European Journal of Operational Research 303 (2022) 1151–1167

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Production, Manufacturing, Transportation and Logistics

Evaluating human behaviour in response to AI recommendations for judgemental forecasting

Naghmeh Khosrowabadi[a], Kai Hoberg[a,*], Christina Imdahl[b]

[a] Kühne Logistics University, Hamburg, Germany
[b] Eindhoven University of Technology, Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 12 May 2021
Accepted 9 March 2022
Available online 14 March 2022

Keywords:
Supply chain management
Forecasting support systems
Judgmental adjustment
Machine Learning
Empirical data

ABSTRACT

Various advanced systems deploy artificial intelligence (AI) and machine learning (ML) to improve demand forecasting. Supply chain planners need to become familiar with these systems and trust them, considering real-world complexities and challenges the systems are exposed to. However, planners have the opportunity to intervene based on their experience or information that the systems may not capture. In this context, we study planners' adjustments to AI-generated demand forecasts. We collect a large amount of data from a leading AI provider and a large European retailer. Our dataset contains 30 million forecasts at the SKU-store-day level for 2019, plus variables related to products, weather, and holidays. In our two-phase analysis, we aim to understand the adjustments made by planners and the quality of these adjustments. Within each phase, we first identify the drivers of adjustments and their quality using random forest, a well-known ML algorithm. Next, we investigate the collective effects of the different drivers on the occurrence and the quality of the adjustments using a decision tree approach. We find that product characteristics such as price, freshness, and discounts are important factors when making adjustments. Large positive adjustments occur more frequently but are often inaccurate, while large negative adjustments are generally more accurate but fewer in number. Thus, planners do not contribute to accuracy on average. Our findings provide insights for the better use of human knowledge in judgmental forecasting.

1. Introduction

Accurate demand forecasting is essential for the reliable planning and execution of all activities in the supply chain. Better forecasts enable supply chain planners to address increasing consumer expectations, shorter lead times, and limited resources (Boone et al., 2018). However, forecasting is difficult for numerous reasons, including the availability and quality of data, the presence of outliers and market shifts, and economic developments (Fildes et al., 2009). This difficulty is exacerbated by the large number of forecasts that must be generated and periodically updated for many products, various locations, and different forecast horizons. Therefore, in practice, many forecasts are initially generated by a computerized forecasting support system but can be later adjusted by human planners, so-called "judgmental adjustments of forecasts" (Lawrence et al., 2006).

An increasing number of these systems are applying state-of-the-art artificial intelligence (AI) and machine learning (ML) methods to facilitate demand forecasting processes and improve forecasting performance (Ferreira et al., 2016). These technologies offer the opportunity to revolutionize planning and radically change the role of the supply chain planner (Alicke et al., 2019). However, the use of AI systems creates new challenges, as AI algorithms are not easy to explain or interpret (Shrestha et al., 2019). In addition, AI systems can utilize abundant variables (e.g., product category, price, weather, or holidays) to generate forecasts, making it impossible for planners to precisely understand how their AI system has generated a forecast and which factors and mechanisms are driving a particular forecast. As a consequence, it is difficult for planners to decide whether adjustments are needed or if they should follow the system. Typically, it is also difficult to rely exclusively on the system, as it might not be able to access important information that is unavailable in structured form (e.g., new product introductions, competitor activities, or local events). Thus, planners can add value to machine-generated

* Corresponding author.
E-mail address: kai.hoberg@the-klu.org (K. Hoberg).

https://doi.org/10.1016/j.ejor.2022.03.017
0377-2217/

forecasts based on their experience and knowledge of the complex environment (Sanders & Manrodt, 2003).

Making judgmental adjustments of forecasts requires considerable effort and time from supply chain planners, and not all of the adjustments improve the forecasts' accuracy (Trapero et al., 2013). Planners aim to distinguish real changes from noisy signals when deciding to make an adjustment (Kremer et al., 2011). In particular, planners are prone to different biases, including optimism, anchoring, and overreaction (Eroglu & Croxton, 2010). Additionally, humans tend to mistrust algorithms and deviate from system advice (Bolton et al., 2012; Dietvorst et al., 2015). Therefore, it is essential to understand under which circumstances adjustments improve forecast accuracy. Baker (2021) suggested the use of a threshold level of forecast value added in distinguishing beneficial adjustments from detrimental ones. Several studies have evaluated the properties of valuable adjustments, such as their magnitude and direction. Fildes et al. (2009) found that large and negative adjustments are more likely to improve forecast accuracy. Harmful adjustments may be triggered by emphasizing single, isolated past events, such as positive adjustments due to optimism regarding past promotions (Fildes et al., 2019).

While a large body of the literature on the drivers of judgmental adjustments of forecasts investigates behavioral effects (e.g., Eroglu & Croxton, 2010; Franses & Legerstee, 2009), the impact of product characteristics has received limited attention. Chen and Li (2019) identified a significant effect of the product margin on the quality of newsvendor decisions. For judgmental adjustments of order recommendations, Sachs (2015) showed the benefit of integrating different product characteristics, such as product-specific costs and demand variability, in an empirical study on bakery sales data. In the field of demand forecasting, however, prior studies have either concentrated on single-product settings or forego considering specific product characteristics. Identifying the effective factors driving the occurrence and accuracy of adjustments can provide insights for improving forecast quality.

We contribute to this stream of research by leveraging a large real-world dataset to better understand when and how humans can improve forecast accuracy when working with AI systems. We specifically aim to address two research questions: (i) When[1] are forecasts adjusted by the planner? (ii) When do forecast adjustments add value? To answer these questions, we have partnered with a leading AI provider and a large European retailer. The forecasts are generated by the AI provider, and the planners of the retailer can adjust the forecasts as needed. On average, over 200 million forecasts are generated and updated each day in this setting. We focus on a sample of ten stores with over 30 million demand forecasts at the SKU-store-day level for 2019. Our main focus is on the effect of product-related factors on the adjustments made and the adjustments quality. Two well-known ML methods are employed to address the research questions: random forest for the former and decision trees for the latter.

We find that the planners adjust only a relatively small fraction of the forecasts. In particular, our random forest analysis indicates that the planners tend to adjust forecasts for expensive, fresh, fast-moving and discounted products, as well as the forecasts made for days with abnormally high temperatures. Only approximately half of these adjustments improve the demand forecast. We also find that price, freshness, direction and magnitude of the adjustment play an important role in the adjustment quality. Interestingly, the occasional events such as high temperatures and discounts embedded in the system only play a secondary role in predicting the adjustment quality. In line with the literature, we find better performance of negative adjustments. We illustrate the collective effects of the different factors on the occurrence and quality of the adjustments using a decision tree approach. For instance, negative adjustments yield better quality when made for fresh and slow-moving products.

The remainder of this paper is structured as follows: Section 2 reviews the related literature. Section 3 describes the research setting, and Section 4 outlines the methodology used. Section 5 reports the results for the forecast adjustments made by planners, and Section 6 reports the results for the quality of these adjustments. In Section 7, we discuss our findings and conclude.

2. Literature review

Various processes in a supply chain rely on demand forecasts, e.g., production planning and replenishment or distribution scheduling. The overall performance of these activities depends on the accuracy of the forecast. In practice, many forecasts are initially generated automatically by a computerized system (Fildes et al., 2009). AI/ML systems work with massive data and make fairly rapid and relatively accurate forecasts, but their output is difficult to interpret. Over the past decade, the development of data collection systems has made a large amount of data available, which has led to better forecasts by AI/ML algorithms (Cui et al., 2018; Feng & Shanthikumar, 2018; Makridakis et al., 2021). However, due to the high uncertainty and complexity of the real world, most companies do not solely rely on the system but rather see the use of it as a decision support system with expectations on its validity, legitimacy, and usefulness (Meinard & Tsoukiàs, 2019). Planners are typically allowed to adjust system-generated forecasts (Fildes & Petropoulos, 2015; Fildes et al., 2009; Ibrahim et al., 2021; Perera et al., 2019; Sanders & Manrodt, 2003; Weller & Crone, 2012). This process is called "judgmental adjustments of forecasts".

Our research relates to two streams of literature around judgmental forecasting. The first stream of relevant literature takes an in-depth perspective on *the reasons for judgmental adjustments of forecasts*. Based on the existing literature, there are two main types of drivers for judgmental adjustments: behavioral drivers/biases and product-related drivers. One of our goals is to detect drivers of the adjustments made and adjustments quality, particularly by focusing on the latter type of drivers, i.e., product-related. Nevertheless, some evidence can be relatable to behavioral drivers or biases. Thus, we review both types of literature to have a comprehensive view of the prior studies.

Adjustments may mirror individual or functional biases caused by informational blind spots or other organizational misalignments in the supply chain Oliva and Watson (2009). Prior research has identified different biases specific to forecasting performance: Systematic affective misprediction (Sevdalis & Harvey, 2007); random error (Fischer & Harvey, 1999); the tendency to under-forecast during the promotional periods and over-forecast during normal condition (De Baets & Harvey, 2018); the tendency to make random errors in judgment or to ignore the value of information in the correlation structure between lower-level demands (Kremer et al., 2016). Additionally, the pull-to-center effect in making order decisions can be reduced by separating the forecast point from the error forecast and splitting the product choice from the order decision (Tong & Feiler, 2017). Humans generally underreact to changes in an unstable environment with obvious signals and overreact in a stable environment with noisy signals, which is called overreaction bias (Eroglu & Croxton, 2010; Kremer et al., 2011). Additionally, Petropoulos et al. (2016) found that big losses undermine future judgments and recommend using simple strategies such as the Blattberg-Hoch: 50–50 decision rule to mitigate the adverse effect of a big loss. Other behavioral factors in the forecasting context are optimism (Eroglu & Croxton, 2010; Fildes et al., 2009;

[1] Under which conditions, this also holds true throughout this paper.

Franses & Legerstee, 2009; Kirshner & Shao, 2019) and overconfidence (Kirshner & Shao, 2019; Ren & Croson, 2013).

In contrast to comprehensive studies on behavioral aspects, researchers have paid little attention to product-related drivers, particularly to multiple-product settings and product characteristics. Since most systems use only historical data, many cannot capture additional novel information, such as future promotions or upcoming events. Franses and Legerstee (2009) found that adjustments mainly emanate from planners' optimism regarding promotions. Embedding such events into the system is not easy, as they occur infrequently (Fildes et al., 2022). Hence, judgmental forecasting is usually adopted to capture these effects (Fildes & Goodwin, 2007), but it can degrade forecasts by focusing on single, isolated pieces of information from past events or overweighting the impact of an event, such as a celebrity's support (Fildes et al., 2019), or considering too much weight for human contributions (Franses & Legerstee, 2010). We investigate the role of price promotions in the adjustments made and their quality.

There is a similar investigation of drivers of judgmental adjustments to order decisions. Chen and Li (2019) experimentally showed that the product groups assigned to newsvendors should be homogeneous in terms of product margin, i.e., a combination of high- and low-margin products in the same group results in different adjusting behavior. In a study on bakery sales data, Sachs (2015) demonstrated the benefit of integrating different product characteristics (demand variability and cost-related information) into their forecasting model to reduce the group aggregated bias. Van Woensel et al. (2007) realized higher substitution rates for perishable products (particularly bakery bread) than for non-perishables, which may affect the forecast adjustments. Van Donselaar et al. (2010) found that planners adjust order recommendations due to a misalignment of the system's objective and the planners' incentives; they advance orders from peak to non-peak days by considering product characteristics, such as case pack size, net shelf space, product variety, demand uncertainty, and seasonality error. We study the role of product-, holiday-, and weather-related variables in the adjustment made and adjustments quality in the demand forecasting context.

The drivers of adjustment decisions have also been studied outside the forecasting context. Elmaghraby et al. (2015) explained the salespeople's response to price recommendations from a decision support tool by a two-stage decision model: a binary decision to change the price and a continuous response for the magnitude of change. They found customer-specific and salespeople-specific market factors as moderators of the relationship between the system and salespeople. Concerning the product price, buyers acquire more high-value products when an inventory shortage is costly (Foster et al., 2019). Hence, if firms can anticipate the buyers' behavior, they can adjust the price to earn better profit. Our study accounts for the role of price in the adjustments made and their quality.

The human tendency towards adjusting or non-adjusting system recommendations may also be related to algorithm aversion. For instance, people lose their trust in an algorithm faster than they lose trust in a human after seeing the algorithm/ the human making errors (Dietvorst et al., 2015). In the newsvendor problem, a system recommendation based on the optimal order quantity may pronounce human biases, such as algorithm aversion and regret aversion (Feng & Gao, 2020). While disclosing the underlying process of generating the recommendations is supposed to increase human trust in the system, increased transparency of the complex AI algorithm can even harm the use of algorithmic advice (Lehmann et al., 2020). Recent research using field experiments yields a better image of the decision makers' propensity to follow the algorithmic advice when: investigating the adjustments benefit for different product life cycles (Kesavan & Kushwaha, 2020), and considering the sales effects (Kawaguchi, 2021). While algorithm aversion is not the main focus, nevertheless we can infer some insights about the planner's thought process, based on the forecasts they choose to adjust.

The second stream of relevant literature investigates the *impact of judgmental adjustments on forecast quality*. Several studies aim to understand which adjustments have a positive impact on the forecast quality. While large adjustments have a greater chance of adding value, as they are expected to account for substantial changes or reliable information about upcoming events that are not captured by the system, small adjustments are known to be less effective (Fildes et al., 2009). Planners generally perform better when making negative adjustments than when making positive adjustments, as the latter are often made in the wrong direction due to reasons such as optimism, as mentioned in the first stream (Davydenko, 2012; Fildes et al., 2009). Despite this characteristic, Davydenko and Fildes (2013) found that positive adjustments are more frequent and twice as large as negative adjustments. To distinguish valuable from detrimental adjustments, Baker (2021) introduced the idea of using a threshold level of forecast value added. He predicted the threshold using classification trees based on the adjustments' features such as size and direction, and product forecastability. In an empirical study, Syntetos et al. (2016) detected a small benefit of human adjustments on statistically derived inventory decisions (order-up-to levels). In line with the literature above, they found better performance of adjustments in the negative direction, but, contrary to others, not a better performance of adjustments of larger magnitude. Hence, it remains unclear whether positive or negative, large or small adjustments are more beneficial, and it may be context-dependent. While not the main focus of our analysis, we study the effect of the direction and magnitude of the adjustment on the adjustment quality. We therewith provide one more piece of evidence that is necessary for a holistic view on this question.

The impact of human actions on decision quality is also of interest outside the forecasting context. In an empirical study, Käki et al. (2019) realized that notwithstanding some efficient adjustments, planners often worsened the performance of the studied production model. They explain this negative impact based on model inaccuracies and simplifications as well as challenges in input data and proposed ideas such as input-adjusting instead of output-adjusting to improve the process. Based on the setting characteristics, they generalized their findings to other fields such as inventory management and operations planning. The decision speed or time required to adjust further affects the quality of judgments (Dane & Pratt, 2007; Moritz et al., 2014).

We contribute to both streams above, as we aim to know when an adjustment is likely to happen (the first stream) and when an adjustment is likely to contribute to forecast quality (the second stream). In summary, our contribution is threefold: First, we add to the literature on judgmental forecasting by studying different product-related drivers that are more in number and variety than ever before. In other words, our study consolidates several factors studied separately before. We investigate the individual and collective impact of the factors, such as price, product category, discount, and shelf life, on the adjustments made and adjustments quality. Second, we analyze a setting where the number of forecasts is massive, and planners have limited time to adjust individual forecasts manually. Consequently, they need to know when their adjustments can improve the forecast quality. Third, we analyze how planners react in complex settings to forecasts generated by advanced AI algorithms. In contrast to other settings studied in the prior literature, the system already considers many variables, such as weather forecasts, promotions, or national/local holidays, to create forecasts. Moreover, we add to the stream of literature that applies empirical data to obtain insights. To the best of our

knowledge, this is the first study in our field that aims to evaluate human responses to AI recommendations by ML techniques.

3. Empirical setting and data

In this section, we provide a comprehensive description of the research setting and then introduce the variables of our analysis. In addition, we show some descriptives of the forecast adjustments.

3.1. Research setting

In this research, we analyze data from a prominent AI provider and a large European retailer that prefer to stay anonymous. The AI provider is a market leader in retail analytics and offers AI and ML for different supply chain challenges. The company uses advanced analytics methodologies that leverage hyper-scalable cloud computing solutions and consider hundreds of variables to create insights. The retailer has more than 500 stores and offers a large product portfolio with approximately 40,000 different SKUs. Both parties engage in the demand planning and demand forecasting process, and the forecast is then used to automatically calculate replenishment orders to the stores based on MRP logic, considering factors such as planned lead times, regular delivery appointments, pack sizes and shelf life information.

In the demand planning process, the AI provider first gathers information on each product including its product group, past sales, prices, past promotions, and events, from the retailer. Based on this and additional data, such as weather forecasts, planned promotions or upcoming holidays, the demand forecasts are generated using ML/AI methods for each SKU in each store up to 21 days ahead. Hereafter, whenever we use the term "forecast", we refer to a forecast that is made for day d, product p, and store s. Overall, the AI provider generates (and updates) more than 200 million demand forecasts on the SKU-store-day level each day for this single retailer. As the demand forecasts are created every day, they are fed into the retailer's ERP system. The AI provider aims to create an accurate demand forecast with minimal deviation from actual demand.

Subsequently, a central team of about 200 planners at the retailer can adjust the provided demand forecasts arbitrarily, i.e., upwards and downwards. The planners are generally organized by product teams but also exchange positions and responsibilities regularly. To facilitate these adjustments, the planners can apply any adjustment not only to a single product in a single store but also to entire product groups or categories or across stores. For single products, an adjustment can be made by changing the absolute value. To change multiple forecasts simultaneously, adjustments are typically specified as percentages of the AI forecast. If a forecast is adjusted, the adjustment overwrites the current forecast and further forecast updates do not impact the adjusted forecast.

With the same rationale as the AI provider, the objective of the planners has been to adjust the forecast to match actual demand, i.e., actual sales plus potential lost sales. Notably, in this setting, the planners have not received any information on the adjustment quality, i.e., there is no feedback loop between the two parties (the AI provider has been in the process of designing a feedback loop; but this was not available to the planners at the retailer at the point where the analysis is carried out). The AI provider offers planners training on the AI system, including the onboarding training, i.e., the instructions, guidance and training the planners get once they take over the planning role. In addition, a learning portal with online videos was introduced to better educate the planners. In addition, new planners receive training from their senior colleagues. Planners are informed that AI-generated forecasts are generally highly accurate but do not have a solid background in the black-box algorithms used to create those forecasts.

The setting we study is well suited for exploring the drivers of the forecast adjustments. First, data were made available on a large variety of products with diverse characteristics that could play an important role for planners when adjusting the forecasts. Second, knowing the original and adjusted forecasts when compared with actual sales has allowed an analysis of the quality of the adjustments.

3.2. Data description

We collected data from the AI provider's data repository for the case retailer from various sources, e.g., product groups, stores, calendar, and adjustments. Ultimately, we attained approximately 30.6 trillion final forecasts for the complete year 2019 (based on the number of stores, the average number of SKUs per store and the operating days in the year). Since we only had access to data for the actual sales, all forecasts in our study have non-zero sales. However, based on the missing dates among all SKU-store-day combinations, we could identify the days with zero sales.

Given the massive data volume and the implications for computational requirements, we randomly selected ten stores from the 500 total stores to obtain a tractable problem size. This approach provided us with approximately 31.5 million forecasts at the SKU-store-day level for the entire year. We inspected our sample to ensure that it is representative of all stores. We found that our sample matches the overall geographical distribution and has a similar distribution of store sizes as the overall sample. We also compared the adjustment percentage, i.e., the portion of adjustments in all forecast decisions, in our sample vs. the full dataset. They match fairly closely; the planners adjusted approximately 5.1% of all forecasts on the SKU-store-day level. Due to the confidentiality, we are not able to reveal data in detail, as the AI provider and the retailer like to stay anonymous. Nevertheless, we report essential information and findings in an anonymous form in the rest of the document whenever needed.

Fig. 1 illustrates the daily percentage of adjustments for 2019. To address the Christmas season, the retailer used a different approach to that offered by the AI provider and fixed volumes with its suppliers well before Christmas. Hence, nearly all forecasts of the AI system were overwritten. Since considering this period could distort the analysis, we excluded the last two weeks of December (approximately 90.0% of forecasts are adjusted). The elimination of any anomalies via data cleansing processes (e.g., duplications, exceptional excess adjustments at the end of the year, and sales outliers) left approximately 30.1 million forecasts for 2019. According to our initial descriptive findings, the percentages of adjustments varied across days of the month and months of the year. Thus, we appended the holiday-related and weather-related variables to our data, elaborated in the next section.

3.2.1. Dependent variables

We conduct our analysis based on two dependent binary variables: (i) adjustment made ("adjusted") and (ii) adjustments quality ("valuable").

The dependent variable "adjusted" is a binary variable indicating whether the planners made an adjustment. Our data includes the original forecast made by the AI system and the adjusted forecast made by planners at the retailer. Accordingly, the variable "adjusted" is equal to 1 if the adjusted forecast differs from the AI forecast at the specific SKU-store-day level and 0 otherwise. We use this as the output variable to investigate when planners adjust the AI forecasts.

The dependent variable "valuable" serves as a binary output variable to determine the adjustment quality, which we define based on the absolute deviation of the AI forecast and the planners' adjusted forecast from the actual demand. Thus, "valuable"

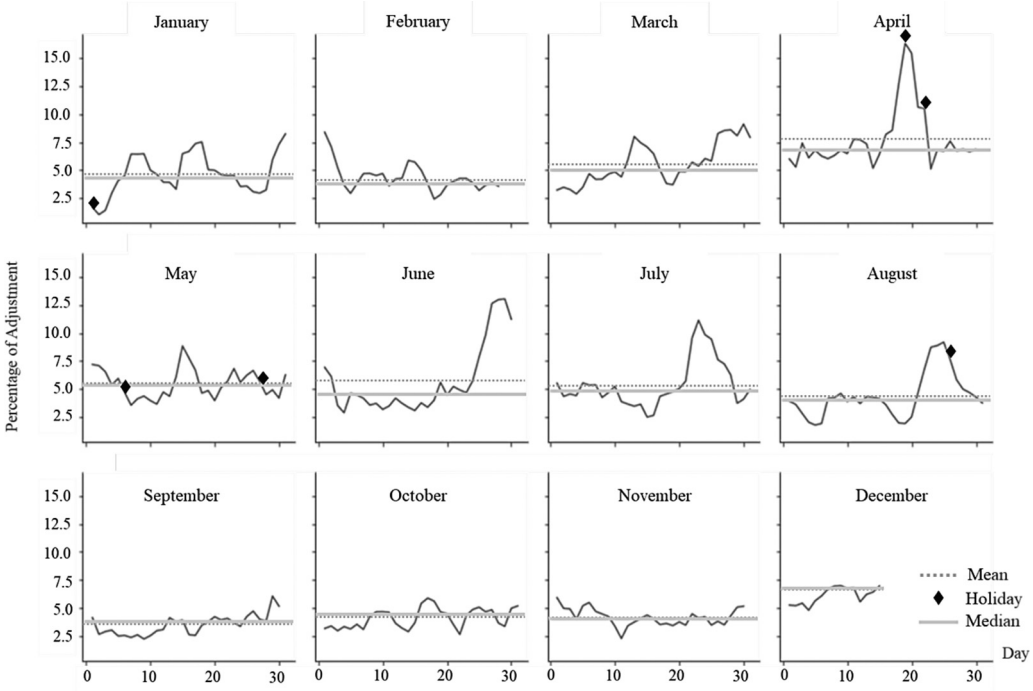

Fig. 1. The percentage of adjustments over time in 2019 (sampled dataset).

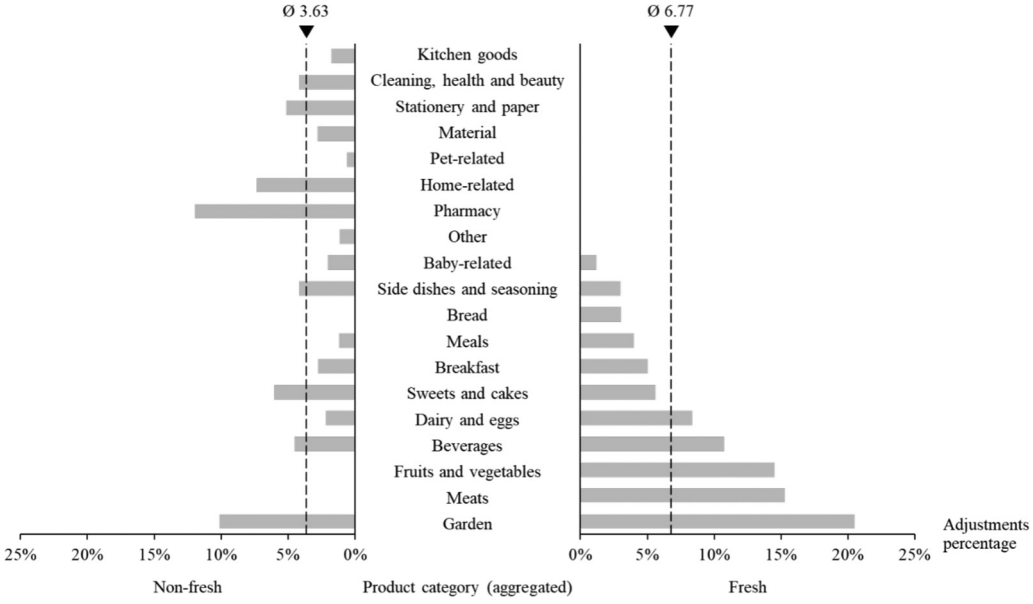

Fig. 2. The frequency of adjustments by product category (in aggregated level).

is equal to 1 if the adjusted forecast is closer to the actual demand and 0 otherwise. As a proxy for actual demand, we utilize the associated sales quantity for each forecast. Since we cannot access stock-out or inventory information, we cannot consider potential stock-outs to account for the exact actual demand and have to deal with censored demand data. However, we would assume that these effects do not have a strong impact on planners' actions, as they are not available at the time of the forecast adjustment.

3.2.2. Independent variables

Directly from the data, we can obtain information regarding the SKUs, stores, and dates for which forecasts are made. We consider different types of variables: numeric (price and sales quantity), categorical (month, store, day of week, and unit type), and binary (freshness and price type). Fig. 2 exhibits the adjustment frequency per product category, which shows a greater share of adjustments for fresh products. We employ the "freshness" variable to distinguish this perishability-related effect.

Moreover, to investigate the adjustment quality, we create two new variables based on adjustment-related information: "adjustment direction" and "adjustment magnitude". The first indicates the sign of an adjustment. The sign is "positive" when the adjusted forecast is larger than the AI forecast and "negative" when the adjusted forecast is smaller than the AI forecast. We define magnitude based on the discrepancy between the adjusted forecast and AI forecast, normalized by seasonal mean sales. We normalize by the seasonal mean sales of the SKUs so that we can compare the magnitude between different SKUs. An adjustment may be large in quantity but small in relation to the average sales of the specific SKU.

Table 1
Description of variables.

Variable	Description
Dependent variables	
Adjusted	Binary variable indicating if there is an adjustment to the system forecast for day d for product p in store s
Valuable	Binary variable indicating if an adjustment increases the forecast accuracy (based on the closeness to real demand) for day d for product p in store s
Independent variables	
Product-level variables	
Product ID	Unique identifier of product p
Product category	High-level aggregation of the product group of product p
Price	Price of product p on day d in store s
Price type	Categorical variable indicating the price state on day d for product p in store s; it is reduced when the product is on promotion and normal otherwise
Unit type	Categorical variable indicating product unit; e.g., kg, pieces
Freshness	Binary variable indicating if a product is fresh or non-fresh based on the product category
Store-level variables	
Store ID	Unique identifier of store s
Sales quantity	Amount of sales on day d for product p in store s
Date-related variables	
Date	Date for which the forecast is made
Month	Month for which the forecast is made
Day of week	Day of week for which the forecast is made
Adjustment-related variables	
AI forecast	Forecast that is generated by the AI system for day d for product p in store s
Adjusted forecast	Forecast that is generated by planners' adjustment to an AI forecast for day d for product p in store s
Adjustment direction	Binary variable indicating if an adjustment is made in the upward direction
Adjustment magnitude	Absolute difference between adjusted forecast and AI forecast normalized by seasonal mean sales
Large magnitude	Binary variable indicating if an adjustment belongs to the 4th quartile of the adjustment magnitude
Small magnitude	Binary variable indicating if an adjustment belongs to the 1st quartile of the adjustment magnitude
Holiday-related variables	
Public holiday	Binary variable indicating if there is a public holiday on the forecast date, based on regional differences
Day before public holiday	Binary variable indicating if there is a public holiday the day after the forecast date, based on regional differences
Weather-related variables	
Daily temperature	Daily temperature for the location of each store s on forecast day d
Mean temperature	Monthly mean temperature for each store location based on the associated weather station over the past 30 years (1989–2018)
Temperature difference	Difference between the temperature on day d for store s and the monthly mean temperature over the past 30 years (1989–2018) for store s
High temperature	Binary variable indicating if the temperature difference for day d at store s is in the top 10% of the highest differences
Low temperature	Binary variable indicating if the temperature difference for day d at store s is in the top 10% of the lowest differences

In addition, we use external information concerning holidays and weather conditions to examine their effects. We create two holiday-related variables, "public holiday" and "day before public holiday", based on the stores' locations and their regional calendar. For the weather-related variables, we add two variables pertaining to the temperature at the store, as temperature is known to impact the sales (Badorf & Hoberg, 2020). First, we obtain the monthly mean temperature for each store location based on the associated weather station over the past 30 years (1989–2018). Then, we calculate the temperature differences between the temperature on day d for store s and the monthly mean temperature (over the past 30 years) for store s. Last, we create two binary variables based on the temperature difference, "high temperature" and "low temperature", indicating whether day d belongs to the 10% hottest days or the 10% coldest days compared to the past 30 years, respectively. Table 1 provides an overview of the variables. The research setting inspires the choice of variables, i.e. we mostly only include variables that were directly available to the planner. Other settings may inspire a different choice of variables (Mathews & Diamantopoulos, 1989a, 1989b).

4. Methodology

4.1. Machine learning approaches

The analysis of the dataset requires a method capable of accounting for the different types of variables available and of considering their ambiguous relationships. Semi- or non-parametric methods from ML commonly perform well in these settings compared to traditional techniques, such as linear models (Cui et al., 2018). A generalized linear model such as a logistic regression may appear applicable for our case, as we have binary dependent variables. Logistic regression uses inputs to model the probability of a binary output and investigates each variable's impact on the odds ratio of the output. An important limitation of this method is the linearity assumption between the dependent variables and the log odds of the independent variables.

To capture non-linear relations between variables on a large scale, we use decision trees (DTs) and random forest (RF), considering their particular advantages and disadvantages (see Lantz, 2013, for a detailed discussion). We reviewed some of the ML-related literature in Section 2 to show the potential superiority of these techniques for forecasting, especially for larger datasets. Here, we elaborate on the two methods applied and the overall research methodology.

Decision trees are well-known ML algorithms that convey relationships between the independent variables and a dependent variable. DTs use the notion of recursive partitioning, in which the sample is divided into subsamples at each step to create the subsamples that are most homogeneous with respect to the dependent variable. Data flow through decision nodes at the root node; then, partitioning is executed by a heuristic search for the best split across all variables according to their values or categories. Thus, the branches of a DT show how the partitions split the data.

Table 2
Overview of analysis phases.

	Phase 1	Phase 2
Research question	When are forecasts adjusted by the planners?	When do forecast adjustments by the planners add value?
Dependent variable	Adjusted forecasts: "Adjusted"	Valuable adjustments: "Valuable"
Sample	30.1 mn forecasts (on SKU-store-day level) of which 5.1% are adjusted	1.5 mn forecasts (on SKU-store-day level) of which 49.9% are valuable
Analysis steps	(1) Identify the required sample size	(1) Use the full sample size
	(2) Identify drivers of adjustments/quality by random forest	
	(3) analyze the effect of adjustments/quality drivers by marginal distribution plot	
	(4) Identify predictors of adjustments/quality by decision tree	

Each sequence of partitioned independent variables from the root node to the final (terminal) nodes reveals a rule that governs the data (Lantz, 2013; Shmueli & Yahav, 2018).

There are various possible decision trees based on the criteria by which a DT selects and branches the variables while avoiding overfitting the data. We use the C5.0 model from the classic type of decision trees. C5.0 selects the variables at each split based on the node impurity and prunes the partitions that do not contribute substantially to predictive power on a test dataset (Shmueli & Yahav, 2018).

Random forest, first introduced by Breiman (2001a), builds a large number of specialized decision trees and then ensemble their outputs. RF establishes a multitude of decision trees at training time and returns the mode or mean of all trees' predictions as the output for classification or regression problems (Hastie et al., 2009). Based on the commonly built-in cross-validation procedure, RF typically prevents overfitting. The method successfully handles noisy or missing data and can be used for extremely large sets of variables (e.g., Robin et al., 2017). RF works with categorical and numeric variables simultaneously, which is essential for the current research given the mixed types of variables available.

RF aims to investigate the relationship between the dependent variable and the independent variables. Various measures can be used to diagnose variables' importance; we chose to use permutation, which randomly shuffles each of the input variables and tracks changes in model accuracy to assess the impact of the variable on the output, i.e., how important a variable is (e.g., Breiman, 2001b; Nicodemus et al., 2010). The main disadvantage is that permutation can be computationally expensive, particularly for a large number of variables. Regardless, permutation is typically preferred for ensemble methods such as RF (Stringer, 2018).

Despite the many advantages for RF, there is one main disadvantage: the outcome of RF is not self-explanatory (Lantz, 2013). This is why we employ DT for a managerial interpretation of the results obtained by the RF. We rely on RF to identify important factors in the occurrence and quality of the adjustments. In addition, we employ DT to illustrate the collective effects of the independent variables on our two variables of interest, the "adjusted" and "valuable" dependent variables. We next elaborate on our analysis phases and steps.

4.2. Analysis phases

Our study encompasses two analysis phases: the first phase pertains to identifying factors related to the adjustments made, and the second relates to the adjustments quality (see Table 2). Each phase consists of four analysis steps: (i) defining the required sample size, (ii) identifying important drivers of the dependent variables, (iii) analysing the effect of the drivers, and (iv) identifying predictors of the dependent variables. Initially, we employ the RF method for the first and second steps. Then, we directly turn to the data and assess the effects of variables through some basic statistics. Finally, we investigate the multivariate relationships between variables using a DT. We discuss all steps in detail below:

(i) Required sample size: The data volume represents a challenge for our analysis. For the first phase of "adjustments made", the data amount to 30.1 million forecasts. Due to the inefficiency of executing the algorithm for millions of data points, we aim to identify a sample size that makes our analysis computationally feasible while ensuring quality results. Thus, we compare the performance of RF across various sample sizes for the first phase while recording the accuracy and runtime of each run. Beyond a certain sample size, the change in accuracy is negligible, while the diagnosed important variables remain almost the same. For the second phase of "adjustments quality", considering the adjustment rate, the data volume reduces to approximately 1.5 million adjusted forecasts. Given this relatively low data volume, there is no need to further reduce the sample size.

We compare the RF method with a benchmark from traditional models (logistic regression). According to the innate characteristics of the data and the research objectives, logistic regression, which is commonly known for its extensive functionality, can serve as our benchmark. We used AUC[2] to measure the accuracy for both RF and logistic regression. For a sample size of 5 million (20% as a holdout), the AUC of the RF remains fairly constant in comparison to larger sample sizes. We have detailed the analysis in Appendix A. The percentage of the adjustments for this sample size is fairly matched with that of the overall sample. To ensure the validity of our findings beyond the selected stores, i.e., their generalizability to other stores of the case retailer, we replicate our analysis for a different set of stores. Our results attest to robustness and external validity of this analysis.

(ii) Important drivers: Exploring the important variables exposes how much a variable affects the adjustments made or adjustments quality. Using RF, we answer the questions of when planners adjust AI forecasts and when these adjustments add value to the forecasts.

(iii) Effect of important drivers: To better understand RF outcomes, we illustrate the marginal distribution of the variables diagnosed as important drivers of the dependent variable at each phase of our study, neglecting multivariate interactions (investigated in step iv). For this purpose, we compare the percentage of adjustments between different categories of a variable.

(iv) Predictors of the dependent variables: To shed more light on the interplay between the different independent variables, we

[2] AUC is the area under the ROC curve. The better the predictive power of a model is, the greater the AUC. Ideally, AUC equals 1, and, in the worst case, 0. The ROC curve shows the true positive rate (recall) against the false positive rate (1-specificity).

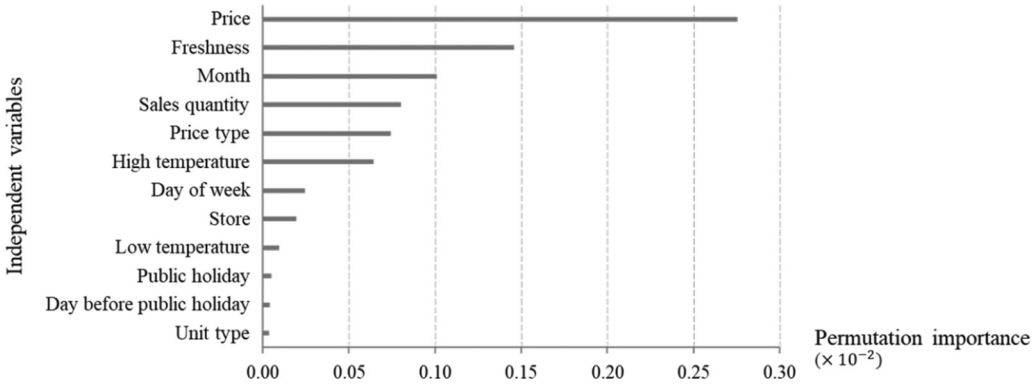

Fig. 3. Important drivers of the adjustments made (based on the training dataset).

visualize and interpret the collective effects of the independent variables on the dependent variable at each phase of our analysis via a DT.

5. Analysis of forecast adjustments

To address our research questions, we first aim to understand when planners adjust AI forecasts (phase 1). Here, we use RF to identify the drivers of adjustments. However, RF analysis shows the importance of variables in isolation and do not immediately reveal interactions. In the next step, we aim to clarify how the variables collectively affect the planners' decision to adjust a forecast. Here, we benefit from the DT methodology, which illustrates the collective conditions of different variables under which an adjustment is more likely to occur. In this phase of our study, the "adjusted" variable serves as the dependent variable for the analysis.

We include all of the independent variables to feed the algorithm, except those that are based on the adjustment itself, such as the adjustment magnitude or direction. Hence, we include price, price type, unit type, freshness, store name, sales quantity, month, day of the week, public holiday, day before the public holiday, high temperature, and low temperature. In line with the methodology presented in Section 4, we ran the analysis using the caret and ranger packages in R. The average runtime on a macOS Mojave Version 10.14.6 (MacBook Pro) computer with a 2.9 gigahertz Intel Core i5 processor and 16 GB 1867 megahertz DDR3 memory was 11 hours and 17 minutes for a dataset with 5.0 million forecasts.

5.1. Drivers of adjustments

Fig. 3 shows the drivers of forecast adjustments according to permutation importance (e.g., Breiman, 2001b; Nicodemus et al., 2010). A higher value for the permutation importance indicates that the variable has a higher impact on the dependent variable. Accordingly, price is the most important factor, followed by the freshness of a product category, forecasting month, the number of products sold, the presence of a promotion or discount (price type), and abnormally warm weather. Notwithstanding the high-temperature variable's rational influence on sales, e.g., for fresh fruits, BBQ meat, or ice cream, a notable decline in temperature does not seem to have a significant impact; unsurprisingly, we do not see low temperature in the top-ranked variables.

The difference in the importance of the variables is interesting. From a planner's perspective, greater attention is paid to expensive items, as a stock-out of such an item is likely to cause a higher loss. Hence, the price is a primary factor guiding the planner's adjustment focus. The product's freshness has a similar association, as perishability increases the mismatch costs. In addition, the planner reacts to uncommon events, such as discounts (reduced price type) and weather-related effects (high temperature). These latter results are surprising, as weather forecasts and promotions are already included in the AI system. Planners might not be fully aware of the capabilities of the AI system. We believe this is a legacy of working with more traditional forecasting systems and can be attributed to the black-box nature of the provided AI forecasts.

Among the identified drivers, "month" may not appear as rational as the others; we believe it may serve more as a control variable. However, we must investigate its effect over the years to judge. This investigation is outside of our research scope, as we only observe one year of data, but it represents a potential direction for future research.

Due to the presence of many categorical variables, RF does not allow for an estimation of effect sizes or effect directions. To interpret the RF outcome and provide some intuition on the effects of the identified important variables, we examine the marginal distribution of the important variables with respect to the dependent variable. Note that this analysis does not address multivariate interactions, which we will return to in Section 5.2.

Price. Fig. 4 shows that the AI forecasts for more expensive products, which constitute a small share of all products, are more likely to be adjusted. This result supports the substantial role of price in making the adjustments.

Likewise, we illustrate the percentage of forecasts and the percentage of adjustments for important binary variables in Fig. 5.

Product Freshness. Fig. 5-a shows that 45.7% of all forecasts relate to fresh products, 6.8% of the forecasts for fresh products are adjusted, and only 3.6% of the forecasts for non-fresh products are adjusted.

Price Type. Fig. 5-b shows that 22.5% of all forecasts are made for products with a price reduction, 6.3% of adjustments are made for those forecasts with reduced prices, and only 4.7% of the forecasts are adjusted given normal prices.

Temperature. In Fig. 5-c, the days with high temperature (warmer days) are those 10% of the days with the highest temperature compared to the past 30 years. Correspondingly, the days without a high temperature (not-warmer days) are the remaining 90% of the days. The percentage of adjustments for warmer days is higher than that for not-warmer days (7.3% vs. 4.8%).

There are apparent discrepancies between the percentage of adjustments per category for product freshness (which has a more balanced frequency distribution) and price type and high temperature (which have fewer reduced prices and fewer days with high temperatures). For example, forecasts of fresh products are adjusted 6.8% of the cases, while non-fresh products are adjusted 3.6% of the cases. These findings confirm that RF's diagnosed variables can provoke planners to adjust the AI forecasts. We further analyze the interactions between important binary variables and the price correlation with these binaries (Appendix B).

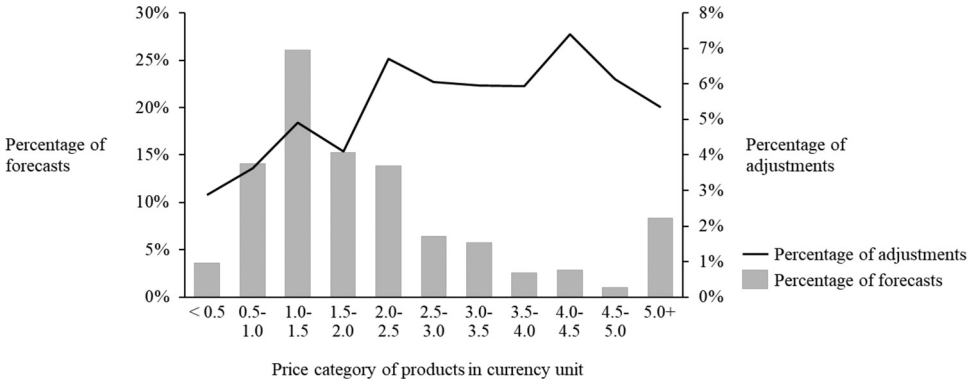

Fig. 4. Adjustments and frequency per price category of products in currency unit.

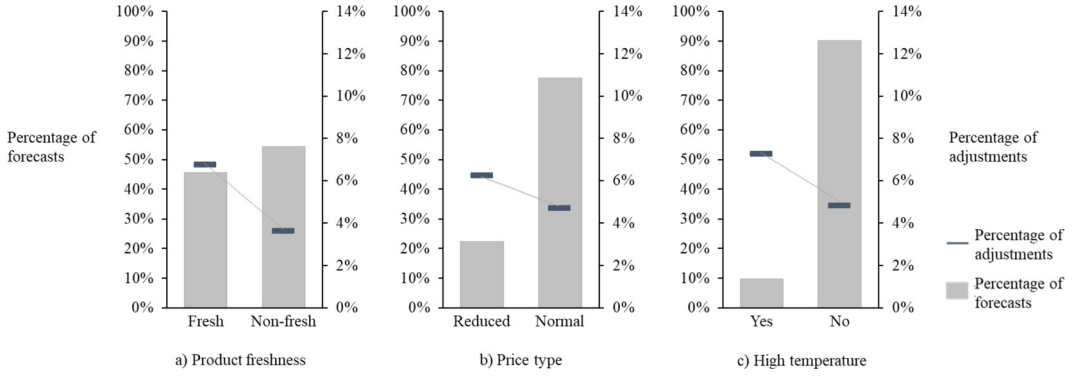

Fig. 5. Effect of the important binary variables on the adjustments made.

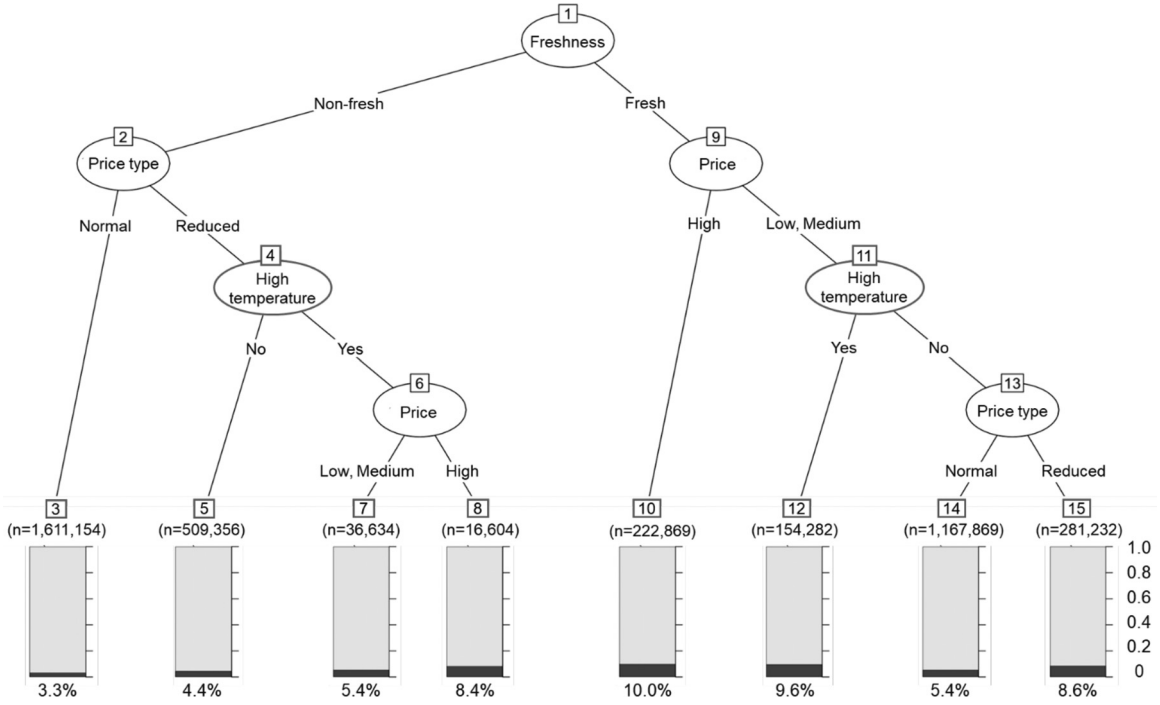

Fig. 6. Decision tree based on adjusted variable with a set of important drivers as predictors on the training set.

Taking the insights collectively, the more important the variables are, the more considerable the discrepancy in the percentage of adjustments will be.

5.2. Predictors of adjustments

In this next step, we aim to understand the collective impact of the different variables using a DT. According to each variable's effect on the adjustments made, we choose a set of important drivers for a tree-based example of the relationships, including price, product freshness, price type, and temperature. To simplify the presentation, we categorize the price into three categories (below 1.5 currency unit, between 1.5 and 3 currency unit, and above 3 currency unit).

Fig. 6 represents a tree-based scheme of the relationships between the "adjusted" variable and its predictors. The DT recursively evaluates the independent variables to select the variable that best splits the data and reduces the impurity at each node. Here, branching is started with product freshness at the root node. The middle nodes show the interactions of the variables, ending with the adjustment probability in the terminal node at the bottom of the DT. Each set of branches from the root to a terminal node reveals a set of product characteristics under which an adjustment is either more or less likely to occur. The DT's ascribed probabilities are interpreted in comparison to the 5.1% overall rate of adjustments.

The comparison of different branches in the DT yields some interesting insights. For instance, for fresh products on the right side of the DT in Fig. 6, expensive products have the highest odds of being adjusted with an adjustment probability of 10.0% (Node 10). However, fresh products with low or medium prices on not-warmer days and at normal prices have far fewer odds of being adjusted with an adjustment probability of 5.4% (Node 14). This is also the node with the highest number of forecasts for fresh products with ($n = 1{,}167{,}869$). For non-fresh products on the left side of the DT, those at normal prices have the lowest odds of being adjusted, with an adjustment probability of 3.3% (Node 3); less than one third compared to node 10. This is also the node with the most observations ($n = 1{,}611{,}154$). The results relate well to our insights regarding RF. Analogous to our insights from Section 5.1, the freshness of a product, the price, a price discount, or a high temperature increase the likelihood of adjustments, also when considering multivariate interactions.

6. Analysis of the quality of forecast adjustments

We next aim to understand when an adjustment is valuable (phase 2). In line with Section 5.1, we first use RF to identify the characteristics of valuable adjustments. As the RF analysis shows the important variables in isolation, in the next step, we illustrate how the different variables collectively affect the adjustment quality. In line with Section 5.2, we utilize the DT methodology to identify the collective conditions of different variables under which an adjustment is more likely to add value and contribute to forecast accuracy. In this phase of our study, the "valuable" variable serves as the dependent variable for the analysis.

In addition to the variables used in Section 5, we include variables that relate to the direction and magnitude of the adjustment made. As mentioned earlier, we focus on only 1.5 million adjusted forecasts. The computation setup remains the same as before.

Before proceeding, we first discuss the direction and magnitude of adjustments with respect to their frequency and quality Table 3. provides a summary of the frequency of small/large and positive/negative adjustment for all adjustments made. There are 70.2% adjustments in positive direction (increases from the AI forecast) and 29.8% adjustments in negative direction (decreases from the AI forecast). In line with Section 3.2.2., we defined large- (small-)magnitude binary variables based on membership in the upper (lower) quartile of an adjustment's magnitude; thus, each of these binary variables is defined as 1 in 25% of the cases for the highest (lowest) relative adjustments. Ultimately, the direction and magnitude must be considered together. While some prior studies believe in the better performance of large adjustments than small adjustments (Fildes et al., 2009), there is also evidence for their worse performance (Syntetos et al., 2016). Hence, we need to investigate the effect of magnitude in interaction with direction, as we find positive adjustments to be much more frequent than negative adjustments. The number of large adjustments made in the upward direction is 4 times greater than the number made in the downward direction, while this ratio is only 1.5 for small adjustments. Taken collectively, the results indicate that (i) planners make more positive adjustments and (ii) positive adjustments are more likely to be of large magnitude.[3] We next discuss more nuanced insights into the interaction of the direction and magnitude.

Table 3
The frequency of all adjustments with the specified magnitude and direction.

		Direction		
		Positive (%)	Negative (%)	Total (%)
Magnitude	Large	19.7	5.3	25.0
	Medium*	35.2	14.8	50.0
	Small	15.3	9.7	25.0
	Total	70.2	29.8	100.0

* Medium covers all the adjustments between the small and large magnitude.

Table 4
The percentage of valuable adjustments by magnitude and direction.

		Direction		
		Positive (%)	Negative (%)	Total (%)
Magnitude	Large	39.8	47.3	41.4
	Medium	54.6	42.6	51.1
	Small	66.6	38.9	55.9
	Total	53.1	42.3	49.9

We initially compare the MAPE of the forecasts made by the system and the forecasts adjusted by the planners (note that we cannot report the absolute forecast accuracy due to a non-disclosure agreement with the research partner). The average system (pre-adjustment) accuracy for adjusted forecasts is approximately 5.0% lower than that for the unadjusted forecasts. Accordingly, we believe that planners generally select the right AI forecasts to adjust. Upward adjustments result in an average 22.7% decline in accuracy, while downward adjustments enhance the accuracy by 11.3%.[4] Thus, negative adjustments typically contribute more to accuracy than positive adjustments do, as expected from prior research (Fildes et al., 2009).

Overall, approximately 49.9% of the adjustments are valuable; in other words, only half of the adjustments improve forecast accuracy. Earlier, we indicated a considerable average accuracy loss for positive adjustments. A high number of the positive adjustments with a large magnitude does not seem to provide value. For more clarification, Table 4 shows the percentage of valuable adjustments for the different magnitudes and directions. For instance,

[3] Note that the magnitude of negative adjustments has a lower bound at zero, whereas positive adjustments are unbounded. Accordingly, positive adjustments have the potential for larger adjustments.

[4] Mean values of relative error measures such as the MAPE may be distorted in the presence of slow-moving items (Davydenko and Fildes, 2013). Our findings are qualitatively robust though using the AvgRelMAE, showing that upward adjustments result in 37% decline in accuracy, while downward adjustments enhance the accuracy by 15%.

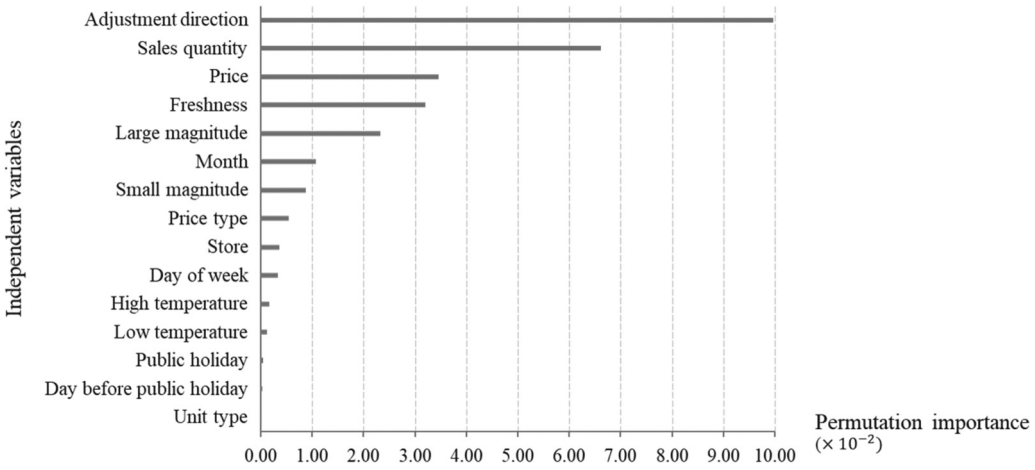

Fig. 7. Important determinants of the adjustments quality.

only 39.8% of the large positive adjustments are valuable, whereas 66.6% of the small positive adjustments improve forecasting accuracy. By comparing the different combinations, we find that (i) the frequency of valuable adjustments is greater for small and positive adjustments than for large and positive adjustments and (ii) the frequency of valuable adjustments is greater for large and negative adjustments is greater than for small and negative adjustments. We infer that valuable positive adjustments are mostly small in magnitude, while valuable negative adjustments are mostly large. The larger an adjustment is, the greater the change in the forecast accuracy. These findings justify the previous conclusion regarding the different accuracies of adjustments in the positive and negative directions. A few valuable large adjustments could not enhance the overall accuracy in the positive direction. In contrast, there are more valuable negative adjustments with large magnitude, yielding better accuracy in this direction.

Here, we have gained insights related to the adjustments' quality, particularly in terms of the magnitude and direction. Nonetheless, the drivers of quality and its predictors have yet to be identified.

6.1. Drivers of quality

Fig. 7 shows the determinants of the adjustments' value using the permutation importance. The adjustment direction is the most relevant variable linked to the quality of that adjustment. The number of products sold, price, product freshness, and large magnitude of adjustments are also important drivers. Returning to our prior discussion, we already observed more frequent positive adjustments, resulting in a considerable impact on quality. The high importance of the sales quantity points to the distinction between products with different sales frequencies. The adjustments for fast- and slow-moving products yield different quality. Additionally, the greater importance of large magnitude adjustments compared to small magnitude adjustments is aligned with our prior intuitions. We also checked the validity of this result, when focusing on a particular direction (either the positive or negative direction), and found the lists of important variables similar.

Comparing these results with those from phase 1, price and freshness remained among the important variables. This result can be interpreted in two ways. First, expensive and fresh products are adjusted more frequently than their corresponding categories, i.e., low- and medium-priced and non-fresh products. Thus, within each category of the variable, the quality of the adjustments may also vary. Second, each category of the price and freshness variables has a specific property that affects the adjustment quality. For instance, adjusting the forecasts of fresh products is challenging and may not necessarily add value. By contrast, we expect planners to perform better in adjusting non-fresh products. Our findings are viable for application in normal market conditions. An unexpected incident can confound the whole process.

In addition to identifying the drivers of the quality, the degree of their importance is notable. Based on Fig. 7, the adjustment direction is more important than the sales quantity, price is more important than freshness, and all of these are more important than large magnitude. Interestingly, in addition to being important for the decision to make an adjustment, price and freshness are important predictors of the quality of the adjustment. By contrast, although important for the decision to make an adjustment, price type and high temperature do not drive quality. This supports our interpretation that their importance for the adjustment decision is an artefact of working with traditional forecasting systems that often neglect these factors. Thus, planners may decide to adjust the related forecasts, but their adjustments are not consistently improving the system's recommendation.

Any variable less important than large magnitude has little influence; thus, we have focused on the first five variables at the top of the figure.

As in the first phase, we look into some of the important drivers to investigate their effect direction and assess whether the RF's outcome is in accordance with the basic intuition elicited directly from the data.

Price. In this phase, a forecast merely refers to an adjusted forecast Fig. 8. illustrates that adjustments for more expensive products, which constitute a lower share of all products, are more likely to be valuable. Hence, not only are more expensive products adjusted more often, but also those adjustments are more likely to provide value. This confirms the overall important role of price in the adjustment process.

Likewise, we show the percentage of forecasts and the percentage of valuable adjustments for important binary variables in Fig. 9.

Direction: Fig. 9-a shows that a total of 70.2% of all adjusted forecasts are in positive direction. A total of 53.1% of positive adjustments are valuable, while only 42.3% of negative adjustments add value. Hence, planners have a higher tendency to adjust upwards and have a higher likelihood of adding value with upward adjustments. However, note that we earlier also looked at the interaction between direction and magnitude (see Table 6) and found that while the upward adjustments add value, there seem to be more valuable small upward adjustments and more non-valuable

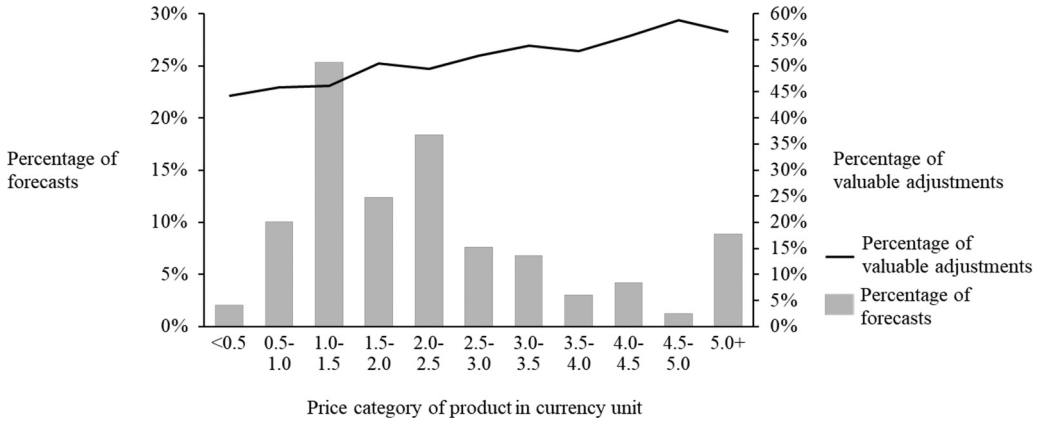

Fig. 8. Valuable adjustments and frequency per price category of product in currency unit.

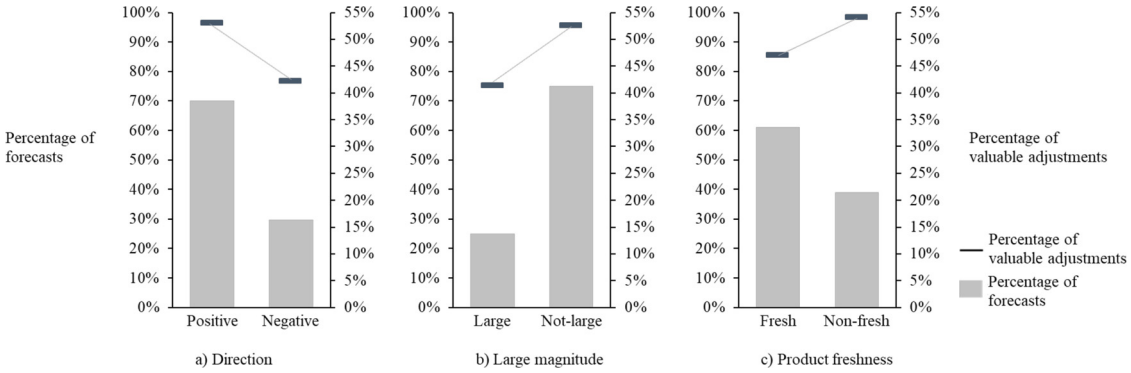

Fig. 9. Effect of the important binary variables on the adjustments quality.

large upward adjustments. Thus, on average, upward adjustments (as outlined in Section 6) reduce accuracy.

Large Magnitude: Fig. 9-b shows that large adjustments are the top 25% of adjustments in terms of magnitude, while correspondingly, not-large adjustments are the remaining 75%. The percentage of valuable adjustments among large magnitude adjustments is lower than that for not-large adjustments (41.4 vs. 52.7%).

Product Freshness: Fig. 9-c shows that a total of 61.0% of all adjustments relate to fresh products. However, only 47.1% of the adjustments for fresh products are valuable, compared to 54.2% of the adjustments for non-fresh products. Hence, planners adjust more for fresh products, but these adjustments are less likely to add value.

6.2. Predictors of quality

In this section, we identify the collective impact of different variables on quality using a DT. According to each variable's impact on the quality found in the previous step, we choose a set of important variables for a tree-based example of relationships, including the variables adjustment direction, sales quantity, price, product freshness, and large magnitude. For a better presentation, we categorize (i) the price into three categories, similar to the previous phase, and (ii) the sales quantity into two categories (fewer than three products sold and more than three products sold) to refer to slow- and fast-moving products, respectively.

Fig. 10 shows a tree-based scheme of the relationships between the "valuable" variable and its predictors. The DT starts with the large magnitude variable at the root node. The variables' interaction and the probability that the adjustment is valuable are in the middle nodes and at the terminal node, respectively. Every navigation route from the root to a terminal node shows a set of conditions under which an adjustment is (not) likely to add value. For instance, for large adjustments on the right side of the DT, adjustments of slow-moving products in the positive direction have the lowest odds of adding value, with a probability of 31.1% (Node 19). This node includes the highest number of large adjustments with ($n = 82{,}755$). However, large adjustments for slow-moving and fresh products in the negative direction have far higher odds of adding value, with a probability of 66.2% (Node 22). On the left side of the DT, not-large adjustments in the positive direction for slow-moving and non-fresh products with medium or high prices have the highest odds of adding value, with a probability of 69.4% (Node 16). Node 9 has the lowest probability of 30.9% (which is less than half of the value add compared to node 16). Node 9 and 16 only differ in the direction in which the adjustment is made. The adjustments in the positive direction provide notably more value than the ones in the negative direction (for not-large, slow-moving, non-fresh products with a medium or high price).

The DT provides insights in line with our findings in Section 6.1. For instance, we find that large negative and not-large positive adjustments are more likely to be valuable than their counterparts, also when considering multivariate interactions. Moreover, we are able to attain some novel insights to understand when an adjustment is more likely to add value when considering product characteristics such as product freshness, sales frequency, and price.

7. Discussion and conclusion

The development of AI/ML systems has shifted practitioners' attention towards intelligent automated demand forecasting. Due to real-world complexities, the human planner will, for the foreseeable future, continue to play an important role in AI-enabled fore-

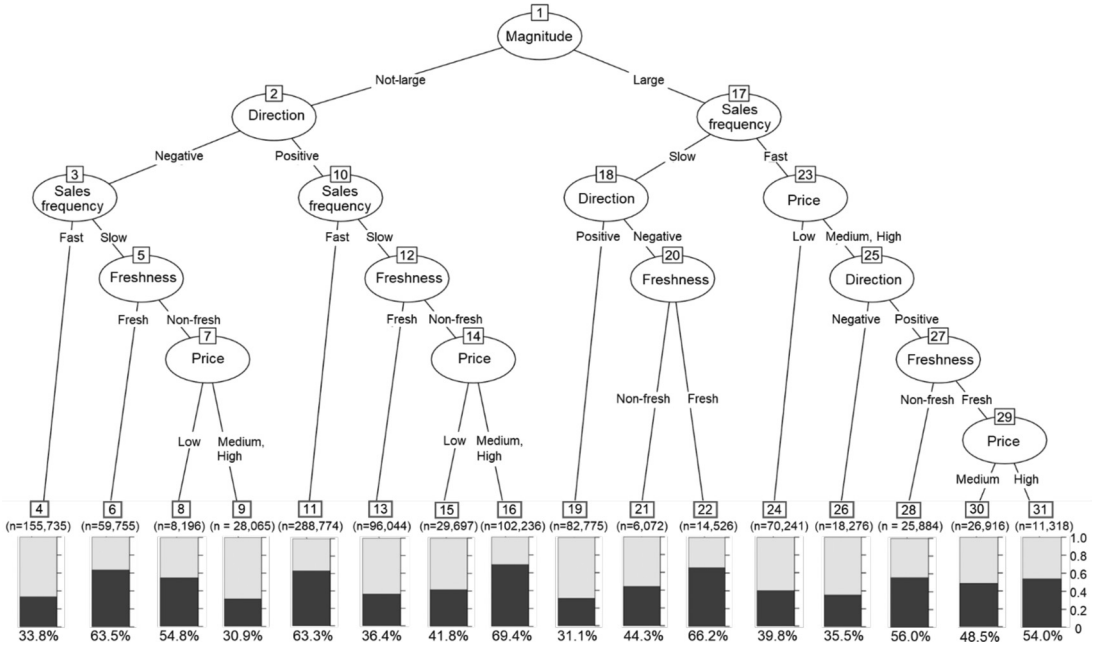

Fig. 10. Decision tree based on valuable variable with associated important predictors on the training set (the bars show the frequency of valuable adjustments in its associated subsample).

cast processes. The availability of high-quality input data is not always guaranteed, and certain information will not be available to the AI system in a structured form. For example, in grocery retailing settings with frequent product introductions, local events, and local competitive actions, fully autonomous forecasting processes may not yet be technologically possible and economically viable for a long time to come. However, any forecast adjustment requires time, and not all adjustments are valuable. In this paper, we have analyzed the human-machine interaction in the context of judgmental adjustments of forecasts, considering forecasts that were made by an AI system.

Our study focuses on two important issues with respect to planners' adjustments of AI forecasts: (i) when the adjustments are made by planners and (ii) when the adjustments are valuable and improve forecasting accuracy. For (i), we find that different product-related variables, such as price, freshness or price discounts, drive the frequency of adjustments. The DT analysis highlights that planners more frequently adjust forecasts under a combination of specific factors, for example, if a forecast is made for a fresh product at a higher price. The probability for an adjustment can be more than three times higher based on the collective effect of multiple important factors, i.e., yielding a 10.0 vs a 3.3% adjustment rate. For (ii), we find that product-related variables (e.g., sales quantity, price, or price type) as well as adjustment-related factors (e.g., adjustment direction, magnitude) are important in regard to the quality of the adjustments. These factors can be considered collectively to identify situations where it is likely that the planner adds value, e.g., for smaller, positive adjustments with higher sales frequency. Here, we find that the probability for an improvement can be more than double for a combination of factors, i.e., yielding a 69.4 vs. 30.9% improvement rate.

Another interesting aspect to consider as a potential driver of human adjustments is the importance and the demand volatility of a product. The company currently does not use an ABC or XYZ analysis on their parts. Therefore, we rather include continuous features such as sales quantity and price instead of an aggregate version. As a robustness check, however, we have classified the products into nine product classes based on the commonly applied ABC-XYZ methodology (Errasti et al., 2010; Nahmias & Olsen, 2015)

Table 5
The frequency of adjustments for different product classes.

	A (%)	B (%)	C (%)	Total (%)
X	8.5	5.0	2.7	6.1
Y	10.1	6.1	3.1	5.1
Z	8.7	7.0	3.3	4.0
Total	8.9	5.7	3.2	5.1

Table 6
The percentage of valuable adjustments for different product classes.

	A (%)	B (%)	C (%)	Total (%)
X	43.9%	44.6%	43.7%	44.1%
Y	46.5%	50.1%	52.4%	49.8%
Z	47.9%	52.6%	63.8%	59.9%
Total	45.0%	48.2%	58.4%	49.9%

Appendix C. provides more details about the approach we used for classifying the products as it is not used by the company.

Tables 5 and 6 show the effect of the different product classifications on the frequency and value of the adjustments, respectively. In line with expectations, "more important" (class A) products have a greater chance of adjustments than "less important" (class C) products (see Table 5). In addition, X products with more stable demand see more adjustments than Z products with more volatile demand. However, the planners do not necessarily add more value for the products for which they adjust more. As shown in Table 6, adjustments for C products add more value than adjustments for A products. Similarly, adjustments for Z products add more value than adjustments for X products. The largest difference can be seen for the subordinate combinations, i.e., AXE vs. CZ with a value-add of 43.9 vs. 63.8%. While the difference between the groups is less compared to the different nodes in the decision trees, the results are consistent. The planners seem to touch more forecasts for products that they feel are important and in fact damage the majority of these forecasts.

As an additional analysis, we check the robustness of our findings when considering the magnitude of the adjustments' value along with two main adjustments' characteristics: direction and

Table 7
Frequency of large/ small deterioration/ improvement per each direction-magnitude of adjustments.

		Large deterioration < -24.2	Small deterioration $(-24.2, 0.0)$	Small improvement $(0.0, 14.6)$	Large improvement > 14.6	Total
Positive adjustment	Small	4.8%	28.6%	**61.9**%	4.7%	100%
	Large	**52.2**%	8.0%	8.3%	31.6%	100%
Negative adjustment	Small	0.5%	**60.6**%	29.1%	9.8%	100%
	Large	28.7%	24.0%	4.6%	**42.7**%	100%

Note: The numbers on the top show the quartile of the value of adjustments.

magnitude. The magnitude of value of the adjustment is calculated by the percent improvement/ deterioration per realized demand. As nearly half of the adjustments improve and half damage the forecasts' accuracy, we classify the obtained variable into four categories: below the first quartile (large deterioration), between the first quartile and 0 (small deterioration[5]), between 0 and the third quartile (small improvement), and above the third quartile (large improvement).

According to the quartile-based classification for the improvement/ deterioration variable, each cell of Table 7 should be compared with 25%. Respectively, we see four prominent cells per each row and column. These again attest to the prior findings; such as the detrimental role of large positive adjustments and the beneficial role of large negative adjustments, while these roles are exchanged between the directions of adjustments for those of small magnitude.

Our research is distinct from prior studies in a variety of ways. Most previous research has studied forecasts made by traditional systems using traditional methods. To the best of our knowledge, this is the first study to analyze human adjustments on forecasts made by AI systems and to apply ML methods to do so. Thus, there are two novel elements compared to prior studies. First, the AI provider uses an AI system to generate different forecasts for different products. We find that product characteristics play an important role in the decision processes of planners in regard to their decision to adjust system forecasts. Planners treat products with a higher price or different freshness differently when spending time adjusting forecasts. They might feel that they should adjust the AI system more frequently for such products, but we find that those adjustment decisions are not always associated with a higher quality. In particular, variables such as weather or discounts that are already considered by the AI system are not strong indicators for the quality of the adjustment. As such, planners overcompensate for these effects in what could be considered "double accounting" (Dietvorst & Simonsohn, 2019): First, the AI systems consider the effect (e.g., of the weather or price discounts) in the system forecast, and then the planner considers the effect again when adjusting the forecast. AI-generated forecasts differ from the forecasts made by traditional models in terms of data volume, transparency and explainability, novelty, and accuracy, and these factors may affect planners' reliance on the system and make it more difficult to identify situations in which one can add value.

The second new element in our research relates to the use of ML in our analysis. Using ML methods, we can analyze a large amount of data efficiently and fairly quickly. ML methods are not limited by the assumptions commonly applied in traditional methods, such as linearity; hence, they can provide more accurate results. Nevertheless, ML results are not self-explanatory and require additional explanations for understanding and interpreting outputs. We have established our study on empirical data with various product-related and decision-related variables. We used RF to identify important variables and enhance this result by applying DT analysis to further interpret the collective effects of different variables of interest.

We conducted our analysis based on the collaborative setup of a leading AI provider and a large European retailer. Our results provide some managerial implications for both parties. The AI provider can consider this information to better train the AI system. Additionally, the retailer can benefit from leveraging this information and training the planners to improve the adjustments. They can also provide their planners with an informed ranking of potential decisions based on predictions of the adjustment and the adjustment quality (Imdahl et al., 2021). At the time of the analysis, planners could not learn how to make better adjustments due to the absence of a feedback process. In this respect, a feedback loop should be implemented between the AI provider and the retailer (De Baets & Harvey, 2020), something the AI provider has been also working on. Nevertheless, designing an efficient feedback process is challenging due to several complexities, e.g., the opaque process of generating AI forecasts due to the black-box nature of the system and the large volume of AI forecasts given the system's ability to generate millions of forecasts. Additionally, explainability plays an important role in making AI forecasts understandable. When planners understand how forecasts are generated, they should better judge whether including their contextual information will truly be valuable. While some systems have started to add explainability, it is essential to find a smart way to execute it for the millions of forecasts generated by an AI system. Based on the forecasts chosen to adjust, we can infer that planners are averse to some AI recommendations. While this is not a new topic, we conjecture that the explainability issue can reduce the algorithm aversion in our setting when planners judge AI recommendations.

This paper opens up multiple interesting avenues for further research in the context of judgmental forecasting with AI-generated forecasts. First, it is important to learn more about the differences in the planner responses to AI-generated models compared to traditional models. Our results suggest that humans may transfer their approach from dealing with traditional systems to dealing with AI systems, thereby neglecting the additional information that an AI model takes into account in the forecast generation process. This double accounting effect can harm forecasting performance when special events or abnormal weather are considered in both the forecast and the adjustment. Further research needs to investigate whether this is part of a change process (i.e., it only relates to planners who are used to traditional systems) or if it is a natural human tendency to try to add value based on some additional information. In addition, this effect could be moderated if the planner understands the factors considered by the AI system. Hence, the role of explainability and interpretability in overcoming double accounting needs to be further studied.

Second, additional research should investigate how to blend AI forecasting with human adjustments and consider product- and planner-specific aspects. Van der Staak et al. (2020) provide an overall framework for automating certain parts of the forecasting task, but we show that including product-related information

[5] The actual median is 0.1 but we restrict it to 0 to keep the meaning of improvement/deterioration.

can partially explain the adjustment frequency and quality. Hence, blending this information into a forecasting automation algorithm would be valuable. Imdahl et al. (2021) utilize product-related information for automating order decisions and control for individual planner behavior. As we have no information on the planners who make forecast adjustments, we cannot identify individual differences in their decision-making processes. It would certainly be interesting to also consider planners' performance differences and treat them as input to the AI forecasting process.

Third, the current study was conducted for a setting with normal market conditions. The outbreak of the COVID-19 pandemic and the effect it had on sales and demand planning (Alicke & Hoberg, 2020) raise the question of how fast AI forecasting can adapt to a change in the environment and how planners deal with the resulting volatility. There have been different dynamics in the adjustments and their quality during this unique period. Certainly, there is a learning period during which AI is not able to fully adapt to the new situation, and planners' adjustments could be valuable. However, at a certain stage, the AI system could become well adapted to the new demand reality, and planners' adjustments are likely to reduce quality. The drivers of adjustments and the quality of those made during this period might be very different from those found in our study. A new study targeting the pandemic period is likely to add interesting new insights into the interplay between humans and AI algorithms in times of change.

Fourth, it is also important to better understand potential feedback mechanisms that can improve planners' actions and stimulate their learning. As mentioned above, there is no feedback mechanism in our research setting, and the level of aggregation, the frequency and the type of feedback used to inform planners about the results of their actions is unclear. Given the huge number of forecasts and the time lags between an adjustment and the observation of demand, the feedback approach needs to be carefully designed. Simply sharing the latest outcome may damage performance due to the focus on the most recent observation and capturing the noises (Fildes et al., 2006; Hammond et al., 1973). While this can be studied in the field, lab experiments can offer an alternative research methodology to test different approaches before validating them in real-world settings.

Finally, the planner can be given "nudges" to achieve a more favourable adjustment behavior (Baker, 2021). For example, a planner can receive a prediction, whether (s)he is likely to add value with a certain type of adjustment. Likewise, a planner who adjusts too many items may receive the average amount of adjustments made among planners as a guideline. Studying which nudges and which feedback affects the behavior of the planner in a favourable way is an interesting future research question.

Acknowledgment

We would like to thank our industry partner team that has provided the data and spent a lot of time discussing our research. In addition, we would also like to thank the review team for their high-quality input and in particular their very valuable guidance towards relevant literature.

Appendix A

Comparison of random forest and logistic regression

Table A1 and Fig. A1 compare RF with logistic regression for different sample sizes. Here, "adjusted" serves as the dependent variable that indicates whether an adjustment is made. We find that the AUC for RF is higher than the AUC for logistic regression for the sample size. For RF, the AUC improves slightly for very large sample sizes, while the runtime grows much more rapidly. We find that the change in accuracy is negligible when the sample size increases from 5 million to 10 million forecasts, while the runtime is more than two times higher for the larger sample. Thus, we considered 5 million forecasts (80% train- 20% test) for our analysis.

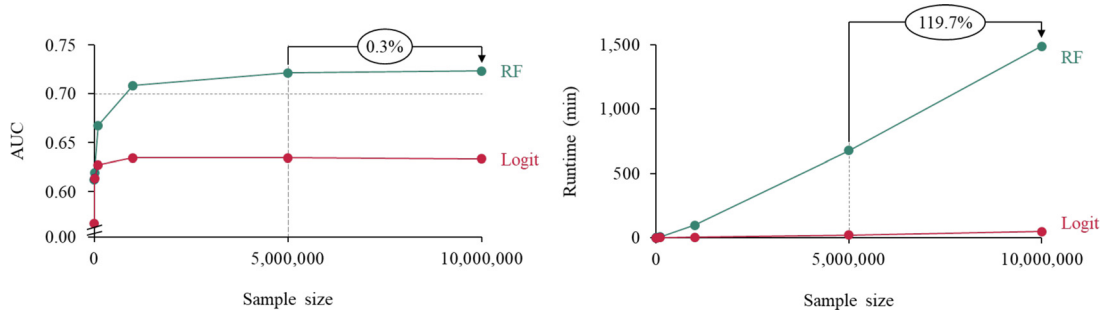

Fig. A1. RF and logistic regression comparison.

Table A1
Comparison of accuracy and runtime for RF and logistic regression with different sample sizes (within-sample robustness).

Sample size		5,000	10,000	100,000	1,000,000	5,000,000	10,000,000
RF	AUC	0.612	0.619	0.667	0.708	0.721	0.723
	Runtime (min)	0.194	0.412	6.949	97.445	676.520	1486.276
Logistic	AUC	0.567	0.613	0.627	0.634	0.634	0.633
Regression	Runtime (min)	0.071	0.087	0.403	4.416	22.073	47.810

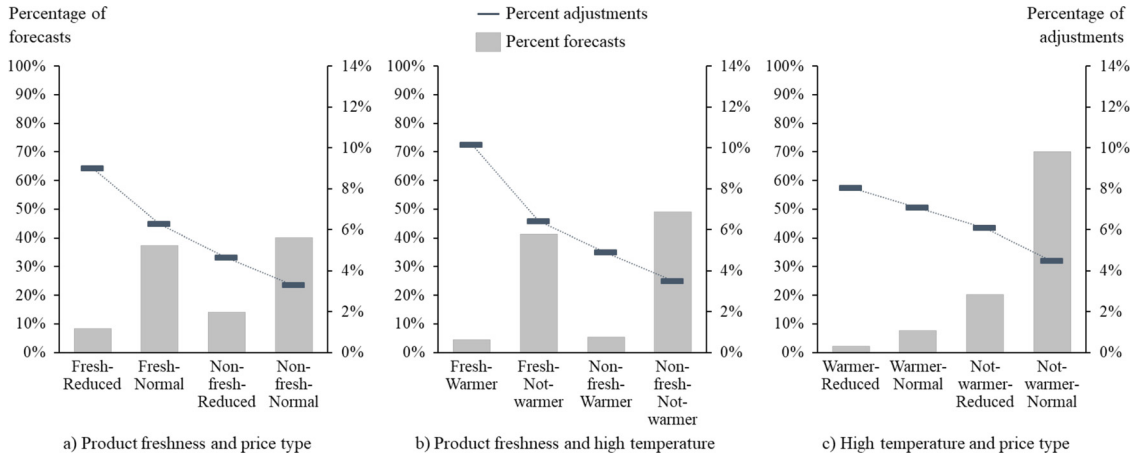

Fig. B1. Interaction between important binary drivers of the adjustments made.

Table B1
Interaction of price with three important adjustments' drivers (price type, high temperature, freshness).

	Price								
Price category	Low-price			Medium-price			High-price		
Adjustment%	4.33			5.48			5.92		
Frequency%	43.81			35.54			20.65		
Binary variables' categories	Reduced price	Warmer	Fresh	Reduced price	Warmer	Fresh	Reduced price	Warmer	Fresh
Adjustment%	6	5.8	5.6*	6.5	8.1	7.1	6.2	8.3	9.7
Frequency%	23	10	51**	21	10	50	24	10	27

* 5.6% of low-priced fresh products are adjusted.
** 51% of all products are low-priced and fresh.

Appendix B

Interaction charts for the drivers of adjustments

By considering the two-way interactions in Fig. B1 and Table B1, we find some insights consistent with those reported earlier. Individual effects exist, i.e., the freshness of a product, a discount, or an abnormally high temperature increases the adjustment probability. A greater discrepancy exists between the percentages of adjustments for different combinations when freshness is involved, e.g., fresh and reduced price, or fresh and warmer, which shows the higher importance of freshness than price type or high temperature, in line with the RF ranking in Section 5.1.

Appendix C

Procedure of ABC-XYZ analysis

We classified all the products in SKU-store level into nine product classes, known as ABC-XYZ classification. For the first part (ABC), we used the sales and price of SKUs per store, and subsequently multiplied them to yield the revenue of each SKU-store combination. After sorting the revenue, A products are 20% of products that account for the highest portion of the annual revenue, B products are 30% of products that account for the next portion of the annual revenue, and finally, C products are the remaining 50% of products that account for the remaining lowest revenue. For the second part of this analysis (XYZ), we classified the products based on the cut-offs of the coefficient of variation of the sales. We had access only to non-zero sales days; however, we were able to identify the days with no sales so that we considered them in the calculation of the standard deviation and mean of the sales. Dividing the standard deviation by the mean of the sales then yields the coefficient of variation. There are different ways of considering the cut-offs, we used tertiles (i.e., 33%–33%–33%) of the distribution of the coefficient of variation of the sales, as above: Tables C1 and C2

Table C1
Summary for coefficient of variation of the sales.

Min	1st Tert.	**Median**	**Mean**	2nd Tert.	Max
0.00	0.95	**1.86**	**2.04**	2.78	7.62

Table C2
The frequency of different product classes.

	A (%)	B (%)	C (%)	Total (%)
X	14.0	14.3	4.7	33.0
Y	4.6	11.6	17.2	33.4
Z	1.4	4.1	28.1	33.6
Total	20.0	30.0	50.0	100.0

References

Alicke, K., & Hoberg, K. (2020). Demand planning in a crisis. *Supply Chain Management Review, 24*(September/October), 42–47.
Alicke, K., Hoberg, K., & Rachor, J. (2019). The supply chain planner of the future. *Supply Chain Management Review, 23*(3), 40–47.
Badorf, F., & Hoberg, K. (2020). The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores. *Journal of Retailing and Consumer Services, 52*, Article 101921.
Baker, J. (2021). Maximizing forecast value added through machine learning and nudges. *Foresight: The International Journal of Applied Forecasting, 60*, 8–15.
Bolton, G. E., Ockenfels, A., & Thonemann, U. W. (2012). Managers and students as newsvendors. *Management Science, 58*(12), 2225–2233.
Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (2018). Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting, 35*(1), 170–180.
Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.
Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231.

Chen, K.Y., .& Li, S. (2019)., "The behavioral traps in making multiple, simultaneous, newsvendor decisions", available at SSRN 2817126. Available at SSRN: https://ssrn.com/abstract=2817126.
Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018). The operational value of social media information. *Production and Operations Management, 27*(10), 1749–1769.
Dane, E., & Pratt, M. G. (2007). Exploring intuition and its role in managerial decision making. *Academy of Management Review, 32*(1), 33–54.
Davydenko, A. (2012)., Integration of judgmental and statistical approaches for demand forecasting: Models and methods, Doctoral dissertation. Lancaster University.
Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting, 29*(3), 510–522.
De Baets, S., & Harvey, N. (2018). Forecasting from time series subject to sporadic perturbations: Effectiveness of different types of forecasting support. *International Journal of Forecasting, 34*(2), 163–180.
De Baets, S., & Harvey, N. (2020). Using judgment to select and adjust forecasts from statistical models. *European Journal of Operational Research, 284*(3), 882–895.
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.
Dietvorst, B. J., & Simonsohn, U. (2019). Intentionally "biased": People purposely use to-be-ignored information, but can be persuaded not to. *Journal of Experimental Psychology: General, 148*(7), 1228.
Elmaghraby, W., Jank, W., Zhang, S., & Karaesmen, I. Z. (2015). Sales force behavior, pricing information, and pricing decisions. *Manufacturing & Service Operations Management, 17*(4), 495–510.
Eroglu, C., & Croxton, K. L. (2010). Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting, 26*(1), 116–133.
Errasti, A., Chackelson, C., & Poler, R. (2010). An expert system for inventory replenishment optimization. In *Peoceedings of the international conference on information technology for balanced automation systems* (pp. 129–136). Springer.
Feng, Q., & Shanthikumar, J. G. (2018). How research in production and operations management may evolve in the era of big data. *Production and Operations Management, 27*(9), 1670–1684.
Feng, X., & Gao, J. (2020). Is optimal recommendation the best? A laboratory investigation under the newsvendor problem. *Decision Support Systems, 131*, Article 113251.
Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management, 18*(1), 69–88.
Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *INFORMS Journal on Applied Analytics, 37*(6), 570–576.
Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems, 42*(1), 351–361.
Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting, 25*(1), 3–23.
Fildes, R., Goodwin, P., & Önkal, D. (2019). Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting, 35*(1), 144–156.
Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*.
Fildes, R., & Petropoulos, F. (2015). Improving forecast quality in practice. *Foresight: The International Journal of Applied Forecasting, 36*, 5–12.
Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International journal of forecasting, 15*(3), 227–246.
Foster, J., Deck, C., & Farmer, A. (2019). Behavioral demand effects when buyers anticipate inventory shortages. *European Journal of Operational Research, 276*(1), 217–234.
Franses, P. H., & Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting, 25*(1), 35–47.
Franses, P. H., & Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting, 29*(3), 331–340.
Hammond, K. R., Summers, D. A., & Deane, D. H. (1973). Negative effects of outcome-feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance, 9*(1), 30–34.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.
Ibrahim, R., Kim, S. H., & Tong, J. (2021). Eliciting human judgment for prediction algorithms. *Management Science, 67*(4), 2314–2325.
Imdahl, C., Hoberg, K., & Schmidt, W. (2021)., "Targeted automation of order decisions using machine learning", Working paper, Available at SSRN: https://ssrn.com/abstract=3822131.
Käki, A., Kemppainen, K., & Liesiö, J. (2019). What to do when decision-makers deviate from model recommendations? Empirical evidence from hydropower industry. *European Journal of Operational Research, 278*(3), 869–882.
Kawaguchi, K. (2021). When will workers follow an algorithm? A field experiment with a retail business. *Management Science, 67*(3), 1670–1695.
Kesavan, S., & Kushwaha, T. (2020). Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science, 66*(11), 5182–5190.
Kirshner, S. N., & Shao, L. (2019). The overconfident and optimistic price-setting newsvendor. *European Journal of Operational Research, 277*(1), 166–173.
Kremer, M., Moritz, B., & Siemsen, E. (2011). Demand forecasting behavior: System neglect and change detection. *Management Science, 57*(10), 1827–1843.
Kremer, M., Siemsen, E., & Thomas, D. J. (2016). The sum and its parts: Judgmental hierarchical forecasting. *Management Science, 62*(9), 2745–2764.
Lantz, B. (2013). *Machine learning with R*. Packt Publishing Ltd.
Lawrence, M., Goodwin, P., O'Connor, M., & Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting, 22*(3), 493–518.
Lehmann, C. A., Haubitz, C., Fügener, A., & Thonemann, U. W. (2020). Keep it mystic? The effects of algorithm transparency on the use of advice. In *Proceedings of the ICIS*.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2021). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2021.07.007.
Mathews, B. P., & Diamantopoulos, A. (1989a). Factors affecting subjective revision in forecasting A multi-period analysis. *International Journal of Research in Marketing, 6*(4), 283–297.
Mathews, B. P., & Diamantopoulos, A. (1989b). Judgemental revision of sales forecasts: A longitudinal eetension. *Journal of Forecasting, 8*(2), 129–140.
Meinard, Y., & Tsoukiàs, A. (2019). On the rationality of decision aiding processes. *European Journal of Operational Research, 273*(3), 1074–1084.
Moritz, B., Siemsen, E., & Kremer, M. (2014). Judgmental forecasting: Cognitive reflection and decision speed. *Production and Operations Management, 23*(7), 1146–1160.
Nahmias, S., & Olsen, T. L. (2015). *Production and operations analysis*. Waveland Press.
Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC bioinformatics, 11*(1), 1–13.
Oliva, R., & Watson, N. (2009). Managing functional biases in organizational forecasts: A case study of consensus forecasting in supply chain planning. *Production and operations Management, 18*(2), 138–151.
Perera, H. N., Hurley, J., Fahimnia, B., & Reisi, M. (2019). The human factor in supply chain forecasting: A systematic review. *European Journal of Operational Research, 274*(2), 574–600.
Petropoulos, F., Fildes, R., & Goodwin, P. (2016). Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research, 249*(3), 842–852.
Ren, Y., & Croson, R. (2013). Overconfidence in newsvendor orders: An experimental study. *Management Science, 59*(11), 2502–2517.
Robin, G., Jean-Michel, P., & Christine, T. (2017). Random forests for big data. *Big Data Research, 9*, 28–46.
Sachs, A. L. (2015). Empirical newsvendor decisions under a service level contract. *Retail analytics* (pp. 79–101). Springer Cham.
Sanders, N., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega, 31*(6), 511–522.
Sevdalis, N., & Harvey, N. (2007). Biased forecasting of postdecisional affect. *Psychological Science, 18*(8), 678–681.
Shmueli, G., & Yahav, I. (2018). The forest or the trees? Tackling simpson's paradox with classification trees. *Production and Operations Management, 27*(4), 696–716.
Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review, 61*(4), 66–83.
Stringer, S. (2018)., "Feature importance: What's in a name?", available at: Https://medium.com/bigdatarepublic/feature-importance-whats-in-a-name-79532e59eea3 (accessed 23 July 2018).
Syntetos, A. A., Kholidasari, I., & Naim, M. M. (2016). The effects of integrating management judgement into OUT levels: In or out of context? *European Journal of Operational Research, 249*(3), 853–863.
Tong, J., & Feiler, D. (2017). A behavioral model of forecasting: Naive statistics on mental samples. *Management Science, 63*(11), 3609–3627.
Trapero, J. R., Pedregal, D. J., Fildes, R., & Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting, 29*(2), 234–243.
Van der Staak, B., Basten, R.J.I., Van de Calseyde, P.P.F.M., Demerouti, E., & De Kok, A.G. (.2020). Some-touch forecasting: A novel method to combine human judgment with statistical algorithms. Working paper. Submitted for publication. Not available online.
Van Donselaar, K. H., Gaur, V., Van Woensel, T., Broekmeulen, R. A., & Fransoo, J. C. (2010). Ordering behavior in retail stores and implications for automated replenishment. *Management Science, 56*(5), 766–784.
Van Woensel, T., Van Donselaar, K., Broekmeulen, R., & Fransoo, J. (2007). Consumer responses to shelf out-of-stocks of perishable products. *International Journal of Physical Distribution & Logistics Management, 37*(9), 704–718.
Weller, M., & Crone, S.F. (.2012)., "Supply chain forecasting: Best practices & benchmarking study". Lancaster Centre of Forecasting.