Introduction

The dataset classified was the NFL combine dataset from nflsavant.com. The classifier used was the round a prospective NFL player was drafted, from 0 - 7. The classifier was then binned such that rounds 1-7 became a '1', and '0' stayed at 0, meaning that a player with a '1' was selected in the NFL draft, and a player with '0' was not selected. This classifier was attempting to classify if a player would be drafted in the NFL draft based off combine performance.

A total of three classification methods were used on this dataset. These were K nearest neighbor, decision trees, and ensembles. The ensemble method with random forests had the highest predictive accuracy at 68% +/- 2%.

Data Analysis

The attributes in this dataset are the various physical evaluations/tests that athletes complete in the NFL combine. All of the attributes that were used in the classifiers are the continuous values, since all of the important tests were measured continuously. The categorical data is mostly identity information for the athlete, and not so much about their combine performance. These values include name, college attended, etc. The attributes before any pre-processing contained in Table 1.

Continuous	Continuous	Categorical
Height feet	TenYD	Year
Height inches	TwentySS (shuttle)	Full Name
Weight	Three cone (drill)	Position
Arms	Vertical	College
Hands	Broad	First name
FortyYD	Bench	Last name
TwentyYD	Pick	Round **(label)**
HeightInchesFeet	Pick Total	Pick Round
ID **(key)**		Wonderlic

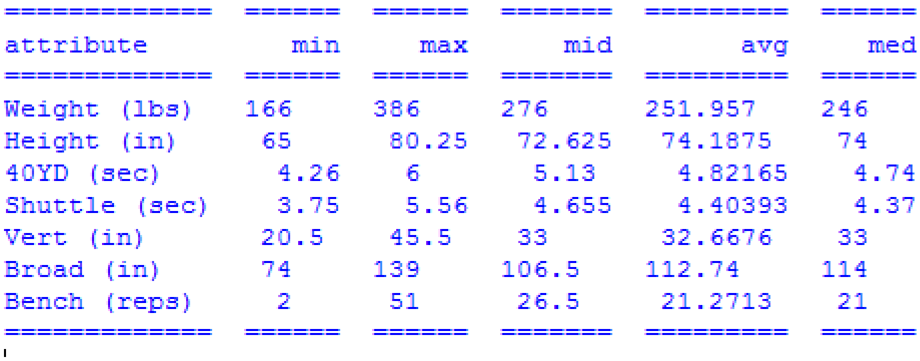
Table 1 - Attributes before pre-processing

Pre-Processing

The highlighted values in Table 1 are the attributes removed after preprocessing. Yellow values indicate that It was removed for containing too many NULL values, >= 80%. The blue values represent superfluous data that can be extracted elsewhere. For example, "Height feet" and "Height inches" are both implicitly contained in 'HeightInchesFeet", allowing me to trim these two attributes. The class label used was "Round", indicating the round (0-7) of the NFL draft the combine participant was taken in. This left the dataset with a total of 18 attributes from the original 26.

After the attributes were removed, each instance with a NULL value had to be resolved. The first method used was to average out the attribute against other players of the same position. This means that a Wide Receiver with a NULL value in the Forty Yard Dash would be assigned the average value among other Wide Receiver's 40-yard dashes. However, this method created a predictive accuracy lower than the second method. The second method will simply remove each instance with a NULL value. This method had a greater predictive accuracy, yielding a total of 2509 instances from the original 4600. Also, positions with either too little information or too little instances were removed. These positions included kickers (K), punters (P), nose tackles (NT), and long snappers (LS). Centers were changed to offensive centers (OC).

A total of 7 continuous attributes were chosen to be used in classification. The eight attributes and their summary statistics after pre-processing are displayed in Figure 1.



```
=============  ======  ======  =======  =========  ======
attribute         min     max      mid        avg     med
=============  ======  ======  =======  =========  ======
Weight (lbs)   166     386     276      251.957    246
Height (in)     65      80.25   72.625   74.1875    74
40YD (sec)       4.26    6       5.13     4.82165    4.74
Shuttle (sec)    3.75    5.56    4.655    4.40393    4.37
Vert (in)       20.5    45.5    33       32.6676    33
Broad (in)      74     139     106.5    112.74     114
Bench (reps)     2      51      26.5     21.2713    21
=============  ======  ======  =======  =========  ======
```

Figure 1 - Summary Statistics for Attributes

A trend is seen in most of these attributes. As the round increases from 1 to 7, there is a noticeable trend in the mean performance for the athletes. The 1st and 2nd rounds contain the highest performance athletes, as one would expect. As the rounds decrease, so does the average performance. When the round is at 0 (undrafted), it is noted that this has the highest range of averages. The box and whisker plots for the 40-yard dash and the vertical are seen in Figure 2 and Figure 3 respectively.

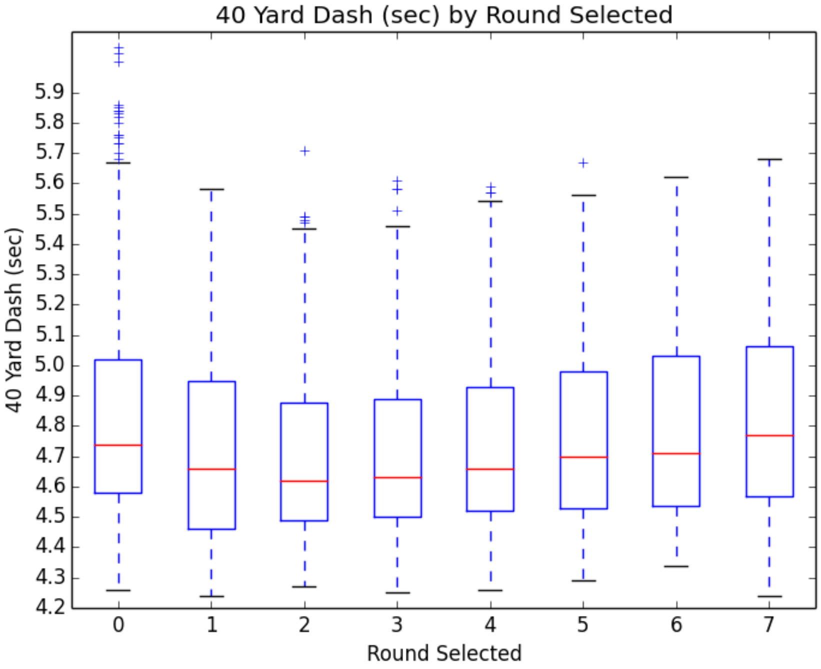


Figure 2 - Round drafted w/ respect to 40 yard time

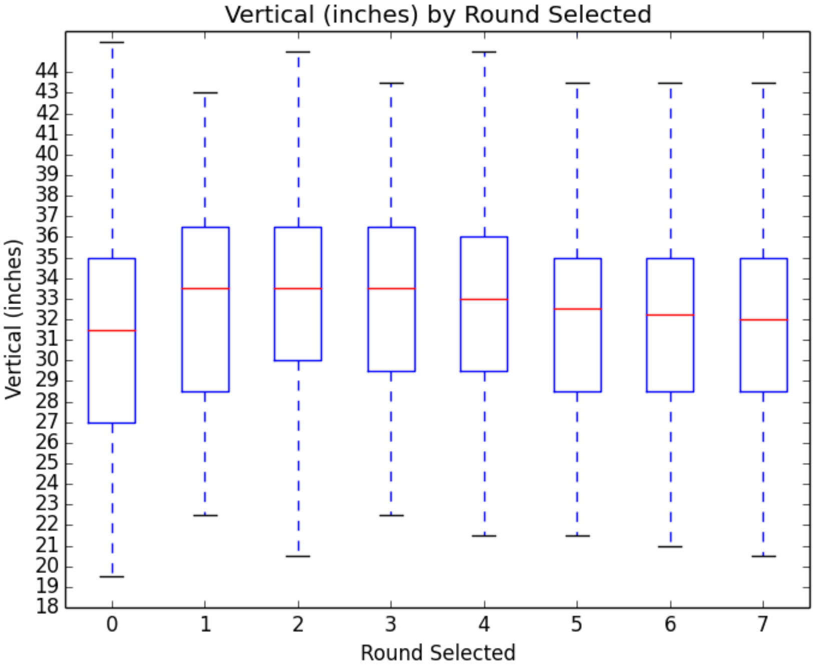


Figure 3 - Round drafted w/ respect to vertical height

The class label distribution over the 2509 attributes is reasonable for classification. About 37% of the athletes were undrafted, while drafted players comprised the other 63%. This is seen in Figure 4. Rounds 1 - 7 have been binned together for classification purposes.

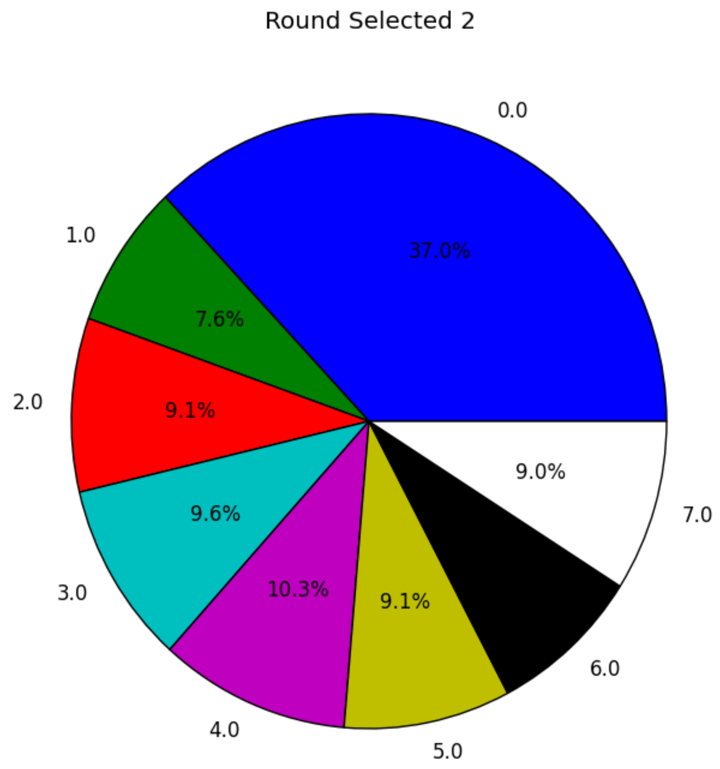
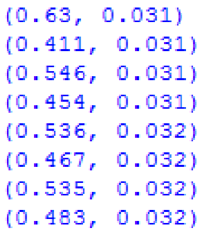


Figure 4 - Round picked distribution

Classification Results

KNN

The first classification method used was K-nearest-neighbor, which also happened to be the worst classifier. Overall, this method yielded a max of 63%, which is non-coincidentally the same percentage as guessing "yes" the whole time. Increasing the value of K will continually increase the predictive accuracy, until it stabilizes at 63% where it guesses “yes” the entire time. So theoretically, the highest value of K that gave me the best predictive accuracy was K = 2509 (length of table), yielding 63%. Figure 5 displays the first 8 K values in KNN classification, which are all very different. After around 20, the predictive accuracy would increase linearly.



```
(0.63, 0.031)
(0.411, 0.031)
(0.546, 0.031)
(0.454, 0.031)
(0.536, 0.032)
(0.467, 0.032)
(0.535, 0.032)
(0.483, 0.032)
```

Figure 5 - KNN Accuracy with K = 1 - 8

A few different approaches were used in the normalization for KNN. The first was using the regular normalization technique over each column. The second would take into account the position the athlete played, since a high quality wide receiver will surely have a faster time in speed drills than a high quality defensive tackle, for example. This technique would normalize by position played. On average, this technique yielded about 1% less predictive accuracy each trial than the regular normalization method, which was unexpected.

Decision Tree

Next, a decision tree was implemented yielding a high of 65% predictive accuracy. This high was achieved by discretizing using split point discretization based on entropy. The attributes that gave the highest entropy were weight, vertical, and bench. Another approach was used that involved binning each attribute into 5 equally spaced bins, discretizing attributes from 0 to 4. The accuracy for this approach stayed at around 63% +/- 2%.

Random Forests

The approach with the highest predictive accuracy was a random forest ensemble. The highest predictive accuracy that came from this technique was about 68% +/- 2%, with N = 40, M = 5 and F = 3. Similar to the decision tree, the trees in the forest were made with both split point discretization and binning. The split point entropy discretization yielded the best predictive accuracy

Each classifier was tested using stratified 10-fold validation and random-subsampling with 10 holdout partitions.

Conclusion

The classifiers performed poorly, at best only doing 5% better than just guessing yes (63%) over the dataset. In Figure 2 and Figure 3, there is not a large difference between players who are drafted in round 0, and those drafted through rounds 4-7. This is where I believe the confusion lies in the classifier. The distinction between players who were drafted in the last few rounds is very small in their mean physical performance. The classifier was binned in different ways, but the 0 - 1 discretization provided the best results. At first, no binning was done at all and the attributes were left untouched, which yielded the worst results.

The results may also yield an improved predictive accuracy if the dataset was split by position played for each classification method. Each position has different physical requirements, one cannot directly compare values from one position with another in some cases. An offensive lineman can run an extremely slow 40-yard dash (5.5+ seconds) and still get drafted, as they are not primarily graded on speed compared to a wide receiver.

To improve the performance, I believe the class label needs to be binned differently. Rounds 6 and 7 in the NFL draft are very similar to not being drafted at all (in terms of mean performance), so this can be labeled as drafted in round 0 during pre-processing. This should make the distinction between a '1' and '0' greater, yielding a better predictive accuracy.