İZMİR BAKIRÇAY UNIVERSITY

GRADUATE SCHOOL

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS

PROGRAMME IN BUSINESS INTELLIGENCE AND DATA ANALYTICS

ESTIMATION OF FOOTBALL PLAYER VALUATIONS USING MACHINE LEARNING TECHNIQUES: THE CASE OF TURKISH FOOTBALL SUPER LEAGUE

MASTER OF SCIENCE (NON THESIS) TERM PROJECT

Erkan ÇETİNYAMAÇ

Supervisor: Asst. Prof. Dr. Serhat PEKER

İZMİR

February 2022

İZMİR BAKIRÇAY UNIVERSITY

GRADUATE SCHOOL

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS
PROGRAMME IN BUSINESS INTELLIGENCE AND DATA ANALYTICS

ESTIMATION OF FOOTBALL PLAYER VALUATIONS USING MACHINE LEARNING TECHNIQUES: THE CASE OF TURKISH FOOTBALL SUPER LEAGUE

Master of Science (Non Thesis) Term Project

Erkan ÇETİNYAMAÇ

SUPERVISOR
Asst. Prof. Dr. Serhat PEKER

İZMİR

FEBRUARY 2022

FINAL APPROVAL FOR TERM PROJECT

This term project titled “Estimation of football player valuations using machine learning techniques: The case of Turkish Football Super League” has been prepared and submitted by Erkan Çetinyamaç in partial fulfillment of the requirements in “İzmir Bakırçay University Directive on Graduate Education and Examination” for the degree of Master of Science (Non Thesis) in the Business Intelligence and Data Analytics program/Department of Management Information Systems has been examined and approved on …/…/20...

.....	.....
Danışman	Lisansüstü Eğitim Enstitüsü Müdürü

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my supervisor, Asst. Prof. Dr. Serhat Peker for his precious support, valuable guidance and motivation. I appreciate all his contributions throughout my term project.

A support decision system is tried to build for football teams by means of machine learning techniques for the transfer making decisions. We strongly believe our experimental work can contribute the football clubs both financially and sportive success perspectives.

Finally, I would like to thank my family for their continuous spiritual and motivational support during this process. I dedicate this project to them.

ABSTRACT

ESTIMATION OF FOOTBALL PLAYER VALUATIONS USING MACHINE LEARNING TECHNIQUES: THE CASE OF TURKISH FOOTBALL SUPER LEAGUE

Erkan ÇETİNYAMAÇ

Department of Management Information Systems
Programme in Business Intelligence and Data Analytics

Izmir Bakircay University, Graduate School, February 2022

Supervisor: Asst. Prof. Dr. Serhat PEKER

In the 20th century, a sport known as "football" became the most popular sport on the planet. Although, the biggest step of popularity football has been taken in the 21st century. This development was reflected in sponsorship agreements, television broadcasting agreements, and most importantly, in the market value of football players. Players' transfer values are still very judgmental and ambiguous in today's football era. Even the richest and biggest clubs in the world can transfer players for fees that they will regret later. Transfer decisions are becoming increasingly vital for football clubs to be financially and competitively successful in this competitive market. As a result, sports analytics' importance is being very vital and widely used around the world. One of the most important problems in this area is the valuation of football players. The goal of this research is to develop models to estimate market prices for football players using machine learning techniques to assist sport clubs in their transfer actions. recommended techniques will be based mainly on the particular statistics of the players of the Turkish Super League for the 2020-2021 season.

Keywords: Data science; Machine learning; Data mining: Sport analytics.

10/02/2022

ÖZET

MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE FUTBOLCULARIN TRANSFER PİYASA DEĞERLERİNİN TAHMİN EDİLMESİ: TÜRKİYE FUTBOL SÜPER LİGİ ÖRNEĞİ

Erkan ÇETİNYAMAÇ

Yönetim Bilişim Sistemleri Anabilim Dalı
İş Zekâsı ve Veri Analitiği Programı

İzmir Bakırçay Üniversitesi, Lisansüstü Eğitim Enstitüsü, Şubat 2022

Danışman: Dr. Öğr. Üyesi Serhat PEKER

20. yüzyılda "futbol" adı verilen bir spor dalı tüm dünyaya yayılmış ve en popüler spor dalı haline gelmiştir. Ancak, futbolun bir eğlence sektörü olarak büyük ölçüde gelişimi 21. yüzyılda olmuştur. Bu gelişim sponsorluk anlaşmalarına, televizyon yayın anlaşmalarına ve en önemlisi futbolcuların piyasa değerine yansımıştır. Bununla birlikte günümüz futbol dünyasında, oyuncuların transfer değerleri hala oldukça yargılayıcı ve kırılgandır. Dünyanın en zengin ve en büyük kulüpleri bile sonradan pişman olacakları ücretler karşılığında oyuncu transfer edebilmektedir. Bu rekabet ortamında futbol kulüplerinin hem finansal hem de rekabet açısından başarılı olabilmeleri için transfer kararları giderek daha fazla önem kazanmaktadır. Bu nedenle spor analitiği, dünya çapında popülaritesi ve uygulaması gün geçtikçe artan bir alandır. Bu alandaki en önemli sorunlardan biri de futbolcuların değerlemesidir. Bu çalışmanın amacı, futbol kulüplerinin transfer kararlarını desteklemek için makine öğrenmesi tekniklerini kullanarak, futbolcular için piyasa değerlerini tahminleme adına model oluşturmaktır. Önerilen modeller, esas olarak 2020-2021 sezonu, Türkiye Süper Ligi'ne ait oyuncuların bireysel istatistiklerine dayanacaktır.

Anahtar Sözcükler: Veri bilimi; Makine öğrenmesi; Veri madenciliği; Spor analitiği.

10/02/2022

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

I hereby truthfully declare that this term project is an original work prepared by me; that I have behaved in accordance with the scientific ethical principles and rules throughout the stages of preparation, data collection, analysis and presentation of my work; that I have cited the sources of all the data and information that could be obtained within the scope of this study, and included these sources in the references section; and that this study has been scanned for plagiarism with Turnitin scientific plagiarism detection program used by İzmir Bakırçay University, and that "it does not have any plagiarism" whatsoever. I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

.....

Erkan ÇETİNYAMAÇ

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ML : Machine Learning

TP : True Positive

TN : True Negative

FP : False Positive

FN : False Negative

KNN : K-Nearest Neighbor

RF : Random Forrest

SVM : Support Vector Machines

XGB : Extreme Gradient Boosting

LR : Logistic Regression

GBM : Gradient Boosting Machines

1. INTRODUCTION

With the beginning 21st century, the football and other major sport branches evolved to huge industry. Astronomical transfer fees are paid to famous players. Paying the right amount of money to the right player has become more important at this time. Considering invested money to players' importance has been increased dramatically both from a sporting point of view and from a business standpoint. With the correct determination of the transfer fees of the football players, the sport clubs may make a profit. Therefore, in this paper, an AI model is offered to contribute the issue.

The main aim of the project is to evaluate performance metrics of an end season statistics of players to discover underpriced players despite their performance and contribute low or mid-level budgeted teams regarding creating a team according to their budgets. By means of suggesting method, these teams will be competitive with a low budget while forming a team. This concept is very similar to Moneyball strategy used by Billy Beane at Oakland Athletics Baseball team in the 2002 season. The statistical analyzes is used to get new players with a reasonable budget.

In the research community for sports analytics, a couple of ML algorithms have been recommended in order to predict transfer market value of football players. The algorithms whose performances have been above certain level are Decision Tree, KNN, Logistic Regression, Gradient Boosting Machine, XGBoost, Random Forrest, Support Vector Machine.

In this paper, we focus on predicting transfer market value of football players who plays in Turkish Football Super League, using statistical modeling techniques. The process has three steps. Firstly, collecting the data of an end season statistics of players. Secondly, several ML models deploying, testing and evaluating. Lastly, predictions of players' transfer market value with the best model and accuracy.

2. LITERATURE REVIEW

Rory P. Bunker and Fadi Thabtah (2019) are interested in a smart methodology that has shown promising results in the categorization and prediction fields. Because of the huge monetary stakes involved in betting, sport prediction is one of the fastest expanding fields that requires excellent predictive accuracy. In addition, club managers and owners are looking for classification models to aid them in understanding and developing successful strategies. These models are based on a number of game-related elements, such as previous match results, player performance indicators, and opponent data. This study offers a comprehensive assessment of the machine learning literature, focused on the application of Artificial Neural Networks to predict sports results. They go over the learning methodologies that were employed, the data sources that were used, the proper model evaluation methods, and the particular challenges that came as a result of predicting sport results. As a result, they provide a one-of-a-kind sport prediction framework that uses machine learning as a learning method. Those performing future research in this sector will hopefully find our findings educational and valuable.

Football is a well-known sport that also happens to be a huge source of amusement, according to scholarly work by Oliver Müller, Alexander Simons, and Markus Weinmann (2017). Player transfers are the most important decisions that team managers make from a managerial sense, hence concerns surrounding player valuation, particularly the assessment of transfer fees and market valuations, are extremely important. Market values are estimates of transfer fees, or the possible prices paid for a player on the football market, and hence play an important role in transfer negotiations. These valuations have traditionally been established by football specialists, but crowdsourcing has become a more popular means of evaluating market value. Despite the fact that studies have found good correlations between estimated market values and real transfer prices, the process underlying crowd judgments is opaque, crowd estimates are not replicable, and they are updated infrequently due to the high number of people engaged. As a result, data analytics could be a suitable alternative or complement to crowdsourced market value estimations. The regression results suggest that data-driven market value assessments can overcome several of the crowd's practical constraints while still producing comparable values. Data analysis provides for exact, impartial, and dependable

market value projections that can be updated at any time, which is important for football coaches and player scouts.

Dibyanshu Patnaik, Harsh Praharaj, and Kartikeya Prakash (2019) are aware that for a long time, the global transfer market in football has been devoid of technology. During the previous two decades, the economic dynamics have radically changed, and the amount of money pushed has dramatically increased. Player values are challenging in and of themselves for any physical entity with a well-defined set of parameters, and they are dependent on a variety of circumstances. It's far more difficult to do the same for a single person with varying parameters depending on the data source. They hope to cover the gap between estimated and final costs by determining the most efficient technique of acquiring data on football players and applying the proper model to it in order to extract meaningful information. They do it by crowdsourcing data and using regression-based approaches in conjunction with the Opta index. Then also employ Neural Networks to try to predict the outcomes, and they conclude by comparing our models.

The value of a football player is a challenging issue for Rade Stanojevic and Laszlo Gyarmati (2016). Estimating a player's worth is vital not only for scouting, bidding, and negotiating, but it also piques the interest of the media and spectators. Because of the complexity that results from the fact that the player pool is scattered across many different football leagues and many different playing positions, many teams hire topic specialists to evaluate the value of potential players. Human-based scouting, they argue, has a number of drawbacks, including a hefty charge, inability to expand to hundreds of active football players, and the inevitability of subjective biases. In this paper, they present a methodology for calculating data-driven player market value that addresses these problems. To evaluate the quality of the proposed methodology and demonstrate that the ML-based estimates outperform the commonly used Transfermarkt football player market pricing estimate in predicting team performance.

Sports analytics is a booming field with a wide range of applications around the world, as Ahmet Talha Yiğit, Barış Samak, and Tolga Kaya (2020) know. One of the unaddressed challenges in this market is the valuation of football players. The purpose of this study is to create a football player value assessment model that uses machine learning techniques to help football clubs make transfer decisions. The proposed models will mostly be based on the key qualities of individual players as

they are presented in the Football Manager video game. R was used to construct all of the models. The performance of the models is compared using their mean squared errors.

Emre Horasan and Emrah Keleş (2017) analyze the impact of sporting success and recognition on market values in this study. For this, the most recent market valuations of the most valued players in the Turkish Football Super League were examined, as well as the impact of sporting performance and media recognition. According to the findings, sports performance is the most important factor of football players' market worth, and recognition has a substantial impact on market value. The density of Google searches was used to indicate the athletes' popularity in the study. The recognition effect was described as the residuals generated from the regression of popularity based on sporting performance, and this effect, independent of performance, explains market value in a significant way. The study's most remarkable finding is that the identification of high-market-value football players in the Turkish Super League has a greater impact on market values than a linear relationship. That is, the higher the level of recognition, the higher the market value, as explained by Adler's Superstar Theory. The study adds to the literature by giving positive evidence for the superstar effect from the Turkish Football Super League, which also uses the quantile regression method.

According to new research from Jana and Hemalatha (2021) on a project, they purpose to create a machine learning-based model that could anticipate football player transfer fees using a range of data sources. Artificial intelligence principles are currently widely used in almost every industry. In order to improve their earnings, many football clubs use such tactics. The paper demonstrates one method for developing such a strategy. The initial phase in the project is to collect data in order to develop a high-precision model. Following that, a variety of machine learning methods were applied, including linear regression and others. A set of criteria was used to assess the models. Findings and interpretations were made after the model evaluations.

3. METHODOLOGY

3.1. Formulation of Player Valuations

The importance to paying optimal transfer fees to get a player holds crucial importance to football club's financial managements. In this paper, to the contribute football clubs' financial operations, a classification model developed in order to get valuation of a football player depending on his one whole season performance statistic indicators [Table 3.1]. The target variable was determined by considering player's transfer market values as;

$$PlayerValuation = \frac{currentValue - preSeasonValue}{preSeasonValue} * 100 \qquad (3.1)$$

After altering the target variable in a standard form, we tried to transform the case from a regression problem to classification. Equal-Frequency Discretization method was applied in order to transform continuous target values to discrete intervals. Therefore, the concept is evolved to a classification problem. We mention discretization process and discretization intervals in pre-processing chapter.

3.2. Indicators of Market Value

Predictor variables are mostly football players whole season performance statistics. There are three features that are nominal. Two of these nominal features holds the information of players' whether recently joined team and league or not. The other nominal feature indicates whether the player foreign or local. All of the features and descriptions are given in the table 3.1 below. The aim is a train a model that predicts whether a player increased or decreased his transfer market value according to a player's whole season performance. Therefore, according to these purpose, 639 rows of data who played Turkish Super Football League in 2020 – 2021 season collected, including meaningful statistics of footballers that can be used to building ML models such as goals, assists and accurate passes.

Table 3.1. The feature descriptions

Indicator	Type	Description
apps	Numerical	Refers to a player's total # of appearances in the matches.
firstStart	Numerical	Refers to a player's total # of being in starting XI in the matches.
minsPlayed	Numerical	Refers to minutes played at league matches.
isForeign	Categorical	Refers whether the player is foreign or local.
newToTeam	Categorical	Refers whether the player joined the team recently or not.
newToLeague	Categorical	Refers whether the player joined the league recently or not.
goal	Numerical	Refers to the # of goals a player has scored in the season.
goalShotRatio	Numerical	Refers to how many of a player's total shots were scored.
passSuccess	Numerical	Refers to a player's successful pass percentage in the season.
keyPassPerGame	Numerical	Refers to a player's # of key passes per game.
assistTotal	Numerical	Refers to the # of assists a player has made in the season.
totalPassesPerGame	Numerical	Refers to the total # of passes a player has made per game.
accurateCrossesPerGame	Numerical	Refers to a player's accurate cross per game.
accurateLongPassPerGame	Numerical	Refers to a player's accurate long pass per game.
accurateThroughBallPerGame	Numerical	Refers to a player's accurate long pass per game.
redCard	Numerical	Refers to the # of red cards received by a player.
dribbleWonPerGame	Numerical	Refers to a player’s successful ball maneuvers per game.
foulGivenPerGame	Numerical	Refers to # of fouls committed on a player per match
offsideGivenPerGame	Numerical	Refers to a player's # of being offside status per game.
aerialWonPerGame	Numerical	Refers to a player's # of aerial ball duels won per game.
tacklePerGame	Numerical	Refers to success rate of a player’s tackles per game.
interceptionPerGame	Numerical	Refers to success rate of a player’s interception per game.
foulsPerGame	Numerical	Refers to a player's # of fouls committed.
offsideWonPerGame	Numerical	Refers to a player's # of offside won per game.
wasDribbledPerGame	Numerical	Refers to a player’s # of get passed by the opponent with ball maneuvers per game.
outfielderBlockPerGame	Numerical	Refers to # of blocks made by players per game.
clearancePerGame	Numerical	Refers to # of clearances made by players per game.

3.3. Machine Learning Classifiers

In this paper, to achieve our prediction purpose, seven ML algorithms are used that are Extreme Gradient Boosting, Gradient Boosting Machine, Support Vector Machine, K-Nearest Neighbor, Decision Tree -CART, Random Forest and Logistic Regression. The majority classifier (MC) is used as a baseline to compare with learning algorithms that are mentioned before. MC assigns as based simply on the proportions found in the training data.

3.3.1. XGBoost classifier

Extreme Gradient Boosting (XGBoost) is a flexible and upgraded version of Gradient Boosting designed for model execution, adequacy, and computing performance. It's an open-source library that's part of the community of distributed machine learning. XGBoost is a unique combination of software and hardware capabilities designed to improve current boosting techniques with pinpoint accuracy in the least amount of time.

3.3.2. Gradient boosting machine classifier

One of the most efficient algorithms in the field of machine learning is gradient boosting machines (GBM). The errors in the ML algorithms are known to be classified two categories that are variance error and bias error. As a boosting algorithm GBM is used to lessen bias error of the model. GBM algorithm is used for both as a regressor and as a classifier. The cost function would be log loss when it is used as a classifier.

3.3.3. Logistic regression

Logistic regression is a classification algorithm even though its name. It is a straightforward and effective technique for linear and binary classification issues. It is extremely efficient algorithm for dealing with linearly separable cases at the machine learning literature. As a statistical based technique for specifying binary classes also logistic regression can be applied to multiclass classification problems.

3.3.4. Support vector machine classifier

Similar to Logistic Regression, Support Vector Machine (SVM) is a classification algorithm. The main goal is to determine the optimal line that divides the two classes. The algorithm ensures that the drawn line passes from the farthest point to the elements in two classes.

3.3.5. KNN classifier

K-Nearest Neighbor (KNN) is one of the most widely used and simplest algorithms for finding patterns in classification and regression problems. It is a supervised machine learning algorithm and is also known as a lazy learning algorithm. It calculates the distance of each test observation from all the observations in the training dataset, then tries to find the K nearest neighbors. The process is performed for each test observation and similarities in the data are found according to this procedure. The current distance metrics used by the KNN algorithm to calculate the distances between observations are Euclid, Manhattan, Minkowski, Chebyschev, Hamming, Cosine, and Jaccard distance metrics.

3.3.6. Decision tree classifier

The decision tree algorithm is one of the most well-known and often used algorithms. Its unrivaled adaptability and straightforwardness are its biggest assets, owing to which it has gained notoriety. In operations research and decision analysis, decision trees are used to solve problems. The examples are ordered from the bottom of the root to the leaf/terminal node in a decision tree categorization.

3.3.7. Random forest classifier

Random Forest (RF) is a supervised machine learning algorithm that uses decision trees in conjunction with the bagging technique. In both regression and classification scenarios, the RF algorithm is applied. The RF's major advantage is that it doesn't have a tendency to overfit.

3.4. Prediction Modelling

Totally 14 prediction models were developed, and they are given in the following Table 3.2. The developed models are both result of clustering analysis and separately according to players' positions. These models are also varying according to whether the target has two or three classes to estimate players' market values.

Table 3.2. Proposed models

Classifier	Positions				Clusters			Class Labels	
	Goalkeeper	Defender	Midfielder	Forward	1	2	3	2	3
1	✓							✓	
2	✓								✓
3		✓						✓	
4		✓							✓
5			✓					✓	
6			✓						✓
7				✓				✓	
8				✓					✓
9					✓			✓	
10					✓				✓
11						✓		✓	
12						✓			✓
13							✓	✓	
14							✓		✓

The importance level of our predictor features varies to players' position in the pitch. According to common football knowledge and correlation map between features, suitable predictors are chosen to deploy ML models. The table 3.3. below shows which predictor variables we chose for position-based model building scenario. In the cluster-based models 9-14, all of the predictor features that are used in position-based models are included.

Table 3.3. Performance indicators for the models

Indicator/Classifiers	1-2	3-4	5-6	7-8	9-14
apps	✓	✓	✓	✓	✓
firstStart	✓	✓	✓	✓	✓
minsPlayed	✓	✓	✓	✓	✓
isForeign	✓	✓	✓	✓	✓
newToTeam	✓	✓	✓	✓	✓
newToLeague	✓	✓	✓	✓	✓
goal		✓	✓	✓	✓
goalShotRatio				✓	✓
passSuccess		✓	✓	✓	✓
keyPassPerGame			✓		✓
assistTotal			✓	✓	✓
totalPassesPerGame			✓		✓
accurateCrossesPerGame			✓		✓
accurateLongPassPerGame	✓		✓		✓
accurateThroughBallPerGame			✓		✓
redCard				✓	✓
dribbleWonPerGame			✓	✓	✓
foulGivenPerGame		✓		✓	✓
offsideGivenPerGame				✓	✓
aerialWonPerGame		✓		✓	✓
tacklePerGame	✓	✓			✓
interceptionPerGame	✓	✓			✓
foulsPerGame		✓			✓
offsideWonPerGame		✓			✓
wasDribbledPerGame		✓			✓
outfielderBlockPerGame		✓			✓
clearancePerGame	✓				✓

4. CASE STUDY

4.1. Data collection and description

Gathering the data is a crucial process for a data science project. Sufficient amount of data is required to deploy successful ML models. The data that are used in this work are scraped by means of Beautiful Soup library of python.

The collected data that contains football players' features who plays in Turkish Super Football League and their whole season performance stats and transfer market values of the 2020-2021 season. The data gathered from WhoScored and Transfermarkt websites. Players' performance values mainly gathered from WhoScored and transfer market values and other personal features was obtained from Transfermarkt. 639 rows of observation's statistical descriptions are given table 4.1 below.

Table 4.1. Statistical feature descriptions

Variable	Mean	Median	Std.	Min	Max
newToTeam	-	-	-	0	1
newToLeague	-	-	-	0	1
isForeign	-	-	-	0	1
firstStart	14.05	12	11.42	0	39
apps	19.17	19	11.61	1	40
minsPlayed	1259.64	1102	984.78	1	3424
goal	1.68	0	2.91	0	22
foulGivenPerGame	0.67	0.58	0.52	0	3.35
tacklePerGame	0.84	0.78	0.68	0	4
interceptionPerGame	0.68	0.54	0.63	0	4
foulsPerGame	0.77	0.75	0.54	0	3
offsideWonPerGame	0.10	0	0.20	0	1.50
wasDribbledPerGame	0.51	0.45	0.41	0	2.03
outfielderBlockPerGame	0.15	0.06	0.22	0	1.30
aerialWonPerGame	0.87	0.59	0.88	0	5.67
keyPassPerGame	0.46	0.36	0.45	0	2.87
dribbleWonPerGame	0.49	0.37	0.50	0	3.93
assistTotal	1.11	0	1.84	0	17
totalPassesPerGame	24.86	23.50	15.05	0	76.97
accurateCrossesPerGame	0.27	0.08	0.36	0	2.53
accurateLongPassPerGame	1.77	1.14	1.88	0	12
accurateThroughBallPerGame	0.01	0	0.03	0	0.25
clearancePerGame	0.98	0.56	1.14	0	6.04
offsideGivenPerGame	0.10	0	0.18	0	2
passSuccess	77.52	80.27	13.58	0	100

4.2. Data Preprocessing

After examining the data, we realized that there were players who played very little time at the matches in the data set. or were transferred to other leagues as soon as the league started. These players either consist of players who were transferred to other leagues when the league just started, or they stayed at the team for the whole but played very little time. In our opinion, these players should not have been included to models for achieving better accuracy. Therefore, we applied two-step cluster analysis to eliminate these group of players. Hierarchical clustering applied with Ward Linkage and a dendrogram was plotted as shown figure 1 below. After the k value is determined, k-means clustering was used with respect to k=4 value.

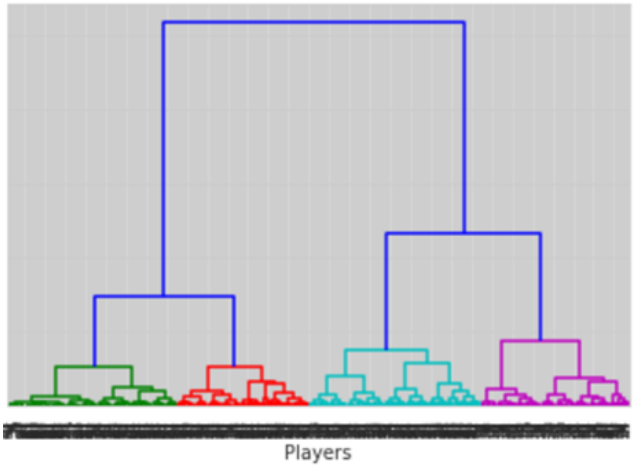

Figure 4.1. Dendrogram of the players

As result of the clustering process which based on minsStart, apps and firstStart features, four clusters were obtained. The clusters' average values of their feature values that based on basically playing time and appearance numbers in the matches is shown in the table 4.2. below.

Table 4.2. Clusters' feature averages

Cluster	Sample Size	Average apps	Average firstStart	Average minsPlayed
1	131	34.04	31.48	2773.50
2	166	26.13	19.03	1682..39
3	164	15.94	8.47	798.25
4	178	4.71	1.70	176.36
Average	160	19.17	14.05	1259.64

One of the clusters contained the players that played not sufficient minutes in the matches or not even played or get transferred to another team before transfer window closed which is cluster four. This cluster considered as outlier players. We eliminated cluster four and determined the final data set to build ML models.

Categorical feature encoding was not applied before building ML models because our three categorical features' values were binary and there was not a NA value in the data therefore missing value imputation also was not applied.

As a feature engineering process, goalShotRatio predictor feature was created by means of goal and totalShot features which are shown below equation.

$$goalShotRatio = \frac{Goals}{totalShots} \tag{4.1}$$

Another step is converting our numerical continuous valued target variable to categorical status. In order to do this process, Equal Equal-Frequency Discretization method was applied into target variable which is playerValuation. Dividing all possible values into 'N' bins, each containing the same number of observations. The intervals are shown in the table 4.3. below.

Table 4.3. Target discretization intervals

Label Count	Intervals
2	[-∞, 0]
	[0, ∞]
	[-∞, -17.32]
3	[-17.32, 25.00]
	[25.00, ∞]

Final step of pre-processing is standardizing the numerical features. It's a common prerequisite to standardize a dataset for many ML algorithms. Our predictor variables were standardized by removing the mean and scaling to unit variance where mean of samples zero and standard deviation of the samples one just before feeding data into ML algorithms.

4.3. Experimental Setup

During the data-preprocess, we discovered some crucial insights of data. We begin the ML models building phase which was shown at the table 3.2.

To conduct experiments, Sample based train – test split method was used for the data. While determining the training and test sets, firstly, cluster four which includes outlier observations extracted from the data. After the extracting players that played less minutes and participate less matches from the data, remained observations are determined as 70% train and 30% of the data as a test set to evaluate machine learning algorithms.

After these processes, we begin to deploy our ML models which will be trained 2020-2021 statistics of football players that plays in Turkish Super League to predict transfer market value of football players. The models' performances will be compared with majority classifier base model in the result chapter.

4.3.1. Software platform

Python is an open-sourced programming language with a dynamic structure and easy syntax. Python makes coding very simple and easy to implement. The biggest plus of python is its documentation. Moreover, it is compatible with a variety of libraries where "machine learning" applications can be developed. Python 3.9 was chosen to be used in this investigation because of its numerous benefits.

Sklearn is a machine learning package for the Python programming language. Sklearn gives the user a lot of options and a lot of machine learning algorithms to choose from. Sklearn comes with a lot of documentation and all of the algorithms to a data science project.

Pandas is a robust Python-based data analysis and manipulation package. When working with a huge data set, Pandas makes it simple to conduct operations like

filtering, column and row deletion, insertion, and modification. The Pandas library is used because of all of these benefits.

Matplotlib and Seaborn are Python packages that provides data visualization. The graphics for the study were created using matplotlib and Seaborn libraries.

NumPy, a Python toolkit that allows you to conduct mathematical and logical operations in computations fast and conveniently, was employed in this research.

4.3.2. Hardware platform

Runtime is a criterion for evaluating machine learning algorithms. Altough, uptime may change based on the computer's performance. As a result, the technical parameters of the computer that will be utilized in the program will be disclosed. The following are the technical specifications of the computer utilized during the application phase:

Central processing unit: Kaby Lake Intel (R) Core (TM) i7-7700HQ CPU @ 2400MHz

RAM: 16 GB - 240 GB SSD

OS: Windows 10 Pro 64-bit

Graphics Processing Unit: 4GB GDDR5 nVIDIA® GeForce® GTX1050

However, we recommend to run our code in a cloud based environment such as Google COLAB.

4.4. Performance Measures

The confusion matrix is a visual representation of the performance of a machine learning classifier. A table that shows how well a classification model performs on a set of test data for which the actual values are known. Allows for quick identification of class confusion. Almost all performance measures are usually calculated from it. The confusion matrix provides a summary of the classification problem's estimation outcomes. True Positive, False Positive, False, and True Negative definitions for more than one class can be derived from the complexity matrix.

The accuracy of a model is an indicator of its performance in classification.

One of the criteria used to evaluate classification models is accuracy. The equation describes how the ML algorithm's performance is measured.

$$Accuracy = \frac{(TP+TN)}{(\text{TP+TN+FP+FN})} \tag{4.2}$$

Precision and recall are two more key metrics. Precision denotes a high probability of success. It's a measure of how many true positives the model claims vs how many positives it claims. Recall, on the other hand, calculates how many Actual Positives our model captures (True Positive). Using the same logic, we can deduce that False Negative has a high cost,

$$Precision = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \tag{4.3}$$

$$Recall = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \tag{4.4}$$

F1-Score is the last metric we evaluated while evaluating performance. The F1-Score is very crucial statistic to evaluate classifiers' performance It's a weighted average of a model's precision and sensitivity. The equation gives the F1 score formula for a particular class.

$$F1 = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \tag{4.5}$$

4.5. Parameter Settings

The ML algorithms that were used in this experimental work has been mentioned in the chapter 3.3. The default parameters were considered to build ML models. The only differences from the default model parameters were at SVM and Logistic Regression. SVM's kernel parameter chosen linear for 2 classes target models and logistic regression's multi_class parameter determined as multinomial for 3 classes target models. The parameters are given in table 5 below.

Table 4.4. Algorithms' parameters

Algorithm	Parameter	Value
SVM	Gamma	'scale'
	C	1
	Kernel	'linear' for 2 label
DT-CART	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	2
	Criterion	'gini'
RF	N_estimators	100
	Criterion	'gini'
KNN	Number of neighbors (k)	4
	Metric	'minkowski'
LR	C	1
	Multi_class	'multinomial' for 3 label
XGB	Maximum number of iterations	2000
	Learning rate	0.1
	Max_depth	3
	Min_child_weight	1
	N_estimators	100
GBM	Learning_rate	0.1
	Loss	'deviance'
	N_estimators	10
	Subsample	1
	Max_depth	3

4.6. Results

In this chapter, we evaluated our models in terms of macro precision, recall and f-measure values. The results of the models that numbered between 1 and 8 are position-based models. Other model group is numbered between 9 and 14 which are cluster-based models. These models groups are given in two different topics. The main evaluation metric of our work was macro f-measure when we consider the ML algorithms' outputs.

4.6.1. Classifiers for player positions

To determine valuation of football players', it is crucial to consider them based on their position in pitch. Since we will be comparing the players' at the same position, the valuation will be meaningful. For instance, the features that are indicative for goalkeepers might not be very effective to evaluate forward position-based players. Therefore, evaluating players according to their position in pitch is first analyzing scenario.

Table 4.5. The model results for goalkeepers

Algorithm	Recall		Precision		F-measure	
	2-label	3-label	2-label	3-label	2-label	3-label
SVM	0.67	0.50	0.75	0.41	0.58	0.34
LR	0.67	0.61	0.75	0.76	0.58	0.54
GBM	0.54	0.50	0.55	0.25	0.49	0.30
DT	0.52	0.50	0.55	0.25	0.49	0.30
KNN	0.67	0.61	0.75	0.52	0.58	0.51
RF	0.71	0.56	0.70	0.60	0.70	0.41
XGB	0.62	0.67	0.62	0.28	0.60	0.39
MC	0.20	0.07	0.50	0.33	0.29	0.11

Since the number of goalkeepers in the data is not high enough, the results of the models we have established for goalkeepers may have given biased results. This factor must be considered. Nevertheless, 0.70 f-measure achieved with RF algorithm among goalkeeper-based models with 2-labelled target.

Table 4.6. The model results for defenders

Algorithm	Recall		Precision		F-measure	
	2-label	3-label	2-label	3-label	2-label	3-label
SVM	0.56	0.36	0.56	0.36	0.56	0.34
LR	0.51	0.41	0.51	0.46	0.48	0.42
GBM	0.51	0.28	0.51	0.29	0.51	0.29
DT	0.48	0.23	0.48	0.21	0.48	0.22
KNN	0.43	0.38	0.43	0.34	0.43	0.35
RF	0.50	0.42	0.50	0.42	0.50	0.42
XGB	0.48	0.31	0.48	0.31	0.48	0.31
MC	0.29	0.13	0.50	0.33	0.37	0.18

We can see that for the models that are established for defenders' f-measures are similar for each algorithm for 2-labelled target. SVM is getting one step ahead among classifier in this case. For 3-labelled target option RF and LR both achieved 0.42 f-measure.

Table 4.7. The model results for midfielders

Algorithm	Recall		Precision		F-measure	
	2-label	3-label	2-label	3-label	2-label	3-label
SVM	0.59	0.37	0.59	0.43	0.59	0.35
LR	0.59	0.46	0.61	0.55	0.59	0.47
GBM	0.58	0.31	0.59	0.37	0.58	0.31
DT	0.47	0.34	0.46	0.34	0.46	0.34
KNN	0.55	0.31	0.56	0.41	0.55	0.32
RF	0.57	0.37	0.57	0.41	0.57	0.37
XGB	0.54	0.28	0.54	0.36	0.54	0.30
MC	0.28	0.10	0.50	0.33	0.36	0.15

SVM and LR algorithms perform overall better than others both 2-labelled and 3-labelled options for midfielders in order to evaluate player valuation.

Table 4.8. The model results for forwards

Algorithm	Recall		Precision		F-measure	
	2-label	3-label	2-label	3-label	2-label	3-label
SVM	0.63	0.45	0.64	0.29	0.61	0.33
LR	0.59	0.53	0.60	0.65	0.58	0.47
GBM	0.49	0.50	0.49	0.54	0.48	0.46
DT	0.59	0.51	0.60	0.52	0.58	0.50
KNN	0.59	0.41	0.61	0.55	0.57	0.41
RF	0.52	0.48	0.53	0.46	0.50	0.39
XGB	0.48	0.50	0.48	0.50	0.48	0.46
MC	0.24	0.08	0.50	0.33	0.33	0.13

For the forward position, we get the best f-measure as 0.61 with SVM algorithm and 2 labelled target option. Both 2 and 3 labelled scores are the closest to each other comparing to other models.

To sum up, SVM and LR algorithms performed overall better than others at position-based models. Although, the best f measure achieved by RF algorithm at

goalkeeper model with 2-labelled target.

4.6.2. Classifiers for player clusters

Another approach is to evaluate the players who take a similar amount of time among themselves is another smart method that we propose. The result of the models numbered between 9 and 14 are given the tables below.

Table 4.9. The model results for cluster 1

Algorithm	Recall		Precision		F-measure	
	2-label	3-label	2-label	3-label	2-label	3-label
SVM	0.50	0.34	0.50	0.28	0.49	0.26
LR	0.56	0.37	0.56	0.36	0.56	0.36
GBM	0.71	0.40	0.70	0.40	0.70	0.40
DT	0.66	0.37	0.65	0.36	0.64	0.36
KNN	0.59	0.32	0.59	0.31	0.58	0.31
RF	0.67	0.35	0.67	0.34	0.67	0.33
XGB	0.65	0.25	0.65	0.25	0.65	0.25
MC	0.33	0.17	0.50	0.33	0.39	0.22

The players that are in the cluster 1 have the highest average playing minutes among other cluster's players. In other words, these players played at most of the matches during the season. Therefore, this cluster's evaluation metrics are vital to our case study. The base model's macro f-measure is 0.39. If we compare our machine learning models with this score, we can say that we have achieved a very good performance in general for 2-labelled target. Despite this, we see that the models in the 3-labelled target option did not give a good performance in order to estimate player valuation. Nevertheless, for this cluster group, GBM algorithm that applied 2-labelled target has the best macro f-measure score which is 0.70.

Table 4.10. The model results for cluster 2

Algorithm	Recall		Precision		F-measure	
	2-label	3-label	2-label	3-label	2-label	3-label
SVM	0.52	0.30	0.52	0.29	0.52	0.26
LR	0.49	0.25	0.48	0.25	0.48	0.25
GBM	0.51	0.42	0.51	0.41	0.51	0.41
DT	0.48	0.26	0.47	0.26	0.46	0.26
KNN	0.47	0.25	0.47	0.20	0.47	0.20
RF	0.44	0.28	0.41	0.28	0.40	0.27
XGB	0.59	0.42	0.59	0.42	0.59	0.42
MC	0.28	0.11	0.50	0.33	0.36	0.16

The table above represents ML performance measures for cluster 2. This cluster consists of players who take average playing time in Turkish Super League 2020-2021 season.

For 2-labelled target, XGB algorithms has the highest macro f-measure. For 3 labelled target, GBM and XGB algorithms have roughly 0.40 f-measure which is successful according to majority classifier's f-measure.

Table 4.11. The model results for cluster 3

Algorithm	Recall		Precision		F-measure	
	2-label	3-label	2-label	3-label	2-label	3-label
SVM	0.70	0.42	0.76	0.30	0.71	0.35
LR	0.63	0.55	0.68	0.57	0.64	0.54
GBM	0.45	0.33	0.42	0.31	0.42	0.30
DT	0.51	0.30	0.52	0.31	0.51	0.29
KNN	0.53	0.39	0.54	0.31	0.52	0.34
RF	0.51	0.38	0.52	0.40	0.46	0.35
XGB	0.42	0.49	0.36	0.46	0.38	0.45
MC	0.32	0.08	0.50	0.33	0.39	0.13

The ML model evaluation results that are given the table above for cluster 3. This cluster has players who play less time in matches than cluster 1 and cluster 2. In this case, SVM algorithm has the best f-measure score for 2-labelled target with 0.71 f-measure.

5. CONCLUSION

In this project, the aim was to analyze seasonal performance of football players in order to predict whether increased or decreased a football player's transfer market value by means of established machine learning models.

The case study based on 639 players who played in Turkish Super Football League 2020-2021 season. The estimation of player valuation is demonstrated by means of our unique collected data set from whoscored and transfermarkt web sites. Totally 14 models are established that are cluster and position based. If we compare our established ML models result to based model scores (MC), we achieved remarkable estimation performance. The highest macro f-measure score achieved at cluster 3 player group by means of SVM algorithm.

With the proposed models, estimation of the change in the transfer value of the football players would use as a decision support system and contribute to football clubs both financially and sportive success perspectives. The methodology that we propose can be adapted not only to the Turkish league, but also to other football leagues and sports branches. We expect that the football industry will increasingly adopt data analytics to support player recruitment and transfer negotiations decisions.

REFERENCES

Behravan, I., & Razavi, S. M. (2020). A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*, *25*(3), 2499–2511. https://doi.org/10.1007/s00500-020-05319-3

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-019-6413-7

Hashemi, S., Anthony, N., Tann, H., Bahar, R. I., & Reda, S. (2017). Understanding the impact of precision quantization on the accuracy and energy of neural networks. *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*. https://doi.org/10.23919/date.2017.7927224

Horasan, E., & Keleş, E. (2017). Kazandıran Sadece Futbol Mu? Futbolcu Tanınırlığı İle Futbolcu Piyasa Değerleri Arasındaki İlişki: Türkiye Futbol Süper Liginin Yatay Kesitsel Analizi. *M U Iktisadi ve Idari Bilimler Dergisi*, 157–169. https://doi.org/10.14780/muiibd.329917

Jana, A., & Hemalatha, S. (2021). Football Player Performance Analysis using Particle Swarm Optimization and Player Value Calculation using Regression. *Journal of Physics: Conference Series*, *1911*(1), 012011. https://doi.org/10.1088/1742-6596/1911/1/012011

Juba, B., & Le, H. S. (2019). Precision-Recall versus Accuracy and the Role of Large Data Sets. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 4039–4048. https://doi.org/10.1609/aaai.v33i01.33014039

Junker, M., Hoch, R., & Dengel, A. (1999). On the evaluation of document analysis components by recall, precision, and accuracy. *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318)*. https://doi.org/10.1109/icdar.1999.791887

Lemenkova, P. (2019). PROCESSING OCEANOGRAPHIC DATA BY PYTHON LIBRARIES NUMPY, SCIPY AND PANDAS. *Aquatic Research*, 73–91. https://doi.org/10.3153/ar19009

Liu, X., & Pedrycz, W. (2007). The development of fuzzy decision trees in the framework of Axiomatic Fuzzy Set logic. *Applied Soft Computing*, *7*(1), 325–342. https://doi.org/10.1016/j.asoc.2005.07.003

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. https://doi.org/10.1016/j.ejor.2017.05.005

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). PlayeRank. *ACM Transactions on Intelligent Systems and Technology*, *10*(5), 1–27. https://doi.org/10.1145/3343172

Patnaik, D., Praharaj, H., Prakash, K., & Samdani, K. (2019). A study of Prediction models for football player valuations by quantifying statistical and economic attributes for the global transfer market. *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*. https://doi.org/10.1109/icscan.2019.8878843

Stanojevic, R., & Gyarmati, L. (2016). Towards Data-Driven Football Player Assessment. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. https://doi.org/10.1109/icdmw.2016.0031

Yiğit, A. T., Samak, B., & Kaya, T. (2019). Football Player Value Assessment Using Machine Learning Techniques. *Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making*, 289–297. https://doi.org/10.1007/978-3-030-23756-1_36