Predictive models of the performance of professional football players

Mafalda Teixeira Costa da Conceição

Dissertation written under the supervision of Viktor Pekar

Dissertation submitted in partial fulfilment of requirements for the International MSc in Management with major in Strategic Marketing at Universidade Católica Portuguesa and for the MSc in Business Analytics at Aston Business School, September 2021.

Acknowledgments

I would like to express my deep gratitude to Mr. Viktor Pekar, my supervisor, who provided me with extremely insightful advice and consistently support in all my doubts. I would also like to thank all my professors in the MSc Business Analytics who taught me valuable skills, knowledge and methods that I used throughout my research and will use in my future career. A special thanks also to Procter & Gamble Portugal which gave me the flexibility to develop my thesis while working full-time. Lastly, but not least, I would like to thank all my friends and family for the endless support and motivation during both my Master's degree and the dissertation.

Table of Contents

List of Tables

List of Figures

Abstract

The aim of this present study was to predict professional player performance, based on a set of features, including team-related ones and its effect on performance. The predictions were made for two player roles: attackers and midfielders and two distinct independent variables were used: goals total and goals assists. The dataset used corresponded to 4523 players from season 2018-2019, from ten different top European leagues. Some individual performance features were used, like passes accuracy, shots on, duels won were used, as well as some player features like age, height and weight, some club performance features like club market value and club goals total and even popularity features like google search and twitter average likes. The team-related features were calculated by taking the average of a variable for the whole team, excluding the player itself. The results showed that *goals_assists_team*, *goals_total_midfielder_team* and *market_value_opponents* were found to be the most important variables and statistically significant (p-value < 0.05) when predicting goals total.

At the same time, *goals_assists_team*, *passes_accuracy_midfielder_team*, *duels_won_defender_team* and *market_value_opponents* were the most important team-related variables when predicting goals assists and they were all statistically significant (p-value < 0.05). Stochastic Gradient Descent Regressor was the most suitable Machine Learning (ML) model to predict goals total, with RMSE of 1.3543, whereas the Ridge Regression achieved RMSE of 1.054 to predict goals assists. Clubs and players should be aware of these team factors that affect goals and assists, to increase knowledge about the best player-team fit and therefore, improve performance.

Keywords: Feature importance; Feature selection; Feature significance; Hyperparameter tuning; Performance prediction (goals total and goals assists); Supervised Machine Learning models; Team-related variables

1. Introduction

1.1. Motivation

Football is one of the most popular sports in the world, and as the years pass more information has become available from different countries and leagues which rose the interest of data analysts for this field. "Sports Analytics" comprises the "segment of data collection, management, predictive modelling and computational methods" with the objective to find useful information to facilitate decision making (Sarlis & Tjortjis, 2020). One of the most interesting subjects within sports analytics is the use of player and game data to make predictions about player performance and composition of teams. This is particularly interesting as there are large monetary amounts involved in scouting and club managers strive to understand and formulate winning teams.

Some applications of analytics in football industry are: match strategy analysis (McCabe & Trevathan, 2008), identify player's styles (Anthony A, Jayalakshmi, & D, 2021), injury prediction (Hughes, Sergeant, van der Windt, Riley, & Callaghan, 2018), performance prediction (Pantzalis & Tjortjis, 2020), match outcome (Hucaljuk & Rakipović, 2011) and market value prediction (Stanojevic & Gyarmati, 2016), among many others.

Further, the interest of the public in advanced data analytics and how they can serve football has rose, putting more pressure in this area (Pantzalis & Tjortjis, 2020).

Football is a very competitive sport with thousands of professional players in the ecosystem. Due to this highly competitive market and the fact that careers in football are considerably short, the focus of players is on performance maximization. On one hand football players want to find a club where their skills can be elevated and with good compensation, while clubs want to scout the players that would make the team shine, delivering the best results. Specialists have agreed that having the highest performing players does not necessarily mean the best team, as there could exist some conflicts between teammates due to lack of team fit (Dizdari & Seiler, 2020). Nowadays, experts, like scouts, use their knowledge to ensure these matches, however, this is highly challenging so the use of data analytics in this field can open many doors (Stanojevic & Gyarmati, 2016).

Another important point is that market value is not the best measure of a player performance, nor should be the baseline to look into for recruitment purposes, as market value is "influenced by player's age, competitiveness of the league, club budget, how famous a player is or even release clauses of the players" (Pariath, Shah, Surve, & Mittal, 2018).

1.2. Aim and Research Questions

This dissertation main aim is to study the effect of team-related variables on the prediction of an individual football player's performance. With this aim in mind, three research questions were formulated:

1. What are the most suitable learning algorithms for this problem?
2. What are the most suitable variables for this prediction problem?
3. What is the contribution of team-related variables to the performance of the prediction models?

To answer these questions, some specific objectives were defined:

- Based on a review of previous research, identify and evaluate a number of regression algorithms that have the ability to predict the performance of football players given the dataset available for this thesis
- Implement several feature selection methods to choose the features that have the possibility to exert higher influence on the target variable
- Compare the performance of the ML models to find the best-performing model on an automatically selected subset of features
- Study the importance of the features to understand which ones are predictive of a football player's performance
- Study the contribution of team variables to the quality of the model

Therefore, with this thesis, the performance of professional football players was predicted using ML models with not only individual characteristics, club characteristics and popularity as independent features, but also team-related variables. These last ones were created specifically for this problem, by taking the averages of existing variables in the dataset. Performance is therefore the dependent variable in this problem, and two variables were chosen as performance indicators: goals total and goals assists. Also relevant to mention is the fact that this work directs its attention to attackers and midfielders only, as there was no concrete measure of performance of defenders and goalkeepers constituted only a small portion of the players, so not enough data was gathered from players in this position.

By focusing on the effect of team-related variables, our study can potentially address a number of important questions about previous unstudied factors affecting an individual player's performance. For example, given that the goalkeeper has a high save score, or the defenders have high tackling scores or duels won, how will these affect the goals and assists of a given

attacker/midfielder? Will the age of other teammates affect the goals or assists of a given attacker/midfielder in the same team?

This thesis is of high relevance, as by studying the effect of the performance of other players in the same team, clubs can understand how team statistics impact the individual player performance. More generally, this project could be relevant to studies on organizational psychology, on how to screen candidates for their ability to fit well with the team. In fact, Bowen, Ledford and Nathan studied a new hiring model based on person-organization fit, rather than on the conventional selecting process based on person fit for a specific job description (Bowen , Ledford Jr., & Nathan, 1996).

When clubs are choosing how their teams should be constituted, some doubts come into play. Are heterogeneity teams more productive? When it comes to production activities, heterogeneous teams benefit from mutual learning, meaning, more skilful members teach less skilful how to execute tasks better and more quickly thereby enhancing team productivity (Hamilton, Nickerson, & Owan, 2003). But is this the case in football?

Ultimately, the emergence of new ML techniques has enabled the development of better predictive models for performance and this thesis proposes a new approach to individual performance prediction, one more focused on identifying the key characteristics to achieve a good team fit.

1.3. Key directions of prior research

There is a considerable number of papers that spent a lot of effort in studying whether salary is correlated with performance prediction. Stanojevic and Gyarmati (2016) found that the Gradient Boosting Regressor was the best algorithm to predict market value, whereas Muller, Simons and Weimann (2017) used Multilevel Regression analysis. Also, predicting match results has also been a topic of utmost interest, with McCabe and Trevathan (2008) using Multi-Layer Perceptrons for these predictions. Pantzalis and Tjortjis (2020) studied the characteristics that make defenders top players and Pariath, Shah, Surve and Mittal (2018) also developed a model to predict player overall performance. Finally, Torgler and Schmidt (2011) introduced team-related variables as an indicator to predict football player performance, which was defined as goals and assists of a player. However, to the best of knowledge, there was not a lot of previous studies that aimed to model the team fit of a player and analyse its effect on the player's performance.

1.4. Contributions

From past literature review, the area of sports analytics lacked studies of how performance of professional athletes in football is affected by players of the same team, so this is the main novelty of this work. This approach tackles on the human psychology in football, by coming up with a methodology to study the effect of team-mates on individual professional players. At the same time, characterizing performance as goals total and goals assists is something not widely seen in literature.

The feature selection methods and ML models used are the ones mostly seen in other literatures, however the approach chosen is different from the ones taken by past papers. As in this case, different subsets, each one with a distinctive set of variables (obtained from a set of feature selection algorithms) were fed into a serious of models and the best combinations of feature selection and ML algorithms were chosen.

1.5. Thesis structure

This dissertation is organized as follows: 2. Literature review-gives a background of past studies into the topic of sports predictions; 3. Methodology-describes the dataset, pre-processing steps followed, the overall approach of this dissertation and an overview of feature selection methods and ML models used; 4,5,6 and 7. Results and Discussions-details the experiments conducted, the results obtained, and the discussion of these results; 8 and 9. Final remarks and Conclusion-details the closing remarks, future work and limitations.

2. Literature review

The first part of the Literature review is related to the performance of teams in general and what has been found to be the factors leading to higher levels of performance. Following this, papers about predicting performance of teams in football will be discussed, before shifting the attention to past papers that aimed to predict market value, match results and which players should a team be constituted by. Lastly, the focus will be on the prediction of performance of individual players, first in other sports and then in football.

2.1. Performance of teams

What principles should one follow to put a team of professionals together? How can team composition affect performance? Should teams be more homogeneous?

Hamilton, Nickerson and Owan (2003) explore these questions in detail, being the empirical context of this study "weekly production reports from a Koret Corporation garment manufacturing facility in California" (Hamilton, Nickerson, & Owan, 2003).

The final results of this study indicated that how the team is composed of, has a large influencer over productivity. More heterogeneous teams are more productive, thus having a mix of high and low ability workers is better than a set of workers with identical technical abilities.

Further, it was found that the stronger members have a higher impact on productivity than the weaker ones. When replacing one member by a more productive one, productivity increases more if the most-able member is replaced by someone better rather than replacing the least-able member for a better one. This poses an intriguing conclusion that can be applied to a soccer team, when deciding which players to sign and which ones they should sell/loan to another club. Team composition is key, as discussed by Hamilton, Nickerson and Owan (2003), team production benefits from both technical and collaborative ability whereas individual production benefits only from technical skills.

Further, Cooman, Matthijs Bal, Lub and Vantilborgh (2015), found evidence, among 121 participants of 30 teams, that the teams that shared high perceptions of supplementary and complementary fit, in general had a superior performance compared to those that did not (Cooman, Matthijs Bal, Lub, & Vantilborgh, 2015). According to the optimal distinctiveness theory (Brewer, 1991), humans have not only the need for belonging thus seeking similarities among team members, but they also have the need for uniqueness, seeking therefore complementary.

2.2. Predict Performance of teams in football

As mentioned by Myatt and Wallace (2008), the addition of an extra "bad apple" player can "spoil the barrel" by destabilizing successful teams (Myatt & Wallace, 2008). However, the recruitment departments of football companies pay little emphasis on team balance and chemistry when scouting for potential signings, whereas the greater focus is on individual qualities and performance. Bransen and Haaren (2020) took the first step in the topic of "How well does a team of soccer players get together?" (Bransen & Haaren, 2020), by predicting chemistry between players that have never played together. This is of high relevance for scouting purposes, as teams need to decide who to sign to better fit with the players that are already in the team (Bransen & Haaren, 2020). This paper made use of match event data (on-the-ball action from each player at each match) from Wyscout website with a selection of 106 domestic and international competitions since the start of the 2015/2016 season. According to

Bransen and Haaren (2020), mutual chemistry is contemplated in player's performance. The assumption made in this paper is that when two players perform at a high level, they have high mutual chemistry, whereas when two players have low mutual chemistry, they perform at a lower level maintaining all other things equal. This paper introduced two novel metrics to measure the joint performance of a pair of players: Joint Offensive Impact (JOI) and Joint Defensive Impact (JDI). The former is about joint actions to increase the likelihood of scoring a goal, and the latter joint actions that decrease the likelihood of conceding a goal.

Further, two ML models were pursued in an attempt to predict the JOI and JDI for players that never played together. The training of the models was done using CatBoost gradient boosting toolkit, the tuning of hyperparameters was done on the validation set, and the appropriate features were selected by optimizing the Root Mean Square Error (RMSE) on the validation set. Both models outperformed the baseline, with RMSE of 0.04464 and 0.88906 for JOI and JDI respectively. When the authors looked for the importance of the features in the ML models, it was found that the scores for a number of player roles have a significant impact on predictions. In addition, the "number of matches that two players had played together ahead of the season has a considerable impact on the predictions of both metrics for a pair of players" (Bransen & Haaren, 2020). Features like if the pair had the same nationality, originated from the same region, or has same mother tongue has a significant weak predictive power in both models, thus a pair of players having complementary playing styles contributes more to the chemistry that similar cultural backgrounds.

Further, Lombardi (2017) focused his attention in the MLS (Major League Soccer), which is the primary league in the United States and North America (Lombardi, 2017). The aim of this paper was the development of a player performance index to "classify the players in the MLS according to their contributions to team performance" (Lombardi, 2017). After, with this index, it was identified how efficient teams were. The EA Sports Player Performance Index was used as base and the statistics from each player position were placed in a production function to determine team efficiency. Determining team performance only based on goals is gather unrealistic, since each game the goals scored is relatively low. As such, looking at shots, shots on net and key passes makes absolute sense as they are actions that led to the goals itself. Thus, the player performance index models goals scored as the outcome of the "number of shots and shot effectiveness" (Lombardi, 2017). It was found that teams that diversified their talent amongst three positions, forwards, defenders and midfielders, were the most efficient. Interestingly this study shed light on the fact that midfielders and defenders contribute the most

to a team's ability to get a shot, which can lead to a goal, when compared to forwards. With this in mind, the target values used by this dissertation were also predicted for midfielders.

2.3. Predicting Market value using player's Performance

Several studies have been conducted to predict player's market value based on performance indicators.

Stanojevic and Gyarmati (2016) derived a method to "assess the player's market value using player's performance data" (Stanojevic & Gyarmati, 2016). The performance data used was obtain from InStat, whether the market values for each player were obtained through transfermarkt.com. Both datasets were intersected, and the data used was composed of 12858 records from the 2014/15 season. This data was divided into four folds, in which three were used for training and the rest for testing (4- fold cross validation). Some performance features used were minutes played per game, assists, passes, shots, dribbles, shots accuracy, passes accuracy, among others. Adding up to this, player features like age, height and position were also used and one key point of this paper is that team features were also included. Average co-player Transfermarkt Market Value Estimates (TMVE) was also derived, which captures "the average TMVE of all the other players from the same team" (Stanojevic & Gyarmati, 2016) and represents how strong the team is. The opponent TMVE was also calculated representing the strength of the opposing team and finally the average points per game was derived to represent the successfulness of the team. Some of these team variables were also derived for this dissertation.

TMVE were used as labels whereas the Performance-Driven Market Value Estimates (PDMVE) were what the study wanted to predict. Different supervised algorithms were use like Random Forest, Gradient Boosting Tree Regression (GBTR) and generalized Linear models. From all of these, the GBTR was the one achieving the lowest error. Grid-search was used to select the most appropriate hyperparameters and only 10% of the training data was used as a validation dataset for the hyperparameter selection. The mean and median differences between TMVE and PDMVE were around 60% and 34% respectively. Interestingly this data-driven approach allocated higher PDMVE to less established players who in the next season (2015/16) saw their TMVE almost double. Thus, this indicates that this analysis was able to pick up this increase by looking at 2014/15 data alone. As such, clubs can make use of this analysis by identifying those players whose PDMVE "outweighs their TMVE" (Stanojevic & Gyarmati, 2016) and pursue them at a more competitive cost.

Muller, Simons and Weimann (2017) also estimated market value using Multilevel Regression analysis (Müller, Simons, & Weinmann, 2017). They divided the predictors of market value in the following categories: player characteristics (age, height, position etc), player performance (goals, assists, playing time, dribbling, among others) and player popularity (news and internet links). The authors collected data from the top five European Leagues and excluded goalkeepers. After obtaining the market value predictions, the differences between these estimates and the actual transfer fees were calculated and the RMSE ended up 3.4% higher versus the RMSE obtained with crowd estimates for market value (transfermarkt). Thus, it can be concluded that this data-driven approach to estimate market value produced very close estimates to the crowd sourcing ones, however, applying significantly less effort.

2.4. Predicting match results

Another topic that has been continuously getting attention in sports analytics is the prediction of outcomes of sports contests. However, predicting these outcomes is a difficult problem due to the number of factors which must be considered that cannot be quantitatively valued or modelled (Hucaljuk & Rakipović, 2011) like the weather conditions or even the psychological impact of fans in the stands.

McCabe and Trevathan (2008) made use of Multi-Layer Perceptrons (MLP), a type of neural network model, to predict the results of sports matches given some basic information (McCabe & Trevathan, 2008). In fact, two learning algorithms were used, back-propagation and the conjugate- gradient method. This paper used data from four league sports: the Australian National Rugby League (NRL), the Australian Football League (AFL), Super Rugby (Super 12 and Super 14) and English Premier League football (EPL) (McCabe & Trevathan, 2008). The following features were the ones collected for each team for a specific competition: Points-for, points-against, overall performance, home and away performance, performance in previous game, performance in previous n games, team ranking, location, player availability, points-for in previous n games and points-against in previous n games. The MLP calculated an "output value for each team based on that team's feature set" (McCabe & Trevathan, 2008). When this output value took a closer value to one, it indicated that there is high confidence that the team was going to outperform rivals and win the next match. When the value was close to zero, the confidence was at a lower level. After two teams face each other, the output value was calculated and the team with the highest value (there was bigger confidence on that team win) was predicted as the winner. The best result was in the Australian National Rugby League with a 58% success, compared to success rates of about 60,65% achieved by human experts. Thus,

predicting probabilistic events, given the broad range of elements that may influence the results, was concluded to be extremely challenging.

2.5. Selection of best players for a match

Shahriar, Islam and Amin (2019) proposed a classification technique to select the best eleven players automatically within a soccer team for a match (Shahriar, Islam, & Amin, 2019). The attributes used to classify a player were matches played, goals scored, assists, yellow card, red card, minutes played, goals conceded and clean sheets. The data was collected from www.footystats.org and players were categorized by their positions: forward, midfielder, defender and goalkeeper. Further, Naïve Bayes Classifier, Decision Tree, Support Vector Machine (SVM), KNN Classifier and Random Forest Classifier were the supervised learning algorithms used, in which the Decision Tree was found to be the best performing one with a RMSE of 0.41. This paper exhibits another great advancement in sports analytics, by providing "an effective, non-biased and scientific approach" (Shahriar, Islam, & Amin, 2019) for coaches to choose the players with best performances.

2.6. Player Performance Prediction in Cricket

Passi and Pandey (2018) focused on predicting the number of runs scored by each batsman and number of wickets taken by each bowler based on their characteristics and statistics from the website cricinfo (Passi & Pandey, 2018). These predictors were classified into ranges, and thus four classifiers were used: Naïve Bayes, Random Forest, Multiclass SVM and Decision Tree. The size of the training and test set differed, from 60% training data and 40% test to 90% train and 10% test. Overall, the Random Forest achieved the highest accuracy for both runs and wickets, one of 90.74% and 92.25% respectively. In general, for all the classifiers the accuracy increased when the size of training data was larger, so the best results were achieved with 90% train and 10% test. One interesting part of this paper was that since the different measures of performance like batting average, strike rate etc had more importance than others, the authors weighted each feature "according to relative importance over other measures" (Passi & Pandey, 2018). To operationalize this, Analytic Hierarchy Process - AHP (Saaty, 1987) was the tool used which allowed for pairwise comparisons.

Further, Anik, Yeaser, Hossain and Chakrabarty (2018) used Linear Regression and Support Vector Machine to predict the runs scored by a batsman and by a bowler. These authors used Recursive Feature Elimination and Univariate Selection as feature selection methods to find out

the five most important features, which have greater impact on the outcome. This study demonstrated that the Linear Regression model achieved a 91.5% accuracy for batsman Tamim, while the Support Vector Machine achieved one of 75.3% for bowler Mahmudullah (Anik, Yeaser, Hossain, & Chakrabarty, 2018).

2.7. Player Performance Prediction in Football

Pantzalis and Tjortjis (2020), focused their scope of attention in "long-term team and player performance prediction" (Pantzalis & Tjortjis, 2020). This paper used team statistical features from the season that ended (e.g. wins, shots, possession percentage etc) and historical data (e.g. performance indicators from previous seasons) as well as some financial data like transfer fees and team salaries. Furthermore, this paper had a deeper focus on evaluating defender's performance as opposed to attackers, as the latter are usually graded higher, even if they are not as essential in the team strategy. The data used had teams from four European football national leagues from 2015 to 2018 and had more than 40 features.

Two experiments were conducted, in which the first was focused on team performance prediction. In the first part the aim was to predict whether a team was going to achieve a better season, meaning, gather more points or not, homologous to the previous year (this was the target attribute). This experiment was handled as a binary classification problem and after the deployment of several classifiers, Random Forest was the one achieving highest accuracy (number of teams with correct performance prediction divided by number of total teams), one of more than 70%.

The second experiment was around player performance prediction, more precisely the characteristics and statistics in comparison with their season rating, that make a central defender, a top-class player. The data had only 59 central defender players, playing in the English Premier League. Player attributes, playing positions and some demographic features were collected. Every numeric value of the data was normalized to range between zero and one. The model chosen was the Multiple Regression one with player rating (calculated by domain experts) as the dependent variable. Interceptions followed by team overall rating were found to be the most important characteristic. Jumping reach, versatility, acceleration and first touch on the ball were discovered to be the player's best attributes. The second approach was splitting the features in:

1. Player's characteristics and attributes.
2. Player statistics.
3. Team statistics.

Three multiple regression models were built with backward elimination. For the first model only seven features were proven as influential. For the second model twelve features were included and the adjusted R squared suffered an improvement versus the first one. The third model, with team statistics did not produce an appropriate model with only team rating as being significant. As such this variable was included in the second model which produced a slightly higher adjusted R squared.

To sum up, some of the following variables were found to be the most important in predicting the performance of central defenders: interceptions, aerials won, tackles, jumping reach, age, passing, strength, minutes played, fouls, key passes, goals, team's rating, among others. Interestingly, we can see that attacking skills (e.g. key passes, goals) are amongst these variables and thus have an impact on central defenders rating, so we can see changes in playing habits of these defenders.

On the topic of player performance prediction, Pariath, Shah, Surve and Mittal (2018) developed a model to estimate a player overall performance based on skills and characteristics of that player (Pariath, Shah, Surve, & Mittal, 2018). The authors also decided to categorize the players into their respective positions (goalkeepers, midfielders, defenders and attackers), thus creating separate models for each one of these positions, with attackers having features like shot power and finishing affecting their performance while midfielders have acceleration, crossing, among others. 36 player attributes were extracted from sofifa website. However, the most relevant attributes when it comes to player performance were identified as being dribbling, shot power, attacking, finishing, head accuracy, acceleration, crossing, skill, curve and ball control. A Linear Regression algorithm was used, achieving an accuracy of 84.34% and RMSE of 2.67 for the midfielder dataset, with the other categories achieving similar accuracy. Such a simple model could have extremely high benefits in football recruitment by allowing scouts to identify high performance and talented players who are probably not as well-known as some of their peers.

Apostolou and Tjortjis (2019) conducted experiments in an attempt to predict player's position, number of goals scored and number of shots attempted by a player (Apostolou & Tjortjis, 2019). Regarding player's position, a Random Forest classifier was used to classify players into four roles: forwards, goalkeepers, midfielders and defenders and the accuracy achieved was one of 81.5%. The second and third experiment used Messi and Suarez players, to predict the number of goals they would score in season 2017/18 and the number of shots they would attempt in a match. The dataset had data scraped from whoscored.com up to season 2016/17 and Random Forest was again the algorithm of election providing the most accurate results.

2.8. Performance prediction in soccer using team variables

Torgler and Schmidt (2011) studied several factors that could affect player performance in soccer, like income level, relative income (versus the other team players) and team effects (Schmidt & Torgler, 2011). The novelty presented here, was that the scope was not in the prediction of salary, like many other studies, but rather on the determinants of individual player performance, which is a rather scarce topic among previous studies. The data used was from the Bundesliga, the top German soccer league, with approximately 2833 observations. Whereas performance data was collected by the firm IMP, the authors used the market value from kicker sportmagazin, a prominent soccer magazine, as it was proven to be highly correlated with salaries retrieved from transfermarkt, thus representing a good proxy for player salary.

The authors chose to center their studies in three research questions. The first extended Simmons and Forrest (2004) studies and focused on understanding whether the non-linear effect between wage and team performance, would remain when studying the individual level (Forrest, Simmons, & Szymanski , 2004). Thus, authors Torgler and Schmidt included a squared term of player's salary, apart from the salary from previous season (Schmidt & Torgler, 2011).

The second research question studied was based on the conjecture that humans compare themselves to others, and soccer players are no different. As such, "player's individual salary relative to that of other team-mates may impact their performance" (Schmidt & Torgler, 2011). To study this, the authors measured "relative income position in terms of the difference between team-mates' average salaries and players' individual salaries" (Schmidt & Torgler, 2011). Due to this study, a team variable to represent market value of the other teammates of a player, was also created for this dissertation.

The third and last research question was focused on the impact of team affects. Idson and Kahane (2000) and Torgler (2007) measured team effect on an individual i by "calculating the average values for the team-mates (excluding the player i)" (Schmidt & Torgler, 2011). This way, one can study if team-mates average age, goals, salary, etc affect players individual performance. This approach of calculation represented a good benchmark for how team variables could be created, on the existing dataset of this dissertation.

One interesting aspect that these authors pursued, was the choice of not only one, but two dependent performance variables: goals and assists. Thus, it was possible to study how team effects work in different indicators. This approach, served also as basis for this dissertation. One variable also used, was if a player had change teams or not, which may indicate that a player joined another club to win more money, which would lead to higher motivation and thus

performance. However, changes can bear the burden of adaption, which may affect performance negatively. To account for the fact that the relationship between player's age and performance might be nonlinear, authors used age and age squared as variables, thus allowing to model the effect of different ages rather than assuming the effect is linear for all ages. Further, the player's position was controlled for (defence, midfield and attack whereas goalkeepers were neglected), to account for heterogeneity.

Two different empirical methodologies, within multiple regression, were presented: the pooling and fixed effect estimations. To sum up, salary was found to have a statistically impact and nonlinear effect on individual performance, however, the direction of this relationship was still unclear. A strong impact was observed for the relative income position and when the individual salary is controlled for, "an even larger difference in the average team-mates' salary leads to" a lower individual performance (Schmidt & Torgler, 2011). Team effects were also found to play a significant part in individual performance, specifically exchanges (number of times a player is brought to the pitch in exchange for another player) and sending offs (number of times player leaves the game as punishment). Keeping all the variables the same, players on a team with more exchanges perform better, while sending offs negatively affects performance. One finding that is worth pointing out, was that there was a tendency for team effects to be more important for goals rather than assists. Age and its square were only statistically significant in some regressions, more precisely in ones focusing on goals. It was found that until 31 years performance rises with age but at a decreasing rate, and beyond this point, the deterioration of the physical condition outweighs the greater experience and knowledge, and the ability to score goals decreases. Changing teams was found to be a variable with no impact on performance, whether the position where the player plays, had indeed an impact. It was also showed that when it comes to goals, attackers have the best chance, whereas on assists, both midfielders and attackers have equal chance of performing better.

Below, a summary of previous findings for the papers discussed throughout the Literature review can be found.

Authors	Title	Dataset	Dependent variable	Independent variables	Model/Method
Bransen and Haaren (2020)	Player Chemistry: Striving for a Perfectly Balanced Soccer Team	Wyscout website	Predict the JOI and JDI for players that never played together	Time in the match, start and end location, result (e.g., successful, or unsuccessful), and the body part that the player used to perform the action, for each on-the-ball action, age, height, weight, nationality, mother tongue, region, preferred foot, position, etc	CatBoost gradient boosting toolkit
Nick Lombardi (2017)	Evaluating Player Performance and Team Efficiency in Major League Soccer	WhoScored website	Development of a player performance index	**Player features:** Passes, crosses, dribbles, clearances, losses of possession due to a tackle or interception, red and yellow cards. **Club Features:** team's goals for, goals against, shots, shots on target, opponent blocked shots, passes, crosses, dribbles, opponent interceptions, opponent tackles, opponent clearances, opponent red cards, opponent yellow cards and opponents' fouls.	Index is formulated using the main component of the EA Sports PPI developed by McHale et al. (2012).
Rade Stanojevic and Laszlo Gyarmati (2016)	Towards data-driven football player assessment	InStat and transfermarkt websites	Assessing player's market value using player's performance data	**Performance features:** games, minutes played per game, goals total, assists, accurate passes, accurate challenges, accurate key passes, accurate shots, accurate dribbles, accurate tackles, accurate crosses, assists accuracy, passes accuracy, challenges accuracy, goals per minute, assists per minute, passes per minute **Player info features:** age, position, height **Team info features:** average co-players TMVE, average opponent TMVE, average points per game, etc	Random Forests, Gradient Boosting, Trees Regression and generalized Linear models
Oliver Müller, Alexander Simons, Markus Weinmann (2017)	Beyond crowd judgments: Data-driven estimation of market value in association football	Internet sources like Google, Reddit, Transfermarkt, WhoScored, Wikipedia, and YouTube	Estimating player's market value	**Player characteristics:** age, height, position, nationality, etc **Player performance:** goals, assists, playing time, dribbling, etc **Player popularity**: news and internet links	Multilevel Regression
Josip Hucaljuk and Alen Rakipović (2011)	Predicting football scores using machine learning techniques	Various websites	Predicting the results of football matches	Current form of teams in the last six games, outcome of the previous meeting of the teams that play the game, current position in the rankings, number of injured players from the first team, average number of goals scored/conceded per game.	Naive Bayes, Bayesian networks, LogitBoost, KNN, Random forest, Artificial neural networks.
Alan McCabe and Jarrod Trevathan (2008)	Artificial Intelligence in Sports	Several sources	Predicting of sports outcomes	Points-for, points-against, overall performance, home and away performance, performance in previous game, performance in previous n games, team ranking, location, player availability, points-for in previous n games and points-against in previous n games.	Multi-Layer Perceptrons
Md. Tanzil Shahriar, Yashna Islam, Md. Nur Amin (2019)	Player Classification Technique Based on Performance for a Soccer Team Using Machine Learning Algorithms	Footystats website	Selecting the best players automatically to form a soccer team	Matches played, goals, assists, yellow cards, red cards, minutes played, goal conceded and clean sheet	Naïve Bayes Classifier, Decision Tree, Support Vector Machine, KNN Classifier and Random Forest Classifier
Kalpdrum Passi and Niravkumar Pandey (2018)	Increased prediction accuracy in the game of cricket using Machine Learning	Scraped from cricinfo website	Predicting the number of runs scored by each batsman and number of wickets taken by each bowler	No. of innings, batting average, strike rate, centuries, zeros, highest score, overs, bowling average, bowling strike rate and four/five wicket haul	Naïve Bayes, Random Forest, Multiclass SVM and Decision Tree
Anik, Yeaser, Hossain, & Chakrabarty (2018)	Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms	Howstat and espncrickinfo websites	Predicting the runs of a batsman and runs of a bowler	Number of balls played, player position, opposite, ground (home or away), pitch (condition) and overs	Linear Regression and Support Vector Machine
Victor Chazan-Pantzalis and Christos Tjortjis (2020)	Sports Analytics for Football League Table and Player Performance Prediction	Scrapped from internet and from expired online competition	1. Team Performance Prediction (achieve more points than last year); 2. Player performance prediction (rating as dependent variable)	**1.** Team performance indicators from previous season, team statistical features (possession percentage, shots, etc) and data not measurable by team performance (financial, etc) **2.** Interceptions, jumping reach, versatility, acceleration, first touch on the ball, clearances, aerials won, tackles, age, passing, vision, strength, minutes played, fouls, inaccurate short passes, key passes, goals, team's rating	1. Naïve Bayes, Decision Tree, Random forest, KNN, SVM and XGBost 2. Multiple Linear Regression
Richard Pariath, Shailin Shah, Aditya Surve, Jayashri Mittal (2018)	Player Performance Prediction in Football Game	Scraped from sofifa website	Predict player's overall performance and market value	Dribbling, shot power, attacking, finishing, head accuracy, acceleration, crossing, skill, curve and ball control	Linear regression
Konstantinos Apostolou and Christos Tjortjis (2019)	Sports Analytics algorithms for performance prediction	Sofifa, whoscored and understat websites	Predict player's position, number of goals scored and number of shots attempted by a player	Player's appearances, minutes played for a whole season, number of goals scored, assists, completed passes, red and yellow cards, etc	Random Forest, Logistic Regression, MLP classifier and Linear SVC
Benno Torgler and Sascha L. Schmidt (2011)	What shapes player performance in soccer? Empirical findings from a panel analysis	Firm IMP and Kicker Sportmagazin	Prediction of performance (goals and assists)	Market value, squared market value, relative market value (difference between team-mates' average values and players' individual values), player's age, squared age, new to a team, player's position, average age value of team-mates, average number of exchanges of team-mates and average number of sending-offs of team-mates	Multiple Regression

3. Methodology

3.1. Dataset

The dataset[1] used included information of 4523 players from season 2018-2019, from ten different leagues- Eredivisie (Netherlands), Premier League (United Kingdom), Ligue 1 (France), Bundesliga (Germany), Serie A (Italy), Primeira Liga (Portugal), Premier Liga (Russia), La Liga (Spain), First Division A (Belgium) and Super Lig (Turkey). The dataset is composed of 89 variables, which can be grouped as:

- **Individual performance features:** games minutes played, goals assists, goals total, dribbles, passes total, passes accuracy, shots on, duels won, goals conceded, among others.
- **Player info features:** nationality, age, dominant foot, weight, height, position and rating.
- **Club performance features:** club goals conceded, club market value, club goals total, club match won, club match lose, among others.
- **Popularity features:** google search, twitter average likes, Wikipedia page views in English, YouTube average likes, among others.

One of the variables-the market value-was obtained from transfermarkt[2], however these values were from the end of the season 2018-2019, and as such, to study how market value affects a player's performance, market values at the end of season 2017-2018 or at the beginning of 2018-2019 needed to be obtained. Using the data from a GitHub repository[3], the transfer fees from the season 2017-2018 were retrieved. Muller, Simons and Weimann (2017) stated that "Market values can be understood as estimates of transfer fees - that is, prices that could be paid for a player on the football market" (Müller, Simons, & Weinmann, 2017), so in this study some transfer fees from the season of 2017-2018 and 2016-2017, were used to replace the original market values of the season 2018-2019.

The first issue encountered, was that two leagues-First Division A and Super Lig were missing from the GitHub repository, so the transfer fees of the past season could not be obtained. In order to avoid losing information and data, instead of deleting records from these leagues, the original market values of the season 2018-2019 were maintained, and it was assumed that these

[1] Suprihanta C. (2020). Player Characteristics, Player Performance, and Player Popularity in Football Player Market Valuation. MSc dissertation. Aston University.

[2] https://www.transfermarkt.com/

[3] https://github.com/ewenme/transfers

values were close to the ones from the past season. For all the other leagues, the replacing process was the following:

- If a specific player present in the Suprihanta (2020) dataset was found on the GitHub of season 2017-2018 and his transfer fee was different than zero, then his market value was replaced with the market value of season 2017-2018 (found in GitHub folder 2017)
- If this was not the case, the player was searched on the GitHub of season 2016-2017, and if he was found and his transfer fee was not zero, the player market value was replaced with the one from season 2016-2017 (GitHub folder 2016).
- Finally, if a player had transfer fee of zero in both of these seasons or if it was absent from these lists (did not change clubs in these seasons), the original market value in the dataset, one of the season 2018-2019, was kept.

Transfer fees of zero could not be kept, as if a player's contract runs out, his transfer fee will be zero, however, that does not mean the player's market value is zero.

To obtain the GitHub transfer fees, the match between the original dataset and the GitHub ones was based on players' names and, in case the name was the same, based on players' clubs and ages. Another issue to keep in mind was that the GitHub had players' names without diacritics (accented, letters, etc), so to mitigate this issue, accents were removed from the players 'name on the original dataset as well. All in all, from the original 4523 records, 788 players suffered a change in their original market values.

3.2. Train-Test Split

The train test split in this study was not done using stratified sampling, as not only every league needed to be present in both train and test, but also, every player of the same club needed to be either on the train or on the test. This is what this study wanted to simulate, a possible real-world use case, where a model is trained on some teams and then applied to some completely different teams. If the split was done not guaranteeing that all the players of a team were either on the test or on the train, team variables would be wrongly calculated. This would be the case as players of the same club will be divided, so when calculating the average of a feature of a specific team in the train set, some players will be present in the test set (and vice versa), and as such the averages will not account all the teammates. Hence, the split was made guaranteeing that not only every league would be present in both train and test, but also, all the players of the same team were put together in the train or test set. After this split, from the 4523 records, 78% were kept on the train (3511) and 22% on the test (1012). Regarding player position (defender,

midfielder, attacker and goalkeeper), market value and popularity (Wikipedia page views) the distributions of these variables were similar both in the train and test set.

3.3. Data Pre-Processing

Since the dataset had originally a considerable high number of variables, the ones less likely to hold any influence on player performance were deleted. These included variables related to reddit platform and Wikipedia statistics in languages different from English.

This dataset had also two variables somewhat related, games minutes played and games appearances, thus it was necessary to check if these values were in accordance. To validate this, it was checked if:

$$Total\ minutes\ played\ \leq game\ appearances\ * 120$$

In which 120 corresponds to the maximum duration of a game. This was the case for all the records, so no further action was taken.

Note that, on the further data pre-processing, some stages were implemented using scikit-learn library in Python.

3.3.1. Missing values

In the dataset five variables has missing values (Table 1):

Table 1. Variables with missing values

Column name	Num missing values
dominant_foot	32
google_search	187
height	37
rating	340
weight	190

Instead of deleting records or even columns, SimpleImputer was used, which is nothing more than an imputation transformer for completing missing values. Since dominant foot is a categorical variable, the missing values were imputed using the most frequent value. The other four numerical variables had their values imputed using the respective median.

3.3.2. Feature Engineering

The novelty of this work is the study of the performance of other players in the same team in the performance of each individual player. Hence, eleven team variables were created:

- Market value of teammates: players sometimes compare themselves to their peers, so it would be relevant to average market values of their peers and see how it affects player performance.
- Age of teammates: are teams with older players negatively influencing player performance?
- Teammate's performance variables: variables like goals assists of teammates, goal total of the attackers of the team, goals total of the midfielders of the team, passes accuracy of the midfielders of the team, tackles total of the defenders of the team, duels won by the defenders of the team, goals conceded by the goalkeepers of the team, were also computed. These correspond to important performance indicators of the players of a team, thus it would be interesting to analyse, for example, given that the goalkeeper has a low value of goals conceded, and the defenders have very high tackling and duels won scores, how will these affect the goals and assists of a given attacker?
- Market value of the opponents: it could be also relevant to study if high or low market values of opponent teams in the same league, can affect the motivation and therefore player's performance. Stanojevic and Gyarmati (2016) also used opponent market value to represent the strength of the opposing team. This was calculated by taking the average of market values for players in the same league as player i, excluding the player i and all his teammates.
- Diversity of the team: calculated by summing the unique nationalities in the team and dividing it by the number of players. The larger this value, the more diverse the team is in terms of nationalities of the players. It may be interesting to see if greater diversity improves or damages players' performance.

3.3.3. Dummy variables creation

Further, the categorical variables like dominant foot, league and position were transformed into dummy variables using OneHotEncoder, which encodes the categorical variables as a one-hot numeric array.

3.3.4. Log of Skewed predictors

Moreover, histograms for a set of variables were plotted and it was found that google search, club market value, market value and market value of the team, were heavily right skewed (Appendix 1). Just like Muller, Simons and Weimann (2017) did in their studies, the natural

log transformation function of NumPy was used to log transform the values of these variables, in an attempt to make those as normalized as possible.

3.3.5. Feature Scaling

Many ML algorithms do not perform well if the independent numerical variables do not have similar distributions. As such, standardization was applied to all the independent variables: each variable was transformed to have a zero mean and unit variance. The target values were kept unchanged to be easier to interpret during error analysis. The standardization:

$$Z = \frac{(x_i - \mu)}{\sigma}$$

was implemented in StandardScaler.

In these three important pre-processing steps - missing values, creation of dummy variables and feature scaling - the SimpleImputer, OneHotEncoder and StandardScaler classes were fitted on the training data and then used to perform the transformation both on the train and test data.

3.4. Approach

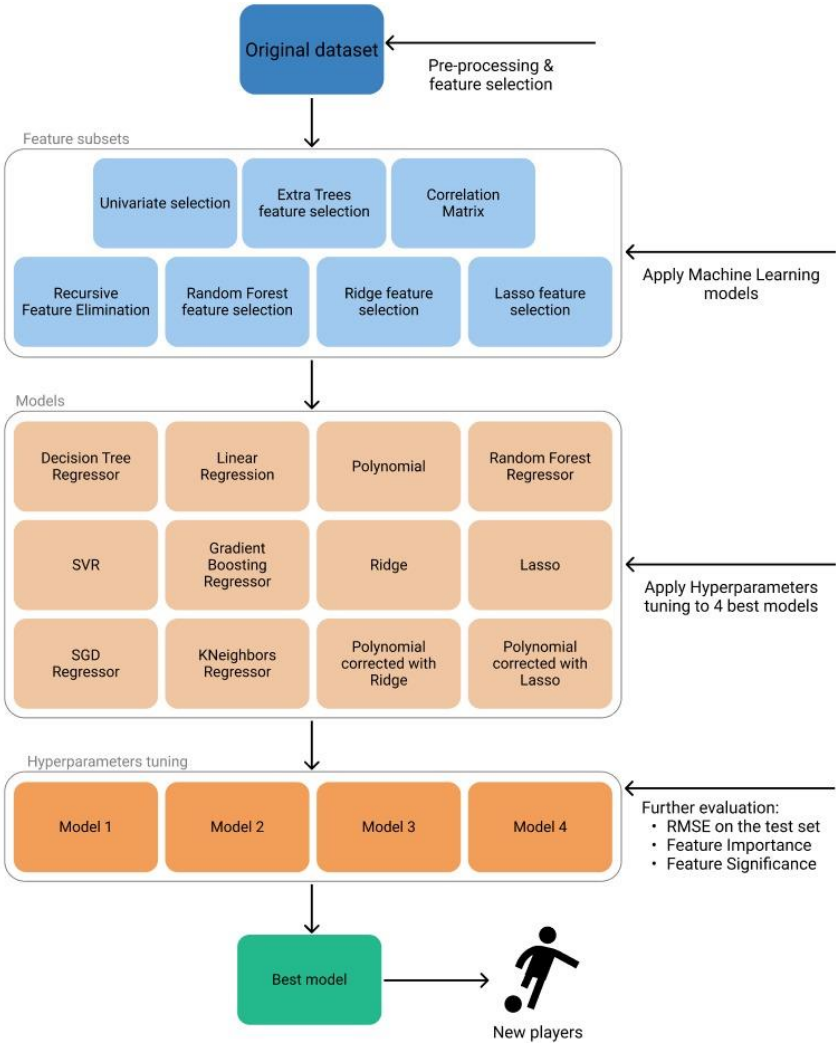

Figure 1. Diagram of Approach

Since the dataset used had a lot of variables and some of them might not be important to explain number of goals total nor number of goals assist of each player (chosen as performance indicators), a series of feature selection methods were used. The process went as follows (Figure 1):

- To the original dataset, seven feature selection techniques were applied, from which seven distinctive feature sets were created, which may or may not overlap on the selected features.
- These seven datasets, plus the original one with all the features and one with only the team variables, were fed into twelve different ML algorithms. This approach was inspired by the paper wrote by Spencer, Thabtah, Abdelhamid and Thompson (2020) that predicted heart diseases.

- To evaluate these twelve models in the training set, a 10-fold cross-validation was used, in which the training data was split into ten parts, the model was trained on nine parts and evaluated on the 10th part. This process was then repeated ten times and the RMSE reported was averaged over the ten runs. Cross-validation is an important practice as if predicted values are evaluated on the training set, even the best models can overfit. Hence, the goal of cross-validation is to test a model's ability to predict completely new data that was not used in estimating a model, to tackle problems like overfitting . It can also provide insights on how a model will generalize to an independent dataset (i.e., an unknown dataset).
- To use cross-validation, cross_val_score, a helper function was called on each estimator and dataset. The cv parameter, which determines the splitting strategy, was not pass as ten, an integer number, but ShuffleSplit was used (with n_splits equal to ten and test_size equal to 30%) to make sure the split is always the same between executions. This will also allow comparisons to be made before and after tuning the hyperparameters.
- Once the best combinations of dataset and model were derived, the top four models were chosen for tuning hyperparameters.
- To make easier the choice for the best hyperparameter combination in the space of hyperparameters, GridSearchCV was used. This is nothing more than an exhaustive search of the values of specified parameters of an estimator.
- The best models found (with the best combinations of values in their parameters) were then analysed and applied to the test set. After evaluating the error, feature importance and significance of each model, the most appropriate model was discovered and the essential conclusions were drawn from it.

3.5. Evaluation Metric

When deciding which models performed better, one evaluation metric was used, the RMSE. Its formula can be seen below:

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

This metric was the focus on this work for several reasons:

- Compared to MSE it presents lower values thus being easier to compare.

- Compared to MSE and MAE, RMSE is a square root, so it ends up having the same units as the dependent variables, thus making interpretations easier.

3.6. Feature Selection techniques

The quote "less is more" (Ludwig Mies van der Rohe) is also applicable in ML. When a dataset has a lot of features, not all of those are important, and keeping these unnecessary variables can lead to some handicaps when training a model. It can lead to reduced accuracy, greater complexity, longer time to execute and even reduce the generalization that the model is capable off, leading to overfitting. Hence, feature selection has three major objectives (Guyon & Elisseeff, 2003):

- Improving the prediction performance of the ML models.
- Leading to faster and more cost-effective ML models.
- Providing better knowledge of the "underlying process that generated the data" (Guyon & Elisseeff, 2003).

As such, feature selection is of utmost importance and in this work, three different techniques were used:

3.6.1. Filter Methods

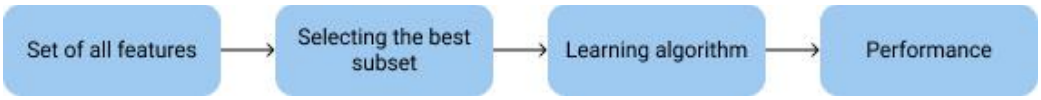

Figure 2. Filter Method Process

These correspond to methods in which the features from a dataset are selected regardless of the ML algorithms used (Figure 2). Usually, they are applied in the pre-processing stage before any model is trained and features are selected based on individual scores in various statistical tests for their correlation with the target variable. Compared to other methods, like wrapper, "filter methods are usually faster" (Guyon & Elisseeff, 2003) and since they can be used when pre-processing the data to "reduce space dimensionality and overcome overfitting" (Guyon & Elisseeff, 2003). However, this selection of features is individually, which can lead to the method missing features that might be useful, not on their own, but when combined with others. From this category, two different filter methods were used:

- Univariate filter selection (based on mutual info regression scoring function)
- Correlation's coefficients (based on Pearson's coefficient)

3.6.2. Wrapper Methods

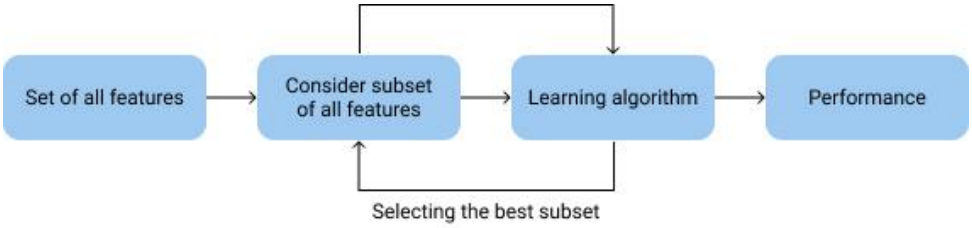

Figure 3. Wrapper Methods Process

Another name for these methods is greedy algorithms, as they train a model with a subset of features, draw conclusions from the previous model and then proceed to adding or removing features from the subset, thus making this an iteratively process (Figure 3). Since these methods provide the optimal set of features for training a model, their accuracy is usually superior compared to the filter methods. However, these methods are "often criticized requiring massive amounts of computation" (Guyon & Elisseeff, 2003).

From this category, Recursive Feature Elimination (with Logistic Regression estimator) was used, with the objective to find the best performing subset of features. The estimator is first trained with a set containing all the features, and their feature importance is retrieved. The ones with the least importance are removed, and this process is repeated until the subset is left with the required number of features.

3.6.3. Embedded Methods

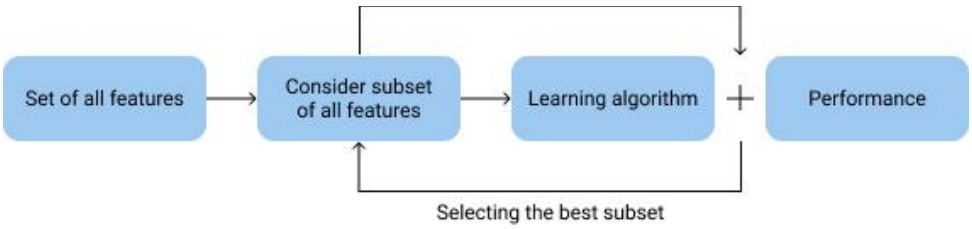

Figure 4. Embedded Methods Process

These are methods in which the algorithms have their own built-in feature selection method, thus "incorporating variable selection as part of the training process" (Guyon & Elisseeff, 2003) (Figure 4). Embedded methods may be more efficient as there is no need to "split the training data into training and validation set" (Guyon & Elisseeff, 2003), thus, they better exploit the data available. Moreover, they are faster "by avoiding re-training a model from scratch for every variable subset investigated" (Guyon & Elisseeff, 2003). In this work Ridge, Lasso and Tree-based algorithms (Random Forrest and Extra Trees Regressor) were the choices. For Ridge and Lasso coefficients were used as the metric for feature selection, while for the Tree-based algorithms, feature importance was used.

When applying all these methods and their algorithms described above, to this dataset, the base number of features selected was always 20. Further, the team variables previously created were always all included in the different datasets. This was the case regardless of the algorithms identifying them as being the most important features or not. This approach chosen goes in line with what this thesis focus heavily on: the influence of the team variables in player performance. Thus, to study this in more detail, all the team variables had to be included in the different datasets derived from all the feature selection methods.

3.7. Baseline Model

To understand the quality of the models produced, a baseline model was created to allow relative comparisons. In this case, for each test instance it was predicted the median target value observed on the training set. Hence, it can then be concluded that if the models produced, outperform this baseline one, they have learned something about this regression problem. If this is not the case, then the models produced might be too simplistic and suffer from underfitting.

3.8. Supervised Machine Learning Models

Like Figure 1 describes, an extensive number of models were applied to each dataset obtained from all the feature selection methods. Nonetheless, five models achieved the higher results and thus the attention will be focused on: Multiple Linear Regression model, Ridge Regression, Polynomial degree 2, Gradient Boosting Regressor and Stochastic Gradient Descent Regressor.

3.8.1. Multiple Linear Regression (MLR)

In the case of this work two Multiple Linear Regressions (prediction of goals total and goals assists) will be produced, as the aim is to "predict the result of an answer variable, using a number of explanatory variables" (Maulud & Abdulazeez, 2020). Being the independent variables x and the dependent variable y, the basic model for each MLR is:

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_m x_m + \varepsilon$$

Nevertheless, this model has a strong bias since if the independent variables and the dependent have a non-linear relationship, this model can underfit the data leading to incorrect predictions.

3.8.2. Ridge Regression

Ridge is a simple regularization technique, that helps to reduce model complexity, multicollinearity as well as overfitting. So, Ridge Regression "penalizes the L2-norm of the

coefficients in linear regression" (Rokem & Kay, 2020), reducing the variance of the estimates. The penalty term, denoted by alpha, "controls the amount of regularization" (Rokem & Kay, 2020), thus being a very important parameter to tune.

3.8.3. Polynomial Regression

A non-linear model created by adding to the linear equation, the original variables taken to a power:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 \ldots + \varepsilon$$

The parameters of the equation are estimated using the ordinary least squares, just like in Linear Regression. Using PolynomialFeatures, a new feature matrix consisting of all polynomial combinations and respective interaction terms was generated. Interaction terms consider that several features can amplify or diminish each other's effect on the target variable. After creating the polynomial features from the original ones, a Ridge or a Lasso model can be fitted if there is a need to regularize the model, thus preventing overfitting. It is also important to note that a high degree Polynomial model, can lead to overfitting as well. For this reason, choosing the right degree is key to obtain a robust Polynomial model.

3.8.4. Gradient Boosting Regressor

Considering this complex dataset, Gradient Boosting will be used due to its prediction speed and accuracy when using datasets with these properties. This method starts by fitting an initial model to the data, followed by a second model with special focus on predicting with accuracy the cases in which the first model under performs. This process is then repeated many times in which each successive model tries to overcome the shortcomings of the combined boosting ensemble of all the others. Thus, this model is based on the fact that successive models when combined with previous, will minimize the overall prediction error.

3.8.5. Stochastic Gradient Descent Regressor

Another model used was the Stochastic Gradient Descent Regressor, that is nothing more than a Linear model fitted by minimizing a regularized empirical loss. For both goals total and goals assists, the Elastic Net model was the one fitted using the Stochastic.

4. Results - Goals Total

It is important to refer that due to the limitations, in terms of computation time, of personal computers, it was decided to use Google Colab to run the experiments of this dissertation. Colab is nothing more than a free Jupyter notebook environment, which runs entirely in the cloud.

Performance of players was defined as goals total and goals assists achieved by an individual player over one season. As such, the same techniques and ML models were used for both dependent features separately. Thus, it was possible to study not only, which variables were important when predicting both these dependent features, but also how team variables affected these two predictors.

The features chosen as the most important, the models with respective results, the hyperparameter tuning and the evaluation on the test set will be first discussed for the variable goals total, and only after, will goals assists be analysed.

4.1. Features Analysis

In order to study which variables affected goals total the most and whether team variables play in fact, a significant role, the features chosen by feature selection algorithm will be presented and discussed.

In the table below (Table 2) the top five features given by each algorithm are presented:

Table 2. Top five features selected from feature selection methods

Univariate Selection	Score
shots_on	0,619
shots_total	0,495
games_minutes_played	0,347
games_appearences	0,333
games_lineups	0,297
Correlation Matrix with Heatmap	
shots_on	0,840
shots_total	0,770
games_minutes_played	0,590
games_lineups	0,580
rating	0,580
Ridge Regression	
shots_on	3,120
games_minutes_played	3,101
club_total_goals	0,887
penalty_success	0,709
youtube_avgview	0,627

Lasso Regression	Score
shots_on	2,360
games_minutes_played	1,100
penalty_success	0,672
passes_total	0,542
club_total_goals	0,375
Random Forest Regressor	
shots_on	0,720
games_appearences	0,022
penalty_success	0,018
games_minutes_played	0,015
club_total_goals	0,012
Extra Trees Regressor	
shots_on	0,418
shots_total	0,167
Midfielder	0,054
games_appearences	0,048
penalty_success	0,040

Recursive Feature Elimination	Score
shots_on	N/A
shots_total	
games_minutes_played	
games_appearences	
games_lineups	
passes_total	
passes_key	
penalty_success	
rating	
passes_accuracy_midfielder_team	
passes_accuracy	
duels_won	
duels_total	
fouls_committed	
tackles_interceptions	
goals_assists_team	
substitutes_in	
substitutes_bench	
twitter_avglikes	
twitter_avgretweets	

It is relevant to point out that it was not possible to extract the scores from the Recursive Feature Elimination method, so all the top 20 features selected were presented to facilitate comparisons. One very important point to notice is that *shots_on*, is agreeably the feature considered the most important when predicting goals total, as it was chosen as the top one feature by all the methods. This intuitively makes all the sense, as the number of shots to the goal that players make, is for sure related to how many goals they score.

At the same time, *shots_total* was seen to be very relevant in four algorithms out of seven and *games_minutes_played* appeared also to be important when predicting this performance indicator, only being left out by the Extra Trees Regressor algorithm.

At first sight it is noticeable that team variables are not present in the top five of features chosen, however, in the chosen 20 features from the Recursive Feature Elimination technique, two team variables are displayed: the passes accuracy of the midfielders (*passes_accuracy_midfielder_team*) and the goals assists of the teammates of a given player (*goals_assists_team*). These may be indications pointing to the fact that these variables can potentially influence player performance.

These feature sets were used to create seven distinct datasets, with the top 20 chosen features adding also all the team variables, due to the problem in question.

4.2. Models

When it comes to model's performance, a baseline model was created, by assigning the median of goals to all the observations' predictions. This is a very simplistic approach, only used for comparisons, so the RMSE achieved was a substantial high one, of 4.62. This means that the predictions made by this model are off by 4.62 goals on average. It is therefore, expected, that the ML models developed and tuned produce better results, meaning, lower RMSE than this baseline one. When applying a 10-fold cross-validation to twelve models (Table 3) with their default parameters, the mean RMSE scores were the following:

Table 3. Mean RMSE scores for each combination of dataset and model

	All features	Univariate feature selection	Feature Importance	Correlation Matrix	Recursive Feature Elimination	Ridge feature selection	Random Forest feature selection	Lasso feature selection	Team features
Decision Tree Regressor	2.58386	2.63612	2.58782	2.63856	2.68042	2.6065	2.60149	2.53763	4.18257
Linear Regression	1.70087	1.72474	1.8048	1.83822	1.72495	1.88951	1.81841	1.67705	4.08577
Random Forest Regressor	1.83025	1.91165	1.84831	1.9041	1.92923	1.89922	1.8331	1.80882	3.20452
SVR	2.48941	2.42849	2.39787	2.47038	2.38847	2.54069	2.41752	2.35042	4.33412
Gradient Boosting Regressor	1.74866	1.83445	1.73845	1.87563	1.85091	1.84497	1.74883	1.71951	3.77443
Ridge Regression	1.69594	1.72558	1.80562	1.83982	1.72479	1.88925	1.81814	1.67697	4.08549
Stochastic Gradient Descent (elastic net)	1.74052	1.75318	1.84877	1.87574	1.73993	1.88826	1.84037	1.67927	4.08901
KNN	2.48192	2.28948	2.1294	2.35762	2.18899	2.44841	2.15467	2.16237	3.84713
Lasso Regression	2.43705	2.43705	2.43705	2.43705	2.43705	2.43705	2.43705	2.43705	4.1804
Polynomial Regression	10.9649	2.98223	6.73098	4.85139	5.64354	2.9282	5.97497	2.06182	4.0937
Polynomial corrected with Ridge	5.7674	2.2138	3.10494	3.09464	2.42696	2.74061	3.51586	1.97073	4.08643
Polynomial corrected with Lasso	2.46559	2.43981	2.46154	2.46335	2.43264	2.46435	2.4622	2.46235	4.1804

One of the first things to notice is that the ML models applied to a dataset with solely team variables (column named 'Team features') to predict goals, performed poorly overall, achieving very high values of error. This is expected, as a big part of the performance of a professional player is related to his individual performance. As such, it was not expected to validate that team variables will be fully responsible for the goals total of a player.

One thing to reference is the fact that not only a 2nd degree polynomial was run, but also, the same model was regularized with Lasso and Ridge in an attempt to reduce overfitting.
Even though, regularization helped, as the RMSE is lower than in the simple Polynomial model, these ones are still not achieving the best results.
After taking a closer look at the scores (Table 3), it is noticeable that the best results were achieved with the subset that has the features chosen by the Lasso algorithm in the feature selection phase. More specifically, by using this dataset, four models achieved the lowest RMSE:

- Linear Regression - 1.67705
- Ridge Regression - 1.67697
- Gradient Boosting Regressor - 1.71951
- Stochastic Gradient Descent Regressor - 1.67927

Since these four models displayed by far the lowest RMSE scores, the process of tuning the hyperparameters was pursued only on these ones, as an attempt to try and improve the errors even more.

4.3. Hyperparameter Tuning

4.3.1. Linear Regression

The parameters of the linear regression are all Boolean, as such only two values can be chosen in each: True or False. In this case the default Boolean values were the ones leading to a lower RMSE values, one of 1.6137 in the training and 1.6806 in the validation. Despite this being a very simplistic model, it achieved a one of the best four RMSE, when compared to other models, so its use was not disregarded. The final model is therefore the following (Figure 5):

```
LinearRegression(copy_X=True, fit_intercept=True,
                 normalize=False, n_jobs=-1)
```

Figure 5. Linear Regression best model

By looking at the learning curve, it can be understood how overfitting depends on the amount of training data, as a learning curve shows the model's performance on the training and validation sets as a function of the amount of training data. When looking at the plot (Figure 6), it can be seen that below 200 observations the overfitting was a major handicap, yet, as the observations increased, both curves converged leading to the considerable improvement when

it comes to this problem. At 1000 instances the training curve seems to have a slightly lower RMSE while the validation curve stabilized, as such, overfitting increased faintly.

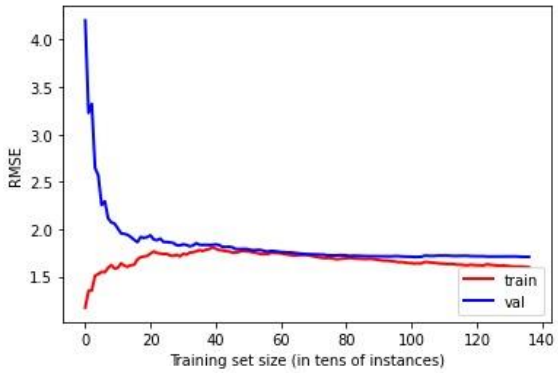

Figure 6. Linear Regression learning curve

4.3.2. Ridge Regression

For this model three parameters were tuned: *alpha*, *max_iter* and *solver*. Alpha corresponds to the regularization strength while *max_iter* to the maximum number of interactions the model should run if it does not converge before. Higher alpha values were used but the RMSE achieved was always worse, as such low values of alpha between 0.5 and 3 were experimented with. Alongside this, every *solver* value was used in the tuning, while the maximum number of interactions started in 1000 until 4000 with step of 1000, including the default value *None*. After grid search, the best model found achieved a validation RMSE of 1.68054 with the following parameters (Figure 7):

```
Ridge(alpha=1.8, copy_X=True, fit_intercept=True,
      max_iter=None, normalize=False, solver='svd',
      tol=0.001, random_state=7)
```

Figure 7. Ridge Regression best model

The learning curve (Figure 8) shows that below the 200 observations there is a clear overfitting with validation set fitting the data poorly. From the 400 observations until the 800, overfitting is almost eliminated, and the training and validation curves converge completely. However, as training size keeps increasing, the train RMSE reduces while the validation is kept on the same level, leading to the overfitting problem appearing again.

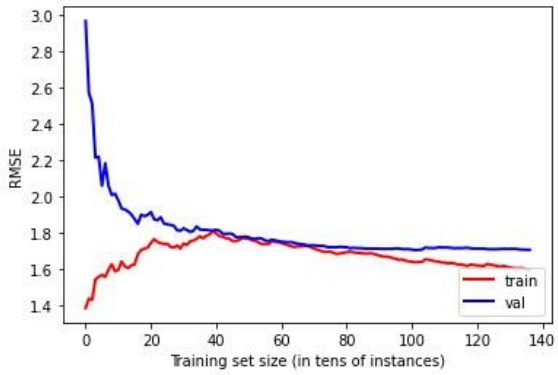

Figure 8. Ridge Regression learning curve

4.3.3. Gradient Boosting Regressor

In this model some parameters were chosen to be fine-tuned. It is important to understand that the parameters of the Gradient Boosting, an ensemble model, can be split in three blocks:

- Tree-specific ones, which are the ones affecting each individual tree.
- Boosting parameters that impact the boosting operation (sequential process in which every new model tries to solve the errors of the previous one on subsets of the dataset).
- Miscellaneous parameters corresponding to all the others used by this algorithm.

To achieve the best model, it was decided to tune a few parameters simultaneously in each step, keeping the others constant. Once their best value was found, these parameters kept that value in the subsequent steps while tuning the rest. This was performed in the following way:

Step 1

The following parameters started with the below initial values, while *n_estimators* was the first to be tuned:

- *learning_rate* = 0.1 (default value)
- *min_samples_split* = 20, approximately 1% of total values
- *min_samples_leaf* = 50
- *max_depth* = 5, chosen based on the number of observations
- *max_features* = 'sqrt'
- *subsample* = 0.8
- In this first step, the optimum number of trees, which corresponds to the *n_estimators,* was tuned, with grid search to test values from 20 to 200 with steps of 10.
- The model with number of estimators 200 achieved the best validation RMSE, one of 1.8052. Higher values of the number of estimators were experimented with, but the very small improvements did not justify the longer computational time.

Step 2

From now on, the tree-specific parameters were the ones tuned. With this optimal number of estimators, the maximum depth, which limits the number of nodes in the tree and the minimum samples split, which corresponds to the minimum number of samples (observations) required in a node to be considered for splitting, were tuned.

- *max_depth* ranging from 3 to 15 with steps of 2 and *min_samples_split* ranging from 2 to 10 with steps of 1.
- The optimal values were found to be 5 for *max_depth* and 9 for *min_samples_split*. The validation RMSE achieved was approximately 1.7295, a very significant improvement.

Step 3

Setting these parameters to their optimal values, the *min_samples_leaf,* which defines the minimum number of observations required to be left in a terminal node, was tuned. From 1 to 40, with steps of 5, the default value 1 was chosen. The validation RMSE did not suffer any significant changes.

Step 4

The last tree-specific parameter was tuned, the *max_features,* which is the number of features to consider when looking for the best split. In a range from 7 to 42 with steps of 2, the value chosen was 13 and validation RMSE decreased to 1.6954.

Step 5

With the most important tree-parameters tuned, the boosting ones were next. Several values smaller than 1 were tried in the *subsample* parameter, but the default 1 gave the best RMSE (1.6887).

Step 6

Now, the learning rate was lowered while the number of estimators was increased, proportionally. The *learning_rate* determines the impact of each tree on the final outcome and usually, lower values should be preferred as the model would be more robust to tree characteristics, thus allowing it to generalize well.

- First *learning_rate* was decreased by half (taking the value of 0.05) and *n_estimators* increased by the double (new value of 400). This led to a RMSE of 1.6780, still an improvement versus last steps
- Then the *learning_rate* original value was decreased by 1/10 (new value of 0.01) while *n_estimators* increased by 10 times, reaching the value of 2000. This led to a RMSE of 1.6533, a very significant improvement.

- Lastly, *learning_rate* decrease to 1/20 of the original value (new value of 0.005) while *n_estimators* increased by 20 times, reaching the value of 4000. This led to a RMSE of 1.6549, slightly worse than the last one, while the algorithm took longer to converge.

Therefore, the best parameters for Gradient Boosting Regressor were found to be the following (Figure 9), leading to a validation RMSE of 1.6533:

```
GradientBoostingRegressor(learning_rate=0.01, n_estimators = 2000,
                          max_depth = 5, min_samples_split = 9,
                          min_samples_leaf = 1, max_features= 13,
                          subsample = 1.0, random_state=7)
```

Figure 9. Gradient Boosting Regressor best model

Below the learning curve for this model can be observed (Figure 10). In the plot below it can be seen the model fits the validation set very poorly, having a clear overfitting, especially below 200 instances. Overall, even when both curves seem to plateau, there is still some significant overfitting present as both curves do not converge. This can also be seen in the difference between the RMSE of the training and the validation which is 0.3425 and 1.6533 respectively. This is a significant difference which lead to the conclusion that the model suffers from overfitting.

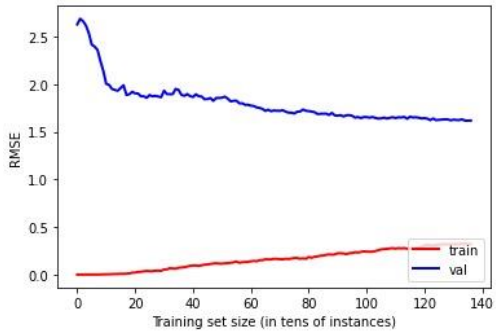

Figure 10. Gradient Boosting Regressor learning curve

4.3.4. Stochastic Gradient Descent Regressor

In this model parameters, it was noticed that a relation existed between the *learning_rate* and some other parameters, like *alpha* and *eta0*. For that reason, it was decided in a 1st instance to keep the values of each *learning_rate* fixed, while tuning the parameters from which they depend upon. After seeing which *learning_rate* and respective parameters achieved the lowest validation RMSE, some other parameters were fine tuned.

Step 1

Learning rate was fixed at "constant", while the *eta0* values (the initial learning rate: eta= eta0) was varied between 0.0001 and 0.02 with 0.0002 steps. Higher values for the *eta0* were also

tried but the RMSE was always worse. The *eta0* chosen was 0.0003 and the validation RMSE was 1.6810.

Step 2

The learning rate "optimal" depends upon the alpha parameter, as the *t0* is given:

$$eta = \frac{1}{(alpha * (t + t0))}$$

Thus, the *alpha* values were varied between 0.0001 and 2 with steps of 0.001. The best validation RMSE achieved was 1.9622, a very high error compared to the one obtained on the previous step.

Step 3

On this step, the learning rate was set to "invscaling", that depends on *eta0* and *power_t* parameters:

$$eta = \frac{eta0}{pow\ (t, power\_t)}$$

While *eta0* was varied between 0.0001 and 0.02 with 0.0002 steps, *power_t* was set to values between 0.1 and 1 with 0.15 steps. The best model was found to be the one using *eta0* of 0.0045 and *power_t* of 0.25 (default value), leading to an error of 1.67869, which poses a substantial decrease versus the other two steps.

Step 4

Finally, the "adaptive" learning rate was fixed, and the parameters *eta0*, *n_iter_no_change* and *early_stopping* were fine tuned. The best model found, meaning the one with the lowest validation RMSE had 0.0051, 1 and True on these parameters respectively. However, the validation RMSE was 1.6792, a higher value than the one achieved on step three with "invscaling".

Step 5

Keeping fixed the learning rate "invscaling" - as it was the one in which the obtained validation RMSE was lower - and the respective values obtained on the 3rd step, the rest of the stochastic parameters were fine-tuned: *max_iter*, *n_iter_no_change, early_stopping, loss* and *penalty.*

The model that chose the best values for these parameters, obtained an improved validation RMSE of 1.67868. Nonetheless, since the penalty chosen was "elasticnet", the *l1_ratio* parameter had to be tuned, as this parameter is only used with this kind of penalty type.

Step 6

The *l1_ratio* parameter was varied between 0 and 1 with 0.15 steps, being the chosen value 0.899. The final validation RMSE was 1.67866, which represents a very small improvement,

but still the best error achieved with the Stochastic model. All in all, the best model found had the following parameters (Figure 11):

```
SGDRegressor(learning_rate='invscaling', eta0=0.0045,
             loss='squared_loss', penalty='elasticnet',
             power_t=0.25, max_iter=1000,
             n_iter_no_change=5, early_stopping=False,
             l1_ratio=0.8999, random_state=7)
```

Figure 11. Stochastic Gradient Descent best model

The learning curve in this model (Figure 12) presents the same behaviour as the Ridge one.

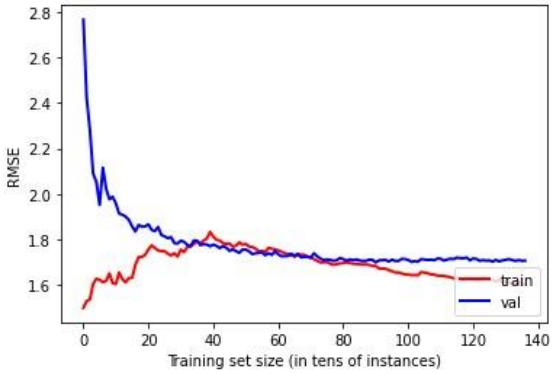

Figure 12. Stochastic Gradient Descent Regressor learning curve

4.4. Comparison of final models

All in all, the table below (Table 4) provides an overview of train and validation RMSE of the four best ML models, ordered by the best validation RMSE to the worst, after the hyperparameter tuning was applied:

Table 4. Comparison of best models for goals total

Model	RMSE Train	RMSE Validation	Used dataset
Gradient Boosting Regressor	0.3425880791521038	1.653297683399031	Lasso feature selection
Stochastic Gradient Descent	1.6175899878408615	1.6786655116689357	Lasso feature selection
Ridge	1.6138547372291452	1.6805450825570707	Lasso feature selection
Linear Regression	1.61366514260722	1.6806231461221328	Lasso feature selection
Baseline	4.6216003422321315	N/A	All features

It was observed that all the models executed performed significantly better than the baseline one. Moreover, the Gradient Boosting Regressor is the one presenting the lowest validation RMSE. However, when looking at the training RMSE, the value is even lower, 0.3426, so the considerable difference between the two indicates the presence of overfitting.

On the other hand, all the other three models present very small differences between training and validation RMSE so overfitting is reduced. More specifically, Linear and Ridge Regressions present very similar performances. Due to the similar and relatively good

performance of all these four models, showing that the predictions are off by less than two goals on average, further analysis will be pursued for each one of these models.

4.5. Analysis of the most prominent models

In order to study each model in more depth several analyses were conducted:

1) The Feature importance was studied for all models. The idea is to have a score for each feature based on their importance when predicting the output. If features are highly responsible to predict the target variable, their scores will be higher. For the parametric algorithms, like Ridge and Linear and also for the Stochastic one, which natively do not support feature importance, permutation_importance was used. Permutation Importance measures feature importance by calculating how the model's performance is affected after changing the features' values. A feature is considered important when the error of the model increases if the feature values are shuffled. On the other hand, if shuffling the values of a feature bears no impact on model performance, then this feature is not important. This measurement was introduced by Breiman (2001).
2) Coefficients for the Ridge, Stochastic and Linear models were extracted, as well. By looking at those, the effect of the predictor on the target variable, for example, the sign or magnitude of the effect, can be studied with more precision.
3) Feature significance to tell if team-related variables are statistically significant when predicting the outcomes of this dissertation or not. For the Linear model the p-values were extracted, whether for the other models, a paired sample T-Test was used.
4) Graphs to visualize predictions made by the models, in which from the training data, 90% was used to fit the models, while 10% to validate and generate the predictions (Appendix 2 and Appendix 3).

5. Discussion - Goals Total

5.1. RQ1: Most suitable learning algorithms

When evaluating these four best models on the test set, the results were the following (Table 5):

Table 5. Evaluation of four best models on the test set

Model	RMSE Train	RMSE Validation	RMSE Test	Used dataset
Gradient Boosting Regressor	0.3425880791521038	1.653297683399031	1.2822653812017957	Lasso feature selection
Stochastic Gradient Descent	1.6175899878408615	1.6786655116689357	1.3542923786407413	Lasso feature selection
Ridge	1.6138547372291452	1.6805450825570707	1.3607219507248154	Lasso feature selection
Linear Regression	1.61366514260722	1.6806231461221328	1.3622306160922737	Lasso feature selection

Gradient Boosting Regressor achieved the lowest RMSE, one of 1.2822, whereas Linear Regression was the model with the highest RMSE, of 1.3622.

Interestingly, these results are significantly better than the ones obtained on the validation set, as such it can be concluded that all these models perform relatively good when presented with a new dataset. As such, any of these learning algorithms seem to be suitable for this problem, which answers the first of the research questions.

Even though Gradient Boosting constantly achieved the best RMSE values, Stochastic has significantly less overfitting and the team-related variables had higher importance and higher accordance with other models than Gradient Boosting. Thus, Stochastic Gradient Descent, with RMSE of 1.3543, was the final model chosen to pursue with this dependent variable prediction (goals total). It is also relevant to say that, since the features were scaled, some predictions will be negative, however, in this model only 26 observations out of 604 were predicted as a negative value.

5.2. RQ2: Most Suitable variables for this problem

The Linear Regression model indicates that *shots_on, games_minutes_played, club_total_goals*, *passes_total* and *penalty_success* are the top five important features when predicting goals total. Further, *shots_on* takes the place of the most important variable, however, in this model, the minutes a player played in a game presents a very similar importance, assuming also a leading role when it comes to predicting the target variable (Table 6). When looking at the p-values (Figure 13) it can be seen that all of these five variables have p-values of zero, which is less than 5%, as such, the null hypothesis (H0: coefficients are equal to zero) was rejected and all of these variables have a statistically significant impact on predicting goals total.

Table 6. Feature importance for Linear Regression model

Feature	Importance
shots_on	10.42272490384158
games_minutes_played	10.085481635122598
club_total_goals	1.2051738360142614
passes_total	1.1017837111285544
penalty_success	1.0301198414123776
goals_total_attacker_team	0.5224333359619872
rating	0.43305733008384023
passes_key	0.4296721663766198
tackles_interceptions	0.3298327123761781
fouls_drawn	0.31841501743773165
goals_assists_team	0.311063328703321
youtube_avgview	0.20885157738110807
Midfielder	0.17943892070887274
tackles_total	0.14652367432849583
goals_total_midfielder_team	0.1199825464549856
passes_accuracy_midfielder_team	0.08919477215472287
club_match_played	0.0680283509793691
dribbles_attempts	0.06378578347496404
market_value_opponents	0.05666686687123201
substitutes_out	0.03034396506399446
First Division A	0.02214665905144435
diversity_team	0.018177013398833885
games_lineups	0.018055036691970595
duels_won_defender_team	0.008945755630135688
market_value_team	0.008579415565138415
age_team	0.0048235142987035125
tackles_total_defender_team	0.0021470615749179965
height	0.0002701896609712762
weight	-3.356310449609268e-06
goals_conceded_goalkeeper_team	-0.0002157529493374888

OLS Regression Results

Dep. Variable:	goals_total	R-squared:	0.852
Model:	OLS	Adj. R-squared:	0.849
Method:	Least Squares	F-statistic:	381.1
Date:	Fri, 20 Aug 2021	Prob (F-statistic):	0.00
Time:	12:12:28	Log-Likelihood:	-3852.9
No. Observations:	2024	AIC:	7768.
Df Residuals:	1993	BIC:	7942.
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.8977	0.036	79.676	0.000	2.826	2.969
shots_on	2.2681	0.080	28.192	0.000	2.110	2.426
games_minutes_played	2.2346	0.414	5.400	0.000	1.423	3.046
penalty_success	0.7227	0.044	16.411	0.000	0.636	0.809
passes_total	-0.7443	0.093	-8.000	0.000	-0.927	-0.562
club_total_goals	0.7820	0.098	7.982	0.000	0.590	0.974
games_lineups	-0.0928	0.423	-0.220	0.826	-0.922	0.737
Midfielder	-0.2986	0.044	-6.742	0.000	-0.385	-0.212
tackles_interceptions	-0.4052	0.082	-4.938	0.000	-0.566	-0.244
rating	0.4678	0.060	7.772	0.000	0.350	0.586
youtube_avgview	0.3292	0.043	7.675	0.000	0.245	0.413
First Division A	-0.0907	0.226	-0.401	0.688	-0.534	0.353
tackles_total	-0.2706	0.081	-3.345	0.001	-0.429	-0.112
fouls_drawn	-0.3949	0.067	-5.882	0.000	-0.527	-0.263
passes_key	-0.4762	0.082	-5.811	0.000	-0.637	-0.316
club_match_played	-0.1951	0.074	-2.646	0.008	-0.340	-0.050
passes_accuracy_midfielder_team	-0.2266	0.221	-1.025	0.305	-0.660	0.207
dribbles_attempts	-0.1811	0.065	-2.803	0.005	-0.308	-0.054
height	0.0197	0.054	0.366	0.714	-0.086	0.125
substitutes_out	0.1223	0.056	2.187	0.029	0.013	0.232
weight	0.0014	0.053	0.025	0.980	-0.103	0.106
goals_conceded_goalkeeper_team	-0.0130	0.049	-0.269	0.788	-0.108	0.082
duels_won_defender_team	-0.0802	0.111	-0.725	0.469	-0.297	0.137
tackles_total_defender_team	0.0220	0.107	0.206	0.837	-0.188	0.232
market_value_opponents	0.1565	0.096	1.631	0.103	-0.032	0.345
diversity_team	0.0919	0.041	2.232	0.026	0.011	0.173
goals_total_midfielder_team	-0.2452	0.052	-4.736	0.000	-0.347	-0.144
goals_total_attacker_team	-0.5172	0.067	-7.744	0.000	-0.648	-0.386
goals_assists_team	0.3865	0.107	3.625	0.000	0.177	0.596
age_team	-0.0400	0.043	-0.933	0.351	-0.124	0.044
market_value_team	-0.0741	0.083	-0.893	0.372	-0.237	0.089

Omnibus:	567.760	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5038.005
Skew:	1.060	Prob(JB):	0.00
Kurtosis:	10.433	Cond. No.	45.8

Figure 13. Feature significances for Linear Regression model

Additionally, for this model, the linear assumptions were checked to see how adequate this model was. Firstly, the scatterplot of the standardized residuals can be seen below (Figure 14):

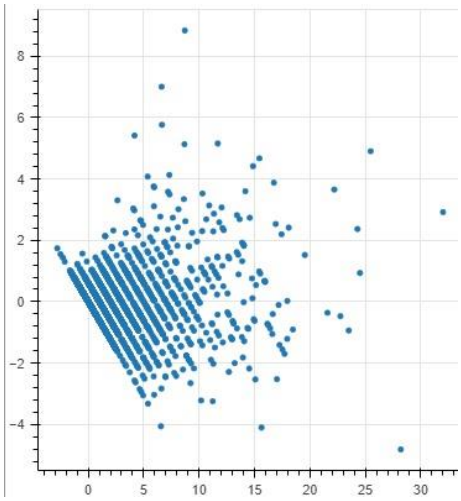

Figure 14. Scatterplot of standardized residuals for goals total

The residuals are randomly scattered around zero, and there is no apparent pattern to the residuals, which suggests that a linear model is appropriate for this relationship (if there was a pattern, such as a U-shape, this would indicate a non-linear relationship). Further, the histogram below suggests the errors are approximately normally distributed (Figure 15).

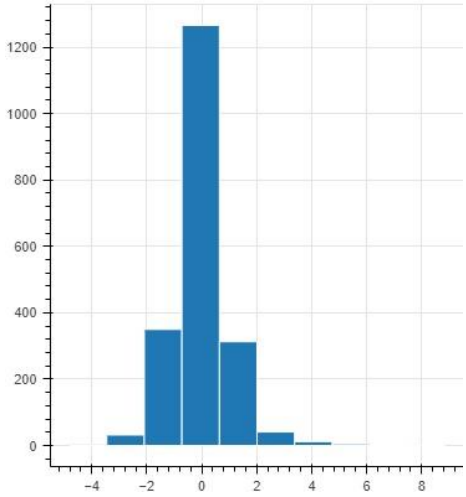

Figure 15. Histogram of residuals for goals total

Furthermore, the Durbin-Watson tests (Figure 13) if the values of the residuals are independent of each other, and for this assumption to be met, the value needs to be as close to two as possible. In this case the value is 2.006 and therefore this assumption will be taken as being met. Moreover, the assumption of independence of the errors and homoscedasticity are seldom violated in data where there is no temporal dimension, which is the case with this one, so this will also be assumed as met.

When it comes to the features considered most important by the Ridge Regression model, the same exact variables and order of the Linear Regression are presented here. Just like in the previous model, *shots_on* assumes a preponderant role with a substantial higher score versus its peers (Table 7). Shifting the attention to the coefficients of these variables, four out of the five described above have positive coefficients (Table 8). However, *passes_total* has a negative coefficient, one of -0.7439. The coefficients define the impact that a unit change of the independent variable (in this case, passes) has on the dependent variable (goals scored by a player). If this coefficient is statistically significant, it can be said that this variable has a negative effect on goals, meaning that an increase in the number of passes will lead to a decrease in the goals total.

Table 7. Feature importance for Ridge Regression model

Feature	Importance
shots_on	10.422001691040705
games_minutes_played	8.235324378355076
club_total_goals	1.187749588012395
passes_total	1.1016010123121904
penalty_success	1.0249831397557512
goals_total_attacker_team	0.5147406989342402
rating	0.42965696991695895
passes_key	0.4203878775890379
tackles_interceptions	0.3306630534055004
fouls_drawn	0.310765709565589
goals_assists_team	0.31075153604518746
youtube_avgview	0.2094967663204652
Midfielder	0.18484354489549748
tackles_total	0.14636597221734315
goals_total_midfielder_team	0.11758956218413355
passes_accuracy_midfielder_team	0.08153142664544628
club_match_played	0.06758076764394957
dribbles_attempts	0.06206438637708862
market_value_opponents	0.05627985245433385
games_lineups	0.029991361598039567
substitutes_out	0.025021074288340238
diversity_team	0.018102677420031375
First Division A	0.016896791207311867
market_value_team	0.008229974819567998
duels_won_defender_team	0.00812396196916403
age_team	0.004731195930439069
tackles_total_defender_team	0.001984401940264435
height	0.0003744077434043369
weight	-2.426121767173939e-06
goals_conceded_goalkeeper_team	-0.00021461032760599074

Table 8. Coefficients for Ridge Regression model

Feature	Coefficient
shots_on	2.2670048082071927
games_minutes_played	2.018650086029632
club_total_goals	0.7760437315536826
penalty_success	0.7205969103989852
rating	0.46568099001351804
goals_assists_team	0.3861269524346507
youtube_avgview	0.3295324489403726
market_value_opponents	0.15580187774520302
games_lineups	0.12347736323035385
substitutes_out	0.1109155946935251
diversity_team	0.09171420184602423
height	0.02148860422320888 3
tackles_total_defender_team	0.020855916336330584
weight	0.0006338184180964797
goals_conceded_goalkeeper_team	-0.013011908174953872
age_team	-0.03961098017247275
market_value_team	-0.0727331567923799
duels_won_defender_team	-0.07708408159747661
First Division A	-0.07744702571494855
dribbles_attempts	-0.1785514138046607
club_match_played	-0.19436515029169038
passes_accuracy_midfielder_team	-0.21732669314627776
goals_total_midfielder_team	-0.2426570420014105 8
tackles_total	-0.2702611964576464
Midfielder	-0.30295984577870055
fouls_drawn	-0.3898760390018814
tackles_interceptions	-0.4055392246794401
passes_key	-0.4710374385582822
goals_total_attacker_team	-0.5131094282165953
passes_total	-0.7439181745069416

In the Gradient Boosting Regressor, when looking at which variables the model found to be the most important when predicting goals total, the Table 9 was analysed. It can be seen that *shots_on, penalty_success, games_minutes_played, games_lineups and Midfielder* were found to be the most relevant variables. The last one indicates that the goals scored by the player depends upon his position, if he is an attacker or midfielder.

It is also important to point out that *shots_on* presents much higher importance when compared to all the other variables, being therefore highly relevant to predict goals total.

Table 9. Feature importance for Gradient Boosting Regressor model

Feature	Importance
shots_on	0.49129458908950685
penalty_success	0.08917416772272634
games_minutes_played	0.08185289021403093
games_lineups	0.057726718173580854
Midfielder	0.045824775709575605
rating	0.040841919774697984
club_total_goals	0.020907219734861884
youtube_avgview	0.02016621000094383
tackles_interceptions	0.018170931462612873
tackles_total	0.015684985972007327
substitutes_out	0.014288515739284679
goals_assists_team	0.013908620194835748
dribbles_attempts	0.01142779072345358
passes_key	0.011348415742112768
passes_total	0.010606180990892618
goals_conceded_goalkeeper_team	0.00582421520697462 1
market_value_opponents	0.0055636596688615 09
passes_accuracy_midfielder_team	0.0053343651928731564
goals_total_attacker_team	0.005061150710912588
height	0.00495197775504443
goals_total_midfielder_team	0.004892890503055052
weight	0.004695177849095415
age_team	0.004646928557481631
market_value_team	0.003917742788590147
fouls_drawn	0.0034937071105388684
duels_won_defender_team	0.0032521904000055056
tackles_total_defender_team	0.0018168567574896634
diversity_team	0.0017596743890724057
First Division A	0.001058626748505605
club_match_played	0.0005069051163754642

In the Stochastic Gradient Descent model, *shots_on, games_minutes_played, games_lineups, passes_total, club_total_goals* and *penalty_success* are by far the six variables that exert the highest importance in the target variable (Table 10). All these variables were found to be the ones with the highest importance whether in the Gradient Boosting (Table 9) or on the Linear/Ridge Regression (Table 6 and Table 7 respectively). Just like in the Ridge, *passes_total* is the only one out of the six described that has a negative coefficient, and with very similar value to the one on the Ridge. As such, if statistically significant, the same interpretation applies and more passes that a player makes, the less goals he scores (Table 11).

Table 10. Feature importance for Stochastic model

Feature	Importance
shots_on	10.586600565880454
games_minutes_played	2.875574137905948
games_lineups	1.9635756765825068
passes_total	1.133769847905296
club_total_goals	1.0659740495614813
penalty_success	1.015450976787795
goals_total_attacker_team	0.4833940569898697
rating	0.42565713717714393
passes_key	0.3810412755379016
goals_assists_team	0.3490940529531738
tackles_interceptions	0.31510972945438454
fouls_drawn	0.27732088940938476
youtube_avgview	0.21729141975341698
Midfielder	0.2036474091666835
tackles_total	0.13421582057090106
goals_total_midfielder_team	0.1082395919106141
club_match_played	0.06188885741599211
dribbles_attempts	0.05572542290285396
market_value_opponents	0.04799309665120592
passes_accuracy_midfielder_team	0.023268068777569262
diversity_team	0.017344631094962315
substitutes_out	0.00852163907636716
market_value_team	0.004474914840070854
age_team	0.003935987391367579
duels_won_defender_team	0.0033381453661264883
tackles_total_defender_team	0.0020604023367385427
height	0.0005197731049112253
weight	5.3734339091215946e-05
First Division A	-7.697641538868538e-05
goals_conceded_goalkeeper_team	-0.00027100830168618194

Table 11. Coefficients for Stochastic model

Feature	Coefficient
shots_on	2.2754621107395665
games_minutes_played	1.1811749121571344
games_lineups	0.9782359168506095
club_total_goals	0.728588210489558
penalty_success	0.7130234477062557
rating	0.4545494774383587
goals_assists_team	0.4045026873635778
youtube_avgview	0.3335544901602708
market_value_opponents	0.13984314074565476
diversity_team	0.09158659698875746
substitutes_out	0.05791862060262916
First Division A	0.04298270787503227
tackles_total_defender_team	0.019602749509617556
height	0.01870480231859661
weight	-0.008033200508472645
goals_conceded_goalkeeper_team	-0.008270692236741274
age_team	-0.03363886886286477
duels_won_defender_team	-0.05853650999841843
market_value_team	-0.06107089757354184
passes_accuracy_midfielder_team	-0.12958999802984922
dribbles_attempts	-0.17863440186615775
club_match_played	-0.19056483658325107
goals_total_midfielder_team	-0.2338702992902145
tackles_total	-0.2679484674470861
Midfielder	-0.32080942716146227
fouls_drawn	-0.3778076308578656
tackles_interceptions	-0.4043264026128439
passes_key	-0.460023302282962
goals_total_attacker_team	-0.5001228119799915
passes_total	-0.7660928973089222

After all these analyses, when shifting the attention to the RQ2 (the most suitable variables for this prediction problem), *shots_on, penalty_success, games_minutes_played, club_total_goals*, *passes_total* and *games_lineups* were agreeably, the top six variables with the highest importance when predicting goals total.

5.3. RQ3: Contribution of team-related variables

For the Linear Regression, the team variables that take the top places are *goals_total_attacker_team, goals_assists_team, goals_total_midfielder_team* and *passes_accuracy_midfielder_team* (Table 6). In the Linear model the p-values were observed

directly (Figure 13), and for the first three variables, the p-values are zero. This means that these variables are indeed statistically significant when it comes to predict goals total. Surprisingly, the only other team variable that was statistically significant was *diversity_team.* In this model, the adjusted R-squared takes the value of 0.849, which means that 84.9% of the total variability of goals total is explained by this model, corresponding to a somewhat good explanatory power. In the Ridge Regression, the *goals_total_attacker_team* is the most relevant variable, similar to what was found in the Linear Regression. At the same time, *goals_assists_team, goals_total_midfielder_team* and *passes_accuracy_midfielder_team* also assume a role in explaining the target variable (Table 7).

When it comes to coefficients, it can be seen that some of these team variables have a negative effect on the target variable (Table 8):

- When the mean age of the teammates of a player is higher, this player will have lower number of goals scored. This can intuitively make sense, as his team may have lower performance due to the elevated average age which impacts individual performance as well.
- When a player has attackers and midfielders on his team that score a high number of goals (*goals_total_attacker_team* and *goals_total_midfielder_team)*, his performance, meaning goals scored, decreases. This can happen due to a higher competitive team environment and the fact that high performance players will be more likely to "monopolize" the goals' opportunities leaving less chances for others to score. The same goes for *passes_accuracy_midfielder_team* which also takes a negative coefficient.
- Midfielders have a negative relation with goals total, so attackers would score more goals than midfielders, which intuitively makes all the sense.

On the other hand, the assists made by the team have a positive relation with the goals total scored by a player. This again appears reasonable as when a team is working towards the same objectives, players will have no fear in allowing team members to score if that will lead to the success of the team. As such, more assists will positively affect individual players goals.

To understand how significant the team variables were, a paired sample T-Test was conducted, more specifically, a two-sided test. For this test, the process was as follows:

- The model was trained with the dataset including all team variables and then also trained with another dataset excluding the specific variable that its effect wanted to be studied.

Then both models were applied to the test set and those predictions were used as arguments for the paired test.

- The null hypothesis for the two-sided test is the following:

$$H_0: u_1 = u_2$$

Thus, the significance of four different team-related variables was tested: *goals_total_attacker_team, passes_accuracy_midfielder_team, goals_assists_team* and *goals_total_midfielder_team*. For the last two variables the null of the two-sided test was rejected, which means that the mean of the predictions with and without each of these variables is different, thus these team-related variables do have an impact on the goals total predictions (Table 12).

Table 12. Paired Sample T-Test for Ridge Regression

	Ttest_relResult	
	statistic	p-value
Excluding goals_total_attacker_team feature	-0.8397677	0.401371455
Excluding goals_assists_team feature	-6.42920597	2.610003e-10
Excluding goals_total_midfielder_team feature	8.0806197	3.529585e-15
Excluding passes_accuracy_midfielder_team feature	0.7505942	0.453189687

In the Gradient Boosting Regressor, the most relevant team variable was found to be *goals_assists_team*, followed by *goals_conceded_goalkeeper_team*, *market_value_opponents* and *passes_accuracy_midfielder_team*. Further, some team variables like diversity of the team, the tackles and duels won by defenders, the age and the market value seem to be less useful when predicting the dependent variable, and as such, in future studies they could be replaced by others with the potential to exert higher influence on goals total (Table 9). Adding to that, unlike what happened on the Ridge/Linear, in which the most important team variable assumed the 6th position on the importance chart, here, team variables appear a few positions below.

In the Gradient Boosting regressor, paired-sample T-Tests were done four times, to test the significance of four different team-related variables: *goals_assists_team*, *goals_conceded_goalkeeper_team*, *market_value_opponents* and *passes_accuracy_midfielder_team*.

For the *goals_conceded_goalkeeper_team*, *market_value_opponents* and *passes_accuracy_midfielder_team* the p-value of this test was below the 5% threshold, and therefore the null was rejected, which means that the mean of the predictions with and without each of these variables was different, thus these team-related variables have each one an impact on the goals total predictions (Table 13).

Table 13. Paired Sample T-Test for Gradient Boosting Regressor

	Ttest_relResult	
	statistic	p-value
Excluding goals_conceded_goalkeeper_team feature	-3.0946313	0.00206187
Excluding goals_assists_team feature	-1.0851457	0.27829075
Excluding market_value_opponents feature	8.5796712	8.05555e-17
Excluding passes_accuracy_midfielder_team feature	5.4144445	8.88799e-08

For the Stochastic Gradient Descent *goals_total_attacker_team, goals_assists_team*, *goals_total_midfielder_team* and *market_value_opponents* were found to be the most useful variables to predict the target variable (Table 10). These are almost the same ones found by the Ridge model (Table 7). Moreover, the coefficients of these variables have the same direction than on the Ridge model, thus the interpretations on the Ridge are transversal to the Stochastic. This similarity of results is a positive indication of the reliability of these results. However, when it comes to the coefficient (Table 11) for the *market_value_opponents,* it takes a positive value, meaning that when the average market value of players of the same league is higher, the player scores more goals, which may indicate that the player could have more motivation to perform at a higher level in an attempt to also earn more money.

For this model, the paired sample T-Test was applied to all these four variables, and excluding *goals_total_attacker_team,* it can be said that for the other three variables, the mean of the predictions with and without each of these variables was different, thus these team-related variables do have an impact on the goals total predictions (Table 14).

Table 14. Paired Sample T-Test for Stochastic Gradient Descent Regressor

	Ttest_relResult	
	statistic	p-value
Excluding goals_total_attacker_team feature	0.10828224	0.91380783
Excluding goals_assists_team feature	3.00001352	0.00281111
Excluding goals_total_midfielder_team feature	5.33380958	1.36217e-07
Excluding market_value_opponents feature	-25.7899169	2.06028e-99

After all these analyses, when shifting the attention to the RQ3 (What is the contribution of team-related variables to the performance of the prediction models), *goals_assists_team* and *goals_total_midfielder_team* were the ones found to have an impact on the predictions in both Ridge Model, Stochastic and Linear model. At the same time, *market_value_opponents* was also found to have an impact on the predictions on the Gradient Boosting and Stochastic models. Further, one interest point to make, is that the market value of teammates has a very low importance across all models, and as such, its effect on predicting goals total is substantially small. Interestingly, this goes against what Torgler and Schmidt (2011) found in their studies, in which a strong impact was observed for the relative income position.

5.4. Stochastic Gradient Descent Regressor further evaluation

Since Stochastic Gradient Descent Regressor was the model that obtained a significant low RMSE with less overfitting and with the team-related variables having higher importance and higher accordance with other models, some further evaluations were made. The Table 15 exemplifies some predictions made, where it can be seen that the predictions are actually very close to the original values. Out of these ten observations, four of them, if the predictions are rounded, would be correctly predicted. The worst overshoot here was in second line in which a player actually scored zero goals, but the model predicted a score of five goals. Even though the last line predicted 13 goals, when the original value was 14, this can be considered a very positive prediction given that the training dataset had only 156 players with ten or more goals scored, which reduced the quality of the training.

Table 15. Predicted vs true values

	goals_total	predicted_goals_total
0	2.0	2.411589
1	0.0	4.567359
2	1.0	1.291346
3	0.0	0.418599
4	0.0	0.532736
...	...	...
599	0.0	-0.348301
600	0.0	-0.523725
601	1.0	-0.077437
602	7.0	6.458196
603	14.0	13.145177

It is also interesting to see regarding position and league what is the difference in the RMSE obtained. Is the model predicting the goals of a player in a certain position or league more accurately than in the others?

On the table below (Table 16) it can be concluded that the test RMSE is substantially lower for midfielders than for attackers. As such, the model predicts goals of midfielders with greater precision. This can be the case as there is more data of midfielders in this dataset, more specifically 1232 midfielders in the trainset versus 792 attackers.

Table 16. Test RMSE of Stochastic model for attackers and midfielders

Model	Attacker - goals total - RMSE	Midfielder - goals total - RMSE
Stochastic Gradient Descent	1.4056896179657066	1.3178132462234786

Doing the same analysis for each league, the respective RMSE can be observed below (Table 17):

Table 17. Test RMSE for Stochastic model in different leagues

League name	League - goals total - RMSE
Primeira Liga	0.9540009192930556
Eredivisie	1.0796268812318428
Serie A	1.093492454195121
Premier Liga	1.1185773633776779
La Liga	1.192448346871633
Bundesliga	1.21255490174483
Premier League	1.321563685755364
Ligue 1	1.3865841643065338
Super Lig	1.44778582738339
First Division A	2.182335460100086

This model clearly predicts better players from the "Primeira Liga" versus all the other leagues. In fact, the model poorly predicts players from First Division A, as the RMSE is above two. One reason that may lead to this, may be related to the fact that this league had less accurate or trusted data.

6. Results - Goals Assists

6.1. Features Analysis

For the goals assists target variable, the Table 18 presents the top five features given by each algorithm:

Table 18. Top five features selected from feature selection methods

Univariate Selection	**Score**
passes_key	0,546
goals_assists_team	0,508
shots_total	0,377
dribbles_attempts	0,369
shots_on	0,351
Correlation Matrix with Heatmap	
passes_key	0,820
shots_total	0,690
shots_on	0,670
dribbles_attempts	0,670
dribbles_success	0,660
Ridge Regression	
passes_key	1,544
games_minutes_played	0,869
club_total_goals	0,728
shots_on	0,275
dribbles_attempts	0,261

Lasso Regression	**Score**
passes_key	1,467
dribbles_attempts	0,169
shots_on	0,167
club_total_goals	0,165
rating	0,096
Random Forest Regressor	
passes_key	0,676
rating	0,031
club_total_goals	0,022
goals_total	0,011
dribbles_attempts	0,010
Extra Trees Regressor	
passes_key	0,368
shots_total	0,068
rating	0,067
shots_on	0,056
dribbles_attempts	0,042

Recursive Feature Elimination	**Score**
shots_on	N/A
games_minutes_played	
games_appearences	
games_lineups	
First Division A	
dribbles_attempts	
dribbles_success	
rating	
passes_key	
passes_accuracy	
duels_won	
duels_total	
fouls_drawn	
fouls_committed	
tackles_interceptions	
tackles_total	
club_mv	
market_value_team	
twitter_avglikes	
twitter_avgretweets	

It is relevant to point out that it was not possible to extract the scores from the Recursive Feature Elimination method, so all the top 20 features selected were presented to facilitate comparisons. One very important point to notice is that *passes_key*, is agreeably the feature considered the most important when predicting goals assists, as it was chosen as the top one feature by all the methods. This intuitively makes all the sense, as the number of important passes that players make, is for sure related to how many assists they perform.

At the same time, *shots_total* and *dribbles_attempts* appeared also to be very important when predicting goals assists, being in the top five on the majority of the methods. Just like for goals total, here team variables are also not present in the top five of features chosen, however, in the chosen 20 features from the Recursive Feature Elimination technique, one team variables was displayed: *market_value_team*. This may be an indication pointing to the fact that this variable can potentially influence player assists.

These feature sets were used to create seven distinct datasets, with the top 20 chosen features adding also all the team variables, due to the problem studied by this dissertation.

6.2. Models

Just like in goals total prediction, a baseline model was created, by assigning the median of assists to all the observations' predictions. This is a very simplistic approach, so the RMSE achieved was a substantial high one, of 2.44. This means that the predictions made by this model were off by 2.44 assists on average. Curiously, the RMSE of the goals total baseline model was almost the double, one of 4.62 versus the one obtained here for the goals assist target variable.

When applying a 10-fold cross-validation to the same previous twelve models (Table 19) with their default parameters, the mean RMSE scores were the following:

Table 19. Mean RMSE scores for each combination of dataset and model

	All features	Univariate feature selection	Feature Importance	Correlation Matrix	Recursive Feature Elimination	Ridge feature selection	Random Forest feature selection	Lasso feature selection	Team features
Decision Tree Regressor	1.79126	1.8147	1.77416	1.8292	1.80944	1.75882	1.75858	1.77623	2.22524
Linear Regression	1.29378	1.28334	1.26678	1.2837	1.29726	1.28064	1.26114	1.25355	2.31104
Random Forest Regressor	1.29683	1.30971	1.28587	1.31205	1.30546	1.29463	1.28388	1.28264	1.73639
SVR	1.55147	1.43231	1.44854	1.47139	1.48443	1.51122	1.45607	1.41901	2.42909
Gradient Boosting Regressor	1.2879	1.30889	1.28032	1.31649	1.31354	1.28993	1.28401	1.27325	2.1686
Ridge Regression	1.28668	1.28287	1.26643	1.28314	1.29652	1.28053	1.26101	1.25352	2.31101
Stochastic Gradient Descent (elastic net)	1.28659	1.28294	1.26818	1.28272	1.29425	1.28458	1.26356	1.25679	2.31207
KNN	1.656	1.50452	1.51164	1.53713	1.53584	1.56098	1.51434	1.55709	2.11541
Lasso Regression	1.7031	1.7031	1.7031	1.7031	1.7031	1.7031	1.7031	1.7031	2.39599
Polynomial Regression	5.94736	1.7835	1.72563	1.836	6.62922	2.70706	1.94098	1.25341	2.31555
Polynomial corrected with Ridge	4.6977	1.59067	1.58665	1.64189	2.33328	2.54515	1.8318	1.26956	2.31083
Polynomial corrected with Lasso	1.76341	1.75059	1.75161	1.75443	1.75089	1.75357	1.75075	1.7503	2.37519

One of the first things to notice is that just like for the goals total, the ML models applied to a dataset with solely team variables to predict assists, performed poorly overall, achieving higher values of errors. After looking at the scores (Table 19), it is noticeable that the best results are achieved with the subset that has the features chosen by the Lasso algorithm in the feature selection phase. More specifically, by using this dataset, four models achieved the lowest RMSE:

- Linear Regression - 1.25355
- Ridge Regression - 1.25352

- Stochastic Gradient Descent Regressor - 1.25679
- Polynomial Regression degree two - 1.25341

It is important to refer that these values are considerable smaller when compared to the ones obtained when the target variable was goals total. They are not only smaller, but the difference between these errors and the baseline RMSE is smaller. The same process of tuning the hyperparameters for these four best models was pursued for goals assists, as an attempt to try and improve the errors.

6.3. Hyperparameter Tuning

6.3.1. Linear Regression

The default parameters of the linear regression were the ones delivering the best errors, one of 1.2461 in the training and 1.2542 in the validation. The final model is therefore the following (Figure 16):

```
LinearRegression(copy_X=True,fit_intercept=True,
                 normalize=False, n_jobs=-1)
```

Figure 16. Linear Regression best model

When it comes to the learning curve (Figure 17), below 200 observations the overfitting was a major handicap, yet, as the observations increased, both curves seem to converge leading to the considerable improvement when it comes to the problem of overfitting.

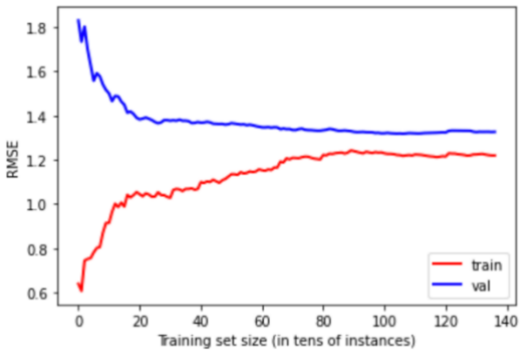

Figure 17. Linear Regression learning curve

6.3.2. Polynomial Regression

For the Polynomial, the parameters tuned were the same ones of the Linear model. However, in this case, the best parameters were (Figure 18):

```
LinearRegression(copy_X=True,fit_intercept=True,
                 normalize=True, n_jobs=-1)
```

Figure 18. Polynomial best model

This model achieved, therefore, a training RMSE of 0.657 and validation RMSE of 1.255, a significant difference that can be explained by overfitting being present.

6.3.3. Ridge Regression

For this model three parameters were tuned: *alpha*, *max_iter* and *solver*. Some values of alpha between 2.5 and 50 were experimented with and every *solver* value was used in the tuning, while the maximum number of interactions started in 1000 until 4000 from thousand to thousand, including the default value *None*. After grid search, the best model found (Figure 19) achieved a validation RMSE of 1.2542 with the following parameters:

```
Ridge(alpha=3.0, copy_X=True,fit_intercept=True,
      max_iter=None, normalize=False, solver='svd',
      tol=0.001, random_state=7)
```

Figure 19. Ridge Regression best model

The learning curve (Figure 20) shows that below the 200 observations there is a clear overfitting with validation set fitting the data poorly. From the 800 observations onwards, overfitting is almost eliminated, and the training and validation curves converge significantly. However, there is still a difference between the curves leading to a small overfitting.

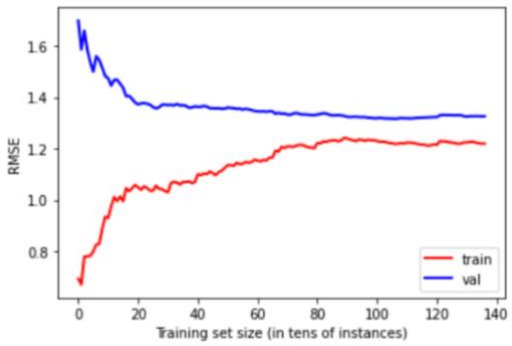

Figure 20. Ridge Regression learning curve

6.3.4. Stochastic Gradient Descent Regressor

For this model the same step-by-step tuning process done for goals total was done for this second target variable. After following this process, the parameters achieving the best RMSE, one of 1.2528 were (Figure 21):

```
SGDRegressor(learning_rate = 'adaptive', eta0 = 0.0069,
             loss= 'squared loss', penalty = 'l1',
             power_t=0.25, max_iter=1000,
             n_iter_no_change=4, early_stopping=True,
             random_state=7)
```

Figure 21. Stochastic Gradient Descent best model

This represents an improvement, however a small one, versus the model with the default parameters. Regarding the learning curve below (Figure 22), it can be observed that from the 800 instances onwards the validation curve decreases its error and both curves seem to converge and plateau. There is a still a difference between them, thus a small overfitting seems to exist.

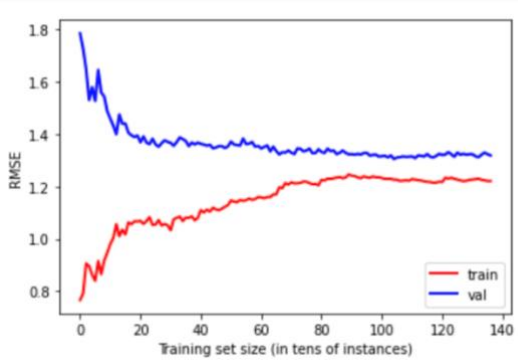

Figure 22. Stochastic Gradient Descent Regressor learning curve

6.4. Comparison of final models

All in all, the Table 20 provides an overview of the RMSE of train and RMSE of validation of the four best ML models, ordered by the best validation RMSE to the worst, after the hyperparameter tuning was applied:

Table 20. Comparison of best models for goals assists

Model	RMSE Train	RMSE Validation	Used dataset
Stochastic Gradient Descent	1.2490279756766778	1.2528120667823979	Lasso feature selection
Ridge	1.2462662633411314	1.2541949773468433	Lasso feature selection
Linear Regression	1.2461560942003476	1.254260373262732	Lasso feature selection
Polynomial degree 2	0.6570804372239855	1.2552840084955317	Lasso feature selection
Baseline	2.445351320976999	N/A	All features

It is observed that all the models executed performed significantly better than the baseline one. With this target variable, Stochastic Gradient Descent was the one presenting the lowest

validation RMSE, however overall, all the four models present very similar performances. Like it was referred, these errors are significantly smaller than when predicting goals total. Further analysis will be pursued for the first three models to obtain the responses for the research questions posed at the start of this thesis. For the Polynomial model, its features will not be analysed as a 2nd degree polynomial features were created, which increased the number of variables even more, adding to the fact that the naming is in a specific format (x1, $x1^2$, etc), which would complicate their interpretation.

7. Discussion - Goals Assists

7.1. RQ1: Most suitable learning algorithms

When evaluating these three described models, plus the polynomial one on the test set, the results were the following (Table 21):

Table 21. Evaluation of four best models on the test set

Model	RMSE Train	RMSE Validation	RMSE Test	Used dataset
Ridge	1.2462662633411314	1.2541949773468433	1.0544931498398662	Lasso feature selection
Linear Regression	1.2461560942003476	1.254260373262732	1.059740826516169	Lasso feature selection
Stochastic Gradient Descent	1.2490279756766778	1.2528120667823979	1.0802243559757763	Lasso feature selection
Polynomial degree 2	0.6570804372239855	1.2552840084955317	31.81720189584637	Lasso feature selection

Ridge Regression achieved the lowest RMSE, one of 1.054, whereas the Polynomial was the model with the highest RMSE, one of 31.82. This is a very big difference versus all the other errors, indicating that this model overfitted the data by a large amount when applied to a new dataset. From the other three models, the RMSE is significantly better than the ones obtained on the validation set, as such it can be concluded that these models perform relatively good when presented with a new dataset. As such, any of these learning algorithms seem to be suitable for the problem of predicting goals assists, which answers the first of the research questions. Being Ridge the one that achieved the lowest RMSE values, and since all models reached a consensus in terms of team variables, the Ridge Regression was the final model chosen to be pursued by this problem.

7.2. RQ2: Most Suitable variables for this problem

The Linear Model identified *passes_key* as being by far the most important variable, followed by *club_total_goals*, *goals_assists_team*, *games_minutes_played* and *passes_accuracy_midfielder_team* (Table 22). When looking at the p-values for all these variables, they are all zero, which means that these variables are all statistically significant

(Figure 23). The exception is *games_minutes_played,* which presents a higher than 5% p-value, thus being considered as not having statistically significant impact on goals assists.

Table 22. Feature importance for Linear Regression model

Feature	Importance
passes_key	4.62816390904738
club_total_goals	0.8779250994334591
goals_assists_team	0.5398981805181419
games_minutes_played	0.27955192443142207
passes_accuracy_midfielder_team	0.17772517537034727
market_value_opponents	0.17503343893428636
dribbles_attempts	0.11133292872448675
duels_won_defender_team	0.10318333083705447
games_lineups	0.09610845213238492
club_match_played	0.08000197908723179
tackles_total_defender_team	0.07765820919109574
club_match_lose	0.06753579454131775
market_value_team	0.06471271858479355
rating	0.05357263754833952
tackles_total	0.051115952033464106
shots_on	0.04814420147168756
Serie A	0.03737738923829277
goals_total_attacker_team	0.028896829014991864
goals_total_midfielder_team	0.026127001217374945
market_value	0.0228896828407384
club_goals_conceded	0.019575843216954804
goals_total	0.01149054241028431
games_appearences	0.010772087217350568
age_team	0.008250440370970846
age	0.0017627614310410378
captain	0.0016254517633070397
club_match_draw	0.00124812189053789
passes_accuracy	0.0010980105606581913
duels_won	0.0005750427684005022
goals_conceded_goalkeeper_team	4.577396360732067e-05
diversity_team	-0.0001561230876724550

OLS Regression Results

Dep. Variable:	goals_assists	R-squared:	0.724
Model:	OLS	Adj. R-squared:	0.719
Method:	Least Squares	F-statistic:	168.4
Date:	Sat, 21 Aug 2021	Prob (F-statistic):	0.00
Time:	15:20:01	Log-Likelihood:	-3313.7
No. Observations:	2024	AIC:	6691.
Df Residuals:	1992	BIC:	6871.
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6151	0.028	57.950	0.000	1.560	1.670
passes_key	1.5237	0.060	25.270	0.000	1.405	1.642
dribbles_attempts	0.2359	0.050	4.699	0.000	0.137	0.334
shots_on	0.1573	0.073	2.152	0.032	0.014	0.301
club_total_goals	0.6731	0.109	6.156	0.000	0.459	0.888
rating	0.1655	0.047	3.503	0.000	0.073	0.258
goals_total	0.0778	0.065	1.204	0.229	-0.049	0.205
Serie A	-0.1326	0.033	-4.015	0.000	-0.197	-0.068
age	-0.0220	0.033	-0.672	0.502	-0.086	0.042
captain	-0.0289	0.031	-0.933	0.351	-0.090	0.032
club_goals_conceded	-0.0977	0.075	-1.303	0.193	-0.245	0.049
club_match_draw	0.0216	0.051	0.423	0.672	-0.079	0.122
club_match_lose	0.1822	0.116	1.566	0.118	-0.046	0.410
club_match_played	-0.1979	0.069	-2.887	0.004	-0.332	-0.063
duels_won	-0.0145	0.073	-0.197	0.844	-0.158	0.130
games_appearences	-0.0723	0.090	-0.806	0.421	-0.248	0.104
games_lineups	-0.2166	0.305	-0.709	0.478	-0.816	0.382
games_minutes_played	0.3728	0.354	1.054	0.292	-0.321	1.066
market_value	-0.1053	0.050	-2.123	0.034	-0.203	-0.008
passes_accuracy	-0.0195	0.055	-0.354	0.723	-0.128	0.089
tackles_total	-0.1586	0.057	-2.798	0.005	-0.270	-0.047
goals_conceded_goalkeeper_team	0.0125	0.039	0.319	0.750	-0.064	0.089
duels_won_defender_team	0.2312	0.084	2.745	0.006	0.066	0.396
tackles_total_defender_team	-0.1912	0.085	-2.246	0.025	-0.358	-0.024
market_value_opponents	0.3006	0.079	3.790	0.000	0.145	0.456
diversity_team	0.0088	0.032	0.273	0.785	-0.055	0.072
goals_total_midfielder_team	0.1177	0.041	2.854	0.004	0.037	0.199
passes_accuracy_midfielder_team	0.3047	0.076	3.997	0.000	0.155	0.454
goals_total_attacker_team	0.1261	0.052	2.419	0.016	0.024	0.228
goals_assists_team	-0.5130	0.081	-6.319	0.000	-0.672	-0.354
age_team	0.0649	0.035	1.871	0.061	-0.003	0.133
market_value_team	-0.1758	0.077	-2.274	0.023	-0.327	-0.024

Omnibus:	274.874	Durbin-Watson:	1.934
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1031.876
Skew:	0.632	Prob(JB):	8.53e-225
Kurtosis:	6.262	Cond. No.	47.5

Figure 23. Feature significances for Linear Regression model

Additionally, just like for goals total, the linear assumptions were checked to see how adequate this linear model was to model the data. Firstly, the scatterplot of the standardized residuals can be seen below (Figure 24):

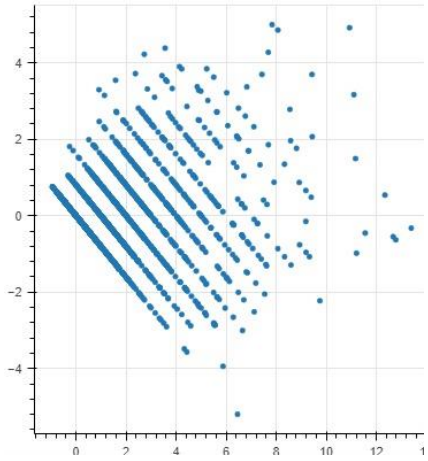

Figure 24. Scatterplot of standardized residuals for goals assists

The residuals are randomly scattered around zero, and there is no apparent pattern to the residuals, which suggests that a linear model is appropriate for this relationship. Further, the histogram below (Figure 25) suggests the errors are approximately normally distributed.

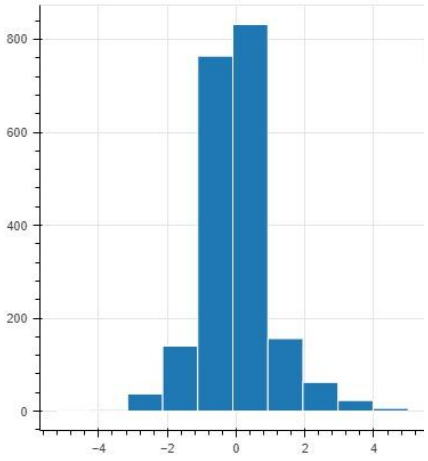

Figure 25. Histogram of residuals for goals assists

Furthermore, the Durbin-Watson tests (Figure 23) if the values of the residuals are independent of each other, and for this assumption to be met, the value needs to be as close to two as possible. In this case the value is 1.934 and therefore this assumption will be taken as being met. Moreover, the assumption of independence of the errors and homoscedasticity are seldom violated in data where there is no temporal dimension, which is the case with this one, so this will also be assumed as met.

For the Ridge Regression it is observable in Table 23, that *passes_key* is by far the variable that seems to be more useful in predicting goals assists. At the same time *clubs_total_goals*, *goals_assists_team*, *passes_accuracy_midfielder_team* and *market_value_opponents* are also

important variables for this prediction. Shifting the attention to the coefficients of these variables, four out of the five described above have positive coefficients (Table 24). However, *goals_assists_team* has a negative coefficient, one of -0.4970. If this coefficient is statistically significant, it can be said that this variable has a negative effect on goals assists, meaning that an increase in the number of the assists of the teammates will lead to a decrease in the goals assists of a player. This intuitively can make sense as the more assists players in a team make, the less opportunities there are for a specific player to make goals assists.

Table 23. Feature importance for Ridge Regression model

Feature	Importance
passes_key	4.580485675035694
club_total_goals	0.806501759607686
goals_assists_team	0.5081955873159777
passes_accuracy_midfielder_team	0.166954148222256155
market_value_opponents	0.16159589422810078
games_minutes_played	0.1439665775839684
dribbles_attempts	0.11412462093923222
duels_won_defender_team	0.09496154923305364
club_match_played	0.07132006794956078
tackles_total_defender_team	0.07038648194731587
market_value_team	0.06005528213917335
rating	0.05597034639900935
club_match_lose	0.05308957570928512
tackles_total	0.05022217182149773
shots_on	0.049940768533477266
Serie A	0.036958435631178774
games_lineups	0.03252634337505891
goals_total_attacker_team	0.028958860877608173
goals_total_midfielder_team	0.024696647231635293
market_value	0.02278925359623383
club_goals_conceded	0.014698413559305656
goals_total	0.011533663354521394
age_team	0.008027301085438254
games_appearences	0.006972118916003423
age	0.0018309110632524605
captain	0.0015938887490790199
passes_accuracy	0.0009981299964427759
club_match_draw	0.0006600316015539454
duels_won	0.0004913383656366177
goals_conceded_goalkeeper_team	4.1644410227203996e-05
diversity_team	-0.00015422320241211817

Table 24. Coefficients for Ridge Regression model

Feature	Coefficient
passes_key	1.5145538830161327
club_total_goals	0.644902289786222
passes_accuracy_midfielder_team	0.29528946349064467
market_value_opponents	0.2887635025817158
games_minutes_played	0.2676685572377386
dribbles_attempts	0.23868317402971614
duels_won_defender_team	0.22179665370446272
rating	0.1690066994467881
club_match_lose	0.16120033736277148
shots_on	0.16006346686095466
goals_total_attacker_team	0.12617726371233295
goals_total_midfielder_team	0.11450585335744103
goals_total	0.07789083338858532
age_team	0.06383797805847059
club_match_draw	0.014853436697435357
goals_conceded_goalkeeper_team	0.012329111400963226
diversity_team	0.008699113570890154
duels_won	-0.01321086947817423
passes_accuracy	-0.01842268901019717
age	-0.022495809306734636
captain	-0.028576216840962093
games_appearences	-0.057919655676825416
club_goals_conceded	-0.08435735169723907
market_value	-0.10500836125833717
games_lineups	-0.12528971930261343
Serie A	-0.1317227112770281
tackles_total	-0.15701036561042284
market_value_team	-0.1690306218919348
tackles_total_defender_team	-0.18159007521462106
club_match_played	-0.18654632969459978
goals_assists_team	-0.496997191698575

The Stochastic Gradient Descent model found that *passes_key* is, as well, the most important variable to predict the target one with a significant leverage versus all the others (Table 25). The other four most important variables are the same as the one from the Ridge. The coefficients (Table 26) are all positive except the *goals_assits_team* one which is negative, leading to a possible negative effect on the target variable, just like it happened on the preceding model.

Table 25. Feature importance for Stochastic model

Feature	Importance
passes_key	4.395374377242435
club_total_goals	0.6914134204278912
goals_assists_team	0.41489572487823967
market_value_opponents	0.26379575023336754
passes_accuracy_midfielder_team	0.18930709738618914
duels_won_defender_team	0.1348995726726775
dribbles_attempts	0.1347564721938726
tackles_total_defender_team	0.1292986450180604
club_match_played	0.08342239329326115
market_value_team	0.07776865234803279
rating	0.07205147222753654
shots_on	0.05833242087289807
tackles_total	0.04993895950344014
Serie A	0.03576642806459949
market_value	0.03044176298825376
games_minutes_played	0.021262743895710125
goals_total_midfielder_team	0.020436861058534306
goals_total_attacker_team	0.018594651230153934
goals_total	0.012444794789436142
club_match_lose	0.011142035469334921
age_team	0.005970635798196788
age	0.004591521458983094
club_goals_conceded	0.0018684122487413247
games_lineups	0.0017875624691226165
captain	0.0012230374809437006
games_appearences	0.0005558523125806847
diversity_team	0.0003974403323427023
passes_accuracy	0.00023350057193536334
club_match_draw	1.2053617602170163e-05
duels_won	-8.381302210702657e-06
goals_conceded_goalkeeper_team	-7.27550710119651e-05

Table 26. Coefficients for Stochastic model

Feature	Coefficient
passes_key	1.47244726392166
club_total_goals	0.5975760742266452
market_value_opponents	0.36196816364875434
passes_accuracy_midfielder_team	0.31204566390776556
dribbles_attempts	0.25849536106448695
duels_won_defender_team	0.2561911708681147
rating	0.18621394678982614
shots_on	0.17230463569960866
goals_total_midfielder_team	0.10584518414072892
goals_total_attacker_team	0.09775733850778053
games_minutes_played	0.09400190302069203
goals_total	0.08198568837509683
club_match_lose	0.07210478643782772
age_team	0.04966655307125735
diversity_team	0.025640146201045988
games_lineups	0.022899163710023
goals_conceded_goalkeeper_team	0.0037058114271539275
club_match_draw	-0.007397558762960387
duels_won	-0.007667219135719191
passes_accuracy	-0.01029578233601389
club_goals_conceded	-0.023855401246339673
games_appearences	-0.026719257045466082
captain	-0.04008714038003636
age	-0.05798702765143281
market_value	-0.1311095162114254
Serie A	-0.13213496152465098
tackles_total	-0.16889388931805208
market_value_team	-0.19992646366045624
club_match_played	-0.21012949533911573
tackles_total_defender_team	-0.2592333182821168
goals_assists_team	-0.4471842423068604

After all these analyses, when shifting the attention to the RQ2, *passes_key, clubs_total_goals*, *goals_assists_team*, *passes_accuracy_midfielder_team* and *market_value_opponents* were agreeably, the top five variables with the highest importance when predicting goals assists.

7.3. RQ3: Contribution of team-related variables

Regarding the team-related variables found to be the most relevant by the Multiple Linear Regression, *goals_assists_team*, *passes_accuracy_midfielder_team*, *market_value_opponents* and *duels_won_defender_team* were found to be the most important (Table 22), and when looking at the p-values, it is observable that all of those have a statistically significant impact on goals assists (Figure 23). Additionally, according to these p-values, the only team-related variables who do not have an impact on assists are *goals_conceded_goalkeeper_team*, *diversity_team* and *age_team.* However, it is important to reinforce that this last one becomes significant if the significance level used was 10% instead of 5%. In this case, the adjusted R-squared takes the value of 0.719, which means that 71.9% of the total variability of goals assists is explained by this model, corresponding to a medium explanatory power, and worse than the one obtained in the model predicting goals total. Just like in the goals total case, the *market_value_team,* surprisingly, is not amongst the top five team variables with higher importance when predicting goals assists.

For the Ridge model, the most important team variables were the same as above: *goals_assists_team*, *passes_accuracy_midfielder_team*, *market_value_opponents* and *duels_won_defender_team* (Table 23). Just like talked in the previous section, *goals_assists_team* is the only with a negative effect on goals assists (Table 24).

Further, the significance of these four different team-related variables was tested. For all these variables the null of the two-sided test was rejected, which means that the mean of the predictions with and without each of these variables was different, thus these team-related variables do have an impact on the goals assists predictions (Table 27).

Table 27. Paired Sample T-Test for Ridge Regression

	Ttest_relResult	
	statistic	p-value
Excluding passes_accuracy_midfielder_team feature	-10.04768	4.569286e-22
Excluding market_value_opponents feature	-25.41246	2.127787e-97
Excluding goals_assists_team feature	-6.695144	4.9383209e-11
Excluding duels_won_defender_team feature	-12.48743	5.4491480e-32

For the Stochastic model the same exact conclusions of the Ridge can be drawn, not only because the most important team-related variables were the same (Table 25 and Table 26), but the significances of the paired sample T-Test were also the same (Table 28).

Table 28. Paired Sample T-Test for Stochastic Gradient Descent Regressor

	Ttest_relResult	
	statistic	p-value
Excluding market_value_opponents feature	0.10828224	1.44074e-98
Excluding passes_accuracy_midfielder_team feature	3.00001352	9.59167e-24
Excluding goals_assists_team feature	5.33380958	1.06488e-08
Excluding duels_won_defender_team feature	-25.7899169	6.00228e-31

After all these analyses, when shifting the attention to the RQ3, *goals_assists_team*, *passes_accuracy_midfielder_team*, *market_value_opponents* and *duels_won_defender_team* were the variables found to have an impact on the goals assists predictions in Ridge, Stochastic and Linear models.

7.4. Ridge Regression further evaluation

Since Ridge Regression was the model that obtained the lowest RMSE, some further analysis was done. The Table 29 exemplifies some predictions made, where it can be seen that the predictions are actually very close to the original values. Out of these ten observations, five of them, if the predictions are rounded, would be correctly predicted. The worst overshoot here was in the first line in which a player did four goals assists, but the model predicted six goals assists. Overall, the predictions were quite assertive.

Table 29. Predicted vs true values

	goals_assists	predicted_goals_assists
589	4.0	5.918686
590	0.0	-0.277829
591	0.0	1.617415
592	5.0	5.187862
593	1.0	0.292789
594	0.0	1.173905
595	1.0	0.432389
596	1.0	0.664008
597	1.0	0.570772
598	0.0	-0.279590

It is also interesting to see regarding position and league what is the difference in the RMSE obtained. Is this model predicting the goals assists of a player in a certain position or league more accurately than in the others?

On Table 30, it can be concluded that the test RMSE is substantially lower for attackers than for midfielders, as such the model predicts goals assists of attackers with greater precision. This is the exact opposite of what was happening in the goals total case.

Table 30. Test RMSE of Ridge Regression for attackers and midfielders

Model	Attacker - goals assists - RMSE	Midfielder - goals assists - RMSE
Ridge Regression	0.9868727324845031	1.0985484692646497

Doing the same analysis for each league, the respective RMSE can be observed below (Table 31):

Table 31. Test RMSE for Ridge Regression in different leagues

League name	League - goals assists - RMSE
First Division A	0.4811538269963663
Primeira Liga	0.8780024675036872
Bundesliga	0.9591576508191142
Premier Liga	1.0390223019570874
La Liga	1.0515320506557433
Super Lig	1.0731256566862086
Serie A	1.0991050557466082
Eredivisie	1.1064730777717242
Ligue 1	1.2928694421945894
Premier League	1.3838445747951638

This model clearly predicts with less error players from the "First Division A" versus all the other leagues, which again poses the exact opposite of what was happening on the goals total prediction. This model predicts with the worst accuracy players from "Premier League", as the RMSE is 1.3838.

8. Final remarks

Overall, the Ridge Regression model for goals assists applied to the test set, achieved a lower RMSE than any model for goals total. Not only that, but for goals assists four team-related variables were found to have an impact on predictions, across all models, so this consensus can be a positive point in the reliability of these results. Moreover, for goals assists, there were more team-related variables in the top important ones for predicting the target variable, thus it is worth pointing out that there seemed to be a tendency for team effects to be more important for assists rather than for goals. This contradicts what was found by Torgler and Schmidt (2011).

9. Conclusion

Altogether, this dissertation touched upon the most important features and most suitable algorithms to predict two target variables: goals total and goals assists of individual football players in a series of European leagues. It was found that Gradient Boosting Regressor, Stochastic Gradient Descent Regressor, Ridge Regression and Linear Regression were the most prominent models to predict goals total, whether for goals assists, the ML algorithms were the same last three plus a Polynomial Regression with degree two. In order to predict goals total it

was found that *shots_on, penalty_success, games_minutes_played, club_total_goals*, *passes_total* and *games_lineups* were found to be the most important variables, whereas *passes_key, clubs_total_goals*, *goals_assists_team*, *passes_accuracy_midfielder_team* and *market_value_opponents* were the most important variables when predicting goals assists. Interestingly, these last three variables are all team-related ones and all statistically significant. Thus, as the midfielder teammates of a player do more accurate passes and as players in the same league have higher market value, a player's goals assists increase. At the same time, as the average of assists of a team increases, the player performs less goals assists, which intuitively makes sense. More high achiever players in the team, decrease the performance of a player. Also, *duels_won_defender_team* was found to be statistically significant with positive coefficient, so when the defenders won more duels, the player increases its goals assists.

At the same time *goals_assists_team*, *goals_total_midfielder_team* and *market_value_opponents* were found to be statistically significant when predicting goals total. Moreover, *goals_total_midfielder_team* is the one with negative coefficient, meaning that more goals scored by midfielders' teammates of a player, the less goals the player scores, as his teammates represent more competition.

Further, one interest point to make, is that the market value of teammates has a very low importance across all models in both these predictions. It was also discovered that for goals total, Gradient Boosting Regressor achieved the lowest RMSE, one of 1.2822, and even though the Stochastic Gradient Descent had a RMSE of 1.3543, it was chosen as the most suitable model to predict goals total. Not only because it had significantly less overfitting, but also because the team-related variables had higher importance and higher accordance with other models. On the other hand, the Ridge Regression was the model most suitable to predict goals assists, with a RMSE of 1.054. Moreover, it was found that the model for goals total derived better predictions for midfielders, whereas attackers had better predictions of goals assists. Also, goals total was predicted with lower error for players in Primeira Liga, whereas goals assists was predicted with lower error for players in First Division A. Overall, the models for goals assists achieved a lower RMSE and there were, not only, more team variables with higher importance, but also statistically significant, for assists rather than for goals.

9.1. Limitations

Regarding the limitations of this thesis, the first one to be noticed is the number of instances the dataset had, which were only 4523 before the pre-processing steps. If more player data was available, the ML models would have more training data, which would potentially increase the

quality of the predictions, decreasing the RMSE. Evidence of this was seen in the learning curves, in which more training data reduced the overfitting. Another limitation to point out is the fact that player performance in some games could have been affected by external factors or recent incidents, which were not proper represented in the dataset used.

9.2. Future work

This dissertation showed that team features do exert a considerable amount of influence in predicting performance of football players. Thus, in the future, to try and improve these predictions, some more relevant predictors could be added, like if it was the 1st season of a player in a specific club, or not, or even with how many teammates of that club, had that player played with before. These could be indications of how used to the club and of his peers, a player is, which could influence his performance. Additionally, sentiment analysis could be used, applied to internet news and article, to derive pertinent features to be used as popularity indications to predict performance. It would be also interesting to understand if these conclusions could be extended to other sports like basketball, cricket, baseball among others.

Further, with a more powerful machine, deep and thorough analysis in hyperparameter tuning could be pursued, in an attempt to find even better combinations of parameters for the ML predictive models.

Lastly, this dissertation opened many doors for investigating how these conclusions can be applied for businesses and their respective workplaces. It would be interest to investigate whether teams with high performers lead to a healthy competition among team members or if dilutes individual performance. All in all, this dissertation showed that professional football sports studies can offer an insightful opportunity for further empirical investigation and can have valuable implications for businesses in general.

References

Sarlis, V., & Tjortjis, C. (2020). Sports analytics — Evaluation of basketball players and team performance. *Elsevier Ltd*.

McCabe, A., & Trevathan, J. (2008). Artificial Intelligence in Sports Prediction. *Fifth International Conference on Information Technology: New Generations.*

Anthony A, S., Jayalakshmi, D., & D, A. (2021). Player Stats Analysis Using Machine Learning. *International Journal of Innovative Science and Research Technology*.

Hughes, T., Sergeant, J., van der Windt, D., Riley, R., & Callaghan, M. (2018). Periodic Health Examination and Injury Prediction in Professional Football (Soccer): Theoretically, the Prognosis is Good. In *Sports medicine (Auckland, N.Z.)* (Vol. 48, p. 2443).

Pantzalis, V., & Tjortjis, C. (2020). Sports Analytics for Football League Table and Player Performance Prediction. *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA).*

Hucaljuk, J., & Rakipović, A. (2011). Predicting football scores using machine learning techniques. *2011 Proceedings of the 34th International Convention MIPRO.* Opatija, Croatia: IEEE.

Stanojevic, R., & Gyarmati, L. (2016). Towards data-driven football player assessment. *16th International Conference on Data Mining Workshops.* IEEE.

Pariath, R., Shah, S., Surve, A., & Mittal, J. (2018). Player Performance Prediction in Football Game. *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA).* Coimbatore, India: IEEE.

Bowen , D., Ledford Jr., G., & Nathan, B. (1996). Hiring for the Organization, Not the Job.

Hamilton, B., Nickerson, J., & Owan, H. (2003). Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation. *Journal of Political Economy, 111*, 465-497.

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research, 263*(2), 611-624.

Cooman, R., Matthijs Bal, P., Lub, X., & Vantilborgh, T. (2015). Creating Inclusive Teams Through Perceptions of Supplementary and Complementary Person– Team Fit: Examining the Relationship Between Person–Team Fit and Team Effectiveness. *Group & Organization Management*.

Myatt, D., & Wallace, C. (2008). When Does One Bad Apple Spoil the Barrel? An Evolutionary Analysis of Collective Action. In *The Review of Economic Studies* (Vol. 75, pp. 499–527).

Bransen, L., & Haaren, J. (2020). Player Chemistry: Striving for a Perfectly Balanced Soccer Team. *MIT Sloan Sports Analytics Conference.*

Lombardi, N. (2017). *Evaluating Player Performance and Team Efficiency in Major League Soccer.*

Shahriar, M., Islam, Y., & Amin, M. (2019). Player Classification Technique Based on Performance for a Soccer Team Using Machine Learning Algorithms. *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT).* IEEE.

Passi, K., & Pandey, N. (2018). INCREASED PREDICTION ACCURACY IN THE GAME OF CRICKET USING MACHINE LEARNING. *International Journal of Data Mining & Knowledge Management Process (IJDKP), 8*(2).

Saaty, R. (1987). The analytic hierarchy process—what it is and how it is used. *Mathematical Modelling, 9*(3), 161-176.

Apostolou, K., & Tjortjis, C. (2019). Sports Analytics algorithms for performance prediction. *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA).* Patras, Greece: IEEE.

Forrest, D., Simmons, R., & Szymanski , S. (2004). Broadcasting, Attendance and the Inefficiency of Cartels. *Review of Industrial Organization volume*, 243–265.

Idson, T., & Kahane, L. (2000). Team Effects on Compensation: An Application to Salary Determination in the National Hockey League. *Economic Inquiry, 38*, 345-57.

Torgler, B. (2007). Determinants of success at the Tour de France. *Journal of Sports Economics*.

Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digital Health*.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*.

Maulud, D., & Abdulazeez, A. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends, 1*(4), 140 – 147.

Rokem, A., & Kay, K. (2020). Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. *GigaScience, 9*(12).

Schmidt, S. L., & Torgler, B. (2011). What shapes player performance in soccer? Empirical findings from a panel analysis. In *Applied Economics* (Vol. 39).

Anik, A., Yeaser, S., Hossain, A., & Chakrabarty, A. (2018). Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms. *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT).*

Dizdari, H., & Seiler, R. (2020). Key players in sport teams. An exploratory study on the effects of attachment styles on intra-team relational networks. *Psychology of Sport & Exercise, 51*.

Brewer, M. (1991). The social self: On being the same and different at the same time. *Personality and Social Psychology Bulletin,*(17), 475–482.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

Appendices

Appendix 1. Histograms of Skewed predictors

1. Google Search

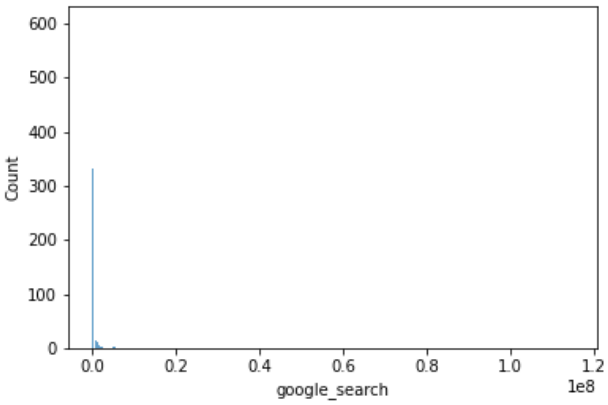

Figure 26. Google search histogram before Log

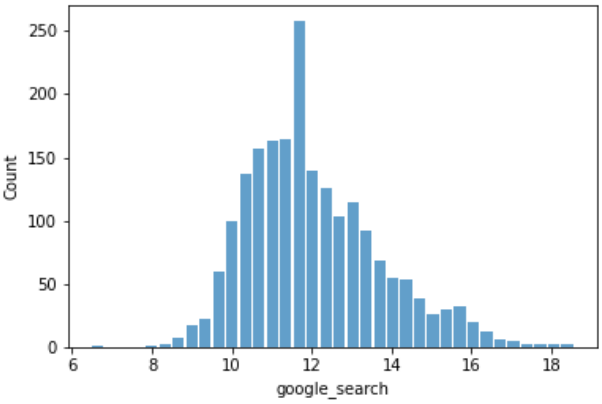

Figure 27. Google search histogram after Log

2. Club market value

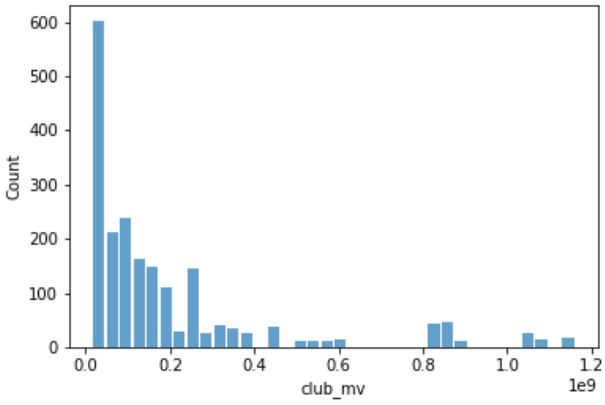

Figure 28. Club market value histogram before Log

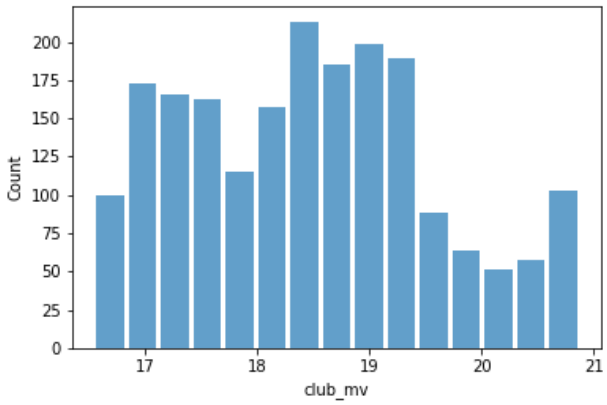

Figure 29. Club market value histogram after Log

3. Market value

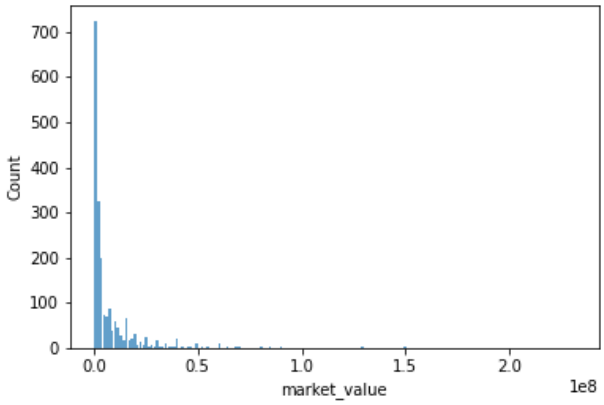

Figure 30. Market value histogram before Log

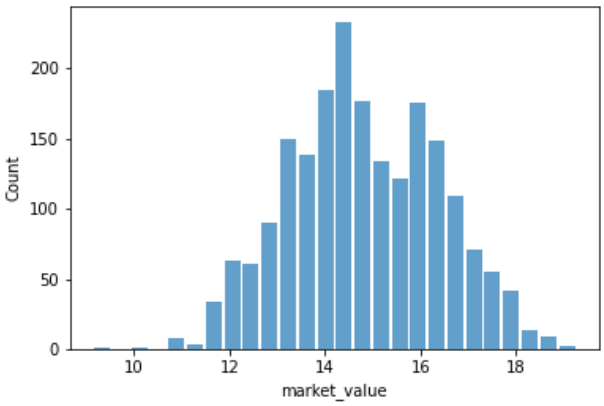

Figure 31. Market value histogram after Log

4. Market value of the team

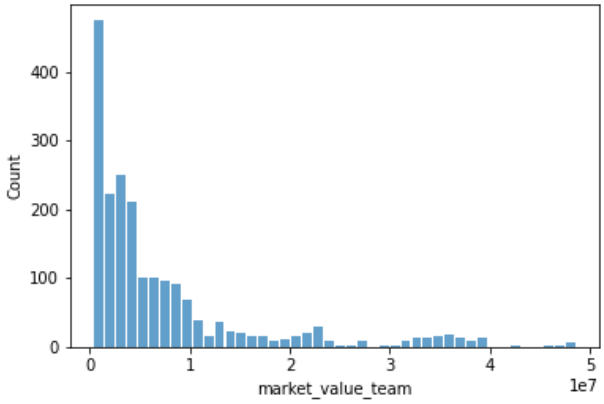

Figure 32. Market value team histogram before Log

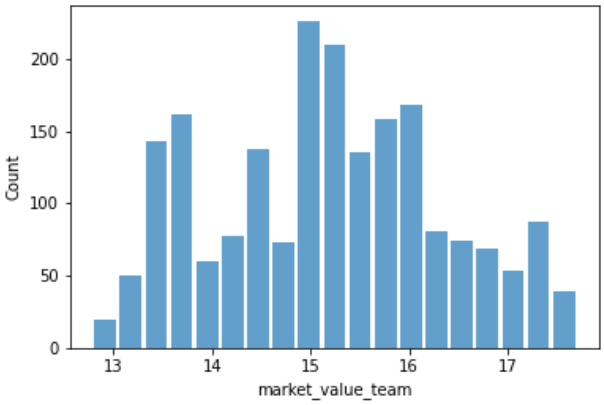

Figure 33. Market value team histogram after Log

Appendix 2. Goals total

Gradient Boosting Regressor

On Figure 34, it can be visualized the predictions made by the Gradient Boosting Regressor. This model was fitted on the first 90% of the training set and generate predictions on the last 10%. There were 2024 training instances, so 1822 were used for training and the rest for testing. Altogether it can be seen that the predictions seem to be quite assertive, except for a few cases where the prediction either overshot or undershot the true value by a significant amount.

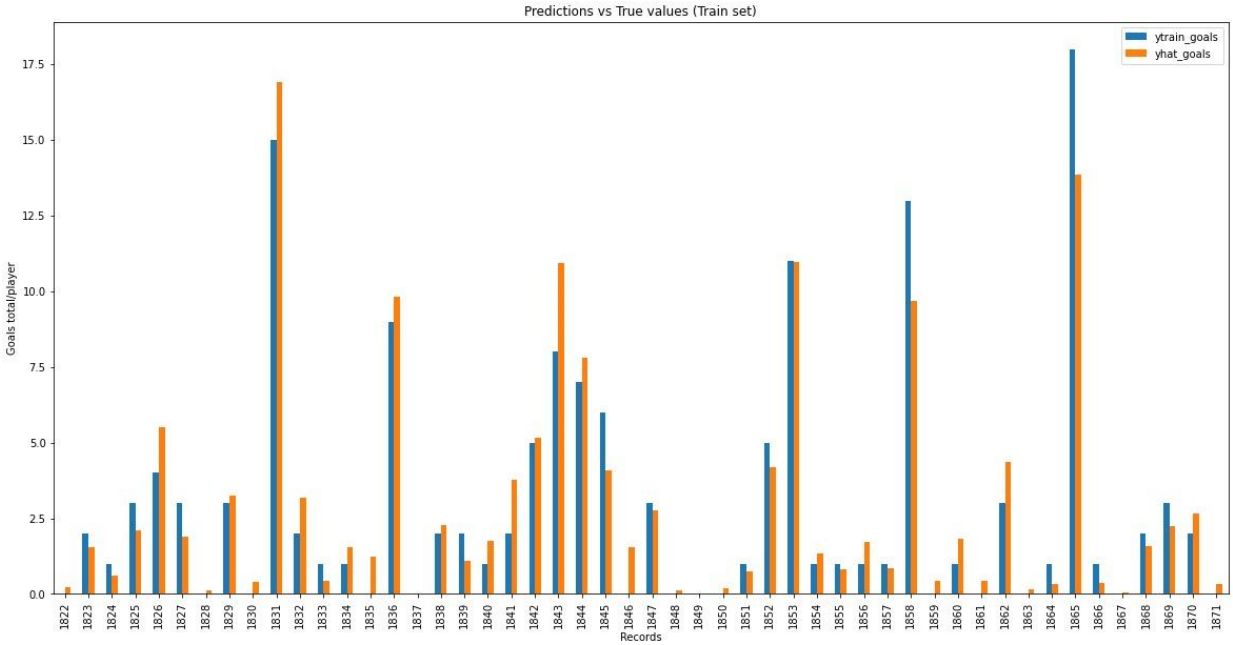

Figure 34. Predictions versus true values for Gradient Boosting

Ridge Regression

Just like in the Gradient Boosting, in Figure 35 the predictions made by this model can be observed. All in all, predictions are quite assertive, as they do not overshoot or undershoot the real values by a major amount. However, it can be seen some cases in which the goals total were zero and the model predicted a value different than zero. Some predictions were even negative, a normal event in cases like this one, in which the features were scaled in the pre-processing stage.

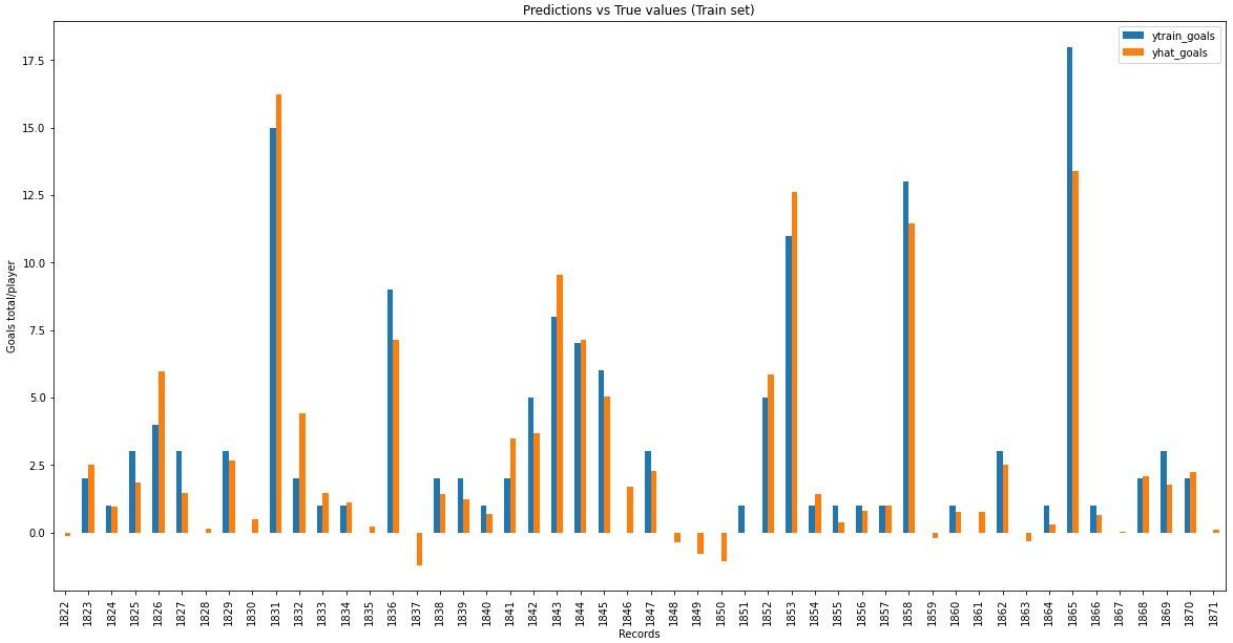

Figure 35. Predictions versus true values for Ridge

Stochastic Gradient Descent

This figure plotted very similar results to the one obtained in the Ridge Regression.

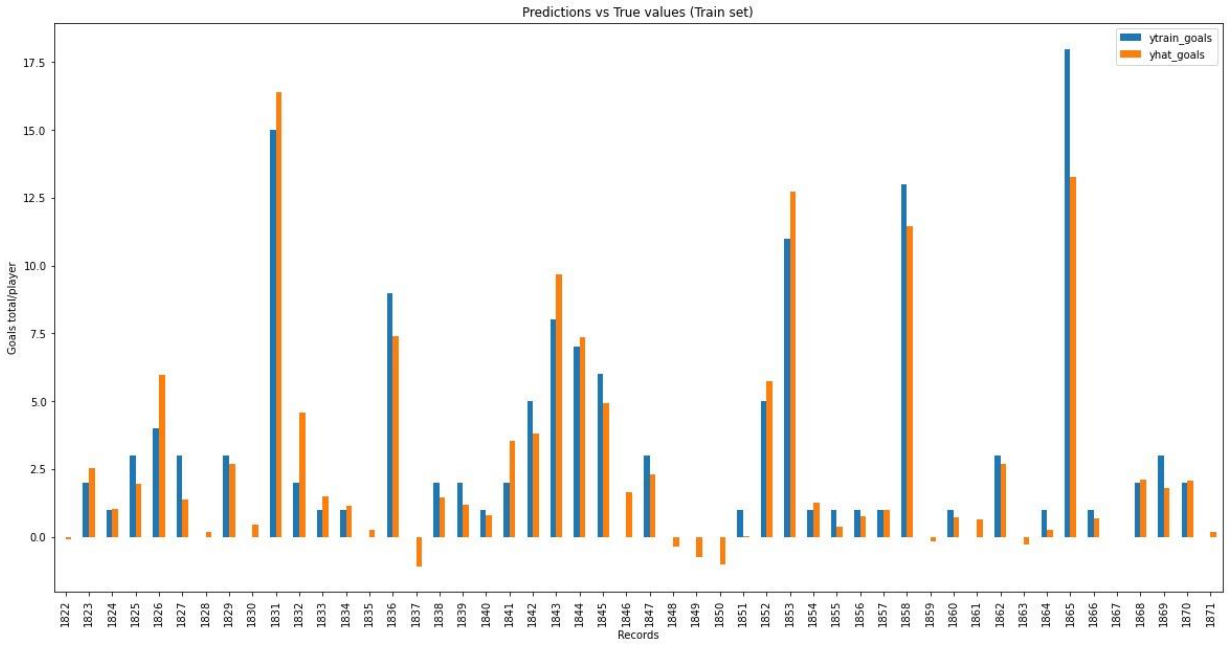

Figure 36. Predictions versus true values for Stochastic

Linear Regression

Even though the Linear regression is historically the simpler model, the predictions below are quite similar to most advanced models like Ridge and the Stochastic.

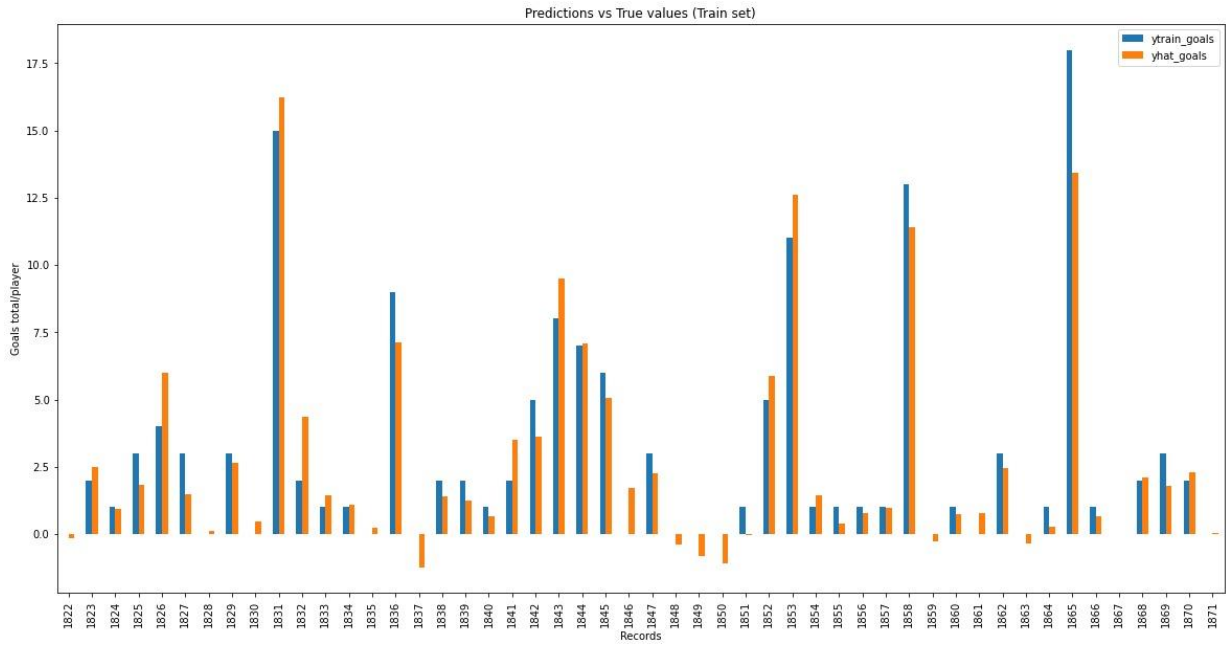

Figure 37. Predictions versus true values for Linear Regression

Appendix 3. Goals assists

Linear Regression

For goals assists, it can be seen that overall, the predictions are very similar to the true values. However, in some instances the true values were quite high, whereas the predictions undershoot by a significant amount.

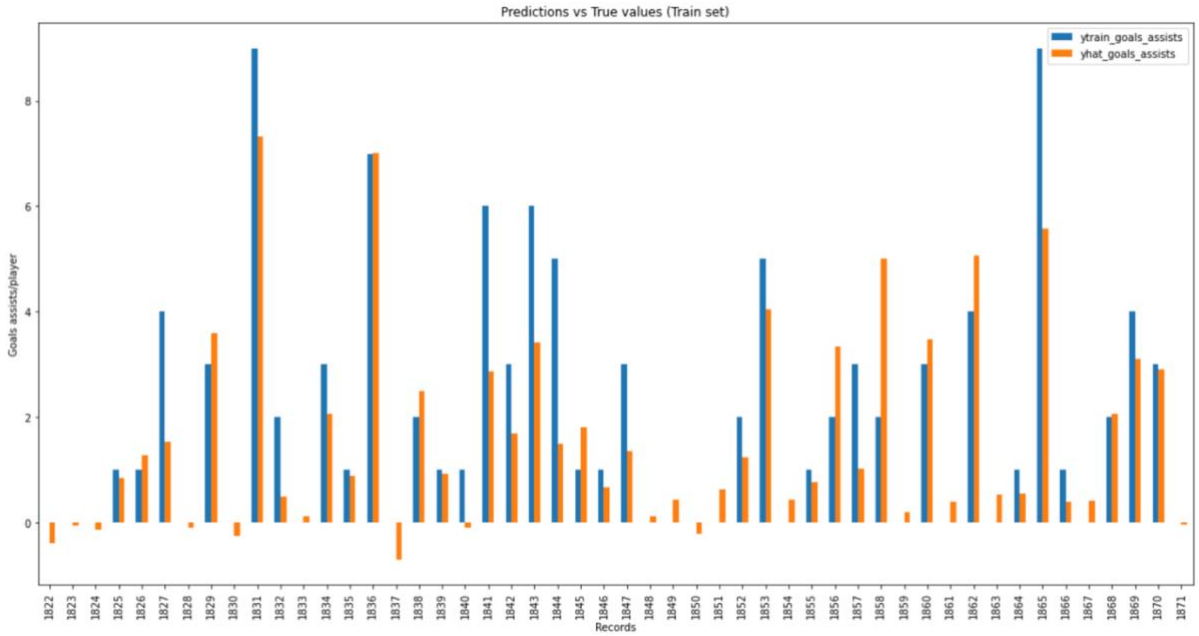

Figure 38. Predictions versus true values for Linear Regression

Ridge Regression

Just like in the other cases, predictions by this model were very similar to the Linear ones.

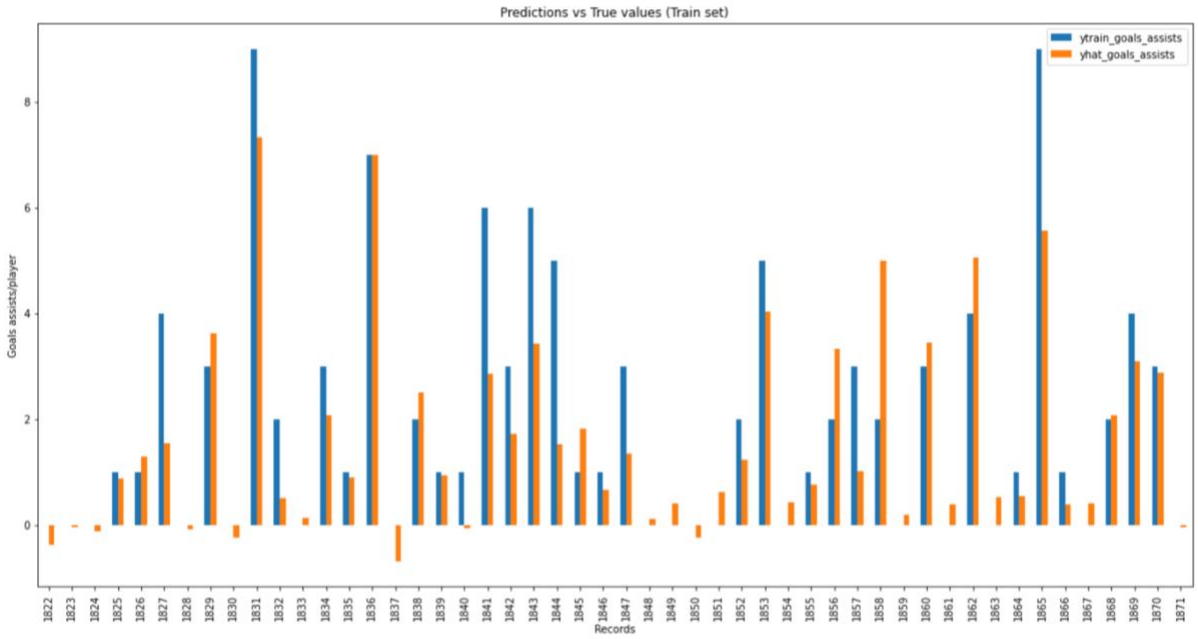

Figure 39. Predictions versus true values for Ridge

Stochastic Gradient Descent

Once again, it can be observed the similarity amongst all these model's predictions, which contributes to the reliability of the results.

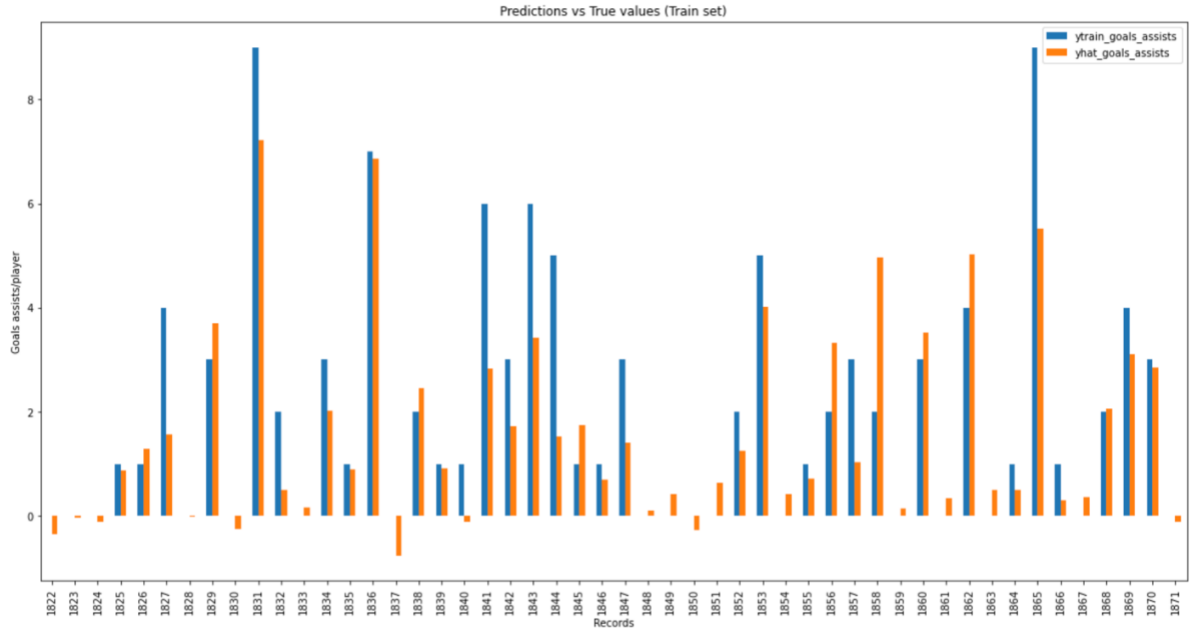

Figure 40. Predictions versus true values for Stochastic