Can AI Beat Vegas?

Student: Michael Wynn

Research Advisor: Dr. Alex Rudniy

Abstract: This project aimed to construct a machine learning model that predicts the outcome of NBA games, and wagers that may be placed on games though online sportsbooks. To achieve this, a myriad of game [2, 4] and sports betting [3] data was collected, processed, and engineered to become acceptable input for classification algorithms [1]. The original thesis was that the inclusion of gambling point spreads, due to their recent abundance in digital form and constantly updating nature, would be helpful in creating a predictive model for game outcomes. This quickly failed, and the objective shifted to optimizing predictive models. The models, or classifiers, were used to predict two things: whether the home team will win the game, and whether they will beat the spread. Classifiers generated were tested on holdout data, then evaluated using accuracy, precision, and recall. Many models outperformed Vegas when predicting outcome outright; the best of which scored an accuracy of 73% for win predictions. However, models performed worse when predicting gambling outcomes. The best model for spread predictions scored 57% accuracy, which is better than it sounds, given that Vegas intentionally sets point spreads to encourage equal distribution of bets.

Keywords: Machine learning, artificial intelligence, neural networks, NBA game outcomes, gambling point spreads.

Definitions

- Artificial Intelligence - the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.
- Machine Learning - the use and development of computer systems able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.
- Neural Network - Artificial neural networks, usually simply called neural networks, are computing systems vaguely inspired by the biological neural networks that constitute animal brains.
- Point Spreads - a forecast of the number of points by which a stronger team is expected to defeat a weaker one, used for betting purposes.

Methods

General

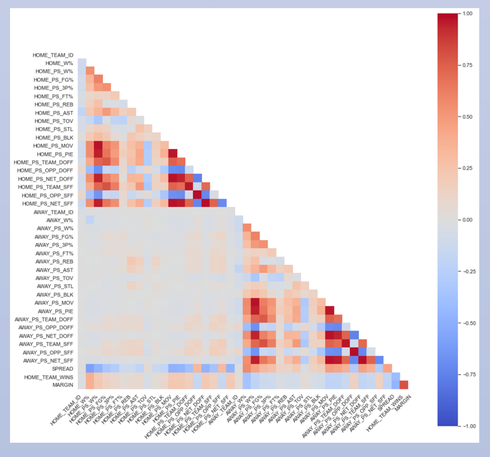
- This project is formatted as a series of experiments, each with a variable dataset or machine learning model. The goal of each experiment is to maximize accuracy of predictions.

Datasets [2, 3, 4]

- The dataset used in machine learning is very important to the results; a famous data science saying goes, “Garbage in, garbage out.” Therefore, creation and analysis of datasets is essential to the success of a predictive model.
- This project employs a correlation matrix, as well as the Feature Importance tool in Scikit-Learn’s Decision Tree classifier, to identify which attributes are valuable in determining the outcome of games.
- 4 total datasets were created. 2 were based on a semi-arbitrary separation of attributes into ‘basic’ and ‘advanced’ categories, both containing the classification target *wins*. 2 kept only the most important features, one having target *wins*, and the other having target *beat_spread*.

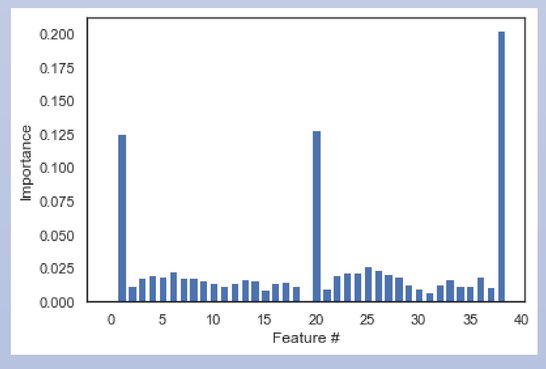
Correlation Matrix

- Importance of values is determined by how strongly they correlate, either positively or negatively, with the classification targets (bottom 2 rows)



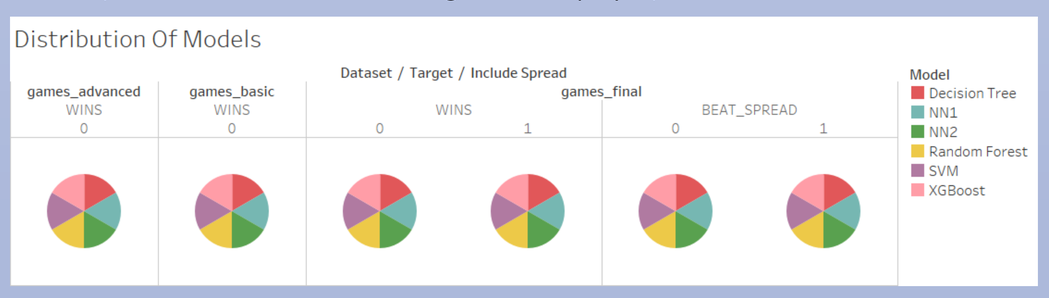
Feature Importance

- Importance of values is determined by how effectively they separate test data in a decision tree, performed automatically via Scikit-Learn
- Most important features, numbered 1, 20, and 38 here, are the home team’s winning percentage, away team’s winning percentage, and gambling point spread, respectively



Models [1]

- 36 total models were created: 24 to predict wins, 12 to predict gambling results
- Models are classified by the data used and classification target they predict, both described above, as well as the classification algorithm employed, described below.



Decision Tree - In this algorithm, the sample (population) is split into two or more sub-population sets, decided by the most significant splitter or differentiator in the input variables. The training process resembles a flow chart, with each internal (non-leaf) node a test of an attribute, each branch is the outcome of that test, and each leaf (terminal) node contains a class label. The uppermost node in the tree is called the root node.

Random Forest – An ensemble algorithm that combines purposefully dissimilar decision trees and determines results via a voting mechanism.

SVM (Support Vector Machine) – This algorithm attempts to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

XGBoost (Extreme Gradient Boost) – An ensemble algorithm that combines gradient boosted decision trees. During training, features that produce good results are strengthened, or *boosted.* This algorithm is called extreme because it maximizes computer resources.

NN1 (Neural Network 1)

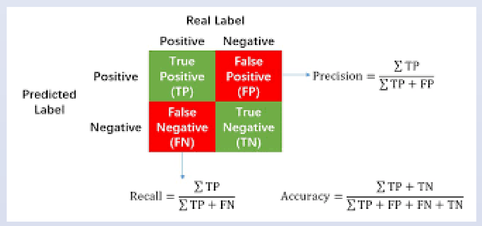
- 3 layers: 2 hidden, 1 output
- Node count/layer: 44, 22, 1
- Activation functions/layer: relu, relu, sigmoid
- Optimizer – Stochastic Gradient Descent
- Loss Function – Binary Cross-Entropy
- Trained with a batch size of 32 over 500 epochs

NN2 (Neural Network 2)

- 4 layers: 3 hidden, 1 output
- Node count/layer: 64, 256, 256, 1
- Activation functions/layer: relu, relu, relu, sigmoid
- L2 regularization – 0.01 per layer
- Dropout – 0.3 per layer
- Optimizer – Stochastic Gradient Descent
- Loss Function – Binary Cross-Entropy
- Trained with a batch size of 32 over 100 epochs

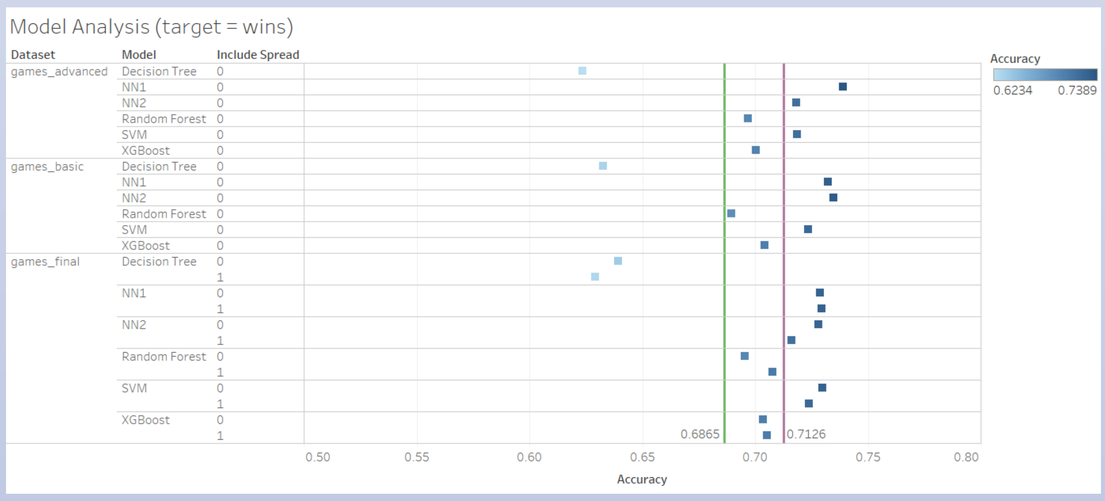
Evaluation

- Classification models are evaluated with metrics that measure the percentage of predictions that are right vs wrong. A confusion matrix holds the number of true positive and negatives (correct predictions), as well as false positives and negatives (incorrect predictions).
- Metrics such as accuracy, precision, and recall are calculated directly from the confusion matrix.

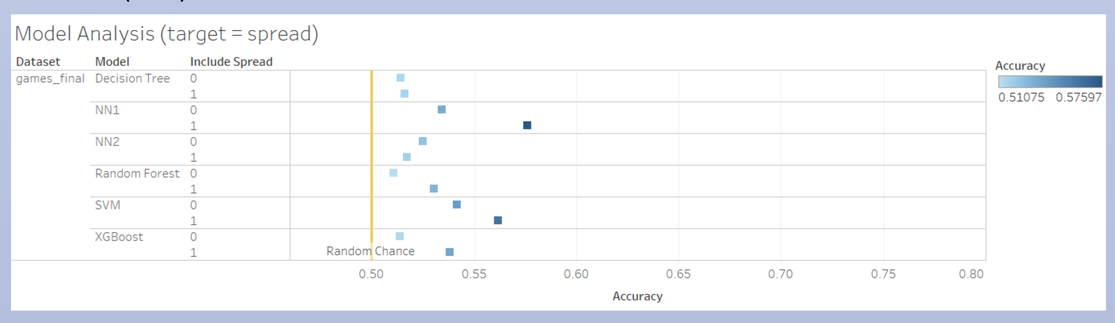


Results

- My original thesis, that gambling point spreads would help improve win prediction accuracy, was false. 50% of models supplemented with spread data performed worse. However, spread data was helpful in gambling outcome prediction, improving 83% of those models.
- To frame the effectiveness of win prediction models, 2 benchmarks were created representing the accuracy of predictions made with very simple methods. The first benchmark made predictions based on the assumption that the team with a higher winning percentage always won – called ChooseHigherW% (71.2%). The second benchmark assumed that the team Vegas favored always won – called ChooseSpreadFavorite (68.6%).



- To evaluate the effectiveness of gambling result prediction models, a much simpler metric is used. Oddsmakers purposely set the spread so that betting evens out on either side, and the house profits regardless of game result. Therefore, any model that can outperform random chance (50%) is notable.



Conclusion

- After examining results, the next step is to choose the dataset and the model with the best performance, then continue adding more data, hopefully improving accuracy.

Performance of datasets was generally similar –with games_final holding a slight lead.

For each dataset, NN1, NN2, and SVM consistently outperform other models.

In the model comparison, it is clear that NN1 outperforms all other models. The average performance is nearly 1% better than NN2 and SVM.

- In the future, the games_final dataset paired with neural network 1, which performed the best overall, will continue to be improved with additional data.

References

[1] Brownlee, Jason. *Machine Learning Mastery*, 19 Sept. 2019, www.machinelearningmastery.com/.

[2] Lauga, N. (2019, December). NBA games data, Version 5. Retrieved February 18, 2021 from www.kaggle.com/nathanlauga/nba-games.

[3] Qui, E. (2020, April). NBA Odds and Scores, Version 3. Retrieved February 18, 2021 from www.kaggle.com/erichqiu/nba-odds-and-scores.

[4] Sullivan, Ryan. “NBA's Most Valuable Statistic Discovered: How To Predict Team Wins With 95% Accuracy.” *Sports Gambling Podcast*, 20 Apr. 2020, www.sportsgamblingpodcast.com/2020/04/20/nba-most-valuable-statistic/.