ID Cover Page

Summary of WP Student Team

FPF FIELD LAB: PREDICTING THE ATTENDANCE OF PORTUGUESE 1ST DIVISION FOOTBALL STADIUMS USING MACHINE LEARNING

Group constitution:

Student Name	Program	Individual Title
Diogo Miguel Vilarigues Passeiro	Business Analytics	FPF Field Lab on stadium attendance – the case of the "Big 3" in Portugal
José Maria C. C. Botelho de Sousa	Business Analytics	FPF Field Lab on stadium attendance – analysis on the occupation rate variable

Work project carried out under the supervision of:

Advisor: Pedro Brinca

A Work Project, presented as part of the requirements for the Award of a Master's degree in

Insert your Program from the Nova School of Business and Economics.

FPF FIELD LAB: PREDICTING THE ATTENDANCE OF PORTUGUESE 1ST DIVISION FOOTBALL STADIUMS USING MACHINE LEARNING

DIOGO MIGUEL VILARIGUES PASSEIRO

JOSÉ MARIA C. C. BOTELHO DE SOUSA

Work project carried out under the supervision of:

Pedro Brinca, PhD

16/12/2022

Abstract

This paper aims to help FPF predict attendance at Portuguese Liga I according to data from the 2015/16 season until 2021/22 to optimize round scheduling of football matches. The dataset contains information about games played, including the teams, stadiums, and weather conditions. Among all the various machine learning regression models tested, the XGBoost provided the overall best results regarding the out-of-sample mean absolute error. Additionally, considering the differences in size and consequently in error between two groups of clubs, narrowing down the scope added value to the project. Besides attendance, the occupation rate was also tested as the target variable.

Keywords: Machine Learning, Prediction, Stadium Attendance, Football

Acknowledgements

We would like to express our full gratitude to our adviser Pedro Brinca, as well as to the Portuguese Football Federation and its collaborators. Effectively, they were the ones who presented us with this challenging topic and provided us with the necessary data to implement in our models. We also want to thank them for their availability in giving us guidance and feedback during all the meetings we had throughout this project.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

1. Introduction

FPF intends to maximize attendance in stadiums in Portugal, more precisely, in leagues under its sphere of influence. For that, they are looking for the main drivers of it besides ticket prices, in a way they can optimize their scheduling choices according to that. This theme has been explored, presenting a wide variety of results, but not applied to Portugal. Therefore, the study focuses on attendance at Portuguese League I matches and evaluates the efficacy of various machine learning regression methods on building prediction forecasts.

2. Literature Review

In this chapter, we have summarized all the research that has given us suggestions for our future model and helped us to outline a structured project so that we stay within the final project objective. First, we looked for studies that indicate factors influencing a fan's attendance at a sporting event. Then, we tried to find out how sports clubs can benefit the most from fans going to stadiums and consequently maintaining a constant presence in the stadium. Finally, to help us with our model, we looked for projects on various sports that used machine-learning techniques to add value to the sport.

2.1 FACTORS THAT INFLUENCE POTENTIAL SPECTATORS' DECISIONS

Based on our research (Borland and Macdonald 2003) "Demand for Sport", where they highlight the determinants of stadium attendance, we selected some of the determinants that we found helpful in further building our predictive model. These determinants that can influence attendance are economic factors, quality of viewing or characteristics of the sporting contest.

2.1.1 Economic factors

Like all sports, watching football also has an associated cost for its fans. Thus, economic factors can limit, to a certain extent, fans' involvement with their club. The most obvious one is the price of attending a match. According to (Borland and Macdonald 2003) there is strong

evidence that not only the ticket price, but the price of admission negatively affects attendance. Moreover, this price of admission usually encloses other factors such as the cost of travel associated with a fan's distance to the stadium, stadium car parking facilities, beverages and food consumption. In contrast, fans' income is positively related to attendance.

As the competition between media channels to obtain television rights from clubs increases, many offer various services depending on the fan's preferences. This deal eventually leads to fans opting to watch the games at home, not contributing to match attendance.

2.1.2 Characteristics of the sporting contest

This determinant of match attendance encompasses several psychological factors affecting a fan's decision to attend a football match. As football matches tend to be most unpredictable (Lo 2011), in some studies, the "uncertainty outcome" is confirmed to affect the match's demand. This measure was used in attendance predictive models (Şahin and Erol 2018) and highlighted as a helpful measure for league directors to impose rules and regulations to promote a "competitive balance" during matches (Borland and Macdonald 2003). This "competitive balance" is also confirmed by (Hautbois, Vernier, and Scelles 2022) to have a "positive influence on attendance".

Increasing competitiveness during a match can also increase "team performance", which can be "positively associated with attendance" (Bradbury 2020) and can explain the stadium attendance demand (Martins and Cró 2018). On the contrary, the quality or reputation of the away team may be positively related to attendance at the stadium (Valenti, Scelles, and Morrow 2020). Furthermore, we have found that fans tend to prefer high-quality opponents (Addesa and Bond 2021) and "are responsive to matches against rival teams and strong opponents" (Watanabe and Soebbing 2017). This can be useful to analyze the impact of rivalry on attendance further.

2.1.3 Quality of viewing

Usually, the quality of viewing during a match influences the fan's final experience. Various factors can measure this quality of viewing level.

Some of these can be the quality of the stadium facilities measured by the access difficulty and the matchday type of weather, which tends to have a negative relationship with attendance and match timing. As stated by (Baimbridge 1997), a match that coincides with a holiday tends to get a higher attendance level.

2.2 THE STADIUM ATTENDANCE IMPACT ON A FOOTBALL CLUB

According to (Uefa 2022), 34% of club revenue comes from television rights compared to just 12% of gate revenue. This difference is even more accentuated in Portugal as it is one of the few countries still with a centralized TV rights distribution system. As a result, compared to other European leagues, the club that collects more TV rights receipts "collects more than nine times the median club" (Uefa 2022). This puts Portugal on top of the unequal countries concerning the monetary distribution of tv rights by broadcasting stations.

Due to this inequality, it is crucial for Portuguese "smaller" clubs that, financially, still depend heavily on stadium revenues, to look for ways to diversify their revenue streams (Schmidt and Holzmayer 2018). If attendance figures increase, they can leverage their impact to generate additional revenues such as stadium tours, club museum tours, hotel stays, or even merchandising (Schreyer and Ansari 2022). In addition, an increase in attendance can serve as a reputation proxy for potential club investors (Gimet and Montchaud 2016).

Finally, it has been proved that maximizing the number of fans in the stadium can create a significant home advantage, creating even greater social pressure on the referee. A fact that could be more penalizing for the opposing teams (Reade, Schreyer, and Singleton 2022).

2.3 HOW IS MACHINE-LEARNING CHANGING FOOTBALL

The idea of creating a system that could improve itself without having to be specially programmed has been tested consistently by many technology enthusiasts. From Arthur's Samuel IBM Checkers program, where an intelligent computer improved every time it made a move, to facial recognition we use to access our smartphones, Machine-Learning has become increasingly more present in our daily lives. The sports industry has been following this technological progress very closely and searching for ways to increase its popularity and add value to every sports fan.

In particular, many have used football to create programs that rely on machine-learning techniques. For instance, academics have constructed Machine-Learning models to understand and predict injuries in football. This way, coaches can adapt "training programs, assess fatigue and recovery, and minimize the risk of injury and illness" (Majumdar et al. 2022). Furthermore, to help coaches' decision-making process before a match, Machine-Learning models can evaluate a player's potential so that coaches can choose the fittest players in the whole squad (Sheng et al. 2021). Concerning players, health Machine-Learning models can also be used to determine the optimal players' diet and training plans so they can maximize their game performance (Alberto Rizzoli 2022).

Moreover, Machine-Learning models can also help plan player transfers (Ćwiklinski, Giełczyk, and Choraś 2021) to assess player market values (Lee, Tama, and Cha 2022) or even a player transfer success. Indeed, all these different machine learning uses may be of particular interest to all those involved in the sport.

3. Project Description

Considering that we are dealing with a company in this project, it is essential to define and clarify the project specificities, giving some context to the work to be developed and discussed further.

3.1 PROJECTS' OBJECTIVES

The main objective of this project is to develop a forecast model to predict stadium attendance for each game during the season with a basis on the main Portuguese football league and then adapt it to smaller ones under the FPF sphere.

Furthermore, it would also be essential to comprehend the main factors that drive people to the stadium besides price, getting a clear notion of what could be the best conditions for FPF to organize a particular game, and understanding the rationales behind the consumers of Portuguese football.

When applied the model, increasing audiences will also allow clubs to get a better environment in the stadium, which enhances fans' experience in the stadium (Cho, Lee e Pyun 2019), and increase the performances of the home team as was reported by (Ferraresi e Gucciardi 2020), noticeable especially during pandemic, and increase revenues through ticket and concession sales (Shajie, et al. 2019).

With this being said, once delivered the project, data-driven insights will be generated and provided to FPF regarding attendance trends and determinants in football stadiums in Portugal that could be further applied to minor leagues as they intend to extend the project to liga3, which they are responsible for. The final goal would always be to have stadiums at their total capacity, increasing, this way, the home advantage given that the performance is expected to be better, revenues for the home team and the whole atmosphere around the event.

3.2 ACTIONS TO BE INFORMED

It is expectable to have two different outputs. Firstly, building some graphic visualizations and basic statistics about the data sourced would provide FPF with a better notion of the main determinants of fans going to stadiums to watch football games. Also, the best predictive model results would support the FPF scheduling strategy, considering the possible combinations and

inputs gotten before the game round to have a more significant number of supporters in the stadiums as possible.

3.3 RISKS, ASSUMPTIONS, CONSTRAINTS AND MITIGATION PLANS

In an early phase, we found it vital not to compromise the project, and enumerate what could be the risks, assumptions and constraints associated with the tasks to define strategies to tackle them and be continuously conscient of the inherent risks, considering the dataset available, project objectives and its implementation.

Data			
Concern	*Impact*	*Mitigation Plan*	*Classification*
Data missing	Data, mainly from previous seasons, are missing	Start the analysis from when it is relevant enough	Risk
Inconsistent or incomplete data	Gaps between game weeks can change predictions for a certain round	Fill missing values with median/average values	Risk
Lack of Id's in datasets	Less accuracy on joining datasets	Change and associate manually each key	Risk
Ticket price	One feature that could be an important driver, but it is not available	Make assumptions on prices according to stadiums, opponents, and average price	Assumption

Table 1 - Risk, Assumptions and Constraints on data

Project			
Concern	*Impact*	*Mitigation Plan*	*Classification*
Predictive uncertainty due to other factors	Unpredictable patterns due to human behavior uncertainty	Gather the maximum determinant factors of stadium attendance possible	Risk
Restricted timeline	Limited time to present results	Define a general timeline to follow	Constraint

Table 2 - Risk, Assumptions and Constraints on project

Implementation			
Concern	*Impact*	*Mitigation Plan*	*Classification*
No further adaptability	The model could not perform as expected when applied to other leagues	Define features adaptable to explore in small scale leagues and different contexts	Risk
Lack of Automatization	Some additional features added manually	Code the web scrapping process to be used based on the required inputs	Risk

Table 3 - Risk, Assumptions and Constraints on implementation

3.4 KEY PERFORMANCE CRITERIA

Setting KPIs would also help to streamline project goals, allowing us to evaluate work progression, its weaknesses, and key factors to succeed.

Target	*KPI*	*Metric*	*Formula*
FPF	Increased Attendance	% Stadium Occupation	$\frac{\#\ occupied\ seats}{\#\ stadium\ seats}$
Football Teams	Maximized Revenue	Tickets Sale	$ticket\ price * \#\ units\ sold$
	Increased Home Wins	% Wins per Game at home	$\frac{\#\ wins\ at\ home}{\#\ games\ at\ home}$
Model	Maximized Accuracy	e.g., MSE, RMSE, MAE	
Data	Reduced Missing Values and Errors	Ratio of Missing Values and Errors	$\frac{\#\ missing\ values + \#\ errors}{\#\ number\ of\ entries}$

Table 4 - KPIs

4. Data

In this section, we intend to define all the data pre-procedures to have a better overview of the inputs available and how to manage them to get new and valuable insights from their raw original format.

4.1 DATA STORIES

Before going into detail on data, we found it essential to get explicit the project's impact on the business model and in the data collection process.

Currently, FPF does not schedule the games bearing in mind the maximization of stadium attendance. This leads to uncertainty and unpredictability in predicting game attendance as clubs face the threat of having empty stadiums (as illustrated in figure 1, in appendix).

The scheduling games process would be optimized by adopting the model in question. Furthermore, the model is expected to perform better as many games FPF collects concerning attendance results (as illustrated in figure 2, in appendix).

4.2 DATA COLLECTION

FPF prepared three different datasets related to the game assistances in stadiums, weather conditions, and the number of viewers of the game broadcasted since the 2015/2016 season in the Portuguese football first division.

The first one we excluded from the scope of the project. However, further research on combining stadium and broadcast attendance can be a topic to be explored.

The weather conditions dataset is composed of 4770916 observations supported by 28 descriptive variables corresponding to the hourly weather status for the different locations of stadiums in the Portuguese league.

The other dataset, and the one that is going to work as a basis of our analysis, is the one related to matches info that contains two tables, the main one with the games' stats composed of 2142 matches and 39 variables, being the stadium name column the one that allows joining with the additional table that contains the city of each of the 32 stadiums.

Nevertheless, we found it essential to search for new variables that could be interesting to explore in the model.

Taking into account the stadium table, more geographic and infrastructure info could be valuable to add to it. Therefore, stadium capacity was scrapped from the Wikipedia, complemented with the respective coordinates (latitude and longitude) and calculations on distance in kilometres and minutes from the city centre gotten from google maps. The level/category of each stadium was also extracted from "Liga Portugal", and the average income level for each municipality from Pordata.

Additionally, assigning a table for each team allowed us to get more details about the teams playing, knowing that it influences fans' decisions to go to the stadiums. Hence, we add each team's market value and previous classification for each season from Transfer Market. Moreover, the number of presences in the first division, the coordinates of the club (base), and a yes or no variable referring to historical rivalry were also considered for analysis. Finally, another table to be merged with the calendar variables is covid, represented in a binary format to represent the periods when measures pro-covid were taken by "Liga Portugal".

4.3 DATA CLEANING AND FEATURE ENGINEER

The data collected for the project was extracted and collected from various resources. The data quality is not guaranteed, which affects the model's performance negatively (Ridzuan and Wan Zainon 2019). Thus, we proceeded with some essential data explorations to get some findings on data that does not have the demanded quality. The first step was to get nulls count from the whole dictionary. Knowing previously from FPF that some game info was missing even though the data got scrapped successfully, we realized afterwards that the season 2015-2016 was incomplete, with data missing in every round. Therefore, the most viable option was to delete it. Still, in the null analysis, another interesting finding was related to the occupation rate variable that contained a few missing values. For example, when multiplied by the attendance, it was not always equal to the capacity for each stadium. This led us to replace that column with the division of the stadium capacity with the attendance. Then we pushed all the

values grouped according to season and home team to the next game since it is a stat that comes after the match (the first game of the season is filled with the mean value).

Regarding the division presences, as a numerical quantity represents it with slight practical differences between close numbers, we turned it into a categorical variable of 4 levels divided according to the respective quartiles.

Besides the cleaning of variables, there was also the opportunity for some data engineering processes to transform and manipulate raw data into features that can be efficient for our model and analyses.

Having the results of each game and not being useful for the model since it is post-prediction, one way to turn them valuable was to build a scoring system for the home team (three points in case of victory, one for draws, and zero if it gets defeated). With this function, a lagged accumulated home performance variable was created for each game and one for the result of the previous match at home (the first match filled with the median value).

With the same rationale behind other stats, it could also be essential to include in the model such as the average of goals scored, goals conceded, an average number for shots, ball possession or even fouls to give the model a continuous notion of performance of the home team throughout the season (first match filled with mean values).

There are different tiers of rivalries in Portugal, mainly due to the clubs' size. Also, regionality plays a vital role in it. Based on that, the main rivalries were signalized in the respective matches.

Since there was information about the precise location of the clubs' base, applying the distance formula ($\sqrt{x^2 + y^2}$) to both coordinates, we got how far clubs are from each other. Also, it was helpful to have more descriptive variables discriminated related to DateTime, namely the year, month, day, weekday or hour.

After joining the weather table to the games, a few columns related to rain and wind presented null values that were filled with the 0 value.

Keeping with the data cleaning process, there were also columns to be eliminated. The ones not related to the project's scope would not have relevance in the model or even penalize it. Hence through figure 3, you can see and have a brief notion on how all the merging procedures would allow us to end up with a single final table with all the essential variables considered for analysis.

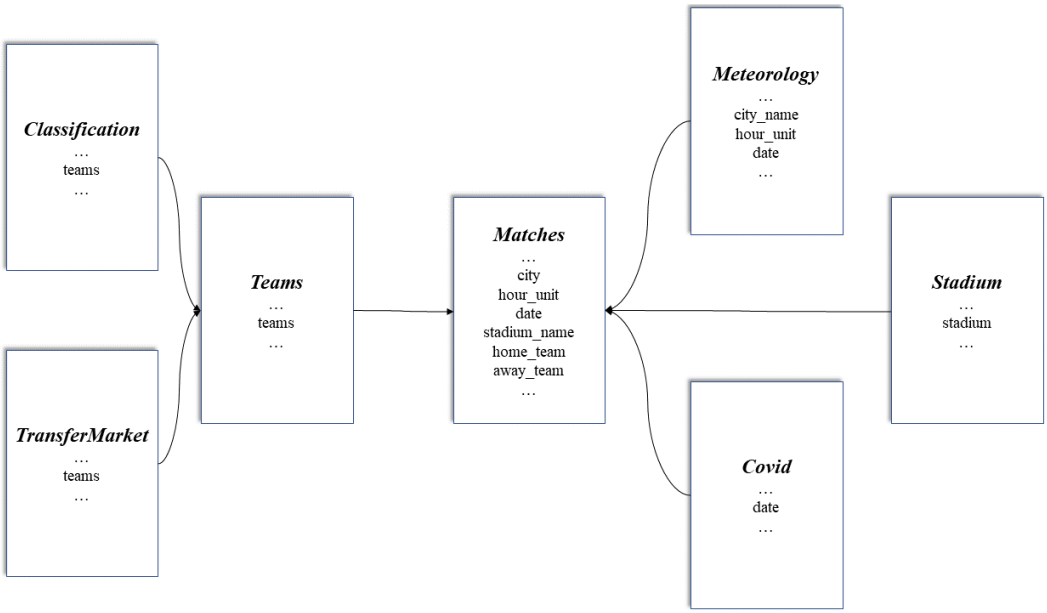


Figure 3 - Simplified database scheme (keys)

4.4 EXPLORATORY DATA ANALYSIS

With this section, the goal is to better understand the data, generate some charts, uncover hidden patterns and withdraw interesting conclusions. Also, to understand which additional data cleaning and preparation tasks are needed. This acknowledgment is important as the final dataset will be the basis of the project.

After completing the data cleaning described previously, we obtained a single data frame, which will be used to perform all the analysis presented on the EDA. Each column’s name and data type are exhibited in table 5, in appendix.

4.4.1 Categorical Variables – univariate analysis

Considering attendance as the project's main scope, the home team and stadium are obvious variables to look at in the first instance. The difference in observations between clubs comes from the presences in the first division being FC Porto, Sporting CP, SL Benfica, SC Braga, Vitoria SC, Boavista FC, Maritimo M, Moreirense FC, Belenenses SAD and CD Tondela the clubs with the most entries in the dataset (graph 1, in appendix). About stadiums, we expect the ones with more observations to belong to those clubs (graph 2, in appendix) with few exceptions, like Belenenses SAD, which played in different stadiums during this period due to bureaucratic issues and disagreements within the institution. Moreover, it is also interesting to check stadium locations on the map and their frequency in observations to conclude about the high concentration of matches in the northern region of Portugal as show in figure 4 (in appendix).

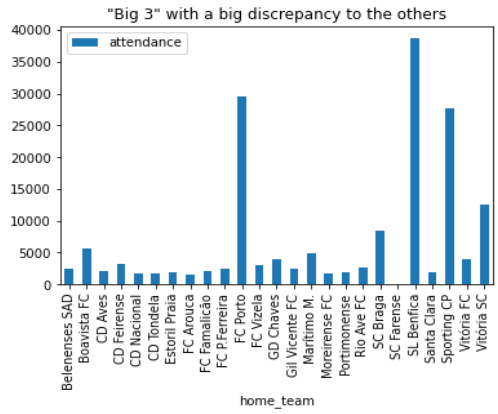
Another crucial determinant of attendance is the time scheduled. Hence, we were interested in finding in what month, weekday or hour most of the games were being listed. On average, January and February are the months when more games are played, with June and July at the bottom of the list (graph 3, in appendix) due to the summer break. However, matches in Portugal used to be in good weather conditions (graph 4, in appendix) during the season despite this distribution with more games in the winter. To attract more people to stadiums, most matches are also during the weekend, especially on Sundays (graph 5, in appendix). The most representative hour is 20:30, with almost 300 games played (graph 6, in appendix).

4.4.2 Numerical Variables- univariate analysis

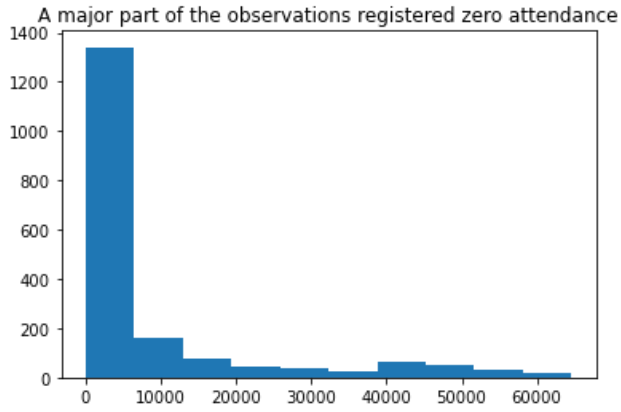
The attendance variable represents the number of fans that watch the game in the stadium. Investigating the attendance numbers in figure 5, in appendix, we observe that the values range between 0 and almost 65000. There is a high concentration of data points below 10000, more than 75% of the observations. This way, it classifies values above 20000 as noise, which may

represent the values the more prominent clubs get us used to in attendance numbers, as graph 7 confirms. Another aspect that stands out is the zero value of “SC Farense”. Considering both shreds of evidence, after analysing the histogram referring to the distribution of attendance values, graph 8, we realize the dataset contains a considerable amount of zeros, which is not desirable since it is our target variable.

Attending these results, it would make sense to test different approaches according to the club size. It is evident the discrepancy between the two groups. And also, take a closer look on attendance zeros, find explanations and decide how to deal with them.



Graph 7 - Attendance per club



Graph 8 - Attendance distribution

Evaluating other aspects besides attendance, we obtained other findings related to the performance of teams when playing at home. And in fact, the influence of playing in front of their fans and getting their support makes the difference in the average of shots made and, more importantly, in goals scored, as shown in figures 6 and 7, in appendix.

4.4.3 Multivariate Analysis

After analyzing individually each numeric and discrete variable, studying the relation between them all may add some value to our research and to find new patterns.

Figure 8 plots the correlation of the numeric variables described in table 6, both in appendix, where we focused our correlation analysis on attendance to figure out the possible main drivers for higher or lower values.

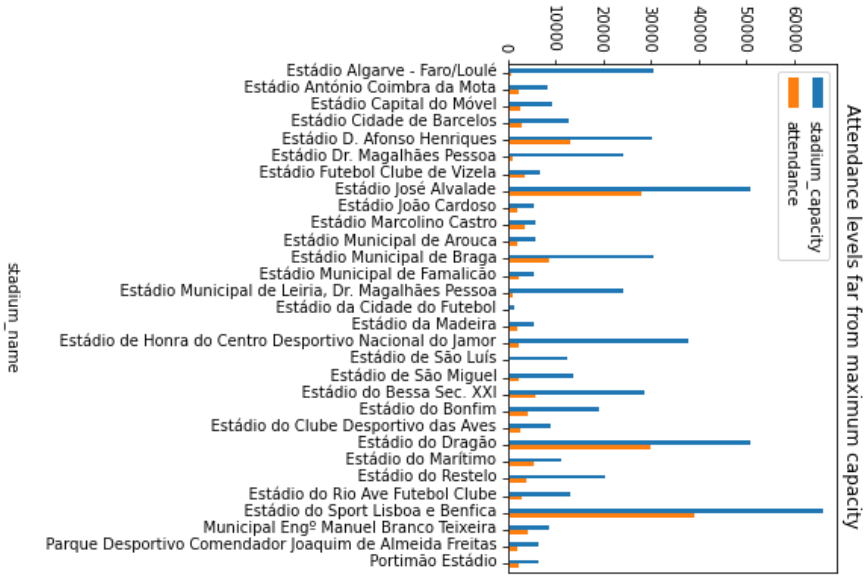
Attendance was already expected to be positively correlated with the occupation rate since they are derived from each other. However, the higher correlation value of 0.61 gives us the confidence to forecast attendance based on previous matches. A theory that is even more supported when looking into the values of the variables referring to the cumulative results and the last result (approximately 0.32 on both). Other ones that also jump out with a strong positive correlation, which means both are expected to follow the same trend pattern, are the "average_home_goal", "market_value_home_team", "avg_home_shots", "avg_home_ball". In the opposite direction, "avg_home_goal_conceded" and "avg_away_shots" give us impressions on the advantages of playing at home in terms of performance. The correlation with the distance to the city center being positive is something that surprised us since it was not expectable stadiums in peripheric locations to attract more people.

On the other hand, attendance is negatively correlated with the year, which may indicate that attendance levels are declining throughout the years. Also, the higher the level of the stadium infrastructure and previous classification (since they represent ordinal variables where one is the best value), the higher the number of spectators in stadiums. Remember that all these correlation results do not guarantee a causality effect, but they are good indicators of the relationship between variables.

Shifting our attention to attendance, throughout this project, we realized every club has its specificities, characteristics, and history, making predicting attendance difficult for each team. They do not start at the same point or position, and in European football, it is mainly like a snowball effect in which the more prominent clubs get more prominent while the smaller ones fight for survival among the sharks, and the Portuguese league does not run out of the rule. The disparity between FC Porto, SL Benfica, and Sporting, followed by SC Braga and Vitoria SC at another level, and the others, is huge. That is evident through the team market value graphs 9 and 10, in appendix. Then the snowball concept is applied when it is traduced in better

attendance and performance as shown through past classifications in graph 11, also in appendix, in the sense that these three variables are expected to be highly correlated with each others.

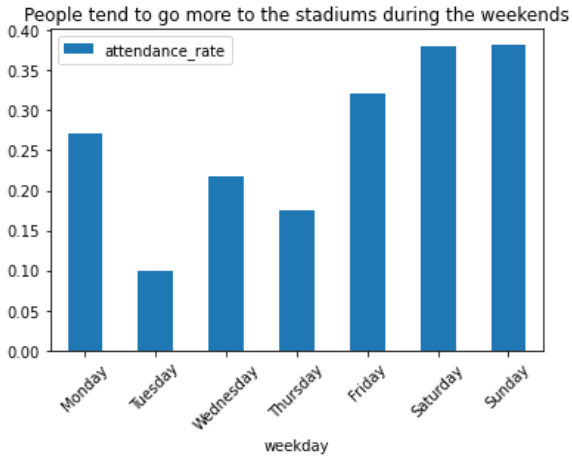
Another aspect that kept our attention is related to the stadium capacity. It can happen when a club registers a large number of fans going to the stadium but with a low occupation rate. This means the club has margins to increase attendance. Alternatively, suppose in the opposite situation, the increase in stadium capacity would probably result in an increase in attendance, which does not appear to be the case since, in Portugal, it is infrequent to have stadiums at total capacity, as shown in the graph 12, below. We cannot forget that covid measures may slightly impact these results.



Graph 12 - Comparison between attendance and stadium capacity

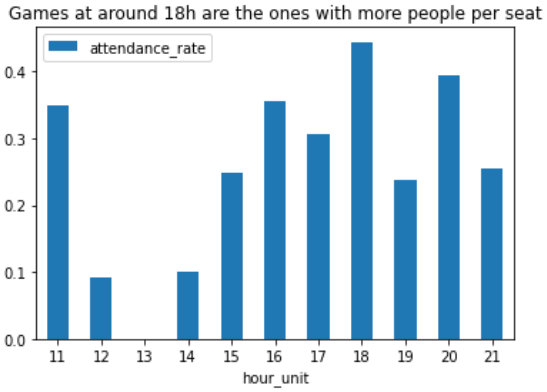
Bearing in mind the differences between the clubs that attracts more supporters to the stadiums and the others, we focused the analysis, this time, on finding patterns among external conditions to the game that could also affect fans' decision to go or not to the game. With this being sad, we selected the day of the week, the hour of the game, and the weather status as possible determinants to attend football matches.

From those graphs, it stands out the propensity for people to watch football in the stadium at the weekends (graph 13).



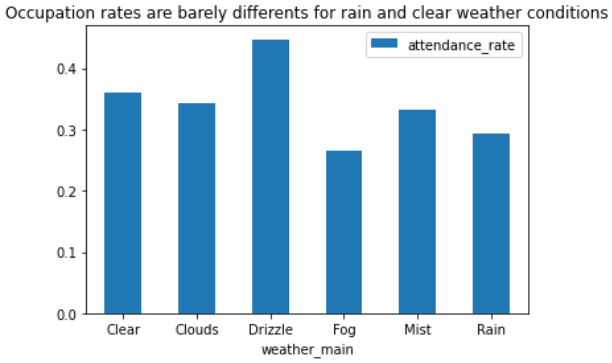
Graph 13 - Occupation rate per weekday

Moreover, games at around 18h are the ones that attract more people to the stadiums, as found on graph 14, despite matches at about 20h being the most representatives (graph 15, in appendix).



Graph 14 - Occupation rate per hour

Weather conditions is a factor that is also connected to the infrastructure of the stadium, since there are stadiums able to protect spectators from rain or even cold and others don't. Considering only "clear", "rain" and "clouds" as the relevant weather status conditions, in graph 16, we got the impression that attendance barely depends on climate condition, with rainy days registering lower values on occupation rate.



Graph 16 - Occupation rate per weather status condition

The influence of the broadcast must be something to consider when selecting the attendance determinants. Nowadays, mainly first-division games, almost every match is streamed and available to football consumers to watch at home. However, this service is not accessible to everyone, with distinct broadcasters charging different amounts of money. The influence of this variable is expected to be more accentuated in inferior divisions with scarce access to broadcasted games, as the only way to see the game is by going to the stadium.

In appendix, in graph 18, we see the divergence between BTV and SportTV broadcasted games in occupation rate. BTV games are directly associated with Benfica's home games which is reflected in the graph since SportTV includes games from every club. Nevertheless, since this project is intended to be adapted to other leagues, this variable could be interesting to incorporate in the analysis or even to do research on combining the maximization of both ways to watch the matches.

4.4.4 Covid Analysis - Diogo Passeiro

Since we got some strange values in attendance, as seen previously, with a large number of zeros, for example, we decided to investigate the cause for that and try to solve it in order to have the best dataset possible to be fed to models.

By general knowledge, we suspected that covid-19 crisis in football could be the cause, as many football leagues suspended activities, and then, stadiums closed doors. So, let's take a look at Portugal's covid seasons and how it affected football:

I. 10/03/2020- Portuguese Football Federation president, Fernando Gomes, decrees games behind closed doors in the 25th round of the 1st and 2nd Leagues and the futsal competitions. At the same time, amateur competitions could only have five thousand spectators.

II. 18/03/2020- A state of emergency is declared in Portugal, with an initial extension of 15 days, determining the closure of facilities intended for sports activities.

III. 07/04/2020- The Portuguese Football Federation cancels the senior non-professional football and futsal championships without awarding titles or enacting divisional climbs and descents. The same measure already had been applied to all younger leagues.

IV. 30/04/2020- After FPF, Benfica, Porto and Sporting meeting with the prime-minister, the deflation plan announced by the Government authorizes the conclusion of the 1st Football League, which still has ten rounds to play, and the Portuguese Cup, whose final will be discussed between FC Porto and Benfica, starting at the end of the last weekend in May. The II League and the modalities practised indoors are excluded from this competitive recovery, while the practice of individual outdoor sports is allowed.

V. 02/05/2020- Portugal transitioned from a state of emergency to a state of calamity. As a result, Vizela and Arouca, clubs with the most points when the Portuguese Championship was suspended, are indicated by the Portuguese Football Federation to move up to the II Liga.

VI. 03/06/2020- The 1st Football League resumes eighty-seven days later, with Portimonense's home triumph over Gil Vicente (1-0), in the first of 90 matches in the last ten rounds of the 2019/20 edition, played with no audience and under a plan approved by the DGS, until July 26th.

VII. 18/09/2020- The 2020/21 edition of the I Liga kicks off simultaneously with the "Campeonato de Portugal", a week after the start of the II Liga.

VIII. 03/10/2020- Eight hundred seventy-three spectators watch the tie between Santa Clara and Gil Vicente (0-0), in the third round of the First Football League, at "Estádio de São Miguel", in the Azores, which was the first professional game with an audience in the stands during the pandemic.

IX. 24/11/2020- Portugal returns to the state of emergency, with an initial extension of fifteen days, authorizing physical and sporting activity in the context of training and competition.

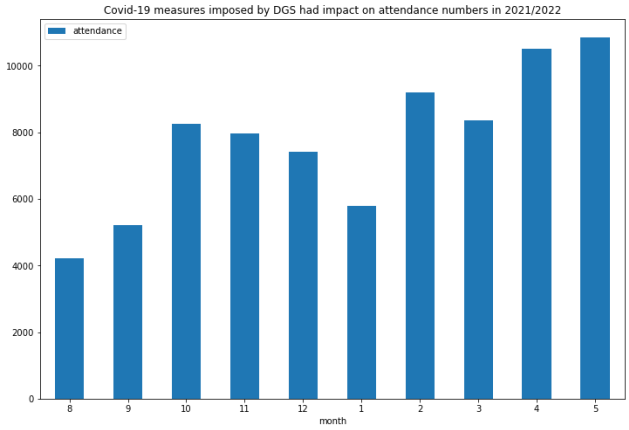
X. 27/02/2021- Nine hundred twenty-four spectators watch the triumph of Santa Clara over Paços de Ferreira (3-0) on the 21st round of the 1st Football League, again at "Estádio de São Miguel". It was the first professional game with fans in the stands in 2021.

In 2020/2021, only six games were played with the public in the stands, with a limited to ten per cent of total stadium capacity, in zones with lower incidence tax.

After seventeen months of silenced matches in stadiums, the government approved attendance in stands for the 2021/2022 season with the condition of filling a maximum of one-third of capacity. Ticket holders were also required to have a negative test or vaccine certification.

Gradually, this capacity limitation passed to fifty per cent and later to one hundred per cent at the beginning of October while keeping masks and test or vaccine certificate presentation mandatory. With another wave threatening Portugal in December, the vaccine certification was insufficient. Those who wanted to go to the stadium had to be tested at least two days before, which demanded a lot of effort from the testing teams to test people in mass for covid-19. During the middle of February, covid measures were eased, and attendance in football got back to normal with all due precautions.

With this being said, now we comprehend better why in 2019/2020 there were 90 matches played with no attendance, 300 in 2020/2021, and only one game registered zero attendance in 2021/2022.



Graph 19 – Monthly attendance in 2021/2022

This graph 19, above, reflects the enumerated measures for the 2021/2022 season on average attendance by month, from the one-third limit capacity to its increase to one-half, until the abolishment of it in October, when the average attendance reached its peak in 2021. The new measures were imposed in December and January, with the second wave that threw attendance numbers again to September levels. Since February, the lifting of measures was visible, reaching numbers in average attendance above ten thousand, getting closer to the values pre-covid. Analyzing independently each match, in figure 9, presented in appendix, especially the ones that registered more attendance, the months when there are more visible differences are August, September and January.

As mentioned previously, we already created a feature named "covid" to signalize the periods affected by covid, expecting the model to learn it and find patterns. However, possessing those values (zeros) in the target variable could be considered inappropriate for such data and may result in biased results (Williford e Aaron 2018).

Bearing this in mind, we eliminated not only all the values from the 2020/2021 season but also every match that registered empty stadiums from the entire dataset. With this action, our final table has gone from 1836 rows to 1439.

4.5 CONCERNS

After all the data cleaning and exploration processes, we finally achieved our final table to apply the pretended model and take out some quantitative insights from it. Despite that, there are still a few concerns regarding the data studied that could become a limitation for the project. For that reason, we found it important to exhibit some of them.

4.5.1 Dataset

We considered the data sourced by FPF not enough to model and predict game assistance with the desired performance. The fact pushed us to find new drivers to predict ticket selling besides the ones already gotten. So we opted for web scrapping some information from public sources on the internet, which was, on the one hand, eased by the characteristics of the data that covers only a few clubs, but on the other hand, it also represents a risk in case of automatization of the model for other leagues.

Another matter of concern is related to the expected model performance since the ticket price does not make part of the scope of the dataset delivered. Price is one of the most critical drivers in every person's action, and watching a football game is no exception (Biscaia, et al. 2013).

Furthermore, as seen in the previous section, there is an apparent discrepancy between SL Benfica, FC Porto, and Sporting CP in attendance and in many other factors, which worries us about its interference with model performance.

The model performance can only be ensured after its display. However, we are in charge of providing the best conditions possible for it to act. In this sense, covid situation adds some dubiety to the model since it could affect the football business and attendance itself in the periods after the outbursts. Therefore, it is up to us and FPF to be aware of it and analyze carefully the before and after pandemic.

4.5.2 Variables

Attending to concerns in more specific cases, there are four issues and assumptions that must be included in the project too.

At the beginning of the 2018/2019 season, the former Portuguese champion "Os Belenenses" lost their rights to own the club to Belenenses SAD, resulting in the split of both, keeping Belenenses SAD the right to play in the first division, but giving up of "Estádio do Restelo" (Amaral 2022). Hence, in the dataset delivered by FPF and to not hinder the analysis, both clubs were considered the same, despite the implications on the team's supporters with this separation that saw themselves also divided.

Additionally, the Portuguese football system implies promotions and relegations of divisions in the league every season, which means new teams are added to the dataset every season while others disappear, hampering model prediction when those teams are in question. Another assumption made was related to the stadium level certification of "Estádio do Algarve" and "Estádio Dr.Magalhães Pessoa" that we classified as level 1 attending their four stars approval by UEFA like "Estádio Cidade de Coimbra", "Estádio do Bessa", "Estádio Municipal de Braga" or "Estádio D. Afonso Henriques". "Estádio da Cidade do Futebol" we classified as level 2 since "Liga Portugal" allowed teams to play there during the pandemic when level 3 stadiums were not allowed to host matches (Pires and Maia 2020).

5. Methodology

Intending to predict attendance for certain conditions correctly, we selected a few regression models within the fields of supervised machine learning since we are dealing with a continuous target variable. Hence, in this section, we pretend to go into detail about how the pipeline was structured for the data to be applied to the models, with the best outcome possible.

5.1 DATA PARTITION

The first process to apply is to split the final table into a training and test dataset, as illustrated in figure 10, in appendix. That is called a partition of a set which consists in gathering all the elements into subsets so that every element is included in exactly one. This way, the training subset will work as a learning mechanism for the machine learning model to discover patterns and correctly predict the output minimizing the error. Once the machine learning model is ready, it needs untrained data to evaluate the algorithm's performance. Therefore, it should represent the original dataset and be large enough to generate meaningful predictions.

If we consider the supervised original dataset as:

$$O = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

Representing x all features and y the target variable. The size of the training dataset is predefined, but its instances are randomly selected. Thus, we end up with two different random datasets:

$$Train = \{(x_1, y_1), (x_2, y_2), \dots, (x_p, y_p)\}, p < k$$

$$Test = \{(x_{p+1}, y_{p+1}), (x_{p+2}, y_{p+2}), \dots, (x_n, y_n)\}, p < k$$

(Dobbin e Simon 2011) state that the most common strategy used for splitting, of $^2/_3$ of instances allocated for training, is considered as the closest to optimal for datasets with n>100. However, pareto principle declares that, for many events, roughly 80% of the effect comes from 20% of the causes (Erridge 2006).

In our case, we decided to do it in such a way as to guarantee an unbiased assigning of the instances into the train and test dataset. Knowing the granularity and scope of the project, our interest is to predict the attendance of a particular team throughout the season. Therefore, we selected bins from each season for each team and sliced them into train and test with a random distribution relation of 85/15, respectively, as shown in table 7.

season	home_team	round	…	attendance	
x_1	a_1	1	…	k_1	
x_1	a_1	2	…	k_2	*Train: 17/20*
…	…	…	…	…	*Test: 3/20*
x_1	a_1	20	…	k_{20}	
x_1	a_2	1	…	k_{21}	
x_1	a_2	2	…	k_{22}	*Train: 17/20*
…	…	…	…	…	*Test: 3/20*
x_1	a_2	20	…	k_{40}	
x_2	a_1	1	…	k_{41}	
x_2	a_1	2	…	k_{42}	*Train: 17/20*
x_2	…	…	…	…	*Test: 3/20*
x_2	a_1	20	…	k_{60}	
x_2	a_2	1	…	k_{61}	
x_2	a_2	2	…	k_{62}	*Train: 17/20*
…	…	…	…	…	*Test: 3/20*
x_2	a_2	20	…	k_{80}	
…	…	…	…	…	

Table 7 - Train and test split

5.2 DATA TRANSFORMATION

Different models present various characteristics and ways of operating. However, all of them ask for the conversion of non-numeric features into numeric (Jain 2020). Also, a few algorithms (e.g., linear, KNN, K-Means, PCA, gradient descent, neural networks) are more sensitives to feature scaling (Roy 2020).

Bearing this in mind, we selected in a first instance the categorical and numeric variables present in our dataset, according to the table 8, in the appendix, and applied different transformations according to the nature and format of the features.

5.2.1 One Hot Encoder

Scikit-Learn library provides the One Hot Encoder to convert integer categorical values into one-hot vectors, turning each numerical value into a binary vector full of zeros with the exception of the respective index, labeled with one. Consistence is also crucial when using this transformation process once it makes it possible to invert our encoding at a later point (Fawcett 2021).

5.2.2 Standard Scaler

One of the most important transformations to apply on data is the feature scaling since very different scales on inputs may affect model performance. Fact that could be useful to attenuate big differences between Portuguese clubs (in the market value or stadium capacity, for example).

The Standard Scaler was the one used to standardize features. It is calculated as:

$$x_{new} = \frac{x - \mu}{\sigma}$$

, which conversely to min-max scaling, it does not restrict values to a specific range, once it may represent a problem when applied to some models such as for neural networks that usually expect an input value ranging between zero and one. Nevertheless, with standardization, the outliers affect less model performance.

5.3 MODEL SELECTION & TUNING

After having all the data prepared to be deployed, to improve the model selection process, model performance needs to be clearly defined and evaluated. Moreover, model parameters should be identified and tested to predict the optimal performance and compare it with all the alternative candidate models (Brooks e Tobias 1996).

5.3.1 Baseline Model – Diogo Passeiro

The creation and application of a baseline model in the first place would allow us to understand if machine learning techniques are necessaire or not on stadiums' occupation forecast, attending to the information already available.

Bearing this in mind, within all the possibilities to declare as the baseline model, our model was built according to the average attendance by each team every weekday in the training dataset. When predictions are compared to the test set, in case it does not recognize any game

of that team on the respective weekday, the minimum attendance value for that same team will be assigned to the predictions.

Note that the datasets input in the model must be the ones before the data transformation step.

5.3.2 Regularized Linear Models

In some cases, when we want to make a model simpler and reduce the risk of overfitting (which occurs when a model accurately predicts the predictions for training data but fails the predictions when it receives new data) we can use two regularized versions of a simple linear model regression. These two versions are called the Ridge and the Least Absolute Shrinkage and Selection Operator (Lasso) Regressions.

5.3.2.1 Ridge Regression

In the Ridge regression model, we include all the features in the model and use an L2-norm penalty. That is, in a model where an input is significantly more relevant (i.e., has a higher weight) in determining the prediction values, a L2-norm penalty is performed to balance all inputs weights. This penalty is done by changing the lambda value to balance all the input weights. This can be useful when having a sample that might not be representative of the true population and we do not want an input to over-influence the predictive values. Below we have the Ridge Regression Error Sum of Squares (SSE) formula.

$$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n} \left(y_i - x_i\hat{\beta}\right)^2 + \lambda \sum_{j=1}^{m} \widehat{\beta_j}^2$$

5.3.2.2 Lasso Regression

The Lasso regression model can be also used to avoid an overfitting of results. The only difference to Ridge Regression models is that it uses an L1-norm penalty that will push certain weights to be exactly zero. This version of a regularized model usually works well when we have a small number of inputs that actually influence the final output and we have others that

have almost no influence in the model result. Below we have the Lasso Regression Error Sum of Squares (SSE) formula.

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|$$

5.3.3 Ensemble Models

Normally, when we are dealing with machine learning models, we select a sample of inputs, use them in a single model and obtain the predicted values. However, if we depend only in the single model results, we may obtain a lower accuracy. If we use ensemble models, predictions from multiple base models are aggregated and used to get a final prediction. This aggregation of several outcomes to obtain a final outcome is usually called "wisdom of the crowd". Ensemble models tend to obtain a better generalization of results and better accuracy than a single model. Each Ensemble model is based on three ensemble methods. These are the Bootstrap Aggregating (Bagging) method, the Boosting method, and the Stacking method.

5.3.3.1 Gradient Boosting

This boosting algorithm is known by combining several “weak” models, so they can be aggregated and used to generate a stronger and more accurate predictive model. This algorithm can be used for both regression and classification problems. Since this model does not require feature preprocessing, it can perform better than random forest models (Silva 2021). However, this model can be more prone to overfitting, can be memory intensive, and requires careful parameter tuning to achieve high accuracy.

5.3.3.2 XGBoost

The Extreme Gradient Boosting model, most known as “XGBoost” is a more regularized version of the Gradient Boosting algorithm, applies L1 and L2 regularization techniques and thus has better generalization capabilities. Its algorithm it is also made to be memory efficient when processing large amount of data and perform parallel computation on a single machine.

5.3.3.3 Light Gradient-Boosting Machine (Light GBM)

One of the models which was tested was the “Light GBM” model. This model can be used for both classification and regression problems. In addition, it has several qualities such as having a fast-training speed and high efficiency, having low memory usage, having better accuracy than other Gradient boosting models, and can process large-scale data (Microsoft 2022).

5.3.3.4 Random Forest

A random forest is a model that works by building multiple decision trees (non-parametric supervised learning models) and merging them together to get a more accurate and stable prediction. Some advantages that come with using this model include ease of use, the fact that it can be trained in parallel on multiple CPU cores, it is not sensitive to outliers, and it has high classification accuracy. In contrast, a random forest model may not perform well in solving regression problems and may not be an easy model to interpret.

5.3.4 Support Vector Regression (SVR)

Support Vector Regression is a supervised learning algorithm model that is used to predict continuous values. While in a simple linear regression the objective is to minimize the error in a Support Vector Regression model the goal is to fit the error inside a pre-determined threshold. This model can handle outliers more easily, it has a good generalization capability with a high prediction accuracy, it is easy to implement but may not be very effective when using large datasets.

5.3.5 Artificial Neural Network (ANN) – Diogo Passeiro

Usually, ANN models suit best for models with numerous features or when the data is continuously scalable, there is the possibility to occur unexpected changes to the inputs, either if there is no need for model interpretation.

The foundations of an ANN are based on nodes connected in a format of a network of artificial neurons and with a correspondent value associated with the strength of connections. Then the network is trained, evaluating all the possible iterations and links to different strength values.

Also, every neuron gets information from others, processes it, and passes the result to another for further processing, like the framework presented below in figure 11.

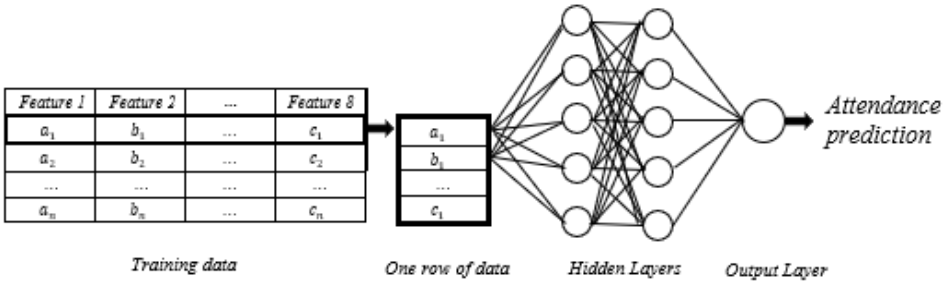


Figure 11 - ANN model illustration

In the framework, as it contains hidden layers (the nodes that perform computations and transfer information from the input to the output nodes), we may call it Multi-Layer Perception (MLP). There, the activation function is applied, making the model capable of learning and performing non-linear tasks. In our case the ReLU (Rectified Linear Unit) is the one in use:

$$R(x) = \max(0, x)$$

This means that for negative inputs, the output is 0.

All the learning process comes from finding and adjusting the optimal combination of weights and biases that minimize the loss function over the training dataset. This can be done through a few known optimization methods besides the backpropagation, which is an algorithm that updates weights according to the partial derivatives of the loss function respective to their own until reaching the minimum value when the derivative becomes zero, as it could cause problems in getting stuck during gradient descent, it could inclusive explode or even running the risk of the neural network overfit to training data.

Other options for optimizers could be, for example, the momentum algorithm that accounts for the direction and speed of the parameters and the Adam that joins the momentum to an adaptive learning rate. Both will be tested, but Adam will be our default option.
Furthermore, batch normalization would also be a valuable technique to improve performance and efficiency in training neural networks. Feature scaling should be applied to both layers, input and hidden ones, normalizing them to speed up the learning process and mitigate the covariate shift that comes from the continuous modifications in the distribution of layers' input and their necessity to adapt to it with the changes in parameters.

Hence, the batch size and the epochs are the other parameters to be adjusted during optimization. The batch size is the number of inputs passed to the network in one go before it changes and updates its weights. To those, we call it one epoch, and so, the number defined of it would correspond to the times the network adjusts weights.

5.3.6 Hyperparameter Tuning & Model Scoring

Next step after model selection is to understand which one is the best and find a way to classify and rank their optimal performances.

In that sense, the application of hyperparameter tuning is the perfect strategy to go through it in searching the ideal model architecture to the best model (Jordan 2017).

There are three methods of hyperparameter tuning in python: Grid search, Random search, and Informed search (Jamal 2021).

In our case, we opted for using the Random search that consists in evaluating a given number of random combinations by selecting a random value for each hyperparameter at every iteration. We choose random search because of its feasibility in computing/processing; however, we are aware that it does not guarantee to find the best score possible.

Before running into the iterations and parameter selection, we need to set the desired score to maximize and the number of folds on cross-validation.

Ideally our model error will be evaluated according four indicators:

a) The MAE (mean absolute error) measures the average extent of errors while forecasting, it is calculated according to the sum of absolute errors divided by the size of the sample.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

b) The MSE (mean squared error) addresses the average squared difference between the observed and predicted values.

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

c) The RMSE (root mean squared error) evaluates the average magnitude of the error giving, at the same time, relatively higher weights to larger errors.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

d) The Standard Deviation estimates how spread the data points are around the mean value.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n}}$$

There are some studies about which one would be a better metric to evaluate model performance. Many researchers choose MAE over RMSE to present their model evaluation, as (Willmott, Matsuura e Robeson 2009) state, while rising some concerns over using RMSE.

Also (Chai e Draxler 2014), in trying to disprove the arguments of the research described above, supporting the use of the RMSE in some cases, they point out that it is more appropriate to use RSME than the MAE when model errors follow a normal distribution which is the opposite of what is expected to happen in our models, once again, due to the disparity on attendance values between teams. Therefore, MAE would be the metric used to evaluate our model performance.

Randomized search also uses cross-validation to select the best parameters, avoid biased evaluations on errors and mitigate overfitting (Yates, et al. 2022). Cross-validation randomly splits the training data into same-sized folds, holding one as the validation set, and training the model on the remaining folds. Then, after repeating the process for each fold, the average is going to represent the selected score for that specific combination, as exemplified in figure 12.

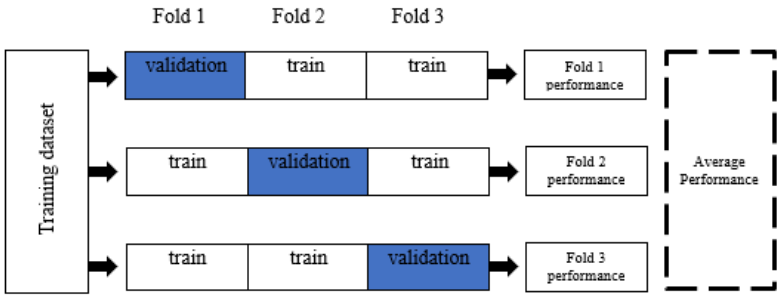


Figure 12 - Cross-validation illustration

In our case the number of samples were set to be 100, the score to maximize was the negative mean average error, according to a cross-validation process of 3 folds, totalizing 300 iterations (*n_iter * cv*).

6. Discussion of Results

After assigning the best parameters for each model to minimize the prediction error as much as possible, we came up with the presented results in table 9 concerning the evaluation metrics applied to both train and test set.

Final Models Evaluation Results

Model	MAE		MSE		RMSE		St.Dev.	
	train	test	train	test	train	test	train	test
Baseline	2698	3372	19535790	37816040	4420	6149	4420	6147
Linear Ridge	2775	3377	15086420	26095890	3884	5108	3884	5106
Linear Lasso	2910	3348	16559200	26183910	4069	5117	4069	5114
Gradient Boosting	1543	2047	5165464	11844370	2273	3442	2273	3439
Random Forest	893	1971	2763675	13619310	1662	3690	1662	3684
SVR	6812	6847	173140560	179013300	13158	13380	12044	12347
Light GBM	1101	1996	3610133	13627980	1900	3692	1900	3686
XGBoost	1222	1927	3131463	11425820	1770	3380	1770	3372

Table 9 - Final results

The model that registered the worse performance, with a significant difference from others, was the SVR model, with an average error of 6847 seats led by its generalizing capacity that predicted incorrectly more significant values on attendance, above the ten thousand spectators, predicting almost every value of that nature below the actual value. This can be explained due to the fact the model associates those values as outliers.

It is followed by the baseline and linear models, both with the L1 and L2 regularization, as the worse models. They registered closed numbers in MAE, but looking at RMSE and the standard deviation associated, the baseline model presents more dispersed data points to the actual value compared with the linear models.

Furthermore, we had access to the coefficient values on Linear Ridge and Lasso regression models, and so, the features that drove it the most to obtain those results. The ones that stood out the most were the season, the round number, the stadium where the match occurred and the hour the match is played. Info related to the market value and previous classification of the home team also represented important factors, emphasizing the importance of those in attendance and the snowball effect, previously described, associated with it, that contributes to the strengthening of the discrepancy between Portuguese clubs.

The ensemble models were the ones with better performance predicting attendance, proving that a single algorithm may not make the perfect predictions for our dataset.

The one that excels is the XGBoost. They all presented an average error of around 2000 spectators when the test set was applied; the Random Forest even had a MAE of below 900 in the training set. But the XGBoost, associated with the respective optimal parameters, was the model that registered the best performance in test regarding all the metrics in evaluation.
Looking into the performance of all the ensemble models in training, we may reject the hypothesis of them being overfitting or underfitting. However, we felt the necessity to evaluate the performance of XGBoost, as the chosen model, on every home team and every season.

Having a look at every team that participated in the predictions in table 10, in appendix, we can take out two main conclusions: Firstly, FC Famalicão is the club with the higher error margin weighted on the average attendance, with the model predicting incorrectly, on average, 2362 stands for the assistance per match of only 2611 spectators per game. Conversely, the "big 3" are again differentiated from the rest. Despite having the higher MAEs, concerning the attendance average, it represented 13.4% for Sporting CP, 16.2% for FC Porto, and 10.7% for SL Benfica. The only exception is Vitoria SC, which, together with these three clubs mentioned before, despite having high numbers in attendance (17341) it registered only 2610 mistakes, on average, predicting occupied seats, resulting in a very low error percentage of 13.4%.

Concerning the table 11, below, referring to the MAE per season, the discrepancy between the 2021/2022 season and the others is clear. This last season accounted for a mean average error of 3208 per game, way above previous seasons, not counting 2020/2021, that we took out from the dataset as mentioned previously, whose values did not surpass the 1700 spectators of error, with exception of 2019/2020 that is still far from the error in 2021/2022.

Season	***2016/17***	***2017/18***	***2018/19***	***2019/20***	***2020/21***	***2021/22***
MAE	1112	1633	1596	2169	-	3208

Table 11 - Error per season

Note in figure 13 that the model overpredicted those values, for that season, which could be explained by the covid measures implemented and uncertainty on attendance associated with it, and constant limitations on stadiums.

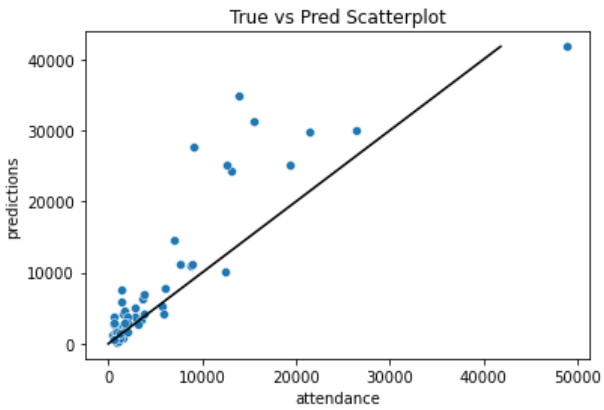


Figure 13 - Precision of predictions for the 2021/22 season

6.1 WHAT IF WE APPLY ARTIFICIAL NEURAL NETWORKS? – Diogo Passeiro

It was expected that ANN would deliver different kinds of results, with a new way to treat the inputs and learn from them on attendance forecast.

After normalizing attendance values, defining the layers, activating the ReLU function, optimize the model according to the Adam method to minimize mean absolute error, we found the optimal combination for the batch size and epochs as 10-100, respectively.

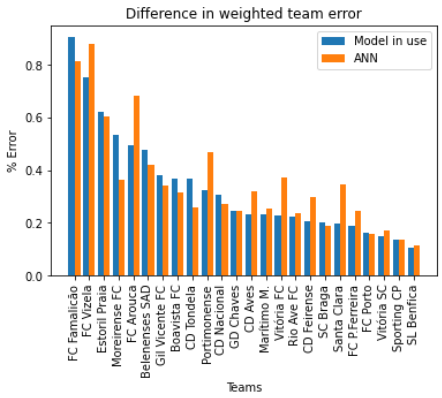
Coming up with the final predictions, we may consider ANN a good contender but not the best model performing.

Model	**MAE**		**MSE**		**RMSE**	
	train	test	train	test	train	test
ANN	493	1987	384125	12431610	620	3525

Table 12 - Results of the ANN model

The results obtained on performance (table 12) are not far from those obtained in the first instance with the best models. However, despite the minor error in training, it was not reflected in the test set, as we see from the significant difference in the MAE or the RMSE.

The explanation for a higher MAE compared to the XGBoost model might be on certain clubs in which it performed significantly poorly, as outlined by graph 20, such as FC Vizela, FC Arouca, Portimonense, CD Aves, Vitoria FC, CD Feirense, Santa Clara or FC P.Ferreira. However, being these clubs might be a good sign to the model in the future since they are not so impactful in absolute terms, not attracting that much public to the respective stadiums. Additionally, CD Aves, Vitoria FC and, CD Feirense are no longer in Liga I.



Graph 20 - ANN vs best model in individualized performance per club

6.2 PROJECT APPLIED TO THE “BIG 3” – Diogo Passeiro

Acknowledging the results obtained throughout this project it is visible two different groups regarding not only attendance but also the other features values, as in previous classifications, market value, and other performance indicators.

This time, as we pretended to avoid significant errors in each prediction, the RMSE was the metric used to optimize and select the best model. However, applying the same process and methods described before for all the teams, we ended up with a lower average error performance for the three clubs, as table 13 indicates. Hence, this time, the Random Forest model was chosen, intending to minimize RMSE.

	Original Project			***“Big3” Project***		
Team	**MAE**	**Attendance**	**%Error**	**MAE**	**Attendance**	**%Error**
Sporting CP	4600	34242	0.134343	4736	33920	0.139625
FC Porto	6083	37557	0.161968	7465	38173	0.195556
SL Benfica	5204	48679	0.106924	5602	477223	0.117393

Table 13 - Original model vs only the "big 3" model

Therefore, we intended to test new solutions to extract the most value possible from the models and obtain better results than previously.

One option was to eliminate from the dataset to be deployed all the periods that were affected by covid and the respective measures in stadium entrance (as were already mentioned in the section about covid). As expected, it changed attendance patterns compared to previous years, even more among the “big 3” that used to attract more people to the stadium. Proof of that is the audit and bias evaluation results between seasons. Season 2021/2022 presented the largest average error by far, followed by the season 2019/2020. This fact led us to think about the exclusion of all matches between April 2020 and March 2022, however, realizing that it would cause interference in training the model for those seasons, as the model would have fewer data points to find patterns, we decided to also exclude from analysis all the data related to the

season 2021/2022, bearing in mind it is still not including the ten rounds played in 2019/2020 with no attendance as we expect less impact on error from it.

With all those changes applied, our training dataset is now composed by 156 data points and the test by 33 (11 per team).

After proceeding with the transformation and deployment of the models according to the new best parameters, new results were achieved, this time, with better outcomes regarding the three clubs under analysis.

We could choose the Light GBM model as it registered the lowest value for MAE with only 3456 missed stand predictions per match with very satisfactory outcomes for the other two clubs. Given that, we obtained more remarkable results in the test, with FC Porto presenting a margin of error of 12%, Sporting CP at only 7%, and SL Benfica at 4%. However, looking closer at predictions for FC Porto, for example, one match forecast registered an error above 15000 (figure 14), which we are trying to avoid in this project. That is the reason why we are prioritizing RMSE this time, considering the similar size between the three clubs.

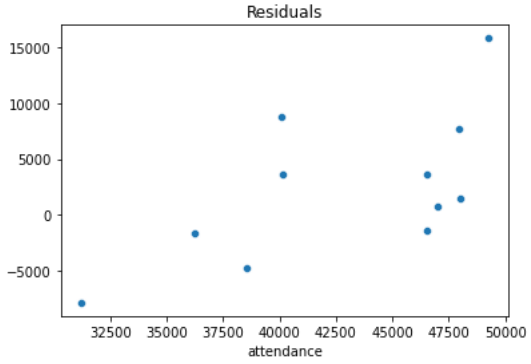


Figure 14 - Difference between prediction and actual values of FC Porto

Taking into account what has been said, XGBoost was considered, once again, the overall best model to be used in this situation. Even despite not presenting such good results for SL

Benfica, as in other models, the general 4419 RMSE value shortened the error range. Figure 15 gives us a good perception of it.

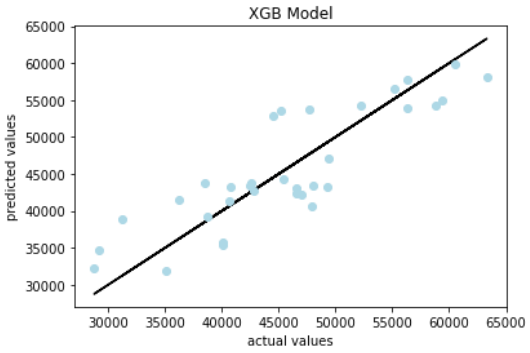


Figure 15 - Prediction results with XGBoost

With this model FC Porto is still presenting a margin of error of 12%, Sporting CP of only 5%, and SL Benfica 7%, as compounded in table 14, the results for all the best models of each approach.

	Original Project			***"Big3" Project***			***"Big3" Project without 21/22***		
Team	**MAE**	**Attendance**	**%Error**	**MAE**	**Attendance**	**%Error**	**MAE**	**Attendance**	**%Error**
Sporting CP	5154	34242	0.15	6191	33920	0.18	1995	39609	0.05
FC Porto	5412	37557	0.14	6167	38173	0.16	5216	42852	0.12
SL Benfica	5400	48679	0.11	6322	477223	0.13	4068	54484	0.07

Table 14 - Difference in errors between the 3 approaches

The results also showed us, again, the potential to improve performances since in the 19/20 season, the last one to be accounted in the model, the average error was larger than the others (table 15), a fact explained by the less quantity of matches as data points in the training set.

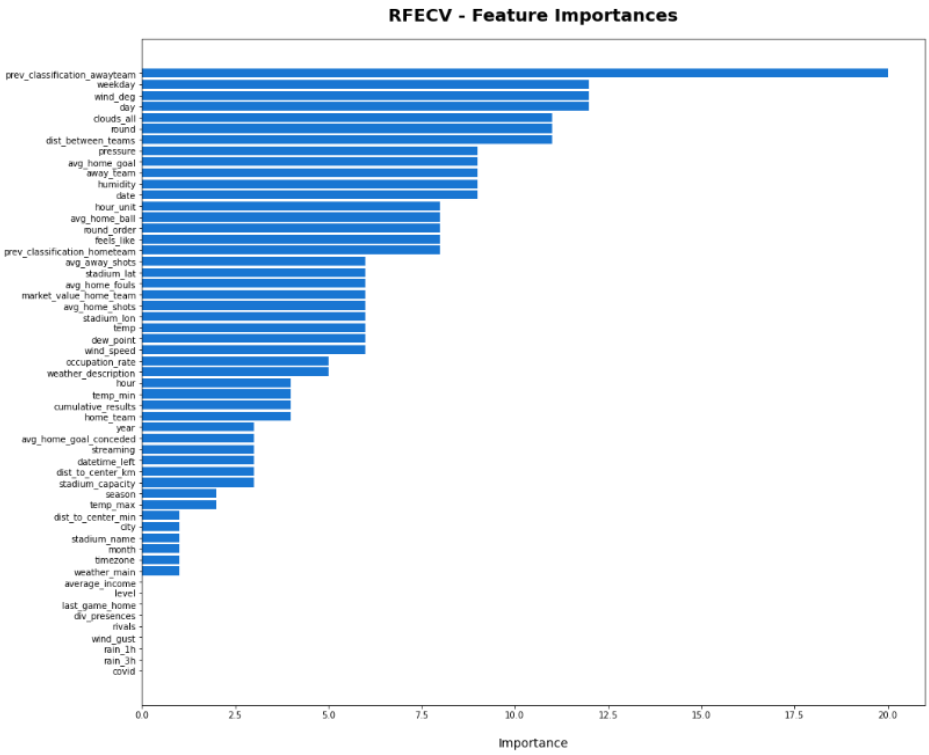
Season	***2016/17***	***2017/18***	***2018/19***	***2019/20***	***2020/21***	***2021/22***
MAE	3609	3241	3358	5364	-	-

Table 15 - MAE per season

We also tried to improve its performance by testing the ANN models applied to this short list of clubs since, as mentioned previously, the "big 3" affected the average error in absolute terms among all the clubs more. However, the results were not satisfactory. It registered an RSME of 5485 and an MAE of 4786, culminating in even larger errors, especially for Sporting

CP, compared with the results shown before, which are going to be the ones that we should stick with for the analysis and attendance forecast of these three clubs.

Another valuable input in this analysis is the feature importance, defined below in graph 21.



Graph 21 - Feature importance

The features that contributed the most to the prediction were the previous classification of the away team, the weekday, the wind and cloudiness, the day of the month, and the round number and distance between clubs which could be helpful and an issue to take into account for FPF in future works of expansion, knowing these are the variables more relevant in forecasting. Moreover, if the project starts to scale to smaller leagues, under the scope of FPF, it would be beneficial to rely on fewer variables and so, to build the model with fewer features while keeping the quality standards, in the sense that it is expectable not to have that large amount and variety of data available.

Also, by examining each of these variables and analyzing the strength of correlation with the target for these groups of clubs, we can infer the strong possibility of them having a causality effect on attendance results. In light of this, recommendations could be made to these clubs to

increase stadium attendance (putting aside ticket price changes). Firstly, as the distance between opponents is a crucial factor, and matches between neighbor clubs are more prone to have more public in stands, which could be in part explained by the inherent rivalry and the emotions that are traduced in fans with that, we would recommend SL Benfica, FC Porto, SC Sporting to present fans when playing against more distant teams that usually attract less public, historical facts between both teams to create empathy on supporters and leverage the rivalry between clubs. The same rationale may be applied to away sides whose previous classification was poor since the model predicts (as the main driver) weaker teams to attract fewer people to stadiums.

Regarding the best day in the month to arrange a match, there is no straightforward correlation with attendance numbers. However, in graph 22, in appendix, it is visible how stadiums tend to be more closed to be sold out in the first fortnight of the month. Nonetheless, playing on Saturdays also seems to be the best option to sway stadium attendance. Additionally, in the middle rounds of the league, clubs are expected to register less adherence to stadiums, as presented in graph 23, in appendix. Therefore, these facts must be considered by clubs that host the match in order to increase ticket sales in off-peak periods.

Attending Cloudiness and wind conditions, they also look to be guilty for many people not going to the stadium, as temperature, which could be fought with investments in infrastructures and stadiums. These recommendations are in line with (Hall, O'Mahony e Vieceli 2010) on attendance at sporting events being affected, not only by the emotion attached to the event, but also, to the perceived quality and availability of facilities.

6.3 AN ANALYSIS ON THE OCCUPATION RATE VARIABLE - José Sousa

After applying the models to the dataset and observing their results, other ways to obtain better results were sought. Although the predictive models applied to the attendance, as the predictive variable, showed satisfactory results, they may not show the full picture of all of the Portuguese clubs. That is because capacity is not the same across all stadiums. Thus, we also

analysed how the models would behave if using the occupancy rate instead. Namely, the division between the attendance by the stadium capacity, as the response variable. Using the occupancy rate as the predictive variable, we can adjust the attendance to the full capacity of each stadium and possibly obtain better results. This approach had already been followed by authors such as (Şahin and Erol 2018).

To implement this approach, we kept the same variables, models and evaluation metrics used in the previous predictive models, with attendance as a dependent variable.

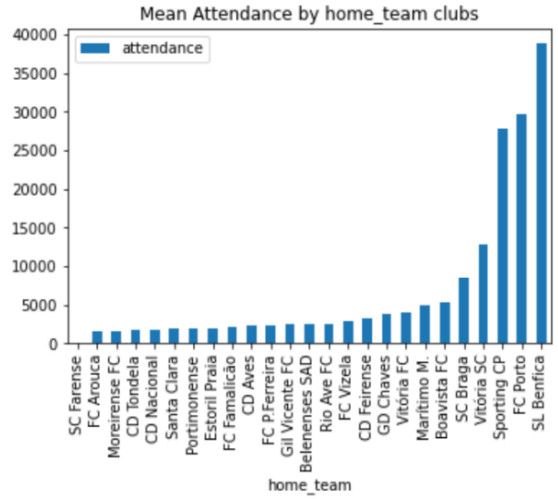
6.3.1 Performed changes in the data and main code

Furthermore, before running the models, solely replacing the attendance output variable with the occupation rate was not enough. Hence, the data used for the models suffered small adjustments. Also, the parameters of the models had to be fine-tuned for the change in data. The new values of these parameters were provided by the cross-validation process made for each model.

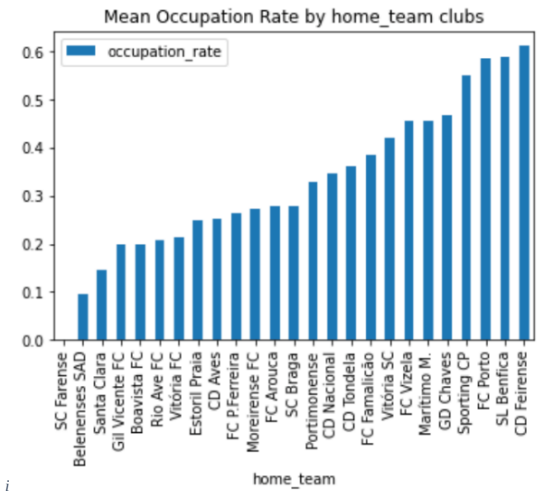
Concerning the data changes, the attendance column had to be moved one row down so that each row kept the past home game attendance as a predictor variable. This action left rows with null values for the variable attendance for every first match of every season. To solve this matter, we calculated the mean, by season, of each team when playing at home and used it to fill all null values accordingly. Regarding the occupation rate column, since there were several clubs with also null values, they were replaced by the average occupation rate of the respective home team.

6.3.2 Resulting changes Data Visualizations

We performed data visualization to better understand the data. Regarding the attendance variable, there is a significant difference in values between some teams. Concerning the occupation rate variable, the differences become smaller across teams.

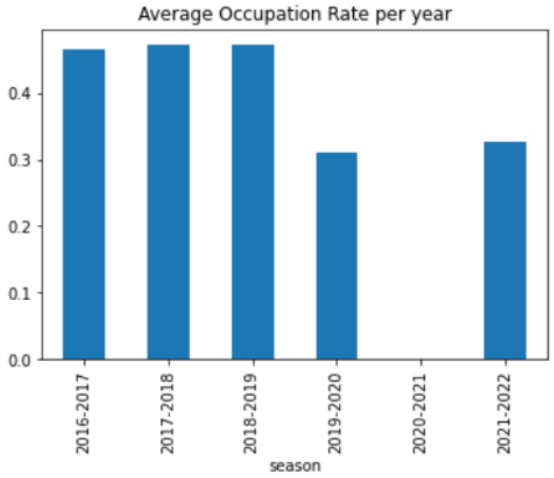


Graph 24 – Mean attendance by home team



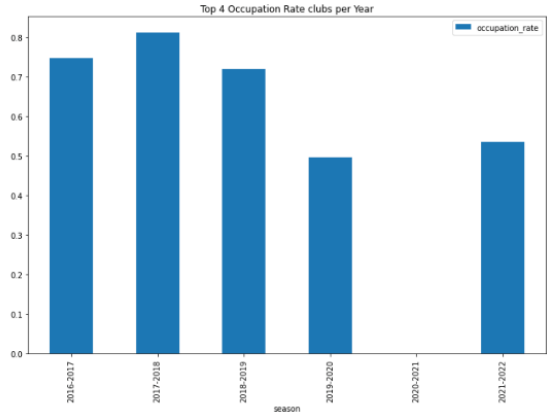
Graph 25 – Mean occupation rate by home team

Moreover, we can take significant conclusions with respect to the Portuguese league attendance rates over the years. As the graph below shows, apart from covid affected years, on average, clubs fill their stadiums slightly more than 40% of their capacity.

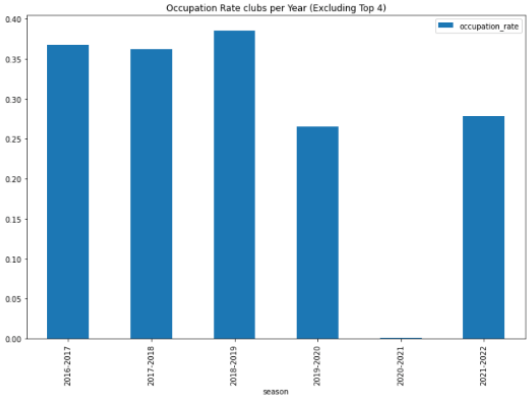


Graph 26 – Mean occupation rate per year

When removing the four clubs with the highest occupation rate (CD Feirense, SL Benfica, FC Porto and Sporting CP), we observe that the average yearly occupation rate for the remaining clubs is below 40%, while the four clubs with the highest occupation rate have all an occupation rate above 70%. This excludes covid affected seasons (2019-2020, 2020-2021, 2021-2022).

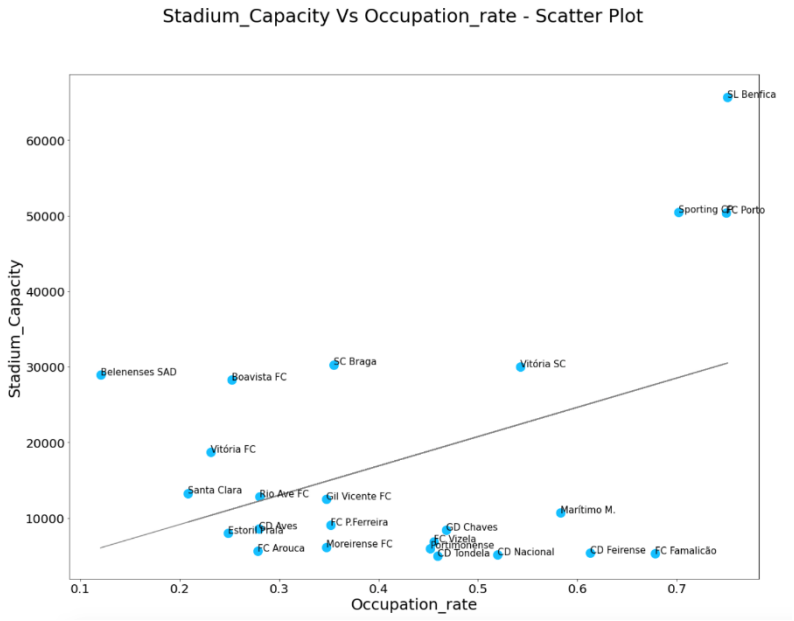


Graph 27 – Top 4 occupation rate clubs per year



Graph 28 – Occupation rate clubs per year (excluding top 4)

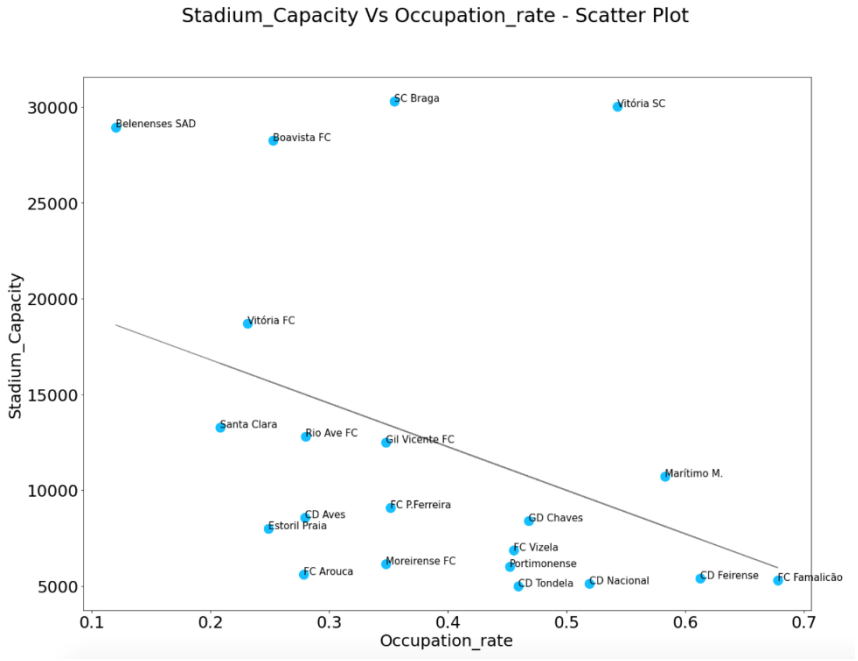
Another visualization we can take from the data concerns the occupation rate and its link with the stadium capacity. As we have seen previously, most stadiums in the Portuguese 1st division usually have a low occupancy rate. Indeed, Portugal is still way below when compared to the clubs in the top five European leagues. To put it into perspective, according to (Batardière 2018), the Portuguese 1st football division stands in 13th place concerning the occupancy rate of all 1st football league divisions of Europe. There are two ways to explain the low occupancy rate for the Portuguese league. Either we enumerate the several factors that influence stadium attendance, as mentioned previously, or we consider that some clubs use stadiums way too big in capacity for their dimension as a club. Through the graphs below, we can learn relevant insights about the latter. In the first graph 29, we can see a positive relationship between the stadium capacity and the occupation rate.



Graph 29 – Occupation rate and stadium capacity

However, looking closely at the data, this conclusion does not seem very accurate. The reason is the existence of three outliers: SL Benfica, FC Porto and Sporting CP. We can see that they have a much greater stadium capacity compared to other clubs in the 1st league and also a very high occupation rate. These positions the three clubs very far from the rest of the points respective to other clubs. In turn, their positions can influence the way the trendline behaves. Hence, we decided to proceed with an analysis similar to the previous one but removing these three clubs. By removing these three outliers, in the graph below**,** the relationship between the stadium capacity and the occupation rate changes and becomes negative. Hence, as stadium capacity increases, the occupation rate contrarywise tends to decrease.

Furthermore, the main conclusion of these two graphs is that clubs in the Portuguese 1st league, except for SL Benfica, FC Porto, and Sporting CP, have a stadium too big for their usual match demand. This reality is even more evident when the mean occupation rate is below 0.6 for all other clubs in the league except FC Famalicão and CD Feirense.

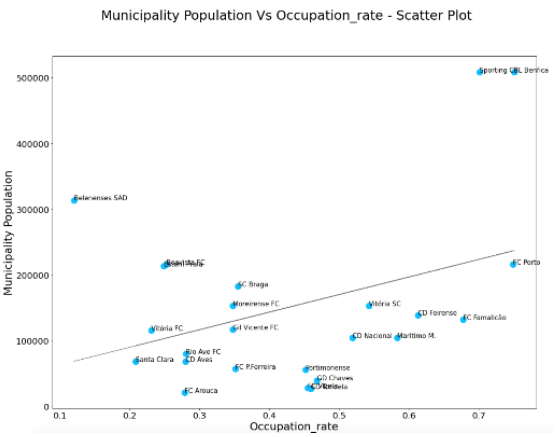


Graph 30 – Occupation rate and stadium capacity (excluding top 3)

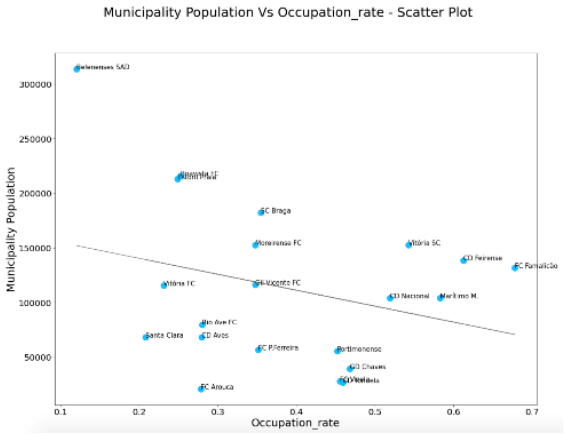
6.3.3 Adding the stadium surrounding area population analysis

We decided to analyse whether the population of municipalities where 1st league football matches take place correlate with the occupation rate of those stadiums on average. As we can see in the graph 31 there is a positive relationship between the average municipality population of each stadium and its occupation rate. Therefore, on average, when a stadium is in a more populated municipality, the occupation rate for that stadium is also higher.

Due to the dimension difference and to be in line with the stadium capacity vs occupation_rate analysis, we also checked the same relationship without SL Benfica, FC Porto, and Sporting CP. The relationship becomes negative (graph 32). This means that if we are not considering these three outliers, on average, when a stadium is in a more populated municipality, the occupation rate for that stadium is smaller.



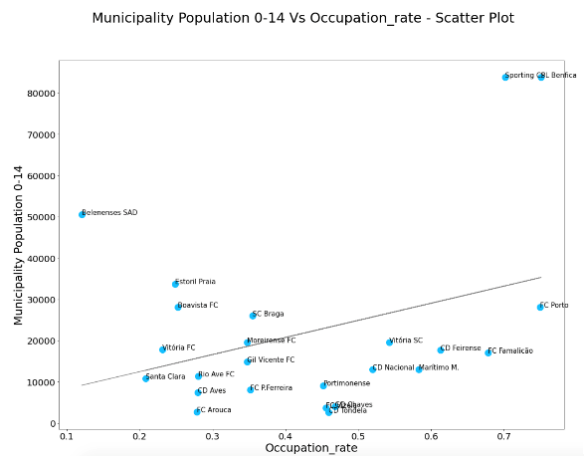
Graph 31 – Population and occupation rate



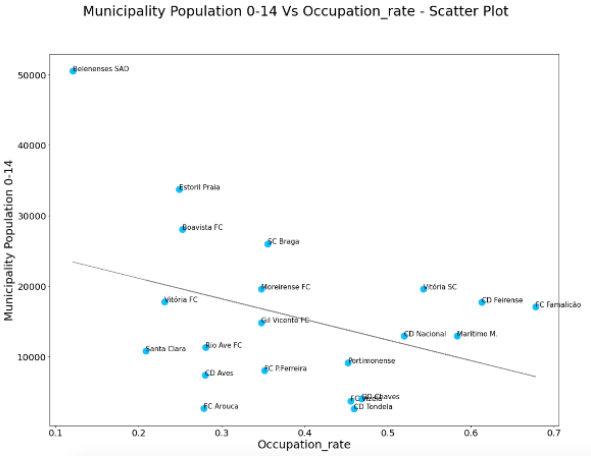
Graph 32 – Population and occupation rate (excluding top 3)

Moreover, we split the municipality population variable into three age groups, 0 up to 14 years old, 15 up to 64 years old, and older than 65 years old. The aim is to find insights into how different generations may behave when deciding whether to attend a football match when occurring in their respective municipality.

Concerning younger people, as we can see in the graph 33, from 0 to 14 years old, there is also a positive relationship between the average municipality population of each stadium and its occupation rate. There is also a negative relationship once the three outliers are removed, as we can see in graph 34. Nevertheless, the increase in the graph 33 is sharper compared to the graph 31 and the decrease in the graph 34 is also sharper than the graph 32. This higher steepness may reveal a possible disengagement of the younger generations in football as the population of the municipalities where the matches take place increases, and also when the three outliers are removed from the analysis.

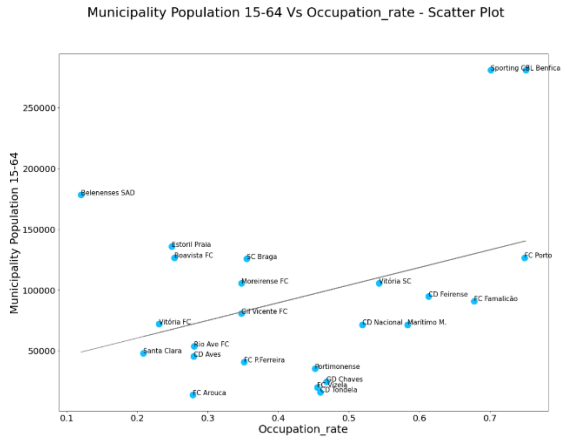


Graph 33 – Population u14 and occupation rate

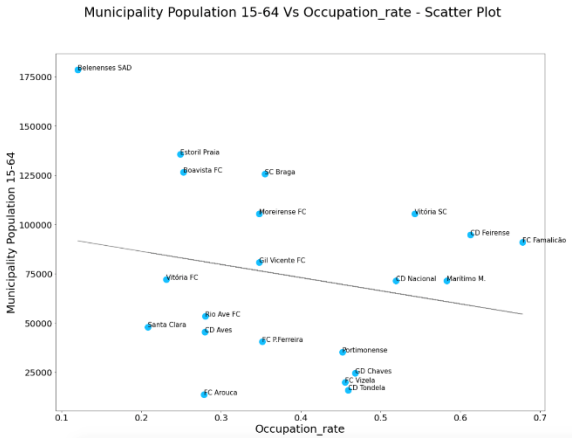


Graph 34 – Population u14 and occupation rate (except top 3)

On the contrary, when examining people between 15 and 64 years old, the relationship is the same as in the previous graphs. Specifically, there is a positive relationship between the average municipality population of each stadium and its occupation rate. Also, there is a negative relationship once the three previously mentioned outliers are removed. Though the increase in graph 35 is smoother compared to graphs 31 and 33, and the decrease in graph 36 is also flatter than the graphs in graph 34 and graph 32.



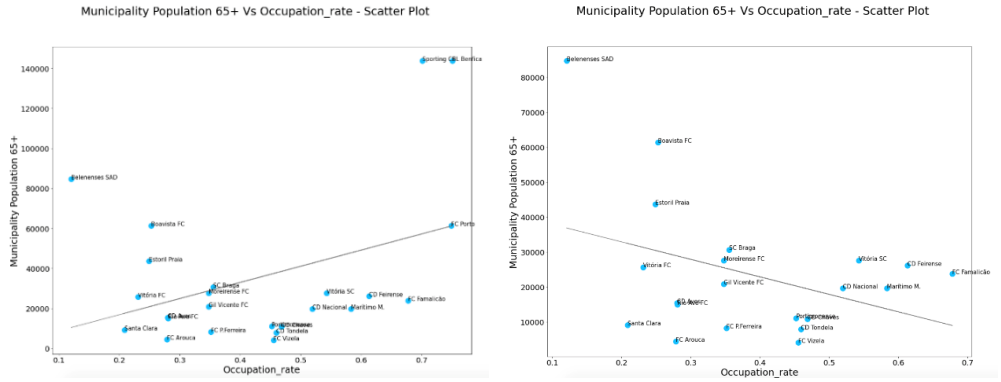
Graph 35 – Adult population and occupation rate



Graph 36 – Adult population and occupation rate (except top 3)

Concerning the age group older than 65 years old, also there is a positive relationship between the average municipality population of each stadium and its occupation rate. As in previous cases, there is also a negative relationship once the three outliers are not taken into consideration. There is a sharper increase in graph 37 compared to graph 35, but smoother than

graph 33 and there is also a steeper decrease in graph 38 compared to graph 36, but smoother than graph 34.



This analysis raises a concern regarding the engagement of younger generations to attend football stadiums and contributes to the occupation rate. This may be caused by a either perceived, or real lack of safety in stadiums or may be related to costs.

6.3.4 The Results

Concerning the performance of the models, we can say that they presented relatively positive results. Through the graph below, we can observe that the model that had the best performance was the Extreme Gradient Boosting Machine Model, with a Mean absolute error of 0.079. In contrast, the model that performed the worst was the linear lasso model with a mean absolute error of 0.135. Surprisingly, the Baseline model, which was the simplest model of all, did not have a much lower Mean Absolute Error (0.125) when compared to the rest of the models. Regarding the Mean Squared Error (MSE), all the models showed similar values. The Gradient Boosting, LightGBM, and XGBoost models showed the highest results 0.013, 0.013, and 0.012 respectively. The worst performing model, concerning MSE, was the Linear Lasso which had a similar performance to the Baseline model. The Root Mean Squared Error (RMSE) metric, showed that the Baseline Model had the highest value 0.172, compared to the XGBoost model RMSE of 0.112. Finally, the last model evaluation metric used, the Standard Deviation (St.Dev) showed relatively close values for all of the models being the XGboost, the model

with the lowest St.Dev of 0.112 compared with the Baseline model which showed the highest standard deviation value of 0.172.

In the table below, one can analyze the overall metrics for each model used.

Final Models Evaluation Results

Model	MAE		MSE		RMSE		St.Dev.	
	train	test	train	test	train	test	train	test
Baseline	0.165	0.125	0.048	0.029	0.220	0.172	0.220	0.172
Linear Ridge	0.077	0.099	0.010	0.018	0.101	0.135	0.101	0.133
Linear Lasso	0.138	0.135	0.031	0.028	0.174	0.167	0.174	0.167
Gradient Boosting	0.047	0.085	0.003	0.013	0.059	0.118	0.059	0.116
Random Forest	0.042	0.094	0.003	0.016	0.058	0.126	0.058	0.125
SVR	0.067	0.099	0.005	0.018	0.076	0.136	0.075	0.132
Light GBM	0.032	0.083	0.002	0.013	0.044	0.115	0.044	0.113
XGBoost	0.013	0.079	0.000	0.012	0.017	0.112	0.017	0.112

6.3.4.1 Clubs first division participations and model error correlation

Not all the clubs were present in the first Portuguese division league for all the years under analysis. Hence, we analyzed the possibility of this case being correlated with the error. Through the graph below, we can conclude that this hypothesis does not hold since the correlation value is -0.449. This is assuming that a strong correlation is 0.8 or above (W. Mulcahy and L. Gregory 2009).

Correlation	*Seasons Participated*
MAE	-0.449

7. Implications

Projects of this type can be beneficial for clubs. The fact that clubs can identify which factors influence the fans' decision to attend a football match is essential for clubs to enhance their relationship with both their supporters and partners.

With such knowledge, clubs can improve the fans' experience during a game. They can collaborate with television broadcasters to better plan the game's schedule, invest in the club more efficiently, and even define better merchandising marketing strategies.

Also, as certain clubs have a small variety of income sources, projects like this could positively impact their revenues.

8. Limitations & Future Work

Although we were able to find interesting results through the models made, some of their findings in this study were subject to some limitations.

The fact that we removed the covid data because they were not fundamental for our analysis constrained some of the accuracies of our models since, typically, the more data we have, the better the results we can get from the models. Adding to this, as we also removed the data relative to the 15/16 season, we were left with only five seasons out of the seven that were initially provided to us. Another limitation we had concerning the data was its non-homogeneity. As in most of the seasons, there are clubs that are relegated and others promoted to the first division. The models could not perform as well for these clubs as those consistently maintained their presence in the first division. Such a limitation left us with only 10 out of 25 clubs that had data for all of the seasons analyzed.

Furthermore, our models could also have given us better results if we had added more features, such as the average ticket price. Although this variable could significantly increase the fans' decision to go to a stadium, we were unable to successfully apply this variable due to the limited time we had for the project.

For future studies on this subject, adding this factor to the model would have to be well-founded. Mainly because it needs to have more information regarding the past ticket prices of many clubs and because an incorrect generalization of the average prices of the clubs that hold this information towards the rest of the clubs could be detrimental to the obtain better results. To replace the significant weight of this variable in the model, one can look for other variables that could influence stadium attendance numbers.

References

Addesa, Francesco, and Alexander John Bond. 2021. "Determinants of Stadium Attendance in Italian Serie A: New Evidence Based on Fan Expectations." *PLOS ONE* 16 (12): e0261419.

Alberto Rizzoli. 2022. "7 Game-Changing AI Applications in the Sports Industry." October 21, 2022.

Baimbridge, Mark. 1997. "Match Attendance at Euro 96: Was the Crowd Waving or Drowning?" *Applied Economics Letters* 4 (9): 555–58.

Batardière, K. 2018. *European Leagues Fan Attendance report*

Borland, Jeffery, and Robert Macdonald. 2003. "DEMAND FOR SPORT." *Oxford Review of Economic Policy* 19 (4).

Bradbury, John Charles. 2020. "Determinants of Attendance in Major League Soccer." *Journal of Sport Management* 34 (1): 53–63.

Ćwiklinski, Bartosz, Agata Giełczyk, and Michał Choraś. 2021. "Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers." *Entropy* 23 (1): 1–12.

Gimet, Celine, and Sandra Montchaud. 2016. "What Drives European Football Clubs' Stock Returns and Volatility?" *International Journal of the Economics of Business* 23 (3): 351–90.

Hautbois, Christopher, Frédéric Vernier, and Nicolas Scelles. 2022. "Influence of Competitive Intensity on Stadium Attendance. An Analysis of the French Football Ligue 1 over the Period 2009-2019 through a Visualization System." *Soccer & Society* 23 (2): 201–23.

Lee, Hansoo, Bayu Adhi Tama, and Meeyoung Cha. 2022. "Prediction of Football Player Value Using Bayesian Ensemble Approach," June.

Lo, Dominic. 2011. "Football, The World' s Game: A Study on Football' s Relationship with Society."

Majumdar, Aritra, Rashid Bakirov, Dan Hodges, Suzanne Scott, and Tim Rees. 2022. "Machine Learning for Understanding and Predicting Injuries in Football." *Sports Medicine - Open* 8 (1): 73.

Martins, António Miguel, and Susana Cró. 2018. "The Demand for Football in Portugal." *Journal of Sports Economics* 19 (4): 473–97.

Microsoft. 2022. "LightGBM."

Reade, J. James, Dominik Schreyer, and Carl Singleton. 2022. "Eliminating Supportive Crowds Reduces Referee Bias." *Economic Inquiry* 60 (3): 1416–36.

Ridzuan, Fakhitah, and Wan Mohd Nazmee Wan Zainon. 2019. "A Review on Data Cleansing Methods for Big Data." In *Procedia Computer Science*, 161:731–38. Elsevier B.V.

Şahin, Mehmet, and Rizvan Erol. 2018. “Prediction of Attendance Demand in European Football Games: Comparison of ANFIS, Fuzzy Logic, and ANN.” *Computational Intelligence and Neuroscience* 2018.

Schmidt, Sascha L., and Florian Holzmayer. 2018. “A Framework for Diversification Decisions in Professional Football.” In *Routledge Handbook of Football Business and Management*, 3–19. Routledge.

Schreyer, Dominik, and Payam Ansari. 2022. “Stadium Attendance Demand Research: A Scoping Review.” *Journal of Sports Economics* 23 (6): 749–88.

Sheng, Bin, Ping Li, Yuhan Zhang, Lijuan Mao, and C. L. Philip Chen. 2021. “GreenSea: Visual Soccer Analysis Using Broad Learning System.” *IEEE Transactions on Cybernetics* 51 (3).

Silva, Francisco. 2021. “A MACHINE LEARNING APPROACH TO PREDICTING STOCK RETURNS.”

Uefa. 2022. “The European Club Footballing Landscape Club Licensing Benchmarking Report Living with the Pandemic.”

Valenti, Maurizio, Nicolas Scelles, and Stephen Morrow. 2020. “The Determinants of Stadium Attendance in Elite Women’s Football: Evidence from the UEFA Women’s Champions League.” *Sport Management Review* 23 (3): 509–20.

W. Mulcahy, John, and Jess L. Gregory. 2009. *A Handbook of Statistics and Quantitative Analysis for Educational Leadership*. University Press of America.

Watanabe, Nicholas, and Brian Soebbing. 2017. “Chinese Super League: Attendance, Pricing, and Team Performance.” *Sport, Business and Management: An International Journal* 7 (2): 157–74.

Amaral, Nuno A. 2022. “Todos os capítulos do divórcio entre o Belenenses e a SAD.” *JN*.

Biscaia, Rui, Abel Correia, António Rosado, João Marôco, e Masayuki Yoshida. 2013. “The role of service quality and ticket pricing on satisfaction and behavioural intention within professional football.” *International Journal of Sports Marketing and Sponsorship* 42-66.

Brooks, R. J., and A. M. Tobias. 1996. “Choosing the best model: Level of detail, complexity, and model performance.” In *Mathematical and Computer Modelling*, 1-14.

Chai, T., e R. R. Draxler. 2014. *Root mean square error (RMSE) or mean absolute error (MAE)? –*. Copernicus Publications .

Cho, Heetae, Hyun-Woo Lee, e Do Young Pyun. 2019. “The influence of stadium environment on attendance intentions in spectator sport: The moderating role of team loyalty.” *International Journal of Sports Marketing and Sponsorship* 276-290.

Dobbin, Kevin, and Richard Simon. 2011. "Optimally splitting cases for training and testing high dimensional classifiers." 4 8: 4-31.

Erridge, P. 2006. “The Pareto principle.” *British Dental Journal* 201-419.

Fawcett, Amanda. 2021. “Data Science in 5 Minutes: What is One Hot Encoding?” *educative.*

Ferraresi, Massimiliano, e Gianluca Gucciardi. 2020. *Team performance and audience: experimental.* Pavia: Società italiana di economia pubblica.

Hall, J., B. O'Mahony, e J. Vieceli. 2010. “An empirical model of attendance factors at major sporting events.” *International Journal of Hospitality Management* 328-334.

Jain, Deepak. 2020. “Feature Handling: Categorical and Numerical.” *Towards Data Science.*

Jamal, Tooba. 2021. “Hyperparameter tuning in Python.” *Towards Data Science.*

Jordan, Jeremy. 2017. “Hyperparameter tuning for machine learning models.” *Jeremy Jordan.*

Pires, Sérgio, and Vítor Maia. 2020. “Dúvidas e certezas: os estádios que vão a jogo.” *MaisFutebol.*

Roy, Baijayanta. 2020. “All about Feature Scaling.” *Towards Data Science.*

Shajie, Kianoosh, Mahdi Talebpour, Seyed Morteza Azimzadeh, Mohammad Keshtidar, e Hadi Jabbari Nooghabi. 2019. “Spectators on the Run: Factors Affecting Football Attendance in.” *Ann Appl Sport Sci.*

Williford, Anne, e Boulton Aaron. 2018. “Analyzing Skewed Continuous Outcomes With Many Zeros: A Tutorial for Social Work and Youth Prevention Science Researchers.” *Journal of the Society for Social Work and Research* 499-797.

Willmott, Cort, Kenji Matsuura , e Scott Robeson. 2009. “Ambiguities inherent in sums-of-squares-based error statistics.” *Atmospheric Environment* 749–752.

Yates, Luke, Zach Aandahl, Shane Richards, e Barry Brook. 2022. *Cross validation for model selection: a review with examples from ecology.* Ecological Society of America.

Appendix

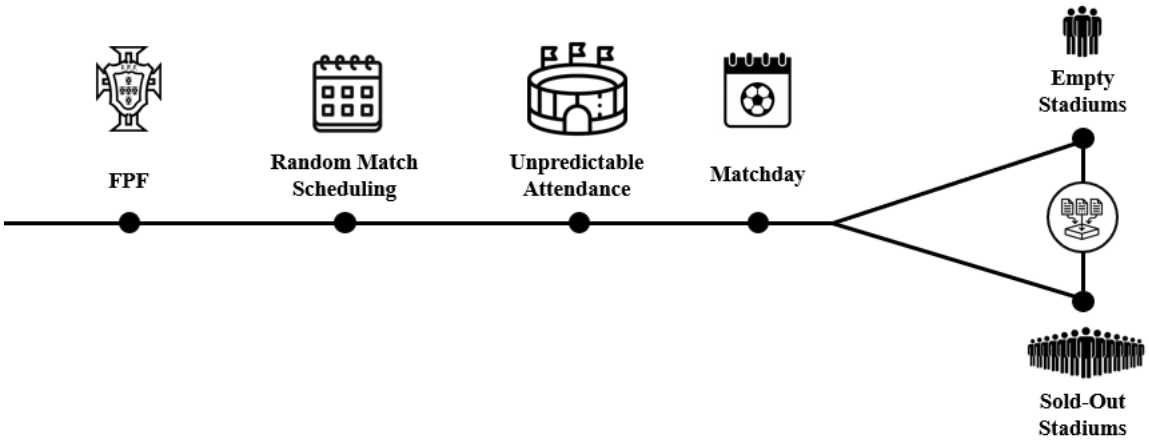


Figure 1 - Data collection process before project

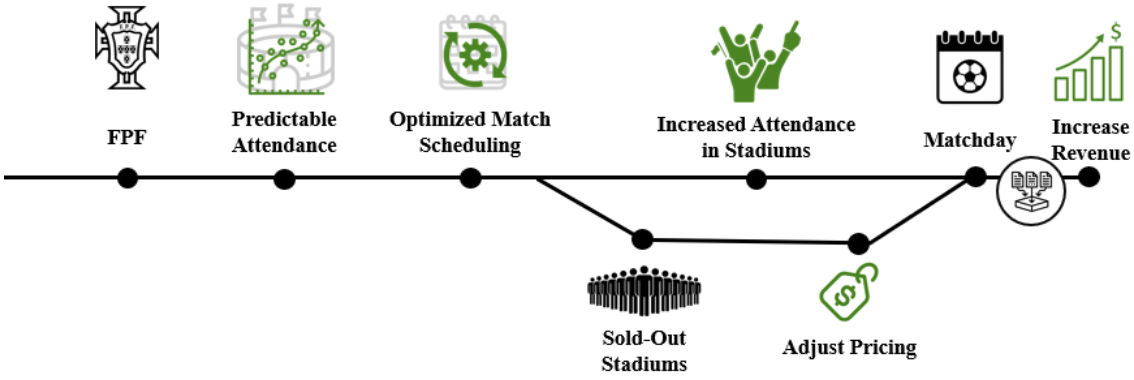


Figure 2 - Data collection process after project

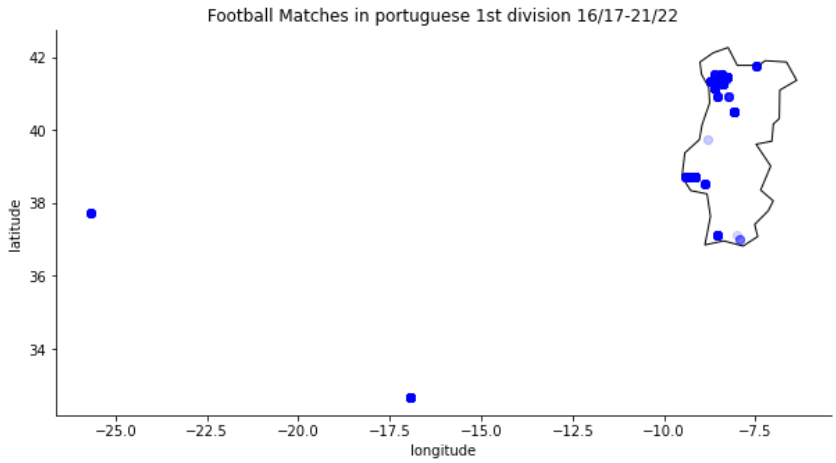
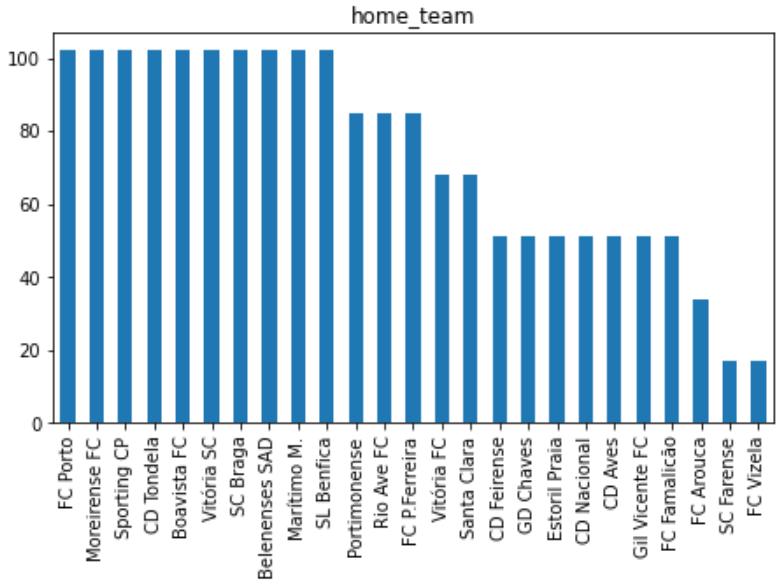
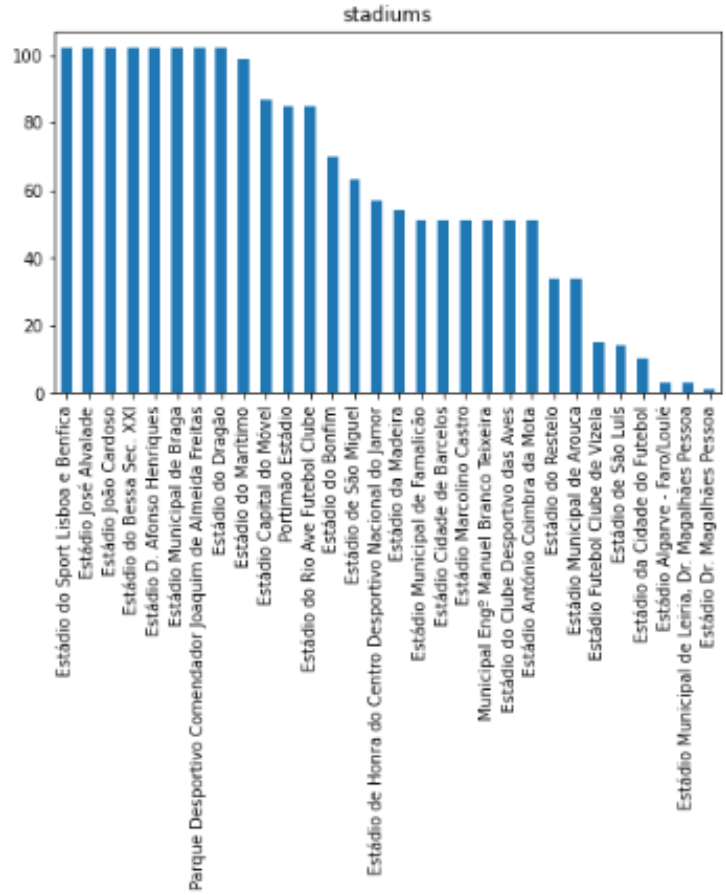


Figure 4 - Geographic representation of matches played



Graph 1 - Home teams' appearances count

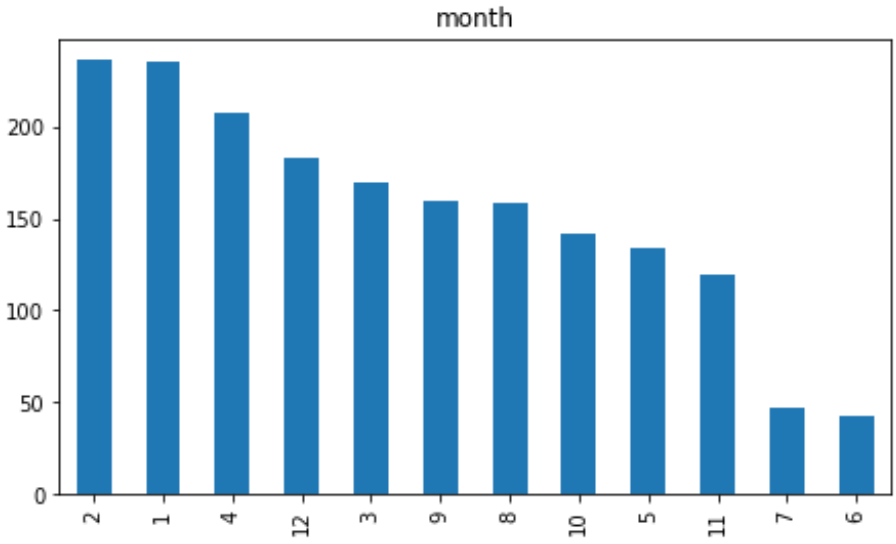


Graph 2 - Stadium's observations count

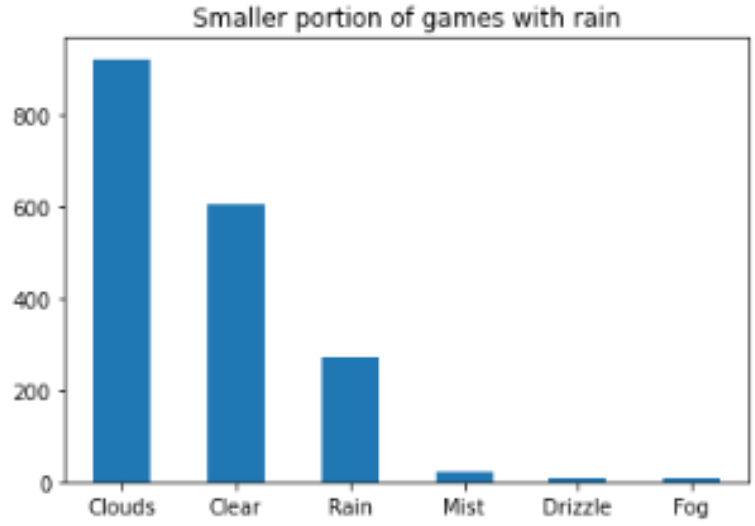
Column	*Data Type*	*Source*
season	object	Proprietary
round	float64	Proprietary
round_order	float64	Proprietary
home_team	object	Proprietary
away_team	object	Proprietary
datetime_left	datetime64[ns]	Proprietary
hour	object	Proprietary
attendance	float64	Proprietary
occupation_rate	float64	Derived
stadium_name	object	Proprietary
streaming	object	Proprietary
div_presences	category	ZeroZero
market_value_home_team	float64	TransferMarket
prev_classification_hometeam	float64	TransferMarket
prev_classification_awayteam	float64	TransferMarket
city	object	Proprietary
stadium_lat	object	Google Maps
stadium_lon	object	Google Maps
stadium_capacity	int64	Wikipedia
dist_to_center_km	float64	Google Maps
dist_to_center_min	float64	Google Maps
average_income	float64	Pordata
level	int64	Liga Portugal
cumulative_results	float64	Derived
last_game_home	float64	Derived
rivals	int64	Quora
avg_home_goal	float64	Derived
avg_home_goal_conceded	float64	Derived
avg_home_shots	float64	Derived
avg_away_shots	float64	Derived
avg_home_ball	float64	Derived
avg_home_fouls	float64	Derived
dist_between_teams	float64	Derived
month	int64	Derived
year	int64	Derived
day	int64	Derived
weekday	int64	Derived
hour_unit	int64	Derived
date	object	Proprietary
timezone	int64	Proprietary
temp	float64	Proprietary
dew_point	float64	Proprietary

feels_like	float64	Proprietary
temp_min	float64	Proprietary
temp_max	float64	Proprietary
pressure	int64	Proprietary
humidity	int64	Proprietary
wind_speed	float64	Proprietary
wind_deg	int64	Proprietary
wind_gust	float64	Proprietary
rain_1h	float64	Proprietary
rain_3h	float64	Proprietary
clouds_all	int64	Proprietary
weather_main	object	Proprietary
weather_description	object	Proprietary
covid	int64	Liga Portugal

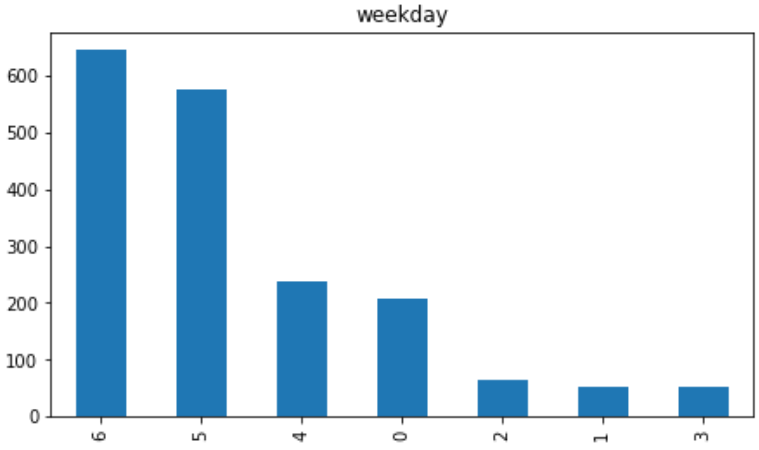
Table 5 - Final variables description



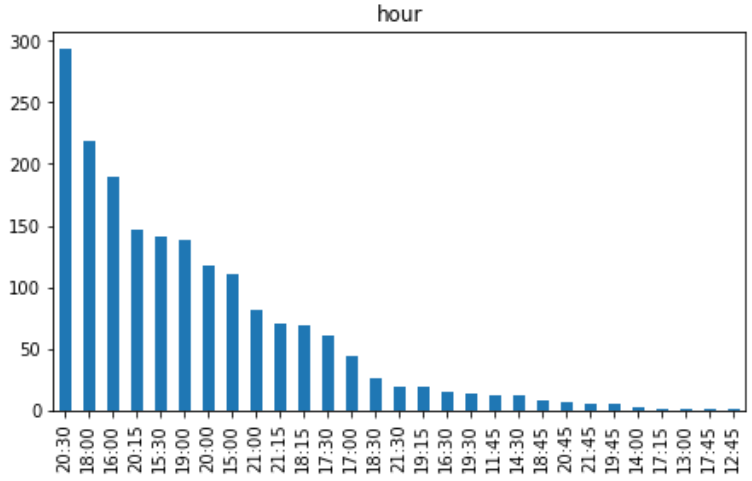
Graph 3 - Count on matches occurred per month



Graph 4 - Count on weather conditions



Graph 5 - Count on weekdays



Graph 6 - Count on matches' hours

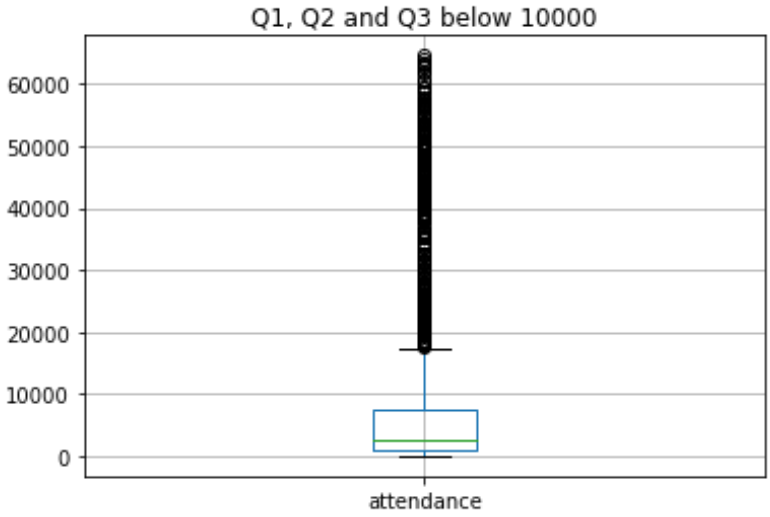


Figure 5 - Attendance distribution

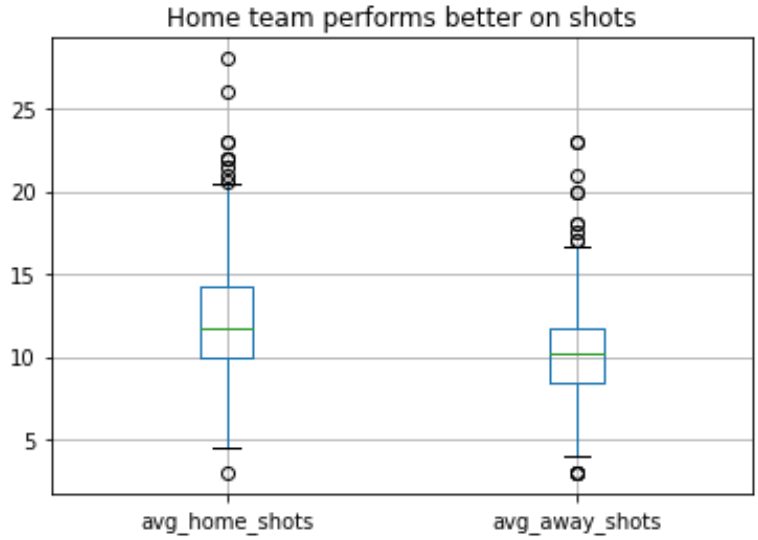


Figure 6 - Difference of playing at home on average shots

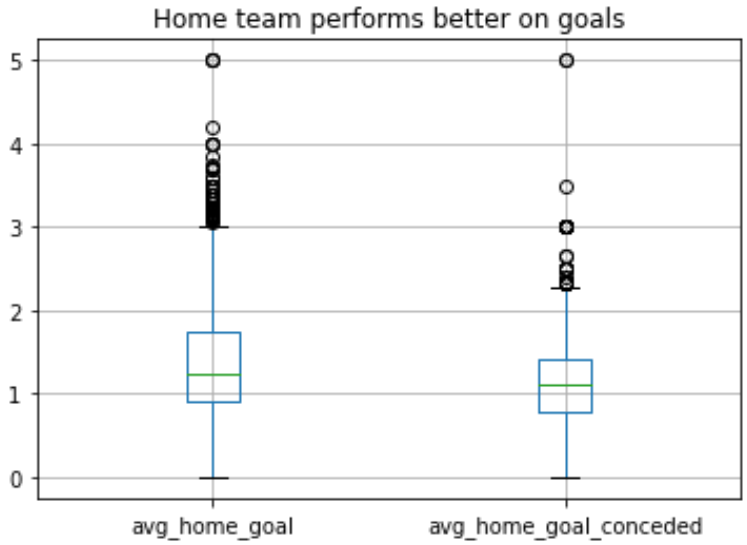


Figure 7 - Difference of playing at home on average goals

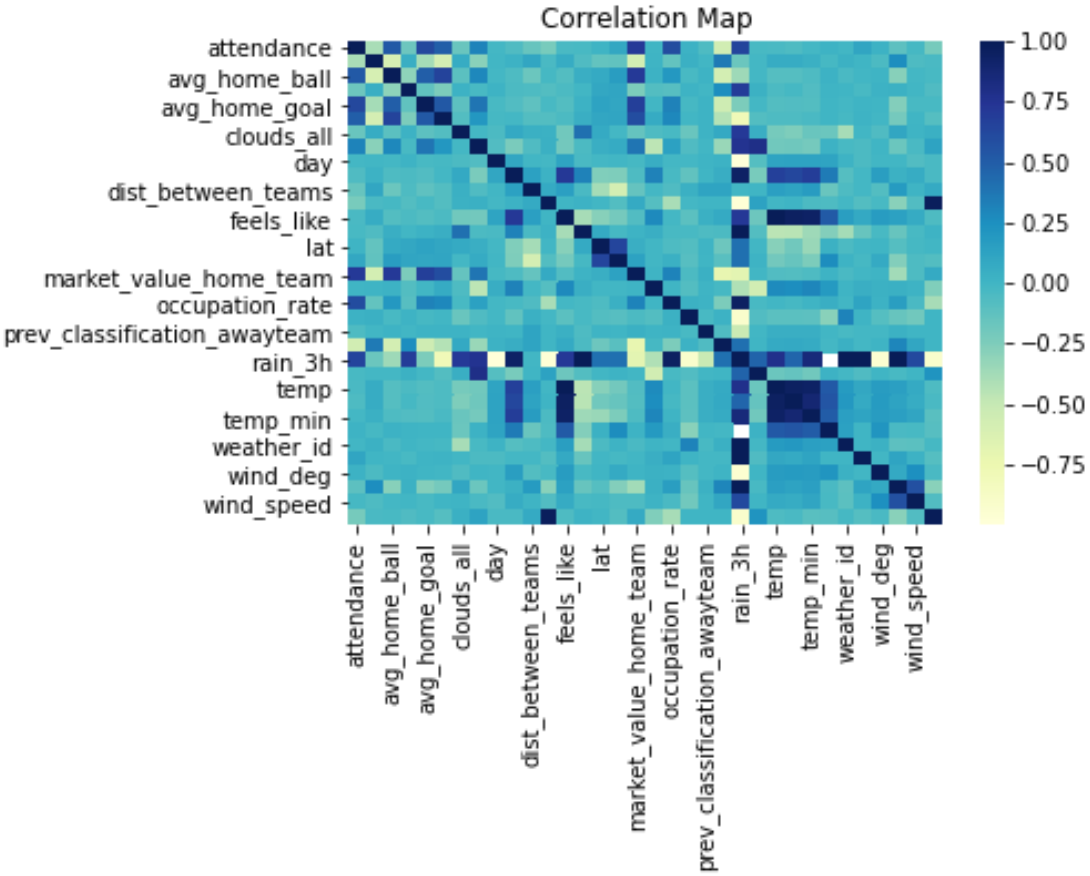
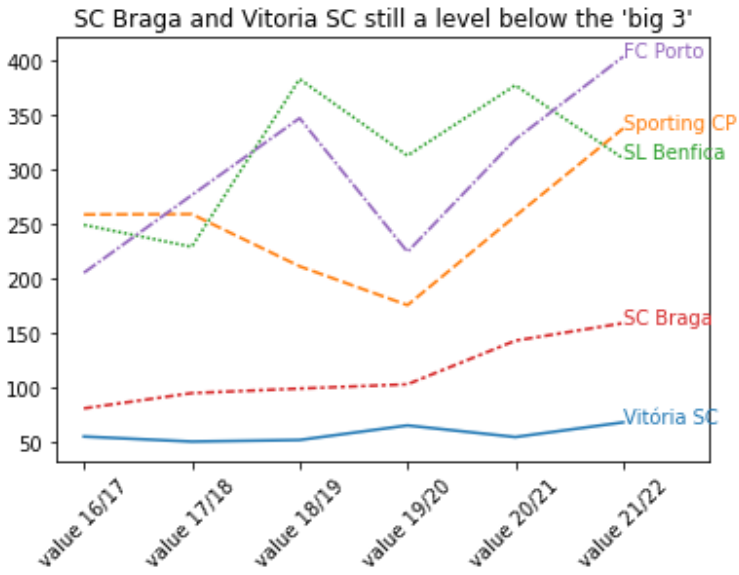


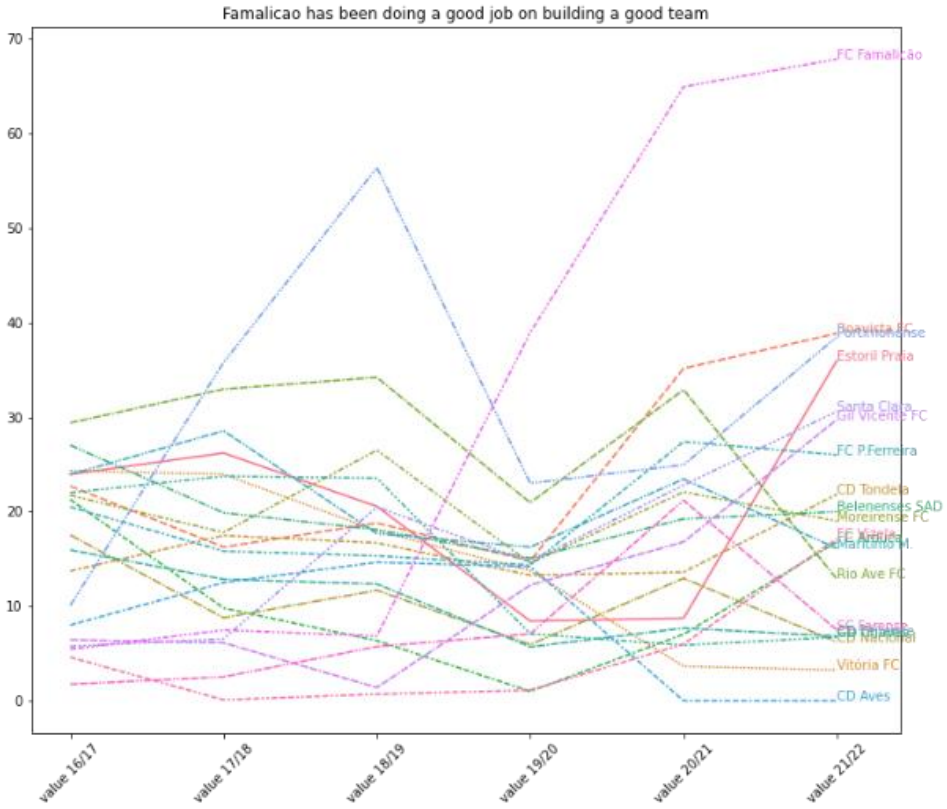
Figure 8 - Correlation map between numeric variables

Variable	*Correlation*
stadium_capacity	0.730102
market_value_home_team	0.712447
avg_home_goal	0.626643
occupation_rate	0.609666
avg_home_ball	0.531727
avg_home_shots	0.495899
dist_to_center_min	0.468223
last_game_home	0.317971
cumulative_results	0.315598
rivals	0.273811
dist_to_center_km	0.265262
hour_unit	0.168185
weekday	0.111624
lon	0.101165
rain_3h	0.03825
wind_deg	0.02718
pressure	0.027129
timezone	0.021772
lat	0.016708
month	0.007069
weather_id	-0.004007
temp_min	-0.008698
day	-0.010645
round_order	-0.01705
wind_speed	-0.018776
round	-0.023644
rain_1h	-0.024738
humidity	-0.028867
temp	-0.029232
feels_like	-0.032528
prev_classification_awayteam	-0.045626
temp_max	-0.051727
dew_point	-0.055708
wind_gust	-0.096533
dist_between_teams	-0.148984
clouds_all	-0.153781
avg_home_fouls	-0.176117
year	-0.213126
dt	-0.218634
avg_home_goal_conceded	-0.371406
avg_away_shots	-0.390322
level	-0.477597
prev_classification_hometeam	-0.563755

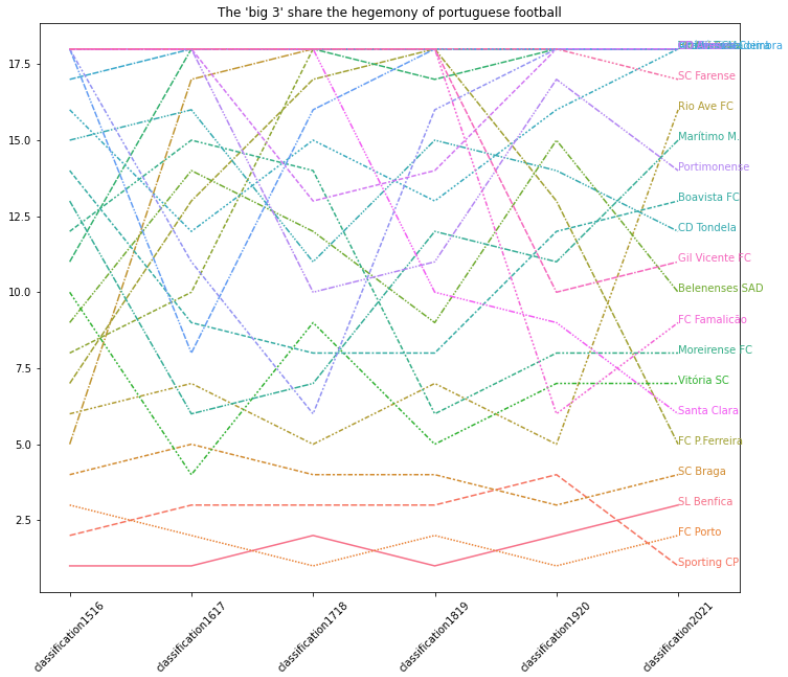
Table 6 - Correlation values on attendance



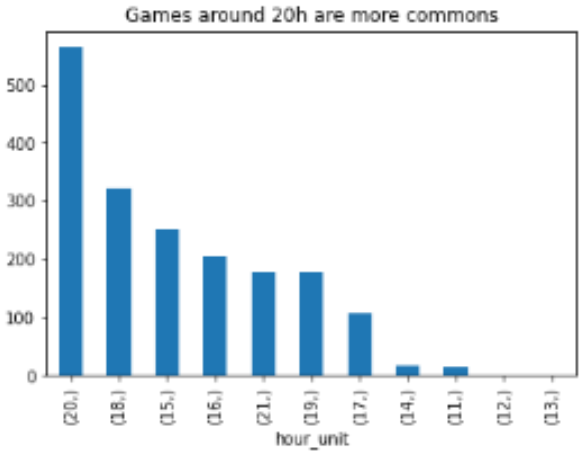
Graph 9 - Market value evolution



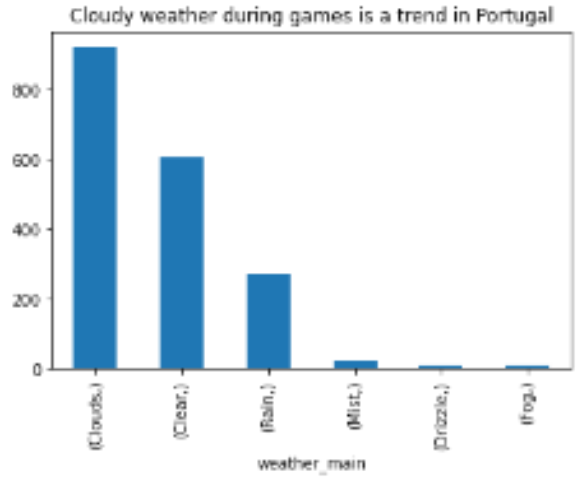
Graph 10 - Market value evolution



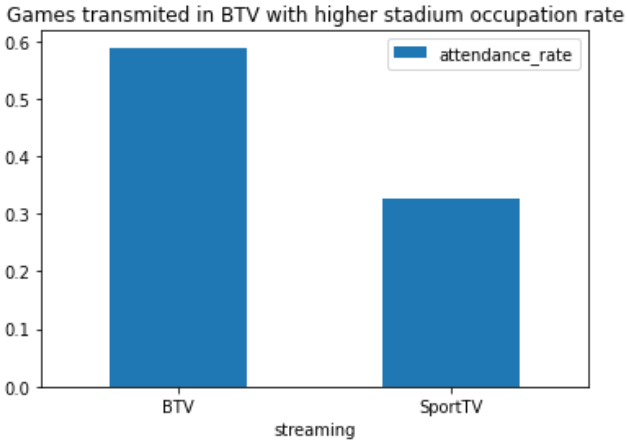
Graph 11 - Evolution on classification during seasons



Graph 15 - Matches played by hour



Graph 17 - Matches played per weather condition status



Graph 18 - Occupation rate per broadcasting channel

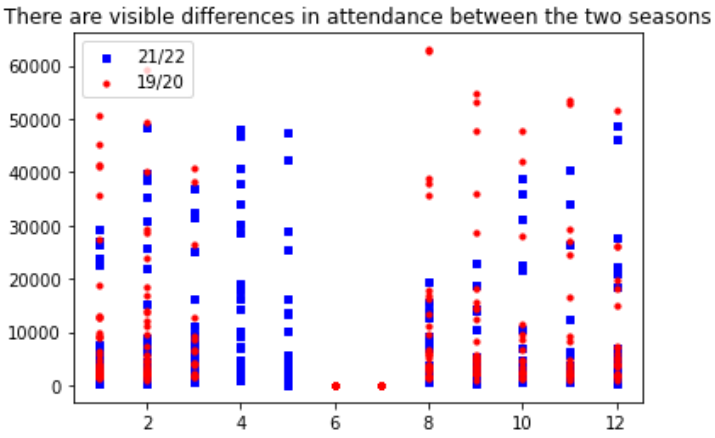


Figure 9 - Comparison between 2019/20 and 2021/22 seasons in attendance

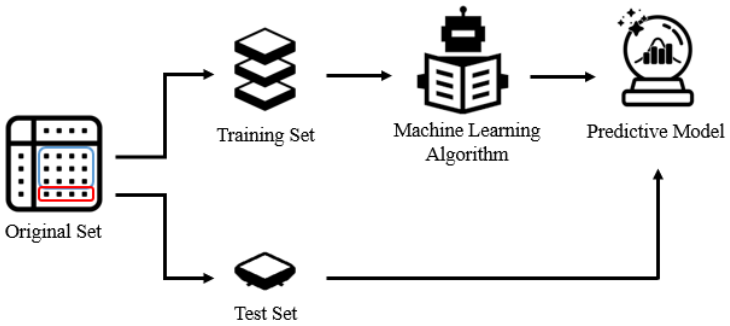


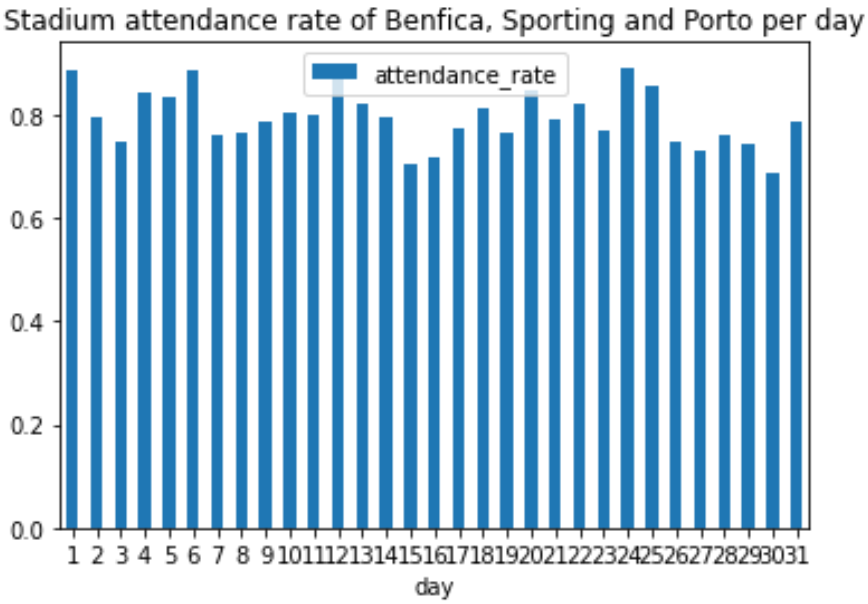
Figure 10 - Data partition illustration

Categorical	Numeric
season	*round*
home_team	*round_order*
away_team	*occupation_rate*
datetime_left	*market_value_home_team*
hour	*prev_classification_hometeam*
stadium_name	*prev_classification_awayteam*
streaming	*stadium_capacity*
div_presences	*dist_to_center_km*
city	*dist_to_center_min*
stadium_lat	*average_income*
stadium_lon	*level*
date	*cumulative_results*
weather_main	*last_game_home*
weather_description	*rivals*
	avg_home_goal
	avg_home_goal_conceded
	avg_home_shots
	avg_away_shots
	avg_home_ball
	avg_home_fouls
	dist_between_teams
	month
	year
	day
	weekday
	hour_unit
	timezone
	temp
	dew_point
	feels_like
	temp_min
	temp_max
	pressure
	humidity
	wind_speed
	wind_deg
	wind_gust
	rain_1h
	rain_3h
	clouds_all
	covid

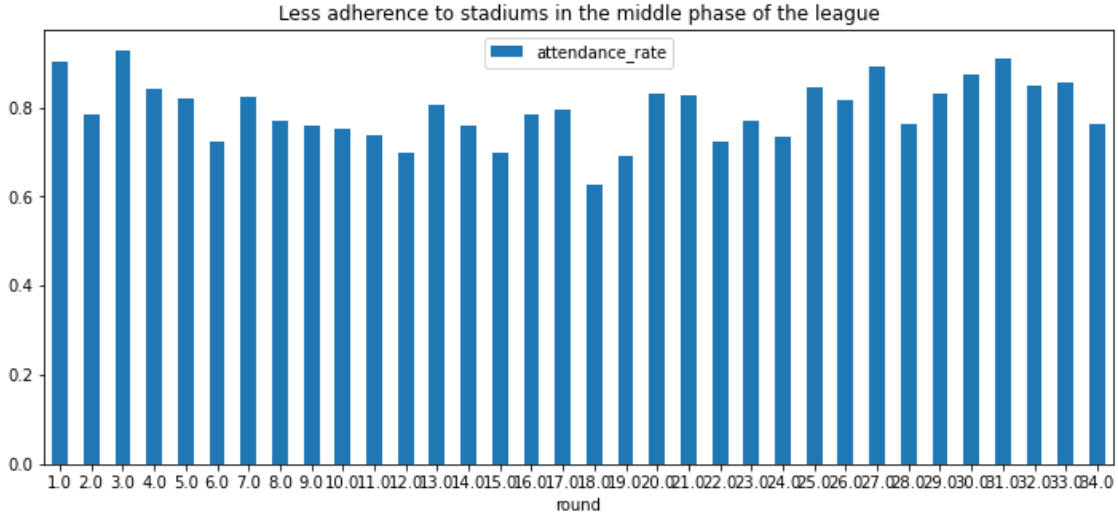
Table 8 - Categorical vs numerical variables

Team	*MAE*	*Attendance*	*%Error*
FC Famalicão	2362	2611	0.904758
FC Vizela	992	1317	0.752961
Estoril Praia	796	1281	0.621602
Moreirense FC	888	1663	0.533792
FC Arouca	821	1658	0.495363
Belenenses SAD	1287	2707	0.475537
Gil Vicente FC	1849	4865	0.380006
Boavista FC	2514	6830	0.368121
CD Tondela	834	2278	0.366055
Portimonense	760	2348	0.323812
CD Nacional	862	2815	0.306117
GD Chaves	1048	4297	0.243911
CD Aves	626	2674	0.234234
Marítimo M.	1289	5517	0.233668
Vitória FC	897	3915	0.229089
Rio Ave FC	886	3942	0.224761
CD Feirense	617	2968	0.208014
SC Braga	2128	10503	0.202593
Santa Clara	470	2395	0.196465
FC P.Ferreira	777	4146	0.187361
FC Porto	6083	37557	0.161968
Vitória SC	2610	17341	0.150541
Sporting CP	4600	34242	0.134343
SL Benfica	5205	48679	0.106924

Table 10 - Error per team



Graph 22 - Attendance rate of the "big3" per day



Graph 23 - stadium occupation per round (big3)

MODELS, PARAMETERS, VALUES AND THEIR DESCRIPTION

Ridge Regression Parameter Grid		
Parameter	**Description**	**Values**
solver	Tries to find the balance between parameter weights that minimizes the loss function	['svd', 'cholesky', 'lsqr', 'sag']
alpha	Constant that controls the regularization strength. Should be non-negative.	[1e-2, 1e-1, 1, 10, 100]
fit_intercept	Tells if we want to add the intercept for the model If set to false, no intercept will be used in the calculations	[True, False]
normalize	If 'fit intercept' is set to False, this parameter is ignored. If True, the variables used to predict a response value will be normalized by subtracting the mean and then dividing it by the L2 norm	[True, False]

Lasso Regression Parameter Grid		
Parameter	**Description**	**Values**
alpha	Constant that multiplies the L2 term, controlling regularization strength. Alpha must be a non-negative float	[1e-2, 1e-1, 1, 10, 100]
fit_intercept	Tells if we want to add the intercept for the model If set to false, no intercept will be used in the calculations	[True, False]
normalize	When set to false, this parameter is ignored. If set to True, the regressors X will be normalized before regression by subtracting the mean and dividing by the l2-norm.	[True, False]

Gradient Boosting Parameter Grid		
Parameter	**Description**	**Values**
n_estimators	Number of iterations performed	[750,1000,1250]
learning_rate	Sets the pace at which the algorithm updates the predictive values	[.001,0.01,.1]
max_depth	Value that limits the number of nodes in the tree	[1,2,4]
subsample	A fraction of samples used to fit individual models	[.5,.75,1]

Extreme Gradient-Boosting Machine Parameter Grid		
Parameter	**Description**	**Values**
max_depth	Value that limits the number of nodes in the tree	[3,6,10]
learning_rate	Sets the pace at which the algorithm updates the predictive values	[0.01, 0.05, 0.1]
n_estimators	Number of iterations performed	[250, 500, 750]
colsample_bytree	Subsample ratio of columns for each level	[0.3, 0.7]

Light Gradient-Boosting Machine Parameter Grid		
Parameter	**Description**	**Values**
num_leaves	Maximum number of leaves per tree. Higher this number, higher the overfitting possibility	(30,45, 60)
feature_fraction	Selects the only features in a dataset to be considered	(0.1,0.5, 0.9)
bagging_fraction	Controls the size of the data	(0.8, 1)
max_depth	Limits a tree depth	(9, 13),
min_split_gain	Minimum loss reduction needed to make a division on a leaf node of the tree	(0.001,0.01, 0.1)
min_child_weight	Minimum data in a leaf or child	(30, 50)

Random Forest Parameter Grid		
Parameter	**Description**	**Values**
Bootstrap	Tells the model whether to use bootstrap samples or not.	[True, False]
max_depth	Maximum depth of a tree	[25, 50, 75, None]
max_features	Specifies the number of features to use when searching for the best split	['auto', 'sqrt']
min_samples_leaf	Minimum of samples required to be at a leaf node	[1, 2, 3]
min_samples_split	Minimum number of samples needed to split an internal node	[2, 5, 10]
n_estimators	Number of trees in a random forest	[750,1000,1250]

Support Vector Regression Parameter Grid		
Parameter	**Description**	**Values**
kernel	Selects the kernel type to be used	['rbf', 'sigmoid']
gamma	Kernel coefficient	['auto','scale']
C	A regularization parameter that must be positive. Also adds an L2 squared penalty to each misclassified data point	[1,5,10,50,100]
degree	Polynomial kernel function degree	[3,5,7]
coef0	Independent term in a kernel function	[0.01,0.1,0.5,1,10]