Rajkumar Kuppuswami

**Deliverable: Milestone 1 - Project Proposal**

**Date: 12/10/2020**

**Introduction**

The National Basketball Association (NBA) is a men’s professional basketball league comprised of 30 teams. Each of these teams have multi-million-dollar budgets that they use in order to build a talented team that will hopefully win enough games to bring home the NBA title. To do this, the teams build their rosters based on talent and budget. While team payrolls are large, they are not unlimited and strategic decisions must be made to develop the most talented team while maintaining budget constraints. The assumption of most teams, and players as well, is that the more money a particular player demands, the better the player. This is not always the case, however. There are numerous instances where players have been drastically overpaid and dramatically underperformed.1-2 Conversely, there have been players that have performed well above their higher-paid counter-parts.3 Ideally, there would be a mechanism where player salary could be determined by consistent player performance. It is reasonable to assume that NBA front offices do indeed subscribe to this premise, but some player salaries would suggest otherwise. The purpose behind this course project is to use predictive analytics to predict what a player’s salary should be by looking at player statistics.

*Background*

With ever increasing team payrolls, owners can usually afford to pay an underachieving player. The real issue is in determining the opportunity cost of this: What could you have gotten for that money instead? How else could you have improved your roster? It should come as no surprise then that teams are looking for alternative methods for determining appropriate pay levels for their athletes. This is where predictive analytics can hopefully shed some needed light on this very expensive problem.

*Problem Statement(s)*

What should an NBA player be paid based on his performance statistics? Which players are drastically overpaid? Which players are drastically underpaid?

*Scope*

For this project, we will be utilizing the NBA player statistics dataset that was obtained from Kaggle.4 The website lists the statistics as originating from 2017-2018 team rosters, but upon further verification, the rosters are more in line with 2016-2017 data.5 The downloaded data contained 88 missing salary values. All but two of these were obtained from HoopsHype.com5 website of player salaries.

The dataset contains 52 different variables. Not all of these are relevant to predicting player salaries and a detailed description of the variables used will be included in the technical approach section below.

*Document Overview*

This proposal will be broken down into five sections: Introduction, Preliminary Requirement, Technical Approach, Expected Results, and Management Approach. Many of these major sections will contain sub-sections that will show the tables, figures, and analysis that went into the predictive model.

**Preliminary Requirement**

We will start by handling missing data in the data set. Like mentioned earlier, our data has some missing salary values that can be sourced from different websites. There are also some features that having missing values that we will attempt to fill with a mean/median replacement. If a feature has too many missing values we may need to remove it from the data set.

We have some columns that are percentages of other columns (e.g., Free\_Throws, Free\_Throws\_Attempted, and Free\_Throws\_Pct). In these instances, we will only consider the Pct (percentage) columns in that they provide a better representation of player performance.

**Technical Approach**

*Analysis*

Our dataset contains 52 different variables that could contribute to our predictive model. We realize this is simply too many, so we have developed a process by which we will select the variables of interest. According to our text, *“Applied Predictive Analytics”*, many predictive algorithms assume the model variables follow a normal distribution. There are inherent advantages to using normally distributed variables, so our approach will focus on columns that closely follow this distribution.

We will determine which variable are normally distributed by conducting the following analyses:

* Summary statistics on all variables—concentrating on mean and standard deviation values
* Skewness of variables—looking for variables with values less than two and as close to zero as possible

*Requirement Development*

Our requirements stem from a thorough understanding of our stakeholder’s objectives in combination with industry data. The objectives were narrowed down to a singular, comprehensive goal which will drive the analysis process. The nature of the data was a key factor in developing our requirements as it alluded to the insights we expect to gain from our analysis. This is a recursive process since our requirements might change based on what we learn from the data. For example, new models might be required, or parts of our data might need to be excluded from the analysis, causing a shift in our approach.

*Model Deployment*

Model deployment will involve conducting a salary prediction on the full dataset using the model that was constructed from the TRAIN dataset. Results will be displayed graphically and conclusions will be based on statistical analysis of model performance.

*Testing and Evaluation*

Our prediction model development will begin with forming a TRAIN and TEST dataset that is split from our main data in a 70-30 TRAIN/TEST ratio. The model will be built with the TRAIN data. When complete, the model will be tested by conducting a prediction of player salaries. These predictions will be saved to the TRAIN data frame. The model will then be tested against our TEST dataset and subsequent salary predictions will be performed and saved.

The first step of our model evaluation will consist of making a scatter plot of actual salaries vs predicted salaries for both the TEST and TRAIN datasets. This will allow us to evaluate our model performance on predicting salaires based on the TEST dataset that the model has never seen. Finally, we can further evaluate the model performance by running it against the full dataset and then look at the summary statistics. Evaluation of its performance will be compared against R2 values, p-values, and significance of model variables.

**Expected Results**

We expect that our model will be able to clearly indicate which athletes are overpaid and by what margin. We anticipate that the number of overpaid players will be rather large. Furthermore, we expect our model to predict which players are underpaid as well. Due to the wide range of salaries in the dataset ($17,224 - $34,682,550), we may need to implement two separate models. One for overpaid athletes and one for underpaid.

**Management Approach**

*Project Plan*

The plan for this effort is to equally divide up the work between the three members of our team. We will have members working on Python code, R code, and document writing. These will not be divided where one person strictly does R code, Python code, and writing. Each team member will have opportunities to do each of the three tasks.

*Project Risk*

The major risk here is that we have too many variables in our dataset and we definitely need to parse those down into a manageable number. Therefore, we need to be sure we accurately pick variables for our model(s) that will return reasonable salary rates. This will likely be a trial and error approach and will be time consuming. The risk we run here is that we either get too bogged down with running too many iterations trying to choose the best variables, or we end up choosing the wrong ones that do not accurately predict player salaries.

1Gleeson, S. (2019, June 19). NBA draft: Ranking the five biggest busts as the No. 1 pick. Retrieved from https://www.usatoday.com/story/sports/nba/draft/2019/06/19/nba-draft-ranking-five-biggest-busts-no-1-pick/1488407001/.

2Davis, S. (2019, June 19). WHERE ARE THEY NOW? The biggest NBA Draft busts of all time. Retrieved from https://www.businessinsider.com/nba-draft-busts-2015-6.

3Ye, D. (2019, July 4). Top 10 Most Underrated NBA Players of the 2018-19 Season. Retrieved from https://howtheyplay.com/team-sports/Top-10-Most-Underrated-NBA-Players-of-the-2018-19-Season.

4Casalan, A. (2018, October 20). NBA player stats 2017-18. Retrieved from https://www.kaggle.com/acasalan/nba-player-stats-201718.

5These are the 2016/17 salaries of all NBA teams. (n.d.). Retrieved from https://hoopshype.com/salaries/2016-2017/.

6Abbott, D. (2014). *Applied predictive analytics: principles and techniques for the professional data analyst*. Indianapolis, IN: Wiley.