Machine Learning Engineer Nanodegree

Capstone Proposal

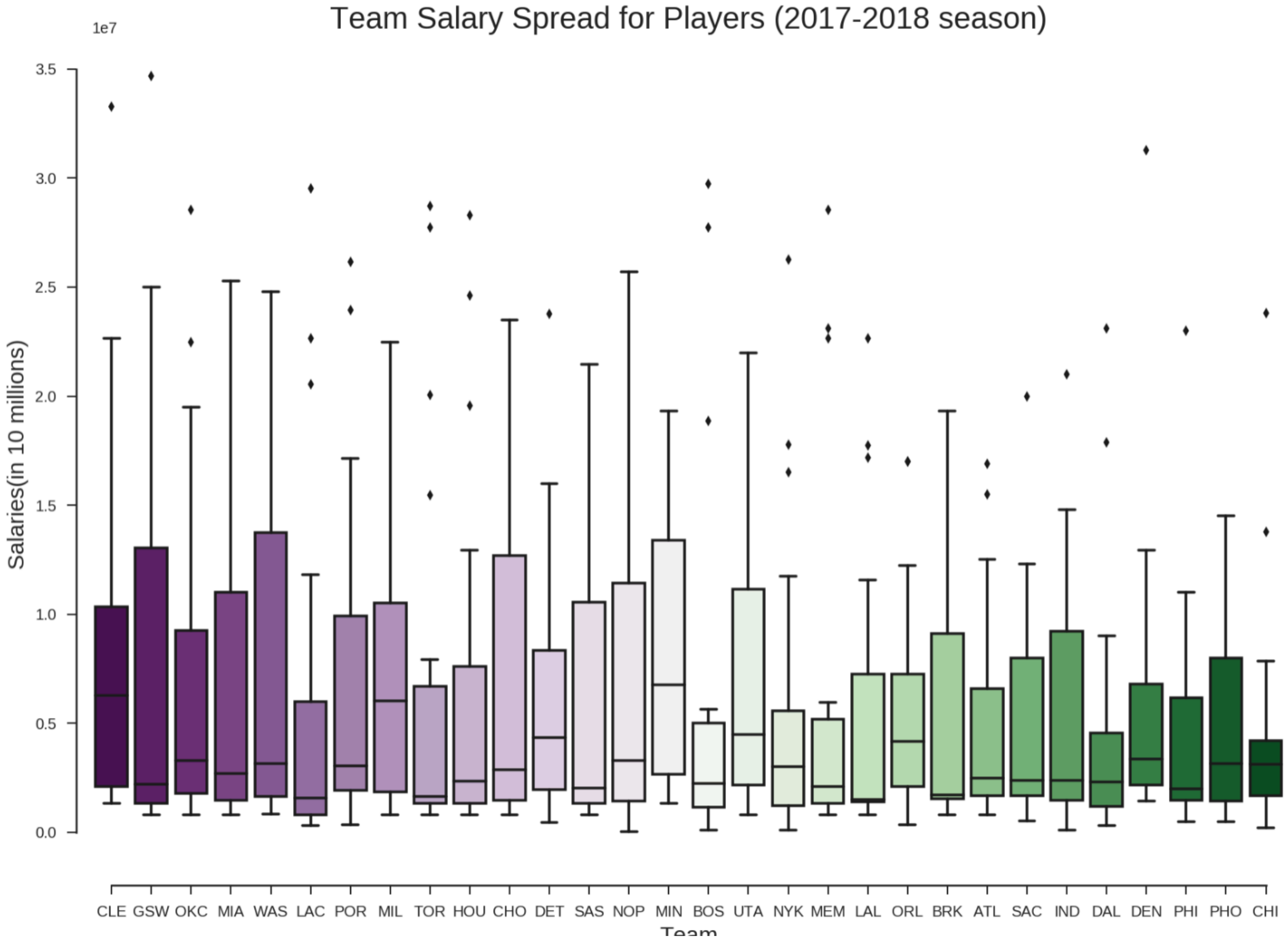
Michael Kim
December 2nd, 2018

Proposal

Domain Background

The National Basketball Association (NBA) is widely considered to be the premier men's professional basketball league in the world and features some of the most well-known athletes. As is such, NBA players are the world's best paid athletes by average annual salary per player. And their increasing salaries have led to salary caps which limit teams' total salaries. This limit is subject to a complex system of rules and exceptions; therefore this is considered a "soft" cap.

The escalation of NBA player salaries has not only been explosive, but it has also created a large earnings gap between players. The box plot below shows the spread of teams' salaries and the spread per player (teams with the largest total salary on the left and lowest on right). As seen below, there are many outliers well beyond the median (as represented by the ticks). For example, the Golden State Warriors are paying Stephen Curry a whopping 34,682,550 (the highest tick mark) this season which is greatly higher than Warrior's teammate Jason Thompson at just 945,126.



Machine Learning Engineer Nanodegree

Problem Statement

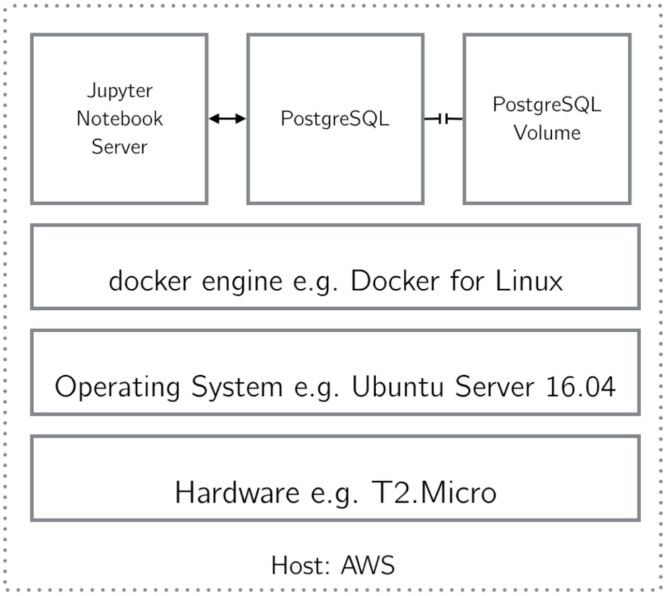
The issue of salaries has been a huge topic of controversy in the NBA and has led to several lockouts over these labor disputes. There were NBA lockouts in the 1995, 1996, 1998 and 2011 NBA seasons and the main issue in each of them was related to players' salaries. Owners felt players were overpaid, and players felt as if their earning power was restricted. And the topic of NBA salaries continues to be a hot topic of discussion today.

Increasingly, NBA players have come under intense scrutiny for their large salaries. Many questions have been asked, including why they are paid what they are and how are salary decisions made. Although these questions are beyond my scope as I don't know anything of negotiating contracts and deals with professional athletes. The topic did peak my interest and I wanted to build out a machine learning model using quantitative data to predict these large salaries.

Datasets and Inputs

Unlike Glassdoor or Indeed, NBA salary information is available online, as well as player statistics. Though there are many datasets different people acquired and put out on the web, none of them were substantial enough to develop a model on. Therefore, I am planning to scrape basketball sites in order to get the data for this project.

I will be using the BeautifulSoup Python library in order to scrape the data and load them into tables within a PostGres database. All the code for the web scraping will be included as a part of my project. In order to manage environments and data storage, I am going to engineer the entire setup using Docker. With setting up Docker containers, we run into the issue of data persistence as when the PostGres container goes down, that data within that container is lost. And so to solve this issue, I created a PostGres volume that houses all the data so that whenever our container goes down, the volume will hold all of it. And when the container comes back up, the volume will operate within that container. All relevant Docker and yml files will be included. A diagram of the environment that I will be using is below.



Machine Learning Engineer Nanodegree

The dependent variable for this study was NBA player salaries and the independent variables were the offensive and defensive statistical categories.

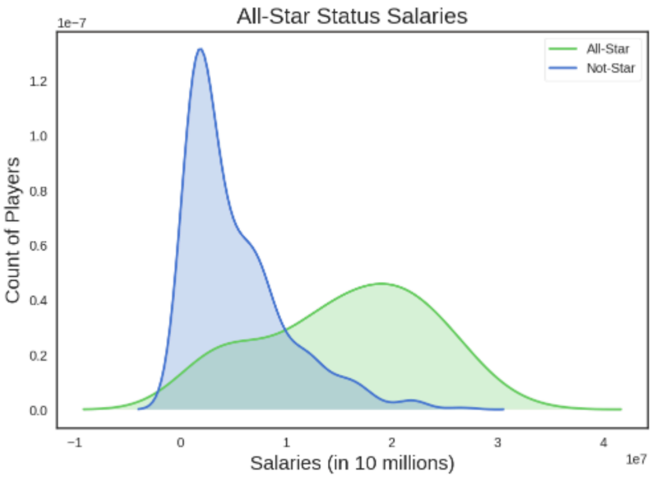
We are going to use the salaries and statistics of 486 NBA players from the most recent season. I decided to only use the statistics from the most recent season as they would be most reflective of the current salary rates. The salary cap for the NBA has been increasing at a rate faster than inflation and so it wouldn't make for a good model to bring in statistics from multiple years.

Solution Statement

In an effort to get an idea of what players are worth, judging strictly on quantative data, I am building a machine learning model to estimate player salaries based on their on-the-court performance.

But why use on-court performance statistics as our predictors? What makes them good variables to use? Aside from the inherent assumption that better performance leads to higher pay, what hard evidence from the data would suggest this? I didn’t want to build out a model based on assumptions, so in my exploration of the data, I tried finding a variable in the data that would encompass all the performance statistics. So I looked at the All-Star status field - this is a binary indicator of whether a player is an NBA all star or not.

NBA all stars are selected based on their high performance on the court. Below is a histogram that shows the salary distributions based on whether a player was an all-star or not. A player's all-star status is usually attributed to their performance on the court. There are other factors as well but by in large, it's based on how well they are performing. So we could see that there is probably a relationship between on-court performance and salaries.

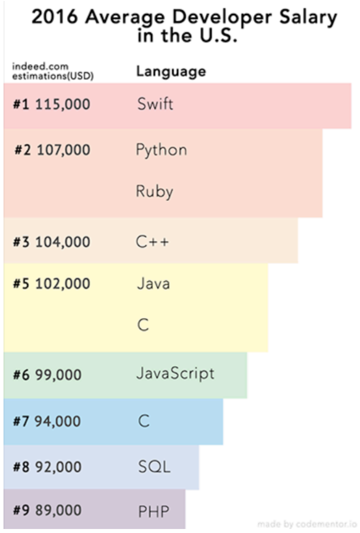


The purpose of this project is to help settle disputes regarding NBA players' salaries and to identify the variables that are most likely to contribute to a player's salary. Unlike coaches who are mainly hired and paid based on a single metric (wins), players are hired and paid based on individual performance, which can be measured by their on-the-court metrics. Though there is a lot of literature regarding what determines NBA salaries, there is very little statistical backing to their claims.

Benchmark Model

The development of this project is based on the hypothesis that a player's performance variables such as points per game, field goals, etc. would be significant contributors to player salaries. The dependent variable for this project was NBA player salaries and the independent variables were the offensive and defensive statistical categories.

In the rest of the job market, one's skill-set and the amount of experience they have using those skills determines salaries and with that information, applications like Glassdoor and Indeed can accurately predict how much a person would make for any given job. The plot below was determined with such information by Indeed.



Likewise with the NBA, I believe there are certain performance determinants that contribute more to higher salaries. Perhaps teams put greater premium on points per game than they do on total rebounds per game. These kind of questions are what I hope to explore and answer. The benchmark model would then be a very basic model where all the independent variables all had equal weight in determining salary. I am forecasting that such a benchmark model will perform poorly in predicting salary as there are probably certain statistical measures that teams are looking for.

Evaluation Metrics

I will be using supervised learning machine learning models to develop the predictive model for salaries. To assess the accuracy of the models that I develop, I will be evaluating the R2 value, mean absolute error and mean squared error. I will also look at coefficients and feature importance in order to understand which independent variables were better predictors of salary.

I will also be assessing P-values to understand which statistical features were relevant to the data. In this case, the null hypothesis would be that there is no relationship between the particular feature and the target variable (salary). So the null hypothesis would assume that on-court-performence statistics are independent of salary. Therefore, if any particular feature has a low p-value, specifically lower than 0.05, then we say that particular feature is statistically significant and we reject the null hypothesis and conclude that there is a relationship between that feature and the target.

Project Design

This project will comprise of five parts: web scraping, data cleaning, exploratory data analysis, supervised learning algorithm development, and unsupervised learning analysis. The unsupervised learning analysis will be something extra that I want to do to see if I could develop a clustering algorithm that would re-define the conventional 5 positions in basketball using basketball statistics.

Tools that are being used for this project are: Python, BeautifulSoup, RegEx, Pandas, Numpy, PostGres, Seaborn, Matplotlib, Docker, AWS, Machine Learning algorithms (linear models, ensemble decision trees, neural networks, k-means and principal component analysis).