# Using Multiple Linear Regression to Predict UFC Career Win-Loss Percentage

Dylan Palmer
Computer and Information Sciences
Virginia Military Institute
Lexington, Virginia, USA
Palmerde245@vmi.edu

Reid Cox
Computer and Information Sciences
Virginia Military Institute
Lexington, Virginia, USA
Coxrc24@vmi.edu

**Abstract**

## 1. Introduction

The Ultimate Fighting Championship (UFC) is the largest MMA promotion in the world [1]. that produces events worldwide that showcase 11 weight divisions (eight men and three women's). The sport has grown in popularity at a consistent pace and was quickly noticed by the sports betting community. UFC surpasses boxing in terms of betting revenues

since 2007 [2]. Blogs, news agencies, betting companies, and statistics companies have used predictions to enhance public fervour in popular fights, increase user pay-per-view, and establish betting odds. The need for accurate pairing of fighters has become a necessary part of the sport, to make a profit and an interesting fight.

## 2. Motivation

We hope that our research strengthens mathematical and predictive analysis using predictive analytics for competitive sports, combining techniques and performance into quantitative variables that can be used to predict the performance of an athlete over the course of their career. There has been very little research done on the use of machine learning algorithms to predict an athlete's career performance. Because of this, we believed that we could use ML through linear regression to predict a win-loss ratio through key physical characteristics of a fighter, in hopes of being able to better predict a fighter's likelihood to be successful in the UFC. Being able to predict performance could be a huge help in the industry, allowing better fighting pairings and making fight cards less of an art and more of a science.

## 3. Literature Review / Background

Throughout our literature review, we discovered that studies conducted on athletics in ways most like our research were mostly done for kinesthetic purposes rather than competitive predictions. This has further motivated our research in hopes to bring nuance to statistical analysis in use cases for athletic performance and competition. Despite the lacking research in our specific field, we found several works that were similar in our research in that they include multifactorial phenomena that contribute to athletic performance such as technique, fitness, physical features, etc.

Loturco, et al. utilized multiple linear regression analysis to predict punching acceleration from selected strength and power variables in elite karate athletes [3]. The paper used multivariate analysis including kinaesthetic tests before and after the impact tests. 1 repetition maximum (1RM) tests for bench press and squat machine exercises were used as determinants for upper and lower body maximal dynamic strength, measuring squat jump and countermovement jump heights, and a 40% body mass load bench throw and jump squat measurement was used to measure propulsive power for the same upper and lower body movements.

Joao et. al. conducted a multiple linear regression analysis to identify the physical fitness variables that best predict Special Judo Fitness Test (SJFT) performance [4]. The Special Judo Fitness Test consists of a test in which the judoka (athlete) must project (throw) his opponents as fast as possible within 15, 30, and 30-second periods with 10 second intervals among them. The executor throws two partners placed 6 meters apart from one another as many times as possible. The athlete's heart rate is recorded immediately after the test and 1 minute later. It is important to note that this is a kinaesthetic test that was studied using multivariate analysis including upper- and lower-body aerobic and anaerobic tests, and those tests alone had a small ability to predict SJFT variables, indicating that further influences such as technique, experience, or physical features such as weight, height, and

wingspan play a larger role in SJFT performance. We account for physical features in this paper.

Hitkul, et. al. explains the potential for an upset in an MMA fight despite mathematical predictions, and the problems that are exacerbated by the lack of well-defined, time series databases of fighter profiles prior to fights. They attempt to develop an efficient model based on machine learning algorithms for the prior predictions of UFC fights [5]. Their multivariate analysis is the closest to what we hope to achieve in this paper. Fight statistics such as significant strike attempts and landed on the ground and in clinch, head strikes attempted and landed, and total strikes attempted and landed are included as the ten highest correlations with their target upset variable. Linear correlations include the time the athlete spends in half-guard control, significant punches attempted and landed on the ground, ground control time, and leg strikes attempted and landed. The paper concluded that the Support Vector Machine (SVM) machine learning model was best at reaching a predicted target variable.

Undergraduate thesis by Clara Chan of the National College of Ireland used different technologies to predict UFC fights [6] and expanded upon a previously mentioned paper [5]. In which it was determined that neural networks outperformed Random Forest machine learning and included variables such as strike differential between fighters in the bout, the age of the fighters, the KO and win/loss differential between fighters, height, and fighting stance. Many of these variables are included in our analysis, but it is important to note that we hope to predict the career win-loss ratio over a data subset, so differentials between opponents were not considered.

McQuaide of Stanford explores the possibilities of analysis with the presentation of over one hundred different features before a UFC fight, and the use of machine learning to predict the fight outcomes [7]. The analysis included variables such as the fighter's height, average distance from opponents, average distance of landed strikes, landed and attempted strikes to different parts of the body, and current win/loss.

## 4. Problem Statement

### 4.1 Problem Statement

#### 4.1.1 
Since UFC is the largest Mixed Marshal Arts contest in the world, there is a need for a reliable way to predict the performance of a fighter over the course of their career. The goal of this paper is to reliably predict the outcome of a fighter's win-loss percentage based on weight, wingspan, wins, loses, draws, etc.

### 4.2 Use Cases

#### 4.2.1 
UFC can use this to match fighters up against each other, and by news organizations to generate predictions based on how they should perform. Having an even fight will make a better fight and better entertainment and will increase betting revenue and sports analysis views.

#### 4.2.2 
Sports betting organizations can use a regression analysis to establish odds for return and increase revenue.

5. Exploratory Data Analysis

5.1 After importing our data of UFC fighter and statistics into pandas as the variable "df". We used the method "info()" to give more details about the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4109 entries, 0 to 4108
Data columns (total 18 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   name                                    4109 non-null   object
 1   nickname                                2255 non-null   object
 2   wins                                    4109 non-null   int64
 3   losses                                  4109 non-null   int64
 4   draws                                   4109 non-null   int64
 5   height_cm                               3812 non-null   float64
 6   weight_in_kg                            4022 non-null   float64
 7   reach_in_cm                             2183 non-null   float64
 8   stance                                  3287 non-null   object
 9   date_of_birth                           2975 non-null   object
 10  significant_strikes_landed_per_minute   4109 non-null   float64
 11  significant_striking_accuracy           4109 non-null   float64
 12  significant_strikes_absorbed_per_minute 4109 non-null   float64
 13  significant_strike_defence              4109 non-null   float64
 14  average_takedowns_landed_per_15_minutes 4109 non-null   float64
 15  takedown_accuracy                       4109 non-null   float64
 16  takedown_defense                        4109 non-null   float64
 17  average_submissions_attempted_per_15_minutes 4109 non-null float64
dtypes: float64(11), int64(3), object(4)
memory usage: 578.0+ KB
```

Figure 1: Dataset Information

5.2 From this command the dataset is shown to 18 columns, ranging from 2255 to 4109 entries, and a various array of entry types such as object, int64, and float64. Since the goal of this paper is to see how the fighters' attributes after their win percentage, many of the useless columns can be dropped. The columns that where drop where ones that provided no meaningful data about the fighters' physical attributes, which are their "name", "date_of_birth" and "nickname."

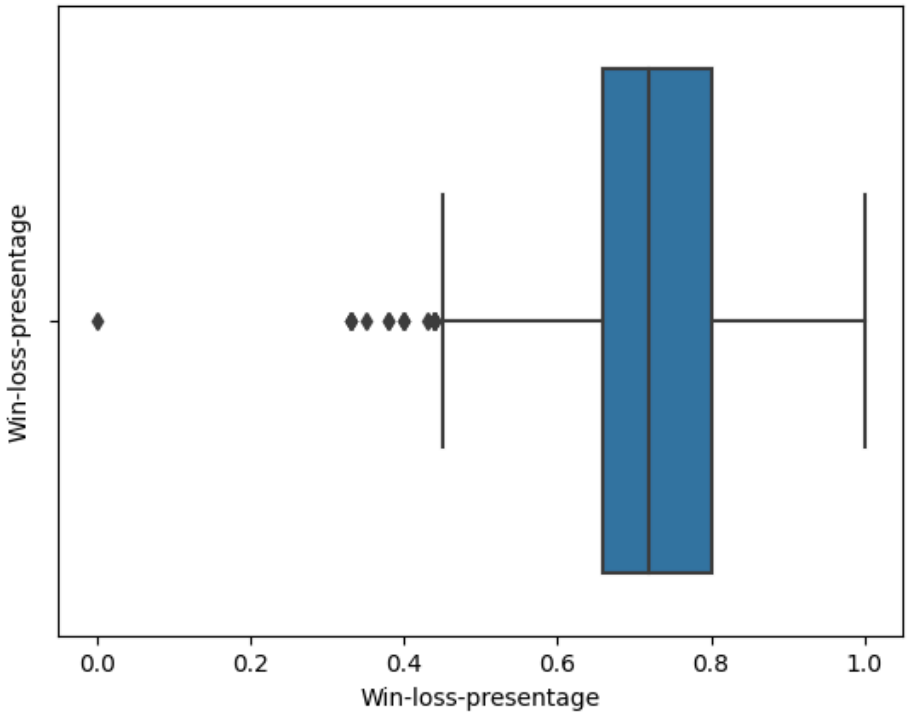


Figure 2: Boxplot of Win-Loss-Percentage

5.3 Looking at our “Win-Loss-Percentage" we can see most of the fighters lie in-between 60% to 80%. We also have two outliers at 100% and 0% meaning either fighter has zero wins or went undefeated during his time in UFC. This data does not account for the number of fights that a fighter has on his record meaning anything outside of our box could be skewed.

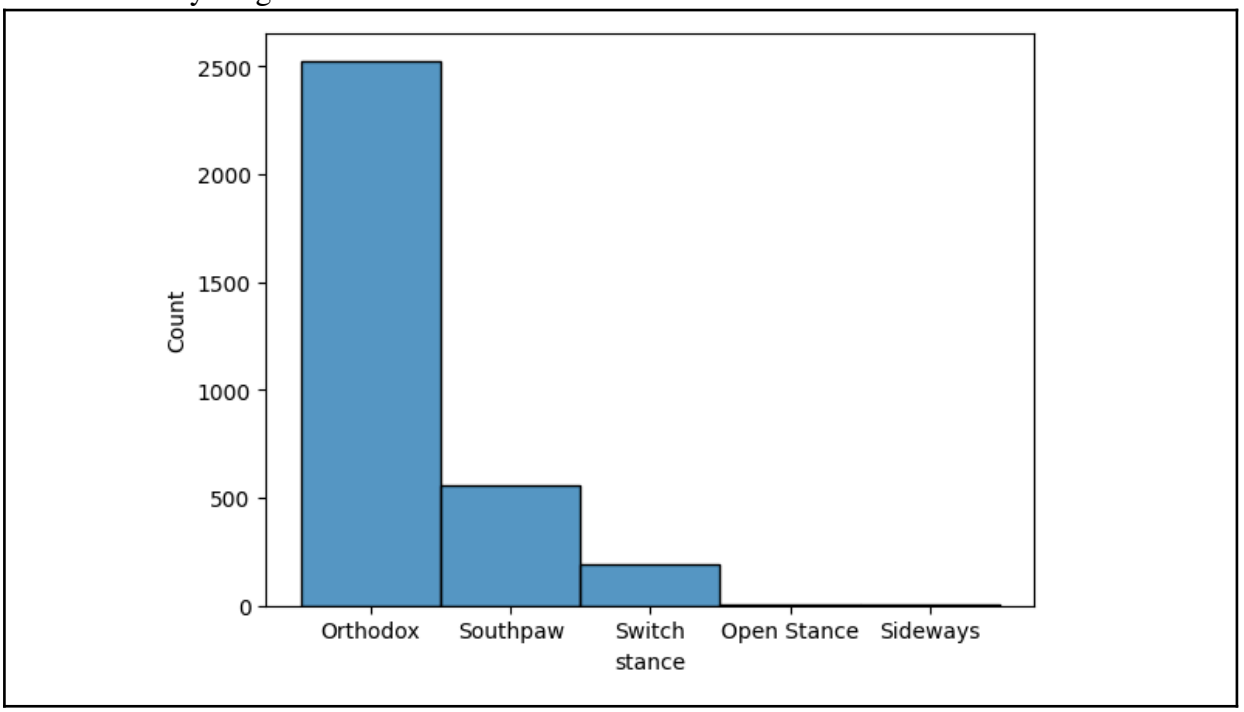

Figure 2: Histogram of Fighter's Stances

5.4 This graph shows the amount of stance that the UFC has had over its long history. As we can see most of these fighters fight in the Orthodox(right-handed) stance with the two second highest stances being Southpaw (Left-handed), and Switch (Right and Left-handed). The last two stances are Open and sideways which range in the single digits of fighters making them have little impact on the data as a whole. We will still include them in our analysis.


Figure 3: Heatmap of Our Variables by Correlation

5.5 The heatmap above shows the correlation between two columns in the data set. We can see that the overall the heatmap does not have a huge correlation to the other variables, but in a fight one attribute does not make a fighter's performance. So, it makes sense that there are small correlations across the heatmap which will hopefully let us make an accurate linear regression model. From this point we will Now from this point we can start our data preprocessing, to make a clean dataset that we can make a linear regression model from.

6. Data Preprocessing

6.1 Missing Value

6.1.1 After doing an exploratory data analysis, the method "df.isnull().sum()" is used to show the sum of each null value per column.

```
wins                                          0
losses                                        0
draws                                         0
height_cm                                   297
weight_in_kg                                 87
reach_in_cm                                1926
stance                                      822
significant_strikes_landed_per_minute         0
significant_striking_accuracy                 0
significant_strikes_absorbed_per_minute       0
significant_strike_defence                    0
average_submissions_attempted_per_15_minutes  0
dtype: int64
```

Figure 4: Null Values per Columns

6.1.2 From this the method “df.dropna(how='any', axis=0)” is used to drop all rows that have these null values. This is done because the four columns that have null values are used to predict the fighter's win ratio making any null entry invalid data and useless to the overall model.

6.2 Inconsistent and Redundant data

6.2.1 In this dataset there are numerous examples of inconsistent data, many of which having the values of zero. Zero for many datasets is not a problem, however for many of the columns having a zero should be impossible. For example, “significant_strikes_landed_per_minute”, “significant_striking_accuracy”, “significant_strikes_absorbed_per_minute”, “significant_strike_defence”, and “average_submissions_attempted_per_15_minutes” should be impossible to have zero if the fighter has had a fight. These numbers can be low with many being with in the hundredth of decimal place but never zero, because of this all rows that process a zero in the any of these columns have been dropped from the dataset.

```
dfresult = dfresult.drop(dfresult[dfresult['significant_strikes_landed_per_minute'] == 0].index)
dfresult = dfresult.drop(dfresult[dfresult['significant_striking_accuracy'] == 0].index)
dfresult = dfresult.drop(dfresult[dfresult['significant_strikes_absorbed_per_minute'] == 0].index)
dfresult = dfresult.drop(dfresult[dfresult['significant_strike_defence'] == 0].index)
dfresult = dfresult.drop(dfresult[dfresult['average_submissions_attempted_per_15_minutes'] == 0].index)
#If the fighter had a fight none of these values could be zeros so we drop those rows adn inconsistant data
```

Figure 5: Dropping inconsistent row

6.2.2 " takedown_accuracy” and “takedown_defense” columns have also been dropped due to having too many zeros, making the dataset too small if we drop all the rows that contain zeros.

6.2.3 After handling all the inconsistent data and missing values the dataset has no redundant values, shown by “df.duplicated().sum()” method producing a value of zero.

6.3 Outliers

6.3.1 The dataset had no outliers that needed to be handled, because all the da6ta was based on real life people and their tangible achievements and attributes. The dataset also has 2137 entries making a outliers no inconsequential.

6.4 Encoding Categorical Data

6.4.1 The dataset to this point only has one column of data that was categorical which was “stance”. Stance columns consist of four values orthodox, southpaw, switch and open stance. To make this column numerical each stance or object value was an assign a number one through four as shown below.

```
[68] dfresult['stance'] = dfresult['stance'].replace({1 : 'Orthodox', 2 : 'Southpaw', 3: 'Switch', 4: 'Open Stance'})
     #make everything nurmeric data
```

Figure 4: Assigning Categorical data Numeric values

7. Model Building - Linear Regression Model/Logistic Regression Model/Clustering Model

7.1 For this analysis, we have established that we will be using a supervised linear regression learning algorithm.

7.2 Multivariate analysis for this research is based off of the changes to many different variables and final calculations for one dependent variable. Because of this, a multiple linear regression model fits our goals best.

7.3 In this paper, we have developed a model that calculates variate coefficients for each column in order to predict a win/loss ratio as our dependent variable across the fighter’s career.

8. Model Performance Evaluation

8.1 Creation of Training and Testing Models

8.1.1 In order to create our X training and testing and Y training and testing models, we utilized Sklearn’s *train_test_split()* function.

```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.30, random_state =1)
```

8.2 R-squared

8.2.1 The R-squared performance that we acquired from our training model was 0.738. The R-squared performance that we acquired from our testing model was 0.772.

8.2.2 This R-squared performance score allows us to determine the accuracy of our model on a percentage scale, from 0-100. We see that on the training model we acquired a 73% accuracy, while on our testing model we acquired a 77% accuracy rate. Given that we are examining an athletic

competition, with many different variables that apply to our fighters, being able to acquire an accuracy rate higher than 60% is notable.

8.3 Adjusted R-squared

8.3.1 The adjusted-R-squared performance that we acquired on our training model was 0.735. The adjusted-R-squared performance that we acquired on our testing model was 0.769.

8.3.2 The adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. Because we do not have an unwarranted and high R-squared, and an adjusted-R-squared higher than 0.7, we can tell that our model is a good fit and is not overfit.

8.4 Root Mean Squared Error (RMSE)

8.4.1 The RMSE performance that we acquired on our training model was 0.057. The RMSE performance that we acquired on our testing model was 0.053.

8.4.2 Root mean square error measures the average difference between the model's predicted value and the actual values as a standard deviation of the residuals. Our very low performance scores, which should be closer to zero, indicate that our model fits the data very well.

8.5 Mean Squared Error (MSE)

8.5.1 The MSE performance that we acquired on our training model was 0.0032. The MSE performance that we acquired on our testing model was 0.0028.

8.5.2 The mean square error is the average of the square of errors. Our performance score is between 0.002 and 0.003, which is very close to zero. This indicates that our model's error magnitude is very small.

8.6 Mean Absolute Error (MAE)

8.6.1 The MAE performance that we acquired on our training model was 0.037. The MAE performance we acquired on our testing model was 0.036.

8.6.2 The mean absolute error is the sum of all the distances between the actual and predicted values of our model, divided by the total number of points in the dataset. It is the absolute average distance of our model prediction. Our model received a score of under 0.04 both times, indicating that our line of best fit is very nearly exact to the data that we used in our analysis.

8.7 Mean Absolute Percentage Error (MAPE)

8.7.1 The MAPE performance that we acquired on our training model was 0.057. The MAPE performance that we acquired on our testing model was 0.052.

8.7.2 The mean absolute percentage error measures the average magnitude of error produced by the model. On our model, we received a performance score between 5% and 6%, indicating that we have predictions that are off, on average, of only 5%-6%.

9. Results and Discussions

In this paper, we hoped to create a multiple linear regression model that could fit existing UFC fighter statistics and use it to generate a predicted win-loss percentage over the course of a fighter' career. In this paper, we performed data preprocessing and exploratory data analysis procedures to format the UFC fighter statistics in a way that applied numerical values to categorical features such as stance and assigned the win-loss percentage as our dependent variable. Using Python and Google Colab libraries for statistics and regression

analysis, we developed a multiple linear regression model that has performed well against training and testing models.

Using these methods, we developed the following multiple linear regression analysis formula:

```
win/loss = [0.72134853] + (0.012506621923289326)*(wins) +
(-0.0321120057881943)*(losses) + (-0.022618754390111198)*(draws) +
(0.000101596651970964)*(height_cm) +
(1.5667221682499205e-05)*(weight_in_kg) +
(-1.8870033893296848e-05)*(reach_in_cm) +
(0.0031669603101046806)*(significant_strikes_landed_per_minute) +
(0.0003460407501996118)*(significant_striking_accuracy) +
(-0.003680627728337594)*(significant_strikes_absorbed_per_minute) +
(-0.0005032755853800423)*(significant_strike_defence) +
(0.0031669603101046728)*(average_submissions_attempted_per_15_minute
s) + (0.004081744934054081)*(stance_Open Stance) +
(-0.002824159497229438)*(stance_Orthodox) +
(-0.002413921839710653)*(stance_Southpaw) +
(0.001156336402886018)*(stance_Switch)
```

10. Conclusions

In this paper, we showed that it is possible to develop a multiple linear regression model that can fit historical UFC fighter data and develop comparable predictions on career win-loss percentage. We have contributed to statistical analysis research related to professional athletic careers and predicting performance over a long period of time, with the assistance of existing research that used machine learning algorithms to determine individual performance in an event or to predict the performance of fighters in a single bout [3][4][5].

Further research should examine the use of multiple linear regression formulas based on greater variance in data and in sports that include more than one opponent, such as American football or rugby. Combining this formula with historic data for each fighter with discretion to opponents and the statistics of each bout should strengthen this model.

This model performed well on all performance metrics that were applied as mentioned in section 8 of this paper. This indicates that the model fits well to the existing data.

**References**


[1] Gift, Paul (March 16, 2022). "UFC Posts Best Financial Year In Company History". *Forbes*. Retrieved April 16, 2023. https://www.forbes.com/sites/paulgift/2022/03/16/ufc-posts-best-financial-year-in-company-history/?sh=3352c852330e

[2] Goff, Justin (July 11, 2007). "UFC set to surpass boxing in betting revenue". *MMAbettingblog.com*. Archived from the original on April 10, 2008. Retrieved March 5,

2008. https://web.archive.org/web/20080410032815/http://www.prweb.com/releases/mma/betting/prweb530718.htm

[3] Loturco, I., Artioli, G. G., Kobal, R., Gil, S., & Franchini, E. (2014). Predicting Punching Acceleration from Selected Strength and Power Variables in Elite Karate Athletes: A Multiple Regression Analysis. *Journal of Strength and Condition Research*, *28*(7), 1826–1832. https://doi.org/10.1519/JSC.0000000000000329. https://journals.lww.com/nsca-jscr/FullText/2014/07000/Predicting_Punching_Acceleration_From_Selected.6.aspx

[4] Martial Arts and Combat Sports Research Group. (n.d.). *Influence of physical fitness on Special Judo Fitness Test... : The Journal of Strength & Conditioning Research*. LWW. https://journals.lww.com/nsca-jscr/abstract/2021/06000/influence_of_physical_fitness_on_special_judo.34.aspx?context=latestarticles

[5] Hitkul, Aggarwal, K., Yadav, N., & Dwivedy, M. (1970, January 1). *A comparative study of machine learning algorithms for prior prediction of UFC Fights*. SpringerLink. https://link.springer.com/chapter/10.1007/978-981-13-0761-4_7

[6] Walsh, G. (2022, May 15). *Predictive analysis of UFC Fights: Technical Report*. NORMA@NCI Library. https://norma.ncirl.ie/5769/

[7] McQuaide, M. (n.d.). *Applying machine learning algorithms to predict ... - stanford university*. Applying Machine Learning Algorithms to Predict UFC Fight Outcomes. https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647731.pdf

[8] Link to Datasets: https://www.kaggle.com/datasets/asaniczka/ufc-fighters-statistics