IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 September 2023, accepted 7 November 2023, date of publication 16 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3333878

RESEARCH ARTICLE

When do Risky Substitutions Improve Match Outcomes? Interpreting Causal Machine Learning Models for Insights

OZDEN GUR ALI[1] AND EMRAH YILMAZ[1,2]
[1]College of Administrative Sciences and Economics, Koç University, Sariyer, 34450 İstanbul, Turkey
[2]FTI Delta Analytics Center of Excellence, Dubai, United Arab Emirates

Corresponding author: Ozden Gur Ali (oali@ku.edu.tr)

ABSTRACT We investigate the impact of risk-taking on football match outcomes within the context of substitutions. The study uses an extensive dataset of 75 thousand substitutions spanning eight seasons across six leagues, encompassing team, match and manager characteristics, and interim results. Given the intricate web of potential drivers and the absence of a well-defined theoretical framework, conventional econometric approaches fall short. Instead, Causal Forest, a black-box causal machine learning (ML) technique, was adopted to estimate the heterogeneous treatment effects, coupled with robustness checks. Its estimates are explained through Generalized Additive Models (GAM), visualizing how different factors affect the risk-taking probability and moderate the treatment effect of risk-taking on match outcomes. Fidelity of the explanation is improved by segmenting the dataset to capture strong interaction effects. The study reveals that risk-taking propensity peaks when a team trails by 2-3 goals and tapers when leading by the same margin. Considering impact of risk-taking on match outcomes, interestingly, younger managers exhibit superior performance compared to middle-aged ones, while older counterparts excel in later substitutions. The manager's tenure with the team increases the impact, tapering in the long run. Earlier substitutions, stronger teams have greater impact when risk-taking. Teams leading by one goal stand to benefit most from risk-taking in substitutions. This research underscores the synergistic potential of black-box causal machine learning and interpretable models, given large representative datasets with all confounders. Insights into football risk-taking dynamics have implications for managerial strategies. Beyond football, it is noteworthy that manager age and tenure with the team are top first and third factors determining the effectiveness of a risky decision.

INDEX TERMS Football analytics, risk-taking, decision making, heterogeneous treatment effect, causal forest, generalized additive models (GAM), interpretability, machine learning, statistical learning, sports.

I. INTRODUCTION

Football is the sport with the largest viewership worldwide https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html, with an economic activity estimated at billions of dollars annually [1]. Aided by the availability of big data and machine learning sports analytics approaches are increasingly used to predict player, team and coach performance [2], [3], [4], [5], [6] Prediction of future performance is very valuable for selecting promising players and coaches, and the accuracy of such predictions can be easily measured with holdout data.

On the other hand, supporting other tactical decisions such as whether and when to engage in a risky substitution, requires estimating the causal effect of risk-taking, controlling for all confounders. The analytics needed to support this decision go beyond the typical machine learning prediction problem, because the incremental effect of the risky substitution on the match outcome is not observable: we observe the match outcome either under the risky or the

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Mei.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

non-risky substitution, which is called the fundamental problem of causal inference [7]. The so-called ''treatment'' effect, which is the difference between these two potential outcomes, can depend on manager, team, and opponent characteristics as well as contextual and interim results available at the time the decision is made. The heterogeneous treatment effect (HTE) estimates are specific to the circumstances, hence they are more valuable for decision support than the average treatment effect (ATE).

While running randomized controlled experiments is the gold standard for treatment effect estimation, it is frequently infeasible: In our case, it would amount to requiring teams to randomly make risky/ non-risky substitution decisions, which could adversely affect the match outcome. Alternatively, observational data that include all confounding factors and exhibit variability in the decisions can be used to estimate treatment effects. Traditionally, researchers have used econometric models with interaction effects of the treatment variable to investigate specific hypotheses about heterogeneous treatment effects. While treatment effects can be estimated with a correctly specified model [8], [9], selecting variables and deciding on the functional form and interactions is a major problem faced by empirical researchers [10]. To hedge against the risk of model misspecification, researchers try to show the insensitivity of the results to different set of controls and functional forms.

This traditional approach falls short of meeting the needs by a) not facilitating discovery of whether and how the treatment effect varies with the many potential heterogeneity dimensions, and b) leading to biased estimates due to incorrect model specification including the functional form of variables.

Recently, in the convergence of economics and machine learning fields, causal machine learning (ML) methods have been introduced to estimate HTEs in high dimensional observational data [11], [12], [13]. These methods provide HTE estimates by automatically detecting interactions and nonlinear effects and controlling for potential confounders, thus addressing the aforementioned problems.

On the other hand, these causal ML models are black boxes and do not provide information about how the treatment effect varies with the heterogeneity dimensions, beyond a variable importance (or feature attribution) metric. Insights into the model are valuable for supporting tactical decisions; further, comparing such findings with background knowledge and previous literature serves to establish trust in the model.

In this paper, we focus on the HTE of risk-taking behavior in substitution decisions on match performance and seek to understand the conditions under which risk taking is worthwhile. To this end, we curate a comprehensive dataset containing manager, team and match characteristics, past performance, and interim results from six countries and eight full seasons with approximately 75 thousand observations. We estimate the HTE using the Causal Forest method. To get insights into the model, we explain the propensity score and HTE estimates with Generalized Additive Models (GAM), visualizing nonlinear additive effects. We employ checks for robustness and seek a high fidelity approximation through cross-validation and segmentation. Finally, we compare our findings with the extant literature.

Our paper contributes to the literature in three ways: First, we illustrate how practitioners and researchers can benefit from big data and causal machine learning methods to estimate heterogeneous treatment effects while gaining insights. Second, we curate a large dataset that is more representative than previous research due to its multi-league, multi-year nature. To the best of our knowledge, manager characteristics are used for the first time for analyzing risk-taking behavior in the football literature. Third, we contribute to the sports analytics literature by providing empirical evidence about when risk-taking in player substitution improves the football match outcome, and when football coaches take risky substitution decisions.

The remainder of this paper is organized as follows. In the next section, we review previous studies on risk-taking in football, next we describe the data. Section IV presents the methodology, followed by the results. We discuss the findings in Section VI and finish with conclusions, limitations, and directions for future research.

II. BACKGROUND STUDIES

We review relevant studies on the determinants of risk-taking and the effects on match outcome in football. Risk-taking behavior is an important element of managerial decision making in sports, economics, and finance. Several studies have investigated the factors contributing to risk-taking and the effect of risk-taking on outcomes, considering manager characteristics and elements of the context within which the decision is made [14], [15].

We refer to literature from other sports to provide motivation for the variables that we curated in our data set. To support heterogeneous treatment effect estimation with observational data, we need to make sure that the data set contains the confounding factors, which are in this case, the determinants of both the match outcome and the risk-taking probability when a substitution is made.

In football, the risk level of the manager's strategy has been investigated in terms of initial team formation (i.e., the number of forward and midfield players in the field) [16], and how the coach substitutes a player (e.g., replacing a midfield player with a forward player or vice-versa) [17], [18].

A. DETERMINANTS OF RISK-TAKING

The determinants of risk-taking behavior in football can be considered in terms of the information available at the beginning of the match (ex-ante) versus information that emerges during the game (interim).

Past performance is one of the ex-ante factors that has been studied previously. Using male soccer data covering eight seasons of the Italian premier league, [16] found that coaches with prior defeats take more risks in the initial formation in the following game. Single and heavy defeats cause coaches

to choose more offensive systems of play, whereas multiple defeats result in less risk seeking. The authors also found that coaches of top teams, those with average salary above 1,000,000 EUR, do not respond to past performance as much as others, and tend to follow more stable strategies by following the preset seasonal objectives of their team.

Reference [18] studied 306 matches of season 2003/2004 German Major League Soccer and found that the difference in ranking, defined as the rank of the team minus its opponent's rank, is positively associated with switching into risky strategies, via initial formation and substitutions. Reference [17] analyzed expectations as reference points in 12 seasons of German and English leagues and found that when teams are behind the expected match outcome, measured by pre-match betting odds, their coaches make significantly more offensive substitutions. Reference [19] reported that coaches of favorites take significantly more risk when their team is behind with two or more goals, whereas coaches of underdogs prefer less aggressive team formation when their team is in the lead.

Reference [18] demonstrated that playing against a team with a similar ranking does not significantly change the aggressiveness of the initial team formation. On the other hand, [20] reported a significant result that playing a game against a rival partner increases risk-taking behavior. To understand the underlying mechanism behind this result, they conducted additional laboratory experiments involving face-to-face competition and showed that rivalry increases risk-taking behavior by increasing promotion focus (i.e., motivation to achieve gains) and physiological arousal.

Decision maker characteristics are shown to have a role in risk-taking behavior. For example, several studies document gender differences in risk-taking in other sports [21], [22]. Investigating the relationship between age and risk-taking behavior, [23] provides a real-life evidence that being older than forty years significantly predicts risk seeking among recreational skiers and snowboarders, but no significant differences in risk-taking behavior on ski slopes across nationalities based on self-reported risk. To the best of our knowledge, this is the first study to use age, tenure, and nationality as potential risk-taking determinants in football.

There are mixed results regarding the influence of where the game is played on managerial risk-taking. Reference [18] document that coaches start a game with a higher number of forward players when they play at home. On the other hand, [19], [24] analyzed the impact of playing at home on risk-taking behavior in weight-lifting and found no significant impact.

A game-theoretical model of the role of interim performance feedback in subsequent risk-taking behavior in tournaments found a Nash equilibrium, in which the leading player chooses the safe strategy, while the one lagged behind opts for the risky strategy [25]. Empirical studies corroborated these theoretical findings. The current score during the game is the first variable in the literature to provide interim information. Reference [18] reported that the probability of risk-taking in substitutions decreases with difference in goals. When a team is ahead, its coach tends to choose less risky team formation to decrease the likelihood of suffering a goal. However, when the team is behind, the coach is more willing to substitute existing players with riskier ones. In support of this finding, [19] documented that trailing teams are more likely to choose riskier strategies than leading teams, based on an analysis of seven seasons in the UEFA Champions League (from 2009/10 to 2015/16). Reference [17] combined the current score with ex-ante expectations and concluded that coaches take more risks if their teams are behind expectations. Evidence from other sports, such as basketball and weightlifting, also indicates that trailing teams take more risks [18], [26], [27]. Reference [24] found that in athletic risk-taking has an inverted-U relationship with the interim rank.

Another interim result is the red card. Reference [19] showed that receiving a red card is negatively associated with risk-taking, which makes sense considering that losing a player increases the likelihood of conceding a goal.

B. EFFECTS OF RISK-TAKING ON THE OUTCOME

We now turn to the literature on the effect of risk-taking in sports on performance.

Several papers have analyzed the average effect of risk-taking on performance. Reference [16] reported that increased risk-taking is not significant in predicting wins. Similarly, [19] found that the cumulative offensive score of each player (i.e., risk-taking) in the field has no significant impact on goal scored and conceded. Reference [18] analyzed the impact of risk-taking substitutions on point advancement and concluded that switching to a strategy of higher risk is less successful than maintaining the initial strategy, based on 612 observations. Reference [17] concluded that the impact of offensive substitutions on performance, defined as the final goal difference and number of points, is negative; and concluded that coaches do not act in a fully rational way. In summary, we do not observe consensus about the overall effect of risk-taking on performance in the football literature. Although insignificant or negative effects have been reported, heterogeneity dimensions have not been investigated. Current studies are limited in terms of the number of observations and representativeness, since they focus on particular leagues and seasons.

Other studies from different domains have investigated the heterogeneity of risk-taking on performance. Reference [26] analyzed the impact of substituting the goalkeeper for an additional field player (i.e., risk-taking behavior) on the probability of game wins (i.e., performance) in handball. They concluded that risk-taking is beneficial for underdogs, whereas favorites lose significantly more games when they take more risks.

TABLE 1. League descriptions with the number of teams and games.

League	Country	# of Teams	# of Matches	Year
Premier League	England	20	38	2013-2020
Bundesliga	Germany	18	34	2013-2020
LaLiga	Spain	20	38	2013-2020
Ligue 1	France	20	38	2013-2020
Serie A	Italy	20	38	2013-2020
Super Lig	Turkey	18	34	2013-2020

C. METHOD TO ESTIMATE EFFECT OF RISK_TAKING ON THE OUTCOME

The majority of the cited studies use regression methods, including ordinary least squares (OLS), and generalized linear models with probit and logit link functions to model categorical outputs. Treatment effect estimation with such econometric models requires a correctly specified model that includes careful selection of control variables, and their functional form. While they assume linear effects, there is some evidence of nonlinear relationships: For example, [24] reports inverted U-shaped risk-taking behavior with interim results, which also makes intuitive sense.

Beyond the difficulty of determining the control variables and their functional form, an important limitation of the econometrics tools employed in the literature is the difficulty in identifying subgroups with significantly different treatment effects. Uncovering such heterogeneity provides important insights for managerial decision-making. However, researchers need to know beforehand which variables create heterogeneity and decide on the specification before running any regression. Especially in situations where the theory is not as well established, it is difficult to find the right variables that could possibly contribute to the heterogeneity of the treatment effect.

Studies estimating the effect of risk-taking on performance involve at most one heterogeneity dimension, while most focus on the overall (average) effect. Hence, it is possible that their finding of no, or negative, significant payoff in risk-taking is due to averaging over positive and negative effects depending on the situation. This would explain the apparent irrational behavior, i.e., that teams engage in risk-taking behavior, which they should not be if risk-taking were not paying off in terms of better outcomes.

The potential heterogeneity dimensions for the effect of risk-taking on the outcome are many: ex-ante information including team past performance, ranking, manager and match characteristics, interim results, as well as match time minutes at the moment of the substitution decision. The methods in the literature are not equipped to deal with the large number of potential heterogeneity dimensions.

TABLE 2. Variable descriptions.

Variable	Description
Goal.Difference	Difference between the goals scored and conceded
Red.Card	Total red cards shown to the team
Red.Card.Opponent	Total red cards shown to the opponent team
Opponent.Risk.taking	Opponent manager's risk-taking decision until now. With one sub, below scores are calculated: 0) Only neutral subs (defender replacing a defender) 1) One increment positive subs (midfielder in - defender out, or forward in - midfielder out) 2) Two increment positive subs (defender out - forward in) -1) inverse of positive 1 case -2) inverse of positive 2 case If there are multiple subs until that time, then summation of above scores might give additional -3, -4, -5, 3, 4, 5 etc.)
Minute.of.the.Game	Minute of the substitution
Manager.Age	Age of the manager
Manager.Foreign	Whether the manager is foreign or local (comparing to the country he is working)
Manager.Tenure	Manager's tenure with the existing team
Relative.Strength	Difference in strength of competing teams. It is the one minus the ratio of total points gathered by each team in the last 5 years.
Current.Ranking	Current ranking of the team in the league ranging from 1 to number of teams competing in the league
UE_Cons_Game_Lost	Number of consecutive matches where they lost the game unexpectedly
UE_Cons_Game_Won	Number of consecutive matches where they win the game unexpectedly
Last_5_Avg_Point	Average points in the last 5 match
Hum_Defeat	Whether the last match result was a humiliating defeat
Big_Victory	Whether the last match result was a big victory
Home.Match	Playing at home vs. away
Derby	3 if the match is recognized a derby match by the authorities (Barcelona vs. Real Madrid). 2 if the teams are big rivalries but not necessarily from the same city (Chelsea vs. Liverpool). O otherwise.
Match.Importance	3 if the match is a derby 2 if either match is a big rivalry or both teams' current rankings are at top 5 or bottom 3 in the league 1 if one of the teams' current ranking is at top 5 or bottom 3 in the league 0 otherwise.
Year	Year

III. DATA AND DESCRIPTIVE STATISTICS

We collected a comprehensive and diverse data set to estimate the heterogeneous treatment effect of risk-taking in substitutions on changing the match outcome, quantifying the many potential confounders and heterogeneity dimensions identified in the previous section.

Our data involve all matches in men's first division in six countries (Germany, France, Italy, England, Spain, and Turkey) that were played in the 8 seasons from 2013 to 2020, as seen in Table 1. After eliminating matches with missing or questionable data we obtained 36,878 team-match pairs. Using the minute-by-minute data on the score, cards shown to each team, substitutions with incoming and outgoing players and their positions, we created a dataset of the substitutions in these matches and added information about the match, manager, teams, and past performance that were available at the beginning of the match. Our data sources are BeinSports

TABLE 3. Descriptive statistics.

	Variable	mean	std	min	max
Interim Results	Goal.Difference	0.0	1.5	-8.0	7.0
	Red.Card	0.1	0.2	0.0	2.0
	Red.Card.Opponent	0.1	0.2	0.0	2.0
	Opponent.Risk.taking	0.1	0.9	-5.0	5.0
	Minute.of.the.Game	69.2	11.0	45.0	85.0
Manager Characteristics	Manager.Age	50.1	7.3	28.6	73.1
	Manager.Foreign	0.3	0.5	0.0	1.0
	Manager.Tenure	1.6	2.2	0.0	26.6
Team Characteristics	Relative.Strength	0.0	0.8	-2.9	2.9
	Current.Ranking	10.3	5.6	1.0	20.0
	UE_Cons_Game_Lost	0.4	0.7	0.0	9.0
	UE_Cons_Game_Won	0.3	0.6	0.0	6.0
	Last_5_Avg_Point	1.4	0.7	0.0	3.0
	Hum_Defeat	0.1	0.2	0.0	1.0
	Big_Victory	0.1	0.3	0.0	1.0
Match Characteristics	Home.Match	0.5	0.5	0.0	1.0
	Derby	0.1	0.5	0.0	3.0
	Match.Importance	0.6	0.7	0.0	3.0
	Year	2016	2	2013	2020

(www.beinsports.com) for the details and results of each game, and TransferMarkt (www.transfermarkt.com) and Wikipedia (www.wikipedia.org) for player- and manager-related information. Explanations of these variables and their descriptive statistics are provided in Tables 2 and 3, respectively. In addition, we have dummy variables for the league and year of the match. This final dataset contains 74550 observations, each describing a substitution situation. Roughly one quarter of the substitutions in the data set are risky, which provides a good number of risky and non-risky decisions for inferring the treatment effect.

The tournament rules for each league are similar as all countries above are members of the UEFA. A match has three possible outcomes for the team: 1) ''Win'' if the team scores more goals than the opponent, gaining 3 points, 2) ''Draw'' if the team scores same number of goals as the opponent, gaining 1 point, 3) ''Loss'' if the team scores less than the opponent, gaining zero points. At the end of the tournament, the team that accumulates the most points becomes the champion, receives a large monetary prize, and qualifies to join the UEFA Champions League which is the most prestigious and profitable contest in Europe. Teams finishing 2nd to 6th also qualify for different prestigious European Cups and a considerable monetary prize. The three or four teams with the lowest points, depending on the country and year, are relegated to the lower division of their league. Furthermore, national football associations offer monetary prizes for each match. Therefore, all teams in the league have an incentive to gain more points in every match, even if they are not contenders for the championship.

The outcome variable, *Match.Points,* is the points the team receives at the end of the match. The treatment variable, *Risk.Taking*, takes the value 1 if the substitution is risky, and 0 otherwise. We consider a substitution risky if the incoming player's role is more offensive than that of the outgoing player, where *Off(role)* indicates the degree of offensiveness of the player's role:

Off(Forward) > Off(Midfielder) > Off(Defender).

IV. METHODOLOGY

We use a well-established causal machine learning algorithm, the Causal Forest [11] that is implemented in the grf R package [28] to estimate the heterogeneous treatment effect of *Risk.Taking* on *Match.Points* using the dataset that we have curated.

The expected heterogeneous treatment effect for an individual observation i, τ_i, is given by:

$$\tau_i = E[O_i|\boldsymbol{X_i}, T_i = 1] - E[O_i|\boldsymbol{X_i}, T_i = 0] \quad (1)$$

where for observation i, O_i is the outcome, which is *Match.Points* and T_i is the treatment, which is *Risk.Taking* in our application, respectively. τ_i provides the incremental points that the team is expected to obtain if a risky substitution is made, rather than a non-risky one.

Covariate vector $\boldsymbol{X_i}$. contains variables that are potential drivers of the treatment decision or the outcome, and moderators (heterogeneity dimensions) of the treatment effect. In particular, we assume that $\boldsymbol{X_i}$ contains potential confounders that affect both the treatment decision and the outcome. The fundamental difference between randomized experiments and observational data is that in the former, treatment is independent of all pre-treatment characteristics, whereas in observational data it is dependent on the covariates. The propensity score, ρ_i, is the predicted probability of treatment given the covariates. Causal Forest provides HTE and propensity score predictions, τ_i and ρ_i, given covariates $\boldsymbol{X_i}$.

$$\rho_i = E\left[P\left(T_i = 1|\boldsymbol{X_i}\right)\right]. \quad (2)$$

Out of sample evaluation of the treatment effect predictions τ_i is not possible because of the fundamental problem of causal inference. Instead, we perform a robustness check by splitting the dataset randomly into two parts, training a Causal Forest model and using it to predict τ_i for the complete data set. We compare the outputs of the two Causal Forest models trained with each random partition of the data. Similar predictions and variable importance ranking provide evidence of the repeatability of the procedure.

Causal Forest provides no explanation for the model beyond the variable importance metric. We use Generalized Additive Models (GAM) that predict τ_i and ρ_i, as a function of covariates $\boldsymbol{X_i}$, to obtain global explanations with additive nonlinear main effects, as well as their statistical significance. We do not use a local method that tailors the explanation around a particular observation, such as the popular Explainable Artificial Intelligence (XAI) method LIME [29], because we would like to compare global patterns to domain knowledge and extant findings in the literature.

A. CAUSAL FOREST

The Causal Forest algorithm is a state-of-the-art method for estimating heterogeneous treatment effects with formal asymptotic guarantees [11]. It has been used in diverse fields to estimate heterogeneous treatment effects, such as e-commerce [30], public health [31], impact of incentives on vaccine uptake [32], finance and student learning [33].

The Causal Forest algorithm consists of an ensemble of causal trees that seek to maximize the treatment effect heterogeneity, where the treatment effect is predicted as the difference in the average outcomes between the treated and control observations in terminal leaves. Therefore, the goal is to find leaves where the treatment effect is constant within the leaf but is different from other leaves (i.e., maximum inter-leaf treatment effect). Causal forest requires unconfoundedness (i.e., all potential confounders are observed). Causal Forest residualizes the treatment and outcome variables before treatment effect estimation to control for confounding, also called double machine learning [12]. The high dimensionality of the covariate space makes classical statistical approaches ill-suited for measuring the effects on the treatment and outcome. Double machine learning alleviates this problem with two prior predictive tasks: 1) predicting the outcome with the controls, 2) predicting the treatment with the controls, which is the propensity score, and using their residuals for treatment effect estimation.

Causal Forest uses these residuals for treatment effect estimation with an ensemble of trees, as in the well-known random forest algorithm [34], which involves randomization of variables and bootstrap samples of observations. Generalized random forests, of which Causal Forest is a member, use the concept of local maximum likelihood, weighting nearby observations more heavily than farther ones using weights based on the fraction of trees in which an observation appears in the same leaf [7]. Half of the data is randomly selected to estimate the tree structure, while the other half is used for effect estimation in each leaf; this so-called honesty condition enables asymptotic results and valid confidence intervals for the estimated parameters [11].

To run the Causal Forest model, we use the Generalized Random Forests *grf* R package [28]. The *grf* package provides a simple variable importance measure based on weighted sum of the number of times feature i was split on for the HTE.

B. GENERALIZED ADDITIVE MODELS

Generalized Additive Models extend multiple linear regression analysis by allowing a nonlinear relationship between each feature and the dependent variable.

$$y = \alpha + \sum\nolimits_m f_m\left(x_m\right) + \epsilon \tag{3}$$

Here, y is the dependent variable, $\mathrm{f_m}$ is a nonlinear smooth function of the variable x_m, α is a coefficient and ϵ is the error term [35]. While the function $\mathrm{f_m}\left(x_m\right)$ can take many forms such as natural or smoothing splines, in this paper we use local smoothing regressions (loess), where the smoothness can be adjusted with the span of the local regression, that is, the percentage of observations in the neighborhood [36].

To fit the GAM models, we used the gam package in R, which uses a backfitting procedure [37]. It provides an analysis of variance table with the significance of parametric and non-parametric effects, and $\mathrm{f_m}$ plots that visualize the effect of each variable on the dependent variable, controlling for the effect of others. While the GAM accounts for nonlinear effects, it does not model interaction effects.

C. OVERVIEW OF THE METHODOLOGY

Our methodology consists of the following steps.

Step1a. Estimate the HTE of *Risk.Taking* on *Match.Points*, given the match, manager, and team characteristics, including past performance, and interim results available at the time of the substitution using the Causal Forest technique with the complete dataset, S. Output the out-of-bag treatment effect estimates τ_i, the estimated propensity to take risk in the substitution situation i, $\rho_i, \forall i \in S$, and the variable importance scores.

Step1b. Perform a robustness check for Step1a by splitting the dataset randomly into two parts, S_1 and S_2, and repeating Causal Forest modeling with each. Compare the resulting treatment effect and propensity score estimates $\forall i \in S$, from both models, as well as variable importance.

Step2a. GAM_ρ: regress ρ_i, against all covariates, using GAM with loess curves on the complete dataset, S. Visualize the functional form with plots. Evaluate the fidelity of this explanatory model with cross-validation accuracy.

Step2b. GAM_τ: regress τ against all covariates, using GAM with loess curves on the complete dataset, S. Explore segmenting the dataset to increase explanatory power while maintaining interpretability. Evaluate the fidelity with cross-validation accuracy. Visualize the functional form with plots.

V. RESULTS

We first provide the results on the propensity score, followed by the heterogeneous treatment effects.

A. PROPENSITY TO MAKE A RISKY SUBSTITUTION

The propensity score ranges from 0.08 to 0.52. Its mean is 0.25 (equal to the population average of *Risk.Taking*), and the

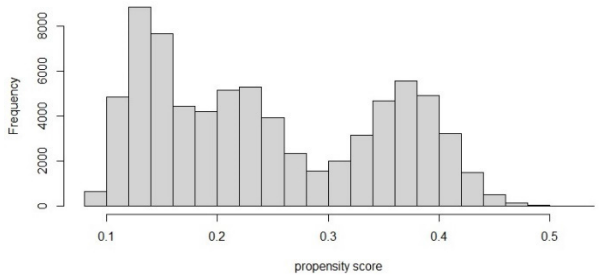


FIGURE 1. Propensity score distribution.

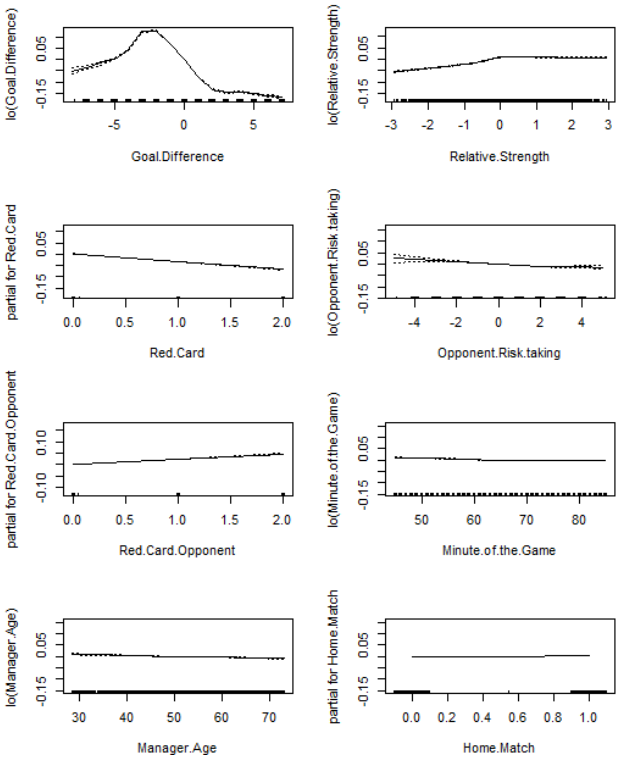


FIGURE 2. GAM plots for the GAM_ρ propensity score model, variables are presented by decreasing impact.

first and third quartiles are 0.15 and 0.35, respectively. This distribution has multiple modes, as seen in Fig.1.

Fig. 2 plots the top 8 GAM_ρ model effects, $f_m(x_m)$ in decreasing order of total F value, using a common scale to facilitate comparison. The ANOVA table, which is provided as Table 5 in the supplementary material posted on ieeexplore.ieee.org indicates significant non-parametric effects. We see that *Goal.Difference* is by far the most important driver of risk-taking in substitutions and has highly nonlinear effect. The five-fold cross validation results show that the nonlinear effects are important in explaining the propensity score rather than overfitting. The validation accuracy of the GAM_ρ model is significantly better than a linear regression with the same covariates, as evidenced by the R-squared and RMSE values 58% and 0.066 for OLS, versus 78% and 0.049 for GAM, respectively.

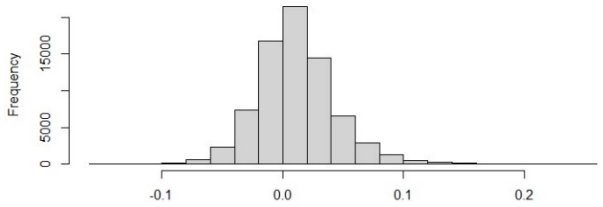


FIGURE 3. Histogram of τ_i, the heterogeneous treatment effect estimates for the complete data set.

We discuss the findings on the drivers of risk-taking in substitutions in section VI.

B. THE HETEROGENEOUS TREATMENT EFFECT OF A RISKY SUBSTITUTION ON MATCH OUTCOME

Based on the Casual Forest analysis, the average treatment effect in this population was estimated to be 0.011. In other words, a risky substitution increases the match points on average by 0.011 compared to a non-risky substitution. The lower and upper quartiles of $\boldsymbol{\tau_i}$ are -0.009 and 0.028, respectively. The histogram is shown in Fig. 3.

The Causal Forest model variable importance scores are provided in Fig. 4, where four variables stand out as important determinants of $\boldsymbol{\tau_i}$: *Manager.Age*, followed by *Minute.of.the.Game*, *Manager.Tenure* and *Relative.Strength*.

Our robustness check shows that the results from the $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ models exhibit reasonable sampling variability, with the average treatment effect estimates based on the models developed using $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ being 0.017 and 0.011, respectively. The comparative statistics are -0.007 and -0.009 for Q1, and 0.037 and 0.029 for Q3, based on the models developed on $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ respectively, and the figures for the entire dataset $\boldsymbol{S}$ are within this range. Another sign of robustness is that all three models have the same top four most important variables.

The simple marginal effects plots (see Fig. 9 in the supplementary material posted on ieeexplore.ieee.org) with these important heterogeneity drivers, that is, Manager Age, Minute of the Game, Manager Tenure, and Relative Strength, indicate highly nonlinear effects that would not be captured by linear models. Next, we fit an initial GAM model with the dependent variable τ_i, and independent variables $\boldsymbol{X}$, including the league and year control dummies. The associated ANOVA table (see Table 6 in the supplementary material posted on ieeexplore.ieee.org) indicates that these are also the top variables according to their F value contributions in this initial GAM model with significant parametric and nonparametric components.

Fig. 5 visualizes the $f_m(x_m)$ associated with the top four variables, explaining how τ_i varies with each variable, while controlling for the nonlinear effect of all other covariates in $\boldsymbol{X}$. We observe that the effect decreases with *Manager.Age* up to age 48, and then increases. The *Minute.of.the.Game* plot shows that after the 57th minute the effect decreases. We observe an increasing effect with *Manager.Tenure* that levels off. *Relative.Strength* has an increasing effect

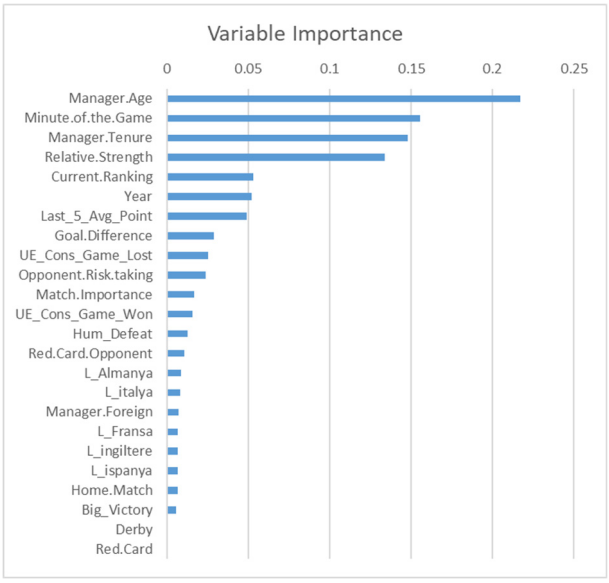


FIGURE 4. Variable importance for the causal forest model.

TABLE 4. Five-fold CV accuracy of models explaining τ_i.

5-fold cross validation results	RMSE		R-Squared	
	mean	std err	mean	std err
OLS	0.0299	0.0001	11.4%	0.1%
Initial GAM	0.0280	0.0001	22.1%	0.2%
GAM by timing	0.0260	0.0001	32.9%	0.5%
GAM by age	0.0273	0.0002	26.2%	0.7%
GAM by trailing	0.0268	0.0001	29.1%	0.2%

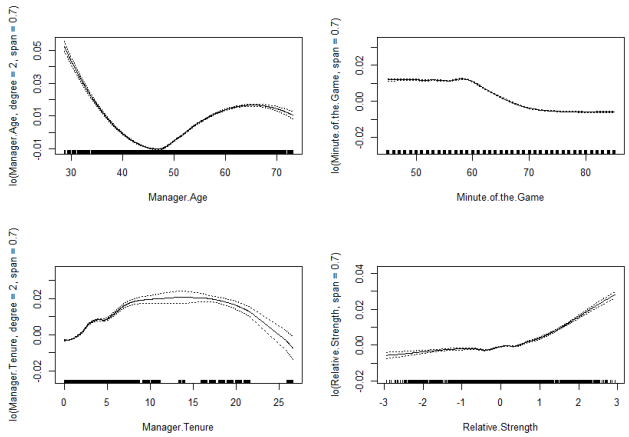


FIGURE 5. Initial GAM_τ model plots for the important variables.

for positive values, but hardly any effect when it is negative.

We evaluated the fidelity of the explanatory model using five- fold cross validation accuracy. As seen in Table 4, this initial GAM_τ model explains 22% of the variability in the Causal Forest HTE estimates, τ_i. By comparison, a linear

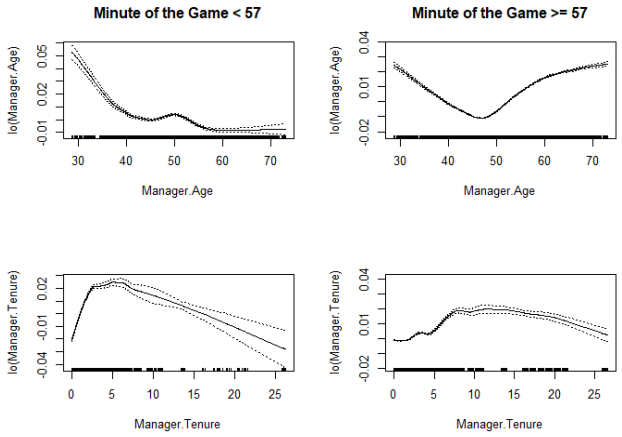


FIGURE 6. Manager characteristics plots, GAM_τ by timing.

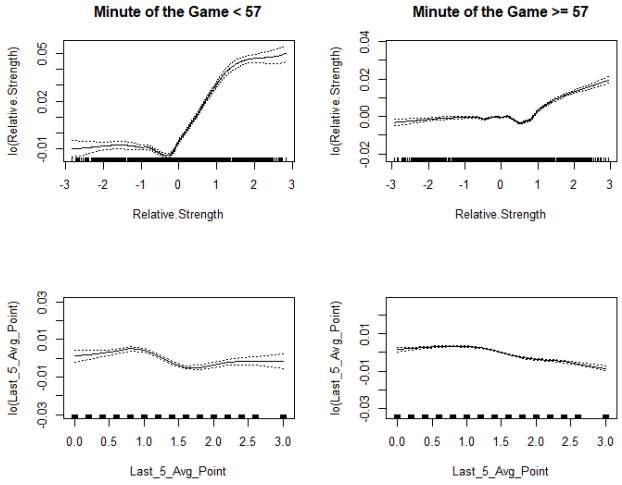


FIGURE 7. Team characteristics plots, GAM_τ by timing.

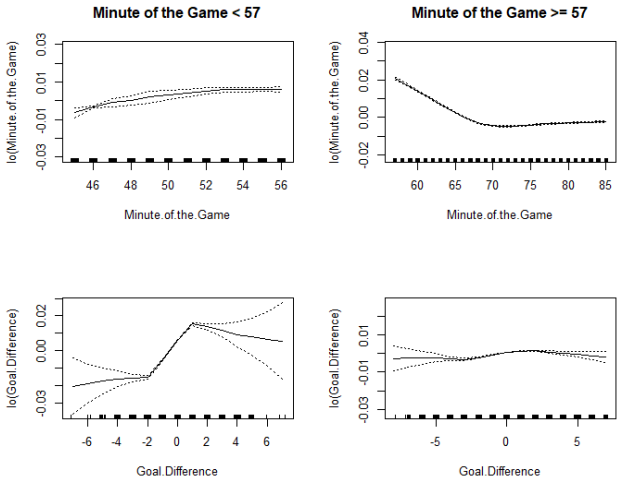


FIGURE 8. Interim information plots, GAM_τ by timing.

regression model with all variables explains only 11%, (13% when quadratic terms are added), which shows that the nonparametric effects are not only statistically significant, but also substantial.

Based on the observed shift in the impact of risk-taking in the marginal effects and GAM plots, we evaluate additional models with binary segmentation: GAM by timing (*Minute.of.the.Game* being before or after 57), and GAM by age (*Manager.Age* being below or above 48). Furthermore, we evaluate a third binary segmentation, GAM by trailing (*Goal.Difference* being negative or not), based on the literature on trailing ([18], [26] as an important moderator of risk behavior.

As seen in Table 4, segmenting the data by timing provides (statistically significantly) the best predictive accuracy, and hence explanatory power, with a cross-validation R-squared of 33%. Tables 7 and 8 in the supplementary material posted on ieeexplore.ieee.org provides ANOVA tables for the GAM models. Figures 6, 7 and 8, visualize the effects of most impactful variables, in terms of manager characteristics, team characteristics and interim results, respectively. Figs. 10–12 in the supplementary material posted on ieeexplore.ieee.org provide plots for the effects of the other variables. All plots explaining τ_i have the same scale to facilitate comparisons.

VI. DISCUSSION AND COMPARISON WITH FINDINGS IN THE LITERATURE

Fig. 2 shows the plots of the propensity score effects, where the scale is kept constant, and the variables are listed in the order of decreasing impact. Our results indicate that goal difference at the time of the substitution is the top driver of the propensity to make risky substitutions. Consistent with football literature [17], we find that trailing teams are more likely to make risky decisions. Our findings also resonate with the findings in basketball and handball that trailing teams take more risk [26], [27]. We go beyond the directional findings in the literature and provide the functional shape of how goal difference at the time of substitution contributes to risk-taking. Controlling for other interim results, match and manager characteristics, we find that the risk-taking propensity is at its peak when the team is losing by two to three goals. It reaches the lowest level when the team is winning by three goals or more, at which point it makes sense to protect the score. When the team is losing by more than three goals, the propensity decreases with the difference, levelling off at a higher level than when the team is winning, probably because the team has not much to lose. This finding is in line with the inverted U-shaped effect of interim ranking on risk-taking in athletics [24].

The second important driver is the relative strength, which is based on the total points of the two teams in the last five years. The weaker the team is relative to the opponent, the lower the propensity for risk-taking in substitutions (see Fig. 2). Findings from the literature [23], [38] also indicate a positive relationship between feeling confident about abilities and making risky choices in rock climbing and skiing. We observe a nonlinear relationship: as the difference in the strengths of the teams increases, the stronger team hardly takes more risk, whereas the weaker team's propensity to take risk decreases. Based on their past performance, the stronger team may believe that a steady course will bring them points.

Unlike the relative strength information considered in the initial game plan, red cards have the potential to change the balance of power between teams during the game. Consistent with [19], we find that managers of teams with red cards are less likely to make risky substitutions, which makes sense given our finding that weaker teams take less risk. When the opposing team is shown a red card, the manager is more likely to opt for a risky substitution, reacting to a new stronger position.

Another interim result of importance is that as the opponent takes more risk, the team's propensity to make a risky substitution decreases.

We observe that the propensity to make a risky substitution decreases from the beginning of the second half to the 70^{th} minute, increasing slightly thereafter. The last increase may be due to managers resorting to bold moves to change the flow of the game within the remaining limited time.

Finally, we observe that risk-taking decreases slightly with the age of the manager and increases slightly in home matches.

Next, we discuss our results on the effect of risk-taking in substitutions on the match outcome. Unlike [18], who declare that risk-taking in substitutions does not lead to better outcomes, we find that the effect of risk-taking in substitutions depends on manager and team characteristics, as well as the interim information. In fact, in our diverse database of six leagues and eight seasons, 63% of the estimated effects were positive (see Fig. 3).

We find that manager characteristics are the most important factors determining the impact of risky substitutions on match outcome. To the best of our knowledge, this study is the first to analyze the role of manager characteristics in the impact of risky substitutions.

The GAM plots for the effect of the important manager characteristics, Fig. 6, show that risky substitutions by young managers result in higher match points than older managers, especially early in the game. Later in the game, we observe a U-shaped relationship, where younger and older managers perform better than their middle-aged counterparts. Generally, tenure with a team increases the impact of risky substitution. It is plausible to think that older managers with 25+ years of experience, have enough experience and confidence to take risks at the right time with the right substitution. It is remarkable that young managers are better in risk-taking. A potential explanation is that becoming a team manager at a very young age may be indicative of top-class abilities. On the other hand, our results suggest that longer manager tenure with the team as of the match date means better risk-taking. This benefit peaks at about 5 years, for early substitutions, and 10 years for late substitutions, with a declining impact thereafter. The tenure effect is not surprising because managers can more effectively take risk

when they already proved themselves within the club, while the decline in football coach performance with longer tenures has been documented in the literature [39].

Among the many variables describing team characteristics, relative strength, and average points of last five games contribute the most to explaining the effect of risky substitutions on match outcome. Fig. 7 suggests that stronger teams achieve better outcomes with risk-taking and the impact increases with the gap in team strength. We see that the impact of relative strength is more prominent in early substitutions than in late ones. While relative strength provides a long-term view of a team's past performance, the average score of the last five games describes the team's current momentum. Interestingly, risky substitutions work less favorably for teams that have been successful recently, which may be explained by some of the findings in the sports literature about increased confidence resulting in complacency over time [40].

The timing of the substitution changes the dynamics leading to the match outcome, as we have seen with manager and team characteristics effects in early versus late substitutions. In Fig. 8, we see that the impact of a risky versus non-risky substitution on match outcome decreases with each minute from the 57^{th} to the 70^{th} minute of the game, presumably due to insufficient time to change the outcome.

We find that the impact of risky substitution on match outcome increases with the goal difference, particularly in early substitutions. These findings are in agreement with [17] in the sense that substituting players in an offensive way while being behind expectations (current match score comparing to ex-ante match expectations) worsens the expected match outcome. Reference [18], on the other hand, found that increased risk-taking negatively impacts the outcome (with no heterogeneity investigation).

VII. CONCLUSION

In this paper we investigated the conditions under which risk-taking pays off in football. We curated a comprehensive dataset with eight seasons from six leagues, which unlike previous studies also includes manager characteristics in addition to team and match characteristics and interim results.

Given the large number of potential drivers of the impact of risk-taking on performance and the lack of a well-established theory about the heterogeneity dimensions and functional form of the effect, traditional econometric methods are not useful. Neither can we employ traditional machine learning methods to predict the outcome and hope that all confounding relationships are accounted for. Instead, we used a causal machine learning method on a large and representative dataset. We explained the resulting black-box model estimates with global interpretable models, which is not straightforward. We used GAM to maintain interpretability while visualizing nonlinear relationships. Encountering strong interaction effects that a GAM model is not designed to capture; we segmented the dataset to increase the fidelity of the explanations. We used cross-validation to select the segmentation dimension and avoid overfitting.

Our findings show that risk-taking propensity peaks when the team is losing by two or three goals, and dips when winning by two or three goals. Weaker teams are less likely to make risky substitutions.

Contrary to some previous literature claiming that managers behave irrationally by making risky substitutions that do not improve match outcomes, we find that the majority of substitution situations in our database have a non-negative impact.

Surprisingly, we find that young managers perform better than middle-aged managers, while older managers perform better in substitutions later in the game. Tenure with the team helps, although the effect wears off with longer duration. Stronger teams perform better with risk-taking. Risk-taking in substitutions before the 57^{th} minute has a better chance of increasing the match points than non-risky substitutions, especially when the team is leading. This result implies that for the leading team, making risky substitutions early in the game has a better chance of keeping the opponent from reversing the match outcome, than playing it safe.

Overall, our results are robust and intuitive, expanding the directional findings of previous literature with nonlinear effects, and adding many new reasonable insights.

The contributions of this work are threefold: First, it introduces a comprehensive dataset on risk-taking in football with approximately 75 thousand observations; second, it serves as an example of using the new black-box causal machine learning methods in conjunction with interpretable models to extract insights beyond feature attribution, and finally it provides new insights into when risk-taking pays off in football, including manager characteristics that can guide policy decisions.

An important assumption in our study is unconfoundedness, indicating that all joint drivers of risk-taking behavior and match outcome are in the dataset. A limitation of our study is that it analyzed only one type of risk-taking behavior in football. Beyond substitutions, managers can take risks with their starting line-up, or instruct a player to be more aggressive on the pitch. Future research can apply causal machine learning combined with interpretable models for global explanations to the effect of manager and player characteristics on match outcomes to support player and coach selection and renumeration decisions.

REFERENCES

[1] M. Jaworska, "European football at the cross-roads," *J. Phys. Educ. Sport*, vol. 20, pp. 1188–1194, Oct. 2020.

[2] R. Baboota and H. Kaur, "Predictive analysis and modelling football results using machine learning approach for English premier league," *Int. J. Forecasting*, vol. 35, no. 2, pp. 741–755, Apr. 2019.

[3] A. Joseph, N. E. Fenton, and M. Neil, "Predicting football results using Bayesian nets and other machine learning techniques," *Knowl.-Based Syst.*, vol. 19, no. 7, pp. 544–553, Nov. 2006.

[4] M. Herold, F. Goes, S. Nopp, P. Bauer, C. Thompson, and T. Meyer, "Machine learning in men's professional football: Current applications and future directions for improving attacking play," *Int. J. Sports Sci. Coaching*, vol. 14, no. 6, pp. 798–817, Dec. 2019.
[5] I. Behravan and S. M. Razavi, "A novel machine learning method for estimating football players' value in the transfer market," *Soft Comput.*, vol. 25, no. 3, pp. 2499–2511, Feb. 2021.
[6] J. Almulla and T. Alam, "Machine learning models reveal key performance metrics of football players to win matches in Qatar stars league," *IEEE Access*, vol. 8, pp. 213695–213705, 2020.
[7] S. Athey and G. W. Imbens, "Machine learning methods that economists should know about," *Annu. Rev. Econ.*, vol. 11, no. 1, pp. 685–725, Aug. 2019.
[8] G. W. Imbens, "Nonparametric estimation of average treatment effects under exogeneity: A review," *Rev. Econ. Statist.*, vol. 86, no. 1, pp. 4–29, Feb. 2004.
[9] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[10] A. Belloni, V. Chernozhukov, and C. Hansen, "Inference on treatment effects after selection among high-dimensional controls," *Rev. Econ. Stud.*, vol. 81, no. 2, pp. 608–650, Apr. 2014.
[11] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *J. Amer. Stat. Assoc.*, vol. 113, no. 523, pp. 1228–1242, Jul. 2018.
[12] V. Chernozhukov, *Double/Debiased Machine Learning for Treatment and Structural Parameters*. Oxford, U.K.: Oxford Univ. Press, 2018.
[13] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 10, pp. 4156–4165, Mar. 2019.
[14] P. Bromiley, "Testing a causal model of corporate risk taking and performance," *Acad. Manage. J.*, vol. 34, no. 1, pp. 37–59, Mar. 1991.
[15] L. Naldi, M. Nordqvist, K. Sjöberg, and J. Wiklund, "Entrepreneurial orientation, risk taking, and performance in family firms," *Family Bus. Rev.*, vol. 20, no. 1, pp. 33–47, Mar. 2007.
[16] A. Bucciol, A. Hu, and L. Zarri, "The effects of prior outcomes on managerial risk taking: Evidence from Italian professional soccer," *J. Econ. Psychol.*, vol. 75, Dec. 2019, Art. no. 102090.
[17] B. Bartling, L. Brandes, and D. Schunk, "Expectations as reference points: Field evidence from professional soccer," *Manage. Sci.*, vol. 61, no. 11, pp. 2646–2661, Nov. 2015.
[18] C. Grund and O. Gürtler, "An empirical study on risk-taking in tournaments," *Appl. Econ. Lett.*, vol. 12, no. 8, pp. 457–461, Jun. 2005.
[19] S. Schneemann, "Risk-taking behavior of heterogeneous contestants and its consequences in dynamic tournaments," Bielefeld Univ., Bielefeld, Germany, Work. Paper, 2017.
[20] C. To, G. J. Kilduff, L. Ordoñez, and M. E. Schweitzer, "Going for it on fourth down: Rivalry increases risk taking, physiological arousal, and promotion focus," *Acad. Manage. J.*, vol. 61, no. 4, pp. 1281–1306, Aug. 2018.
[21] R. Böheim and M. Lackner, "Gender and risk taking: Evidence from jumping competitions," *J. Roy. Stat. Soc. Ser. A, Statist. Soc.*, vol. 178, no. 4, pp. 883–902, Oct. 2015.
[22] E. Feess, H. Müller, and P. J. Jost, "The influence of ability, rank, and gender on risk-taking in contests: Evidence from the ski world cup," *J. Institutional Theor. Econ.*, vol. 172, no. 4, pp. 573–614, 2016.
[23] G. Ruedl, E. Pocecco, R. Sommersacher, H. Gatterer, M. Kopp, W. Nachbauer, and M. Burtscher, "Factors associated with self-reported risk-taking behaviour on ski slopes," *Brit. J. Sports Med.*, vol. 44, no. 3, pp. 204–206, Mar. 2010.
[24] C. Genakos and M. Pagliero, "Interim rank, risk taking, and performance in dynamic tournaments," *J. Political Economy*, vol. 120, no. 4, pp. 782–813, Aug. 2012.
[25] P. Nieken and D. Sliwka, "Risk-taking tournaments–theory and experimental evidence," *J. Econ. Psychol.*, vol. 31, no. 3, pp. 254–268, Jun. 2010.
[26] L. Neuberg and S. Thiem, "Risk-taking in contests with heterogeneous players and intermediate information—Evidence from handball," *J. Sports Econ.*, vol. 23, no. 7, pp. 851–880, Oct. 2022.
[27] C. Grund, J. Höcker, and S. Zimmermann, "Incidence and consequences of risk-taking behavior in tournaments-evidence from the NBA," *Econ. Inquiry*, vol. 51, no. 2, pp. 1489–1501, Apr. 2013.
[28] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *Ann. Statist.*, vol. 47, no. 2, pp. 1148–1178, Apr. 2019.
[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?": Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, 2016, pp. 1–9, doi: 10.1145/2939672.2939778.
[30] X. Luo, X. Lu, and J. Li, "When and how to leverage e-commerce cart targeting: The relative and moderated effects of scarcity and price incentives with a two-stage field experiment and causal forest optimization," *Inf. Syst. Res.*, vol. 30, no. 4, pp. 1203–1227, Dec. 2019.
[31] O. G. Ali, "Targeting resources efficiently and justifiably by combining causal machine learning and theory," *Frontiers Artif. Intell.*, vol. 5, Oct. 2022, Art. no. 1015604.
[32] H. Klüver, F. Hartmann, M. Humphreys, F. Geissler, and J. Giesecke, "Incentives can spur COVID-19 vaccination uptake," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 36, Sep. 2021, Art. no. e2109543118.
[33] S. Athey and S. Wager, "Estimating treatment effects with causal forests: An application," *Observational Stud.*, vol. 5, no. 2, pp. 37–51, 2019.
[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[35] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*. Boca Raton, FL, USA: CRC Press, 1990.
[36] W. Cleveland, E. Grosse, and W. Shyu, "Local regression models," in *Statistical Models in S*, J. M. Chambers and T. J. Hastie, Eds., Pacific Grove, CA, USA: Wadsworth & Brooks/Cole, 1992, Ch. 8, p. 608.
[37] T. Hastie. (2023). *gam: Generalized Additive Models_. R Package Version 1.22–2*. [Online]. Available: https://CRAN.R-project.org/package=gam
[38] D. J. Llewellyn, X. Sanchez, A. Asghar, and G. Jones, "Self-efficacy, risk taking and performance in rock climbing," *Personality Individual Differences*, vol. 45, no. 1, pp. 75–81, Jul. 2008.
[39] M. Hughes, P. Hughes, K. Mellahi, and C. Guermat, "Short-term versus long-term impact of managers: Evidence from the football industry," *Brit. J. Manage.*, vol. 21, no. 2, pp. 571–589, Jun. 2010.
[40] T. A. Gilson, "Outcomes of confidence in sport training settings," *Strength Conditioning J.*, vol. 32, no. 5, pp. 91–96, 2010.



OZDEN GUR ALI received the B.S. degree in industrial engineering from Boğaziçi University and the M.S. and Ph.D. degrees in decision sciences and engineering systems from the Rensselaer Polytechnic Institute. Currently, she is an Associate Professor with the Business School, Koç University. Prior to her academic career, she had ten years of experience in research and development and management consulting in the USA. Her work has been published in journals and conferences, such as IEEE Neural Networks, IEEE Transactions on Knowledge and Data Engineering, *Operations Research*, *European Journal of Operational Research*, *International Journal of Forecasting*, and KDD. Her current research interests include interpretability and causality with AI/ML and big data in retail forecasting, customer behavior modeling, and healthcare.



EMRAH YILMAZ recevied the degree in economics (mathematics) from Sabanci University and the Ph.D. degree from Koç University, with a focus on machine learning and optimization methods. Currently, he serves as the Managing Director of the Analytics Center of Excellence, FTI Delta, which is a global industry-specialized strategy consulting practice. Prior to his role at FTI Delta, he gained valuable experience working with global consulting firms, such as PwC and IBM. Throughout his professional journey, he has successfully overseen the delivery of over 60 projects, with an emphasis on AI and data science. These endeavors have primarily been concentrated in the domains of marketing, sales, and operations, particularly within the telecom, banking, and retail sectors.