UFC Fight Outcome Prediction using Logistic Regression

Angelo Desiderio

2024-03-09

Introduction

The UFC which is a mixed martial arts company stands for the Ultimate Fighting Championship, has a recording of almost every fight ever since it's inception in 1993, which includes variable data such as total number of kicks, number of take-downs, total strikes landed, etc. One particular variable that has been recorded is the fighter's stance that can fall under one of the following categories Orthodox, Southpaw, Open Stance, Sideways, Switch and a certain exception where a empty value was recorded. We wish to use this classification system as our predictor variable. Our goal is to model the relationship of the fighter's stance and the probability of winning the fight. Making the result variable our response variable. The records follow a binary system where a 0 represents the loss for the fighter and a 1 representing a win. And because of this system the implementation of a logistic regression model will be ideal.

Descriptive Statistics

Handling NA values

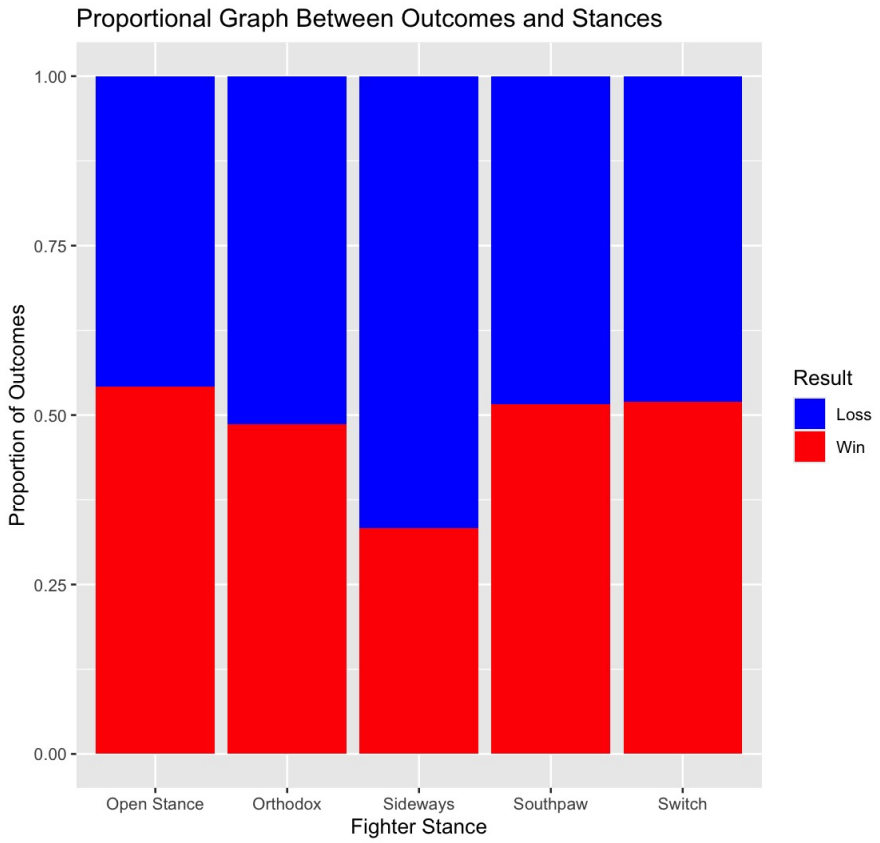
Before some changes were made a total of 96 fights contained an empty stance column, meaning a fighting stance was not recorded for that particular fighter. The 96 fights accounted for less than half a percent of the data set. The dates of these fights were randomly scattered throughout the data set without a consistent pattern being present. This maybe due in part of human error on the UFC's part or perhaps the fighter declined to input a preferred stance, but the true reasoning is unknown. The way in which this class of data will be handled, is by treating it as typical NA values and exclude it from the final regression model as there is a sufficient data size to still work with. Alternatively we could have implemented an imputation method. But roughly 75% of all time UFC fights will most likely have one fighter using Orthodox Stance. Meaning approximately 72 out of 96 of the fighters would have had a Orthodox Stance.

Open Stance	Orthodox	Sideways	Southpaw	Switch
24	10007	6	2637	552

After the exclusion of the non-recorded fights and associating the existing fighting stances with the binary outputs, the following table of data is produced. This is the data that the logistic regression model will be working with. At a quick glance most of the fighting stances are evened out in terms of wins and losses.

	Loss	Win
Open Stance	11	13
Orthodox	5142	4865
Sideways	4	2
Southpaw	1275	1362
Switch	265	287

Visualization:



After we have analyzed, cleaned and visualized the data we can begin building the logistic regression model

Regression Model:

```
Coefficients:
             Estimate Std.     Error  z-value  Pr(>|z|)
      (Intercept) 0.1671     0.4097   0.408    0.683

       Orthodox  -0.2224     0.4102  -0.542    0.588

       Sideways  -0.8602     0.9577  -0.898    0.369

       Southpaw  -0.1011     0.4115  -0.246    0.806

       Switch    -0.0873     0.4184  -0.209    0.835

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18333  on 13225  degrees of freedom
Residual deviance: 18323  on 13221  degrees of freedom

AIC: 18333
Number of Fisher Scoring iterations: 3
```

Results: Each coefficient (e.g., -0.2224 for Orthodox, -0.8602 for Sideways, etc.) in the logistic regression output corresponds to the change in the log odds of winning associated with a one-unit change in the respective predictor variable, while holding all other variables constant. For instance, if we take the coefficient for Orthodox as an example (-0.2224), it suggests that holding other factors constant, being in the Orthodox stance category is associated with a decrease of 0.2224 in the log odds of winning compared to the reference category (which is not explicitly stated here). Additionally the coefficients for the other stance categories provide information about their respective effects on the log odds of winning.

After the implementation of the regression model we are shown the results of the individual predictor variable, as we can see none of them returned with a p-value small enough for it to be considered statistically significant. In the context specified earlier, the log-odds would represent the linear relationship between the stance categories (Orthodox, Sideways, Southpaw, Switch) and the log odds of winning in a logistic regression model.

Null Deviance: Before adding any predictor variables, there was a certain amount of unpredictability in fight outcomes, represented by the null deviance. Residual Deviance: After adding predictor variables, the model was able to explain some of the unpredictability in fight outcomes, but there's still some unexplained variability left, as indicated by the residual deviance.

AIC: This model's AIC value is 18333, which is a measure of model quality. Lower AIC values indicate a better fit, so this model doesn't seem to fit the data particularly well.

Model Assessment

```
Analysis of Deviance Table

 Model 1: result ~ stance

 Model 2: result ~ 1

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1     13221      18323

2     13225      18333 -4  -10.129  0.03831 *

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

Because we are using a logistic model, we would not make use of a T-test or F-test, instead we can compare two models the fitted with the the reduced.

Improvement in Model Fit: The p-value associated with the difference in deviance between Model 1 and Model 2 is 0.03831, which is less than the conventional significance level of 0.05. This suggests that adding the predictor variable 'stance' significantly improves the fit of the model compared to the null model (intercept-only model). Therefore, the inclusion of the fighter's stance as a predictor variable provides valuable information for predicting the outcome variable 'result' (win or loss).

Effectiveness of the Model: The improvement in deviance (Deviance = -10.129) indicates that Model 1, which includes the predictor variable 'stance', provides a better fit to the data compared to Model 2. This suggests that the logistic regression model with the fighter's stance as a predictor has some predictive power in explaining the variability in the outcome variable 'result'.

Practical Significance: While the improvement in model fit is statistically significant, it's also important to consider the practical significance of the predictor variable 'stance'. Even though the inclusion of 'stance' improves the model fit, the magnitude of its effect (as indicated by the coefficient estimates) and its practical relevance should be assessed in the context of the specific research question or application.

Conclusion:

In this output, none of the coefficients for the stance variables are statistically significant at conventional levels ($p > 0.05$), indicating that there is insufficient evidence to conclude that fighter stance has a significant effect on the outcome of UFC matches in this model. To answer the initial question, the adoption of a certain fighting stance will not provide a fighter an advantage and according to these numbers seems to be more of matter of personal preference. Lastly, the analysis of deviance table provides evidence supporting the inclusion of the fighter's stance as a predictor variable in the logistic regression model for predicting outcome of UFC matches.

Citation:

UFC Data set: https://www.kaggle.com/datasets/danmcinerney/mma-differentials-and-elo/data