Predicting the Future Performance of Soccer Players

Vangelis Gkiastas

Master Thesis

September 2022

Presentation Overview

- Introduction
- Background
- Techniques and methods
- Dataset
- Feature significance
- Time Series Forecasting of Player Performance
- Future work

Introduction

Sports Analytics

- Refers to the use of historical data from the past and sophisticated statistics to evaluate performance, make data-driven choices, and forecast future outcomes.
- Applied to many aspects of the game
  - match outcome predictions, injury preventions, team tactics improvement, betting odds estimations and more
- Increasing amount of data and technology advancement
  - Event data, player tracking, optical pitch view and more
- Top tier clubs initiate data science departments
- Data-driven decisions ⟶ De Bruyne new contract

Problem Formulation

Can statistical and machine/deep learning methods predict successfully the future performance of a player in a soccer game?

- Is it possible to create a dataset of soccer players and their games?
- Which attributes are important when focusing on individual performance?
- Can future player performance be predicted using multivariate time-series models?

Presentation Overview

- Introduction
- Background
- Techniques and methods
- Dataset
- Feature significance
- Time Series Forecasting of Player Performance
- Future work

Research and Commercial Landscape

- Sports research from ‘50s
  - Still a fragmented field
  - But with a thriving popularity
- Predicting the results of games or competitions
  - Win/lose probabilities, score estimation
- Examining movements and tactics
  - Positional data, video analysis
- Analyzing and forecasting injuries
  - Fitness attributes, pre-injury, post-surgery profiles

- Sports analytics market is expected to grow from USD 2.1 billion in 2020 to USD 16.5 billion by 2030.
- Smart technology
  - Brings quantitative data
  - Enhances game performance
- Top brands:
  - Opta (StatsPerform)
  - StatSports
  - SportsRadar

Related Work

- Most of the research focuses on performance as a rating system and suggests methods of rating a player using game events
- Only Arndt and Brefeld, in 2016, provided regression-based methods, to forecast future soccer player performances
- Challenges:
  - Psychological condition, interpersonal conflicts, team fitness
  - Stochastic nature of soccer, element of luck

General Terms

- Data mining → a method for obtaining useful information from a bigger collection of any raw data.
- Machine learning → building methods that utilize data to learn to make predictions or judgments, as opposed to traditional rule-based artificial intelligence.
- Deep learning → larger family of ML, based on artificial neural networks in which multiple layers of processing are used to extract progressively higher level features from data.

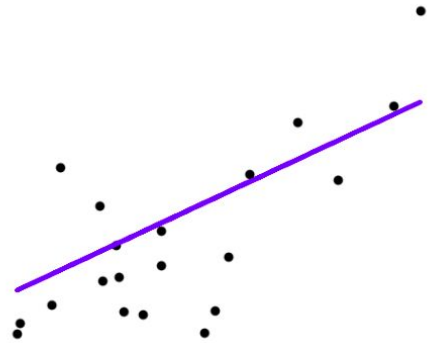
Presentation Overview

- Introduction
- Background
- Techniques and methods
- Dataset
- Feature significance
- Time Series Forecasting of Player Performance
- Future work

Techniques and methods

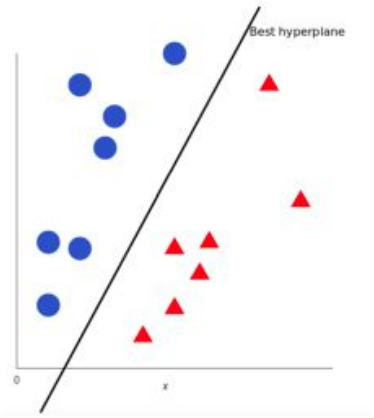
Linear regression

Finds the linear dependency of the dependent variable Y from the independent variable X.



Support Vector Regression

Tries to fit the best line within a threshold value, which is the distance between the hyperplane and boundary line.



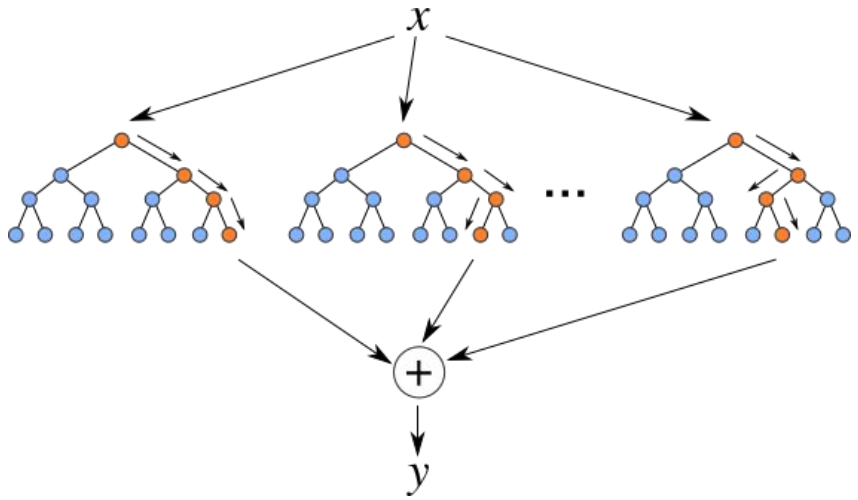
ARIMA

- Autoregressive (AR)⟶ refers to a model with a variable that regresses on its own lagged values.
- Integrated (I)⟶ is the time of differencing applied to the raw data that allows the time-series data to become stationary.
- Moving Average (MA)⟶ indicates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Techniques and methods

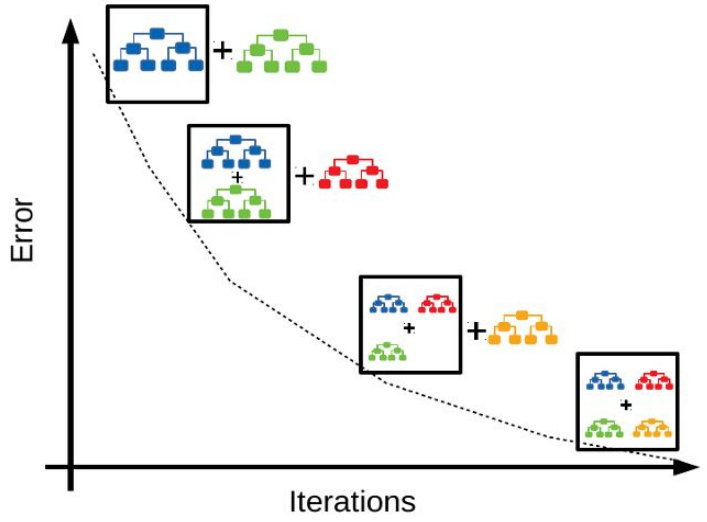
Random Forest

- Parallel ensemble learning
- Bagging: builds decision trees on different samples and takes their majority vote for classification or average in case of regression.



Gradient Boosting Trees

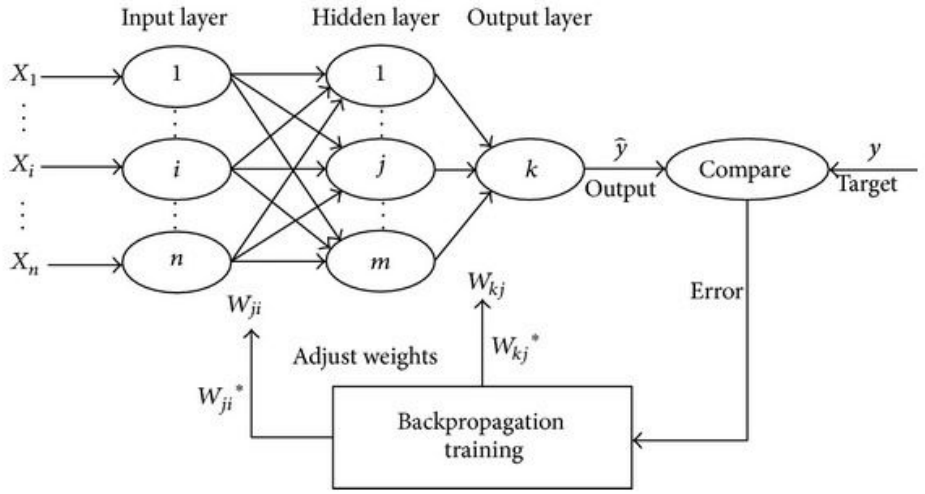
- Sequential ensemble learning
- Boosting: combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy.



Techniques and methods

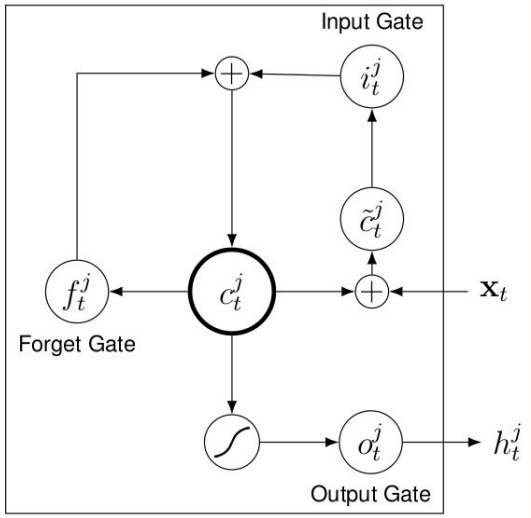
Multi-layer Perceptron

- Artificial neural networks (ANNs), are computerized models of the connections between neurons in the human brain.



Long short-term memory

- Unlike standard feedforward neural networks, LSTM has feedback connections, adding loops between actions or events, enabling the usage of information at a later time, much like a memory.



Presentation Overview

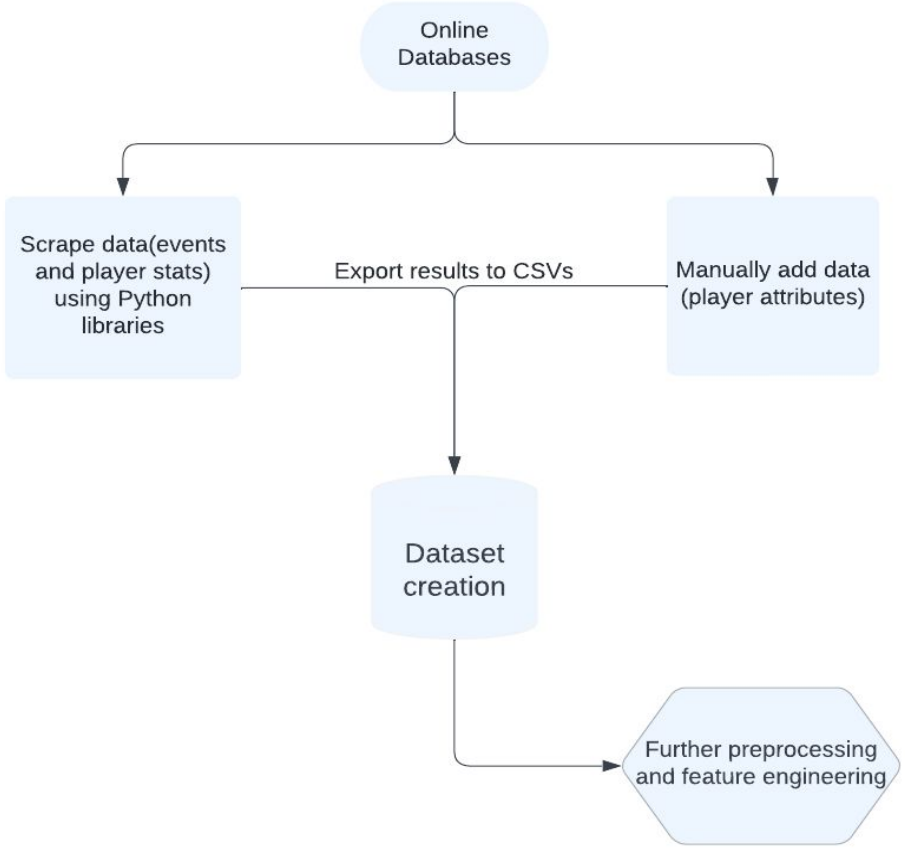
- Introduction
- Background
- Techniques and methods
- Dataset
- Feature significance
- Time Series Forecasting of Player Performance
- Future work

Current Datasets in Soccer

- Not recognized standards for capturing game-related information, including individual performance data.
- Several parties with various priorities are in charge of recording individual player statistics
- Numerous databases with match statistics.
- Player rating systems:
  - Fotmob: 300 distinct Opta stats.
  - SofaScore: 1,500 different events and statistics in every game.

Data Collection

- Define a group to study.
  - 30 nominees for Ballon d'Or 2021
- Get relevant match information:
  - SofaScore
- Limit target data
  - Only national league games



Data Collection

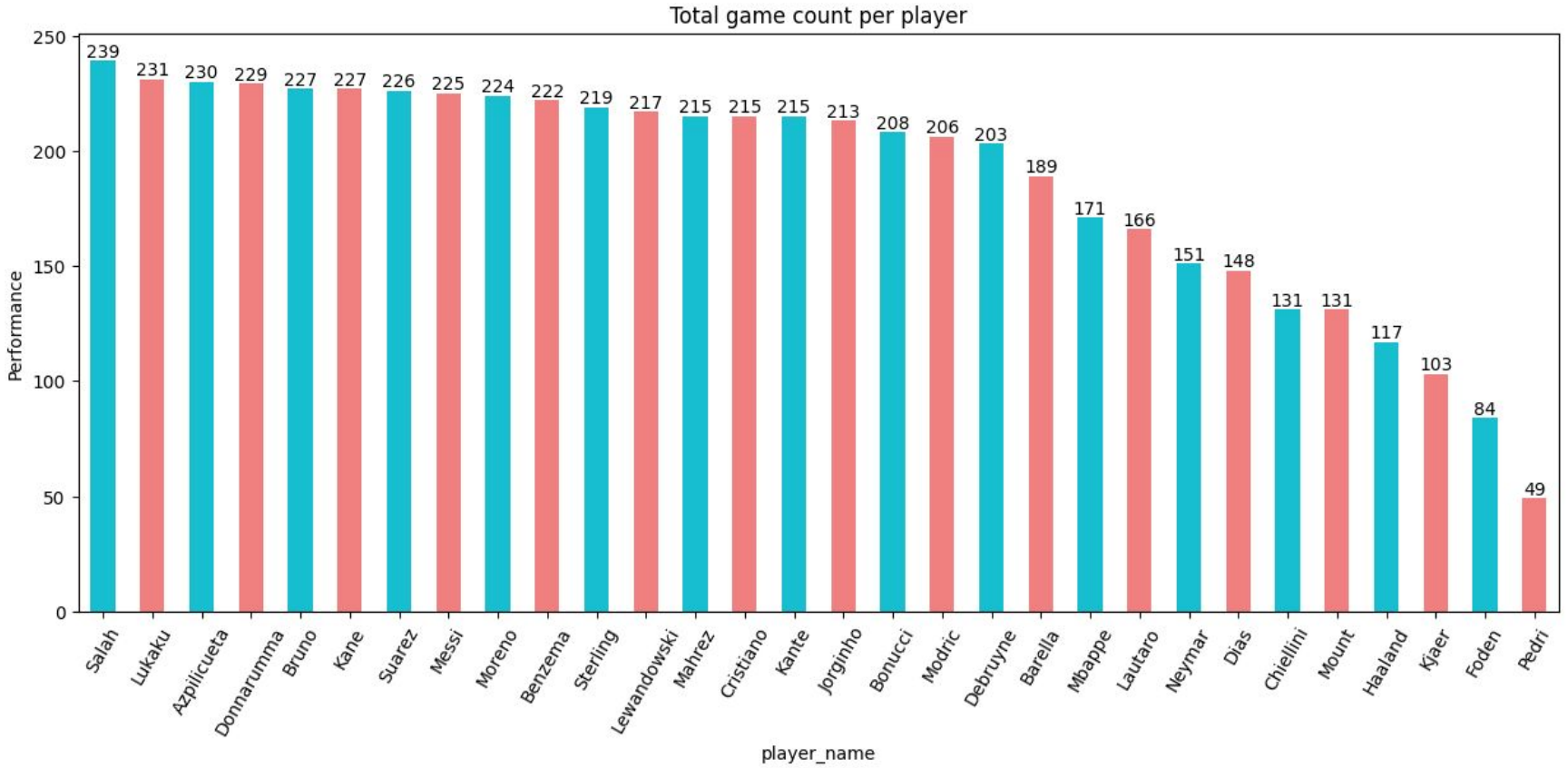
Feature Engineering

- Separate elite clubs from the rest
- Separate home game from away
- Find rest days from previous match
- Record injury event, severeness and recovery days
- Get player skills and potential from FIFA videogame
- Add player attributes like position, strong foot and age

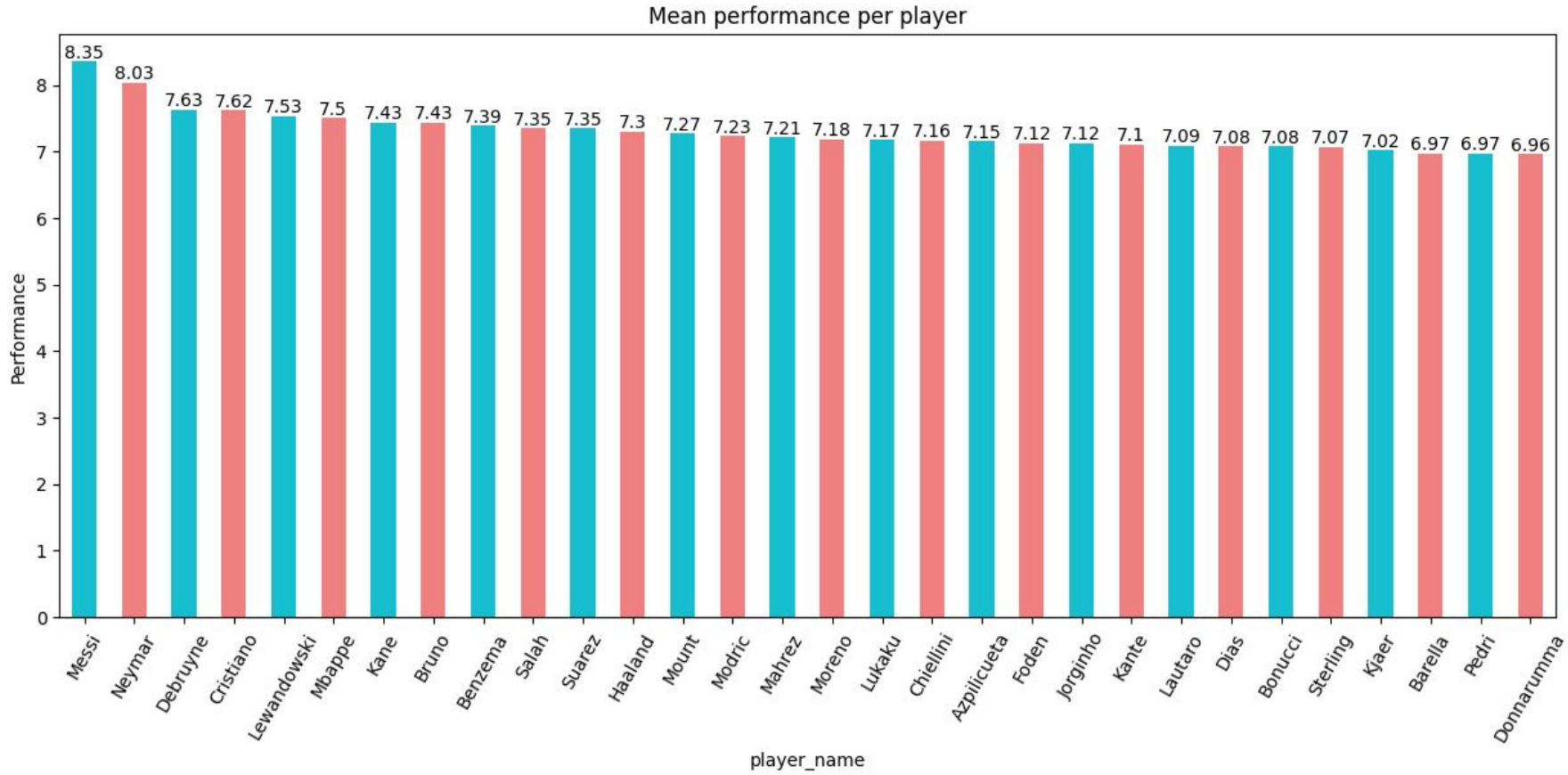
Dataset Overview

- 5631 games, 26 features
- Player attributes:
  - name, position, nationality, birth, foot, height, current age, fifa rating, fifa potential, after injury, injury days, injury type, rest days
- Pre-Game information:
  - match id, season, timestamp, tournament, home game
- Team details:
  - current team, current team category, opponent, opponent category,
- Post-Game information:
  - home score, away score, result, performance

Exploratory Data Analysis

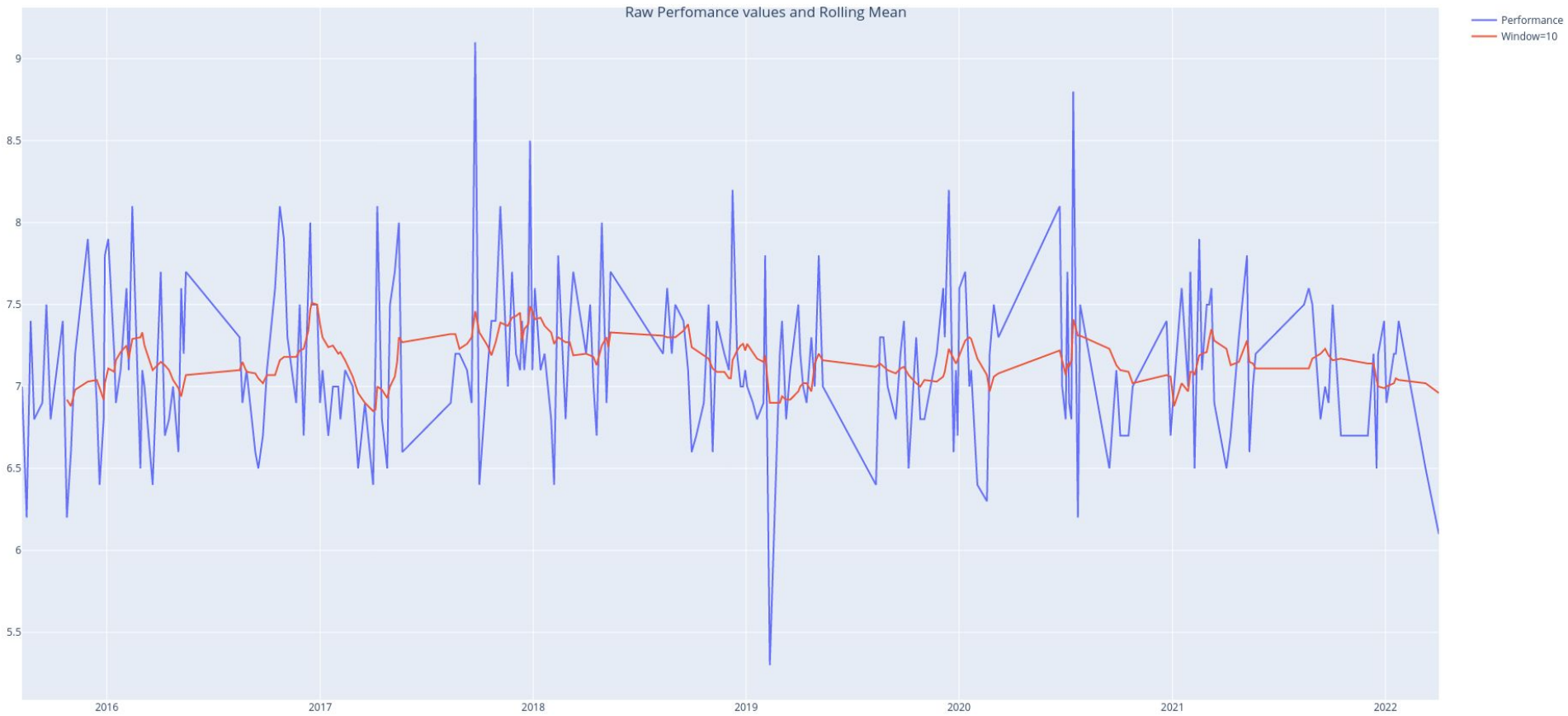


Exploratory Data Analysis



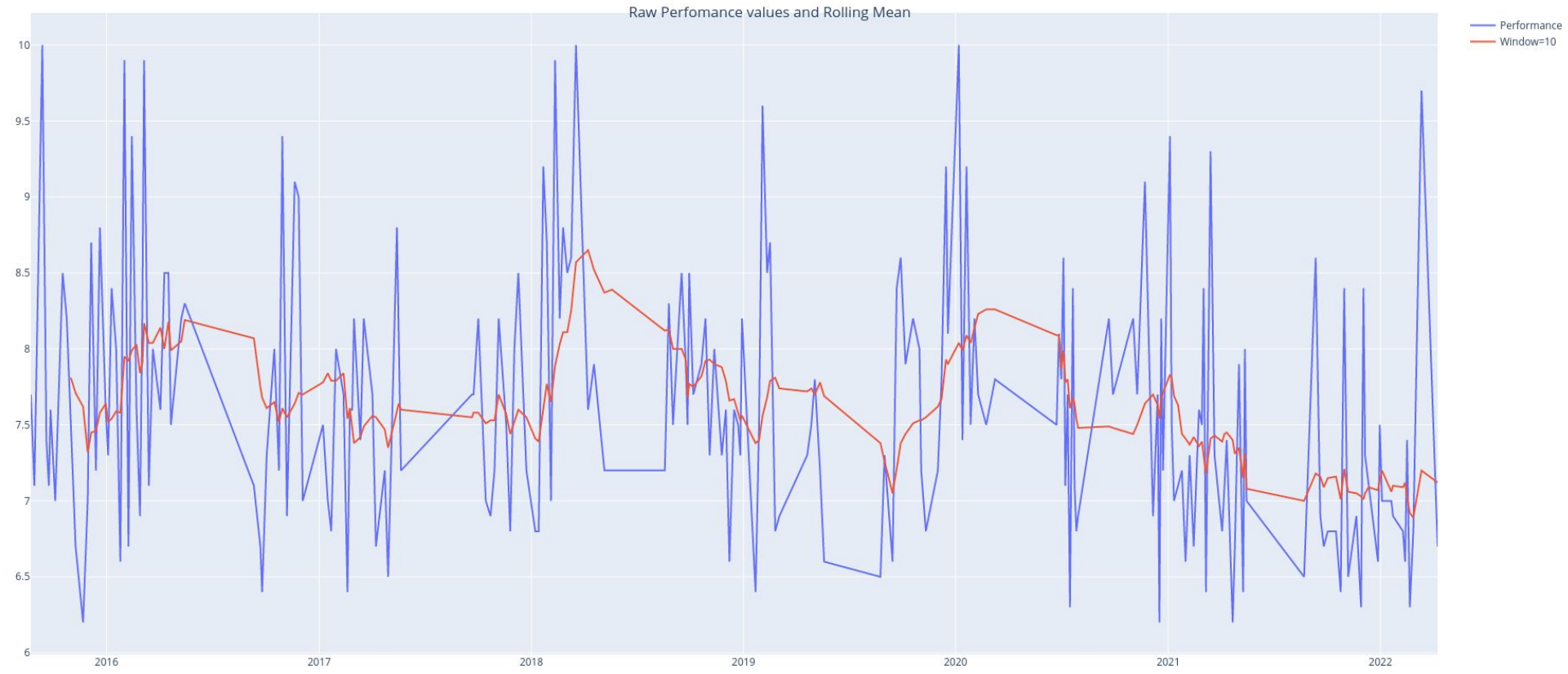
Exploratory Data Analysis

Cezar Azplilicueta performance history



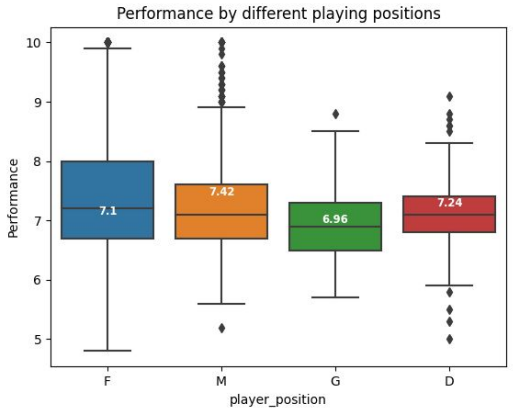
Exploratory Data Analysis

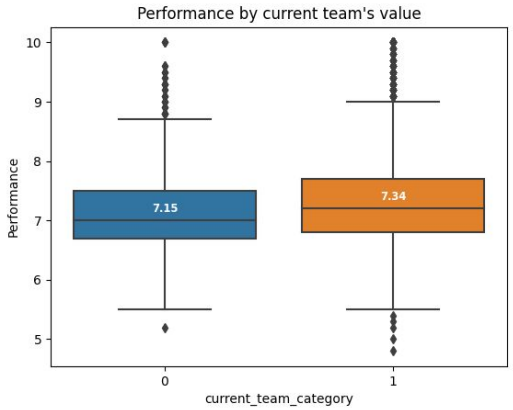
Cristiano Ronaldo performance history

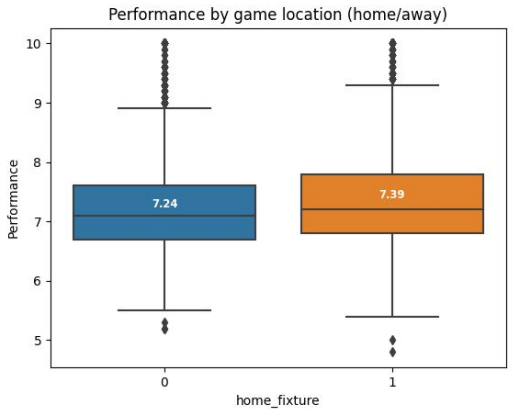


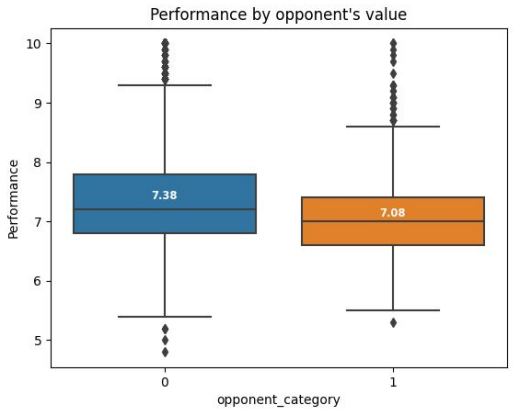
Exploratory Data Analysis

These graphs show the distribution of some indicative variables



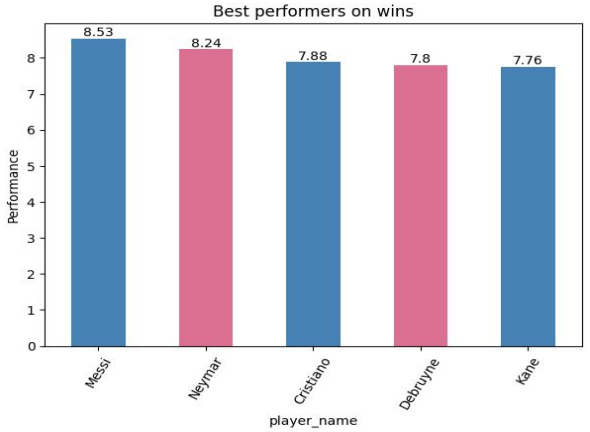


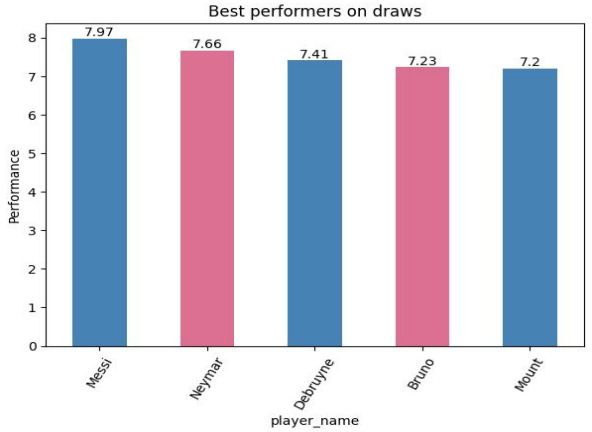


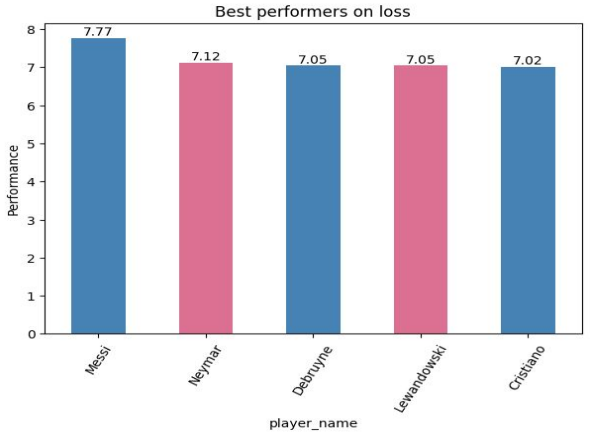


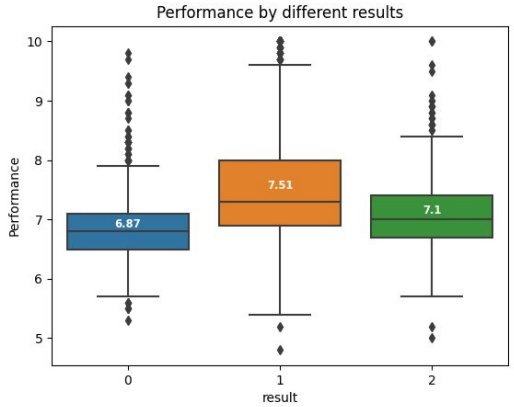
Exploratory Data Analysis

Best players on different outcomes



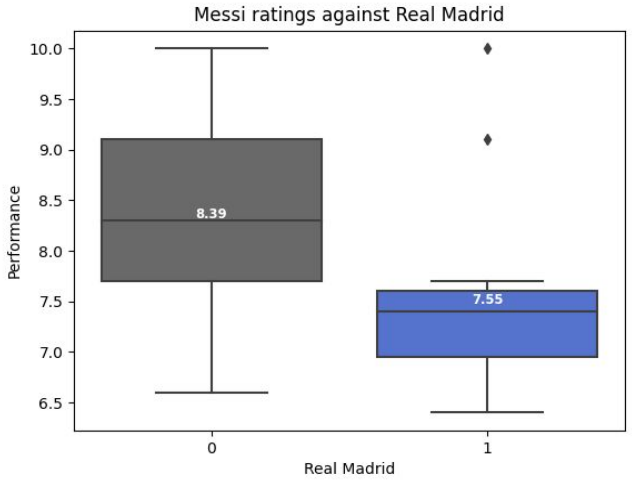


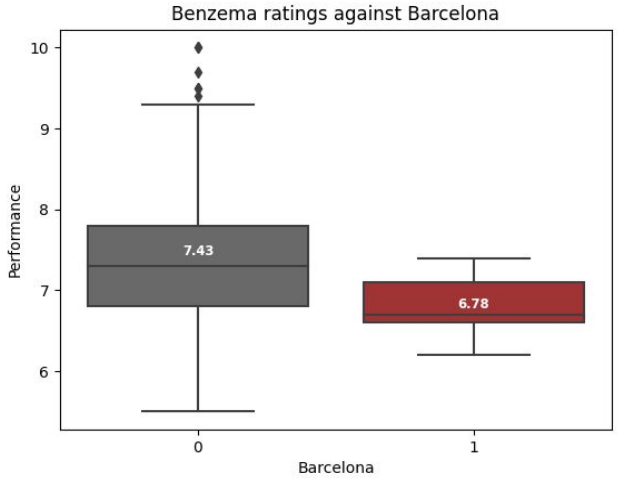




Exploratory Data Analysis

How Messi and Benzema perform on El Classico?



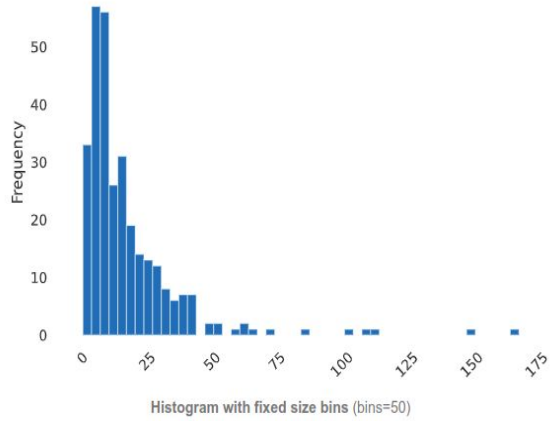


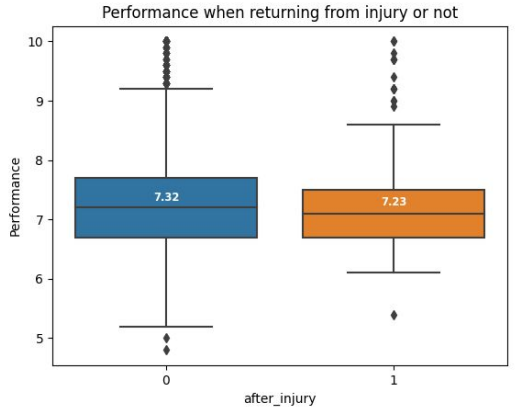
Exploratory Data Analysis

Graphical information of injuries

Common Values

Value	Count	Frequency (%)
Hamstring Injury	33	10.9%
Muscular problems	19	6.2%
Muscle Injury	18	5.9%
Adductor problems	18	5.9%
Ankle Injury	17	5.6%
Corona virus	17	5.6%
Knee Problems	14	4.6%
Thigh Problems	9	3.0%
Knock	8	2.6%
Calf Injury	8	2.6%
Other values (73)	142	46.7%





Presentation Overview

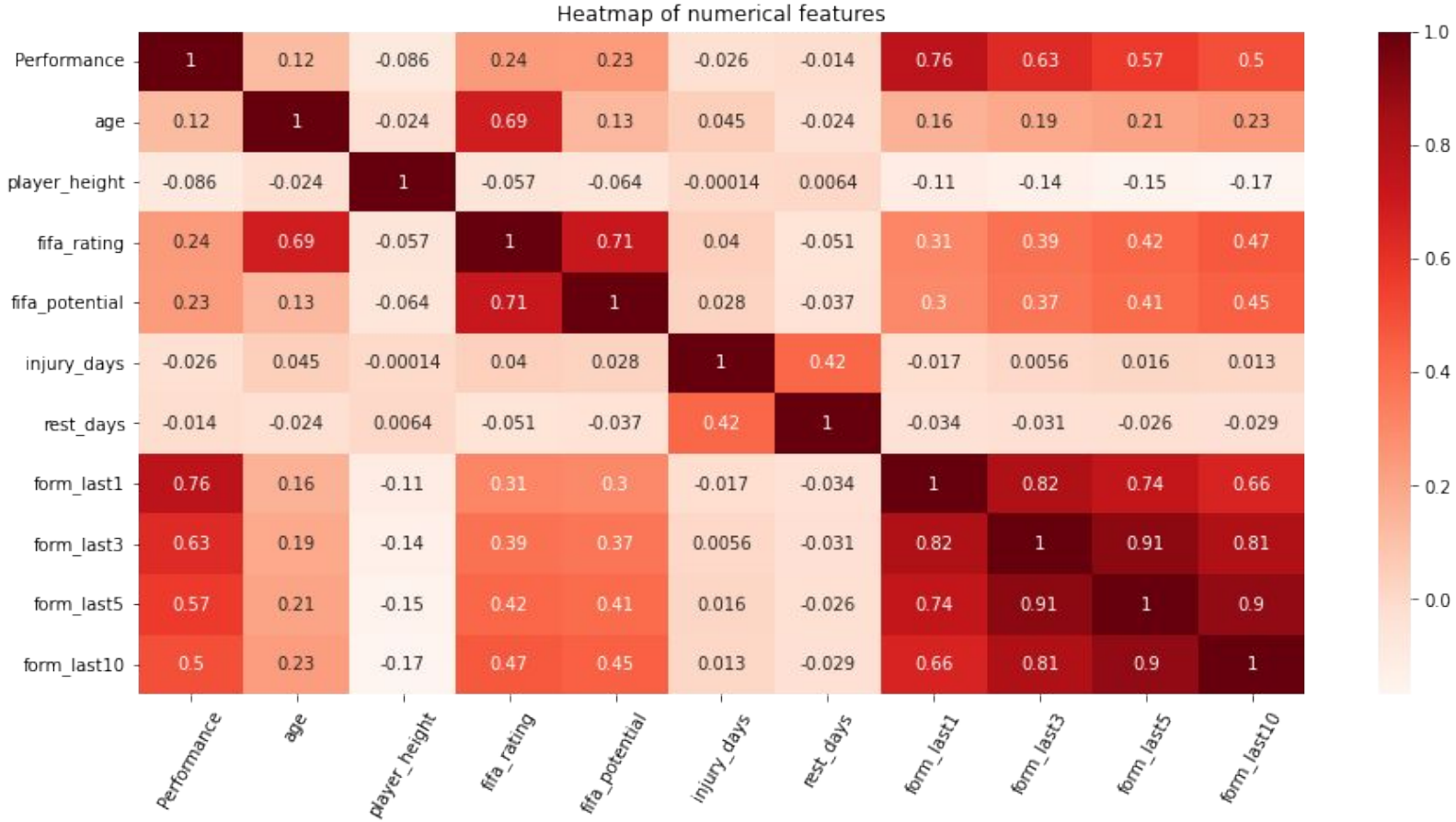
- Introduction
- Background
- Techniques and methods
- Dataset
- Feature significance
- Time Series Forecasting of Player Performance
- Future work

Dataset and Preprocessing

- Calculate player's last games' form.
- Insert fixed effects for each player.
- Dataset consists of 5631 games
  - 50 features
  - 1 dependent variable (Performance)

Variable	Description
player height	The height of each player (Numerical)
age	Age of the player on match date (Numerical)
fifa rating	Overall fifa ranking for this season (Numerical)
fifa potential	Potential growth of ranking through this season (Numerical)
after injury	Whether a player came back from injury or not (Categorical)
injury days	Days of the injury (Numerical)
rest days	Days passed since last game (Numerical)
form last 1	Average player's performance on previous 1 match (Numerical)
form last 3	Average player's performance on previous 3 matches (Numerical)
form last 5	Average player's performance on previous 5 matches (Numerical)
form last 10	Average player's performance on previous 10 matches (Numerical)
current team category	Market value of the current club (Categorical)
opponent category	Market value of the opponent club (Categorical)
home fixture	Whether the game is played in home stadium or not (Categorical)
player position F	Forward (Categorical)
player position M	Midfielder (Categorical)
player position D	Defender (Categorical)
player position G	Goalkeeper (Categorical)
player foot R	Player strong foot (Right) (Categorical)
player foot L	Player strong foot (Left) (Categorical)
player name Messi	If the game is about Messi(1) or not(0) (Categorical)
player name ...	If the game is about ...(1) or not(0) (Categorical)
player name Azpilicueta	If the game is about Azpilicueta(1) or not(0) (Categorical)
Performance	How good the athlete performed on this match, based on rating (Numerical)

Heatmap of Pearson coefficients for numerical variables



Backward Stepwise Feature Selection

Most significant features

Variable	Coefficient	P-value
const	0.0397	0.635
injury days	-0.0033	0.019
rest days	0.0013	0.032
form last 1	0.9891	< 0.001
opponent category	-0.1249	< 0.001
home game	0.1204	< 0.001

Linear Regression

- Full model:
  - $r^2 = 0.5924$
- Reduced model:
  - $r^2 = 0.5917$
- On scaled data:
  - $r^2 = 0.6120$
- Similar or worse results with other regressors.

Presentation Overview

- Introduction
- Background
- Techniques and methods
- Dataset
- Feature significance
- Time Series Forecasting of Player Performance
- Future work

Dataset and Preprocessing

- Transform time series to regularly spaced intervals.
- Drop player attributes like height, foot and position.
- Divide data in 30 batches.

Sample data of Raheem Sterling

	age	fifa_rating	fifa_potential	after_injury	injury_days	rest_days	current_team_category	opponent_category	home_fixture	Performance
2836	20.68	82	88	0	0	86.0	1	0	0	6.8
2837	20.70	82	88	0	0	6.0	1	1	1	6.7
2838	20.72	82	88	0	0	7.0	1	1	0	6.3
2839	20.74	82	88	0	0	6.0	1	0	1	7.2
2840	20.79	82	88	0	0	4.0	1	0	1	6.2
...	...	...	...	...	...	...	...	...	...	...
3050	27.20	88	89	0	0	3.0	1	0	0	9.3
3051	27.22	88	89	0	0	4.0	1	1	1	6.6
3052	27.24	88	89	0	0	7.0	1	1	0	6.5
3053	27.33	88	89	0	0	13.0	1	0	0	7.8
3054	27.36	88	89	0	0	5.0	1	1	1	6.2

219 rows × 10 columns

Dataset and Preprocessing

- Correlate each game information with the previous one.
- 30 datasets containing N games with 19 features and 1 dependent variable (Performance).
  - Training set: N-10 games from August 2015
  - Test set: 10 last games of 2022

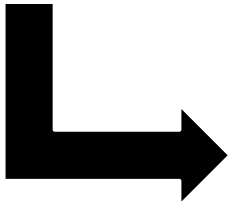
Variable	Time
age	$t-1$
fifa rating	$t-1$
fifa potential	$t-1$
after injury	$t-1$
injury days	$t-1$
rest days	$t-1$
current team category	$t-1$
opponent category	$t-1$
home fixture	$t-1$
Performance	$t-1$
age	t
fifa rating	t
fifa potential	t
after injury	t
injury days	t
rest days	t
current team category	t
opponent category	t
home fixture	t
Performance	t

Methodology

Preprocessing

Isolate games of each player to create a smaller dataset.

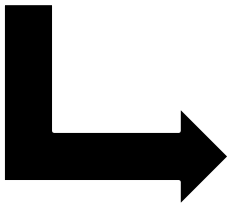
Define input and output sizes.



Training

Define model and run hyperparameter tuning.

Train six learning models (and one naive) on training set with the best parameters.



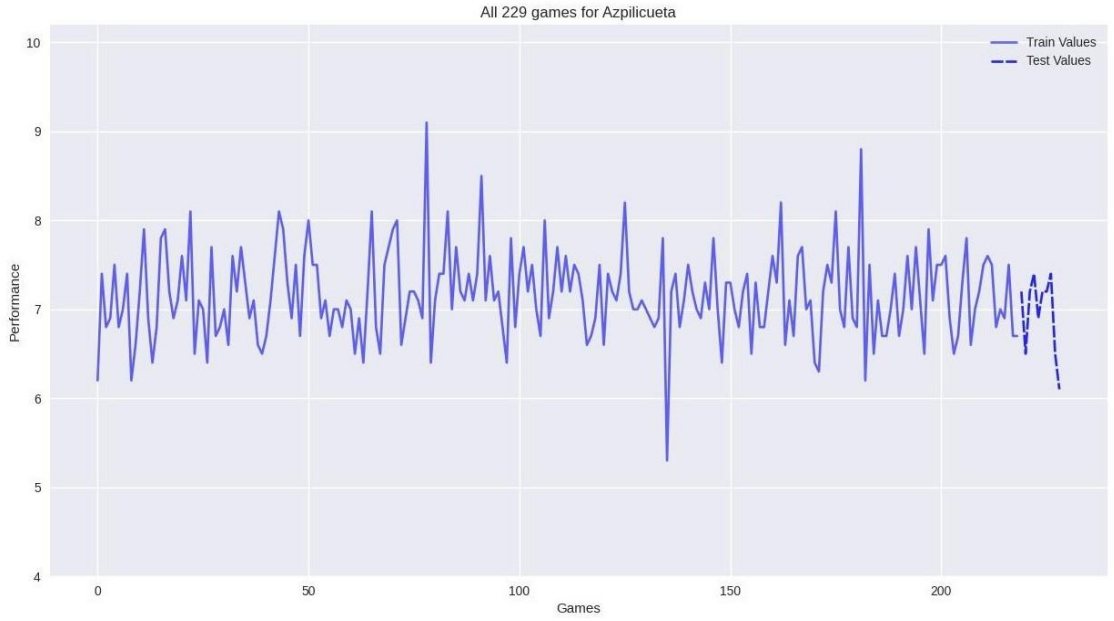
Inference

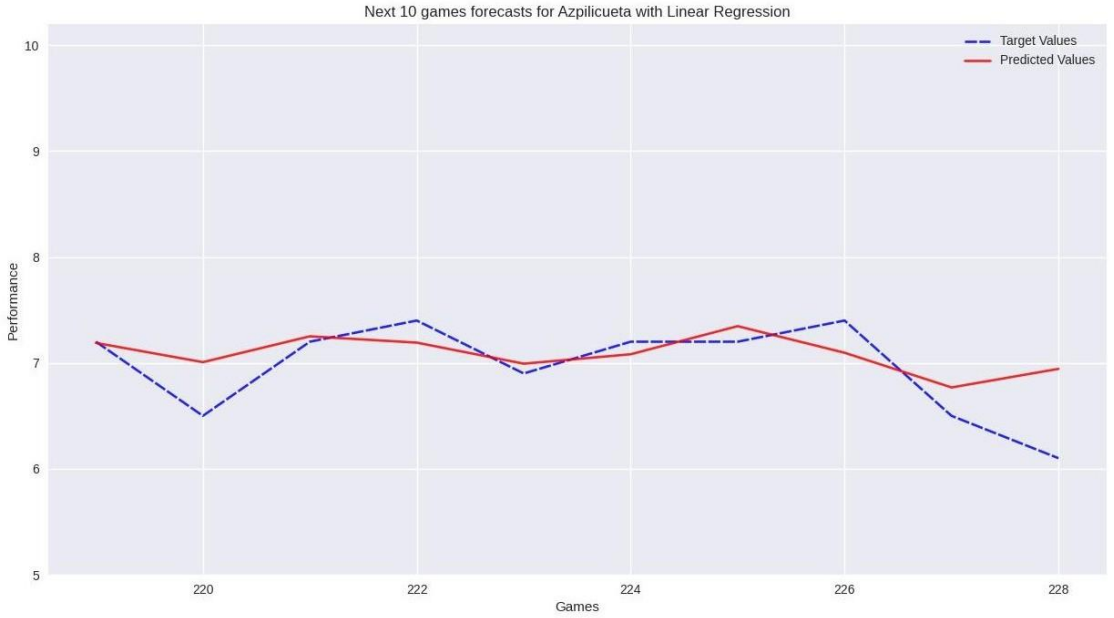
Predict next games

Get metrics and plot forecasts.

Results and Discussion

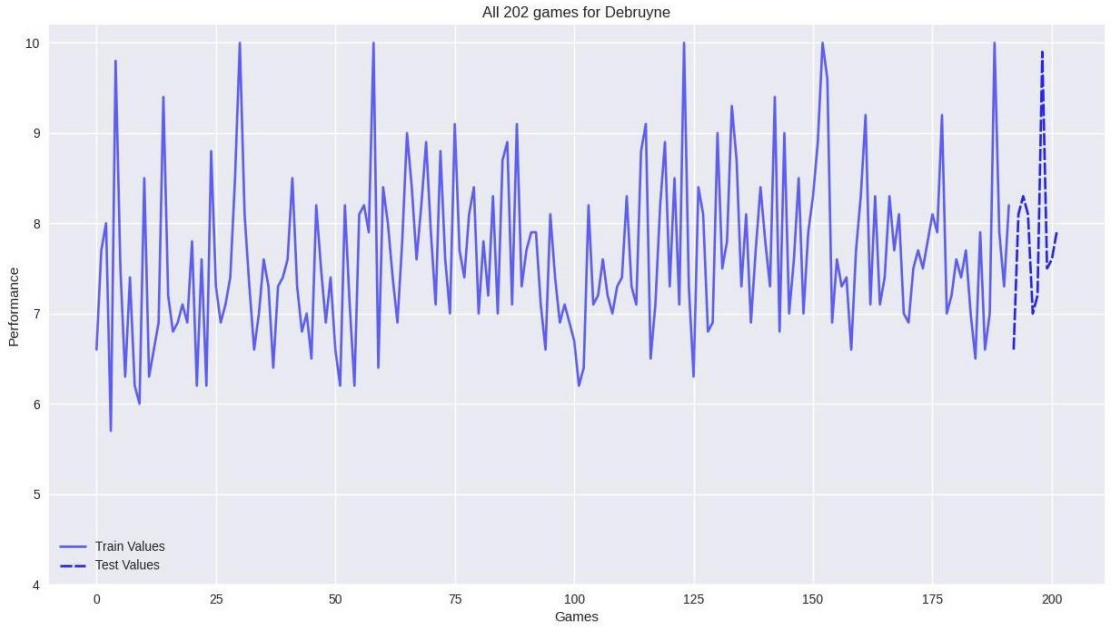
Real values and forecasts for Azpilicueta

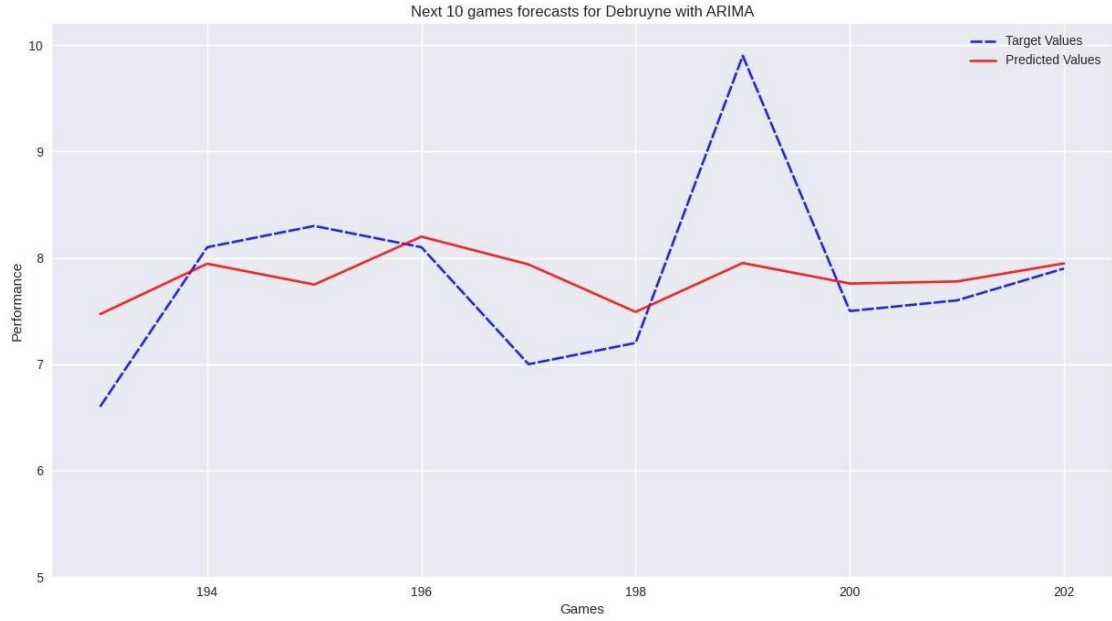




Results and Discussion

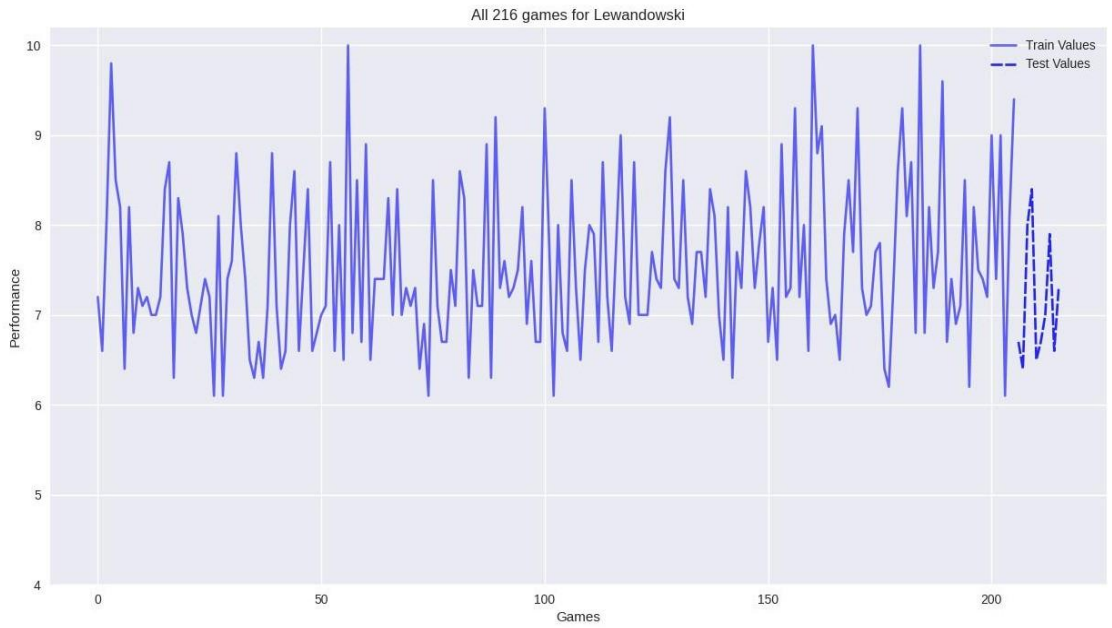
Real values and forecasts for De Bruyne

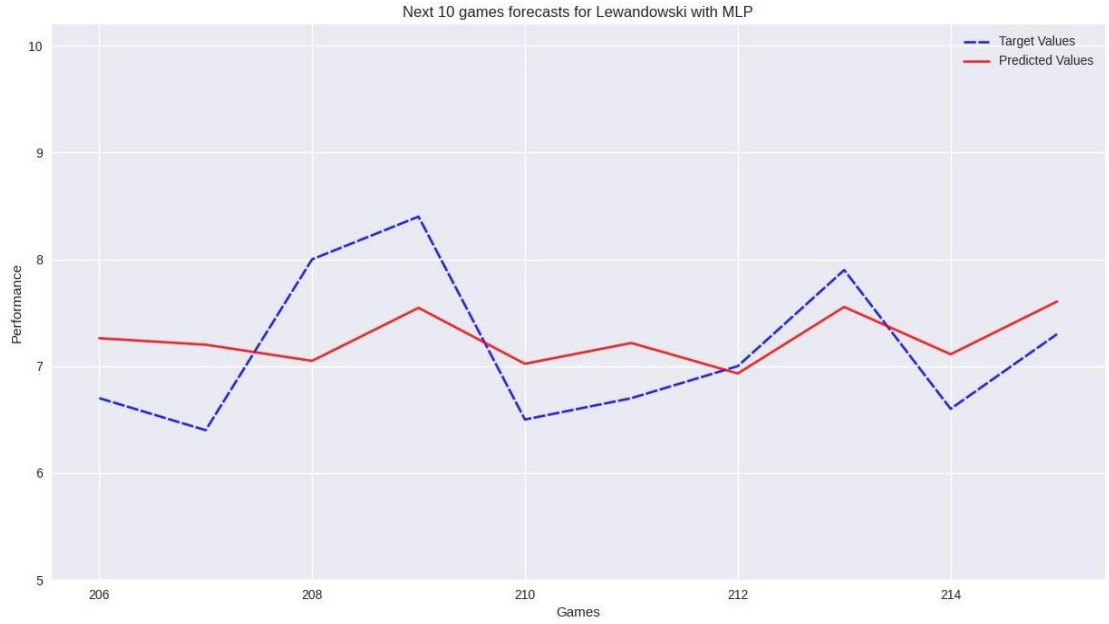




Results and Discussion

Real values and forecasts for Lewandowski





Comparison and evaluation (MAE)

Player	Best Model	MAE	Player	Best Model	MAE	Player	Best Model	MAE
Messi	Linear Reg	0.621	Haaland	Naive	0.657	Bruno	ARIMA	0.569
Lewandowski	MLP	0.727	Lukaku	MLP	0.502	Lautaro	MLP	0.587
Jorginho	Linear Reg	0.406	Chiellini	XGBoost	0.290	Kane	XGBoost	0.601
Benzema	Linear Reg	0.734	Bonucci	XGBoost	0.440	Pedri	Random Forest	0.561
Kante	XGBoost	0.333	Sterling	XGBoost	0.760	Foden	Naive	0.253
Cristiano	Random Forest	0.573	Neymar	Random Forest	0.798	Moreno	MLP	0.922
Salah	Linear Reg	0.640	Kjaer	MLP	0.228	Barella	MLP	0.472
De Bruyne	Linear Reg	0.503	Suarez	Random Forest	0.384	Dias	Naive	0.210
Mbappe	LSTM	0.932	Mount	Linear Reg	0.769	Modric	ARIMA	0.459
Donnarumma	Random Forest	0.327	Mahrez	MLP	0.533	Azpilicueta	Linear Reg	0.255

Average Mean Absolute Error : 0.53

Comparison and evaluation (RMSE)

Player	Best Model	RMSE	Player	Best Model	RMSE	Player	Best Model	RMSE
Messi	Linear Reg	0.712	Haaland	Naive	0.746	Bruno	Linear Reg	0.635
Lewandowski	MLP	0.783	Lukaku	MLP	0.563	Lautaro	MLP	0.885
Jorginho	ARIMA	0.615	Chiellini	Naive	0.342	Kane	XGBoost	0.765
Benzema	SVM	0.855	Bonucci	Random Forest	0.646	Pedri	Linear Reg	0.679
Kante	XGBoost	0.417	Sterling	Linear Reg	0.942	Foden	Naive	0.310
Cristiano	XGBoost	0.890	Neymar	MLP	1.129	Moreno	Linear Reg	1.162
Salah	Linear Reg	0.753	Kjaer	MLP	0.302	Barella	MLP	0.677
De Bruyne	ARIMA	0.772	Suarez	Random Forest	0.486	Dias	Naive	0.282
Mbappe	LSTM	1.170	Mount	Linear Reg	0.921	Modric	ARIMA	0.564
Donnarumma	LSTM	0.342	Mahrez	MLP	0.665	Azpilicueta	Linear Reg	0.350

Average Root Mean Squared Error : 0.67

Comparison and evaluation (Correct predicted directions)

Player	Best Model	Directions	Player	Best Model	Directions	Player	Best Model	Directions
Messi	Linear Reg	8/10	Haaland	Linear Reg	8/10	Bruno	XGBoost	8/10
Lewandowski	LSTM	8/10	Lukaku	SVM	5/10	Lautaro	Linear Reg	6/10
Jorginho	Random Forest	8/10	Chiellini	Random Forest	7/10	Kane	ARIMA	7/10
Benzema	Linear Reg	7/10	Bonucci	MLP	8/10	Pedri	XGBoost	6/10
Kante	XGBoost	8/10	Sterling	Linear Reg	7/10	Foden	ARIMA	8/10
Cristiano	XGBoost	7/10	Neymar	SVM	6/10	Moreno	MLP	6/10
Salah	LSTM	5/10	Kjaer	SVM	7/10	Barella	Linear Reg	7/10
De Bruyne	Linear Reg	7/10	Suarez	SVM	5/10	Dias	MLP	9/10
Mbappe	Random Forest	9/10	Mount	Linear Reg	6/10	Modric	ARIMA	9/10
Donnarumma	XGBoost	6/10	Mahrez	SVM	8/10	Azpilicueta	MLP	8/10

Average correct predicted directions : 7.2/10

Presentation Overview

- Introduction
- Background
- Techniques and methods
- Dataset
- Feature significance
- Time Series Forecasting of Player Performance
- Future work

Future Work

- Wider range of models can be explored
  - A lot of research is being done and new algorithms are constantly being released
- More data would significantly improve the performance of any model
  - Training information → exposure records, coach reports and GPS data
  - Event data → shots, passes, duels won and more
  - Team playing styles, player similarities, teammate interactions
  - Psychological condition
- Study different player groups
  - Investigate if findings from top class players apply to ordinary ones
  - Examine more leagues and divisions

Questions