Outlier Detection and Game Outcome Prediction of NBA Game

Aravind Anantha
aananth3@ncsu.edu

Abinav Pothuganti
apothug@ncsu.edu

Chethan Thipperudrappa
cnanniv@ncsu.edu

Huy Tu
hqtu@ncsu.edu

Abstract

The plausibility of predicting sporting events', simulating games, and estimating the performances of athletes is naturally attractive to businesses, coaches, and researchers. Especially, that plausibility for such major and professional sport league as National Basketball Association (NBA) has influences and fame across countries all around the world. For this milestone of the project, the main objective is to report current progress and experiments' results from using various Machine Learning methods.

1 Introduction

Thirty teams comprise two conferences in the NBA. Throughout the regular season, there are 1230 NBA games in total. It brings in significantly monetary opportunities for businesses, enterprising sport gamblers. The huge fan following and increasing interest motivates us to predict the MVP (Most Valuable Players) and game outcome. This milestone project report would explore prior researches that have been done relatively similar to NBA game outcome and outstanding players prediction.

There are 34 features for our NBA dataset. We use data beginning from the 1980 season since that is when 3 pointers were first introduced. The input is a team feature vector containing the 34 features for all NBA teams for several seasons. Each data point is the performance of each team in a season [1]. Various approaches for feature selections (BIC, PCA, and stepAIC) would be applied along with linear regression model and SVM Regression for team win ratio of per year prediction. Outstanding players are predicted based on the various attributes such as rebounds, steals, three points etc. We are making use of multivariate outliers library in r to perform outlier detection.

Our midterm report is constructed as follow: (2) report on related work has been done in the past, (3) proposed methodologies for features, models, and algorithm selection, (4) results from our experiments with those methodologies, and finally (5) ways to improve our model before the final report.

2 Related Work

There have many work on predicting NBA team wins on per-game level and predicting the win percentage of that team in that NBA season with or without basing on that per-game level prediction. One paper by Evan Giarta and Nattapoom Asavareongchai predicts the win percentage of a team in a season using the per season cumulative stats of all players with PCA for feature selection [2]. The learning model used was a second order polynomial regression model [2]. Moreover, Bernard, L et al.'s paper in 2009 and Pedro et al.'s paper in 2012 used network based algorithm to predict NBA wins with very high success rate [3] [4]. Kevin Wheeler in 2012 used Linear Regression for player performance prediction along with SVM and Naive Bayes for outcome per-game prediction with PCA for feature reduction [6].

Mahalanobis distance is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D. It measures the number of standard deviations from P to the mean of D thus is is unitless and scale invariant.As per the paper "An Overview of Multiple Outliers in Multidimensional Data"[5], Mahalanobis distance of the points always falls in Chi Squared distribution and if a data point is not in Chi Squared distribution then it can be possible outlier. Calculation of Mahalanobis distance of all the observations can be done using the MCD(Minimum Covariance Determinant) estimates. Calculation of MCD(Minimum Covariance Determinant) [3] is a very computationally expensive operation. MCD estimates are then used to compute the Mahalanobis distance of all the observations to detect outliers.

3 Proposed Methods and Experimental Results

3.1 Feature Selection

First, best feature selection models are applied:

- BIC is an estimate of a function of the probability of a model being true, under a certain Bayesian set among to the sets of candidates in which indicates a lower BIC means that a model is considered to be more likely to be the true model. There are forward and backward selection, we applied backward selection. In our experiment, we got 7 features being selected in the end for the lowest BIC(o_fgm, o_ftm, o_3pm, d_fga, d_ftm, d_fta, and d_3pa).
- StepAIC is an estimate of a constant plus the relative distance selecting a model that describe the unknown true likelihood function of the data and the fitted likelihood function of the model which indicates that a lower AIC means a model is considered to be closer to the truth. In our experiment, we got 13 features being selected in the end for the lowest stepAIC (o_fgm, o_fga, o_ftm, o_oreb, o_asts, o_to, o_3pm, d_fgm, d_fga, d_ftm, d_to, d_blk, and d_3pm)
- PCA uses orthogonal transformation to transform a set of possibly correlated variables into a set of linearly uncorrelated variables called principal components. We used the preProcess function from the caret library for the task of PCA. In our experiment we choose principal components that explain 95% of the variance in the data. This came out to choosing the first ten principal components. The Table 1 below shows the results of our PCA.

Table 1: PCA for Team Performance Prediction

PC's	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Variance	0.69	0.14	0.05	0.03	0.02	0.016	0.01	0.006	0.005	0.004
Cumulative Variance	0.69	0.84	0.89	0.92	0.94	0.956	0.966	0.972	0.977	0.981

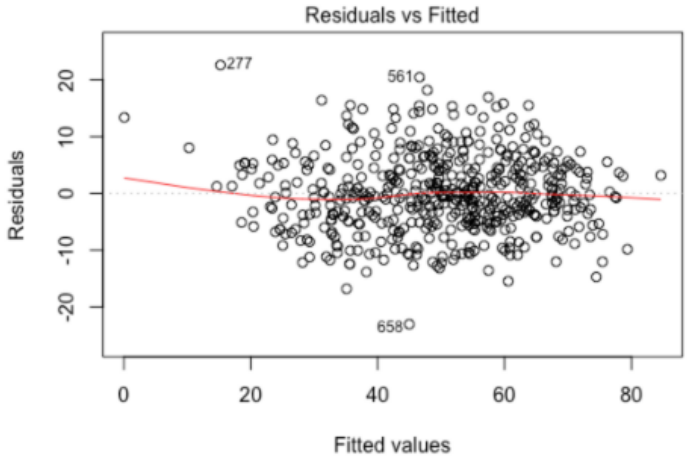
3.2 Team Performance Prediction

Feature reduction methods as stated above were used to reduce the number of dimensions. Different machine learning algorithms were used including linear regression (LR), polynomial regression, logistic regression and support vector regression(SVR) for the task of predicting the wins per season. Our baseline method is the linear regression model that runs on all of the features. Using 75% of the dataset as the training data and the remaining 25% as test data. The data was split into training and test randomly using the sampling method without replacement.

Table 2: Team Performance Evaluation

Root Mean Square Error in %	LR with PCA	LR with stepAIC	SVR with stepAIC	SVR with PCA	SVR with BIC	Baseline Method
Train error	6.785	3.83	4.34	5.03	9.59	77.3
Test error	6.85	4.05	4.62	7.24	10.63	73.5

The results of the experiments that are evaluated with root mean square errors are listed below in table 2 above. Among the methods that were applied, the simple linear model with stepAIC showed the best prediction with lowest root mean square errors in both testing and training data.



The figure above shows the residuals vs fitted values for the linear regression model generated from the data selected using PCA. The residuals vs fitted values curve shows that the errors are all random around the zero line, which shows that the linear assumption is reasonable. Also, we can make out that the variance of the error terms are equal and that the error terms are uncorrelated.

3.3 Team Performance Classification

From win ratio, to give audience better idea of how well the team perform per season, labels from 1-4 will be applied as 1 for win percentage 0% to 35%, 2 for 35% to 50%, 3 for 50% to 70% and class label 4 for a win percentage greater than 70%. The table below shows the precision, recall, f-measure and accuracy of the model after applied SVM model for classification.

Table 3: PCA for Outlier Detection

	Precision	Recall	F - Measure	Accuracy
Class: 1	0.826087	0.52777778	0.6440678	0.7490741
Class: 2	0.56	0.59574468	0.5773196	0.7091627
Class: 3	0.6701031	0.91549296	0.7738095	0.7977465
Class: 4	1	0.05882353	0.1111111	0.5294118

3.4 Outstanding Player Detection

The dataset contains has 5 parts i,e (i) Statistics of players in regular season(19113 records) (ii) Statistics in playoff season(7544) (iii) Statistics in all star season(1463 records) (iv) Statistics in playoff career(2056 records) (v) Statistic in regular season career (3760 records). Though the data corresponds to different seasons the attributes are common among all of them. There are 23 attributes in each dataset and 6 of the attributes are nominal and they are not used in the analysis. Remaining 17 attributes are numerical attributes and they are scaled to lie between 0 and 1 before performing the analysis. We have considered the dataset from the year 1980 as some of the attributes have missing values prior to 1980.

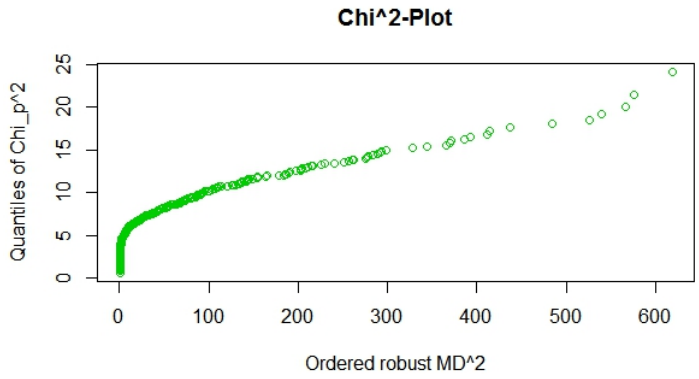
In case of player_regular_season data, we are aggregating the player statistics based on the player name and year and then apply the outlier detection algorithm to find the outstanding players in a season. But we can reduce the number of attributes by applying PCA technique, in particular we used prcomp method in r to find the PCs. The PCA results of the dataset are as below:

From the above list of PCs, 5 Principle Components were considered for the analysis by keeping 95% of the variance in the data. Outliers are the ones which deviate from the other data points the most.

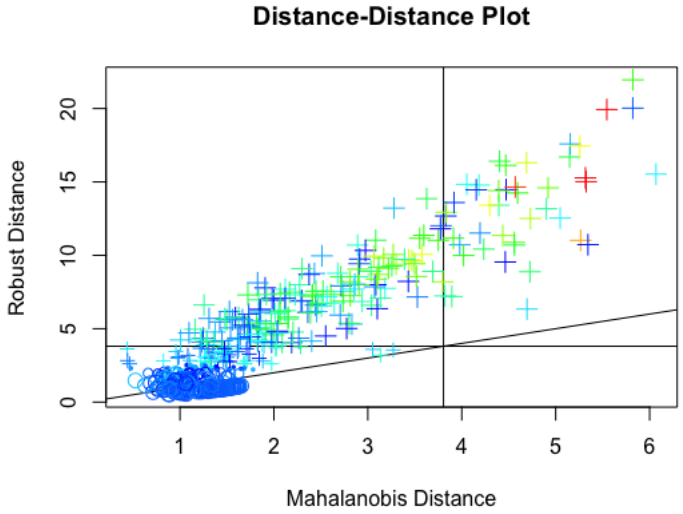
Table 4: PCA for Outlier Detection

PC's	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Variance	0.59	0.12	0.04	0.03	0.03	0.02	0.02	0.02	0.01	0.01
Cumulative Variance	0.59	0.72	0.77	0.8	0.83	0.86	0.89	0.91	0.93	0.95

Chi-squared plot function is used to detect the outliers in the data. The function chisq.plot plots the ordered robust mahalanobis distances of the data against the quantiles of the Chi-squared distribution.



The function dd.plot plots the classical mahalanobis distance of the data against the robust mahalanobis distance based on the mcd estimator.



Method dd.plot is not viable because it is predicting nearly one third of the dataset as the outliers.

4 Appendix

- Game outcome prediction was planned to be done using only classification model but now we are using both regression model and classification model.
- To improve we may use neural network along with SVM.

Table 5: Work Distribution and Future Plan

	Precision	Recall	F - Me
Abinav	Chi-Squared method to detect outliers	Try different models	
Chethan	dd.plot method to detect outliers.	Improve accuracy of outlier detection	
Huy	Step AIC, BIC for Feature Selection with SVR and LR with SVM Classification	Try Neutral Network and other Classification Algorithm	
Arvind	PCA for Feature Selection with LR and SVR with SVM Classification	Try Logistic and Polynomial Models	

References

[1] Basketball Reference. http://www.basketball-reference.com/

[2] Giarta, E. & Asavareongchai, N. (2015) *Predicting Win Percentage and Winning Features of NBA Teams*.

[3] Loeffelholz, B. & Bednar, E. & Bauer, K. W. (2009) *Journal of Quantitative Analysis in Sports*.

[4] Melo, P. O. S. V. D. & Almeida, V. A. F. & Loureiro, A. A. F. & Faloutsos, C. (2012) *Forecasting in the NBA and Other Team Sports: Network Effects in Action*.

[5] Sajesh, T. A. & Srinivasan, M. R. (2013) *An Overview of Multiple Outliers in Multidimensional Data*.

[6] Wheeler, K. (2012) *Predicting NBA Player Performance*.

[7] Yang, S. (2015) *Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics*.