Hi! Paris Data Bootcamp

Final presentation

23rd of August 2024

Group 7

The members of group #7

Théo Vidal

ENSTA IP PARIS

Intermediate track

Manon Truchy

HEC PARIS

Beginner track

Nora Nalbant

HEC PARIS

Beginner track

Simon Haakerud

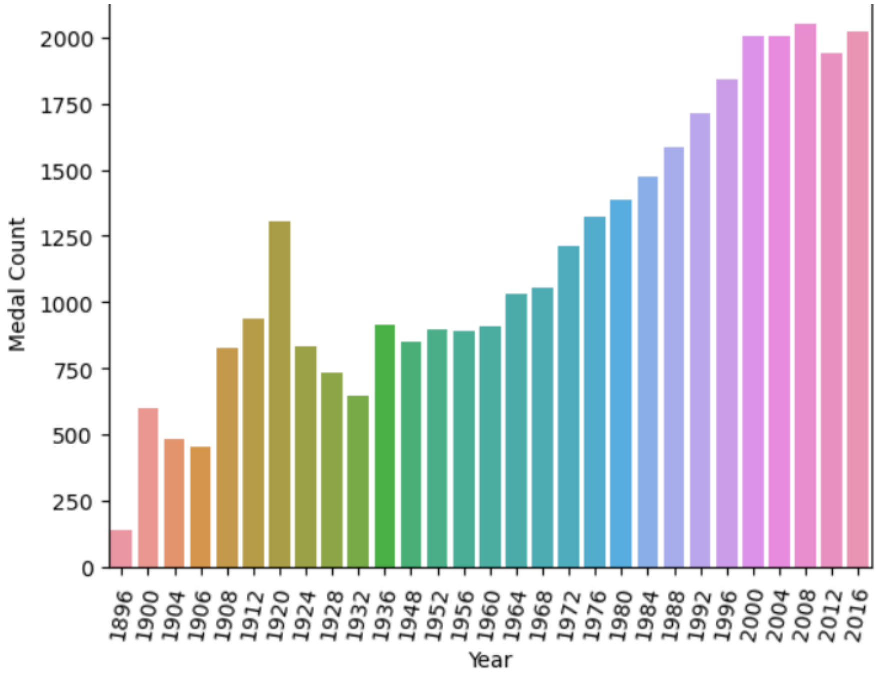
HEC PARIS

Beginner track

Data Visualization: the Olympic Games medals

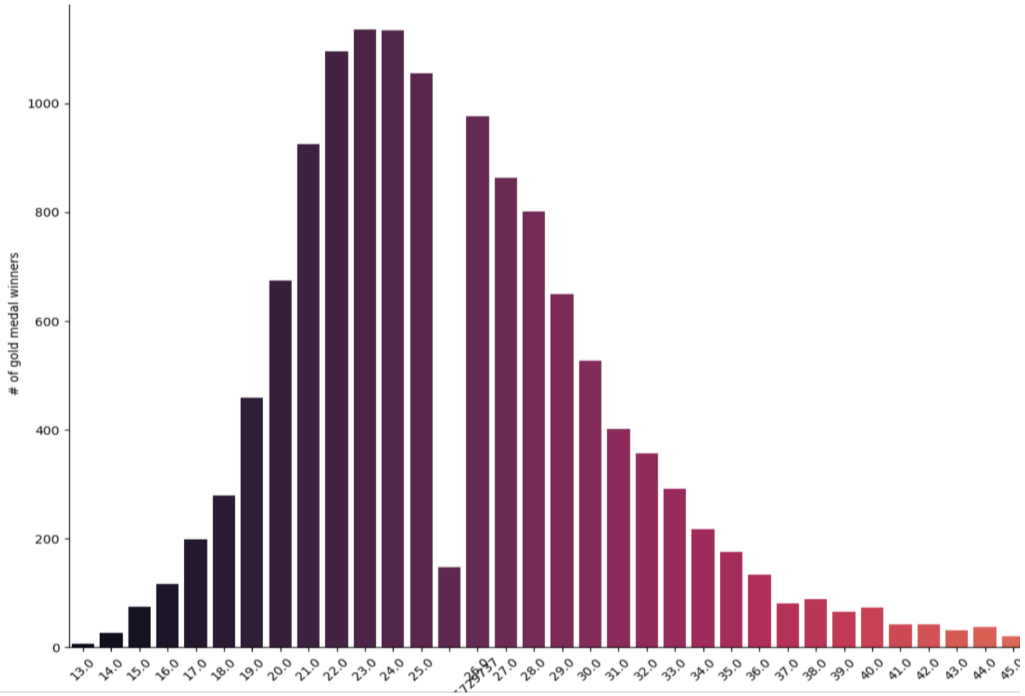
There is a general trend of increasing # medals won in the Olympics, and the winners are usually aged ~19-31

medals won in the Summer Olympics



An increasing number of medals won indicates the introduction of new sports and disciplines

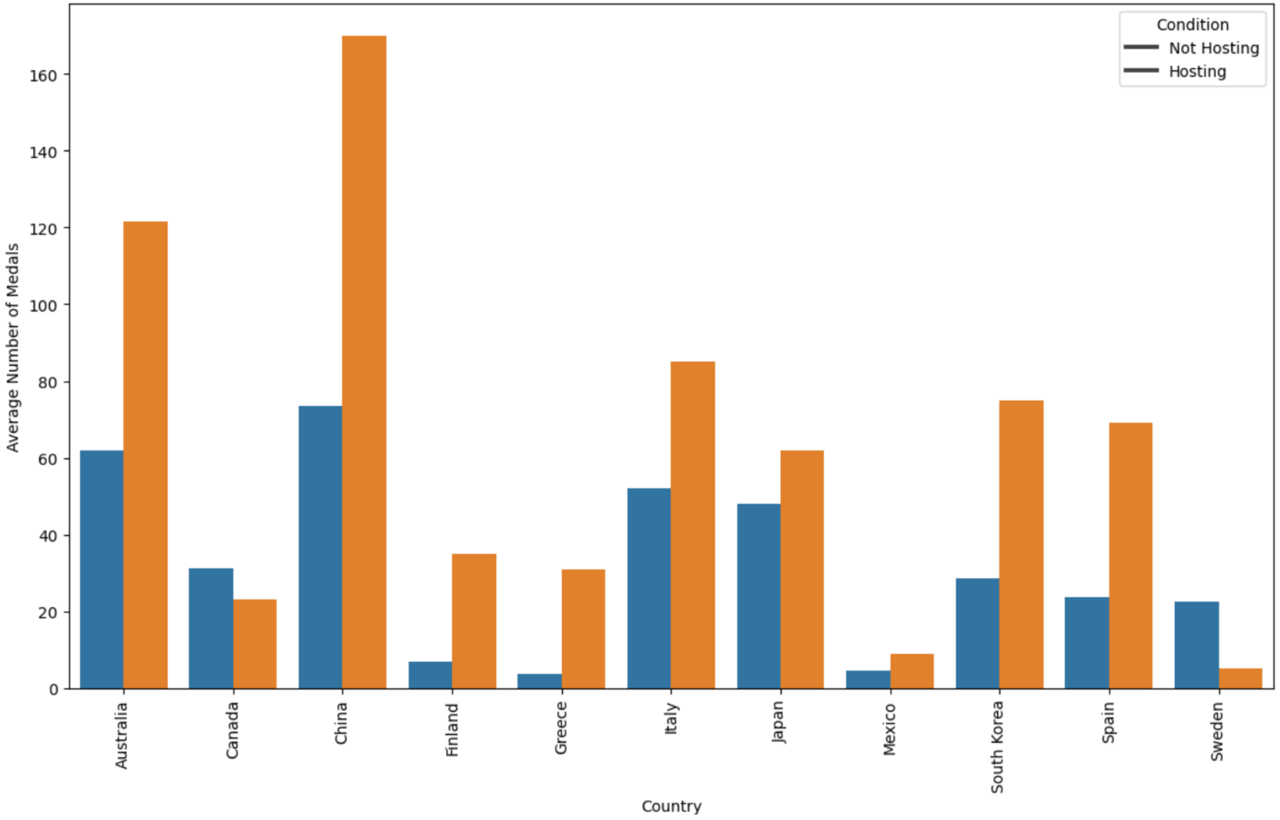
Age distribution for gold medal winners



Gold medal winners in the Olympics are usually aged between ~19-31, with 23-year-olds as the most common

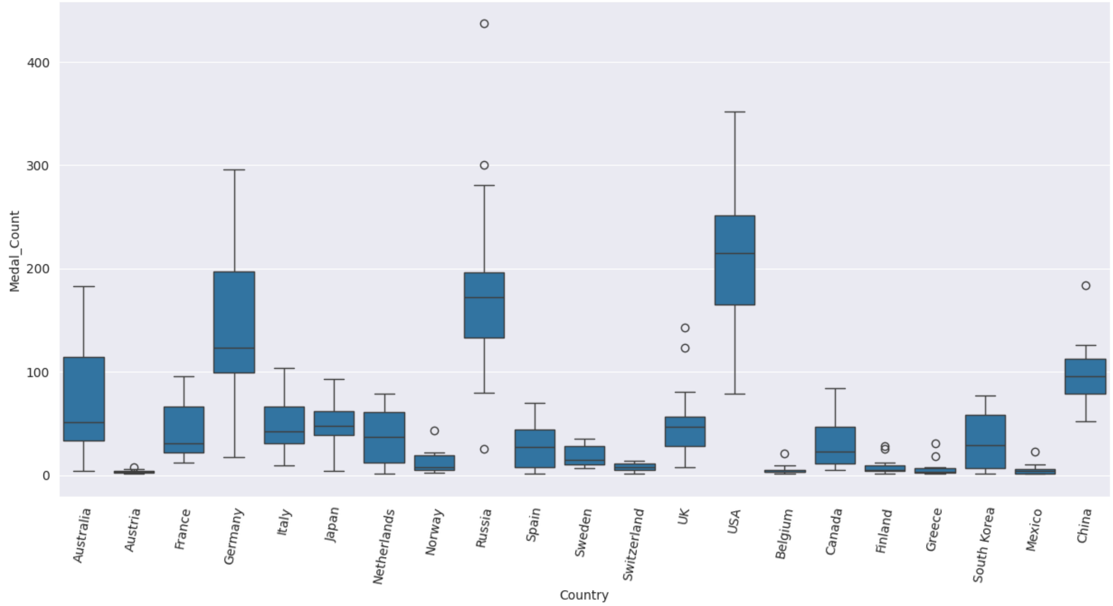
The number of medals won when hosting the Olympics are significantly higher than when not hosting

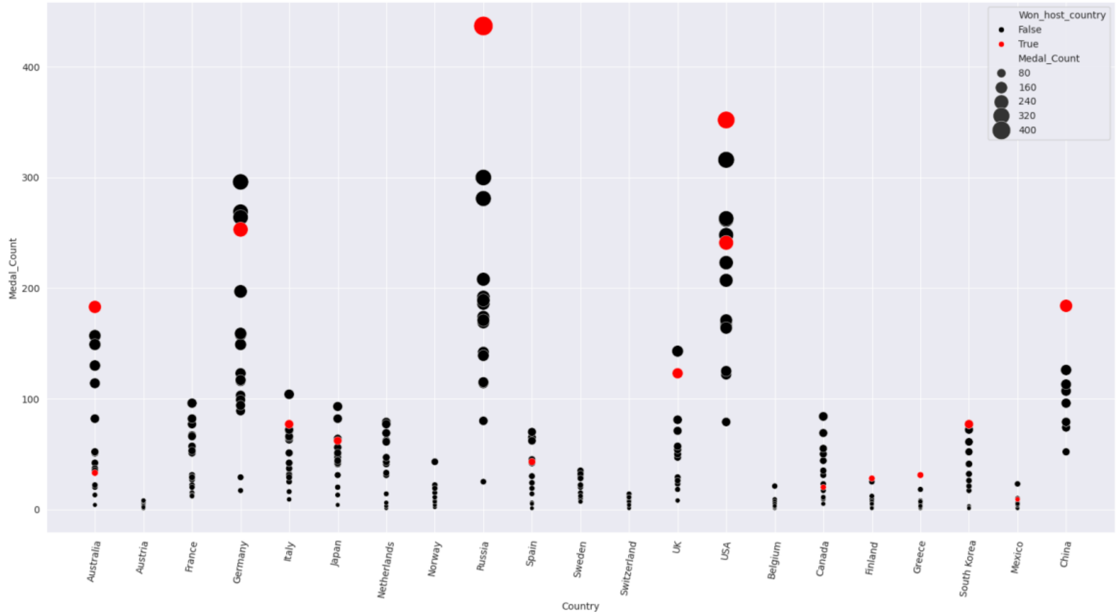
Average number of medals won by countries when hosting vs. not hosting the Olympic games



- The average number of medals won by countries hosting the Olympics are significantly higher than when they are not hosting
- As countries seldom host the games, the average when they host will be based on significantly fewer datapoints, resulting in a less accurate estimate
- The mean for hosting games is likely influenced by extreme values

Number of medals won per edition by countries that have already hosted the Games

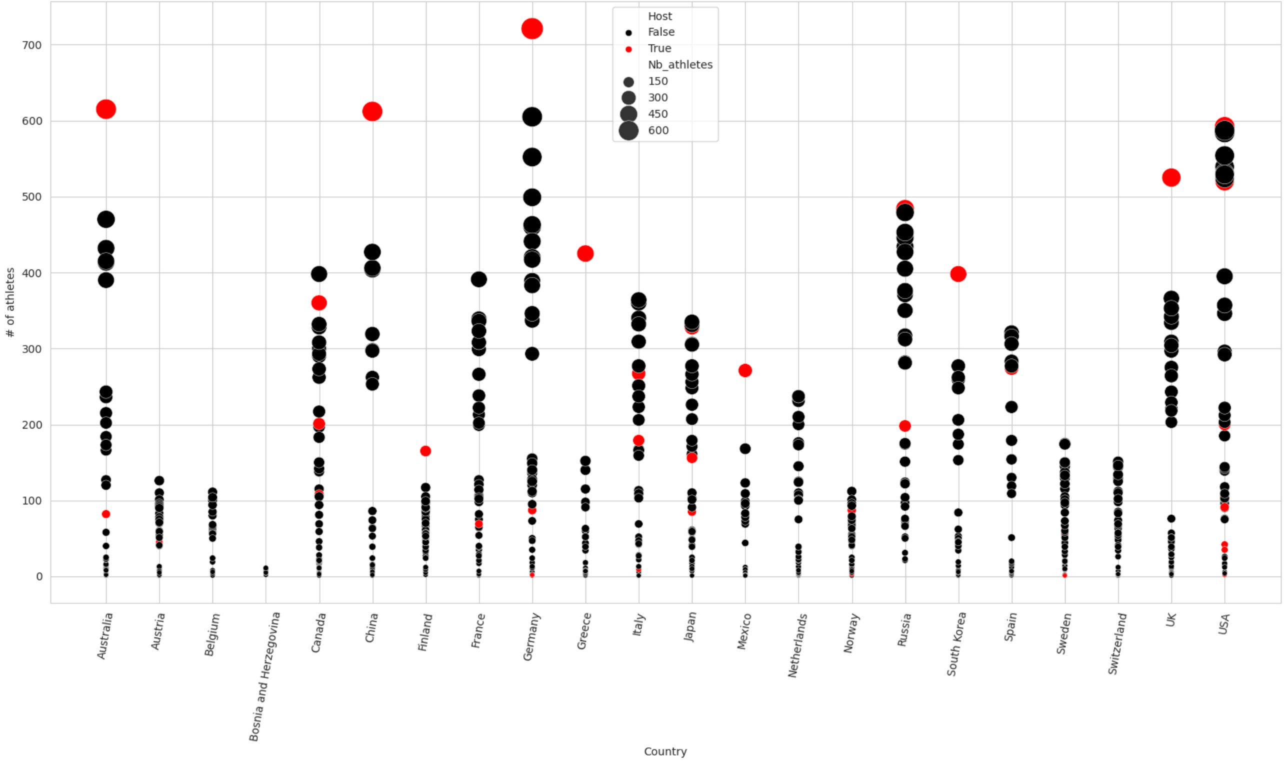




- For many countries, the high number of medals won when hosting the games is more an extreme value or even an outlier, thus confirming the means calculated before
- Therefore, one should always pay attention to the distribution of data, and not draw conclusions over means

The number of medals won when hosting the Olympics can be influenced by the number of athletes

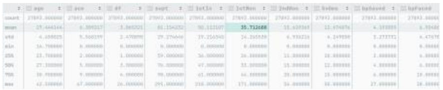
Number of athletes for each edition of the Games, per country

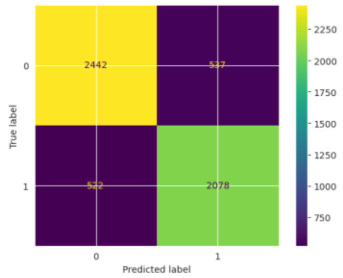


- The average number of athletes at one edition of the Games tend to be higher when the corresponding country is a host
- This might be explained by geographical and financial reasons, as travelling inside one's country is easier than going abroad
- Thus, if the number of athletes is more important, a greater number of them can win medals

Data Science: predicting Tennis players' performance

The Data pipeline of tennis performance prediction





Dataset

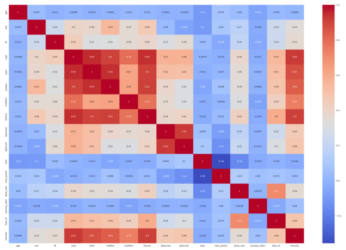
Results & Conclusions

Exploratory Data Analysis

Data Preprocessing

Searching for a model

Hyper-parameter tuning



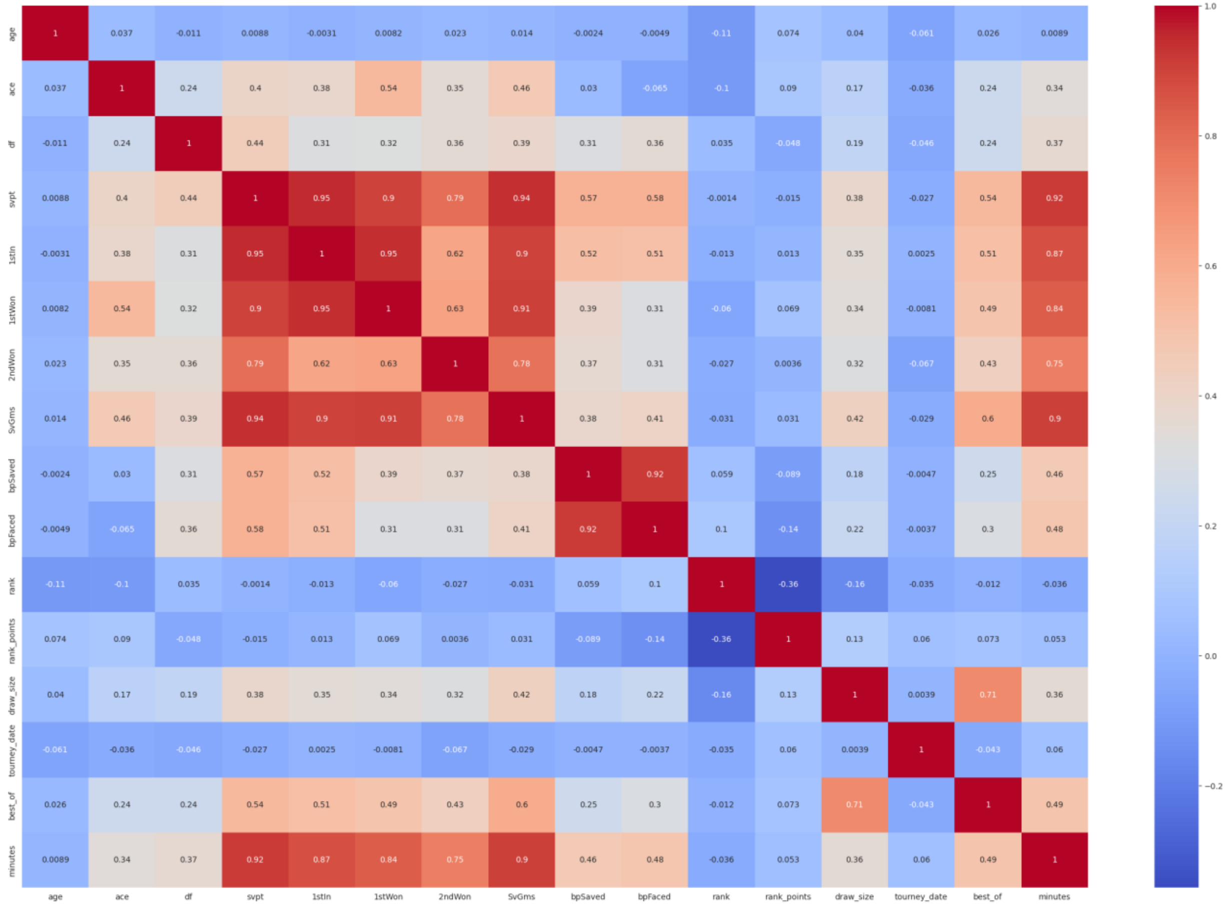
- Cleaning
- One-Hot encoding
- Standardization vs Normalization

- KNN?
- Logisitic Regression?
- Decision tree?

- # of iterations?
- Tolerance?
- Regularization?

Some continuous variables tied to time and services seem highly correlated

Correlation matrix of the tennis performance dataset



- Many variables directly tied to time and service are highly correlated: minutes, percentage of exchanges won on a first or second serve, total number of service games, serve percent
- Indeed, the longer a match lasts, the higher number of services may be needed, especially if a play doesn't contain much exchanges
- Number of breakpoints served and faced is also higly correlated

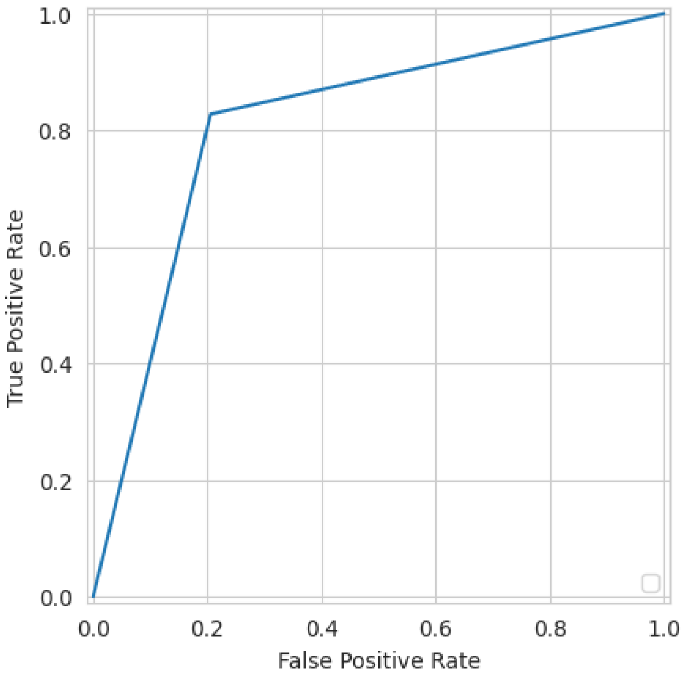
Scaled using standardization, chose the logistic regression model, and used the hyperparameter grid search

Decisions made, model chosen, and optimization realized

- **Class balance**
  - Both classes are quite balanced, with 53% of plays lost by the player involved
  - Thus, all rows of our dataset were used by the model, as class imbalance won't be a problem
- **Data preprocessing: Standardization**
  - The decision to standardize was primarily driven by the distribution of the data. With some features already exhibiting a near-normal distribution, standardization seemed the appropriate choice.
- **Model chosen: Logistic Regression**
  - Among the models tested (KNN, Decision Tree and Logistic Regression), Logistic Regression demonstrated the highest accuracy and F1 score, making it the preferred model
- **Optimization realized: hyperparameters tuning**
  - To enhance the performance of the Logistic Regression model, we conducted a grid search across a range of hyperparameters

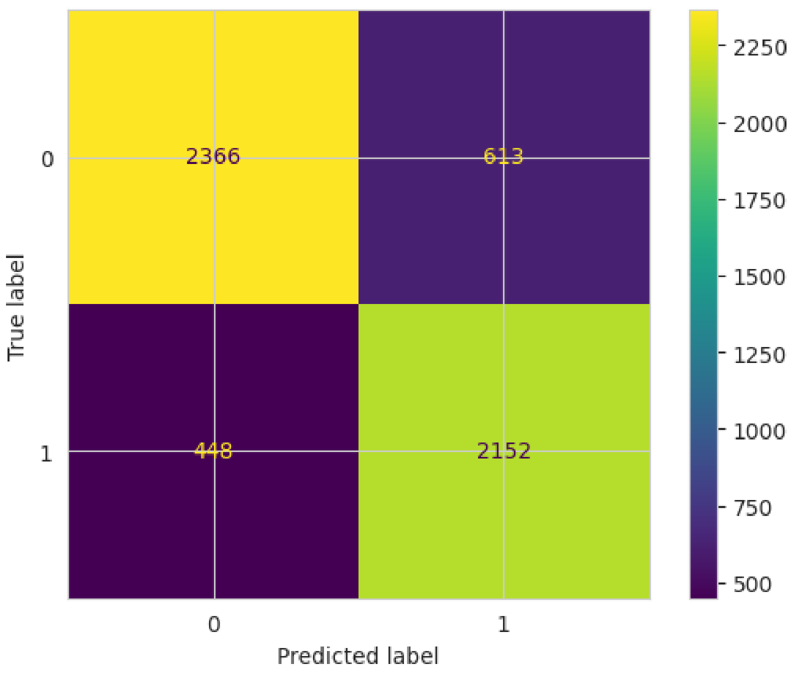
The improved Logistic Regression model achieves 81% accuracy with good classification results

ROC Curve



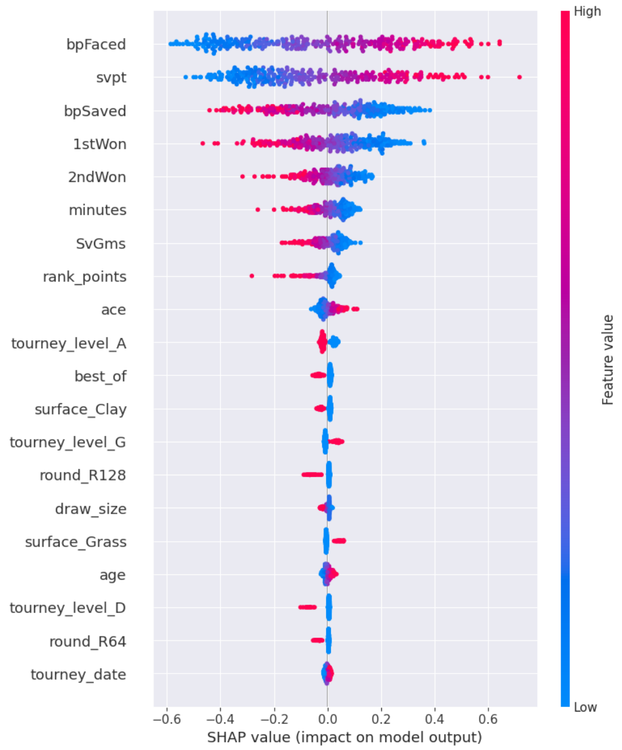
This curve is concave, indicating good classification results in both categories

Confusion Matrix



The model has relatively balanced performance in predicting both classes

Explainability



Breakpoints have the greatest impact, followed by serve percent. Overall, breaks and services have much influence.

Conclusions regarding predicting tennis performance

Key takeaways from the tennis performance prediction exercise

A key element in achieving an accurate prediction is to test for different models

Distance-based models, such as KNN, were not optimal for this exercise due to the presence of categorical values in the dataset, thus the "distance" doesn't make much sense

Tree-based models are not well suited, as our dataset contains many outliers

The logistic regression model predicts the positive class with **81%** accuracy and the negative class with **79%** accuracy, with around **20%** misclassification for each class

The different rates (True Positive, False Positive...) are crucial to analyze the performance of a model, especially if FPs or FNs can have serious impacts (for instance in cancer detection)