**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.

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 (CITATIONS). Its viewership is estimated around, and MMA gambling is estimated at (CITATIONS).

**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 (fightmatrix). These data include.

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).

**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 (CITATIONS), 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.

The cluster analysis revealed class membership by actions taken in prior fights. There is a

By generating a variable that includes class membership of both competitors, this latter concern should be addressed. 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**

Text

**Ridge Regression Model**

Text

**Gradient Boosting Tree Model**

Text

Model Fit

Data Viz

Cluster elbow charts

Class membership

Elo Ratings

Discussion