Matt Nicholson and Jayden Soni
Northwestern University EECS 349 Spring 2018

Using Probabilistic and Decision Tree Classifiers to Predict NFL Games Started and Played

Introduction and Data Collection:

The National Football League is a multi-billion dollar industry responsible for the largest annual televised American sports event: the Super Bowl. Each of the 32 teams that make up the league has a massive impact on its local community, especially those with rich histories of Hall of Fame players. Players inducted to the Pro Football Hall of Fall are seen as the best of the best at their position based on their career stats and overall impact on the game. However, this is not the only measure of a player's contribution to their franchise. Other such measures include the number of Pro Bowls a player will be invited to, games they will play in, and games they will start. Thus, we sought to predict the value players will bring to their teams (using these different measures) over the course of their careers based on year-by-year playing statistics.

To get our data, we scraped the website pro-football-reference.com. This website has game, season and career level statistics for every player that has ever played in the NFL and there are tens of thousands of these players. Because we used career level statistics as target values and to also limit the amount of data to a large, but manageable subset, we only considered players that had their full careers between 1985 and 2005 (since players are only Hall-of-Fame eligible after 5 years since the end of their careers, we wanted to allow ample time for that possibility). After a few iterations of attribute-target representations, we decided to use the playing statistics from a year of a players career (93 numeric attributes such as completion percentage, fumble recoveries, and average punting distance) to predict the number of games played and games started that player would have for the rest of their career (these targets were numeric attributes discretized of equal frequency into 5 and 3 groups respectively, numbers chosen so that the "worst" category would contain only all of the examples classified as 0 games played/started).

We hypothesized that a logistic regression would work well for our data set, since it could handle examples with many 0-value attributes (e.g., most quarterbacks have never kicked a field goal) and work well with numeric inputs and a limited number of classification options. We decided to compare these results against ZeroR as a benchmark, Naive Bayes (a different probabilistic classifier), Decision Trees (to analyze attribute selection and overfitting in our data set), and Random Forests (to see if we could improve upon the Decision Tree accuracy after a surprisingly good result). Working with 14,136 examples, we decided to use 3-fold cross validation in Weka to train and test our data, maintaining a balance between having large-enough training sets and a realistic idea of classifier accuracy. For every player, we made sure the years of their career all fell within the training set or test set for any given fold to maintain independence between the training and test sets.

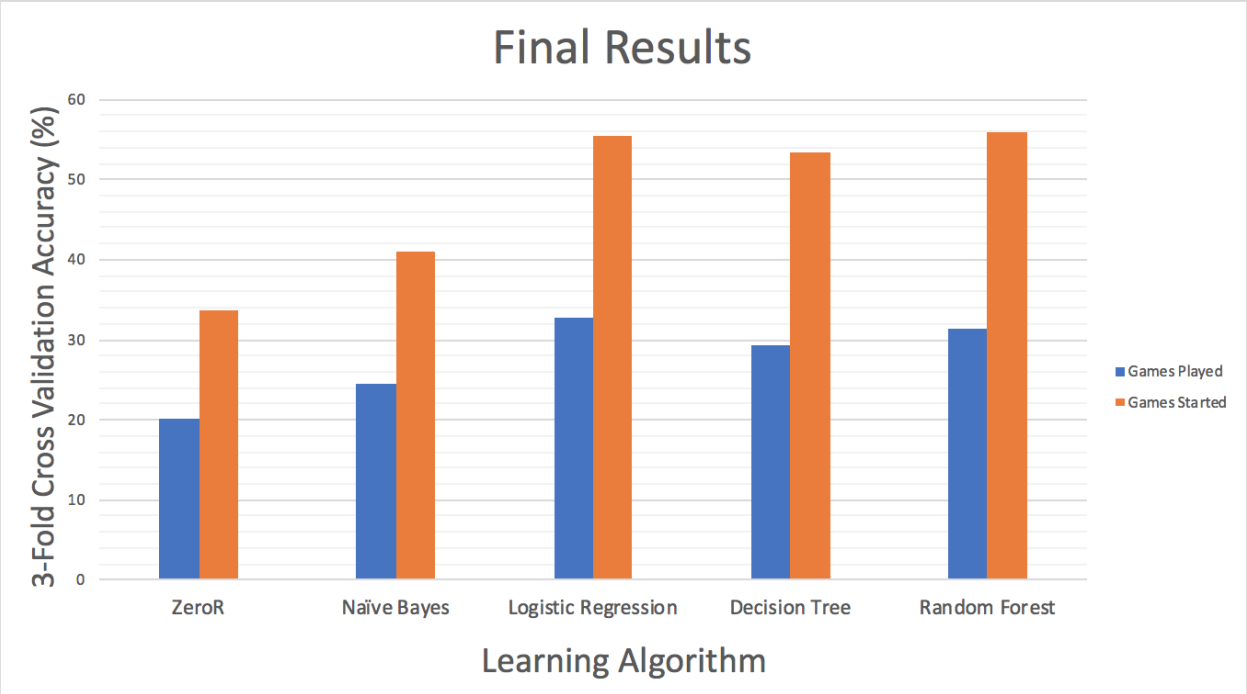
Investigations and Results:

We performed a number of different investigations to try to put together a classifier that gets at the notion of talent for a player, each with varied levels of success.

Investigation 1: Will a player make the Hall of Fame based on their statistics in their first four years in the league (the length of a typical rookie contract)? Will a player make any Pro Bowls in the remainder of their career, based on their statistics in their first four years in the league?

We tried a number of different models, all with results identical or worse than ZeroR. Upon performing the investigation, we found that instances of Hall of Fame careers were too few and far between to create a classifier that had a chance to beat the 99.7% benchmark set by ZeroR. As a brief follow up to this initially investigation, we also examined the efficacy of creating individual classifiers for each position in our data set. We started by partitioning our data by position, but found that the problem of very few Hall of Famers to be exacerbated, as there were still hundreds of players at each position and no more than five Hall of Famers at any position in our data set. The results were nearly identical in the case of using whether a player will make the Pro Bowl as a target, as the proportion of Pro Bowlers in the data set to still be roughly 99.3%.

Investigation 2: Given a player's statistics in a given year, how many games will that player play in the remainder of their career? Given a player's statistics in a given year, how many games will that player start in the remainder of their career?



With this, we sought target variables that would lead to more balance in our data set. We initially tried a binary output of whether a player would ever play again, but after trying a number of model, we did not consistently beat ZeroR using 3CV because roughly 80% of examples played at least one more game. We then tried to make classifiers to predict at a more granular

level for both game started and games played. In both cases, we binned the target variable into equal sized classes such that the size of the class was the number of observations that had target values of 0. This led to 5 classes for the games played case ([0], [1,18], [19,44], [45, 77] and [77, infinity)). Note that this roughly breaks down into no seasons left, 1-2 seasons left, 3-5 seasons left, more than 5 seasons left. In the games started case, this binned lead to 3 classes ([0], [1,28] and [29, infinity)), which roughly corresponds to no more starts left, up to 2 seasons left and more than 2 seasons left as a starter. An analysis of our decision tree classifier showed that age was one of the most important variables in terms of information gain, as well as attributes that could sort by position (e.g., passes_defensed > 0 to filter out defensive players, and punt_long > 0 to filter out punters).

Other Investigations: Pro-football-reference has its own generated statistic, “AV,” that is described as their attempt to attach an “approximate value” to a player’s year, summed up in a single integer. It is an interesting concept, one which we decided to test using linear regression. We also figured age would have a large impact, since it influences the length of time left in a player’s career, and ran regressions with both variables, as well as all attributes.

Career Games Played	
Attribute(s)	R^2
AV	0.040
Age	0.066
Both	0.149
All	0.222

Career Games Started	
Attribute(s)	R^2
AV	0.138
Age	0.030
Both	0.224
All	0.289

In general, the results of these regression are uninspiring, showing a only a small amount in the variation of the target variables can be explained with linear models only including age and/or AV. It was interesting to note, however, that age played a larger role than AV in determining career games played, while it was the other way around for games started.

Discussion:

Jayden and Matt worked collaboratively at each step of the project. Much of the time was spent preparing the data set. First, we wrote a scraper in python using the BeautifulSoup library. The scraper was then run on an Amazon EC2 instance, since each run would take roughly three hours to complete. We then did aggregated each player’s career accomplishments from the scraped data to gather the targets for Investigations 2-5. The resulting csv files were used in Weka to analyze different learning algorithms.

Prior to our attempts to build classifiers, we explored the data. We found that vast majority of players have very short careers (around 4 years on average) and that there are

different tiers of players. Upon looking at the distribution of the more balanced targets we considered (games played and games started in remaining years), we see that the distribution is bimodal, with the majority of the point on either zero games or 16 games. This leads us to believe that there could be an issue of survivorship bias in our data set. For example, consider the case of a veteran backup quarterback QB2 and a rookie starting quarterback QB1. It is strong possibility that QB2 will not play many snaps in a this year and thus, will not have impressive statistics for that season. However, that does not necessarily imply that QB2 will not play or start many more games in their career, as players frequently get injured. Conversely, it is also likely that QB1 gets hurt in the preseason of their first year, but goes on to have a long career. These points could be obfuscated in our data and account for a number of the misclassifications we see across models. Even so, our classifiers trained on games played and games started in remaining years consistently beat random chance and rarely mistake the best targets for the worst targets and vice versa.

Directions for Future Work:

Given that our results for the games played and games started cases beat ZeroR significantly, we see these as the most promising target variables to include in future work. One possible direction would be to more critically select relevant features from our data set to use in the classification task. Additionally, we believe that there could be strong interyear dependencies that determine whether a player will continue to have a long career as a starter in the NFL. As such, we would repeat the initial analysis we did by aggregating each player's statistics from a subset of their career, rather than treating separate years as separate observations. This would reduce the number of examples we had to train on, but increase the available information in each one. We could also increase the size of our data set by expanding our original range of 1985-2005 to the present day, now that we are no longer concerned with the Hall of Fame (as long as a player has retired, we know their career games started and player).