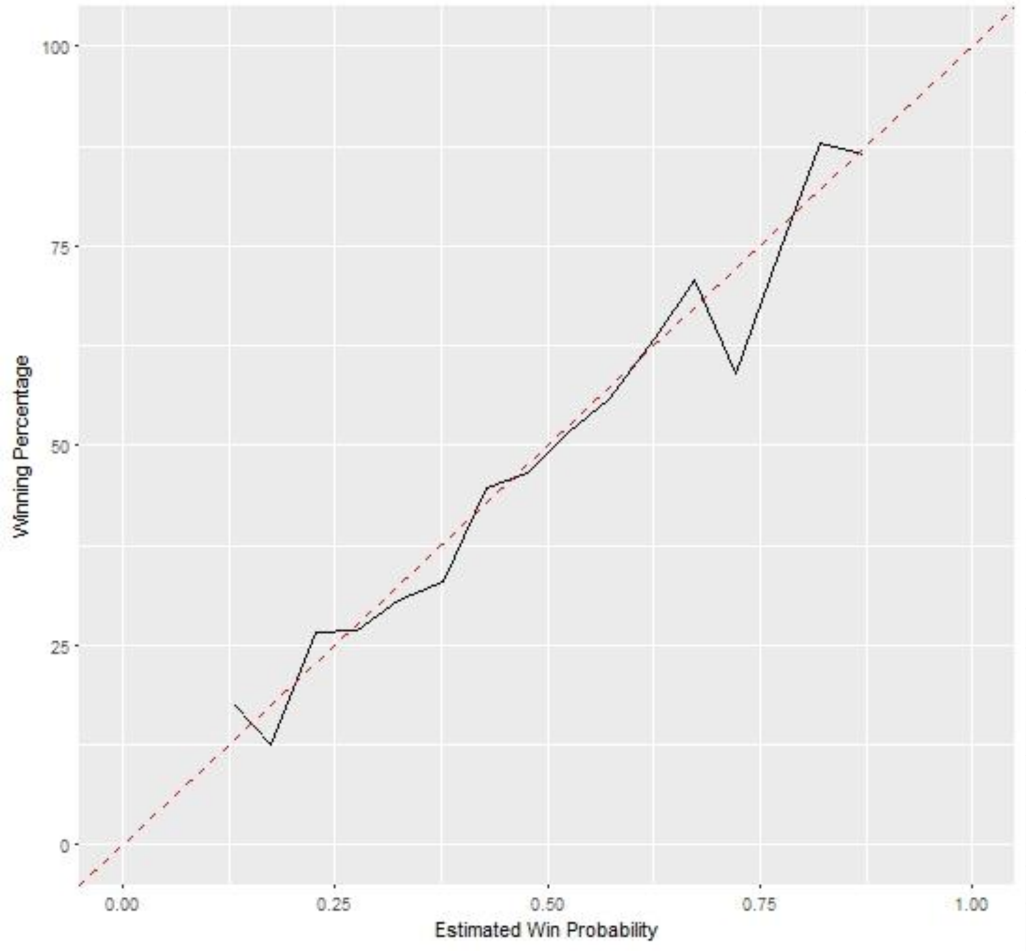
- As far as predicting winners goes, both of those claims are equivalent

- We can bin the games into 5% estimated win probability increments and calculate the winning percentage of the teams in each bin
  - For example, the method using a GLM with a logit link and variable set 2 gave 23 teams a win probability between .10 and .15
  - Of those 23 teams, 4 of them won, for an actual winning percentage of 17.39%

GLM Probit using Variable Set 7 (Total Accuracy: 63.92%)



GLM Logit using Variable Set 2 (Total Accuracy: 62.85%)



Conclusion

- We used 32 methods to either estimate a Team’s win probability or predict a Team’s point differential for 28 seasons
- Overall accuracy at predicting winners using our 32 methods ranged from 62.85% to 63.92%
- 24 methods that predict win probability proved to perform well at assigning win probabilities to games

Future Work

- Somehow incorporate unranked teams in the analysis
- Compare different ranking systems
- Compare different Eras
- Look further into a “Week” effect
- Use the points in the AP Poll instead of the rankings

RK	TEAM	REC	PTS
1	Clemson (43)	12-1	1506
2	Oklahoma (18)	12-1	1474
3	Georgia	12-1	1409
4	Alabama	11-1	1307
5	Ohio State	11-2	1300
6	Wisconsin	12-1	1162
7	Auburn	10-3	1123
8	USC	11-2	1101
9	Penn State	10-2	1008
10	UCF	12-0	983

- The difference between Alabama and Ohio State is only 7 points
- The difference between Ohio State and Wisconsin is 138 points