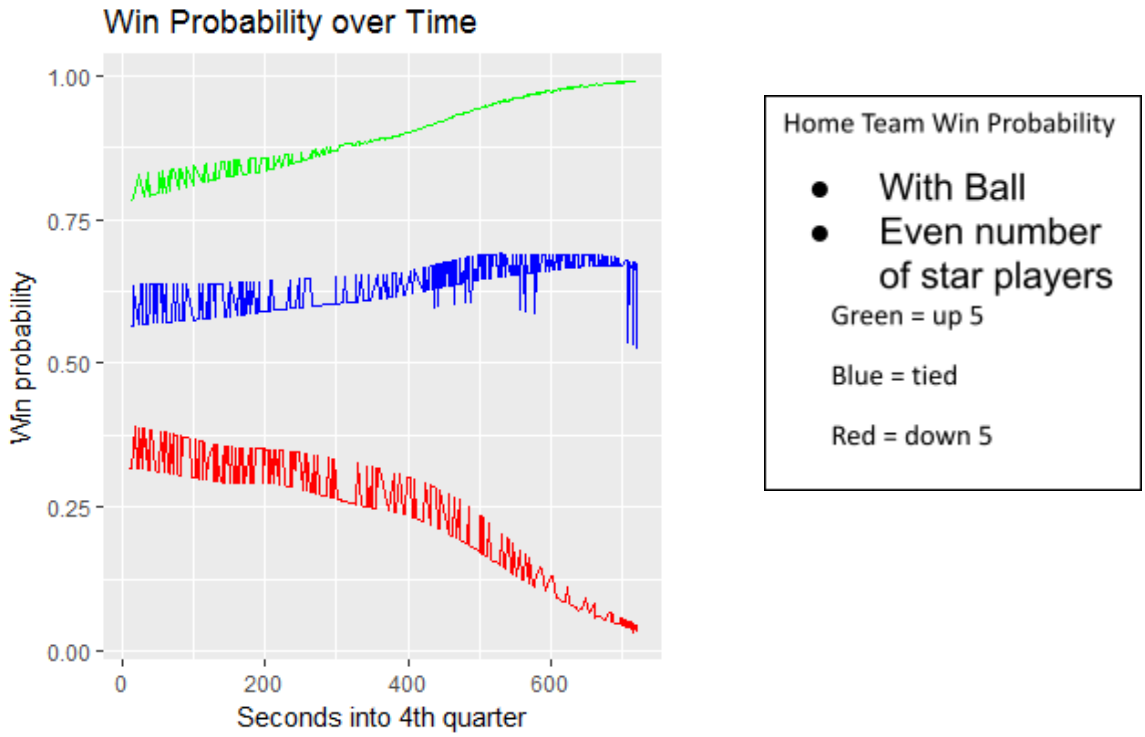
Project

I decided that I wanted to base the model on five parameters:

- Score difference
- Possession of the ball
- Seconds remaining in the game
- Number of star players on the home team
- Number of star players on the road team

To attain proper values for the first three parameters for each row in the file, I wrote some python and used pandas dataframes to pre-process the data and infer possession of the ball by the event data and play descriptions. Star players were determined using Sports Illustrated's list of the top 25 current players in the NBA.

I explored a variety of classic techniques to solving binary response problems such as logistic regression, random forests, and support vector machines, but these methods did not do a sufficient job of capturing the temporal aspect of the in-game win probability model. Therefore I decided to use a non-parametric method called locally weighted logistic regression, which weights points by their proximity to the current point and then performs regression on those weighted points. I trained this model on a training set that composed of about 70% of the data and then tested it on the rest of the data.



This visualization graphs the trends of win probabilities in three separate cases for home teams with the ball and an equal number of star players on the court as the away team. Those three cases are: up five points (green), tie game (blue), and down five points (red). This plot demonstrates the relative importance of score differential at various points in the fourth quarter. It was made by filtering the test data for cases that satisfy the given conditions and then graphing the predicted win probability vs. seconds into the fourth quarter using R's ggplot2 package. This methodology could easily be adjusted to analyze another scenario, e.g. the probability of a win over time for when a home team that is tied that has one less, equal, or one more star than the away team.

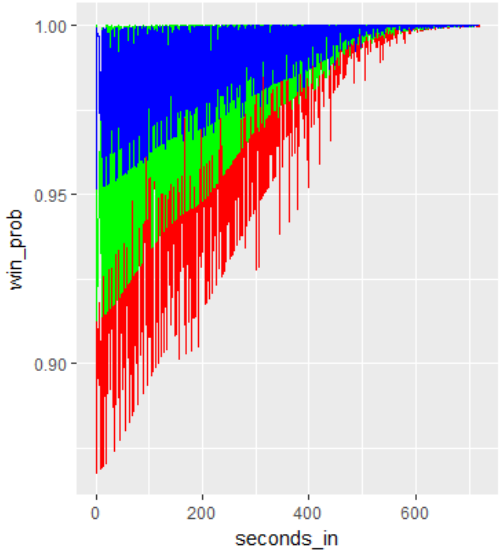
Game Scenario

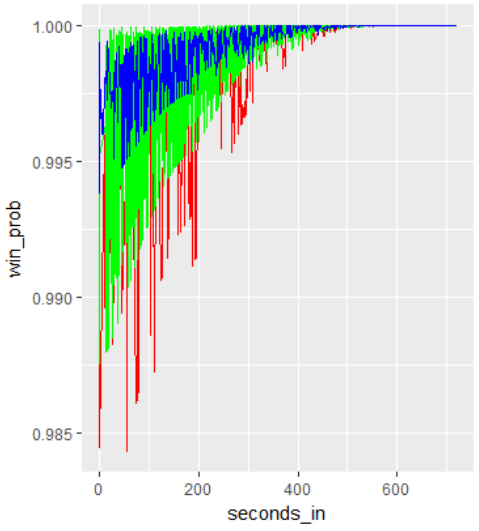
a. **We're down 106-105 with 29 seconds to go. They get a defensive rebound. Should we foul in that situation?**

   Below are the possible results from each scenario, along with the probability of a win given that event:

   - Foul (leaving 29 seconds to go)
     - Down 3, probability = P(2 made free throws) = $(0.8)^2$ = .64
       - Probability win = .208
     - Down 2, probability = .32
       - Probability win = .345
     - Down 1, probability = .04
       - Probability win = .516
     - **Cumulative probability of win = 26.4%**
   - Don't foul (leaving ~5 seconds to go, assuming the other team uses the whole shot clock)
     - Down 4, probability = P(3 point attempt) * P(3 points | 3 point attempt) = .33 * .35 = .116
       - Probability win = .094
     - Down 3, probability = P(2 point attempt) * P(2 points | 2 point attempt) = .67 * .5 = .335
       - Probability win = .180
     - Down 1, probability = .549
       - Probability win = .495
     - **Cumulative probability of win = 34.3%**

   Suggestion: Do not foul because your chance of winning in this case is approximately 34.3% versus a chance of 26.4% if you do foul.





Win probability vs. seconds into 4th quarter with home team up by 20 or more.