NBA Game Outcome Predictor

Yuxi He, Sai Praneeth Kolli, Yang Zhang

Introduction:

In this project, we analyzed the NBA game statistics and made predictions on game outcomes for the 2013-2014 season.

Professional sports, along with sports betting and lottery have been affecting people's life spiritually and financially for decades. Major, professional sports leagues such as the NBA, NFL, and MLB contain a significant amount of easily accessible data whose outcomes tend to be randomly distributed and offer attractive data for analytical purposes. Predicting the outcomes of sporting events represents a natural application for machine learning. Above all, machine learning can make hidden data trends come to surface, which will help coaches and management levels make better decisions and strategies. Moreover, precise game outcome prediction is also critical to sports bettors for obvious reasons. In a nutshell, people's personal opinions, professional or not, may differ from what the data says and what the truth is. As a consequence, using machine learning on statistics has a big chance to beat the NBA odds for a competitive or financial advantage.

Dataset description:

We extracted and processed the raw data from www.basketball-reference.com. Our training datasets came from the 911 games played before March 6, 2013 in season 2013-14. We used 'win-loss' as a binary classification where '1' represents the home team won the game and '0' represents the visitor team won. We selected various attributes to learn the training datasets, which will be elaborated in the analysis section. We collected the outcomes of the 89 games played between March 6, and March 17 as testing data.

Analysis:

Initially we did analysis and evaluation only on data containing attributes that we chose with expert knowledge, such as team's defensive rating, pace, etc. It made sense for us to choose these attributes as they could be the prime factor deciding a win.

Then we took data with all the attributes possible. As a first step, we evaluated on all attributes possible. A Chisquare test was used to analyze a set of 140 attributes of team statistics to determine the subset of features to be used in the prediction models. Using a Chisquare value of 0.05 as a cutoff (all statistics below this value were subsequently disregarded), 20 features were then selected by this method. We performed an analysis on data with these 20 attributes. Finally we removed attributes checking whether their deletion led to improvement in accuracy. This left us with 18 attributes. The following is list of attributes used.

h_MOV - Home team Margin of Victory
h_PL - Home team Pythagorean Losses (i.e expected losses based on points scored and allowed)
h_PW - Home team Pythagorean Wins (i.e expected wins based on points scored and allowed)
h_SRS - Home team Simple Rating System (i.e a team rating that takes into account average point differential and strength of schedule)
v_PL - Visitor team Pythagorean Losses (i.e expected losses based on points scored and allowed)

t_h_FGPercentage - Home team Field Goal Percentage
h_ORtg - Home team Offensive Rating
t_h_2PPercentage - Home team 2-point Field Goal Percentage
v_SRS - Visitor team Simple Rating System (i.e a team rating that takes into account average point differential and strength of schedule)
v_MOV - Visitor team Margin of Victory
v_PW - Visitor team Pythagorean Wins (i.e expected wins based on points scored and allowed)
t_h_PTS - Home team total Points
h_SOS - Home team Strength Of Schedule
h_off_eFGPercentage - Home team
t_h_DRB - Home team Defensive Rebounds
h_AST - Home team Assists
v_DRB - Visitor team Defensive Rebounds
h_DRtg - Home team Defensive Rating

The plot matrix in Figure. 1 represents how our data is distributed and visualized in respective to a few attributes. Figure. 2 is clear representation of one block from Figure. 1.
We used Naive Bayes classifier and decision tree learning algorithms on our data set. We tried all possible classifiers and these two tended to perform well and also scaled well. Hence we chose these two classifiers as our learning algorithms.
Naive Bayes outperforms all classifiers clearly for our task. We used cross-validation with 10-folds for our validation strategy to validate our classifier. Then we used 'future' data i.e game outcomes from March 6 to March 17, as our test set and tested our model on it. The average accuracy observed for Naive Bayes was "68.25%", followed by Decision Tree with an average accuracy of "64.71%". Figure. 3 depicts more clearly how our classifiers perform and scale.

Results:
Using manually chosen attributes our accuracy was only "62 %" on average. But with attributes chosen by Chisquare test we were able to achieve accuracy of "65%" on average and by removing attributes that enhanced accuracy finally we were able to achieve around "67%". This proves that selecting attributes using learning algorithms perform well compared to expert knowledge. Even though there is a similarity in attributes chosen by algorithm and expert knowledge.
Also from the chosen attributes and increasing in accuracy, we noted two important features. Home team statistics play an important role in deciding who wins or not. As most of the attributes chosen are related to Home team's performance. This makes sense because home team has a bit of advantages and plays a major role in determining result.
The data consists of 140 attributes taken from three different statistics "Home performance", "Opponent performance" and "General Performance" statistics of a team. From our analysis we noticed that most of features selected are from "General Performance" statistic like "Defense rating", "Offense rating", etc. This inference tends to be correct as some betting websites also tend to use only these statistics to set the odds. And these attributes tend to be more meaningful in determining the result.

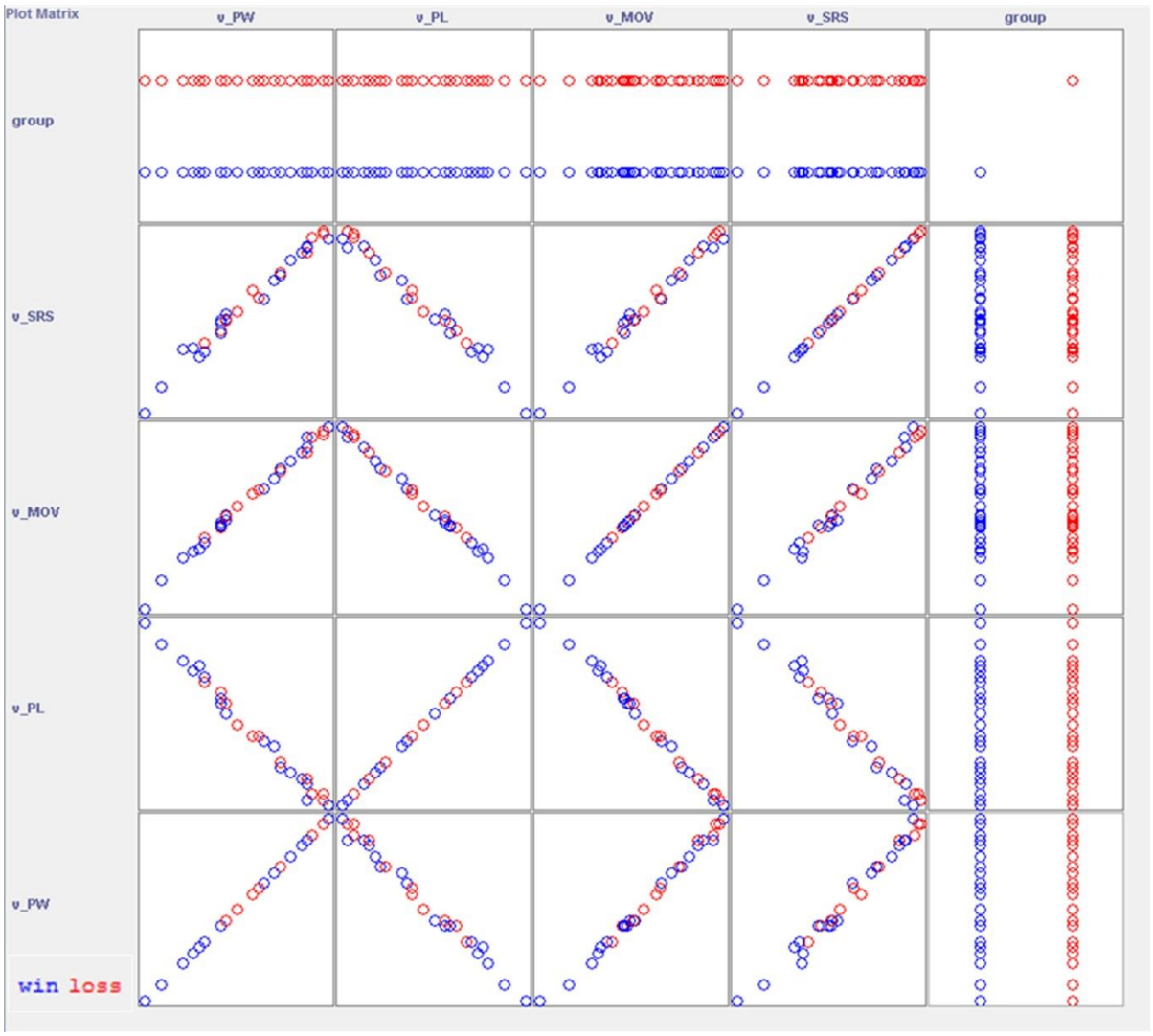
Plot Matrix
v_PW
v_PL
v_MOV
v_SRS
group
group
v_SRS
v_MOV
v_PL
v_PW
win loss

Figure. 1

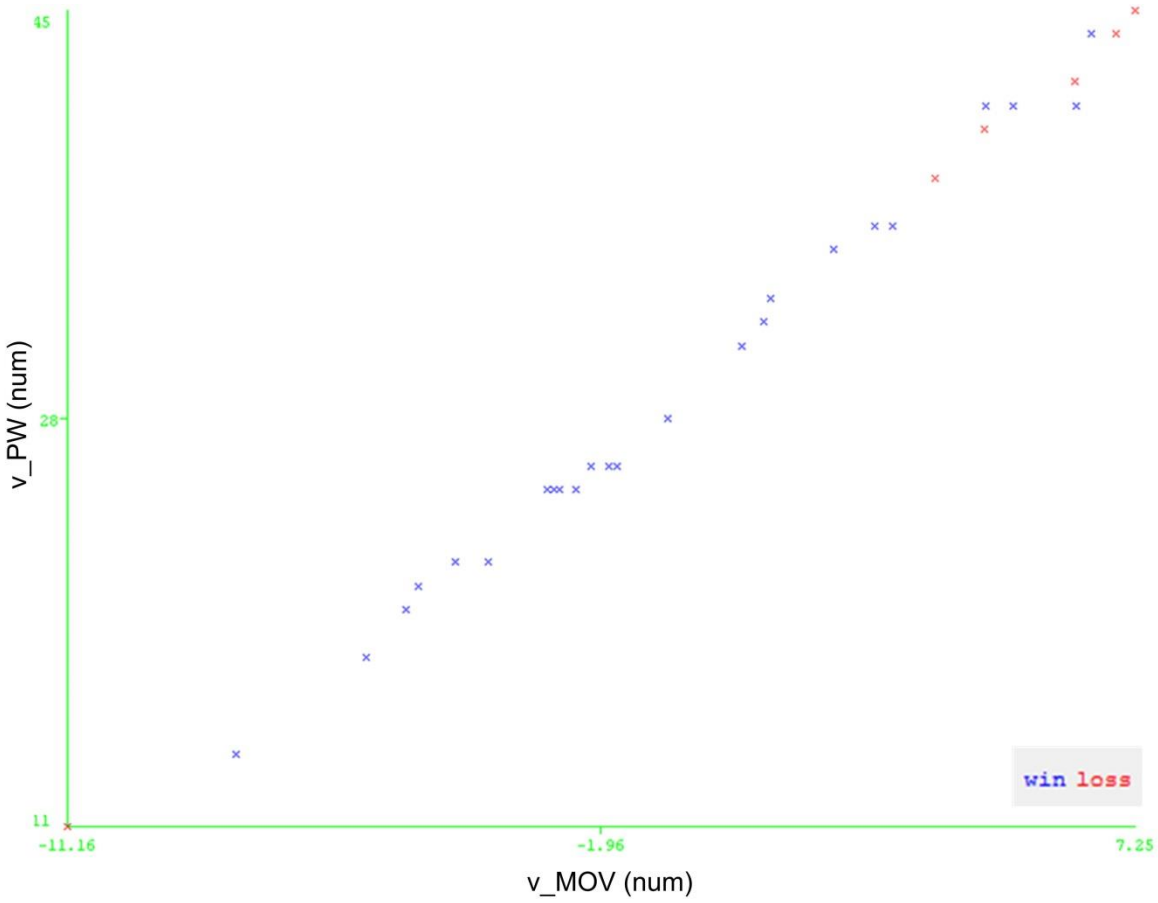

Figure. 2

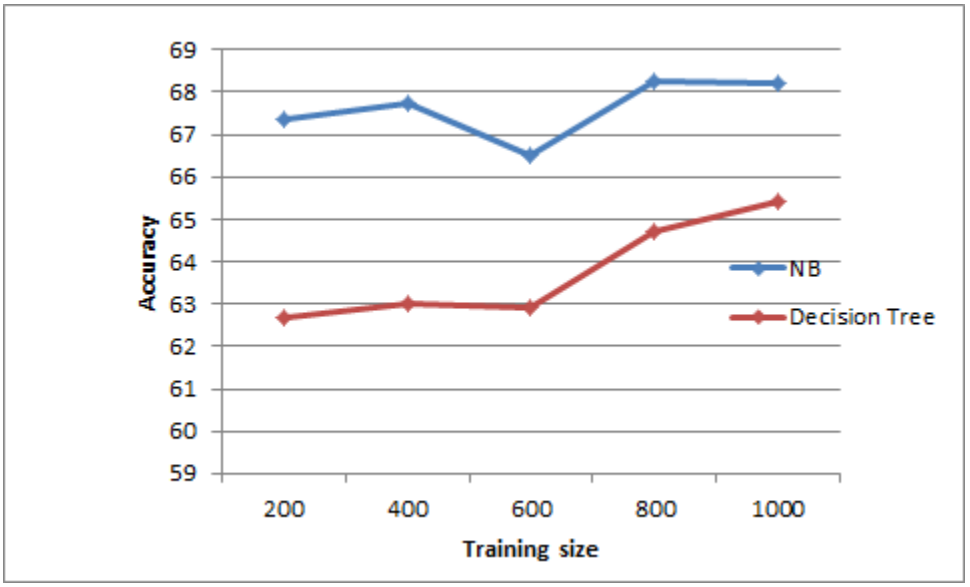

Figure. 3