# **NBA Salaries Prediction**

Jonathan Hicks, Fernando Flores, Mays Mansoor, Walaa Alani

**5/14/2021**

**Background**

When it comes to the topic of basketball and the NBA, one of the most commonly heard questions about players is how much money they make. Being able to predict a players NBA salary can never be completed with 100% accuracy, with some factors like previous injuries, how common they are injured, whether they are a star, etc. due to the fact that not all players are consistent and many teams do not always make the most sound decision when changing their roster for the season. Teams will often negotiate to receive a star NBA player for a ton of money, then build a team around them with players who are valued less but coaches believe could work well with the star on the court based on personalities and chemistry. Of course, there are many teams that go for a star when they have a solid roster to begin with, and want to increase their chances of securing a ring. Based on the main statistics that people see on sites like ESPN, and have the highest trendlines with salary, I created a model to predict a players salary. *The goal of the project was to predict a players salary using machine learning, and creating a web page to allow users to input a player's stats and receive a predicted salary based on the model.*

**Variables for Prediction**

* PPG (Points per Game)
* MPG (Minutes Played per Game)
* APG (Assists per Game)
* SPG (Steals per Game)
* BPG (Blocks per Game)
* WS (Win Shares)
* Age

**Exploratory Analysis**

Using Tableau, some analysis was done on the data for the 2017-2018 NBA Season Statistics in an attempt to find relationships within the dataset. Some findings based on the analysis was that the Average Salary for an NBA player in the U.S. is not even close to being at the top of the list when compared to some other countries. This is due to the fact that a large proportion of players in the NBA are born in the U.S, with only few of them actually being viewed as a star when they are drafted. Other analysis that were used before creating the machine learning model was trendlines for columns in the dataset to see which factors played the biggest role in determining players salaries, with each of the variables showing a strong trend, aside from Age. FG also showed a strong trendline for salaries, however it decreased the accuracy of the model so it was dropped from the regression model. The tableau workbook acts as an interactive dashboard for people to view player and team stats in an efficient manner.

**Building the Model**

After doing some preprocessing on the dataset to drop null rows and convert a few column types in a jupyter notebook, the cleaned data was read into a new jupyter notebook file. The first step was taking a log of the salary column in order to end up with a better predictive model, creating a new column for these values which would become the y target of the multiple linear regression. Next, five new columns were made, being the first 5 variables as they were given (PTS, MP, AST, STL, BLK) and dividing them by the number of games played for each player. At this point, the X targets were set to the 7 variables listed above, and the y target set to Salary\_log.

Using sklearn, the data was split into train and test sets and then fitted to a linear regression to receive the salary prediction from the X\_ test data.

After fitting the model, matplotlib was used to construct a residual plot to show a visual representation of the model. After the scatter plot was constructed, sklearn was utilized to find the mean squared error and r-squared score of the model.

*MSE: 0.6326738743239875, R2: 0.5523769265005738*

At this point, the model needed to be referenced in an app.py file to connect the user inputs from a web page to to predict a salary. In order to do this, pickle.dump was used to pass the model into a .pkl file that could be read into app.py.

The app.py file uses flask to set up two routes for the web page; the first being the loading page which renders the index.html on load, which has user input boxes for each of the 7 variables in the model, and the second being a prediction route. The predict route uses request.form.get for each of the user input values after the predict button is pressed, and then passes the values into a numpy array as float values. Once these inputs are passed into an array, model.predict passes them into the regression, and numpy.exp reverses the log for the Salary\_log column so the value given to the user outputs in the correct form. Finally, it rounds the value for the predicted salary and then returns the value to the user on the same web page under the predict button.

**Deployment**

The app is deployed with heroku, as it can be connected directly via github using flask in the app.py.

**Tools Used**

* Tableau
* Pandas, Numpy, os, Matplotlib, Pickle, Sklearn
* Python, Flask
* HTML/CSS

**Datasets Used**

* <https://www.kaggle.com/aishjun/nba-salaries-prediction-in-20172018-season>
* <https://www.kaggle.com/drgilermo/nba-players-stats> (to extract the pts column)