Daniel Aybar CS225 11/30/22

**NBA Player Salary Prediction**

**Introduction:**

American professional sports are becoming more lucrative than ever thought possible. In 2019, the last year unaffected by the COVID-19 pandemic, the National Basketball Association (NBA) saw total revenue climb to $8.8 billion with an average growth rate of 8% since 2010 [1]. Even with limited fan attendance and shortened seasons over the past 2 years due to COVID-19, the NBA has seen revenues of well over $6 billion annually [1]. In 2022, with a full season of games and ticket sales, the NBA could be in for its biggest single year growth in its history. The driving factor for this growth is the players. The performance of the players determines both the marketability and quality of each team. Unlike other sports, such as baseball and English football, NBA teams are restricted in how much they can spend on their players’ salary in a given year. Therefore, it is essential that players are paid what they are worth.

While the NBA has grown significantly in revenue, their ability to evaluate players has seemed to stay constant over the years. There are countless times where players are given salaries that do not match their production on the basketball court. With this being said, some individual teams have seen improvements in player scouting and development pay big dividends. The Golden State Warriors were valued at around $315 million in 2010. Since then, they have went on to increase their team valuation to over $5.6 billion, win 4 NBA championships, and move into a new $1.4 billion stadium [2], [3]. Their success comes from finding players such as Stephen Curry, Klay Thompson, and Draymond Green, who were all signed at a bargain price for the majority of their best playing years. These players not only contribute to winning games, but play a big role in sales and broadcasting revenues.

This project will use Naive Bayes classification to predict NBA player salaries using performance statistics from a player’s previous contract period. The data contains player statistics and salaries for 2,980 contract periods from 1995 to 2017.

**Methodology:**

The project was made up of 4 phases: 1) Data Collection; 2) Feature Selection; 3) Model Creation; 4) Model Testing and Evaluation

***Data Collection:***

NBA player statistics and salary data was obtained from separate Kaggle datasets [4], [5]. Missing data was filled in using Basketball Reference [6] and Hoops Hype [7]. Contract signing periods were found using Pro Sports Transactions [8]. Players in their last contract period (i.e retiring players) were not included in this data set, as they did not have a new contract salary to predict. Players who received a new contract after 2017 were not included, as there was no new contract data in the data set.

***Feature Selection:***

The original data set contained over 70 features, some of which were non-descriptive, such as name, college, and year. The target feature, Next\_cont, was measured using a player’s dollar salary as a percentage of total team salary cap. This was used to adjust for both dollar-value inflation, and the growth of salaries due to increased revenues across the league. A categorical salary variable (sal\_type) was then created using 4 separate brackets of the Next\_cont variable:

first\_option: Next\_cont ≥ 0.25;

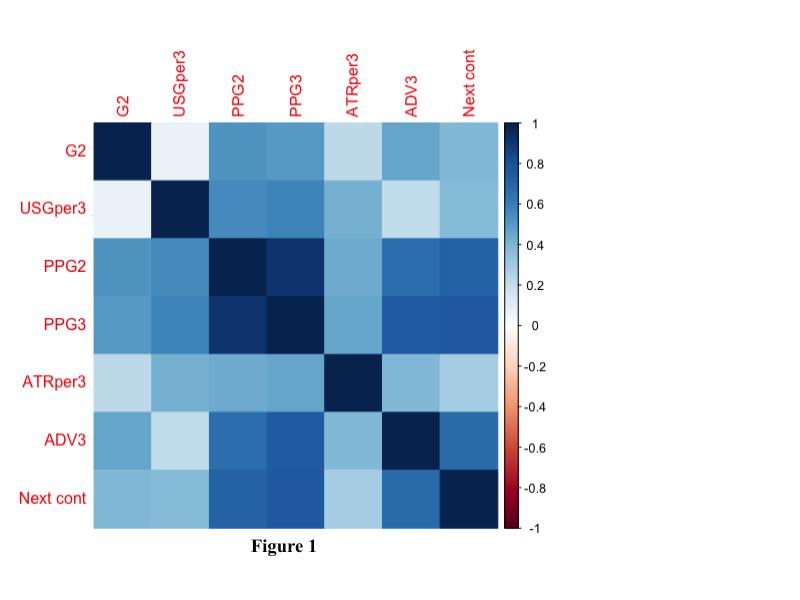
second\_option: 0.15 ≤Next\_cont< 0.25;

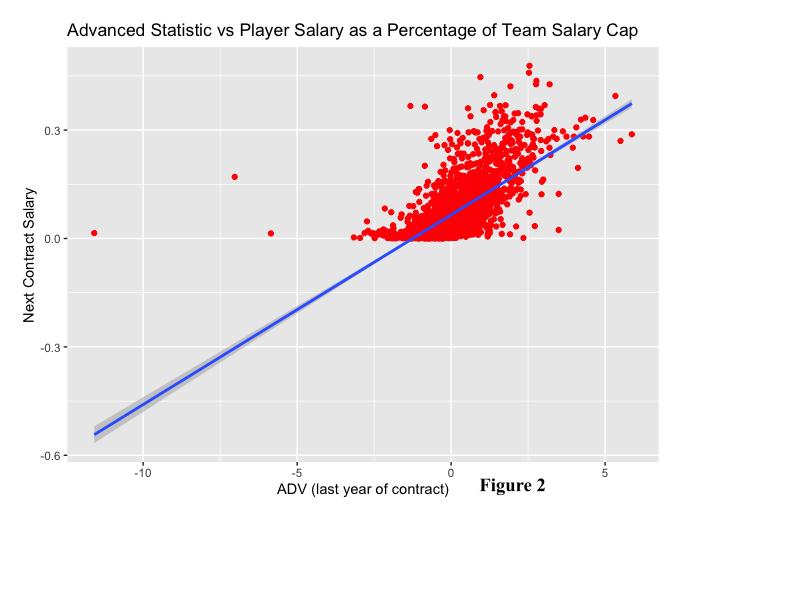
role\_player:0.05≤Next\_cont<0.15;

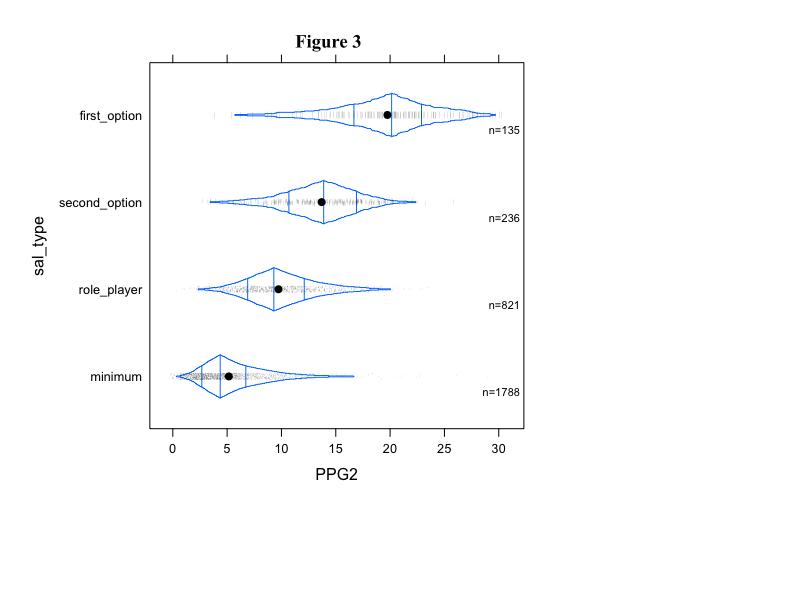
minimum: Next\_cont ≤ 0.05.

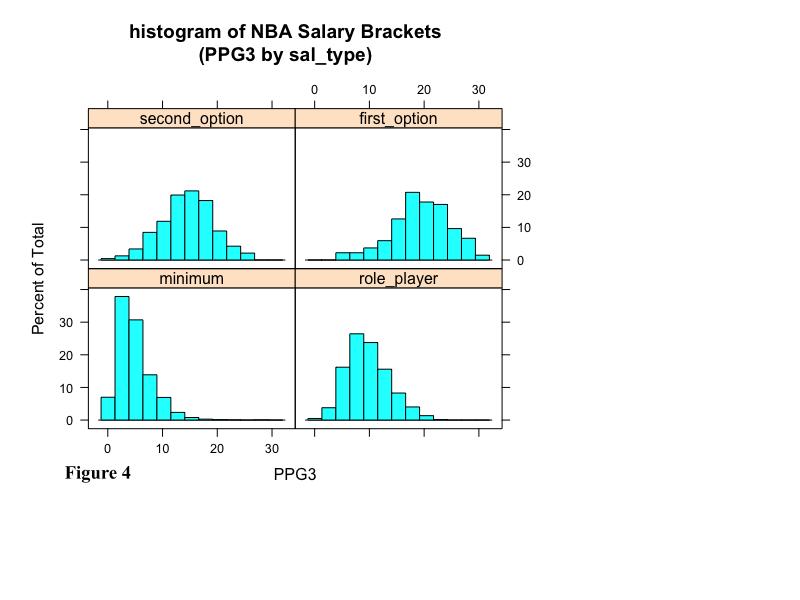
Four different types of feature selection were implemented to find the best set of descriptive features: 1) Forwards Greedy Search; 2) Backwards Greedy Search; 3) Hill Climbing Greedy Search; 4) Exhaustive Search. Although the exhaustive search is more computationally expensive, it is also a method that often leads to the best model. Because of this, the feature subset chosen by the exhaustive method was used for the model. The final set of descriptive features included ADV3, PPG2, PPG3, ATRper3, G2, and USGper3. Figure 1 shows a correlation plot for all of the variables used as well as the numeric salary variable (Next\_cont). ADV3 is a standardized combination of 4 commonly used advanced statistics (VORP, WS, BPM, and PER) for a players last year or their contract. A scatter plot showing the relationship between ADV3 and Next\_cont using a line of best fit is shown in Figure 2. PPG2 is the player’s points per game over the middle years of their contract, while PPG3 is a players points per game for the last year of their contract. Figure 3 shows a box-and-whiskers plot for PPG2 and the different salary brackets of sal\_type. Figure 4 depicts separate histograms for PPG3 and the four groups of sal\_type. ATRper3 is the ratio of a player’s assist percentage to their turnover percentage for the last year of their contract. USGper3 is the player’s usage percentage during the last year of their contract. Finally, G2 is the amount of games played during the middle years of the player’s contract.

Infinite values and missing values for ATRper were replaced with zeroes, while other missing values were replaced with variable medians. This is because infinite and missing values for the ATRper variable typically meant that the player never or extremely rarely turned the ball over, indicating that they most likely did not play enough to do so. Additionally, all players who made more than 50% of their teams salary cap were excluded from the data as outliers.



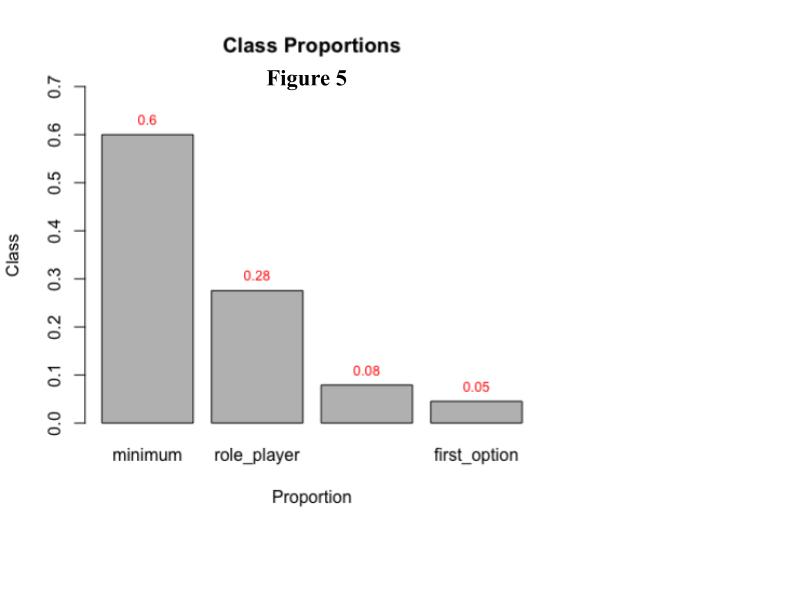






***Model Creation:***

5 separate models were created and assessed to try to predict either the sal\_type (nominal) or Next\_cont (numeric) variables. In the end, a Naive Bayes Classifying Model was selected. One of the main reasons this model was chosen was because it accounted for class imbalance. As seen in Figure 5, the sal\_type data is very skewed. Roughly 60% of the players in the data set were in the minimum bracket, while only around 5% of players were in the first\_option bracket. The model adjusts for this by establishing prior probabilities associated with each level of the target feature (sal\_type). This, in turn, makes the model aware of the class imbalance, and helps adjust the proportion of the class predictions to be more representative to the real world data.

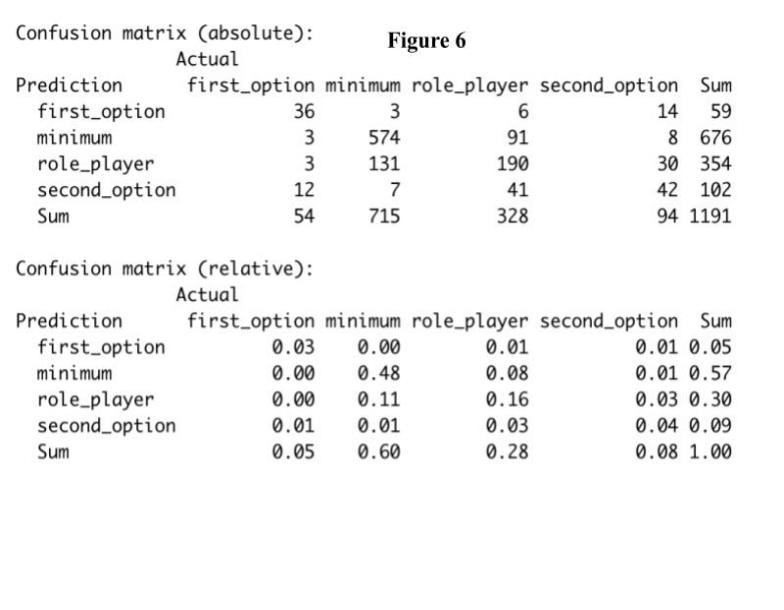


***Model Testing and Evaluation:***

When creating the model, 60% of the data was assigned as training data while the other 40% was assigned as testing data. As previously mentioned, the Naive Bayes uses prior probabilities to help address the class imbalance problem. In addition, it requires that the proportion of each level of the target feature is the same for both the training and testing data. This is to say that because roughly 60% of players are in the minimum classification, roughly 60% of both the training and testing data should also be in the minimum class. This is true for all levels of the target feature.

When predicting the testing data, the model returned an overall accuracy score of 70.7%. Figure 6 shows the resulting confusion matrix. The model returned a precision of 61% and a true positive rate of 66.67% for first\_option, which was the best of all the models. The accuracy of the first\_option predictions is the most important for teams, as it defines which players are the face of their respective franchises. The model also performed well when classifying players in the minimum class, with a precision of 84.91% and a true positive rate of 80.28%.

Although the model did relatively poorly at classifying role\_players (53.67% precision and 57.92% true positive rate) and second\_option players (41.17% precision and 44.68% true positive rate), this is less worrying for a number of reasons. Firstly, these players are less important for the organizations, both for winning and advertising. Secondly, the difference in performance between these two groups is not as starkly different as the first\_option and minimum groups. The first\_option players tend to be in the first\_option category because they are significantly better players than everyone else. On the other hand, minimum players tend to be in their category because they are significantly worse than contributing players. As for the role\_player and second\_option players, not much is separating them. Lastly, many of these players are close to the cutoff regions for their classes. This makes it increasingly difficult to separate players who may be in separate classes, but close together in terms of Next\_cont. In theory, if this is the case, the model should generally misclassify players into their neighboring groups. For example, if a player was in the top level of the second\_option group, it would make sense that the model classified them as first\_option. It would not make sense, though, if the model classified them as minimum. Thus there should be only be a small number of cases where players are incorrectly classified into a non-neighboring group. This tends to be the case. Of the 164 misclassified players in the role\_player group, only 3 were placed in the non-neighboring group (first\_option). For second\_option, 7 of the 60 misclassified players were put in the non-neighboring group (minimum). This implies that even when the model is incorrect, it is generally close to the actual results.



**Conclusion:**

The Naive Bayes model is relatively successful at classifying NBA players into salary brackets based on their playing statistics. The overall accuracy rate was 70.7%, the estimated reduced error was 26.68%, and the p-value was < 0.1 x 10^-10. This implies that the model does a significantly better job than guessing.

In general, though, there is still a lot of unaccounted error, which is difficult to encapsulate. As previously mentioned in the Model Testing and Evaluationsection, the model could be under-performing due to instances that are close to cutoff values. Another potential problem is that the model does not factor for player position. The roles of players among various positions are significantly different in basketball. It would be interesting to subset the data by position and compare the results of each ensuing model. One problem with that potential approach is that it significantly reduces the number of instances in each model. This can be especially detrimental when considering the class imbalance problem.

Another potential problem with the model is that it does not account for factors outside of player performance such as popularity and injury frequency. This paper initially tried to account for popularity by attempting to use jersey sales as a potential proxy variable, however, there was not any available data. As for injuries, it is very difficult to find a player’s injury potential. Even as modern medicine continues to improve and pick up on potential health problems in athletes, there will always be a random element in sports injuries.

In all, sports are very challenging to predict because there are many factors that are difficult to control for. For example, while this model attempts to correctly classify which salary group a player was in, there is no guarantee that the player deserved to be there. In addition, a player’s situation could drastically change their potential output. A player could be performing poorly because they are struggling with personal issues, teammate and coach issues, or because they do not have equal opportunity to succeed. While other fields can be relatively black and white, there is no guarantee in sports. Because of this, it becomes increasingly difficult to predict the value of a player. While data will help the process of evaluating players based on performance, there will always be factors that cannot be accounted for.

Sources

1. D. Curcic, “NBA Revenue Statistics (2001-2022),” *Athletic shoe reviews*, 01-Nov-1970. [Online]. Available: <https://runrepeat.com/nba-revenue-statistics.>

[2] C. Gough, “Golden State Warriors franchise value 2022,” *Statista*, 08-Dec-2022. [Online]. Available: <https://www.statista.com/statistics/194654/franchise-value-of-the-golden-state-warriors-of-the-nba-since-2006/.>

[3] A. Gaus, “Perspective: Warriors past and future: Chase center still a 'Great night out' for this Oracle Arena Fan,” *The San Francisco Standard*, 26-May-2022. [Online]. Available: <https://sfstandard.com/business/warriors-past-and-future-chase-center-still-a-great-night-out/.>

[4] O. Goldstein, “NBA players stats since 1950,” *Kaggle*, 27-Apr-2018. [Online]. Available: https://www.kaggle.com/datasets/drgilermo/nba-players-stats.

[5] F. Blanco, “NBA - player salary (1990-2017),” *Kaggle*, 06-Jul-2018. [Online]. Available: <https://www.kaggle.com/datasets/whitefero/nba-player-salary-19902017.>

[6] “Basketball Statistics & History of every Team & NBA and WNBA players,” *Basketball Reference*. [Online]. Available: <https://www.basketball-reference.com/.>

[7] “These are the salaries of all NBA players,” *HoopsHype*. [Online]. Available: <https://hoopshype.com/salaries/players/.>

[8] *Search Basketball Transactions*. [Online]. Available: <https://prosportstransactions.com/basketball/Search/Search.php.>