4-24-2023

Money Makes the Diamond: A Creation of a Predictive Model to Forecast On-Field Performance Through Usage of Past Financial Data

Jacob Bowman

SENIOR THESIS APPROVAL

This Honors thesis entitled

“Money Makes the Diamond”

written by

Jacob Bowman

and submitted in partial fulfillment ofthe requirements for completion of the Carl Goodson Honors Program meets the criteria for acceptance

and has been approved by the undersigned readers.

_____

Dr. James Files, Ph.D., thesis director

_____

Dr. Marshall Horton, Ph.D., second reader

_____

Dr. Kathy Collins, Ed.D., third reader

_____

Dr. Barbara Pemberton, Honors Program director

Money Makes the Diamond:

A Creation of a Predictive Model to Forecast On-Field Performance Through Usage of Past Financial Data.

ABSTRACT

This paper examines the feasibility of using statistics to predict win values for major league baseball. Definite correlations were discovered between a Major League organization's finances and on-field performance. Stated correlations are used to generate a predictive model that will predict on-field outcomes. Using regression analysis, such a model is construed, and successfully predicted win ratios for Major League Baseball organizations using only available past financial data.

I. Introduction

1.1 Objective of this Thesis

The primary goal of this paper is to determine if a professional baseball organization's financial metrics can be utilized to forecast said professional organization's performance in the next competitive season. Forecasts frequently utilized by sports analysts, team managers, and the general public take into account a player's on-field statistics, such as batting average, earned-run average, and fielding percentage, as well as other complex formulas regarding past performance on the field, to predict with varying degrees of accuracy, next season's results. In layman's terms, successful evidence proving this paper could result in a "stock market atmosphere" of sports clubs, because the performance in the box office will be just as important as the performance on the field.

1.2 Background of Research Problem

This paper chose Major League Baseball (MLB) due to a variety of factors. The MLB's already heavy involvement with statistics, coupled with the extensive amount of data and history that comes with being the United States' oldest professional sports league, place it in a unique position well suited for study. Despite being placed in a well-suited position for study, this particular research problem has not been entertained in the Major League Baseball context. Several incursions into the statistical analysis of baseball have been made, however. For example, the movie *Moneyball* demonstrates how the Oakland A's used smart budgeting and proper finances to propel their team to the longest winning streak in baseball in the modern era. Another instance is the story illustrated by the book *Astroball*, of the Houston Astros' rise from the bottom of the MLB to winning the World Series using analytics and data-driven management techniques, essentially picking up where *Moneyball* left off. These two real-world occurrences

showed a re-engineering of a way to look at the game of baseball accomplished in real time by the box office. The new methods they gained definitely have an opportunity to influence the sport, and have slightly. However, the statistical atmosphere surrounding baseball remains largely unchanged.

1.3 Need for Study

This problem merits study for several reasons. There are several occurrences of Major League teams that, based on statistics, did not belong in the postseason. The most recent example is the 2022 Phillies, who had a negative Defensive Runs Saved (DRS) and Ultimate Zone Rating (UZR) for their entire team. DRS quantifies a player's entire defensive performance by measuring how many runs a defender saved. UZR quantifies the same thing as DSR, but takes into account errors, range, double-play ability, and arm strength. The 2022 Phillies were in the bottom 7 of the league in both ratings. However, they made it to the World Series. This was in part a result of their assembling of one of the league's best offenses. However, no one knew it at the time. In the preseason, the Phillies were picked to finish third in their division, drastically different from being the next to last team standing. This disparity of picks shows a lack of awareness by outsiders of the future performance of the team. Having knowledge on how financial ratios relate to baseball success could provide a better picture in April of the teams people will see in October. Another example we see of foggy future predictions is the 2005 Yankees. In a league of 30 teams, the Yankees comprised 10% of the league's combined payroll. Shockingly, they were statistically the worst team in the history of the aforementioned statistics above. They embodied the risky philosophy that outspending everyone else will lead to success (SportStorm, 2023). That box-office strategy worked, with the 2005 Yankees making it to the postseason. However, they fell to the Anaheim Angels in 4 games. The Yankees are historically

the largest team in baseball financially; so, they were able to weather the roughly 200 million in salaries and luxury tax. Even with that amount of financial resources, they had failed to accomplish anything of note in the postseason. Hypothetically, if one was able to examine the Yankee's financial data in the preseason, they could have pointed out glaring errors or noticed correlations that led to a less than desirable effect, thereby helping the Yankees perhaps survive more than the first round of the playoffs.

1.4 Methodology

This inquiry into financial standings of baseball organizations being used as a predictive tool for future performance requires several stages of research. The first stage of research will involve the gathering of relevant data and compiling it into a workable format, as the data originated from several sources. Data sources came primarily from *Forbes* magazine, and baseball reference, an online, verified encyclopedia of every statistical occurrence in the history of baseball. Once enough data is gathered, the second stage could begin, which was a data analysis on multiple relationships between an organization's financial data and field performance to determine any heavy correlation. The data used to determine correlation was collected on all 32 teams from the five-year period of 2016-2021. Several formulas were developed in house, as well as the use of the data analysis tools provided by Microsoft Excel. Using what is learned of the correlations, the third stage of the research sees correlations being used on other time periods to determine the reliability of the discovered correlations in predicting final standings at the end of the regular season. Should the final stage prove conclusive, this paper will conclude that it is indeed possible to predict baseball.

1.5 Organization of Study

The next section will be a literature review over materials used in this thesis. The third section will be an overview on the selection of data for the model, and the construction of the model. It will explain the usage of specific variables, as well as the usage of specific sources for those variables. It will also explain the arithmetic behind the model that produces future results. The fourth section will be a testing of the model, including descriptive statistics of the variables and regression statistics confirming the validity of the model. It will also show testing of the completed model by running the model using data out-of-sample. The fifth and final section will be the conclusion to this paper and commentary on future implications.

Literature Review

A few sources were used in the development and research of this thesis. First, Pinnuk and Potter's (2006) *Impact of On-Field Football Success on the Off-Field Financial Performance of AFL Football Clubs,* was a review of factors that contribute to the financial performance of clubs. The authors of this paper examined clubs in the Australian Football League. Particularly, they focused on how match attendance was positively related to both the short-term and long-term success of the clubs, as well as the expected level of competition that was set to occur during the match. Level of competition in this manner refers to if the expected outcome was too close to tell or if it was going to be in sports terms, a blowout. Pinnuk and Potter, in writing this paper, essentially wrote the reverse of what this paper is attempting to accomplish. Their findings were used to help provide sort of a rubric for what was needed to attempt to prove that financial data impacts the field performance of professional sports teams, and more generally, Major League Baseball.

In Ecer and Boyukaslan's (2014) study, *Measuring Performances of Football Clubs Using Financial Ratios: The Gray Relational Analysis Approach*, authors measured financial performance of football clubs in Turkey, as well as identified important financial indicators that measure the financial performances. The authors concluded that liability indicators provide the most important indicator of a club's economic perspective. The methods they used to obtain these results are what greatly influenced this paper. Using what is known as the Gray Relational Analysis Approach, by giving each ratio a variable, the conductors of this study were able to use a matrix-based approach to determine which club, according to the variables they selected, had the best financial performance. This analyzing of data to find relevant correlations is the premise of this paper and will be incorporated into this thesis to find relevant correlation of off-field financials to on-field performance.

Atkinson, Stanley, and Tschirhart's (1988) study, *Revenue Sharing as an Incentive in an Agency Problem: An Example from the National Football League*, examined how well the proposition of revenue sharing is used to encourage desired behavior from member teams in relation to their league. The league is considered the principal, and the teams, or team owners, are considered the agents. The league must do everything in its power to get the team owners to cooperate, yet is powerless to actually do anything, as direct intervention (talent distribution) will violate the integrity of that league. They further extrapolate and say that there is a private non-monetary benefit of winning that enforces competitiveness between agents. In the conclusion of their study, they view that revenue sharing's effectiveness in sports is dependent on the league's having a fixed supply of talent, and the owners, or agents acting as profit maximizers, ignorant of other benefits. This study provided some insight into the behaviors of owners. Additionally, it provided excellent literature on the process of estimating revenues. Revenues are a key point of

the model used to indicate correlation of past financial performance with current on field success.

Research involving MLB Team Finances has always, and will always be for the forseeable future, scarce. Major League Teams are actively encouraged to keep their financial information private. This problem is compounded for this thesis as a result of the MLB and Major League Baseball Players Association (MLBPA), which is the union of professional baseball players, engaging in collective bargaining around the same time a majority of the data was researched for this thesis. Therefore, because of the scarcity of primary source information, discussions and studies on MLB's financial data will always have an asterisk by it. That asterisk is the fact that the best available source, Forbes' annual baseball valuations, is an estimate, and not exact. However, Forbes has stated that it's committing their due diligence. Regardless, this thesis will not be exact, as an exact thesis would infer a knowledge of private financial statements. That knowledge would be worth significant interest by the MLBPA, thus, MLB keeps indiviudal team finances' private.

III. Beginnings of a Predictive Model to Project Future Performance

For a predictive model to be succesful for this thesis, it must be able to predict an organization's playoff eligibility based on previous data. The major question is what previous data is useful and pertinent to the model. An excellent answer to this question can be to look at what others have done for different leagues. Pinnuk and Potter (2006), gave an excellent insight with their paper. They first examined the anecdotal evidence that society in general realizes that the teams' winning and losing is a direct reflection on the box office and a general manager's spending habits. However, Pinnuk and Potter, who were examing the Australian Football League, focused primarily on match-day attendance, ticket sales, and overall fan attitude towards

a specific club (2006). This focus enabled them to create a model to show the relationships between fans and financial performance. However, an important finding from their study showed that fans attend games for one primary reason: the team is winning (Matt Pinnuk, 2006). Pinnuk and Potter received results from their model that showed attendance to be a function of short-term and long-term team success. In other words, Pinnuk and Potter have indirectly proved the reverse of this thesis to be true. While this thesis is concerned with past financial data predicting future team performance, Pinnuk and Potter proved past team performance correlate with future team finances indirectly through attendance of fans (2006). Therefore, to create the model for this thesis, inspiration will have to be taken from Pinnuk and Potter and into the workings of a viable model.

There is some merit in a matrix-based approach to this model, as done by Ecer and Boyukaslan (2014). The Gray Relational Analysis is a method which assists in decision making processes containing many criteria by ordering them as to relation grade (Ecer, 2014). The problem with this approach however, is that the model is going to be an attempt at predicting one single variable, not a ranking of several variables. The Gray Relational Approach will simply not provide the outputs needed to generate a sufficient model; it did however, rank financial variables to a degree of importance in determining a measurement of performance. The results from their study showed financial ratios converging on the topic of liabilities were important to consider, as they determine an organization's ability to invest its funds into performance-increasing measures such as new players (Fatih Ecer, 2014). An example of this can be seen in Major League Baseball's salary cap, which is a restriction on how much one organization can spend on filling their rosters. Usually, an organization does not have its entire salary cap available, as most of it is tied up in obligations, i.e. liabilties, to other players in long-term

contracts. An inclusion of liabilities into the model is essential, as it will provide a factor representing, in essence, free cash on hand.

The model then, should take financial data from Major League organizations that provides information about liabilities. This owing to an underlying issue of “restricted” money that the organization is not free to spend on itself. Additionally, the model must have some data that can be used to measure the scale of performance. Two excellent metrics that are rather basic are the win ratio, expressed as a decimal, and whether a team makes the playoffs, being expressed as a “0” for no, and a “1” for yes. These two “performance statistics” can provide a basis for the model to cross-examine against the financial data and generate a predicition.

3.1 Discussion of Financial Indicators Available for MLB Organizations and their Origins.

Financial indicators are metrics that summarize given financial data and provide a means to measure the performance of an organization relative to its competitors. Common indicators can be simple things such as profit margin, return-on-equity, value change, or the debt/value ratio. Regardless of what the metric or ratio is, each one tells a unique perspective on an organization, and provides information on the performance of the organization.

MLB is in a unique position in that it does not release its financial data. There are several reasons for this, but the primary reason comes down to leverage. The league, and the owners of the team, must meet every few years and form a collective bargaining agreement with the players’ union to establish salary minimums, regulations, salary caps, and other items necessary for the continuation of a professional sports league. Should the financial data be released, it would only provide the players with more leverage to “hardball” the league into higher salaries, thus forcing the signing of “union-buster” players, who would then drive the talent of the league

down and cause a reduction in revenue and profits for the entire league. Therefore, in the interest of maintaining an elite talent-level, and a competitive league, Major League Baseball decided to keep financial data of member teams private to the maximum possible extent.

Despite a large degree of secrecy remaining around team finances, there are still plenty of data to analyze and create estimates of an MLB team's finances. *Forbes* realized this, and has created an annual MLB valuation blog. This list, known as "The Business of Baseball," ranks teams according to financial strength (Forbes, 2022). They determine financial strength by analyzing five financial ratios. These five ratios are Current Value, One Year Value Change, Debt to Value Ratio, Revenue, and Operating Income. Forbes does not provide an explanation as to how the arrive at these values, however, their method seems to be appropriate. For example, the owner of the Kansas City Royals, John Sherman, purchased the Royals for $1 billion USD in 2019. That same year, before the purchase, Forbes valued the franchise at $1.015 billion. Another example can be seen in owner John Stanton of the Seattle Mariners, who purchased the club for $1.2 billion in 2016 (Digiovanna, 2022), with Forbes valuating it at $1.1 billion. So while the *Forbes* valuation is not the exact amount that an owner paid for the club, it is realistically close for estimates, and therefore can be trusted as a rubric for the purposes of creating a predictive model.

3.2 Integration of Historical Major League Data to Build a Model

The first steps in the creation of the model is to establish the time frame from which we will gather our data. It was decided that the period from 2016-2022 would be used, as it is rather recent and contains enough data for the model to return what is a fair estimate and not one skewed by the long-term volatility of baseball statistics. Additionally, this period has *Forbes* valuations for each and every year. For the baseball performance data, which is Win Ratio and

Playoff Appearances, that information was obtained from *BaseballReference.com,* an online encyclopedia of baseball statistics that are verified by the Major League Baseball Office. Now that the performance and financial data were located in the same workbook, it was time to match the financial data to its respective performance data.

This step would prove exceedingly difficult, because while *Forbes* did provide a date of publishing, the date typically coincided with MLB's opening day by a week, so the question was, "If the financial data provided by *Forbes* was for the year it was posted, or the year that had already occurred?" After careful analysis, it was determined that *Forbes*' values provided were current at the date of publishing, meaning they accounted for the previous season, but not the season that was approaching. Therefore, the financial data provided by *Forbes* were an analyst's current prediction of their value before the beginning of the season.

Now that the timeframe was established, a pairing of the *Forbes* data and the *Baseball Reference* statistics was required to proceed with the model. An important factor in the pairing was the heart of the thesis, which was if past financial indicators could provide insight into future on-field performance. With this in mind, the *Forbes* data were paired with season statistics of the season that occurred after the revelation of the valuations. This could allow for completion of the objective of the model.

With the basis of the model now determined, all that is left is to actually bring the pieces together into a cohesive unit and create the proper formulas. Luckily, Microsoft Excel has an excellent add-in that allows for regression analysis. Regression analysis is a series of processes that statistically estimate relationships between one singular dependent variable and one or more independent variables. Our dependent variable in this case was the Win Ratio for the future year,

or the value that we want to predict. Our independent variables were the *Forbes* valuation metrics, as well as the previous year's win ratio, and the previous year's playoff appearances.

3.3 Analysis of Chosen Independent and Dependent Variables.

The first *Forbes* valuation metric provided is the current value of the franchise in billions of dollars. This metric is an independent variable in the regression analysis, and was expressed as 1.000 for one billion. Table (1) shows descriptive statistics for the variable, with the average team over the selected time period having a value of 1.71 billion dollars. This value was chosen first for its availability, but also because of what the variable represents. Current value represents total capital that the organization could possibly spend on self-improvement in the field. It is worth noticing that current value tended to increase for each team each year. This tends to suggest that the organization, like other successful corporations, increases retained earnings consistently while not necessarily increasing spending on development of its team. This occurrence of accumulating wealth rather than developing teams could possibly be a result of revenue-sharing; which is a result of the league dividing revenue amount teams to encourage desired behavior (Scott E. Atkinson, 1988). Owners may be spending all of their profits. They do not allocate to the revenue sharing program in-season. However, this would show minimal change in valuation, and would suggest increases in valuation are a direct result of box office inactivity. Regardless, valuation is an excellent metric for the model because the old truth of "bigger teams win" is prevalent in the Majors, with the league being divided into "small" and "big" market teams.

The next variable that was provided by *Forbes* was the change in valuation over one year, or one year value change. Expressed as a percentage, teams from the 2016-2022 time period experienced on average a fourteen percent change in value (Table 1). However, this variable

varied greatly, as the maximum change seen in one year was 100 percent by the 2016 San Francisco Giants. There were very few instances in our sample of organizations losing value over a year. There are some arguments that this variable is not necessary to include owing to valuation already being in the model, but in the interest of thoroughness it was included in the base model and initial regression.

Then, *Forbes* provided the Debt to Value Ratio as the next independent variable for the model. As stated earlier in the paper, it was necessary for the model to include some component of liability (Fatih Ecer, 2014). Debt to Value is an excellent ratio for this purpose. Providing a picture of the organization's standing with its liabilities is necessary to understand how the organization is able to spend its money. A higher ratio would imply a tighter restriction of funds, whilst a lower ratio implies either an under-utilization of debt as a tool to expand, thereby being a box-office fault, or a lack of money tied up in contracts to players, signaling a weaker team.

The next ratio to be included in the model is the yearly total revenue, provided by *Forbes*. Table 1 shows some basic descriptive statistics on it. For the model, it was decided to post revenue numbers as millions, with the highest number being 683 million by the 2020 Yankees. It was chosen from the *Forbes* list because it, along with debt-to-value, provides a small picture as to the working capital of a Major League Baseball team.

The last *Forbes* value to be included in the model would be Operating Income, also by millions of dollars. This was a precise picture of the working capital of an organization, and can help show the range of action available for the box office to take. Descriptive statistics seen on Table 1 on this variable show the data to posses quite a large range, of 102 to -192 respectively. With a mean of 19.2 million, Operating Income tends to be on the positive side, but only slightly

for Major League Clubs. This variable could be a sign of the disparity between teams with better financial resources and teams with worse.

Another variable being used in the model was the previous year's Win Ratio. This allowed us to have context behind the financial numbers. As the financial numbers were for the past year, knowing the Win Ratio that resulted from those financial numbers is very helpful for the model to correlate the relationship between the variables. It was obtained from *BaseballReference.com*, and, as seen in Table 1, held a mean of .501, with a range of .427. This is important because it shows that for every winning team there is most likely a losing team that can pair with it.

The last variable being used is our dependent variable. This is the current year's Win Ratio. This is the value we want to be able to predict upon completion of the model. To do that, the model must first understand all of the provided variables and the relationships between them. The significant point of this variable is that it is the Win Ratio for the season that is after the *Forbes* data is released.

3.4 Creation of the Model

Valuation data and performance statistics were collected for all 32 teams over the 7-year period. Then, the first task was to run a regression analysis on it with Win Ratio as the dependent variable. The results, posted in Table 2, were then analyzed to determine any strong correlations between the dependent and independent variables. The criteria that signified whether a correlation existed in this study was a p-value, which is the probability of that result under an assumption of no effect or difference of obtaining a result equal or more extreme than what

actually is observed. For our model, we determined that a p-value less than .10 was statistically significant, and a value less than .05 was extremely significant.

Two regression analyses were computed. The first one used every independent variable that had been compiled to see which variables correlated strongly with future win ratio. Unsurprisingly, the strongest correlation by far was whether a team had made it into the playoffs for the year in which the win ratio is the dependent variable. This was not acceptable for several reasons; however, the primary reason owing to playoff appearance being a direct result of win ratio. A team cannot make the playoffs in Major League Baseball without having a good win ratio. Therefore, the future year playoff appearance variable is not independent, it is another dependent variable, which can be substituted by our current dependent variable of future win loss ratio. For this reason, it will be removed from the next regression analysis.

The variables that are chosen for the second regression analysis are the variables that presented the most statistically significant p-values in the first regression analysis. To be considered for the second regression analysis, a significance level of .010 was deemed appropriate. These variables were the team's current value, their debt-to-value ratio, and their previous year's win percentage. The next closest variable was a teams' one-year change in value; however, it was not statistically significant, posting a p-value of .173. Current Value posted a p-value of .0869 and was the largest of our three variables included. With a p-value of .0728, debt-to-value ratio was the second most statistically significant variable. The most significant variable of the first regression analysis however was the previous year's win percentage, with a p-value of .00056. One extremely significant variable not included was whether or not the team made the playoffs that year or the previous year. This is due to the team making the playoffs being an uncontrollable variable that finances really have no impact on, teams may make or miss the

playoffs based on the actions of other teams. Therefore, playoff appearance was determined to be an inconsistent variable. Additionally, playoff appearance is a direct reflection of win ratio for that year, good win ratio's result in playoff appearances, and vice versa.

Having a low p-value is significant because it allows for a replication of the experiment with similar results. It also allows for a model to be built, theoretically, a model that predicts outcomes will use independent variables that have a p-value of 0. This would be a model that perfectly predicts any outcome. However, in baseball it is impossible to perfectly predict any outcome. There are trends, such as the Oakland A's winning twenty games in a row, or the Yankees consistently having winning seasons, but there are no guarantees. Therefore, it is highly unrealistic to expect a perfect p-value of 0 from any of the independent variables; however, this comes with a cost. The model will have a degree of error in its predictions owing to not having a perfect p-value. Therefore, should the model work, and predict a win loss percentage for each team, it is expected it will be inexact, with slight error.

One of the pivotal steps in the creation of this prediction model was the second regression analysis, seen in Table 3. Upon completion of this analysis, a model could be generated that could predict the win ratio of teams. Using the aforementioned independent variables with statistically significant p-values, a refined regression analysis, known as ridge regression, was performed. The regression analysis itself predicts the win ratio that it expects based on what the model has learned about the tendencies of data. This prediction is seen in the analysis' residual output. Understanding the formula that the analysis computes will be the final key to the model. By examining the coefficients of the data and the natural intercept, the formula with which we can use to predict a team's future win ratio appeared. The final model is:

$$Future\ Win\ Ratio = \alpha + \beta_1 Current\ Value + \beta_2 Debt\ to\ Value\ Ratio + \beta_3 Win\ Ratio\ Previous\ Year$$

Where α is the coefficient of the intercept, β_1 is the coefficient of Current Value, β_2 is the coefficient of Debt to Value Ratio, and β_3 is the coefficient of the Previous Year's Win Ratio.

IV. Testing of the Model

To test the model, there are three key things that must be understood. The first is the regression statistics. These statistics tell how well the calculated linear regression equations fit with the sample data. This then leads to the second part of the model, which is to determine the margin of error that the model has. Third, is the using of the formula with data that originates from outside the sample data. Through multiple trials using outside data, more can be learned about the real-world reliability of the model.

4.1 Regression Statistics

There are several measures that tell whether a model fits the sample data appropriately or not. The first measure is known as the Multiple R. This is the correlation coefficient that measures the depth of a linear relationship between two variables. This coefficient may be any value between -1 and 1, with the midpoint, 0, indicating minimal levels of relationship. The lower limit, -1, represents a perfect negative relationship while 1, the upper limit, represents a perfect positive relationship. The first regression analysis had a Multiple R of .821, which is indication of and strong positive relationship. The second, however, had a Multiple R of only .577. This score indicates a slight positive relationship, and that the model cannot explain some

variabilities in the sample data. However, this concern is mitigated to the reality of the sport, baseball itself is extremely unpredictable, with lesser teams commonly winning games against stronger teams. Therefore, while Multiple R does not prove the model is a perfect fit, it does fail to disprove it.

A regression statistic that does signal the model to be an excellent fit for the data is the Significance F statistic. This is the p-value for the F-test of overall significance, and determines whether or not the model with all of the independent variables associated explains the variability better than a model with no independent variables. For the second regression analysis, the Significance F statistic was 4.5428 E-12, as seen in Table 3. This is extremely statistically significant, and therefore indicates that the model provided, and the independent variables used, are sufficiently able to explain the variability of the independent variable, which is the current year's win ratio.

4.2 Standard of Error for the Model

A measure that shows the model to be an excellent fit is the standard error, or the standard error of the estimate. As the name implies, this measure is a measure of the average deviation of the errors. The errors in this case are a difference in the predicted value and actual values. The model for the second regression analysis has a standard error of .070. This means that on average, the model was roughly 11 games away from the actual score. This means that the variables in the model correctly predict 151 of the 162 games in the schedule, with some being much closer than that, while others being slightly larger. This is an acceptable standard of error for the model, as it would be extremely hard to predict a win ratio for a 162-game schedule, with a guess on win ratio having only a .006% chance of being correct. This chance is derived from simple math, in a 162-game schedule, there are only 81 counts of separate whole positive

integers that total 162. Then, accounting for the fact that there are two separate columns, a win and loss column, you must then multiply by 2. This gives a total of 162 different ratios that can be selected as the win loss ratio. If one singular guess is made, your chances are statistically 1/162, or .006%, However, the higher concentration of numbers in the .500 range mean that the odds are slightly higher than .006. This model is now a known way to "shrink" the range even further.

For example, let us say the model predicts team "A" to finish with a .574 win ratio. This is the equivalent of a 93-69 record. If we take into account the standard of error, we now realize that, on average, the model is off by .70. Therefore, it is statistically likely that the win ratio for team "A" will be within the range of .644 and .504. This is the equivalent to 22.68 games. If one were to use this model to predict win ratios, they would be able to narrow down their field considerably, and have a 5% chance of correctly predicting the win ratio, at minimum.

One last thing to consider about the standard of error is that it will get smaller as the sample size gets larger. Due to time constraints, the sample size for this model had to be limited, but should it continue to be updated, the standard of error will continue to regress until it is minute. At which point the model will be extremely precise in its estimates.

4.3 Real-World Testing of the Model

The last step to confirm the validity of the model is to test the model using data that originates from outside the sample data set. The goal is to test the repeatability of the model and ensure that the model properly replicates its results while accounting for the known standard error. To do this, several teams are selected from various periods, and the predicted value will be

compared to the actual value. Teams are selected on a random basis, but each selection was made with the intent of not selecting a team that was present in the sample data.

The first team selected was the 2012 Seattle Mariners. The Mariners finished with a 75-87 record, enough for them to take 4th place in the American League (AL) West, but not enough to qualify for postseason play. The threshold for a Wild Card spot in the AL for the Mariners was a win ratio of .574; however, their actual win ratio was a dismal .460 (Baseball Reference, 2023). The question is could someone have used the model to determine whether a team is likely to make it to the playoffs? The answer is yes. If one takes Seattle's current value at the time (585 Million Dollars), their Debt-to-Value ratio from that time (30%), and their win ratio from the previous year (0.414), and insert it into the current year, the model will predict a win ratio for the current year of .434. In regards to their actual win ratio, the model was only .030 off, which is the equivalent of five baseball games. This is also well within the acceptable standard of error, so the model passes its first real-world test.

The next real-world example to bring into the model would be the 2014 Pittsburgh Pirates. The Pirates, having finished with 88 wins in the current year, boasted a preseason valuation of 572 million, with a Debt-To-Equity ratio of 16%, and a previous year's win ratio of .540; received a predicted win ratio of .525 from the prediction model. This prediction was .020 off from their actual win ratio, which was .525 (Baseball Reference, 2023).

The final real-world testing of the model was seen in the 2014 St. Louis Cardinals. This team finished first in the National League (NL) Central with a record of 90-72, giving them a win ratio of .555 (Baseball Reference, 2023). This team would end up in the postseason, losing the National League Championship Series to the San Francisco Giants. Regardless, they had a valuation of 820 million, a Debt-to-Value ratio of 35%, and a previous year win ratio of .598.

When these values are inserted into the given formula, along with the given coefficients for these variables seen in Table 2, the win ratio predicted for them is .533. This is extremely close to their actual value, and regardless would have predicted a postseason finish, as teams above .500 in the win ratio segment tend to be in the postseason.

The success of these real-world tests signifies an acceptance of the model as a proper prediction tool. The promising outlook is that as more data is compiled into the model, the model will only improve. Regardless, as long as the predicted win ratio continues to be within the standard of error, the model can be considered appropriate for predicting future on-field success based on past financial performance.

V. Conclusion

The viability of creating a model that correctly forecasts on-field performance based on past financial data is extremely strong. The model created used a basic regression analysis, with a relatively small time frame from which to draw the sample data, and was able to be within 11 games of the actual record on average. Some predictions were further off, while others were almost exact, but, the fact that this has occurred shows major potential ramifications for the sport of baseball and the surrounding social culture.

With regards to the standard error, the outlook regarding the model is optimistic, as more data is added to the sample, the equation will become more precise, resulting in a decreasing standard error. Should enough data be compiled into the model that the model reflects perfectly the effect each variable has on win ratio, the evaluation of managers and owners will become much easier, with owners now able to understand exactly how many wins to expect.

The social culture will also be greatly impacted should the model continue to improve, as stated, expectations will be set. When expectations are set, betting odds tend to reflect those examples, therefore, should a perfect model appear, the betting market with regards to wins and losses will cease to exist. No gain or at the very most, minimal gains, would be theoretically possible if the future outcome is already known. This would cause an increase in props, or propositions, betting that would focus parts of the game that have nothing to do with the final outcome, such as the performance of a single player.

Another aspect of social culture influence would be the financial data itself. Should the outcome already be known for the fans, attendance would either diminish or increase, leading to affected revenues of that organization. Fans tend to attend games for one of several reasons. They are: uncertainty of the match outcome, short-run success, long-term success, socioeconomic variables, and future success (Matt Pinnuk, 2006). Four of the variables identified as reasons for attendance would be directly influenced by knowing the future outcome, for either better or worse.

In conclusion, this model is an excellent tool for the prediction of professional sports leagues as a whole, not just baseball. Certain variables will have to be adjusted, and additional tests may be warranted, but the preliminary model suggests a future that is known. The question that remains to be seen is whether the knowledge of the outcome is a knowledge that would benefit or harm the general populace. Arguments have been made that publication of this model would eliminate gains from sports betting, hurting a segment of the economy. However, all the model currently does, is increase the chances of guessing a team's record. It does not tell you which game the team will win, it does not tell you which team will win the World Series, and it does not tell you specific player props. Therefore, while initial harm may come to the sports

betting area, the impact would be mitigated by the realization that the model still leaves some things up to uncertainty. Regardless, the money on the diamond has now been proven to directly relate to the performance on the diamond, therefore, money makes the diamond.

Table 1

Descriptive Statistics of the Variables

	Current Value (Billions)	1-Yr Value Change	Debt/Value Ratio	Revenue (Millions)	Operating Income (Millions)	Win % Ratio (Previous Year)	Win % Ratio
Mean	1.713	0.142	0.123	284.948	19.203	0.501	0.501
Standard Error	0.062	0.012	0.007	7.585	3.303	0.006	0.006
Median	1.395	0.080	0.110	269.000	25.750	0.500	0.499
Mode	1.300	0.020	0.000	266.000	68.000	0.420	0.574
Standard Deviation	0.894	0.177	0.097	109.910	47.864	0.082	0.085
Sample Variance	0.799	0.031	0.009	12080.213	2290.933	0.007	0.007
Kurtosis	4.075	4.236	1.904	1.276	1.273	-0.381	-0.526
Skewness	1.861	2.049	1.188	0.812	-0.818	-0.085	-0.038
Range	5.375	1.040	0.460	587.000	292.000	0.427	0.427
Minimum	0.625	-0.040	0.000	96.000	-190.000	0.290	0.290
Maximum	6.000	1.000	0.460	683.000	102.000	0.717	0.717
Sum	359.750	29.910	25.860	59839.000	4032.600	105.269	105.157
Count	210.000	210.000	210.000	210.000	210.000	210.000	210.000

Table 2

1st Regression Analysis Statistics and Correlations

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.821053401
R Square	0.674128688
Adjusted R Square	0.661158685
Standard Error	0.049648181
Observations	210

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	1.024942307	0.128117788	51.97598763	6.42041E-45
Residual	201	0.495453317	0.002464942		
Total	209	1.520395624			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.344985268	0.030750994	11.21867043	5.39168E-23
Current Value (Billions)	0.01169493	0.006798937	1.720111436	0.08695131
1-Yr Value Change	0.027685501	0.020282609	1.36498719	0.173783004
Debt/Value Ratio	-0.067600539	0.037492156	-1.803058216	0.072876812
Revenue	-8.54916E-05	6.7992E-05	-1.257377009	0.210076133
Operating Income (Millions)	0.00011894	0.000120363	0.988184116	0.324250544
Playoffs (Y/N)				
Previous Year	0.007929519	0.010564579	0.750575977	0.453785637
Win % (Previous Year)	0.232303785	0.066324116	3.502553824	0.000567816
Playoffs	0.113848751	0.00796451	14.29450836	1.92E-32

Table 3

2nd Regression Analysis Statistics and Correlations

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.577749522
R Square	0.33379451
Adjusted R Square	0.324092488
Standard Error	0.070121113
Observations	210

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	0.507499712	0.169166571	34.40463445	4.54929E-18
Residual	206	1.012895912	0.00491697		
Total	209	1.520395624			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.213299335	0.031939093	6.678315257	2.20512E-10
Current Value (Billions)	0.014435227	0.005923963	2.43675183	0.015669062
Debt/Value Ratio	-0.023185724	0.052541475	-0.441284233	0.659470051
Win % (Previous Year)	0.529791768	0.062591517	8.464274327	4.80687E-15

Works Cited

Baseball Reference. (2023). Retrieved from https://www.baseball-reference.com/

Digiovanna, M. (2022, 2 28). Here are the Billionaire Team Owners Who Rule Baseball Amid the MLB Lockout. *Los Angeles Times*. Retrieved from https://www.latimes.com/sports/story/2022-02-28/mlb-billionaire-team-owners-roster-2022-lockout

Fatih Ecer, A. B. (2014). *Measuring Performances of Football Clubs Using Financial Ratios: The Gray Relational Analysis Approach.* ResearchGate. Retrieved from https://www.researchgate.net/publication/259799919_Measuring_Performances_Of_Football_Clubs_Using_Financial_Ratios_The_Gray_Relational_Analysis_Approach

Forbes. (2022). *The Business of Baseball*. Retrieved from https://www.forbes.com/mlb-valuations/list/

Matt Pinnuk, B. P. (2006). *Impact of On-Field Football Success on the Off-Field Financial Performance of AFL Football Clubs.* The Authors Journal Compilation.

Scott E. Atkinson, L. R. (1988). *Revenue Sharing as an Incentive in an Agency Problem: An Example from the National Football League.* Wiley. Retrieved from https://www.jstor.org/stable/2555395

SportStorm. (2023, 26 1). The 2005 Yankees Experiment Nearly Broke Baseball. Retrieved from https://www.youtube.com/watch?v=SpIaldPnDWw