Predicting UFC Bouts

Purdue University — Krannert School of Management

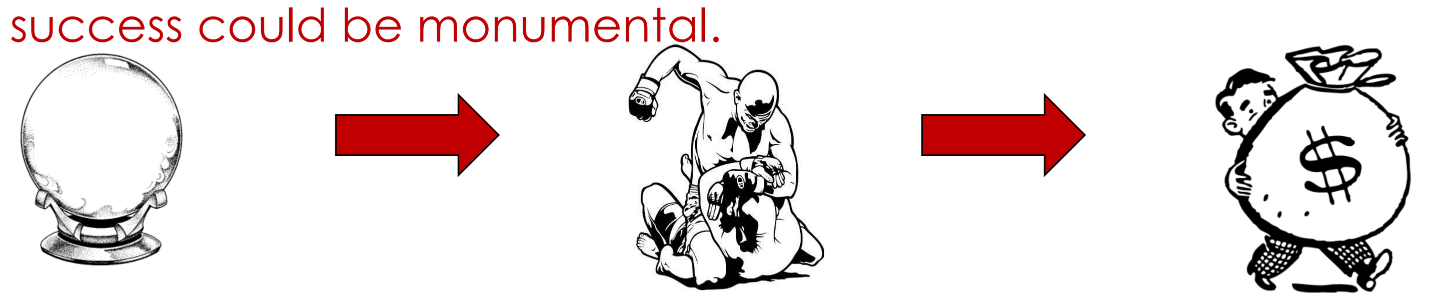
Lucas Booras, Gianni Frasca, David Hamilton, Matthew McTighe, Cameron Schitter, Matthew A. Lanham
Purdue University, Krannert School of Management
lbooras@purdue.edu; gfrasca@purdue.edu; hamil114@purdue.edu; mmctigh@purdue.edu; cschitte@purdue.edu; lanhamm@purdue.edu

ABSTRACT

The purpose of our study is to develop a classification model to predict the winner of any UFC bout using historical data on previous matchups. In developing this model, we hope to profit via placing wagers with sports books. By having a statistically effective predictive model, we will be able to generate positive expected value over time. By making numerous wagers with positive expected values, profit is mathematically guaranteed over infinite trials with unlimited capital.

INTRODUCTION

Due to lifting of federal regulations across the United States, the sports gambling industry has skyrocketed in popularity. A key component of the success to this industry is the Ultimate Fighting Championship, or UFC. As things currently stand, only 23 states have some form of legalized sports gambling. The industry is projected to be worth over $155 billion and will grow about 9% annually until 2024. This makes investments in the industry, like a model to predict winners on fighting bets, a worthwhile proposition and success could be monumental.



MODEL BUILDING

Our model is based off a classification-type problem and is created using binary logistic regression method. Our problem necessitates high accuracy as the model decides between two fighters in each matchup. We used accuracy to evaluate our predictive model.

Research Questions:

- What key drivers influence the outcome of fights and how accurately can these drivers predict the fight?
- Can a model combined with active bankroll management be used to place profitable wagers in the long run?

LITERATURE REVIEW

Study	Similarities to Our Model	Novelty of Our Model
Johnson (2012)	Utilized a logistic regression model and cross-validation	Used K-fold validation
Lessman, et al. (2010)	Random forest classifiers to generate decision trees for classification	Used RF classifiers as a part of stacked ensemble model
McQuaide (2019)	General Linearized Model, K-fold validation, training data set	Used GLM as a part of stacked ensemble model
Miljković, et al. (2010)	Regression model using K-fold validation	Differnt k values; used stakced ensemble, deep learning and logistic models

METHODOLOGY

We need to predict UFC fights at a high accuracy to keep a profitable average over the long run. To do this we focused on three goals 1) Pre-process the data 2) Pick strong predictor variables 3) Train multiple models and compare. With these steps we were able to create a data frame with our desired variables and apply it to h2o – a software that will automatically train your model based on your preferences. Through h2o we created two models classify and provide probabilities on which fighter will win.

1 **Data Pre-Processing**

Master Dataset – Columns that are not relevant – Imputation & Omission – One-Hot Encodings = A dataset that is ready to be analyzed!

2 **Variable Selection**

All predictor variables + Feature removal based on zero variance & high collinearity + Backwards Selection = Robust dataset with strong predictors!

3 **Cross Validation and Model Selection**

Data Partition (Holdout Method) 70% Train 30% Test

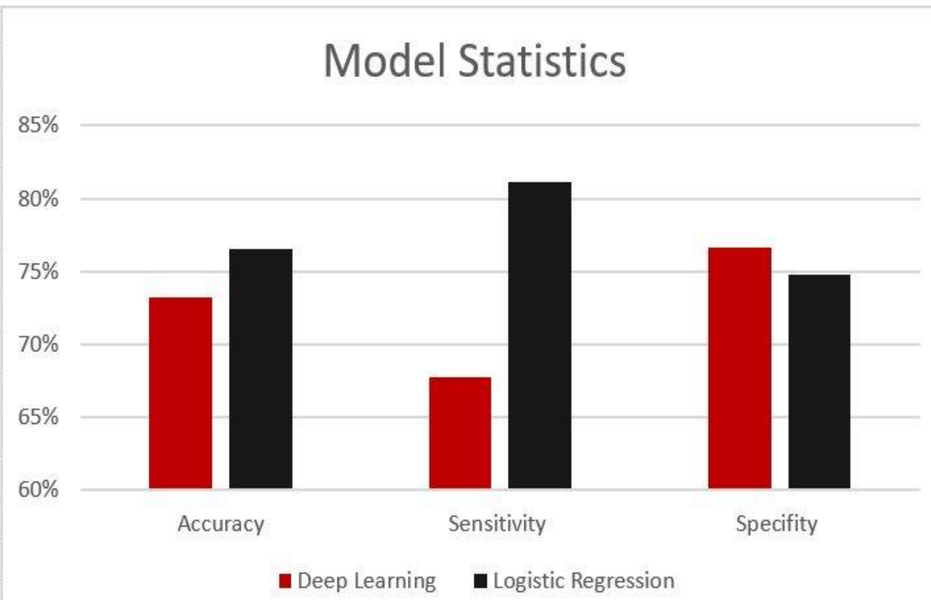
Utilized h2o software to create deep learning and logit model

H2O

Result: Logit Model with 76.05% Accuracy on test set

STATISTICAL RESULTS

Since our Logistic Regression Model displays higher accuracy and sensitivity than our Deep Learning Model, we can assume that this will be our optimal model for predicting UFC bouts. We are confident that the Logistic Regression Model will give us a higher predictive accuracy and increase the expected value for profit.



EXPECTED BUSINESS IMPACT

In the world of algorithmic sports betting, positive expected value (EV) is the name of the game. Just having a winning model is not enough. You must also identify profitable wagers by assessing the odds set by Vegas sportsbooks. To determine whether a wager may be profitable or not we can use the following equation to calculate expected value:

E = p(profit) + (1-p) * (loss)

Where p = probability of a win and (1-p) = probability of a loss. We derive profit and loss from given sportsbook odds. Since expectations are additive, we can create a chain of separate expectations representing various wagers adding up to one final expected value. If our result is positive, we should make the wager.

E(x1+x2+...xn) = E(x1) + E(x2) + E(xn)

Normal model			Our model			bet unit:
prob win	70.00%		prob win	76.05%		$ 100.00
prob loss	30%		prob loss	24.0%		
win amount	40%	40	win amount	40%	40	
loss amount	-100%	-100	loss amount	-100%	-100	
Expected Value	-2.00		Expected Value	6.47		

With a win percentage of 76.05% vs an average win percentage of 70% in similar models, our model makes the difference between winning money or losing money in the long run. Assumptions for the above table include a standard betting unit of $100 with our loss amount set over 2x that of our win amount (typical when betting on a favored fighter). With the given inputs, our model would mathematically guarantee a positive expected return over time while betting on fights with similar payouts. Compared to the model that is 6.05% less accurate, this model would return a negative expected value and mathematically guarantee to lose the user 100% of his or her bankroll in the long run.

CONCLUSIONS

Some of the key drivers in our model were signature strike percentage, total strikes landed, a fighter's reach, the number of knockouts, and many more.

By trusting the laws of probability and mathematics the union of a highly accurate model and an expected value calculator will yield the user of these systems positive returns over time. Placing wagers on fights with positive expected values with a 76.05% degree of accuracy will turn a profit in the long run given the user can sustain statistical swings (losing streaks).

ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham and our TAs Xinyu Wang and Theo Ginting for their guidance and support on this project.