Supervised Machine Learning Methods for Predicting NBA Playoff Contention

Emanuel Azcona[1], Professor Ivan Selesnick[1], Professor Sundeep Rangan[1,2]

NYU TANDON SCHOOL OF ENGINEERING

1. NYU Tandon School of Engineering, Department of Electrical & Computer Engineering *2. NYU Wireless*

NYU WIRELESS

Abstract

Sports data analytics, data mining, and predictive analysis of sporting game outcomes are increasingly becoming more comprehensive as data collection methods continue to advance. Before data mining, sports organizations traditionally depended on human experience from coaches, scouting agents, managers, and players for predictive analysis. As the scope of data continually becomes more complex, research in machine automated predictive analysis draws a wide concern for research in predicting the outcomes of sporting events.

Our work focuses on using supervised machine learning algorithms to model and predict NBA-Playoff contention. The algorithms involved in our studies include: Simple Logistic Classification, Support Vector Machines, and Random Forest Classification. For convincing results, our data includes individual player and team statistics from 2001 through 2016 (2001-2002 through 2015-2016 seasons).

After processes of data collection, model training, and prediction, the Random Forest Classification and Simple Logistic Classification algorithms yielded the best results with respective best result accuracies of 87.5% each and the Radial Basis Support Vector Machine algorithm yielded the worst respective accuracy of 50%.

Introduction / Background

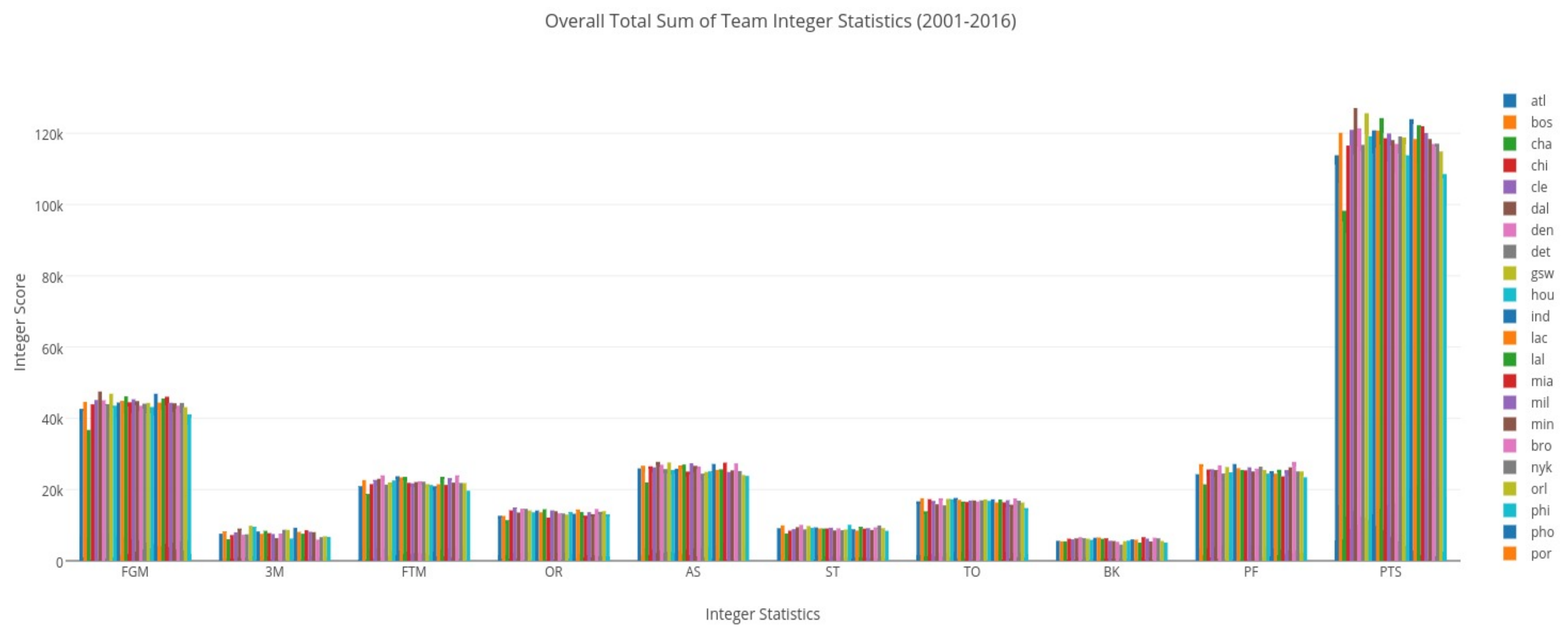
Although today's databases can carry data in sizes of up to hundreds of millions, there is still not a mature solidified technique to help us understand, analyze, and covert data into useful predictions of sporting event outcomes. In sports, relying on expert (coaches, players, etc.) experience is continues to grow less feasible as data continues to grow.

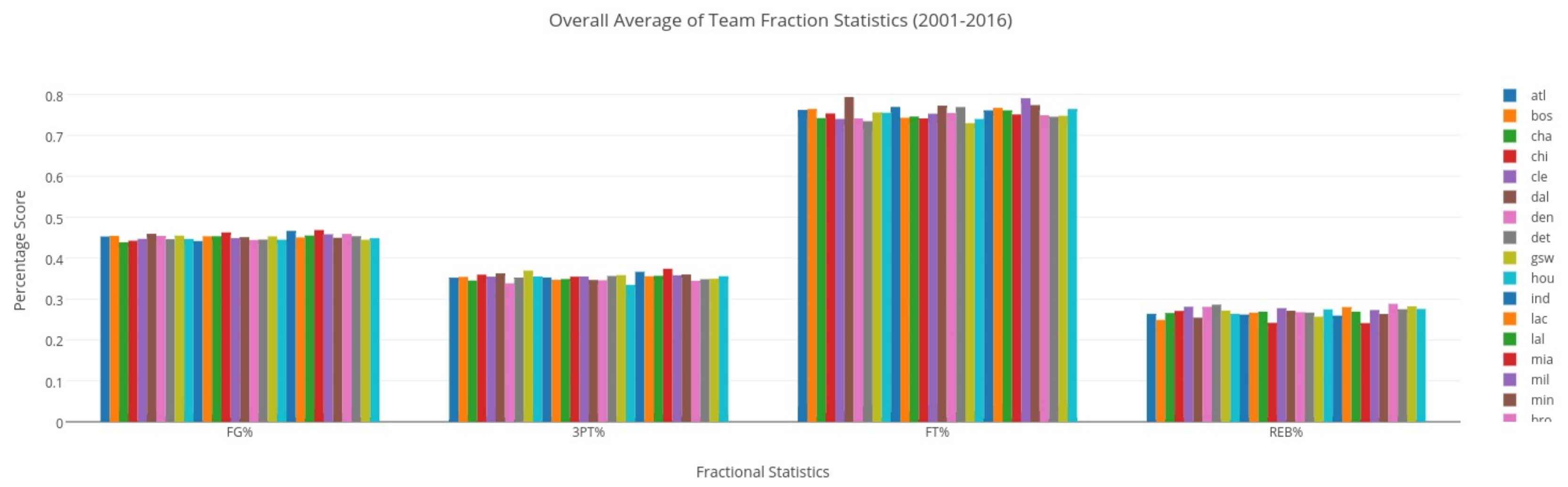
Expert knowledge can not discover all the value and potential of all collected data. Unlike digital lottery, which focuses on luck and chance, sporting event outcomes can be predicted through historical data that tells what features in a sport influence the outcome.

Our work uses pre-existing Python implementations of supervised classification algorithms and analyzes the accuracy in their predictions. The machine learning algorithms used in our study are:

1. Logistic Regression Classification
2. Support Vector Machines
   - Linear SVM Classification
   - Radial Basis SVM Classification
   - Polynomial SVM Classification
   - Sigmoid SVM Classification
3. Random Forest Classification

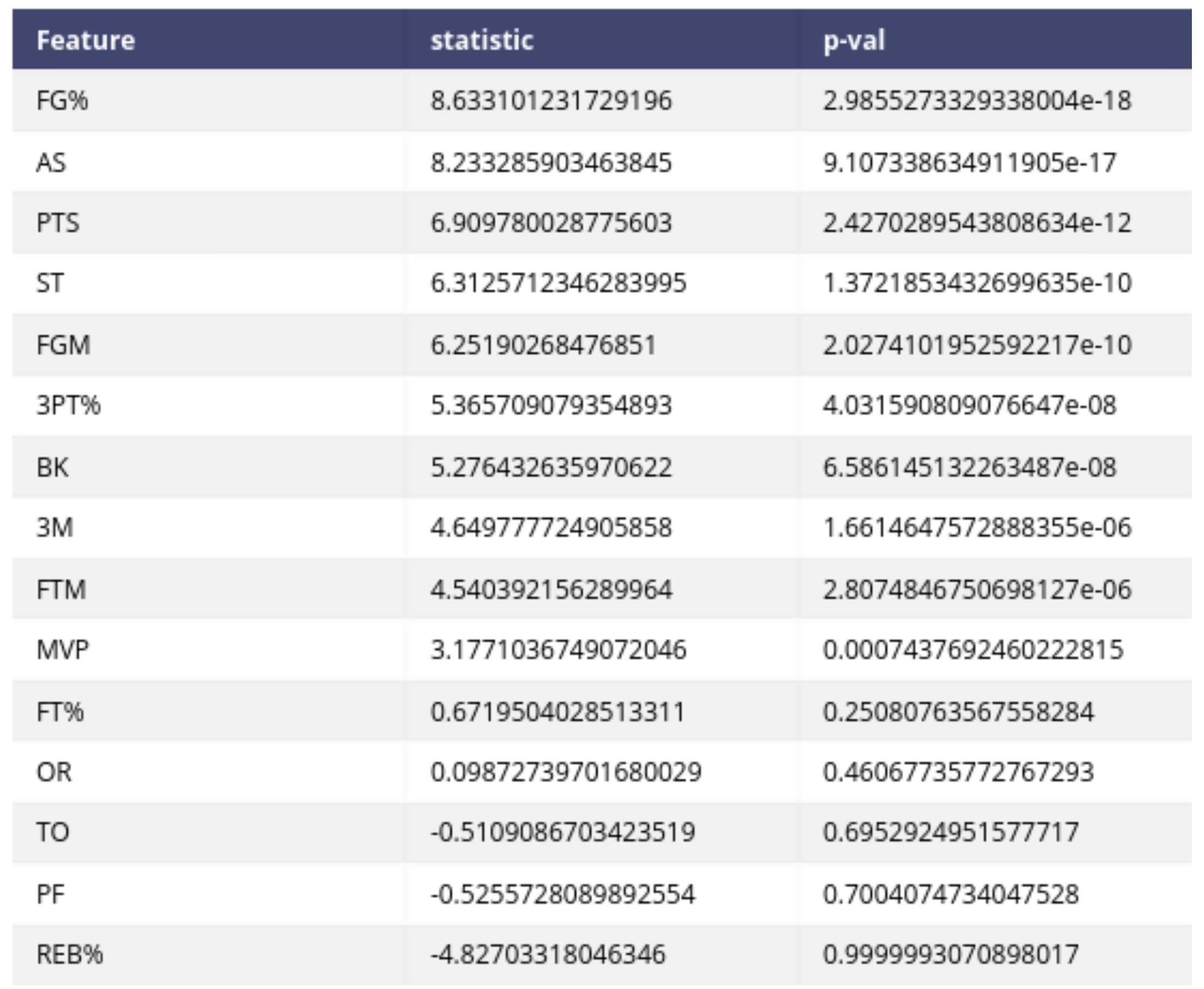
Overall Team Statistics


Overall Total Sum of Team Integer Statistics (2001-2016)


Overall Average of Team Fraction Statistics (2001-2016)

Statistical Significance of NBA Team Features

Prior to classification modeling, our analysis consisted of determining the statistical significance of predictors in our feature matrices of NBA team statistics. Based on the p-values of our test of statistical significance for our predictors, we remove features with p-values greater than or equal to 0.1 since they indicate a rejection of the null hypothesis.

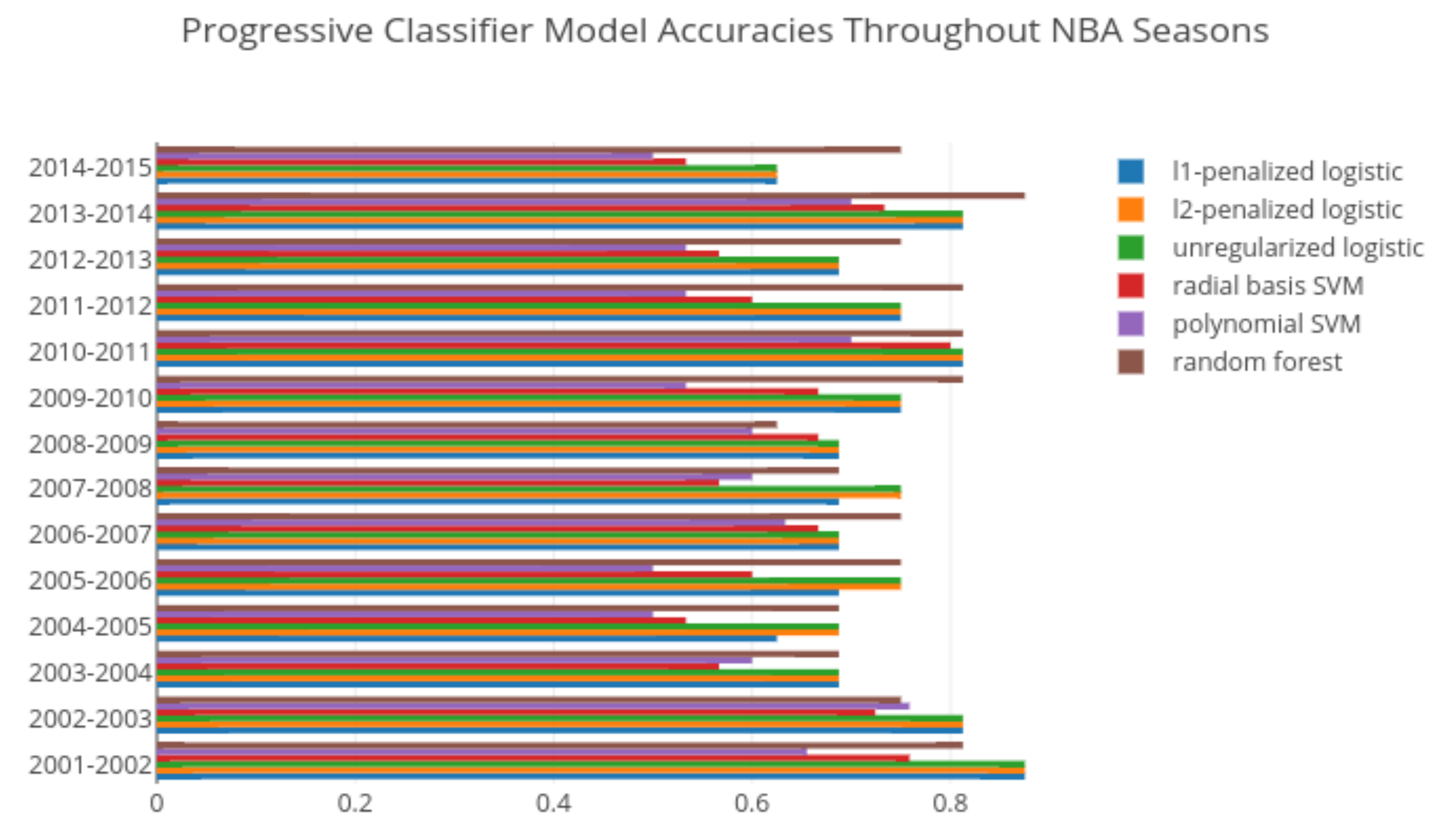


Feature	statistic	p-val
FG%	8.633101231729196	2.9855273329338004e-18
AS	8.233285903463845	9.107338634911905e-17
PTS	6.909780028775603	2.4270289543808634e-12
ST	6.3125712346283995	1.3721853432699635e-10
FGM	6.25190268476851	2.0274101952592217e-10
3PT%	5.365709079354893	4.031590809076647e-08
BK	5.276432635970622	6.586145132263487e-08
3M	4.649777724905858	1.6614647572888355e-06
FTM	4.540392156289964	2.8074846750698127e-06
MVP	3.1771036749072046	0.0007437692460222815
FT%	0.6719504028513311	0.25080763567558284
OR	0.09872739701680029	0.46067735772767293
TO	-0.5109086703423519	0.6952924951577717
PF	-0.5255728089892554	0.7004074734047528
REB%	-4.82703318046346	0.9999993070898017

Results/Conclusion

Our analysis shows that the following supervised learning algorithms predict NBA Playoff contention with the highest accuracy amongst their competitors:

1. *l1*-Penalized Logistic Sparse Classifier
2. *l2*-Penalized Logistic Sparse Classifier
3. Random Forest Classifier


Progressive Classifier Model Accuracies Throughout NBA Seasons

Further Work

Other classification algorithms, such as Naïve Bayes, can be investigated and compared to the current results or our analysis. After an observation, of the length time it took to train the SVM models, analyzing the time it took to train each of these models would be interesting to tabulate and compare. A lengthy, but goal within reach would be to implement each of these classification algorithms in Python and compare their performance to the Python sklearn implementations.

Works Cited

[1] Plotly Technologies Inc. "Collaborative data science" *Plotly Technologies Inc*., Montreal, QC (2015); https://plot.ly
[2] Cao, C. "Sports Data Mining Technology Used in Basketball Outcome Prediction" *Dublin Institute of Technology*, Montreal, Ireland (2012) http://arrow.dit.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis
[3] Predregosa, F., et. al "Scikit-learn: Machine Learning in Python" *Journal of Machine Learning Research*, 2825-2830 (2012) http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
[4] Breiman, L., Cutler, A. "Random Forests" *University of California, Berkeley* (2004) https://www.stat.berkeley.edu/~breiman/RandomForests/cc_papers.htm