Using ML for NBA Player Salary Predicition Based on Player Statistics

Bruke Amare
B.S. in Computer Science &
Data Science
College of William & Mary
200 Stadium Drive, United
States of America
btamare@wm.edu

webpage code: Visit https://brukeamare.github.io

github code: Visit https://github.com/brukeamare

Abstract— **In this project, we investigate the use of machine learning to predict NBA players' salaries by analyzing a wide range of player performance statistics. The study employs an extensive dataset to discern patterns and draw correlations that could lead to accurate salary predictions [1]. Emphasizing Support Vector Machines (SVM) and their various iterations, we explore their suitability for this complex task. Our results provide new insights into the quantifiable links between players' on-court achievements and their salaries, contributing valuable perspectives to the understanding of economic valuation in professional basketball. Through this endeavor, we offer a view of how data-driven techniques can interpret and possibly anticipate the financial aspects of athletic performance.**

I. INTRODUCTION

The fascinating world of professional basketball, particularly the NBA, offers a unique blend of athletic prowess and financial strategy. Our research dives into this realm, aiming to leverage the power of machine learning in predicting NBA players' salaries, an endeavor that combines the excitement of sports with the rigor of data analytics [2]. This study is not merely an academic exercise but a venture with substantial practical implications, offering insights that could guide teams and agents in player valuation and contract negotiations.

Existing literature in this domain often focuses on player performance metrics, but rarely do these studies make a direct connection with the financial aspects. This literature goes as far back as the 1980s with computer vision of tennis with IBM and statistical analysis of Baseball and has developed as far as Player Performance Prediction, Game Strategy Optimization, Injury Prediction, Fan Engagement and Betting. this research aims to fill this gap by correlating player statistics directly with their salaries. I start by compiling a comprehensive data set, primarily sourced from a Kaggle project [3], that includes a wide array of player statistics. This data set, chosen for its detail and alignment with current NBA trends, forms the cornerstone of our analysis, allowing us to apply and evaluate various machine learning techniques.

The journey through this project was not without challenges. I grappled with data inconsistencies and the intricate task of integrating a diverse range of statistical measures into a coherent analytical framework. This paper details our approach, findings, and the lessons learned, providing a comprehensive view of the potential and limitations of using machine learning in the high-stakes world of NBA salary prediction.

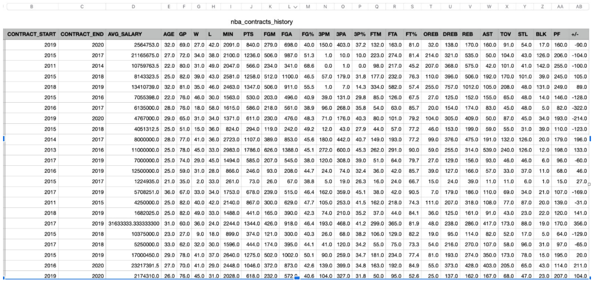
II. THE DATA

The data set used in this study is comprised of an array of NBA player statistics which reflect a diverse range of performance data of each player in the NBA. Some of the key statistics include points scored per game, assists, rebounds, shooting accuracy, and minutes played. In addition to this, each player's salary is also included, which is a crucial component of the data set. In order to streamline the analysis of the data, certain variables that were less relevant to the prediction model were removed. Some of these excluded variables include the names, contract start and end dates, and the average salary which provided no benefit to our model.

This ample data set is was sourced from a Kaggle project [7], where the project author has attempted to predict NBA salaries using strictly Random Forest for their machine learning prediction model. The decision to utilise this data set was largely based on how detailed the data set was as the original plan to scrape player statistics from HoopsHype [6], a website renowned for its comprehensive coverage of NBA news, player and team information, and financial aspects like salaries, proved to be inadequate. The Kaggle data set also aligned much better with current NBA trends [4], making it more suitable for the development of our prediction model. Unlike, the original plan to web-scrape, which would have required extensive cleaning and pre-processing [5], this new data set was well-structured and required little cleanup which enabled more focus on the analytical aspect of the project.

But we didn't just take this data at face value. We sifted through it, recalculating each salary to its 2023 equivalent, giving us a clear lane to the basket. This was anticipation—setting up the play to ensure that every variable was effectively being considered by the salaries.

Some modification were performed to the data, specifically, normalization techniques as a means to scale the data. This ensures a balanced contribution from all the variables present in the prediction model. However, it was found that the normalization step was not a significant factor



nba_contracts_history

CONTRACT_START	CONTRACT_END	AVG_SALARY	AGE	GP	W	L	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	+/-
2019	2020	2564753.0	32.0	69.0	27.0	42.0	2091.0	840.0	279.0	698.0	40.0	150.0	403.0	37.2	132.0	163.0	81.0	32.0	138.0	170.0	160.0	91.0	54.0	17.0	160.0	-90.0
2015	2017	21165675.0	27.0	72.0	34.0	38.0	2100.0	1236.0	506.0	987.0	51.3	1.0	10.0	10.0	223.0	274.0	81.4	214.0	321.0	535.0	50.0	104.0	43.0	126.0	206.0	-104.0
2011	2014	10759763.5	22.0	80.0	31.0	49.0	2047.0	566.0	234.0	341.0	68.6	0.0	1.0	0.0	98.0	217.0	45.2	207.0	368.0	575.0	42.0	101.0	41.0	142.0	255.0	-100.0
2015	2018	8143323.5	25.0	82.0	39.0	43.0	2581.0	1258.0	512.0	1100.0	46.5	57.0	179.0	31.8	177.0	232.0	76.3	110.0	396.0	506.0	192.0	170.0	101.0	39.0	245.0	105.0
2018	2019	13410739.0	32.0	81.0	35.0	46.0	2463.0	1347.0	506.0	911.0	55.5	1.0	7.0	14.3	334.0	582.0	57.4	255.0	757.0	1012.0	105.0	208.0	48.0	131.0	249.0	89.0
2015	2016	7055398.0	22.0	76.0	46.0	30.0	1563.0	530.0	203.0	496.0	40.9	39.0	131.0	29.8	85.0	126.0	67.5	27.0	125.0	152.0	155.0	65.0	48.0	14.0	146.0	-128.0
2016	2017	6135000.0	28.0	76.0	18.0	58.0	1615.0	586.0	218.0	561.0	38.9	96.0	268.0	35.8	54.0	63.0	85.7	20.0	154.0	174.0	83.0	45.0	48.0	5.0	82.0	-322.0
2019	2020	4767000.0	29.0	65.0	31.0	34.0	1371.0	611.0	230.0	476.0	48.3	71.0	176.0	40.3	80.0	101.0	79.2	104.0	305.0	409.0	50.0	87.0	45.0	34.0	193.0	-214.0
2015	2018	4051312.5	25.0	51.0	15.0	36.0	824.0	294.0	119.0	242.0	49.2	12.0	43.0	27.9	44.0	57.0	77.2	46.0	153.0	199.0	59.0	55.0	31.0	39.0	110.0	-123.0
2014	2017	8000000.0	28.0	77.0	41.0	36.0	2723.0	1107.0	389.0	853.0	45.6	180.0	442.0	40.7	149.0	193.0	77.2	99.0	376.0	475.0	191.0	132.0	126.0	20.0	179.0	196.0
2013	2016	11000000.0	25.0	78.0	45.0	33.0	2983.0	1786.0	626.0	1388.0	45.1	272.0	600.0	45.3	262.0	291.0	90.0	59.0	255.0	314.0	539.0	240.0	126.0	12.0	198.0	133.0
2017	2019	7000000.0	25.0	74.0	29.0	45.0	1494.0	585.0	207.0	545.0	38.0	120.0	308.0	39.0	51.0	64.0	79.7	27.0	129.0	156.0	93.0	46.0	46.0	6.0	96.0	-60.0
2016	2019	12500000.0	25.0	59.0	31.0	28.0	866.0	246.0	93.0	208.0	44.7	24.0	74.0	32.4	36.0	42.0	85.7	39.0	127.0	166.0	57.0	33.0	37.0	11.0	68.0	46.0
2015	2017	1224935.0	21.0	35.0	2.0	33.0	261.0	73.0	26.0	67.0	38.8	5.0	19.0	26.3	16.0	24.0	66.7	15.0	24.0	39.0	11.0	11.0	6.0	1.0	15.0	27.0
2017	2019	5708251.0	36.0	67.0	33.0	34.0	1753.0	678.0	239.0	515.0	46.4	162.0	359.0	45.1	38.0	42.0	90.5	7.0	179.0	186.0	110.0	69.0	34.0	21.0	107.0	-169.0
2011	2015	4250000.0	25.0	82.0	40.0	42.0	2140.0	867.0	300.0	629.0	47.7	105.0	253.0	41.5	162.0	218.0	74.3	111.0	207.0	318.0	108.0	77.0	87.0	20.0	139.0	-31.0
2018	2019	1682025.0	25.0	82.0	49.0	33.0	1488.0	441.0	165.0	390.0	42.3	74.0	210.0	35.2	37.0	44.0	84.1	36.0	125.0	161.0	91.0	43.0	23.0	22.0	120.0	141.0
2017	2019	31633333.333333300	31.0	60.0	36.0	24.0	2244.0	1344.0	426.0	918.0	46.4	193.0	468.0	41.2	299.0	365.0	81.9	48.0	238.0	286.0	417.0	173.0	88.0	19.0	170.0	356.0
2015	2018	10375000.0	23.0	27.0	9.0	18.0	899.0	374.0	121.0	300.0	40.3	26.0	68.0	38.2	106.0	129.0	82.2	19.0	95.0	114.0	82.0	52.0	17.0	5.0	64.0	-129.0
2017	2018	5250000.0	33.0	62.0	32.0	30.0	1596.0	444.0	174.0	395.0	44.1	41.0	120.0	34.2	55.0	75.0	73.3	54.0	216.0	270.0	107.0	58.0	96.0	31.0	97.0	-65.0
2015	2019	17000450.0	29.0	78.0	41.0	37.0	2640.0	1275.0	502.0	1002.0	50.1	90.0	259.0	34.7	181.0	234.0	77.4	81.0	193.0	274.0	350.0	173.0	78.0	15.0	195.0	20.0
2016	2020	23217391.5	27.0	70.0	41.0	29.0	2448.0	1046.0	372.0	873.0	42.6	139.0	399.0	34.8	163.0	192.0	84.9	55.0	373.0	428.0	403.0	205.0	65.0	43.0	114.0	211.0
2019	2020	2174310.0	26.0	76.0	45.0	31.0	2028.0	618.0	232.0	572.0	40.6	104.0	327.0	31.8	50.0	95.0	52.6	25.0	137.0	162.0	167.0	68.0	47.0	23.0	207.0	104.0

Fig. 1. This CSV snippet represents a portion of the initial data set, showcasing the variety of statistics collected for each player, such as points per game, assists, rebounds, and other performance indicators.

in preventing any single variable from disproportionately influencing the model's predictions down the road. This key insight truly guided the steps for data preparation, allowing the integrity of the original data characteristics to be kept and making it error free.

The data was divided into two subsets: a training set to build the machine learning models and a separate testing set to evaluate the accuracy of these models. Our main goal was to create a strong system that can accurately predict a player's salary based on their performance on the court. This prediction has great potential in player valuation and contract negotiations in the field of sports analytics.

To achieve accurate predictions, we utilized a wide range of techniques, starting from basic linear regression models to more advanced machine learning algorithms. We paid special attention to exploring different variants of Support Vector Machines (SVM) [8] due to their effectiveness in handling complex and multidimensional data. With this diverse set of methodologies, we aimed to identify the most effective approach for accurately predicting NBA salaries, considering the complexity of the data set.

III. MACHINE LEARNING TECHNIQUES

In our exploration of machine learning techniques to predict NBA salaries, we identified five methods each with unique characteristics and strengths. These techniques range from basic linear models to more complex ensemble methods, catering to various aspects of the underlying data.

- **Linear Regression:** This is a fundamental technique that models the linear relationship between independent variables and the dependent salary variable. It provides a baseline understanding of the data.
- **Ridge Regression:** Building upon linear regression, ridge regression introduces regularization [10] to mitigate the risk of over-fitting, enhancing model reliability.
- **Random Forest:** An ensemble method that utilizes multiple decision trees to increase the robustness and accuracy of predictions. It's particularly effective for handling non-linear data.
- **Gradient Boosting:** Another ensemble technique, gradient boosting focuses on improving models sequentially [11] by correcting previous errors, often leading to high accuracy.

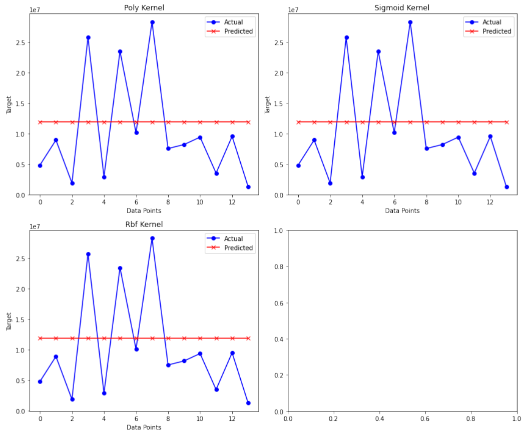


Fig. 2. The series of plots represent the performance of four different Support Vector Machine (SVM) kernels: Polynomial (Poly), Sigmoid, Radial Basis Function (RBF), and an unnamed fourth kernel. Each plot contrasts actual salaries (in blue) with predicted values (in red) across a subset of data points. The polynomial and sigmoid kernels exhibit significant variance between predicted and actual values, while the RBF kernel shows a closer alignment, suggesting better prediction accuracy. The fourth kernel's graph remains blank, indicating either a lack of data or an error in visualization. These visual comparisons underscore the variability in the effectiveness of different kernels as explored in our paper, with implications for model selection in predictive salary modeling.

- **Support Vector Machine (SVM):** This method is instrumental in classification problems, aimed at finding the optimal hyperplane for separating different salary levels.

A. SVM TECHNIQUES

Support Vector Machines (SVM) were given special attention in our study due [9] to their effectiveness in classification problems. We explored various SVM types, each offering distinct advantages depending on the data structure and the complexity of the relationships within the data.

- **RBF SVM:** The Radial Basis Function (RBF) SVM is adept at handling non-linear relationships, making it suitable for complex data sets where the relationship between variables is not linear.
- **Kernel SVM:** This variant offers more flexibility in defining decision boundaries, which can be crucial for data sets with intricate patterns.
- **Polynomial SVM:** The Polynomial SVM is effective for capturing complex and higher-order relationships within the data, though it requires careful tuning of its parameters.

When implementing these models I see that none of them seem to not be distinctively predicting the data different from each other. The results from the Support Vector Machine (SVM) models employing different kernel functions—polynomial (poly), sigmoid, and radial basis function (RBF)—indicate a notable challenge in accurately predicting NBA salaries. Each model's performance, as measured by Root Mean Square Error (RMSE) and R-squared (R^2) values, suggests significant discrepancies between the predicted and actual salaries. Specifically, the RMSE values for all three kernels (approximately 8.7 million) are exceptionally high, implying a substantial average deviation of the model's predictions from the actual salary figures.

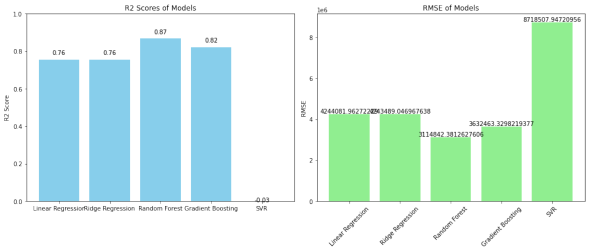


Fig. 3. The left graph illustrates the R2 scores for various models, with Gradient Boosting and Random Forest outperforming others, indicating a higher proportion of variance in salaries explained by these models. The right graph compares the Root Mean Square Error (RMSE) across models, where lower values represent more accurate salary predictions. SVM exhibits the highest RMSE, suggesting less predictive accuracy compared to ensemble methods like Gradient Boosting, which shows the lowest RMSE. These metrics are crucial in evaluating the effectiveness of each model [14] as described in the paper, guiding the selection of the most appropriate model for predicting NBA players' salaries.

Additionally, the negative R^2 values for these kernels, all hovering around -0.03, signal that the models perform worse than a simple horizontal line representing the average salary. This negative R^2 is a clear indicator of poor model fit, suggesting that the models, with their current configurations and the data set used, are unable to capture the complex relationships inherent in the salary data effectively. Such results point towards a need for significant model refinement or reconsideration of the data and approach used in predicting NBA player salaries.

B. BEST TECHNIQUE(S)

Through rigorous testing and validation, we observed that Gradient Boosting and Random Forest demonstrated the most promising results in terms of accuracy and generalization to unseen data. Ultimately, the Random Forest was chosen for its superior performance [12] in handling the complex nature of our data set. It balanced accuracy with the ability to generalize well on validation data.

A graph illustrates the comparative performance of the different machine learning techniques. It clearly demonstrates the varying levels of accuracy and error rates across these methods. There is also a graph that compares the different models cost over iterations, and it shows that Gradient Boost and Random forest are the most exceptional. Each technique interacted differently with the data. For instance, linear models struggled with the non-linear aspects, while ensemble methods and SVMs were less adept at capturing intricate patterns‘ in the data. The inconsistencies and occasional low accuracies in our tests were hypothesized to stem from various factors, such as the inherent unpredictability of sports salaries, limitations in the data set, and the complexity of accurately capturing human performance through statistical measures alone.

In the intricate dance of machine learning and NBA salary prediction [13], not all moves make the cut. We played the game of feature selection with a critical eye, aiming to avoid the trap of multicollinearity where the interplay between features—like a well-rehearsed team play—could lead to redundancy and skew our model's insights. Key players like Wins, 3-point percentage, field goal percentage, and free throw percentage were sidelined after the coach's review showed they were passing the ball amongst themselves more than shooting for the hoop. Their direct interaction, a statistical pick-and-roll, risked inflating our model's confidence without improving its game.

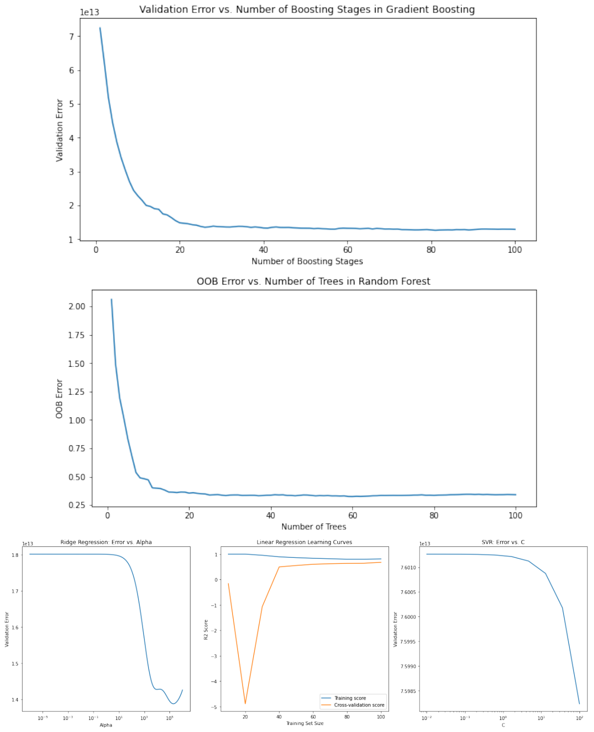


Fig. 4. The first graph showcases the Out-of-Bag (OOB) error trend in a Random Forest model, illustrating the diminishing returns in error reduction with the increase in the number of trees. The second graph displays the validation error against the number of boosting stages in a Gradient Boosting model, indicating a similar trend of sharp improvement that plateaus as stages increase. The third set of graphs presents a triptych analysis: Ridge Regression's validation error across different alpha values, highlighting the optimal regularization parameter; Learning Curves for Linear Regression, contrasting training and cross-validation scores to assess model performance; and the impact of varying the penalty parameter C on the validation error in Support Vector Regression (SVR), providing insights into the model's sensitivity to the regularization strength. Together, these visualizations synthesize the comparative performance and complexity of the employed models in predicting NBA player salaries.

But the playbook didn't end there; we went for a Hail Mary with feature expansion, driving towards a second-order model that promised a slam dunk in statistical performance. We were envisioning a symphony of interactions, where each feature's contribution was amplified, not just added—a basketball ballet that soared beyond linear correlations to capture the artistry of the game in the hard numbers of predictive analytics.

The reality check was a buzzer-beater. This ambitious expansion, while a strategic maneuver, didn't drop the statistical change we were hoping for. The scoreboard of our analysis didn't light up with the significant leaps in accuracy that would justify such complexity. It was a hard-fought lesson that sometimes, in the quest for the perfect predictive model, more isn't always better. Like a team recognizing the power of simplicity over a flashy, crowded offense, we learned that the key to a winning strategy might lie not in more elaborate plays but in executing the fundamentals.

IV. COMPARISON OF RESULTS

For testing purposes, the chosen Random Forest model was applied to a designated subset of data. This step was instrumental in gauging the model's predictive accuracy and its capacity to generalize to data it had not previously encountered. The results of the model were assessed by comparing the predicted salaries against the actual salaries from the test subset of NBA player data. This side-by-side comparison yielded insights into the precision of the model, highlighting both its successes and shortcomings in the context of salary prediction.

Despite showing potential, the model was characterized by certain constraints. The Mean Absolute Percentage Error (MAPE) stood at 44.21%, which, although it provides an understanding of prediction errors in terms of percentage, also reflects the average error's proportion to the actual salary values. Additionally, the R-squared value of 0.65 suggests that the model was able to explain 65% of the variability in salaries, indicating a reasonable fit to the data but also room for improvement. The remaining variability suggests the influence of factors not captured by the model.

Deeper analysis has underscored the complexity of predictive modeling in salary prediction. The study reaffirmed that not all influential factors can be fully captured or quantified by a machine learning model. For example, non-quantifiable elements such as a player's marketability, team financial strategies, and individual negotiation prowess, though critical to salary outcomes, present a challenge to integrate into predictive models. This realization points towards the necessity for a more nuanced approach that can accommodate such complex determinants in salary predictions.

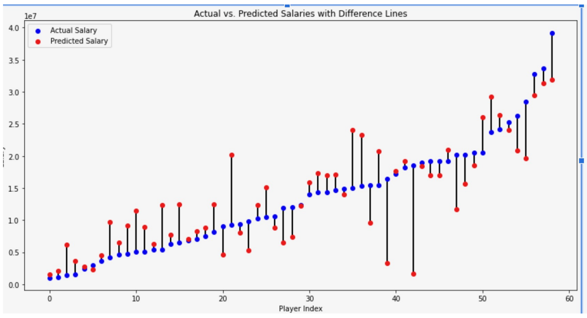


Fig. 5. This scatter plot with connecting lines illustrates the discrepancy between the actual salaries (in red) and the model's predicted salaries (in blue) across a sample of NBA players, indexed for anonymity. The vertical lines represent the difference for each player, providing a visual representation of the model's predictive accuracy.

V. DEEP LEARNING OPTIMIZATION WITH PARTICLE SWARM OPTIMIZATION (PSO)

In addition to traditional machine learning models, this study used a Particle Swarm Optimization (PSO) algorithm to optimize a neural network model for predicting NBA player salaries. The data was preprocessed to include a 'performance score', aggregating various performance metrics such as points, assists, and rebounds. All features were scaled using 'StandardScaler' to standardize them.

To improve predictive accuracy, feature engineering was performed initially. A new feature, 'performance score', was derived based on various performance statistics like points, assists, and rebounds. Other key features included shooting percentages, minutes played, and games played. Feature scaling using 'StandardScaler' ensured standardized contributions from each feature. The dataset was split using KFold cross-validation with five folds to ensure consistent evaluation, leading to more robust model performance and reduced overfitting.

A custom neural network architecture ('SalaryNet') was defined specifically for salary prediction, consisting of multiple layers with ReLU activations and dropout for regularization. Inspired by AlexNet Neural network from Daniel Vasiliu . A training function was implemented to evaluate the neural network's performance based on batch size and learning rate, using R-squared (R2) and Mean Squared Error (MSE) as evaluation metrics.

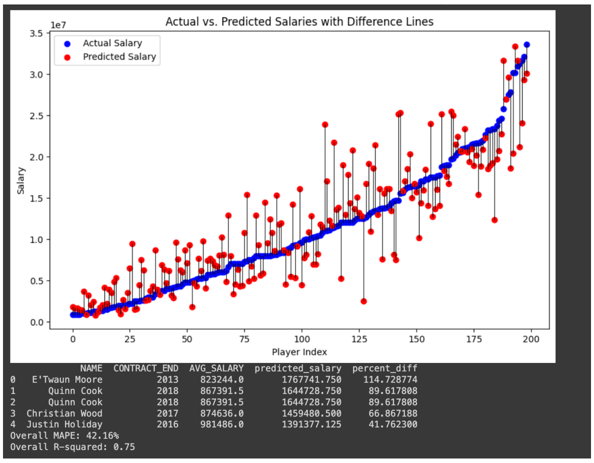
The PSO algorithm iteratively adjusted particle positions based on personal and global best positions, also inspired by Daniel Vasiliu, seeking to minimize the negative R-squared value. The final optimized model was trained and evaluated using the best batch size and learning rate. Predicted NBA player salaries were compared to the actual salaries, and the results were sorted in ascending order of actual average salary.

The comparison between actual and predicted salaries was visualized using scatter plots with difference lines. Key evaluation metrics, including the Mean Absolute Percentage Error (MAPE) and R-squared, were calculated and displayed. Despite promising results that were better than inital test by 10 percent increase of accuracy, the analysis highlighted the need for further refinement.

In summary, this approach effectively combined PSO with deep learning to provide accurate NBA player salary predictions. The sorted comparison tables and scatter plots offered clear insights into the model's performance, making this an effective solution for estimating player salaries based on historical contract and performance data.

VI. FUTURE IMPROVEMENTS

Building on the insights and findings from our intensive study on machine learning techniques for NBA salary prediction, we have identified several avenues for future enhancements to our model. This progression is rooted in the need to address the intricacies and dynamic nature of the data more effectively.



```
                 NAME  CONTRACT_END  AVG_SALARY  predicted_salary  percent_diff
0   E'Twaun Moore          2013    823244.0       1767741.750    114.728774
1      Quinn Cook          2018    867391.5       1644728.750     89.617808
2      Quinn Cook          2018    867391.5       1644728.750     89.617808
3  Christian Wood          2017    874636.0       1459480.500     66.867188
4  Justin Holiday          2016    981486.0       1391377.125     41.762300
Overall MAPE: 42.16%
Overall R-squared: 0.75
```

Fig. 6. This scatter plot with connecting lines illustrates the discrepancy between the actual salaries (in red) and the Neural Net model's predicted salaries (in blue) across a sample of NBA players, indexed for anonymity. The vertical lines represent the difference for each player, providing a visual representation of the model's predictive accuracy.

Our primary focus is on the exploration of higher-order models. Given the complex and often non-linear relationships in our data set, advancing to more sophisticated algorithms appears imperative. These advanced models are expected to capture the nuances of player salaries with greater precision, offering insights beyond the scope of our current model. Moreover, diversifying our machine learning techniques is another critical area of improvement. By incorporating a variety of methods, including advanced ensemble techniques like boosting and blending different algorithms, we anticipate a significant enhancement in the model's accuracy and robustness. Such a diversified approach could counterbalance the limitations we observed in linear models and single-method frameworks.

Further refinement is also planned in our validation strategies. Adopting randomized cross-validation methods will provide a more thorough assessment of the model's performance, particularly its generalizability across different data sets. This approach aligns with our observation that ensemble methods like Random Forest and Gradient Boosting displayed superior performance in our tests, especially in terms of accuracy and adaptability to unseen data.

In addition, adjustments in data pre-processing, such as further normalization or standard scaling, are being considered. These adjustments are crucial in dealing with the range and distribution of data, a challenge that was evident in our study where different techniques interacted variably with the data set. Furthermore, the inclusion of additional variables is under contemplation. For instance, considering factors like a player's history with different teams each season could unveil deeper insights and add layers to our understanding of salary dynamics.

Lastly, a more comprehensive account of external factors is imperative. Our current model, while robust in many respects, may not fully capture external influences significantly impacting salary predictions. These factors, such as a player's marketability and team financial strategies, though challenging to quantify, are crucial for a holistic and accurate prediction model. This realization emerged from our analysis, highlighting the gap between the model's predictive capability and the multifaceted reality of NBA salaries.

Overall, these future improvements are geared towards refining our model to not only enhance its predictive accuracy but also to ensure it remains relevant and reflective of the complex nature of sports analytics. This endeavor, we believe, will bridge the gap identified in our initial findings, where despite promising results, certain constraints like the high Mean Absolute Percentage Error and the unaccounted variability in salaries pointed towards the need for a more nuanced modeling approach.

VII. CONCLUSION

Our journey through the complex terrain of predicting NBA player salaries using machine learning techniques culminated in a wealth of insights and an appreciation for the challenges inherent in this task. The focus on Support Vector Machines (SVMs), along with other methods, brought to light the intricate balance between model complexity and predictive accuracy. While we achieved a degree of success, the project illuminated the multifaceted nature of salary prediction, where statistical data interacts with a system of unseen or unquantifiable factors.

Our exploration across different SVM types (RBF, Kernel, Poly) [15] and other techniques highlighted the variability in performance, each method presenting its unique strengths and weaknesses. The choice of our final model was a careful consideration of accuracy, computational efficiency, and the ability to generalize across diverse data sets.

The project, however, did not shy away from acknowledging the limitations we encountered. The Root Mean Square Error (RMSE) revealed that while our model could identify salary trends, it fell short of delivering precise predictions. This shortfall stems from the inherent challenge in capturing all the variables that influence an NBA player's salary, particularly those external to the data set, like marketability and team financial strategies.

The road ahead for this research is one of continuous improvement and expansion. Future iterations of this study could benefit from incorporating a broader spectrum of data, including non-statistical factors, and experimenting with more advanced or hybrid machine learning techniques. Such efforts could pave the way for more accurate, reliable models that not only predict salaries but also offer deeper insights into the economic dynamics of professional sports. Ultimately, this research stands as a testament to the potential of machine learning in sports analytics, starting further exploration into this exciting intersection of data science and sports.

REFERENCES

[1] G. Pastorello, "Predicting NBA salaries with machine learning," Medium, https://towardsdatascience.com/predicting-nba-salaries-with-machine-learning-ed68b6f75 (accessed April. 18, 2024).

[2] J. Xu, "Predicting NBA player salary with Data Science," Medium, https://betterprogramming.pub/predicting-nba-player-salary-with-data-science-c5702caa3f2e (accessed April. 18, 2024).

[3] J. Cirtautas, "NBA players," Kaggle, https://www.kaggle.com/datasets/justinas/nba-players-data (accessed April. 18, 2024).

[4] V. Vinco, "2022-2023 NBA player stats," Kaggle, https://www.kaggle.com/datasets/vivovinco/20222023-nba-player-stats-regular (accessed April. 18, 2024).

[5] "Basketball (NBA) datasets," Kaggle, https://www.kaggle.com/discussions/general/52669 (accessed April. 18, 2024).

[6] "These are the salaries of all NBA players," HoopsHype, https://hoopshype.com/salaries/players/ (accessed April. 18, 2024).

[7] "ESPN NBA Salaries," ESPN, https://www.espn.com/nba/salaries (accessed April. 18, 2024).

[8] "1.4. Support Vector Machines," scikit, https://scikit-learn.org/stable/modules/svm.html (accessed April. 18, 2024).

[9] The radial basis function kernel - University of Wisconsin–Madison, https://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/svms/RBFKernel.pdf (accessed April. 18, 2024).

[10] "5.1 - ridge regression," 5.1 - Ridge Regression | STAT 897D, https://online.stat.psu.edu/stat857/node/155/ (accessed April. 18, 2024).

[11] Written by: M. Simic, "Gradient boosting trees vs. random forests," Baeldung on Computer Science, https://www.baeldung.com/cs/gradient-boosting-trees-vs-random-forests (accessed April. 18, 2024).

[12] "Sklearn.ensemble.randomforestclassifier," scikit, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed April. 18, 2024).

[13] NBAPlayerSalary According to race and stats) - university of connecticut, https://stamford.econ.uconn.edu/wp-content/uploads/sites/1361/2016/05/NBA-Salary-Race-Stats-Powerpoint.pdf (accessed April. 18, 2024).

[14] "NBA players' pay and performance: What counts?," The Sport Journal, https://thesportjournal.org/article/nba-players-pay-and-performance-what-counts/ (accessed April. 18, 2024).

[15] "Can statistics be used to determine an NBA player salary," Singapore Travel Guide, https://www.streetdirectory.com/travel_guide/41055/recreation_and_sports/can_statistics_be_used_to_determine_an_nba_player_salary.html (accessed April. 18, 2024).

[16] Github, https://github.com/dvasiliu/AAML (accessed April. 18, 2024).