SCORE_Project

Avi Benki and Ryan Henry

2024-04-21

1. Learning Goals To conduct a multivariable analysis of a binary outcome, we must be able to:

a. Read data from a CSV file

b. Select key variables and format them to be computed

c. Create a logistic regression model

d. Analyze p-values and summarize results

2. Introduction Can someone predetermine which fighter will win a fight in UFC? Is it stance, height, weight, age, number of strikes landed, or some combination? Every fighter and their style is unique so what plays the most significant role in helping a fighter win a fight? Analyzing each UFC fight and its winner, we will used a logistic regression model to explain what factor(s) play the most significant role in determining who is more likely to win a given fight.

3. Data The data set that will be used is the "fights" data set. This includes over 6000 fights dating back to 1994 and summarizes all stats for each fighter during the fight such as height, weight, reach, stance, style as well as each fighters previous average number of take downs, significant strikes which is further broken down to shots to the head, body, shots in the clinch, etc. It also gives how each fight ended whether that be by knockout, split decision, unanimous decision, how many rounds the fight was scheduled for and what round it ended in. A few rows have been included below.

Downloading data

Use the code below to download the the UFC data to your computer. We will save the fight data in a data frame called fights, and information about the fighters in a data set called fighters. For simplicity the following tables make up all of the stats for the "Blue" fighter, the data also includes all of the same stats for the "Red" fighter. We have included a couple columns below, but feel free to check out the rest of the data in your fights data set.

```r
library(tidyverse)
library(broom)
library(knitr)
library(rsample)
fights <- read.csv("C:/Users/avaneesh.benki/Documents/data.csv")
```

R_fighter	B_fighter	Referee	date	location	Winner	title_bout	weight_class
Adrian Yanez	Gustavo Lopez	Chris Tognoni	2021-03-20	Las Vegas, Nevada, USA	Red	False	Bantamweight
Trevin Giles	Roman Dolidze	Herb Dean	2021-03-20	Las Vegas, Nevada, USA	Red	False	Middleweight
Tai Tuivasa	Harry Hunsucker	Herb Dean	2021-03-20	Las Vegas, Nevada, USA	Red	False	Heavyweight

B_avg_KD	B_avg_opp_KD	B_avg_SIG_STR_pct	B_avg_opp_SIG_STR_pct	B_avg_TD_pct
0.0	0	0.42	0.50	0.33
0.5	0	0.66	0.31	0.30
NA	NA	NA	NA	NA

4. Methods and Instructional Content It will be useful while constructing our SCORE Module to reference other work that has been done on teaching statistics, especially in mixed martial arts. The first source that we found is a Kaggle tutorial at this link : https://www.kaggle.com/code/olawaleoladipo/ufc-data-analysis-visualization-beginner . Although the tutorial works through python and not R, it is useful as an example of how to introduce statistical concepts to individuals who might not have much experience with statistics, or even MMA. A very interesting part of the tutorial worth looking at is the last section in which many different MMA statistics are visualized in a variety of models. This will be useful to emulate when making our SCORE module so that the student can help portray their findings in a digestible way.

Another useful reference to use, can be found at this link : https://www.datacamp.com/tutorial/generalized-linear-models. This is a very accessible and interesting tutorial on Generative Linear Models in R that starts with the expectation that one knows almost nothing about them. It is written and implemented in a very understandable and fun way, which is how we want our own module to end up as well.

The fights data set has a fight in each of row, with fight statistics given in terms of Blue corner fighter (B_fighter) and Red corner fighter (R_fighter). Each row in the data set represents a match-up, and includes the statistics for both fighters. The statistics are given as an average of how they have done in their fights up until the date of the fight in any given row as well as how well their opponents have done, so if the Blue corner fighter landed 8,4, and 6 head kicks each in their only three fights, which option below would we expect to see in the data frame?

a. B_avg_opp_HEAD_landed = 4

b. B_avg_HEAD_att = 6

c. B_avg_HEAD_landed = 6

d. B_avg_HEAD_landed = 8,4,6

The answer is c.

Data cleaning

Now that we have a basic understanding of how the data is laid out, lets combine the datasets to create a tailored data frame to make our analysis easier. The first thing we can do is remove every time the fight ended in a draw or a no contest. This can be done by using the filter function.

First we need to figure out what the possible options are in the “Winner category”

```r
fights %>%
  select(Winner) %>%
  unique()
```

```
##    Winner
## 1     Red
## 4    Blue
## 13   Draw
```

Now we can take out the Draws.

```r
fights_new <- fights %>%
  filter(Winner != "Draw")
```

In order to continue to hone our dataset, lets restrict the data set to only include the columns that we will be using in our analysis. This can be done using the select function.

```r
fights_new <- fights_new %>%
  select(B_fighter,R_fighter, Winner, date,  B_avg_SIG_STR_pct, B_avg_TD_pct,
         B_avg_SIG_STR_landed, B_avg_opp_SIG_STR_landed, B_avg_TD_landed,B_wins,
         B_losses, B_Stance, B_Height_cms, B_Reach_cms, B_Weight_lbs,
         R_avg_SIG_STR_pct, R_avg_TD_pct, R_avg_SIG_STR_landed,
         R_avg_opp_SIG_STR_landed, R_avg_TD_landed,R_wins, R_losses,
         R_Stance, R_Height_cms, R_Reach_cms, R_Weight_lbs, R_age, B_age)
```

Lets take a look at the reduced data set. Like before, navigate to your own data set to view the full data.

```r
tfnew1 <- fights_new %>%
  select(B_fighter, R_fighter, Winner, date, B_avg_SIG_STR_pct,
         B_avg_TD_pct) %>%
  head(5) %>%
  kable()
tfnew2 <- fights_new %>%
  select(B_avg_SIG_STR_landed, B_avg_opp_SIG_STR_landed, B_avg_TD_landed,
         B_wins) %>%
  head(5) %>%
  kable()
tfnew3 <- fights_new %>%
  select(B_losses, B_Stance, B_Height_cms, B_Reach_cms, B_Weight_lbs,
         R_avg_SIG_STR_pct) %>%
  head(5) %>%
  kable()

tfnew1
```

B_fighter	R_fighter	Winner	date	B_avg_SIG_STR_pct	B_avg_TD_pct
Gustavo Lopez	Adrian Yanez	Red	2021-03-20	0.420000	0.330
Roman Dolidze	Trevin Giles	Red	2021-03-20	0.660000	0.300
Harry Hunsucker	Tai Tuivasa	Red	2021-03-20	NA	NA
Montserrat Conejo	Cheyanne Buys	Blue	2021-03-20	NA	NA

B_fighter	R_fighter	Winner	date	B_avg_SIG_STR_pct	B_avg_TD_pct
Macy Chiasson	Marion Reneau	Blue	2021-03-20	0.535625	0.185

```
tfnew2
```

B_avg_SIG_STR_landed	B_avg_opp_SIG_STR_landed	B_avg_TD_landed	B_wins
20.0000	45.0000	1.0	1
35.0000	16.5000	1.5	2
NA	NA	NA	0
NA	NA	NA	0
57.9375	28.4375	1.5	4

```
tfnew3
```

B_losses	B_Stance	B_Height_cms	B_Reach_cms	B_Weight_lbs	R_avg_SIG_STR_pct
1	Orthodox	165.10	170.18	135	0.5000000
0	Orthodox	187.96	193.04	205	0.5768750
0	Orthodox	187.96	190.50	241	0.5389063
0	Southpaw	152.40	154.94	115	NA
1	Orthodox	180.34	182.88	135	0.4030762

Which of the following statistics is not considered by our new dataset?

a. average takedowns by the opponents of the Blue Corner fighter

b. Red corner fighter's average reversals

c. Blue corner fighter's stance

d. Average significant strike percentage by red corner fighter

answer b.

In order to simplify our analysis, we look at the data as differences between the red and blue corner fighters. Using significant strikes as an example, if the red corner averages 30 per fight and the blue corner averages 45, the new data point would be -15. This will be useful when trying to find a regression line that considers the winner of the fight with the differences in different aspects of fighting. By analyzing the differences in statistics between fighters, we make it easier to determine if there is a positive or negative linear correlation to winning. The assumption is that typically the taller fighter with more reach who lands the most significant strikes will likely be the winner. We'll look at the differences of Red - Blue and see if as the height gap, gap between the fighters number of take downs, and other variables grows, so does the probability of the red fighter winning. One important thing to mention is that for all the fights recorded in this data set is that typically the current champion or fighter with more wins is represented by the red corner. To do this, we will utilize the mutate function.

```
fights_new <- fights_new %>%
  mutate(diff_avg_SIG_STR_pct = R_avg_SIG_STR_pct - B_avg_SIG_STR_pct,
         diff_avg_TD_pct = R_avg_TD_pct - B_avg_TD_pct,
         diff_avg_SIG_STR_landed = R_avg_SIG_STR_landed - B_avg_SIG_STR_landed,
         diff_avg_opp_SIG_STR_landed = R_avg_opp_SIG_STR_landed - B_avg_opp_SIG_STR_landed,
         diff_avg_TD_landed = R_avg_TD_landed - B_avg_TD_landed,
         diff_age = R_age - B_age,
         diff_Reach_cms = R_Reach_cms - B_Reach_cms,
         diff_height = R_Height_cms - B_Height_cms,
         diff_wins = R_wins - B_wins,
         diff_losses = R_losses - B_losses,
         diff_fights = (R_wins + R_losses) - (B_wins + B_losses))

fights_diff <- fights_new %>%
  select(R_fighter, B_fighter, Winner,date, diff_avg_SIG_STR_pct,
         diff_avg_TD_pct, diff_avg_TD_landed, diff_wins,
         diff_avg_SIG_STR_landed, diff_avg_opp_SIG_STR_landed, diff_age,
         diff_Reach_cms, diff_height, diff_losses, diff_fights)
```

We now have all the same data, but now in the form of red fighter minus blue fighter. By running a logistic regression on all the times that the red fighter won, we will be able to discover how much the differences between the fighters impact a win.

Lets make one final change and remove all the rows that contain N/A for any value.

```
Final_df <- na.omit(fights_diff)
```

Lets take one last look at the data set.

```
tfinal1 <- Final_df %>%
  select(R_fighter, B_fighter, Winner, date, diff_avg_SIG_STR_pct,
         diff_avg_TD_pct) %>%
  head(5) %>%
  kable()
tfinal2 <- Final_df %>%
  select(diff_avg_SIG_STR_landed, diff_avg_opp_SIG_STR_landed,
         diff_avg_TD_landed) %>%
  head(5) %>%
  kable()
tfinal3 <- Final_df %>%
  select(diff_age, diff_Reach_cms, diff_height, diff_wins, diff_losses,
         diff_fights) %>%
  head(5) %>%
  kable()
tfinal1
```

	R_fighter	B_fighter	Winner	date	diff_avg_SIG_STR_pct	diff_avg_TD_pct
1	Adrian Yanez	Gustavo Lopez	Red	2021-03-20	0.0800000	-0.3300000
2	Trevin Giles	Roman Dolidze	Red	2021-03-20	-0.0831250	0.1062500
5	Marion Reneau	Macy Chiasson	Blue	2021-03-20	-0.1325488	0.3267188
6	Leonardo Santos	Grant Dawson	Blue	2021-03-20	0.0501563	-0.0979688
7	Song Kenan	Max Griffin	Blue	2021-03-20	0.0338477	-0.3221875

```
tfinal2
```

	diff_avg_SIG_STR_landed	diff_avg_opp_SIG_STR_landed	diff_avg_TD_landed
1	-3.00000	-39.000000	-1.0000000
2	8.15625	11.093750	-0.7187500
5	-13.57520	56.117188	-0.2382812
6	1.43750	6.921875	-1.9921875
7	-26.61328	-24.048828	-1.4531250

```
tfinal3
```

	diff_age	diff_Reach_cms	diff_height	diff_wins	diff_losses	diff_fights
1	-4	7.62	5.08	0	-1	-1
2	-4	-5.08	-5.08	2	2	4
5	14	-10.16	-12.70	1	5	6
6	14	7.62	5.08	3	1	4
7	-4	-12.70	2.54	0	-5	-5

Nice! Now that we have our data all cleaned up, lets begin analyzing what qualities makes a fighter more likely to win. Look at all the columns and make a prediction as to what factor(s) are the most significant to winning a UFC fight.

ANS: (ex. height and significant strikes landed)

One thing statisticians do when analyzing data is splitting it into training and test groups. This means that a model is trained with only the data in the training group, and then its accuracy is checked against the test group. For this dataset, we will be using a 75/25 split, where the data will be trained on a randomly selected 75% of the rows.

```
dt = sort(sample(nrow(Final_df), nrow(Final_df)*.75))
traindiff<-Final_df[dt,]
testdiff<-Final_df[-dt,]
```

Logistic Regression

To analyze the data with the training set, we will use a linear regression model. This means that we are assuming all dependent variables have a linear relationship with winning. For example, as the difference in height increases so does the likelihood of a fighter winning. Here is an example using the linear regression model for one variable. We can apply the following logistic regression model to our data. Let Y be a random variable for whether the fighter in the Red corner won.

$$Y = \beta_0 + \beta_1 \text{dif.Height}$$

where β_0 is fighters are the same height and β_1 is the effect on the likelihood of winning given one cm greater difference in height between the fighters.

```
height_model<- traindiff %>%
  glm(Winner == "Red" ~ diff_height,
      data = .,
      family = "binomial")
  height_model %>% tidy() %>% kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.5259269	0.0385816	13.63154	0.0000000
diff_height	0.0138036	0.0060544	2.27994	0.0226112

What does the p-value say about the difference in height? Provide a brief description below on whether a difference in height leads to a better chance of winning a UFC fight.

ANS

It appears that there is very strong evidence that a positive difference in height contributes to a better chance of winning because of the low p-value and the positive slope.

Now lets look at takedowns and strikes (qualities not controlled by genetics.)

$$Y = \beta_0 + \beta_1 \text{dif.avg.SIG.STR.pct} + \beta_2 \text{dif.avg.TD.pct} + \beta_3 \text{dif.avg.SIG.STR.landed} + \beta_4 \text{dif.avg.TD.landed}$$

```r
skill_model <- traindiff %>%
  glm(Winner == "Red" ~ diff_avg_SIG_STR_pct + diff_avg_TD_pct +
        diff_avg_SIG_STR_landed + diff_avg_opp_SIG_STR_landed +
        diff_avg_TD_landed,
      data = .,
      family = "binomial")
summary(skill_model)
```

```
##
## Call:
## glm(formula = Winner == "Red" ~ diff_avg_SIG_STR_pct + diff_avg_TD_pct +
##     diff_avg_SIG_STR_landed + diff_avg_opp_SIG_STR_landed + diff_avg_TD_landed,
##     family = "binomial", data = .)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  0.529305   0.039094  13.539  < 2e-16 ***
## diff_avg_SIG_STR_pct         0.025118   0.252671   0.099 0.920811
## diff_avg_TD_pct             -0.184424   0.137432  -1.342 0.179618
## diff_avg_SIG_STR_landed      0.010846   0.001862   5.824 5.76e-09 ***
## diff_avg_opp_SIG_STR_landed -0.012310   0.001930  -6.377 1.80e-10 ***
## diff_avg_TD_landed           0.092314   0.026688   3.459 0.000542 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3810.0  on 2885  degrees of freedom
## Residual deviance: 3740.7  on 2880  degrees of freedom
## AIC: 3752.7
##
## Number of Fisher Scoring iterations: 4
```

Does anything stand out to you? Take a closer look at the p-value for average significant strikes landed by an opponent. Does the significance of this variable surprise you?

Lets create a model for the genetic variables and the fighter’s current record at the time of the fight.

$$Y = \beta_0 + \beta_1 \text{dif.Age} + \beta_2 \text{dif.Reach} + \beta_3 \text{dif.Height} + \beta_4 \text{dif.Wins} + \beta_2 \text{dif.Losses}$$

```r
gen_model <- traindiff %>%
  glm(Winner == "Red" ~ diff_age + diff_Reach_cms + diff_height +
        diff_wins + diff_losses,
      data = .,
      family = "binomial")
summary(gen_model)
```

```
##
## Call:
## glm(formula = Winner == "Red" ~ diff_age + diff_Reach_cms + diff_height +
##     diff_wins + diff_losses, family = "binomial", data = .)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.565582   0.041772  13.540  < 2e-16 ***
## diff_age       -0.062987   0.008437  -7.466 8.29e-14 ***
## diff_Reach_cms  0.017641   0.006266   2.815  0.00488 **
## diff_height    -0.014084   0.008149  -1.728  0.08393 .
## diff_wins       0.037435   0.014067   2.661  0.00779 **
## diff_losses    -0.099215   0.020916  -4.744 2.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3810.0  on 2885  degrees of freedom
## Residual deviance: 3675.9  on 2880  degrees of freedom
## AIC: 3687.9
##
## Number of Fisher Scoring iterations: 4
```

These variable seem to carry much more significance than the previous variables. Why do you think that is? Finally lets complete out model with all 11 variables (combine 2 models above).

```r
final_model <- traindiff %>%
  glm(Winner == "Red" ~ diff_avg_SIG_STR_pct + diff_avg_TD_pct +
        diff_avg_SIG_STR_landed + diff_avg_opp_SIG_STR_landed +
        diff_avg_TD_landed + diff_age + diff_Reach_cms + diff_height +
        diff_wins + diff_losses,
      data = .,
      family = "binomial")
summary(final_model)
```

```
##
## Call:
## glm(formula = Winner == "Red" ~ diff_avg_SIG_STR_pct + diff_avg_TD_pct +
##     diff_avg_SIG_STR_landed + diff_avg_opp_SIG_STR_landed + diff_avg_TD_landed +
##     diff_age + diff_Reach_cms + diff_height + diff_wins + diff_losses,
##     family = "binomial", data = .)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.570490   0.042089  13.554  < 2e-16 ***
## diff_avg_SIG_STR_pct        -0.012420   0.257711  -0.048 0.961561
## diff_avg_TD_pct             -0.224615   0.140132  -1.603 0.108959
## diff_avg_SIG_STR_landed      0.007420   0.001928   3.849 0.000119 ***
## diff_avg_opp_SIG_STR_landed -0.008640   0.001998  -4.325 1.53e-05 ***
## diff_avg_TD_landed           0.084284   0.027447   3.071 0.002135 **
## diff_age                    -0.058609   0.008508  -6.888 5.65e-12 ***
## diff_Reach_cms               0.016111   0.006338   2.542 0.011019 *
## diff_height                 -0.011994   0.008234  -1.457 0.145245
## diff_wins                    0.025000   0.014324   1.745 0.080924 .
## diff_losses                 -0.073998   0.021453  -3.449 0.000562 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3810.0  on 2885  degrees of freedom
## Residual deviance: 3642.9  on 2875  degrees of freedom
## AIC: 3664.9
##
## Number of Fisher Scoring iterations: 4
```

Provide a brief summary of which variables are significant and why you think these variables are important. Look at both the p-value and the estimated effect on the model.

Now that we have generated a model from our data, lets try to utilize it to predict winners of fights that have not happened. Before we do this, it might be helpful to visualize what our models means. Lets take a look at the first model that we created which tried to determine the winner based only on height. By creating a list of several possible height differences, and utilizing our model to predict the odds of victory for these heights, we will get a look at what the prediction curve looks like. (activity modeled from https://www.theanalysisfactor.com/r-glm-plotting/)

It is important to note that the Y that results from this equation is not probability. It is log odds, which takes the natural logs of the odds-ratio. Normally we would have to do a short mathematical process to convert the log-odds into probabilities, but we can get around this by using the R predict() function, and setting the type as "response". This tells R to output the probability of event Y, after making its prediction on a fitted logistic regression model.

Lets make a list of height differences from -15 cm to 15 cm, with a step size of 0.1 cm.

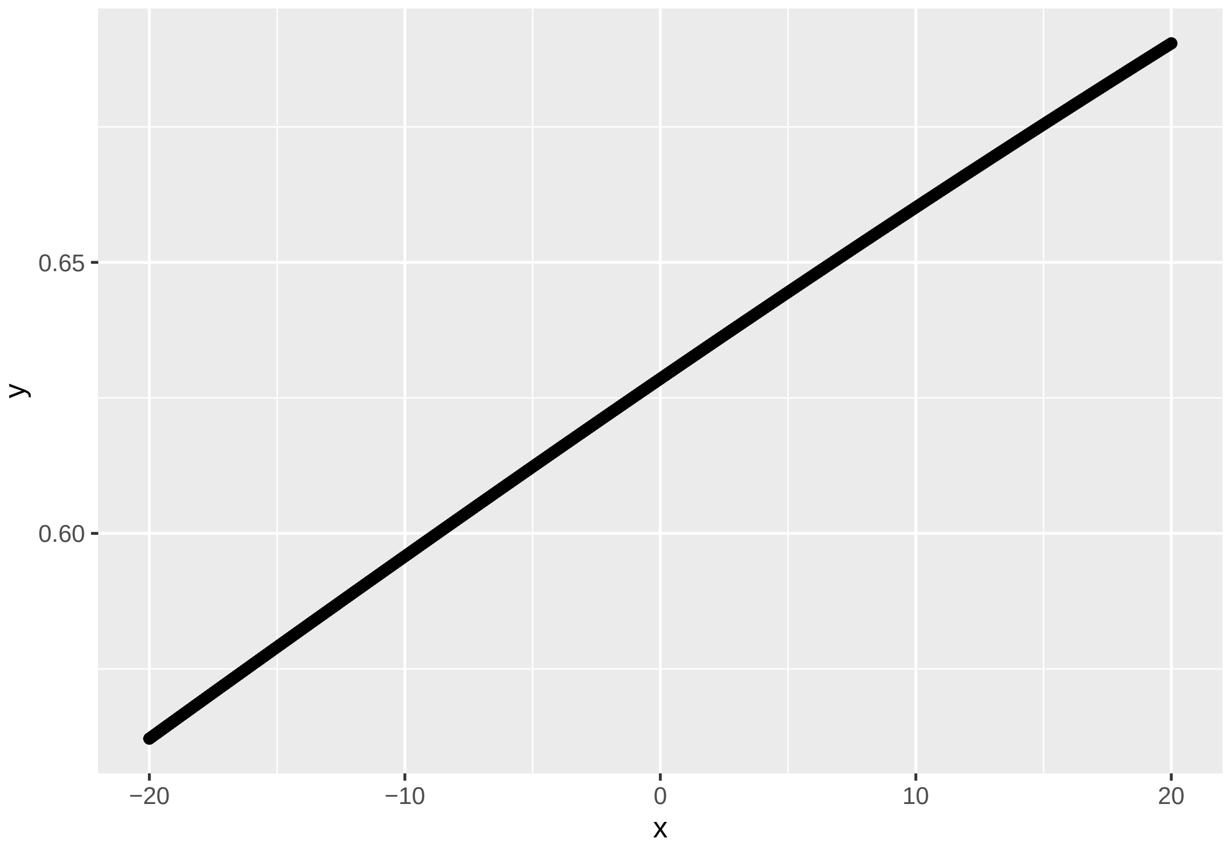
```
heights  <- seq(-20,20,0.1)
```

The predict function will use the intercept and slope that we found above to determine a win probability for each of these heights.

```
predheights <- predict(height_model, list(diff_height = heights),
                       type = "response")
```

Now we can plot the results.

```
heightsdata <- data.frame(x = heights, y = predheights)
ggplot(heightsdata, aes(x,y))+
  geom_point()
```



As expected, as the height difference goes up, so does the expected probability of a win. Furthermore, a height difference of 0 gives a probability of around 0.625 which makes sense as neither fighter has the height advantage, and the red corner is typically given to the more popular fighter or the defending champion.

Okay, now let's try to complete the same process for the model that we created that looks at all the factors. We will create a data frame to simulate a mismatch, in which one fighter will have a big advantage in every category that we are looking at. Let's see how it goes.

```
mismatchdata <- data.frame(diff_avg_SIG_STR_pct = 0.2, diff_avg_TD_pct = 0.2,
                           diff_avg_SIG_STR_landed = 30,
                           diff_avg_opp_SIG_STR_landed = -30,
                           diff_avg_TD_landed = 2, diff_age = -7,
                           diff_Reach_cms = 6, diff_height = 6, diff_wins = 6,
                           diff_losses = -6)
```

Now lets predict!

```
mismatchprediction <- predict(final_model, newdata = mismatchdata,
                              type = "response")
mismatchprediction
```

```
##         1
## 0.9004635
```

Wow! It looks like our model predicts about an 88% chance for the fighter with all the advantages to win! Now lets go back and use the test data set to find how many times the model's predicted winner matched up with the actual winner. The code below will create a new data frame containing the red fighters win probability based on the chosen variables.

```
testing <- final_model %>%
  augment(type.predict = "response",
       newdata = testdiff)
```

Now, we will assign a 1 for every time the actual winner of the fight matched the fighter which our model gave a better chance of winning (>50%).

```
testing2 <- testing %>%
  mutate(PredWin = ifelse((.fitted >= .5), "Red", "Blue"))
testing2 <- testing2 %>%
  mutate(correct = ifelse(Winner == PredWin, 1, 0))
```

Now we can find a proportion of times that our model gave the advantage to the person who ended up winning.

```
score <- testing2 %>% pull(correct)
correct <- sum(score)
total <- length(score)
correct/total
```

```
## [1] 0.6392931
```

It seems like our model was right about 66% of the time. Not Bad!

To wrap up our module we will do some quick analysis on which of the models that we created were the most accurate. If you need a refresher, go back above and take a look at the four models that we trained based on different combinations of variables. What were these four models?

ANS:

height, genetic/record, skill, final/everything

The two metrics that we will use for this analysis are AICs and BICs. AICS (Akaike Information Criterion) and BICs (Bayesian Information Criterion) both measure how well the model fits the data and put this analysis into a single score. The difference between the two is the BICs more heavily penalizes complexity in the model since complexity can introduce more opportunities for error. A lower AICS and BICS score indicates a better fitting/ more accurate model.

```
AICs = c("AICheight" = AIC(height_model),"AICgenetics" =AIC(gen_model),
         "AICskills" = AIC(skill_model),"AICfinal" = AIC(final_model))

BICs = c("BICheight" = BIC(height_model),"BICgenetics" =BIC(gen_model),
         "BICskills" = BIC(skill_model),"BICfinal" = BIC(final_model))

AICs
```

```
##   AICheight AICgenetics   AICskills    AICfinal
##    3808.798    3687.930    3752.733    3664.875
```

```
BICs
```

```
##   BICheight BICgenetics   BICskills    BICfinal
##    3820.733    3723.736    3788.539    3730.519
```

Based on the results, which of our models were the most fitting?

ANS: Choose the models with the lowest scores for each.

6. Conclusion In examining fights throughout the history of the UFC, we sought to determine what factors makes a fighter more likely to win. We observed 11 different factors and found the most significant were the fighters current number of wins and losses, height, age, and the average number of significant strikes an opponent lands on them per fight. We found that you fighters who have a height advantage against their opponent are typically more successful. A fighters ability to avoid strikes also leads to a fighters likelihood of winning. Most importantly a fighters current record is the most relevant variable in determining which fighter will win a UFC fight. The more wins the better the chance and the less losses the better the chance. We found that the final model that included the most factors was the most accurate, and gave the advantage to the actual winner in the red corner about 66% of the time.

EXTRA:

Just for fun, lets look at one of the biggest upsets of all time, Matt Serra vs George St. Pierre. The money line for this fight was -1300 to +840, one of the largest differentials in UFC History. This fight occurred on April 7th 2007 at UFC 69. GSP was considered untouchable at this time, and Serra had a couple of bad losses on his record already. Looking at the table below can show just how better GSP had been in his previous fights.

```
#isolate fight
ma <- Final_df %>% filter(date == "2007-04-07") %>% filter(R_fighter == "Matt Serra")

#Put GSP in red corner because he was the favorite
ma [1] <- "GSP"
ma [2] <- "Matt Serra"

 #Switch differences
ma <- ma %>%  mutate(diff_avg_SIG_STR_pct = diff_avg_SIG_STR_pct *-1,
              diff_avg_TD_pct = diff_avg_TD_pct *-1,
              diff_avg_TD_landed = diff_avg_TD_landed*-1,
              diff_wins = diff_wins*-1,
              diff_avg_SIG_STR_landed = diff_avg_SIG_STR_landed*-1,
              diff_avg_opp_SIG_STR_landed = diff_avg_opp_SIG_STR_landed*-1,
              diff_age = diff_age*-1,
              diff_Reach_cms = diff_Reach_cms*-1,
              diff_height= diff_height*-1,
              diff_losses= diff_losses*-1,
              diff_fights = diff_fights*-1
              )
ma %>% kable()
```

R_fighter	B_fighter	Winner	date	diff_avg_SIG_STR_pct	diff_avg_TD_pct	diff_avg_TD_landed	diff_wins	diff_avg_SIG_STR_landed	diff_avg_opp_SIG_STR_landed	diff_age	diff_Reach_cms	diff_height	diff_losses	diff_fights
GSP	Matt Serra	Red	2007-04-07	0.04625	0.604375	0.2070312	2	26.61328	-20.65234	-7	20.32	12.7	-3	-1

```
#calculate prediction
prediction <- predict (final_model, newdata = ma, type = "response")
prediction
```

```
##         1
## 0.8435403
```

Our model gave George St.Pierre about a 85% chance of victory!!