Predicting Teams in the World Series Using KNN, LDA, and Logistic Regression

Jon Southam

November 2023

Introduction

Every year baseball fans get the opportunity to enjoy hundreds of professional baseball games. Some people have a favorite team they have followed for years, others have a favorite player they follow, and others might just like the game. Whatever the reason people follow the game, the month of September gets more intense because the regular season is wrapping up and teams are getting closer to the World Series championship. At this time the question on fans' minds becomes "Who will make it to the World Series?" This is the question we are attempting to answer. We will make a classification model to predict whether or not a team will make it to the World Series. We will create three different classification models using linear discriminant analysis, logistic regression, and K-nearest neighbors.

Description of Data

The data we will use come from *Baseball-Reference.com*. They call themselves "The complete source for current and historical baseball players, teams, scores and leaders." In terms of Major League Baseball (MLB), the website has all the standard baseball statistics for each season going back to 1876. Prior to 1969 MLB did not have a playoff series, meaning that the top team in each of the American and National Leagues ascend to the World Series for the championship. The top team was determined by the team with the best win-loss record in each league. In 1969, playoff series were introduced adding excitement for the fans and another layer of statistics for us to analyze. Figure 1 displays the structure for the post season games. For this project, we will years 1969 to the present, giving us 55 seasons for analysis. In 1981, there was a baseball strike that resulted in the cancellation of 38% of the regular season games. In August 1994, there was a baseball strike that cancelled the remainder of the season including the regular season. Due to Covid-19, only 60 regular season games were playing in 2020. For these reasons, we will exclude these three seasons from our data. Since the 2023 World Series recently concluded with the Texas Rangers winning the title against Arizona Diamondbacks, we will also remove the recent 2023 season so we can see how our models categorize the this year's results. This gives us 51 seasons of statistics for training and testing.

Figure 1: **Playoff Structure**

There are nearly 80 variables when we consider pitching, hitting, and fielding so we need to decide which to include in our analysis. In order to determine which variables to consider we used the logistic regression model function in R to regress "Being in the world series" (WS) on all variables and noted

Table 1: **Variables for Analysis**

Variable Name	Description
SB	Stolen Bases
OPS	On-base Plus Slugging Percentage
SO.Batting	Strikeouts (Batting)
tSho	Total Shutouts
cSho	Complete Game Shutouts

the significant variables at $\alpha = 0.05$ level of significance. We first split the data in half into training and testing datasets. The significant variables are given in Table 1 which are based on the training set.

SB is the number of bases a team gets without a hit. OPS is the sum of the teams batting average and its slugging percentage. Slugging percentage is the total number of bases a team records per at-bat meaning it incorporate how many bases they get per hit. OPS is useful as a metric because it accounts for a teams ability to hit for power. SO.Batting is the number of strikeouts a team gets while batting. A team wants this to be low, as opposed to strikeouts while pitching which a team wants to be high. The last two statistics, tSho and cSho are similar in that they both measure how many times a team did not allow any runs during a game. The difference is that tSho measures how many times one or more pitchers did not allows any runs during a game, while cSho measure how many time a single pitcher pitched an entire game and did not allow any runs. We can see in the correlation plot in Figure 2 that there is a $y = x$ maximum on the scatter plot because it is impossible to have a higher cSho than tSho. Since this project is looking at a team's statistics we will remove cSho from our analysis. Figure 3 shows the correlation plots for the remaining four variables. Based on the plots we do not see correlation between the variables.

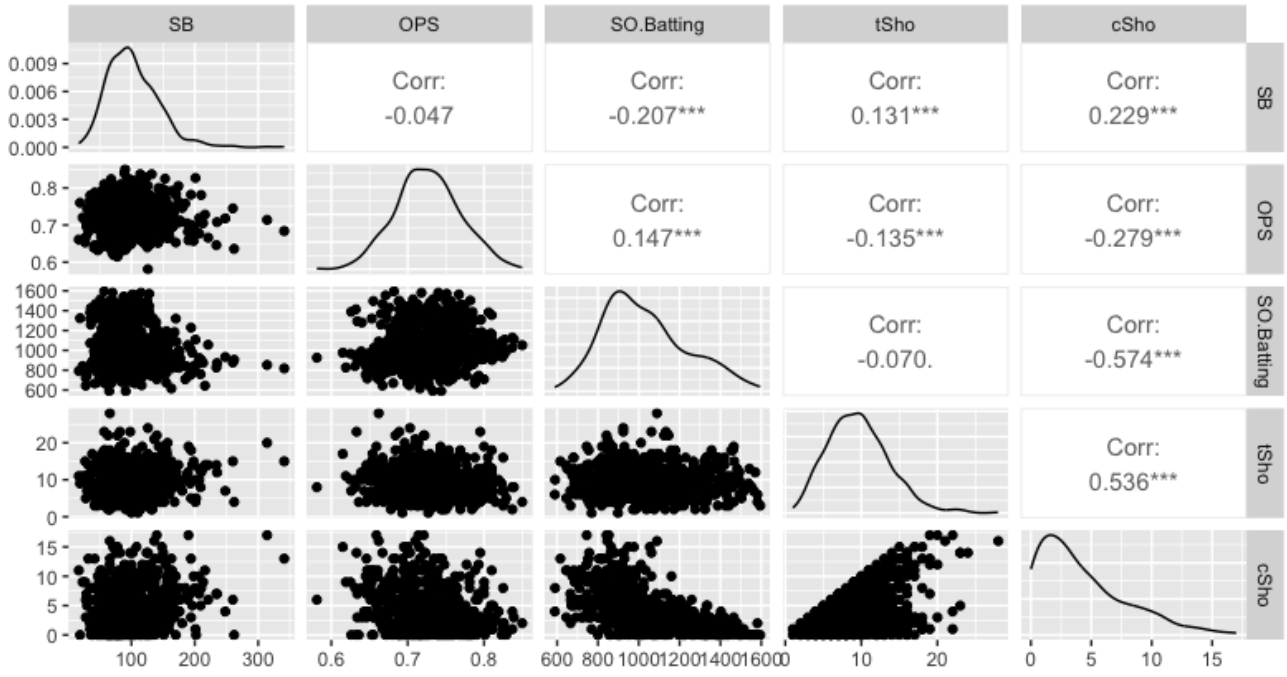

Figure 2: **Correlation plot with all significant variables**

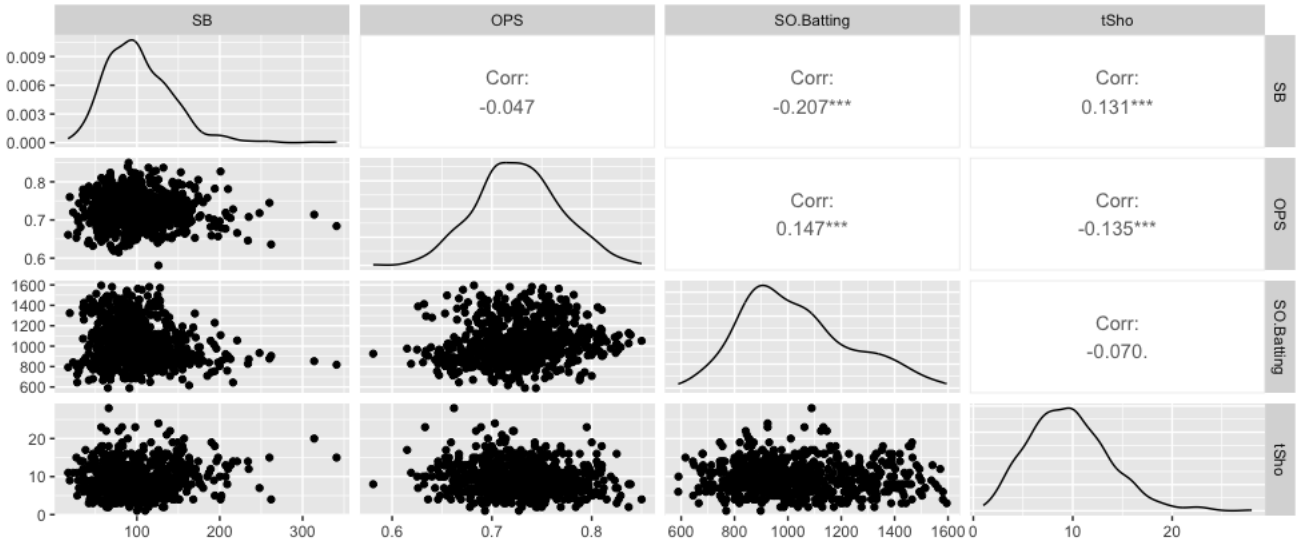

Figure 3: **Correlation plot with selected significant variables**

We do see some points within stolen bases that may be outliers and could he high leverage points. To asses this we will look at the Leverage plot in Figure 4. We see that a few points do have high leverage, but they are within Cook's distance and so we will not consider them wielding significant weight on the coefficients. Let's begin our analysis with our LDA model.

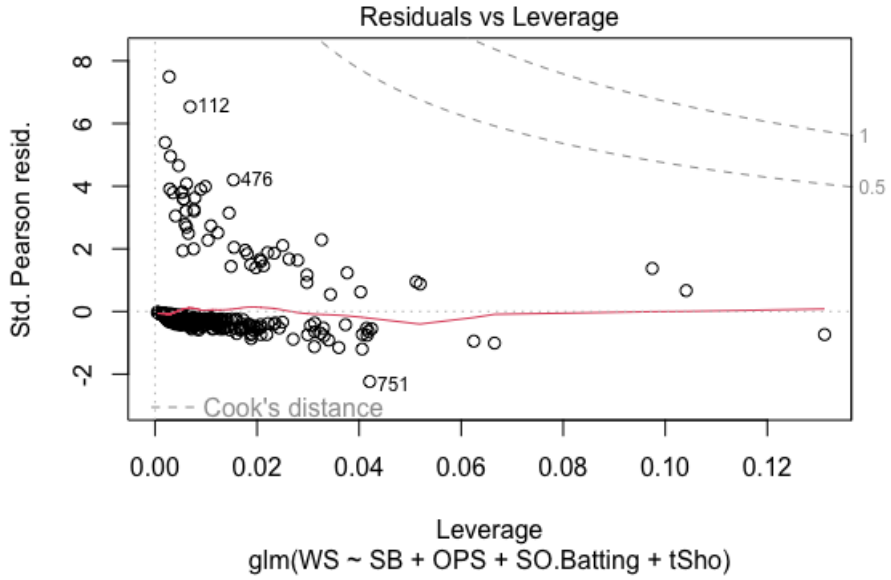

Figure 4: **Leverage Plot**

Linear Discriminant Analysis

There are three main assumptions for LDA. The measurements are independent from each other, distributions are normal, and the covariances are equal for each class. The measurements are independent when we take the data set in its entirety because any team results across seasons would not impact another teams' from different seasons. However, this brings up a concern within seasons because the team results are affected by other teams because a team with a high tSho statistic would impact a teams OPS because a strong bullpen (another name for a teams' set of pitchers) would perform better at preventing a team from getting hits and therefore decreasing that other teams' OPS. Also a team with a higher

tSho statistic would inflict more strike outs on a team. But when we look at these statistics over the 51 seasons these influences would be less severe. For this reason we will assume independence across the seasons. Based on the correlation plot above we see that the variances appear relatively symmetric. We see that SB, SO.Batting, and tSho are slightly skewed right but we will make the normal assumption due to the variable being approximately normal.

Below are the covariance matrices for the teams that did not make it to the world series ("Not in WS") and for those that did ("In WS"). Because of the nature of our domain and there being a high level of variability on these statistics we will lower our alpha to 0.01. When we perform a Box's M test on the matrices we see that the matrices are not significantly different at 0.01 level of significance ($\chi^2 = 21.32, df = 10$).

Covariance of "Not in WS"

$$\begin{bmatrix} \text{SB} & \text{SO.Batting} & \text{OPS} & \text{tSho} \\ 1567.59 & -299.88 & -0.13 & 14.68 \\ - - - & 54035.76 & 1.31 & 39.71 \\ - - - & - - - & 0.00 & -0.04 \\ - - - & - - - & - - - & 17.02 \end{bmatrix}$$

Covariance of "In WS"

$$\begin{bmatrix} \text{SB} & \text{SO.Batting} & \text{OPS} & \text{tSho} \\ 2184.43 & -1468.20 & -0.31 & 2.72 \\ - - - & 39235.28 & 1.35 & 41.64 \\ - - - & - - - & 0.00 & -0.07 \\ - - - & - - - & - - - & 19.89 \end{bmatrix}$$

Based on the training data, the prior probabilities for being in the World Series (In WS) and Not in the World Series (Not in WS) are 0.075 and 0.925, respectively. It is a significantly higher probability of not being in the World Series than being in it. Table 2 displays the variable means for each group. We see that the mean number of stolen bases, OPS, and shut outs are higher and the strikeouts are lower which is a reasonable observation because we would expect a team to be more success given those four criteria. We see in the LDA coefficients in Table 3 that OPS contributes more discriminatory power than the other three variables.

Table 2: **Group Means**

	SB	OPS	SO.Batting	tSho
Not in WS	100.7049	0.7245	1034.9373	9.3349
In WS	120.0943	0.7503	979.6981	13.0755

Table 3: **Coefficients of Linear Discriminants**

Variable	LD1
SB	0.0069
OPS	15.5902
SO.Batting	-0.0010
tSho	0.2008

When we run our testing data through our model we find it are 92.5% accurate ([True Positive + True Negative] / Total). The precision is 16.7% which means that of the teams that were predicted to go the World Series, our model correctly predicted 1 in 6. (True positive / [True positive + False Positive]) The recall is 2%. (True Positive / [True positive + False Negative]) This tells us that of teams that went to the world series, the model correctly predicted only 2 percent as going to the championship. The high accuracy comes form the high prior probability of not going to the World Series. Table 4 shows the confusion matrix. Looking at the ROC curve (Figure 5) we see that while our model performs better than a random classifier with AUC = 0.77.

Table 4: **LDA Confusion Matrix**

	Predicted	
Actual	**Not in WS**	**In WS**
Not in WS	653	5
In WS	48	1

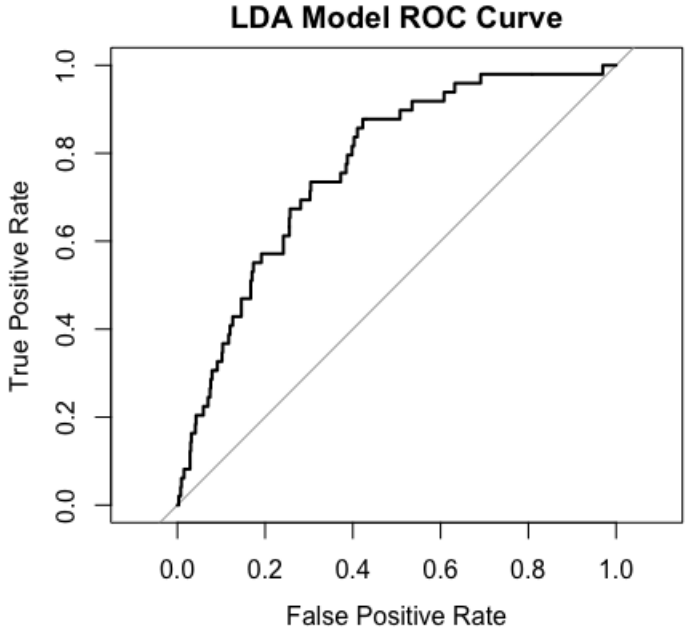

Figure 5: **LDA ROC Curve**

When we input the 2023 season data, our model only predicts the Atlanta Braves will go to the World Series. While the Braves did make it to the play offs, the did not make it to the championship. Using the posterior probabilities that the LDA model provides we select the team with the next highest probability to go against the Braves which in this case was the Tampa Bay Rays which were another team in the playoffs.

Logistic Regression Analysis

Now let's explore our Logistic regression model. The assumptions of logistic regression are 1) the dependent variable is binary, 2) the observations are independent of each other, 3) there is no multi-collinearity between the independent variables, 4) linearity between independent variables and log odds, and 5) we need a large sample size. The first two assumptions are clearly met. We noted above that the variables are not collinear, and we do have a large sample size. We will make the assumption between independent variables and log odds.

The model coefficients and deviances are displayed in Table 5. Similar to the LDA model, OPS carries a higher contribution to discriminate between our two categories than the other variables. We also see that the Residual deviance is lower than the Null deviance which tells that the model performs better with our four variables than with the intercept alone. This is good for our model.

The logistic regression model's accuracy is 88.4%, its precision is 33.3%, and its recall is 6.1%. It is marginally stronger than our LDA model in terms of its precision and recall. Similar to the LDA model, when we input the 2023 season data, our model only predicts the Atlanta Braves will go to the World Series. To determine their predicted rival we use the posterior probabilities our model provides and select the team with the next highest posterior probability. That team was the Los Angeles Dodgers, who also were in the playoffs. The confusion matrix is displayed in Table 6. We can see that the ROC curve is similar to that of the LDA model above with AUC = 0.78.

Table 5: **Logistic Regresssion Coefficients**

Variable	Estimate	Std. Error	z value	Pr(>z)
(Intercept)	-21.24	3.314	-6.411	1.45e-10 ***
SB	0.00735	0.00348	2.111	0.0347 *
OPS	23.01	4.117	5.589	2.28e-08 ***
SO.Batting	-0.00182	0.0009148	-1.993	0.0463 *
tSho	0.2492	0.03886	6.413	1.43e-10 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Null deviance: 376.54 on 706 degrees of freedom Residual deviance: 296.40 on 702 degrees of freedom AIC: 306.4				

Table 6: **Logistic Regression Confusion Matrix**

	Predicted	
Actual	**Not in WS**	**In WS**
Not in WS	652	6
In WS	46	3

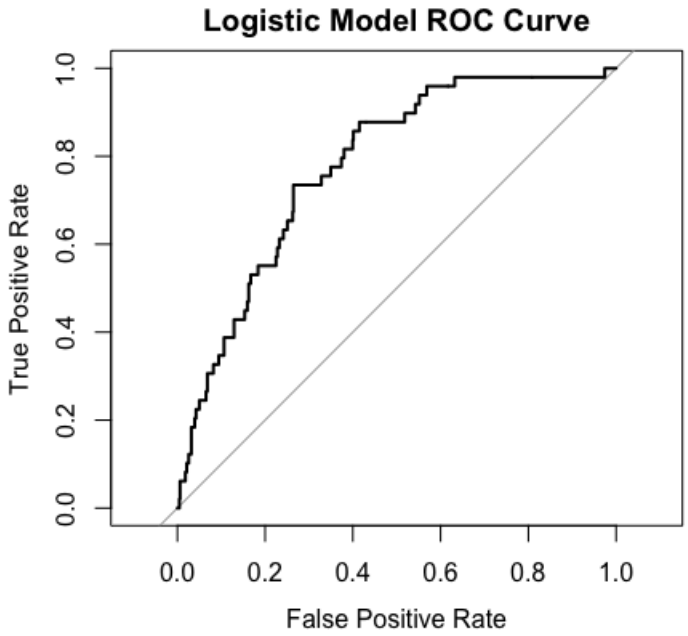

Figure 6: **Logistic Model ROC Curve**

K-Nearest Neighbors

K-Nearest neighbors (KNN) does not assume anything about the distribution of the data, it just assumes that similar points will be close to each other. To asses the KNN model we need to know how many nearby points, k, we need to consider. We begin with k = 1 and observe the model's recall and precision percentages. We see in the table below that after k = 2 our model is not effective as its recall and precision are 0. The confusion matrix is displayed in Table 7. At k = 2 our recall and precision are marginally better than the other models, but nothing that indicates that it is a more useful model.

Table 7: KNN k-values and performance

k	Recall	Precision	Accuracy
k = 1	6.1%	5.6%	86.3%
k = 2	8.2%	7.8%	86.0%
k = 3	0%	0%	91.5%
k = 4	0%	0%	91.4%

Table 8: KNN Confusion Matrix

	Predicted	
Actual	**Not in WS**	**In WS**
Not in WS	611	47
In WS	45	4

When we run our KNN model on our 2023 season with k = 2, our model correctly predicts the Arizona Diamondbacks making it to the World Series. However it only predicted the one team making it to the championship.

Discussion

Using classification models to predict a team making it to the World Series has is difficult because it is not nicely linear nor easily divided into clusters. Something to note is that all three models were greater than 80% accurate because so few teams make it to the World Series and predicting "Not In WS" is relatively easy because the prior probability is high and likely accurate. There are 30 teams vying for the championship and only two get the opportunity to play for it each season. Of the teams that the models predicted to go to the World Series, few of them actually did. While the models were accurate due to the high number of "Not in WS", the low recall and precision rates tell us the models are not practically able to make the types of classification we want. Thus our significant predictive capability of our models are low.

The structure of this analysis that was problematic is that they did not account for there needing to be 2 teams each season in the World Series. A 1 team championship would not be interesting to watch. While we are able to get around this issue by using the posterior probabilities provided by the LDA and logistic model to determine the top 2 teams, the KNN model is not able to accomplish this because it discriminates by Euclidean distance rather than a probabilities statement. The issue arises from the models being trained on all teams from any season making it to the World Series rather than looking at two teams from each season to go so far. When we look at all 51 seasons' statistics we need to first discriminate by season, then by which teams were in the championship. Based on the data we also need to consider a non-linear model. A next step would be to transform the data into higher dimensions and then implement a support vector machines model.