Predicting NCAA Men's Basketball Win Percentage and Close Game Percentage by Time and Points Difference

Executive Summary

As a spectator watching a basketball game, I always want to know how close my team is to winning the game when the team is having the lead. On the other hand, when the team is trailing I want to know the probability of coming back under this situation, or even in a game that is not that close I always wonder whether the game will come down to the wire or not. These numbers and information sometimes could be estimated by some experienced fan's intuition, but for a spectator that does not have that much experience watching the game would see this information very valuable.

The main product of this project is a dashboard that could let the users interact. The dashboard could let the users choose the margin of points and time left on the play clock (20:00, 16:00, 12:00, 8:00, 4:00), then the probability of winning and the game becoming a close game would come out. The reason I choose these time periods is because these are the NCAA media timeouts. In this research, the definition of a close game is under 4 minutes and the margin of points is within four. The probability of the game winning by the home team is based on logistic regression and the probability of it becoming a close game is based on a 3-layer neural network. The data I used is all the conference games from the 2023-2024, 2022-2023, and 2021-2022 seasons. The reason I only used conference games is because these games are more evenly matched. It will have fewer games that are won by 20 or more points. so the result of the model is more valuable and accurate.

Methodology

Data Collection

I started this project by pulling the API from the website that is given from the NCAA. The first step is to find the season ID for the three seasons I mentioned earlier, and then I used these three seasons to find all the conference game IDs. Lastly put all the conference game ID into a list and then insert it into the play-by-play API to acquire all the conference game's play-by-play.

Data Cleaning

After getting the play-by-play data I filtered the data to only have 20:00, the closest to 16:00, 12:00, 8:00 and 4:00, end of game, and the smallest point margin under 4:00 for each game. The end of the game is to determine the win/loss of the game, and the smallest point margin under 4:00 is to distinguish whether it is a close game or not. The definition of a close game is at any point after 4:00, if the points difference is equal or less than 4 then it will be determined as a close game.

After I mutated two columns that contained win/loss, and close game I deleted the row that represented the end of game and the smallest point margin under 4:00. The data frame only contains the time of the game at 20:00, 16:00, 12:00, 8:00, 4:00 and the win loss and whether it is a close game or not.

Implementing Model

With all the information I needed I started to build my model. The first model I build is to predict the winning percentage at 20:00, 16:00, 12:00, 8:00 and 4:00 with a given margin of points. I choose logistic regression, because logistic regression is commonly used in predicting the percentage if the feature is binary, in this case the result is either win or loss, so logistic regression is a perfect model to predict the win/loss outcome. When I am predicting the percentage of becoming a close game I used a 3-layer Neural Network with a Softmax activation function. Because of using the Softmax activation the result will be in percentage form. I started by using a logistic regression model but the output is not linearly distributed. The largest and smallest margin of points has the lowest chance of becoming a close game and if the margin of points is closer to 0 then it is more likely to become a close game. Due to this phenomenon, the outcome of the logistic regression is very inaccurate, and the Neural Network will be a perfect fit to predict the close game probability.

Result

Win/Loss Percentage

The result of the logistic regression is what we are thinking of. The coefficient is positive and increases as it gets close to the end of the game (*Table 1*). Higher coefficient indicates that in the same amount of point difference the team has a higher chance of winning that game.

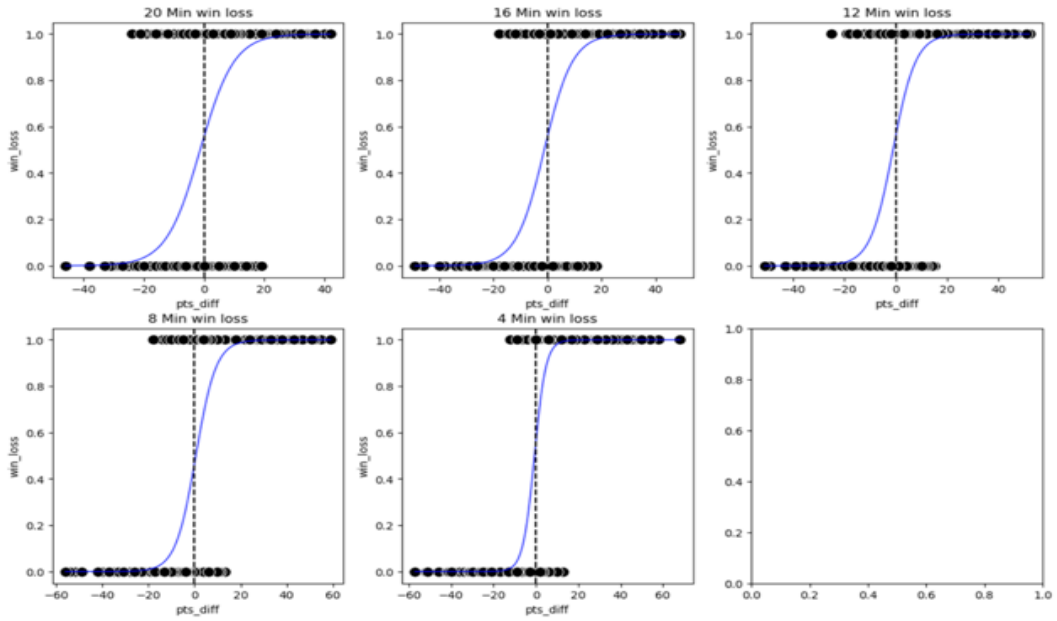
This is because when a team is leading by the same amount of points but with less time left on the clock the team is more likely to win.

The input of the logistic regression is home team score minus away team score, so this model is interpreting the winning percentage for the home team. Because the input is based on the home team we can see that on *Figure 1*, the sigmoid line is shifted to the left a little bit. This indicates that when the home team has a 50% chance of winning the game, is when they are in a trailing situation instead of a tie game. This shows that home court advantage actually does exist.

Table 1

	20:00	**16:00**	**12:00**	**8:00**	**4:00**
Intercept	0.2267	0.2273	0.2437	0.2104	0.2445
Coefficient	0.1772	0.1916	0.2265	0.2667	0.3703

Figure 1



Model Evaluation

	20:00	16:00	12:00	8:00	4:00
Log-likelihood	-2565.1	-2240.7	-1967.9	-1657.8	-1191.9
Deviance	5130.2	4481.4	3935.7	3315.6	2383.8
Pseudo R-squ	0.3375	0.3929	0.4634	0.5175	0.5891

This is the evaluation of the logistics model. The performance in the 20:00 isn't really ideal. The metrics indicate that the accuracy of the prediction in the 20:00 is really moderate.

The Log-likelihood is closer to 0, deviance is decreasing and Pseudo R-squ is closer to 1. But the model does improve as it gets closer to the end of the game.

Close Game Percentage

For the close game percentage, I build a three-layer neural network. The first two hidden layers I used a Relu activation function and the output layer I used a Softmax function. The reason I used a Softmax function is because I want the output to be in probability format. Since the neural network is by classification, it does not have an equation like the logistic regression where I could interpret the coefficients.

When we are looking at *Figure 2* we can see the same trend as *Figure 1*, which is that when it gets closer to the end of the game the slope is getting steeper and the highest percentage of going to a close game is shifting to the left. This indicates that every point is going to be more valuable towards the end of the game.

Model Evaluation

Figure 3 specifies that this model is well trained because the loss is trending downward as the epoch starts to increase. Lower loss means that the model is predicting it more accurately which is what we are pursuing.

Figure 2

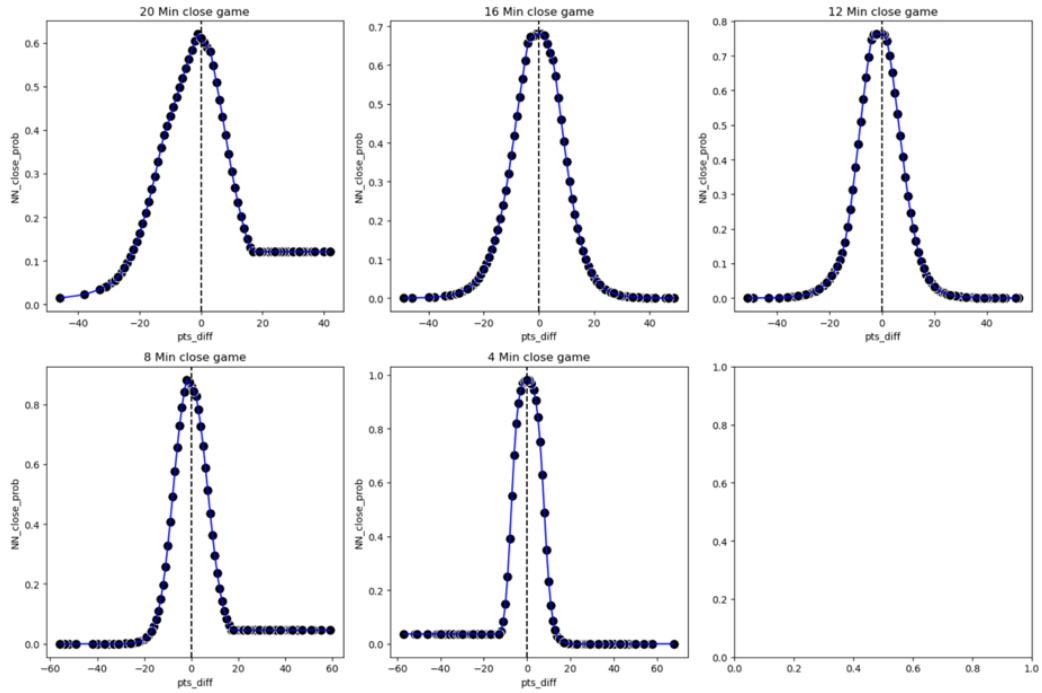
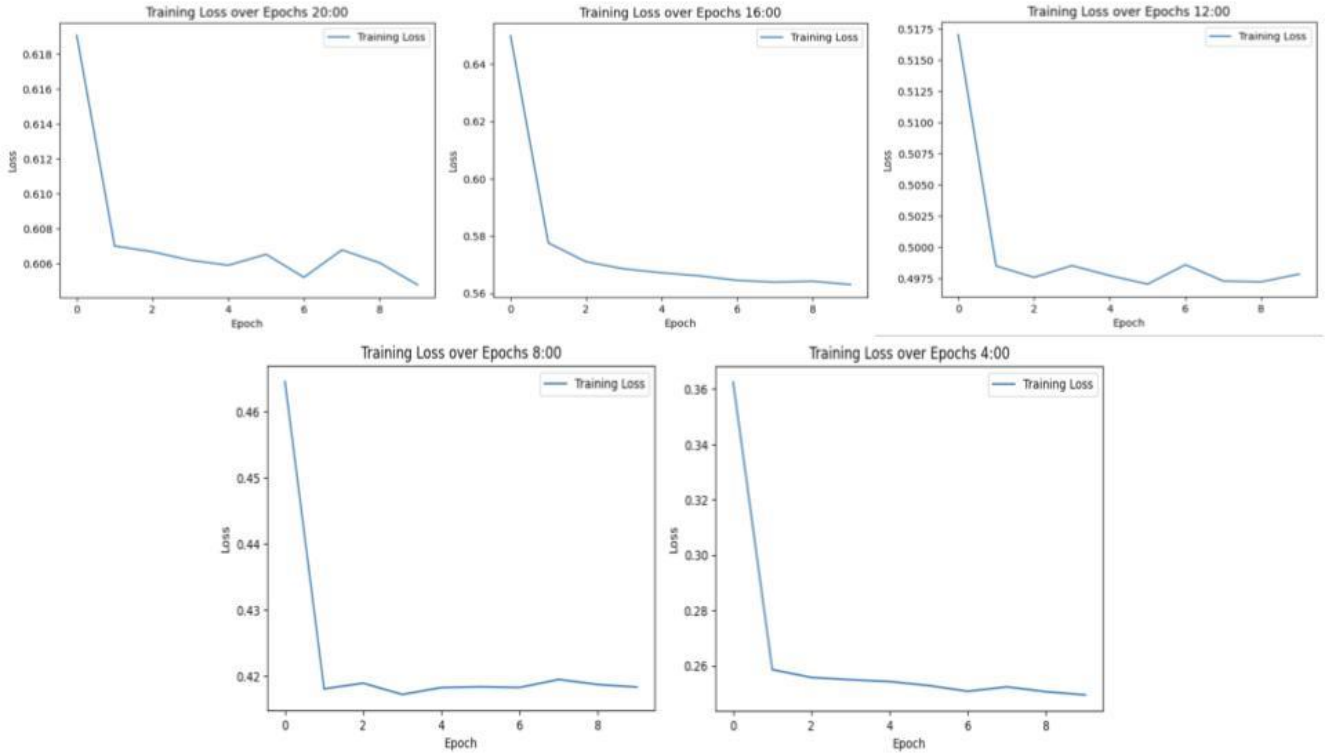

20 Min close game
16 Min close game
12 Min close game
8 Min close game
4 Min close game
NN_close_prob
pts_diff

Figure 3


Training Loss over Epochs 20:00
Training Loss over Epochs 16:00
Training Loss over Epochs 12:00
Training Loss over Epochs 8:00
Training Loss over Epochs 4:00
Training Loss
Loss
Epoch

Dashboard

Figure 4

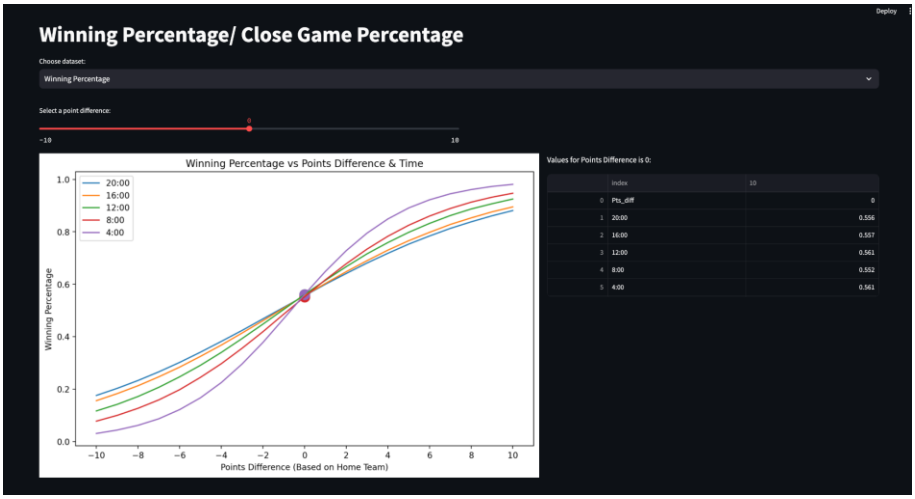
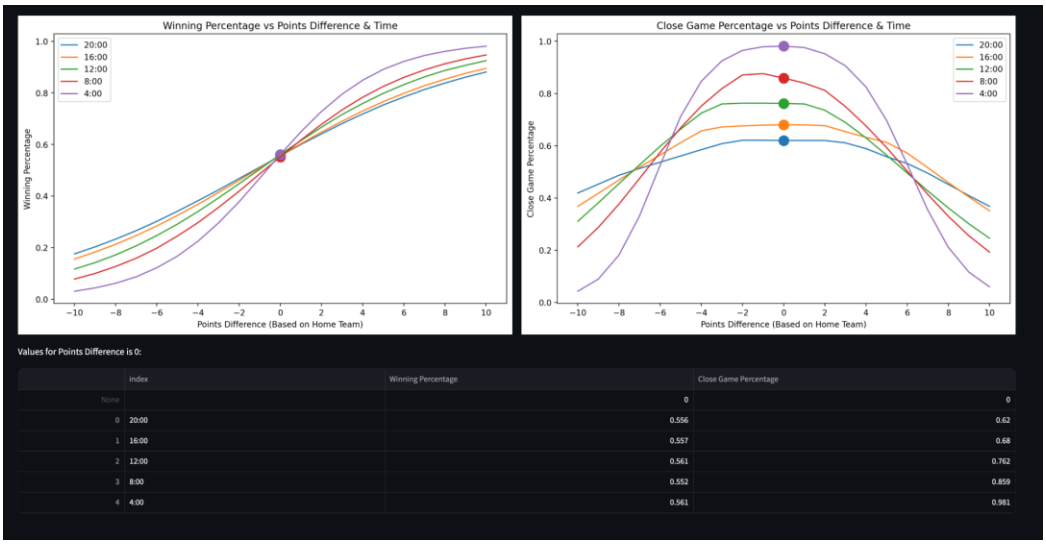


Figure 5



In the dashboard, the users can choose whether they want to look at only the winning percentage or both the winning percentage and close game percentage at the top. The difficult setting is only the winning percentage. *Figure 4,* is the result when the user chooses only the winning percentage. On the top there is a slider for the user to choose the points difference

that the user is looking for. The highlighted point is the point of the points difference. On the right side of the plot, the table shows the winning percentage in that situation. *Figure 5* is the result when the users choose both the winning percentage and close game percentage. It could also use the slider to choose the points difference. The table for the percentage will be shown below the plot.

Insights/ Recommendations

Figure 6

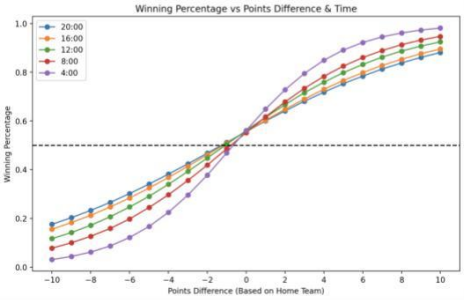


Table 2

	Pts_diff	20:00	16:00	12:00	8:00	4:00
0	-10	0.176	0.156	0.117	0.078	0.031
1	-9	0.203	0.183	0.142	0.100	0.044
2	-8	0.233	0.213	0.172	0.127	0.062
3	-7	0.266	0.247	0.207	0.159	0.087
4	-6	0.302	0.284	0.247	0.198	0.122
5	-5	0.341	0.325	0.291	0.245	0.167
6	-4	0.382	0.368	0.340	0.297	0.225
7	-3	0.424	0.414	0.393	0.356	0.296
8	-2	0.468	0.461	0.448	0.419	0.378
9	-1	0.512	0.509	0.504	0.486	0.469
10	0	0.556	0.557	0.561	0.552	0.561
11	1	0.600	0.603	0.615	0.617	0.649
12	2	0.641	0.648	0.667	0.678	0.728
13	3	0.681	0.690	0.716	0.734	0.795
14	4	0.718	0.730	0.759	0.783	0.849
15	5	0.753	0.766	0.798	0.825	0.891
16	6	0.784	0.798	0.832	0.860	0.922
17	7	0.813	0.828	0.862	0.889	0.945
18	8	0.838	0.853	0.887	0.913	0.961
19	9	0.861	0.876	0.907	0.932	0.973
20	10	0.881	0.895	0.925	0.947	0.981

In *Figure 4*, it clearly shows that the blue line has the flattest curve and the black line has the steepest curve. This indicates that each point will matter more later in the game. What is interesting to me is that when the point difference is 0 the probability almost overlaps at 55% meaning that at any point in the game when the game is tied the win percentage is going to be over 50% which is slightly in favor of the home team.

In *Table 2*, we can see that even though at a tie game the home team is only 5% above 50%, but when the home team is leading by 1 points their win percentage are at 60%, 60.3%, 61.5%, 61,4%, 61.7%, 63.4%, and 65% when there is 20:00, 16:00, 12:00, 8:00,and 4:00 left on the clock. This shows that home court advantage does exist, even when the home team is down by 1 they still have a win percentage of 51.2%, 51%, 50.4%, 50%, 49%, 47.3%, and 47%.

This makes me want to do a deep research on how the team is performing when under 4 min because of how much steeper the slope at the 4:00 mark is compared to other times. In *Figure 5,* we can see that with under 4 minutes to play and in a close game situation the teams tend to score more points than any other 4 minutes span in the second half. This is very counter

intuitive, since I think that in clutch situations a team will play more physically therefore it will affect their performance offensively, but this isn't the case from the data that I collected.

In *Figure 6,* I took all the 4:00, 3:00, 2:00, 1:00 that are in a close game situation to run a logistic regression. It clearly showed that the slope of winning percentage is the steepest when the margin of points is around 3. This is the most obvious at the one-minute mark because each point matters more towards the end of the game. In *Figure 6,* I also draw a dash line that represents 50%. On the graph, it clearly showed that when the game is tied the win percentage is above 50% for the home team.

Figure7

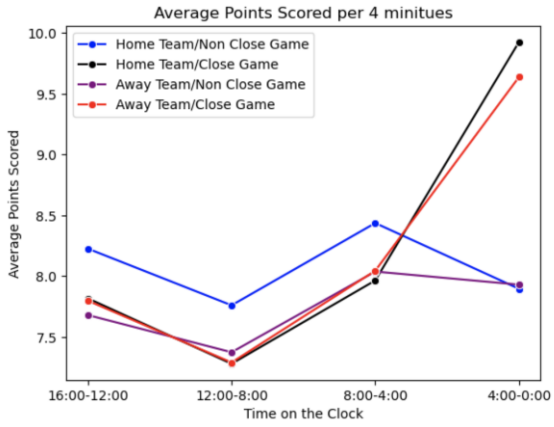


Figure 8

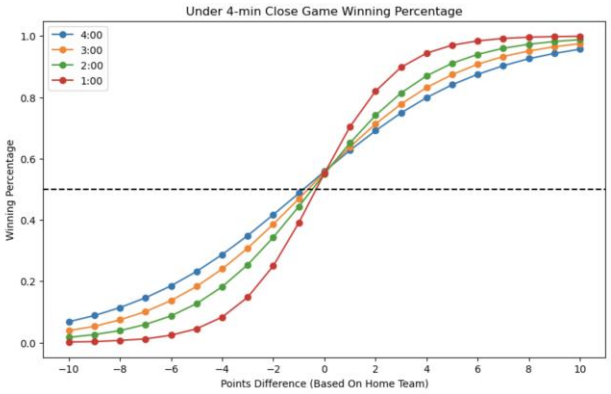


Table 3

	Pts_diff	4:00	3:00	2:00	1:00
0	-10	0.068	0.039	0.017	0.002
1	-9	0.088	0.053	0.026	0.003
2	-8	0.114	0.074	0.039	0.007
3	-7	0.146	0.101	0.059	0.012
4	-6	0.185	0.137	0.087	0.024
5	-5	0.232	0.183	0.127	0.045
6	-4	0.287	0.240	0.182	0.083
7	-3	0.349	0.308	0.254	0.148
8	-2	0.417	0.386	0.343	0.250
9	-1	0.487	0.469	0.444	0.391
10	0	0.558	0.555	0.550	0.552
11	1	0.627	0.638	0.651	0.704
12	2	0.691	0.713	0.741	0.820
13	3	0.749	0.778	0.814	0.898
14	4	0.799	0.832	0.870	0.944
15	5	0.841	0.874	0.911	0.970
16	6	0.875	0.908	0.940	0.984
17	7	0.903	0.933	0.960	0.992
18	8	0.926	0.951	0.973	0.996
19	9	0.943	0.965	0.982	0.998
20	10	0.957	0.975	0.988	0.999

Figure 9

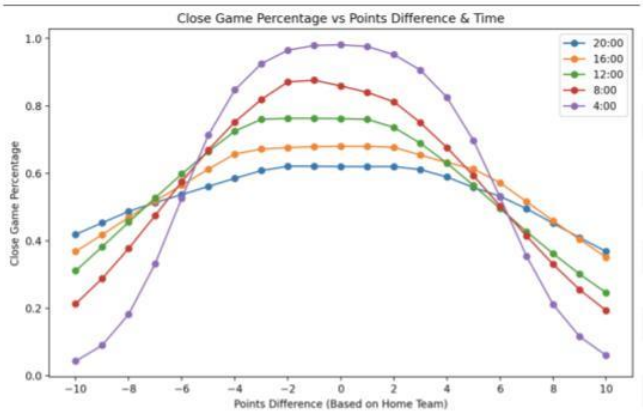


Table 4

	Pts_diff	20:00	16:00	12:00	8:00	4:00
0	-10	0.419	0.368	0.311	0.213	0.043
1	-9	0.453	0.418	0.382	0.288	0.090
2	-8	0.487	0.469	0.455	0.377	0.182
3	-7	0.513	0.518	0.527	0.475	0.332
4	-6	0.537	0.565	0.598	0.575	0.526
5	-5	0.561	0.612	0.664	0.669	0.713
6	-4	0.585	0.657	0.725	0.752	0.847
7	-3	0.608	0.672	0.760	0.819	0.925
8	-2	0.621	0.676	0.763	0.871	0.965
9	-1	0.621	0.679	0.763	0.876	0.979
10	0	0.620	0.680	0.762	0.859	0.981
11	1	0.620	0.680	0.760	0.840	0.976
12	2	0.620	0.677	0.736	0.812	0.952
13	3	0.611	0.654	0.689	0.750	0.906
14	4	0.589	0.632	0.630	0.676	0.825
15	5	0.558	0.613	0.564	0.593	0.697
16	6	0.532	0.572	0.496	0.503	0.529
17	7	0.495	0.516	0.427	0.414	0.354
18	8	0.452	0.459	0.362	0.330	0.211
19	9	0.409	0.404	0.301	0.255	0.116
20	10	0.368	0.351	0.246	0.193	0.060

Just like *Figure 4,* in *Figure 7* with less time left on the clock the curve is getting steeper.

One thing that is interesting to me is that at almost every point the mark of the line almost

mirrors themselves meaning that whether the home team is leading or trailing the game has the same amount of chance of going into a close game when the points difference is the same. The only one expectation I see is 16:00. The highest percentage is skewed to the left. In *Table 4*, I observed that the 50% chance of being in a close game is when the home team is trailing around 8 to 7 or leading 6 to 7. Express that when the home team is trailing more points then they are leading still have the same percentage that the game will be in a close game.

In Table 4, I observed that the largest drop in the model at the 4:00 is from -5 to -7 and and from 5 to 7. This means that when the points difference is at 5 it is still very likely to be a close game, but when the points difference becomes 6 the percentage dropped around 17% no matter if the home team is trailing or leading and it will drop another 18% if the points difference is at 6.

These 6 time points I listed are pretty much mirrored each other with less than 3% of deviation. -5 is 1.6% higher than 5, -6 is 0.3% lower than 6, and -7 is 2.2% lower than 7. Even though I showed so much evidence that the home team has a slight advantage of winning the game, the away team actually has a really similar percentage of pulling the game away to the home team when they are leading in the same amount of points.

Conclusion

We always hear a lot of spectators in sports discussing home court advantage, this report clearly shows that home court advantage does exist. Using a logistics regression model to predict winning percentage and a Neural Network to predict whether the game is going to be a close game or not, the result is favoring the home team. What is even more interesting is that

in all the different times I input into the model, when the game is tied the home team has a 55% chance of winning the game. This can be seen on the graph showing that the winning percentage is leaning towards the home court team even if they are trailing. Even though the home team has an absolute advantage on the winning percentage, either when they are trailing from 5 to 7 points or leading 5 to 7 points the chance of getting into a close game is very similar. Indicating that the home team won't just run away from their opponents. Their opponents still have the ability to keep the game close, but just came up short of winning the game. This would be even more distinct once you used the dashboard and dragged the bar to see how the percentage shifts.

I believe that this product will give basketball spectators new information when they are watching the game. They could use the dashboard and enter the scenario that their team is under and would show a game winning percentage and the percentage that the game is going to be a close game. For users like sportsbooks, they could implement this into their risk management system and make the odds more accurate. On the other hand, amateur bettors could use this dashboard to help them make better decisions on whether to place money in certain situations.

Reference

Nadj, M., Maedche, A., & Schieder, C. (2020). The effect of interactive analytical dashboard features on situation awareness and task performance. *Decision Support Systems*, *135*, 113322. https://doi.org/10.1016/j.dss.2020.113322

Hong Qiao, Jigen Peng, Zong-Ben Xu, & Bo Zhang. (2003). A reference model approach to stability analysis of Neural Networks. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *33*(6), 925–936. https://doi.org/10.1109/tsmcb.2002.804368