Prediction and Analysis of NFL Game Results Based on R Language:

Comparison among Classification Methods

Jilan Lang, Jiahao Wei, Han Bao, Linhan Cai, Jiaqi Zhao

Statistic and Operation Research Department, UNC-Chapel Hill

STOR 565: Machine Learning

Dr. Zhengwu Zhang

May 1, 2022

Appendix

1. Motivation and Goals
2. Dataset Overview
3. Exploratory Data Analysis
4. Data Cleaning
5. Model Construction
    - Logistic Regression
    - LDA/QDA
    - KNN
    - SVM
    - Neural Network
6. Reflect

Motivation and Goals

The National Football League is a professional American Football League that consists of 32 teams. It is one of the major North American professional sports leagues and the highest professional level of American football globally. The NFL has a very long history since its creation in the 1800s. It is not only popular in the USA but also has popularity worldwide, especially in Europe, Asia, Africa, and Australia. Moreover, it is also one of the favorite sports for betting. According to the current estimates, the NFL rakes in a massive 2.58 billion dollars of legal gambling and an additional 380 billion of illegal gambling annually. So our team believes that this industry is very intriguing and worth studying. The goal of this project is to build a system to predict the outcome using machine learning techniques such as logistic regression, Linear Discriminant Analysis(LDA) /Quadratic Discriminant Analysis(QDA), K-Nearest Neighbors Algorithm(KNN), Support Vector Machine(SVM), and Neural Networks(NN). We are attempting to answer the question: How can the variables in our dataset be used to train a model that can predict the game result. Which machine learning techniques could predict the game result with the highest accuracy or AUC value? We will use Principal Component Analysis (PCA) to catch the direction that explains the most data variation. Logistic regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and K-Nearest Neighbors will be used as the old supervised learning methods. The Support Vector Machine, as the new method, will be used to classify the result type by locating linear decision boundaries between these two subtypes. Neural Network will be used to classify data when they have a labeled dataset to train on.

We now describe several important terms that we use in this paper.

Accuracy: The proportion of correct predictions in all predictions made;

Specificity: The proportion of actual negative cases that have gotten predicted as negative by our model;

Sensitivity: The proportion of actual negative cases that have gotten predicted as negative by our model;

AIC: An estimate of prediction error. It estimates the quality of the model, relative to other models;

AUC: Area Under Curve. The value of AUC ranges from 0 to 1

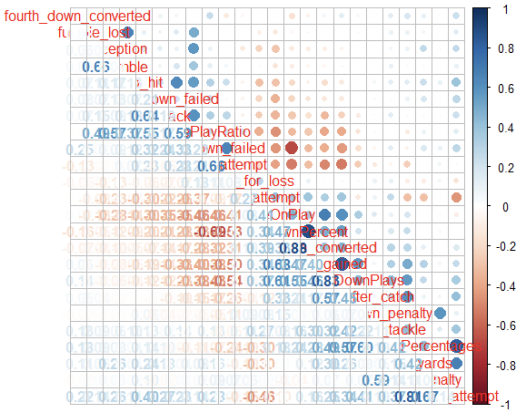
Dataset Overview

This dataset comes from an R library called nflfastR. There are over 300 variables and hundreds of thousands of data records. The dataset includes play-by-play data, such as play_id, home_team, away_team, etc. Some valuable variables such as EPA and results contain scores and statistics important in our prediction.

Exploratory Data Analysis

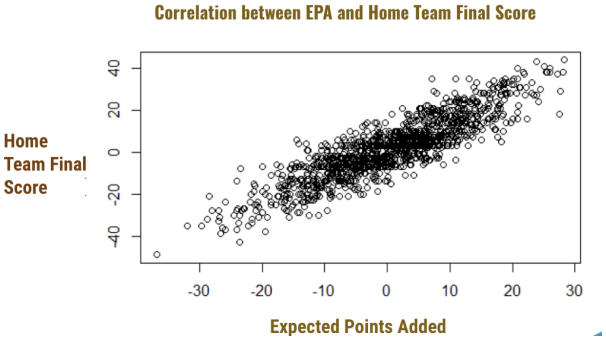
First, to visualize the relationship across all variables, we created a correlation plot for all variables. However, since our raw dataset contains too many variables, it's challenging for us to extract insightful findings from this graph. So we decided to take a step back and observed the relationship between only two variables and conducted repetitive work throughout all variables. We discovered a strong positive correlation between EPA and the Home Team Final Score.

Image 1: Correlation Graph of All Variables



EPA stands for the Expected Points Added. It's the difference between a team's Expected Points at the end of a play and their Expected Points at the beginning of a play.

Image 2: Correlation Graph between EPA and the Home Team Final Score



Because our research interest is to predict future game results, we want to see win and loss ratios across years and teams before we conduct a deeper analysis. We randomly picked

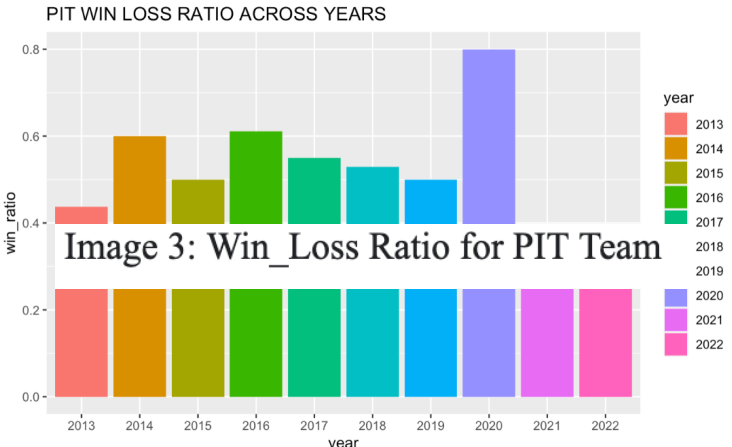


Image 3: Win_Loss Ratio for PIT Team

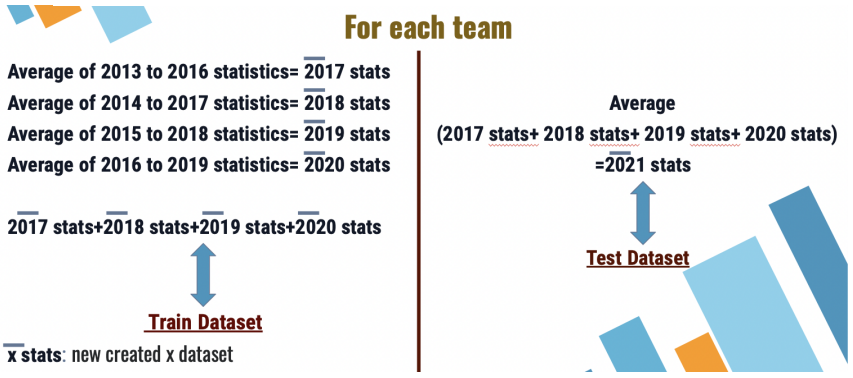
the “PIT” Team as an example and utilized the raw dataset to calculate its win-loss ratio as a home team. In this graph, we can see that before the pandemic(2020), the win-loss ratio for PIT was relatively stable. We could ignore the statistics for 2022 because our dataset only contains plays in January, so it’s not the overall win-loss ratio for the whole year. Though there’s a vast difference between 2020’s win-loss ratio and 2021’s, it’s probably a consequence of COVID19. We observed the same significant variance between 2020’s results and 2021’s for all teams. The reason for the large gap between the 2020s and the 2021s statistics was unknown but provided a possible direction for future study. In conclusion, this graph helped to support the argument that it’s reasonable for us to use historical data to predict future outcomes.

Data Cleaning

Before we started to construct models, we needed to solve two major problems. First, The entire dataset has nearly 430 thousand rows and about 370 variables. To be more concise, we decided only to include rows that contain information about every game instead of every play and variables related to games. The original PBP dataset shows data for each play. So we only chose the rows with zero remaining time of a game. Then, we regarded the variables containing the word “total” in that row as the final stats of this game, respectively.

Besides, EPA is generated during each game. As a result, it has a high correlation with the game result, but we do not know the EPA of a game that has not started. Therefore, we can’t use EPA directly to predict the game result. Therefore, we decided to use the average value of each EPA variable as the predictor because it can be calculated before the game. We calculated the average values of the actual stats of all competitions that one team played in 4 seasons before the aiming season, which were 64 games. And we viewed those values as the stats of this team for all games it would play in the aiming season. At the same time, we kept the actual result of each game as the response. In this way, we reduced our dataset to only about 1300 rows and 29 variables(including Home

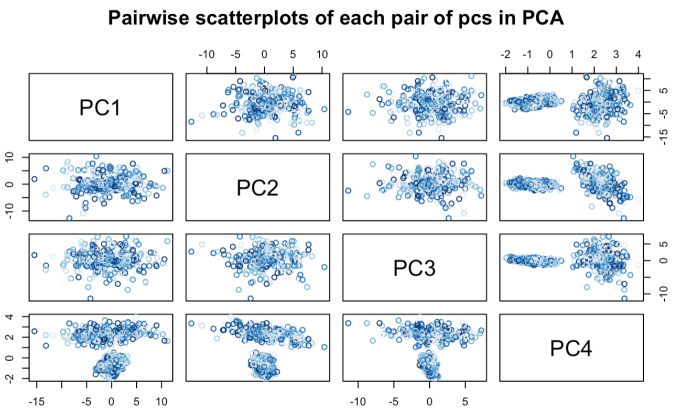
Image 4: Data Cleaning Flow Diagram for EPA Related Varia

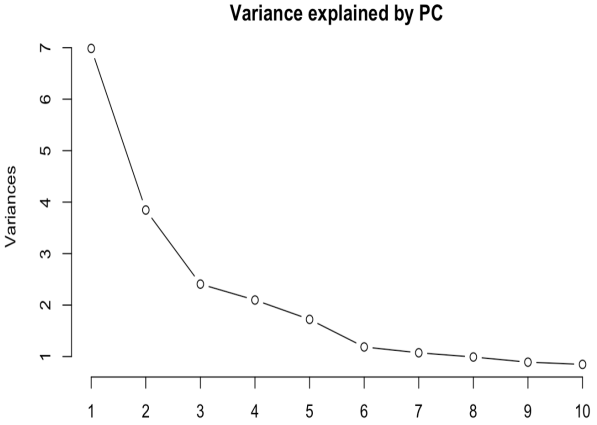


and Away teams' Names). Finally, we utilized the new datasets of seasons from 2017 to 2020 in the Training data and 2021 as the Testing data.

Principal Component Analysis

One of the most essential steps in EDA is to detect the separability of the dataset. We





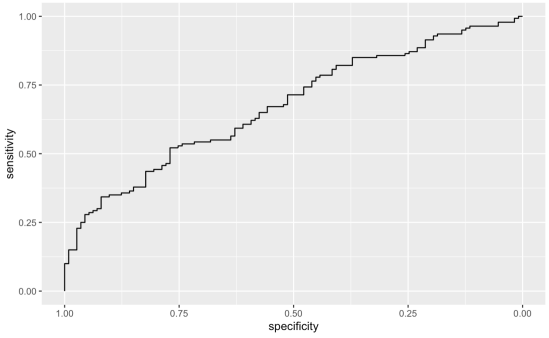
adopted Principle Component Analysis(PCA) to reveal this. “Pairwise scatter plots of each pair of pcs in PCA” show that our dataset is linearly separable as there is a clear boundary between light blue and dark blue dots in the first two principal components. The first four PCs explain most of the variance from the dot plot of variance explained by PCs. So PCA is a helpful technique to help us reduce the dimension and explore more insights into the data.

Model Construction

Logistic Regression

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. (George Lawton, Ed Burns, Linda Rosencrance, 2022) A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. We built two models due to the significance of the variables. For the first model, we included all the variables, and we received a model with an accuracy of 0.62, a sensitivity of 0.57, a specificity of 0.65,

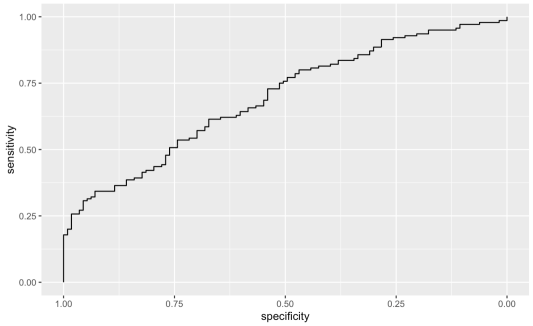
Image 7: ROC curve for first logistic regression model



and an AIC value of 1030; The area under the curv3(AUC) we obtained through this model is 0.676;

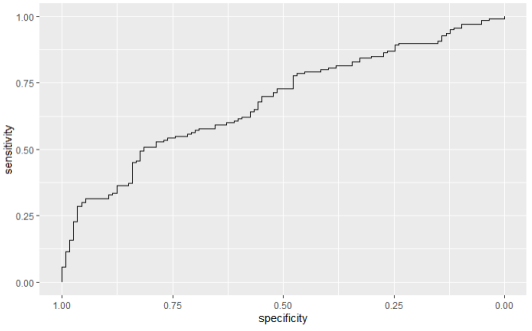
We then constructed a new logistic regression model with higher accuracy by removing less significant variables. The new model had an accuracy of 0.636, a sensitivity of 0.60, a specificity of 0.66, and an AIC value of 1011. The area under the curve (AUC) we obtained through the new model is 0.695; Comparing the two logistic regression models, we could easily observe that the second model had higher accuracy and more balanced sensitivity and specificity with significant variables selected. We could build the model easier by avoiding unnecessary variables and yielding similar prediction accuracy. According to the principle of parsimony, simple models with fewer variables are preferred over complex models with many variables.

Image 8: ROC curve for second logistic regression model

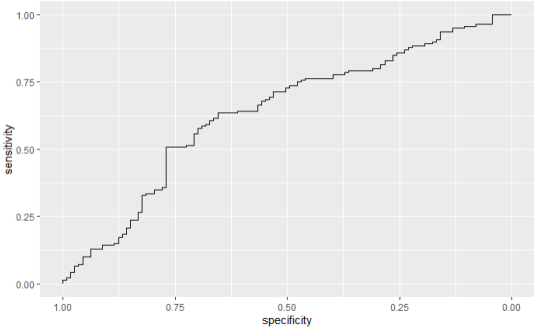


LDA/QDA

Linear discriminant analysis (LDA) is a linear classification and dimensionality reduction model. It is most commonly used for feature extraction in pattern classification problems, perfectly fitting our objective - predicting win or loss. By applying LDA to our dataset, we could obtain a model with an accuracy of 0.621, a sensitivity of 0.522, and a specificity of 0.7. The area under the curve (AUC) we obtained through LDA is 0.6834.



Besides, quadratic discriminant analysis (QDA), assuming conditional multivariate normal distribution and using quadratic surfaces to separate the data, would apply to our dataset. We got the QDA model with an accuracy of 0.585, a sensitivity of 0.327, and a specificity of 0.793. The AUC we calculated here is 0.6376. Compared with these two models, we could find



LDA has a higher AUC, which on the other hand, reflects the solid linear correlation we observed initially from EPA and Home Team Final Score.

KNN

K Nearest Neighbors is a machine learning technique and algorithm used for both regression and classification tasks. K-Nearest Neighbors examines the labels of a chosen number of data points surrounding a target data point to make a prediction about the class that the data point falls into. When we performed our KNN models, we restricted the value of K from 1 to 10. Besides, we selected Specificity, Accuracy, and Sensitivity as the measure of performance. After running all ten models, we visualized the result in this graph.

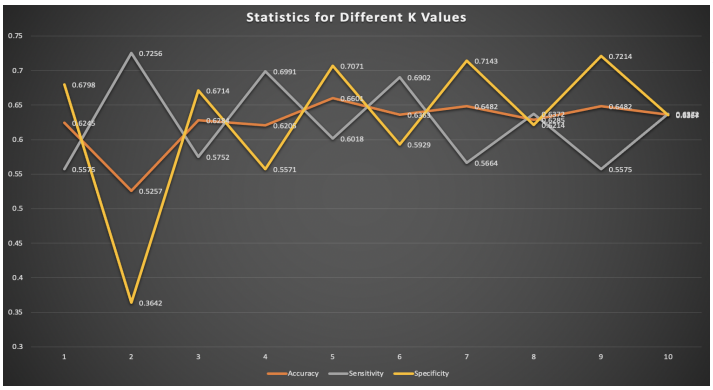
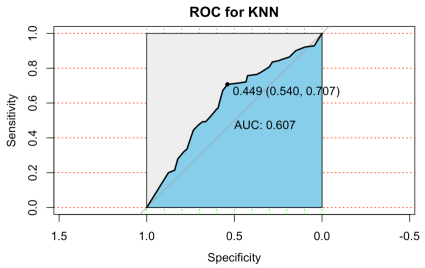


Image 11: KNN Model Results

When K equals 5, we achieved the highest accuracy, which is 0.66.

Besides, we took a step further by measuring the AUC for our optimal KNN model. This helped us conduct comparisons and contrast at the end of our project. The value of the KNN model when K equals 5 is 0.607.

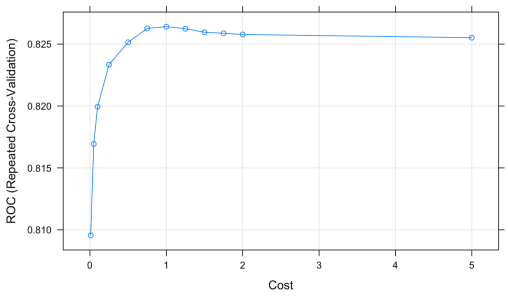
Image 12: AUC Result for KNN Model

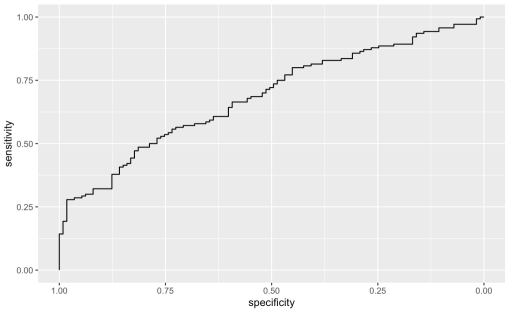


SVM

Support Vector Machine is the new supervised learning method we adopted. Different from Quadratic Discriminant Analysis or LDA, which requires some statistical assumptions, SVM tries to understand the data from a geometric point of view, meaning that we are trying to locate the decision boundary by maximizing the distance between the boundary and margins. The first thing we need to decide in SVM is the kernel function. After comparing

the performances in both the train and test sets, we found that linear was the most appropriate kernel for this situation. It scored 0.62 in accuracy and 0.69 in AUC.





Neural Network

The neural network is a technique in machine learning that simulates the circuit of biological neurons. In other words, an artificial neural network. In our project, there are many variations we have to consider. Among all the neural networks we learned, while CNN is most commonly applied to analyze visual imagery, RNN usually deals with sequential or time-series data, we have decided to dig into NN and DNN. For both models, we had to decide how many units to add to each hidden layer, and for DNN, we had to determine how many layers to apply. From what we learned in class and doing the research ourselves, we decided to set the units around the size between the input and output layers. Since we have twenty-five nodes (variables) in the input layer and one node in the output layer, we tried around thirteen hidden layer units. It turned out that thirteen units give the best prediction accuracy. Layer-wise, we tested our model by increasing the number of layers gradually. Our model performed worse when adding layers. Thus, thirteen hidden units and one layer became our best model.

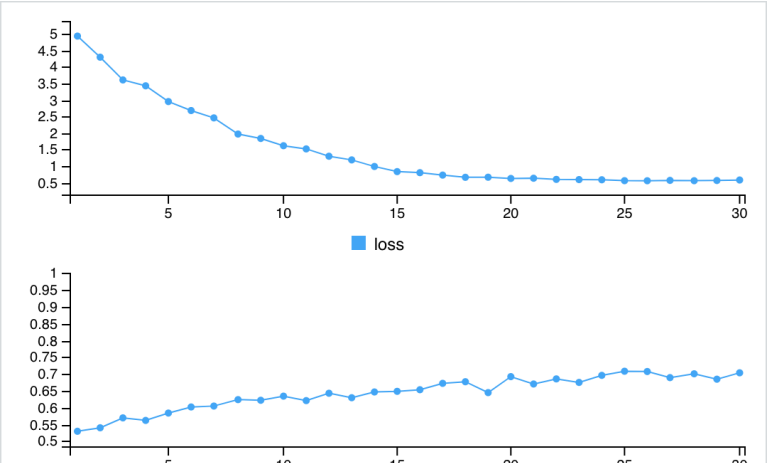


Image 15: Neural Network Accuracy and Loss during 30 Epochs

Furthermore, we studied and tried out different activation functions. An activation function is a function that defines how the weighted sum of the input is transformed into an output from the nodes in the defining layer. For the hidden layer, we chose the rectified linear activation function (ReLU). This is because ReLU is simple to implement and effective in

most cases. Among the regular functions we tested, such as Sigmoid and Tanh, ReLU turns out to be the best option. For the output layer, we tested Sigmoid and Softmax (especially for DNN), and Sigmoid gives us the best model. Eventually, the single-layer NN model with thirteen units applying ReLU and Sigmoid activation functions performs the best, giving an accuracy of around 0.65. The AUC is 0.6945.

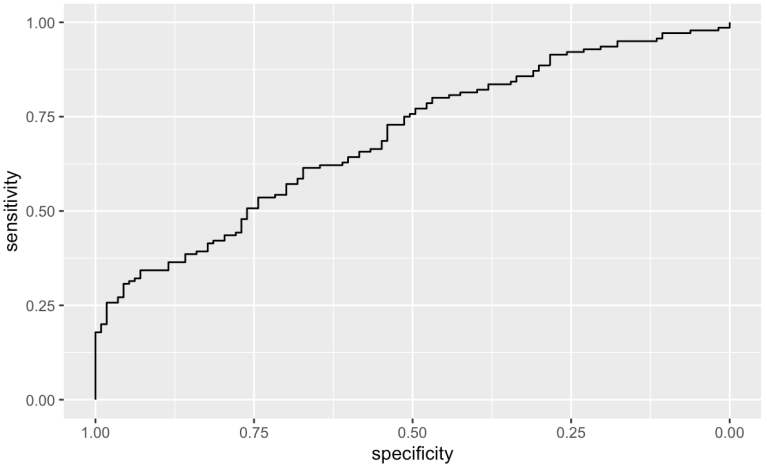


Image 16: ROC of NN model

Discussion and Future Work

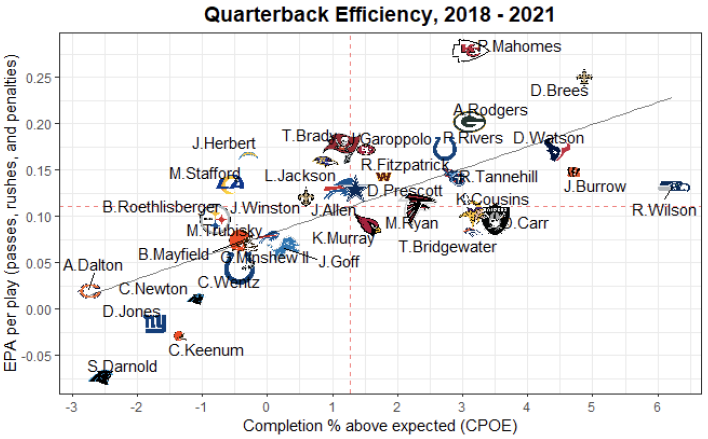
As we know, accuracy looks at fractions of correctly assigned positive and negative classes. If our problem is highly imbalanced, we get a high accuracy score by simply predicting that all observations belong to the majority class. (Jakub Czakon, 2021) In our problem, ROC and AUC are especially good at ranking predictions. By comparing these five models, we could conclude that the Neural Network model has the highest AUC of 0.6945 and relatively good accuracy of 0.65. This is not surprising because neural networks simulate how a human thinks. The hidden layer provides more accurate predictions than other classifier and regression methods.

Admittedly, our prediction remains space for improvement.

First, EPA, our decisive variable, may lose its credibility in recent periods. As we employ the average EPA to evaluate the performance (or say competitiveness) of a team, the most straightforward proof of a team's competitiveness is its win-loss ratio. However, according to the graph "Win_Loss Ratio for PIT Team" shown initially, the win-loss ratio from 2013 to 2019 is quite stable but fluctuates a lot after 2020. The explosion of the pandemic can generally explain this instability, and we believe the win-loss ratio would become balanced in the following years, as people adapt to living with the existence of the

COVID and each team goes back to its routine mode. Consequently, it is indispensable to keep studying “Postseason” data to further check the universality of our methods.

Moreover, while our research strictly focused on the application of EPA, quarterback rating (QBR) would be another essential indicator for predicting game results.



From this graph, we can see that sometimes excellent quarterback players may primarily affect the EPA rating. The joining of QBR in the future could complement our logic of methods and strengthen the accuracy and persuasiveness of our prediction.

At the end of this paper, we would like to thank professor Zhengwu Zhang, who supports us with solid fundamental knowledge in machine learning and guides us through the research direction and concrete analysis. Your insightful feedback pushed us to sharpen our thinking and brought our work to a higher level. We would also like to thank two IAs Brian White and Adrian Allen who provided us with the tools that were needed to choose the right direction and successfully complete our research.

Reference

1. George Lawton, Ed Burns, Linda Rosencrance. “Logistic Regression”. 2022. What is Logistic Regression? - Definition from SearchBusinessAnalytics
2. Jakub Czakon. “F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose?”. 2021. https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc