Using The machine learning algorithm to predict the Win-loss percentage and ERA for the team from Korean Baseball Organization(KBO)

Min Set Aung st122825,Albin Raj Maskey st123469

1.Introduction

In this project we predict key performance metrics in baseball, such as Earned Run Average (ERA) and Win-Loss Percentage, which is a compelling endeavor that delves into the intricate dynamics of the sport. Baseball, renowned for its statistical richness, offers a fertile ground for leveraging historical data and advanced analytics to gain insights into team performance.

Earned Run Average (ERA) serves as a pivotal indicator of a pitcher's effectiveness, encapsulating their ability to prevent opposing teams from scoring runs. On the other hand, Win-Loss Percentage provides a broader perspective on team success, reflecting the proportion of games won relative to the total played. Both metrics are crucial in assessing the overall competitive prowess of players and teams in the Korean Baseball Organization(KBO)

In this pursuit, the main objective is to develop predictive models using machine learning to anticipate a pitching team's average ERA and a Win-Loss Percentage. By scrutinizing a wealth of historical data, encompassing factors such as team statistics, we find patterns, trends, and contributing factors that influence these critical metrics.

2.Problem statement

The objective of this study is to develop predictive models that accurately forecast the Win-Loss Percentage and ERA of KBO teams based on historical data. By leveraging machine learning algorithms and relevant baseball statistics, we aim to create models that enhance the ability to anticipate a team's performance in terms of wins and losses during regular seasons.win-loss percentage and ERA is the important factors for pitching teams and team management.Team management can evaluate the pitching team performance and they can think how to get into the No1 seed in the league to get into playoff and how can they improve the pitching team to get better results in the season.

The win-loss percentage in baseball is a measure that indicates the success of a team in terms of the ratio of games won to the total number of games played. It is a commonly used statistic in baseball to evaluate a team's performance over a season. The formula for win-loss percentage is:

$$\text{Win-loss percentage}=\frac{number\ of\ win}{total\ games\ played}$$

ERA stands for Earned Run Average, and it is a key statistical measure used in baseball to evaluate the effectiveness of a pitcher. The Earned Run Average is calculated based on the average number of earned runs a pitcher allows per nine innings pitched

$$\text{ERA}=\frac{Earned\ runs}{inning\ pitches}\times 9$$

3.Related work

Sports outcome prediction has gained significant attention due to its relevance in various fields, including sports analytics, betting, and fan engagement. Stekler et al. (2010) highlight the growing popularity of sports result prediction, particularly in the context of sports betting. In the realm of baseball, Major League Baseball (MLB) stands out as a multi-billion dollar industry, prompting researchers to explore predictive systems for specific game outcomes (Baumer & Zimbalist, 2014).Sabermetrics, the empirical analysis of baseball statistics, has become integral to understanding and predicting baseball outcomes (Wolf, 2015). While sabermetrics has been extensively used for player evaluation, its application in predicting game results poses unique challenges. Studies by Menéndez et al. (2015) and Yang and Swartz (2004) attempted to measure the impact of variables associated with baseball games, emphasizing the need to evaluate team performance comprehensively.Despite the abundance of baseball data, predicting game outcomes remains a challenging task (Sykora et al., 2015). The difficulties arise from the intricate selection of factors influencing game results and the potential overlap of data elements in predictive models. The application of machine learning techniques to baseball data has witnessed active research, with ongoing efforts to enhance predictive accuracy (Ockerman & Nabity, 2014; Robinson, 2014).While previous studies have made strides in understanding baseball through sabermetrics and economic analysis, there exists a research gap in systematically applying advanced data mining methods to predict win-loss percentages and Earned Run Average (ERA) in MLB games. This study aims to contribute to the existing body of knowledge by evaluating the effectiveness of machine learning models in forecasting these critical baseball metrics.

4.Methodology

4.1 Exploratory Data Analysis (EDA)

In this project we use random forest,linear regression and KNN method to predict the winning percentage and ERA of the team.After we download data source, we started the etl(Extract,Transform and Load .We use ETL process to clean the NaN(Not a number) value and find the data type of each column and drop the unnecessary columns(feature) in the data source to clean the data which will make

EDA process easier.After that we use label encoder in team column and add team number column to make model training convenient. We choose the specific columns such as(year','team_number','wins','losses','games','saves','runs_per_game','innings_pitched','earned_runs','win_loss_percentage','ERA')to start the data training.We also use correlation matrix to see the impacts of features to our targets.

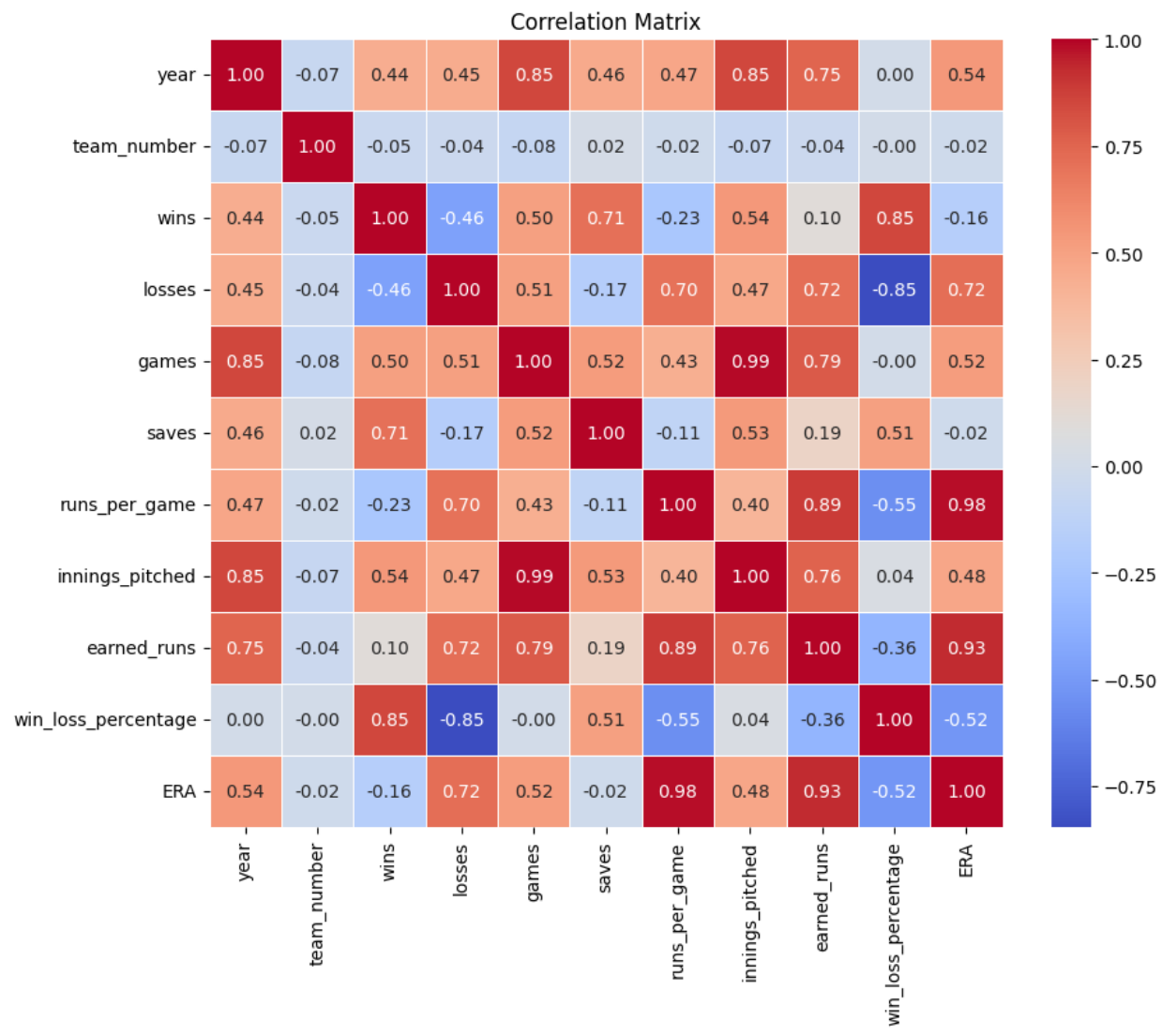

Fig 3.1 Correlation matrix of features

4.2 Data training and model evaluation

We extract relevant columns for predicting 'win_loss_percentage' and 'ERA' from the DataFrame and the columns we chose for 'win_loss_percentage' prediction are 'year', 'team_number', 'wins', 'losses', 'games', and 'saves' and for 'ERA' prediction are 'runs_per_game', 'innings_pitched', 'earned_runs', 'year', and 'team_number'.After that we build the standard scalar function from scikit-learn to build the pipeline for using KNN for ERA and Linear Regression for win-loss percentage prediction and then we split the

data into training data and test data and train the model on pipeline that we created with training data.After this process we used test data 0.0001918736721810483 8set to evaluate the performance of the model that we created.

5 Results and discussions

	Linear Regression	Random Forest
Cross validation Score (Negative means square)	0.00019	0.00036
Mean square error	0.00013	0.00023

Table 4.1 Mean Squared Error for Win-loss prediction

	Linear Regression	Random Forest
Cross validation Score (Negative means square)	0.00019	0.00036
Mean square error	0.04519	0.0151

Table 4.2 Table Mean Squared Error for ERA prediction

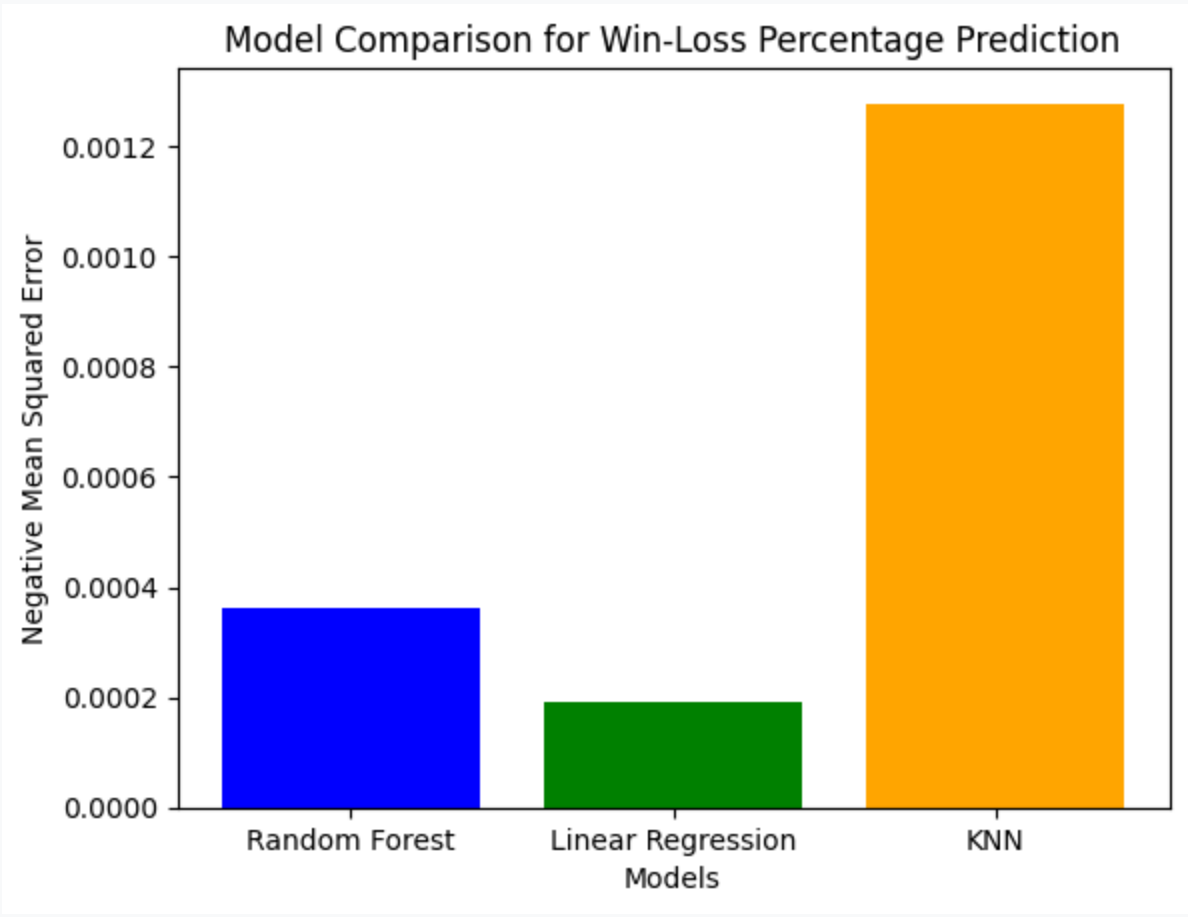

Fig 4.1 Comparison between Negative MSE of three models

As we can see in Table 4.1and 4.2,we can see the scores for Win-loss percentage and ERA prediction The Random Forest model has a lower cross-validation score (negative mean squared error), suggesting that it performed better in terms of prediction accuracy during cross-validation compared to Linear Regression.The mean squared error values also indicate that the Random Forest model has a lower error on average compared to Linear Regression.In summary, based on the provided results, the Random Forest model appears to be more suitable for the given prediction task, as it shows better performance in terms of both cross-validation scores and mean squared error.

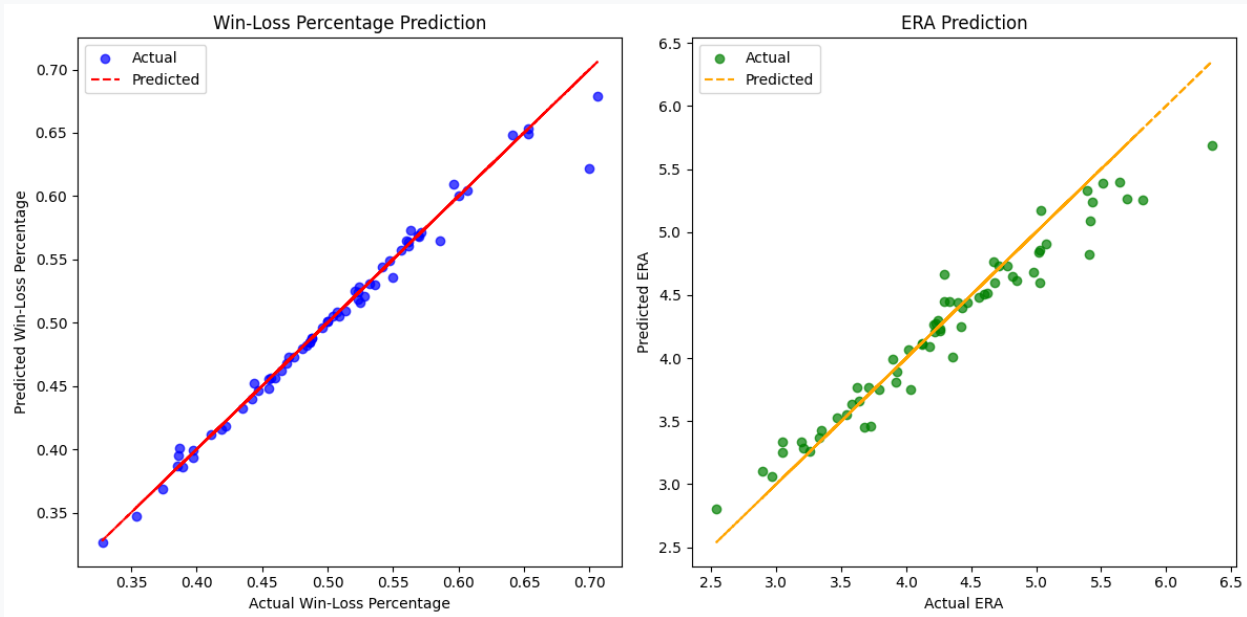

Fig 4.2 The result of prediction for Win-loss percentage and ERA

In the Fig 4.2 we can see the results of prediction for Win-loss percentage and ERA,we can see that results are quite accurate because of smaller MSE results.

6. Conclusions and future works

In conclusion we build two different models to predict the win-loss percentage and ERA of the pitching teams of the KBO.because of the small data set,the models work efficiently and MSE results are small enough to evaluate that our prediction is quite accurate.We need more business understanding to use some of the features to build complex models and get better results.Because of the small data source we have limitation in feature selection and find difficulties because of our business understanding.To predict the team win-loss percentage we need batting and fielding team.We can also build complex deep learning models like CNN and ANN to obtain the better results for win-loss percentage and ERA.