WNBA Playoffs Machine Learning Predictions

M.EIC- Machine Learning

Luca Di Pietro (up202401432) IF: 50%
Noemi Messori (up202401607) IF: 50%

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Table of content

01

Domain Description

01 Domain Description

We aim to predict which teams will make the **playoffs** in the Women's National Basketball Association (WNBA).
To do this, we analyzed 10 years of data, including team stats, player performance, and other factors.

In the WNBA, the playoffs are decided by the regular season standings. At the end of the season, the top 8 teams from both conferences combined qualify for the playoffs, based on their win-loss records. If teams are tied, league rules determine the tiebreakers.

02

Data Analysis

02

Data Analysis

We carried out exploratory data analysis using tools from the **pandas** library and by creating various plots. The analysis was performed on each dataset, and here is what we discovered:

- There are columns with no variability like lgID and some statistics;
- There are dead players;
- There are players with height/weight equal to 0;
- There are some columns with empty strings, others with default 0 values and other values that represent null;
- Some columns have binary (confID, playoff) or ternary (finals, semis, firstRound) values;

02

Data Analysis

- The number of games played by each team differs, so they are not directly comparable. Win rate should be used;
- Player attributes do not follow Gaussian Distributions ;
- There are teams that have changed their name;
- In the year 6 there are two winners
  - Connecticut Sun and Sacramento Monarchs
- There are players with no position and no college assigned
- Using some techniques like “zscore” we removed some outliers

03

Predictive Data Mining

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Problem definition

Objective

Create a **machine learning model** to predict which WNBA teams are **likely to qualify** for the playoffs.
The model should use historical data, including team statistics, player performance metrics, and other relevant factors from past seasons.

Specifics

Our purpose is to **predict** in the most correct way if a team is going to qualify for the playoffs.
In order to predict that we created a new dataset that **merged** that table with the features from the others and that was properly cleaned, filtered and prepared with the most relevant attributes.

Data Preparation

- Players who did not participate in any games during the years covered by the dataset were removed.

- Several columns contained **null values** that were not initially recognized as such. Dates formatted as 00-00-00, integers with a value of 0, and **empty strings** were all converted to **None**.

- Normalized attributes that followed a Normalized Distribution. Standardize attributes that didn't, linearly from 0 to 1

- We added up all the scores made by the individual players per team and **replaced** this result with the value of the points of the individual teams

Data Preparation

- We created new specific attributes, for *Teams*:
  Win rates, Total Player Insights (TPI), Numbers of Post Season's match

- We created attributes for the average weight and height of each team based on the players' stats, and we calculated the players' ages

- Some attributes were merged into new ones or converted in numerical format

- We divided the table into **East** and **West** to conduct a targeted study for each individual conference, and then we merged everything back together

- Last thing was adding a new column, called PlayOffNextYear in order to save, for each team, if they were qualified for the playoff of the next year

Experimental Setup

Choosing a prepared dataset

We chose a dataset that combines linearly normalized non-Gaussian attributes with standardized Gaussian attributes, using merged data that averages certain player and teams metrics.

Different classifiers

We experimented with different classifiers:
Logistic Regression, KNN, Random Forest, Naive Bayes, Gradient Boosting, Decision Tree, XGBoosting and SVM.

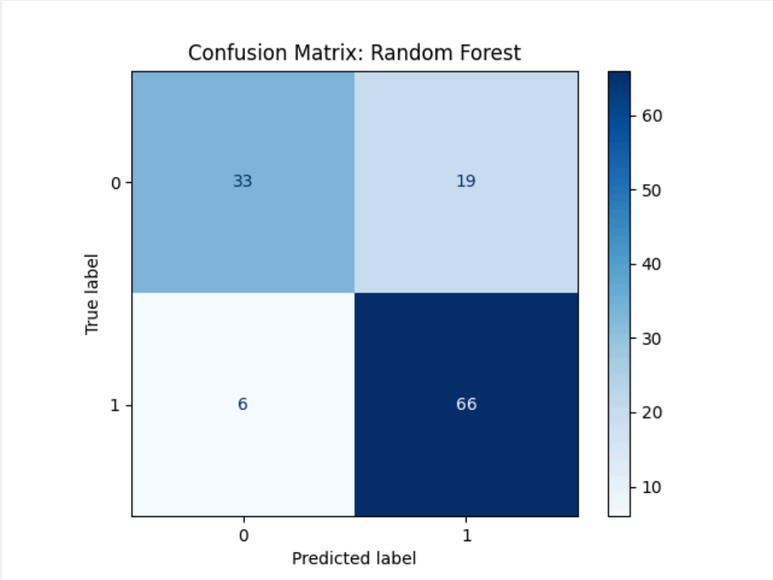
Just 8 teams are going to pass

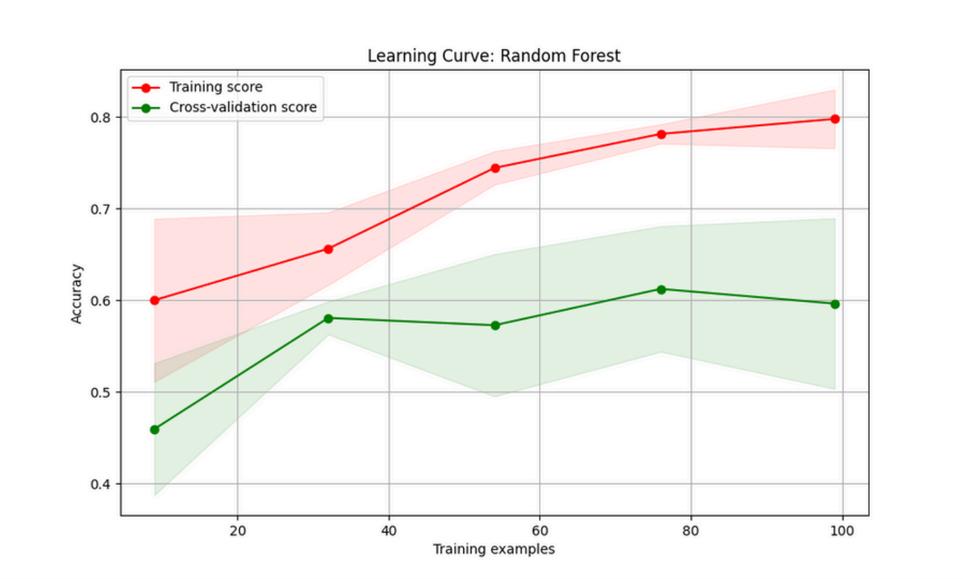
We took 4 teams from each confederation in order to guarantee exactly 8 teams with the 8 highest probabilities to pass.

Results

We selected the best-performing model, the Random Forest, and trained it using data with the approach that incorporates all the relevant data from the previous year.

On the training set the model achieved an accuracy of 81%





04

Future Works

04

Future Works

- **Ensemble Methods**: investigate the use of ensemble techniques to enhance accuracy and stability (e.g. Rolling Window) by combining predictions from multiple models.
- **Advanced Integration of Models**: examine the benefits of integrating different types of models using techniques like stacking or bagging, to achieve a better performance
- **Explainability of Predictions**: another important direction is to explore techniques that provide clear and meaningful explanations for the model's predictions.

05

Real World Applications

05

Real World Applications

- **Application for Bookmakers**
  - **Odds Calibration:** Use predictions to set more accurate odds for teams reaching the playoffs, reducing financial risks.

- **Sports Industry Applications**
  - **Team Strategy Optimization and Performance Analysis:** Coaches and managers can define strategies, player recruitment, and rotations based on the model, also identifying some key factors that can influence the success.

- **Economic Applications**
  - **Sponsorship, Marketing and Market Valuation:** Guide brands in sponsoring high-performing teams, influence ticket pricing and advertising based on predicted results.

06

Conclusions

06

Conclusions

Current state

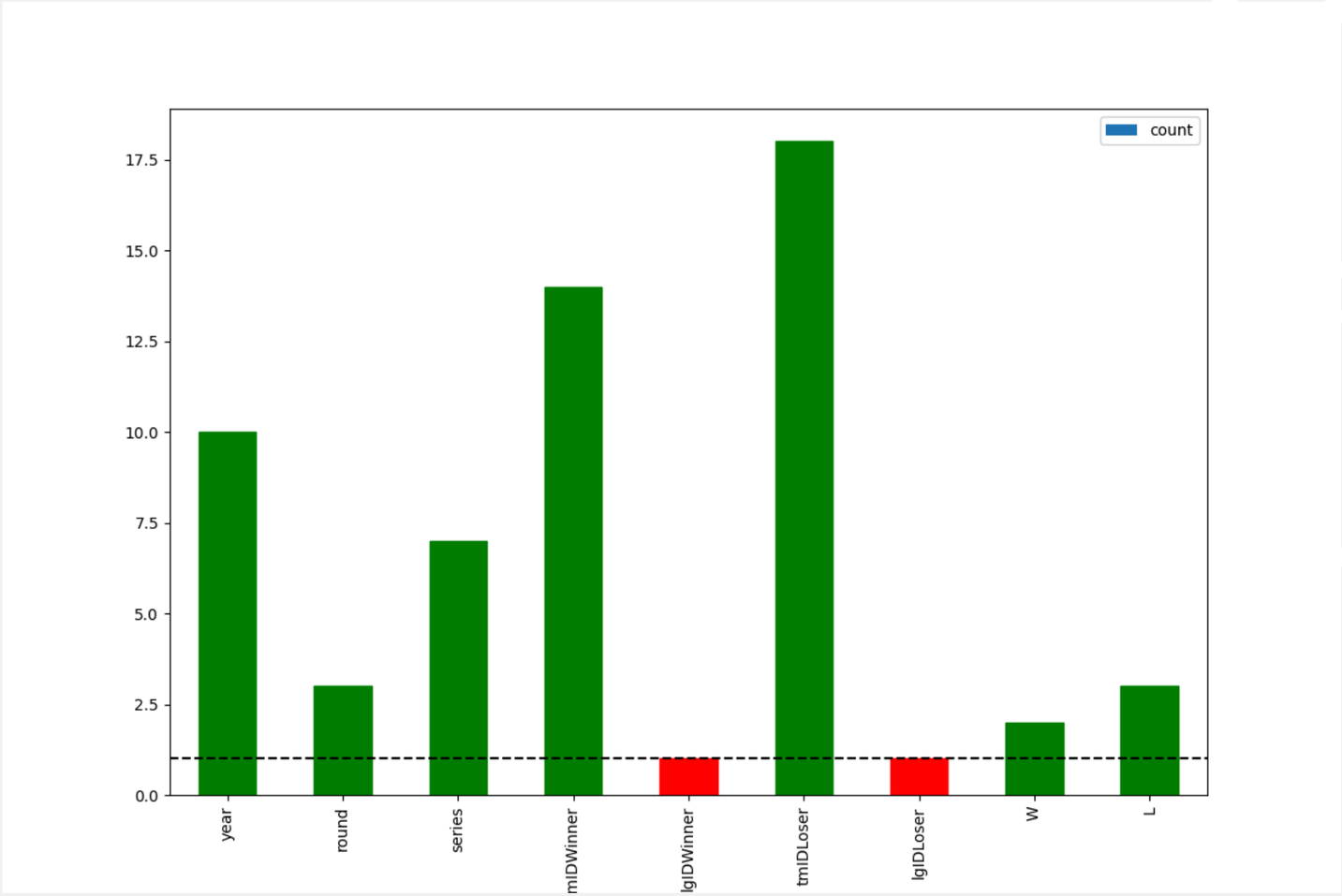
We conducted a thorough analysis and preparation of the dataset, ensuring a comprehensive understanding of both the data and our objectives. Additionally, we evaluated various models to identify the most effective approach for generating accurate predictions.

Current results

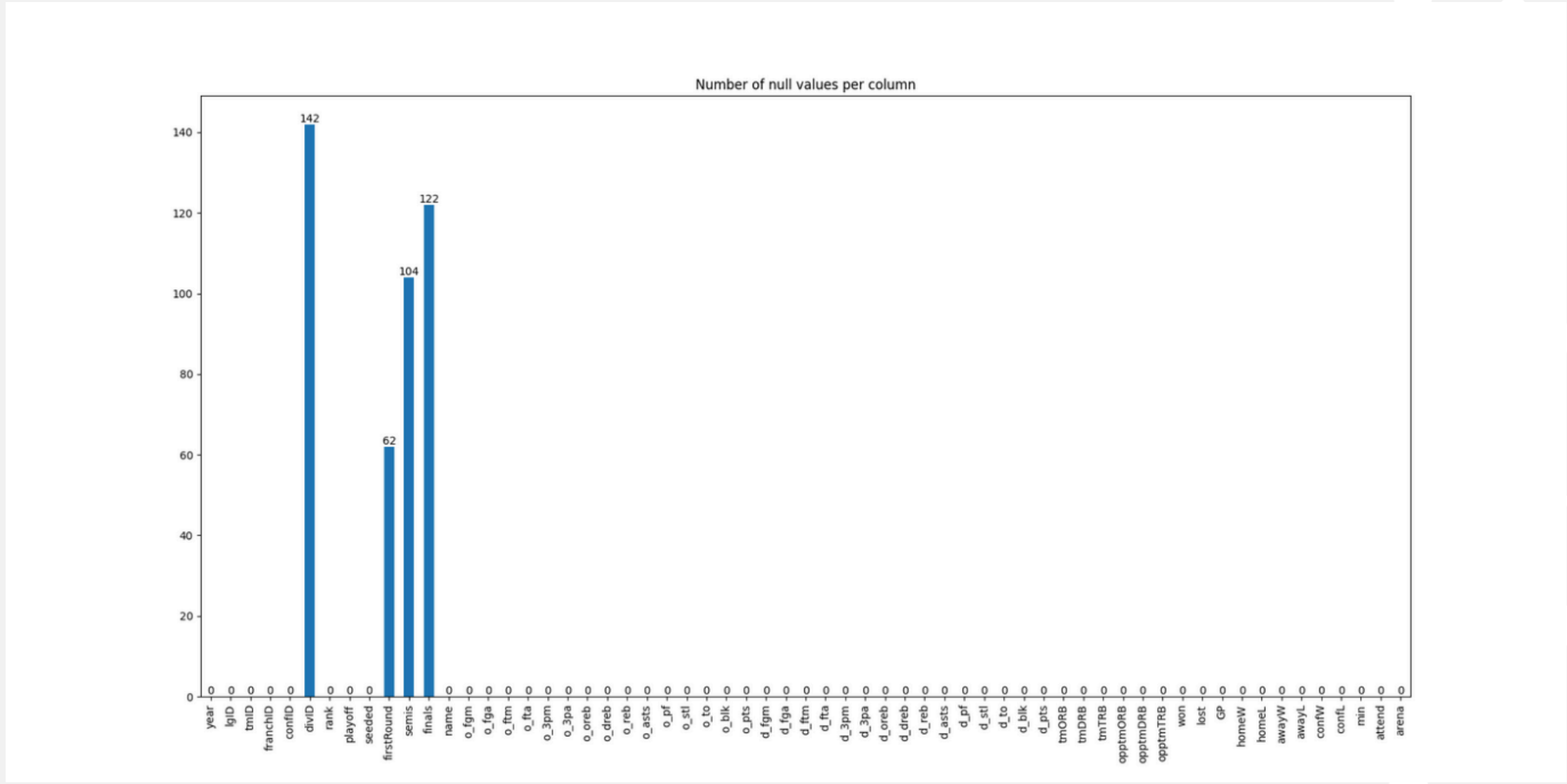
We are pleased with the current results, as the accuracy is quite satisfactory. However, we aim to further enhance the model's performance.

07

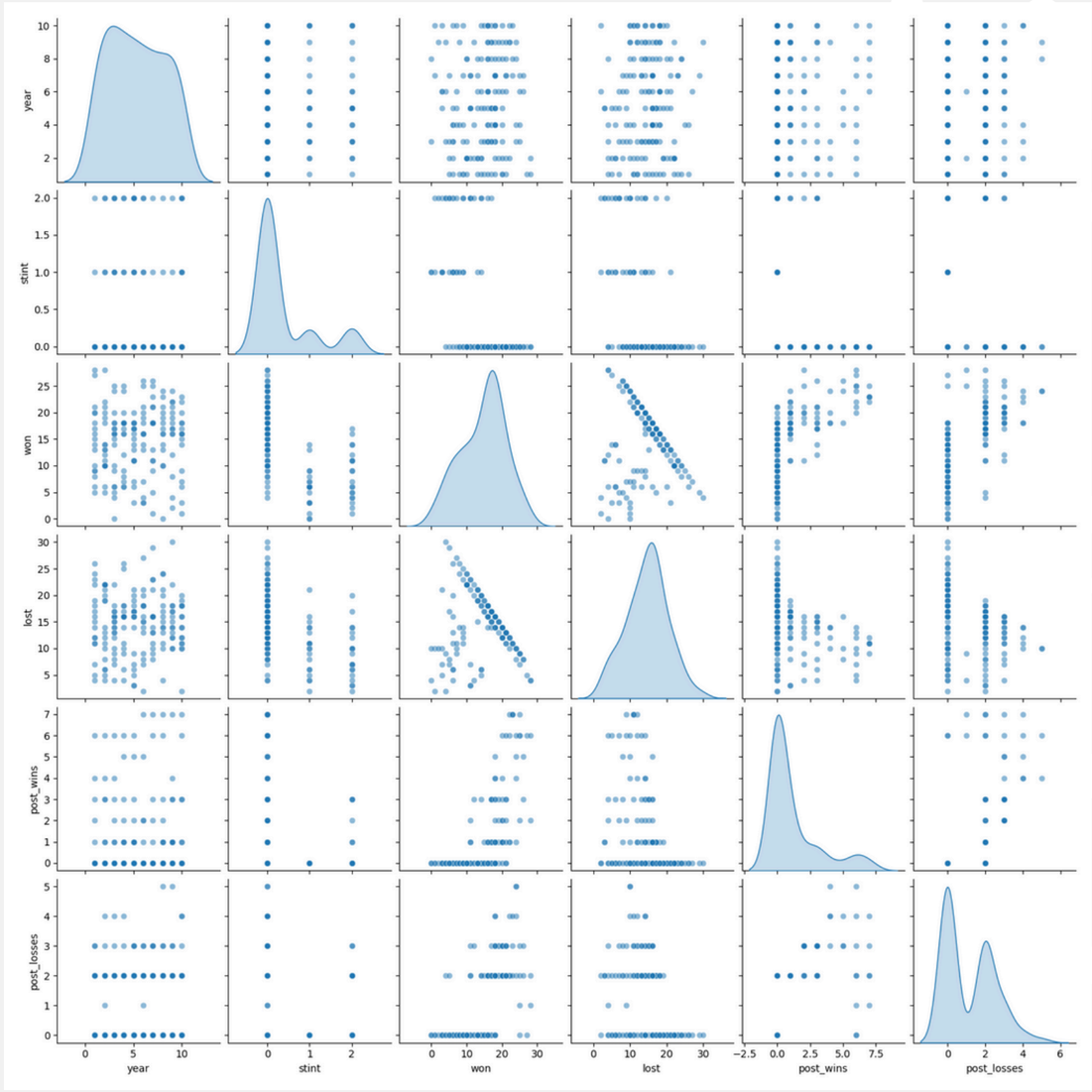
Annexes



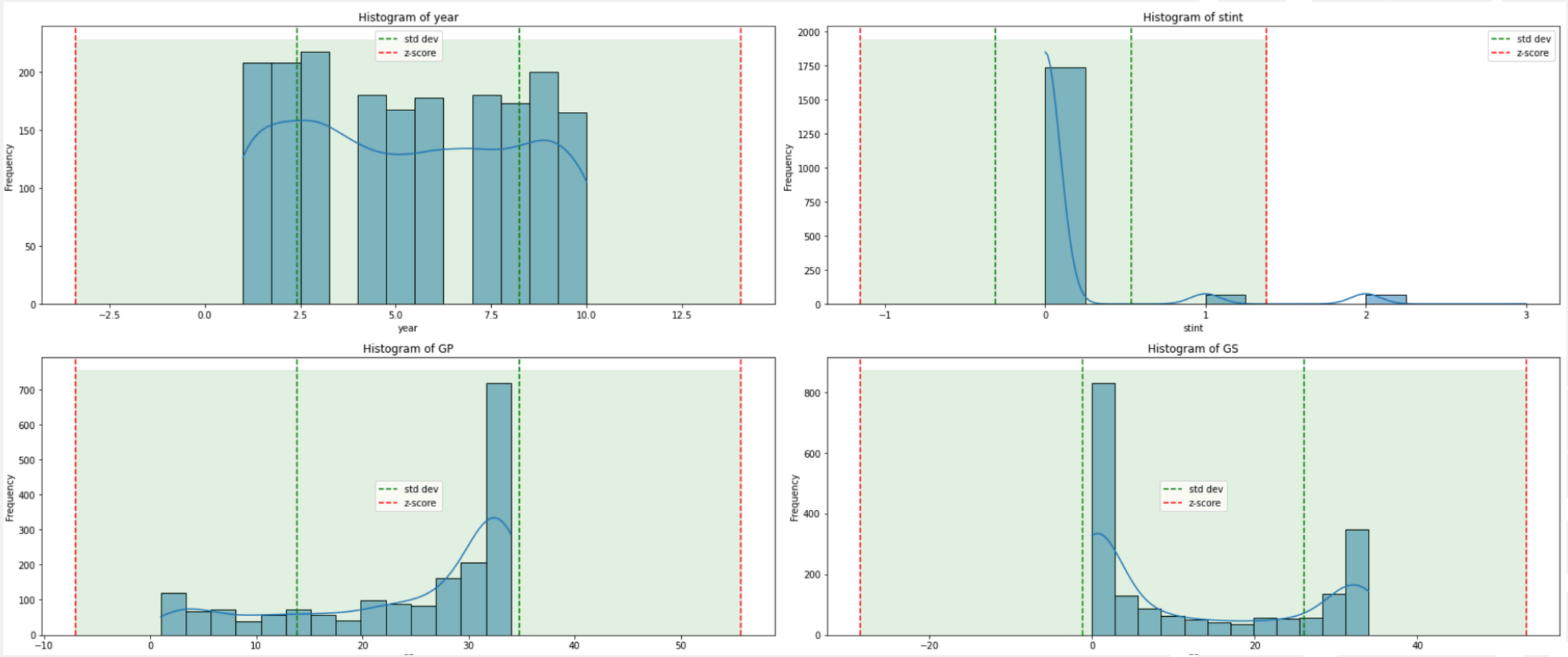
Red columns: attributes with only unique values



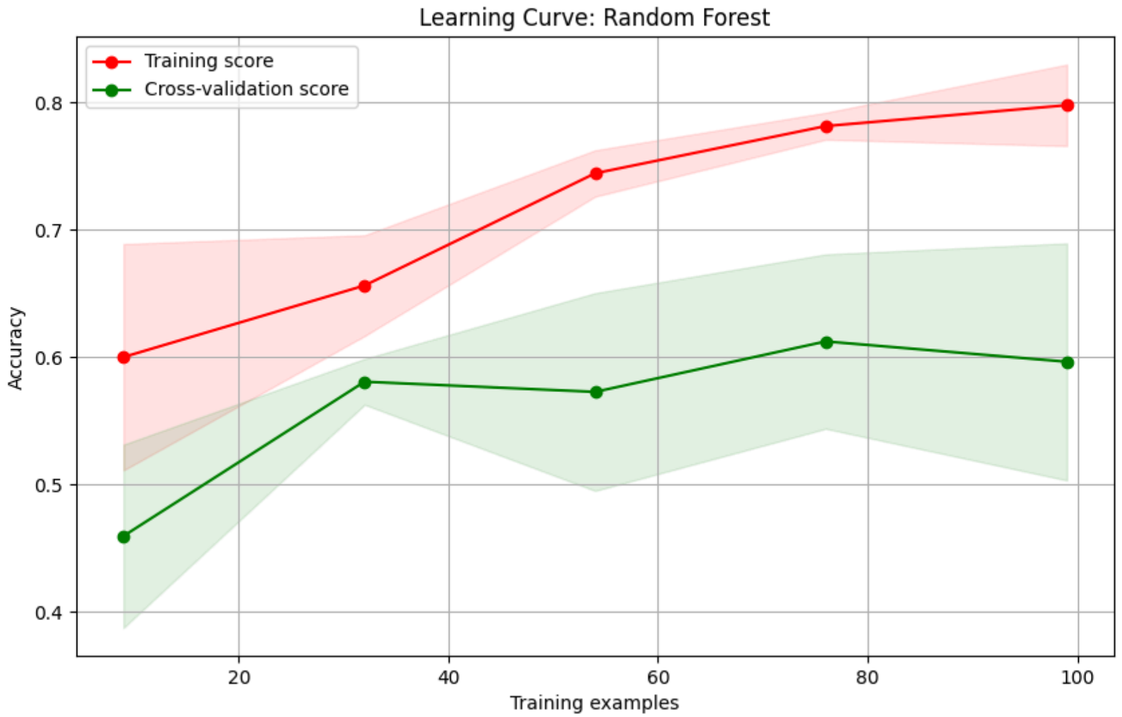
Number of null values per columns

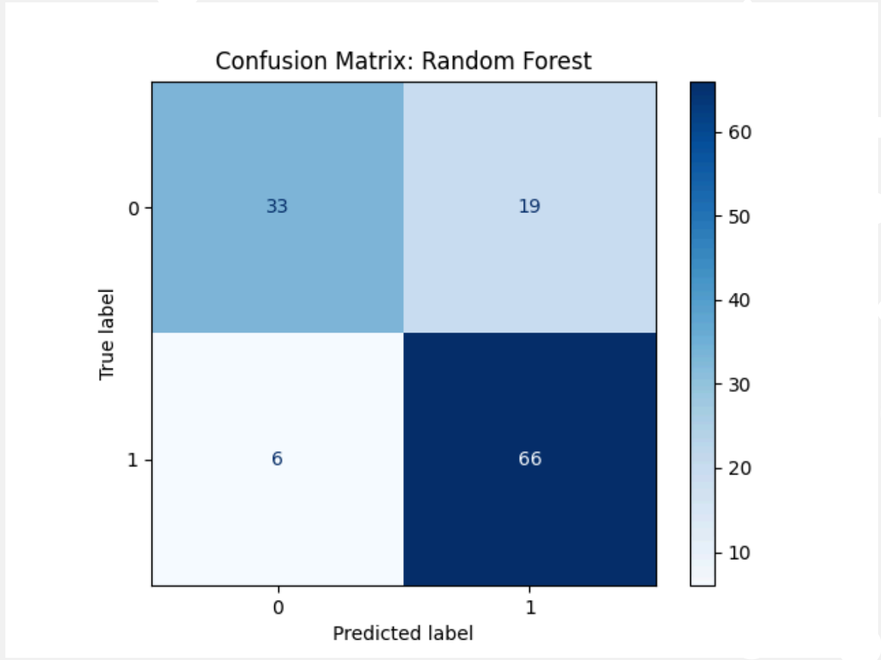


Scatter plot to identify the presence of outliers



z-score to identify the presence of outliers





Learning Curve and Confusion Matrix for Random Forest

Thank You