**Background**

The aim of this project is to develop a machine-learning based approach to predict the outcome of a mixed martial arts (MMA) bout using retrospective data. The sports gambling industry is an international juggernaut, with an estimated total value of approximately $100bn (“Global Sports Betting Market Size 12. 4% Annual Growth,” n.d.). Effectively forecasting the outcomes of sporting events is of considerable interest to gaming stakeholders, as doing so allows them to set an enticing odds line and manage their risk effectively. Of course, such a model would also be of interest to gamblers, especially to the extent that its predictions differ from lines set by sportsbooks (Holmes et al., 2022). Regardless, a predictive model for MMA bouts allows interest parties to effectively engage in the MMA gambling markets.

Mixed martial arts is a combat sport, comparable to boxing or kickboxing, albeit with a relatively limited ruleset that notably allows for more involved grappling exchanges, as well as chokeholds and jointlocks. For details regarding the rules, regulations, and terminology of MMA, refer to the Associated Boxing Commission’s Unified Rules statement (*Unified Rules – Association of Boxing Commissions*, n.d.). Its viewership reach is estimated around 260 million, and MMA gambling is estimated to be 1% of total sport gambling revenue (Post, 2020).

**Description of Data**

These data are drawn from a publicly available dataset on Kaggle. They were compiled by (CITE). The complete data set spans from 1995-2020, and includes bouts contested within the Ultimate Fighting Championship™ (UFC), the premier international MMA promotion, both in terms of valuation (est. 4bn market capitalization), and talent roster (“Current MMA Rankings,” n.d.).

These data include 13322 observations of 530 variables. Only bouts contested since 2010 will be considered. This limits the data to a subset of 10834 observations. Variables are too numerous to exhaustively describe within the main text of this manuscript, but include sufficient data to identify the contest (names of both fighters, date), fighter descriptives (age, height, reach), the result of the contest, and numerous summary statistics of actions taken during the fight (strikes attempted and landed, takedowns attempted and landed, submissions attempted and landed), and similar statistics of each fighter from prior bouts.

As the goal of this endeavor is to accurately predict the outcome of the match using only a priori data, only variables using data accumulated prior to the match will be considered. This will likely limit the predictive accuracy of the model, as statistics of the match itself will be highly deterministic—though not completely so--of the result of the match (by analogy, including the score differential of a game such as basketball or soccer would be perfectly predictive of the outcome, but engender a useless predictive model).

Missingness was assessed to be minimal. Recent pre-competition data was the most missing at approximately 30%. In a sport with a poorly recorded regional competition scene and high turnover, such missingness is expected.

**Extant Variables**

The variables being considered will include all data retrospective at the time of the event. Within the dataset, these are variables with the prefix “precomp”, and include averages of attempted and completed strikes (of various type), takedowns, submissions, and ground control. Similar statistics of the fighters most recent bout (with the prefix “recent”) will also be considered. Age, reach, and the differentials of each between fighters will also be included.

**Bespoke Variables**

As well, several new variables have been generated from these data. Notably, these include a value based on a cluster analysis from the above variables, as well as a skill rating based on the elo rating method. It applies, in effect, a number that allows for ordering of players by skill.

The elo method is a well-understood methodology for rating relative skill in zero-sum games. Its use is perhaps most associated with chess, although it sees use in numerous other sports. By examining past wins and losses within a social network, a single numeric skill rating can be generated for a competitor that will allow for projecting the probability of victory or defeat. While elo ratings in other sports have been demonstrated to be highly predictive of outcome (Angelini et al., 2022; Hvattum & Arntzen, 2010), their application in MMA is more contentious, due to the nature of the sport, which is considered both more stochastic and based on the interaction of the stylistic approach of the respective competitors.

Best summarizing this concern is the common aphorism “styles make fights.” By generating a variable that includes style of both competitors, the concern should be addressed. To do so, a cluster analysis of each fighter was conducted, then the particular style by style matchup was encoded.

The cluster analysis revealed class membership by actions taken in prior fights. Visual inspection of plots for average silhouette and elbow methods of determining optimal clusters suggest 2-3 clusters. This analysis uses 3 clusters. However, a more theoretically derived approach would likely be a more optimal approach. Limiting the number of variables to those salient to determining “style” may improve validity. The distribution of clusters is shown below (Figure 1). A dummy variable indicating the style matchup (cluster membership for each combatant) was generated. Thus, there are a potential 9 variables to indicate matchup.

It is hypothesized that these two bespoke variables, as well as age differential, will be most salient.

**Description of the Models**

Three predictive models were fit to these data: a non-regularized logistic regression model, a ridge regression model, and a gradient boosting tree model. All analyses were conducted in R (R Core Team, 2022). Packages used in these analyses include here, eloRating, cluster, factoextra, tidyverse, finalfit, caret, and recipes (Greenwell et al., 2022; Harrison et al., 2022; Kassambara & Mundt, 2020; Kuhn, 2022; Kuhn & Wickham, 2022; Maechler et al., 2022; Neumann & Kulik, 2020; Thiele & Hirschfeld, 2021; Wickham et al., 2019).

**Non-regularized Logistic Model**

Logistic regression models were used since the outcome of a fight is, for all intents and purposes, a dichotomous variable (win vs loss). This model uses Maximum Likelihood estimation to estimate parameters for predictor variables corresponding to probability of classification. Logistic regression does not make many key assumptions associated with ordinary least squares algorithms: residuals do not need to be normally distributed, nor is homoscedasticity required. This model does not have any hyperparameters to optimize. Model performance will be assessed using area-under-curve (AUC), accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and precision (PRE).

**Ridge Regression Model**

This model will function in a similar manner to the previous model but is regularized to avoid overfitting our data. While n>>p, and we therefore expect minimal incremental improvement over the non-regularized logistic regression model, we will empirically test this hypothesis. While a lasso model would allow for feature selection, as noted above, we envision this as an exploratory work, with feature selection primarily an a priori process. Ridge regression, being more effective in handling correlated features, seems a preferable model, as many of these variables are expected to correlate (Boehmke & Greenwell, 2019). This model optimizes a single hyperparameter, lambda (λ), the penalty term for this model. Model performance will be assessed using area-under-curve (AUC), accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and precision (PRE).

**Gradient Boosting Tree Model**

A gradient boosting machine (GBM) is also tested, primarily due to their success in predictive competitions (Boehmke & Greenwell, 2019), and the role of this paper is to develop and compare successful predictive models. Gradient boosted machines iterate through decision tree models and improve on the previous model using that models residuals as feedback. Numerous hyper-parameters will be tested: number of trees (total trees in the sequence), learning rate (the contribution of each tree on the final outcome), tree depth (longest path from root to terminal node), and minimum number of observations in terminal nodes.

Model performance will be assessed using area-under-curve (AUC), accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and precision (PRE).

**Model Fits**

See table 1 for a comparison of model fit. As hypothesized, regularization does not significantly improve on the non-regularized model. However, the GBM approach does appear to have added incremental value despite the added computational complexity.

Data Viz

Cluster elbow charts

Class membership

Elo Ratings

Discussion

**References**

Angelini, G., Candila, V., & De Angelis, L. (2022). Weighted Elo rating for tennis match predictions. *European Journal of Operational Research*, *297*(1), 120–132. https://doi.org/10.1016/j.ejor.2021.04.011

Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. CRC Press.

Current MMA Rankings. (n.d.). In *Fight Matrix*. Retrieved December 5, 2022, from https://www.fightmatrix.com/mma-ranks/

Global sports betting market size 12. 4% annual growth. (n.d.). In *World Sports Network*. Retrieved December 1, 2022, from https://www.wsn.com/betting/global-sports-betting-market-annual-growth/

Greenwell, B., Boehmke, B., Cunningham, J., & Developers, G. B. M. (2022). *gbm: Generalized Boosted Regression Models*. https://CRAN.R-project.org/package=gbm

Harrison, E., Drake, T., & Ots, R. (2022). *finalfit: Quickly Create Elegant Regression Results Tables and Plots when Modelling*. https://CRAN.R-project.org/package=finalfit

Holmes, B., McHale, I. G., & Żychaluk, K. (2022). A Markov chain model for forecasting results of mixed martial arts contests. *International Journal of Forecasting*, S0169207022000073. https://doi.org/10.1016/j.ijforecast.2022.01.007

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460–470. https://doi.org/10.1016/j.ijforecast.2009.10.002

Kassambara, A., & Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. https://CRAN.R-project.org/package=factoextra

Kuhn, M. (2022). *caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret

Kuhn, M., & Wickham, H. (2022). *recipes: Preprocessing and Feature Engineering Steps for Modeling*. https://CRAN.R-project.org/package=recipes

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2022). *cluster: Cluster Analysis Basics and Extensions*. https://CRAN.R-project.org/package=cluster

Neumann, C., & Kulik, L. (2020). *EloRating: Animal Dominance Hierarchies by Elo Rating*. https://CRAN.R-project.org/package=EloRating

Post, S. / A. (2020). *MMA Online Gambling Industry*. https://mymmanews.com/mma-online-gambling-industry/

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Thiele, C., & Hirschfeld, G. (2021). cutpointr: Improved Estimation and Validation of Optimal Cutpoints in R. *Journal of Statistical Software*, *98*(11), 1–27. https://doi.org/10.18637/jss.v098.i11

*Unified Rules – Association of Boxing Commissions*. (n.d.). Retrieved December 1, 2022, from https://www.abcboxing.com/unified-rules/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686