Members : Kenneth Chan and Chuanzhou Zhang

Project Name : NBA Player Salary Prediction Using Deep Learning

Introduction

The central inquiry of this project revolves around whether NBA players are fairly compensated based on their performance statistics and if future salaries can be predicted using these metrics. The goal is to assess whether players are underpaid, overpaid, or appropriately compensated in the NBA. Employing web scraping techniques, the project aims to gather comprehensive player performance data, encompassing points per game, rebounds, assists, shooting percentages, and other pertinent metrics. By correlating these statistics with current and past player salaries, the project endeavors to identify patterns indicative of fair compensation. Additionally, machine learning models will be deployed to forecast future salaries, providing insights into potential trends and market values for NBA players.

Data Scraping

In this project, our primary goal was to compile extensive NBA player data covering the seasons from 2019-2020 to 2023-2024. We sourced information from two main platforms: the official NBA website for player statistics and HoopsHype for player salaries.

When attempting to scrape player statistics from the NBA website, we initially faced a challenge as the site blocked all scraping activities, resulting in a 403 error. To overcome this obstacle, we temporarily used a paid service from ZenRows to bypass the bot checker. However, for a more sustainable solution, we decided to directly extract player stats from the NBA website. This presented challenges due to the dynamic structure of NBA.com. Standard scraping using requests alone proved insufficient, so we employed the Selenium module. This allowed us to simulate user actions, such as changing display options, to access all player records and filter data for the desired season. Using BeautifulSoup, we then extracted relevant information, excluding unnecessary hidden data based on header names.

For extracting salary information from HoopsHype, traditional scraping techniques were employed since the website had a less dynamic structure. With the assistance of BeautifulSoup, we successfully gathered salary data for all players.This multi-step approach enabled us to overcome scraping challenges and compile a comprehensive dataset encompassing NBA player statistics and salaries for in-depth analysis over the specified seasons.

Data Cleaning

In the process of combining player stats and salary data from the NBA website and HoopsHype, challenges emerged due to discrepancies in sample numbers. Notably, the NBA site listed around 480 players in the 2023 season, while HoopsHype documented approximately 550 players, attributed to factors like 10-day contracts and two-way players not formally listed on the NBA website.

The project's initial attempt involved matching player information based on names, but formatting differences resulted in about 50 unmatched players per season. The second approach improved matches by splitting names and adding columns for the first three letters of the first name and the last name. However, issues persisted with players sharing the same last name and first three letters.

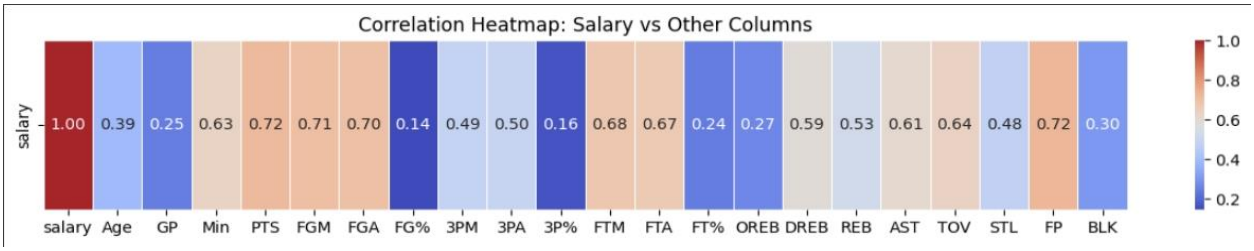
In the final attempt, datasets were merged through an inner join using exact first and last names, significantly improving matching accuracy and reducing unmatched players to around 10 per season, primarily due to name representation differences.

Following data alignment, the dataset was optimized by converting the salary column to a float, adjusting for inflation to enhance model accuracy. Unnecessary columns were eliminated, resulting in a consolidated dataset of 2656 records with 33 columns. This refined dataset serves as the foundation for visualizations and in-depth analyses of NBA player statistics over the specified time frame.

Data Analysis/Visualization

It's vital to recognize that NBA salaries are primarily shaped by player performance, revealing a noticeable correlation between statistics and salary. This suggests a level of predictability in forecasting future player salaries based on on-court performance.

Initially, we leveraged the combined dataframe to generate a correlation matrix between salary and various statistics. By utilizing Python's built-in .corr() method, we produced a visually insightful heatmap illustrating these correlations.

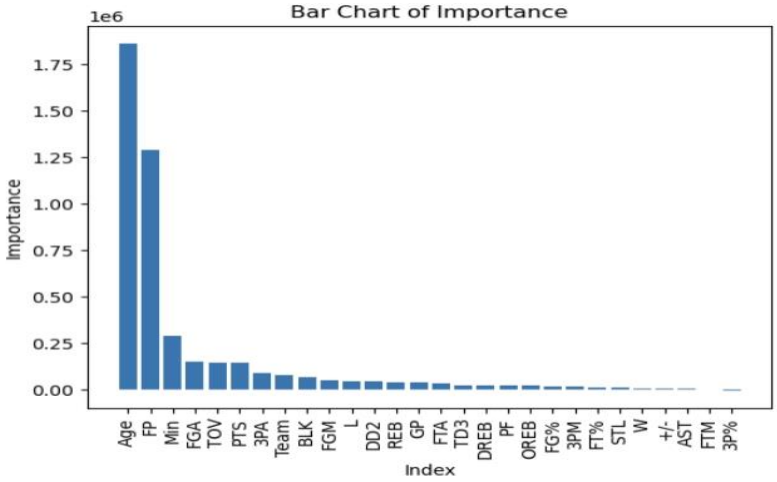


The heatmap indicates that points scored (PTS), fantasy points (FP), and field goals made (FGM) are the most crucial factors in determining salary. This initial observation highlights the significance of these three factors in salary determination.

Deep Learning Study

Subsequently, we employed a machine learning approach using Amazon's Autogluon, a deep learning module. This enabled us to identify the most influential factors in determining player salaries and predict future salaries based solely on our model . The accompanying bar chart illustrates the significance of each statistical category in this analysis.

Based on this method, Age, Fantasy Points, Minutes Played, Field Goal Attempts, Points, and free throw attempts emerged as the pivotal factors influencing salary decisions.



The key difference lies in the importance assigned to age. Our Pearson correlation suggested its lesser significance, whereas this method identifies age as the most crucial factor in salary determination.

ML Step 1: Data Preprocessing

In the initial step of our machine learning process, we combined five seasons of data, totaling 2555 records. During the preprocessing phase, all columns related to player names were excluded to enable the model to predict solely based on player statistics. The dataset was then split, allocating 80% for training and 20% for testing. In the testing dataset, the predictive column ["salary"] was removed to isolate the model from the prediction answer.

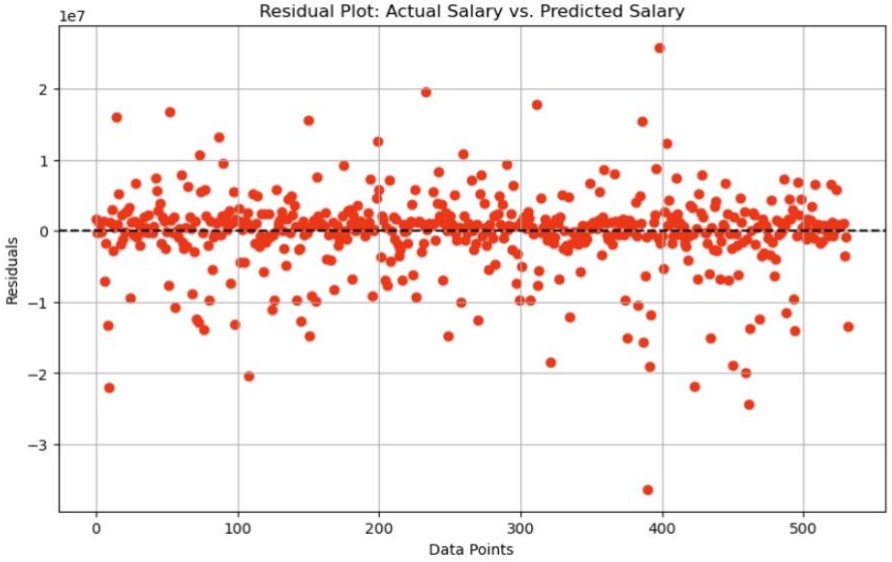
ML Step 2: Model Training

For enhanced training flexibility, we implemented varying training times, ranging from 2 to 20 minutes. Utilizing Autogluon, developers can customize parameters such as bagging, stacking, and others to achieve improved prediction accuracy. Below are four notable instances of our training runs.

Training	Memory/CPU	Training Time	Data	Method	Mean Absolute Error	Median Absolute Error
#1	Low	2 Minutes	1 season	None	5340750	3728067
#2	High	7 Minutes	4 seasons	Bagging	3630347	2755374
#3	Highest	20 Minutes	4 seasons	Stacking	2845665	1652856
#4	Highest	20 Minutes	4 Seasons drop names	Stacking	3351819	1753217

ML Step 3: Conclusion/Finding

Upon excluding player names from the training dataset, a noteworthy increase in mean absolute error was observed, while the median absolute error saw only a slight uptick. Player names can be considered a composite of non-quantifiable factors, including their agent's negotiation efforts and reputation. The substantial difference in mean absolute error between the two datasets indirectly suggests that the salaries of the highest earners are extreme values and cannot be accurately predicted solely based on game statistics. However, the trained model proves effective for predicting salaries of average players.



ML Step 4: Self-Evaluation

While our mean absolute error remains elevated, hindering the precise prediction of NBA players' exact salaries, it's essential to acknowledge the wide salary range among players and the constraints of available data sources. Despite not offering pinpoint accuracy, the model serves a valuable purpose by providing a ballpark estimate. This ensures that players are not undervalued for their services, and it helps teams avoid entering into 'bad contracts' where a player is significantly overpaid.

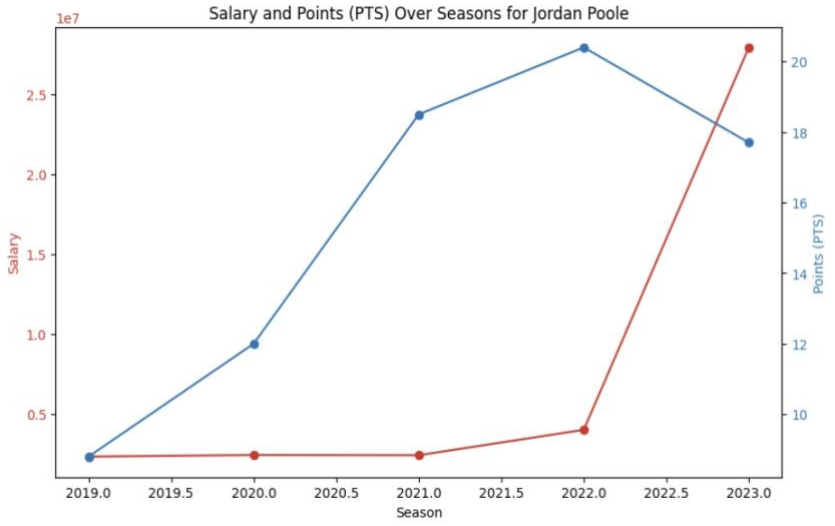
Table: NBA Players Salary Distribution

Min	25%	50%	75%	Max	STD
$6,022	$1,704,661	$3,351,582	$10,957,070	$51,915,620	$10,823,290

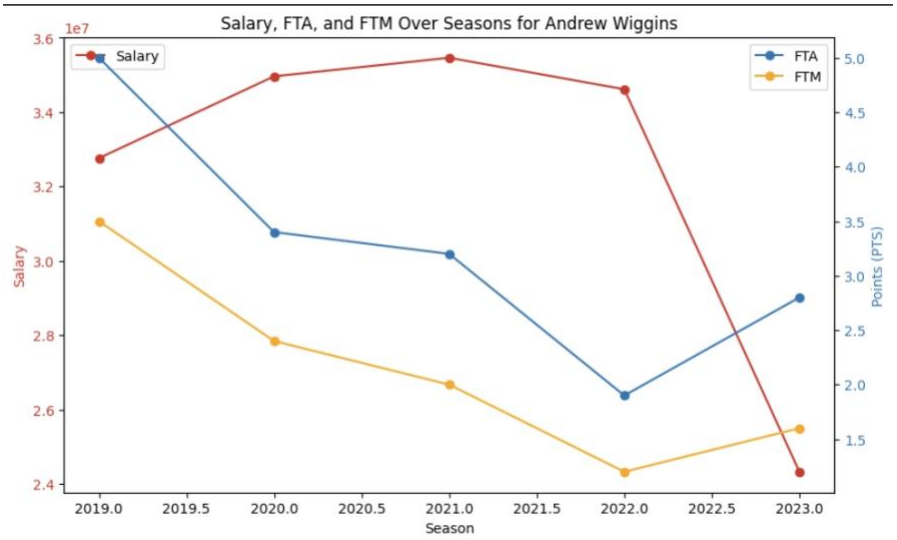
Individual Player Focused Analysis:

In analyzing player statistics and salaries, certain factors impact earnings. While age and fantasy points are beyond a player's control, increasing playing time, scoring more points, and refining free throws can boost the likelihood of a higher salary. Despite the current hype on three-point shots, our model and correlation matrix suggest it's not the most vital factor for players to focus on improving.

Here are some examples of players who secured new contracts in the 2022-2023 season, affirming the significance of these factors in determining salaries.



Presenting the time series graph for Jordan Poole, it's evident that since entering the league in 2019, he has consistently elevated his points per game. This upward trend persisted until the 2022 season when he secured a new contract. The increased points production directly correlated with a substantial rise in salary compared to his previous contract. This validation reinforces the accuracy of our model, emphasizing the importance of points in determining player salaries.



Examining another player, Andrew Wiggins, who secured a new contract after the 2022 season, we observe a consistent decrease in both free throws made (FTM) and attempted (FTA) from 2019 to 2022. This decline led to a new contract in the 2023 season, featuring a considerably lower salary than his previous one. Once more, this aligns with our model's assertion that free throws made and attempted are pivotal in predicting player salaries.

Future Work

Given more resources and time, improving our model's performance and analysis could involve refining our approach to NBA player salaries. Currently, salaries are associated with annual periods, but a more insightful method would be to break down the data by each player's contract term, providing a nuanced understanding of salary fluctuations based on performance.

However, acquiring contract-specific data is challenging due to its private nature and lack of a standardized tabular format. Extracting this information may require advanced techniques, such as Natural Language Processing (NLP), and would involve scraping multiple websites, a resource-intensive process.

Alternatively, enhancing the dataset could involve scraping more data over a more extended period or exploring additional websites for comprehensive player statistics. This would provide our model with a richer set of factors for training, potentially improving accuracy and overall performance.