Beyond the Box Score: Using Psychological Metrics to Forecast NBA Success

Basketball Track
Paper ID: 20251423
Sean Farrell[1], Ethan Laity[2], Dave Laughlin[3], Dean Oliver[4] &
James W. Pennebaker[5]

Abstract

Risk assessment of potential recruits is of prime importance to all NBA franchises. Although most scouting focuses on physical performance statistics, there is general agreement that psychological factors also play an important role in determining success. Accurate psychological assessment of potential recruits can be difficult due to limited access to athletes, the time required to complete the assessments, and the self-reporting nature of traditional psychology questionnaires. In this paper we explore applications of language psychology metrics using machine learning and survival analysis techniques to predict success in the NBA. We found that we could predict which athletes would make it onto an NBA roster with an accuracy of 63% without any physical attributes included in the model. In contrast, a model built just using NCAA playing statistics achieved an accuracy of 78%, and combining physical statistics with the psychological features boosted the performance to 83%. Adding in physical attributes such as age, height and weight, along with the NCAA conference the athlete played in increased the accuracy further to 87%.

Our analyses could also predict the career survival probability (in terms of the number of NBA games played) and starter probability (as in number of NBA games in the starting lineup) using survival analyses with high significance for players who made it onto an NBA team roster. Our models achieved concordance indices of 0.61-0.62 for psychology only models, and as high as 0.84-0.86 when adding NCAA playing statistics, physical attributes, the NCAA conference, and where in the draft they were selected. All of these models perform significantly better than chance. Augmenting physical performance statistics with psychological features can provide added value to recruitment teams from a risk assessment perspective. While this work focused specifically on basketball recruitment, our results have far-reaching implications for identifying emerging talent across all elite sports.

[1] Corresponding author: drsafarrell@gmail.com
[2] The University of Newcastle, Australia
[3] Courtex Performance, https://www.courtexperformance.com/
[4] ESPN
[5] University of Texas at Austin, USA

1. Introduction

Physical performance attributes of aspiring athletes to professional basketball have typically been the prime selection factor in recruitment [1,2,3]. College basketball performance in combination with the metrics produced by the annual NBA draft combine are the most readily available sources of data on physical attributes and performance. Prior studies have shown that college performance statistics can be significant predictors of future NBA performance [2]. However, as athletes can declare for the NBA draft at 19 years of age, many players in each draft cohort may only have limited college playing experience. In addition, assessment of non-NCAA prospects such as international recruits can be challenging, as comparing NCAA statistics against statistics from other leagues of varying quality is problematic. Other studies have found limited evidence that NBA draft combine assessments provide any predictive power with respect to future NBA performance [1, 3]. Identifying other metrics in addition to college performance and combine assessments that show promise in predicting NBA success to augment recruitment decision making could thus provide a valuable competitive edge.

While it is widely believed that the mindset and psychology of athletes are important factors in determining future success [4], these dimensions have historically been difficult to assess [5,6]. Evidence has been presented that personality [7], psychological skill [8], and cognition [5] all have a positive correlation with athlete performance. The Conscientiousness trait in the Five-Factor model of personality was found to be significantly with stronger player game statistics in NCAA Division 1 soccer [7]. The Athletic Coping Skills Inventory has been found to be as capable at predicting MLB batting average variance as physical traits, was better at explaining the variance in pitcher's earned run averages than physical skills and predicted athletes' survival for the two and three years after signing with a club [8]. A significant difference was also observed in Athletic Intelligence Quotient scores between NBA and non-NBA players [5], suggesting that cognitive ability is correlated with NBA performance outcomes.

Research in this area has been hampered by a lack of reliable and objective psychological data. Access to athletes (particularly in the lead-up to the NBA draft) is typically limited, making meaningful psychological assessments challenging. In addition, most traditional assessments rely upon self-reporting, which is known to be problematic as eager recruits can mold their answers to what they think assessors want to hear [9]. In contrast, language psychology techniques allow us to analyze the everyday language of athletes from interviews or writing samples and assess their cognitive, social, and emotional states [6] in a less biased way without providing additional load on the athletes.

Profiles constructed using language are far less open to conscious manipulation and therefore provide a more accurate picture of a candidate, while also bypassing the need to have athletes submit for formal assessments. A person's language usage provides an enormous amount of information about who they are. In particular, a class of words called function words has been found to provide meaningful insights into people's social and psychological states. Function words include pronouns (e.g., he, she, they, I, we), articles (a, the), prepositions (to, of, for), and a small number of other common but almost invisible words that we use at high rates every day. Function words account for less than 1% of our entire vocabulary but represent over 50% or our word usage [10]. Across hundreds of studies, we now know that how you say something (i.e., the linguistic style as measured by function words) is far more revealing of your personality or thinking styles than what you actually say (i.e., content words). Further, it is exceptionally more difficult to consciously

manipulate your function words while being interviewed than to shape the content of your conversation [11].

Linguistic Inquiry and Word Count (LIWC) [12] was developed to extract psychological profiles of individuals from their language. LIWC is a dictionary-based text analysis tool that produces scores that represent the frequency of use of different categories of words as a fraction of the number of words that are present in a sample of language. In this paper we explore the application of LIWC in combination with machine learning algorithms to predict which aspiring college basketballers will be signed by NBA teams. Using survival analysis techniques, we seek to predict NBA career duration and the number of NBA games a player will start in based on the athlete's linguistic psychological profile. For both sets of models, we assess the impact of augmenting these models with NCAA playing performance statistics plus other features such as age, height, weight, and NCAA conference.

2. Methods

2.1. Data

ASAP Sports[6] have been producing high quality transcripts of player interviews and press conferences for a wide range of professional and amateur sports since 1989. At the time of writing, their archive of basketball interviews spans 1992 to 2024 and includes both NCAA and NBA post-game interviews with coaches and athletes. For this research, we scraped a total of 25,805 basketball transcripts from the ASAP Sports archive. Filtering on the keywords "NCAA" in the tournament description field we selected a sample of college basketball player interviews. We then cross-matched this sample against NBA draft records scraped from the official NBA website[7] and NCAA athlete profiles scraped from the Real GM website[8] to identify athletes that were signed by an NBA team (either drafted or offered a contract after the draft) vs those that were not based on name and date matches.

The match-ups between ASAP Sports interviews and the Real GM NCAA player profiles were manually inspected to confirm the accuracy of the matches, and where necessary match-ups were adjusted or discarded. Profile information plus NCAA and NBA playing statistics were scraped from Real GM for each athlete to identify details such as which draft year they nominated for, whether they were drafted or signed by an NBA team, how many NBA games they have played, how many of those games they started in, and whether they were still competing in the NBA. Additional information such as the athlete's date of birth, height, weight, and NCAA conference were also scraped. Any matches where the interview date was > 4 years before the draft year or after mid June on the year the player nominated for the draft were considered spurious and discarded, along with any transcripts that mentioned "NBA" in the interview tournament details. The language in the transcripts were then aggregated by player.

To assure that the language samples provided trustworthy estimates of their psychological states, we excluded any sample with fewer than a total of 100 words across their interviews. The final

[6] http://www.asapsports.com/

[7] https://www.nba.com/stats/draft/history/

[8] https://basketball.realgm.com/

interview sample contains 1,533 athletes of whom 846 were drafted or signed by an NBA team between 1995 and 2023. The remaining 687 athletes were not successful in making it onto an NBA team roster. For the NCAA playing statistics, only those from the year immediately prior to the draft the athlete nominated for were assessed. The final physical performance sample contains 20,723 athletes, where 2,281 were signed by an NBA team and 18,442 were not. We then matched the interview sample against the NCAA physical performance dataset. This combined dataset is comprised of 1,294 athletes, 730 of whom succeeded in making it onto an NBA roster while 564 did not. The sample size is smaller than the interview sample due to two factors: the NCAA statistics sample only goes back to 2003, and not all the interviewed draft prospects played in the NCAA (a small percentage were recruited from international sources).

The interviews with college basketball players were only conducted following championship games, and typically only occurred if the athlete was a stand-out player. This sample is thus constructed from the very best players in the very best teams and is therefore not a general representation of student athletes. While we acknowledge the presence of this bias, NBA scouts are likely to focus their attention on the best players on the best teams so we expect this to be a valid training set for the purposes of building models to predict athlete success and career duration within the NBA.

We ran all language samples through the LIWC-22 software package[9] to obtain LIWC language psychology profiles[10]. We discarded all features related to punctuation as the language was transcribed from audio recordings and therefore may not reflect the athlete's actual language use but could instead represent the transcribers themselves. Finally, we discarded any features with a median LIWC score < 0.5 as such low baseline words are highly unstable due to a high rate of zero scores in the data set. Our final LIWC feature count after feature selection was 64.

Per-game and per-minute NCAA playing statistics were calculated for each player in the season immediately prior to their draft year for all features that were not fractions (e.g. percentage of free throws made). These were combined with raw playing statistics to produce a total of 116 features for the physical models.

2.2. Machine Learning Models

We used the LightGBM machine learning algorithm [13] to build three separate supervised classification models to predict which athletes would be signed by an NBA team using language psychology (LIWC) features, NCAA physical performance statistics, and a combination of both. A fourth model was also built with the addition of age (as of January 1st on the year of the draft each player nominated for), height, weight, and the NCAA conference they played in during the season of their draft. For the purposes of this study we consider athletes who were drafted during the NBA draft and those who were not drafted but later signed by an NBA team to be the same.

LightGBM is a gradient boosting framework that builds an ensemble of decision trees by iteratively optimizing predictions to minimize errors based on a training set with known labels (e.g., "Y" vs "N" representing signed vs not signed). LightGBM was chosen because it is exceptionally fast, highly

[9] https://www.liwc.app/

[10] We discarded the word count (WC) feature as athletes who are performing well are likely to attract more attention from the media and perform more interviews. Thus WC is likely strongly correlated with the outcome variable but not in any way that would be useful to NBA teams.

accurate, and scalable to large datasets. Its flexibility and efficient handling of class imbalances through the use of class weights make it a robust choice for predictive modeling, even with minimal hyper-parameter tuning.

While LightGBM has a large number of hyper-parameters, the most significant impact on statistical performance is provided by tuning the fraction of randomly selected features used at each node within the decision trees (feature fraction), the number of training epochs (number of rounds), and the learning rate. While it is important to tune both the feature fraction and the number of rounds to avoid over-fitting, decreasing the learning rate value reduces over-fitting at the expense of increased computation time. A lower learning rate also increases the number of rounds required, so selecting the lowest number possible that results in a reasonable run-time is standard practice. We built our models in the R programming language using the *lightgbm*[11] library.

We tuned our LightGBM models using 5-fold cross-validation with 5 repeats, performing a grid-search for feature fractions between ½ and 1/(total number of features). For all the models we set the learning rate to 0.1. For the LIWC and combined models we set the maximum number of rounds to 500. For the NCAA physical performance model we set the maximum number of rounds to 2,500 due to the much larger size of that training set. Gradient-based One-Side Sampling (GOSS) was selected to improve computational efficiency and the boosting option of Dropouts meet Additive Regression Trees (DART) was used to reduce over-fitting. The downside of using DART is that we could not employ early-stopping, which increased the computation time. The scale positive weight parameter was set as the ratio of negative (un-signed players) to positive (signed players) to account for the imbalance between class categories. All other hyper-parameters were left at their default values.

For each fold iteration, we trained a model using the 4 other folds and used the 5th fold as a validation sample, so that each time we trained and validated using different data sets. Continuing this process over all 5 folds meant that we eventually trained and validated on the entire data set. Repeating the process 5 times with random splitting into the train and validation sets acts to reduce variance in the performance estimates, providing more robust and stable feedback for hyper-parameter optimization. The optimal feature fraction and number of rounds values were determined by minimizing the log-loss metric calculated on predictions applied to the validation data set. Log-loss measures how well a model's predicted probabilities align with the true labels and is a good metric because it evaluates the calibration and confidence of the predicted probabilities, not just the accuracy of the final class predictions. The optimal hyper-parameters obtained through this process are provided in Table 1, along with the model balanced accuracies. As the sizes of our training sets are relatively small (at least for the LIWC and combined models), we used leave-one-out cross-validation to estimate the final model performances so as to maximize the amount of data used to train each model.

As an ensemble algorithm LightGBM is notoriously difficult to interpret. However, analyzing feature importance can tell us which features provide the most information gain. An additional tool to gain a generalized idea of how the model is undertaking decisions can be obtained via partial dependence plots [14]. For each feature in the model, we iterate through the parameter space between the minimum and maximum value with 100 steps, changing the feature value to the new

[11] https://github.com/Microsoft/LightGBM

value while leaving all other features the same, and re-scoring using the final model. In this way we can analyze how changing a feature impacts the model prediction, providing a useful sanity check. We extracted feature importance values (measured using the Gain metric) and generated partial dependence plots for each of the LIWC features used to construct the model (though partial dependence plots are shown only for the top 6 features of the LIWC and combined models).

Feature	**LIWC**	**NCAA**	**LIWC+NCAA**
Learning Rate	0.1	0.1	0.1
Feature Fraction	0.219	0.044	0.370
Number of Rounds	166	1,517	256
Balanced Accuracy	0.6324	0.78295	0.8299

Table 1 – LightGBM model hyper-parameter values.

2.3. Survival Analysis Models

In addition to predicting the likelihood that a student athlete will be signed by an NBA team, we explored whether we could forecast how successful a career in the NBA each player might have. For this purpose we considered two metrics: the total number of games played in the NBA (quantifying career longevity), and the number of those games where the player was in the starting lineup (indicative of the relative importance of each player on their team's roster).

Of the 730 athletes in our combined sample of psychological and physical performance features who were signed, only 520 have completed their NBA careers, providing a limited sample for building machine learning models. However, a dataset such as this where we have a mix of athletes who have finished their careers and others who are still playing is well suited to survival analysis [15]. Survival analysis is a statistical methodology designed to analyze time-to-event data, providing a framework for understanding the distribution of time until an event of interest occurs. Initially developed in medical research to assess patient lifespans, survival analysis has since found applications across a range of disciplines. Key techniques within survival analysis include the Kaplan-Meier estimator [16], which provides non-parametric estimates of survival functions, and the Cox proportional hazards model (CPH) [17], facilitating the exploration of covariate effects on survival using multivariate features.

Using the R library *survival*[12] we generated Kaplan-Meier career survival curves for our 730 signed athletes, grouping them by draft status: early 1st round (pick 1-10), mid 1st round (pick 11-20), late 1st round (pick 21-30), early 2nd round (pick 31-40), mid 2nd round (pick 41-50), late 2nd round (pick 51-60). An additional "Not Drafted" category contains all athletes who were not selected in the draft but were subsequently signed by an NBA team. Separate Kaplan-Meier curves were produced for the number of NBA games played in (Figure 1) and the number of NBA games started in (Figure 2). The Kaplan-Meier curves support the hypothesis that the NBA draft is highly efficient, with athletes selected early in the draft generally going on to have longer careers and more games in the starting lineup. The anti-correlation between draft status and both career games and game starts continues through to those players who were signed after the draft. There are, however, some interesting overlaps: the differences between mid to late picks in the first round are less pronounced than between early in the first round and later in the first round. Similarly, athletes

[12] https://cran.r-project.org/web/packages/survival/index.html

picked early or in the middle of the second round are hard to separate. The difference between athletes picked late in the second round and those who were not drafted but subsequently signed by a team are, however, minimal.

We next fitted CPH models initially using only 3 sets of features: LIWC features only, NCAA playing statistics only, and a combination of LIWC and NCAA statistics. We then experimented by building two additional models with the addition of the physical attributes (age, height, weight) plus college conference, and the draft status group. While the draft status is not known prior to the draft, our goal was to test whether language psychology features, NCAA playing stats, or which school conference an athlete played in would provide any improvement to the models over where in the draft an athlete was selected.

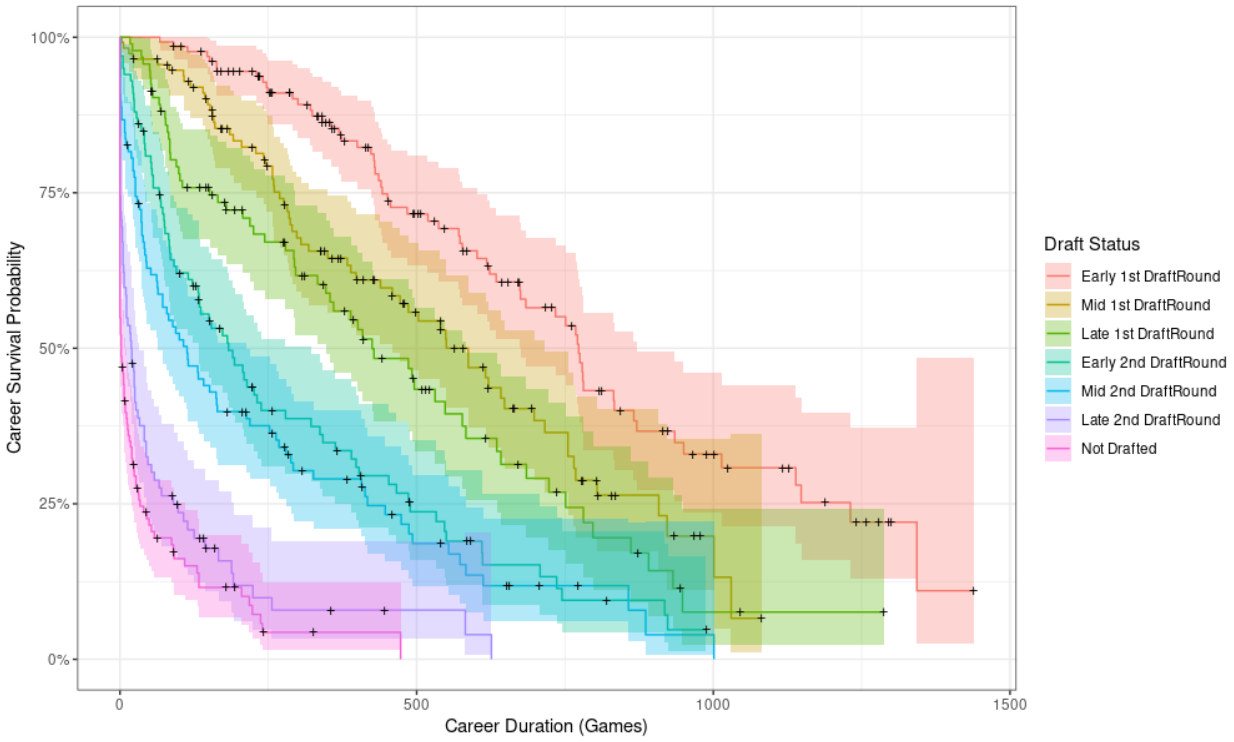

Figure 1 – Kaplan-Meier career survival curves for NBA players in terms of number of NBA games played, grouped by their draft status.

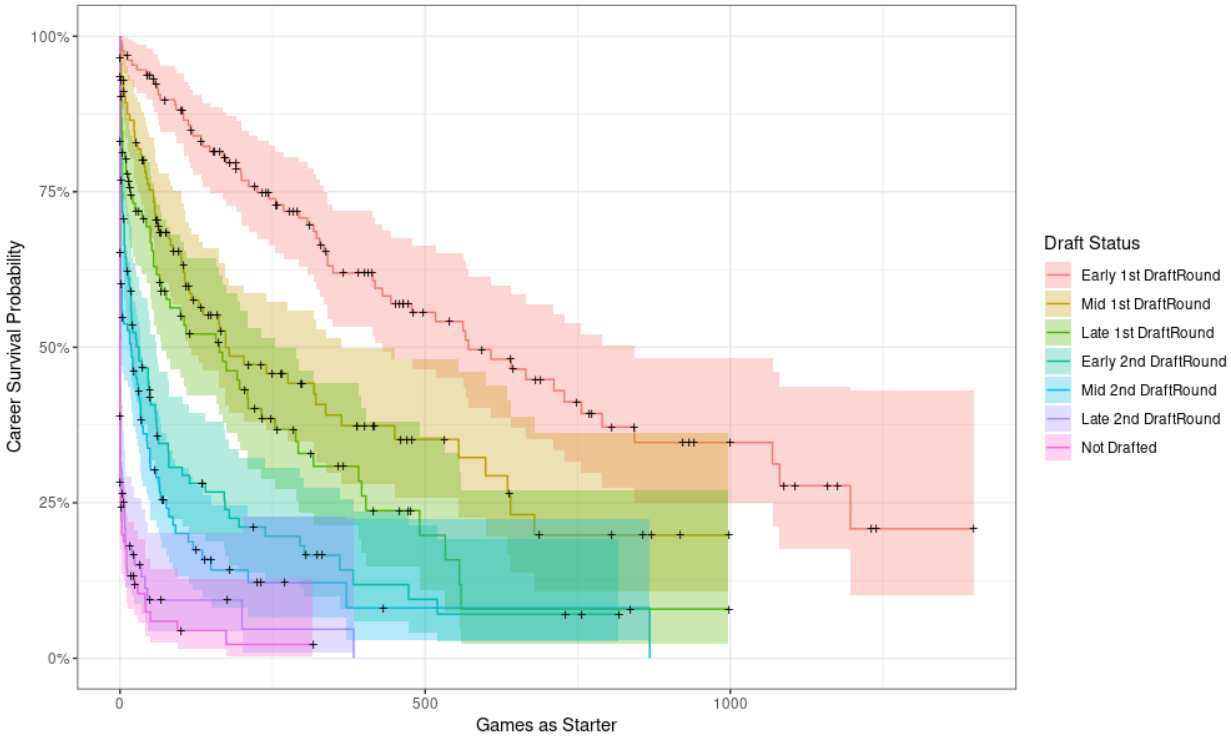

Figure 2 – Kaplan-Meier career survival curves for NBA players in terms of number of NBA games where they were in the starting lineup, grouped by their draft status.

3. Results

3.1. LightGBM Model - Psychological Features

Our cross-validation process for the language psychology only (i.e. LIWC) model achieved a balanced prediction accuracy of 63.2%. Due to the slight imbalance between classes, the no information rate (NIR; the accuracy you would obtain through random selection) was 55.2%, and the p-value for the accuracy exceeding the NIR was 7.6 e^{-13} meaning this result is significant. Our prediction accuracies for both classes were fairly balanced, with the accuracy at correctly predicting the N and Y classes being 61.1% and 66.1%, respectively. Figure 3 shows the confusion matrix for this model.

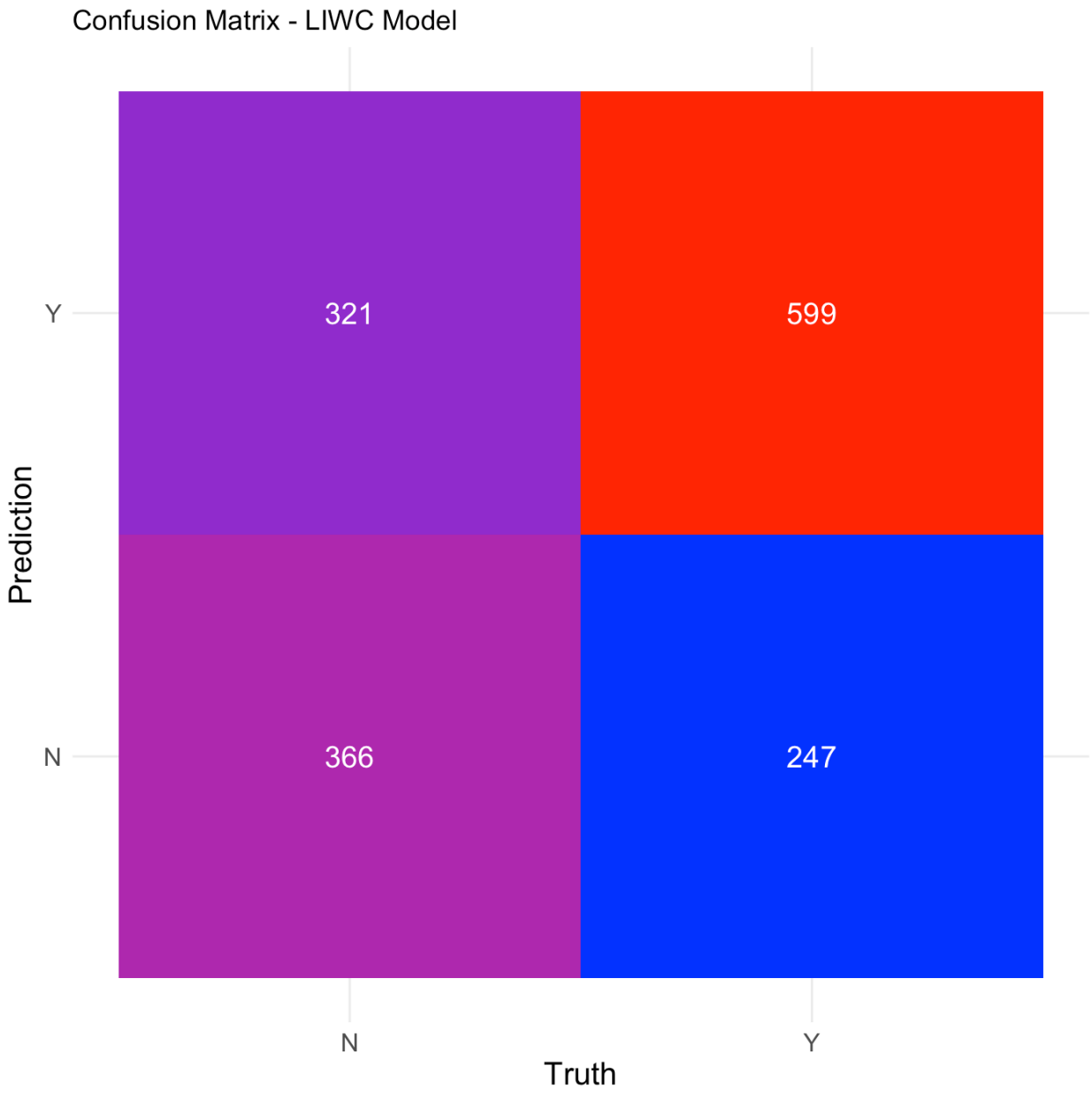

Figure 3 – Confusion matrix for the LIWC LightGBM model from leave-one-out cross-validation (Y = signed by NBA team, N = not signed).

Figure 4 shows the feature importance plot for the final tuned model. The six most important features in the model are future focus (focusfuture), conjunctions (conj), authenticity (Authentic), positive emotions (emo_pos), present focus (focuspresent), and verbs. Figure 5 shows the partial dependency plots for these six features, allowing us to determine in which direction varying values for each feature push the model predictions. For future focus, values between 0.45 and 4.04 lead the model to predict lower probabilities (~2% lower) of success while very low values (< 0.45) lead the model to predict a higher success probability (~4% higher). Values > 5 result in very marginal (< 1%) increase in success probability. Similarly, present focus scores above 5 produce higher probabilities (~4% higher) while lower scores push the model to predict a higher likelihood of

failure (~4.5% lower). High scores in present focus are indicative of thinking about what is happening right now and can be an indicator of mindfulness and a sign of someone who is living in the moment. Higher future focus scores reflect people's thinking about the future, including their goals and plans. Future focus language is also associated with better health at the individual and community level, so is likely an indicator of a balanced and healthy outlook. Present focus in particular could indicate an athlete has a higher likelihood of achieving a flow state, which has been found to have a strong positive relationship with performance in athletes and musicians [18, 19]. In contrast, a higher conjunction score (indicative of more complex language use) leads to the model down weighting an athlete's chance of being signed. Conjunctions are joining words and a high score indicates more complex language usage. The authenticity dimension is based on a group of words associated with being self-focused, honest, and non-defensive. People high in this domain are likely more comfortable in interviews.

Calculating simple Pearson correlations between the LIWC scores and a binary label (i.e. 0 for not signed, 1 for signed) can also provide some insight into what is driving the model's decision making. Table 2 shows the correlation coefficients for those LIWC features where the absolute value of the coefficient is > 0.075.

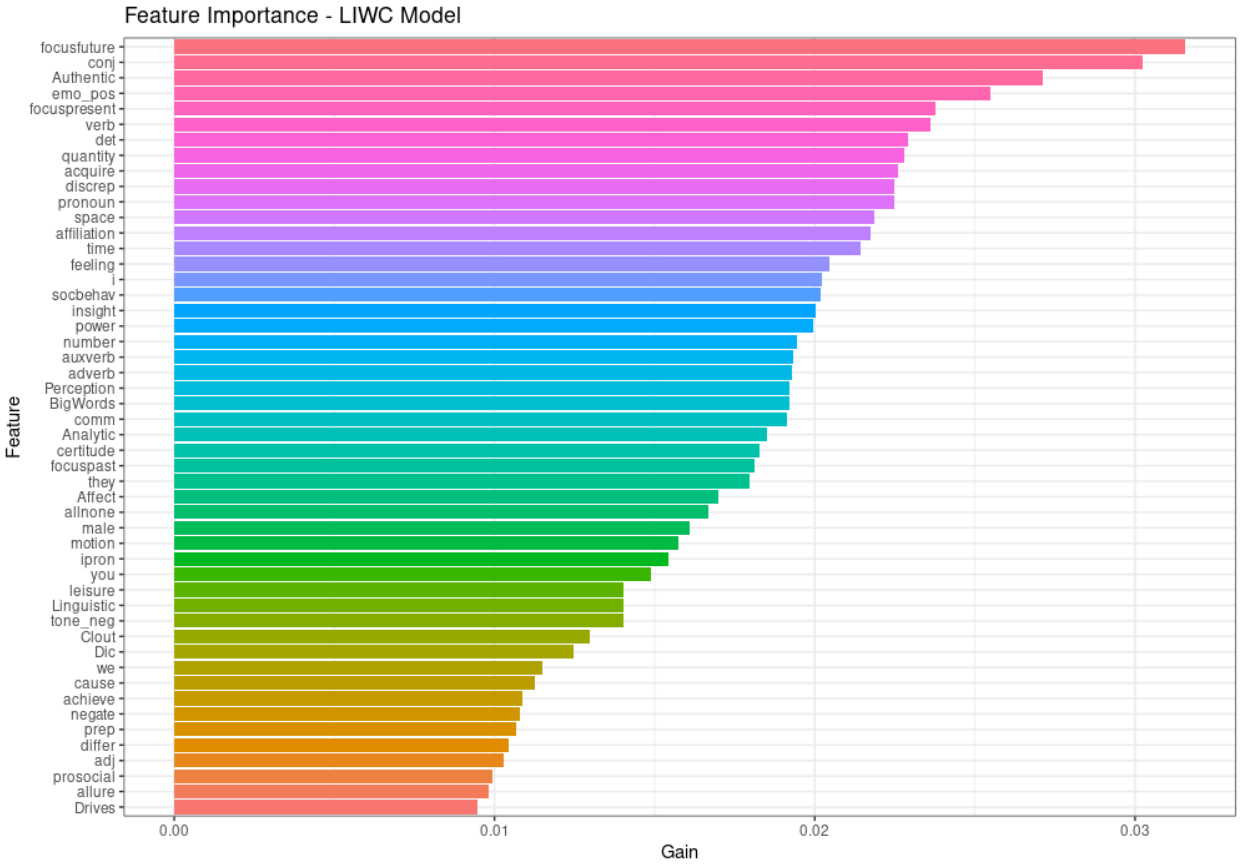

Figure 4 – Feature importance plot for the top 50 features in the LIWC LightGBM model. The higher the Gain, the more important the feature is. See the LIWC-22 manual[13] for definitions of each of these features.

[13]https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf

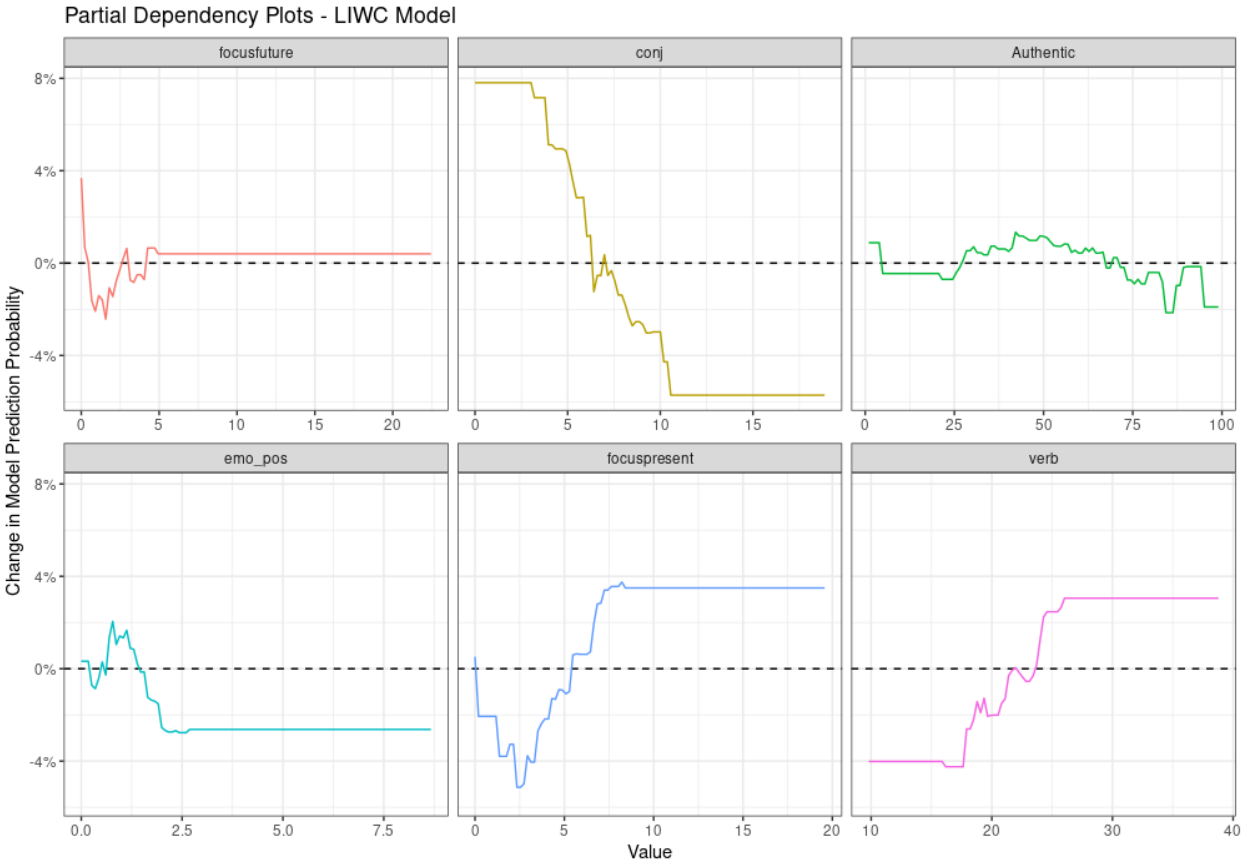

Figure 5 – Partial dependence plots for the top 6 LIWC model features by decreasing importance. An increase in probability correlated with an increase in LIWC score means that higher LIWC scores result in higher model prediction probabilities (and vice versa). See the LIWC-22 manual[14] for definitions of each of these features.

Feature	Correlation
conj	-0.170
ipron	-0.114
tentat	-0.109
adverb	-0.092
feeling	-0.086
function	-0.084
socrefs	0.085
verb	0.104
you	0.109
ppron	0.131

[14]https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf

Table 2 – Pearson correlation coefficients between LIWC features and a binary label (i.e. 0 for not signed, 1 for signed by an NBA team). Only those features with an absolute value of > 0.1 are shown for clarify.

3.2. LightGBM Model – NCAA Playing Statistics

Our cross-validation process for the physical performance (i.e. NCAA) model achieved a balanced prediction accuracy of 78.3%. Unlike the other data sets, the NCAA sample is highly imbalanced with a NIR of 89.0%. However, the p-value for the accuracy exceeding the NIR was < 2.2 e^{-16}. Our prediction accuracies for both classes were highly imbalanced, with the accuracy at correctly predicting the N and Y classes being 95.4% and 53.6%, respectively. These metrics indicate that this model on its own is not particularly useful in terms of being able to accurately forecast future NBA success.

3.3. LightGBM Model – LIWC + NCAA Playing Statistics

The cross-validation process for the combined LIWC plus NCAA playing statistics model achieved a balanced prediction accuracy of 83.0%. The NIR for this sample was 56.4%, and the p-value for the accuracy exceeding the NIR was a highly significant < 2 e^{-16}. Our prediction accuracies for both classes were fairly balanced, with the accuracy at correctly predicting the N and Y classes being 81.1% and 85.0%, respectively. Figure 6 shows the confusion matrix for this model.

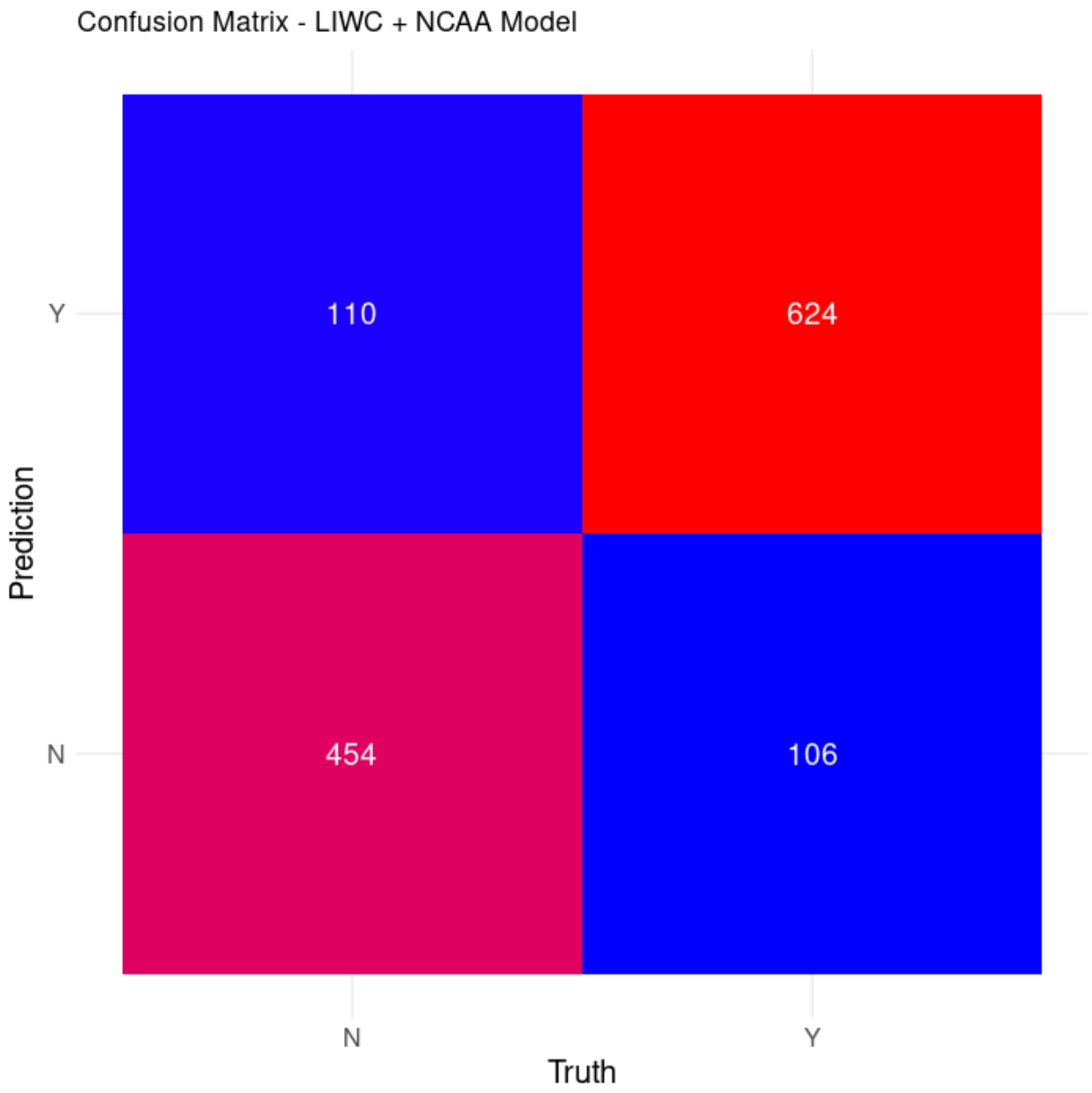

Figure 6 – Confusion matrix for the combined LIWC + NCAA statistics LightGBM model from leave-one-out cross-validation (Y = signed by NBA team, N = not signed).

Figure 7 shows the feature importance plot for the final tuned model. While the model is clearly dominated by the NCAA statistics, a number of LIWC features are present in the top 50 most important features. The six most important LIWC features in the model are words per sentence (WPS), conjunctions (conj), adverbs, negations (negate), feeling words, and absolutist language (allnone).

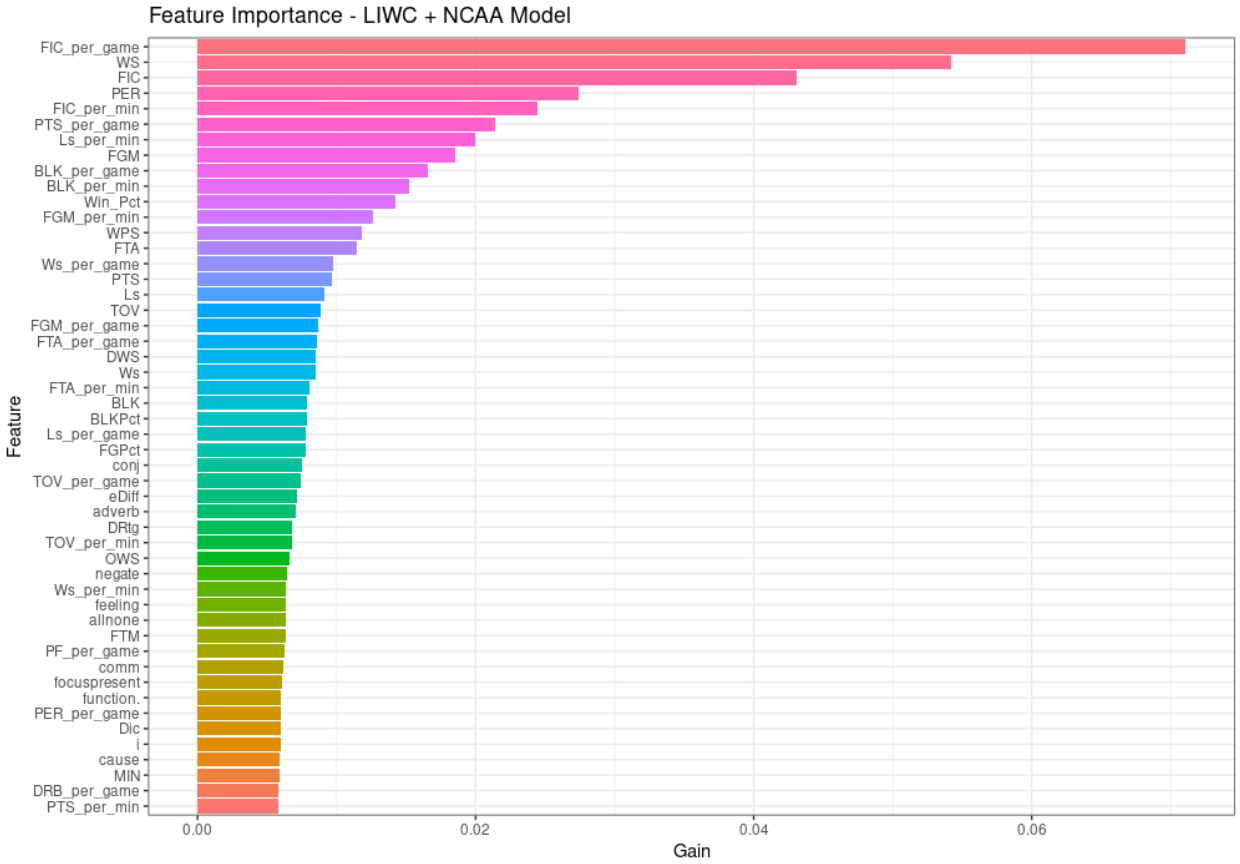

Figure 7 – Feature importance plot for the top 50 features in the LIWC + NCAA statistics LightGBM model. The higher the Gain, the more important the feature is. See the LIWC-22 manual[15] for definitions of each of the LIWC features, and the Real GM website for definitions of the NCAA statistics[16].

Figure 8 shows the partial dependency plots for the top 6 LIWC features in this model. As with the LIWC only model, we can see that more complex language (represented by words per sentence and conjunctions) pushes the model to reduce the likelihood of an athlete making it into the NBA. Similarly, higher rates of negations and feeling words are anti-correlated with success. Feeling words reflect a person's expressing their feelings (e.g. "I feel…", "I sense that you are…") while negations reflect being more defensive and inhibited. Combined, high scores in both of these categories might suggest people who are more insecure. Interestingly, the use of adverbs and absolutist language have non-linear correlations with success. Very low and very high values are associated with higher probabilities, but in between the model predicts lower probabilities of success. Absolutist language includes words such as never, absolutely etc. People who are depressed tend to use these words more often, so the fact that higher scores result in lower success probabilities supports them being predictive of poor performance (the absolutist scores do return

[15]https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf

[16]https://basketball.realgm.com/ncaa/stats

to positive territory above scores of ~2.5, but the change in model probability is < 1% indicating it does not impact the model significantly above this range).

The addition of player age, height, weight and NCAA conference boosted the model accuracy further to 86.7%. For this model, age was the most important feature, with athletes older than ~22.5 years having lower probabilities of success. The NCAA conference was the third most important feature, with athletes from the power conferences (e.g. the South Eastern Conference, Big 12, and Pac-12 conferences) having much higher success probabilities on average than those of lower tier conferences. Height was also an important feature (8th), with students taller than 205cm (~6'8") having higher success probabilities. Weight, however, ranked low in terms of feature importance (81st) and therefore does not have a significant impact on the model.

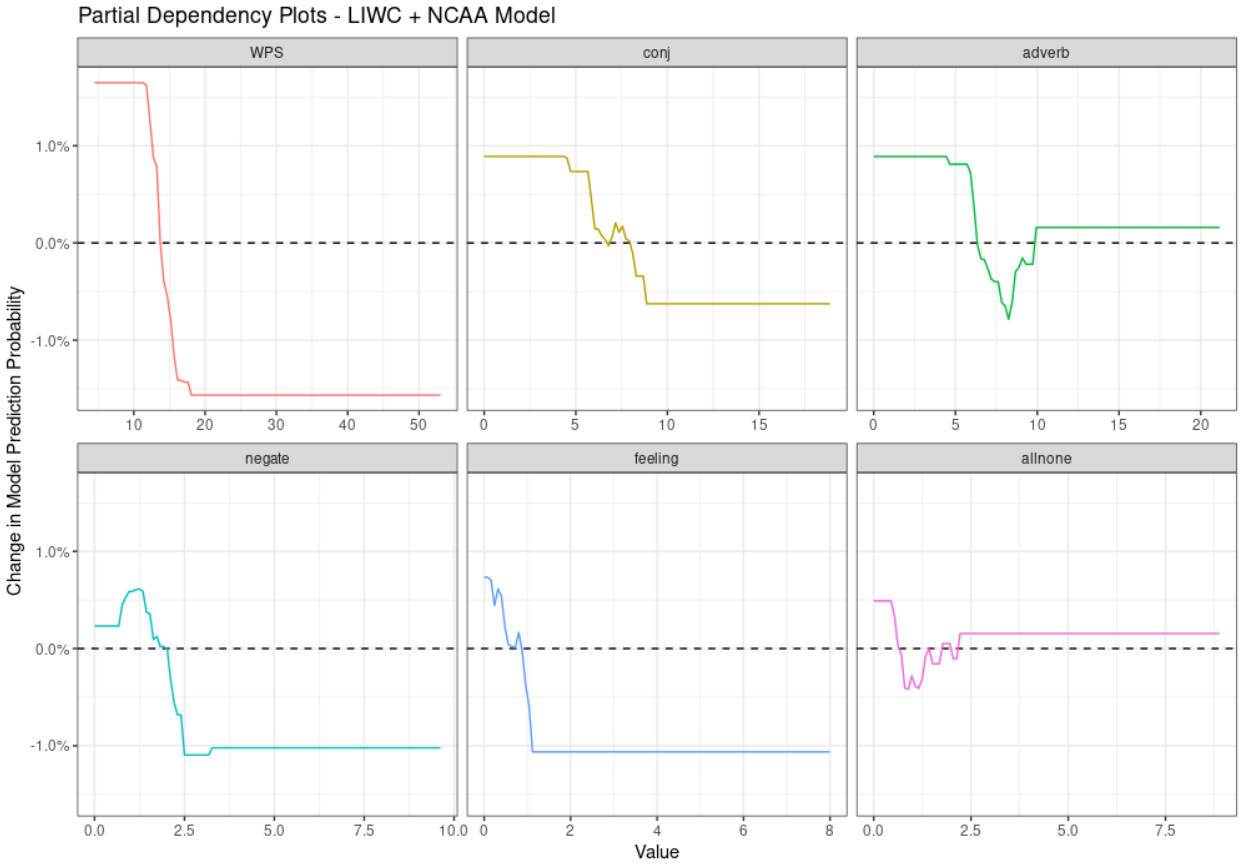

Figure 8 – Partial dependence plots for the top 6 LIWC features in the LIWC + NCAA model by decreasing importance. An increase in probability correlated with an increase in LIWC score means that higher LIWC scores result in higher model prediction probabilities (and vice versa). See the LIWC-22 manual[17] for definitions of each of these features.

[17]https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf

3.4. Survival Analysis – NBA Games Played

To assess the importance of psychology on career duration, we built a CPH model with the LIWC features to predict the probability that an athlete will play a given number of NBA games in their career. The concordance index (C-index) is a commonly used metric in survival analysis to evaluate the performance of a predictive model [20]. A C-index of 0.5 indicates that the model performs no better than chance, while a value of 1 indicates perfect concordance. The C-index for the model built purely from the LIWC measures was 0.613 +/- 0.014, indicating that the language psychology features do provide predictive power. We also built models using just the NCAA playing stats (C-index = 0.704 +/- 0.013), a combination of LIWC + NCAA features (C-index = 0.738 +/- 0.013), and LIWC + NCAA features + age/height/weight/conference (C-index = 0.808 +/- 0.01). Finally, we added draft status to the combination of the all other features and achieved a C-index = 0.842 +/- 0.008. Figure 9 shows a comparison of the C-index values for all five models.

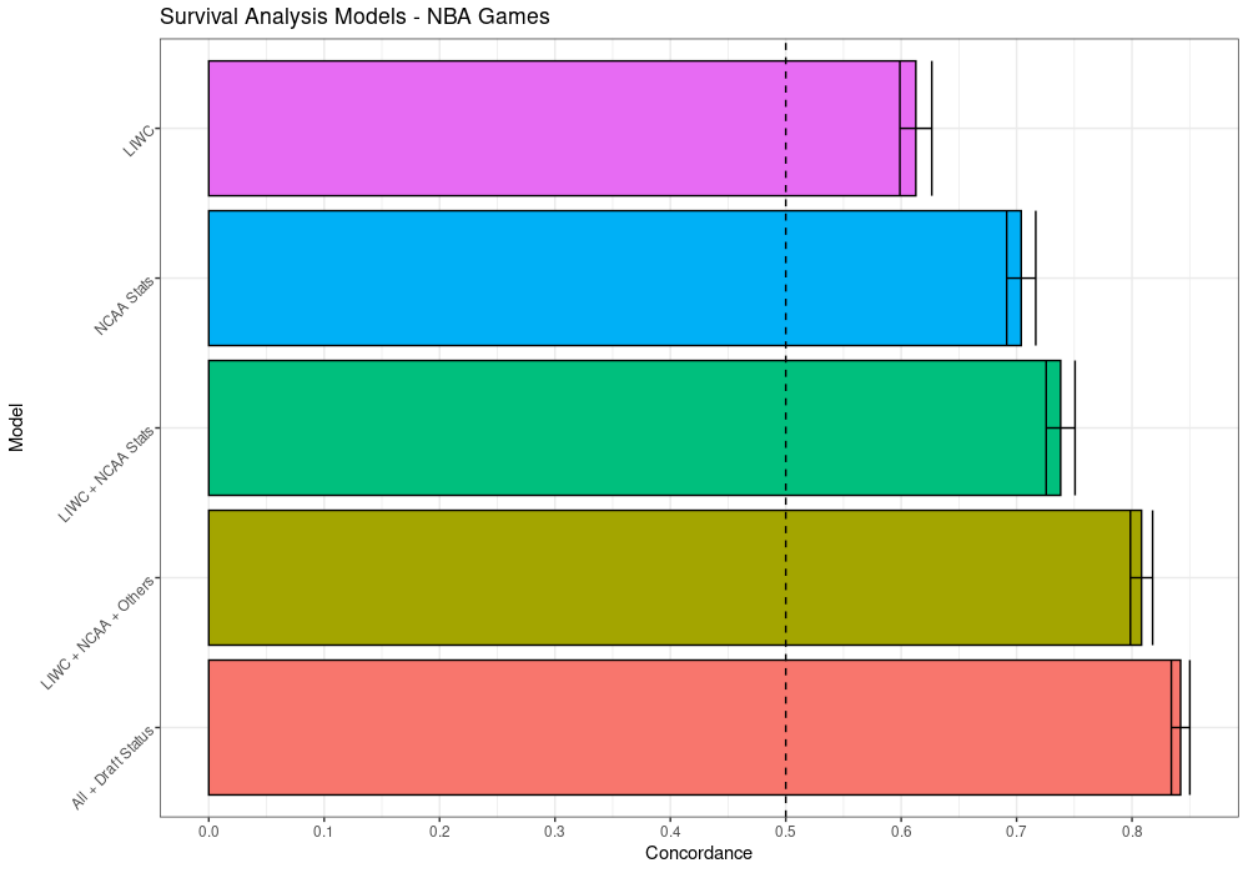

Figure 9 – Comparison of concordance values for Cox proportional hazard models to predict number of NBA games played. The models represented are: LIWC language features, NCAA playing stats, a combination of LIWC + NCAA features, a combination of LIWC + NCAA + age/height/weight/conference, and with all of the above plus draft status. The error bars show the standard error for each concordance value.

3.5. Survival Analysis – NBA Starting Lineup Games

We also built CPH models to predict how many NBA games a player will be in the starting lineup. The C-index for the model built purely from the LIWC measures was 0.617 +/- 0.015. We also built models using just the NCAA playing stats (C-index = 0.712 +/- 0.013), a combination of LIWC + NCAA features (C-index = 0.748 +/- 0.013), and LIWC + NCAA features + age/height/weight/conference (C-index = 0.823 +/- 0.01). Finally, we added draft status to the combination of the all other features and achieved a C-index = 0.857 +/- 0.008. Figure 10 shows a comparison of the C-index values for all five models.

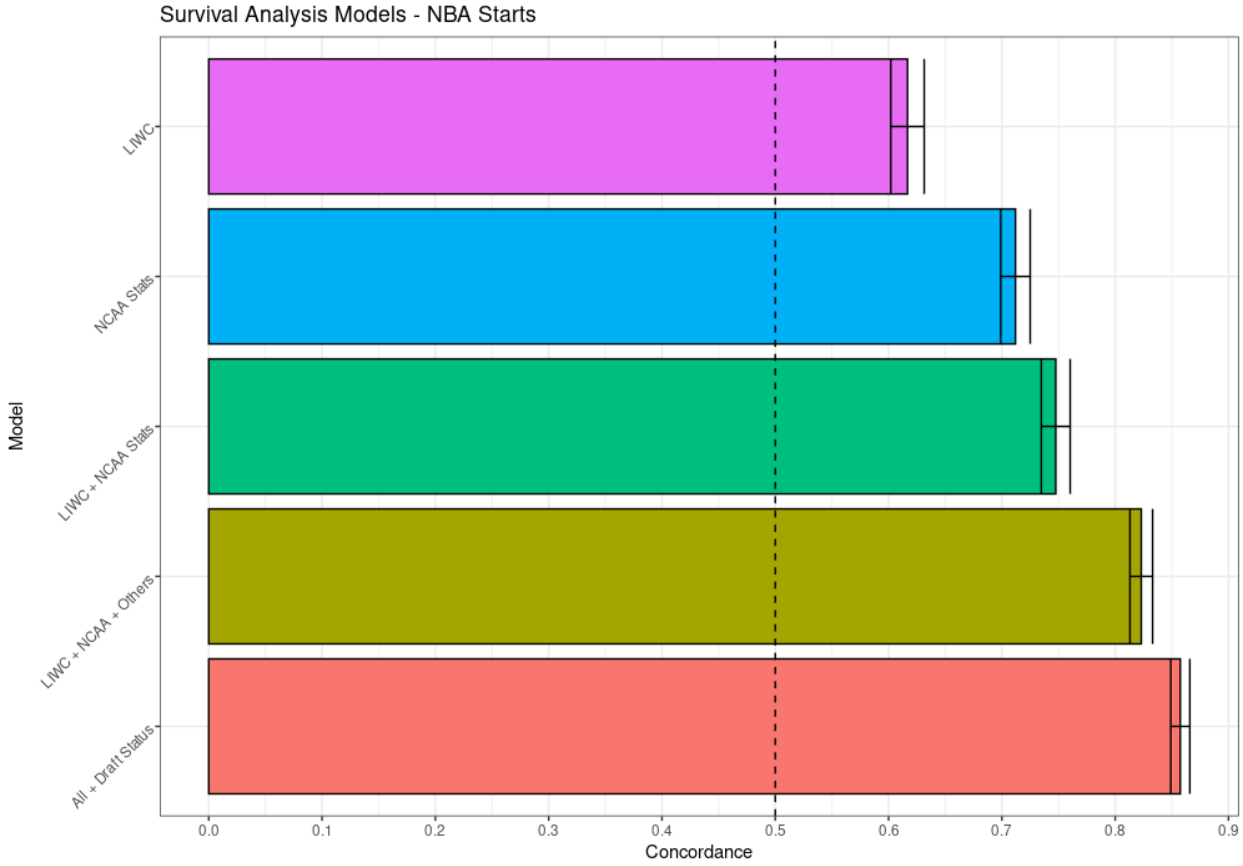

Figure 10 – Comparison of concordance values for Cox proportional hazard models to predict number of NBA games a player will start in. The models represented are: LIWC language features, NCAA playing stats, a combination of LIWC + NCAA features, a combination of LIWC + NCAA + age/height/weight/conference, and with all of the above plus draft status. The error bars show the standard error for each concordance value.

4. Conclusion

Our research suggests that the ways that athletes use words in everyday interviews and conversations can predict the ways they will ultimately perform in professional sports. Through computer-based natural language analyses, we can quantify people's thinking styles and psychological traits that are correlated with success and build models to predict the likelihood that

an athlete will be signed by an NBA team. Meaningful psychological traits have traditionally been difficult to obtain during the pre-draft process. Employing language psychology-based machine learning models addresses two of the biggest barriers to obtaining this data (i.e., access to athletes and potential issues with the accuracy of self-reported psychological characteristics), transforming what has historically been a challenging factor to include in pre-draft models into an easily accessible option. Using the LightGBM machine learning algorithm, we were able to predict with an accuracy 63% which student-athletes will successfully make it onto an NBA roster using language psychology features alone. Augmenting these features with college playing statistics we can achieve accuracies of 83%, and with the addition of physical traits such as age, height, and weight, plus the NCAA conference the student played in, we can boost this further to 87%. We also found that survival analysis models built using language psychology features were able to predict the likelihood that an NBA player's career will last beyond a given number of games far better than by random chance, and that combined with NCAA statistics, age/height/weight/conference, and a player's draft status these models have significant predictive power. Similarly, using survival analysis techniques we were able to predict how many NBA games a player will feature in the starting lineup with high precision. These models would allow NBA teams to provide probabilistic forecasts of player career durations pre-draft, allowing for a significant improvement in risk assessment over traditional methods.

Combining our analysis of the LightGBM feature importance, partial dependence plots and simple correlations, we were able to determine which features are driving the models. We saw evidence that people who are more present or future focused and who use less complex language (possibly an indication of quick decision making as opposed to a longer more analytical thinking style) are more likely to make it on to an NBA roster and have higher probabilities of enjoying longer and more successful careers. Athletes who have more of a present focus may be more likely to achieve a flow state, which is thought to be positively correlated with performance. Those who are more likely to not make it into the NBA use more complicated language and ruminate more on the past than look to the future.

Using language from student athletes prior to the NBA draft in combination with NCAA playing statistics we have shown that we can accurately predict not only which athlete's will be signed by an NBA team, but how long their careers will last and how impactful they will be. These models could be used to forecast the probability of a player making it to a specific milestone, for example past the initial 4-year rookie contract. Such forecasts could be utilized by a range of interested parties beyond the NBA franchises themselves including player agents and sponsors. This work has implications for talent scouting and player development, potentially revolutionizing how teams identify emerging talent. While this research focused entirely on basketball, it can also be applied to any other high-performance sport where athlete language samples are available.

Acknowledgments

We thank Fraser Tully and Andrew Goodwin for their comments, suggestions and general feedback that helped to craft this paper into it's final form.

References

[1] Teramoto M, Cross CL, Rieger RH, Maak TG, Willick SE. (2018). Predictive Validity of National Basketball Association Draft Combine on Future Performance. Journal of Strength and Conditioning Research, 32(2), 396-408.

[2] Moxley, J. H., & Towne, T. J. (2015). Predicting Success in the National Basketball Association: Stability & Potential. Psychology of Sport and Exercise, 16, 128-136.

[3] Ranisavljev, I., Mandic, R., Cosic, M., Blagojevic, M., & Dopsaj, M. (2021). NBA Pre-Draft Combine is the Weak Predictor of Rookie Basketball Player's Performance. Journal of Human Sport and Exercise, 16(3), 493-502.

[4] McNeil, D. G., Phillips, W. J., & Scoggin, S. A. (2023). Examining the Importance of Athletic Mindset Profiles for Level of Sport Performance and Coping. International Journal of Sport and Exercise Psychology, 1-17.

[5] Hogan, S. R., Taylor, D., Boone, R. T., & Bowman, J. K. (2023). The Athletic Intelligence Quotient and Performance in the National Basketball Association. Frontiers in Psychology, 14.

[6] Siemon, D., Ahmad, R., Huttner, J. P., & Robra-Bissantz, S. (2018). Proceedings of the Americas Conference on Information Systems (AMCIS), *[s. l.]*, p. 1–5.

[7] Piedmont, R. L., Hill, D. C., Blanco, S. (1999). Predicting Athletic Performance Using the Five-Factor Model of Personality. Personality and Individual Differences, 27(4), 769-777.

[8] Smith, R. E. & Christensen, D. S. (1995). Psychological Skills as Predictors of Performance and Survival in Professional Baseball. Journal of Sport & Exercise Psychology, 17(4), 399-415.

[9] McDaniel, M. J., Beier, M. E., Perkins, A. W., Goggin, S., & Frankel, B. (2009). An assessment of the fakeability of self-report and implicit personality measures. Journal of Research in Personality, 43(4), 682-685.

[10] Tausczik, Y. R. & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology, 29(1), 24-54.

[11] Pennebaker, J. W. (2011). The Secret Life of Pronouns: What our Words Say About Us. Bloomsbury Press.

[12] Pennebaker, J. W., Boyd, R. L., Booth, R. J., Ashokkumar, A., & Francis, M. E. (2022). Linguistic Inquiry and Word Count: LIWC-22. Austin, TX: Pennebaker Conglomerate, www.liwc.app.

[13] Ke G., Meng, Q., Finley, T. et al. (2017). Lightgbm: A highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, 30, 3146-3154.

[14] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29, 1189-1232.

[15] Klein, J.P. & Moeschberger, M. L. (2003). Survival Analysis: Techniques for Censored and Truncated Data (Vol. 1230). New York: Springer.

[16] Kaplan, E. L. & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53, 457-481.

[17] Cox, D. R. (1972). Regression Models and Life Tables (with Discussion). Journal of the Royal Statistical Society, Series B, 34, 187-220.

[18] Antonini Philippe, R., Singer, Jaeger, J.E. et al. (2002). Achieving Flow: An Exploratory Investigation of Elite College Athletes and Musicians. Frontiers in Psychology, 13, 831508.

[19] Csikszentmihalyi, M., & Nakamura, J. (2002). The Concept of Flow. Handbook of Positive Psychology.

[20] Rahman, M.S., Ambler, G., Choodari-Oskooei, B. et al. (2017). Review and Evaluation of Performance Measures for Survival Prediction Models in External Validation Settings. BMC Medical Research Methodology, 17, 60.