Abstract and Problem Statement

The Elo rating system is a way to quantify the skill level of teams in adversarial games. First introduced for the game of chess[1] by Arpad Elo, it has been applied to a large variety of games since its inception. FiveThirtyEight has published multiple articles applying Elo ratings system to sports leagues such as the NFL[2], MLB, and NBA[3], the latter of which will be the focus of this report. The main objectives of this project will be to replicate the results obtained by the FiveThirtyEight NBA Elo rating model and then improve upon those results by building machine learning models to better predict the game outcomes.

Related Work

While FiveThirtyEight describes their methods, they are not particularly transparent about the specific math used for the NBA Elo rating model[4], nor do they provide the code to precisely replicate the results they obtained. This made it challenging to perfectly match their outcomes since there were model decisions left out of the article where the methods were described. They did, however, include their Elo ratings as a part of the data set. It was thus possible to treat their Elo rating as a sort of label to compare other models against.

With the help of external resources[5] aimed at deciphering the math under the hood of the FiveThirtyEight model, I was able to accurately reproduce their results. The math behind the Elo rating model looks as follows.

$$\mathrm{ELO}_{i+1} = \mathrm{ELO}_i + \mathrm{K} * (\mathrm{S} - \mathrm{E})$$

$$\mathrm{K} = 20 * \frac{(|\mathrm{Pts}_{home} - \mathrm{Pt}_{away}| + 3)^{0.8}}{7.5 + 0.006 * |\mathrm{Elo}_{home} + 100 - \mathrm{Elo}_{away}|}$$

$$\mathrm{S} = \{1 \text{ if the team won, } 0 \text{ otherwise}\}$$

$$\mathrm{E} = \frac{1}{1 + 10^{\frac{\mathrm{Elo}_{away} - (\mathrm{Elo}_{home} + 100)}{400}}}$$

As the equations show, the Elo calculation is fairly basic. At its core, the model updates a given team's Elo rating based on how real game outcomes compare to expected game outcomes. With enough games played it is reasonable to expect Elo to converge on a value that accurately represents a team's comparative strength against other teams. The key feature of FiveThirtyEight's model is how it accounts for margin of victory, which is with the term *K*. This multiplier is used to scale the amount of Elo rating points a team is adjusted so it is proportional to the number of points that team won or lost each game by.

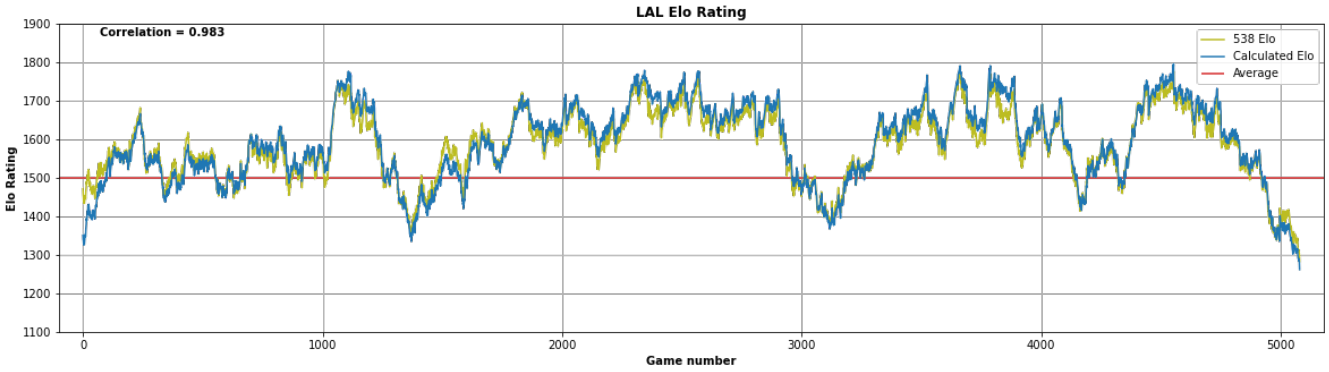

Figure 1: Los Angeles Lakers Elo rating model

As *Figure 1* shows, the model created very closely resembles the FiveThirtyEight model. See the included project notebook to explore how the two models compare for any past or present NBA team. The Pearson correlation coefficients were also calculated for each team as a way to quantify the model similarity. The inter-model correlation is displayed in the upper left corner of the team Elo comparison plot.

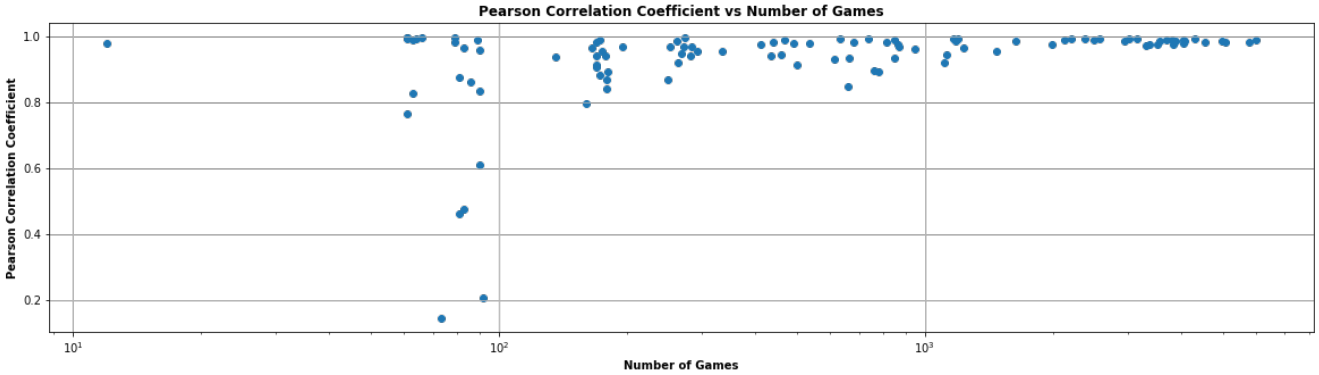

Figure 2: Relationship between number of games played and correlation coefficient

An interesting note is the relationship between the number of games a team played and the Pearson correlation coefficient between the models. As *Figure 2* shows, when a team has less than one hundred games played the correlation between the two model tends to be low and variable. With more than one hundred games played every team has a correlation coefficient around or above 0.80. With more than a thousand games played every team has a correlation coefficient greater than 0.90. The overall inter-model correlation coefficient, weighted by the number of games each team has played, is 0.978. This is strong evidence that the Elo rating model created accurately reproduces the FiveThirtyEight Elo rating model.

Methods

The first attempt to create a more accurate model than the base FiveThirtyEight Elo rating model revolved around using hyperparameter optimization. The parameters of the margin of victory scaling factor *K*, described previously, were tuned in an attempt to create a model that would more accurately predict game outcomes. Since this optimization is computationally expensive, it was run on Michigan State University's High Performance Computing Center and took over eight hours to complete. As will be detailed in the Results section, the hyperparameter tuning alone did not significantly improve the accuracy of the model.

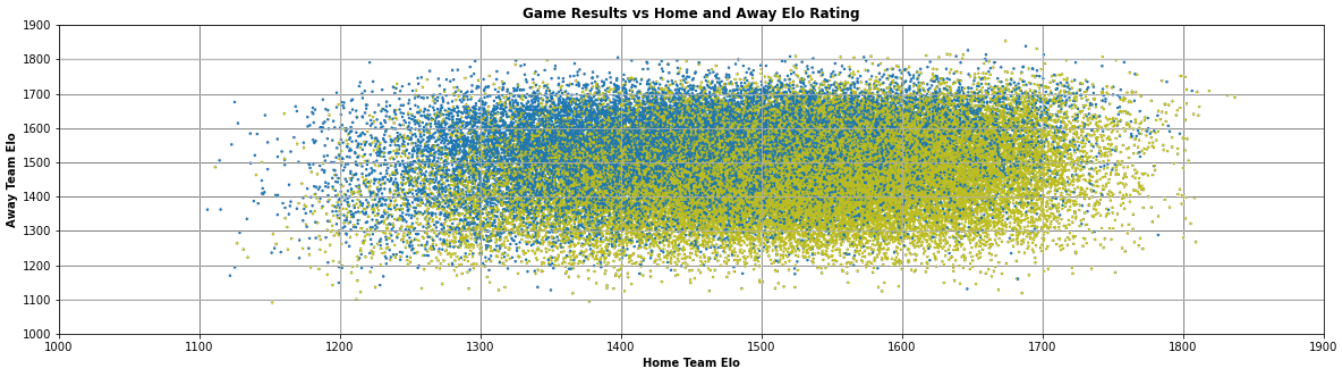

Figure 3: Games outcomes based on the Elo ratings

This is where additional methods were applied to predict game outcomes. The Python scikit-learn library[6] was used to construct these models. The data these models attempted to predict is partially visualized in *Figure 3* and contains the Elo rating of the home team, the Elo rating of the away team, the difference in Elo rating, and whether the game was a playoff game for each game in the original FiveThirtyEight data set. These features are admittedly quite limited as a result of the limited nature of the FiveThirtyEight data set.

Results

The provided FiveThirtyEight Elo rating model had an accuracy of 64.3%. The model that was created to replicate the FiveThirtyEight model had an accuracy of 64.2%. While both of these accuracies are above random chance, they are not particularly high performing. This is where additional analysis was used to create improved models.

The first attempt to do so was aimed at tuning the hyperparameters of the Elo rating equations to create a model that more accurately predicted game outcomes. The adapted equation for the best result of this hyperparameter optimization is shown below.

$$K = 20 * \frac{(|Pts_{home} - Pt_{away}| + 1)^{0.9}}{7.5 + 0.005 * |Elo_{home} + 100 - Elo_{away}|}$$

The tuned model resulted in different values for each of the parameters except 20 and 7.5. This model had a correct outcome prediction accuracy of 64.3%, an improvement over the previous accuracy but still no better than FiveThirtyEight's base Elo rating model.

This is where the data described previously was used to create various other machine learning models based on some additional game information and the rating values generated by the FiveThirtyEight Elo rating model. It was this additional information that allowed for more accurate outcome predictions than the standalone Elo rating model. The table below shows the training accuracies, testing accuracies, and complete data accuracies of various classification methods applied to the NBA game and Elo rating data.

Model	Train Accuracy	Test Accuracy	Overall Accuracy
Logistic Regression	0.682	0.678	0.681
KNN n = 1	1.000	0.593	0.898
KNN n = 5	0.752	0.633	0.722
KNN n = 10	0.716	0.645	0.698
KNN n = 20	0.698	0.663	0.689
LDA	0.682	0.679	0.681
QDA	0.590	0.598	0.592
Decision Tree	1.000	0.590	0.898
Random Forest	1.000	0.626	0.906
Gradient Boosted	0.688	0.678	0.686
SVC Linear	0.683	0.677	0.681
SVC RBF	0.678	0.675	0.677
K-Means Clustering	0.355	0.354	0.355

Of the wide variety of classifiers used to predict game outcomes it appears as though predictions generated by logistic regression, linear discriminant analysis, gradient boosted tree classification, and support vector classification with a linear kernel performed the best with test accuracies of 67.8%, 67.9%, 67.8%, and 67.7% respectfully. From the observed results, it appears as though the decision boundary is somewhat linear as the methods which produce linear decision boundaries seem to have performed particularly well. The outlier in this trend is the performance of the gradient boosted classifier, which has the capacity to produce nonlinear decision boundaries.

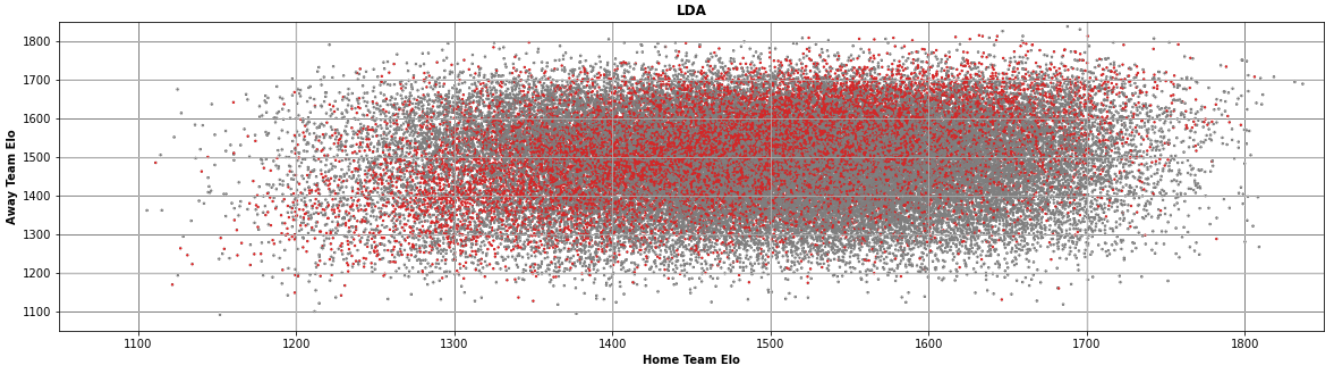

Figure 4: Linear discriminate analysis classification errors

To visualize the classification errors made by each of the methods, plots similar to *Figure 3* were created after model training and testing. In these plots, correct predictions are plotted as grey and incorrect predictions are plotted as red. *Figure* 4 is one such plot showing the classification errors made by linear discriminate analysis, the model found with the largest test accuracy. As one may expect, there were many correct predictions made when large Elo rating discrepancies existed between teams and many incorrect predictions made when the two teams were close in rating.

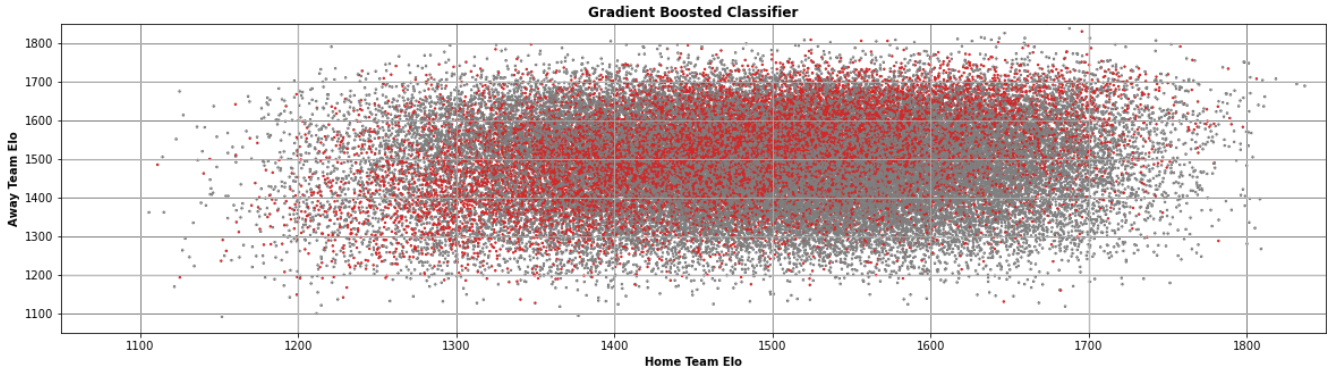

Figure 5: Gradient boosted classifier classification errors

The gradient boosted classifier, shown in *Figure* 5, displays results very similar to those obtained by linear discriminate analysis. In fact, all the high performing models, and particularly those that produce linear decision boundaries, look similar to *Figures 4* and *5.* The classification error plots were produced for each model and can be found in the corresponding project notebook.

Discussion and Conclusion

This project attempts to replicate the results obtained by FiveThirtyEight's NBA Elo rating model and expand on those results by creating a machine learning model which can accurately predict game outcomes based on relatively limited game information. Techniques including logistic regression, k-nearest neighbors, support vector machines, and decision trees were used as classifiers to make these predictions. Among the models tested, linear discriminate analysis was the most accurate, successfully capturing 10.1% of the remaining accuracy the base Elo rating model was unable to. This is a particularly impressive improvement as little supplementary information was included in the linear discriminate analysis model that was absent from the base Elo rating model.

There are many additional steps that could be taken to further study the NBA Elo rating model. These could range from simply applying additional modeling methods in an attempt to improve prediction accuracy to incorporating advanced analytics data into either the Elo rating or classification models. The performance of each team could also be studied in detail to see if any perform particularly when at home or away, and how performance changes for each team during the playoffs.

References

[1] Elo Rating System. *Chess.com*, https://www.chess.com/terms/elo-rating-chess .

[2] How Our NFL Predictions Work. (2018). *FiveThirtyEight*. https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/.

[3] Silver, N. The Complete History of the NBA. (2021). *FiveThirtyEight*, https://projects.fivethirtyeight.com/complete-history-of-the-nba/.

[4] Silver, N. How We Calculate NBA Elo Ratings. (2015). *FiveThirtyEight*, https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/ .

[5] Replicating Nate Silver's NBA Elo Algorithm. (2017). *Ergo Sum*, https://www.ergosum.co/nate-silvers-nba-elo-algorithm.

[6] *scikit-learn: Machine Learning in Python*, https://sklearn.org.