Predicting NBA Career Duration Based on Rookie Season

Andy Chuang
Department of Statistics
University of Virginia

Kevin Nguyen
Department of Statistics
University of Virginia

***Abstract*—This report investigates the association between in-game statistics during a player's rookie season and their career duration of more than four years in the NBA. The dataset was centered and scaled, duplicate players were aggregated by mean, and predictors were converted from a per-game basis to a per-minute basis. We employed three classification methods, including logistic regression, random forests, and gradient boosting, with the response variable `target_4yrs` and predictors `gp`, `pts`, `ft`, `fg`, `X3p`, `reb`, `ast`, `stl`, `blk`, `tov`, and `min`. All three methods were trained on a training set and then ran on the test set to predict a rookie's career duration. The logistic regression model achieved an accuracy rate of 72.9730%, with `gp`, `min`, and `reb` being the top three most important predictors. The random forests model achieved an accuracy rate of 71.8147%, with `gp`, `min`, and `fg` identified as the top three most important predictors. The gradient boosting model achieved an accuracy rate of 72.2008%, with `gp`, `fg`, and `stl` being the top three most important predictors. Although the gradient boosting model does not have the highest accuracy, it has the lowest false positive rate, which would save NBA franchises the most money in the context of this problem. We recognize that the accuracy of our models is not particularly high, but we are limited by the computational power of our devices, as well as by the imbalanced binary response variable. Future work could involve stratified sampling to address the slightly unbalanced response variable, performing feature selection on the logistic regression model to potentially improve accuracy, and raising the classification threshold to reduce the false positive rate, improve interpretability, and obtain more accurate results.**

I. Introduction

A. General Background

In the context of the National Basketball Association (NBA), franchises place significant emphasis on assessing rookie players for potential recruitment at a relatively low cost. To this end, it would be beneficial for these franchises to have a predictive model that can forecast the likelihood of a rookie player having a productive and sustainable career in the league, based on their performance statistics during their inaugural season.

The dataset chosen for this study contains in-game statistics for NBA players during their rookie season, encompassing variables such as points per game, minutes played, free throws, and career status after four years. Given that rookie contracts typically have a duration of approximately four years, including team options, this dataset represents an optimal resource for NBA teams to evaluate a player's potential longevity in the league.

B. Variable Descriptions

Variable Name	Description
Career Duration (`target_4yrs`)	Whether a rookie's career lasted 4 or more years (0 if no, 1 if yes)
Games Played (`gp`)	Number of games played in rookie season
Points per Minute (`pts`)	Average points scored per minute
Free Throws (`ft`)	Average number of free throws made per minute
Field Goal Percentage (`fg`)	Average ratio of baskets made to baskets attempted
3-Point Percentage (`X3p`)	Average ratio of 3-pointers made to 3-pointers attempted
Rebounds (`reb`)	Average number of rebounds made per minute
Assists (`ast`)	Average number of assists made per minute
Steals (`stl`)	Average number of steals made per minute
Blocks (`blk`)	Average number of blocks made per minute
Turnovers (`tov`)	Average number of turnovers made per minute
Minutes Played (`min`)	Average number of minutes played per game

II. Objective

The research objective of this report is to investigate the in-game statistics of a player during their rookie season and identify the key performance indicators that exhibit the strongest association with the player's career duration exceeding four years in the NBA. Our aim is to provide valuable insights into the factors that are indicative of sustained success in the league for rookie players.

III. Approach to Project

A. Data Cleaning

In the initial stage of our approach, we imported the dataset into the R environment for analysis. However, prior to addressing our research question, we performed data cleaning to ensure data integrity. Specifically, we encountered instances of duplicate records for several NBA rookies in our dataset. To address this, we employed an aggregation strategy that involved calculating the mean of these duplicate rows, thereby retaining all data points. Moreover, as our variables were expressed on a per-game basis, it was necessary to standardize the data to avoid potential confounding effects resulting from unequal minutes played by different players. We achieved this standardization by dividing each player's statistics by their respective minutes played. Additionally, to prevent any particular predictor from unduly influencing the analysis, we centered and scaled our predictors.

B. Exploratory Data Analysis

We conducted a preliminary assessment of the distribution of our binary response variable.

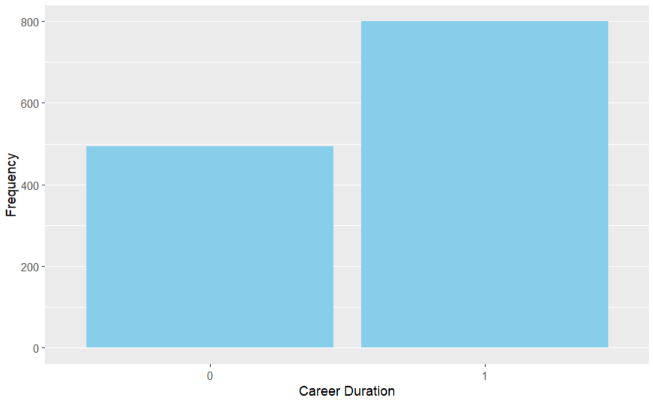


Fig. 1. Distribution of Career Duration.

From Figure 1, we note a distribution of approximately 65% and 35%, which is acceptable given the context of our research problem. Subsequently, we generated histograms of each predictor variable using the `hist.data.frame()` function available in the `Hmisc` package.

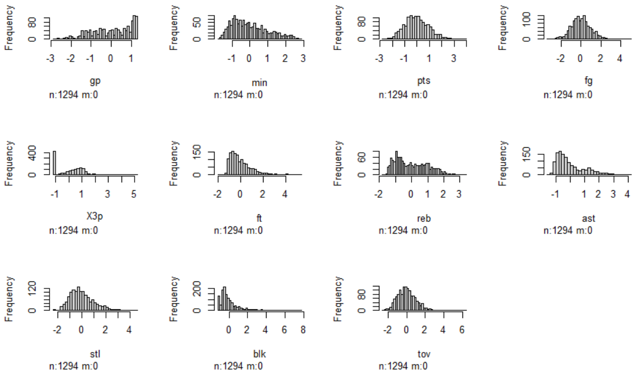


Fig. 2. Histograms of Predictor Variables.

Upon examination of Figure 2, we observed that the majority of the predictor variables exhibited a roughly normal distribution, while a few predictors displayed some degree of skewness. However, considering the problem context, such skewness is reasonable and does not necessarily warrant remedial action. Furthermore, to address the issue of non-normality, we opted for analytical methods that do not rely on normality as a prerequisite assumption, as detailed in the subsequent section.

C. Methodology

To address our research objective, we have selected three methods: logistic regression, random forests, and gradient boosting. Prior to analysis, we partitioned the dataset into training and testing sets utilizing an 80% and 20% split ratio. The training subset was used for model training purposes, while the test subset was reserved for forecasting rookies' career duration.

Logistic regression is a widely-used classification algorithm, suitable for predicting a binary categorical variable based on a set of independent variables. It offers easy interpretability and allows us to assess the significance of each predictor by examining the Wald test statistic. However, the assumption of a linear relationship between the predictors and the log odds of the response variable may not hold in some cases.

To mitigate this issue, we have chosen random forests, a robust and flexible algorithm that is less prone to overfitting and offers useful feature selection capabilities. The method assesses variable importance through the mean decrease in accuracy and mean decrease in Gini impurity. However, the risk of overfitting remains, as the algorithm constructs a large decision tree on our dataset.

Our final method, gradient boosting, offers a solution to this issue. It is an ensemble method that learns from previous iterations, automatically selects the most important features, and assesses the relative influence of each predictor. It is a flexible and robust approach that can help us address our research objective. All three methods we have selected will contribute to achieving our research goals.

IV. Results

A. Logistic Regression

We initiated our analysis by performing logistic regression. Specifically, we employed the `glm()` function from the `glm` package to build our model, using all predictors in the dataset to predict the `target_4yrs` response variable. The `family = "binomial"` argument was utilized, and we opted to include all variables in our model, as each predictor was deemed relevant to addressing our research question. Our model yielded an accuracy rate of 72.9730%, a false positive rate of 45.0550%, and a false negative rate of 17.2619%. Moreover, we generated a variable importance plot to visualize the contribution of each predictor to the model.

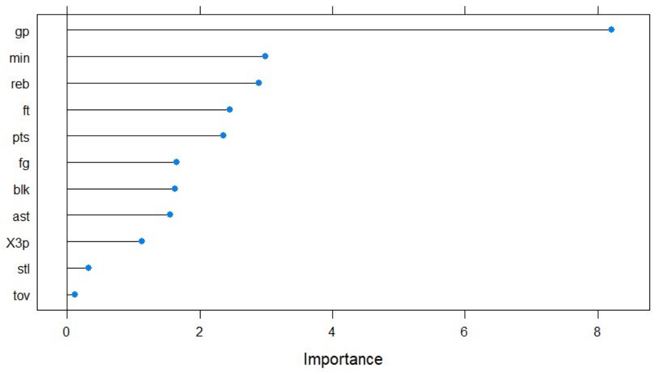


Fig. 3. Variable Importance Plot for Logistic Regression.

Figure 3 reveals that `gp`, `min`, and `reb` are the most important predictors for the logistic regression model.

B. Random Forests

We proceeded with the random forest model by conducting a grid search using the `train()` function, with the aim of finding the optimal tuning parameters. Specifically, we explored a range of `mtry` values from 2 to 8, and `ntree` values from 100 to 500 (incremented by 100), with 10-fold cross-validation. This search yielded the best `mtry` and `ntree` values of 5 and 400, respectively. Upon identifying the optimal tuning parameters, we trained our random forest model using the `randomForest()` function from the `randomForest` library with all the predictors as previously specified. The model achieved an accuracy rate of 71.8147%, a false positive rate of 37.3626%, and a false negative rate of 23.2143%. Again, we generated variable importance plots to visualize the contribution of each predictor to the model.

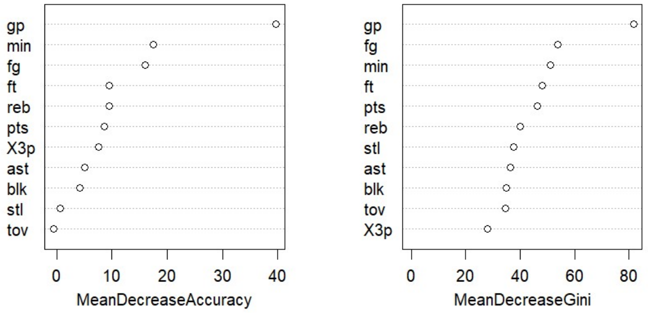


Fig. 4. Variable Importance Plots for Random Forests.

As depicted in Figure 4, the plot based on the mean decrease in accuracy identifies `gp`, `min`, and `fg` as the top three predictors, while the plot based on the mean decrease in Gini impurity identifies `gp`, `fg`, and `min` as the top three predictors. However, despite the slight variation in the ranking order, both plots consistently highlight these three predictors as the most important in the model.

C. Gradient Boosting

Lastly, we proceeded with the boosting model, employing the same methodology as before by conducting a grid search to find the optimal tuning parameters using the `train()` function. The grid search was performed over `ntree` values ranging from 100 to 500 (incremented by 100), `interaction.depth` values between 2 and 5, and `shrinkage` values of 0.01, 0.05, and 0.1, while applying 10-fold cross-validation. This search yielded the best `ntree`, `interaction.depth`, and `shrinkage` values of 400, 4, and 0.05, respectively. Upon identifying the optimal hyperparameters, we fit the boosting model using the `gbm()` function from the `gbm` library with all the predictors as previously specified. The model exhibited an accuracy rate of 72.2008%, a false positive rate of 38.4615%, and a false negative rate of 22.0238%. Once more, we generated a variable importance plot to visualize the contribution of each predictor to the model. In order to address the issue of some variable names not being included in the variable importance plot, we incorporated the corresponding relative influence values for each predictor.

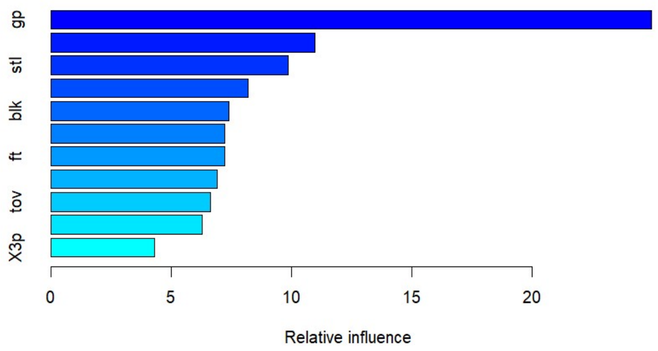
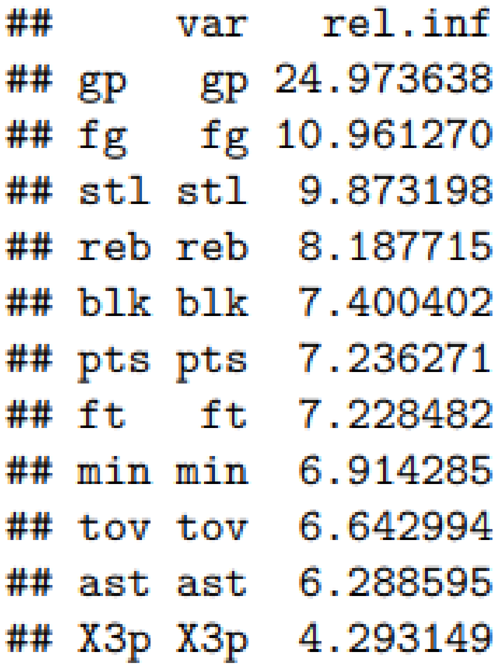


Fig. 5. Variable Importance Plot for Gradient Boosting.



```
##      var   rel.inf
## gp   gp  24.973638
## fg   fg  10.961270
## stl stl   9.873198
## reb reb   8.187715
## blk blk   7.400402
## pts pts   7.236271
## ft   ft   7.228482
## min min   6.914285
## tov tov   6.642994
## ast ast   6.288595
## X3p X3p   4.293149
```

Fig. 6. Relative Influence Values for Each Predictor.

Based on Figures 5 and 6, we observe that the three most important predictors are `gp`, `fg`, and `stl`.

V. Conclusion

A. Summary

In terms of accuracy, logistic regression exhibits the best performance among all models evaluated. However, the model displays the highest false positive rate, which in the context of the problem, is unfavorable. False positives indicate that the NBA team's model would classify a player as having a career beyond the rookie contract period, leading the team to sign the player. But soon, the team realizes that the player is not a promising prospect, resulting in a waste of funds on this rookie. False negatives occur when the model fails to predict a successful player, resulting in the team not signing the rookie. From our perspective, false positives are more costly than false negatives; therefore, a model with the highest false positive rate is not ideal. Therefore, we recommend the gradient boosting model, which exhibits the lowest false positive rate among all models, despite a slight decrease in accuracy.

Regarding the variable importance, all models have identified `gp` as the most significant predictor in forecasting the rookie's career duration beyond four years. Furthermore, the random forest and boosting models highlight `fg` as a significant predictor.

B. Limitations

As mentioned earlier, our binary response variable exhibits a slight imbalance with a 65% to 35% split. This factor could be the reason why all models demonstrate a higher false positive rate than the false negative rate, limiting our capability to interpret the results accurately. Additionally, our computational resources restrict us from performing a broader grid search on our random forests and gradient boosting models, which could have led to a better-performing iteration of the model.

C. Future Work

For future work, we propose to incorporate stratified sampling, which is a technique that guarantees the acquisition of a representative sample of a population by selecting a subset of data containing a proportional representation of different classes as the entire population. Stratified sampling is particularly useful when dealing with imbalanced datasets, as it addresses the problem of the slightly unbalanced response variable. Additionally, we would perform feature selection on our logistic regression model by eliminating variables with little significance, which could enhance our model's performance. Lastly, we would raise our classification threshold, given that we consider false positives more costly than false negatives. This adjustment would reduce our false positive rate, improve interpretability, and further aid in obtaining more accurate results.

References

[1] "Official NBA Stats," NBA Stats. https://www.nba.com/stats.