Decision Analytics Journal 8 (2023) 100296

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression

Matt Gifford [a], Tuncay Bayrak [b,*]

[a] Fingerpaint Group, Saratoga Springs, NY, USA
[b] Western New England University, College of Business, 1215 Wilbraham Rd. Springfield, MA, USA

ARTICLE INFO

Keywords:
Predictive analytics
Decision tree
Logistic regression
National Football League
Sports analytics
Statistical models

ABSTRACT

Sports analytics has gained rapid popularity in recent years and will likely continue to evolve. In this study, we construct predictive analytics models to forecast the NFL games outcomes in a season using decision trees and logistics regression. Several variables are used as predictors (independent variables). The binary win–loss outcome measure is used as a target (dependent) variable. Decision tree and binary logistic regression models are constructed to describe the relationships between the predictors and football game outcomes in the NFL.

1. Introduction

Sports analytics has emerged as a field of research and has had a significant impact on transforming sports-related data into valuable insights for better decisions such as analyzing athlete performance, predicting the outcome of a given game, evaluating the strengths and weaknesses of opponents, and optimizing team performance. In other words, sports analytics allows coaches to understand the science behind athlete performance and game prediction.

Sports analytics may be defined as "*the process of using sports-related data to find meaningful patterns and communicate those patterns to help make decisions*", [1]. This definition suggests that sports analytics employs the application of descriptive and predictive analytics to sports.

Recent advances in data analytics and data management tools have allowed team managers and coaches to improve decision-making by constructing predictive models to anticipate team and player performance [2]. As argued by Metulini and Gnecco [3], professional sports team managers and staff face the need for extracting useful information for the monitoring of the performance of their teams, as well as their athletes. Similarly, Steinberg [4] holds that sports analytics has enabled sports entities and their players to be more efficient, and it will impact every aspect of high school, collegiate, and professional sports.

Sports Analytics plays a key role in supporting teams, players, and coaches to improve performance [5]. Further, valuable insights generated through sports analytics can be used for team composition, athlete career assessment and improvement, and future predictions [6]. Consequently, as pointed out by Steinberg [4], sports analytics is the present and future of professional sports.

Analytics continues to grow in the National Football League (NFL), and almost every major professional sports team either has an analytics department or an analytics expert on staff [4]. Therefore, through the development of descriptive and predictive analytics models, one can identify measurable accuracy in how various team statistics interact with the outcome of a game. Which team statistics are most significant — offensive stats such as passing yards, defensive stats such as turnovers forced, or other factors such as how many wins a team has that season prior to a game? [7].

While data analytics helps coaches and managers evaluate and assess the performance of their athletes to improve the team performance, as argued by Gifford and Bayrak [7], there is limited evidence available regarding the construction of predictive models to determine which factors and more specifically which team statistics have the largest impact on winning. In other words, sports forecasting models may involve restrictions. For instance, most of the data used to predict the outcome of any given game is derived from relatively small data sets. Our study on the other hand involves a relatively large data set. Additionally, constraints such as time in possession, playing home vs. away, and the complexity and volume of sports data may also create problems with identifying the most important variables with the largest impact on winning. In this study, through decision tree and binary logistic models, we were able to identify the most important team statistics with the largest impact on the target variable of win/loss. As argued by Davis et al. [8], the predictive models themselves may also have to be evaluated in terms of their accuracy which may be complicated by the non-stationary nature of sports data. Our models were able to predict the outcomes of the NFL games with impressive

* Corresponding author.
E-mail addresses: matthewgifford17@gmail.com (M. Gifford), tbayrak@wne.edu (T. Bayrak).

https://doi.org/10.1016/j.dajour.2023.100296
Received 29 October 2022; Received in revised form 16 July 2023; Accepted 31 July 2023
Available online 7 August 2023

accuracy. Therefore, using the 2002–2018 NFL data, this study focuses on developing and constructing predictive models to quantify the influence of team statistics on regular season wins, assess, and evaluate the constructed models to determine the best predictive model, and validate the best predictive model by applying it to the 2018 NFL regular season games and comparing it to the actual season standings. Throughout this analysis, we used SAS Enterprise Miner, an advanced data analytics tool.

2. Background

Advancements in data collection and data analytics have fueled data-driven decision-making and the use of analytics in sports. Hence, as pointed out by Fury et al. [2], the proliferation of advanced analytics has changed the understanding of individual and team performance in the 21st century. One would argue that advanced statistics and sports analytics have become vital in the analysis of statistics pertaining to various games and teams. Moreover, the rise of Data Science and related fields of Big Data, Machine Learning, and Deep Learning has transformed the sports analytics landscape and has changed the way various sports are played to different degrees. Thus, in recent times, sports institutions and clubs have given increased importance to such research that will ultimately help them have a competitive edge over rivals [9]. Consequently, Goes et al. [10] suggest that sports analytics may assist sports science in addressing data management challenges and finding new insights.

Over the years, teams and coaches have relied on the use of analytics and employed sports analytics to facilitate decision-making. For instance, using the data collected between 1995 and 1997, Joseph et al. [11] investigated the performance of expert-constructed Bayesian Networks (BNs) and compared it with other machine learning techniques for predicting the outcome (win, lose, or draw) of matches played by Tottenham Hotspur Football Club. The authors conclude that the expert BN is generally superior to the other techniques for this domain in predictive accuracy. Nunes and Sousa [12] conducted a similar study in which they used data association rules, classification, and visualization techniques to find patterns in datasets from several European championships. They maintain that their exploratory work confirmed several well-known patterns in football, and found that among the several techniques used visualization produced the best results.

A study carried out by McCabe and Trevathan [13] employed artificial intelligence for the prediction of sporting outcomes. Their model attempted to capture the quality of various sporting teams. The authors claim that their system performed well. Davoodi and Khanteymoori [14] applied Artificial Neural Networks (ANNs) to horse racing prediction. The authors utilized real horse racing data collected from AQUEDUCT Race Track in NY, which included 100 actual races from 1 January to 29 January, 2010. Their results show that ANNs are appropriate methods in the horse racing prediction context. Using eight years of data and three popular data mining techniques namely artificial neural networks, decision trees, and support vector machines, Delen et al. [15] have developed both classification and regression-type models to assess and evaluate the predictive abilities of different methodologies and techniques. They argue that the classification-type models predict the game outcomes better than regression-based classification models, and of the three classification techniques, decision trees produced the best results.

Maszczyka et al. [16] ran a study that involved a group of 116 javelin throwers to make a comparison between regression and neural models with respect to their accuracy in predicting sports results. Their analysis shows that the neural model does better at predicting sports results than the regression model. Using thirteen seasons of Dutch Eredivisie match data, Tax and Joustra [17] employed several combinations of dimensionality reduction techniques and classification algorithms to predict match outcomes. The authors hold that the highest prediction accuracy was achieved by using a combination of Principal Component Analysis (PCA) with a Naive Bayes or Multilayer Perceptron classifier. To evaluate the impact NBA players have on their teams' chances of winning, Deshpande and Jensen [18] employed a win probability framework. The authors propose a Bayesian linear regression model to estimate an individual player's impact, after controlling for the other players on the court.

Young et al. [19] modeled the relationships between player actions and match outcomes in Australian Football by utilizing a wide range of performance indicators (PIs) and a longer time frame for the development of predictive models. In their study, the authors used the categorical Win–Loss and continuous Score Margin match outcome measures as dependent variables, and Ninety-one team PIs from the 2001 to 2016 Australian Football League seasons as independent variables. Kapadia et al. [20] adopted machine learning techniques including Naïve Bayes, Random Forest, and K-Nearest Neighbor (KNN) to predict cricket match results based on historical match data. They claim that tree-based models particularly Random Forest performed better in terms of accuracy and precision when compared to probabilistic and statistical models.

More recently, using what is called Goal Impact Metric (GIM), Liu et al. [21] employed a Deep Reinforcement Learning (DRL) model to rank and measure soccer players' overall performance in the English Football League Championship. The authors claim that GIM is a temporally stable metric, and its correlations with standard measures of soccer success are higher than that of computed with other state-of-the-art soccer metrics. In a different study, using the event and tracking data of soccer matches, Rahimian and Toka [22] utilized the same DRL technique to discover the optimal actions in different situations of a soccer game. They argue that what is called an "optimization framework" will aid players and coaches in understanding the optimal actions in any given soccer game. Additionally, by employing a large dataset that includes all of the teams from the top five European national championship leagues, including detailed player and team statistics, Toma and Campobasso [23] used an ordered logit model to better understand how a team's performance can be predicted based on its tactical and financial choices. Their study concluded that the results differ between the top-tier and medium-to-bottom tier teams.

Nguyen et al. [24] performed deep learning, regression, and classification analysis to predict basketball players' future performance. Their results suggest that scoring by the primary players is the most important factor for any team. Teeselink et al. [25] conducted a study involving large samples of Australian football, American football, and rugby matches to see whether being slightly behind increases the likelihood of winning. The authors find no evidence of such an effect on the three sports. By analyzing the 2018 European Men's Handball Championship games, Romero et al. [26] conducted a study to evaluate team performance based on the weighted aggregation of statistical indicators. The authors ran a principal component analysis (PCA) to examine the relationship between each game's statistical indicators and used a fuzzy multi-criteria decision-making method to predict the player of the match in any given game. A study carried out by Duran [27] investigated how sports analytics can be employed to tackle the main challenges and problems facing sports scheduling in Latin American countries.

Several studies examined how sports data visualization techniques can be utilized to analyze human behavior patterns in sports games. For instance, Du and Yuan [28] conducted a comprehensive study to demonstrate how sports analytics can be used to analyze various sports data visualization techniques. The authors claim that they attempt to provide guidelines in helping readers to find appropriate techniques for different sports data. Similarly, to investigate players physically, Li et al. [29] employed artificial intelligence and data visualization techniques to analyze a group of diving athletes to improve the recognition of motion effects based on image recognition technology.

With respect to using various statistical models to predict the outcomes of NFL games, in one of the earliest studies [30] employed

Table 1
Variable names and labels.

Variable	Description
def_1st_down	Total 1st downs allowed by defense
def_pass_yds	Total passing yards allowed by defense (includes loss sack yardage)
def_rush_yds	Total rushing yards allowed by defense
def_total_yds	Total yards allowed by defense
def_turnovers	Turnovers gained by defense
losses	Number of previous losses for the season
off_1st_down	Total 1st downs gained on offense
off_pass_yds	Total yards gained by passing (includes loss sack yardage)
off_rush_yds	Total yards gained by rushing
off_total_yds	Total yards gained by offense
off_turnovers	Team turnovers lost
opp_score	Points allowed
tm_score	Points scored
wins	Number of previous wins for the season

linear models to develop a procedure for predicting the outcomes of National Football League games. The predictions for 1,320 games played between 1971 and 1977 had an average absolute error of 10.68. A similar study by Boulier Bryan and Stekler [31] evaluated power scores as predictors of the outcomes of NFL games for the 1994–2000 seasons. The evaluation involved a comparison of forecasts generated from probit regressions based on power scores published in The New York Times with those of a naive model, the betting market, and the opinions of the sports editor of The New York Times. They concluded that the betting market is the best predictor followed by the probit predictions based on power scores.

Steinberg [4] argues that the origin of sports analytics can be traced to Oakland Athletics' General Manager, Billy Beane who employed data analytics to discover value players when constructing the 2002 Athletics roster. He postulated that a team comprised of players with high on-base percentages was more likely to score runs thus translating into more wins. The popularity of data-driven decision-making in sports has significantly grown since then and is now making headway in professional NFL Football. For instance, the Philadelphia Eagles currently have an analytics team that provides guidance on making critical football decisions. By employing a variety of data analytics tools and models, the team can make more informed decisions on when to attempt for a two-point conversion after a touchdown or keep the offense on the field on fourth down [32].

Bill Barnwell, who is a sportswriter for ESPN.com, contributed to the sports analytics domain by determining which team statistics better evaluate the quality of a team and the probability of future wins aside from their current season record. He hypothesized that one of the best metrics was Defense-adjusted Value Over Average (DVOA), which compares the success of a team during a given play with the expected result when factoring the situation, down, distance and the opponent [33]. A study carried out by Rudy [34] took a different approach, focusing on turnovers and how the season as a whole is impacted. The author concluded that 44% of the variation in a team's winning percentage was explained by the means of a team's turnover differential. To determine the importance of passing and rushing in comparison to team performance, Feng [35] developed scatterplots based on efficiency during the regular season. Efficiency refers to the yards per play gained on offense subtracted by yards per play allowed on defense. Feng (n.d.) argues that passing efficiency is more significant since many playoff and Super Bowl contenders excelled in passing while rushing efficiency is extremely scattered. Approximately 88% of playoff teams from 2003–2012 gained more yards per play than they allowed [35].

Using data from the 2000–2011 NFL seasons, a team of data scientists at MIT developed a logistic regression model to understand the most influential factors of field goal success. The team concluded that almost all environmental factors had a significant impact on field goal success. Moreover, the same study revealed that all situational or psychological factors were insignificant including whether the kick was attempted in a regular-season or postseason game, a home or away game, a high pressure or low-pressure situation, and whether a timeout was called before the kick was attempted [36]. A similar study conducted by Pelechrinis and Papalexakis [37] produced a simple predictive model that can quantify the impact of various factors on the probability of winning a game of American football. Finally, using the JMP software, a data scientist developed a neural network model to predict wins based on a number of input variables including statistics from the previous seasons, and the salary teams pay for each position. The study concluded that the team's number of wins and point differential from the previous season were the top predictors [38].

3. Research method

3.1. Sample

Through the use of Pro Football Reference and its extensive historical database of football metrics and statistics, we compiled a sample of sixteen seasons worth of data to analyze and evaluate. The sample was compiled from all regular season games from the 2002–2017 seasons. The data in the sample starts with the 2002 season as it marks the introduction of the Houston Texans as the 32nd team into NFL team. This allows for all data which matches the current makeup of the NFL while still remaining relevant to the way the game is played and the strategies employed in 2017. This sample provides the same number of teams competing and games occurring in each year of data. In total, the sample includes a total of 4096 games. Table 1 lists the variable names and their labels used in this study. Sample statistics pertaining to the variables employed in this study are summarized in Table 2.

3.2. Data preparation

Before the creation of the two models, precautionary measures were taken to ensure the data was properly cleaned for accurate analysis. These steps included removing ties from the data set, converting the target variable, and creating necessary dummy variables. Throughout the 4096 games over the sixteen seasons, only seven ties occurred. Since this only accounted for approximately 0.17% of the total sample, they were ruled insignificant and removed from the data set. Additionally, the target variable of outcome, which was previously marked as a "W" or "L" was converted into a binary, 1 or 0 with "1" signifying a win. Lastly, in order for proper analysis through SAS Enterprise Miner, dummy variables were developed for variables such as Overtime and Home (whether the location of the game was home or away).

3.3. Model creation

In order to determine which team statistics have the greatest impact on the outcome of a game, we selected two models to build, one classification model and one estimation/prediction model. The first

Table 2
Sample statistics.

Variable	Role	Mean	Standard Dev.	Non-missing	Missing	Min.	Median	Max.	Skewness	Kurtosis
def_1st_down	INPUT	19.13	4.99	8178	0	3	19	40	0.08	−0.019
def_pass_yds	INPUT	221.20	77.74	8178	0	−7	217	522	0.26	−0.006
def_rush_yds	INPUT	113.89	51.59	8178	0	−18	107	378	0.77	0.882
def_total_yds	INPUT	335.04	84.93	8178	0	26	335	653	0.03	−0.05
def_turnovers	INPUT	1.62	1.35	8178	0	0	1	8	0.84	0.586
losses	INPUT	3.74	2.96	8178	0	0	3	15	0.71	−0.089
off_1st_down	INPUT	19.13	4.99	8178	0	3	19	40	0.08	−0.019
off_pass_yds	INPUT	221.18	77.74	8176	2	−7	217	522	0.26	−0.007
off_rush_yds	INPUT	113.89	51.59	8178	0	−18	107	378	0.77	0.882
off_total_yds	INPUT	335.02	84.93	8178	0	26	334	653	0.93	−0.051
off_turnovers	INPUT	1.62	1.35	8178	0	0	1	8	0.84	0.058
opp_score	INPUT	21.91	10.28	8178	0	0	21	62	0.3	−0.053
tm_score	INPUT	21.91	10.28	8178	0	0	21	62	0.3	−0.053
wins	INPUT	3.74	2.95	8178	0	0	3	15	0.71	−0.085

model created was a decision tree. A decision tree is a powerful tool that can assist in determining which teams in the NFL are more likely to win based on the analyzed data. The analysis does so by determining which attributes a team should possess in order to have the best chance of winning a game. Therefore, by understanding the most important attributes to winning and which team possesses them, we can conclude the outcome of a particular game. A decision tree takes the collection of data and runs it through individual variables determining a cutoff value for each one and then splits the data according to the cutoff value [39].

With the completion of the decision tree, we generated a binary logistic regression as our estimation/prediction model. Since the target variable, winning, is binary, as it is limited to winning and losing, a logistic regression is a suitable model to run. Logistic regression is basically an extension of multiple regression in situations where the dependent variable (DV) is categorical or discrete and may have as few as two values such as membership or non-membership in a group or completion or non-completion of an academic program [40]. In logistic regression, instead of predicting the value of a variable Y from a predictor variable $X1$ or several predictor variables (Xs), the probability of Y occurring given known values of $X1$ or Xs is predicted [41]. In a typical logistic regression analysis, there will be one binary dependent variable and a set of independent or predictor variables that may be either dichotomous or quantitative or some combination thereof [42,43].

The general form of the logistic regression equation with n number of independent variables (predictors) is as follows (see Meyers et al. [43]):

$$ln(odds) = a + b_1X_1 + b_2X_2 + \ldots\ldots + bnXn$$

The second stage of generating the predicted probability of target group membership is to use the log odds in the following expression to generate the predicted chance of a case being in the target group. In other words, the logistic regression equation from which the probability of Y is predicted is given by (see Meyers et al. [43]):

$$P(Y) = \frac{1}{1 + e^{-(a+b1X1+b2X2+\cdots\ldots+bnXn)}}$$

where $P(Y)$ is the predicted probability of target group membership, e is the base of natural logarithms, and the other coefficients form a linear combination much the same as in multiple regression (Meyers. The resulting value from the equation varies between 0 and 1 [41,43]. Mathematically, rather than the least-squares estimation procedure, logistic regression uses a maximum likelihood estimation procedure, which selects coefficients that the observed values are most likely to have occurred.

4. Data analysis and results

4.1. Decision tree

We first generated and built a decision tree model to determine which teams in the NFL are more likely to win based on a number of predictors and input variables, and to represent the outcomes of the games.

Table 3
Variables and their log worth values.

Target Variable: outcome			
Variable	Variable description	-Log(p)	Branches
off_turnovers	off_turnovers	122.30	2
def_turnovers	def_turnovers	112.30	2
off_rush_yds	off_rush_yds	96.67	2
def_rush_yds	def_rush_yds	95.29	2
off_total_yds	off_total_yds	72.05	2
def_total_yds	def_total_yds	64.14	2
def_1st_down	def_1st_down	58.37	2
off_1st_down	off_1st_down	55.90	2
home	home	18.55	2
wins	wins	13.14	2
losses	losses	9.42	2
off_pass_yds	off_pass_yds	7.14	2
def_pass_yds	def_pass_yds	5.97	2
overtime	overtime	0.40	2

In predictive modeling, as a common practice, the data source is often partitioned into training and validation sets to assess the quality of model generalization. While the training data set is used to create and fit the preliminary model, the validation data set is employed to measure overall performance and optimize the selected model. In this case, we are assigning 50% of the data for training, and 50% of the data for validation. This means that 50% of the data will be used for fitting the model to the data, and 50% of the data will be used to optimize and fine-tune everything to improve the model's generalization.

A portion of the decision tree generated in this analysis is illustrated in Fig. 1.

By looking more closely at the tree we can determine which variables are the most important in this predictive analysis. Since the variable "off-turnovers" is at the top of the tree and has the highest log-worth value of 122.30 (Table 3), it is the most important variable in predicting the winner. This also means that this was the variable used for the first split.

To determine the importance of input variables in predicting or classifying the target variable and validate the conclusions about which variables are the most significant in generating the predictive decision tree model we can look at the "variable importance" values tabulated in Table 4. As seen, through the decision tree model, we determined that the most important team statistic in determining the winner of an NFL game is *off_turnovers*. The four most important variables in order include *off_turnovers* (turnovers lost), *def_turnovers* (turnovers forced), *off_total_yds* (total yards on offense), and *def_total_yds* (total yards allowed on defense) (Table 4).

A decision tree discovers the relationship between the target variable and the input variables and generates a set of decision rules to predict the values of new observations. In other words, the decision

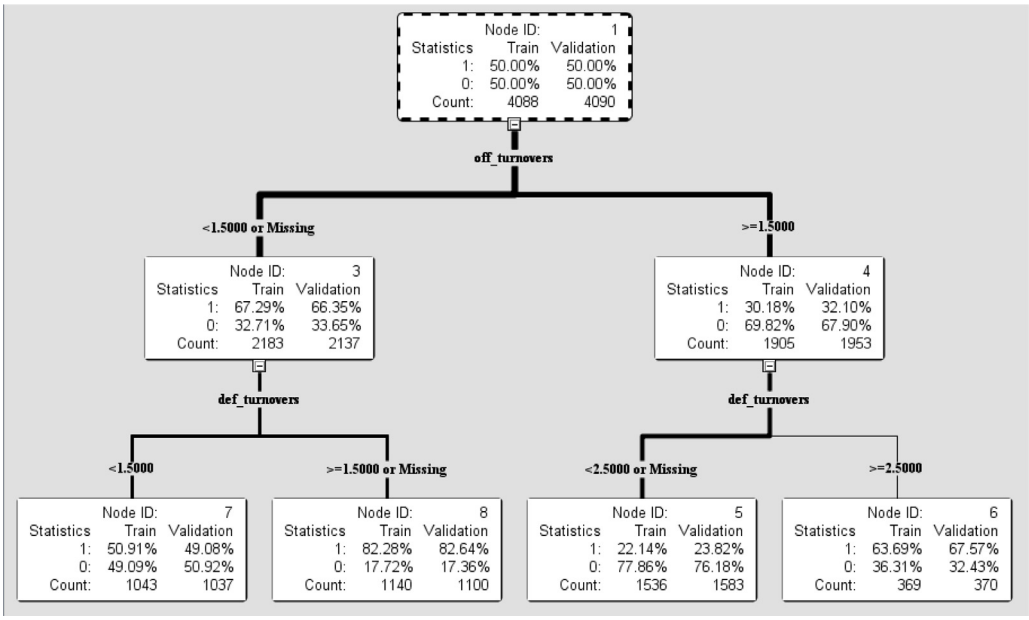

Fig. 1. Decision tree.

Table 4
Full tree input variables and their importance.

Variable Name	Number of splitting rules	Train importance	Validation importance
off_turnovers	4	1.0000	1.0000
def_turnovers	5	0.8548	0.9942
off_total_yds	5	0.6099	0.4557
def_total_yds	3	0.5328	0.5560
def_rush_yds	4	0.4722	0.4494
off_rush_yds	1	0.2501	0.1498
off_pass_yds	2	0.2261	0.1915
def_1st_down	1	0.1990	0.1633
home	0	0.0000	0.0000
off_1st_down	0	0.0000	0.0000
def_pass_yds	0	0.0000	0.0000
losses	0	0.0000	0.0000
wins	0	0.0000	0.0000
overtime	0	0.0000	0.0000

rules are employed to assign new observations from the data set to a node. Values in each node can also be interpreted using the decision rules. For instance, looking at node 4, according to the model, only 32% of teams won a game when turning the ball over at least twice on offense. Additionally, approximately 83% of teams won if they turned the ball over on offense no more than once and forced two or more turnovers on defense (Fig. 1, and Table 5).

Since our target variable is a binary variable, model performance or model fit can be judged by its misclassification rate, which measures the fraction of cases where the decision does not match the actual target value. In other words, the misclassification rate measures how often the predictive model makes an incorrect prediction [44]. Since this study involves a binary classification problem (winning or losing a game), misclassification is an appropriate measure of model performance. As argued by Zhou et al. [45], the misclassification rate is an important index for the evaluation of classification algorithms since the ultimate goal of classification is to reduce the misclassification rate of testing data and produce accurate predictions.

To determine the accuracy of this decision tree, we can look at the "model fit statistics" summarized in Table 6, and "subtree assessment plot" depicted in Fig. 2.

The subtree assessment plot depicts the proportion misclassified versus the number of leaves. As seen, the lowest misclassification rate occurs in leaf 26. According to this graph, the current model has a

Table 5
Node 4 statistics.

Statistics	Train	Validation
Count	1140	1100
Prediction	1	1
% with target = 1	82.28%	82.64%
% with target = 0	17.72%	17.36%
% correctly predicted	82.28%	82.64%

Table 6
Model fit statistics.

Target	Statistics label	Train	Validation
Outcome	Misclassification rate	0.1854	0.2166

roughly 21% misclassification rate. This means that it is 79% accurate at predicting whether or not a team will win a game.

Fig. 2 shows the misclassification rate corresponding to each tree in the sequence as the data is sequentially split. As seen, the performance on the training sample becomes better as the tree becomes more complex. However, the performance on the validation sample improves up to a tree of twenty-six leaves, and then diminishes as model complexity increases. In other words, precision slowly diminishes as complexity increases past leaf twenty six. The validation performance

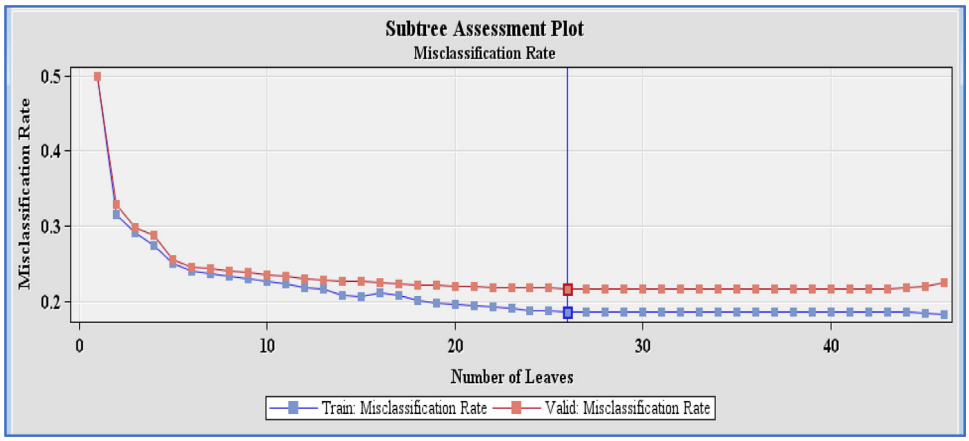

Fig. 2. Subtree assessment plot.

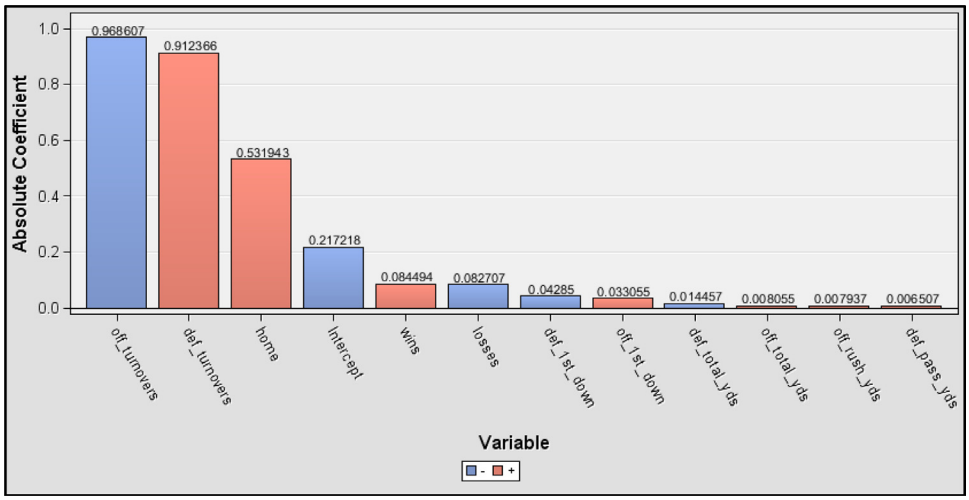

Fig. 3. Variables and their coefficients.

under misclassification rate shows that the optimal tree appears to have twenty-six leaves.

As seen, the accuracy of the model does not get better with the number of leaves beyond twenty-six leaves. Thus, the optimal tree is the one with twenty-six leaves.

4.2. Binary logistic regression

As pointed out before, when the dependent (target) variable has two possible outcomes, a binary logistic regression analysis is often employed to calculate the probability of the event. In running this regression, we were able to determine which variables or team statistics have the largest impact on determining the target variable of winning. In the process, this analysis displays which variables are meaningless and negligible and therefore should not be included in the model. Thus, the result is a refined model consisting of important variables which can aid us in prediction.

As summarized in Table 6, the maximum likelihood estimation method for this dataset yielded the following predictors along with their *b* coefficients, *p*-values for testing whether a particular variable is significantly associated with the target variable, and *odds ratios* (exp estimates). Odds ratios indicate the amount of change expected in the log ratios when there is a 1-unit change in the predictor variable with all the other variables in the model held constant. As seen in Table 6, the *p*-value of all variables listed is less than 0.05, which is our chosen alpha, therefore they will all be included in the final model. As a note, the variables listed below are a subset of the original list of variables. Through the estimates column, SAS displays the coefficients for each variable within the model which are all part of the equation to determine the predictive value.

Based on the refined model created in the analysis, we can predict the outcome of a regular-season game from the team statistics. Through the binary logistic regression, the most important variable is offensive turnovers which has a heavy negative effect (−0.9686) on the outcome while defensive turnovers is slightly less important and has a strong positive impact (0.9124). Other notable takeaways include that offensive rushing yards and defensive passing yards are included in the final model but offensive passing yards and defensive rushing yards are not and home field advantage has an impact (Table 7).

In addition, we can also look at the "(Exp (Est)" "or "Odds Ratios" column to determine which variable produces the greatest outcome and has the greatest impact on improving the probability of the target variable. A general rule is that the variable with the highest point estimate value is considered to have the greatest impact on the dependent variable. In this case, the variable *def-turnovers* has the highest point estimate value of 2.49. This means that a one-unit change in this variable contributes the most in determining whether or not a team will win any given game in comparison to the other input variables.

Fig. 3 depicts the input variables along with their coefficients in terms of their absolute values. As seen, variables *off_turnovers* and *def_turnovers* have the highest coefficients, suggesting that they are the most important variables when it comes to predicting which team would win any given game. As pointed out before, the general form of

Table 7
Logistic regression variables and coefficients.

Analysis of maximum likelihood estimates							
Parameter	DF	Estimates	Standard error	Wald Chi-Square	Pr > Chi-Square	Standardized estimates	Exp(Est)
Intercept	1	0.0488	0.3258	0.44	0.5049		0.805
def_1st_down	1	−0.0428	0.0158	7.35	0.0067	−0.1186	0.958
def_pass_yds	1	0.00651	0.00108	36.28	<.0001	0.2782	1.007
def_total_yds	1	−0.0145	0.00131	121.83	<.0001	−0.676	0.986
def_turnovers	1	0.9124	0.0435	440.87	<.0001	0.6785	2.49
home	1	0.5319	0.0923	33.19	<.0001	0.1467	1.702
losses	1	−0.0827	0.0165	25.12	<.0001	−0.1333	0.921
off_1st_down	1	0.0331	0.0162	4.16	0.0415	0.091	1.034
off_rush_yds	1	0.00794	0.00106	55.89	<.0001	0.2291	1.008
off_total_yds	1	0.00806	0.000991	66.13	<.0001	0.378	1.008
off_turnovers	1	−0.9686	0.0444	475.9	<.0001	−0.7211	0.38
wins	1	0.0845	0.0161	27.55	<.0001	0.1387	1.088

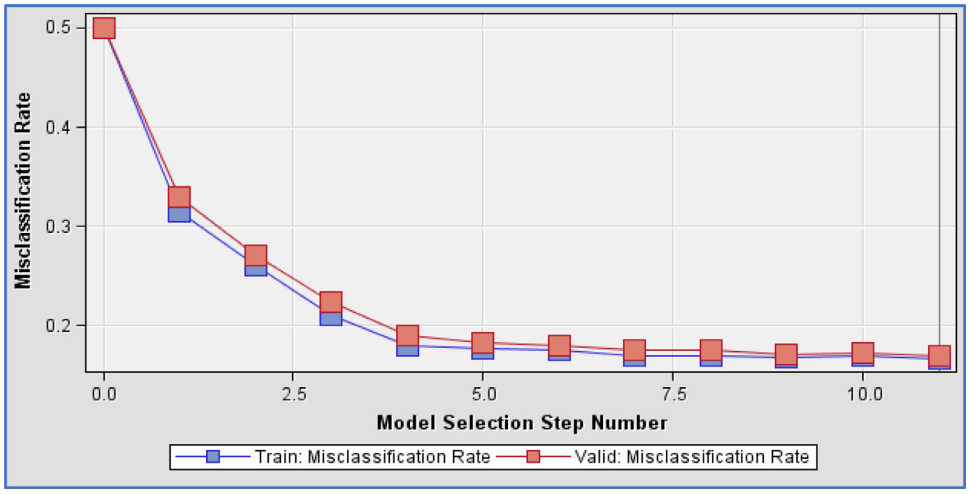

Fig. 4. Summary of stepwise selection method.

the logistic regression equation with n number of independent variables (predictors) is as follows:

$$ln(odds) = a + b_1X_1 + b_2X_2 + \ldots\ldots + b_nX_n$$

The equation, in this case, would be:

$$\begin{aligned} ln(odds) = {} & 0.0488 - 0.0428 * \text{def\_1st\_down} + 0.00651 * \text{def\_pass\_yds} \\ & - 0.0145 * \text{def\_total\_yds} + 0.9124 * \text{def\_turnovers} \\ & + 0.5319 * \text{home--0.0827} * \text{losses} + 0.0331 * \text{off\_1st\_down} \\ & + 0.00794 * \text{off\_rush\_yds} + 0.00806 * \text{off\_total\_yds} \\ & - 0.9686 * \text{off\_turnovers} + 0.0845 * \text{wins}. \end{aligned}$$

Then by plugging any given values for a particular team into the following equation a decision can be made as to whether or not that team will win the game. If the final result is closer to 1, then the team will likely win the game. If the final result is closer to 0, then the team will not likely win the game.

$$P(Y) = \frac{1}{1 + e^{-(a+b_1X_1+b_2X_2+\cdots+b_nX_n)}}$$

In this analysis, we run a "Regression" node using the stepwise selection model. Fig. 4 depicts the stepwise selection method that was employed in the regression model. In total, the model went through 11 steps in order to determine which variables were the most significant in determining whether or not a team would win a game. After the 11 steps were completed, the optimal model was achieved. Neither adding or dropping a variable would improve the accuracy of the model after step 11.

Fig. 5 shows the variables and the steps in which they were added to the final predictive model. As seen, variable *off-turnovers* was added to the model in step one and was retained in the model up until the final step. Similarly, variable *def-turnovers* is the second most important variable that was added to the model in step 2. Less important variables were added to the model in later steps.

Table 8
Fit statistics.

Target	Fit statistics	Train	Validation
Outcome	Misclassification Rate	0.167	0.169

The final aspect to look at for this regression model is the accuracy of it. Since "win" is a binary variable, we are making a decision. With decisions, it is important to look at the *misclassification rate*. Table 8 shows that the logistic regression model developed in this study has a misclassification rate of 16.9%. This means that the model can correctly predict winning teams about 83.1% of the time.

4.3. Model comparison

After comparing the Decision Tree and Binary Logistic Regression models, the Binary Logistic Regression is deemed to be the final model since it displays a lower misclassification rate. The misclassification rate for the Decision Tree model is approximately 0.216 while the Binary Logistic Regression is 0.169 (Table 9).

4.4. Model validation

Upon completion of the Binary Logistic Regression, the same fourteen metrics from the 2018 NFL regular season were collected. Using

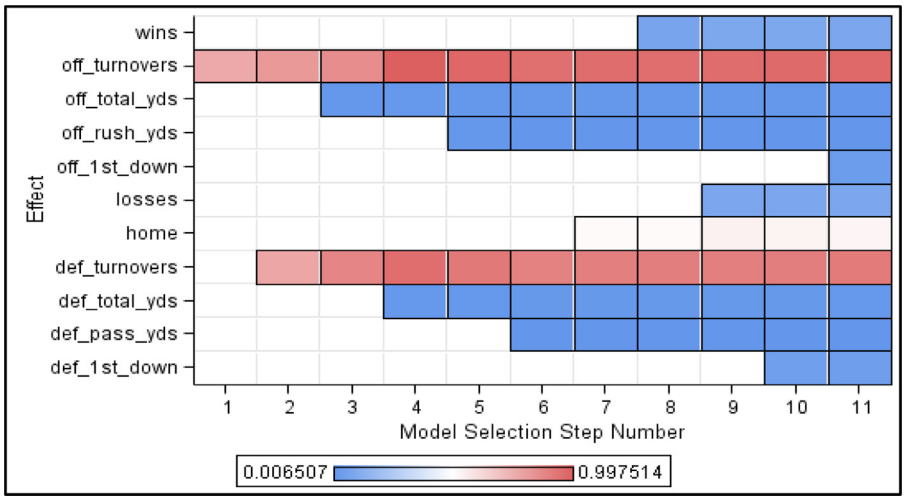

Fig. 5. Variable selection plot.

Table 9
Model comparison fit statistics: Model selection based on valid misclassification rate.

Selected model	Model description	Train misclassification rate	Valid misclassification rate
Y	Binary Logistic Regression	0.167	0.169
	Decision Tree	0.185	0.216

Table 10
Model validation.

Team	Actual			Predicted		Team	Actual			Predicted	
AFC East	W	L	T	W	L	NFC East	W	L	T	W	L
New England P*	11	5	0	10	6	Dallas Cowboys*	10	6	0	10	6
Miami Dolphins	7	9	0	7	9	Philadelphia E+	9	7	0	9	7
Buffalo Bills	6	10	0	9	7	Washington R	7	9	0	9	7
New York Jets	4	12	0	3	13	New York G	5	11	0	6	10
AFC North						**NFC North**					
Baltimore R*	10	6	0	11	5	Chicago Bears*	12	4	0	14	2
Pittsburgh S	9	6	1	9	7	Minnesota Vikings	8	7	1	10	6
Cleveland B	7	8	1	9	7	Green Bay Packers	6	9	1	5	11
Cincinnati B	6	10	0	5	11	Detroit Lions	6	10	0	5	11
AFC South						**NFC South**					
Houston T*	11	5	0	13	3	New Orleans S*	13	3	0	9	7
Indianapolis C+	10	6	0	7	9	Carolina Panthers	7	9	0	9	7
Tennessee T.	9	7	0	8	8	Atlanta Falcons	7	9	0	8	8
Jacksonville J	5	11	0	7	9	Tampa Bay B	5	11	0	4	12
AFC West						**NFC West**					
Kansas City C*	4	0	9	7	4	Los Angeles R.*	13	3	0	12	4
Los Angeles C+	4	0	9	7	4	Seattle Seahawks+	10	6	0	11	5
Denver Broncos	10	0	8	8	10	San Francisco 49ers	4	12	0	4	12
Oakland Raiders	12	0	4	12	12	Arizona Cardinals	3	13	0	3	13

the equation, which was developed through the model, each team within each game was assigned a Win Value. Win Values were then compared for every game within the season and the team with the higher Win Value was assigned a win with the other team being assigned a loss. After all games were completed, regular season records were determined and placed in league standings, abiding by NFL Tiebreaking Procedures. Table 10 displays the final result with the comparison of the model's predicted standings with the actual standings and results from the 2018 regular season. With two ties occurring in the 2018 season, we used 254 of the 256 games for comparison. After comparison, we determined that the results of the games were correctly predicted approximately 83.07% of the time. Additionally, 25 of the 32 teams were properly ordered within the standings.

Table 11 displays the comparison of the model's predicted standings for the more recent 2022 regular season with the actual standings. With two ties occurring in the 2022 season, 269 of the 271 games were used for comparison. After comparison, we determined that the results of the games were correctly predicted approximately 81.4% of the time.

It is important to note that 272 games were scheduled for the season, but the Jan 2nd matchup between the Buffalo Bills and Cincinnati Bengals was canceled due to a player emergency on the field.

The NFL team owners voted in March 2021 to institute an expanded 18-week 17 game regular season which began with the 2021 NFL season.

5. Future research directions

Although the models in this research indicate the relationship between team statistics and winning, the results unearth other considerations in future model creation.

Table 11
Model validation with 2022 regular season data.

Team	Actual			Predicted		Team	Actual			Predicted	
AFC East	W	L	T	W	L	NFC East	W	L	T	W	L
Buffalo Bills*	13	3	0	11	5	Philadelphia Eagles *	14	3	0	12	5
Miami Dolphins+	9	8	0	6	11	Dallas Cowboys+	12	5	0	12	5
New England Patriots	8	9	0	11	6	New York Giants+	9	7	1	10	7
New York Jets	7	10	0	6	11	Washington Commanders	8	8	1	5	12
AFC North						**NFC North**					
Cincinnati Bengals*	12	4	0	11	5	Minnesota Vikings*	13	4	0	9	8
Baltimore Raven+	10	7	0	11	6	Detroit Lions	9	8	0	9	8
Pittsburgh Steelers	9	8	0	8	9	Green Bay Packers	8	9	0	8	9
Cleveland Browns	7	10	0	6	11	Chicago Bears	3	14	0	6	11
AFC South						**NFC South**					
Jacksonville Jaguars*	9	8	0	7	10	Tampa Bay Buccaneers*	8	9	0	8	9
Tennessee Titans	7	10	0	7	10	Carolina Panthers	7	10	0	7	10
Indianapolis Colts	4	12	1	5	12	New Orleans Saints	7	10	0	7	10
Houston Texans	3	13	1	2	15	Atlanta Falcons	7	10	0	6	11
AFC West						**NFC West**					
Kansas City Chiefs*	14	3	0	11	6	San Francisco 49ers*	13	4	0	13	4
Los Angeles Chargers+	10	7	0	8	9	Seattle Seahawks+	9	8	0	9	8
Oakland Raiders	6	11	0	7	10	Los Angeles Rams	5	12	0	4	13
Denver Broncos	5	12	0	8	9	Arizona Cardinals	4	13	0	6	11

While the decision tree and binary logistic regression models account for the correlation between winning and team statistics, it is also possible that team statistics can be derived from causation. The models demonstrate that turnovers are correlated to winning, but it might also be the case that winning causes turnovers. When football teams hold a lead, especially in the 4th quarter, play-calling becomes more conservative, and more run plays are used. Conversely, teams in losing situations are often more inclined to take risks, including more aggressive play-calling and high-danger throws, for a chance at taking the lead, causing the probability of a turnover to increase.

Another consideration for future research is the impact of team statistics throughout a game. This study recognizes each game by the team statistics following a result, but the value of each statistic could change by quarter. How does rushing and passing yardage at halftime impact winning? Is a turnover in the 1st quarter easier to overcome than one in the 3rd quarter?

This research can be further improved by segmenting the data by year or smaller groupings of years to understand how the impact of team statistics has changed over time. It could be hypothesized that passing yards has had an increasing impact while rushing yards has had a decreasing impact as passing efficiency has improved, play-calling philosophies have changed. Additionally, it can be hypothesized that team statistics such as turnovers will always have an impact on winning.

A final consideration for future research is developing a model to predict the outcome of future games. The models in this study were built retroactively to recognize the impact of individual team statistics on winning a game. Although this is important in achieving a better understanding of the game, it does not allow for the prediction of a game's outcome without the final statistics. Through the utilization of performances from previous games and a team's makeup (skill level of individual players and positional groups), one could predict the results of a game before it happens and determine the win probability for each team.

6. Discussions and conclusions

One would argue that data is a key driver for the success of sports teams. Using various data analytics tools and statistical methods coaches can identify effective strategies in sports, and make useful decisions to track, enhance and predict individual athlete and team performance. Both descriptive and predictive analytics in sports provide coaches with new opportunities and discoveries, improve their decision-making, enable them to plan better and innovate faster, and help them use everything at their disposal to reach their full potential.

Although analytics in sports is a relatively new concept, it has gained rapid popularity in recent years, and will likely continue to evolve as a field. Major professional sports teams have been employing various data analytics applications and tools as they can provide competitive advantages to decision-makers and teams. Team managers and coaches can employ the use of data to provide insights, make data-driven decisions, and measure performance.

As demonstrated in this study, predictive models such as decision trees and logistic regression analysis may be used to determine with measurable accuracy how different team statistics interact with the outcome of a game. Both models developed in this study illustrate how combinations of the predictors are associated with wins.

In this study, a number of variables were used as predictors (independent variables). The binary win–loss outcome measure was used as a target (dependent) variable. Decision tree and binary logistic regression models were created to describe the relationships between the predictors and football game outcomes in the NFL. While the decision tree model predicted win–loss with up to 79% accuracy, the binary logistic regression model predicted outcomes or win–loss with up to 83% accuracy. The variables that are most predictive of wins include offensive and defensive turnovers. Therefore, one of the best ways to increase a team's chance of winning a game is to limit the occurrence of turnovers on offense and force turnovers on defense.

Turnovers are a significant element in football for two reasons, it marks a loss of opportunity to score and it is a disruptor of team momentum. Losing the opportunity to score is crucial as NFL teams average an estimated, eleven possessions per game. Every loss of possession significantly impacts the number of points that can be scored, making it more difficult to win a game. The shift in team momentum from a turnover can also change the makeup of a game. A lengthy drive downfield which results in a turnover can cause frustration and become detrimental to player performance. Multiple turnovers can begin to create a lack of confidence in the team and individual player abilities. At the same time, the opposing offense can have a surge of confidence after seeing their defense succeed.

This research has contributed to filling the gap that exists in the current literature on developing a predictive model in sports analytics. For instance, most of the data utilized in earlier studies to predict the outcome of any given game is derived from relatively small data sets. This study, however, analyzed a total of 4096 games which is one

of the largest data sets ever used to develop a predictive model in sports analytics. Therefore, this study extends and improves on previous studies by employing a longer time frame for the development of predictive models to demonstrate additional insights about important predictors of wins in the NFL.

The models and methods used in this study can be implemented by other sports analysts to generate further insights that support the strategic decision-making processes of coaches and their sports organizations.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Data availability

Data used in this study is available online on the NFL website.

References

[1] V. Stolbunov, Introduction to sports analytics, 2014, Retrieved from October 10, 2021 from https://sportsanalytics.sa.utoronto.ca/2014/12/11/introduction-to-sports-analytics/.

[2] M.S. Fury, L.S. Oh, M.E. Berkson, New opportunities in assessing return to performance in the elite athlete: unifying sports medicine, data analytics, and sports science, Arthrosc. Sports Med. Rehabil. 4 (5) (2022) e1897–e1902.

[3] R. Metulini, G. Gnecco, Measuring players' importance in basketball using the generalized Shapley value, Ann. Oper. Res. (2022).

[4] L. Steinberg, Changing the game: The rise of sports analytics, 2015, Retrieved May 2, 2019 from https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/#4a82a5ca4c1f.

[5] V. Sarlis, V. Chatziilias, C. Tjortjis, D. Mandalidis, A data science approach analyzing the impact of injuries on basketball player and team performance, Inf. Syst. 99 (2021).

[6] V. Sarlis, C. Tjortjis, Sports analytics: Evaluation of basketball players and team performance, Inf. Syst. 93 (2020).

[7] M. Gifford, T. Bayrak, What makes a winner? Analyzing team statistics to predict wins in the NFL, in: Americas Conference on Information Systems, AMCIS, 2020, pp. 10–14.

[8] J. Davis, L. Bransen, L. Devos, W. Meert, P. Robberechts, J. Van Haaren, M. Van Roy, Evaluating sports analytics models: Challenges, approaches, and lessons learned, in: AI Evaluation Beyond Metrics Workshop at International Joint Conference on Artificial Intelligence, Vol. 3169, 2022, pp. 1–11.

[9] P. Thakkar, M. Shah, An assessment of football through the lens of data science, Ann. Data Sci. 8 (2021) 823–836.

[10] F.R. Goes, L.A. Meerhoff, M.J.O. Bueno, D.M. Rodrigues, F.A. Moura, M.S. Brink, M.T. Elferink-Gemser, A.J. Knobbe, S.A. Cunha, R.S. Torres, K.A.PM. Lemmink, Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review, Eur. J. Sport Sci. 21 (4) (2021) 481–496.

[11] A. Joseph, N.E. Fenton, M. Neil, Predicting football results using Bayesian nets and other machine learning techniques, Knowl.-Based Syst. 19 (7) (2006) 544–553.

[12] S. Nunes, M. Sousa, Applying data mining techniques to football data from European championships, 2006, Retrieved May 28, 2020, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.867.8080&rep=rep1&type=pdf.

[13] McCabe, J. Trevathan, Artificial intelligence in sports prediction, in: Fifth International Conference on Information Technology: New Generations, IEEE, Las Vegas, NV, USA, 2008, pp. 1194–1197.

[14] E. Davoodi, A. Khanteymoori, Horse racing prediction using artificial neural networks, recent advances in neural networks, Fuzzy Syst. Evol. Comput. (2010) 155–160.

[15] D. Delen, D. Cogdell, N. Kasap, A comparative analysis of data mining methods in predicting NCAA bowl outcomes, Int. J. Forecast. 28 (2) (2012) 543–552.

[16] M. Maszczyka, A. Gołaś, P. Przemysław, R. Roczniokа, A. Zając, A. Stanula, Application of neural and regression models in sports results prediction, Procedia - Soc. Behav. Sci. 117 (2014) 482–487.

[17] N. Tax, Y. Joustra, Predicting the dutch football competition using public data: A machine learning approach, 2015, Retrieved May 28, 2020 form https://www.researchgate.net/publication/282026611_Predicting_The_Dutch_Football_Competition_Using_Public_Data_A_Machine_Learning_Approach.

[18] S.K. Deshpande, S.T. Jensen, Estimating an NBA player's impact on his team's chances of winning, J. Quant. Anal. Sports 12 (2) (2016) 51–72.

[19] C.M. Young, Wei Luo, P. Gastin, J. Tran, D.B. Dwyer, The relationship between match performance indicators and outcome in Australian football, J. Sci. Med. Sport 22 (4) (2019) 467–471.

[20] K. Kapadia, H. Abdel-Jaberb, F. Thabtaha, W. Had, Sport analytics for cricket game results using machine learning: An experimental study, Appl. Comput. Inform. 18 (3/4) (2022) 256–266.

[21] G. Liu, Y. Luo, O. Schulte, T. Kharrat, Deep soccer analytics: Learning an action-value function for evaluating soccer players, Data Min. Knowl. Discov. 34 (2020) 1531–1559.

[22] P. Rahimian, L. Toka, A data-driven approach to assist offensive and defensive players in optimal decision-making, Int. J. Sports Sci. Coach. (2023).

[23] P. Toma, F. Campobasso, Using data analytics to capture the strategic and financial decision-making of Europe's top football club, Technol. Forecast. Soc. Change 186 (Part A) (2023).

[24] N.H. Nguyen, D.T.A. Nguyen, B. Ma, J. Hu, The application of machine learning and deep learning in sport: Predicting NBA players' performance and popularity, J. Inf. Telecommun. 6 (2) (2022) 217–235.

[25] B.K. Teeselink, M. Assem, D. Dolder, Does losing lead to winning? An empirical analysis for four sports, Manage. Sci. 69 (1) (2022) 513–532.

[26] F.P. Romero, C. Lozano-Murcia, J.A. Lopez-Gomez, E.A. Sanchez-Herrera, E. Sanchez-Lopez, A data-driven approach to predicting the most valuable player in a game, Comput. Math. Methods 3 (4) (2021) 1–11.

[27] G. Duran, Sports scheduling and other topics in sports analytics: A survey with special reference to Latin America, TOP Off. J. Span. Soc. Stat. Oper. Res. 29 (1) (2021) 125–155.

[28] M. Du, X. Yuan, A survey of competitive sports data visualization and visual analysis, J. Vis. 24 (2021) 47–67.

[29] H. Li, A. Manickam, R.D.J. Samuel, Automatic detection technology for sports players based on image recognition technology: The significance of big data technology in China's sports field, Ann. Oper. Res. (2022) 1–18.

[30] D. Harville, Predictions for national football league games via linear-model methodology, J. Amer. Statist. Assoc. 75 (371) (1980) 516–524.

[31] L. Boulier Bryan, H.O. Stekler, Predicting the outcomes of national football league games, Int. J. Forecast. 19 (2) (2003) 257–270.

[32] T. McManus, Mathletes: Eagle's analytics team has an in game-line to doug pederson, 2017, Retrieved April 4, 2019 from http://www.espn.com/blog/philadelphia-eagles/post/_/id/22272/math-movement-eagles-analytics-team-has-direct-line-to-doug-pederson-in-game.

[33] B. Barnwell, The NFL stats that matter most, 2017, Retrieved March 15, 2019 from http://www.espn.com/nfl/story/id/20114211/the-nfl-stats-matter-most-2017-offseason-bill-barnwell.

[34] K. Rudy, A statistical look at how turnovers impacted the NFL season, 2014, Retrieved April 25, 2019 from http://blog.minitab.com/blog/the-statistics-game/a-statistical-look-at-how-turnovers-impacted-the-nfl-season.

[35] E. Feng, How passing and rushing affect winning in the NFL, 2019, Retrieved April 14, 2019 from https://thepowerrank.com/2014/01/10/which-nfl-teams-make-and-win-in-the-playoffs/.

[36] T.K. Clark, A.D. Johnson, A.J. Stimpson, Going for three: Predicting the likelihood of field goal success with logistic regression, in: 7th MIT Sloan Annual Sport Conference, March 1–2, 2013, 2013, Retrieved July 25, 2020 from http://www.sloansportsconference.com/wp-content/uploads/2013/Going%20for%20Three%20Predicting%20the%20Likelihood%20of%20Field%20Goal%20Success%20with%20Logistic%20Regression.pdf.

[37] K. Pelechrinis, E. Papalexakis, The anatomy of American football: Evidence from 7 years of NFL game data, PLoS One 11 (12) (2016) http://dx.doi.org/10.1371/journal.pone.0168716.

[38] JMP, Are you ready for some football...predictive models?, 2017, Retrieved April 3, 2019 from https://community.jmp.com/t5/JMP-Blog/Are-you-ready-for-some-football-predictive-models/ba-p/43546.

[39] J. Wu, K. Lin, C. Wu, An integrated model of scenario planning and decision tree analysis for new product development financial planning: A case of smart phone development project in Taiwan, Int. J. Ind. Eng. Theory Appl. Pract. 22 (1) (2015) 616–627.

[40] C.A. Mertler, V.R. Reinhart, Advanced and Multivariate Statistical Methods: Practical Applications and Interpretations, sixth ed., Routledge, New York, NY, 2017.

[41] A. Field, Discovering Statistics using IBM SPSS Statistics, Sage Publications, Thousand Oaks, CA, 2016.

[42] S.J. Yeoum, Y.H. Lee, A study on prediction modeling of Korea military aircraft accident occurrence, Int. J. Ind. Eng. Theory Appl. Pract. 20 (9–10) (2013) 562–573.

[43] L.S. Meyers, G. Gamst, A.J. Guarino, Applied Multivariate Research: Design and Interpretation, third ed., Sage Publication, Thousand Oaks, CA, 2017.

[44] S. Jaggie, K. Lertwachara, A. Kelly, L. Chen, Business Analytics: Communicating with Numbers, first ed., McGraw Hill, New York, NY, 2021.

[45] X. Zhou, X. Wang, C. Hu, R. Wang, An analysis on the relationship between uncertainty and misclassification rate of classifiers, Inform. Sci. 535 (2020) 16–27.