Sydney Scaccia

4/30/2021

DSC 609: Machine Learning

**Final Project Paper**

**Introduction**

The purpose of this analysis was to determine if machine learning algorithms could predict and classify potential National Football League (NFL) prospect’s position based on the scores and outcomes of each of the NFL combine events. The NFL combine is a series of standardized athletic tests designed to assess the current college football player’s skills. This is an invitational event to showcase their athletic ability to NFL scouts with the hopes of being drafted for the upcoming season. The motivation behind this project was to determine if potential prospects could have their position predicted correctly based on these events. Ultimately, the goal was to begin by classify the existing players by position based on the combine event results and put the different types of algorithms to the test to determine if these combine events truly showcase the athlete’s potential. The research question at hand is: *Can the combine results accurately predict potential professional athlete’s positions correctly?*

**Literature Review**

Prior to any analysis, a formal literature review was completed to determine if any prior studies answered the research question. The three articles included within the literature review both refute and support the idea that the combine can accurately depict a player’s position and whether or not it is a true evaluation of their abilities. Combining these articles will set a solid base for this research project because it will help eliminate any hidden biases.

The first article titled, “The NFL Combine: Does It Predict Performance in the National Football League?”, refutes the idea that the combine can accurately depict a player’s position by the event results. The authors, Frank Kuzmits and Arthur Adams, state that since “the combine enjoys a significant degree of media hype” it “clearly lacks any meaningful degree of validity” (Kuzmits & Adams, 2008, p. 1725). This article mentions that since the combine is such a huge event within the world of football and athletes prepare around the clock for this day, anything can happen during the actual events. These scores could be significantly better than the athlete has ever preformed throughout college, or the scores could severely downplay the athlete’s potential giving the combine itself no real value.

The second article by Marcus Woo titled “How Much Does the Combine Reveal About Future NFL Players?” expresses that the combine “gauge[s] the general athleticism” of the players so it will correctly predict the position, but he expressed some concerns (Woo, 2019). This is ultimately because the drills isolate individual skills of the athletes. It is quite reasonable to believe that someone who is a stronger athlete would be better at football, but the drills in the combine do now show how these athletes will react in certain situations on the field. Woo suggests that instead of using the combine event results to predict the player’s position or if they will be drafted to the professional level, analysts should use full season statistics and real game situations to evaluate the athlete in all aspects of the game, physically and mentally.

The last article, “Does the NFL Combine Have Any Predictive Value to NFL Success?”, by Tyler Schlater once again refutes the idea that the combine is a solid predictor of the athlete’s potential. Throughout the article, Schlater give examples of now professional athletes who have performed strong at the combine but were drafted very late, and of athletes who performed poorly at the combine but are some of the most successful players in the game today. Tom Brady is a prime example supporting this idea because he performed terribly at the combine. This could be because he was having an off day, the events do not measure the true athleticism of a quarterback, or because of all the pressure put on the athletes. On the other hand, Bruce Campbell had an outstanding performance at the combine during his draft year but was still drafted very late because when analysts compared his season statistics to the combine outcome, it was a fluke for him to perform that well.

All these articles allude to the idea that the combine is nothing but a glorified event for the players to meet the coaches. According to the multitude of studies on this football event, it does not predict the player’s true potential or depict how well of a player they are. It has turned into a few days of competition that are now mainly used for gathering the player’s medical information and the interaction between the athletes and coaches and between the athletes and major NFL sponsors. This literature review will lead into a very interesting study to determine whether these machine learning algorithms can produce a high accuracy to refute the previous results on these research studies.

**Data Sources**

Looking at the data itself, the data set was found on reddit and was available to the public to be downloaded. The data that is contained within the table is very accurate and has information that was scraped from the NFL combine reference website. Within the table there is a record of every player who has entered the combine beginning with the year 2000 and ending with the year 2019. All these players together create a data set containing over five thousand football players. The table has the player’s position, what college they attended, the player’s weight, their height, the player’s forty-yard dash time, their vertical jump height, the number of bench press repetitions, the player’s broad jump distance, their three-cone drill time and their shuttle run time. It also includes the player’s name and what year they entered the combine along with what team they were drafted to in the NFL or if they were drafted at all.

The variables selected for this project will aid in the determination of whether the machine learning algorithm can classify these players in their correct positions based upon their combine event results. The dependent variable here is each player’s position. Within this table there are eighteen different positions that have entered the combine throughout the past nineteen years. The independent variables are their height, their weight, their forty-yard dash time, the number of repetitions the player completed on the bench press, their broad jump distance, their three-cone drill time, and their shuttle run time. By including all the main events, the machine will have the best chance at predicting the player’s position correctly.

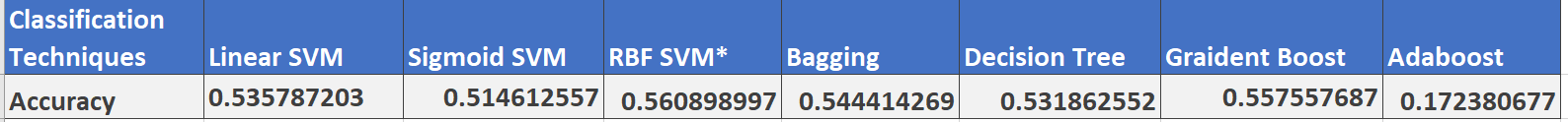
**Methodology**

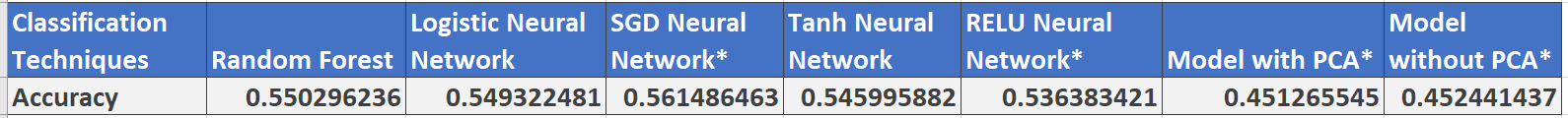
Before the analysis can begin, the data needs to be combed through and preprocessed to ensure the best machine learning algorithm result. The data set itself contains a considerable amount of “NA” values meaning that these players did not participate in that specific combine event. These values are an issue because if they are simply changed to a zero, the analysis will be skewed. This is because some of these players could receive a legitimate zero for some of the combine events. For example, the player could have not been able to bench press the amount of weight on the bar for repetitions resulting in the score of a zero, or they could have faulted on the event. Because of this, these values needed to be removed from the data set before the analysis. The second preprocessing task is scaling the data. This is another necessary step before beginning the analysis because right now all the variables are on their own scale. Without scaling the data, it will be difficult to compare the influence of each variable on the player’s position.

After preprocessing the data, the analysis can begin. Originally, only a decision tree model and a random forest algorithm were going to be implemented, however in the final analysis multiple support vector machines, neural networks, ensemble algorithms, and a principal component analysis were used to determine which algorithm would produce the highest accuracy and be declared as the best classifier. By including all these classifiers and computing their accuracy statistics, it will not only produce the best result possible and the most accurate classifier, but it can then be used again in the future to determine whether this algorithm can also predict if the player will be drafted.

**Results**

Table 1 displays the results from each of the machine learning algorithms accuracies once ran with the specified variables above. Almost all the classification models produced an accuracy around fifty percent, except for one whose accuracy was extremely low.

**Table 1: Results from the Classification Techniques**

**\*** RBF SVM **=**Radial Based Function Support Vector Machine

**\*** SGD **=** Stochastic Gradient Descent, RELU = Rectified Linear Activation, PCA = Principal Component Analysis

Instead of selecting just one of the support vector machines, all three types, linear, sigmoid, and the radial based function, were ran. The same goes for the neural networks. The two best models produced an accuracy of about fifty six percent. These models were the radial based function support vector machine and the stochastic gradient descent neural network. The best model is the stochastic gradient descent neural network. It has the higher accuracy of the two, but they only differ by about five hundred thousandths of a decimal point.

The next best classifiers were the gradient boost algorithm and the random forest algorithm which produced an accuracy of about fifty five percent unrounded. The remaining algorithms all had an accuracy of less than fifty five percent. The bagging classifier, the logistic neural network, and the tanh neural network all had an accuracy of about fifty four percent. The linear support vector machine produced an accuracy of about fifty three percent, the same accuracy as the decision tree algorithm finally, the RELU neural network and the sigmoid support vector machine produced an accuracy of about fifty one percent.

The algorithms with the lowest accuracies are classified as the worst models. In the scope of this study, these are the classification algorithms that produced an accuracy of under fifty percent. The two algorithms that had an accuracy of about forty five percent are the partial component algorithms. One of the models included the principal component and the one is without the principal component. The one with the partial component included produced a lower accuracy than the one without the partial component. These models once again differed by one hundred thousandths of a decimal point. The worst model over all that did not even produce an accuracy of twenty percent which was the adaboost algorithm. This model produced an accuracy of about seventeen percent which is very low. Figure 1 displays each of the outputs from the machine learning algorithms compared to a trend line represented in red. All the algorithms whose bar passes the red trend line have an accuracy of over fifty percent. It also displays the comparison of the models against one another from the analysis.

**Figure 1:** Model Accuracies Compared to a Trend Line

**Discussion**

Although this analysis has a solid foundation and multiple models were ran to support the findings, there are still some limitations. One issue within this analysis could be that the events within the combine do not accurately represent some of the positions the athletes play. Take a punter for example. Oftentimes the punter is not sprinting down the field, so his forty-yard dash time could alter the analysis if it is extremely slow. Another possible limitation is the number of positions being analyzed at once. Within this analysis there are eighteen positions being evaluated at one time using the same variables. To improve this analysis within the future or for future subsequent projects, isolating the positions and selecting only a few variables that pertain to that position would probably improve the analysis. In doing so, this will allow the machine to learn the algorithm more effectively and efficiently. Currently within the program, some of the algorithms can compute the accuracy within a few seconds while one algorithm took up to a half hour to compute. This is due to the volume of data included within the analysis, but by evaluating each position individually, it will cut down on the amount of data included and will allow the algorithm to produce a better result. Another basic alteration that could be made to the code itself is the changing of the parameters. Right now, being that the program took a very long time to run, changing the parameters time after time would not be the most time efficient and effective way to change the accuracies, cutting down on the data would be the best option.

In the future, this algorithm could be used to determine if a player with similar results from the combine will get drafted to the NFL and when based upon their position. Based off of past combine results and past players, the same algorithm that is used to predict the player’s position to determine if the athlete’s combine results are strong enough to be drafted.

**Conclusion**

The purpose of this analysis was to determine if a machine learning algorithm could correctly classify potential professional football player’s positions by their combine event results. Throughout this analysis, multiple machine learning classification algorithms were applied to determine which would produce the highest accuracy resulting in the best classification model. The model that produced the highest accuracy was the stochastic gradient descent neural network with the radial based function support vector machine as a close second. Both models produced an accuracy of about fifty six percent unrounded creating a decent model. Changing the parameters within the analysis and limiting the amount of data that is included in subsequent projects would improve the models and create a more accurate prediction. As it stands now, the machine learning algorithm can only predict a player’s position correct fifty six percent of the time.

**Work Cited/ References**

Kuzmits, F. E., & Adams, A. J. (2008). The NFL Combine: Does It Predict Performance in the National Football League? *Journal of Strength and Conditioning Research*, *22*(6), 1721–1727. https://doi.org/10.1519/jsc.0b013e318185f09d

Schalter, T. (2017, October 3). *Does the NFL Combine Have Any Predictive Value to NFL Success?* Bleacher Report. https://bleacherreport.com/articles/1535606-does-the-nfl-combine-have-any-predictive-value-to-nfl-.

Woo, M. (2019, March). *How Much Does the Combine Reveal About Future NFL Players?* Inside Science. https://www.insidescience.org/news/how-much-does-combine-reveal-about-future-nfl-players.