Predicting Winning Percentage of MLB Teams

April 6, 2023

1 Introduction

Data analysis has become an increasingly important part of baseball in recent years, as teams and players look for ways to gain a competitive edge. The use of data and statistics to analyze performance and make strategic decisions is known as "sabermetrics", and it has become a major part of the game. Sabermetrics involves using data to evaluate players and teams in order to identify strengths and weaknesses, and to make decisions about things like player personnel and game strategy. This can include methods like analyzing a player's statistics to determine their value to a team and using data to determine the best lineup or defensive alignment for a particular game.

For this project, we hoped to determine which factors are most predictive of the success of a baseball team as defined by their winning percentage in the next season, using MLB team and player statistics using data from 1997 to 2022. We specifically looked at predictors that do not directly measure runs scored in order to gain better insights into the type of team that leads to success. We decided to explore this from two points of view: using overall team data and using individual player data. To determine the factors that contribute most to a team's winning percentage, we fit various types of models to evaluate which set of predictors generalize the best to out-of-sample data and explain the highest proportion of variance in the winning percentage response variable. We started with a baseline linear model that enforced our prior for which predictors would be important then further explored using regularization as well as decision trees, random forests, and mixed effects models. In addition to evaluating predictive performance, we chose models that were easily interpretable so that we could explore the association between important features and the response variable. Through this iterative process of improving our various types of models, we were able to better understand the relationships between each of the predictors and the response variables as well as create a ranking of the relative importances of the predictors. In addition to finding the exact ranking of variables in terms of importance in predicting the response variable, one other hypothesis that we wanted to assess was whether batting, pitching, or fielding statistics are most important.

2 Exploratory Data Analysis and Baseline Model

We scraped batting, pitching, and fielding statistics from `www.baseball-reference.com`. For each year T, we use individual players' statistics from year $T-1$ in order to predict the team's performance in year T. In particular, we collected data from each year from 1997 to 2022 to use $T = 1997, \ldots, 2021$. Then, we aggregated the individuals' statistics to get overall

team statistics, which we then used as the predictors for the next year's winning percentage. We also included some data that do not directly measure in-game performance, such as average age and quality and quantity of free-agent acquisitions, as part of our predictor set.

The source data listed each player's statistics individually and split the data into three categories: batting, pitching, and fielding. Thus, for each year of data we grouped the data by team and took the mean of the team's players' statistics weighted by the number of games each player played, making sure to drop rows such as league averages that do not refer to specific individuals. We then merged these data sets in order to get a final data frame where each row corresponds to a pair of team and year. Furthermore, missing data were filled in using mean imputation, though the exact method of imputation is likely unimportant since the missing data generally corresponds to individuals who play very few games and thus have low weights in our weighted averages. Similarly, we dropped rows with any predictor values that were `Inf`, as these corresponded to players who would have otherwise ended up with low weights in the data set otherwise. In practice, this meant that we dropped 11 pitchers who had 0 innings pitched, meaning that their `ERA`s, calculated as $9 \cdot \frac{\texttt{earned runs}}{\texttt{innings pitched}}$, was reported as `Inf`. Finally we split the data by random sample into a training set and a test set, where the train set is 80% of the total data set which we collected, giving us a training test size of $n_{train} = 536$.

2.1 EDA

First, we explored the response variable of interest, which is the winning percentage of a team. As expected, this response variable is symmetric and close to normally distributed and centered approximately at 50% (mean = 50.14%, median = 50.6%). There were also no clear outliers, so we did not transform the winning percentage.

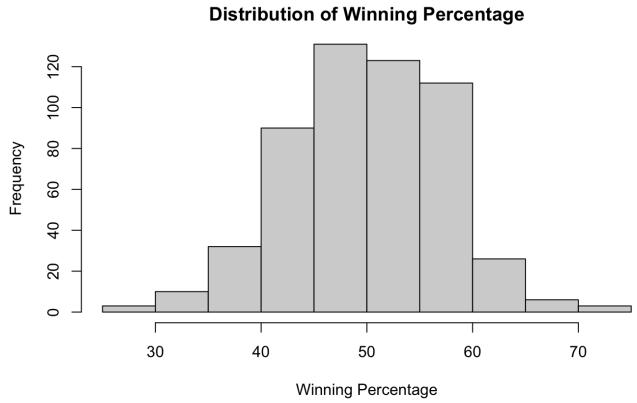

Figure 1: Distribution of winning percentage across all teams over the training dataset from 1997-2022.

We next explored the relationships between the response and some of the predictor variables. Specifically, we looked at variables which our prior tells us should be correlated with

our response variable While none of the plots of the response versus predictors are clearly linear, the signs of the associations between the data make sense intuitively. For example, batting averages and pitchers walks and hits per inning pitched (WHIP) both seem to have a positive association with winning percentage. We generally also tried to ensure that we used predictors that measured each of batting, pitching, fielding, and other team characteristics. However, we did not analyze predictors that directly measure runs scored, such as runs and earned run average. These two predictors are measures of the team's expected offensive and defensive strength, respectively, and thus we would expect these associations to be positive. Note that the signs of these correlations are not trivial since we are using the players' statistics from the previous year to predict the winning percentage of the current season; the fact that these associations are reasonable gives us reassurance that linear regressions will likely be a useful tool for a team's future performance. While this decision will likely reduce the predictive capabilities of future models, it will help to give interpretability beyond the obvious fact that scoring more runs and allowing less runs will lead to more wins.

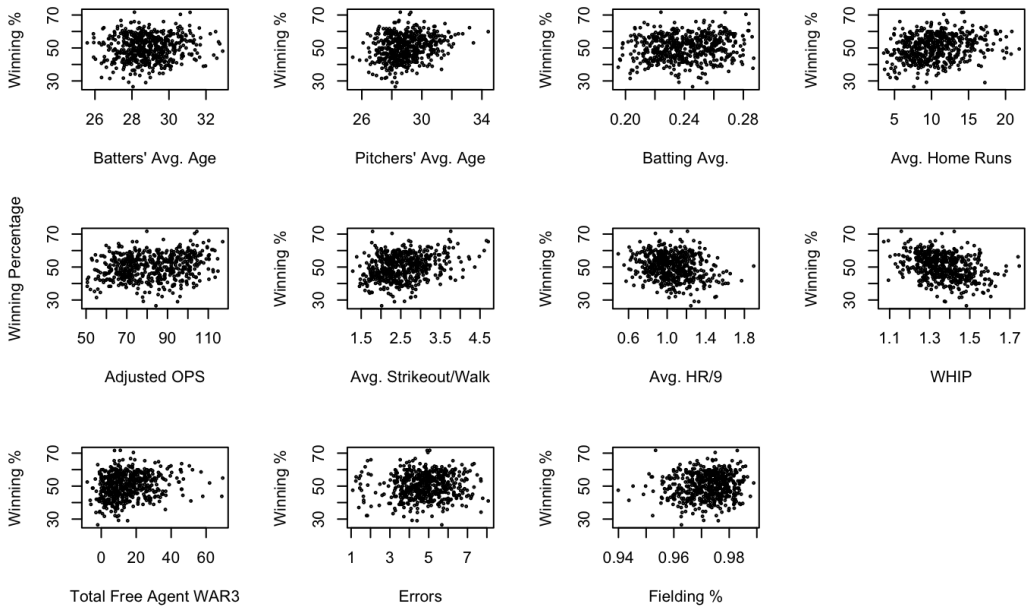

Figure 2: Scatter plots of the winning percentage vs. each of the baseline predictors.

Notably we find that most of the distributions of the response variable given a linear function of each predictor seems to roughly follow normal distributions with minimal outliers, and thus we probably do not need to transform these predictors.

2.2 Baseline Model

For our baseline model, we fit a multiple linear regression model on the 11 predictors we explored in the Exploratory Data Analysis.

The 11 predictors included in the multiple regression model are batter age, pitcher age, batting average (batters), home runs (batters), OPS (batters), strikeouts over walks (pitchers), HR/9, walks and hits per inning pitched (pitchers), total free agent WAR over the past 3 years, errors, and fielding %. These predictors are a mix of offensive and defensive statistics to encompass each of the important portions of the game. While there are many more options, we chose this set of predictors both because they seem intuitively important and because they each measure a different aspect of a team's players' skill sets.

In the model summary, we see that the coefficient estimates on `BA`, `OPS`, `HR9`, `WHIP`, `tot_fa_war3`, and `Fld%` are significant at the $\alpha = 0.05$ level. This validates our prior belief that most of these predictors would have a significant impact on the team's success. This baseline model has an R-squared of 0.293 and an adjusted R-squared of 0.278; the unadjusted value means that our linear model predicts 27.2% of the variance of the response variable. Since we do not anticipate transforming the response variable, we will compare all future models' performances to these values.

From the plots, we see that normality is roughly satisfied, though the residual plot shows some evidence of nonlinearity influenced by a few high outliers. Constant variance is also roughly satisfied. The data are not quite independent (an individual team is often similar between consecutive years), so we will use care when interpreting significance tests and build mixed effects models that control for this. However, this independence violation will not bias our estimates, so we will still be able to make meaningful predictions from our models.

When looking at the predictors, we find that the correlations between them are generally fairly low, with most being below 0.5 in magnitude. In particular, the correlations between the predictors within a predictor group (batting stats, pitching stats, or fielding stats) are largely similar in magnitude to the correlations between the predictors from different predictor groups. We expect that the offensive and defensive statistics should be independent of each other when conditioned on factors that affect the whole team, such as payroll. Since the within-group and between-group correlations are comparable in magnitude, this gives us some assurance that these predictors were chosen reasonably. In future analysis, we will look into the effect of other total team statistics and examine whether controlling for certain other factors reveals any insights into the association between the response variable and our current predictor set.

There is one pair of predictors, `BA` and `OPS`, which do have a high correlation (0.948), which makes sense given the similarities between the formulas used to calculate them. For future models, we will investigate whether we should continue to use both of them, only use one, or replace them with more fundamental measures of a team's ability to get on base and their power.

3 Methods and Results

3.1 Linear Regression

For our baseline model, we chose several predictors that we hypothesized would provide predictive power for winning percentage. In this section, we explore more sophisticated methods for feature selection in our multiple linear regression model that we hope will result in a better performance. We began by fitting a full model with all main effects and a full model with all main effects and the interaction terms. The main effects included all predictors in the data frame except for columns that we do not expect to provide predictive power, such as those related to `year`, and columns that are directly related to our response variable. These include wins (`W`), losses (`L`), and those related to `runs`, such as `RBI`, `ERA`, and `Rtot`. In Table 1, we see that the Full Main Effects Model performed worse on the test set when we compare the RMSE and R^2 with that of the Baseline model. It is also noticeably overfit as indicated by the RMSE being lower on the train than test set. The Full Interactions model performs extremely poorly on the test set, which is expected due to the very large number of predictors overfitting to the train set giving an extremely small train RMSE and 100% R^2. We chose RMSE to assess model performance on out-of-sample data and R^2 to provide interpretation into the proportion of variance in winning percentage that is explained by the predictors in the model. R^2 was chosen instead of adjusted R^2 in order to compare performance with the baseline and full predictor models.

The first method we used to alleviate the overfitting seen in the Full Main Effects and Full Interactions models was regularization with Ridge and LASSO. For both Ridge and LASSO, we created a well-tuned model by utilizing cross validation to hypertune the λ parameter, using values ranging from 10^{-4} to 10^4. The best λ values that were determined are listed under each model in Table 2. It makes sense that the penalty is higher for the models with interactions due to the greater model complexity. The train and test RMSE values for these models indicate that these models perform similarly and overfitting was reduced, but the Full Main Effects model with Ridge regression performed slightly better with a test RMSE of 6.929 and an R^2 value of 0.236.

Model	**Train RMSE**	**Test RMSE**	**Train R^2**	**Test R^2**
Baseline	6.274	7.059	0.293	0.207
Full Main Effects	5.669	7.189	0.423	0.178
Full Interactions	6.104×10^{-9}	157.610	1.000	-394.247
Full Main Effects with Ridge	5.796	6.929	0.397	0.236
Full Interactions with Ridge	5.742	6.962	0.408	0.229
Full Main Effects with LASSO	5.799	6.948	0.396	0.232
Full Interactions with LASSO	5.657	6.985	0.425	0.224
Stepwise Both Directions	5.729	7.082	0.411	0.202

Table 1: Linear Regression Model Comparison Metrics

In Table 2, we show the variable importance ranking for each of the models, which is

indicated by the predictors with the highest coefficients. The top 10 predictors seem to share several similarities, namely the main and interaction terms containing on-base percentage (`OBP`), fielding percentage (`Fld%`), and slugging percentage (`SLG`). These predictors make sense since they are expected to contribute to scoring more runs or performing better defensively.

The other method we used to reduce overfitting was stepwise variable selection using Akaike information criterion (AIC). We fit this using the Full Main Effects model and set the lower scope to be the intercept-only model and the upper scope to be the Full Interactions model. The final formula included 25 predictors, where the top 10 are listed in the last column of Table 2. The top predictors are slightly different compared to the Ridge and LASSO models, with the top three predictors being total wins above replacement from free agents signed in the past three years (`tot_fa_war3`), at bats (`AB`), and plate appearances (`PA`). These predictors make sense intuitively because they indicate high player value from free agents and more opportunities to score runs when at bats or appearing at the plate.

	Main Ridge $\lambda = 1.259$	**Interactions Ridge** $\lambda = 19.953$	**Main LASSO** $\lambda = 0.1$	**LASSO** $\lambda = 0.126$	**Stepwise** Both Directions
1	`OBP`	`Fld%`	`OBP`	`OBP:Fld%`	`tot_fa_war3`
2	`Fld%`	`BA:OBP`	`SLG`	`SLG:Fld%`	`AB`
3	`SLG`	`OBP:Fld%`	`WHIP`	`OBP`	`PA`
4	`OPS`	`OBP:SLG`	`Fld%`	`SH:SHO`	`num_fas`
5	`WHIP`	`OBP`	`SHO`	`IBB.pitch:BK`	`X3B`
6	`SHO`	`BA:SLG`	`BK`	`SF:HR9`	`H.bat`
7	`RF/G`	`OBP:OPS`	`RF/G`	`OBP:Age.pitch`	`WHIP`
8	`BK`	`BA:Fld%`	`H9`	`WHIP:SO9`	`PO`
9	`HR9`	`SLG:Fld%`	`X3B`	`WHIP:RF/G`	`A`
10	`H9`	`BA:OPS`	`HR9`	`X3B:HBP.pitch`	`Ch`

Table 2: Top 10 Selected Features

Overall, based on Table 1, we see that the best performing multiple linear regression model was the Full Main Effects model with Ridge regularization, with a test RMSE of 6.929 and an R^2 value of 0.236. Unlike LASSO, Ridge does not perform feature selection and sends unimportant predictor coefficients to near zero rather than zero. We output the coefficients for only the top ten predictors for this model seen in Table 2.

Predictor	Coefficient
(Intercept)	38.006
OBP	38.478
Fld %	18.457
SLG	15.601
OPS	10.724
WHIP	6.301
SHO	4.142
RF/G	2.620
BK	2.396
HR9	1.382
H9	1.274

Table 3: Coefficients for the Full Linear Regression Model with Ridge Regularization

Table 3 illustrates that the coefficients seem to drop off after only about 4 or 5 predictors. These top few predictors seem to provide the strongest associations with the response variable and explain most of the variability. This model has an intercept of about 38.006, which indicates that if all predictors are 0, then the expected winning percentage for a team is on average about 38.006%. The predictor with the largest predictor by far is `OBP`. This indicates that an increase in on-base percentage by 0.01 is on average associated with a $0.01 \cdot 38.5\% = 0.385\%$ increase in winning percentage. One possible explanation for why this predictor has the most predictive power is that it may be correlated with other predictors. One piece of evidence that supports this is that for the top predictors in the Interactions Ridge model in Table 2, many of the interaction terms include `OBP`, for example `BA:OBP`, `OBP:Fld%`, and `OBP:SLG`. Since the on-base percentage captures many metrics in its formula by measuring how frequently a batter reaches base per plate appearance, $\texttt{OBP} = \frac{\texttt{H+BB+HBP}}{\texttt{AB+BB+HBP+SF}}$, it makes sense that it would have a higher predictive power.

3.2 Decision Trees and Random Forest

We have also attempted to predict the winning percentage of a team in the next year using decision trees and ensemble methods. The reason for this is that one of the main advantages of decision trees and random forests are their ability to handle both continuous and categorical data, and to handle complex interactions between variables. They are particularly useful in situations where you need to make a prediction based on numerous, complex input variables, like baseball data. For all of our tree models, we followed a process where we create a simple one at first, and then fit a more tuned model using different hyper parameters.

Our first model was a simple decision tree (`tree1`), with a `cp` equal to 0 and a `maxdepth` of 20. The predictors used were the same from the full linear regression model, as they contain the most important predictors determined by the correlation matrix without the use of "run" related ones. The train RMSE for this data was 3.951, and the test RMSE for 8.432. This

indicates that our model was extremely overfit, as our train data was significantly lower than the test data. In addition, our test RMSE performed much worse than our baseline model, which shows that our model can be improved. Therefore, we pruned the `tree1` model using the best `cp` found, and we saw a significant improvement. The train RMSE was 6.461, while the test RMSE was 7.490. The model is less overfit, and our test RMSE, which we want to optimize, has decreased. However, since the test RMSE is still worse than our baseline model, we fit more models using ensemble methods.

To begin, we used a bagged model (`bag`) with the same predictors as the previous models. We see a big improvement from the decision trees, as our train and test RMSEs were 6.397 and 7.110, respectively. This shows that the model is not overfit or underfit, while maximizing its predicting abilities. To see if we can do better, we fit a random forest model (`rf1`) as well. We tuned for hyperparameters, such as `mtry` and `maxnodes`. The train RMSE for this model was 6.378, and the test RMSE was 7.138, using `maxnodes` of 100 and `mtry` of 15. Even though we expected the random forest model to perform better, it was actually slightly worse than the bagged model, in terms of the test RMSE. Furthermore, this model was adequately fit to the training data, similar to the bagged model. It was disappointing to see that none of these models performed better than the baseline model.

Despite not performing to our expectations, we still analyzed which variables were most important in determining for each model. The top 5 variables for the decision tree were `FIP`, `H9`, `W.L..sameyear`, `WHIP`, and `IBB.bat`, . For the bagged and the random forest model, the most important predictors were `tot_fa_war3`, `WHIP`, `SV`, `H9`,and `W.L..sameyear`. We were surprised to see `IBB.bat` in all 3 models, as we did not expect the number of intentional base on balls of the batters on a team to have a significant effect on predicting the win-loss ratio for the next year. The other variables were not as surprising, since they are directly correlated how well a team performs in a given year.

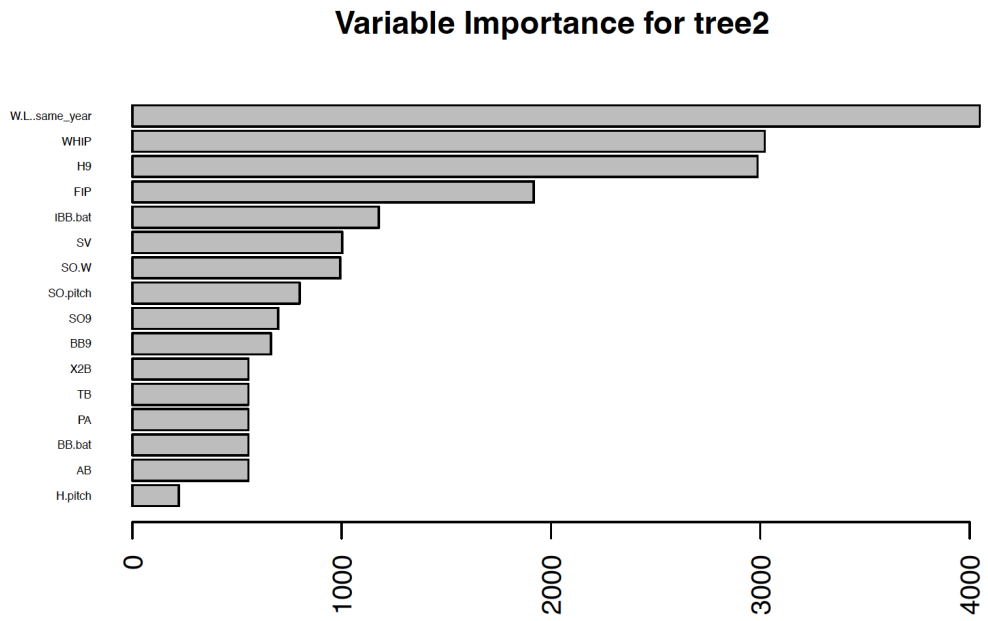

Figure 3: Variable importance plot for `tree2` model

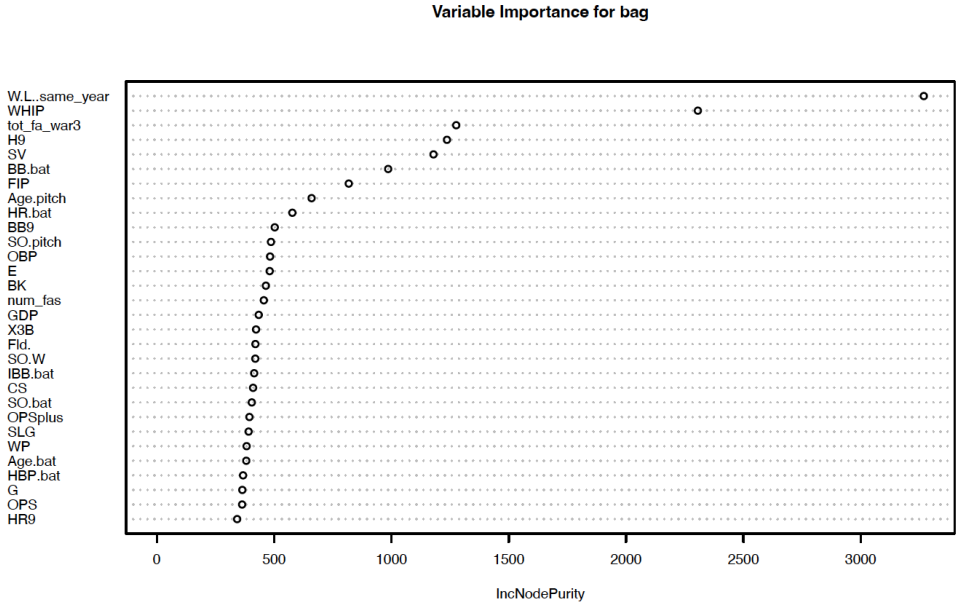

Figure 4: Variable importance plot for `bag` model

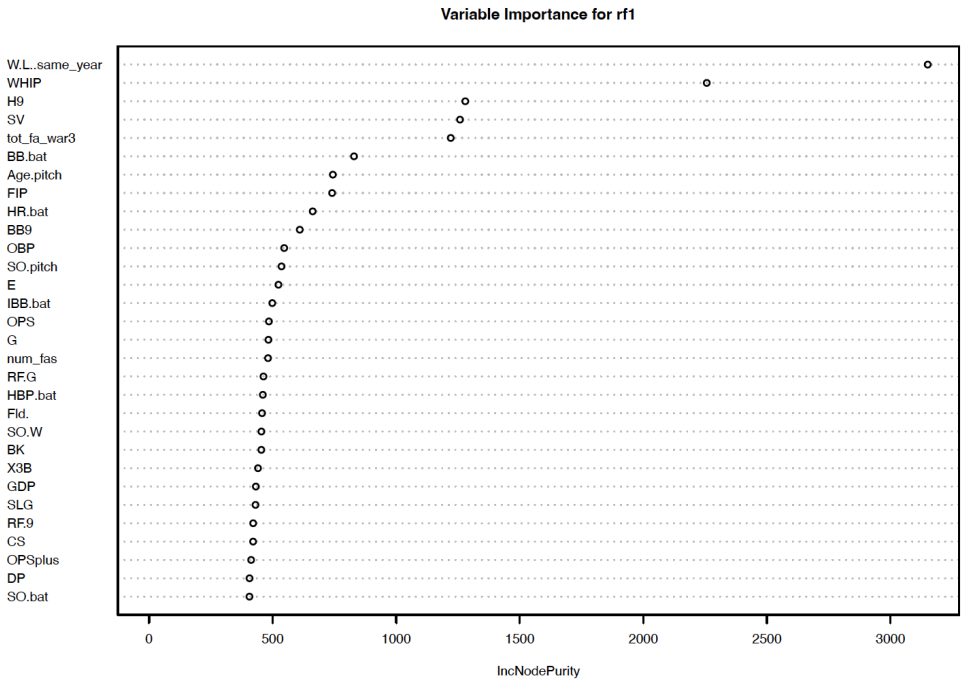

Figure 5: Variable importance plot for `rf1` model

In our last efforts to create a model that was better than our baseline, we re-trained our decision tree (`tree4`) and random forest (`rf2`) models with the top 5 important predictors from the previous random forest model: `tot_far_war3`, `H9`, `SV`, `WHIP`, and `W.L..sameyear`. However, these models all performed worse than our existing ones. For reference, the table below contains all of the train and test RMSEs of the models we have tested:

Model	**Train RMSE**	**Test RMSE**
`tree1`: Full Main Effect Tree	3.951	8.432
`tree2`: Full Main Effect Pruned Tree	6.461	7.490
`bag`: Full Main Effect Bagged	6.397	7.110
`rf1`: Full Main Effect Random Forest	6.378	7.138
`rf2`: Top 3 Variable Random Forest	7.138	7.484
`tree4`: Top 3 Variable Pruned Tree	6.686	7.653

Table 4: Decision Tree and Random Forest Model Comparison Metrics

3.3 Mixed Effects Model

We used multiple years of baseball data in our training data set to fit our models. Since we have multiple rows of observations that correspond to each of the MLB teams, there inherently exists a clustering in our data. Given that team and player statistics will tend to be correlated within each team, these observations are likely no longer independent.

Therefore, we decided to fit several mixed effects models with team as a group in order to control for independence.

For the mixed effects model, we attempted to predict the winning percentage of a team in the next year using the top ranked variables marked as most important by the previous ridge, lasso, random forest, and decision tree regressions. For the four different models, we created a mixed effects model using the most important variables in each respective model as fixed effects and a random intercept and a mixed effects model with both random effects and fixed effects as the most important variables.

We used the following variables for each of the mixed effects models: `OBP`, `Fld%`, `SLG`, `OPS`, and `WHIP` ("Ridge Full"); `Fld%`, `BA:OBP`, `OBP:Fld%`, `OBP:SLG`, and `OBP` ("Ridge Full Interaction"); `OBP`, `SLG`, `WHIP`, `Fld%`, and `SHO` ("LASSO Full"); `OBP:Fld%`, `SLG:Fld%`, `OBP`, `SH:SHO`, and `IBB.pitch:BK` ("LASSO Full Interaction"); `tot_fa_war3`, `AB`, `PA`, `num_fas`, and `X3B` ("Step"); `WHIP`, `W.L..same_year`, `SV`, `tot_fa_war3`, and `H9` ("Bagged").

In addition, we compared the train and test RMSEs of each mixed effects model with and without random effects in Figure 6. The LASSO Full mixed effects model performs the best with the lowest test RMSE of 6.831 which preforms better than the baseline model. In addition, the difference between the train and test RMSE of the LASSO Full mixed effects model is 0.898, thus it is on the less over-fit side in comparison to all the mixed effects models.

Models	**Train RMSE**	**Test RMSE**
Ridge Full (Fixed + Random Coefs.)	5.490	6.967
Ridge Full (Fixed + Random Intercept)	5.947	6.929
Ridge Full Interaction (Fixed + Random Coefs.)	5.924	7.082
Ridge Full Interaction (Fixed + Random Intercept)	6.212	7.018
LASSO Full (Fixed + Random Coefs.)	5.436	6.938
LASSO Full (Fixed + Random Intercept)	5.933	6.831
LASSO Full Interaction (Fixed + Random Coefs.)	5.843	6.929
LASSO Full Interaction (Fixed + Random Intercept)	6.158	7.061
Step (Fixed + Random Coefs.)	5.763	7.124
Step (Fixed + Random Intercept)	6.141	7.103
Bagged (Fixed + Random Coefs.)	5.562	7.370
Bagged (Fixed + Random Intercept)	6.004	7.395

Figure 6: Mixed Effects Model Comparison Metrics

For the model that performed the best on the test set, Figure 7 shows the relevant coefficients and intercept of the fixed effects. Furthermore, the optimal random effect for the intercept had a standard deviation of 2.742.

Predictor	Coefficient
(Intercept)	33.292
OBP	48.052
SLG	33.469
WHIP	-21.499
Fld %	18.524
SHO	6.184

Figure 7: Coefficients of the Fixed Effects in the best performing model: the mixed effects model fixed effects for the best predictors identified by the LASSO model trained on the main effects of all the predictors and a random effect for only the intercept.

The intercept of this model is 33.292 which means that if all predictors are 0, then the expected winning percentage for a team will be on average 33.292%. The largest predictor in this model is OBP which indicates that a one percentage point increase in on-base percentage will result in a 48.052 increase in winning percentage. However, OBP does not increase by one percentage point and will usually increase by some fraction. For example, an increase in OBP by 0.01 percentage points will result in a 0.4805 increase in winning percentage. Similar to what we observe in the full linear regression model with ridge regularization, this is likely because OBP is highly correlated with other predictors.

4 Conclusion and Discussions

From the results in Figure 6, we saw that the best performing model out of all the models we explored was the LASSO full model with all fixed effects and only a random intercept. In particular, this model started as a linear regression using all predictors, from which the top few were chosen to build the mixed effects model. Furthermore, the grouping variable used was the teams so that we had a linear model where the intercept was different between teams and the coefficients on each of the predictors were the same between teams.

The fact that using only the intercept as a random effect performed better than using all predictors leads to a convenient and intuitive interpretation. Namely, each team has a different baseline ability, which encompasses factors such as the quality of ownership and management, but that each team is affected equally by a change in the other predictors. For example, in from Figure 7 we see that an increase in the team's average slugging percentage by 0.01 is expected to increase their winning percentage by $0.01 \cdot 33.5\% = 0.335\%$, regardless of which team's success we are trying to predict. This both makes sense – we expect an improvement in the quality of players to have a relatively similar effect on all teams – and also is helpful in generalizability of the interpretation of the model effects to other scenarios. For example, if a new team joins the league, while we will not know their baseline random intercept, we will be able to predict the effect that a change in the important predictors will have on the team's winning percentage. Thus, the answer to our original question – which predictors are most important in projecting the team's winning percentage – can still be

applied to previously unobserved teams.

We also see that the directions of the coefficients make sense: good players are generally expected to have high values of `OBP`, `SLG`, `Fld %`, and `SHO`, while good pitchers usually have lower values `WHIP`. Furthermore, the distribution of the random intercept effects, which has standard deviation of 2.742, indicates that the difference between the top team's intercept (roughly top 2.5%) and the bottom team's intercept (roughly bottom 2.5%) is expected to be $1.96 \cdot 2.742\% - (-1.96) \cdot 2.742\% = 10.75\%$. In practice, the best and worst teams generally have a larger gap between their winning percentages as good teams often have good management and other non-player factors (high intercept) as well as good players (predictor values that maximize winning percentage prediction).

One surprising observation was that none of the decision trees nor the random forest models performed better than the baseline models. We believed that the pruned trees or the well-tuned random forest model would perform better than the baseline multi linear regression model at the minimum. One of the possible underlying reasons for this phenomena is that our response variable that we are trying to predict, `W.L..next.year`, is a continuous variable that ranges from 0 to 1. Therefore, this value fluctuates based on linear relations between your features, rather than having clear cut "splits" that determine if you have a higher or lower win-loss ratio.

Lastly, from our results, it is difficult to conclude whether fielding, batting, or pitching statistics are the most important in determining whether a team performs well. As we have seen from our linear model coefficients and variable importance plots for tree models, there is a wide variety of predictors that contributed to win percentage. One commonality is that fielding, batting, and pitching statistics all tell us a story about whether the team is scoring more runs or giving up more runs. For example, a team may look poor on paper if they have a high `WHIP`, but they might be able to compensate for that by scoring more runs through a high `OBP`. As a result, it is difficult to pinpoint which aspect of baseball is the most important from this analysis, but it is safe to say that being a well-rounded team in all three areas is crucial in winning games.

A Glossary of Terms

A.1 Batting Statistics

- `BatAge` – Batters' average age
- `Weighted` by AB + Games Played
- `R/G` – Runs Scored Per Game
- `G` – Games Played or Pitched
- `PA` – Plate Appearances
- `AB` – At Bats
- `R` – Runs Scored/Allowed
- `H` – Hits/Hits Allowed
- `2B` – Doubles Hit/Allowed
- `3B` – Triples Hit/Allowed
- `HR` – Home Runs Hit/Allowed
- `RBI` – Runs Batted In
- `SB` – Stolen Bases
- `CS` – Caught Stealing
- `BB` – Bases on Balls/Walks
- `SO` – Strikeouts
- `BA` – Hits/At Bats
- `OBP` – (H + BB + HBP)/(At Bats + BB + HBP + SF)
- `SLG` – Total Bases/At Bats or (1B + 2*2B + 3*3B + 4*HR)/AB
- `OPS` – On-Base + Slugging Percentages
- `OPS+` – OPS+ $100 * [OBP/lgOBP + SLG/lgSLG - 1]$ Adjusted to the player's ballpark(s)
- `TB` – Total Bases = $\texttt{Singles} + 2 \cdot \texttt{Doubles} + 3 \cdot \texttt{Triples} + 4 \cdot \texttt{Home Runs}$
- `GDP` – Double Plays Grounded Into

- `HBP` – Times Hit by a Pitch.
- `SH` – Sacrifice Hits (Sacrifice Bunts)
- `SF` – Sacrifice Flies
- `First` tracked in 1954.
- `IBB` – Intentional Bases on Balls
- `LOB` – Runners Left On Base
- `#Bat` – Number of Players used in Games

A.2 Pitching Statistics

- `#P` – Number of Pitchers used in Games
- `PAge` – Pitchers' average age
- `Weighted` by 3*GS + G + SV
- `RA/G` – Runs Allowed Per Game
- `W` – Wins
- `L` – Losses
- `W-L%` – Win-Loss Percentage $\texttt{W}/(\texttt{W}+\texttt{L})$
- `ERA` – $9 \cdot \texttt{ER}/\texttt{IP}$
- `G` – Games Played or Pitched
- `GS` – Games Started
- `GF` – Games Finished
- `CG` – Complete Game
- `tSho` – Shutouts by a team No runs allowed in a game by one or more pitchers.
- `cSho` – Shutouts No runs allowed and a complete game.
- `SV` – Saves
- `IP` – Innings Pitched
- `H` – Hits/Hits Allowed

- `R` – Runs Scored/Allowed
- `ER` – Earned Runs Allowed
- `HR` – Home Runs Hit/Allowed
- `BB` – Bases on Balls/Walks
- `IBB` – Intentional Bases on Balls
- `SO` – Strikeouts
- `HBP` – Times Hit by a Pitch.
- `BK` – Balks
- `WP` – Wild Pitches
- `BF` – Batters Faced
- `ERA+` – ERA+ $100 * [\texttt{lgERA}/\texttt{ERA}]$ Adjusted to the player's ballpark(s).
- `FIP` – Fielding Independent Pitching
  - this stat measures a pitcher's effectiveness at preventing `HR`, `BB`, `HBP` and causing `SO`
  - $(13 * \texttt{HR} + 3 * (\texttt{BB} + \texttt{HBP}) - 2 * \texttt{SO})/\texttt{IP} + \texttt{Constant}$
  - The constant is set so that each season major-league average `FIP` is the same as the major-league avg `ERA`
- `WHIP` – (BB + H)/IP
- `H9` – 9 x H / IP
- `HR9` – 9 x HR / IP
- `BB9` – 9 x BB / IP
- `SO9` – 9 x SO / IP
- `SO/W` – SO/W or SO/BB
- `LOB` – Runners Left On Base

A.3 Fielding Statistics

- `Rk` – Rank
- `Name` – Player Name
- `Age` – Player's age at midnight of June 30th of that year
- `Lg` – League
- `AL` - American League (1901-present)
- `NL` - National League (1876-present)
- `AA` - American Association (1882-1891)
- `UA` - Union Association (1884)
- `PL` - Players League (1890)
- `FL` - Federal League (1914-1915)
- `NA` - National Association (1871-1875)
- `ANL` - American Negro League (1929)
- `ECL` - Eastern Colored League (1923-1928)
- `EWL` - East-West League (1932)
- `NAL` - Negro American League (1937-1948)
- `NNL` - Negro National League (1920-1931)
- `NN2` - Negro National League 2 (1933-1948)
- `NSL` - Negro Southern League (1932)
- `G` – Games Played or Pitched
- `GS` – Games Started
- `CG` – Complete Game
- `Inn` – Innings Played in Field
- `Ch` – Defensive Chances = `Putouts` + `Assists` + `Errors`
- `PO` – Putouts
- `A` – Assists

- `E` – Errors Committed
- `DP` – Double Plays Turned
- `Fld %` – Fielding Percentage = (`Putouts` + `Assists`)/(`Putouts` + `Assists` + `Errors`)
- `Rtot` – Total Zone Total Fielding Runs Above Avg
- `Rtot/yr` – Total Zone Total Fielding Runs Above Avg per 1,200 Inn
- `Rdrs` – BIS Defensive Runs Saved Above Avg
- `Rdrs/yr` – BIS Defensive Runs Saved Above Avg per 1,200 Inn
- `Rgood` – BIS Good Plays/Misplays Runs Above Avg
- `RF/9` – Range Factor per 9 Inn = $9 * ($`Putouts` + `Assists`)/`Innings Played`
- `RF/G` – Range Factor per Game = (`Putouts` + `Assists`)/`Games Played`
- `Pos Summary` – Positions Played