Win Probability Model

Engel, Alec

7/26/2021

pbpfull <- pbpfull %>%   
 mutate(winteam = ifelse(result > 0,home\_team, ifelse(result<0,away\_team,"tie")))

pbpfull <- pbpfull %>%  
 mutate(poswins = ifelse(winteam == posteam,"PosWins", "PosLoses"))

pbpfull <- pbpfull %>%  
 mutate(poswins = fct\_relevel(poswins, "PosLoses"))

pbpfull <- pbpfull %>%  
 mutate(posspread = ifelse(posteam == home\_team, spread\_line, -1\*spread\_line))

cols = c("qtr","down","poswins")  
pbpfull = pbpfull %>% mutate\_at(cols,as\_factor)

pbpfull = pbpfull %>% drop\_na(yardline\_100,game\_seconds\_remaining,down,posspread,score\_differential)

pbpfull = pbpfull %>% filter(qtr != 5)

pbpfull = pbpfull %>% filter(winteam != "tie")

mod1 = glm(poswins ~ yardline\_100 + game\_seconds\_remaining + down +  
ydstogo + posspread + score\_differential, data = pbpfull, family = "binomial")  
options(scipen = 999)  
summary(mod1)

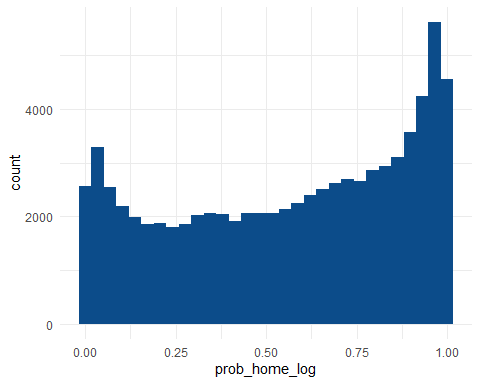
##   
## Call:  
## glm(formula = poswins ~ yardline\_100 + game\_seconds\_remaining +   
## down + ydstogo + posspread + score\_differential, family = "binomial",   
## data = pbpfull)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.77967 -0.70990 0.03902 0.72009 2.75136   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.89297229 0.03658383 24.409 < 0.0000000000000002 \*\*\*  
## yardline\_100 -0.00764009 0.00040078 -19.063 < 0.0000000000000002 \*\*\*  
## game\_seconds\_remaining -0.00001609 0.00000888 -1.811 0.070072 .   
## down2 -0.07916587 0.02317970 -3.415 0.000637 \*\*\*  
## down3 -0.18613450 0.02710723 -6.867 0.00000000000657 \*\*\*  
## down4 -0.36305542 0.03365284 -10.788 < 0.0000000000000002 \*\*\*  
## ydstogo -0.01285875 0.00245459 -5.239 0.00000016175409 \*\*\*  
## posspread 0.11797229 0.00163133 72.317 < 0.0000000000000002 \*\*\*  
## score\_differential 0.18525467 0.00152764 121.269 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 108615 on 78352 degrees of freedom  
## Residual deviance: 68623 on 78344 degrees of freedom  
## AIC: 68641  
##   
## Number of Fisher Scoring iterations: 5

This looks to be a strong model with all predictors except for game\_seconds\_remaining being statistically significant. We can see that as posspread and score\_differential increases, the probability of the team in possession winning also increases. It also makes sense that as yardage, game time and downs decrease, the probability of the team in possession in winning decreases because that team with posession may be running out of time, being backed up on the field, or running closer and closer to 4th down opportunities.

predictions\_log = predict(mod1, type = "response")

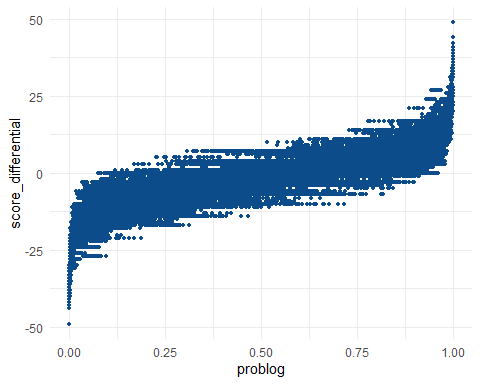
pbpfull = pbpfull %>% mutate(problog = predictions\_log) %>%  
mutate(prob\_home\_log = ifelse(posteam == home\_team, problog , 1-problog))

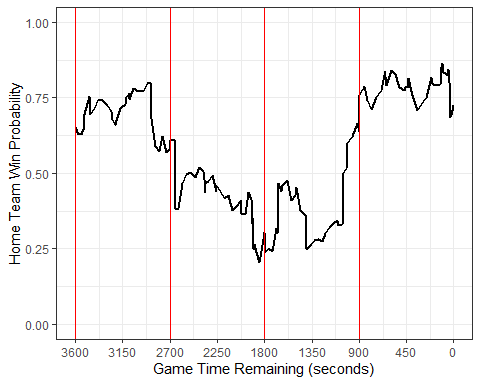
#esquisser()  
ggplot(pbpfull) +  
 aes(x = prob\_home\_log) +  
 geom\_histogram(bins = 30L, fill = "#0c4c8a") +  
 theme\_minimal()



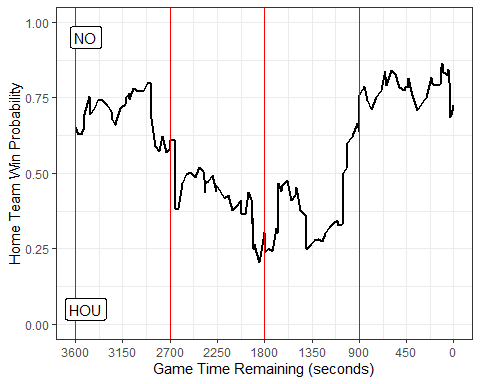
This distribution of “prob\_home\_log” looks to be a bit bimodal with either much higher tails of win or loss probabilities rather than 50/50 chances.

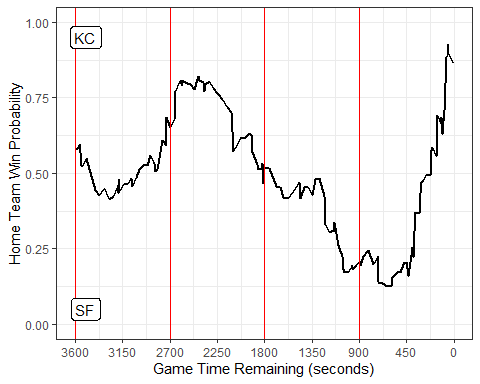
#esquisser()  
ggplot(pbpfull) +  
 aes(x = problog, y = score\_differential) +  
 geom\_point(size = 1L, colour = "#0c4c8a") +  
 theme\_minimal()

 This comparison chart of win probability of the team in possession to score differential aligns with what I would imagine. The better a teams score differential or the more lead they have, the higher probability that the team with possession will win as they are most likely running out the clock. The same works in opposite direction because the more behind a team is, the less possession they may have since they are trying to score quickly and not run out the clock.



ESPN’s chart ratio not only measures the game time, but also highlights specific plays that may increase or decrease probabilities. Our win probability model only includes the Saint’s win probability. While we can assume the inverse probability would be the Texans, it is much more visually appealing for a viewer to see both ends of the spectrum. The dip at the 36 second mark when Houston lead seemed a bit less drastic in our chart however which proves to be okay since realistically the Saint’s actually still had a decent shot at driving down the field and kicking a field goal even with that amount of time remaining.





As a 49ers fan, I am regrettably taking a look back at Super Bowl LIV in 2019 against the Kansas City Chiefs. Our created chart does show a glimmer of hope in the 1st quarter with higher 49ers probability although the ESPN chart does not show San Francisco with a higher probability of winning until the 3rd quarter. ESPN’s chart looks to sync up with ours from that point onward as the 49ers had higher probability of winning until a steep decline at roughly 5 minutes left when Mahomes gets the ball down by 3 points. Then KC had higher probability the rest of the game once they scored a touchdown to go up 24-20 with 2 minutes and 44 seconds left in the game.Both charts definitely tell the game story but I do enjoy the additional vertical lines separating out quarters as we see the 49ers leading win probability quarters one and four, while the Chiefs led in probability for quarter two and then of course the most important time in the game which was the last several minutes and end.

