ISYE 6420: Project

Spring 2024

Syed Muhammad Husainie

Scope

Herein, a Bayesian logistic regression model was implemented for Ultimate Fighting Championship (UFC), a North American Mixed Martial Arts (MMA) organization. For this, fighter data (boxing, wrestling, and submission prowess) for 29 random UFC fighters were analyzed. The resulting model aimed to predict a new UFC fighter's potential in addition to predicting upcoming fights.

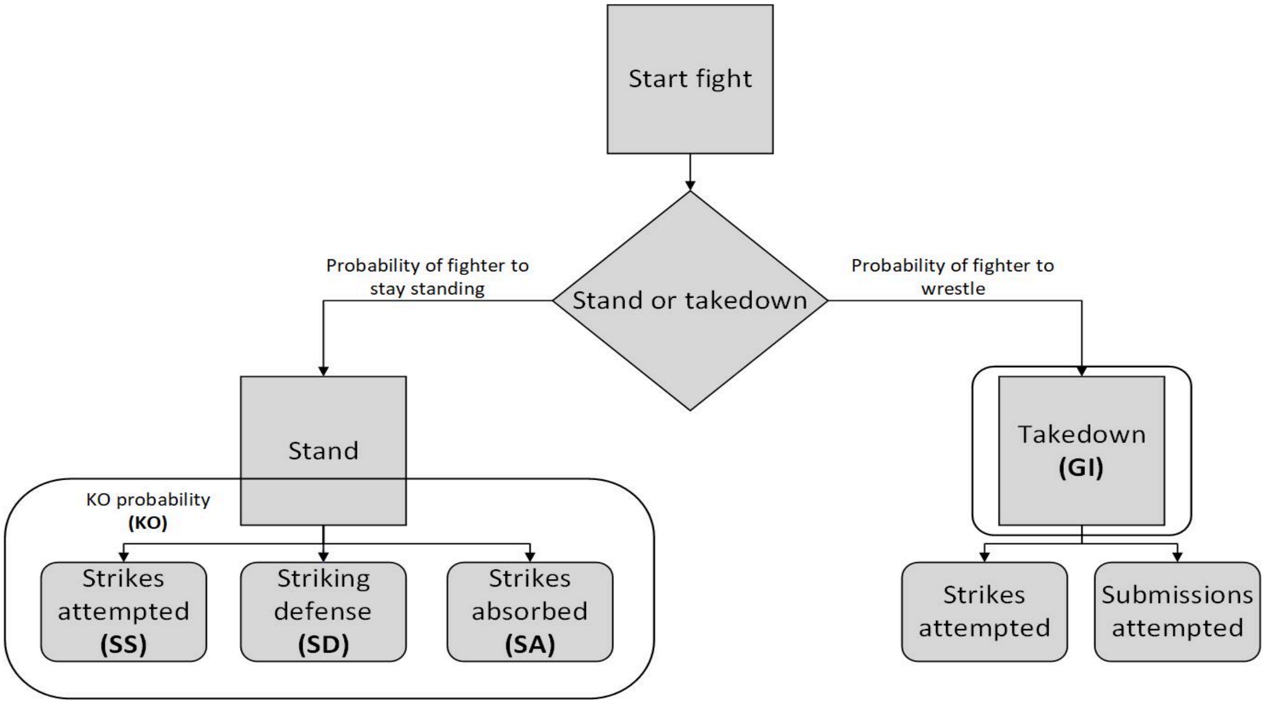
Model Preparation

To prepare variables for regression, each fighters boxing, punching power, wrestling, and submission skills were recorded. This data was retrieved from UFC's website which includes statistics for each fighter during their entire MMA career.

The following variables were selected in order to run logistic regression:

Variable	Description
SA	Strikes absorbed
SD	Striking defense
SS	Standing strikes
KO	Knockout/ technical knockout average, punching power
GI	Wrestling damage/ impact (Ground Impact)

The first three variables were extracted from the UFC stats website. To record 'SS' and 'GI', 10 fights for each of the fighter were analyzed to see their standing versus ground strikes, and the strikes proportion was averaged. The decision tree below highlights how these variables interplay:



Where,

SS = Proportion of standing strikes × Strikes attempted per minute

GI = (Takedowns landed per minute × Proportion of ground strikes × Strikes attempted per minute) + (Takedowns landed per minute × Average attempted submissions)

For the purpose of this model, each fighter's number of fights won was normalized from a total of 26 fights. Below is a summarized list of the data modeled:

Fighters	SA	SD	KO	SS	GI	Y (Number of observations)
Khabib Nurmagomedov	1.75	0.65	0.28	1.77	1.11	26
Conor Mcgregor	4.66	0.54	0.68	3.55	0.03	20
Dustin Poirer	4.36	0.53	0.39	3.67	0.16	21
Israel Adesanya	3.11	0.56	0.59	3.99	0.00	23
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Fighters	SA	SD	KO	SS	GI	Y (Number of observations)
Sean O Malley	3.52	0.62	0.63	4.01	0.01	25
Georges St Pierre	1.4	0.72	0.29	2.92	0.63	24
Tony Ferguson	4.41	0.55	0.34	3.85	0.03	19

Background on Logistic Regression

To carry out logistic regression, a Binomial event was used with 26 observations (total number of fights). The response variable, y, is given as:

$$\boldsymbol{y = 1}, \text{ for win}$$

$$\boldsymbol{y = 0}, \text{ for loss}$$

The odds of a win are given by:

$$log(\frac{p(x^{var})}{1 - p(x)^{var}}) = \alpha + \beta\ x_{SA} + \beta_1 x_{SD} + \beta_2 x_{KO} + \beta_3 x_{SS} + \beta_4 x_{GI}$$

The probability of success, p_{var}, for logistic regression is defined by the logistic function:

$$\text{logistic}(x_{\text{SA}}, x_{\text{SD}}, x_{\text{KO}}, x_{\text{SS}}, x_{\text{GI}}) = p(x_{var}) =$$

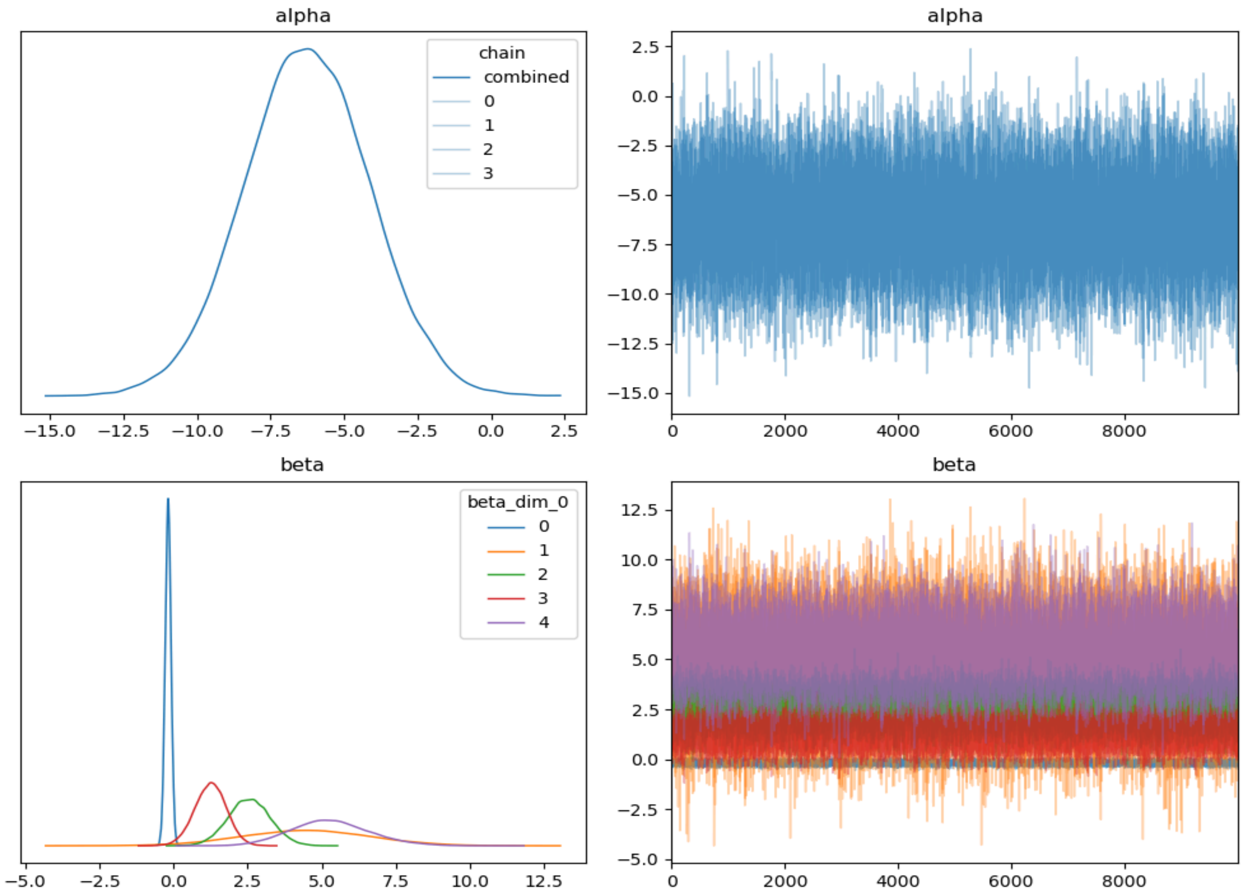
$$\frac{1}{1 + e^{-(\alpha + \beta_0 x_{SA} + \beta_1 x_{SD} + \beta_2 x_{KO} + \beta_3 x_{SS} + \beta_4 x_{GI})}}$$

Where, α is the intercept and β_0, β_1, β_2, β_3, β_4 are the coefficients of the SA, SD, KO, SS, GI variables, respectively.

Bayesian Approach

Non-informative normal priors were used for the βs as each variable may have a positive or negative correlation to the fighter's winrate.

4 chains were run with 10,000 iterations each and 1,000 tuned:



The model results showed the intercept coefficient had a mean value of -6.3 with a Confidence Interval (CI) of [-10,-2.1]. This seems accurate as it gives a p value, probability, of 0 (keeping all variables as 0), indicating that if a fighter came in with no skills whatsoever, they would have 0 probability of winning the fight.

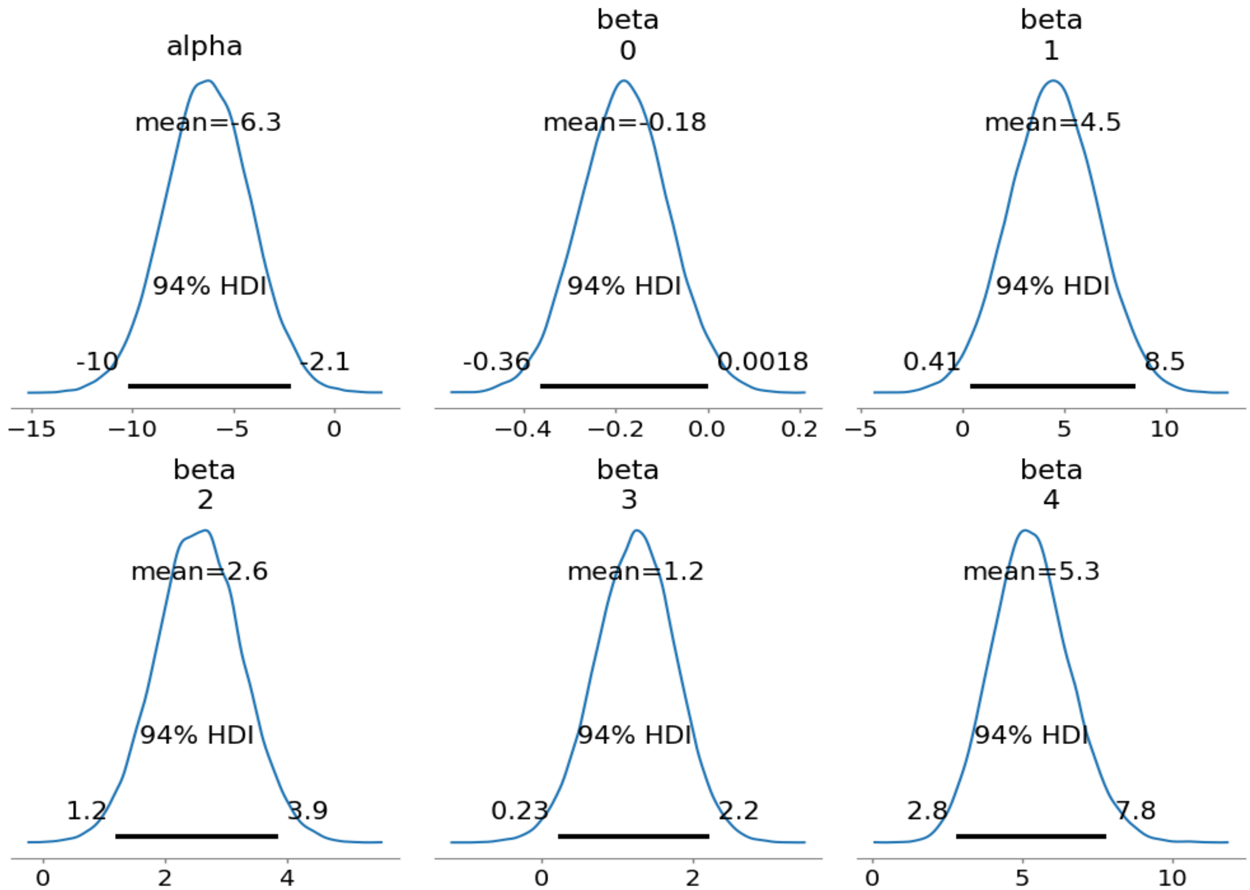
Stikes Absorbed, SA, coefficient had a mean value of 0.18 and CI of [-0.36,0.0018] indicating that strikes absorbed has little to no impact on the probability of the fight.

Strikes Defended, SD, mean coefficient was 4.5 with CI of [0.41,8.5]. The results indicate that there is a positive correlation between a fighter's stiking defense and their probability of winning a fight. However, the CI is quite large showing low reliability on that data alone.

Knockout Power, KO, with a mean of 2.6 shows a strong correlation between a fighter's punching/ knockout power and their win %. The smaller CI of [1.2,3.9] adds further strength to this argument, and, real life performances of Conor McGregor, Sean O Malley, and the likes of such fighters leaves little to no doubt that punching power is a major factor in winning fights.

Similar to knockout power, strikes while standing, SS, with coefficient mean 1.2, and CI [0.23,2.2] also has a strong correlation with fight win %.

Lastly, ground impact, which is a combination of a fighter's wrestling skills, ground strikes, and submission skills has a cofficient mean 5.3 with CI [2.8,7.8]. This shows a very strong relationship of high GI vs. win %. However, the wide confidence interval also indicates non-informity, and this is likely due to the fact that wrestling intrinsically is a very difficult sport to master and there have been very few fighters who have been able to exploit the benefits of wrestling to a high level in MMA.



To further analyze the effectiveness of each variable, the model was rerun five more times while removing one of the five variables in each. The 'goodness' of the model fit and significance of each factor was measured using the R^2 values obtained:

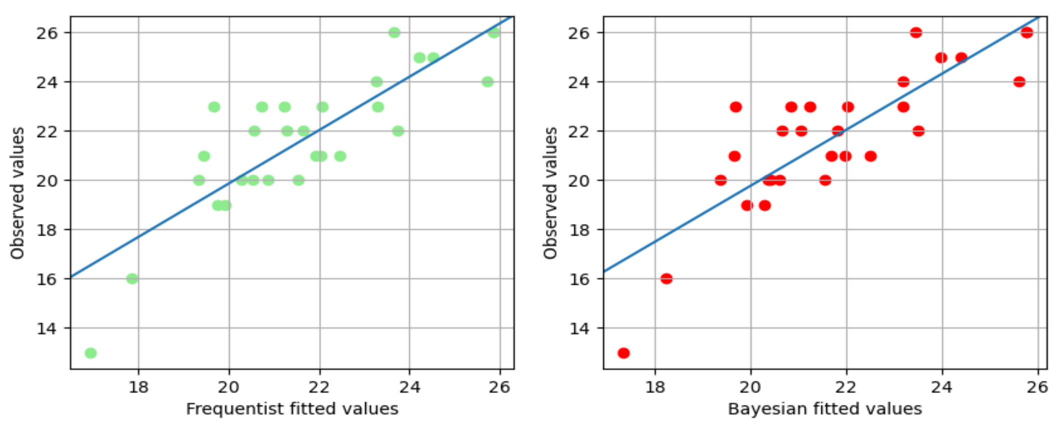
Model Variables	$\mathbf{R^2}$
SA, SD, SS, KO, GI	0.64
Removing Strikes Absorbed	0.57
Removing Strike Defense	0.57
Removing Standing Strikes	0.52
Removing KO power	0.42
Removing Ground Impact	0.35

These results indicate that it is imparative to know the fighters punching power and ground impact in order for the model to predict a fighter's outcome effectively.

Frequentist Approach

To compare the Bayesian regression model to a frequentist approach, python's GLM method was used. The results indicate a similar model fit using both methods. Althought the coefficient mean values were similar, the CI was much more wide compared to the Bayesian approach. This is as expected as Bayesian approach provides more reliability when working with small datasets.

Model Coefficient	Mean	CI
α	-7.40	[-30.36, 15.55]
β_0	-0.17	[-1.163, 0.81]
β_1	5.56	[-18.61, 29.735]
β_2	2.64	[-4.63, 9.91]
β_3	1.36	[-4.13, 6.85]
β_4	5.51	[-8.99, 20.012]



Model Sensitivity and Specificity

To test the accuracy of the Bayesian model, each fighter's true positive and negative values were recorded and the sensitivity and specificity

Model sensitivity is given by:

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} = \frac{21.35}{21.35 + 0.38} = 0.98$$

Model specificity is given by:

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives} = \frac{3.42}{3.42 + 0.86} = 0.80$$

Predicting the Success of Other Fighters

Win amounts for six random fighters, outside of the fighters used for the model, were predicted to measure their success after 26 fights. The model worked fairly well, however, it leaned towards overestimating fighter wins. This doesn't seem unsurprising though, as the first three fighters used for this prediction had very long careers resulting in deminishing performance past their 'prime'. On the other hand, there were also other factors such as starting MMA at a very young age and improving tremendously over the years. This is particularly true for fighters like Charles Oliveira, Arman Tsarykyan, and Merab Dvalishvili:

Fighter	Y Predicted	Y Actual	False Positives	False Negatives
Nick Diaz	23	19	4	0
Tony Ferguson	20	19	1	0
Josh Thompson	23	19	4	0
Charles Oliveira	24	21	3	0
Sean Brady	24	25	0	1
Arman Tsarukyan	17	23	0	6
Merab Dvalishvili	25	22	3	0

Predicting Fight Outcomes

What good is a model that cannot predict fights? To make prediction possible, the impact of each fighter within a fight was measured against their respective opponent by taking ratio of the response variable y, for the difference in skills between fighter A and B and dividing by the total response for difference in skill with respect to fighter A and fighter B:

$$Impact\ of\ fighter\ A\ in\ fight\ with\ fighter\ B = \frac{Predicted\ response\ (X_{varA} - X_{varB})}{Predicted\ response\ (X_{varA} - X_{varB}) + Predicted\ response\ (X_{varB} - X_{varA})}$$

Fighter A vs. Fighter B	Fighter A Impact	Fighter B Impact	Likely Winner	Actual Winner
Petr Yan vs. Aljamin	60.0%	40.0%	Aljamin Sterling	Aljamin Sterling

Fighter A vs. Fighter B	Fighter A Impact	Fighter B Impact	Likely Winner	Actual Winner
Sterling				
Conor Mcgregor vs. Max Holloway	71.0%	29.0%	Conor Mcgregor	Conor Mcgregor
Khabib Nurmagomedov vs. Justin Gaethji	99.5%	0.5%	Khabib Nurmagomedov	Khabib Nurmagomedov
Islam Mackhachev vs. Alexander Volkanovski	81.5%	18.5%	Islam Mackhachev	Islam Mackhachev (2/2)
Kamaru Usman vs. Leon Edwards	62.5%	37.5%	Kamaru Usman	Leon Edwards (2/3)
Robert Whittaker vs. Dricus Du Plessis	21.6%	78.4%	Dricus Du Plessis	Dricus Du Plessis
Khabib Nurmagomedov vs. Georges St. Pierre	83.0%	17.0%	Khabib Nurmagomedov	Never took place
Jon Jones vs. Georges St. Pierre	5.0%	95.0%	Georges St. Pierre	Never took place

Conclusions

This project aimed to use Bayesian Logistic Regression to predict an MMA fighters potential and predict fights. This was done successfully with some deviation in the results.

To improve the prediction power of this model, it is recommended to make individual models for each weight class of fighters and to limit the age group of the regressed fighter data. Age and weight class makes a significant difference in power, speed, and stamina of the fighter and the resulting models will likely be more accurate predictors of fighter outcomes.

Reference

http://www.ufcstats.com/statistics/events/completed