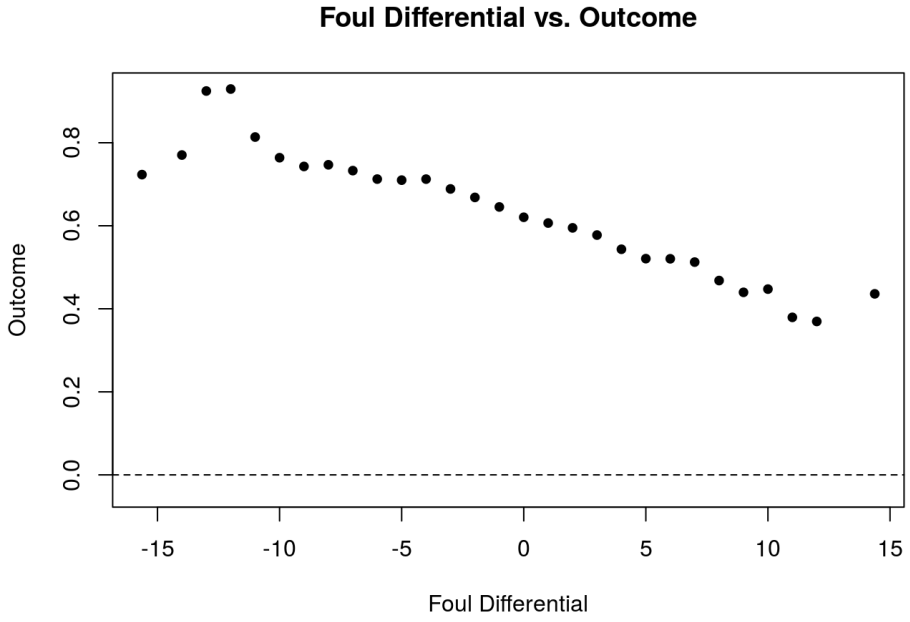
Stats Perform Data Science Evaluation

Dylan Zumar (Duke University)

Exploratory Data Analysis:

First, I mutated the existing variables to mirror the already-created *HT_SCORE_DIFF* by taking the difference of each statistic from the home team and away team. For instance, I created *HT_AST_DIFF* which took the total assists from the away team and subtracted it from the total assists from the home team, as the statistic only holds significance in relation to the other team. I did this for each relevant statistic.

Before creating the model, I used binned plots to determine whether there were any non-linear relationships between the potential explanatory variables and the outcome of the games. As in the plot below, I did not determine that there were any explanatory variables that should immediately be removed from consideration.



As you can see above, foul differential appears to have a clear linear relationship with the game's outcome.

Model Creation:

Because the task at hand was to model a binary outcome of win (1) or loss (0), I decided to create a logistic regression model because it limits the response variable to between zero and one by using a logit function.

I then put all of the differential statistics and game time in a model, then I used backwards selection to filter for the significant predictors. Backwards selection did not remove any of the

explanatory variables, so I proceeded with the full model. I then interacted each variable with *GAME_TIME*, as the statistics hold varying weights depending on the remaining time left in each game.

Finally, I used a nested F test to assess the significance of various other variable interactions. I inferred which statistics would potentially have interactions, and the nested F test confirmed that there were significant interactions between steals and fouls (there is high value in defending well without fouling) in addition to assists and turnovers (a high assist to turnover ratio is widely considered to be a contributor to team success).

Model and Interpretation:

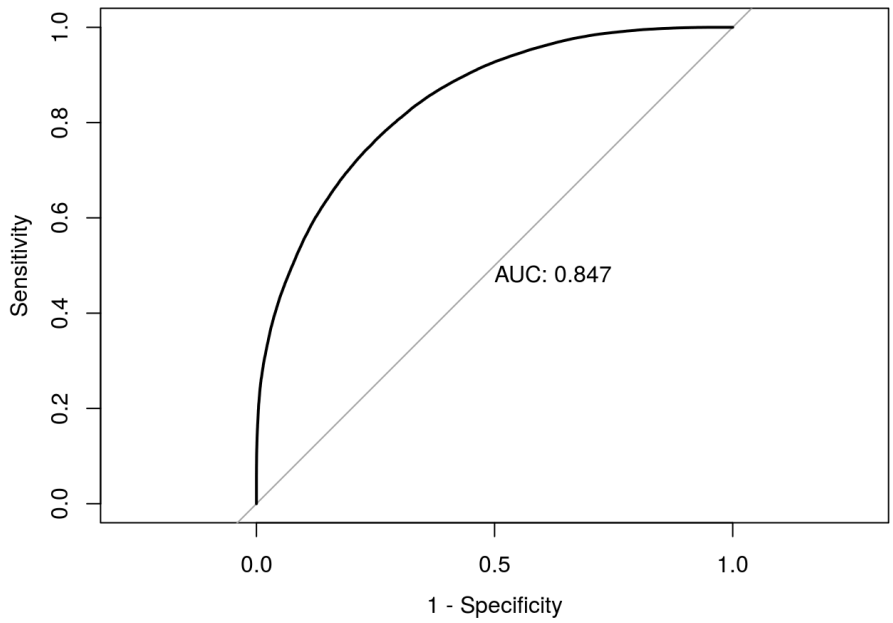
My model is as follows:

$$\begin{aligned}P(\text{OUTCOME} = 1 \mid X_1, X_2\ldots) = {} & 0.5566494 \\ & + 0.0008281 * \text{HT\_SCORE\_DIFF} - 0.1294901* \text{HT\_FOUL\_DIFF} \\ & - 0.0177948 * \text{HT\_ORBD\_DIFF} + 0.0831552 * \text{HT\_DRBD\_DIFF} \\ & - 0.0268101* \text{HT\_TRBD\_DIFF} + 0.0611942 * \text{HT\_AST\_DIFF} \\ & + 0.0279202 * \text{HT\_STL\_DIFF} + 0.0717196 * \text{HT\_BLK\_DIFF} \\ & - 0.1163436 * \text{HT\_TOV\_DIFF} - 0.0001866 * \text{GAME\_TIME} \\ & + 0.0000957 * \text{HT\_SCORE\_DIFF:GAME\_TIME} \\ & + 0.0000284 * \text{HT\_FOUL\_DIFF:GAME\_TIME} \\ & + 0.0000037 * \text{HT\_ORBD\_DIFF:GAME\_TIME} \\ & + 0.0000130 * \text{HT\_DRBD\_DIFF:GAME\_TIME} \\ & + 0.0000156 * \text{HT\_TRBD\_DIFF:GAME\_TIME} \\ & - 0.0000161* \text{HT\_AST\_DIFF:GAME\_TIME} \\ & - 0.0000049 * \text{HT\_STL\_DIFF:GAME\_TIME}\end{aligned}$$

Take *HT_SCORE_DIFF* as an example for interpreting the model's coefficients. As the home team scores 1 more point than the away team, the log odds of the team winning increases by 0.0008281.

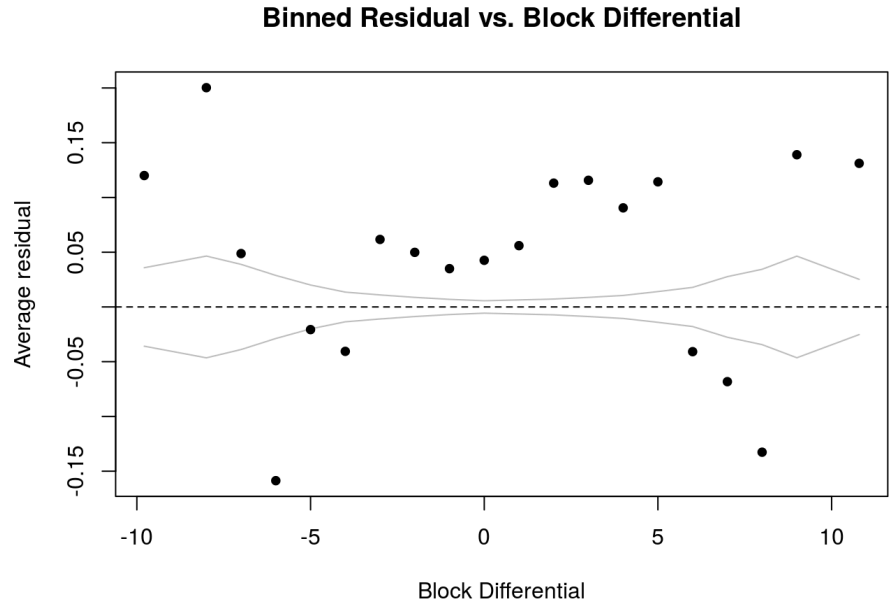
Assessing Model Fit:

To assess the model's fitness, I created a Receiver Operating Characteristic (ROC) curve which assesses the model's false positives and true negatives.



Because the area under the ROC curve is 0.847 and close to 1, it suggests that that my model is fairly accurate in distinguishing between a home team win and loss.

Additionally, I created binned residual plots against my model's predicted outcome and against all of the model's explanatory variables. Ideally, the binned residual plots would show random scatter and no patterns.



While not each explanatory variable had binned residual plots that were randomly scattered, the above plot using block differential and its binned residuals showed random scatter.

Finally, I filtered out some observations to see where my model works and where my model needs improvements. First, I checked the highest win probabilities for each unique game to see if my model works when it is quite certain that the home team will win.

GAME_NUM <int>	OUTCOME <dbl>	GAME_TIME <dbl>	HT_SCORE_DIFF <dbl>	predicted <dbl>
1034	1	2880	51	1.0000000
174	1	2856	44	0.9999998
18	1	2880	45	0.9999997
130	1	2854	45	0.9999996
82	1	2742	48	0.9999996
938	1	2637	41	0.9999980
171	1	2743	39	0.9999975
402	1	2880	36	0.9999967
695	1	2880	40	0.9999944
243	1	2685	36	0.9999937

As we can see, when the game is nearing the end of regulation (*GAME_TIME* = 2,880) and the home team has a significant lead, my model becomes most certain that the home team will win. This contextually makes sense, which is a good sign for the model.

Next, I checked cases where my model was most certain of a home team win despite the home team actually losing.

GAME_NUM <int>	OUTCOME <dbl>	GAME_TIME <dbl>	HT_SCORE_DIFF <dbl>	predicted <dbl>
1092	0	2160	23	0.9977464
1269	0	1627	17	0.9938272
804	0	2449	15	0.9924020
503	0	2591	13	0.9902450
1270	0	1977	15	0.9895706
360	0	2264	14	0.9889124
1265	0	2395	14	0.9887052
280	0	1795	15	0.9865853
1017	0	1720	17	0.9823938
136	0	2088	15	0.9813231

My model seems to misjudge when the home team has a double-digit lead fairly late in the game. While this makes sense, there seem to be some cases where the away team has ample time to make up the deficit, yet my model deems the lead to be more or less unassailable.

After analyzing the false positives, I checked the true negatives wherein my model predicts that there is a minute probability of the home team winning despite them going on to win.

GAME_NUM <int>	OUTCOME <dbl>	GAME_TIME <dbl>	HT_SCORE_DIFF <dbl>	predicted <dbl>
269	1	2190	-27	0.006444161
895	1	2279	-17	0.007407305
1094	1	1961	-19	0.008553571
456	1	2025	-17	0.009079619
1020	1	2278	-15	0.017625271
981	1	1735	-19	0.019583226
1213	1	1508	-16	0.019712061
166	1	2321	-14	0.021036373
1000	1	1800	-14	0.023744689
54	1	2010	-14	0.026144015

Again, my model does not work well with double digit leads late into games. While it is expected that the win probability would be low in these cases, some of the figures such as 0.7% win probability for a 17 point deficit with roughly 9 minutes to go seems to be far too low.

Conclusion/Model Improvements:

Overall, my model seems to follow practical reasoning given circumstances where the winner and loser should be evident. Moreover, the model holds up relatively well according to the ROC plot and binned residual plots. The model is, however, relatively complex due to the number of explanatory variables. In theory, a model that simply uses time and point differential could likely achieve similar results like my model.

One improvement that I would make to my model would be to incorporate time series into the prediction. While the data set included a variable called *EVENT* which provided each significant play throughout the game, I did not incorporate the variable into my model due to the quantity of levels to the variable. However, to improve the model, it would be possible to divide the events into positive (i.e. scoring a basket) and negative events (i.e. turning the ball over) for the home team, and a consecutive series of positive and negative events could be used to predict outcome. This would definitely be helpful in predicting outcomes, as basketball is a sport based heavily on momentum, so having some sort of measurement of positive and negative momentum would greatly benefit the model.

Additionally, the model could likely be improved with various transformations. Some binned residual plot showed patterns which suggests some sort of non-linear relationship with the variable and the outcome of the game, so a transformation would potentially allow the model to predict better.

Finally, the model could also be fit to each individual team better. For instance, if a team with a given ID had the best overall record, then the team's win probability from tipoff should be higher than the average team. My model failed to account for pure success of a team, though I know each team's history is used extensively in real-life win probability models.