International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-12 Issue-3, February 2023

Determine the Undervalued US Major League Baseball Players with Machine Learning

Lu Xiong, Kecheng Tian, Yuwen Qian, Wilson Musyoka, Xingyu Chen

Abstract*: Baseball is a sport of statistics. The industry has accumulated detailed statistical data on both offence and defence for over a century. Experience has shown that data analysis can provide a competitive advantage over teams that do not utilise such analysis. Over the last two decades, the development of machine learning and artificial intelligence has enabled the creation of more advanced algorithms for analysing data in baseball. In the following research, we will run different ML models using sci-kit-learn and H2O on Colab, and the Caret package on RStudio to examine the datasets (hitting dataset and salary dataset) and determine the undervalued players by predicting the number of runs scored in the following year. We will compare machine learning regression algorithms and ensemble methods, providing comprehensive explanations of the results. The suggestion of which model is superior in terms of prediction accuracy will be determined.

***Keywords*: Sports Analytics, Machine Learning, Ensemble Methods, Deep Learning.**

I. INTRODUCTION

The Oakland Athletics gained fame for their innovative use of a limited budget, as depicted in the 2011 film "Moneyball." The management of the Major League Baseball (MLB) team was the first documented instance of using statistical analysis to inform player acquisition decisions. The Athletics utilised advanced analytics to identify players with low salaries who could perform competitively. Coaches of professional baseball teams in the United States have long relied on the conventional rule that statistics like a player's Batting Average (AVG) are what determines a player's salary. Billy Beane, the general manager of the Oakland Athletics during the period that the movie Moneyball covers in the team's history, took a new approach by using On-Base Percentage (OBP) to select players.

Manuscript received on 27 December 2022 | Revised Manuscript received on 01 February 2023 | Manuscript Accepted on 15 February 2023 | Manuscript published on 28 February 2023.
*Correspondence Author(s)
Lu Xiong*, Assistant Professor, Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA. Email: Lu.Xiong@mtsu.edu, ORCID ID: https://orcid.org/0000-0003-2471-1256
Kechen Tian, Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA. Email: kt5s@mtmail.mtsu.edu, ORCID ID: https://orcid.org/0000-0002-5481-0012
Yuwen Qian, Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA. Email: yq2c@mtmail.mtsu.edu, ORCID ID: https://orcid.org/0000-0002-3210-5961
Wilson Musyoka, Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA. Email: wwm2p@mtmail.mtsu.edu, ORCID ID: https://orcid.org/0000-0002-6102-4597
Xingyu Chen, Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA. Email: xc2k@mtmail.mtsu.edu, ORCID ID: https://orcid.org/0000-0001-5716-2121

OBP represents a player's ability to get on base, whether that is a hit, walk, or hit-by-pitch. Over the past two decades, the use of data analytics has shown a resurgence in professional sports, business, and government. This resurgence was partly attributed to Moneyball, which provides audiences with an understanding of advanced baseball data analytics and how they can improve player selection and game management. Professor Tom Davenport is widely thought to be the founder of analytics and describes data analytics as "the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions" [1]. Baseball is well-suited to study the effects of data analytics because the precise point in time when teams adopted advanced data analytics can be determined. The first recorded baseball game was played between the New York Knickerbockers and the New York Nine in 1846. Being a historic sport means that a vast amount of data is available for analysis, which is another compelling reason for choosing baseball analytics as the subject of this paper. MLB has the most significant number of games in professional sports in North America. A team needs to play 162 regular games in a season, which runs from April to September. The more games played, the larger the dataset that can be collected. With the advancement of machine learning and artificial intelligence, we can utilise more sophisticated algorithms to analyse baseball data. In this paper, we will utilise the players' hitting and salary data to identify underpaid players using various machine learning algorithms. Recommendations will be made to determine which model is superior in terms of predictive accuracy. The predictions made by this paper can be more accurate if more advanced machine learning algorithms are used, rather than the statistical analysis currently employed in the sports analytics industry.

II. LITERATURE REVIEW

Sabermetrics is the investigation of historical and statistical questions in baseball [2], [3]. Bill James wrote a book called The Bill James Abstract. James states the first axiom of Sabermetrics: the objective of every team is to contribute to winning games while avoiding actions that contribute to losing games [3]. Contributions to winning or losing games include these variables: runs, walks, runs allowed, hits, home runs, strikeouts, errors, etc. [3]. Based on the article [3], many analytical research papers have been done. [2] reviewed past research comprehensively and pointed out that the styles of researching baseball data fall into three categories: Regression, Binary Classification, and Multiclass Classification. Baseball data sets can result in modelling and explaining the variation in players' performances [4] and salaries [5]. This process requires the implementation of techniques in regression [4], [5].

Another popular method for evaluating game outcomes is support vector machines (SVM).

SVM is well-studied and widely used for both classification and regression problems as a type of supervised machine learning in baseball analytics [6], [7]. After comparing the artificial neural network (ANN), SVM, linear regression (LR), and one-dimensional convolutional neural network (1DCNN), [6] suggested SVM is the most accurate in predicting the games' outcome.

Numerous studies have been done to predict the salary of the players and the outcome of baseball games [6], [8]. A few studies examined the number of runs scored as a target variable. This study will focus on modelling the number of runs scored by players as a function of their performance using various machine learning methods. The results will determine the players' possible future contribution to teams. This will be of great use to any players and teams, especially those who intend to hit unrestricted free agency.

III. METHODOLOGY

Figure 1 below shows the significant steps of our research. The datasets are preprocessed using Google Colab. We used Python and R to visualize the data. Machine learning regression algorithms and ensemble methods will be presented and evaluated. We will provide comprehensive interpretations of the results in the end.

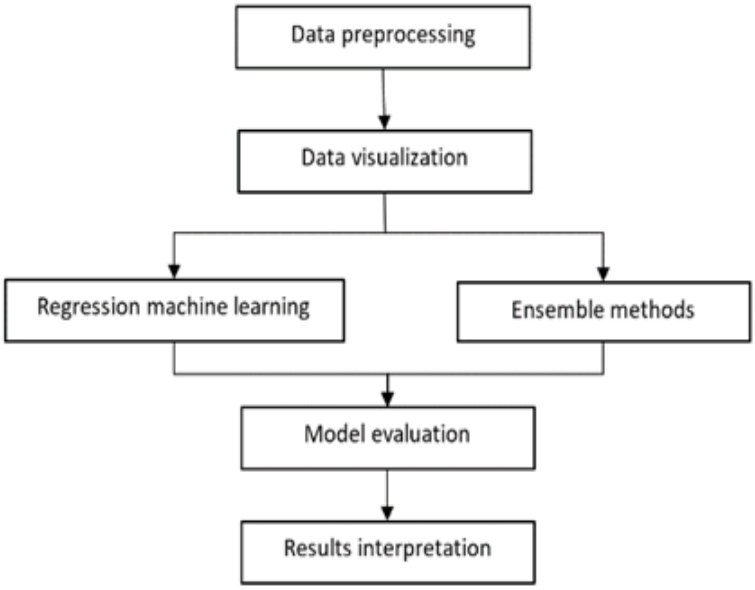

Fig. 1. Major steps of this research.

A. Regression Algorithms

SVM

In machine learning, there are four approaches that one can take: namely, supervised, unsupervised, semi-supervised, and reinforcement learning [9]. Classification is a good example of supervised machine learning, where the system is fed data and generates an output or classification for the data. The output generated is typically derived from previous examples or samples of data that the system was fed during the training phase. Support Vector Machine (SVM) is a computer algorithm that follows this idea; it operates by receiving examples and then outputs labels for those examples. SVM works by creating a hyperplane of best fit that separates the data into two distinct classes in higher dimensions. The aim is to maximise the distance (margin) between the hyperplane and the point closest (support vector) to the hyperplane from each class. A good example of the use of SVM is in the banking industry, where SVM learns to detect credit card fraud by examining thousands of fraudulent and non-fraudulent credit card transactions. Although SVM is a tool widely applied in classification due to its balanced accuracy and reproducibility, it has also been occasionally used for regression. Similarly to classification, the aim is still to minimise the sum of distances between the many points and the hyperplane. Although in regression we must solve a quadratic programming problem, which makes it more challenging to use.

B. Ensemble Methods

Random Forest

Random forest (RF) is an ensemble learning method for classification and a machine learning technique. It belongs to the Bagging (short for Bootstrap A Ggregation) method in ensemble learning [10].

When the classification task is performed, new input samples enter and let each decision tree in the forest judge and classify them separately, each decision tree will get a classification result of its own; whichever one of the classification results of the decision tree has the most classifications, then the random forest will take this result as the final result. Every single tree that the random forest algorithm generates is weak, but all of them combined are powerful. The random forest can determine the interaction between different features, making it challenging to overfit. Random forests can handle quantities whose attributes are discrete values and quantities whose attributes are continuous values.

Gradient Boosting Regression (GBR)

Gradient Boosted Regression Tree is a more powerful model built by merging multiple decision trees [11]. This model can be applied to both regression and classification tasks.

Retrieval Number: 100.1/ijitee.B94060112223
DOI: 10.35940/ijitee.B9406.0212323
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Unlike the random forest approach, gradient boosting constructs trees sequentially, with each tree attempting to correct the errors of the previous tree. Gradient boosting trees typically use trees with small depths (between 1 and 5) to minimise memory usage and improve prediction speed. The main idea behind gradient boosting trees is to combine many simple models (called weak learners), such as trees with small depths. Each tree can only make good predictions for a portion of the data, so adding more and more trees allows for constant iteration and improves performance.

XG Boost

XGBoost, also known as eXtreme Gradient Boosting, is a supervised learning algorithm that can be used to solve regression and classification problems [12]. XGBoost is developed from GBDT (Gradient Boost Decision Tree), which also uses an additive model with a forward stepwise algorithm to achieve the optimisation process of learning, but is different from GBDT (Gradient Boost Decision Tree). The main differences include the objective function, optimisation method, handling of missing values, and prevention of overfitting, which lead to better results with fewer computational resources in the shortest time using XGBoost.

Bagging vs Boosting Ensemble

The bagging method and boosting method are two primary methods in ensemble learning [13]. The main difference between these two methods is the way of training [14].

Bagging, which usually considers homogeneous weak learners, learns them independently in parallel and combines them according to some deterministic averaging process [15]. The most common method of bagging is the use of bootstrapping to create multiple samples (often called "weak learners"). Bootstrapping samples have representativeness and independent features. Next, various models are aggregated in parallel in different ways [16]. In the case of regression, the average of the output values of all independent classification results is selected. In the case of classification, the class with the highest number of accepted votes will be chosen.

Boosting, which also considers homogeneous weak learners, sequentially fits multiple weak learners in a very adaptive manner (iteratively fitting the model so that the model fitted later is influenced by the models fitted in previous steps) and combines them according to a deterministic strategy. Each iteration of fitting has a greater effect on fitting more difficult observations, and as the number of iterations increases, a stronger learner with lower bias characteristics can be obtained. Once the learner is selected, there will be many methods to be sequentially fitted, two of which are AdaBoost and gradient boosting [17].

Neural Network and Deep Learning

As an algorithm in Machine Learning (ML), Neural Networks were initiated in computer science to mimic the structure and processes of the human brain. A neural network is also known as Artificial Neural Networks (ANNs). ANN is a mathematical model comprising a large number of nodes (or neurons) interconnected with each other. Each node represents a specific output function, called an activation function. The connection between each pair of nodes represents a weighted value for the signal passing through the connection, known as a weight, and the neural network simulates human memory in this manner. The output of the network depends on its structure, the way it is connected, the weights, and the activation function. The network itself is typically an approximation of a specific algorithm or function. The operation of biological neural networks inspires the concept of constructing neural networks. The ANN combines the understanding of the biological neural network with the mathematical statistical model, and realizes it with the help of mathematical statistical tools [18]. If a neural network contains more than three layers, it is considered a deep learning algorithm.

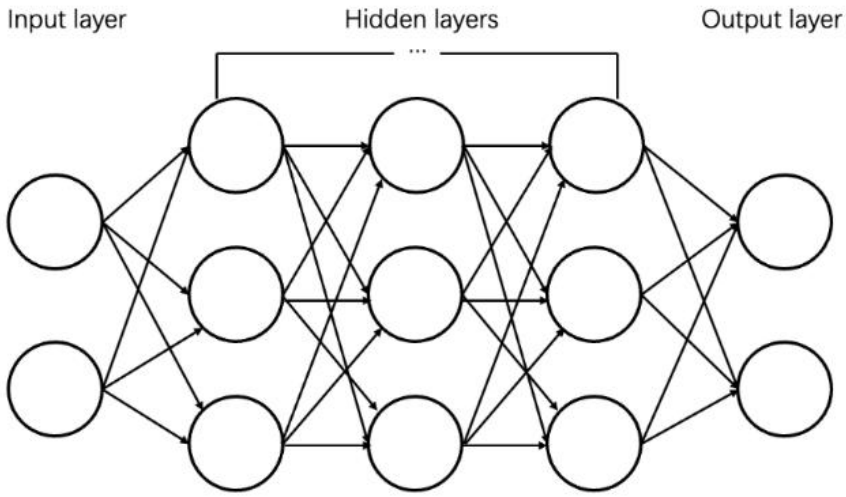

Fig. 2. Deep Learning Neural Networks.

IV. DATA DESCRIPTION AND EXPERIMENT

For this project, we obtained our data from Kaggle.com. Kaggle is a well-known data hub for data analysts; hence, we can be sure that the data being used will be, at the very least, reliable and consistent. The original database was created by Sean Lahman, a journalist and baseball researcher, and was licensed under a Creative Commons Attribution license.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

 The main modifications the data underwent from its original form were changing the column names to have consistent formatting and naming, and creating an SQLite database in addition to the already existing CSV files. We will use two different datasets for this project. The first will consist of baseball hitting statistics dating back to 1871 and extending to 2015. The second data set will comprise player salaries from 1985 to 2015, along with their respective teams. The hitting data set has 22 columns and 101,333 rows, while the salaries data set has 5 columns and 25,576 rows. The salary data set only dates back to 1985; therefore, we will need to exclude some of the earlier data from the batting records for the two data sets to ensure a match. The misalignment in data limits our experiment, but it is a necessary process for the experiment to make sense.

The hitting data consists of the following variables:

- PlayerID: The unique identifier of the player
- YearID: Number of years
- TeamID: The name of the team
- IgID: Which league did the player play in
- G: Number of games played
- AB: Number of at-bats
- R: Number of runs scored
- H: Number of base hits
- 2B: Number of doubles
- -3 B: Number of triples
- HR: Number of home runs
- RBI: Number of runs batted in
- SB: Number of Stolen Bases
- CS: Number of times caught stealing
- BB: Number of bases on balls
- SO: Number of strikeouts

The salary data set consists of:

- YearID: Number of years
- TeamID: The name of the team
- IgID: Which league did the player play in
- PlayerID: The unique identifier of the player
- Salary: The Amount of money each player made during a given season.

The two data sets are combined based on four shared categorical variables: player, year, team, and league. After mixing the two data sets, we have 22,750 columns and 17 rows. From this combined data set, the highest salary is $33,000,000, and the lowest salary is $0. The highest salary found in the study belonged to Alex Rodriguez, who was the highest-paid player in MLB from 2001 to 2013, except for 2004. The standard deviation of salaries is $3,437,933.71, and the average is $2,144,725.57, indicating a relatively low level of dispersion in the salary dataset. According to Figure 3, most players' salaries are below $5,000,000. In the 'Number of Games Played' column, the highest value is 163, and the lowest value is 1. There are two categorical variables: team ID and Ig ID. MLB is divided into two leagues, the American League (AL) and the National League (NL). In the IgID column, there are two leagues: AL and NL. There are thirty teams in total, divided into two leagues: 15 teams in the NL and 15 teams in the AL.

We will use the number of runs as our target variable, while having different predictors. This would enable us to evaluate the players' salaries based on their on-field performances from previous years. This prediction will indicate whether the player is being fairly compensated, whether they are underpaid or overpaid.

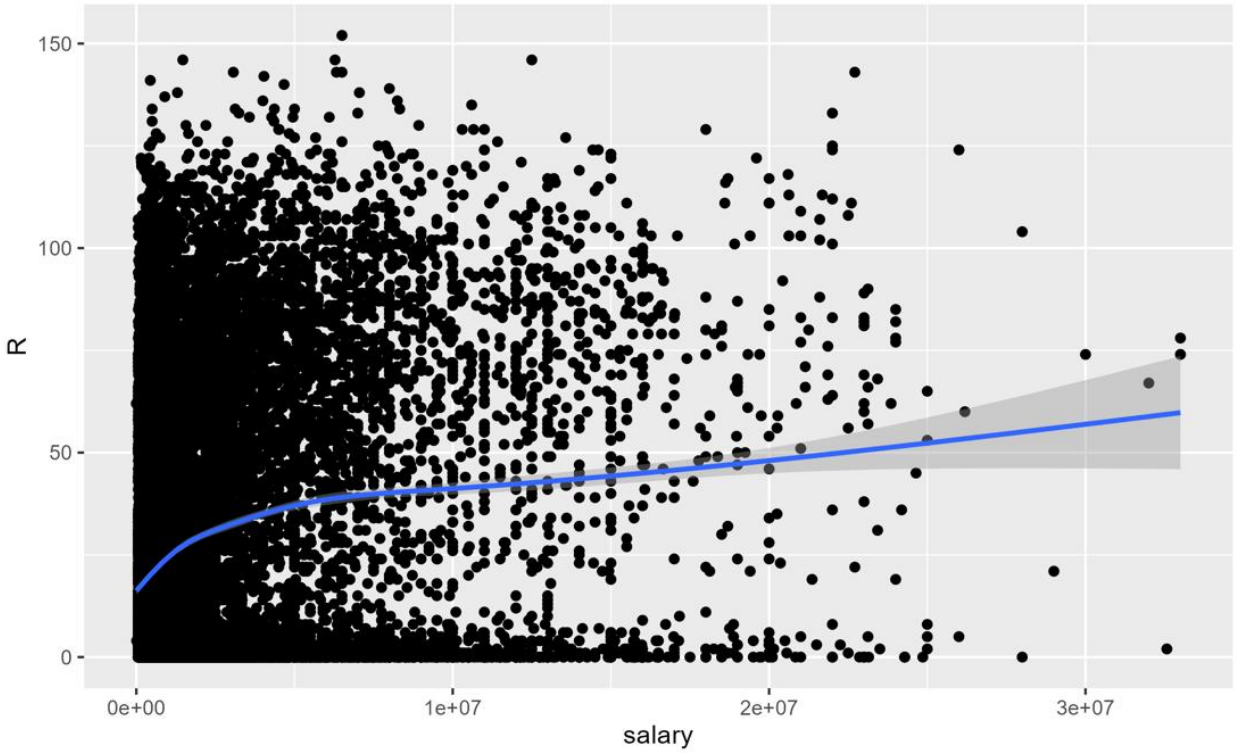

Fig. 3. Scatter Plot of Salary vs. Score with the Fitted Curve and 95% Confidence Interval.

To explore the relationship between the salary and the number of runs scored, we plotted Figure 3. Each dot in this figure represents the data of one player in a particular year. The blue line represents the regression line, illustrating the trend of the data, and the grey band indicates the 95% confidence interval of the regression value. According to this figure, when the salary variable increased, the score showed a logarithmic trend of rapid growth. Then it slowed down, indicating that in the early stage, when the salary level was not high, there was a large room for improvement. However, when the salary reached a high level, the score in the competition had almost reached its limit, and growth was limited. When the salary range is below $15 million, a large amount of sample data is available for regression, and the confidence interval of the obtained fitted curve is narrow, indicating that the prediction accuracy is higher for low-salary levels compared to high-wage levels. However, we still note that when the salary range is below $15 million, there is still a lot of real data outside the 95% prediction confidence interval, such as those with lower salaries but higher runs scored. These are the "undervalued" players we are looking for.

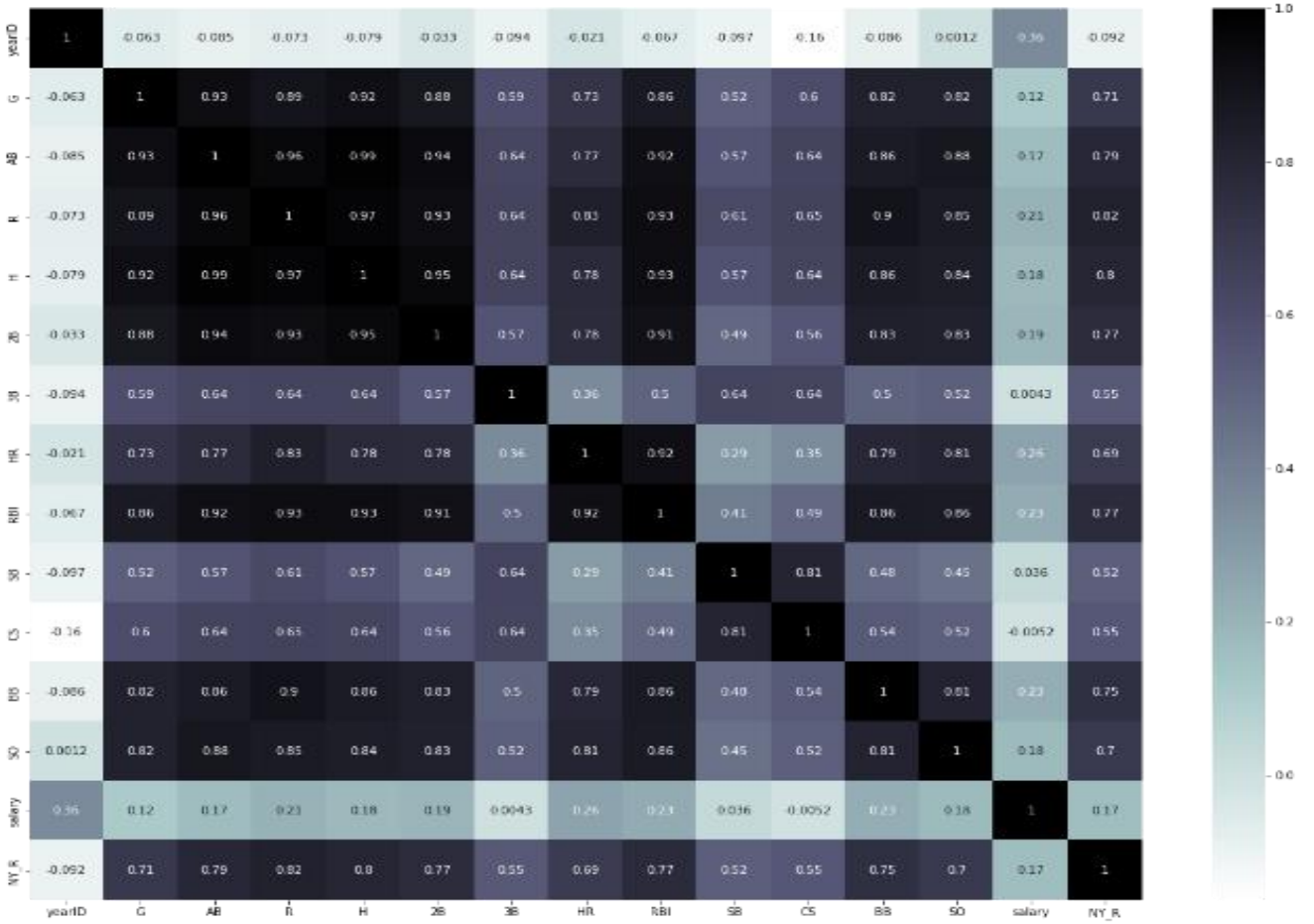

Fig. 4. Heat Map of the Correlation Between the Variables.

We are also interested in the correlations between the variables, so we plotted a heat map, as shown in Figure 4. We first discuss the relationship between salary and individual variables. The correlation between salary and year is the highest among all variables, and that's because salary generally increases over the year due to inflation. All the remaining variables showed a positive correlation with salary, except for the number of stolen bases. Among them, the number of runs scored, home runs, RBIs, and on-base percentage showed a higher correlation with salary. This suggests that when evaluating the value of players, people tend to focus more on these variables. Especially, the home run variable, as the most difficult hitting action in baseball, requires the batter to hit the ball out of the boundary in one hit, and baseball players who can hit home runs also reflect their high level, so based on such psychological expectation, the players' salary level will be relatively higher. The number of RBI, baseball on base, and runs scored are essential variables related to whether a baseball player can win a game or not. When these variables increase, the chance of winning a game rises, and so does the salary. From Figure 4, we can also see that the numbers of at-bats, runs scored, and base hits are highly correlated, with a correlation of over 95%, indicating that as the number of at-bats increases, the number of hits increases, and so does the number of runs scored. This aligns with what one would expect. The remaining variables with correlations as high as 90% are the number of games and number of at-bats, the number of runs scored, and the number of RBIs in various situations, which indicates that as the number of at-bats increases, the chance of scoring runs also increases. Python and R will be the programming languages used in this experiment, with CoLab and RStudio as the programming environments. CoLab is a cloud-based Google Research program that enables anyone with internet access to run Python code.

Of the total data available, we will use 80% for training and the remaining 20% for testing. Some of the information on the hardware used in CoLab is as follows: Intel(R) Xeon(R) CPU @ 2.20GHz, with a CPU MHz of 2.20 GHz and a cache size of 56.32 MB.

V. RESULTS

A. Evaluation Metrics

When evaluating machine learning models, it is essential to select the most suitable evaluation metrics. There are diverse types of evaluation criteria in the real world, and sometimes it is even necessary to create new evaluation metrics that are appropriate for business problems [19]. Evaluation metrics are often used to measure the performance of a model and to detect the stability of its operation.

Next, we will introduce some standard evaluation metrics for machine learning algorithms. Specific notations will be used to represent the information in our dataset. We assume that there are n rows in this data set, their values of the predictors are $x_1, x_2, \ldots, x_i, \ldots, x_n$, the corresponding values of the target variable are $y_1, y_2, \ldots, y_i, \ldots, y_n$, the predicted values of the target variable are $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_i, \ldots, \hat{y}_n$. The mean of the target variable is denoted by $\bar{y}\bar{y}$. e_i denotes the prediction error.

Mean Absolute Error (MAE) [20]: the average of all absolute errors, which is used to measure the average absolute distance between the predicted and actual target values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (1)$$

Mean Squared Error (MSE): the mean squared error between the predicted and actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \quad (2)$$

Root Mean Squared Error (RMSE) [20]: RMSE is defined as the standard deviation of the prediction errors. The RMSE measures the dispersion of the errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1} e_i^2} \quad (3)$$

Mean Squared Logarithmic Error (MSLE): MSLE is the mean of the relative squared difference between the log-transformed actual and predicted values.

$$MSLE = \frac{1}{n}\sum_{i=0}^{n}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2 \quad (4)$$

Root Mean Squared Logarithmic Error (RMSLE): RMSLE is the square root of the MSLE.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2} \quad (5)$$

Mean Percentage Error (MPE): MPE is the average percentage of errors of each entry in a dataset.

$$MPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{e_i}{y_i} \quad (6)$$

Mean Absolute Percentage Error (MAPE): MAPE can be calculated as the average of the absolute percentage errors of each entry in a dataset.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| \quad (7)$$

R2 Score: R2 Score indicates how well the model fits the data. An R-squared value close to 1.0 indicates that the model fits the data well, while a value close to 0 indicates that the model does not fit very well. R-squared may be negative when the model predictions are highly unrealistic.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

B. Results

Several machine learning algorithms, including non-ensemble learning, ensemble learning, and deep learning, are run on our datasets. To compare the advantages and disadvantages of each algorithm for this prediction problem, we list Table 1. The table includes seven models with the eight metrics we discussed earlier.

Table 1: Prediction Accuracy Measured by Various Metrics for Different Algorithms.

Model Name	MAE	MSE	RMSE	R-Square	MAPE	MPE	MSLE	RMSLE
XGBoost	11.70	329.01	18.14	0.70	112.11%	-80.53%	0.67	0.82
SVM	26.80	1246.37	35.25	-0.12	222.30%	-138.90%	3.40	1.84
GBR	11.7	329.42	18.15	0.70	112.30%	-80.63%	0.67	0.82
Random Forest	11.42	326.72	18.08	0.71	55.07%	-39.46%	0.57	0.75
Deep Learning	11.25	313.38	17.70	0.72	109.61%	-78.49%	11.89	0.717
Gaussian process regression	29.36	1972.97	44.42	-0.78	19.61%	18.99%	8.50	2.92
GLM	12.77	335.2587	18.6384	0.6866	126.57%	-98.69%	0.1931	0.4394
M5 Decision Tree	11.67	319.581	18.1445	0.7030	116.88%	-83.10%	0.1103	0.3322

According to the results, the R-squared value for prediction using the SVM model is close to 0, indicating that the final prediction results of the model are less satisfactory. Other metrics of SVM also deviate significantly from the optimal criterion compared with those of the different models, so the SVM model is not the best choice for prediction. Among the ensemble learning methods, XG Boost, GBR, Random Forest, Deep Learning and M5 Decision Tree all achieved good fitting results in terms of the R^2 over 0.7. We also observed that these five models achieved relatively small prediction errors in the MAE. And the 5 models also achieve better performance in terms of MSE. MAPE and MPE measure the percentage error, and the Gaussian process regression (GPR) model obtained the best results (19.61% MAPE and 18.99% MPE), with the RF model following closely behind. We also utilise MSLE and RMSLE metrics to measure the prediction accuracy of skewed, distributed data more accurately. Model GLM and M5 got the best results in terms of these 2 evaluation metrics. Overall, the ensemble model has better prediction accuracy than

single models.

The reason is that many learning algorithms work by performing some form of local search, which can get stuck in a local optimum. The ensemble algorithm formed by running local searches from many different starting points can provide a better approximation to the actual unknown function than any single classifier. As a demonstration, we utilise the best model, which is deep learning, as shown in Table 1, to make a prediction using the latest year in the data. This illustrates how readers can apply our research results to select undervalued baseball players. The latest data available is from 2014, which includes each player's salary for that year. We input it into the deep learning model obtained above to predict the performance of each player in the next year, 2015. Figure 5 is the scatter plot we generated, which shows the expected player's performance in the next year versus the player's salary in the current year. The players in the upper left corner of the figure are the most economical choices for coaches or team managers to invest in, as they are paid a lower salary but are predicted to have high performance in the following year.

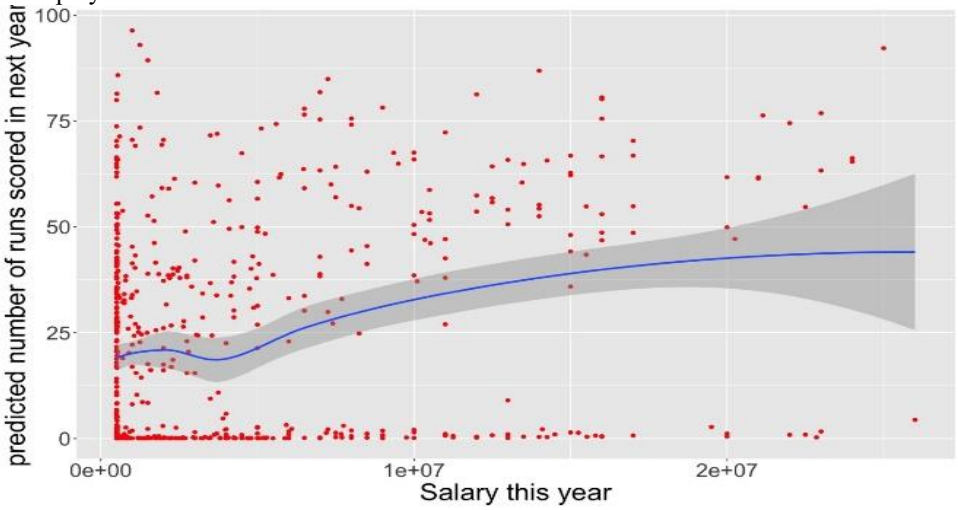

Fig. 5. Scatter plot of Current Salary Vs Predicted Performance.

VI. CONCLUSION

We used different machine learning methods to model the players' number of runs scored based on their past years' available data. We also merged two original data sets, which are hits and salaries, to build the model more effectively. After running different ML models using scikit-learn and H2O on Colab, the Caret package on RStudio and evaluating the results, we concluded that Neural Network-Deep Learning (DL) was superior to other models. Besides, we can achieve data visualization using regression machine learning and ensemble methods. During this process, we also need to set the number of runs scored as the target variable to create a graph. Through our analytic process, undervalued players can be determined. In detail, baseball team managers can use the number of runs scored to identify which players with excellent performance are underpaid. Players, especially those who intend to hit unrestricted free agency, can also use our research to evaluate themselves, allowing them to negotiate a more favourable price in their upcoming contracts.

Like any research study, this one also has some limitations. One of the limitations is that our research is based on the massive amount of data collected in the past seasons. Some individual cases are not discussed in this paper, such as age and injuries. If a great player is injured for several seasons, his record would be affected, which would alter the predicted variable numbers of runs scored next year. However, these can be improved in future studies. We should acquire the latest data and consider age as a variable. Then, the baseball teams and players can provide a superior two-way selection.

ACKNOWLEDGMENT

We thank Haiting Cai, Xintong Cao, Jiyao Luo and Yitong Meng for their contribution during various stages of the paper preparation.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval or consent to participate, as it presents evidence that is not subject to interpretation.
Availability of Data and Material/ Data Access Statement	The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.
Authors Contributions	Lu Xiong: the overall research strategy, direction, ideas, team management, and writing. Kechen Tian: coding and results summary. Yuwen Qian: literature review and citations. Wilson Musyoka: data description and general introduction of SVM. Xingyu Chen: Model evaluation and interpretation.

REFERENCES

1. T. H. Davenport and J. G. Harris, "Competing on Analytics, Updated, with a New Introduction: The New Science of Winning.," Harvard Business School Press Books, p. 1, 2017, [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=124794328&site=ehost-live&scope=site%0Ahttp://hbr.org/product/a/an/10157-HBK-ENG
2. K. Koseler and M. Stephan, "Machine Learning Applications in Baseball: A Systematic Literature Review," Applied Artificial Intelligence, vol. 31, no. 9–10, pp. 745–763, 2017, doi: 10.1080/08839514.2018.1442991. [CrossRef]
3. B. James, The Bill James Abstract. 1980.
4. M. R. Watnik, "Pay for Play: Are Baseball Salaries Based on Performance?," Journal of Statistics Education, vol. 6, no. 2, pp. 1–6, 1998, doi: 10.1080/10691898.1998.11910618. [CrossRef]
5. Y. Han, J. Kim, H. Keung, and T. Ng, "Logistic Regression Model for a Bivariate Binomial Distribution with Applications in Baseball Data Analysis," Entropy, 2022. [CrossRef]
6. S. Li, M. Huang, and Y. Li, "Exploring and Selecting Features to Predict the Next Outcomes of MLB Games," Entropy, 2022.
7. K. Koseler and M. Stephan, "Machine Learning Applications in Baseball: A Systematic Literature Review," Applied Artificial Intelligence, vol. 31, no. 9–10, pp. 745–763, 2017, doi: 10.1080/08839514.2018.1442991. [CrossRef]
8. M. Huang, "Use of Machine Learning and Deep Learning to Predict the Outcomes of Major League Baseball Matches," Applied Sciences, 2021. [CrossRef]
9. H. Wang and D. Hu, "Comparison of SVM and LS-SVM for regression," in International Conference on Neural Networks and Brain Proceedings, 2005, vol. 1, pp. 279–283. doi: 10.1109/icnnb.2005.1614615. [CrossRef]
10. S. Wan and H. Yang, "Comparison among methods of ensemble learning," in Proceedings - 2013 International Symposium on Biometrics and Security Technologies, ISBAST 2013, 2013, pp. 286–290. doi: 10.1109/ISBAST.2013.50. [CrossRef]
11. G. Rong et al., "Rainfall Induced Landslide Susceptibility Mapping Based on Bayesian Optimized Random Forest and Gradient Boosting Decision Tree Models—A Case Study of Shuicheng County, China," Water (Basel), no. 3066, p. 12, 2020, doi: 10.3390/w12113066. [CrossRef]
12. S. Dey, Y. Kumar, S. Saha, and S. Basak, "Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting," in PESIT South Campus, 2016. doi: 10.13140/RG.2.2.15294.48968.
13. E. Bauer and R. Kohavi, "An Empirical comparison of voting classification algorithms: bagging, boosting, and variants," Mach Learn, vol. 36, no. 1, pp. 105–139, 1999, doi: 10.1023/a:1007515423169. [CrossRef]
14. T. G. Dietterich, Ensemble methods in machine learning. 2000. doi: 10.1007/3-540-45014-9_1. [CrossRef]
15. R. K. Dhanaraj et al., "Random Forest Bagging and X-Means Clustered Antipattern Detection from SQL Query Log for Accessing Secure Mobile Data," Wirel Commun Mob Comput, vol. 2021, 2021, doi: 10.1155/2021/2730246. [CrossRef]
16. F. Petropoulos and E. Spiliotis, "The Wisdom of the Data: Getting the Most Out of Univariate Time Series Forecasting," Forecasting, vol. 3, no. 3, pp. 478–497, 2021, doi: 10.3390/forecast3030029. [CrossRef]
17. V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki, "Fair adversarial gradient tree boosting," in Proceedings - IEEE International Conference on Data Mining, ICDM, 2019, vol. 2019-Novem, pp. 1060–1065. doi: 10.1109/ICDM.2019.00124. [CrossRef]
18. IBM Cloud Education, "Neural Networks | IBM." 2020. [Online]. Available: https://www.ibm.com/cloud/learn/neural-networks
19. J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," Electronics (Switzerland), vol. 10, no. 5, pp. 1–19, 2021, doi: 10.3390/electronics10050593. [CrossRef]
20. W. Wang and Y. Lu, "Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model," IOP Conf Ser Mater Sci Eng, vol. 324, no. 1, 2018, doi: 10.1088/1757-899X/324/1/012049. [CrossRef]

AUTHORS PROFILE

Lu Xiong is an assistant professor in actuarial science at the Department of Mathematical Sciences and a faculty member of the Computational and Data Science Ph.D. program at Middle Tennessee State University. He's also a credentialed actuary who earned an associate membership of the Society of Actuaries (ASA) in 2021. His research interests are business data analytics, actuarial science, and computational science.

Kechen Tian is currently studying in the Department of Mathematical Sciences at Middle Tennessee State University and has worked on machine learning projects using Python in his senior year of college. His major is Actuarial Science, and he also minors in Risk Management and Data Science. He was born and brought up in China. He completed his first two years of college at Ningbo University in 2021 and came to MTSU the same year to finish his degree.

Yuwen Qian is currently a graduate assistant in the Department of Mathematical Sciences at Middle Tennessee State University, TN, US. She worked on a project examining the impact of COVID-19 on insurance rates as part of a summer internship. By utilising spreadsheets, the relationship between COVID-19 cases and insurance rates was analysed. Yuwen was born and grew up in Taiyuan, Shanxi, China. She finished her undergraduate degree in Taiyuan and her graduate degree in Wuhan, China.

Wilson Musyoka is currently a student and teaching assistant in the Department of Mathematical Sciences at Middle Tennessee State University, TN, US. He has previously worked in the Casualty Actuarial Society (CAS) summer intern program, where, along with other students, he developed a phone warranty plan. Wilson was born and grew up in Mombasa, Kenya. He graduated with a bachelor's in mathematics from Cumberland University and is currently on course to complete his master's degree in Actuarial Science.

Xingyu Chen is currently studying Actuarial Science at Middle Tennessee State University and participated in a machine learning project using Python, R, and other data analysis methods. She completed her bachelor's degree in Economics from Guangxi University in Nanning, Guangxi, in 2022 and finished her secondary education at Nanning No. 3 High School in Nanning, Guangxi, in 2018.