Understanding and Modeling Contributing Factors to Earned Run Average in Modern Major Leauge Baseball

Final Project: Part 3

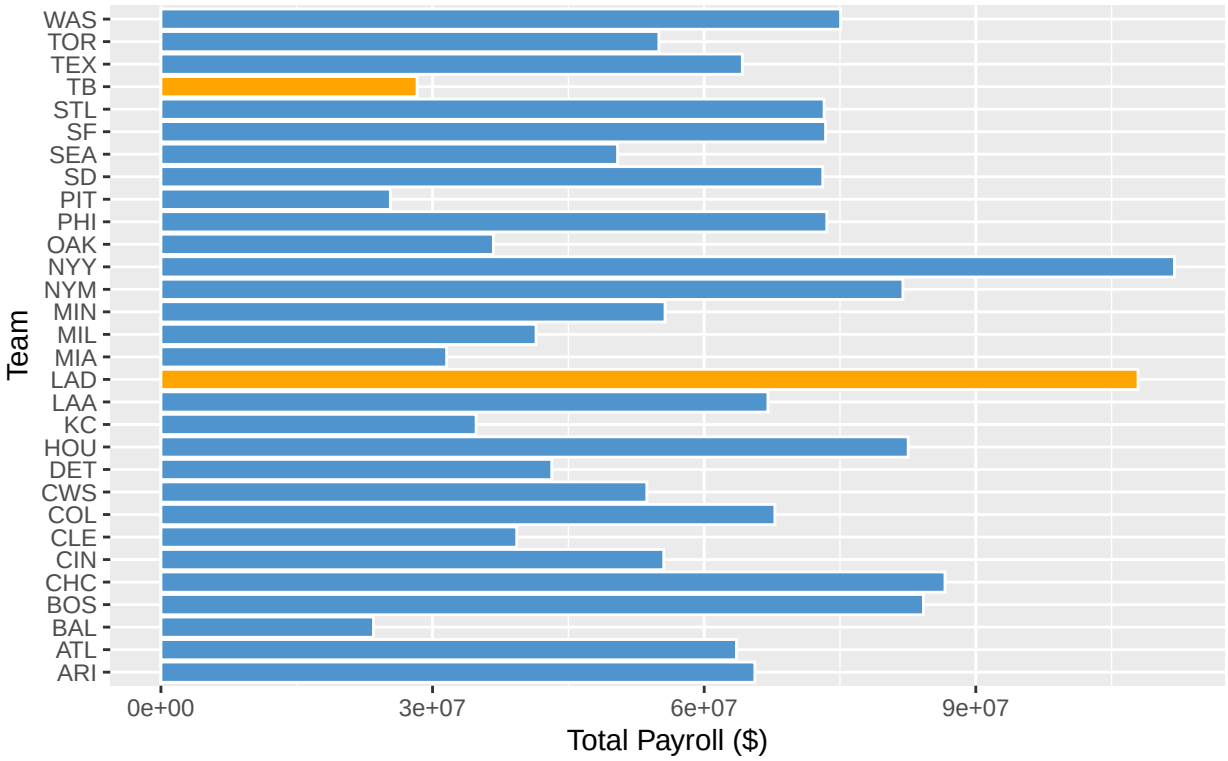
Carson Slater

5/4/2022

Introduction

The 2020 World Series was a reminder to the world that the best Major League Baseball team is more than just the team that possessed the highest payroll. That year, the 60-game shortened regular season led to a postseason like no other, culminating in a World Series with the National League team possessing one of the highest payrolls in baseball (`LAD`) and the American League team possessing one of the lowest payrolls in baseball (`TB`). Figure 1 shows the payrolls of all 30 MLB franchises.

Figure 1: Total 2020 Payroll of MLB Teams



Payroll data taken from https://www.spotrac.com/mlb/payroll/

From Figure 1 the two teams with orange bars are the teams that made it to the World Series. People have begun to wonder what methods teams like the Tampa Bay Rays (`TB`) are using to optimize their payroll for

performance. What measurable metrics are teams looking at in their scouting, player development, lineup and pitching decisions? Hence, the goals of this paper are twofold:

1. to analyze 21 years (2000-2020) of Major League Pitching data from the `Lahman` baseball database, looking at what measurable statistics have the biggest impact on `ERA`, or "earned run average," and
2. to build the simplest but most accurate regression model to predict `ERA`.

This paper assumes that the goal of a pitcher is to prevent runs, so the baseline performance metric to assess a pitcher will be the classic statistic, `ERA`. `ERA` is the average amount of earned runs a pitcher has given up over a nine-inning span.

The plan to determine which alternative pitching statistics are the most effecting in predicting `ERA` is to run a multiple regression model using 10-fold cross validation across the 12,548 pitching stints over the past 21 years (2000-2020). These data are from the Lahman baseball database, a database that contains complete batting and pitching statistics from 1871 to 2020. A pitching 'stint' is a given amount of time a pitcher has spent on the roster of an MLB club. In theory, these data can include multiple stints of the same pitcher, even within the same year, but all for different teams. We are going to assume the independence of these stints due to different teams having different common opponents (National League teams more often play other National League teams), different pitching coaches and possibly different roles. Unfortunately, statistical tests for independence are beyond the scope of this paper, and are encouraged to be be pursued upon in further research and predictive modeling. One additional aspect to note is that these data possess all pitching stints where a pitcher has thrown at least nine innings within that stint.

Throughout the paper, we will initially be using the using the following variables as predictors:

Table 1: Table 1: Predictors utilized in the modeling process for ERA.

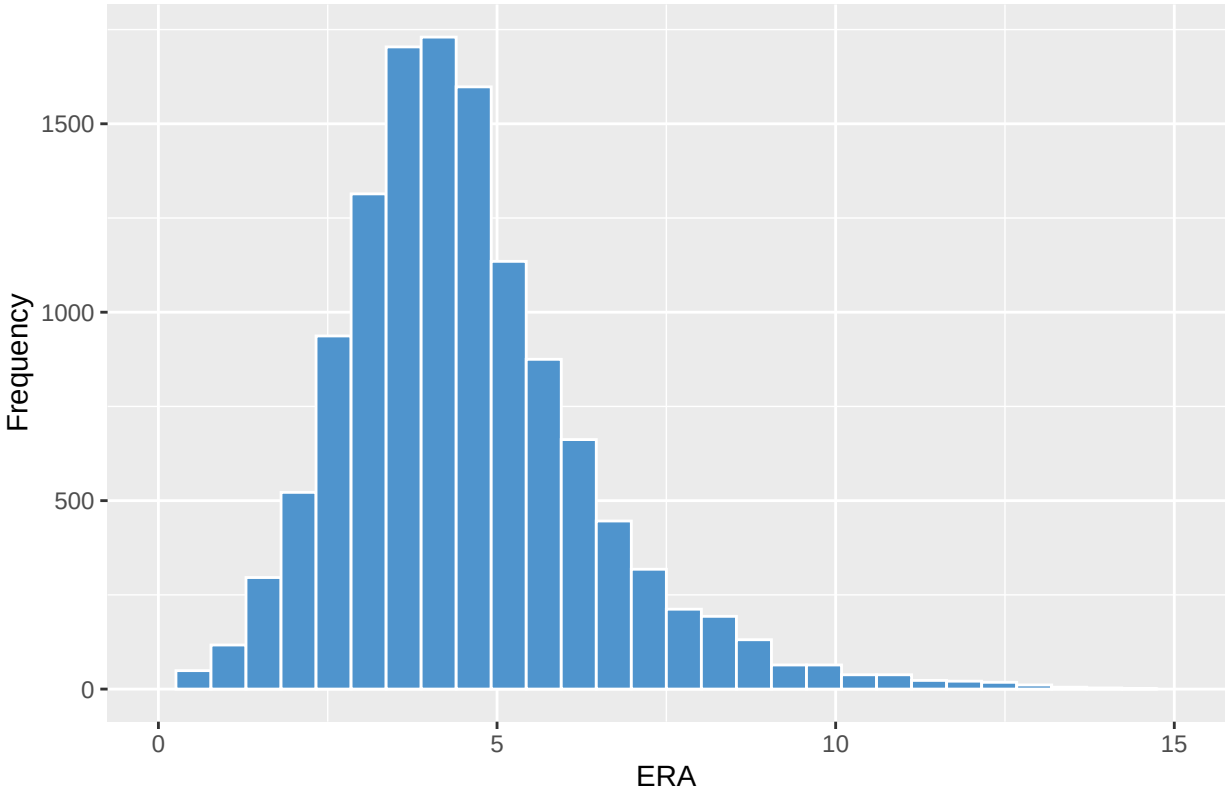
Variable	Statistic	Variable	Statistic	Variable	Statistic
`W`	Wins	`SO`	Strike-Outs	`SF`	Sacrifice Flies by Opposing Batters
`L`	Losses	`BAOpp`	Opponent Batting Avg.	`GIDP`	Double Plays Induced
`G`	Games	`IBB`	Intentional Walks	`WHIP`	Walks + HIts per Innings Pitched
`GS`	Games Started	`WP`	Wild Pitches	`OBPOpp`	Opponent On Base %
`CG`	Complete Games	`HBP`	Batters Hit	`FIP`	Fielding-Independent Pitching
`SHO`	Complete Game Shutouts	`BK`	Balks	`SOp9`	Strike-Outs per Nine Innings
`SV`	Saves	`BFP`	Batters Faced by Pitcher	`BBp9`	Walks per Nine Innings
`H`	Hits Allowed	`GF`	Games Finished	`BBpct`	Walk %
`HR`	Home Runs Allowed	`R`	Runs Allowed	`BABIP`	Batting Avg. for Balls in Play
`BB`	Walks	`SH`	Sacrifices by Opposing Batters	`LOBpct`	Left-On-Base %

Some of the variables like `FIP`, `SOp9`, `SOpct` , `BBp9`, `BBpct`, `BABIP`, and `LOBpct` are advanced statistics, while the rest of the statistics are traditional statistics. One noteworthy point is that the `FIP` statistic factors in a league-wide FIP constant. For simplicity, we chose the 2019 FIP constant for this project (3.214). It usually hovers between 3.1 and 3.2 each year.

Exploratory Data Analysis

What predictors best correlate with a good `ERA`? Would it be a variable that measures damage control, like `LOBcpt`? Could it be variable that indicates most common type of contact, like `BABIP`? Is it how dominant a pitcher is (`SOp9`)? Perhaps it could even be how efficient a pitcher is (`GIDP`). These questions remain unanswered, but some exporatory data analysis can inform our thinking.

Figure 2: Distribution of Pitcher ERAs (2000−2020)



Before we try to tackle those questions, we want to look at the outcome of interest. From Figure 2 we observe the distribution of `ERA` is skewed to the right a little bit, and for good reason. In Major League Baseball, it is a lot easier to obtain a higher earned run average, whereas a lower `ERA` requires consistent, good pitching. From Table 2, we can see that the sample mean is greater from the median, confirming our suspicion that the distribution is in fact right skewed. For the most part, it still possesses a normal Gaussian shape. This is really convenient, because it implies these outcome data are well-behaved and will likely be easier to build a model on.

Table 2: Table 2: Summary Statistics for ERA

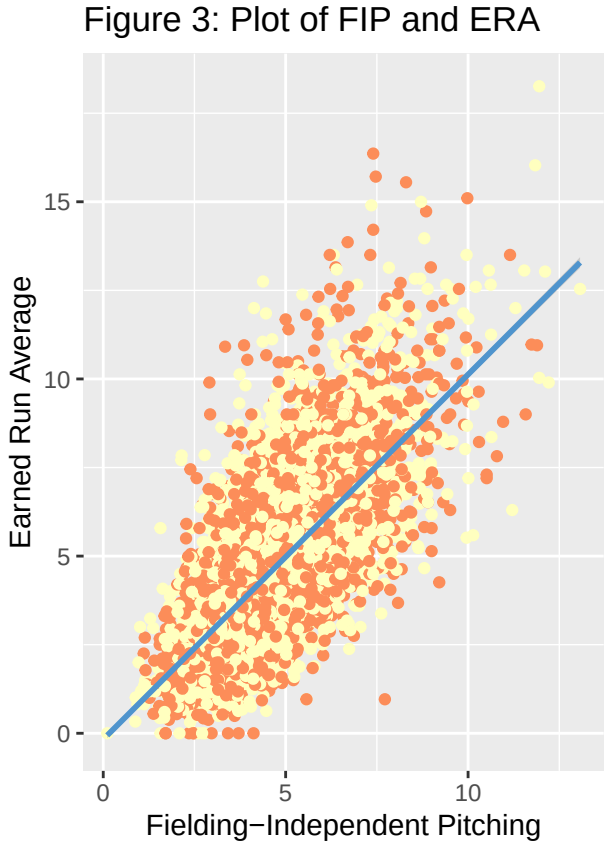
	Minimum	1st Quartile	Median	3rd Quartile	Maximum	Mean	Standard Deviation
ERA	0	3.32	4.26	5.43	18.26	4.526	1.859

Table 3: Table 3: Table of Correlation Coefficients

Statistic	Correlation with ERA
Walks + Hits per Innings Pitched	0.8116
Opponent On-Base %	0.7801

Statistic	Correlation with ERA
Opponent Batting Avg.	0.7506
Fielding-Independent Pitching	0.7177
Batting Average for Balls in Play	0.5375
Walks %	0.2489
Double Plays Induced	-0.1521
Saves (Closing Pitcher Only)	-0.2098
Wins	-0.2558
Strike Out %	-0.4024

Table 3 lists the most notable correlation coefficients of possible predictors with `ERA`, ordered fromt the highest positively correlated to the highest negatively correlated. One of the surprising observations from Table 3 is the correlation coefficient of fielding-independent pitching (`FIP`). `FIP` assesses a pitcher based on the three 'true' outcomes of a plate appearance: strike-outs, walks, and home-runs. What intuitively makes sense is that the metric `BABIP`, which measures frequency of times a batter reaches base without accounting for any of the three true outcomes, would be more correlated with `ERA` than `FIP` would. We found this to be surprising. The question still remains as to which predictor is stronger for `ERA`.



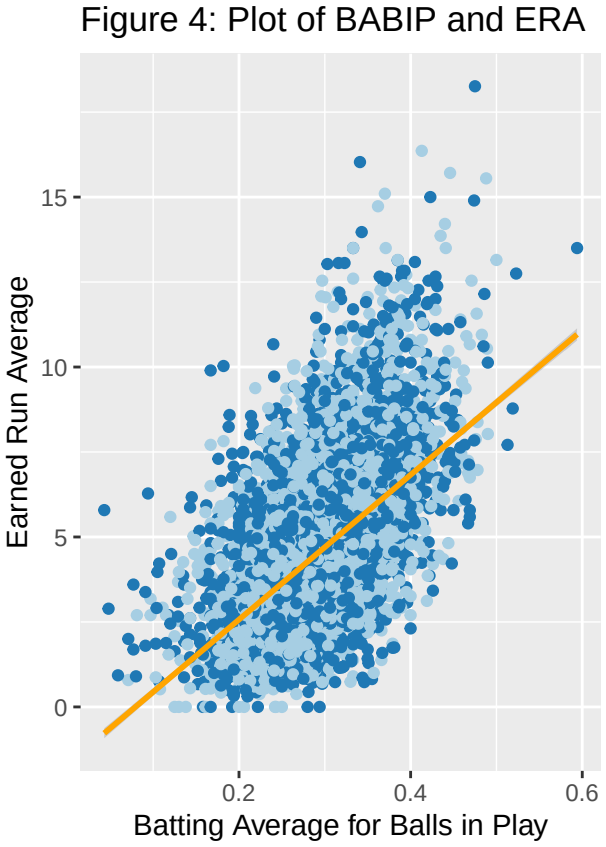


Figure 3 and Figure 4 show the linear relationship between both `FIP` and `ERA` as well as `BABIP` and `ERA`. Both have noticable linear relationships. Also, `FIP` not only has a slightly stronger correlation with `ERA` than `BABIP`, but also has a slightly more positive linear relationship as well. As we build our models we will pay close attention to these two predictors.

Methods

Upon building a neural network, we must consider efficiency when building a model. Because neural networks take copious amounts of time to fit, and are a 'black box' model, it is in our best interest to eliminate unnecessary predictors. To begin this process, we will run a multiple regression with all 34 predictors including the `Lahman` default pitching performance metrics and the additional advanced ones we were able to fabricate from the `Lahman` data. Casually referred to as 'the kitchen sink method,' this model will most likely have a high R^2, as it possesses the most information of any model we will build, but it is also possible that it will model some noise. We split the data 80/20 for training and testing respectively, and trained the linear model on ten-fold cross-validation.

After building the first linear 'kitchen sink' model, we are able to narrow the set of predictors down to fourteen predictors, from 34 initial predictors. We run this regression to test the model accuracy, and then decide the performance is good enough to build a neural network.

For our neural network, we use the `nnet` engine in `tidymodels`. Considering the 14 predictors remaining after our multiple regression trials, we tune the number of hidden units from 1 to 10, the penalty from 0.001 to 0.1, all with 1000 epochs. Using the same recipe as the regression models the neural network is fitted with same 10 folds used to train the regression models, and finally test it on the remaining 20 percent of the data.

After building a neural network, we also build a support vector regression model. Although this model is also a supervised black box model, and thus highly uninterpretable, we wanted to see if this model - typically used for classification - could be used for regression. We used the `kernlab` engine in `tidymodels`, with a radial kernel. By running a radial kernel as opposed to a polynomial kernel, we run the risk of over-fitting the model, but the size of our training data gave us confidence to proceed. A support vector regression essentially builds a hyper-plane in n-dimensional space, and can be transformed by a kernel to create the shape that best fits the data.

Results

The Kitchen Sink Linear Model

Table 4: Table 4: Kitchen sink model metrics.

Metrics
0.4895
0.9281

As expected, we observe from Table 4 the high R^2 of the first multiple regression model. We observe the following predictors all possess p-values above the $\alpha = 0.05$ significance level, rendering them as possible hindrances to model accuracy and precision: losses (`L`), game appearances (`G`), complete games (`CG`), shutouts (`SHO`), intentional walks (`IBB`), wild pitches (`WP`), hit batsmen (`HBP`), balks (`BK`), games finished (`GF`), strike-outs per nine innings (`SOp9`), and walks per nine innings (`BBp9`). We remove these predictors from our next multiple regression model.

In the linear model recipe, we also remove very highly correlated predictors with an absolute correlation coefficient above .95. The correlation filter removed the cumulative statistics: hits allowed (`H`), walks (`W`), strike outs (`SO`), batters faced by pitcher (`BFP`), and regrettably, the non-cumulative statistic `WHIP`. `WHIP` is a linear function of walks, hits, and innings pitched, so although this predictor is highly correlated (see Table 3) with the outcome of interest, we will proceed without it in hopes that our single layer neural network will uncover a similar relationship from the remaining predictors.

The Refined Linear Model

Table 5: Table 5: Comparison of linear model metrics.

Model 1 Metrics	Model 2 Metrics
0.4895	0.4919
0.9281	0.9273

After removing 11 predictors, Table 5 demonstrates the simplified model essentially possessed the same performance metrics as the kitchen sink model. We have successfully produced a simplified version of the initial model without sacrificing too much performance.

Table 6: Table 6: Coefficients for second linear model.

Term	Estimate	Std. Error	P-value
(Intercept)	2.3225	0.1033	0.00000
GS	0.0110	0.0012	0.00000
SV	0.0042	0.0008	0.00000
HR	-0.0113	0.0023	0.00000
BAOpp	16.9872	0.8278	0.00000
R	0.0017	0.0008	0.03350
SF	-0.0272	0.0039	0.00000
GIDP	-0.0187	0.0017	0.00000
OBPOpp	3.3974	0.7833	0.00001
FIP	0.6838	0.0146	0.00000
SOpct	8.2961	0.2356	0.00000
BBpct	0.5642	0.5176	0.27574
BABIP	-3.3660	0.7164	0.00000
LOBpct	-9.6847	0.0723	0.00000

Observing the refined linear model coefficients in Table 6, we note the remaining coefficients absorbed some of the noise from the dropped predictors; however all of these coefficients are statistically significant, with the exception of walk percentage (`BBpct`). Even though the coefficient for walk percentage in this cleaner linear model is not statistically significant, we proceed with these fourteen predictors for the neural network.

One of the primary purposes of this paper is to find pitching performance metrics that carry the most weight in predicting `ERA`. This is difficult to do with a neural network, which is highly uninterpretable as a 'black box method.' We instead examine the coefficients of the second linear model, and observe that `BAOpp` has the biggest coefficient. We must remember that the variable itself is a percentage. For this particular model, we know that if every batter a pitcher faces gets a hit off the pitcher three times out of ten, we can assume the model will say that a pitcher's `ERA` will have 5.1 earned runs added to whatever their `ERA` is given all other parameters are held constant. Although `FIP` had the strongest linear relationship with `ERA` it was not the biggest positive relationship by magnitude, which was a noteworthy observation. Unsurprisingly, the metric with the biggest negative linear relationship with `ERA` given constant all other variables in the model was `LOBpct`. This intuitively makes sense, because pitchers will often give up a hit or two in a given inning, and if they can prevent runners on base from scoring, their low earned run average stay in tact. This phenomena is often referred to in the baseball world as 'damage control,' where a player gives up a lot of baserunners but manages to keep runs from scoring.

The surprising metric that might potentially be a weakness of this model is the coefficient for strike out percentage, `SOpct`. Generally speaking, a pitcher is thought perform better when their strike out percentage is high, but our model tells us otherwise. Our model has a relatively high, positive coefficient for `SOpct`,

which may be attributable to low variation ($\sigma^2_{SOpct} = 0.0631$) within the statistic. Small changes in `SOpct` can lead to bigger changes in `ERA`. The question remains, why is the coefficient positive? Intuitively, this does not make coherent sense, so it is most likely a flaw of the linear model.

Primary Model: A Single Layer Neural Network for Regression

We observe from Table 7 that our neural network performed exceptionally well, but there was not too much room for improvement beyond our linear model to begin with.

We observe that the neural network is able to explain an extra 2% of the variation between of `ERA`. Hence we can conclude that the neural network is marginally better than the linear models.

Figure 5: Tuning the Neural Network

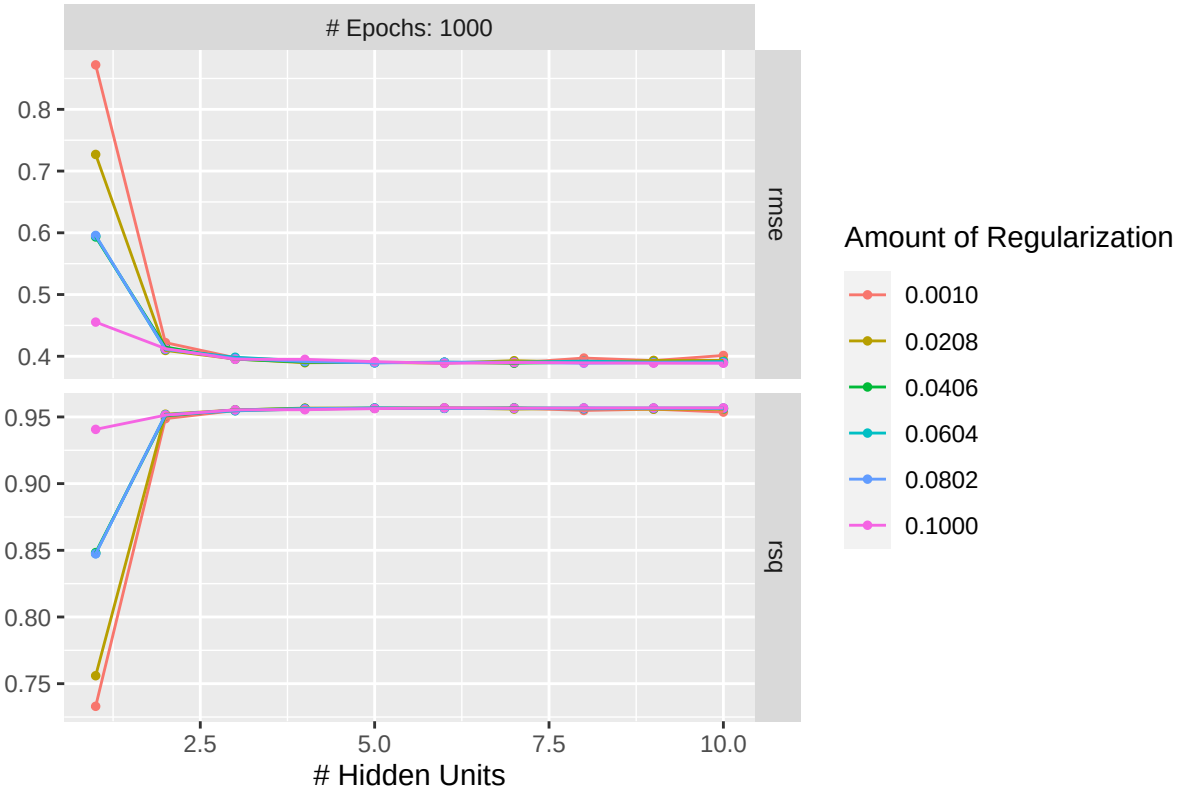


Figure 5 indicates the biggest R^2 for tuning occurred with a 0.1 L1 penalty and 6 hidden units. It must be noted the model performance stays fairly consistent from four hidden units to nine hidden units. With the testing set, we know the model scored just below this particular R^2 value in Figure 5. Hence, we have successfully built a single-layer regression neural network to predict `ERA` for major league baseball pitchers.

Additional Model: Radial Support Vector Regression

Table 7: Table 7: Comparison of all model metrics.

Linear Model 1	Linear Model 2	Neural Network	Radial Support Vector Machine
0.4895	0.4919	0.4091	0.4141
0.9281	0.9273	0.9498	0.9496

As a model, the radial support vector regression took a substantially longer time to tune. From Table 7 we observe that radial support vector machine performed marginally worse than the neural network. With an essentially identical R^2 and a slightly larger RMSE, we do not gain any predictive power by using a support vector machine. It appears a neural network was able to uncover more nonlinear relationships than the radial support vector machine was able to.

Discussion

Summary

From our analysis we can conclude with certainty that batting average and left-on-base percentage are truly key contributors for marginal `ERA` differences in Major League Baseball. Major League Baseball pitching can be characterized as what is often called 'well-behaved' data, as it is fitted well with only a multiple regression model, leaving little room for improvement to begin with for machine learning models.

With our single-layer neural network, we are able to build an efficient, 14-predictor model that explains 95% of the variation in `ERA`, and also keeps residuals within .40 standard deviations from the estimate. Our support vector machine performed very similarly to the neural network as far as performance.

Weaknesses of This Paper

Critiques of our model would be that we did not stratify our folds or training/testing data by year of stint. Although this may have impacted our model tuning and performance, we wanted our model to account for the true 'randomness' with equal probability of each observation being in either the training or testing set, and subsequently equal probability that it would end up in any fold.

Additionally, it could be said that the dynamics of baseball, especially hitting philosophy and pitcher development, have drastically changed within the past 10 years (2012-2022), rendering our training data from any year before 2012 as outdated. This most definitely can be viewed as a potential weakness of our model; however we proceeded with this selection of data because the cost of only using recent data, was less than the benefit of using slightly outdated data. It is not as if pitchers were not also trying to prevent runs in 2000. The role of a pitcher remains the same.

As previously mentioned, our data set does not contain career statistics for pitchers, but rather divides the pitcher's performance metrics into 'stints' with teams. A potential weakness of this paper, as well as a further area of study, is the testing of independence for each of these observations. Without independence, there can exist bias in our model.

Lastly, we recognize that by using a neural network and a support vector regression leaves little room for model interpretability. In order to analyze isolated impact of a neural network, we needed to outsource that task to a multiple regression model. This is most definitely a weakness, especially when we remain curious about non-linear relationships between predictors.

Further Areas of Study

One potential area of study would be to discover the non-linear relationships between predictors. Alternatively if there were a way to measure impact on winning outcomes, and the relationships between added win probability of a pitcher per game and some performance metrics, that might give players, general managers and coaches more useful information on day-to-day operations. Lastly, a potential research opportunity would be to analyze minor league performance metrics for pitchers and see how they predict major league performance, and if there are any discrepancies between predicting performance from major and minor league performance.