MEF UNIVERSITY

PREDICTING OUTCOMES AND IMPROVING GAME MODELS FOR FOOTBALL MATCHES

Capstone Project

Murat Göçer

İSTANBUL, 2019

MEF UNIVERSITY

PREDICTING OUTCOMES AND IMPROVING GAME MODELS FOR FOOTBALL MATCHES

Capstone Project

Murat Göçer

Advisor: Asst. Prof. Hande Küçükaydın

İSTANBUL, 2019

MEF UNIVERSITY

Name of the project: Predicting Outcomes and Improving Game Models For Football Matches

Name/Last Name of the Student: Murat Göçer

Date of Thesis Defense: 09/09/2019

I hereby state that the graduation project prepared by Murat Göçer has been completed under my supervision. I accept this work as a "Graduation Project".

/ /

Asst. Prof. Hande Küçükaydın

I hereby state that I have examined this graduation project by Murat Göçer which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

/ /

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of Murat Göçer and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof. Hande Küçükaydın	……..
2. ……..	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Murat Göçer 09/09/2019

Name	Date	Signature

EXECUTIVE SUMMARY

PREDICTING OUTCOMES AND IMPROVING GAME MODELS FOR FOOTBALL MATCHES

Murat Göçer

Advisor: Asst. Prof. Hande Küçükaydın

SEPTEMBER, 2019, 27 Pages

This study is conducted to predict the results of the 2017/2018 English Premier League football matches and show the teams what they should pay attention to in order to win. In this study, classification algorithms are used and the algorithm that gives the best results is applied to real matches. After evaluating the results, some suggestions are made for similar future studies and for the teams to develop their game models.

Key Words: Predicting Sports Competitions, Betting On Football Matches, Classification Algorithms

ÖZET

FUTBOL MAÇLARINDA SONUÇ TAHMİNİ VE OYUN MODELLERİNİN GELİŞTİRİLMESİ

Murat Göçer

Tez Danışmanı: Dr. Öğr. Üyesi Hande Küçükaydın

EYLÜL, 2019, 27 Sayfa

Bu çalışma 2017 / 2018 yılı İngiltere Premier Ligi futbol maçlarının sonuçlarını tahmin etmek ve takımlara galibiyet almak için nelere dikkat etmelerini göstermek üzere yapılmıştır. Sınıflandırma algoritmaları kullanılan bu çalışmada, en iyi sonucu veren algoritma gerçek maçlar üzerinde uygulanmıştır. Ortaya çıkan sonuçlar değerlendirildikten sonra ilerde yapılacak benzer çalışmalar ve takımların oyun modellerini geliştirmeleri için bazı tavsiyelerde bulunulmuştur.

Anahtar Kelimeler: Spor Karşılaşmaları Tahminleri, Futbol Maçları Üzerine Bahis Yapmak, Sınıflandırma Algoritmaları

TABLE OF CONTENTS

.

1. INTRODUCTION

Football is a sports competition consisting of 11 people and played between 2 teams. It is known worldwide as football or soccer. Football is played with body parts other than hands and arms, it is free to play by hand only for the goalkeeper within the penalty area. The team who scored the most goals to the opposing team within 90 minutes wins the game. A coach leads a football team and the team is generally composed of 4 areas. These are goal, defense, midfield and forward areas. Players play in these areas according to their characteristics and abilities. In addition to national leagues, international leagues that are organized every year and the world cup that is organized once in 4 years have made the football a permanent and a major sector. The national leagues are organized by the football federations of each country, while the international leagues (Champions League, European Cup, etc.) are held by football leaders of continents such as the Union of European Football Associations (UEFA), and the Fédération Internationale de Football Association (FIFA) organizes the world cup affiliate. It is estimated that there are approximately 250 million licensed football players worldwide and 1.3 billion people are interested in this sport.[1]

The study examines the prediction of results of soccer matches and game models that can help to improve the game style of a football team through some classification algorithms. The aim of the project is to see if it is possible to predict the outcome of soccer matches with good precision and also to develop a good and dynamic model that can change the course of the match. It will be done by analyzing the feature of players and match scores of the English Premier League. For this purpose, some algorithms are developed and later the algorithm that gives the best prediction result is selected. This algorithm first tries to predict the match result under normal circumstances. It then predicts the proposed game models to change the result if there is an undesirable result. In this section, the objectives and scope of the study is dwelled upon in detail along with the brief literature review that highlights the significance of the issue.

[1] History of football. Retrieved from: https://www.britannica.com/sports/football-soccer

1.1. Literature Survey

The prediction system generally establishes a mathematical and statistical relationship between the events in the past and tells the results of future events with certain margins of error. In addition to the technical statistics of the match, the use of non-technical statistics such as weather conditions at the match day, the psychological conditions of the players, the socio-economic status of the city in which the match is played may strengthen the prediction model. However, the increase in technical and non-technical knowledge will increase complexity of the model, and as the complexity grows, the model may exit being a machine learning model and becomes an artificial intelligence model. As the complexity increases, using artificial intelligence methods and techniques gives better results (Gevaria et al., 2015). For the sake of increasing complexity, non-technical data are not used in the model.

It is a very difficult problem to predict results in a people-oriented game. Many studies have been carried out in the literature to solve prediction problems. Not only academic studies but also betting companies (also called bookmakers) have been working on match prediction for years and turned this field into a sector. Especially in terms of football, hundreds of matches are held each week around the world. The working logic of betting companies simply works as follows: if a bettor wagers 10 units of money on a bet with odds 1.20, they will win 10*1.20=12 with net profit 12-10=2 units. If the prediction of a bettor is incorrect, the bookmaker gets all units of money (Buursma, 2011).

It is known that companies are employing many statisticians and experts in determining the odds. These experts work with many features such as the power of the teams, the probability of scoring, the probability of winning, etc. and the data coming from the past. Obviously, it is beneficial to use these odd ratios in the model that is developed.

Buursma (2011) worked on a method for predicting the results of football matches, using odds set by bookmakers. In his work, he used some major features such as goals scored and conceded by teams, the average number of points gained by teams. He used several machine learning algorithms: MultiClassClassifier, RotationForest, LogitbooST, BayesNet, Naive Bayes and Home Wins. The accuracy of the best algorithm was 55%. However, with this accuracy, in the long run, an optimum gain could not be achieved. The major conclusion of his work is that the prediction of football matches will always be difficult.

Farzin et al. (2013) worked on 2008-2009 Spanish League. They chose one of the best team of the league, FC Barcelona, and they used Bayesian Network Model (BN) to predict the results of football matches of FC Barcelona. Their data was different from the others mentioned before. They used non-physiological factors such as weather report, statistics of five previous matches, results of the matches; and physiological factors such as age of the players, injury status of main players, for their project. They used NETICA software, which is used for BN models. They calculated an accuracy of 92% for the 2008-2009 season matches.

Baio and Blangiardo (2010) proposed a Bayesian hierarchical model to predict football results. They used Italian Seria A league as the data set and their work based on attack power and defense power of the teams. They applied Markov chain Monte Carlo (MCMC) procedure to predict the main effect, which is correlated with the scoring rate. Their predictions were 95% accurate which is higher than the others. Their work has yielded good results on teams that tend to score goals or concede goals. In other words, it gave good results on teams with high attack power or high defensive power.

Tüfekçi (2016) presented a model that predicts the results of football matches. She used the Turkish Super League as the data set. The data set includes 70 features and a 4-year period matches between 2009 and 2013. Three machine learning classification methods, which are Support Vector Machines (SVMs), Bagging with REP Tree (BREP), and Random Forest (RF), are made use of. Before applying these classification methods, some feature selection methods were used to find the most effective features in the data set. The number of the features was reduced to 38 from 70. After some pre-processing steps the best prediction accuracy was calculated as 70.87% by using the RF method.

Predicting outcomes are not only popular in football matches but also in other sports such as Major League Baseball (MLB), the National Basketball Association (NBA), or the National Football Association (NFL), since there is only a few teams and their features do not vary too widely. For this reason, statistics are more useful for this kind of sports. There are also many academic works on the prediction of the outcomes of these sports.

Moorthy et al. (2013) predicted the results of National College Athletics Association Basketball (NCAAB). There are more than 300 teams and each team plays about 30 games per season. Before applying a model, all teams were examined according to attack and defense power. All features were clustered into attack or defense groups and normalized.

Then, a method called “Four Factors” is applied for each team. The Four Factors method is used to determine 4 statistics (Effective field goal percentage, Turnover percentage, Offensive rebound percentage, Free throw rate) that are very important indicators of a team’s success. As machine learning methods, Decision trees, Rule learners, Multi-layer perceptron, Naïve Bayes and Random Forest methods were applied the data set. Their best accuracy results were between 74% – 75% that were changed according to the seasons.

1.2. The Data Set

Several companies sell data sets to the teams. One of them is Opta Sports[2] which is the biggest sport data company in the world. They collect all data of the players and teams. English Premier League (EPL) 2017/2018 season data set is shared by Opta Sports. The data set consists of 10,449 rows and 271 columns. Each row represents a player in a match and the columns represent the individual performances of each player in the match. The data set divides the positions of the players into 4 zones which are 1 = Goalkeeper, 2 = Defense, 4 = Midfield, 6 = Forward. All 271 columns are mixed according to the positions, and at the same time, each column has a different importance according to the game zone. For example, scoring is very important for the forward players, but a successful pass is more important for the midfielder. The defender needs to remove the ball, while saving the shot is important for the goalkeeper. All features consist of continues values, but “Result” feature is a kind of discrete value that includes 1, 0, 2 where 1 represents that the home team wins, 2 the away team wins, and 0 means a draw. All measurements were made while players were playing with the ball. There is no measurement in the area without the ball. There are no missing values in the data set. Within this project some classification models are applied to the data set and the model that gives the highest results is selected as a base model.

[2] Opta Sports Corporate Identity. Retrieved from: https://www.optasports.com/

2. PROJECT DESCRIPTION

In this section, the objective and the scope of the project are discussed by highlighting the best features that are playing a role on the target field “Result”. Following that, the methodology consists of data analysis and model deployment steps, respectively.

2.1. Problem Statement

The main questions of this research are: how football matches can be effectively predicted and how the game models can be developed. To answer these, the following questions must be answered first.

- What features are important in predicting the outcome of a match?
- What features are important to change the course of the match?
- How to generate the test set before the match starts?

To summarize, the most effective features are needed to be found to predict the outcome and also the most effective feature is needed to be found for developing a game model or changing the course of the game.

2.1.1. Project Phases

The primary aim of the project is to develop a model that predicts the outcome of an unplayed match. This way, the teams are able to see whether the opponent team wins before the match. Based on this information, the teams can find a feature that can improve the game model to win or score points, or develop a working model that will improve the player's performance in training.

Project phases are as follows:

- Analyzing the data set provided by Opta Sports,
- Pre-processing the data set,
- Selecting necessary features to reduce the complexity of the model, and
- Measuring the performance of the algorithm using the results of the played matches

3. METHODOLGY

In order to accurately predict match outcomes, a detailed data set is needed. This data set should collect information about the detailed performance of footballers and should collect match results of previous seasons. The more data we have, we get better results. The data set from Opta Sports is a very well data set based on footballer performances, but the weak side of the data set is that it contains only 2017/2018 season. Too many features (271 features) with a small number of samples (380 matches) may not give good results. It may cause "overfitting". Overfitting refers to a model that learns very well on training data with over 95% accuracy, but shows poor performance on a new test set under 70% - 60% accuracy. Overfitted models learn all details of training set including noises and it negatively impacts test performance. The main reason for overfitting is a minimum number of samples with a huge number of features. This huge number of features causes complexity and complexity leads to overfitting. To reduce the effects of overfitting, number of features should be reduced and the number of samples should be increased.

Because of league fixture, the test set will change every week. To change the test set an algorithm is coded. To fill the features of the test set (unplayed matches) features of the first 19 weeks were averaged. These average performance features are matched with the names of the teams which are unique for every team like an ID number. To calculate and visualize the correlation between features, "Seaborn" library of Python is used. However, 271 features are too large for visualizing. Before visualization, the number of features is reduced. Benefits of reducing the number of features are both visualizing and reducing overfitting. Finally, some of the most important machine learning algorithms are used and the best of them is chosen and is applied to all data set.

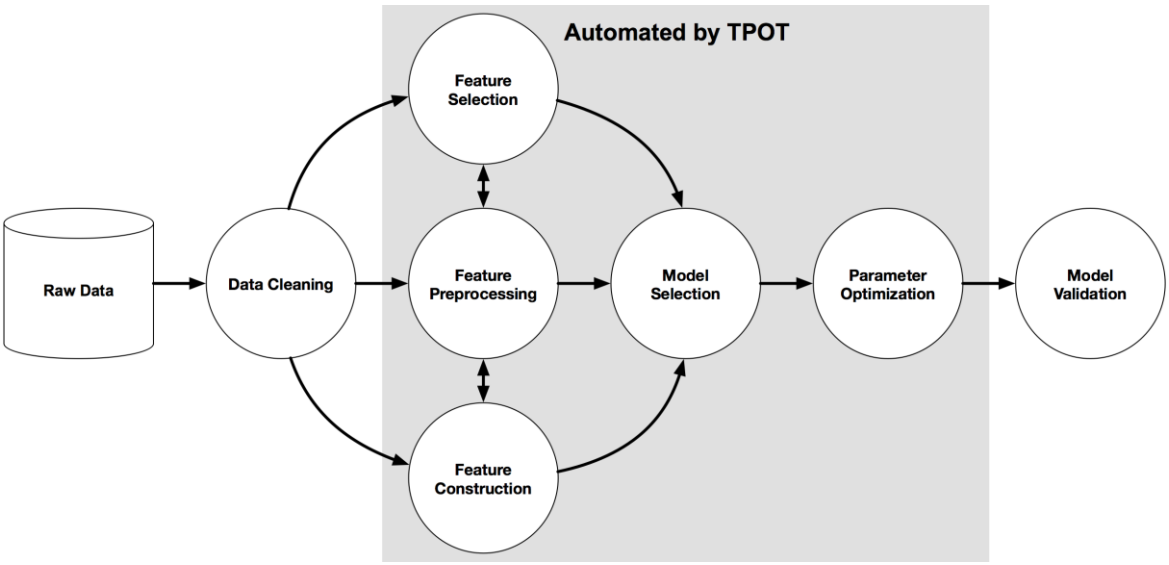
In line with the problem statement, classification models of machine learning are preferred, since the target takes discrete values 0, 1 or 2.

The data set is composed of 10,449 rows and 271 columns. However, 271 columns belong to one team and since two teams are compared, this number of columns should be written on both teams. So, for each match, there are totally 542 columns. The result column is the common column for both teams, so, we obtain 541 columns to consider for each match. Also, in order to compare the weekly performance of the teams, it is required

to get the sum of the data values of the players playing that week. Thus, there are 14 rows including the three substitution players when comparing two teams.

The pre-processing steps are developed in Python interpreter (Spyder) by mostly using Pandas and Numpy libraries. The classification algorithms are developed using Sklearn library. Besides, "Tpot" library, which is recently popular, is used. Tpot is a kind of automated machine learning library in Python. It uses classification algorithms with all parameters and tries to find the best model. It also provides a pipe-line and the codes of the best model. (Randal et al., 2016). In Figure 1, a standard machine learning pipeline and automated steps of the pipeline by Tpot are shown.

Figure 1: An example of Tpot machine learning pipeline[3]



As mentioned before, the number of features is too high which results in overfitting and prevents visualization. For these reasons, domain knowledge and forward feature selection methods are used to reduce the number of features before creating the model.

3.1. Exploratory Data Analysis

During the exploratory data analysis phase, the most effective 20 features are determined based on domain knowledge and forward feature selection method. The selected features are scaled with "Minimum maximum scaler". The reason of scaling is the weight of features. While the number of passes is measured with hundreds, the number of

[3] Tpot library. Retrieved from: http://epistasislab.github.io/tpot/

shots is just between 5 and 15 and the number of red cards is usually 0 or 1. So, the scaler is used to eliminate these weights of the features.

When the forward feature selection method is executed, it selects the 20 most effective columns among 541 columns as in the Table 1.

Table 1: Selected features with forward feature selection method

Feature	Definition
Penalties Saved	Total number of penalties saved for the home team
Direct Free-kick Off Target	Total number of direct freekicks off target for the home team
Shots Off Target Outside Box	Total number of shots off target from outside the box for the home team
Headed Shots On Target	Total number of attempts with a header on target for the home team
Attempts from Corners on target	Total number of attempts from corners on target for the home team
Attempts from Penalties off target	Total number of attempts from penalties off target for the home team
Total Successful Passes All	Total number of successful passes, including goal kicks for the home team
Unsuccessful Crosses Corner	Total number of unsuccessful crosses and corners for the home team
Touches	Total number of touches, inclusive of; attempts (all), pass (all), tackles, clearances, saves, claims, punches, smothers, take ons, take-on overruns, unsuccessful touches, freekicks (foul and dangerous play), good

	skills, interceptions and ball touches for the home team
Red Cards	Total number of straight red cards and second yellow cards received for the home team
Saves Made	Total number of saves made for the home team
Drops	Total number of catches dropped for the home team
Clean Sheets	Total number of clean sheets for a player for the home team
Shots On Conceded	Total number of goals and saved attempts conceded for the home team
Penalties Not Scored.1	Total number of penalties missed and saved for the away team
Unsuccessful Dribbles.1	Total number of dribbles attempted, resulting in a tackle and contact with the ball for the away team
Unsuccessful Crosses Right.1	Total number of unsuccessful crosses from the right side of the pitch for the away team
Unsuccessful Lay-Offs.1	Total number of unsuccessful one-touch passes made to a teammate whilst facing away from the opponents goal for the away team
Clean Sheets.1	Total number of clean sheets for a player for the away team

Sweeper Unsuccessful	Total number of failed goalkeeper attempts to come off his line and/or out of his penalty area to clear the ball for the home team

Some of these features affect the target as “Home Win”, while some of them affect “Away Win”. So, the new data set consists of 380 rows (matches) and 20 features with a target.

19 weeks of the league is taken as working set which makes a total of 190 matches. 80% of the data set is used as the training set with 152 matches and the rest serves as a test set with 38 matches.

As mentioned before, there is no missing value in the data set and, in order to be sure, “Missingno” library is used to check the “NaN” values and visualize them.

Figure 2: Checking “NaN” values with Missingno library

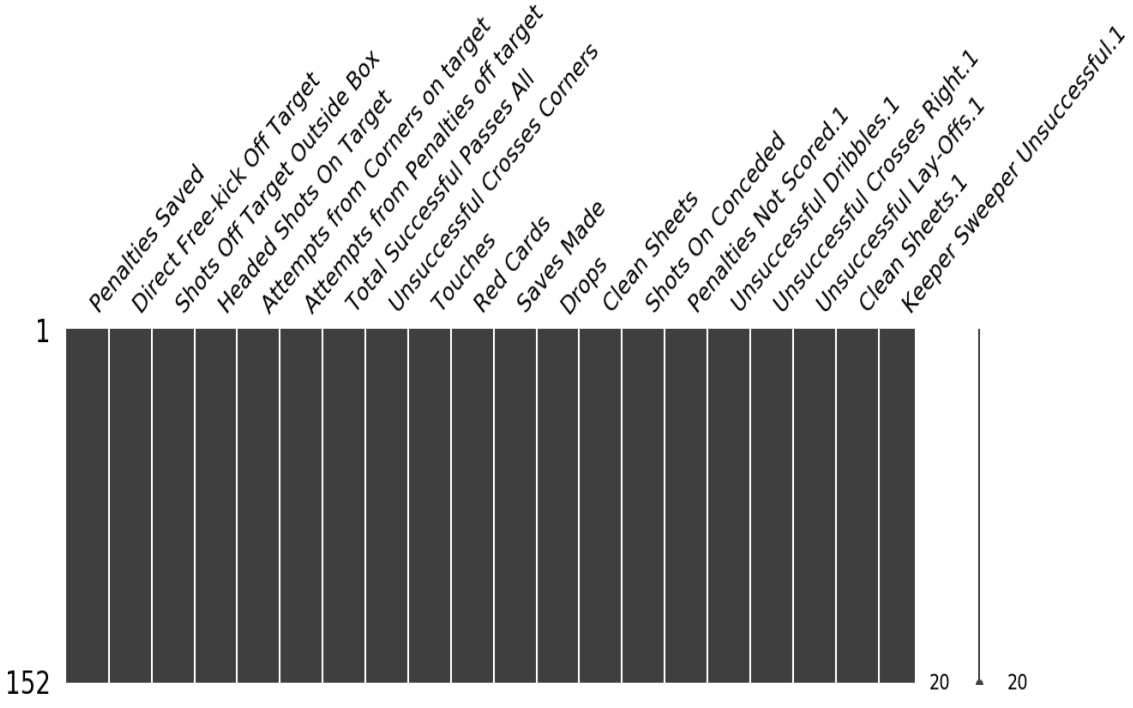
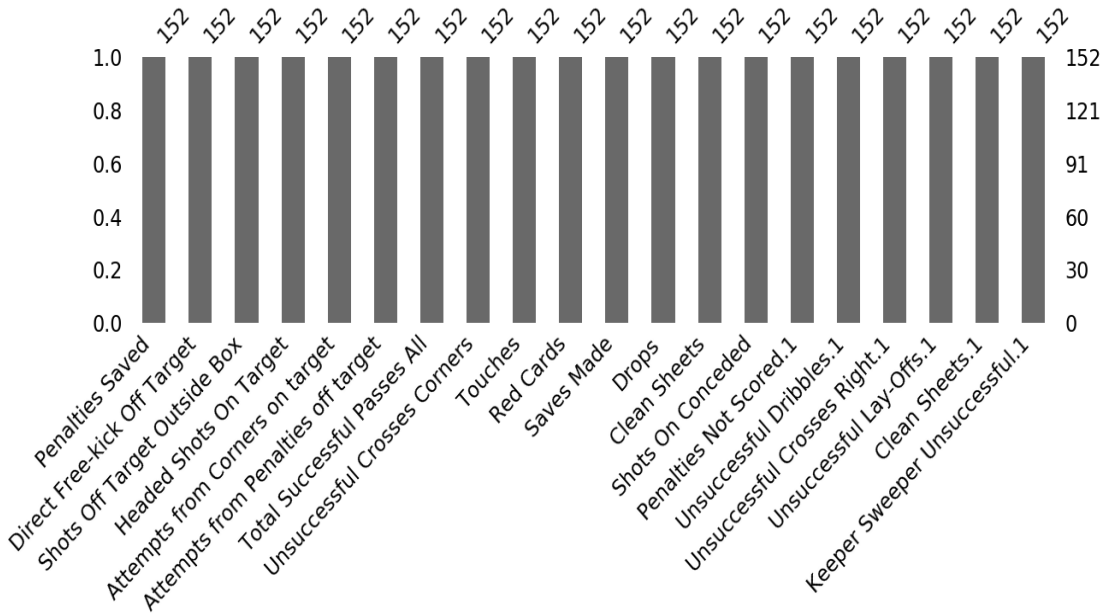


Figure 3: Checking the number of "NaN" values with Missingno library



As can be seen from Figure 2 and Figure 3, there is no missing value in the data set. These figures show the number of missing values for each column. While new data set consists of 152 samples, the bars are fully shown.

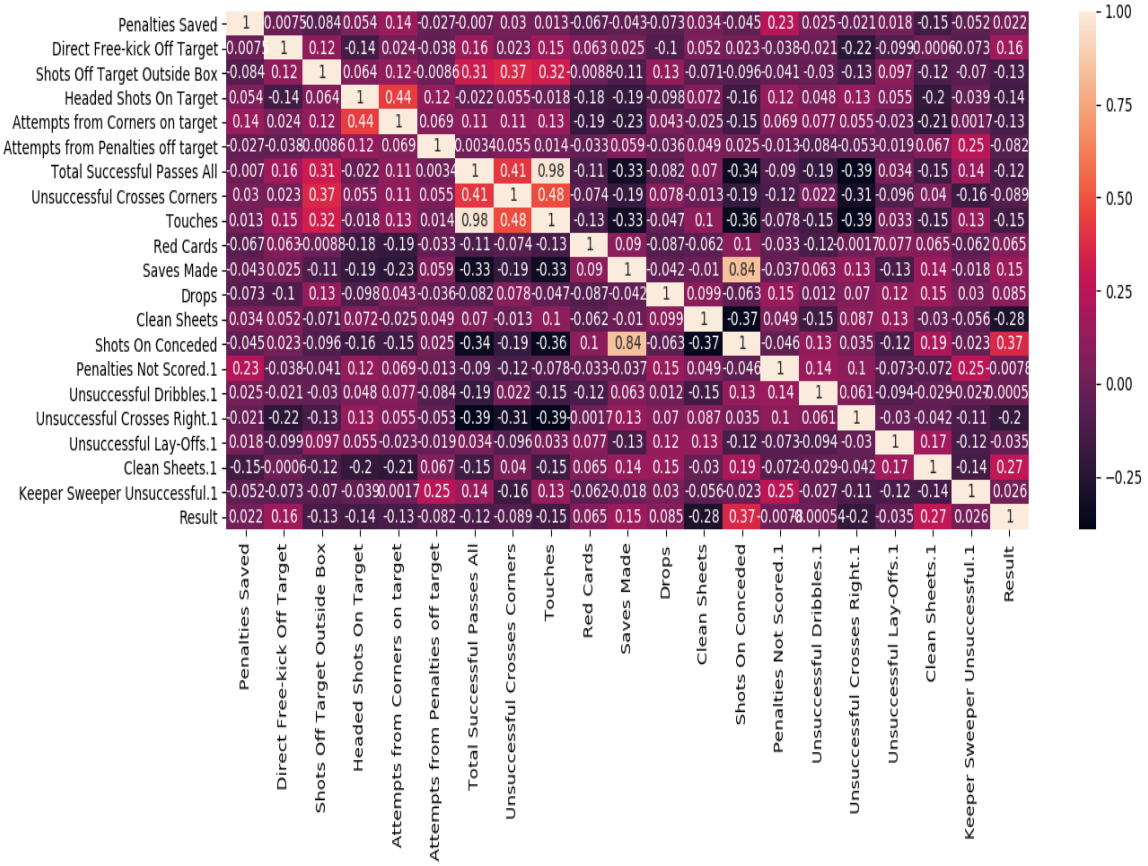
The importance of features and the relation between features can be seen from the full correlation matrix. As shown by Figure 4, a correlation matrix is a table showing correlation coefficients between variables.[4] It shows the relation between all features and also the effect of features on target value which is the outcome of the match. The best side of a full correlation matrix is that the effect of features can be seen with numbers. For instance, the correlation between "Touches" and "Successful Crosses Corners" is equal to 0.48, while the correlation between "Touches" and "Penalties Saved" is equal to 0.013. In Figure 5, a half correlation matrix is shown. In this figure, there are no numbers and the relations are represented by colors. Effect of the colors are shown in the bar on the right side of the matrix.

The most effective feature on the result columns is "Shots On Conceded" column. Scoring goals determines the match result and to score goals the ball is needed to shot. Therefore, a correlation of 0.37 makes sense. The other most effective features are "Clean

[4] What is a correlation matrix? Retrieved from: https://www.displayr.com/what-is-a-correlation-matrix/

Sheets". It means that the team prevents their opponents from scoring any goals during an entire match. This is a defensive feature with 0.27 and – 0.28 correlation with the result column. One of them has an effect on "Home Win", while the other one affects "Away Win".

Figure 4: Correlation matrix of the data set



In Figure 5, the most effective features are depicted with on orange color. Orange colored features on the result column are "Shots On Conceded", "Clean Sheets", "Clean Sheets.1" and "Direct Free-kick Off Target" columns.

Half matrix and full matrix correlation tables show same results in different ways, but instead of using colors, numbers always show relations clearer. So, checking the full correlation matrix table gives clear and absolute relations.

Figure 6 shows the features associated with the "result" column, respectively. The most effective features are represented by the longest bars in the bar plot. For instance, the shots on conceded column is the longest bar and stands out as the most effective feature on the result for the home team. On the other hand, the clean sheet column, which is the other

long bar, is negatively correlated with the result column, which means that this column is important to the away team. It is one of the commonly used plots for a quick idea.

Figure 5: Half correlation matrix of the data set

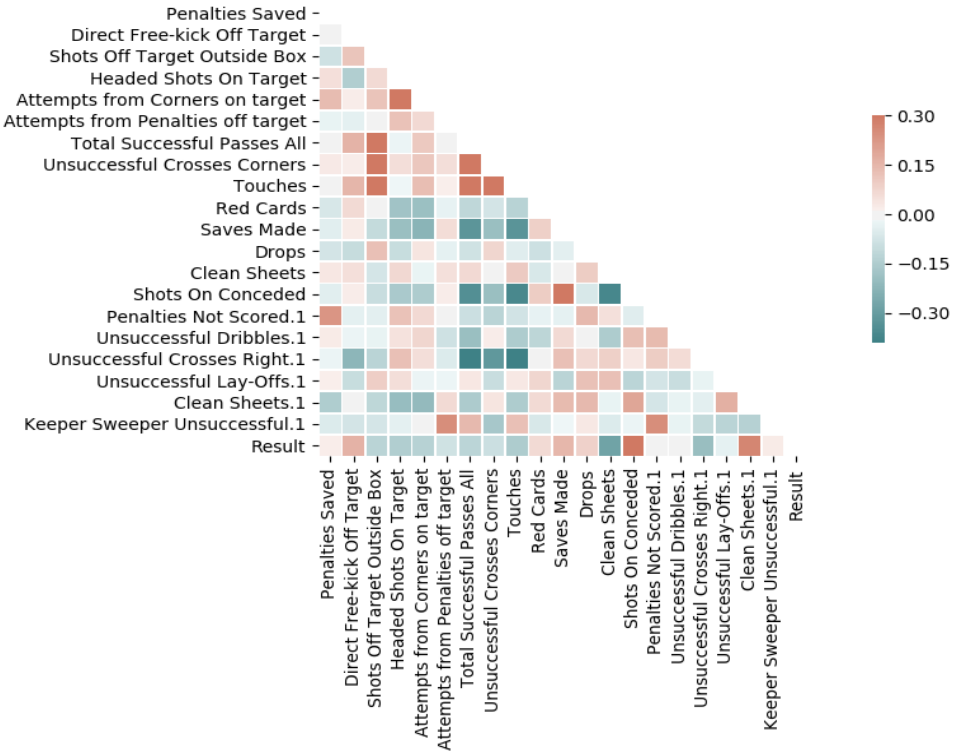
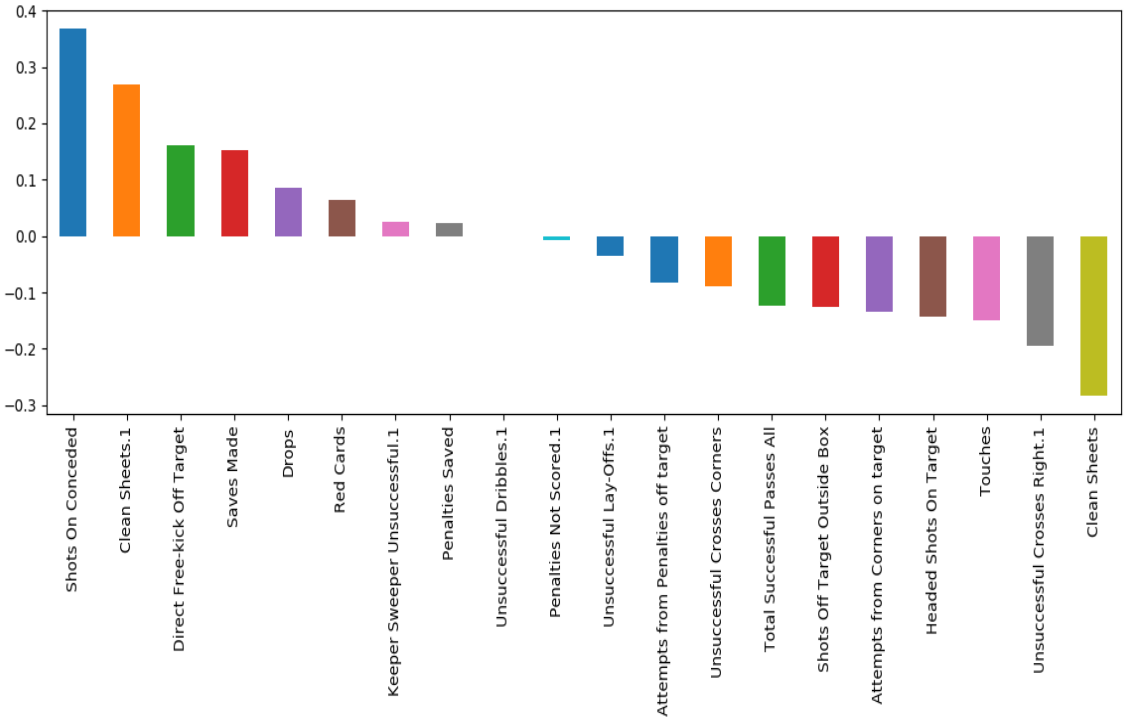


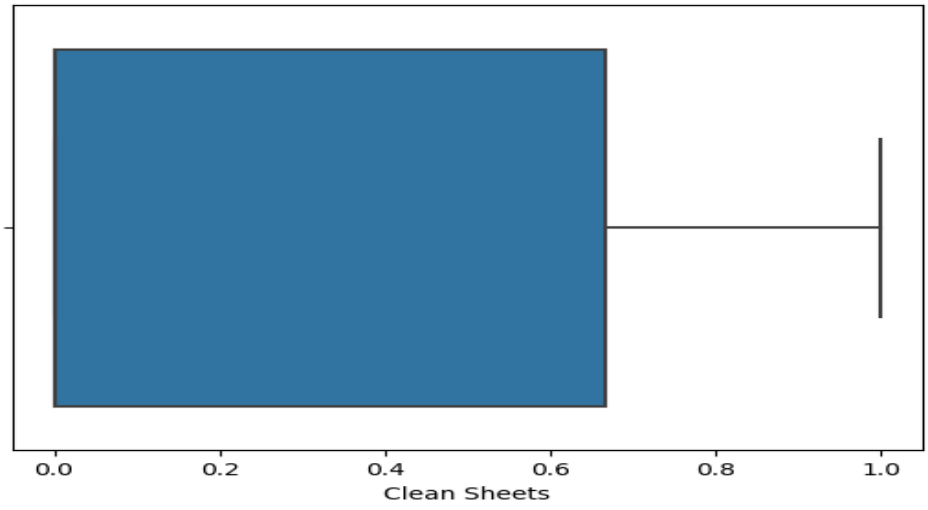
Figure 6: Bar plot of the correlation matrix

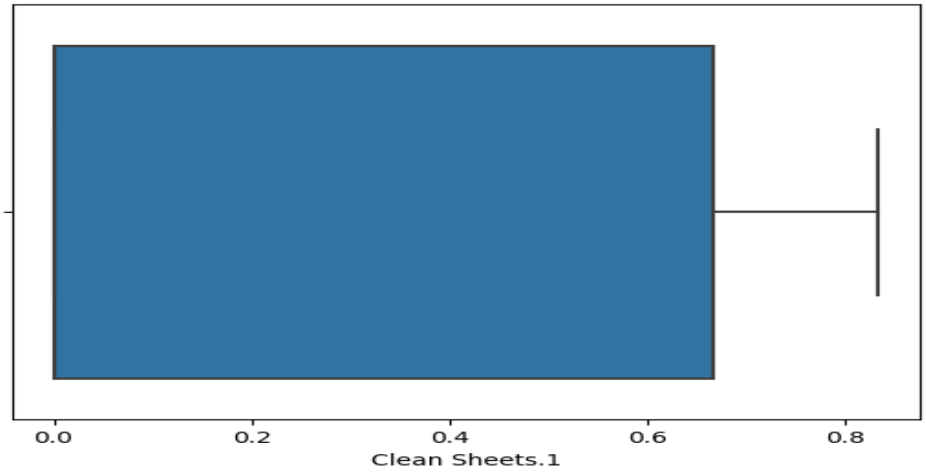


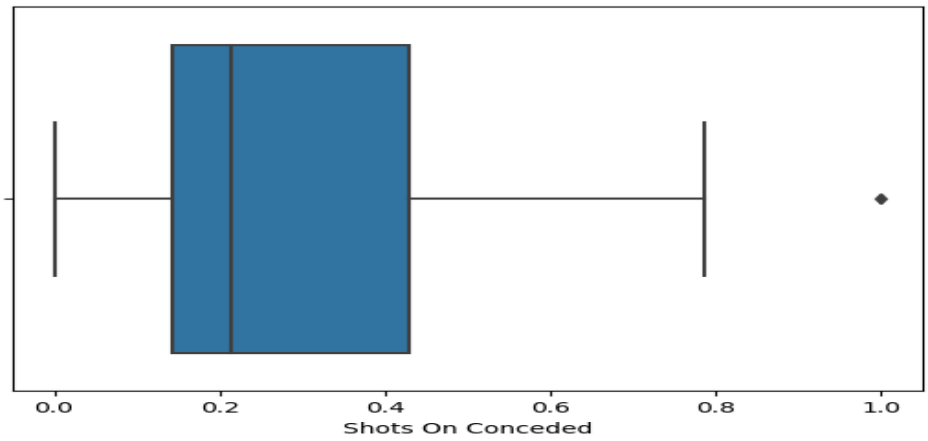
One of the most important steps in pre-processing is detecting the “outliers”. An outlier is a sample that is distant from all other samples. Outliers are not following the same pattern than the others. In sport competitions, performances may vary from day to day and from team to team. So, calling a variable as an outlier is not easy. That variable can be the best performance or the worst performance of the team. For instance, not scoring any goals is not an outlier or shoting 30 times in a match is not an outlier, it is a good performance. However, the outlier control is made for the most correlated features by using “box-plot” and “scatter-plot” tools in Figure 7 and Figure 8.

Figure 7 indicates that the feature clean has an outlier, but as mentioned before, it is not an outlier, it is more like a good performance.

Figure 7: Box plots of most correlated features for checking outliers

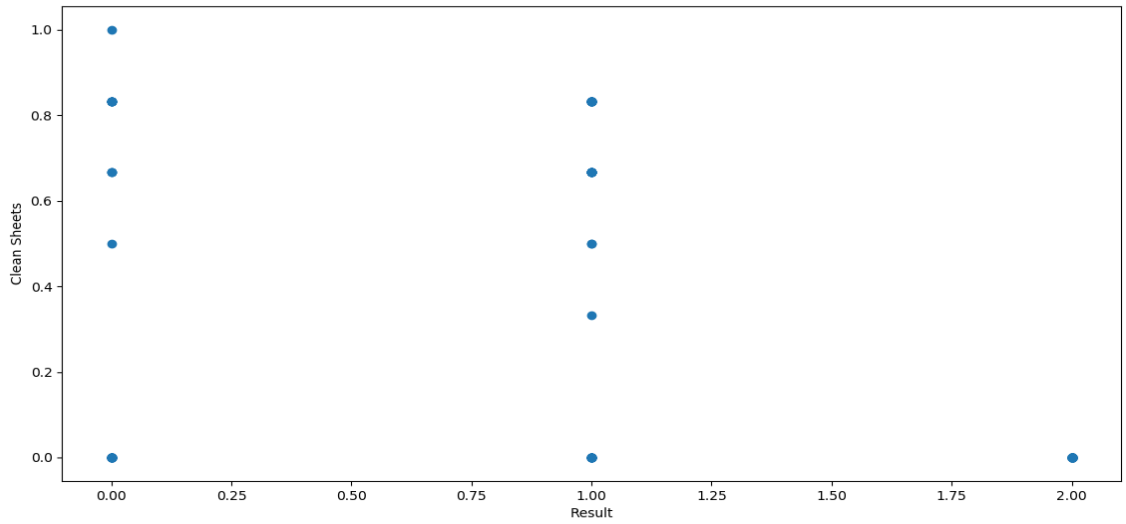


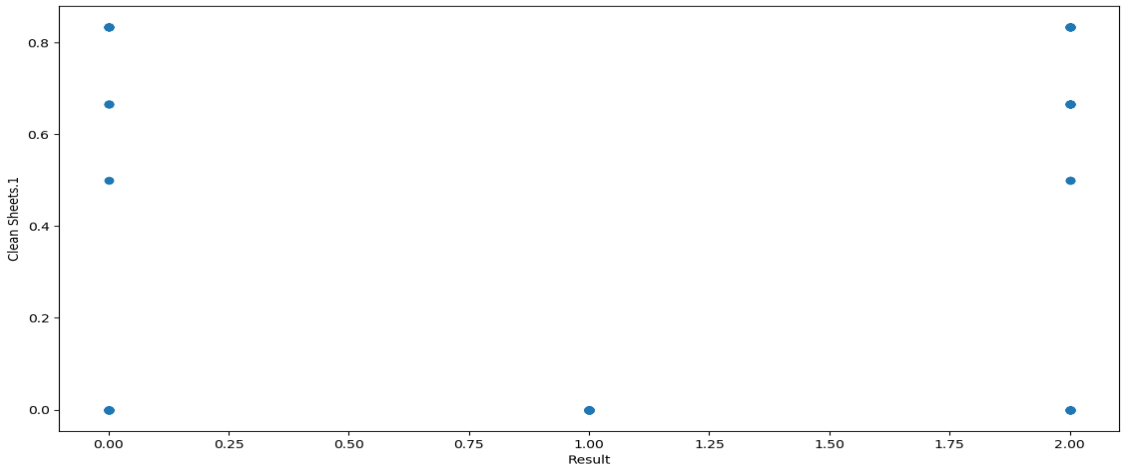


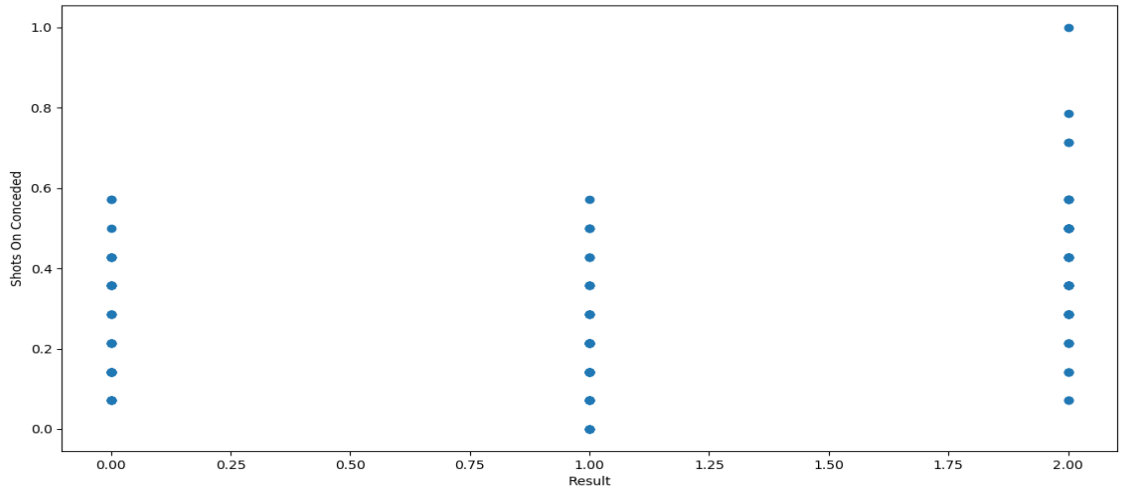


In Figure 8, the variables are shown according to classes 0, 1, and 2. All variables are close to each other. It can be deduced that there is no outlier.

Figure 8: Scatter plots of most correlated features for checking outliers

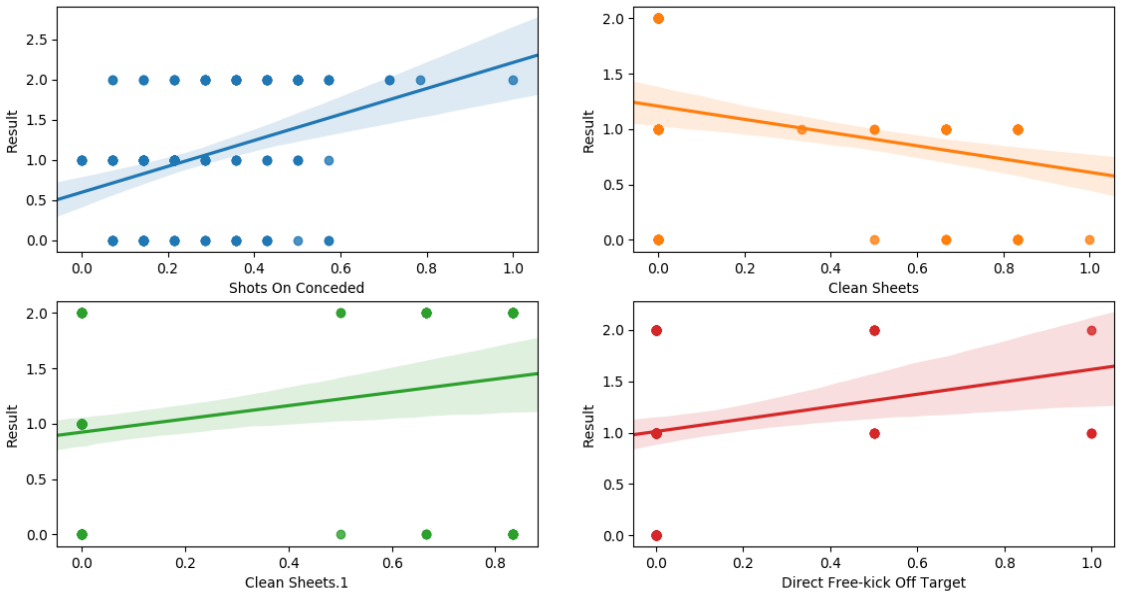






After detecting that there is no outlier, the relation between the result column and the most correlated column is displayed by Figure 9. When equal number of balls are shoted, the result of the match can be observed in favor of the opposing team. It is because that when the away team scores a goal, it starts to play a more defensive game to protect the score. Cleen Sheets and Cleen Sheets.1 columns show defensive power of the teams, so the first one is for the home team and the second one is for the away team. Direct Free-kick Off Target column shows that when the free-kicks do not hold the goal, the opposing team comes closer to the win. Thus, it can be concluded that the home team wins more free-kicks.

Figure 9: Scatter plots of most correlated features with Result column



3.2. Model Building

The model is based on the first 19 games of 2017/2018 season to learn. 80% of these 19 weeks serves as the train set, where the rest as a test set. Then, “Random Forest Classifier”, “Gradient Boosting Classifier”, ”Support Vector Classifier”, “KNN Classifier”, “Multi Logistic Regression Classifier” classification algorithms are applied to the data set and the algorithm that gives the best prediction performance is used to predict the remaining weeks of the model. In addition, the previously mentioned TPOT library is also used and the results are compared with the performance of other models. To find the best performance, also “ten-fold cross-validation” is used. Ten-fold cross-validation splits the data into ten portions (that is why it is called ten-fold) and trains the data with the nine portions while dedicating one portion for testing. It repeats this process ten times and changes train and test sets in every process. Then, the test step starts. During this step, a prediction for all matches are compared to actual results. Ten-fold cross-validation is a very powerful method for selecting the right algorithm, but it lasts too long. Thus, if time is not important in a model, it is beneficial to use it, but if the time is valuable it can be reduced to five-fold cross-validation (Buursma, 2011).

After deciding on the use of cross-validation, 5 different classification algorithms, “Random Forest Classifier”, “Gradient Boosting Classifier”, ”Support Vector Classifier”, “KNN Classifier”, “Multi Logistic Regression Classifier” and an additional TPOT library are made use of. Just looking at the “Accuracy” metric is not reliable. To select an appropriate model, “Precision”, “Recall” and “F1 Score” metrics should be checked, too. These metrics measure the success of the model. Recall signifies sensitivity. It states that it is more important to correctly detect some anomalies than to produce false alarms. Precision provides certainty by finding out how true the so-called truths are. Both methods have negative aspects. An example for negative aspects of recall and precision metrics can be given from the cancer detection. Suppose that in a city of 10000 inhabitants, 10 people are known to have cancer, and they want to be identified. Everyone is called to hospital, which makes the recall 1, and all the cancer patients are determined. However, calling all city inhabitants to the hospital means that most of the alarms are wrong. In fact, only one person is determined to be ill and that person is really ill, which makes the precision 1, but the remaining 9 people cannot be identified. These examples show that both metrics have weaknesses. After using these two metrics, another metric that is called "F1 Score" should

be checked. F1 Score is a metric calculated by taking the harmonic average of precision and recall. A curve is generated using the possibility of correctly alarming people with cancer and using the possibility of giving false alarms to people without cancer (ROC Curve). The area under this curve is equal to the F1 Score. A high F1 score indicates a good model.[5]

The "Margin of Error" is a metric for measuring the error rate of the models. The metric shows the error margin of models within a certain confidence interval. As the metric increases, the distance of the predicted event from the actual event increases. In other words, the probability of error of the model increases.[6]

When Table 2 is examined in accordance with the information given above, a high "5-fold Mean Accuracy" and a low "Margin of Error" model should be selected. Thus, it can be said that the distance between the estimated values and the actual values should be low. The Precision metric shows how many of the predicted matches are really true, and the higher the number, the more successful is the model. The Recall metric shows the correct prediction rate of the predicted matches, and a high score indicates that the model is successful. To cope with the trade-off between Recall and Precision metrics, the F1 Score metric is used. The higher this metric is, the higher is the success of the model.

As Table 2 summarizes the performance of each used algorithm, the best result is obtained by the Tpot classifier where the high rate of accuracy and F1 score, and low rate of margin of error.

[5] Success Criteria in Classification Models. Retrieved from: https://medium.com/data-science-tr/s%C4%B1n%C4%B1fland%C4%B1rma-modellerinde-ba%C5%9Far%C4%B1-kriterleri-2d86488799c6

[6] How Does Margin of Error Work? Retrieved from: https://www.statisticssolutions.com/how-does-margin-of-error-work/

Table 2: Performance metrics of the classifier algorithms

Algorithm / Model	5-Fold Mean Accuracy	Margin of Error	Precision	Recall	F1 Score
Random Forest Classifier	0.67	(+/- 0.24)	0.66	0.64	0.65
Gradient Boosting Classifier	0.72	(+/- 0.18)	0.83	0.84	0.83
Support Vector Classifier	0.77	(+/- 0.17)	0.82	0.84	0.80
KNN Classifier	0.70	(+/- 0.20)	0.81	0.81	0.76
Multi Logistic Reg. Classifier	0.64	(+/- 0.28)	0.78	0.76	0.77
Tpot Classifier	0.82	(+/- 0.09)	0.89	0.92	0.89

The reason why Tpot classifier is more successful than the other algorithms is that it uses more accurate parameters. It starts with all available parameters, selects the best one and saves a pipe-line that shows the parameter and the algorithm uses it.

In order to compare the predicted results with the test results obtained by the Tpot classifier, we make use of a confusion matrix. In general, a confusion matrix, is a table facilitating this comparison by showing the number of accurate and inaccurate estimates.[7] The confusion matrix of the Tpot classifier, which is a 3x3 matrix, is given in Table 3, where the columns correspond to predictions and the rows to the actual values. The values 6, 12, 10 displayed on the main diagonal of Table 3 represent the number of the correct predictions. The first row is the number of actual draw-finished matches. While the model predicts all of these 6 matches correctly, 3 matches, 2 of which must be home win and 1 of

[7] Confusion Matrix in Machine Learning. Retrieved from: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

which must be away win, are incorrectly predicted. The model correctly predicts 12 of the 14 home win matches, while it incorrectly predicts 2 matches as draw. 10 out of a total of 11 away win matches are correctly predicted, while the remaining 1 is incorrectly predicted. In the test set, 28 out of a total of 31 games are correctly predicted. This means that the accuracy is close to 90% and that the model works well.

Table 3: Confusion matrix of the Tpot classifier

		PREDICTED		
		Draw	**Home Win**	**Away Win**
ACTUAL	**Draw**	6	0	0
	Home Win	2	12	0
	Away Win	1	0	10

Tpot uses Random Forest Classifier and uses “criterion = entropy”, “min_samples_leaf = 3”, “min_samples_split = 18”, and “n_estimators = 100” as the best parameters to find the best score. Once the model has been decided, it is necessary to try out this model on real data, i.e. the data that never goes into the train set, and it finds out how many of the matches are correctly predicted.

4. EVALUATION OF THE OUTCOMES

In this section, the real match results are compared with the predictions made by the model that gives the best results on the train set. Later, the results are discussed and ideas are shared on what kind of improvements can be made.

4.1. Evaluation of the Algorithm

Table 4 shows the comparison of the predictions and actual results of the 20th week, after the Tpot is applied to the actual test set.

Table 4: Performance of the classifier algorithm for the 20th week

Week	Home Team	Away Team	Result	Prediction
20	Bournemouth	West Ham United	0	0
20	Chelsea	Brighton and Hove Albion	1	1
20	Huddersfield Town	Stoke City	0	0
20	Liverpool	Swansea City	1	0
20	Manchester United	Burnley	0	1
20	Tottenham Hotspur	Southampton	1	1
20	Watford	Leicester City	1	0
20	West Bromwich Albion	Everton	0	0
20	Newcastle United	Manchester City	2	2
20	Crystal Palace	Arsenal	2	2

As it can be seen from Table 4, 7 of the 10 matches are correctly predicted for the 20th week, which has a prediction accuracy of 70%. 70% is a high rate for the human-oriented sport and a model using only technical data. However, in the following weeks,

this rate is gradually decreasing. Table 5 shows the 27th week matches, results and predictions.

Table 5: Performance of the classifier algorithm for the 27th week

Week	Home Team	Away Team	Result	Prediction
27	Everton	Crystal Palace	1	0
27	Manchester City	Leicester City	1	1
27	Stoke City	Brighton and Hove Albion	0	0
27	Swansea City	Burnley	1	0
27	Tottenham Hotspur	Arsenal	1	2
27	West Ham United	Watford	1	0
27	Huddersfield Town	Bournemouth	1	0
27	Newcastle United	Manchester United	1	2
27	Southampton	Liverpool	2	0
27	Chelsea	West Bromwich Albion	1	1

In Table 5, the accuracy reduces to 30% from 70%. Finally, the total accuracy of the 38 weeks equals to (90/190)*100 = 47%. As the weeks go by, "accuracy" gradually decreases, starting at 70%, the total fell to 47%. Accuracy scores for each week can be seen from Table 6.

Table 6: Performance of the classifier algorithm for the real life data set

Week	Accuracy (%)
20	70
21	60
22	60
23	50
24	50
25	50
26	40
27	30
28	50
29	50
30	40
31	40
32	50
33	50
34	40
35	50
36	50
37	30
38	40

4.2. Conclusion

In this study, the data of the 2017/2018 season of the English Premier League are used. Based on the first 19 games of this season, the results of the next 19 games are predicted. To make these estimates, 6 classification models are utilized and compared. The best result is obtained by Tpot Classifier with a performance of 82%.

While the Tpot classifier succeeds with an 82% accuracy on the train set, the test set provides 70% accuracy on the 20th week. The accuracy rate decreases to 30% in the following weeks. Overall accuracy on the test set is calculated as 48%.

The following items can be mentioned as the reasons why the accuracy rate, which starts at 70% as of the 20th week, decreases to 48% overall:

- The performance of the match played in the 20th week has 70% accuracy, since it is close to the performance data of the previous 19 weeks. However, the closeness of the performances in the other weeks to the performances in the baseline weeks cannot be that much accurately predicted. Accuracy is 60% in the weeks close to baseline but, decreases to 40% in the following weeks.
- Small sample size in the data set and higher number of features may have caused overfitting. Although the number of features has been reduced to 20, the small sample size may not have reduced the overfitting issue sufficiently.
- There are also non-technical data that are not included in the data set, but this affects the performances of the players. There are many non-technical variables before and during the match. Among these weather report, referee effect, psychological conditions of players can be mentioned. This non-technical data has led to unpredictable results on performance.

In order to improve this model, the following ideas can be developed for future studies: instead of 19 weeks based, performance data of previous weeks for weeks of test set may be added to the train set. Instead of averaging 19 weeks, accuracy can be increased by considering different mathematical formulas. The number of seasons in the train set can be increased and more suitable results can be obtained from the test set. This is also expected to prevent overfitting. If some of the non-technical data can be calculated or predicted with a certain prediction, the model can be made more realistic.

In addition to the model comments, when the performance data is examined, the following suggestions can be listed to increase the number of wins of a team: when Figure

6 is examined, if a team wants to win the match, they must increase the number of shots. It would be beneficial to develop programs to improve this. Again, when Figure 6 is examined, the second important factor is the number of cleen sheets which represents the defensive power of the team. In this case, the increase in the performance of blocking the goal to win the match is one of the other effective features to win. Another important factor, the direct free-kick off target, is very effective on the result for home teams. Freekicks which can not be scored for the away team are an advantage for homeowners. For away teams, more organized and more studied scenarios can be developed on this feature.

5. REFERENCES

- Baio, G., Blangiardo, M. (2010). Bayesian hierarchical model for prediction of football results. *Discovery, University College London.*, 2010.
- Buursma, D. (2011). Predicting sports event from past result: Towards effective betting on football matches. *Preceding 14thTwente Student Conference on IT. University of Twente, Faculty Electrical Engineering, Mathematics and Computer Science, Netherlands*, 2011.
- Confusion matrix in machine learning. (2019). Retrieved from: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
- Farzin, O., Parinaz, E., Faezeh, S. M. (2013). Football result prediction with Bayesian network in Spanish league-Barcelona team. *International Journal of Computer Theory and Engineering*, *5*(5), 2013, 812-815.
- Football. (2019, July 8). Retrieved from: https://www.britannica.com/sports/football-soccer
- Gevaria, K., Sanghavi, H., Vadiya, S., Deulkar, K. (2015). Football match winner prediction. *International Journal of Emerging Technology and Advanced Engineering, Volume 5, Issue 10, October 2015.*
- How does margin of error work? (2019). Retrieved from: https://www.statisticssolutions.com/how-does-margin-of-error-work/
- Moorthy, S., Shi, Z., Zimmerman, A. (2013). Predicting college basketball match outcomes using machine learning techniques: Some results and lessons learned, (2013), arXiv preprint: 1310.3607.
- Randal S. Olson, Ryan J. Urbanowicz, Peter C. Andrews, Nicole A. Lavender, La Creis Kidd, and Jason H. Moore (2016). Automating biomedical data science through tree-based pipeline optimization. *Applications of Evolutionary Computation*, pages 123-137.
- Success criteria in classification models. (2018, May 12). Retrieved from: https://medium.com/data-science-tr/s%C4%B1n%C4%B1fland%C4%B1rma-modellerinde-ba%C5%9Far%C4%B1-kriterleri-2d86488799c6
- Tpot. (2016, March 15). Retrieved from: https://epistasislab.github.io/tpot/

- Tüfekçi, P. (2016). Prediction of Football Match Results in Turkish Super League Games. In A. Abraham, K. Wegrzyn-Wolska, E. A. Hassanien, V. Snasel & M. A. Alimi (Eds.), Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015 (pp. 515-526).
- What is a correlation matrix?. (2018, August 16). Retrieved from: https://www.displayr.com/what-is-a-correlation-matrix/