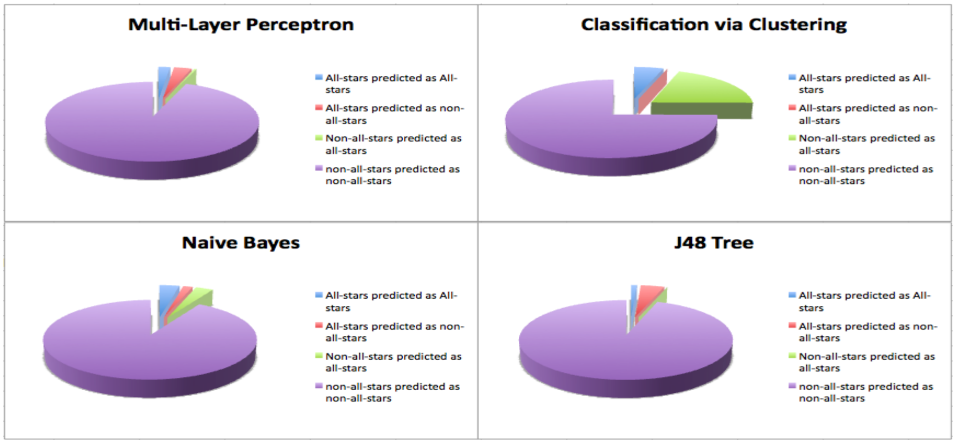
Methodology

All of our statistics, every player in the NBA from 1951-52 to 2011-12 season, were accessed from CSV files available on basketball-reference.com. The general algorithm, machine learning implementation aside, was to use the previous year only to predict the next years all –star team. We used this strategy, as opposed to using multiple years as a basis, because of a logical issue we encountered early in the project's conceptual development. This specific dilemma was that playing styles change in the NBA on a somewhat regular basis, at least every decade. Using, for example twenty years of statistics to predict the current years' team could have possibly increased prediction accuracy but also would have grossly ignored the relative importance of certain statistics. For example, in the 1980s decade, there is less variance in the "Big Three" statistics of players (Points per game, Rebounds per game, Assists per game) than in the 2000s.

Once the general idea was established, we set out to find the best machine learning algorithm to attack the issue. As mentioned briefly on the website, we initially used a decision tree, similar to Homework 2, to classify the players while training. This intuitively seemed like a solid idea to start, and the test results were encouraging, with estimated accuracy of over 90%. Yet, this metric was too good to be true. After the first roll out of Decision Tree on unlabeled csv files of players from the 2011-12 Season, the decision tree was extremely inaccurate and unreliable. We attributed this to the fact that all of our statistics were purely numerical in nature. As a response, we added a nominal team attribute into the data set. This gave a new, previously unaddressed dimension to the data set. It accounted for possible discrepancies in the numerical data such us a slightly better or popular player on the same team or nearby market that prevented all-star selection.

The low accuracy of the tree was still persistent after further experimentation. We were not able to identify any deeply seeded bugs in the decision tree classification. As a response to this obstacle, we decided to experiment with other machine learning methods. These methods included Naïve Bayes, clustering, and multilayer perceptron. We found that each of these models had their own respective weaknesses and strengths, but that after reviewing all of the options holistically, Naïve Bayes was the most reliable method. Below are prediction accuracies after training on 2009-10 and 2010-2011.



*Note that for the purpose of presentation, we have only included figures and data from recent years (Late 2000s and Early 2010's) on Web page. All-star predictions from earlier decades had similar results in terms of classification accuracy when fed into the model, and therefore would have been superfluous. (See next page for classifications from pre-2000 era)

*For the 2009-2010 and 2010-2011 combined training, if a player was an all-star in both years, their statistics were not combined. In other words our data set for 09-10 and 10-11, a player like Carmelo Anthony who was an all-star each year existed twice in the all-star grouping, one "separate" player for each season

Execution

Once we moved forward with the Naïve Bayes classification, we began to gather quantitative information on the accuracy of the model.

- With training on 2009-2010 and 2010-2011 and test labeling on the 2011-2012 season

17/25 Selected All-Stars were labeled as All-Stars	(68%)
15/453 Non All-Stars Selected as All-Stars	(3.3%)
% Players Classified Correctly/Not	(95.1%...4.9%)

Most Important Quality – Win Shares

Total Mean Win Shares ‘1’ labels	(10.07)
Total Mean Win Shares ‘0’ labels	(3.487)
Margin +/-	(+6.583)
Standard Deviation ‘1’ labels	(3.854)
Standard Deviation ‘0’ labels	(5.345)
Margin +/-	(-1.491)

- With training on 1991-1992 and 1992-1993 and test labeling on 1993-1994 season

26/26 Selected All-Stars were labeled as All-Stars	(100%)
57/377 Non All-Stars Selected as All-Stars	(15.12%)
% Players Classified Correctly/Not	(85.9%...14.1%)

Most Important Quality – Win Shares

Total Mean Win Shares ‘1’ labels	(9.938)
Total Mean Win Shares ‘0’ labels	(2.399)
Margin +/-	(+7.539)
Standard Deviation ‘1’ labels	(3.49)
Standard Deviation ‘0’ labels	(2.55)
Margin +/-	(0.94)

The results for both eras are fairly similar. We attribute this to the fact that we only trained for two years prior the predicted year as opposed to multiple years before. Also note that during the early 1990s, there were less players in the league, accounting for the lower ratio of correctly classified players. This also explains the smaller margins for mean and standard deviation of the statistic that the model identified as most important (Win Shares)

Link for Win Share algorithm: www.basketball-reference.com/about/ws.html

*Note that for the purpose of presentation, we have only included figures and data from recent years (Late 2000s and Early 2010’s) on Web page. All-star predictions from earlier decades had similar results in terms of classification accuracy when fed into the model, and therefore would have been superfluous. (See next page for classifications from pre-2000 era)

*For the 2009-2010 and 2010-2011 combined training, if a player was an all-star in both years, their statistics were not combined. In other words our data set for 09-10 and 10-11, a player like Carmelo Anthony who was an all-star each year existed twice in the all-star grouping, one “separate” player for each season

Conclusion

After completing our research we found that using the statistics of past all-star and non-all-star players as a way of a predicting the all-star players is not as dependable as intuitively expected. After analyzing our results we came to this conclusion for a couple of reasons:

1) We found that players were sometimes voted in for popularity reasons even when they were not statically deserving of the spot. One example is Kobe Bryant who was able to play in only six games this season, yet was voted as an all-star starter. This can severely hamper the model we built as it threshold for the quality of player that should be selected as an all-star.
2) Another issue is that both fans and coaches select the all-star teams. Fans may not make the most informed selections (part 1), and coaches may also be biased towards players they have coached or currently coach.
3) Our model does not take into account either position or the amount of players needed to fill an all-star team. There are supposed to be about 25 all-stars and only a certain number of players from each position. By being unable to set these parameters our predictions become less accurate. We could end up with 15 centers selected or only 10 players selected.
4) Because Win Shares is a statistic that is heavily factored into our model and based upon how many wins a team gets, a player on a bad team may not get as many win shares as a player on a good team. Additionally, a player on a bad team may get a ton of counting stats, but may not be efficient which can be detrimental to a team, but make him seem like a good player.

5) Finally, although statistic analysis may help determine the All-Star potential of a player and therefore the economic profitability / market power of a player, statistics will not completely predict the popularity, profitability, or marketability of a player, which is often highly correlated to NBA all-star selection. As, a future consideration, we would consider developing a model, possibly using Bayes Nets, that would be more comprehensive in terms of attributes, including more nominal traits like position. Additionally, some type of ranking algorithm could be implemented either as a supplement, or as an entire replacement to the Bayes Nets approach we have taken thus far.

*Note that for the purpose of presentation, we have only included figures and data from recent years (Late 2000s and Early 2010's) on Web page. All-star predictions from earlier decades had similar results in terms of classification accuracy when fed into the model, and therefore would have been superfluous. (See next page for classifications from pre-2000 era)
*For the 2009-2010 and 2010-2011 combined training, if a player was an all-star in both years, their statistics were not combined. In other words our data set for 09-10 and 10-11, a player like Carmelo Anthony who was an all-star each year existed twice in the all-star grouping, one “separate” player for each season