2024 MLB Managerial Statistical Analysis

Dani DiTomasso, Data Analytics 2024 Term Project Report

Introduction

Within the sport of baseball at the professional level, managers are a huge determinant on how a major league team will perform for any given season. They make imperative decisions including player subbing, types of plays, overall game strategy, and so much more. It is for this reason that the data focuses on and studies team statistics based on strategic plays called by the manager. Given this information, the hypothesis for this study is the more aggressive offensive plays a manager chooses to take, the more success the team will have overall. Note that "team success" is measured by any team having a Win/Loss percentage that is higher than 53%.

Data Description

For this study, two datasets were compared and combined in order to run a meaningful statistical analysis. Both of these datasets were obtained from *Baseball Reference*, a website that posts real-time statistics about all major facets of baseball, including players, teams, seasons, leaders, scores, playoffs, and many more. This project looks at the 2024 season statistics for all managers of Major League Baseball teams, which includes two main datasets: "2024 Manager Records" and "2024 Managerial Tendencies."

Within the first dataset, "2024 Manager Records", the statistical data mostly contained factors that described the "effectiveness" of the manager, or what the data would be compared to to measure how successful a manager's performance was in regards to the team. This includes manager name, team abbreviation, wins, losses, wins/loss percentage, ties, games played, team finish, postseason wins, postseason losses, postseason wins/loss percentage, challenges, overturned, overturn percentage, and ejections. For this study, the wins/loss percentage was mainly used to be compared against for determining a "successful" team manager compared to an "unsuccessful" one.

Furthermore, for the second dataset "2024 Managerial Tendencies" contained data specifically related to the decisions managers have to make during the 2024 season. Unlike the first dataset, this one contained two layers of headings, with the first layer containing the main groupings of data: these included stealing second, stealing third, sacrificial (sac) bunts, intentional walks, and substitutions. Within each of these headings are subheadings that pertain to the individual data points. For example, under stealing second there are 4 main subheadings which include "Ch" (chances to steal second), "Att" (attempts to steal second), "Rate" (Rate of attempting to steal second) and "Rate+" (League-adjusted steal of 2nd rate). Each major heading has its own group of subheadings that pertain specifically to that managerial decision made.

In total, there are thirty three teams within major league baseball - this means that there were around thirty three rows of data, not including any totals calculated at the bottom or headings of the dataset. All of the data for one manager was contained in a singular row, and the rows were numbered under the "Rk" (or rank) column. Additionally, both datasets were ordered the same way, so combining the data into one dataset didn't pose a significant challenge. A link to the current view of the datasets found on *BaseballReference.com* can be found here.

Preliminary Analysis

The datasets were originally accessed on October 1st, 2024. This means that although the regular MLB season was coming to a close, the post-season statistics were not yet made available since the playoffs to the 120th World Series had yet to start. It is for this reason that the analysis does not include any post-season statistics and these blank columns were dropped from the Manager Records 2024 dataset. However, it is important to note that although these statistics were not used directly, any teams that had made the World Series playoffs were used as a benchmark for a "successful" or "winning" team / manager. Essentially, every team that had progressed far enough within the season to make it to the bracket play for the World Series (twelve teams in total), had a win-loss percentage higher than fifty three percent. When comparing this to the rest of the teams however, only one additional team had above a fifty three percent out of the total of thirty three teams (thirteen in total out of thirty three). It is for this reason that the post-season win-loss percentage was used as a benchmark for deeming a team as either "winning" or "losing" based on whether or not the overall percentage was above or below fifty three percent. It can be argued that only the top performing teams within Major League Baseball get to compete for the World Series title, and it is with this logic that this benchmark was used within this preliminary analysis.

Data Cleaning & Merging

To achieve a usable dataset for statistical analysis, a few steps had to be taken before creating any models and visualizations. Firstly, each of the datasets were read in as a csv file and stored into two separate variables, one named "record" for the 2024 Manager Records dataset, and one named "tendencies" for the 2024 Managerial Tendencies dataset. Then data cleaning was conducted for each dataset before ultimately merging the two together.

For the "record" dataset, the first cleaning action taken was to drop the postseason statistics columns, which were strictly filled with NA values. Once these three columns were erased, no other NA values were found within either datasets. The next step was to remove the last row of the dataset, which contained the overall totals for each of the columns. Although this data could be useful, it would incorrectly skew the data by the predictive models assuming this data was the result of a team or manager, which would largely inflate the results. If any sums of variable data was needed, then it was later calculated using the "sum()" function within the program. And finally, an additional column of data was created for this dataset called "status", in which if a team had above a 0.53 win-loss percentage they would be assigned the value of "winning", where below or equal to 0.53 would be assigned the value of "losing", as discussed in the preliminary analysis above. This column of added data was used in future models to determine overall manager success based on certain variables specified.

Due to the nature of how the "tendencies" dataset was created, a significantly greater amount of cleaning had to be completed. The first step was to remove the last row as it only contained the sum of all the variables, and for the same reasons mentioned above for the "records" dataset this information is not desired. The next step was to fix the headings that were automatically loaded within the dataset – since there were five large headings with subheadings underneath each one, the individual data points were inaccessible. To remedy this issue, the first row of the dataset was dropped (which contained the subheadings) and the actual heading themselves were overwritten manually. It's important to note that within the original dataset, many of the subheadings had the same name since they depict the same type of statistic (for example, "Ch" stands for chances to steal second, chances to steal third, and chances for sacrificial bunts). To avoid confusion, repeating variable names were given a new simplistic naming convention where a number was added to each iteration of the variable ("Ch" for second base, "Ch.1" for third base, and "Ch.2" for sacrificial bunts, for example). In order to keep track of these variable changes, a document was created that lists all of the column headings, along with their meaning (click here). Furthermore, all numeric variables within this dataset were set to be "character" type, which would have made conducting any models impossible to complete. To change this, any "%" symbols found within the dataset were dropped and all numeric columns were changed to the type of "numeric." Finally, the first few columns of the "tendencies" dataset exactly match columns within the "records" dataset. To avoid repeating data, these columns (rank, manager, team, and games played) were dropped from the dataset.

To combine both of these datasets together, it was imperative to ensure that the data for the rows in "records" correspond to the data in "tendencies" for the same manager. Since both of these datasets were obtained through the same source, they were ordered in the same way – meaning that no data manipulation had to occur to get the datasets to match for the rows of the manager data. Each dataset had thirty three rows, and the "tendencies" dataset was binded to the right of the "records" dataset to create a new dataset called "combined." This dataset had all of the data for each manager in one row, unique column names, and no NA values, allowing for ease of model development for the statistical analysis.

Exploratory Data Analysis

To gain a better understanding of the pre-established trends and significant data points found within the dataset, an exploratory data analysis was conducted. Since the variable "W.L." (the Win-Loss percentage rate) was used most frequently to determine the success of a manager, this variable was the first to be visualized in the form of a histogram. As seen in figure 1, most teams fell between a forty to sixty percentage rate, with only a few teams being below forty and above sixty percent. Taking this into account, this variable was then divided into two main groups of "winning" and "losing" based on the 0.53 cutoff, as described in the previous sections. The results of this data can be seen in the adjacent pie chart in figure 1, thirty nine percent of teams have a "winning" record, whereas sixty one percent have a "losing" record. These two visualizations offered a deeper understanding of the variable decided to determine the overall success of a manager, which was used in all models within the data analysis.

For most of the remaining data groups, such as stealing second base, stealing third base, intentional walks and sacrificial bunts, the rate percentage variable was used to display a general overview of each of these variable groups. More specifically, the "Rate" percentage variable is a value that is calculated by the rate of attempting to complete the chosen action (stealing second, intentional walks, etc) divided by the total number of chances to complete that action within the 2024 season. Due to the calculations used to achieve the Rate percentage, this variable appeared to give a general overview of each of the larger data groups and was chosen to provide a deeper understanding of the data in the form of histograms. As seen in figure 2, an interesting observation is the rate of trying to steal second base is around ten percent, whereas the rate of trying to steal third drops between one to two percent, showing a drastic decrease. Both the rates of attempting intentional walks and sacrificial bunts are below one percent, showing it is not a common decision that most managers tend to make during the season for their team.

Model Development

For the bulk of the data analysis, four main models were developed in order to determine the effectiveness of managerial decisions for team success. These models include linear regression models, decision trees models, a random forest model, and a naive bayes model. Each model included taking the "W.L." variable and running it against a specific set of other variables, in order to determine if they had any significant impact on the win-loss percentage outcome.

Looking at the first type of model, six separate linear models were run independently. This was done to determine the most significant variable (the lowest p-value) for each grouping of data (manager records, second base, third base, sac bunts, intentional walks, and substitutions). The linear models included all available variables under each of these data groupings in order to determine which specific value had the most impact on W.L. percentage, regardless whether or not the impact was positive or negative. For the first linear model based on managerial records, the value to have the most significant impact was the “Finish” value which returned a p-value of 1.42e-06, indicating a very small number. Within the dataset, the “Finish” value is representative team finishing rank, which for career totals of managers, these are the average of all years weighted by the number of games played or managed. Teams with a lower Finish value (or a higher rank) overall had a higher W.L. statistic, correlating to the fact that they had won more games. The second linear model compared all second base statistics, and returned the “Ch”, or total chances to steal second base as being the most statistically significant with a p-value of 0.0276. Furthermore, the total number of chances for both stealing third base and sacrificial bunts also had the most statistically significant number, with “Ch.1” returning a value of 1.23e-05, and “Ch.2” returning a p-value of 9.66e-05. These results indicate that the amount of chances available to make a play have more of an impact on the win-loss percentage than the actual decision to make said play. Additionally, for the data concerning intentional walks, two variables including “IBB” (which is the total number of intentional walks) and “Rate.3” (the rate of intentional walks divided by total number of plate appearances) were closely significant to each other, with “IBB” returning a p-value of 0.00188, and “Rate.3” returning 0.0031. Finally, no significant values or values with a p-value less than 0.05 were returned for any variables regarding the data for substitutions made, so these variables were not taken into account for any future model analysis. However, all of the above variables that were singled out for being the most significant were the main variables used within the next three model developments, in order to return results that had the highest possible impact on determining managerial success. The summary of all of the linear regression models created can be seen in figure 3 below.

Moving onto the second model, decision trees, seven total models were created: these include manager records, stealing second, stealing third, sacrificial bunts, intentional walks, and most significant variables both including and excluding the “Finish” variable specifically (the “Finish” variable was excluded purely to yield greater decision tree depth). A decision tree is “a flowchart-like structure used to make decisions or predictions. It consists of nodes representing decisions or tests on attributes, branches representing the outcome of these decisions, and leaf nodes representing final outcomes or predictions” (GeeksForGeeks). These models were utilized within this study to have a more detailed view of how the variables influenced the win-loss percentage, and if the more aggressive decisions made by managers had any significant impact on their teams’ outcome. Summarizing the significant ideas from figure 4 below, which include the decision trees for manager records, stealing second, stealing third, sacrificial bunts, and intentional walks data, if teams attempted to steal third base more than ten times but less than twenty times, then they were likely to have a “winning” record. Additionally, for the manager records, if “Finish” had a value less than or equal to two, then the team would also likely have a “winning” record. However, for intentional walks, if the league-adjusted rate for intentional walks taken divided by the total number of opportunities is less than sixty six, then the team is likely to have a "winning" record, directly contradicting the hypothesis that managers who decide on more aggressive offensive plays will be more likely to have a better record. Moving onto figure 5, these two decision trees show that the most influential variables to having either a "winning" or "losing" variable is either determined by the team's overall finishing rank, and the total amount of chances presented to make a certain play within the season, not any direct decision a manager may take. To further support this argument, a prediction model was created for the last two decision trees, and each returned a mean squared error value of about 0.002 to 0.003, indicating a better model and that the models predictions are very close to the actual values. However, an interesting note is that within the diagrams, the decision trees viewed any "winning" result as having a win-loss percentage being over 0.51, instead of the predetermined 0.53, as can be seen in the final nodes that are highlighted in green. Perhaps this is due to small errors within the predictive model, as the r-squared values were 0.77 and 0.63, indicating some variability despite the extremely low mean squared error values.

The third model, a random forest, was run specifically on variables that were previously chosen to be most significant. A random forest is "a commonly-used machine learning algorithm…that combines the output of multiple decision trees to reach a single result" (IBM). This model was extremely useful for solidifying the findingings of the previous two model types and returned a summary of the most important variables. Within these plots shown in figure 6, the "Finish" variable is by far the most significant variable with %IncMSE (percent increase of mean squared error based on how much the model's prediction accuracy decreases when that variable is removed), and IncNodePurity (increased node purity which reflects how much a variable contributes to reducing the impurity of the nodes in the decision trees of the random forest model). The remaining variables switch from medium to low significance between these two plots, showing low impact on the win-loss percentage variable. Furthermore, a prediction model was run for the random forest, and returned a mean squared error value of about 0.001, and a r-squared value of about 0.85, showing a very accurate representation of the predicted values when compared to the actual values.

Finally, the fourth model run on the most significant variables is a Naive Bayes. A Naive Bayes classifier "assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature" (Ray). This model was particularly useful in determining the effectiveness of the "W.L. variable, without taking into account the other most significant variable determined previously unlike the three previous models. When a prediction model was created and a confusion matrix was returned, the results appeared to be highly accurate, with only four out of thirty three variables being predicted incorrectly. Additionally, as seen in figure 7, the precision recall curve shows an accuracy of about 75%, which can be determined by calculating the area under the curve. This accuracy is further supported by the ROC curve, in which the more the values are closer to the top left corner of the graph, the better the predictive model is at accurately sorting the data into the correct groupings (Bobbitt).

Overall, all four of these models were developed successfully. They identified the most significant variables to the win-loss percentage value, and then based on these findings, allowed the viewer to make interpretations about how successfully (or unsuccessfully), a manager who makes more aggressive play calls might perform based on overall team statistics.

Conclusion

Based on the results of the models developed in this study, it was discovered that the null-hypothesis could not be rejected. In other words, the data analysis showed that aggressive offensive plays called by a team's manager did not yield any greater team success. It was revealed that the most influential variables on a teams win-loss percentage was the overall team finishing rank, and the total number of chances to make plays, such as stealing second, stealing third, and sacrificial bunts. It can be argued that a strong and successful MLB team will make more opportunities and have more chances to steal second, steal third, and make more plays, which may be why these variables have the most significance on the win-loss percentage ratio. However, this is beyond the scope of this data analysis and does not pertain to the hypothesis established at the beginning of this study. Additionally, this is not to say that team managers do not have important jobs, and that their decisions do not impact a team's standings. Within this dataset, it could not be proven that risker / more aggressive plays definitively increase a teams or managers overall success.

Discussions

Within this study, many important lessons were learned due to trial and error to receive the most accurate results possible. For example, one linear model was originally developed with all possible variables, which led to no data types being statistically significant. This prompted the realization that linear models are more accurate and successful when run with smaller chunks of data, hence multiple linear models being developed for different headings of data. Additionally decision trees are an extremely useful data analysis, specifically for determining the individual impact of a given variable, while also being influenced by other data. Within this study, decision trees were crucial in assessing the influence of aggressive plays on win-loss ratios. Finally, and perhaps most importantly, it was found that data cleaning and merging is critical for achieving an effective analysis. Having a clean dataset made the analysis and developing models much easier and faster, and caused less errors later in the process.

Figures

If a better view of each visualization is required, please click here to view a presentation slide deck, with the figures displayed on each slide with more clarity.

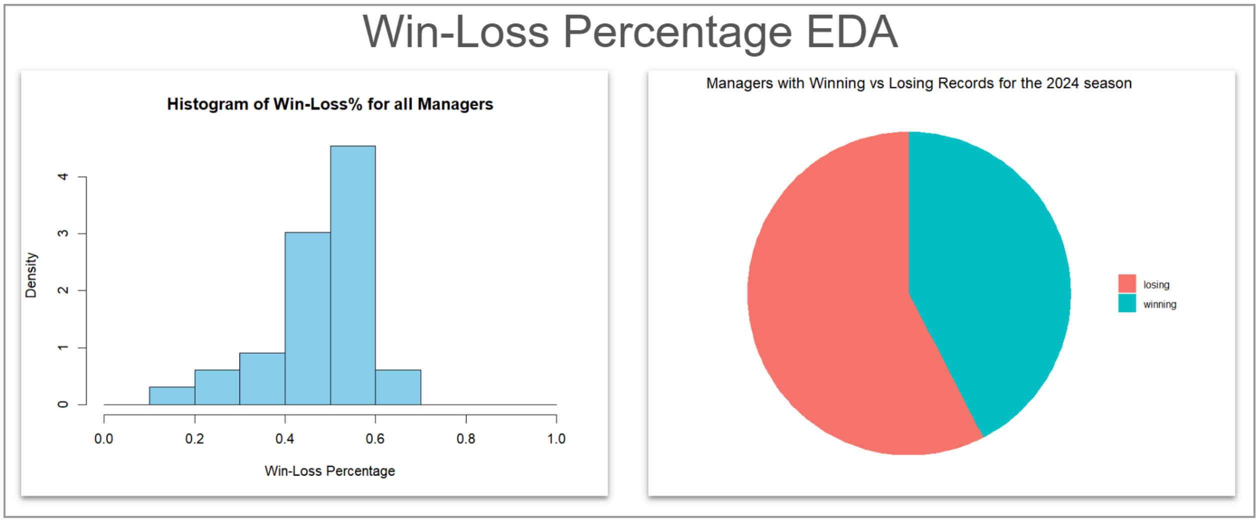


Figure 1: Exploratory Data Analysis for the Win-Loss percentage variable.

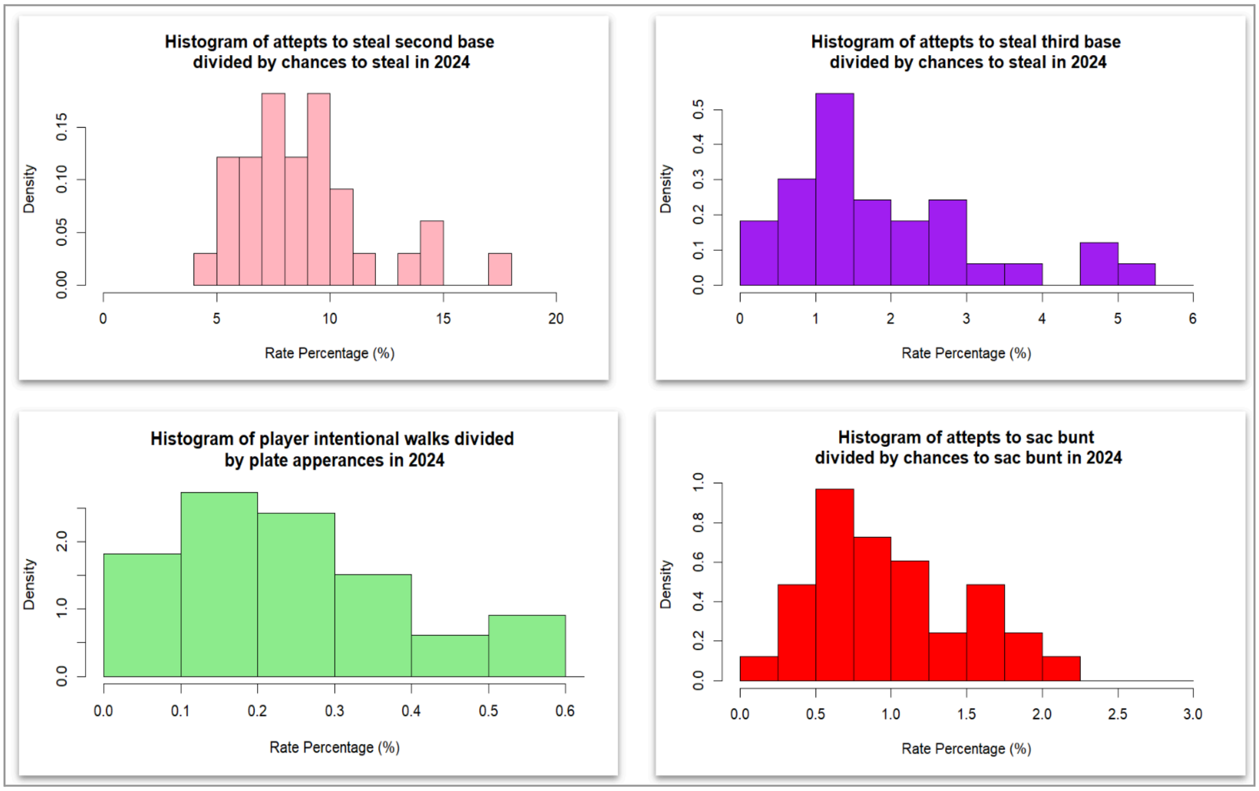
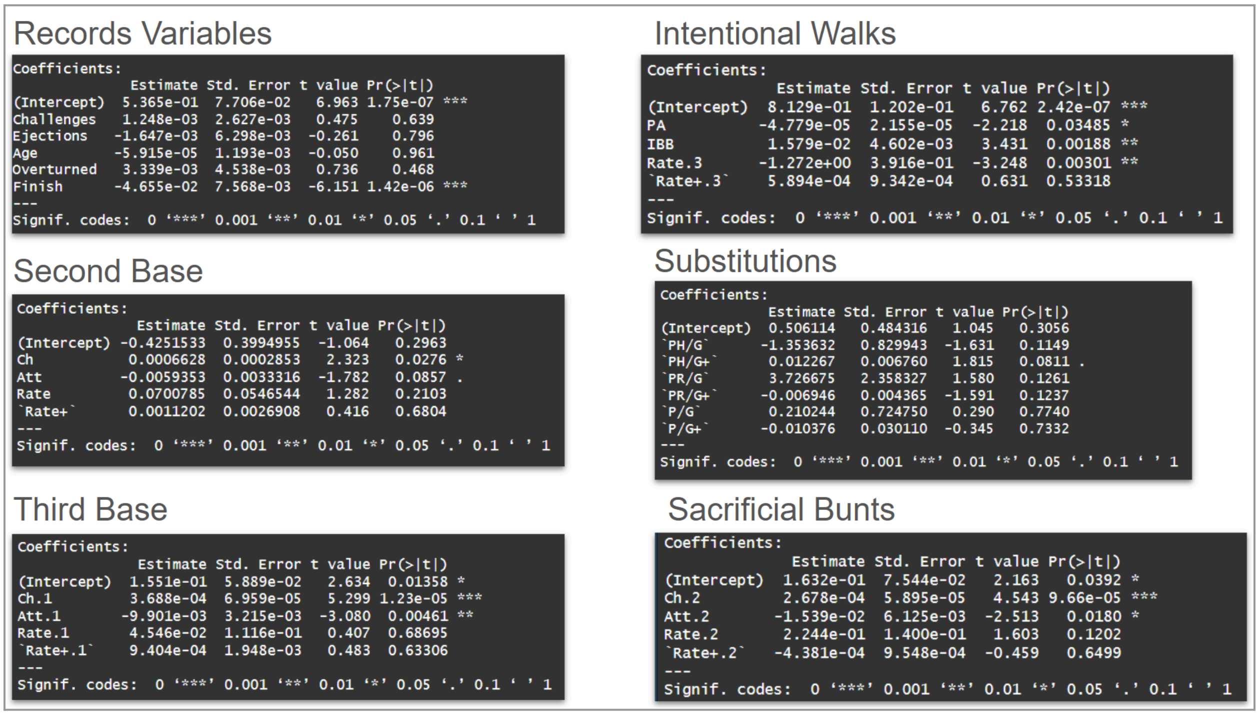


Figure 2: Exploratory Data Analysis the Rates of other main variables and data groupings.



Records Variables

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.365e-01  7.706e-02   6.963 1.75e-07 ***
Challenges   1.248e-03  2.627e-03   0.475    0.639    
Ejections   -1.647e-03  6.298e-03  -0.261    0.796    
Age         -5.915e-05  1.193e-03  -0.050    0.961    
Overturned   3.339e-03  4.538e-03   0.736    0.468    
Finish      -4.655e-02  7.568e-03  -6.151 1.42e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Second Base

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) -0.4251533  0.3994955  -1.064   0.2963  
Ch           0.0006628  0.0002853   2.323   0.0276 *
Att         -0.0059353  0.0033316  -1.782   0.0857 .
Rate         0.0700785  0.0546544   1.282   0.2103  
`Rate+`      0.0011202  0.0026908   0.416   0.6804  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Third Base

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.551e-01  5.889e-02   2.634  0.01358 *  
Ch.1         3.688e-04  6.959e-05   5.299 1.23e-05 ***
Att.1       -9.901e-03  3.215e-03  -3.080  0.00461 ** 
Rate.1       4.546e-02  1.116e-01   0.407  0.68695    
`Rate+.1`    9.404e-04  1.948e-03   0.483  0.63306    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Intentional Walks

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.129e-01  1.202e-01   6.762 2.42e-07 ***
PA          -4.779e-05  2.155e-05  -2.218  0.03485 *  
IBB          1.579e-02  4.602e-03   3.431  0.00188 ** 
Rate.3      -1.272e+00  3.916e-01  -3.248  0.00301 ** 
`Rate+.3`    5.894e-04  9.342e-04   0.631  0.53318    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Substitutions

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)  0.506114   0.484316   1.045   0.3056  
`PH/G`      -1.353632   0.829943  -1.631   0.1149  
`PH/G+`      0.012267   0.006760   1.815   0.0811 .
`PR/G`       3.726675   2.358327   1.580   0.1261  
`PR/G+`     -0.006946   0.004365  -1.591   0.1237  
`P/G`        0.210244   0.724750   0.290   0.7740  
`P/G+`      -0.010376   0.030110  -0.345   0.7332  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sacrificial Bunts

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.632e-01  7.544e-02   2.163   0.0392 *  
Ch.2         2.678e-04  5.895e-05   4.543 9.66e-05 ***
Att.2       -1.539e-02  6.125e-03  -2.513   0.0180 *  
Rate.2       2.244e-01  1.400e-01   1.603   0.1202    
`Rate+.2`   -4.381e-04  9.548e-04  -0.459   0.6499    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Linear Model result summary to determine most significant variables.

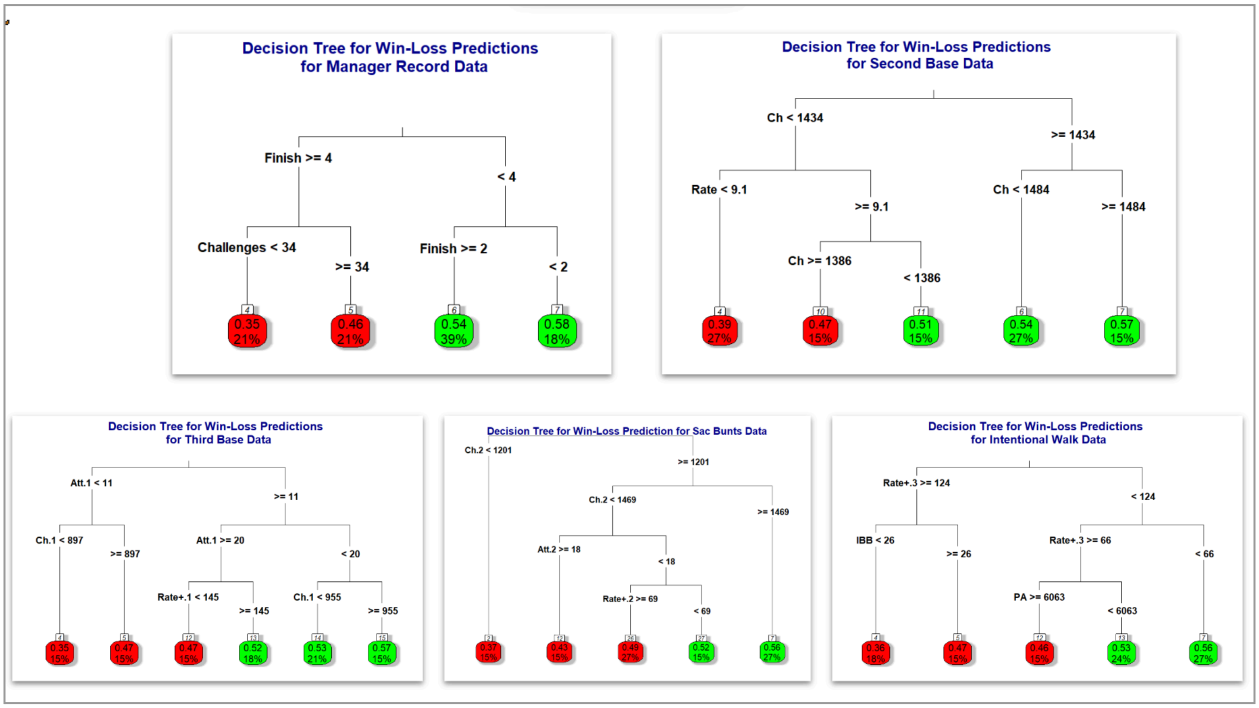
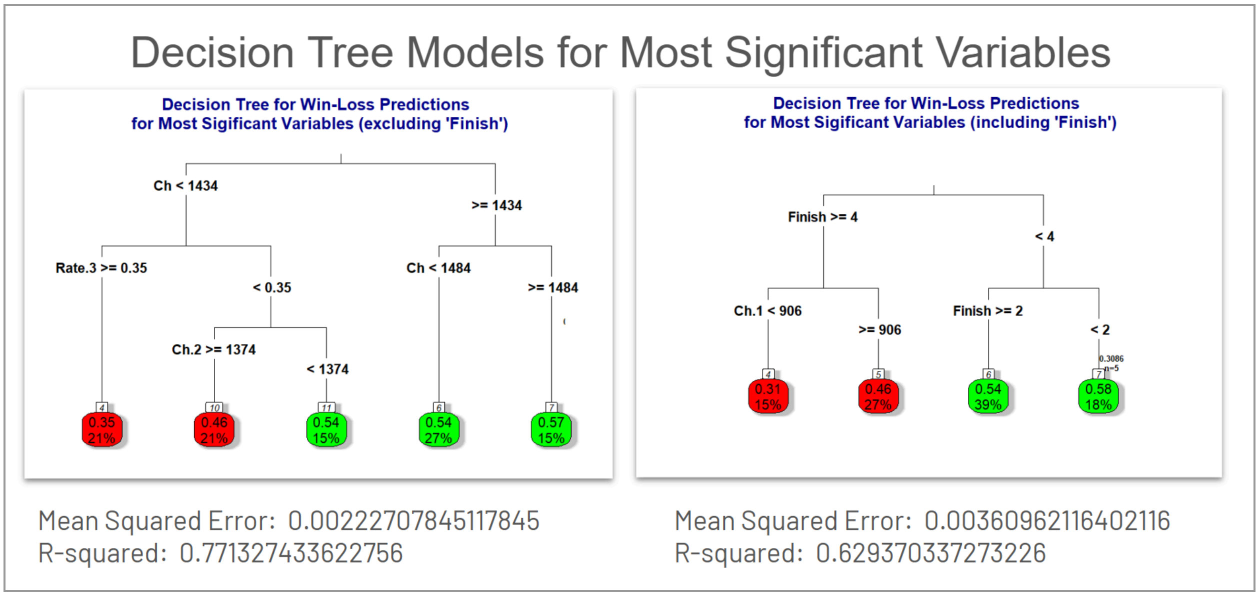


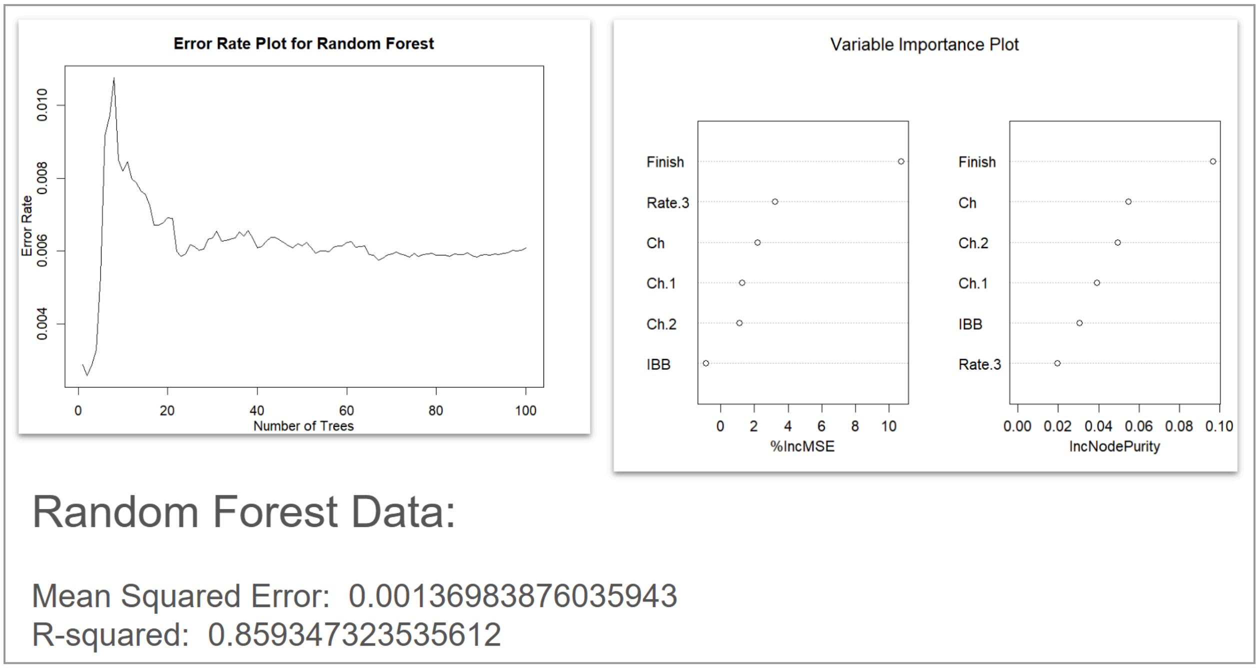
Figure 4: Decision Tree Models for 5 main data groupings.


Decision Tree Models for Most Significant Variables
Decision Tree for Win-Loss Predictions
for Most Sigificant Variables (excluding 'Finish')
Ch < 1434
>= 1434
Rate.3 >= 0.35
< 0.35
Ch < 1484
>= 1484
Ch.2 >= 1374
< 1374
4
0.35
21%
10
0.46
21%
11
0.54
15%
6
0.54
27%
7
0.57
15%
Decision Tree for Win-Loss Predictions
for Most Sigificant Variables (including 'Finish')
Finish >= 4
< 4
Ch.1 < 906
>= 906
Finish >= 2
< 2
4
0.31
15%
5
0.46
27%
6
0.54
39%
0.3086
n=5
7
0.58
18%

Mean Squared Error: 0.00222707845117845
R-squared: 0.771327433622756

Mean Squared Error: 0.00360962116402116
R-squared: 0.629370337273226

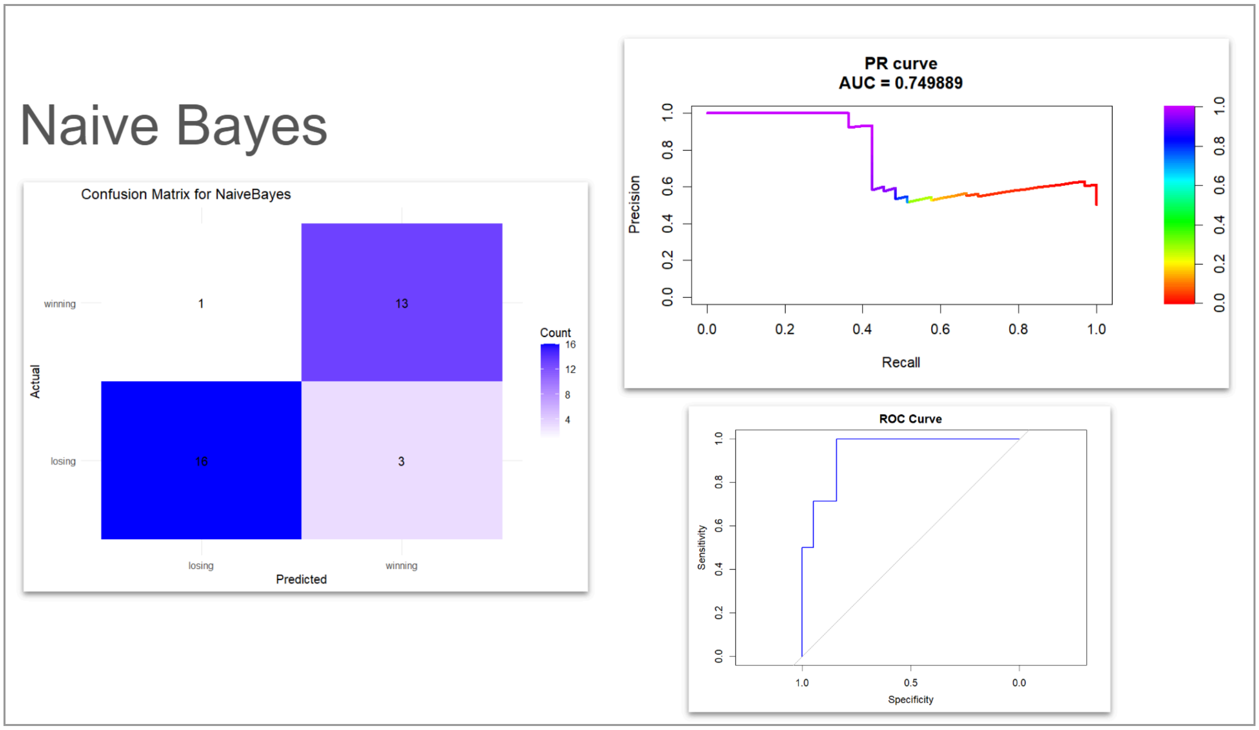
Figure 5: Decision Tree Models for most significant variables determined by the linear models.



Random Forest Data:

Mean Squared Error: 0.00136983876035943
R-squared: 0.859347323535612

Figure 6: Random Forest data visualizations based on most significant variables.



Naive Bayes

Figure 7: Naive Bayes prediction visualizations based on most significant variables.

References

Bobbitt, Zach . “How to Interpret a ROC Curve (with Examples).” *Statology*, 9 Aug. 2021, www.statology.org/interpret-roc-curve/.

GeeksforGeeks. “Random Forest Algorithm in Machine Learning.” *GeeksforGeeks*, 12 July 2024, www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/.

IBM. “What Is Random Forest? | IBM.” *Www.ibm.com*, IBM, 2023, www.ibm.com/topics/random-forest.

Ray, Sunil. “6 Easy Steps to Learn Naive Bayes Algorithm (with Code in Python).” *Analytics Vidhya*, 3 Sept. 2019, www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.