# Money is Motivation: A Random Forest Approach to the Game-Level Analysis of Salary Dispersion and Team Performance in the National Basketball Association

Matthew Senick
senickma/1008220676
University of Toronto
ECO1400, Martin Burda

# 1 Executive Summary

In this paper, game-level data is extensively scraped and analyzed for the National Basketball Association (NBA) to gather the predictive value of salary dispersion and scheduling of salary concerning single-game outcomes. Economic theory states contradicting opinions on how salary dispersion affects firm performance. On the one hand, tighter salary dispersion is thought to lead to higher within-team cohesiveness and thus better results.[1] On the other hand, highly dispersed salary is thought to motivate lesser-paid individuals and higher-paid individuals leading to better results.[2]

To establish predictive power, five measures of salary dispersion are considered simultaneously as input to three different Random Forest models. These inputted dispersion metrics include the maximum salary of a player for a team in a given game, average salary spent per minute of game time, dispersion of salary per minute of game time, dispersion of salary for active players on the roster, and dispersion of salary for all players on the roster. The predicted game-level outcomes for these three models were the ratio of home score over the away score, home score and away score, and win/loss by the home team respectively.

For all three models, it is found that salary dispersion tends to have little value in predicting the outcome of a single-game in the NBA. This result is consistent with other popular works exploring the effect of salary dispersion on game-level outcomes in the NBA.[3] However, the predictive importance for each considered input can be ranked. Overall, average salary per minute of game time is the most important predictor for game-level outcomes, while active and full roster dispersion is the least important. Considering previous research on the effect of salary dispersion on team performance, these results may not generalize across sports.

These results inspire multiple avenues of further research. Most interestingly, research has shown that using previous individual performance to predict single-game outcomes can be highly accurate.[4] As such, using salary dispersion to indirectly measure effects on team performance through differences in individual performance may be a more useful course of action. This falls in line with economic theory in that individuals should be paid at or above their market value and perform differently based on different levels of salary.

Extending generally, these results can be applied to firms in competitive markets. Essentially, NBA teams are used as a proxy for how competitive firms operate towards achieving market results. From the results, decision makers in competitive markets should not be concerned with the dispersion or scheduling of their wages or salaries as a predictor for competitive outcomes, but rather should be focused upon individual production as an accurate predictor of single period performance in the market.

1. David I Levine, "Cohesiveness, productivity, and wage dispersion," *Journal of Economic Behavior & Organization* 15, no. 2 (1991): 237–255.

2. Edward P Lazear and Sherwin Rosen, "Rank-order tournaments as optimum labor contracts," *Journal of political Economy* 89, no. 5 (1981): 841–864.

3. Hajime Katayama and Hudan Nuch, "A game-level analysis of salary dispersion and team performance in the National Basketball Association," *Applied Economics* 43, no. 10 (2011): 1193–1207.

4. Wei Gu et al., "A game-predicting expert system using big data and machine learning," *Expert Systems with Applications* 130 (2019): 293–305.

## 2 Introduction

Across competing firms in a capitalist society, distributional variation in salary is inevitable. Thus, the influence of salary dispersion on firm performance becomes an interesting economic question. Furthermore, the predictive capacity of salary dispersion on firm performance is important for forward-looking firms who are trying to maximize future market performance.

North American professional sports provide an excellent natural experiment for this phenomenon, as teams are given a fixed minimum and maximum payroll, with freedom as to how to distribute these funds in acquiring talent which will directly impact their competitive performance. As an example of these distributional differences, compare the case where one team spends lavishly on a "star" player, keeping salaries of other players on the team much lower, to the case where a team spends uniformly on all players. When considering North American professional sports, it can be seen that there is no consensus on whether teams prefer more to spread out or to condense salary distributions.[5]

Economic literature interestingly provides conflicting opinions on the preferable approach. On the one hand, it is argued that more uniform salary dispersion increases team cohesion, leading to improved performance.[6] But on the other hand, greater disparity might provide incentivization for both lower-paid players and higher-paid players to increase their productivity to earn more or prove their value respectively.[7] Therefore, empirical application appears to be encouraged for this topic.

Here, I extend the existing literature for salary dispersion's effect on team performance in the National Basketball Association (NBA). Specifically, regular season NBA games from the 2010-2011 season until the 2017-2018 season are considered, for which game-level statistics and player salary amounts were extensively scraped from online resources. From this collected data, various within-team salary disparity metrics were produced matching existing causal literature on the topic. These metrics included the maximum salary (*maxsal*) of a player participating for a team in a given game, the average salary per minute (*avesal*) for a team during a given game, variation in salary per minute (*disper1*) during a given game, salary variation in "active"[8] players on a team (*disper2*), and salary variation in all players on a team roster (*disper3*).

Models were then constructed to predict relative (home/away) score results, home and away team point totals respectively, and game-level win/loss results. Considering predictive capacity, it is natural to apply machine learning methods, and here Random Forest approaches to these three prediction tasks are explored. Previous literature focuses heavily on establishing causal relationships between these disparity metrics and team performance providing a mixture of conflicting results. This paper takes the opposite approach, attempting to find predictive power to these metrics.

The results of this research are useful to basketball team decision makers in a few ways. First

5. Katayama and Nuch, "A game-level analysis of salary dispersion and team performance in the National Basketball Association."

6. George A Akerlof and Janet L Yellen, "The fair wage-effort hypothesis and unemployment," *The Quarterly Journal of Economics* 105, no. 2 (1990): 255–283.

7. Fredrik Heyman, "Pay inequality and firm performance: evidence from matched employer–employee data," *Applied Economics* 37, no. 11 (2005): 1313–1327.

8. An active player is defined as one who plays in at least half of a teams games in a given season

off, basketball teams often have set "rotations" or schedules for players during their games. Thus, teams may be interested in how varying this schedule of play will affect the prediction of their team performance given other teams' schedules are approximately known or previously revealed. This is helpful for example, to give specific players rest or when injuries occur. Secondly, this is an important consideration for decision makers within the organization to decide on roster construction given salaries are approximately known. By way of example, will signing a very expensive player improve or hurt predicted performance? This sort of question could lead to optimizing behaviour from the team with respect to salary disparity structure when roster flexibility is available. Lastly, in recent years, the NBA has seen drastic changes in total salary availability. When salary cap changes occur, and alter salary flexibility, It is helpful for teams to know if their choices in how the allocation of this new flexibility in salary could affect predicted game-level performance.

In this paper, only teams in the NBA are considered, however, these considerations can be brought into general terms for competing firms in any industry. A firm may consider how their relative or absolute performance in the market is predicted by varying wage scheduling within their firm. As such, similar points may be made for the general firm for each of the above points. Additionally, a firm may alter salary distribution by deciding upon employee salary or wage structure. Lastly, when policies are passed that change salary and wage availability, a firm may consider these results when deciding their best response to the policy given predicted market performance.

## 3 Literature Review

Inspiration for this paper was mainly derived from *A game-level analysis of salary dispersion and team performance in the national basketball association* by Katayama and Nuch (2011) which studies NBA games from 2002 to 2006. Katayama and Nuch (2011) provide a Fixed Effects (for comparison to other studies) and a GMM model to causally link game by game outcomes for relative score to *avesal* and a dispersion metric (*disper1, disper2, or disper3*) while controlling for coaching skill by using coach win rates. Notably, the authors account for variable endogeneity which is unique to other similar studies. Their considered fixed effects model has the form:

$$y_{ijt} = \beta_0 + \beta_1 avesal_{ijt} + \beta_2 disper\#_{ijt} + X_{ijt}\gamma + \mu_{ij} + v_{ijt}$$

where i is the home team, j is the away team, t is the time period, $X_{ijt}$ is the relative coach win rate, $avesal_{ijt}$ is the relative average salary per minute, and $disper\#_{ijt}$ for the home team i compared to the away team j in time t where # is equal to 1, 2, or 3. The error term $\mu_{ij}$ is the unobserved fixed effect, and $v_{ijt}$ is the idiosyncratic error. Note, $ij \neq ji$ here since a team cannot play itself.[9] This model is then run through GMM techniques for non-parametric estimation.

Katayama and Nuch (2011) arrive at various conclusions. Firstly, in the Fixed Effects model

9. Katayama and Nuch, "A game-level analysis of salary dispersion and team performance in the National Basketball Association."

they find that the coefficient on *avesal* is positive and significant, the coefficient on *disper* (all three metrics) is negative and significant, and that coaching skill is a determining factor. By traditional methods, one might say that lower dispersion is casually linked with better team performance while also considering coaching skill. However, their GMM model finds that the Fixed Effects estimation is problematic with simultaneity bias. *avesal* is found to be even more positive and significant, however, all dispersion measures are found to not be causally linked with relative score performance and insignificant. The authors conclude that within-team individuals care little about the salaries of others and thus dispersion does not causally affect team performance and that the true performance metric is maximizing profits.

The work in this paper differs from Katayama and Nuch (2011) in that the models studied here flip the direction of inference. Prediction is considered as opposed to causality. Additionally, more recent game-level (as well as much more) data from 2010 to 2018 is alternatively used. This paper does not consider the coaching skill control variable in its analysis, such that the focus is only on the predictive power of salary dispersion. Lastly, all three dispersion measures are utilized at once when predicting game-level output.

Now, consider the theoretical basis of this game-level analysis. On one hand, economic theory lends support for greater salary disparity. For example, Lazear and Rosen's (1981) rank-order tournament model[10] posits that pay should be distributed based on a ranking of employee merit, from high to low implying a larger variance in distribution giving better results. Thus, this incentivizes workers to provide higher effort levels to extract higher pay levels. On the other hand, Akerloff and Yellen (1990) provide an "effort wage variance" model[11] whereby higher labour effort is supported by lower salary dispersion due to better cohesion amongst workers who are paid at or above their wage threshold. Likewise, Levine (1991) finds that coordination between high-skill and low-skill coworkers leads to greater firm output,[12] which is particularly helpful in basketball, given that game scores are positively correlated to "coordination" statistics such as assists and rebounds by McGoldrick and Voeks (2005).[13]

Empirically, as Katayama and Nuch (2011) compare,[14][15] the results in various sporting leagues are just as conflicted. In baseball, multiple studies conclude that smaller dispersion is causally linked with higher dispersion. For example, in Richards and Guell (1998),[16] Jewell and Molina (2004),[17] and Bloom (1999).[18] In other sports the results are not as obvious. Interestingly, literature suggests conflicting results within specific leagues. Considering the NBA, Berri and Jewell (2004) find that season winning percentages are not related to salary dispersion

10. Lazear and Rosen, "Rank-order tournaments as optimum labor contracts."
11. Akerlof and Yellen, "The fair wage-effort hypothesis and unemployment."
12. Levine, "Cohesiveness, productivity, and wage dispersion."
13. KimMarie McGoldrick and Lisa Voeks, ""We got game!" an analysis of win/loss probability and efficiency differences between the NBA and WNBA," *Journal of Sports Economics* 6, no. 1 (2005): 5–23.
14. Katayama and Nuch, "A game-level analysis of salary dispersion and team performance in the National Basketball Association."
15. Full comparison table is available in the appendix
16. Donald G Richards and Robert C Guell, "Baseball success and the structure of salaries," *Applied Economics Letters* 5, no. 5 (1998): 291–296.
17. R Todd Jewell and David J Molina, "Productive efficiency and salary distribution: The case of US Major League Baseball," *Scottish Journal of Political Economy* 51, no. 1 (2004): 127–142.
18. Matt Bloom, "The performance effects of pay dispersion on individuals and organizations," *Academy of Management Journal* 42, no. 1 (1999): 25–40.

metrics.[19] This study differs from all of these studies in that only game-level data and results are considered, whereas all of these papers focus only on season-long metrics and results.

Recently there has been more research into the topic. Tao et al. (2016) explore an analogous Fixed Effects and GMM estimator to Katayama and Nuch (2011) in Major League Baseball (MLB). Tao et al. conclude that the team-cohesiveness hypothesis is supported.[20] Bucciol et al (2014). note that segmented teams are becoming more and more common within organizations, and in their research, a negative effect of pay dispersion within teams on team performance,[21] again supporting team-cohesiveness. Breunig et al (2014). consider both game-level and season-level scopes and again find a negative effect of salary dispersion on team performance at both of these scales.[22] Interestingly, they also build a wage/effort function that corresponds relative wages with relative effort levels.

Finally, consider support for the usage of machine learning methods in predicting sporting game results. Consider Gu et al. who build highly accurate models which predict NHL game results with over 90% accuracy.[23] Of course, the models considered in this paper differ in that only NBA games are included, which perhaps have different stochasticity.[24] Now, concerning NBA total game scores, Chen et al (2021). implement various algorithms, such as KNN, ELM, etc. to establish the most important variables in predicting NBA game scores accurately. They establish that among these, cohesion statistics such as assists and rebounds are included.[25] However, they do not consider salary statistics.

# 4 Model and Data

Data for this project was sourced exclusively from basketball-reference.com.[26] Data on salary information for NBA players in all studied seasons was sourced from data.world,[27] which in turn scraped this data from basketball-reference.com. However, for game-level and season level statistics, a custom web-scraping algorithm was produced for this paper to scrape game-by-game statistics in the studied seasons as well as season level statistics for every team.[28] All in all, data from over 9000 games was scraped over seven seasons, which included data points for how many minutes a player played during a given game, what team they played for, how many games they played in a given season for that team, etc. which was then linked to the

19. David J Berri and R Todd Jewell, "Wage inequality and firm performance: Professional basketball's natural experiment," *Atlantic Economic Journal* 32, no. 2 (2004): 130–139.

20. Yu-Li Tao, Hwei-Lin Chuang, and Eric S Lin, "Compensation and performance in Major League Baseball: Evidence from salary dispersion and team performance," *International Review of Economics & Finance* 43 (2016): 151–159.

21. Alessandro Bucciol, Nicolai J Foss, and Marco Piovesan, "Pay dispersion and performance in teams," *PloS one* 9, no. 11 (2014): e112631.

22. Robert Breunig et al., "Wage dispersion and team performance: a theoretical model and evidence from baseball," *Applied Economics* 46, no. 3 (2014): 271–281.

23. Gu et al., "A game-predicting expert system using big data and machine learning."

24. Sears Merritt and Aaron Clauset, "Scoring dynamics across professional team sports: tempo, balance and predictability," *EPJ Data Science* 3, no. 1 (2014): 1–21.

25. Wei-Jen Chen et al., "Hybrid Basketball Game Outcome Prediction Model by Integrating Data Mining Methods for the National Basketball Association," *Entropy* 23, no. 4 (2021): 477.

26. This is a trusted resource that is utilized in peer-reviewed published works such as Katayama and Nuch (2011)

27. https://data.world/datadavis/nba-salaries

28. This web-scraping code is available upon request

salary data mentioned above.

Given game-by-game player statistics, dispersion metrics were calculated for each team involved in every given game. First off, the maximum season-long player salary for players participating in a given game g for a given team i and was called $maxsal_{ig}$. Secondly, the average salary, per minute of game-time was calculated for each team i in every game g using:

$$avesal_{ig} = \frac{\sum_k^{K_{ig}} sal_{ki} mp_{kig}}{\sum_k^{K_{ig}} mp_{kig}}$$

where $K_{ig}$ is the set of all participating players on team i in game g, $sal_{ki}$ is player k's season-long salary, and $mp_{kig}$ is how many minutes player k for team i played in game g. Thirdly, *disper1*, was calculated using:

$$disper1_{ig} = \frac{\sqrt{varsal_{ig}}}{avesal_{ig}} \text{ where } varsal_{ig} = \frac{\sum_k^{K_ig}(sal_{ki} - avesal_{ig})^2 mp_{kig}}{\sum_k^{K_ig} mp_{kig}}$$

*disper2* was calculated differently as it is considered on a season-by-season basis. Essentially, the Gini coefficient for salary was calculated for all players on a given team that have played greater than or equal to half of the team's games in a given season.[29] This was calculated using:

$$disper2_{ig} = \frac{\sum_{k_1}^{K_i} \sum_{k_2}^{K_i} |sal_{k_1 i} - sal_{k_2 i}|}{2|K_i|^2 \bar{k_i}}$$

and was applied to every game the team was involved in that season. Finally, *disper3* was also a season level statistic given by the Herfindahl index for the salary of all players on a team's roster, given by the following formula:

$$disper3_{ig} = (\sum_k^{K_i}[(\frac{sal_{ki}}{\sum_k^{K_i} sal_{ki}})100])^2$$

These metrics were utilized to form the input of Random Forest models which can be separated into two classes: regression and classification. The regression models can further be divided into two types. First off, a Random Forest regression model is used to predict the relative score of a given game ("$\Delta$score"). "$\Delta$score" can be interpreted as follows: $\Delta score_g$ *is the ratio of the total amount of points scored by the home team divided by the total amount of points scored by the away team in game g* (ie. relative score). The input to the model includes all of the above statistics for both the home and away teams. Secondly, a Random Forest regression model is used to predict total team point output (for both home and away teams respectively), given only the dispersion metrics for the home team when predicting home point total, and dispersion metrics for the away team when calculating the away point total. Using dispersion metrics for both the away and home teams to predict home/away scores was also considered. Thirdly, a Random Forest classification model was built to predict a binary output which equaled one if the home team won, and zero if they lost, given all of the dispersion metrics for both the home and away teams.[30]

29. To meet this requirement, players had to have played more than 40 games for the team that season, except for the 2011-2012 season, where only more than 24 games were required due to a shortened season

30. Sample trees for each of the model types are given in the appendix (Fig. 1, Fig. 4, Fig. 13)

To test these models, there were multiple considerations. Previous works were concerned with the endogeneity of the dispersion metrics. However, with the Random Forest approach, this concern is erased as endogeneity should not affect the model and does not need to be explicitly tested for. The main testing metric of concern in the regression models was the root mean squared error (RMSE), which was compared against a baseline OLS model of all of the inputs and the relevant distribution statistics for the output. Additionally, the percentage of variance explained, percentage increase in MSE, and increase in node purity metrics were considered. For the classification matrix, testing techniques such as the confusion matrix of the test data and the Receiver Operator Characteristic curve (ROC) were of main concern. Mean decrease in accuracy, as well as mean decrease in Gini, were also considered.

In total, regarding all three models, 9520 games were considered. These games were randomly split into 70% training data and 30% testing data. The training was done in a supervised setting, whereby the predicted value of the output was already known for each studied NBA game, whether it be relative score, team point totals, or binary win/loss field. Additionally, for each model, 500 trees were considered in each Random Forest.

The scraped relative data for the studied team performance and salary dispersion metrics in the 2010-2011 to 2017-2018 seasons is summarized as follows:

Table 1: Summary Statistics

| Statistic | Mean | St. Dev. | Min | Max | Pctl(25) | Median | Pctl(75) |
|---|---|---|---|---|---|---|---|
| home score | 102.533 | 12.022 | 59 | 149 | 94 | 102 | 111 |
| away score | 99.743 | 12.104 | 56 | 148 | 91 | 100 | 108 |
| home win | 0.589 | 0.492 | 0 | 1 | 0 | 1 | 1 |
| $\Delta$score | 1.038 | 0.142 | 0.571 | 1.965 | 0.936 | 1.039 | 1.121 |
| $\Delta$maxsal | 1.026 | 0.409 | 0.113 | 4.413 | 0.740 | 1.000 | 1.245 |
| $\Delta$avesal | 1.155 | 0.659 | 0.113 | 11.338 | 0.736 | 1.013 | 1.399 |
| $\Delta$disper1 | 1.035 | 0.252 | 0.320 | 3.316 | 0.873 | 1.005 | 1.158 |
| $\Delta$disper2 | 1.015 | 0.179 | 0.452 | 2.214 | 0.890 | 0.999 | 1.124 |
| $\Delta$disper3 | 1.094 | 0.498 | 0.198 | 5.052 | 0.774 | 0.996 | 1.292 |

Note: All metrics displayed are ratio values (homeStat/awayStat) except for home score, away score, and home win

There are a few important interpretive details to note. One, the home team generally has better performance than the away team in the set of games that were collected for this study. Ie ($\frac{points_{home}}{points_{away}} > 1$) on average. Two, *avesal* is drastically above one in ratio value ($\frac{avesal_{home}}{avesal_{away}} >> 1$), perhaps meaning that the home team is more willing to spend more on salary at the game-level. Intuitively this makes sense, as teams may be more inclined to play their higher-paid players in front of their home crowd to maximize revenue from the event (maximizing profits). Three, the relative dispersion statistics *disper1*, *disper2*, *disper3*, and relative *maxsal* are slightly above one. This may be interpreted as the home team wanting to please their revenue creating home crowd more by playing higher paid players.

The summary statistics found above, are largely consistent with Katayama and Nuch's (2011) work, but notably, a larger relative value for *avesal* than Katayama and Nuch (2011),

and much lower relative *disper2* value is seen here.[31] Perhaps these differences can be explained through how the NBA has evolved its salary rules and shifted their views on resting important, high-salary players when they are the away team. The other relative values calculated match closely to Katayama and Nuch (2011). Note, relative value for *maxsal* is not considered against Katayama and Nuch (2011), as the authors do not consider this statistic.

# 5 Empirical Results

For the following explanations of the empirical results, plots will be consistently referred to and can be found within the appendix of this paper.

Relative Score Random Forest Regression Model:

In the Random Forest model predicting relative score results of a given game, there are a few key results. When evaluating the trained model, the output suggests 20% of the variance in the relative score for game-level results is explained through the input variables. Relatively speaking, this is a weak result. Predictive models should aim for a larger percentage of explained variance to yield more accurate results, and a value this low suggests poor explanatory value in the input variables and should lead to viewing the input variables as poor predictors of the game-level relative score.

In the testing data, the observed RMSE is a relatively high value of 0.1278. Also note, the error rate was decreasing throughout, but it appears that this is about as low as the error rate will go (Fig. 2). Dividing this by the mean of the output variable, the RMSE is about 12.3% of the mean of this output variable. The first and third quartiles of this output variable fit comfortably within this range, so the significance of this model is skeptical. When compared to the general OLS RMSE, the model is shown to have about a 3% gain in efficiency, but this is still poor. When considering a weak poor percentage of variance explained and a high RMSE value relative to the output, salary dispersion variables can be claimed as poor predictors of relative score at the game level.

Important information can still be taken away from this unproductive model, particularly when observing the plot for percentage increase in MSE per input variable, and the plot for the increase in node purity per each input variable (Fig. 3). From these plots, different salary dispersion variables are ranked based on their predictive importance.

Percentage increase in MSE displays the relative increase in MSE as a result of the specific input variable being permuted. As such, the higher this number is, the greater the effect is on the MSE value, and the more important and integral the variable is to the model's success.

Increase in node purity should also help determine the relative importance of input variables in a different way. Increase in node purity gives information on how input relates to the loss function employed by the Random Forest model to determine rational optimizing splits or segments of the data. In this model, MSE is the loss metric. More important inputs will have a higher increase in node purity value since when choosing node splits, it is desirable to have smaller amounts of intra-node variance and larger amounts of inter-node variance. This would

31. See appendix for Katayama and Nuch's (2011) table containing comparable values

allow the model to more accurately predict outcomes with higher precision.

Given the charts in the appendix that corresponds to the above information, it is clear to see that *avesal*, specifically for the away team, is the most important predictor for relative score outcomes. Additionally, *disper1* and *maxsal* for both teams are moderately important in comparison. Perhaps most importantly, *disper2* and *disper3* provide little importance to the model from these metrics, consistently placing last in these rankings. Interestingly, in both plots, there seem to be clear "jumps" in the data from one level of importance to the next. In other words, inputs that are close in importance seem to be very closely ranked, whereas those who are not close are widely dispersed in ranking. For example, the jump from *disper2* and *disper3* to *disper1* and *maxsal* for the percentage increase in MSE metric. This suggests a large gap in predictive power amongst these salary dispersion measures.

Team Score Random Forest Regression Model:

The team score model can be broken into four scenarios: predicting home team score using home salary dispersion metrics, predicting away team score using away salary dispersion metrics, predicting home team score using both home and away team salary dispersion metrics, and predicting away team score using both home and away team salary dispersion metrics. During this analysis, these models will be referred to as model 2a, 2b, 2c, and 2d respectively.

With respect to percentage of variance explained, model 2a, 2b, 2c, and 2d displayed values of 13.23%, 10.35%, 15.88%, and 13.59% respectively. Across the board, these are poor values. Much like the relative score model, these values simply are not high enough to establish an explanatory relationship between the dispersion metrics and their respective output. Thus, the predictive power of the inputs to any of these models is likely to be low and have little predictive significance.

When looking at RMSE, models 2a, 2b, 2c, and 2d report values of 11.07, 11.23, 12.63, and 11.06 respectively. Note also, from the error rate plots, this appears to be as low as the error rates will go (Fig. 5-8). Analogously to the relative score model, the RMSE values are unfavourably high. Looking at the mean of each model, each of the RMSE values accounts for 10.8%, 11.2%, 12.3%, and 11.0% of the mean for their respective outputs. Considering that the first and third quartile for each model's output falls comfortably within each of these RMSE values, these are poor RMSE values. Additionally, when comparing to the appropriate OLS RMSE, gains of only about 5%, 4%, -9%, and 5% are seen for models 2a, 2b, 2c, and 2d respectively. This is very low, and in the case of model 2c, OLS performs more efficiently. Strong skepticism about the predictive power of these models with respect to total score prediction follows.

Combining the low explained variance and poor RMSE values for each total score model, the predictive value of salary dispersion metrics on either home or away team point totals is not valuable. It is tough to tell whether including both teams' salary dispersion metrics as input to the model helps or hinders performance. This is because each of the explored models in this section performs nearly as poorly as the others. Since explanatory or predictive power in these models cannot be established, it is immature to be ranking how well they fit the data or even how they compare.

In the percentage increase in MSE plots, surprisingly, considering their salary dispersion

metrics (2a and 2b), *disper3* is the most important variable. However, looking at 2c and 2d, *disper3* is relegated back down to the bottom end of the chart, becoming much less important when considering both teams (Fig. 9-12). Generally, the most important variables in 2a and 2b become the least important variables in 2c and 2d. This may be due to the fact that when predicting point totals of a single team, using only their salary dispersion metrics, it is more important and consistent to know what the actual roster structure is. Whereas when using both teams' dispersion metrics these values become much less consistent and having more game-level information, such as how much salary was being spent per minute, becomes more important. In any case, this is a unique phenomenon.

Consider the plots reflecting increase in node purity for each input (Fig. 9-12). Observe that across the board, *disper2* is ranked as the least important variable concerning increasing node purity in models 2a, 2b, 2c, and 2d. Additionally, *disper3* and *maxsal* are relatively unimportant inputs with respect to this node purity. Especially in the cases of 2a and 2b, *avesal* and *disper1* are the most important input variables to be splitting and segmenting on when predicting single team point totals.

With all of this information, predicting the total point production of a home or away team at the game level using either their salary dispersion metrics or both teams' salary dispersion metrics, is not accurate or precise. Salary dispersion metrics do not appear to be a viable method to predict total point production for a given team.

Win/Loss Random Forest Classification Model:

Consider the case of predicting a binary output column that equals one in the event of a home team win and zero in the event of a home team loss. This model takes in salary dispersion metrics from both teams and employs a random forest classification approach.

As this is a classification model, different metrics of success will be considered. First off, observe the "Out-of-bag" estimate of error rate. In simple terms, this represents how often a classification model can be expected to be incorrect on original data. This number is calculated by keeping a specific portion of the training data to estimate how often the model will be incorrect on new data. Reported in the model output, an "Out-of-bag" estimate error rate of 34.68% is seen. This means that on new data, the model can be expected to be incorrect about 34.68% of the time. Since the model is looking to predict a binary classification output, a completely random model could expect to be incorrect about 50% of the time. Thus, the model shows some improvement over a completely random guessing model.

| | Predicted | |
|---|---|---|
| Actual | 0 | 1 |
| 0 | 560 | 410 |
| 1 | 617 | 1269 |

To test this error rate, the model is run against testing data. The results of this test can be seen in the testing "confusion matrix" shown above. The confusion matrix states the true value of the output on the rows, and the models guess at the output along the columns. So, the upper left portion of the matrix can be interpreted as a "true negative" (TN) or a case where the model guessed a loss by the home team and the home team actually lost. Analogously, a

"true positive" (TP) is in the bottom right portion of the matrix and occurs when the model guesses a home win and the home team actually won. Similarly, a "false positive" (FP) is in the top right, where the model guessed for a win, but it was actually a loss, and a "false negative" (FN), where the model guessed for a loss, but it was actually a win. The confusion matrix then states counts for each of these scenarios.

Reading the confusion matrix, the model accurately guessed ($\frac{TP+TN}{TP+TN+FP+FN}$) about 64% of the time, which falls approximately in line with the "Out-of-bag" estimate. On the testing data, the model performed significantly better than a randomly guessing algorithm, which would report 50%. Also, it can be seen that when the output was actually a loss, measuring negative predictive power, the model guessed correctly ($\frac{TN}{TN+FN}$) about 58% of the time. Additionally, when considering output for a win, or the precision of the model, the model guessed correctly ($\frac{TP}{TP+FP}$) about 67% of the time. This means that the model performs slightly worse when the true output is a loss as opposed to a win. This improvement on guessing wins can be seen in the corresponding error rate plot, where the error is progressively being reduced in predicting wins. But, the error rate seems to be squeezing in on 0.5 for losses, or becoming about as accurate as a randomly guessing model (Fig. 14).

Consider also how often the model is correct when it guesses for a win or for a loss. When guessing for a loss, the specificity of the model ($\frac{TN}{TN+FP}$) was only 47%. This suggests the model was probably as good as random in this respect. When guessing for a win, the model sensitivity ($\frac{TP}{TP+FN}$) was a staggering 75%, suggesting the model was much better at predicting for wins. This higher accuracy may be because of the fact that the original data contained about 66% wins versus 44% losses for the home team. When considering in conjunction that about 58% of guesses were for wins, so the model may have been learning to simply predict for a win more often.

Like the previous models, input importance can be evaluated (Fig. 15). When considering mean decrease in accuracy, or how much less correct "out-of-bag" results become when removing an input variable, it is seen that *avesal* and *disper1* are very important. Also, *maxsal* is somewhat important, and both *disper2* and *disper3* are least important to the model. When considering the mean decrease in Gini coefficient, or the gain in node purity when including a variable, a similar ranking is seen to the mean decrease in accuracy.

Finally, consider the ROC curve (Fig. 16). This curve is built by plotting the true positive rate against the false positive rate (the true negative rate against the false negative rate) on a one unit by one unit plot. Along the 45-degree line is the theoretically randomly guessing algorithm with an area underneath the curve (AUC) of 0.5. Now, in the figure, the prediction rate for a loss is in pink, and the prediction for a win is in green. From the output, these curves have an identical AUC of 0.67, meaning that they are slightly better than the randomly guessing algorithm. This AUC may be interpreted as the probability that a random positive instance is ranked higher than a random negative instance and is essentially the predictive accuracy of the model for positive/negative results. Additionally, note that the win and loss prediction curves are extremely similar and have almost no sections where they differ, so the earlier worry about the model simply having better accuracy on wins can be reduced.

The win/loss classification model performs significantly better than a randomly guessing

model by around 14 percentage points in practice but performs about 2 percentage points worse than a model which only predicts for wins. The predictive accuracy of this model is not inspiring. At the game level, salary dispersion metrics can have some level of predictive value on binary game outcomes, but this value is not significant.

# 6 Conclusions

The results of this study are consistent with Katayama and Nuch's (2011) findings.[32] It cannot be determined with certainty that salary dispersion metrics have any predictive value for game-level team performance. This was found to be the case when predicting relative point totals in a game, individual team point totals, and a binary win or loss classification for NBA teams. Even so, it is also found that generally, *avesal* is found to have the most predictive value for these models. Additionally, amongst the dispersion metrics *disper1*, *disper2*, and *disper3*, are the most useful metric tends to be *disper1* when predicting team performance. Ultimately, there seems to be little predictive value for salary dispersion metrics in predicting NBA team performance results. Even with these negative results, interpretive value can still be found in these findings.

Overall, decision makers within a team should not be concerned about salary dispersion within their team having direct predictive implications on their team's future performance. Decision makers should simply be concerned with finding players with the most individual output as individual statistics have more predictive power,[33] while keeping in mind that a higher average salary per minute generally has the most impact on team performance prediction. This means that teams should not be afraid of paying at or above market value for individuals who perform.

If a decision maker was to concern themselves with salary dispersion for game-level predictiveness, it would be wise to consider *disper1* as a first choice. This is likely due to the fact that this variance metric is directly based on the average salary per minute of a team in a game. Also, it is the only game-level dispersion metric that is considered, as *disper2* and *disper3* are calculated at the season-level and thus are intuitively less likely to have as much value on a specific game's outcome.

*disper2* and *disper3* are generally poor dispersion metrics to be using for game-level prediction in the NBA. Interpreting this, the Gini coefficient for "active" players and the Herfindahl index for the entire roster has little predictive value. As such, salary construction of the full team or even salary construction of those who frequently play is not significantly a predictive factor in team performance in any given game. As a result, decision makers within the team should not be worried about how dispersed their roster is when considering salary for how their team will perform in a single game.

Interestingly, *maxsal* has little predictive value, but still generally has more value at the game-level than *disper2* and *disper3*. This means that the "star player effect"[34] has more value

32. Katayama and Nuch, "A game-level analysis of salary dispersion and team performance in the National Basketball Association."

33. Chen et al., "Hybrid Basketball Game Outcome Prediction Model by Integrating Data Mining Methods for the National Basketball Association."

34. Jacqueline Agesa, Richard U Agesa, Maria Toshkova, et al., "NBA salaries: role players and superstars.,"

than "active" and full roster salary dispersion. Essentially, given that a team has an extremely well-paid "star" player is a more important predictive fact than knowing how spread out the salaries are on a given roster. However, this effect still has little predictive power.

These results provide motivations for further research. For example, research on the NBA has had relatively mixed results for the establishment of a causal relationship between salary dispersion and team performance. Perhaps considering another sport or league which has had more success in determining this causal relationship, such as the MLB, would yield more accurate predictive results. Additionally, since the NBA only has five players playing at any given moment, perhaps sports that have more players playing at once would yield more distinctive dispersion and thus higher predictive accuracy. A couple of examples of such observable leagues could be the National Football League (NFL) or MLB.

Another interesting line of research could be the combination of salary metrics along with individual performance metrics to establish more accurate predictive measures. As referred to in the literature review, highly accurate predictive models for game-level team performance can be built simply from individual performance metrics. Adding salary dispersion metrics to these models could enhance the predictive granularity with which team performance is segmented. Salary dispersion could be used alongside individual performance metrics to further segment team performance without overfitting. Along a similar vein, salary dispersion could be utilized in causal models which relate individual performance to team performance. Since it has been established that salary dispersion alone does not have significant causal[35] or predictive value in evaluating team performance in the NBA, perhaps salary dispersion could be used as an instrument to individual performance metrics. Specifically, the dispersion could be utilized as an instrument to "cohesion statistics" such as assists and rebounds, or even to all individual performance metrics.

Additional predictive models could also be explored. However, this is limited by how little predictive value was found for these dispersion metrics to have in using a random forest approach. Random Forests seem to have the most value in this context since decision makers within a team are looking for maximal interpretive value to react accordingly and make changes to their roster and player schedules to predict better team performance. Using a complex model, such as an Artificial Neural Network, would tell be able to predict how a team is going to perform but would not be able to tell a decision maker the most important metrics that are driving performance when attempting to increase predicted team performance.

The value of predicting team performance based on salary dispersion may be called into question. Given that salary contracts generally cannot be altered within a given season for any player without cost, and the fact that adapting salary dispersion on a game-by-game basis would be very difficult for a team, it may be the case that it is not particularly useful to know how salary dispersion affects game-level team performance. Essentially, the managerial overhead and lack of flexibility in salary distribution in the short run for a team may not be worth the low predictive power on a game-level granularity. Perhaps considering prediction on the season level, where teams have more salary dispersion flexibility and decision-making capacity would

*The Sport Journal* 8, no. 2 (2005).

35. Katayama and Nuch, "A game-level analysis of salary dispersion and team performance in the National Basketball Association."

be a more interesting topic to consider. Additionally, teams may care more about profit margins when building salary amounts, which may be a more telling metric to predict.

All in all, it is found that salary dispersion is not helpful in the prediction of team performance at the game level. Notice that redefining team performance in three different ways does not aid in the predictive performance of the five considered salary dispersion metrics. These results make sense considering previous research on the NBA, but, the generalizability of the lack of results found here is uninspiring. Thus, there are numerous future avenues to be explored to find predictive capacity in salary dispersion on team performance. Importantly, the relationship between game-level team performance and salary dispersion may stem from differences in player skill, in which higher-skilled players get paid more and win more. As such, possibly the most useful case for salary dispersion metrics may be as an additional explanatory aid for more predictive metrics that are more greatly attached to team outcomes, such as assists and rebounds. Alone, salary dispersion is not significantly predictive for game-level team outcomes. Thus, salary may be distributed in various ways to reach favourable team results.

## References


Agesa, Jacqueline, Richard U Agesa, Maria Toshkova, et al. "NBA salaries: role players and superstars." *The Sport Journal* 8, no. 2 (2005).

Akerlof, George A, and Janet L Yellen. "The fair wage-effort hypothesis and unemployment." *The Quarterly Journal of Economics* 105, no. 2 (1990): 255–283.

Berri, David J, and R Todd Jewell. "Wage inequality and firm performance: Professional basketball's natural experiment." *Atlantic Economic Journal* 32, no. 2 (2004): 130–139.

Bloom, Matt. "The performance effects of pay dispersion on individuals and organizations." *Academy of Management Journal* 42, no. 1 (1999): 25–40.

Breunig, Robert, Bronwyn Garrett-Rumba, Mathieu Jardin, and Yvon Rocaboy. "Wage dispersion and team performance: a theoretical model and evidence from baseball." *Applied Economics* 46, no. 3 (2014): 271–281.

Bucciol, Alessandro, Nicolai J Foss, and Marco Piovesan. "Pay dispersion and performance in teams." *PloS one* 9, no. 11 (2014): e112631.

Chen, Wei-Jen, Mao-Jhen Jhou, Tian-Shyug Lee, and Chi-Jie Lu. "Hybrid Basketball Game Outcome Prediction Model by Integrating Data Mining Methods for the National Basketball Association." *Entropy* 23, no. 4 (2021): 477.

Gu, Wei, Krista Foster, Jennifer Shang, and Lirong Wei. "A game-predicting expert system using big data and machine learning." *Expert Systems with Applications* 130 (2019): 293–305.

Heyman, Fredrik. "Pay inequality and firm performance: evidence from matched employer–employee data." *Applied Economics* 37, no. 11 (2005): 1313–1327.

Katayama, Hajime, and Hudan Nuch. "A game-level analysis of salary dispersion and team performance in the National Basketball Association." *Applied Economics* 43, no. 10 (2011): 1193–1207.

Lazear, Edward P, and Sherwin Rosen. "Rank-order tournaments as optimum labor contracts." *Journal of political Economy* 89, no. 5 (1981): 841–864.

Levine, David I. "Cohesiveness, productivity, and wage dispersion." *Journal of Economic Behavior & Organization* 15, no. 2 (1991): 237–255.

McGoldrick, KimMarie, and Lisa Voeks. ""We got game!" an analysis of win/loss probability and efficiency differences between the NBA and WNBA." *Journal of Sports Economics* 6, no. 1 (2005): 5–23.

Merritt, Sears, and Aaron Clauset. "Scoring dynamics across professional team sports: tempo, balance and predictability." *EPJ Data Science* 3, no. 1 (2014): 1–21.

Richards, Donald G, and Robert C Guell. "Baseball success and the structure of salaries." *Applied Economics Letters* 5, no. 5 (1998): 291–296.

Tao, Yu-Li, Hwei-Lin Chuang, and Eric S Lin. "Compensation and performance in Major League Baseball: Evidence from salary dispersion and team performance." *International Review of Economics & Finance* 43 (2016): 151–159.

Todd Jewell, R, and David J Molina. "Productive efficiency and salary distribution: The case of US Major League Baseball." *Scottish Journal of Political Economy* 51, no. 1 (2004): 127–142.

# Appendix

Result comparison table from *A game-level analysis of salary dispersion on team performance in the National Basketball Association*:

**Table 1. The effect of salary dispersion on team performance in US professional sports**

| League | Author | Performance measure | Dispersion measure | Estimation | Effect |
|---|---|---|---|---|---|
| MLB | Richards and Guell (1998) | SWP, DT, LC, WC | Variance | OLS, Probit | − significant for SWP; − insignificant for DT, LC and WC |
| | Bloom (1999) | SWP, FP, FA1 | Gini coefficient | OLS | − significant for SWP and FA1; + significant for FP |
| | Depken (2000) | SWP | Herfindahl | FE, RE | − significant |
| | Frick *et al.* (2003) | SWP | Gini coefficient | FE, RE | − significant |
| | Jewell and Molina (2004) | SWP | Gini coefficient | SF | − significant |
| | DeBrock *et al.* (2004) | SWP, FA2 | Herfindahl | OLS | − significant for SWP and FA2 |
| | | SWP, FA2 | Conditional Herfindahl | Two-stage method | − insignificant for SWP; − significant for FA |
| NBA | Frick *et al.* (2003) | SWP | Gini coefficient | FE, RE | + significant |
| | Berri and Jewell (2004) | ΔSWP | Herfindahl 1 | FE, RE | + insignificant |
| NHL | Sommers (1998) | SEPT | Gini coefficient | OLS | − significant |
| | Frick *et al.* (2003) | SWP | Gini coefficient | FE, RE | + insignificant |
| | Marchand *et al.* (2006) | TP, PO, DV, CF, SC | Gini coefficient | OLS, Logit | + significant for TP and PO; + insignificant for DV, CF and SC |
| NFL | Frick *et al.* (2003) | SWP | Gini coefficient | FE, RE | − insignificant |

*Notes*: NBA, MLB, NHL and NFL denote the National Basketball Association, the Major League Baseball, the National Hockey League and the National Football League, respectively. SWP denotes Season Winning Percentage. DT, LC and WC refer to Division Title, League Championship and World Championship, respectively. FP represents Finishing Position which is the number of games behind the division leader the team was at season's end. Thus, the higher this value, the poorer the performance. FA1 denotes total home attendance (stadium capacity × the number of home games), whereas FA2 is total home attendance. SEPT denotes Season-Ending Point Totals (two points for a win, one for a tie and no points for a loss). TP, PO, DV, CF and SC denote team points, whether the team made playoffs, whether the team won the division, whether the team won the conference and whether the team won the Stanley Cup, respectively. Herfindahl denotes the Herfindahl index. Herfindahl 1 refers to the variable that represents how many SDs above or below the league average Herfindahl index a team's salary dispersion is located. FE, RE, SF and OLS denote the Fixed Effects estimator, the Random Effects estimator, the Stochastic Frontier model, and Ordinary Least Squares, respectively. This table does not cover financial performance measures used in Bloom (1999) and DeBrock *et al.* (2004).

Relative salary dispersion metric results from *A game-level analysis of salary dispersion on team performance in the National Basketball Association*:

**Table 2. Summary statistics**

| Variable | Mean | SD | Min | Max |
|---|---|---|---|---|
| Ratio of final points (*y*) | 1.049 | 0.142 | 0.559 | 1.746 |
| Ratio of average salaries (*avesal*) | 1.122 | 0.573 | 0.154 | 7.827 |
| Ratio of salary dispersion 1 (*disper*1) | 1.078 | 0.452 | 0.234 | 4.71 |
| Ratio of salary dispersion 2 (*disper*2) | 1.386 | 1.372 | 0.048 | 20.757 |
| Ratio of salary dispersion 2 (*disper*3) | 1.027 | 0.243 | 0.447 | 2.239 |
| Ratio of coaches' records (*Coaches' record*) | 8.878 | 56.131 | 0.001 | 1938 |
| Ratio of coaches' experience (*Coaches' experience*) | 1.071 | 0.41 | 0.163 | 6.196 |

*Notes*: These statistics are calculated using 4176 unique games. *avesal* refers to the ratio between opposing teams of their minute-adjusted average salaries. *disper*1, *disper*2 and *disper*3 refer to the ratio between opposing teams of their minute-adjusted coefficient of variation of salaries, Herfindahl index of salary dispersion and Gini coefficient for salaries, respectively. *disper*2 is constructed using only players who played more than half of their team's games, whereas *disper*3 is constructed using team rosters that include all players who had participated in at least one game during the season. The 'Ratio of coaches' records' refers to the ratio between opposing teams of their coaches' previous losing records, where a losing record is the ratio of losses to total games. The 'Ratio of coaches' experience' is the ratio between opposing teams of their coaches' experience measured in games coached.

Plots accompanying the relative-score regression model:

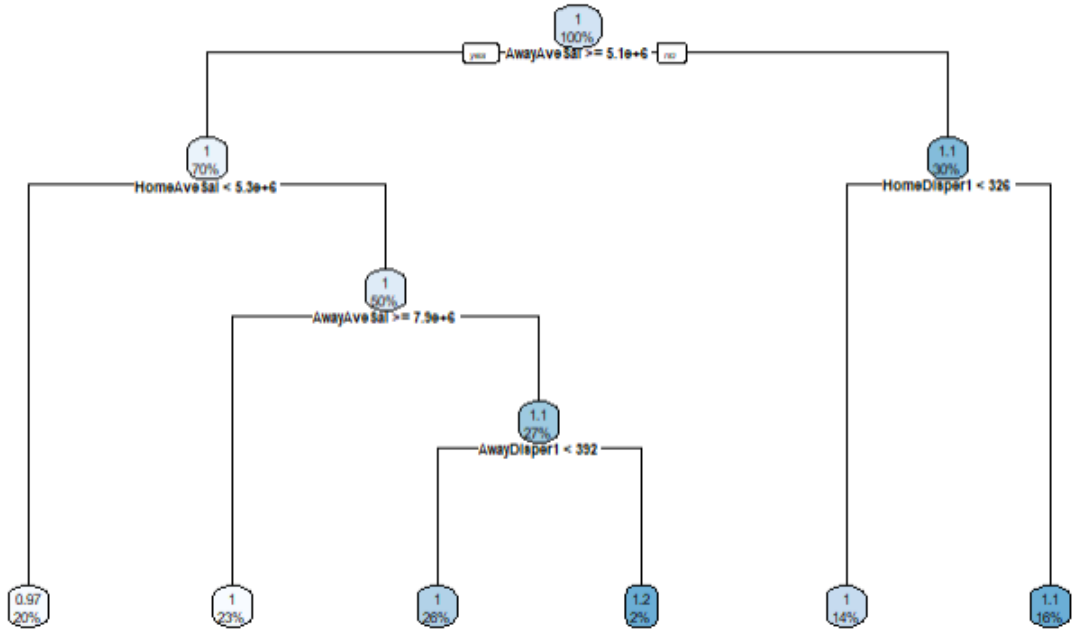


Figure 1: Sample tree for the relative score regression model

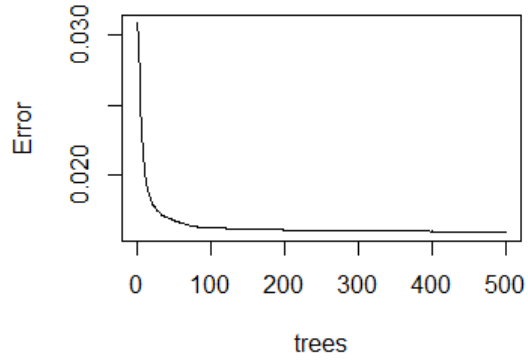


Figure 2: Error rate over time for the relative score regression model

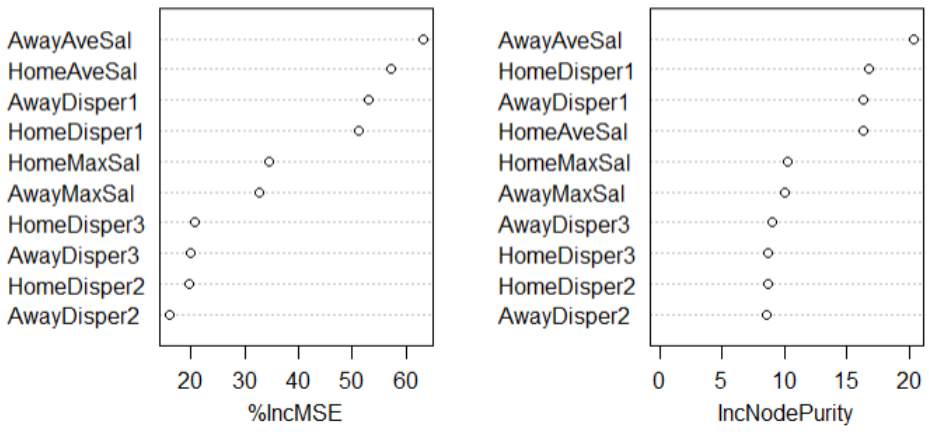


Figure 3: Variable importance for the relative score regression model

Plots accompanying the total-score regression models:

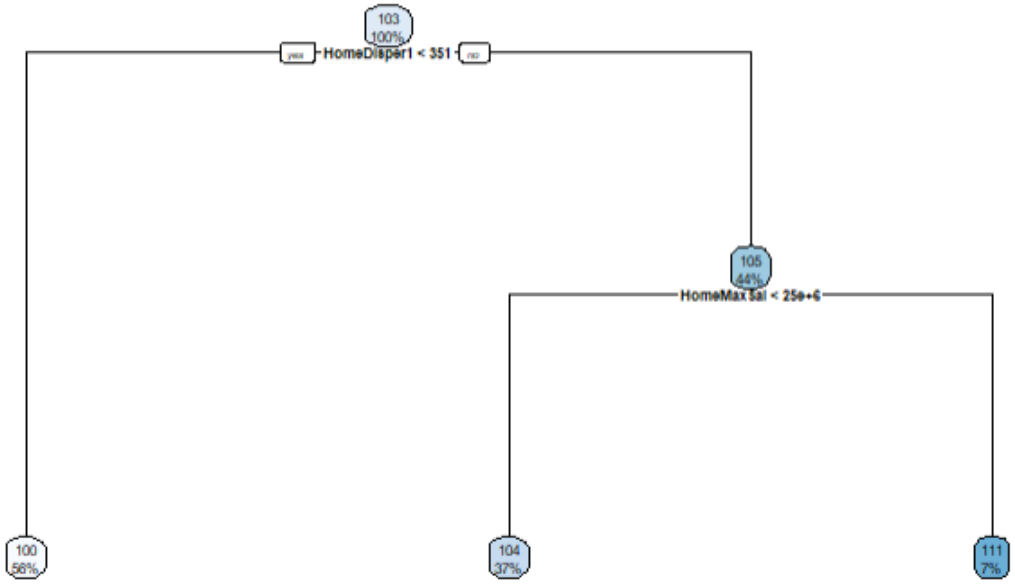


Figure 4: Sample tree for the total score model 2a (2b, 2c, and 2d follow similarly)

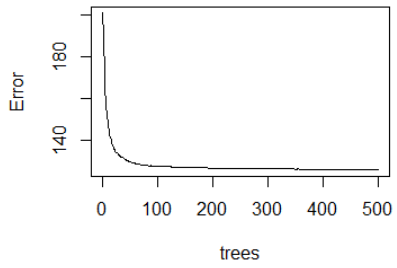


Figure 5: 2a error rate

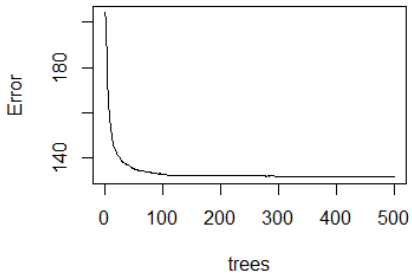


Figure 6: 2b error rate

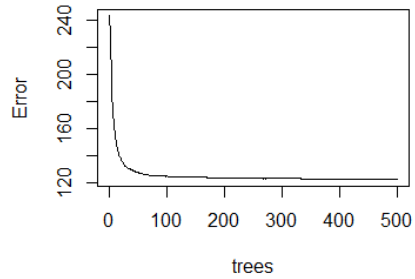


Figure 7: 2c error rate

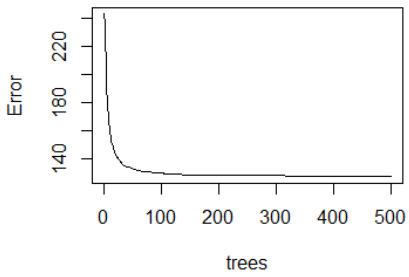


Figure 8: 2d error rate

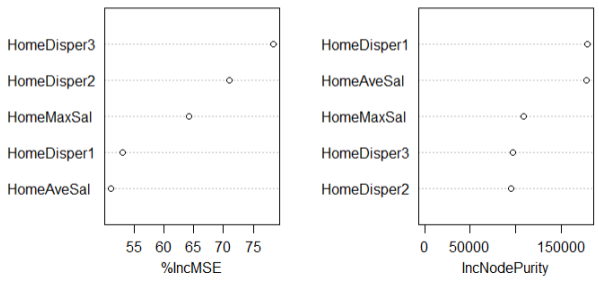


Figure 9: 2a variable importance

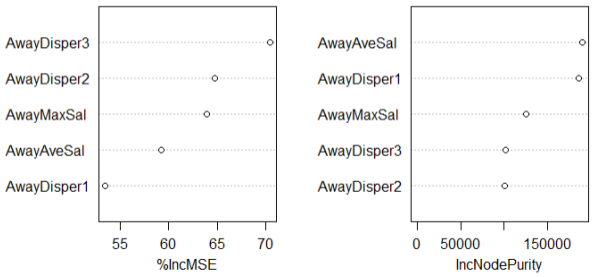


Figure 10: 2b variable importance

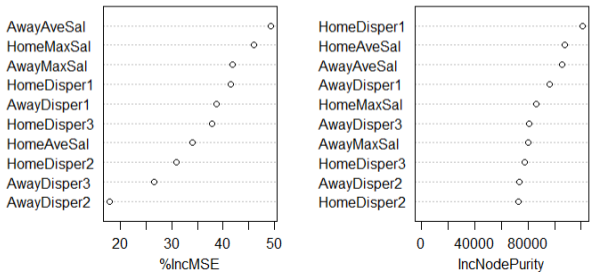


Figure 11: 2c variable importance

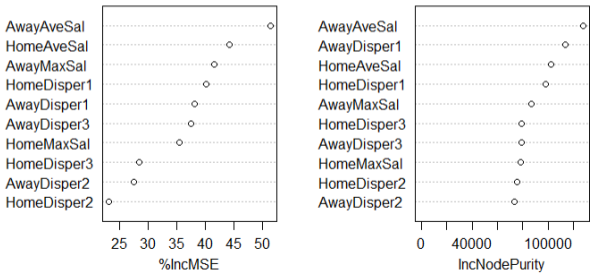


Figure 12: 2d variable importance

Plots accompanying the win/loss classification model:

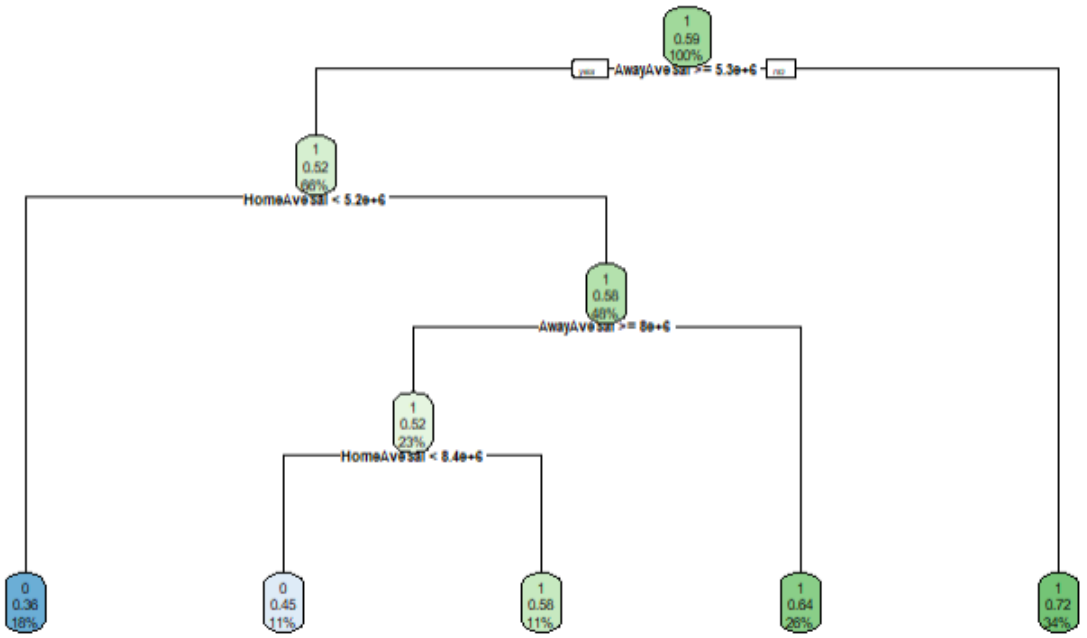


Figure 13: Sample tree for the win/loss classification model

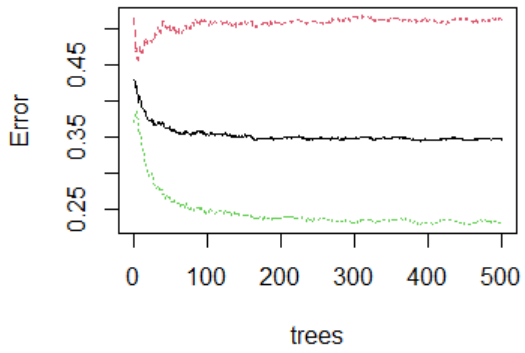


Figure 14: Error rate for the win/loss classification model (red is loss error, green is win error)

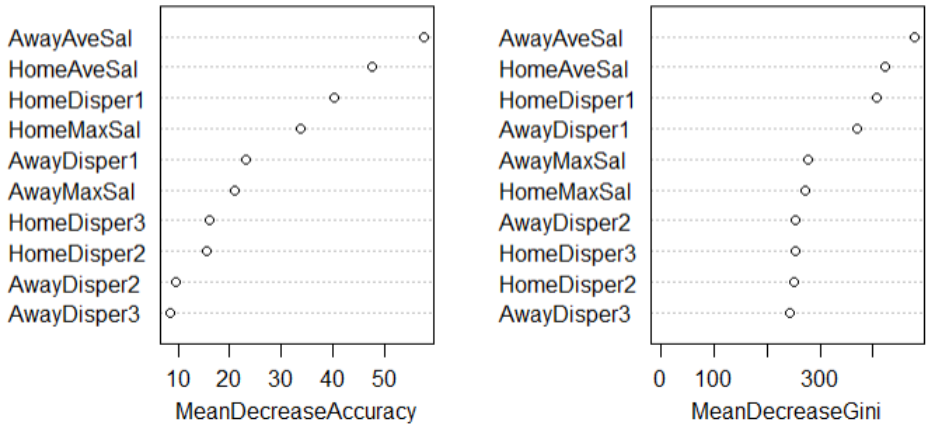


Figure 15: Variable importance for the win/loss classification model

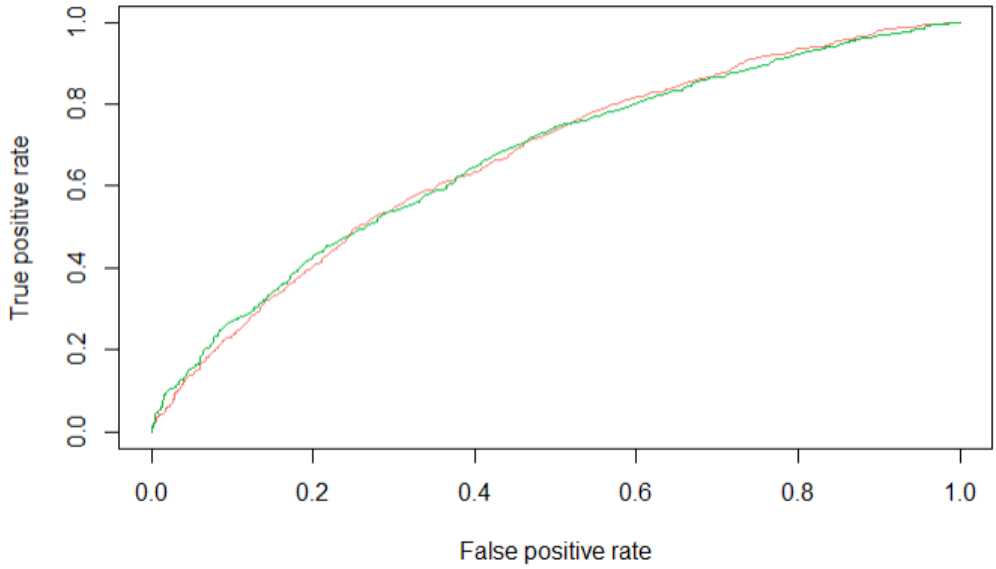


Figure 16: Receiver Operator Characteristic curve for the win/loss classification model (red represents predicting a loss, green represents predicting a win)