UFC Analysis using Bayesian Networks

Norberto Casarin
Master's Degree in Artificial Intelligence, University of Bologna
norberto.casarin@studio.unibo.it

March 11, 2024

Abstract

This project explores the use of Bayesian Network and Decision Tree models to predict the win rate of UFC fighters based on various statistics. The Bayesian Network model outperformed the Decision Tree, providing more accurate predictions.
Through data discretization and inference techniques, the model was able to assess the likelihood of success for fighters in specific matchups. Results showed a significant increase in performance with the Bayesian Network, highlighting the potential for reliable win rate predictions in the world of mixed martial arts. Future testing and refinement are necessary to further validate the model's effectiveness.

Introduction

Domain

In the world of mixed martial arts, accurately predicting the success of a UFC fighter is a constant challenge for analysts, fans, and bettors alike. With numerous factors influencing the outcome of a fight, from physical capabilities to strategic tactics, creating a reliable model to predict a fighter's win rate can provide invaluable insights into their performance potential.
This work sparks from a personal interest in mixed martial arts, as well as being inspired by the dataset used and the work of past students*.
The UFC most important statistics, always present even in betting sites, are the weight class, the height and the fighting stance (which derives from the sparring style and the main martial art used by the fighter).
Although these aren't the only factors that contributes to the match end result, they still play a major role, being the perfect candidates for a prediction model.

Aim

The purpose of this project is to utilize a Bayesian Network model to analyze a dataset containing various statistics of UFC fighters, as well as comparing the results with a Decision Tree classification.
By applying the Bayesian model, we aim to predict the win rate of new fighters based on their individual attributes and performance metrics.
Through the use of inference techniques, we aim to effectively assess the likelihood of success for a fighter in specific matchups and improve the overall accuracy of win rate predictions.

Method

- We utilized the pandas library to discretize every column in the dataset that was used by our model. This process involved converting continuous variables into categorical variables, which is crucial for certain machine learning algorithms like Decision Trees.
- We employed the sklearn library to split the dataset into training and testing sets. We trained our Decision Tree model on the training data and evaluated its performance on the testing data.
- We utilized the pgmpy library to build a Bayesian Network model. Once our Bayesian Network model was constructed, we tested it on the test set of the dataset to evaluate its accuracy in predicting outcomes.
- We used the Bayesian Network model to infer new prediction data. This involved making predictions based on the probabilistic relationships learned from the data, and generating insights on potential outcomes.
  This inference process was tested through three methods, and a final user friendly demo was created.

Results

There's a noticeable increase in performance between the Decision Tree and the Bayesian Network, with very high scoring metrics.
With a fully discrete model, it is entirely possible to have reliable predictions over the performance of a specific fighter, even thought modeling such a problem with full discretization might not be the best solution.
Using such approach was also a great way to practise Machine Learning techniques and smart solutions, as well as discovering new ways to approach a probability problem using the pgmpy library.

Model

In this section we'll different aspects of the model.
Starting with the nodes, which are:

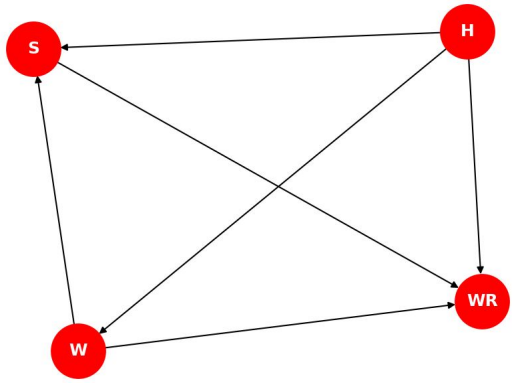

Figure 1: Bayesian network. Image taken from the code notebook

- H: height expressed in centimeters ranges
- W: weight expressed in weight class
- S: fighting stance
- WR: Winrate, in ranges between ”Very Low” and ”Very High”

In this Bayesian Network, the conditional probability relationships are as follows:

- The weight of a player depends on their height.
- The stance of a player depends on their height and weight.
- The winrate of a player depends on their height, weight, and stance

Since the purpose of this project was to not have prior assumptions about the importance of each node in regards to the winrate, the CPT tables have been set to avoid specific beliefs about the distributions of the variables, prioritizing a state where all classes are equi-significant.
For this purpose, a specific Bayesian Estimator has been used to set the conditional probabilities.
Most of the changes applied to the initial database have been made with regards to a previous knowledge about the sport, particularly in the discretization process for the height, weight and stance classes, as well as the transformation of winrate from a percentage to a range of values.

Analysis

Experimental setup

After creating the Bayesian network, we can predict the win rate of a new fighter by estimating probability distributions by utilizing various inference methods. We delve into both approximate and exact inference methods to compare the results. Specifically the ApproxInference, which is expected to give results with a higher variance (in spite of consistency), as well as BeliefPropagation and VariableElimination, which we expect to have very similar, if not the same results.
Additionally, metrics scores for accuracy have been used for the final evaluation of the Decision Tree, while custom scoring functions from the pgmpy library have been used for the evaluation of the Bayesian Network

Results

Given the full discretization approach over the data, as expected the performance on a Decision Tree model was poor, with an accuracy a little over 50%.
When data has been overly discretized, critical information may be lost during the process, leading to less accurate and potentially misleading results when using a Decision Tree model. Additionally, the decision tree may struggle to effectively distinguish between small differences within categories, resulting in sub-optimal predictions.
Instead, this approach lead to a very good performance for the Bayesian Network, with a surprising perfect score on correlation and very good score on log likelihood.
Although the inference hasn't been tested, for lack of new data, the results are reasonable and on the same level of the dataset entries, which makes them reliable enough to gain some insight over their usage, although more testing is needed.

Conclusion

In conclusion, this project successfully utilized a Bayesian Network model to predict the win rate of UFC fighters based on their individual attributes and performance metrics. That being said, much more testing is needed to fully asses the completeness of the results. Finally, a big improvement over the whole application can be found in enlarging the database, by adding more parameters to be added to the Bayesian Network (ex: reach, daily calorie intake, strength benchmarks etc.etc.).

Links to external resources

Link to the dataset

References

*

- Notebook of inspiration for the pre-processing part
- Githubs list of inspiration for the Bayesian Network usage
- Site used to fact check the UFC stats