Project Directions: https://www.seas.upenn.edu/~cis520/lectures/final_project.pdf

Team name: The Winning Team

Group members: Matthew Kligerman and Natalie Maus

Motivation:
We are interested in developing a machine learning model capable of predicting the outcome of boxing matches. Gambling on the outcomes of boxing matches is a highly lucrative market and predicting the winner based on large datasets containing vast features and samples beyond the trivial limits of human calculation is both interesting and profitable.

Data set:
BoxRec

Our dataset is all of the professional boxing matches from BoxRec.com. BoxRec contains every sanctioned professional boxing match in the past few decades and the majority of the professional boxing matches beforehand. Furthermore, it contains the outcomes of each match and the weights of the fighters in the matches, along with the profiles for the fighters with age, height, reach, weight class, etc.

Related Work:
The Sweet Data Science. Applying Data Science to the Sweet… | by Stephen Plainte | Medium

The article above, written by Stephen Plainte, details his work on training a model to predict the outcome of an arbitrary boxing match. To start, Stephen used a simple random forest classifier to predict the outcome of boxing matches. He then computed the feature importance of the classifier and found that the number of years that the fighter had been active, their height, and their reach, ended up being the three most important factors that their model considered. Stephen found that more experienced fighters tend to win, and that when fighters have equal experience, taller fighters with longer reach tend to win. Finally, Stephen removed the least important features and trained a XGBoost Classifier. His final classifier was able to predict the outcome of fights with ~ 68.73 percent accuracy.

We would like to build off this work and train a similar model to predict the outcome of boxing matches. We want to try using different variations of Random Forest and XGBoot Classifiers, as well as other types of models to see which is actually best for this task. We are also curious to see if we find that similar features are most important in predicting fight outcomes (years of experience, height, reach, etc.).

Problem Formulation:
We will use the boxers fighting one another and their respective metadata as features in our model to predict the labels which are who won the fight or if it was a draw or a no contest (but not necessarily how this outcome occurred). We may also consider the judges, promoters,

managers, and/or officials and historical outcomes of fights involving them if we want to get extremely high accuracy. Although features such as height, reach, and age may be useful, it is important that we use these features in the context of historical data regarding the past fights of the fighters, our most important feature. This will be a graph-like model in which we have fighters as nodes in our model and the edges as the fights between them as paths to traverse in training our model.

We may traverse the entire BoxRec database, but depending on computational resources and time constraints, we may instead attempt to train the model using various different depths to traverse. Furthermore, we may consider using the dates of the fights as weights in our model, considering changes to boxing over time in terms of glove size and padding, number of rounds, officiating, and societal factors.

Methods:

Imputation: Linear Regression, KNN

Dimensionality Reduction: PCA

Classifiers: Random Forest/XGBoost, Decision Tree/Ensemble of Models, K-Nearest Neighbor, Recurrent Neural Network

Evaluation:
In evaluating our model, we have a few different benchmarks that we can compare against. First we can compare the accuracy of our model in predicting the winner with the accuracy of predicting the winner based solely on their rankings in BoxRec (assuming that the higher ranking fighter will win). Alternatively, we can compare our predictions with the predictions in the betting odds of the betting site(s); in this case, we could use the odds as weights in calculating a net monetary performance.

Project plan: Provide a rough timeline of your project work schedule, including which team members are responsible for what portions of the project.

- When to get data - by November 26th
- Final final report structure - due November 30th
- Data Exploration - finish initial exploration by november 30th
- Dimensionality reduction - Initial dimensionality reduction by December 2nd
- Baseline models from related literature - December 3rd
- Try out first new model - December 6th
- Try second new model - December 9th
- If we have time, try more models (no deadline)
- Final due date: December 10th 11:59PM