Predicting Win Percentage and Winning Features of NBA Teams

Nattapoom Asavareongchai, Evan Giarta

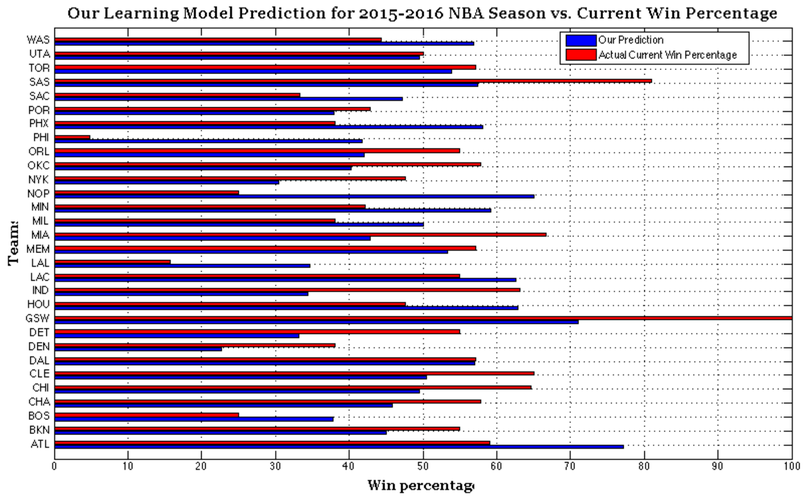
CS 229 (Autumn 2015)

Objective

The NBA is the premier basketball league of the world. It is an enormous business, with an estimated revenue of nearly $5 billion in 2014. Teams within the NBA with more wins will gain popularity and increase income, beneficial to the league and its players. Thus, predicting the number of wins of an NBA team based on the previous performance their roster is quite valuable to general managers, coaches, players, fans, gamblers, and statisticians alike. Moreover, knowing which particular stat or feature is most influential to winning games is also desirable.

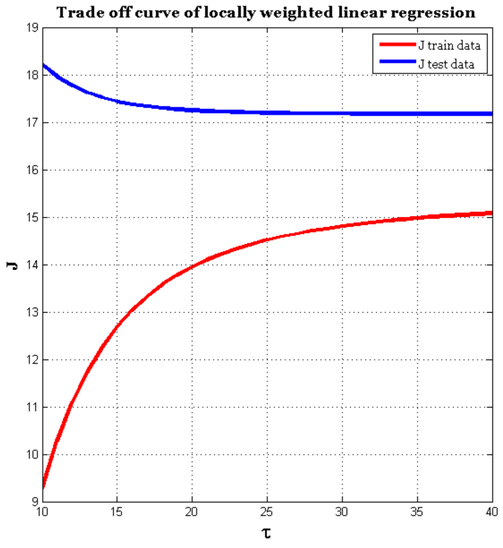
Data and Methods

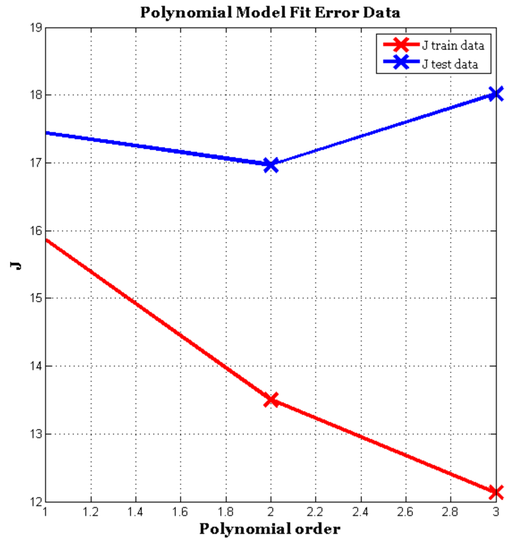
We collected individual player statistics and team win percentage for the past decade to help create a model of predicting a team's success. Our model takes a team's roster from one season and individual player statistics from the previous season to create feature vector of estimated team statistics for that season. We then ran the data on different learning models, including locally weighted linear regression, linear regression and polynomial fit, to derive weighting parameters. We then chose the best model through k-fold validation, choosing the one with the least generalized error. With our best model chosen, we made improvements through feature selections. With detrimental features removed and features reduced, we reran our model to get the final weighting parameters with better results.

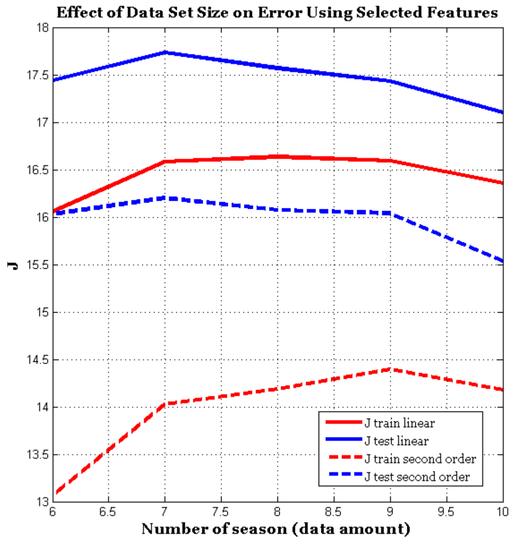


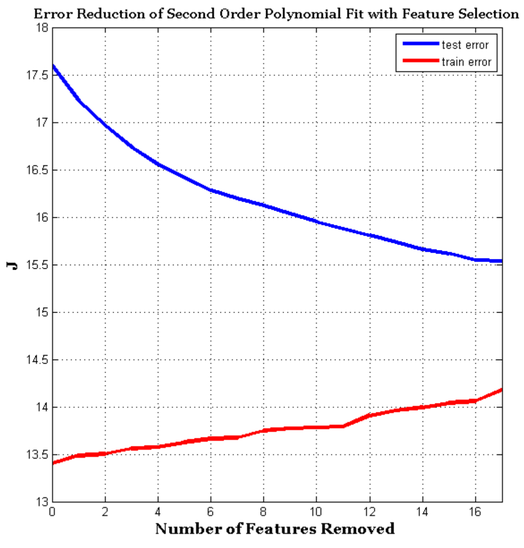
Features Used (x)	First Order Weights	Second Order Weights
GP	0.016	-0.114
W	1.893	-0.031
FGM	-2.873	0.138
3PM	3.758	0.192
3PA	-4.212	-0.047
AST	2.288	-0.006
STL	1.078	0.035
DD2	-2.830	0.178
TD3	-1.705	0.095
+/-	-1.141	0.004
Allstars	15.754	-2.412
Rebounder	6.977	-0.881
Double-double	11.528	-1.731
Triple-double	6.725	-0.821

Results









Features Removed	Order
num_rookies	1
PF	2
team_salaries	3
TOV	4
DREB	5
FGA	6
FG%	7
FT%	8
FTA	9
3P%	10
PTS	11
MIN	12
FTM	13
AGE	14
REB	15
OREB	16
BLK	17

Conclusion and Future Work

After running our statistics and data on three different models, linear regression, locally weighted linear regression and polynomial fits (with different orders), and validating using k-fold validation, we found that using a second order polynomial fit model provided us with the lowest generalized test error. This is therefore our chosen model. The average generalized test RMS error for this model is J = 22.7 compared to around 25 or more with other models. We then increased the size of our dataset, reducing the RMS error to J = 16.9. Feature selection then resulted in 17 features removed from our original 31 feature vector. With the remaining 14 features, our average RMS generalized error goes down to 15.54 with the RMS training error of 14.18. The variance of our model is therefore low. However, the RMS error in general is still relatively high. With more time, future work on the model to reduce the RMS error would include adding in more detailed and relevant features, such as coach statistics, different plays run by different team, etc.

References

Special Thanks to our mentor: Sam Corbett-Davies

List of data sources:
1. http://stats.nba.com/league/player/#!/
2. https://www.eskimo.com/~pbender/

Simulation packages used: Matlab