Using Machine Learning to Correlate NBA Player Performance to Overall Value

John Gomes
500754885
Ryerson University
CPS 803-011
Toronto, Canada
john.nathaniel.gomes@ryerson.ca

Rayaan Meeran
500749720
Ryerson University
CPS 803-011
Toronto, Canada
rmeeran@ryerson.ca

Abstract

The value of an NBA player can be indicated by their salary. Many of these players, however, do not earn a salary that represents their performance. Our goal was to develop a method to determine if an NBA player's performance matches up to their salary. To measure the performance of an NBA player, we used per game and advanced statistics. We chose statistics that best correlates with a player's individual performance and implemented them into regression models to predict the salary that player in the 2019-2020 NBA season should be earning. Regression models used to predict NBA player salaries were Linear Regression, Support Vector Machines, Logistic Regression, and Naive Bayes. With our predicted salaries completed, we calculated the difference in value between a player's actual salary and predicted salary to determine their overvaluation or undervaluation. Finally, we used graph analysis and K-Means clustering to identify potential relationships between our findings and specific statistics. These relationships identify patterns that likely explain valuation of specific players.

1. Introduction

With the increasing popularity of the basketball, the NBA's overall profits are at an all-time high. In 2019, the average NBA team was evaluated to be worth $1.9 billion USD, an increase of 13% from 2018, according to [1]. These profit margins are not only beneficial to the team owners, but also to the players. The NBA salary cap for the 2019-2020 NBA season is projected to be $109 million USD and the luxury tax projected to be set at $132 million USD. This is up from the 2018-2019 NBA season with a salary cap set at $101.9 million USD and a luxury tax set at $123.7 million USD, as mentioned in [2]. This increase in salary cap provides teams with an increased budget to allocate to players, which in turn results in players receiving larger salaries.

Whether or not these salaries correlate with a player's on-court performance, however, is up for debate. Quite frequently, NBA players underperform and fail to live up to the production expected of them when they negotiated their salaries. Underperforming players can become a detriment to a team's salary cap for several years, as contracts are fully guaranteed in the NBA, meaning teams cannot renegotiate salaries once a player has signed. The opposite situation occurs quite frequently as well, with players over performing and going beyond the expectations placed on them when they signed their contract.

This overvaluation and undervaluation of players was something we chose to further investigate. We wanted to find a way to determine how much value a player has to their team. The best way of seeing a player's value is to see what their team is willing to pay them; in other words, their salary. Our goal was to develop a method that would analyze whether an NBA player's on-court performance correlates to their salary and value to a team.

2. Related Works

Machine learning is becoming more prominent in the modern NBA. Analyzing available data and statistics have become key to many teams' successes. Existing works related to machine learning and NBA salary prediction does exist but does not exactly solve the problem we have identified and desire to solve. The related works we found tries to identify what specific NBA players would make in relation to the salary cap as in [3] or tries to analyze what performance metric most impacts a player's salary as in [4]. This differs from what we are trying to accomplish as we want to determine what every player's salary would be in the 2019-2020 NBA season based on various provided statistics.

3. Methodology

To accomplish our goal of correlating NBA player performance to overall value, we will need to apply various machine learning techniques. First, we will need to create training and testing data. Our training and testing data will need to include several numerical player performance and player impact metrics in order to evaluate a player's value. We will also require a single numerical value to equate a player's value to, in this case it will be a player's salary. The testing data will have to be for the 2019-2020 NBA season exclusively as we want to test our method for the current NBA season. Once our training and testing data is formulated, we will require regression models in order to make a prediction for a particular player's salary. After predictions are made for all players, we will need additional machine learning techniques to analyze and identify potential patterns and relationships that exist with player salaries and performance.

3.1 Creating Training and Testing Data

In order to measure an NBA player's on-court performance, we will need NBA player statistics. The NBA tracks and records various statistics for players and teams throughout the league's history, that are accessible through websites such as Basketball Reference.

The specific collection of statistics we are interested in for our regression model are per game and advanced statistics. Per game statistics provides a non-inflated average of a player's performance using basic metrics, such as points per game. In addition, advanced statistics provides a more in-depth numerical evaluation of a player's on-court performance, such as through player efficiency rating. We decided against using statistics including player totals due to the variation in games played for players. We also decided against using per 36 and per 100 possession statistics due to their inflated nature.

For the player statistics required for the training data, we retrieved per game and advanced statistics from Basketball Reference for the 2016-2017 NBA season through [5]-[6], 2017-2018 NBA season through [7]-[8], and 2018-2019 NBA season through [9-10]. Our cutoff point was the 2016-2017 NBA season due to the unprecedented jump the NBA salary cap had from $70 million USD in the previous season to $94 million USD, as stated in [11]. This allowed teams to offer players much larger salaries compared to previous years. Using data prior to the 2016-2017 NBA season would likely underfit the predictions in our regression model due to the much lower salaries in years prior.

Since our goal was for the current season, we retrieved per game and advanced statistics for testing data from Basketball Reference for the 2019-2020 NBA season through [12]-[13]. All statistics from the 2019-2020 NBA season used in our testing data and regression models are accurate as of December 5th, 2019.

With the retrieved statistics, we placed and parsed them into a CSV file corresponding to its NBA season. We cleaned the data by removing columns of data that were not relevant to our experimentation, such as total minutes played. We created 4 CSV files for player statistics, each one relating to a season. A problem that we encountered was that our CSV files were saved in UTF-8 format, and some players names contained non-Unicode characters. An example is the player Ante Žižić. We had to manually change some players names to regular UTF-8 characters in each of the 4 CSV files.

Another part of our training and testing data required was player salary information. This would be the targets that our regression models would fit and predict. We retrieved player salary information for our training data for the 2016-2017 NBA season from [14], 2017-2018 NBA season from [15], and 2018-2019 NBA season from [16]. We also retrieved player salary information for our testing data for the 2019-2020 NBA season from [17]. For each season, we placed and parsed the salary information into a CSV file relating to its season, for a total of 4 CSV files. We cleaned the data by removing unnecessary inflation information.

With our training and testing data, we used Python and Jupyter Notebook to further modify and clean our data. We created a DataFrame that combined a season's player statistics CSV file and player salary CSV file. This was done by matching the player name in each of the files. A problem we encountered was that some players had their names stored differently in each of the files. An example of this is the player J.J. Barea in one file being stored as Jose Juan Barea in the other file. We bypassed this problem by removing these players from the DataFrame. Another problem was that the players traded mid-season appeared numerous times in the player statistics file. We bypassed this by including total statistics for a player only. Further cleaning was done to the data by converting player positions to numerical values (E.g. Point Guard converted to 1), implementing a function to remove accents from player names (E.g. Álex Abrines to Alex Abrines), and converting traditional dollar representations to integers (E.g. $1,000,000 to 1000000). We then combined the DataFrames for each season into a single DataFrame for training and single DataFrame for testing.

3.2 Deciding the Most Impactful Statistics

Once our training and testing data was finalized, we had to determine which player statistics we would include in our regression models. We excluded statistics that were heavily influenced by team performance, as we wanted to maximize a player's individual performance and impact. The majority of the statistics that decided upon including correlate with player performance. In other words, the higher the number a player has in a specific statistic, the more likely it is that they are a high value or high salary player. The statistics we used in our regression models from per game statistics are as follows:

- MP (Minutes Played) - Number of minutes played per game.
- PTS (Points) - Number of points scored per game.
- FG (Field Goals) - Number of 2-point and 3-point field goals made per game.
- FGA (Field Goals Attempted) - Number of 2-point and 3-point field goals attempted per game.
- 3P (3 Point Field Goals) - Number of 3-point field goals made per game.
- 3PA (3 Point Field Goals Attempted) - Number of 3-point field goals attempted per game.
- FT (Free Throws) - Number of free throws made per game.
- FTA (Free Throws Attempted) - Number of free throws attempted per game.

- TRB (Total Rebounds) - Number of defensive and offensive rebounds per game.
- AST (Assists) - Number of assists per game.
- STL (Steals) - Number of steals per game.
- BLK (Blocks) - Number of blocks per game.
- TOV (Turnovers) - Number of turnovers per game.
- PF (Personal Fouls) - Number of personal fouls per game.

The statistics we used in our regression models from advanced statistics are as follows:

- PER (Player Efficiency Rating) – According to [18], the sum of a player's positive accomplishments subtracted by negative, returning a per-minute rating of a player's performance. The formula includes previously seen statistics computed into a single number.
- TS% (True Shooting Percentage) - According to [18], a measure of shooting efficiency that incorporates a player's 2-point field goals, 3-point field goals, and free throws. The formula divides points scored per game by 2 times true shooting attempts per game according to.
- USG% (Usage Percentage) - According to [18], an estimate of the percentage of team plays used by a player while they are on the court. The formula includes previously seen statistics computed into a single percentage.
- BPM (Box Plus/Minus) – According to [18], a box score estimation of points per 100 possessions a player contributed above an NBA-average player, translated to an NBA-average team. The formula includes previously seen statistics computed into a single number.
- VORP (Value Over Replacement Player) - According to [18], a box score estimation of the points per 100 team possessions that a player contributed above a replacement player, translated to an NBA-average team and an 82-game season. A replacement player is valuated at -2.0.

3.3 Regression Models

With the most impactful statistics chosen from our training and testing data, we were to incorporate them into our various regression models. Our plan was to use several regression models to get multiple predictions for a player's salary. We would then use ensemble learning to get the best possible prediction for a player's salary based upon the multiple predictions we generated through our regression models. The regression models to be used for ensemble learning were Linear Regression, Support Vector Machine, Logistic Regression, Naive Bayes, Decision Trees, and Random Forest. With repeated trials and testing, we were able to fine tune the parameters for each regression model into what we felt produced the most accurate and viable results.

3.3.1 Linear Regression

Linear Regression was implemented by importing it from [19]. The parameters used in this regression model were the default parameters. We trained our Linear Regression model by fitting it with our training data. We then used the model to predict player salaries for the current season with our testing data. Once our predictions were made for each player, we placed them in a DataFrame that would contain all predictions from all regression models.

3.3.2 Support Vector Machine

Support Vector Machine was implemented by importing [20]. The parameters we changed in this regression model were the kernel to polynomial, max iteration to 1000, and gamma to automatic. The other parameters used were the default parameters. We trained our Support Vector Machine model by fitting it with our training data. We then used the model to predict player salaries for the current season with our testing data. Once our predictions were made for each player, we placed them in the previously created DataFrame.

3.3.3 Logistic Regression

Logistic Regression was implemented by importing [21]. The parameters we changed were the solver to saga and max iteration to 50. The other parameters used were the default parameters. We trained our Logistic Regression model by fitting it with our training data. We then used the model to predict player salaries for the current season with our testing data. Once our predictions were made for each player, we placed them in the previously created DataFrame.

3.3.4 Naive Bayes

Naive Bayes was implemented by [22]. The parameters we used in this regression model were the default parameters. We trained our Naive Bayes model by fitting it with our training data. We then used the model to predict player salaries for the current season with our testing data. Once our predictions were made for each player, we placed them in the previously created DataFrame.

3.3.5 Decision Trees

Decision Trees was implemented by importing [23]. The parameters we used in this regression model were the default parameters. We trained our Decision Trees model by fitting it with our training data. We then used the model to predict player salaries for the current season with our testing data. Once our predictions were made for each player, we placed them in the previously created DataFrame.

3.3.6 Random Forest

Random Forest was implemented by importing [24]. The parameters we changed were the number of estimators to

24. The other parameters used were the default parameters. We trained our Random Forest model by fitting it with our training data. We then used the model to predict player salaries for the current season with our testing data. Once our predictions were made for each player, we placed them in the previously created DataFrame.

3.4 Ensemble

Once all of the regression models made their predictions, we incorporated each of them to determine a final value for an individual player's salary. We decided against using Decision Trees and Random Forest due to the random nature of the models and the large variations in predictions we had from the other models. Additionally, we felt that Linear Regression and Logistic Regression both slightly underfit the predictions for player salaries, so we attached a weight of 0.85 0.75 to them respectively. To compensate for this change, we added a weight of 1.15 to Support Vector Machine and a weight of 1.25 to Naive Bayes, both of which we felt would help increase the predictions to a more respectable output. We calculated the final prediction for a player's salary in ensemble learning by multiplying the weights by the individual predictions, then adding up the sum of them, and then dividing it by 4. We then stored the final predictions into the DataFrame our regression model predictions. In addition, we stored the difference between the actual salary and predicted salary for each player in the DataFrame. This DataFrame was then converted to a CSV file.

4. Results and Discussion

With predictions made for each player's salary for the 2019-2020 NBA season, we can now begin to analyze the data and examine any relationships or patterns in the data. The data for player salaries that are presented and analyzed are accurate as of December 5th, 2019.

4.1 Results of Predictions

From our model, the 5 NBA players that should be receiving the highest salary in USD are the following:

1. James Harden Prediction = $27,882,409
2. Luka Doncic Prediction = $26,923,089
3. LeBron James Prediction = $26,624,015
4. Giannis Antetokounmpo Prediction = $25,330,620
5. Kawhi Leonard Prediction = $23,694,148

The order of these players that the model predicted is accurate as these are the players that had the best on-court production in terms of statistics that measure player performance.

From our model, the 5 NBA players that should be receiving the lowest salary in USD are the following:

1. Jordan Bone Prediction = ($209,016)
2. Justin Patton Prediction = ($145,731)
3. Dewan Hernandez Prediction = ($75,769)
4. Chimezie Metu Prediction = $22,747
5. Sekou Doumbouya Prediction = $65,910

The order of these players that the model predicted is likely accurate as these are the players that had the worst on-court production in terms of statistics that measure player performance. One source of error, however, is that some players were predicted to earn a negative salary. The NBA has a minimum salary system in place that determines the minimum salary a player can receive, making a negative salary impossible.

From our model, the 5 NBA players that are the most undervalued in USD are the following:

1. Luka Doncic Value Difference = $19,239,729
2. Domantas Sabonis Value Difference = $19,027,906
3. Jarrett Allen Value Difference = $18,278,377
4. Bam Adebayo Value Difference = $17,829,954
5. Pascal Siakam Value Difference = $16,255,294

The order of these players that the model predicted as most undervalued is accurate as the on-court production in terms of statistics that measure player performance of these players greatly outperform their current salaries.

From our model, the 5 NBA players that are the most overvalued in USD are the following:

1. Stephen Curry Value Difference = ($31,157,143)
2. Blake Griffin Value Difference = ($26,859,581)
3. Nicolas Batum Value Difference = ($22,489,153)
4. Mike Conley Value Difference = ($21,999,054)
5. Steven Adams Value Difference = ($21,598,833)

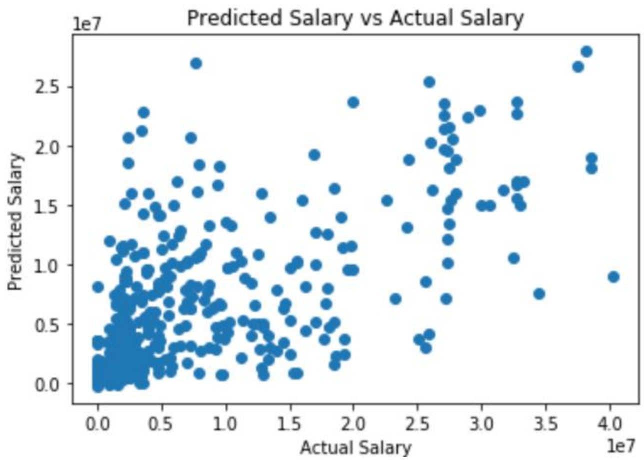
The order of these players that the model predicted as most overvalued is somewhat accurate as the on-court production in terms of statistics that measure player performance of these players greatly underperform their current salaries. A reason why some of these players, such as Stephen Curry and Blake Griffin, are listed is due to injuries. Our model fails to take into account player injuries and players underperforming because of it.

We also determined players that most accurately represent their salaries in terms of their performance statistics. From our model, the 5 NBA players that were most accurately predicted in USD were the following:

1. Cheick Diallo Value Difference = $7,853
2. Johnathan Motley Value Difference = ($99,05)
3. Yuta Watanabe Value Difference = $18,166
4. Jordan Poole Value Difference = ($23,966)
5. Jahlil Okafor Value Difference = $39,667

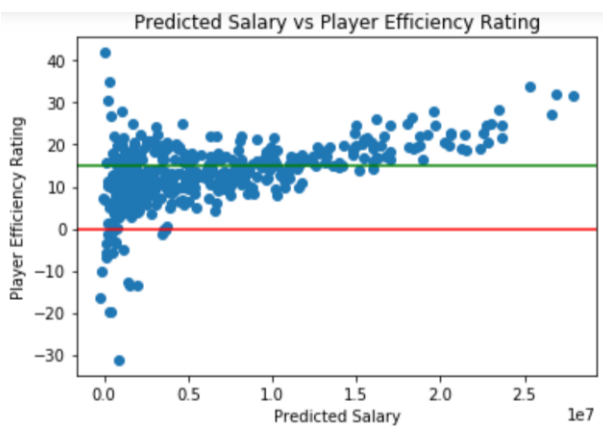
4.2 Key Graphs

With the results of our predictions, we were able to begin analyzing potential relationships and patterns that may exist within the results. The first graph visualized is predicted salary of players from our model against the actual salary of those players.



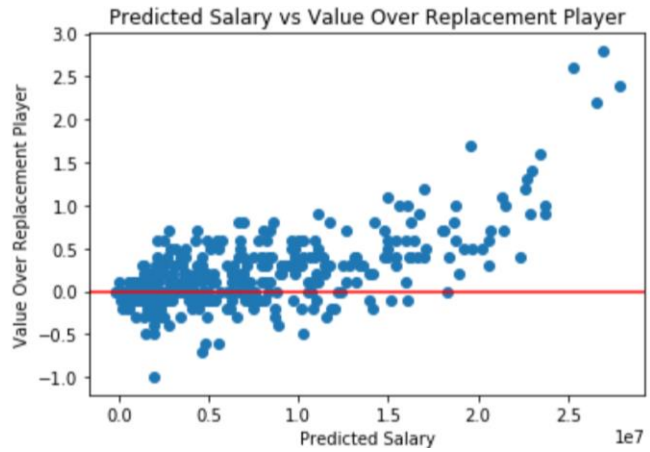
From this graph, several assumptions can be made. First, our model seems to underestimate the salary of some of the higher paid players in the NBA, as evidenced by their locations on the graph. Additionally, a large portion of NBA players are predicted to make $10,000,000 USD or less, similar to a large portion of actual player salaries.

Another relationship that we wanted to visualize was predicted salaries against specific player performance statistics. The statistics we planned on analyzing were catch-all statistics; statistics that evaluate a player's performance or specific aspect of their performance into a single number. This would help us visualize whether our model accurately predicted player salaries relative to performance. The first relationship is predicted salary against player efficiency rating (PER).



The green line indicates the NBA-average PER of 15. Some variance exists in the graph at predicted salaries less than $2,000,000 USD. An explanation for this is due to player's playing poorly in a small number of minutes played. Jordan Bone, the lowest paid player predicted by our model, has a low -16.5 PER in about 2 minutes played per game. Despite these outliers, our model is mostly accurate, as a majority of players predicted to earn more than $15,000,000 USD have an above NBA-average PER (i.e. PER above 15). This indicates that our model predicts that the majority of efficient players deserve high salaries.

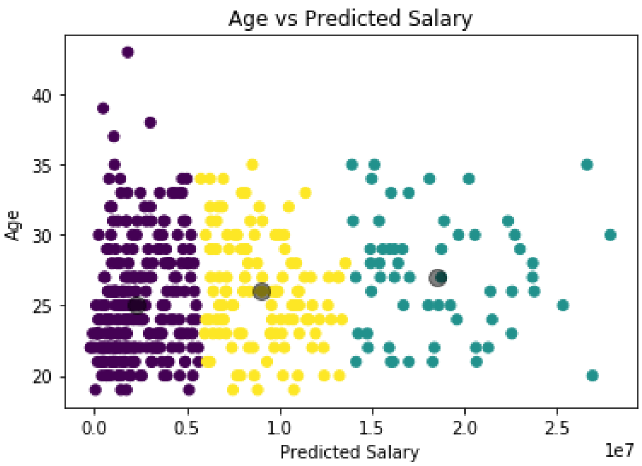
In addition to PER, we wanted to graph the relationship between predicted salary and value over replacement player (VORP).

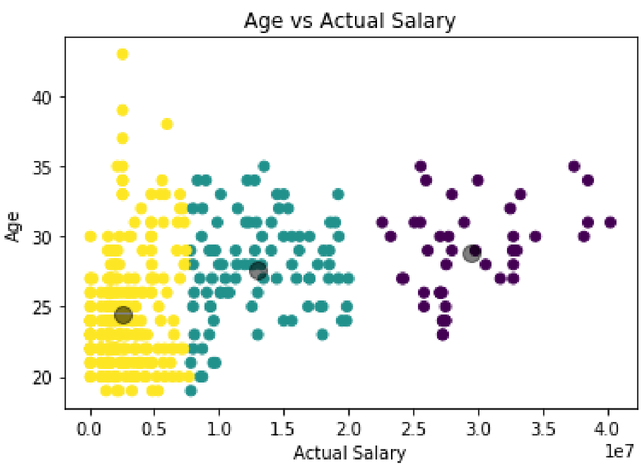


From this graph, our model can be determined as mostly accurate. Similar to the PER graph, a majority of players predicted to earn more than $15,000,000 USD have a positive value above an average replacement player. High earning players should in theory out-perform lower earning replacement players. The 4 evident outliers that exist in the graph are also the players predicted to earn the highest salary. These 4 players: James Harden, Luka Doncic, LeBron James, and Giannis Antetokounmpo have had arguably the best on-court production as of December 5th, 2019. This further supports the predictions that our model made.

4.3 Clustering

To get a better analysis and identification of potential relationships and patterns, we decided to cluster specific datasets from our results. This would help visualize specific groups of related players and possibly reasons for their predicted salary. To cluster the data, we implemented K-Means clustering by importing [25]. We chose 3 as the number of clusters for each dataset for players identified as overvalued, undervalued, or accurately valued. The datasets we clustered first were age against our models predicted salaries and also age against actual salaries.

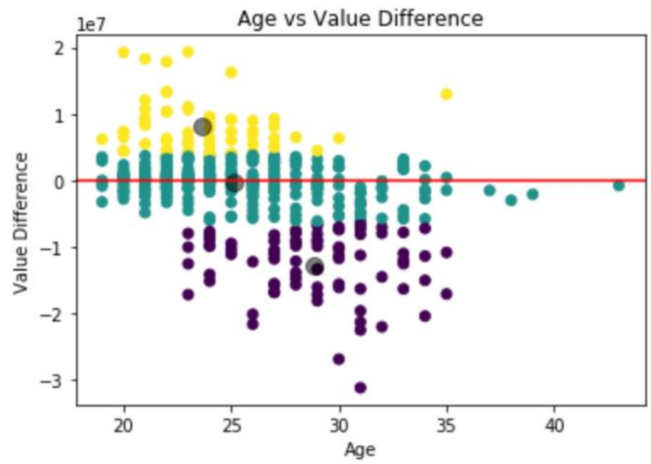




Age is of particular interest due to the way salaries are structured in the NBA. Younger players in the NBA, for example the 19-23 age bracket, are typically on rookie scale contracts, which is a much lower than average NBA salary. A rookie contract is determined by rookie salary scale and by a player's draft position, giving no room for players to negotiate, according to [26]. From the first graph, there are several players in the 19-23 bracket that exceed a predicted $10,000,000 USD salary, but in the second graph, only 3 total players exceed an actual salary of $10,000,000 USD. Those 3 players are also likely no longer on rookie contracts, making them exceptions. This discrepancy between the two graphs demonstrates how our model predicts many young players in the NBA to be undervalued.

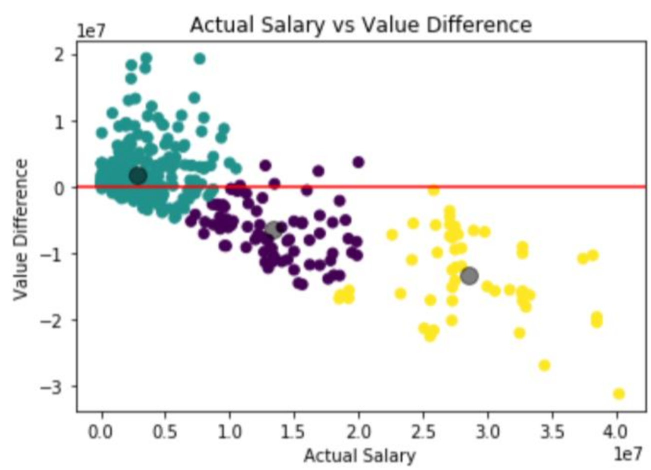
Additionally, in the second graph, there is a clear separation between players actually earning more than $20,000,000 USD and players below this threshold. In the first graph, our model has a much more consistent and compact prediction of player salaries. The predicted highest paid player in the first graph is James Harden at $27,882,409 USD and the actual highest paid player in the NBA is Stephen Curry at $40,231,758 USD. This discrepancy shows that our model predicts that the highest paid players in the NBA are actually overvalued. This discrepancy can also be attributed to the max contract in the NBA. Max contract salaries can be given to players with high on-court production and numerous accolades. Max contract salaries give players, primarily based on the number of years spent in the NBA among several other criteria, a large salary based on a percentage of the salary cap, according to [27]. The common age group for players to have max contract salary is the 24-34 age group.

To further support the observations made in the previous two graphs, we will cluster data on age and predicted player value difference.

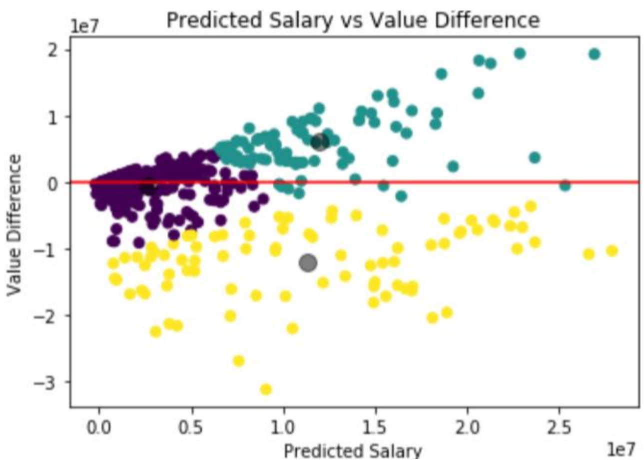


It is evident from this graph that a large portion of the most undervalued players in are younger than the age of 25, further supporting the argument that younger players are undervalued in the NBA. In addition, the most overvalued players in the NBA as visible in the graph are in the 26-34 age bracket, which is around the typical age that a player is eligible to earn a max contract salary. The large number of accurately valued players fall within $5,000,000 USD of the model's prediction and varies across all ages.

Another set of data that we had a particular interest in analyzing was the with the value difference predicted by our model. To see any potential relationships and patterns, we first clustered data between value difference predicted by our model and actual player salary.



This clustering was particularly interesting as it showed that the most undervalued players in the NBA predicted by our model are the players that are paid the least. Players that earn less than $10,000,000 USD made up a majority of the undervalued players. Additionally, players between the range of $10,000,000 USD and $20,000,000 USD are valued accurately or are slightly overvalued. Furthermore, this clustering identified that the most overvalued players in the NBA predicted by our model are players that are paid the most. Players that earn more than $20,000,000 USD made up a majority of the overvalued players. This can further support the argument of the overvaluation of max contract salaries, as all max contract players are listed above this threshold. We can also compare this clustering to the clustering between value difference and predicted player salary.



This clustering has a more scattered representation than the previous clustering and has a less obvious visualization of clusters. The group of undervalued players are dispersed and seem to increase as predicted salary increases. The larger differences for undervalued players are likely younger players on rookie scale contracts. In addition, the more overvalued players seem to be dispersed evenly as predicted salary increases, unlike the previous clustering.

4.4 Sources of Error and Difficulties

During our experimentation and analysis, we had several error sources and difficulties. A source of variance and noise in our predictions is due to the small sample size of the 2019-2020 NBA season. This also includes several players playing limited minutes and players underperforming or overperforming as of December 5th, 2019, that may regress to their typical averages further along the season. Our model also failed to take into account player injuries and players underperforming because of it.

Additionally, the statistics used in our regression models do not 100% represent a player's performance. We looked mainly at individual performance and not a team's performance to evaluate players. There are also aspects of a players' impact that statistics and traditional numbers just can't measure or quantify. This may be an explanation for some players being predicted to be overvalued.

Furthermore, the yearly inflation of the NBA salary cap likely affected the training of our regression models, leading to an overvaluation of some players. The salary cap increases every year, such as the salary cap increasing to $109 million USD for the 2019-2020 NBA season, up from $101.9 million USD the previous season, as mentioned in [2]. With a lower salary cap, players likely earned less money relative to their production in our training data than compared to our testing data.

We also had difficulties with the CSV files due to player names. The trouble was primarily with UTF-8 encoding due to some players having non-Unicode characters in their name. In addition, we had trouble with our 2 sources of data due to differences in player names in the sources. We ultimately had to leave out a small number of players in our training and testing data because of this issue.

5. Conclusion

Despite the errors and difficulties, we were able to accomplish our goal of developing a method to determine whether an NBA player's on-court performance correlates to their salary and value to a team. We managed to create viable training and testing data for use in several regression models. The combined usage of our models into ensemble learning successfully equated a player's performance statistics to a single numerical value representing their predicted salary.

With the results, we were able to observe what players our model predicted to have the highest salaries and lowest salaries. In addition, we were able to observe what players our model believed were the most overvalued, most undervalued, and most accurately paid relative to their performance statistics. This also led to several observations that correlated specific statistics used to train and test our models to the relative ordering of player salaries.

Clustering also led to several observations as to the reasons for differences in our model's predictions and actual player salaries. Age was a large factor impacting player salaries, specifically with young players and also high earning players. Young players in the NBA are restricted due to the nature of rookie scale contracts and the lower salary it provides. Older, high earning players are impacted by age due age's contribution to max contract eligibility. Additionally, our model determined that the most undervalued players in the NBA are the least paid players while the most overvalued players are the highest paid. This is once again likely due to the limitations of rookie scale contracts and the large salaries provided to max contract players. Final conclusions that can be drawn from our experimentation and clustering are that salaries should be distributed more evenly across the league, that rookie scale contract salaries should be increased, and that max contract salaries should be decreased.

References

[1] Forbes.com. (2019). *Forbes Releases 21st Annual NBA Team Valuations*. [online] Available at: https://www.forbes.com/sites/forbespr/2019/02/06/forbes-releases-21st-annual-nba-team-valuations/#11bef90511a7 [Accessed 15 Dec. 2019].

[2] Jones, K. (2018). *Report: NBA sets salary caps for 2019-20, 2020-21*. [online] Sports Illustrated. Available at: https://www.si.com/nba/2018/09/18/nba-109-million-salary-cap-2019-20-season-118-million-cap-2020-21 [Accessed 15 Dec. 2019].

[3] Rosson, J. (2019). *NBA Salary Predictions*. [online] Medium. Available at: https://towardsdatascience.com/nba-salary-predictions-4cd09931eb55 [Accessed 15 Dec. 2019].

[4] Xu, J. (2019). *Predicting NBA Player Salary With Data Science*. [online] Medium. Available at: https://medium.com/better-programming/predicting-nba-player-salary-with-data-science-c5702caa3f2e [Accessed 15 Dec. 2019].

[5] Basketball Reference (2017). *2016-17 NBA Player Stats: Per Game* [online] Available at: https://www.basketball-reference.com/leagues/NBA_2017_per_game.html [Accessed 5 Dec. 2019].

[6] Basketball Reference (2017). *2016-17 NBA Player Stats: Advanced* [online] Available at: https://www.basketball-reference.com/leagues/NBA_2017_advanced.html [Accessed 5 Dec. 2019].

[7] Basketball Reference (2018). *2017-18 NBA Player Stats: Per Game* [online] Available at: https://www.basketball-reference.com/leagues/NBA_2018_per_game.html [Accessed 5 Dec. 2019].

[8] Basketball Reference (2018). *2017-18 NBA Player Stats: Advanced* [online] Available at: https://www.basketball-reference.com/leagues/NBA_2018_advanced.html [Accessed 5 Dec. 2019].

[9] Basketball Reference (2019). *2018-19 NBA Player Stats: Per Game* [online] Available at: https://www.basketball-reference.com/leagues/NBA_2019_per_game.html [Accessed 5 Dec. 2019].

[10] Basketball Reference (2019). *2018-19 NBA Player Stats: Advanced* [online] Available at: https://www.basketball-reference.com/leagues/NBA_2019_advanced.html [Accessed 5 Dec. 2019].

[11] Davis, S. (2017). *A fateful decision by the NBA players' union 2 years ago is coming back to haunt the players*. [online] Business Insider. Available at: https://www.businessinsider.com/nba-free-agency-effect-nbpa-rejecting-cap-smoothing-2017-7 [Accessed 15 Dec. 2019].

[12] Basketball Reference (2019). *2019-20 NBA Player Stats: Per Game* [online] Available at: https://www.basketball-reference.com/leagues/NBA_2020_per_game.html [Accessed 5 Dec. 2019].

[13] Basketball Reference (2019). *2019-20 NBA Player Stats: Advanced* [online] Available at: https://www.basketball-reference.com/leagues/NBA_2020_advanced.html [Accessed 5 Dec. 2019].

[14] HoopsHype. (2017). *2016/17 NBA Player Salaries*. [online] Available at: https://hoopshype.com/salaries/players/2016-2017/ [Accessed 5 Dec. 2019].

[15] HoopsHype. (2018). *2017/18 NBA Player Salaries*. [online] Available at: https://hoopshype.com/salaries/players/2017-2018/ [Accessed 5 Dec. 2019].

[16] HoopsHype. (2019). *2018/19 NBA Player Salaries*. [online] Available at: https://hoopshype.com/salaries/players/2018-2019/ [Accessed 5 Dec. 2019].

[17] HoopsHype. (2019). *2019/20 NBA Player Salaries*. [online] Available at: https://hoopshype.com/salaries/players/ [Accessed 5 Dec. 2019].

[18] Basketball Reference. (n.d.). *Glossary*. [online] Available at: https://www.basketball-reference.com/about/glossary.html [Accessed 15 Dec. 2019].

[19] scikit-learn. (n.d.). *sklearn.linear_model.LinearRegression*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html [Accessed 15 Dec. 2019].

[20] scikit-learn. (n.d.). *sklearn.svm.SVC*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html [Accessed 15 Dec. 2019].

[21] scikit-learn. (n.d.). *sklearn.linear_model.LogisticRegression*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html [Accessed 15 Dec. 2019].

[22] scikit-learn. (n.d.). *sklearn.naive_bayes.GaussianNB*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html [Accessed 15 Dec. 2019].

[23] scikit-learn. (n.d.). *sklearn.tree.DecisionTreeRegressor*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html [Accessed 15 Dec. 2019].

[24] scikit-learn. (n.d.). *sklearn.ensemble.RandomForestRegressor*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html [Accessed 15 Dec. 2019].

[25] scikit-learn. (n.d.). *sklearn.cluster.KMeans*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html [Accessed 15 Dec. 2019].

[26] Jessop, A. (2012). *The Structure of NBA Rookie Contracts*. [online] Forbes.com. Available at: https://www.forbes.com/sites/aliciajessop/2012/06/28/the-structure-of-nba-rookie-contracts/#5275658c7299 [Accessed 15 Dec. 2019].

[27] Honkasalo, M. (2017). *HoopsHype explains: How do maximum contracts work?*. [online] HoopsHype. Available at: https://hoopshype.com/2017/05/13/hoopshype-explains-how-do-maximum-contracts-work/ [Accessed 15 Dec. 2019].