Athens Journal of Sports - Volume 3, Issue 4– Pages 239-252

Predicting Major League Baseball Championship Winners through Data Mining

By Brandon Tolbert[*]
Theodore Trafalis[†]

The world of sports is highly unpredictable. Fans of any sport are interested in predicting the outcomes of sporting events. Whether it is prediction based off of experience, a gut feeling, instinct, simulation based off of video games, or simple statistical measures, many fans develop their own approach to predicting the results of games. In many situations, these methods are not reliable and lack a fundamental basis. Even the experts are unsuccessful in most situations. In this paper we present a sports data mining approach to uncover hidden knowledge within the game of baseball. The goal is to develop a model using data mining methods that will predict American League champions, National League champions, and World Series winners at a higher success rate compared to traditional models. Our approach will analyze historical regular season data of playoff contenders by applying kernel machine learning schemes in an effort to uncover potentially useful information that helps predict future champions.

Keywords: *Baseball, Data Mining, Kernel machine learning, Prediction.*

Introduction

The game of baseball is becoming more reliant on analytical and statistic based approaches to assemble competitive baseball teams. The most well-known application of statistics to baseball is captured in Michael Lewis's book "Moneyball". Applying a simple statistical approach, Moneyball demonstrated that on-base percentage (OBP) and slugging percentage (SLG) are better indicators of offensive success than other standard statistics such as batting average, homeruns, and runs batted in (Lewis 2003). By gauging players success on OBP and SLG, the Oakland A's were able to construct a team of undervalued players to compete against teams with much higher payrolls. There success of using this objective based approach is justified by their playoff runs in 2002 and 2003. With the success of the Oakland Athletics, managerial schemes have shifted from a subjective based approach to the search for objective knowledge. While baseball metrics focus on measuring individual successes, baseball is a team sport i.e. a sum of individual successes. Therefore it is not justified to rely on individual player statistics to determine the team's success rate as a whole. Data mining can integrate multiple individual and team performance statistics to discover potentially useful objective knowledge for projecting a team's overall success. In this paper, the data mining approach will be applied to predict the outcome of whether or not a team is a championship contender or not.

[*] Graduate Student, University of Oklahoma, USA.
[†] Professor, University of Oklahoma, USA.

https://doi.org/10.30958/ajspo.3.4.1 doi=10.30958/ajspo.3.4.1

Data Preparation and Description

Studies of forecasting championship contenders in major league baseball are rare. Only one reference was found that attempted to apply data mining methods to predict World Series champions (Egros 2013). Specifically, our study will attempt to develop classifiers using Support Vector Machines (SVM's) to predict championship winners on three levels: (1) American League champion, (2) National League champion, and (3) World Series champion. The data is partitioned into three subsets because the National and American League are separate divisions that rarely compete with each other during the regular season. Further, each league has a unique set of rules. For instance, in the National League the pitcher is required to be in the hitting lineup. By comparison, the American League has a designated hitter that hits in place of the pitcher and is not required to play defense. The data will be resolved using a binary classification scheme. For instance, the World Series championship consists of two teams, one which is classified as a "loser" and one of which is classified as a "winner". A multivariate data set of statistical measures will be analyzed. The dataset used for our analysis will come from various sources which include baseball1.com, mlb.com, and fangraphs.com. Attributes that will be considered are listed in Table 1. It includes a combination of traditional and sabermetric statistics.

Table 1. *Summary of Attributes Evaluated in this Study*

Attribute	Category
Total Runs Scored (R)	Offensive
Stolen Bases (SB)	Offensive
Batting Average (AVG)	Offensive
On Base Percentage (OBP)	Offensive
Slugging Percentage (SLG)	Offensive
Team Wins	Record
Team Losses	Record
Earned Run Average (ERA)	Pitching
Save Percentage	Pitching
Strikeouts per nine innings (K/9)	Pitching
Opponent Batting Average (AVG)	Pitching
Walks plus hits per inning pitched (WHIP)	Pitching
Fielding Independent Pitching (FIP)	Pitching
Double Plays turned (DP)	Defensive
Fielding Percentage (FP)	Defensive
Wins Above Replacement (WAR)	Baserunning, Hitting, and Fielding

Experimental Procedure

The objective of this study is to apply different SVM algorithms to determine the best predictor of championship winners. The SVM software used in this analysis included Matlab and LIBSVM. Matlab includes many machine

learning algorithms that are easily applied and integrated with decent sized data sets. LIBSVM can address unbalanced datasets which would otherwise be difficult to resolve with standard SVM algorithms in Matlab.

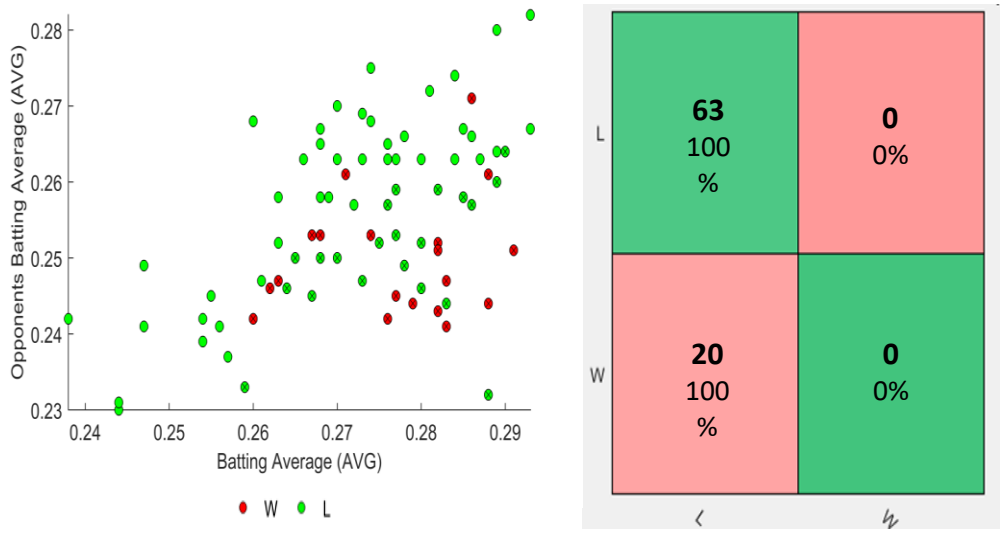
For our experiment, the SVM classifiers will be developed using 10-fold cross validation. This method divides the dataset into a training set to develop a model and a testing set to evaluate the performance measure of the model. Specifically our analysis will attempt to determine two features from the 16 shown in Table 1 that develop the most accurate classifier for the given datasets without overfitting the data. The SVM algorithms used in our experiment include: (1) linear kernel SVM, (2) quadratic kernel SVM, (3) cubic kernel SVM, and (4) Gaussian kernel radial basis function (RBF) SVM. The results from the different classification models will be used to predict championship winners for the 2015 postseason.

Some of the datasets we will encounter in this study will be unbalanced, i.e. one class contains a lot more examples than the other. To account for the imbalance in data different costs for misclassification for each class will be assigned through a trial and error approach. The method of handling unbalanced data with the LIBSVM software package is addressed in Chang and Lin (2011).

American League Pennant Model Selection

The American League pennant race dataset consisted of 20 years of data with two class labels: American League champion (W) and loser (L). 20 teams were labeled as American League champions and 63 teams were labeled as losers. The dataset is unbalanced, i.e. 24% of the data is classified as American League champions and 76% are classified as losers. When basic support vector machine algorithms are used to classify the unbalanced dataset the results were not useful as demonstrated in Figure 1.

Figure 1. *Standard SVM Algorithms to Classify the Unbalanced Data*



As can be observed from Figure 1, the losing class (L) was predicted with 100% accuracy and the American League champion class (W) was predicted with 0% accuracy. When classifying imbalanced datasets, basic support vector machine algorithms are often biased towards the majority class. Further, a prediction cannot be made by this model because the entire domain of the feature space is classified as the majority class (L). To address the bias towards the majority class, one must modify the basic support vector machine algorithm by assigning weights to each individual class. This ensures that the minority class can be correctly classified for at least a few instances by the model.

Determining the optimal gamma (γ) of the Gaussian kernel radial basis function (RBF) SVM was another challenge encountered when developing classifiers. The gamma parameter defines the area of influence of a training instance (Ben-Hur and Weston 2010). A high gamma value indicates that the area of influence is very small. A low gamma value indicates that the area of influence is very large. Figure 2 shows the effect of the gamma parameter of the Gaussian kernel RBF for a constant soft margin value (C). Table 2 compares the accuracy of the different Gaussian RBF classifiers given in Figure 2 and includes the cost values assigned to the minority (W) and majority (L) classes**.** Note that the costs or weights were determined by a trial and error approach.

Figure 2. *The Effect of the Gamma Value for a Fixed Soft Margin Constant (C)*

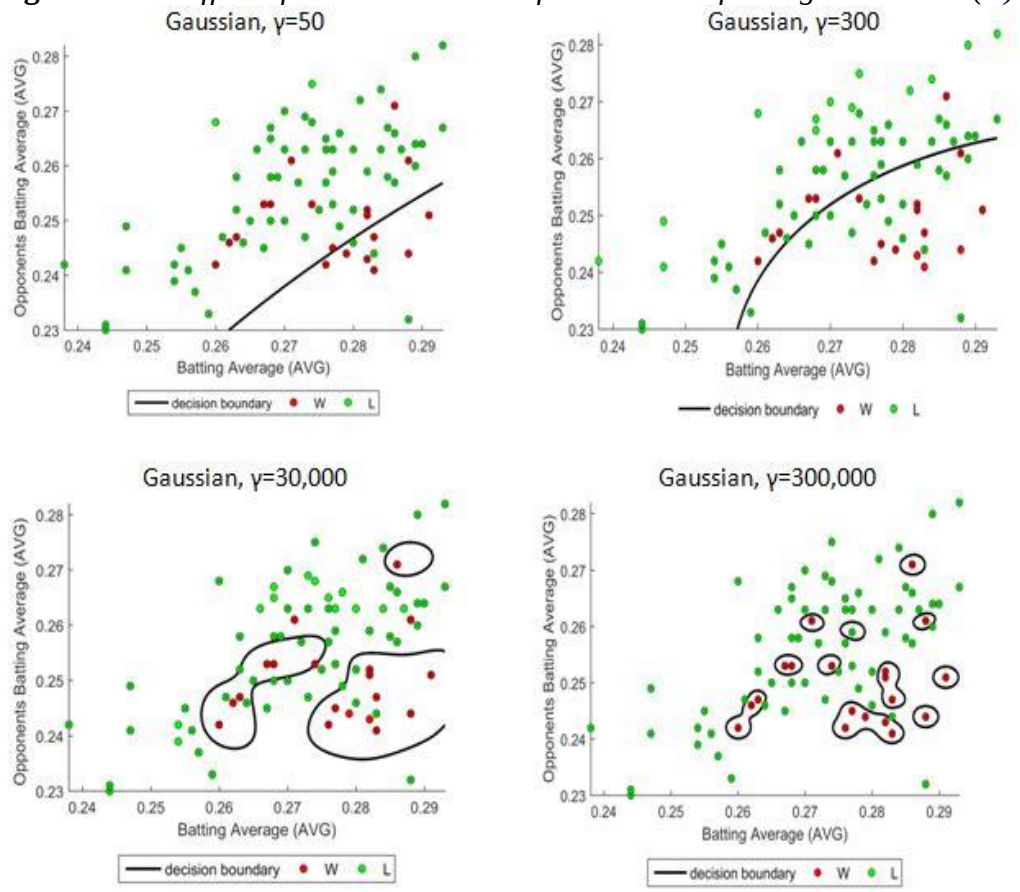
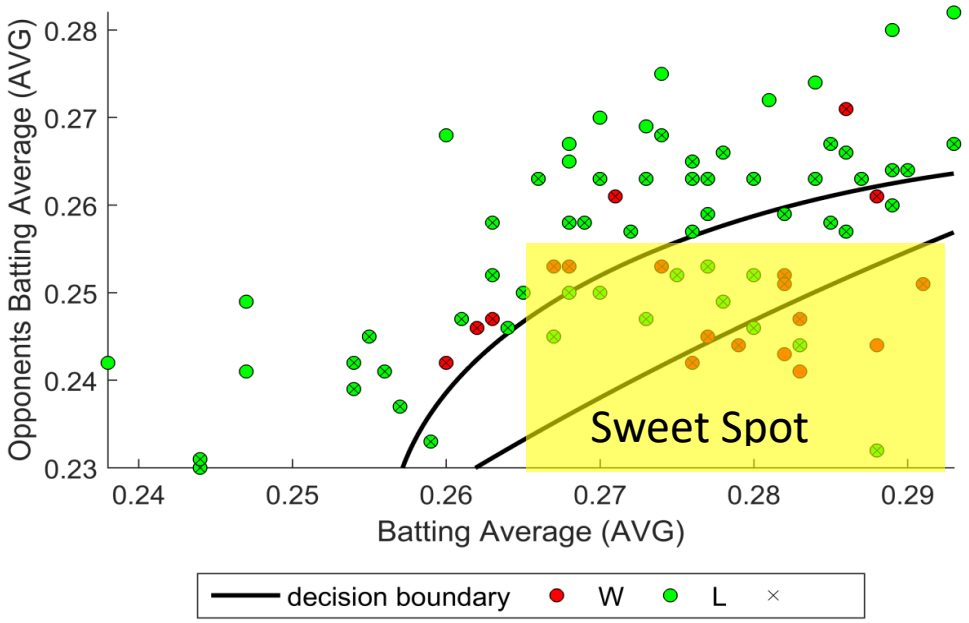


Table 2. *Accuracy Comparison of Different Gaussian RBF Classifiers Shown in Figure 2*

Model	Cost for Majority (L)	Cost for Minority (W)	Gamma (γ)	Accuracy (%)
Case 1	1	3	50	80.7
Case 2	1	3	300	71.8
Case 3	1	3	30,000	86.7
Case 4	1	3	300,000	98.8

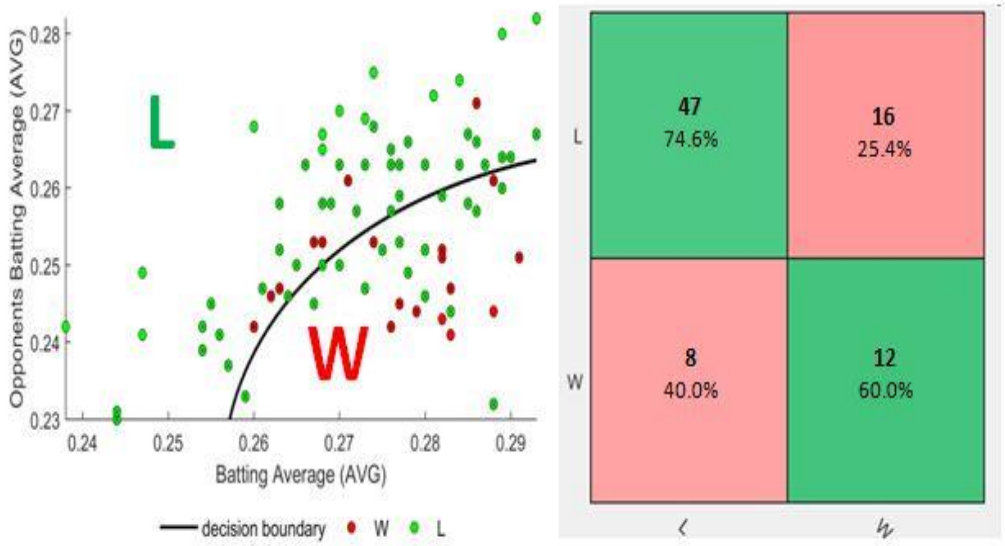
Table 2 suggests the best classifier is the model with the highest gamma value, i.e. Case 4, because it classifies the data points with the highest accuracy of all four cases. Further, it suggests Case 2 is the worst classifier which has a gamma value of 300. However, selection of the appropriate model cannot be justified based on accuracy values alone. Figure 2 clearly shows for increasing gamma values, flexibility of the decision boundary increases. However, for high gamma values, overfitting of the data occurs. This is demonstrated in Cases 3 and 4. Further for low gamma values the model is too constrained and the decision boundary cannot fit the complexity of the data. This is demonstrated by Case 1. Based on visual inspection of the four models, Case 2 was determined to be the best classifier because it did not overfit the data and it aligned with common baseball knowledge. In baseball, your top tier teams are considered to be those teams that hit for average really well and prevent opposing teams from getting base hits. So if given a crossplot with opponent batting average vs. batting average you would expect the championship contenders to fall within a specific region of the chart. This region is highlighted in Figure 3.

Figure 3. *Highlight of the Region One Would Expect Championship Contenders to Fall Into Based Off of Common Baseball Knowledge - the Decision Boundary for Cases 1 and 2 Are also Shown*



It is well recognized from Figure 3 that Case 2 captures more of the highlighted region that one would expect American League champions to be contained in. Based on this result, Case 2 was chosen as the most appropriate classifier to be used for prediction even though case 1 demonstrated a higher accuracy value. A crossplot of Case 2 with its confusion matrix is given in Figure 4.

Figure 4. *The Classifier Selected to Perform Prediction on the American League Pennant Race*



National League Pennant Model Selection

The National League pennant race dataset consisted of 20 years of data with two class labels: National League champion (W) and loser (L). 20 teams were labeled as National League champions and 63 teams were labeled as losers. The dataset is unbalanced, i.e. 24% of the data is classified as National league champions and 76% are classified as losers. Crossplots of the features for the National League presented inconsistences with common baseball knowledge. Figure 5 shows examples of these inconsistencies.

Figure 5 shows that the majority of the National League championship winners (W) do not fall within the region one would expect. Further the data is sporadic. Crossplotting all the attributes listed in Table 1 presented similar inconsistencies. There was no identifiable trend that agreed with common baseball knowledge. After a trial and error approach, an appropriate classifier was developed from the Gaussian kernel RBF SVM. The results are given in Figure 6. The SVM parameters used to develop the classifier and the classifiers' overall accuracy is given in Table 3.

Figure 5. *Highlight Regions One Would Expect National League Champions to Fall Into Based Off of Common Baseball Knowledge*

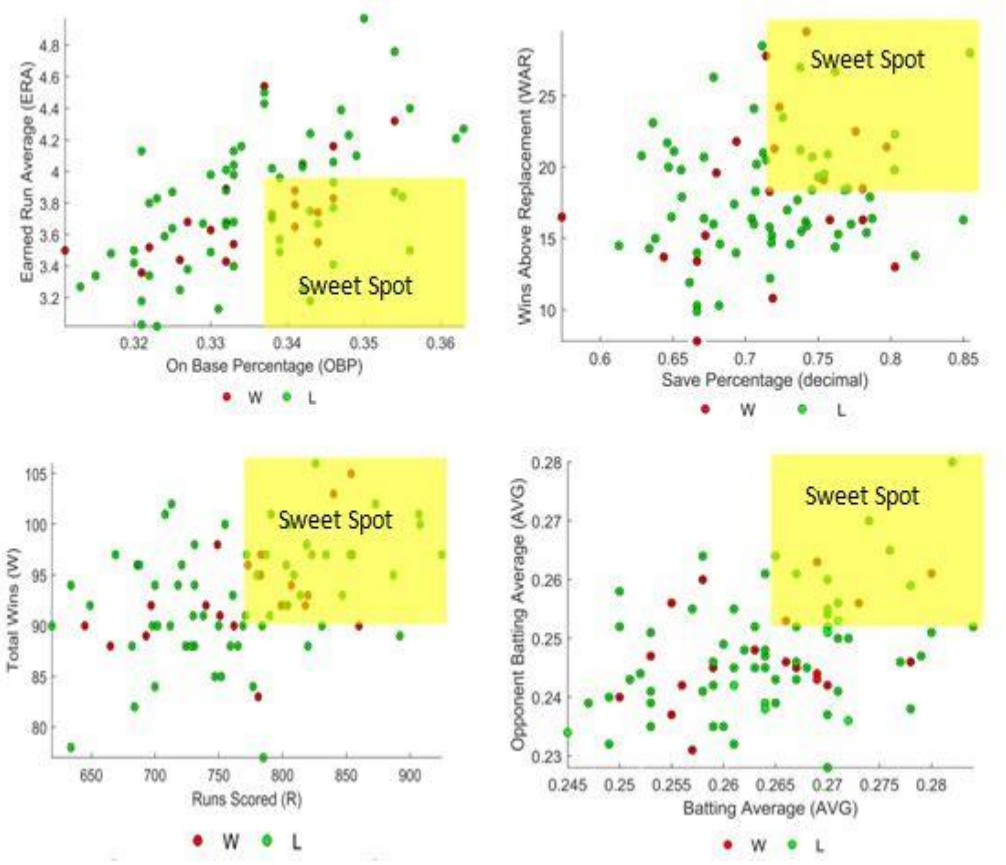


Figure 6. *The Classifier Selected to Perform Prediction on the National League Pennant Race*

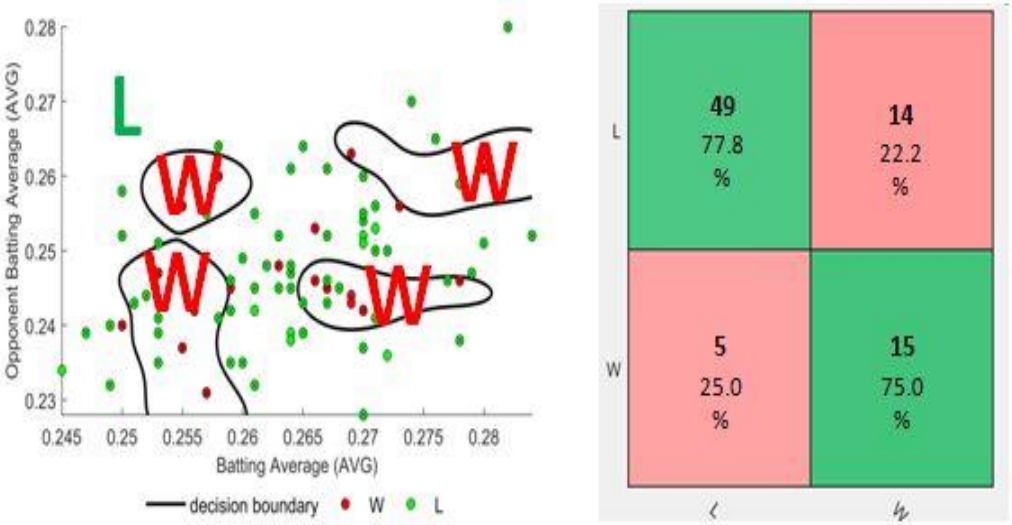


Table 3. *Parameters Used to Develop the Classifier in Figure 6 and Classifier Overall Accuracy*

SVM Model	Cost for Majority (L)	Cost for Minority (W)	Gamma (γ)	Accuracy (%)
Gaussian RBF	1	3	30,000	77.1

World Series Model Selection

The World Series dataset consisted of 42 years of data with two class labels: World Series champion (W) and World Series loser (L). 42 teams were labeled as World Series champions and 42 teams were labeled as World Series losers. The dataset is balanced between the two class labels. Multiple SVM algorithms were compiled to determine which two attributes from Table 1 provided the most accurate classifier. The results for the linear kernel, quadratic kernel, cubic kernel, and Gaussian kernel RBF SVM algorithms are given in the Figures 7-10. Table 4 compares the best classifier between separate SVM algorithms containing two attributes from Table 1.

Figure 7. *The Best Classifier Acquired from the Linear Kernel SVM Algorithm*

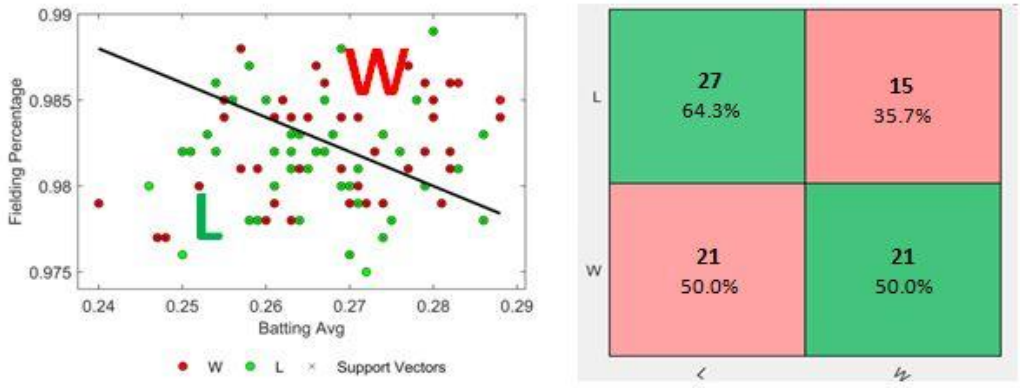


Figure 8. *The Best Classifier Acquired from the Quadratic Kernel SVM Algorithm*

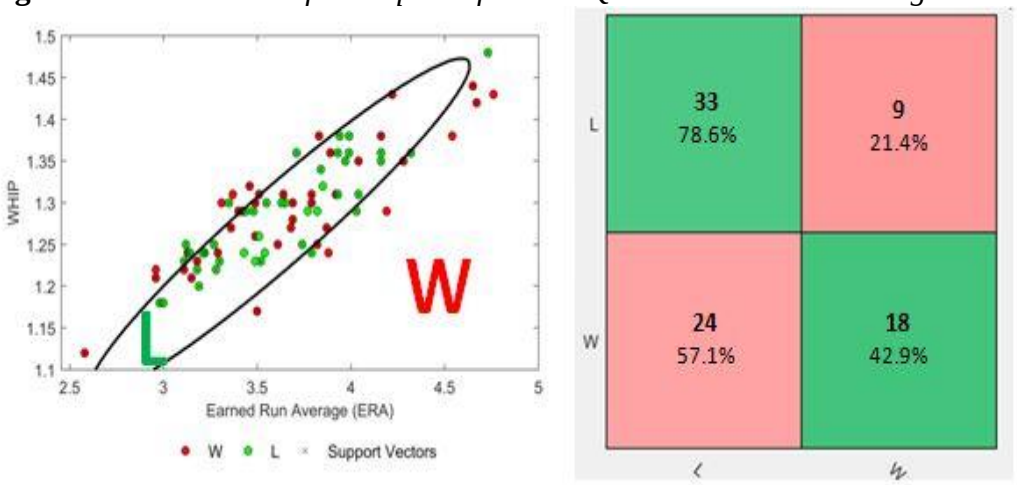


Figure 9. *The Best Classifier Acquired from the Cubic Kernel SVM Algorithm*

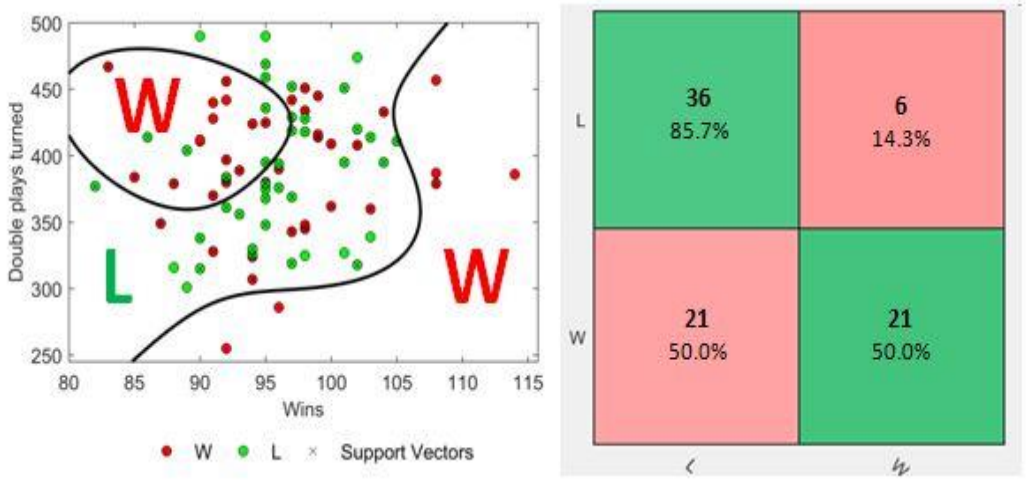


Figure 10. *The Best Classifier Acquired from the Gaussian Kernel RBFSVM Algorithm*

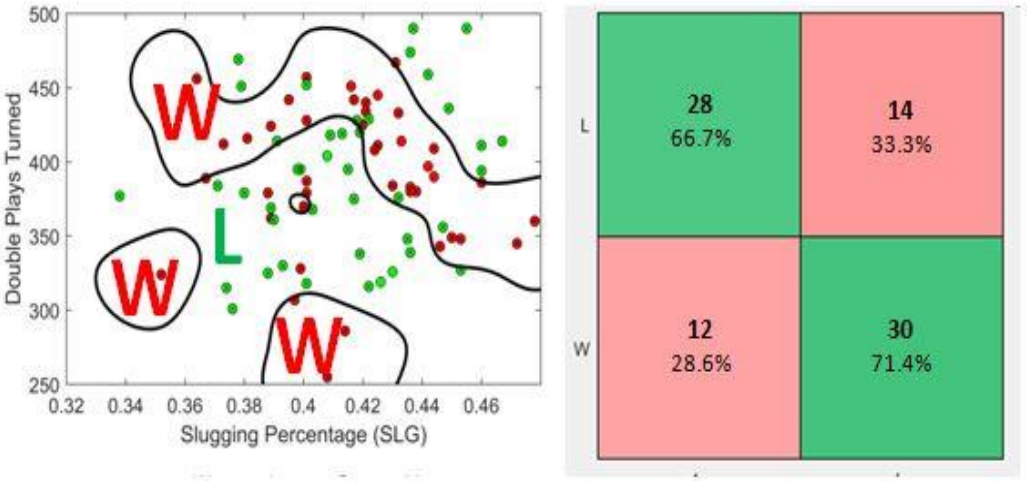


Table 4. *Accuracy Comparisons of the Best Classifier for Each SVM Algorithm*

Machine Learning Algorithm	Model Accuracy	Attribute (1)	Attribute (2)
Linear Kernel SVM	57.1%	Fielding Percentage	Batting Average
Quadratic Kernel SVM	60.7%	WHIP	ERA
Cubic Kernel SVM	67.9%	Double Plays turned	Wins
Gaussian Kernel RBF	69%	SLG	Double plays turned

Table 4 suggests that the least accurate classifier is developed by the Linear Kernel SVM algorithm. Further, the most accurate classifier is produced by the Gaussian Kernel RBF SVM algorithm. This agrees with intuition because the decision boundary becomes more flexible as the degree of a polynomial increases. Thus given data that is not linearly separable as demonstrated in Figures 7-10, one would expect the accuracy of a classifier to increase with increasing polynomial degree.

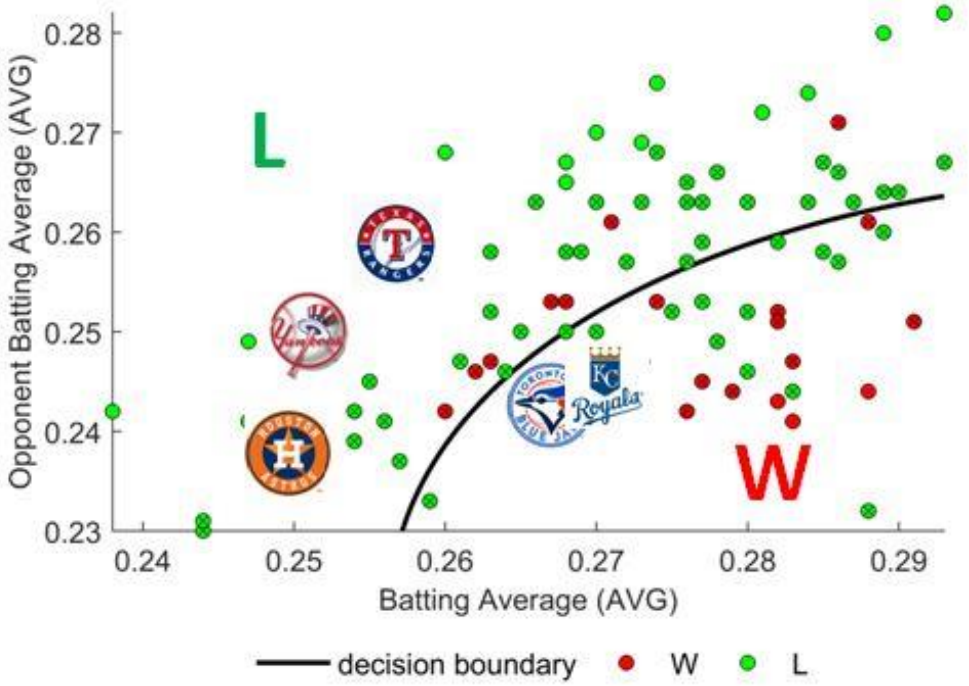
Results/Prediction

The playoff results for the 2015 major league baseball season are shown in Figure 11. As observed in Figure 11, the American League champion was the Kansas City Royals, the National League champion was the New York Mets, and the World Series champion was the Kansas City Royals for the 2015 season. We will evaluate the classifiers developed previously to see if we achieve similar results.

Figure 11. *Playoff Results of the 2015 Major League Baseball Season*

The two dimensional classifier selected to best represent the American League was used to make a prediction of the results for the 2015 American League pennant race. The teams competing in the American League playoff included the New York Yankees, Houston Astros, Texas Rangers, Toronto Blue Jays, and Kansas City Royals as shown in Figure 11. The teams competing in the 2015 American League pennant race were plotted on the classifier developed for prediction. The results are shown in Figure 12.

Figure 12. *Plots the Teams Competing in the 2015 American League Pennant Race on the Best Classifier Developed Previously*

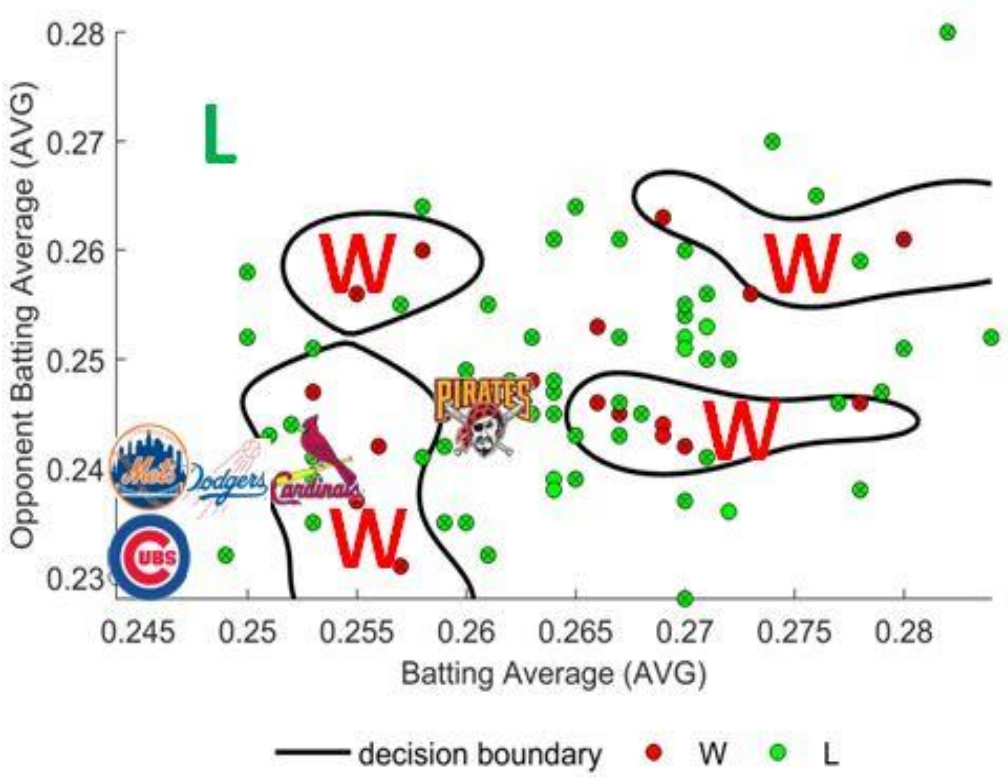


As shown in Figure 12, the classifier predicted two American League champions and three losers. The model successfully predicted the American League champion, the Kansas City Royals, and three of the losing teams. Although the model incorrectly classified the Toronto Blue Jays as a American League champion, they did make it to the American League championship series as shown in Figure 11. Further, the Kansas City Royals and Toronto Blue Jays had identical stats, practically overlaying each other on Figure 12.

The two dimensional classifier selected to best represent the National League was used to make a prediction of the results for the 2015 National League pennant race. The teams competing in the National League playoff included the New York Mets, Pittsburgh Pirates, Los Angeles Dodgers, Chicago Cubs, and St. Louis Cardinals as shown in Figure 11. The teams competing in the 2015 National League pennant race were plotted on the classifier developed for prediction. The results are shown in Figure 13.

As shown in Figure 13, the classifier predicted one National League champion and four losers. The model incorrectly identified the National League champion to be the St. Louis Cardinals. Further, it incorrectly classified the actual National League champion, the New York Mets, as a loser. The prediction from the model shown in Figure 13 did a poor job classifying National League champions and losers from the results of the 2015 season. This was expected because the data did not align with common baseball knowledge for any of the attributes listed in Table 1. Thus for future research it is suggested to evaluate different attributes other than those listed in Table 1 to see if a better result can be achieved for predicting National League champions given the historical data.

Figure 13. *Plots the Teams Competing in the 2015 National League Pennant Race on the Best Classifier Developed Previously*



The SVM models developed from the linear, quadratic, cubic, and Gaussian RBF algorithms were used to make a prediction of the results of the 2015 World Series. The two teams competing in the 2015 World Series included the Kansas City Royals and New York Mets as shown in Figure 11. Prediction results from the SVM models developed previously are shown in the Table 5.

Table 5. *Prediction Results of the 2015 World Series for Several Classifiers*

Machine Learning Algorithm	Model Accuracy	World Series Winner (W)	World Series Loser (L)
Actual Result	-	Royals	Mets
Linear Kernel SVM	57.1%	Royals	Mets
Quadratic Kernel SVM	60.7%		Royals, Mets
Cubic Kernel SVM	67.9%		Royals, Mets
Gaussian RBF Kernel SVM	69%		Royals, Mets

From Table 5, the linear kernel SVM was the only model that correctly classified the 2015 World Series results. All other models classified both the Kansas City Royals and New York Mets as losers. The classifiers inability to classify the results correctly could be attributed to the time span over which the data was evaluated. Over a 42 year time span, the game has evolved and changed. For instance, in the 90's major league baseball experienced an increase in offensive output unparalleled to previous decades. It is believed that steroids and other performance enhancing drugs (PEDs) contributed to the increased offensive output. Since 2003, Major League Baseball has implemented league wide PED testing and enforced severe consequences for players that use steroids or other forms of PEDs to try to gain an unfair advantage when competing. Subsequently, the quantity of players taking steroids in today's game has decreased. Future research should consider a smaller subset of data that more closely resembles the way the game is played today to see if a better classifier can be developed to predict World Series winners.

Conclusion

In this work we applied SVMs to develop a model to predict championship winners and losers for Major League baseball. In summary, the following conclusions were derived from this work:

1. The challenge of unbalanced data was addressed by assigning different costs values for misclassification of each class. Appropriate cost values was determined to be three for the minority class (W) and one for the majority class (L) for both the National and American League pennant races.
2. An appropriate value of gamma for the Gaussian kernel was determined through a trial and error approach. When the gamma value was too large, the model overfit the data. When the gamma value was too small, the model was too constrained and could not capture the complexity or shape of the data.
3. The classifier developed for the American League pennant race agreed fairly well with common baseball knowledge and the prediction was acceptable.
4. The classifier developed for the National League pennant race did not agree with common baseball knowledge and the prediction results were poor.
5. The classifiers developed for the World Series game produced mixed results. The classifier that produced the highest accuracy was the Gaussian Kernel RBF SVM.

While SVM models provided mixed results, future research should consider introducing newer baseball statistics to the models in an effort to develop better classifiers. Additional data analysis can determine what modifications can be made to improve the results. This can include tweaking the dataset or tuning parameters that influence the decision boundary. Through a trial and error approach a more appropriate classifier most likely can be developed.

Acronyms

L	Losing Class
OBP	On-Base Percentage
PED	Performance Enhancing Drugs
RBF	Radial Basis Function
SLG	Slugging Percentage
SVM	Support Vector Machine
W	Winning class

Nomenclature

γ	Gamma value
C	Soft margin constant

References

Ben-Hur A, Weston J (2010) A User's Guide to Support Vector Machines. In O Carugo, F Eisenhaber (eds.), *Data Mining Techniques for the Life Sciences* (pp. 223-239). Humana Press.

Chang, C-C, Lin C-J (2011) LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3): 1-27.

Egros R (2013) *Forecasting MLB World Champions Using Data Mining*. *Fox Sports Dallas*. Retrieved from goo.gl/IwIpm2

Lewis M (2003) *Moneyball: The Art of Winning and Unfair Game*. W.W. Norton.