Metrics Impacting NBA Player Contract Values

Executive Summary

Background:

An NBA agent hypothetically hired our team to analyze the factors impacting contract values. The agent will earn a 5% commission for each contract; higher contract values result in higher commissions for the agent. Both the agent and the players stand to gain from a growing NBA market.

Objectives:

- A specific list of factors likely to increase or decrease contract values.
- Expected contract values of newly signed players based on college performance.
- A list of schools to focus recruitment efforts.

Data Sources:

Explain the sources

Data Manipulation Methods:

A blurb data cleaning steps

Analysis:

A blurb on insights that includes visuals

- Two levels (EDA – Descriptive)
- Regression coef blurb 1
- Regression coef blub 2
- Comparison of regression output

Limitations:

A blurb that explains limitations

Inspiration

To Help Players and Agents Capitalize On A Growing NBA Market

Background:

For this project, we have created a hypothetical scenario in which an agent of an NBA player management firm hired us to deliver a report on the leading factors impacting an NBA player's guaranteed contract value. The agent receives a 5% commission on each contract he negotiates; the higher the value, the more money he earns. The agent will use the analysis to advise players and prioritize performance improvement metrics.

Context:

Like most other sports, the NBA sets a salary cap that determines how much a team can spend on salaries. Since some franchises reside in more significant, popular cities, they generate more revenue than smaller market teams. The cap aims to keep the game competitive by limiting the amount of money big market teams can spend on their roster of players. The cap is partly determined by the league's total revenue, which has doubled in the last decade. As the NBA makes more money, the players and their respective agents stand to gain financially. Figure 1. on the right shows salary projections reaching more than 80 million annually by 2029.

Similar Studies:

An article titled "NBA Player Salary Analysis based on Multivariate Regression Analysis" was published by *Highlights in Science, Engineering and Technology*. The authors conducted a similar study but focused on ways to balance disproportionate salaries between the players. This study gave us an idea of what variables to investigate for our project. Additionally, we referred to a prior study from Koki Ando, who also conducted a regression analysis, but his goal was to predict a player's future salary; from that study, we got an idea of how to structure our experiment.

Objectives:

For this project we aim to deliver the following to the agent:

- A specific list of factors likely to increase or decrease contract values.
- Expected contract values of newly signed players based on college performance.
- A list of schools to focus recruitment efforts.

NBA Supermax Salary Projections

SEASON	SALARY CAP PROJECTION	MAX SALARY
2023-24	$136.021 mil	$47.6 mil
2024-25	$149.623 mil	$52.37 mil
2025-26	$164.585 mil	$57.6 mil
2026-27	$181.044 mil	$63.37 mil
2027-28	$199.148 mil	$69.7 mil
2028-29	$219.063 mil	$76.67 mil
2029-30	$240.969 mil	$84.34 mil

Figure 1. Table of supermax salary projections. *NBA salaries keep going up. Prepare to have your mind blown in the future* by Mike Vorkunov, 2023, The Athletic. https://theathletic.com/4740069/2023/08/03/nba-salary-cap-rise-jaylen-brown/

Name	Description & Important Variables	Size	Format	Links
Player Career Stats		419 KB		https://github.com/swar/nba_api/blob/master/docs/nba_api/stats/endpoints/playercareerstats.md
Player IDs		99 KB	Pandas DF	https://github.com/swar/nba_api/blob/master/docs/nba_api/stats/static/players.md
Draft History				https://github.com/swar/nba_api/blob/master/docs/nba_api/stats/endpoints/drafthistory.md
NBA Teams		312 B		https://github.com/swar/nba_api/blob/master/docs/nba_api/stats/static/teams.md
Contract Values		42 KB 476 rows		2023-24 NBA Player Contracts
College Stats		~12.4 MB		SR-CBB

Data Manipulation Methods

Clean and Prepare NBA Data for Regression Analysis

How did we need to manipulate the data:

The default file type for the player contract data is .xls, and this generates a value error because pandas cannot determine the file format. In this case, we opened the downloaded file, renamed it, and saved it as a .xlsx file. Additionally, we executed the following manipulations:

- Convert player names to lowercase characters.
- Replace accented characters with their English versions.
- Split the player column into first and last name columns.
- Rename columns to be more representative.
- Drop duplicate rows.

The player ID dataset needed the following manipulations:

- Convert first and last name columns to lowercase.
- Convert the player ID column to string datatype.
- Update specific player names to allow for joining.

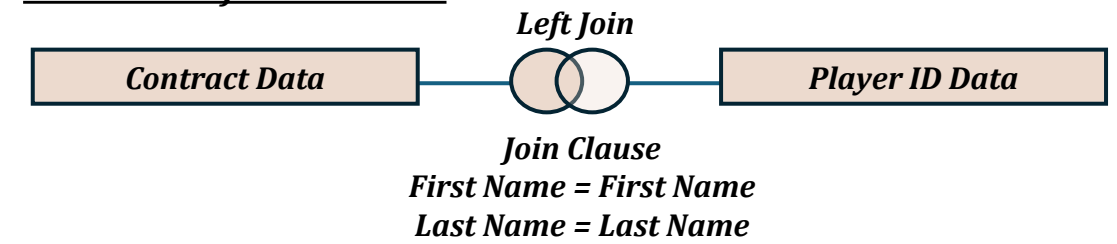
The career stats needed the following manipulations:

- Convert the season column to datetime truncated to year.
- Drop rows corresponding to the 2024 season.
- Limit career stats to the last five years.

How did we address data quality concerns and challenges:

The 2024 season is in progress and contains a partial season's metrics. We removed rows corresponding to the year 2024 from the dataset. Our dataset did contain null values because one player did not have a contract value, and several players are rookies competing in their first season this year; these players were dropped from the dataset because they did not fit the scope of our data goals. Several players needed name corrections because our data sources recorded names differently. The basketball reference website stores a player's name with accented characters, while the NBA_API stores the name without the accents. This was the most challenging data quality issue because it prevented the accurate joining of the two datasets. We addressed this using code that corrected specific names.

How did we join the data:



COLLEGE STATS MANIPULATION

Data Retrieval & Filtering

The goal of the associated notebook was to combine draft history variables with college statistics for each active player with an identified contract value who was drafted from a college/university. A pre-aggregated dataset was used as the primary source for players. To minimize the number of requests to the Sports Reference server, data needed to be retrieved in a particular sequence:

1.Extract team IDs from NBA API

2.Using team IDs, extract draft history for each NBA team and combine into one df

3.Keep only players in source list

4.The Draft History Endpoint imports all draft picks for each NBA team. Filter draft history to include only players who were drafted from a college/university.

5.Modify name values to match Sports Reference HTML links:

•Remove spaces between last name and suffix

•Remove apostrophes

•Manual replacement of name values that don't follow typical naming format

6.Remove players without college record (attended but did not play or drafted from international college)

7.Scrape player profiles from Sports Reference

Once the name values were consistent with HTMLs, we extracted the entire player profile from Sports Reference for each player:

•For each player, pull HTML file and store locally (to avoid re-sending requests)

•Open each HTML file, parse and extract the "Players Totals" table into a df for each player

•Concatenate all player dfs together

Data Cleaning & Processing

The Players Total table contains a row for each college season and a totals row (to represent total stat values across all seasons). The provided totals row is dropped for the dataset. To mimic the aggregation of NBA stats, the season totals were summed by player – each player represented by exactly one row. Additionally, we engineered two additional variables from college stats:

•"Season Count": Count of college seasons

•"Team Count": Count of unique college teams

Merging Data

Finally, the college stats df was joined with draft history variables and contract values on player name.

Barriers to Data Manipulation

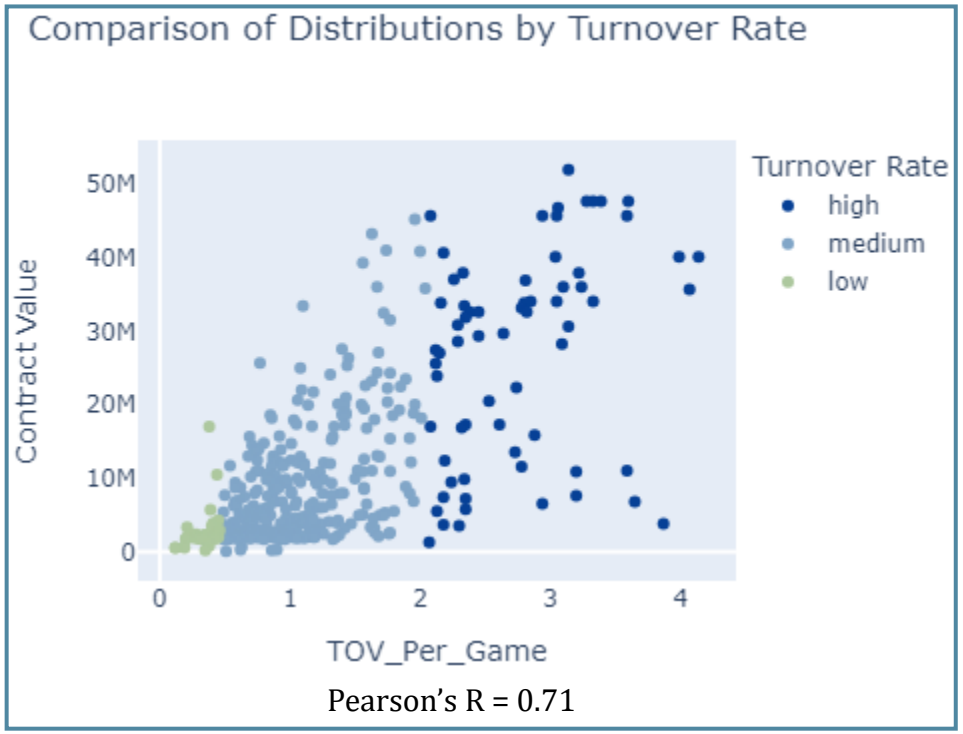
•The unique identifiers across dfs are not consistent (NBA API data contain unique player IDs while Sport Reference data are limited to players' names.

•Checking player name values against HTMLs was initially a very manual process – each time a request would fail, we would have to search the players name on the website and replace the value accordingly. Then we needed to re-run the request using the correct name value and manually delete the invalid HTML file.

•Even with proper HTMLs, the requests to the website server were extremely time-consuming – with execution times between 20 – 30 minutes per 15 players.

NBA Exploratory Analysis

Descriptive Statistics and First Insights



Pearson's R = 0.71

Interesting Insights and relationships:

- A moderately strong positive correlation exists between turnovers and a player's contract value. However, this case of correlation ***does not*** equal causation because turnovers hurt the game. When a player turns the ball over, his team loses an opportunity to score points. In the context of this dataset, players with higher contract values also have higher turnover rates because they usually have the ball more. It is ***not*** recommended that players increase their turnover rates.
- The NBA regular season consists of 82 games, and the dataset contains metrics from the last five years, not including 2024, so the maximum number of games one can play is 410. The median games played across the 419 players in our dataset is just 206 games or 2.5 seasons. Players making more than $10 million per year tend to play more than 250 games over a 5-year period.
- Steals per game have a weak positive correlation to contract values. Steals are obtained when a player assumes a defensive position and prevents the opposing player from scoring by taking the ball without fouling. Since this metric is associated with preventing the other team from scoring, it is surprising that players who steal the ball more do not have higher contract values.

What didn't work, and why?:

In similar studies, additional metrics such as age, position, and years of experience are used to conduct the analysis. Unfortunately, I needed to timebox the time I spent looking for data sources and did not find one with the above stats. Additionally, I could not experiment with different regression algorithms for similar timebox constraints. Given more time, I would assess advanced NBA metrics and try different regression algorithms.

NBA Regression Analysis

Inferential Insights

NBA Linear Regression Process:

1. Collect and pre-process data.
2. Conduct EDA to identify underlying trends.
3. Validate that the dataset meets the assumptions for regression analysis.
   - Independent variables are linear to dependent variable.
   - There is no multi-collinearity between independent variables.
   - The observations in the dataset are independent of each other.
4. Scale independent variables.
5. Fit the model.
6. Remove insignificant features.
7. Fit the final model.

Final Model and Performance

We removed three points made, offensive rebounds, defensive rebounds, and steals per game because they were statistically insignificant. The final model uses field goals, assists, blocks, and personal fouls per game to explain the variation in contract values. Next, we used R-square, which measures how well the independent variables explain the variability of the dependent variable in a regression model to assess how well the final model explains contract values. An R-Square of 70% indicates that our model does not account for 30% of the variation in contract values.

Factors Impacting Contract Values

Standardize Feature	Interpretation
FGM_Per_Game: Field goals per game	Increasing average field goals by 1 ***adds nearly 9 million dollars to contract value*** when controlling for other factors.
AST_Per_Game: Assist per game	Increasing average assists by 1 ***adds 1.3 million dollars to contract value*** when controlling for other factors.
BLK_Per_Game: Blocks per game	Increasing average blocks by 1 ***adds 1.2 million dollars to contract value*** when controlling for other factors.
PF_Per_Game: Personal Fouls per game	Increasing average fouls by 1 ***subtracts 9.8 million dollars from contract value*** when controlling for other factors.

COLLEGE REGRESSION ANALYSIS WITH VISUAL(S)

- * link notebook

COMPARISION

- A
- * link notebook

CONSIDERATIONS & LIMITATIONS

- D

References & Statement of Work

placeholder text

References

- Vorkunov,M. (2023). Vorkunov: NBA salaries keep going up. Prepare to have your mind blown in the future. *The Athletic*. https://theathletic.com/4740069/2023/08/03/nba-salary-cap-rise-jaylen-brown/
- Feng, X., Wang, Y., & Xiong, T. (2023). NBA Player Salary Analysis based on Multivariate Regression Analysis. *Highlights in Science, Engineering and Technology*, *49*, 157-166. https://doi.org/10.54097/hset.v49i.8498
- Ando, K. (2018). NBA Players' Salary Prediction Using Linear Regression Model. https://rstudio-pubs-static.s3.amazonaws.com/371407_e21330910f3c4bd2b6e19440013ea793.html#