PREDICTING OUTCOMES OF MIXED MARTIAL ARTS FIGHTS WITH NOVEL FIGHT VARIABLES

by

JEREMIAH DOUGLAS JOHNSON

(Under the Direction of Daniel B. Hall)

ABSTRACT

In this study, I attempt to forecast the win/loss outcomes of mixed martial arts bouts with fight data. Both basic 'count' variables and newer, constructed variables are considered. These novel measures are then used to predict wins and losses using a logistic regression model, and this model is compared to baseline models in terms of predictive ability. The final model contains both 'count' variables and constructed variables and is found to have significantly greater predictive ability than baseline models. Cross-validation and discriminant analysis confirm these results.

INDEX WORDS: Sports Statistics, MMA, Logistic Regression, Sports Forecasts

# PREDICTING OUTCOMES OF MIXED MARTIAL ARTS FIGHTS WITH NOVEL FIGHT VARIABLES

by

JEREMIAH DOUGLAS JOHNSON

Bachelor's of Business Administration, University of Georgia, 2009

A Thesis Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2012

# PREDICTING OUTCOMES OF MIXED MARTIAL ARTS FIGHTS WITH NOVEL FIGHT VARIABLES

by

JEREMIAH DOUGLAS JOHNSON

Major Professor: Daniel B. Hall

Committee: Lynne Seymour
Jaxk Reeves

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2012

TABLE OF CONTENTS

Page

LIST OF TABLES ..... 2

1 INTRODUCTION ..... 4

2 BACKGROUND ..... 5

i) MMA BACKGROUND ..... 5

ii) STATISTICAL BACKGROUND ..... 6

3 DATA SUMMARY ..... 9

4 METHODS ..... 13

5 RESULTS ..... 16

6 LIMITATIONS ..... 29

7 CONCLUSIONS ..... 31

REFERENCES ..... 33

# LIST OF TABLES

Page

Table 1: First level ‘count’ variables ..... 10

Table 2: Second level created variables ..... 10

Table 3: Second level created variable formulas ..... 11

Table 4: Summary Statistics ..... 16

Table 5: First level variable correlations ..... 17

Table 6: Second level variable correlations ..... 18

Table 7: Second level variable univariate models ..... 19

Table 8: Second level variable full model ..... 20

Table 9: Second level variable full model diagnostics ..... 20

Table 10: Second level variable reduced model ..... 21

Table 11: Second level variable reduced model diagnostics ..... 21

Table 12: Winning percentage model ..... 23

Table 13: Winning percentage model diagnostics ..... 23

Table 14: First level variable univariate models ..... 24

Table 15: First level variable full model ..... 24

Table 16: First level variable full model diagnostics ..... 24

Table 17: First level variable final model ..... 25

Table 18: First level variable final model diagnostics ..... 25

Table 19: Combined final model ..... 26

Table 20: Combined final model diagnostics ..... 26

Table 21: Discriminant Analysis..... 27

Table 22: Cross Validation..... 28

Table 23: Cross Validation diagnostics..... 28

# 1. INTRODUCTION

In recent years, the use of statistical analysis of sports data has become more mainstream than ever before. What used to be an extremely niche field is now a major area of research among both academics and professionals in the sporting world.

The sport of mixed martial arts, however, has not seen a high level of statistical analysis. Mixed martial arts (or MMA) is a relatively young sport. The first Ultimate Fighting Championship (or UFC), which marks the beginning of high-level MMA as a sport in North America was held in 1993, less than 20 years ago. In that time, the sport has evolved from a fringe, no-rules spectacle into a regulated mainstream sport with a dedicated fan-base, and the UFC (now MMA's largest league/organization) has become a billion dollar company. However, despite MMA's great advances as a sport there has been very little detailed statistical analysis of the sport in the same way that major sports such as baseball, basketball and football have been analyzed. In fact, there has been precious little statistical analysis of the sport at any depth. The use of any type of statistics to analyze matchups and predict outcomes is not widespread, even among dedicated fans.

This paper aims to develop novel statistics which will be utilized to create a model for predicting win/loss outcomes in MMA fights. These models based on novel statistics will be compared against other, more basic models for predicting outcomes of fights, such as models based on winning percentage.

# 2. BACKGROUND

## i) MMA BACKGROUND

In order to more accurately follow the arguments presented in this paper, it will be necessary to have at least a baseline understanding of MMA.

Mixed martial arts originated as a test of "Which martial arts are the most effective?" Initially, there were few rules and no time limits or rounds, and combatants were typically representatives of a single style of martial art (e.g. judo, boxing, karate, jiu-jitsu, etc.). Today, MMA is regulated under a unified rule set in 46 of the 48 US states that have athletic commissions, and is similar in structure to boxing. Time and round limits are nearly universal and certain fouls are prohibited such as head-butting, kicking a downed opponent, etc.

MMA fights consist of a certain number of rounds, just as in boxing. Most fights last either 3 or 5 rounds, with 5 minute rounds as the standard round length. During each round, fighters attempt to outfight their opponent using the three main areas of MMA – striking, wrestling, and grappling. Mixed martial arts fights resemble a mixture of kickboxing, amateur wrestling, and submission fighting. While standing, fighters attempt to effectively strike one another in a manner similar to a kickboxing bout. Common techniques include punches, kicks, knees and elbows, and fighters utilize techniques from boxing, kickboxing, Muay Thai and karate among many martial arts. Fighters are also permitted to utilize takedowns to bring the fight to the ground, and many fighters use techniques from freestyle wrestling, Greco-Roman wrestling and judo to take their opponents down to the mat. Once on the mat, fighters are free to strike their opponents

or to attempt submissions. Submission techniques include armlocks, leglocks and chokeholds which are designed to make the opponent submit or 'tap out', and techniques from Brazilian jiu jitsu and catch wrestling are common.

There are three common ways to win an MMA fight. First, one can win via knockout or technical knockout. A KO/TKO occurs when the referee deems one fighter is unable to continue due to strikes. Second, a fight can end via submission, when a fighter taps out from being caught in a submission hold, conceding the fight. Referees may also end the fight by way of submission if a fighter visibly breaks a limb caught in a submission hold or loses consciousness due to a choke, which is called a technical submission. Finally, if neither fighter is able to knockout or force submission of their opponent over the duration of a fight, the fight will be determined by judges (as in boxing) who will render a decision to determine the winner of the fight. According to the Unified Rules of Mixed Martial Arts, judges will determine the winner of a bout based on clean strikes, effective grappling, Octagon control, and effective aggressiveness.

## ii) STATISTICAL BACKGROUND

The central goal of this paper is to assess the ability of certain novel variables created from recorded fight data to forecast win/loss fight outcomes. In order to assess the quality of such a model, a baseline level of comparison is necessary. The literature concerning statistical analysis of sports is essentially entirely devoid of any mention of the sport of MMA, but similar scenarios for predicting outcomes can be found in other sports. In particular, there have been several papers that attempt to model the win/loss

outcomes of the Division 1 NCAA men's basketball tournament games, among other sports.

Boulier and Stekler (1999) provide an example of a paper which attempts to forecast win/loss outcomes. Boulier and Stekler are primarily interested in whether or not seedings, or rankings, in various sports tournaments (specifically, tennis tournaments and the NCAA basketball tournament) are accurate predictors of win/loss outcomes. In order to examine this question, they first examine how often the higher seeded participant defeated the lower seeded participant. After observing results from this basic step, they next build a probit model which attempts to predict the win/loss outcome using difference in seeding/ranking as an explanatory variable. Of note is that they describe their probit model as essentially similar to a logit model, saying that "Results based upon the logit model are inconsequentially different from those based on the probit." Boulier and Stekler build their model, and observing the sign and significance of the coefficient of the explanatory variable 'difference in seeding' conclude that the probability of winning does indeed depend on the difference in seeding.

Next, Boulier and Stekler observe that they have built a model which is estimated from the data for the entire time period they measured (1985-1995). Not wishing to predict past events based on future data, they took the step of creating models which predict win/loss outcomes based cumulative totals: that is, on only those matches/games which have already been played thus far. With this method, they undertake the same modeling steps as before and conclude that the models are still significantly predictive. Boulier and Stekler also use the Brier score, a score function which assesses the accuracy

of probability forecasts, as a measure of model comparison. The Brier score will be discussed further in depth in the Results section.

Stekler and Klein (2012) provide another good example for how to examine such a model. Their model attempts to predict win/loss outcomes in the NCAA tournament via a created 'consensus rankings' which encompass several human and computer polls. In order to examine the efficacy of their model, which was created in a similar fashion to Boulier and Stekler, Stekler and Klein compare their results to a baseline model which predicts wins and losses based solely on a team's seed in the tournament. Stekler and Klein examine whether or not their more in-depth model involving consensus rankings can outperform the simple model using only seeding.

There is no perfect analog to seeding when considering MMA bouts. However, a common-sense approach in the spirit of creating a simple, obvious baseline model is to examine winning percentage. It seems intuitive that fighters with high winning percentages compared to their opponents will win more often than fighters with low winning percentages compared to their opponents, and that the difference in winning percentage between two combatants might be a good predictor of who will win that fight. This baseline model will be examined in more depth in the Results section.

# 3. DATA SUMMARY

The data set for this project comes from FightMetric LLC. FightMetric is the official statistics provider to the UFC, and has a database of 3638 MMA fights between 1868 different fighters for which they have recorded statistics. Broadly speaking, FightMetric has collected data from nearly all high level MMA promotions and/or events beginning with the sport's unofficial inception in 1993 at the first UFC event. While the criterion for what constitutes a 'high level' promotion or event is unavoidably grey, FightMetric is an industry leader and has a very extensive fight database. FightMetric has catalogued 565 events over the last 19 years, and these events include every UFC event, as well as every event from other high-level promoters such as PRIDE FC, World Extreme Cagefighting, King of the Cage, StrikeForce, etc. and other events deemed to be significant.

For each fight, FightMetric measured dozens of different categories, such as length of the fight, strikes thrown and landed for both fighters, takedowns attempted and landed for both fighters, submission attempts for both fighters, win/loss, etc. The collected data are grouped in many ways, but the three most important distinctions are data for a fighter, data for a fight, and data for a fighter within a fight. The particular variables of interest in this study are listed in Table 1.

Table 1: First level ‘count’ variables

| | |
|---|---|
| Striking Variables | Total Strikes Landed, Knockdowns |
| Wrestling Variables | Takedowns Attempted, Takedowns Landed |
| Grappling Variables | Advance to HalfGuard, Advance to Side, Advance to Mount, Advance to Back, Sweeps/Reversals, Submission Attempts |
| Fight Variables | Fight Outcome, Method of Victory, Total Time, Event Date |

These variables can mostly be described as ‘count values’, which is to say that they simply count the number of instances a certain event occurs. Takedowns Attempted counts the number of takedowns attempted by a fighter in a given fight, while Sweeps/Reversals counts the number of sweeps or reversals from the bottom position by a fighter in a given fight, and so forth. These variables can be computed at the fighter, fight, and fighter within a fight level.

Using these count variables, we will create a second level of more descriptive variables which describe fighter behavior on a deeper level. Similarly to the count variables, these new descriptive variables can be computed at the fighter, fight, and fighter within a fight level.

Table 2: Second level created variables

| | |
|---|---|
| Striking Variables | Striking Ratio, Strike Differential per Minute, Power Rating |
| Wrestling Variables | Total Takedown Percentage |
| Grappling Variables | Ground Activity per Takedown |

These measures are computed from the basic count variables given in Table 1, and are meant to convey a fighter's success or failure in the different main areas of MMA discussed in the Background section.

Table 3: Second level created variable formulas

| | |
|---|---|
| Striking Ratio | (Total Strikes Landed)/(Opponent's Total Strikes Landed) |
| Strike Differential per Minute | (Total Strikes Landed – Opponent's Total Strikes Landed)/(Minutes Fought) |
| Power Rating | (Knockdowns + Knockouts/Technical Knockouts)/(Total Strikes Landed) |
| Total Takedown Percentage | (Takedowns Landed + (Opponent's Takedowns Attempted – Opponent's Takedowns Landed))/(Takedowns Attempted +Opponent's Takedowns Attempted) |
| Ground Activity per Takedown | (Advance to HalfGuard + Advance to Side + Advance to Mount + Advance to Back + Sweeps/Reversals + Submission Attempts)/(Takedowns Landed + Opponent's Takedowns Landed) |

These second-level variables in Table 3 attempt to measure different aspects of the effectiveness of a fighter within a fight. Striking Ratio and Strike Differential per Minute measure 'volume striking', or which fighter is landing more blows. Power Rating measures 'power striking', or how likely a fighter is to knock down or knock out his opponent with a strike. Total Takedown Percentage measures overall wrestling/takedown ability. Ground Activity per Takedown measures grappling activity level and skill. Using these constructed variables corresponding to various areas of MMA, we will attempt to create a predictive model for win/loss outcomes in fights.

In order to create relevant models, significant data manipulation is required. The most basic grouping of data in the original dataset is on the fighter within a fight level. Using these data to regress against wins would not provide us with any valuable information; it seems rather obvious that the fighter who lands more strikes in a fight, for instance, would be very likely to win that particular fight. But it isn't helpful at all in predicting fights before they happen, because we don't know which fighter will land more strikes in this fight before the fight starts. Career totals are likewise ineffective, because using career totals would be predicting past events based on future information (which was not available at the time of the fight in question).

Instead, the fight-level data were transformed into cumulative totals which represent all the fights in a fighter's career up to the point of his current fight. This approach uses the maximum amount of data available at the time of the fight to predict that fight. In addition, the variables mentioned will all be differences in these cumulative totals between fighters. It is intuitively important to compare the data for both fighters participating in a fight, using their differences rather than simply using one fighter's cumulative totals. Therefore, each variable used in our analysis will be a difference between fighters in a cumulative version of that variable. For the purposes of succinctness, variables will retain their simpler names, i.e. Striking Ratio rather than Cumulative Difference in Striking Ratio.

Creating cumulative variables in this manner reduces the size of the data set, because every fighter will have no accumulated fight data for their first fight. However,

the sample size will be in the thousands regardless and should remain more than sufficient for our purposes.

# 4. METHODS

The goal of this paper is to forecast win/loss outcomes in MMA fights, and to examine the efficacy of these new, created variables in predicting those outcomes. In accordance with that goal, several natural hypotheses present themselves.

1. The model will be significantly better than random chance in predicting outcomes for fights.
2. The model will be significantly better than a simple model based on winning percentage.
3. The model will be significantly better than a model based on the 'first level' variables.

In order to forecast wins and losses using our new variables, we will use logistic regression. The binary predicted variable will be the win/loss outcome of each fight. It should be noted that some fights are ruled as draws or no-contests, but these cases make up only around 1% of the data and have been removed from the data set. The explanatory variables will be the five second-level statistics mentioned in the data summary. In addition, the natural log of striking ratio will be examined in addition to striking ratio.[1]

[1] Striking Ratio has non-linear characteristics, considering that an additional strike can cause either very small changes or large changes to Striking Ratio. A single strike changes a 24 to 3 result (SR=8) to 24 to 4 (SR=6), while in a different fight, a single strike would only change a 14 to 13 result (SR=1.08) to 14 to 14 (SR=1). Using the natural log of striking ratio provides a way to reduce this effect. This also has the added benefit of providing fight level data points which mirror one another, as two opponents in a fight will always have reciprocal Striking Ratios, and the sum of the natural log of two reciprocal numbers is always zero.

The general model for binary logistic regression is

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

$$p_i = Probability\ of\ an\ event\ Y_i$$

$$\beta_0 = \text{ the value of } \log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right), \text{ or the log odds of the event (e.g., winning) when all the X's are 0.}$$

$$\beta_i = \text{ the expected change in the log odds of a success } (assuming\ all\ other\ X_j's\ are\ constant)$$

$$X_i = Predictor\ variable.$$

In order to evaluate the performance of the various logistic models presented in the Results, we will examine a number of relevant statistics. The first, most basic statistic is the percent concordant/disconcordant. Percent concordant/disconcordant simply measures how often the model accurately predicts an event, in our case the win/loss outcome of the fight. We will also examine the rescaled R-Squared of the model and the Brier score. Brier scores are a measure of model comparison which asses the forecasting ability of model. Brier scores measure the mean squared deviation between predicted probabilities of a set of events and the actual event outcomes, and are written as

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

Where $f_t$ is the forecasted probability, $o_t$ is the actual outcome (0 or 1), and N is the number of forecasting occurrences. A lower Brier score indicates higher model accuracy in predicting outcomes.

In addition to these diagnostic statistics, we will also use several statistical techniques to ensure that our final model for predicting fight outcomes is statistically sound. Specifically, cross validation will help to verify our model's usefulness, and discriminant analysis will be employed to support the model choice. Discriminant analysis attempts to classify data into categories based on independent variables. It is similar to logistic regression in some ways, but uses different methods and requires different assumptions. Some of the assumptions necessary for discriminant analysis limit the applicability of this technique, as will be discussed in the Limitations section, but it remains a somewhat useful tool for examining which independent variables may affect win/loss outcomes. Cross validation involves creating a model from one portion of your data, and then testing that model on a different section of your data to see if it still performs in the same manner.

Using these different statistics and techniques, we will create a model from secondary, created variables and compare this model against models from random chance, winning percentage and first level variables.

# 5. RESULTS

Table 4 presents summary statistics for our first-level statistics.

Table 4: Summary Statistics

| | Mean | Standard Deviation |
|---|---|---|
| Total Strikes Landed | 41.76 | 41.33 |
| Knockdowns | 0.20 | 0.48 |
| Takedowns Attempted | 2.42 | 3.41 |
| Takedowns Landed | 1.05 | 1.66 |
| Positional Advances[2] | 1.37 | 2.30 |
| Submission Attempts | 0.60 | 1.08 |
| Total Time | 535.58 (seconds) | 389.05 (seconds) |

Of more interest than simple averages is how correlated differences in cumulative first level statistics are, because the cumulative differences are what will ultimately comprise our models.

[2] Positional Advances = Advances to HalfGuard + Advances to Side Control + Advances to Mount + Advances to Back + Sweeps/Reversals. These variables are clumped together both by logical consequence of how they are used in MMA and for brevity's sake.

Table 5: First level variable correlations

| | Total Strikes Landed | Knock downs | Takedowns Landed | Positional Advances | Submission Attempts |
|---|---|---|---|---|---|
| Total Strikes Landed | 1.00000 | 0.47239<br><.0001 | 0.65747<br><.0001 | 0.68986<br><.0001 | 0.52468<br><.0001 |
| Knockdowns | 0.47239<br><.0001 | 1.00000 | 0.10838<br><.0001 | 0.19888<br><.0001 | 0.14978<br><.0001 |
| Takedowns Landed | 0.65747<br><.0001 | 0.10838<br><.0001 | 1.00000 | 0.75851<br><.0001 | 0.43321<br><.0001 |
| Positional Advances | 0.68986<br><.0001 | 0.19888<br><.0001 | 0.75851<br><.0001 | 1.00000 | 0.63246<br><.0001 |
| Submission Attempts | 0.05246<br><.0001 | 0.14978<br><.0001 | 0.43321<br><.0001 | 0.63246<br><.0001 | 1.00000 |

Many of our first level variables are highly positively correlated with one another. In fact, every single pairing has a significant correlation of some degree. It seems that fighters who perform well in one area of MMA tend to perform well in all areas of MMA. This is also largely true when we examine the correlations for the differences in the cumulative second level variables below in Table 6, with one exception.

Table 6: Second level variable correlations

| | Log of Striking Ratio | Striking Ratio | Strike Differential per Minute | Ground Activity per Takedown | Total Takedown Percentage | Power Rating |
|---|---|---|---|---|---|---|
| Log of Striking Ratio | 1.00000 | 0.70072<br><.0001 | 0.91221<br><.0001 | 0.13483<br><.0001 | 0.37750<br><.0001 | -0.22532<br><.0001 |
| Striking Ratio | 0.70072<br><.0001 | 1.00000 | 0.59634<br><.0001 | 0.17167<br><.0001 | 0.24044<br><.0001 | -0.05523<br>0.0016 |
| Strike Differential per Minute | 0.91221<br><.0001 | 0.59634<br><.0001 | 1.00000 | 0.09667<br><.0001 | 0.34280<br><.0001 | -0.19684<br><.0001 |
| Ground Activity per Takedown | 0.13483<br><.0001 | 0.17167<br><.0001 | 0.09667<br><.0001 | 1.00000 | 0.02625<br>0.1498 | -0.07148<br><.0001 |
| Total Takedown Percentage | 0.37750<br><.0001 | 0.24044<br><.0001 | 0.34280<br><.0001 | 0.02625<br>0.1498 | 1.00000 | -0.07717<br><.0001 |
| Power Rating | -0.22532<br><.0001 | -0.0552<br>0.0016 | -0.19684<br><.0001 | -0.07148<br><.0001 | -0.07717<br><.0001 | 1.00000 |

Our created, second level variables are almost all correlated to some degree with one another. Additionally, they are all positively correlated excepting Power Rating, indicating that a fighter who is has high values in one area is likely to have high values in many areas (excepting Power Rating, which shows the opposite).

Table 7 shows the univariate results for the six tested second-level variables when modeled against wins using logistic regression.

Table 7: Second level variable univariate models

| | Rescaled R-Square | % Concordant /Disconcordant | Associated p-value | Brier Score |
|---|---|---|---|---|
| Strike differential per minute | 0.0573 | 61.9/37.5 | <.0001 | 0.2389 |
| Striking Ratio | 0.0245 | 61.6/36.6 | <.0001 | 0.2445 |
| Log(Striking Ratio) | 0.0683 | 62.7/36.8 | <.0001 | 0.2371 |
| Power Rating | 0.008 | 49.3/44.6 | 0.0001 | 0.2486 |
| Total Takedown Percentage | 0.0195 | 57.0/42.0 | <.0001 | 0.2463 |
| Ground Activity per Takedown | 0.0016 | 51.7/43.6 | 0.0686 | 0.2497 |

These univariate models provide us with some useful information before we begin combining the variables in a multivariate model. We can see that all the variables with the exception of Ground Activity per Takedown are individually significant at an $\alpha = 0.05$ level. In terms of percent concordant, the three 'volume striking' statistics seem to perform the best while Power Rating performs the worst. None of the models has even a moderately high R-Square.

Combining these six variables into a multiple logistic regression model produces the model below in Tables 8 and 9.

Table 8: Second level variable full model

| Parameter | Estimate | Standard Error | Chi-Square | Pr > ChiSq | Standardized Estimate |
|---|---|---|---|---|---|
| Intercept | 0 | 0.0384 | 0.0000 | 1.0000 | |
| Ground Activity per Takedown | 0.0236 | 0.0245 | 0.9227 | 0.3368 | 0.0209 |
| Power Rating | -0.8247 | 0.8784 | 0.8816 | 0.3478 | -0.0257 |
| Strike Differential per Minute | 0.0138 | 0.0239 | 0.3335 | 0.5636 | 0.0354 |
| Striking Ratio | -0.0629 | 0.0198 | 10.0727 | 0.0015 | -0.0981 |
| Log(Striking Ratio) | 0.4368 | 0.1231 | 12.5950 | 0.0004 | 0.2425 |
| Total Takedown Percentage | 0.4991 | 0.1433 | 12.1380 | 0.0005 | 0.0807 |

Table 9: Second level variable full model diagnostics

| % Concordant/Discordant | Rescaled R-Squared | Brier Score |
|---|---|---|
| 62.4/37.0 | 0.0678 | 0.2373 |

The full model involving all six created second-level variables has three which appear at first glance to be significant, and three which appear to be rather insignificant. Indeed, by removing variables until every variable left is significant at$\alpha = 0.01$, we arrive at the following model shown in Tables 10 and 11.

Table 10: Second level variable reduced model

| Parameter | Estimate | Standard Error | Chi-Square | Pr > ChiSq | Standardized Estimate |
|---|---|---|---|---|---|
| Intercept | 0 | 0.0384 | 0.0000 | 1.0000 | |
| Striking Ratio | -0.0495 | 0.0180 | 7.5931 | 0.0059 | -0.0809 |
| Log(Striking Ratio) | 0.4820 | 0.0552 | 76.1438 | <.0001 | 0.2772 |
| Total Takedown Percentage | 0.3486 | 0.1303 | 7.1560 | 0.0075 | 0.0601 |

Table 11: Second level variable reduced model diagnostics

| % Concordant/Discordant | Rescaled R-Squared | Brier Score |
|---|---|---|
| 62.3/37.2 | 0.0638 | 0.2380 |

We are left with only three variables in the final model – Striking Ratio, Log(Striking Ratio), and Total Takedown Percentage. The estimates of Log(Striking Ratio) and Total Takedown Percentage are both positive, which indicates that as values for difference in Log(Striking Ratio) and Total Takedown Percentage increase, the log odds of winning also increase. This result seems intuitive. Striking Ratio has a negative estimate, which seems counter-intuitive until we remember that Striking Ratio and Log(Striking Ratio) both measure the same aspect of the fight (although one is transformed by a log scale). If we examine both components of Striking Ratio in our model, we can see that the equation

$$-0.0495x + 0.482 * \ln(x)$$

is increasing from 0 until about 9.7, meaning that as difference in Striking Ratio increases from 0 to ~9.7, the log odds of winning increase. This range covers the vast majority of all fights.

These results seem a bit strange at first glance. In terms of both percent concordance, rescaled R-Square and Brier Score, the best model is the univariate model which predicts outcomes based on the Log(Striking Ratio) variable. When Log(Striking Ratio) is put into a multiple logistic regression with other terms, the concordance, rescaled R-Squared and Brier Score all worsen. Even when we reduce from the full model to a model where every variable is significant at $\alpha = 0.01$, the univariate model for Log(Striking Ratio) still outperforms that model.

This somewhat odd result leaves us with an interesting choice between the simpler univariate model of Log(Striking Ratio) and the more complex model with Striking Ratio, Log(Striking Ratio) and Total Takedown Percentage. If we choose the simpler model, we are discarding two variables which are significant at the $\alpha = 0.01$ level. If we choose the latter model, we are adding two variables which actually slightly decrease the predictive power of the model. I feel that a persuasive argument could be made for both sides, but ultimately because Log(Striking Ratio) is highly correlated with both Striking Ratio and Total Takedown Percentage, I will opt to remove those two variables and utilize the univariate model with only Log(Striking Ratio).

Next, we will measure our final model involving novel, second-level variables against our three hypothesis listed in the Methods section. For our first and most basic

test, it is clear that our model is better than random chance in predicting fight outcomes. The difference between predicting 62.7% of fights correctly and only 50% of fights correctly is easily significant at $\alpha = 0.01$ (p <0.0001) given our large sample size.

The next comparison for our final model is against a model which attempts to predict fight outcomes from winning percentage. As with other variables, winning percentage was calculated as a difference in cumulative winning percentage between opponents.

Table 12: Winning percentage model

| Parameter | Estimate | Standard Error | Chi-Square | Pr > ChiSq | Standardized Estimate |
|---|---|---|---|---|---|
| Intercept | 0 | 0.0349 | 0.0000 | 1.0000 | |
| Winning Percentage | 1.07 | 0.0963 | 123.405 | <0.0001 | 0.2235 |

Table 13: Winning percentage model diagnostics

| % Concordant/Discordant | Rescaled R-Squared | Brier Score |
|---|---|---|
| 59.5/37.2 | 0.0503 | 0.2405 |

These statistics reveal that the model based on winning percentage is inferior to our model using the second level variable Log(Striking Ratio). The rescaled R-Square is lower, the Brier score is higher, and the model predicts fewer outcomes correctly. The difference in percent concordant between the two models is significant at $\alpha = 0.05$ (p =0.02). Based on these results, we conclude that our model predicting fight outcomes

from secondary variables is significantly better than a model which predicts fight outcomes from winning percentage.

Our final measuring stick for our second level variable model is to compare it to a model comprised of first level variables. Using the first level variables from Table 1, we first create univariate logistic models to predict win/loss outcomes.

Table 14: First level variable univariate models

| | Rescaled R-Square | % Concordant /Disconcordant | Associated p-value | Brier Score |
|---|---|---|---|---|
| Total Strikes Landed | 0.0528 | 62.6/34.0 | <.0001 | 0.2397 |
| Knockdowns | 0.0203 | 45.0/28.6 | <.0001 | 0.2461 |
| Takedowns Landed | 0.0660 | 60.8/29.8 | <.0001 | 0.2371 |
| Positional Advances | 0.0630 | 60.7/30.9 | <.0001 | 0.2375 |
| Submission Attempts | 0.0357 | 52.4/33.1 | <.0001 | 0.2432 |

Every univariate model from first level variables is significant to a high degree. Next, we combine these five variables into a multivariate logistic model.

Table 15: First level variable full model

| Parameter | Estimate | Standard Error | Chi-Square | Pr > ChiSq | Standardized Estimate |
|---|---|---|---|---|---|
| Intercept | 0 | 0.0276 | 0.0000 | 1.0000 | |
| Total Strikes Landed | -0.0001 | 0.0001 | 0.5576 | 0.4552 | -0.0209 |
| Knockdowns | 0.0823 | 0.0138 | 35.3193 | <.0001 | 0.1159 |
| Takedowns Landed | 0.0367 | 0.00527 | 48.5546 | <.0001 | 0.1915 |
| Positional Advances | 0.0109 | 0.00429 | 6.4400 | 0.0112 | 0.0783 |
| Submission Attempts | 0.0160 | 0.00577 | 7.6920 | 0.0055 | 0.0602 |

Table 16: First level variable full model diagnostics

| % Concordant/Discordant | Rescaled R-Squared | Brier Score |
|---|---|---|
| 64.9/32.0 | 0.0851 | 0.2334 |

The full model for first level variables outperforms all previous models, including our models from novel, second level variables. This full model has a higher percent concordant, a higher rescaled R-Squared, and a lower Brier score than any previous model. The difference between the percents concordant of the full first level variable model and the best second level variable model (64.9 to 62.7) is statistically significant at $\alpha = 0.05$ with a p-value of 0.0209.

This full model has every variable significant other than Total Strikes Landed, which appears to be clearly insignificant. Removing Total Strikes Landed, we arrive at the final model

Table 17: First level variable final model

| Parameter | Estimate | Standard Error | Chi-Square | Pr > ChiSq | Standardized Estimate |
|---|---|---|---|---|---|
| Intercept | 0 | 0.0276 | 0.0000 | 1.0000 | |
| Knockdowns | 0.0770 | 0.0119 | 41.6626 | <.0001 | 0.1085 |
| Takedowns Landed | 0.0353 | 0.00490 | 51.8519 | <.0001 | 0.1840 |
| Positional Advances | 0.0101 | 0.00417 | 5.9178 | 0.0150 | 0.0729 |
| Submission Attempts | 0.0152 | 0.00566 | 7.2086 | 0.0073 | 0.0571 |

Table 18: First level variable final model diagnostics

| % Concordant/Discordant | Rescaled R-Squared | Brier Score |
|---|---|---|
| 64.7/31.7 | 0.0850 | 0.2335 |

This final model using only first level variables has essentially the same rescaled R-Squared and Brier Score. It loses 0.2 percent concordance, but also removes 0.3 percent discordance thus increasing the ratio of concordance to discordance while removing an insignificant variable.

Because both the first level variable model and the second level variable model were significantly better than both random chance and winning percentage, and because the models included differing significant terms, we next attempt to create a model which combined first level and second level variables. To do so, we start with a full model using all six created second level variables and all five first level variables, and then use backwards selection to remove any variables not significant at $\alpha = 0.05$.

Table 19: Combined final model

| Parameter | Estimate | Standard Error | Chi-Square | Pr > ChiSq | Standardized Estimate |
|---|---|---|---|---|---|
| Intercept | 0 | 0.0393 | 0.0000 | 1.0000 | |
| Knockdowns | 0.0610 | 0.0149 | 16.8510 | <.0001 | 0.1121 |
| Takedowns Landed | 0.0355 | 0.00470 | 57.1120 | <.0001 | 0.2294 |
| Total Strikes Landed | -0.00042 | 0.000162 | 6.7485 | 0.0094 | -0.0946 |
| Submission Attempts | 0.0217 | 0.00556 | 15.1403 | <.0001 | 0.0999 |
| Striking Ratio | -0.0426 | 0.0198 | 4.6385 | 0.0313 | -0.0664 |
| Log(Striking Ratio) | 0.4794 | 0.0628 | 58.3300 | <.0001 | 0.2661 |
| Total Takedown Percentage | 0.4402 | 0.1522 | 8.3639 | 0.0038 | 0.0712 |

Table 20: Combined final model diagnostics

| % Concordant/Discordant | Rescaled R-Squared | Brier Score |
|---|---|---|
| 66.8/32.9 | 0.1211 | 0.2274 |

This model keeps seven variables; three of our novel, created variables and four first level variables. It appears to be superior to any earlier effort to forecast bouts, as it has a higher percent concordant and a higher rescaled R-Squared than any previous model, as well as a lower Brier Score than any previous model. The difference in percent concordant between this combined model (66.8%) and the model from first level variables (64.7%) is statistically significant at $\alpha = 0.05$ with a p-value of 0.0251.

Performing a discriminant analysis on the data set provides some additional support for this combined, seven variable model. A discriminant analysis using backwards selection and $\alpha = 0.05$ finds the following variables to be significant.

Table 21: Discriminant Analysis

| Parameter | Partial R-Square | F-Value | Pr > F |
|---|---|---|---|
| Knockdowns | 0.0059 | 16.76 | <.0001 |
| Takedowns Landed | 0.0209 | 60.78 | <.0001 |
| Total Strikes Landed | 0.0024 | 6.74 | 0.0095 |
| Submission Attempts | 0.0055 | 15.61 | <.0001 |
| Striking Ratio | 0.0014 | 3.93 | 0.0476 |
| Log(Striking Ratio) | 0.0210 | 60.90 | <.0001 |
| Total Takedown Percentage | 0.0029 | 8.22 | 0.0042 |

The seven selected variables under a discriminant analysis turn out to be the same seven variables chosen by our final logistic regression model. Although this discriminant analysis may be of limited value (see Limitations), it does provide some evidence that our model is a reasonable choice.

In order to further assess this model's ability to forecast fights, we will cross-validate our model. Cross validation helps insure that a model does not over fit the given data. For our cross validation, the data set was split into two, with one data set containing

75% of the data (randomly assigned) and the second data set containing the remaining 25%. The model which fits the first data set is then used to forecast the second data set, measuring whether the model can forecast data which it has not previously seen.

Table 22: Cross Validation

| | Predicted Wins | Predicted Losses |
|---|---|---|
| Actual Wins | 212 (29.73%) | 137 (19.21%) |
| Actual Losses | 124 (17.39%) | 240 (33.66%) |

Table 23: Cross Validation diagnostics

| % Concordant/Discordant | Rescaled R-Squared | Brier Score |
|---|---|---|
| 63.4/36.6 | 0.1518 | 0.2211 |

When cross validating this model, we can see that the results remain highly significant. The cross-validated data set has a lower percent concordant, but actually improves its rescaled R-Squared and Brier Score.

# 6. LIMITATIONS

This project is subject to some limitations which I wish to note. This project's core technique is logistic regression. Our model utilizes logistic regression to attempt to predict the binary outcome "Win/Loss" using cumulative career statistics created for each fighter. The general model for binary logistic regression is given on page 14 of the Methods section. One of the important assumptions of this model is that the data $Y_1, \ldots, Y_n$ are independently distributed across sample units. This assumption leads to a problem with this particular data set. Because the data set comprises major fights from 1993 to roughly the present, a very large number of fighters have multiple bouts in the data set. Many fighters have more than a dozen fights recorded. It may be unrealistic to assume that a fighter's results in one fight are totally independent of his results in a previous fight. A fighter may have a particular skill set which makes his fights correlated with one another, and could be measured with a 'fighter effect'.

However, modeling this type of correlation would prove extremely problematic. A model which accounts for a fighter effect would look like

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

Where $\alpha_j$ is the effect due to the j-th fighter, which might be assumed to be a random variable, as in generalized linear mixed effect models. However, this type of model has its own problems. Assuming that a fighter has a constant fighter effect ignores that fighter's change over time. It would be incorrect to assume that a fighter has the same skillset and abilities as a younger rookie, in his prime, and as he ages and

deteriorates. It is unreasonable to assume that fighters remain static over their careers, and because $\alpha_j$ is constant, this represents a problem.

Ultimately, despite the lack of a complete independence among the data, we will still use logistic regression to model winning outcomes in MMA bouts. Although this method is not perfect, it is still a valuable tool as long as we are careful to remember the potential violation of the model's independence assumption.

After using logistic regression to create a forecasting model, we provided a brief discriminant analysis. Logistic regression and discriminant analysis have many similarities. Both techniques make use of independent variables to attempt to predict a dependent variable which is categorical, not continuous, and often produce similar results. Discriminant analysis, however, requires assumptions that logistic regression does not. Discriminant analysis requires the assumption that the independent variables used to predict categorical grouping are normally distributed. None of the 11 independent variables considered are normally distributed, as determined by the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests for normality. Under logistic regression, the assumption of normality for independent variables is not required, but for discriminant analysis it is. While discriminant analysis can be somewhat robust in the face of certain types of moderate non-normality (Lachenbruch and Goldstein, 1979), our independent variables are unanimously non-normal to a significant degree. Thus, discriminant analysis may be useful in a broad sense to help

support our findings from logistic regression, but its precise usefulness is limited by the data set's violation of normality.

# 7. CONCLUSIONS

The goal of this project was to examine the ability of novel variables created from manipulation of 'count' variables to forecast the results of mixed martial arts fights. These created second level variables were chosen with an eye towards their significance in measuring different areas of MMA skills, and for their relevant interpretations to those familiar with the sport. The data from FightMetric was analyzed using logistic regression, and was to be compared with various baseline measures of predictive ability such as random chance, winning percentage and the first level variables.

Using logistic regression, we were able to create a model for predicting win/loss outcomes for MMA bouts using our created, second level variables. This model was clearly better than random chance, and was also significantly superior to a model which forecast win/loss outcomes from winning percentage. However, this second level model was not superior to first level 'count' variables in predicting outcomes. In fact, the model from first level variables proved to be significantly more predictive than the model from second level variables, with a higher percent concordant, a higher rescaled R-Squared and a lower Brier score.

Combining both first level and second level variables into a single model, we arrived at the best model examined in the entire project. This combined logistic regression model contained a combination of significant first level and second level variables. In all, four first level variables and three second level variables were found to be significant in the combined model. This final, combined model was significantly more predictive than any previous model, including random chance, winning percentage,

and the separate first level and second level models. It had a higher percent concordant, a higher rescaled R-Squared, and a lower Brier score than any previous model, clearly indicating that this model was superior to all earlier models.

A brief discriminant analysis using all eleven of the first level and second level variables together provided additional support for this final, seven-variable combined model. The discriminant analysis identified seven independent variables which were significantly related to win/loss outcomes, and those seven variables were the same seven variables chosen as part of our final combined model. We performed a cross validation of the final model, which showed that the final model performed reasonably well predicting for previously unknown data.

Overall, the project was able to successfully create several significant second level variables which forecast win/loss results of MMA fights. These variables were, in the end, inferior to first level count variables in predicting outcomes when the two categories were held apart. But combined, the second level variables improved on the predictive power and accuracy of every previous model and the combined model was significantly superior to the other models examined.

REFERENCES

Stekler, Herman O. & Klein, Andrew. "Predicting the Outcomes of NCAA Basketball Championship Games." *Journal of Quantitative Analysis in Sports* Volume 8 Issue 1 (2012)

Boulier, Bryan L. & Stekler, H.O. "Are Sports Seedings Good Predictors? An Evaluation." *International Journal of Forecasting* Volume 15 (1999) 83-91.

Lachenbruch, P. A. & Goldstein, M. "Discriminant Analysis." *Biometrics* Volume 35 No. 1 (1979) 69-85.