Group Members: Julia Stiller, Eli Koester, and Josip Kauzlaric

Group name: The Data Devils

Weather and Win Probability in the NFL

**Abstract**

This project attempts to see if weather affects the result of a football game after accounting for the percentage chance that a team is going to win. Prior studies have shown that temperature does affect the outcome of games and whether the home team wins, so we wanted to see if weather would add to accounting for the probability that a team will win. We used Elo probability which assigns each team a rating based on their prior performances and uses that to calculate a percent chance that one team will win over the other. It would be interesting to see if weather can improve Elo which would optimize their calculations while giving bettors an upper hand if they wanted to win money on a game. We collected data from the 2009 to 2013 NFL seasons which included over 1,000 games with the results of the game and the weather at them which included temperature, wind speed, and humidity. We carried out analyses with these variables, individually and including Elo in the model, with the result being if the home team won or not. We did not find any significant results, except that Elo is a significant predictor of the odds of the home team winning (at a 0.05 significance level).

**Introduction**

The NFL (National Football League) is an American football league that has been analyzed closely since it began in 1920. Like any type of sport, the NFL is full of statistics that relate to team metrics of success and individual player metrics of success. The Data Devils are all very interested in the statistics of sports and decided to choose the NFL because this is a common favorite of all group members. When developing the research question, we were hypothesizing possible non-traditional predictors of a home team’s success in the NFL. We were wondering what can contribute to the success of a home team, besides the traditional measures of success; such as win-loss record, playoff appearances, total yards, total touchdowns, etc. We decided to analyze weather and the varying effects that weather has on the success of a home team. We chose weather because many NFL stadiums are open-roofed, and the football game is affected by the outside weather conditions that these open-roofed stadiums cannot control.

We came up with some non-traditional predictors when considering what changes from each NFL team and the differing stadiums that each team plays in. Our predictions were that the home team (home team stadium) will have an advantage given weather during the game. After accounting for each team’s ELO probability, we used variables such as wind speed, temperature, and humidity to assess whether or not the home team has an advantage. An example of weather having an impact on a home team’s success is the Snow Bowl of 2002. This game was played between the New England Patriots and the Oakland Raiders. The New England Patriots are based out of Foxboro, Massachusetts; while the Oakland, Raiders are based out of Oakland, California. These two locations are very different in terms of weather conditions and the game was played in Foxboro, so we would expect the New England Patriots to have an advantage. It turns out, they ended up winning the football game by a field goal, which was kicked by an experienced kicker. The kicker who kicked the field goal was used to the weather conditions and extreme snow, so we believe that this advantage was due to the location of the stadium and the weather associated with the location. Oakland has a very high average temperature and these players were used to playing the dry heat conditions, not heavy snow.

From our statistical analysis, we can test our null hypothesis; which is that none of the non-traditional weather predictors are significant indicators of a home team win. The alternative hypothesis would be that at least one of the weather condition predictors are statistically significant predictors of a home team win. We accumulated data from four NFL seasons, 2009 – 2013, for all games played for all 32 NFL teams. We used multiple sources to construct the data set. We imported the data set from Microsoft Excel into R Studio to conduct some statistical analysis. We used multiple logistic regression with a categorical response variable. This response variable was home\_win, which was coded 1 = home team win; and 0 = home team loss or tie. We analyzed the significance of the non-traditional weather predictors and used graphical analysis to determine the relationship.

**Data**

Sources:

Weather data- <http://nflsavant.com/dump/weather_20131231.csv>

This dataset has the weather as well as the score and which team was home for every NFL game since 1960. We used from the years 2009 to 2013.

Elo data- <https://projects.fivethirtyeight.com/nfl-api/nfl_elo.csv>

This dataset contains every Elo score for every team and their win probability along with who the home and away teams were for every game since 1920. We used only data from 2009 to 2013. These two data sources were then combined in to one dataset in which all the games aligned with each other to give us our complete data.

Sample and population:

There were 1,108 games in our dataset that we used, with each observation being a game played and the outcome being if the home team won or not. The estimated population size of all games played is over 16,000, so with our sample, we will be trying to make inferences for future games to be played. For our purposes, we only used data from 2009 to 2013, because we are trying to predict what will happen in the future, and the further back in time we go, the less applicable it may be to today’s game because things in the NFL change constantly. At the same time, we want to have enough data to analyze and see if there are any patterns. This sampling design is not random, and this is an observational study.

Variables:

home\_win- A categorical variable coded as 1 if the home team won and 0 if they lost. This is our response variable.

temperature- The temperature (in degrees Fahrenheit) at the location of the game. An explanatory quantitative variable ranging from 0 to 96.

extremetemp- The category of temperature at the location of the game. An explanatory categorical variable of temperature that bins the temperatures in to 3 categories: 0 = 30 to 80 degrees Fahrenheit, 1 = 30 degrees Fahrenheit and below, 2 = 80 degrees Fahrenheit and above.

humidity- The percent of moisture in the air as measured on the day of the game played. An explanatory quantitative variable ranging from 0 to 0.98

wind\_mph- The speed of wind at the location of the game. An explanatory quantitative variable ranging from 1 to 31.

elo\_prob1- Probability that the home team was going to win the game. An explanatory quantitative variable ranging from 0.11 to 0.95

Elo is a rating system invented by Arpad Elo in the mid 20th century to calculate the skill of a player in a zero-sum game where one player wins and the other loses. It has been adapted for various sports including football by the organization Five Thirty-Eight. A rating is calculated for each team for each game; the outcomes of the previous games each team played in are accounted for along with home field advantage in an equation that gives the team a score. The scores of the two teams are then used to generate the win probability for each team through a formula developed by Five Thirty-Eight.

**Results**

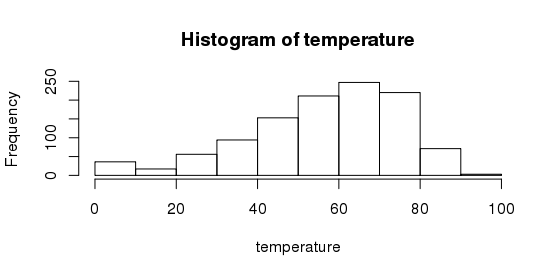
Temperature:

> summary(temperature)

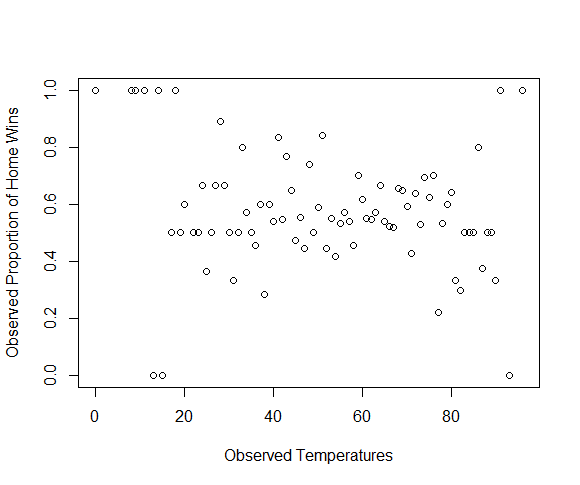
Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 46.00 60.00 56.59 72.00 96.00

> hist(temperature)



With a mean of 57 and a median of 60, we see that the typical game is not too extreme in temperature, but it’s seen in the histogram that there are still some extreme temperature games.



This is the graph of the observed proportion of home team wins for each observed temperature. We see that at the ends of the graph, there are temperatures with so few games that they have 0% or 100% of home team wins.

> model1 <- glm(home\_win~temperature, family = binomial)

> summary(model1)

Call:

glm(formula = home\_win ~ temperature, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.358 -1.297 1.032 1.060 1.096

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.449192 0.226227 1.986 0.0471 \*

temperature -0.002655 0.003724 -0.713 0.4760

The quantitative variable temperature is not a significant predictor of the odds of the home team winning (z-score = -0.713, p-value = 0.476).

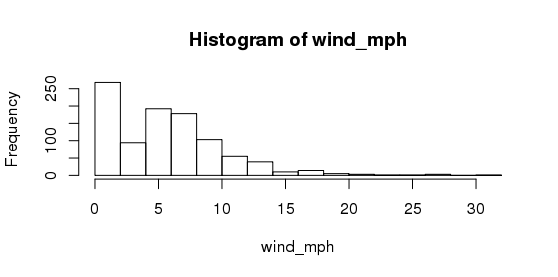
Wind Speed:

> summary(wind\_mph)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 2.000 6.000 6.071 8.000 31.000

> hist(wind\_mph)



Most games are not very windy as seen by the mean and median around 6 and the 3rd quartile only at 8. Since most games have wind speeds below 10, the majority of games are not played in extreme wind conditions.

> model2 <- glm(home\_win~wind\_mph, family = binomial, data = EloWeather)

> summary(model2)

Call:

glm(formula = home\_win ~ wind\_mph, family = binomial, data = EloWeather)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.537 -1.295 1.014 1.064 1.099

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.16757 0.10989 1.525 0.127

wind\_mph 0.02089 0.01476 1.416 0.157

The quantitative variable wind speed is not a significant predictor of the odds of the home team winning (z-score = 1.416, p-value = 0.157).

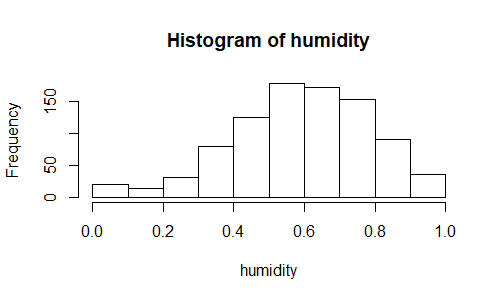
Humidity:

> summary(humidity)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.0000 0.4700 0.6100 0.5947 0.7400 1.0000 204

> hist(humidity)



Most games have a 60% humidity level. The histogram shows that the majority of games are played in an average amount of humidity.

> model3 <- glm(home\_win~humidity, family = binomial)

> summary(model3)

Call:

glm(formula = home\_win ~ humidity, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.324 -1.288 1.050 1.069 1.113

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.1525 0.2117 0.720 0.471

humidity 0.1851 0.3382 0.547 0.584

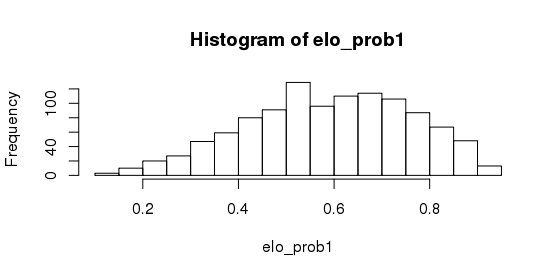
Humidity is not a significant predictor of the odds of the home team winning (z-score = 0.547, p-value = 0.584).

Elo Probability:

> summary(elo\_prob1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.1162 0.4688 0.5981 0.5880 0.7228 0.9431

> hist(elo\_prob1)  


The median and mean are above 0.5 showing that the home team is predicted to win more than 50% of the time. This is due to how Elo calculates the score and factors in home field advantage.

> model4 <- glm(home\_win~elo\_prob1, family = binomial)

> summary(model4)

Call:

glm(formula = home\_win ~ elo\_prob1, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9131 -1.1366 0.7015 0.9920 1.8298

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.1238 0.2418 -8.784 <2e-16 \*\*\*

elo\_prob1 4.1656 0.4047 10.292 <2e-16 \*\*\*

The associated Elo probability of the home team winning is a significant predictor of the odds of the home team winning (z-score = 10.292, p-value = 2\*10^-16). For each 0.1 increase in Elo probability, the odds of the home team winning increases by 51.67% (e^(0.1\*4.1656)=1.5167).

Full Model:

> model5 <- glm(home\_win~elo\_prob1 + temperature + wind\_mph + humidity, family = binomial)

> summary(model5)

Call:

glm(formula = home\_win ~ elo\_prob1 + temperature + wind\_mph +

humidity, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0015 -1.1172 0.6862 1.0011 1.9389

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.4859397 0.4895346 -5.078 3.81e-07 \*\*\*

elo\_prob1 4.3014770 0.4607400 9.336 < 2e-16 \*\*\*

temperature 0.0001463 0.0044350 0.033 0.9737

wind\_mph 0.0285123 0.0163881 1.740 0.0819 .

humidity 0.0988244 0.3888210 0.254 0.7994

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1196.2 on 873 degrees of freedom

Residual deviance: 1094.3 on 869 degrees of freedom

(234 observations deleted due to missingness)

AIC: 1104.3

Number of Fisher Scoring iterations: 4

After accounting for weather (temperature, wind speed, and humidity) and a precalculated Elo probability, the Elo variable was the only significant predictor of the odds of the home team winning. Backwards eliminating variables resulted in the reduced model with just Elo probability (model4 as seen above). To confirm that the variables in the full model are not significant, we calculated a G-statistic using the aforementioned reduced model as well as the full model filtering out the data with any missing values in order to have the same sized data sets.

> anova(model4reduced, model5)

Analysis of Deviance Table

Model 1: home\_win ~ elo\_prob1

Model 2: home\_win ~ elo\_prob1 + temperature + wind\_mph + humidity

Resid. Df Resid. Dev Df Deviance

1 872 1097.5

2 869 1094.2 3 3.2566

> pchisq(G, df=3, lower.tail = FALSE)

[1] 0.3537289

After accounting for Elo probability, there is no evidence that the coefficients of temperature, wind speed, or humidity are significantly different from 0 (G-stat = 3.2566, p-value = 0.3537).

**Conclusions/Inferences**

We had set out to see if weather could add any predictive value to Elo in determining if a home team was going to win an NFL game. From previous studies it was seen that at least temperature on its own was a significant predictor on if the home team won the game and so we wanted to confirm that and then adjust for Elo to see if an even better model could be produced. It turns out that we could not reject the null and could not find any significant results on if weather correlates with the home team winning. Elo on its own was a very significant predictor, meaning that statistic is very accurate in its own right. No weather variables on their own or while factoring in Elo were significant at 0.05 level. This is not to say that weather does not affect the outcome of a football game. As seen in previous studies, at least temperature affects which team wins. Our differences in results could be due to the grouping of data, because the other study we looked at grouped the teams by where the away team normally plays (in a cold climate, warm climate, moderate climate, or in a dome stadium), so this could affect the results.

**References**

Friscojosh. “The NFL's Home-Field Advantage Is Real. But Why?” *FiveThirtyEight*, FiveThirtyEight, 21 Dec. 2018, fivethirtyeight.com/features/the-nfls-home-field-advantage-is-real-but-why/.

natesilver538. “How Our NFL Predictions Work.” *FiveThirtyEight*, FiveThirtyEight, 5 Sept. 2018, fivethirtyeight.com/methodology/how-our-nfl-predictions-work/.

“Pro Football Statistics and History.” *Pro*, www.pro-football-reference.com/.

**ELO Data:**

<http://nflsavant.com/dump/weather_20131231.csv>

<https://projects.fivethirtyeight.com/nfl-api/nfl_elo.csv>

Note: These files are CSV files and a citation could not be created for the actual file.