Williams Consulting

Predicting Wins/Losses and Game Spread with NFL Data

Data Mining

Immanuel Williams
12/20/2014

Contents

Executive Summary

In football, sports analyst and fans are consistently trying to predict which team is going to win and by how much. Most sports analyst discuss key players, match ups and coaching when it comes to a team winning. However, little is said about the utilization of statistical models to predict victories or the point spread. The purpose of this report is to use past game statistics to predict whether a team wins or loses and to predict the spread of a game as well. The data used in this project was extracted from a website www.pro-football-reference.com. The data found at this website was then manipulated so that previous games statistics such as yards, points, point difference, turnovers and average wins were used to predict game outcomes and game spreads. Based on the statistical models used in this paper, the implementation of support vector machine and quadratic discriminate analysis were good methods used to predict the outcomes of games. However, the methods implemented in this paper to predict games spreads did not perform well at all.

Introduction

Exploring what makes a team win is important not only to people who are passionate fans but also to other stakeholders who watch these games religiously. Based on these statistical models, team owners, general managers and coaches will be able to determine the outcome of each game which will allow them to make appropriate adjustments to ensure an upset or maybe a closer game. The National Football League (NFL) will be able to schedule games in such a way that close games (small point spread) will be scheduled during primetime to ensure maximum viewers. Cable companies could also use this information to determine what type of advertisement should be played during certain games due to the fact that if a game is going to be close, the cost of advertisement should be higher compared to when a game is going to be a blowout (large point spread). These techniques could also be applied to other sports to ensure a certain level of watchers.

There are multiple of studies that examine predicting the probability of a team winning and by how much. One study analyzed determining the probability of a favored team beating an underdog team by p points (Stern, 1991). This work only looked at 5 years of data and did not utilize techniques discussed in this paper. There has also been research that evaluated how a community of NFL fans has the ability to predict future game wins (Szalkowski & Nelson, 2012). Other work used twitter as source to predict wins (Sinha et. Al, 2013). However little research has used these variables and statistical models discussed in this paper to predict wins, losses and point spread.

The subsequent section describes the derivation of the data that was extracted from the website. Then the following section discusses the statistical models used to predict game outcomes and point spread. The final section will review the findings and its implications as well as discuss future research.

Data Derivation & Summary

Once the data was extracted from the website, a certain level of cleaning and organizing was done in order to acquire information from the data. This included removing playoff and super bowl games, reformatting the data to include the past 12 years (2002 to 2013). This also included manipulating the data so that the last 6 games of each season were treated as the response data and the current season (previous 10 games) and previous two years of game statistics as the predictors. The exclusion of the playoff and super bowl games was done so the data was not inflated by non-random data. Reformatting and manipulating the data was done for two reasons:

1) Ensure that there was data for all 32 teams (before 2002 there were 31 NFL teams)
2) Utilize the current season statistics and previous 2 seasons data in prediction

Once the formatting was done, 45 predictors were created based on 5 variables. These 5 variables were yards, points, turnovers, point difference and average wins. The derivation of the 45 variables was accomplished by splitting the current season and the previous 2 seasons into three groups. Each group represented the beginning, middle and end of the season affect. The average of the 5 variables was calculated for each group with respect to each team for the last 6 games of each season. This was done to mode teams streakiness (win or lose games consecutively).

A binomial distribution with a probability of 0.5 was used to randomly determine which game was going to be used as a win or a loss. The wins/losses are then used as the response variable. The wins/losses are used to determine the point difference which is used as the game spread response variable.

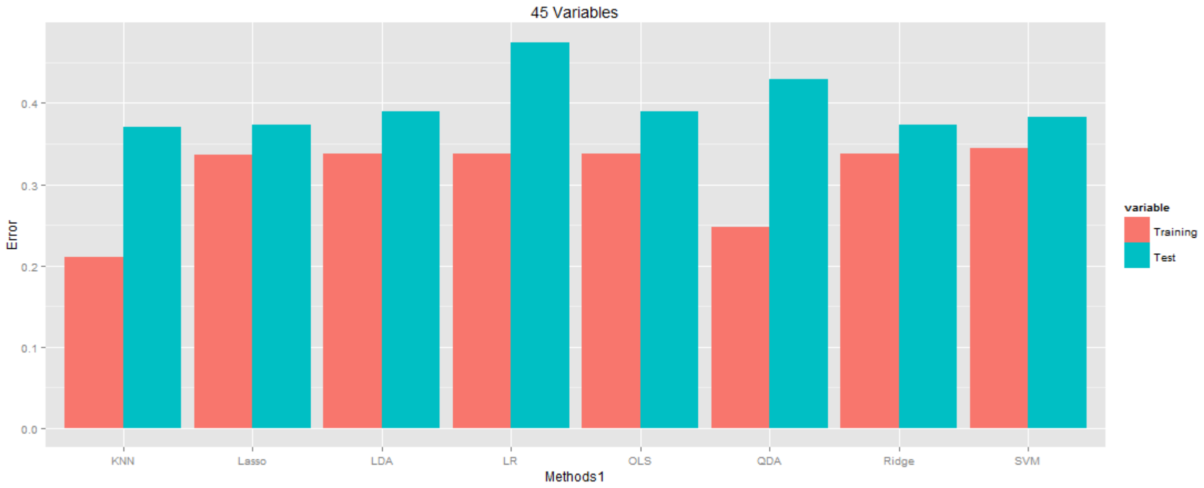
Due to the number of variables, generalized descriptions are given regarding the data. The average and standard deviation of the amount of yards variables seemed to be around 330 and 50, respectively. The point amount variables were generally around 20 for the mean and 5 for the standard deviation. The turnover variables were around 2 for the mean and 0.6 for the standard deviation, whereas the point difference tended to have a small mean around 0.5 and standard deviation around 8. Another way of looking at this fact is to notice that some games are close and some games are blow-outs, thus the small mean difference between points and large standard deviation. The average wins seemed to have a mean around 0.5 and standard deviation around 0.24.

Analysis

Prediction of Outcomes

Before any of the statistical models were used, the data was split into two data sets, training and test. This was done at random using the sample function in R. This was done to verify the statistical methods. The size of the training set was 958 games and the test data set contain 300 games. Once the data was split, the statistical models used were the ordinary least square (OLS), logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), support vector machine (SVM), k-nearest neighbor (KNN), Ridge and LASSO regression to predict game outcomes. The results can we be seen in figure 1.

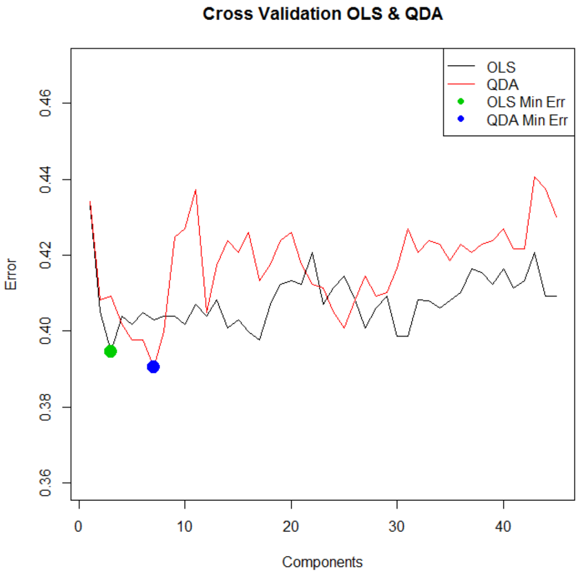
Figure 1.



The misclassification rate was used to measure the performance of each model with respect to game outcome. The results show that KNN and QDA provide the lowest training data set error and KNN, Ridge and LASSO gives the lowest training error. Using SVM the best condition is when the cost function was set at 0.01 and the γ value was set at 0.022. The KNN best condition for the training data set was at N=3 and N=22 for the test data set. With respect to the Ridge and LASSO regression the best tuning parameters for each model is λ=0.011 and λ=0.001, respectively.

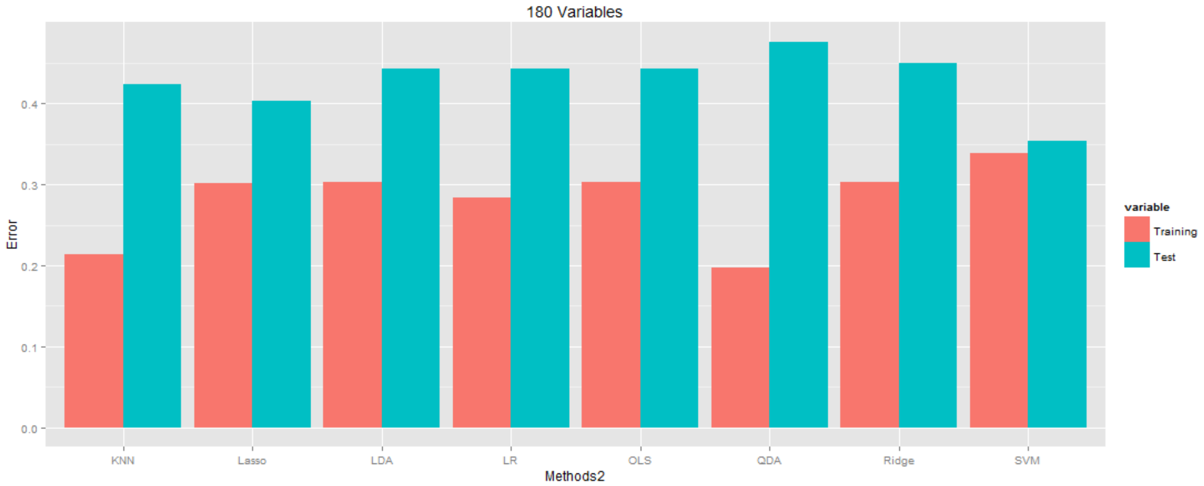
Once this analysis was done a cross validation was completed on the OLS and QDA methods with respect to the number of principal components. Figure 2 shows graphs the number of components and error. The points on the graph denote the minimum error for OLS and QDA which are 0.394 with 3 components and 0.390 with 7 components, respectively.

Figure 2.



A basis expansion was then used to increase the number of variables. This was only done on the current season variables (polynomial equal to 10) which turned the number of predictors into 180. Figure 3 denotes the error found using the same statistical models used within the first analysis. Once, again the QDA and KNN outperforms the other methods with respect to the training data set and the SVM method produces the smallest amount of test error, which used the cost function of 0.001 and γ equal to 0.0005. The KNN used the N=2 for the training data set and N=21 for the test data set. The λ's for the ridge and LASSO regression were 0.0031 and 0.0051 respectively.

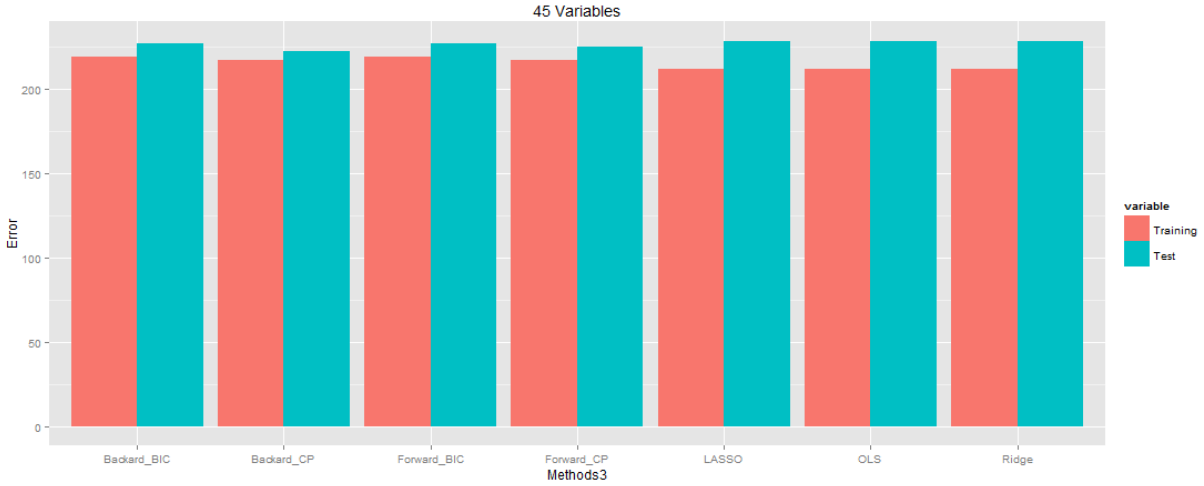
Figure 3.



Prediction Game Spread

As in with the prediction of outcomes the data set is randomly split into two data sets training and test. Instead of using all the statistical methodology used in the previous section, the use of only the OLS, Ridge and LASSO were used in predicting game spread. The other methodologies used were forward selection and backward elimination with respect to the best Mallow's Cp and bayesian information criterion (BIC). Figure 4 highlights the findings of using these statistical models.

Figure 4.



The mean square error was used to measure the precision of each statistical model. Overall, each model did not perform very well for both training and test data sets. The best λ for both Ridge and LASSO regression was set at 0.001. The best number of variables for both backward elimination and forward selection with respect to Cp was 7 and 4 for BIC.

Conclusion

Predicting game outcomes and game spreads are important but difficult tasks. In this line of research not only do the statistical models have to be highly discriminative and predictive but the data has to be derived in such a way that the methods can be useful. This can be said about predicting game outcomes. The KNN, QDA and SVM worked reasonably well when it came to misclassification of outcomes with respect to both training and test data sets. On the other hand, the prediction of game spread was not estimated well using any of the statistical models. There are two reasons why these models probably could not predict game spreads well. The first reason stems from the derivation of the data, one may say the variables used and the way they were organized would not predict the data well. Another reason is that the absolute value of the game spread was not implemented thus, the large mean squared error.

There are many ways to improve this study. One way is to include more types of variables such as number of first downs, number of penalties and number of touchdowns per game. This is important because it will provide more information about how well a team performs which will lead to better predictions. Another way to improve this study is to take the absolute values of the of the game spread response variables. This will allow for better accuracy with respect to prediction. Lastly, once the incorporation of more diverse variables are included into the data set, dimension reduction tools such as principal component analysis and fisher discriminant analysis should be implemented to ensure precision.

References

Szalkowski, G., & Nelson, M. L. (2012). The Performance of Betting Lines for Predicting the Outcome of NFL Games.
Sinha, S., Dyer, C., Gimpel, K., Smith N., A. (2013). Predicting the NFL Using Twitter.
Stern, H., (1991). On the Probability of Winning a Football Game.