Yonatan Kazovsky
Profesor Chelsea Parlett
CPSC 392 - 02
11 December 2023

An analysis of UFC Fighters

(CPSC 392 Final Project)

Introduction

The goal of the following analysis was to find meaningful trends and patterns in ufc fighters' data that will provide insight into what elements drive fighter success and other metrics. The dataset which was sourced from kaggle contains information on over 4000 fighters and provides the following metrics about each fighter in the dataset:

- Name (name of fighter)
- Nickname (nickname of fighter)
- Wins (number of wins in professional mma record)
- Losses (number of losses in professional mma record)
- Draws (number of draws in professional mma record)
- Height_cm (fighter height in centimeters)
- Weight_in_kg (weight of fighter in kilograms)
- Reach_in_cm (reach of fighter in centimeters)
- Stance (The fighting stance of the fighter (Orthodox/Southpaw/Switch))
- Date_of_birth (fighter's date of birth)
- Significant strikes landed per minute (The average number of significant strikes landed by the fighter per minute)
- Significant striking accuracy (The percentage of significant strikes that land successfully for the fighter)
- Significant strikes absorbed per minute (The average number of significant strikes absorbed by the fighter per minute)
- Significant strike defense (The percentage of opponent's significant strikes that the fighter successfully defends)
- Average Takedowns.. (The average number of takedowns landed by the fighter per 15 minutes)
- Takedown Accuracy (The percentage of takedown attempts that are successful for the fighter)

- Takedown Defense (The percentage of opponent's takedown attempts that the fighter successfully defends)
- Average Submissions.. (The average number of submission attempts made by the fighter per 15 minutes)

Before reading, take into consideration the following pitfalls that have affected the analysis:

1. Lots of data was lost when data cleaning as many fighters had null/missing data in various columns including fighters stance.
2. Analysis did not take into account the fact that fighters are assigned a weight class. Henceforth, the analysis does not account for trends that differ by weight class. This is something I would potentially look to improve in the future by analyzing weight classes individually.
3. Win rate is a variable i created myself by performing the following computation using already existing variables: wins/(wins+losses+draws)

The report is organized into 3 distinct questions which will each examine a different aspect of our fighters.

Question 1

***Question adjusted to use different variables which have a much more meaningful relationship with win rate.**

When predicting win rate, which predictor 'Significant strikes landed per minute', 'Significant striking accuracy', 'Significant strikes absorbed per minute', 'Significant strike defense', or 'Stance' improves the R^2 the most when compared to a model with all other variables except itself?

In other words, which of the striking metrics is most important in predicting win rate?

Data Cleaning: Null Values Dropped. Stance variable (categorical) dummied.

Modeling/Computation: A Train/Test split with 80/20 split was used. Continuous variables were z-scored. Six linear regression models were fit to predict win rate. One model used all the predictors, and the remaining 5 used all

but one predictor. R^2 scores were pulled from both train and test set and yielded the following results.

Graphs:

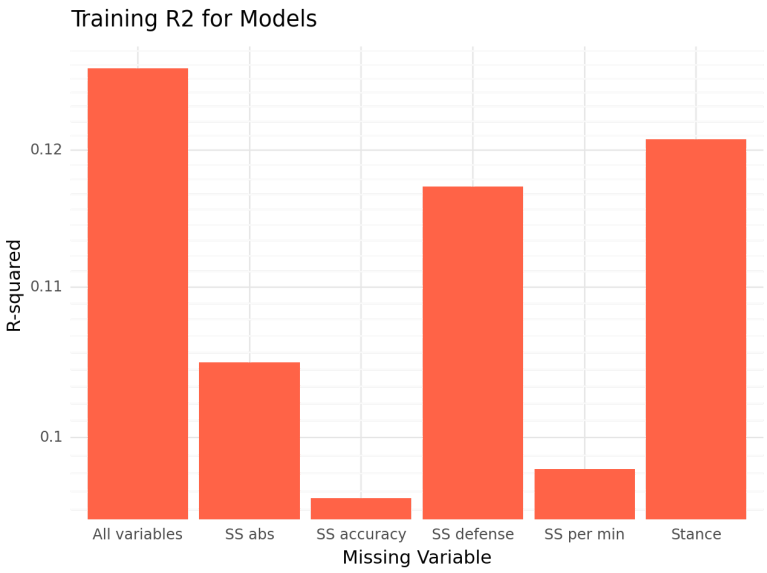


Figure 1
Relationship between missing variables and the corresponding R^2 for regression model on training data

As we can see in this graph, when we don't include the Significant Strike Accuracy and Significant Strikes per minute in our regression model we take a big hit to our R^2.

Figure 2
Relationship between missing variables and the corresponding R^2 for regression model on test data

When we don't include the Significant Strikes per minute in our regression model we take a large hit to the R^2. Somehow our R^2 is actually increased when we don't include Sig. Strike Accuracy which may suggest some issues with our data analysis.

Conclusions: Based on the graphs we can see that the 'significant strikes per min' metric has a huge impact on accurately predicting win rate on both our training and testing data. We can conclude this because in both instances,

regression models that didn't take sig. strikes per minute into account had much smaller R^2 values than the regression models that did include sig. strikes per minute. This means that if I were lets say an agent and I wanted to sign a new prospect fighter, I would weigh their significant strikes landed per minute pretty heavily when trying to predict their win rate and henceforth how successful they may become.

Question 2

When comparing a model using PCA on all continuous variables (other than height_cm) in the dataset and retaining enough PCs to keep 90% of the variance, to a model using all the continuous variables (other than height_cm), how much of a difference is there in mean absolute error when predicting fighter height?

Data Cleaning: Null Values Dropped.

Modeling/Computation: PCA was conducted on continuous variables excluding height_cm, while trying to retain 95% variance with minimal components. Continuous variables were z scored and two models were created. One based on the PCA transformed data and the other using the original variables. The MSE was calculated for both models and compared.

Graphs:

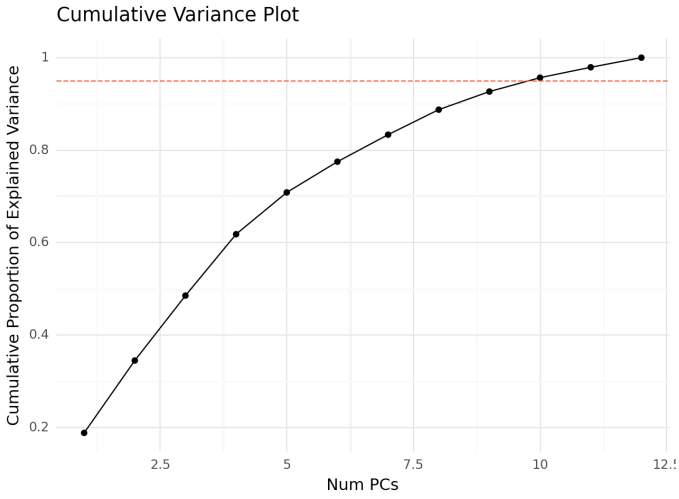
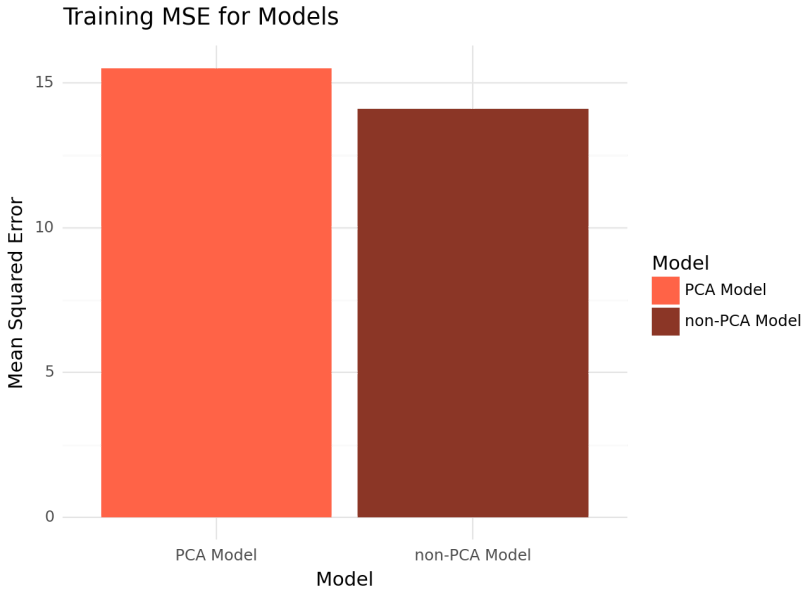


Figure 3
Relationship between number of principal components and explained variance. Dashed line at 95% variance

Based on this plot I will use 10 PCs for my principal component analysis which will allow me to cut down from the 13 variables in the non-PCA model and still maintain 95% variance.

Figure 4
Bar chart displaying Mean Squared Errors for PCA and non-PCA models using the training data.

The PCA model has a slightly higher Mean Squared Error than the non-PCA model.



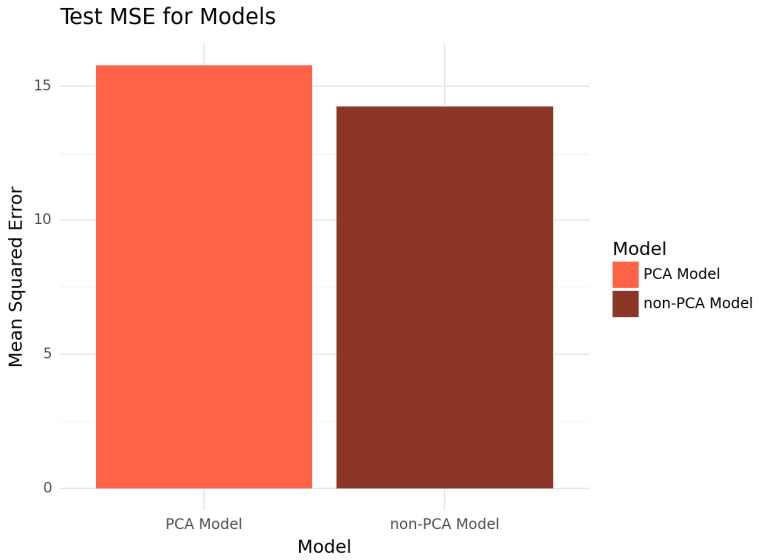


Figure 5
Bar chart displaying Mean Squared Errors for PCA and non-PCA models using the test data.

Once again the PCA model has a slightly higher Mean Squared Error than the non-PCA model.

Conclusions: The graphs both show us that in using PCA we are able to successfully introduce dimensionality reduction while only slightly increasing our MSE. Dimensionality Reduction means we are cutting down the number of variables we are using by essentially meshing them to make the model simpler, while still maintaining the majority of the information. Naturally this will result in an increased error, however in this case it is rather small. This could be used to streamline the process of predicting fighter height, in this case, but the same

concept can be used to optimize the process of predicting any of our continuous variables.

Question 3

When considering height, weight, and reach, what clusters emerge and what characterizes those clusters?

Data Cleaning: Null Values Dropped.

Modeling/Computation: Cluster analysis was performed using height, weight, and reach to identify distinct groups of fighters. Continuous variables were z-scored. The k-means clustering algorithm was used for splitting the data into clusters, with the number of clusters being determined using the elbow method. Then these clusters were compared based on how they performed on other metrics not included in the clustering such as win rate.

Graphs:

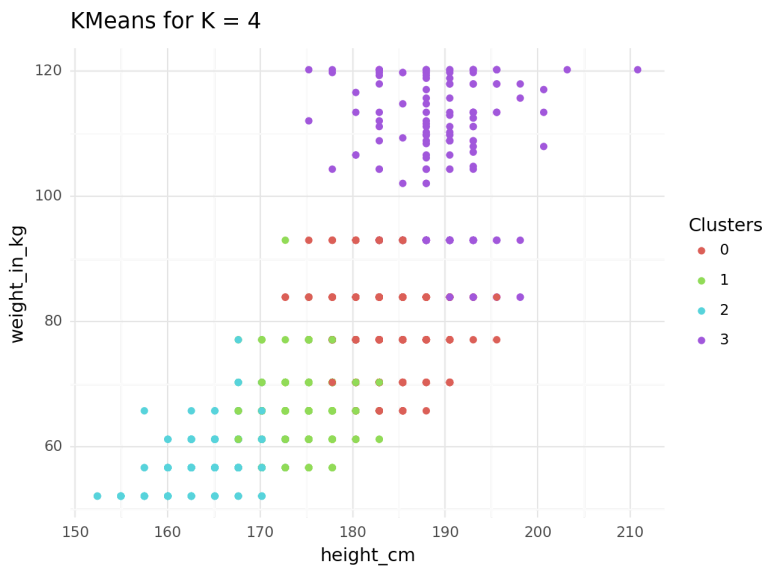


Figure 6
Scatter Plot showing relationship between height and weight. Color-coded based on clusters

We can see that clusters are pretty distinct from one another as expected (taller fighters weigh more). Note how the data below 100 kg is organized neatly into what appears like rows. This is due to the weight-class regulations used in MMA. Cluster 3 is staggered in the way that it is because it represents the heavyweight division which is anywhere below 265 lbs and above 205 lbs. Other

divisions are much more strict such as lightweight 155 lbs (+- 1 lb) or welterweight 170 lbs (+- 1 lb) resulting in the row like structures below the 100kg (220lbs) threshold.

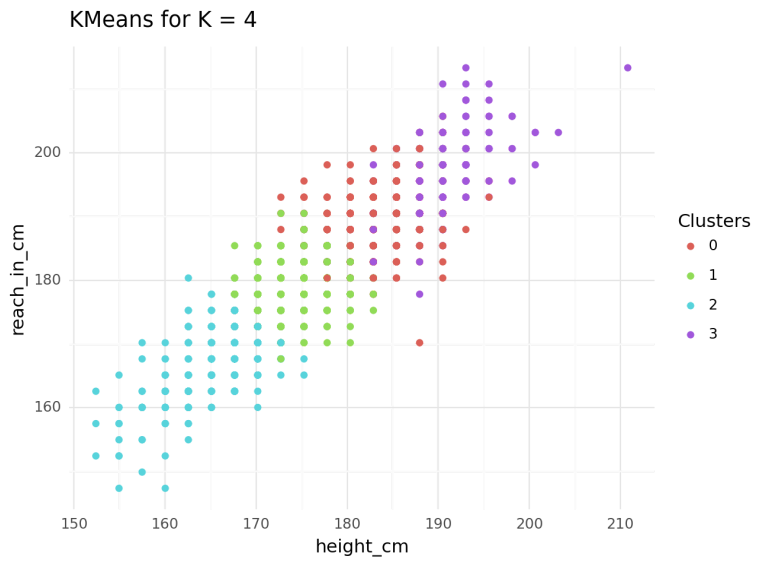
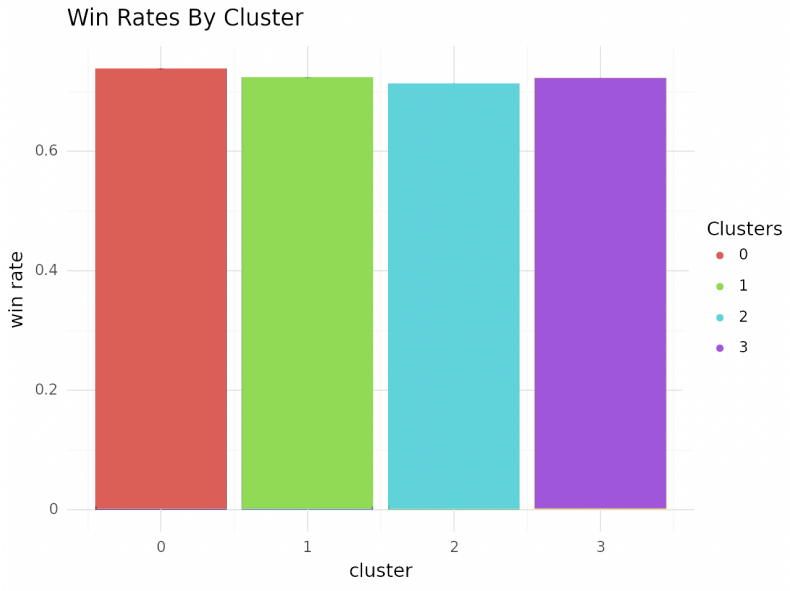


Figure 7
Scatter Plot showing relationship between height and reach. Color-coded based on clusters

Clusters once again are pretty distinct from one-another which makes sense because there is a pretty strong relationship between height and reach as well as weight (weight not pictured in this graph).

Figure 8
Bar chart displaying each clusters' average win rate

As we can see clusters are performing pretty similarly in terms of their average win rate with cluster 2 doing slightly worse than the rest and cluster 0 doing slightly better



Conclusions: Based on Figure 7 one may assume that fighters in cluster 0 are better than those in cluster 2 due to higher win rates. However, because fighters

in respective clusters are fighting other fighters within the same cluster, a lower win rate in said cluster might suggest that the weight division associated with that cluster is simply more competitive. That is also an assertion that cannot be proven with just the information provided above. Again, we would need to do a much deeper analysis where we look into each weight division separately. By doing that and applying the same clustering methods, we can potentially find the ideal height and reach of a fighter for each weight class based on win rate.