Dec 11th, 12:00 AM

Algorithm-Human-Algorithm: A New Classification Approach to Integrating Judgemental Adjustments

Christopher Chen
Indiana University, cch3@iu.edu

Nitish Jain
London Business School, njain@london.edu

Varun Karamshetty
National University of Singapore, varun.karamshetty@nus.edu.sg

Algorithm-Human-Algorithm: A New Classification Approach to Integrating Judgemental Adjustments

Completed Research Paper

Christopher J. Chen
Indiana University Bloomington
107 S Indiana Ave, Bloomington, IN 47405, United States
cch3@iu.edu

Nitish Jain
London Business School
Regent's Park, London, NW1 4SA, United Kingdom
njain@london.edu

Varun Karamshetty
National University of Singapore
11 Research Link, 119391,Singapore
varun.karamshetty@nus.edu.sg

Abstract

Modern-day firms face the predicament of blending the comparative advantages of their two core resources: machines and humans. When forecasting demand (e.g. for a product), extant literature documents that always permitting (or prohibiting) human revision of a machine forecast is beneficial if the humans' private information role is larger (or smaller) than that of machine-accessible public information. We propose and test a complementary framework that shifts the focus to the regulation of each human revision; and in doing so, adjusts for human vulnerability to systematic biases. We collaborate with a European retailer to compile a large dataset (∼ 1.1 mn transactions) on machine-led demand forecasts. Humans revise nearly 38% of these forecasts, but revisions do not always yield an improved final forecast. Compared to the always permit or prohibit strategies, the ideal regulation of each of these revisions could reduce the absolute deviation of the final forecast from realized sales in close to 50% of instances. Any regulation to obtain the best of the machine and human-revised forecasts requires an ability to predict revision quality. In addition to private information indicators, we show that indicators of systematic bias in human judgement (e.g., indicators of cognitive load) can improve the prediction of revision quality. In an out-of-sample analysis, our revision-level regulation approach picks the best of available forecasts in 14% more instances, compared to an always-permit or prohibit strategy at the product-store level. Our paper provides practitioners with a novel approach to combining the machine-human output to improve demand forecasting performance.

Keywords: *Human-Algorithm Interface, Human discretion, Demand Forecasting, Systematic Biases*

Introduction and Literature Review

Firms worldwide are increasingly looking for ways to strengthen their traditional human-led forecasting operations using data-led machine output. Often, these two resources offer comparative advantages. On the one hand, machines (henceforth interchangeably referred to as algorithm) offer consistent application of data-led learning, but the availability of public information (codifiable features) limits this advantage. On the other hand, humans may possess pertinent private information to improve machine outputs (Kesavan & Kushwaha, 2020; van Donselaar et al., 2010), but they may apply it in an inconsistent and biased way (Caro

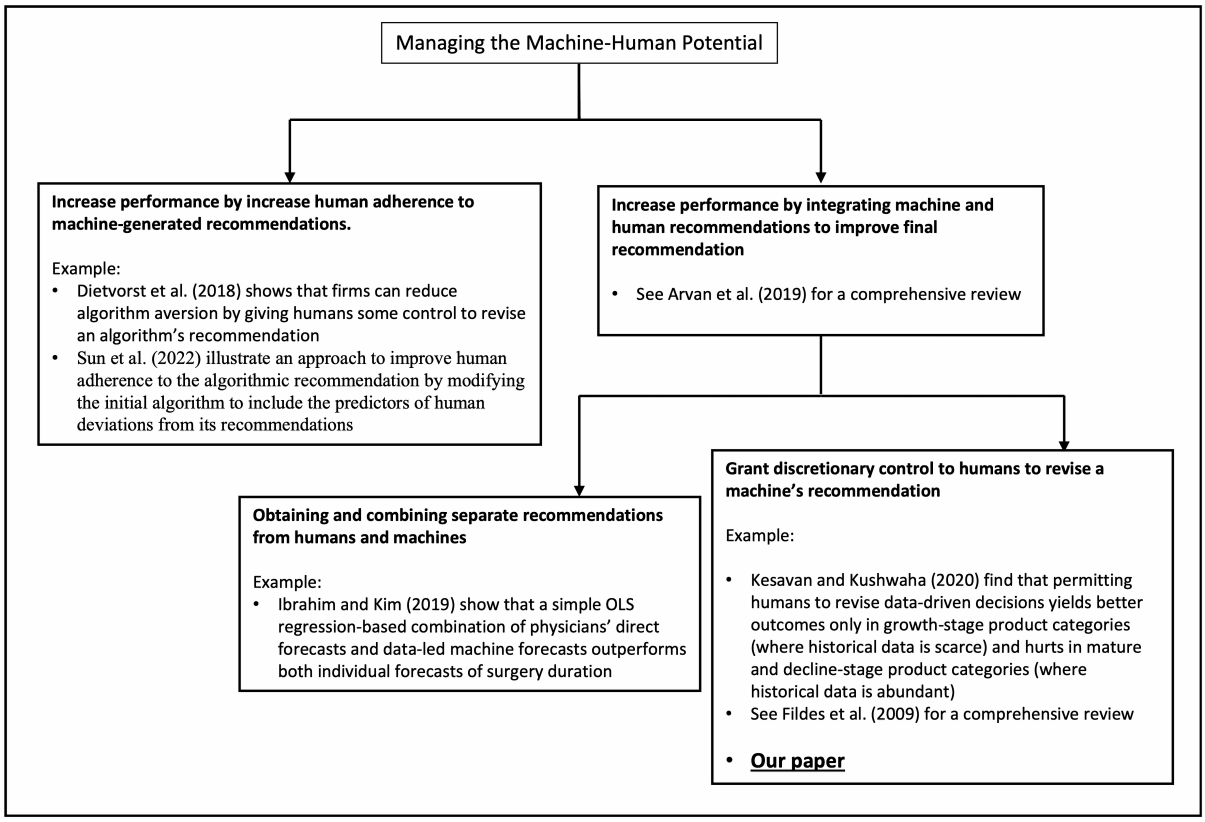


Figure 1. Landscape of literature on managing machine-human collaboration.

& de Tejada Cuenca, 2023; Ibrahim et al., 2021). The complementary pros and cons of these two resources create a predicament for modern-day firms: how to leverage the comparative advantages of machines and humans?

Extant literature on managing the machine-human potential spans multiple fields including Operations Management, Information Systems, and Strategy. The approaches proposed in the literature can be split into two primary streams. The first stream favors the machine-generated output and, thus, searches for ways to increase human adherence to it. For example, Dietvorst et al. (2018) shows that firms can reduce algorithm aversion by giving humans some control to revise an algorithm's recommendation. In another example, Sun et al. (2022) illustrate an approach to improve human adherence to the algorithmic recommendation by modifying the initial algorithm to include the predictors of human deviations from its recommendations. The second stream explores ways to integrate machine and human recommendations to provide an improved final recommendation (see Arvan et al. (2019) for a comprehensive review). Our paper contributes to this second stream of work.

Within the literature stream on the integration of machine and human recommendations, there are two popular approaches. The first focuses on obtaining and combining separate point forecasts from humans and machines. For example, Ibrahim and Kim (2019) show that a simple OLS (Ordinary Least Squares) regression-based combination of physicians' direct forecasts (DF) and data-led machine forecasts outperforms both individual forecasts of surgery duration. In more recent work, Ibrahim et al. (2021) discuss the disadvantage of eliciting and using human DFs. Specifically, the authors argue that since humans are prone to inconsistent application of their private information, a better strategy is to elicit private information adjustment (PIA) — "how much the human thinks the algorithm should adjust its forecast to account for the

information that only the human has." That said, in practice, eliciting PIA is a challenging task.[1]

The second approach grants humans discretionary control to revise an algorithm's point forecast (e.g., see Fildes et al. (2009)). Under this approach, the challenge is determining how to regulate human discretion (i.e., when to permit or prohibit). Using a large field experiment, Kesavan and Kushwaha (2020) find that the strategy of *always* permitting humans to revise the data-driven decisions yields better outcomes only in growth-stage product categories (where historical data is scarce) and hurts in mature and decline-stage product categories (where historical data is abundant). In other words, the strategy of always-permitting human revisions was beneficial when the share of human private information seemed more prominent than that of machine-accessible public information. On the flip side, it is better to adopt the always-prohibiting strategy for human revisions when the share of private information seems lower than that of public information.

We contribute to the literature on human discretionary revision of machine-generated forecasts by proposing a novel framework to regulate human discretion. These discretionary revisions are also known as *judgemental adjustments* in the demand forecasting literature (Arvan et al., 2019). This stream provides extensive evidence to suggest that humans may be inconsistent or biased in their application of private information (Goodwin, 2000; Ibrahim et al., 2021; Lee et al., 2007). For example, Fildes et al. (2009) document a bias toward optimism when applying upward revisions to the algorithm's forecasts. Past studies have focused on attenuating these biases by suggesting various interventions, including additional system support (Lee et al., 2007) and requiring humans to explicitly request and record the reason for revision (Goodwin, 2000) (for a recent review on such interventions see Section 12.5.4 of Chapter 'Forecast Decisions' in the Handbook of Behavioral Operations, Donohue et al. (2018)). In addition, previous studies have shown that often human biases are systematic and, thus, predictable (Kremer et al., 2011; Tversky & Kahneman, 1974). Prior work has found that these biases are driven by factors such as cognitive limitations (Moritz et al., 2022), recent outcomes (Schweitzer & Cachon, 2000), and the organization's salient culture (Caro & de Tejada Cuenca, 2023).

The presence of predictable biases in human revisions (Caro & de Tejada Cuenca, 2023) motivates the first principle of our framework. We propose utilizing the predictability in human bias to regulate revisions and, as a result, improve the forecasting performance. This approach is largely orthogonal to the extant intervention-based approach to de-bias judgmental adjustments for improving the final forecast performance. Until a firm discovers interventions to eliminate all possible biases, it can leverage our approach to utilize the predictability in the (post-intervention) remaining biases to regulate revisions.

The second principle of our framework shifts the extant focus from an always-permitting approach, where each revision is allowed to pass without any regulation, to a revision-level regulation approach – regulate each revision instance, by its merit, in computing the final forecast. This approach is motivated by the possibility of humans producing biased revisions which, in turn, can lead to a poorer forecast in categories for which the always-permitting strategy is adopted.

Building on these principles, we propose and test a framework that predicts the quality of a human revision and subsequently uses the quality information to regulate the revision's role in computing the final forecast. We test this framework in collaboration with a European food retailer. We obtain a large dataset (~1.1 mn) with 24 weeks of data-driven machine forecasts and associated human revisions at the product-store-week level. Human forecasters revise close to 35% of the machine-generated forecasts. Among the revised forecasts, 51% of instances benefited from the human revision – with a lower absolute deviation from realized sales compared to the machine-generated forecast. In other words, in these revised forecast instances, an oracle-like manager would make the best use of the two available forecasts by selecting the human-revised forecast in 51% of the revisions and the machine-generated forecast in the remaining instances. This indicates a substantial potential for improving final forecast performance by utilising human revisions when they are beneficial and discarding them otherwise.

[1]Ibrahim et al. (2021) raise the following limitation in identifying PIA: "In our experiments, private information is easy to identify because the researcher knows all the data that exist in the environment and what data the human can access that the algorithm cannot. In practice, such an exercise is more difficult because information kept private from the algorithm may also be kept from the system designer. In other words, you cannot ask for what you do not know exists."

To construct an ex-ante measure of a human revision's quality, we build a classification model using machine learning algorithms. In our context, we quantify the incremental contribution of the bias-related (BR) predictors when compared against a baseline model that includes only the private information predictors (PI). At the product-category level, we find that the PI+BR model significantly improves the average AUC[2] of the PI-only model by 6.8% and up to 16% in select categories.

The predicted quality information can be used to implement multiple strategies for combining machine and human-revised forecasts. In an out-of-sample analysis, we study the performance of a simple threshold-based strategy that picks a human-revised forecast over the machine-generated forecast when the predicted quality is above an ex-ante specified threshold. Compared to an always permitting or prohibiting strategy at the product-store level, our threshold-based strategy yields a 14% improvement in picking the best of the two available forecasts. This improvement translates to a 7.2% reduction in the absolute deviation between the final forecast and sales when compared to oracle-like deviation. Finally, we find that the BR predictors make a significant incremental contribution to achieving these improvements. Specifically, it increases the PI-only model performance by 6% and 2%, respectively, in picking the best forecast and reducing the absolute deviation.

Our findings imply that the firms can leverage indicators of systematic bias in human judgement to improve their forecasting performance. To the best of our knowledge, ours is the first study to identify and test the usefulness of a revision-level regulation framework motivated by the rich theory on humans' vulnerability to systematic bias.

Study Context and Data

We collaborate with a European food retailer to compile a detailed dataset at the product $(p) \times$ store $(s) \times$ week (w) level, covering 665 products across 110 stores during a 24-week period from June 30th, 2019, to December 8th, 2019. The dataset includes information on algorithmic forecasts (AFs), judgmental adjustments, sales (in units sold), product categories within a four-tier hierarchy[3], in-store shelf life, and store characteristics.

Our engagement with this retailer was initiated by their supply chain coordinator and head of data analytics and motivated by a critical operational issue: the effective blending of human expertise with data-driven forecasts to reduce the uncertainty of demand, thereby mitigating food waste. This challenge not only influenced our research question but also practically shaped the design of our study, leading to the creation of our framework for modulating human input in forecast adjustments. While the problem stems from this specific retailer, the implications are far-reaching, impacting a range of stakeholders from consumers to the environment, and aligning with broader goals in corporate social responsibility and sustainability.

Demand Forecasting Process

The retailer's in-house demand forecasting team implements a two-step process to generate demand forecast f. As an illustration, consider the process of forecasting demand of product p at store s in week w, denoted f_{psw}.

Algorithm-generated forecast, AF, f^a_{psw}. In the first step, the forecasting team produces AFs using a data-driven model that processes available codified features like product and store attributes, seasonality, weather, and historical sales.[4] The team initiates forecasting for a week w in the week $w-3$ i.e., three weeks in advance. The forecasting model can update f^a_{psw} at the daily frequency to reflect any learning of recent demand trends.

Human-adjusted forecast, HF, f^h_{psw}. In the second step, the forecasting team shares f^a_{psw} with a human

[2]Area Under the Curve (AUC) is a standard metric to evaluate the performance of classification models, for details see Fawcett (2004).
[3]At the top-most level (level-2), the 665 products in our sample are grouped into three categories, and at the bottom-most level they are grouped into 85 categories
[4]Forecasts are mapped from the week- to day-level using a set of day-of-week weights that sum up to 1. These weights can vary by product and store but are time-invariant within a six-month selling period (Jan-June and July-Dec).

forecaster for review. Each forecaster is responsible for a cluster of products and stores. The team affords a forecaster complete flexibility in applying her/his judgment to AF. S/he can adjust the AF as soon as it becomes available (i.e., three weeks in advance) and revise the adjustment until the last day of the week w. Further, s/he can apply the adjustment at an aggregate level, such as the product p's category- or store-level which gets proportionally applied at the psw level, or s/he can directly adjust the f^a_{psw}. In practice, we observe that forecasters typically start applying adjustments two weeks in advance and adjust AF at the product-category-store level for the week w. Forecasters revise their adjustment to f^a_{psw} during the week w in fewer than 2% of instances. We denote human-adjusted AF by f^h_{psw}.

Final Demand forecast, f_{psw}. During our study period, the forecasting team followed the policy of setting the final forecast f_{psw} to f^h_{psw} whenever an adjustment is applied and to f^a_{psw} otherwise. In the words of the forecasting team's head, "[the retailer] wants to take advantage of our human forecasters' agility in responding to evolving market trends and unanticipated events."

Multiple teams use forecast f to make decisions ranging from placing procurement-order with suppliers to planning fulfillment of stock to retail stores from local distribution centers. These teams face different lead time constraints for their decisions and, thus, work with the most updated f_{psw} value that is available as per their timeline. The performance of the forecasting team, however, is measured on the accuracy of produced forecasts against the realized sales S and not on the performance of operational decisions they feed in to.

Data Sample

For a forecasting instance at the product-store-week level, the data may include multiple AFs and HFs. We construct our analysis sample using the forecasts available on the preceding Saturday of forecasting week w. For example, for the week commencing 8th July, 2019, only the forecasts available on Saturday 6th July, 2019 are included in the sample. This choice has two advantages. First, it ensures that AFs in our sample reflect the latest information available to the algorithm. Likewise, judgemental adjustments capture humans' up-to-date assessment of AFs. Second, it provides consistency in the store-level analysis since many stores in our data are closed on Sundays.

Our data sample contains a few instances of low-quantity machine forecasts, where the extent of adjustment appears unnaturally large due to scaling effects. For example, the following scaled measure of adjustment size relative to AF, $ScaledAbsRevSize = |f^h_{psw} - f^a_{psw}|/f^a_{psw} \times 100$ can be as high as 422. We exclude such outlier observations from the analysis sample by truncating it at $ScaledAbsRevSize$'s 99^{th} percentile values. Further, we also exclude close to 2% instances where a forecaster revised her/his adjustment to f^a_{psw} during the week w. Our final sample consists of 1,071,729 psw observations.

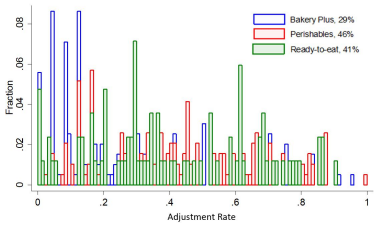
Table 1 reports the sample summary statistics. We report scaled values of sales S and f^a quantities to comply with our collaborator's non-disclosure directives. Specifically, we show standardized sales ($\widehat{S}$) and forecasts (f^a) scaled by S. Sales vary considerably across products, stores, and time (the mid-90% sales range is -0.64 to 1.60). The algorithm, on average, provides an inflated sales forecast, Avg $\widehat{f}^m_{psw}$ = 1.23. Human forecasters make, on average, an adjustment of 7.64% to AF. In terms of forecast accuracy, AFs and HFs are comparable. The scale-adjusted Mean Absolute Deviation (MAD) of AF and HF is 0.36 and 0.34 respectively. For at least half of the product-store combinations, we have the full 24 weeks of data.

In Figure 2 of the Appendix, we show characteristics of the human forecasters' adjustment patterns. The adjustment frequency varies considerably across the three (level-2) product categories (Panel(a)): most frequent interventions occur in the Perishables category, with adjustments applied to 46% of the AFs. However, the frequency of within-category adjustments is comparable with each category having products that receive frequent (intervention rate $\geq$ 80%), moderate (20% to 80%) or rare ($<$20%) adjustments. Humans typically adjust algorithm forecasts by 5%, 10% or 15% (panel (b)) and are much more likely to adjust it upwards (68% times, panel (c)) than downwards.[5]

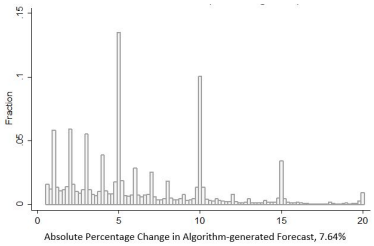
Next, panels (d) to (i) show evidence of human forecasters' ability to provide beneficial adjustments. We

[5]The retailer shared that within the organization, the salient culture is to be more tolerant of surplus stock in stores rather than presenting empty shelves to the customers. In their view, this could explain the observed predilection of forecasters' for upward adjustments.

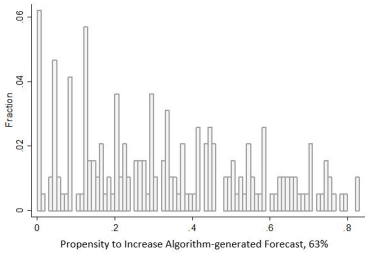
Human Forecasters: Adjustment Pattern



(a) Product-Level Adjustment Rate, by Product Categories



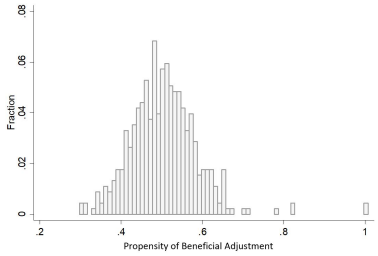
(b) Adjustment Size (Product x Week x Store)



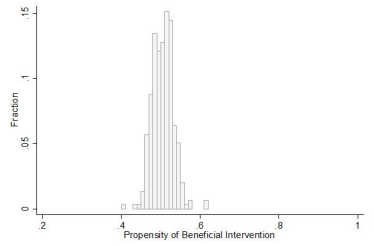
(c) Adjustment Direction (Product-level)

Human Forecasters: Adjustment Quality (51%)

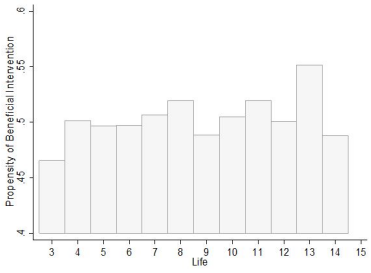
Predictors of Private Information



(d) Adjustment Quality by Product, σ/μ = 0.16

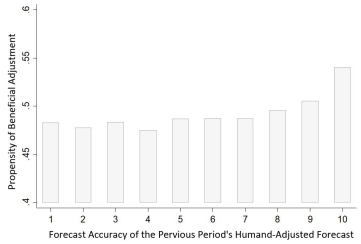


(e) Adjustment Quality by Store, σ/μ= 0.05

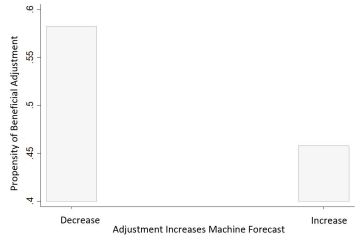


(f) Adjustment Quality by Product's Shelf Life

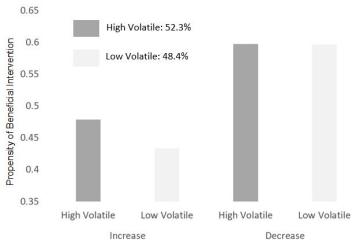
Predictors of Systematic Decision Biases



(g) Adjustment Quality by Recent Performance



(h) Adjustment Quality by Adjustment's Direction



(i) Adjustment Quality by Volatile Forecasting Period

Figure 2. Human Forecasters' Intervention Pattern and Performance

Statistic, Unit	N[1]	Obs	Mean	St. Dev.	5^{th}	50^{th}	95^{th}
Sales[2] $\widehat{S}$, ratio	psw	1,071,729	0	1	-0.64	-0.31	1.60
Algorithm-generated forecast [2] $\widehat{f}^{a}_{psw}$, ratio	psw	1,049,415	1.23	1.09	0.57	1.04	2.38
$ScaledAbsRevSize > 0$, %	psw	409,261	7.64	7.11	1.00	5.07	20.04
Algorithm-generated forecast $\widehat{\text{MAD}}^{3}_{\text{AF}}$, #	ps	48,654	0.36	0.59	0.14	0.29	0.73
Human-adjusted forecast $\widehat{\text{MAD}}^{3}_{\text{HF}}$, #	ps	46,410	0.34	0.37	0.11	0.28	0.70
Count of weeks per product-store, #	ps	48,654	22.03	4.42	10	24	24

[1] The retailer selectively ranges products across stores and over time to match customer preferences, resulting in a sample with 1.1mn observations ($\ll 665 \times 110 \times 24$)

[2] Standardized Sales variable, $\widehat{S} = (S - Avg(S))/Std.Dev(S)$ and Scaled algorithm-generated forecast, $\widehat{f}^{m} = f^{m}/S$. Number of observations $<$ 1.1 mn because our sample contains product-store-weeks with zero sales.

[3] Mean Absolute Deviation (MAD) is computed at the product-store level and scaled by its mean sales.

Table 1. Summary statistics

define an adjustment as beneficial ($binv = 1$) if $\mathbb{I}(|S - f^{h}_{psw}| \leq |S - f^{a}_{psw}|)$, where $\mathbb{I}$ is the *Indicator* function). Across the 409k adjustments, human judgment improved forecast accuracy 51% times. Further, forecasters' ability to make beneficial adjustments is much more heterogenous at the product level than at the store level. The Coefficient-of-Variation (σ/μ, CV) of the beneficial adjustment rate is 0.16 at the product level (panel (d)) and only 0.05 at the store level (panel (e)). The beneficial adjustment rate does not vary much by a product's shelf-life (panel (f), between 45% to 55%).

Finally, similar to past studies, we find preliminary evidence suggesting that a forecaster's ability to make a beneficial adjustment (panel (g)) indicates that the beneficial adjustment rate is weakly correlated with the past-period HF error, and panel (h) shows it is substantially higher (by 14%) when forecasters make downward adjustments than upward adjustments. Interestingly, similar to Kremer et al. (2011), we find forecasters exhibit a much better ability to adjust beneficially in high-volatility instances than in low-volatile ones, and much of this difference is observed while making upward adjustments (panel(i)).

Classification-based Solution Approach to Integrate Human Adjustments

We propose a two-module framework to classify whether a human forecaster's adjustment to AF should be accepted or rejected for the final forecast.

Module 1: Prediction Model for Adjustment Quality.

We build a model to predict the likelihood that an applied adjustment is beneficial (i.e., $binv = 1$). We then use the model output as a measure of an adjustment's predicted quality in improving forecast accuracy. To build such a prediction model, we use the following two classes of predictors:

1. Private Information, PI: Past studies have documented that the ability of humans to improve data-driven decisions is tied to their private information advantage (equivalently, local knowledge) in dimensions such as those related to customer behavior (impact of variety on customer demand (van Donselaar et al., 2010)) or product's PLC (growth-stage versus mature (Kesavan & Kushwaha, 2020)). Motivated by these studies, we include product and store characteristics in our model to proxy for humans' private information advantage.
2. Bias-Related, BR: Humans are vulnerable to producing biased judgment when operating in certain environments - for example, when forecasting under stable environments (Kremer et al., 2011), or when swayed by the organization's salient focus (e.g., inventory rather than revenue (Caro & de Te-

jada Cuenca, 2023)), or past errors (Fildes et al., 2009). To capture the potential of systematic biases (Tversky & Kahneman, 1974), we include several metrics, including those related to the forecasting week and/or past adjustments and of past forecast accuracy.

We describe the PI and BR predictors used in our prediction model in the Appendix.

Module 2: Classification of an Adjustment – Accept or Reject Integration in the Final Forecast.

We use a threshold-based heuristic to incorporate an adjustment's predicted quality (obtained from Module 1) in classifying its integration into the final forecast f_{psw}. Specifically, we set the final forecast to the human-adjusted forecast if the adjustment's predicted quality, conditional on the included predictors, is above a pre-specified threshold and to the algorithm-generated forecast otherwise. Formally,

$$f_{psw} = \mathbb{I}_{\widehat{binv}_{psw} \geq \tau_{psw}} f^h_{psw} + \mathbb{I}_{\widehat{binv}_{psw} < \tau_{psw}} f^a_{psw}, \tag{1}$$

where $\widehat{binv}_{psw}$ denotes the predicted quality of the human forecaster's adjustment to AF for forecasting-instance psw, and $\tau_{psw} \in (0, 1)$ denotes the pre-specified threshold for that instance. A manager can set the threshold level to meet a variety of objectives. For example, different threshold levels can be set for different product-store combinations to reflect the historical performance of the associated human forecasters. Though there is no causal interpretation to the way we have computed an adjustment's predicted quality, the use of a forecaster's historical performance as a threshold might appear fair and increase trust in the firm's approach to integrating the recommended adjustment in the final solution.

Results

In this section, we implement and present the results from our two-module framework described in the previous section, in the context of our collaborating retailer.

Module 1: Evaluating an Adjustment Quality.

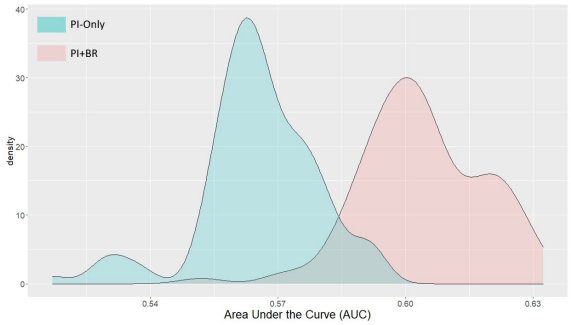
We use the `xgboost` (XGB) package (Chen & Guestrin, 2016) from CRAN to train and test classification models that can predict an adjustment's quality, i.e., probability of $binv = 1$ conditional on included predictors. We selected a gradient boosting methodology given its efficiency and interpretability. As the research question is partly driven by our commercial partner, there was additional emphasis placed on interpretability than in other contexts. Among the most common gradient boosting algorithms, we selected XGBoost over LightGBM because of its longevity and community acceptance.

We train and test classification models on the subsample of forecasting instances in which humans made an adjustment to the AF ($N = 409K$). For robustness, we also fit random forest models, which had similar performance to XGBoost.

We employ a fixed-size rolling window methodology to assess our model performance, using the AUC metric.[6] This approach is a well-accepted method for out-of-sample testing in time-series data (Tashman, 2000). Each rolling window comprises eight weeks of training data and a single week of walk-forward test data; these windows are consistently trained with the same hyperparameters. Specifically, the model that predicts the quality of a judgmental adjustment in week w utilizes data from the prior eight weeks $\{w-8, \ldots, w-1\}$. Rolling the window forward, the model for week $w+1$ is informed by data from $\{w-7, \ldots, w\}$. Analogous to k-fold cross-validation, where one of the k folds serves as holdout data, our methodology reserves week w for out-of-sample testing of model performance.

To study the incremental contribution of the BR predictors, we build and compare two prediction models with the same hyperparameters. The first model includes only PI predictors (PI-only model), while the second includes both the PI and BR predictors (PI+BR model). We compare the AUC of these models using out-of-sample observations, i.e. observations from the walk-forward test weeks of our rolling window

[6] AUC ranges between 0 and 1, with higher values indicating better performance.



(a) Comparison of the AUC Distributions

	PI	PI+BR	Gain[1]
A. *Top 5 Product Categories*			
Healthy Meal	0.50	0.58	16%
Traditional Meal	0.56	0.65	16%
Fish	0.52	0.58	13%
Salads	0.52	0.59	12%
Dairy	0.50	0.55	10%
B. *Bottom 5 Product Categories*			
Desserts	0.54	0.59	8%
Bakery	0.56	0.61	8%
In-store Bakery	0.51	0.54	7%
Meat	0.58	0.62	7%
Vegetables	0.59	0.62	6%

[1] Gain = (1 - PI+BR/BR) × 100

(b) AUC Performance by Product-Category

Figure 3. Performance of Prediction Models.

methodology. We have 143,830 out-of-sample observations. Finally, we use bootstrapped samples for statistical inference on the incremental predictive power of the BR predictors. See the appendix for details of our bootstrap implementation and for discussion on the relative importance of the PI and BR predictors in the two models.

Panel (a) of Figure 3 presents the bootstrap comparison of the AUC values. Compared to the PI-only model, we find that the PI+BR model exhibits a statistically significant increase in AUC, 6.8%*** (95% CI: [6.6%, 7.0%]).[7] Given that the out-of-sample AUC of the PI-only model is 0.565, this improvement in AUC is material. Panel (b) of Figure 3 reports average AUC (across bootstrap iterations) by product categories. Column 1 and 2 report the AUC values of the PI-only and PI+BR models, respectively. When ranking the product categories in descending order of relative gain in their AUCs in the PI+BR model, compared to the corresponding AUC in the PI-only model, we find AUC increases by as much as 16% in select categories with the inclusion of BR predictors.

Module 2: Classification of an Adjustment – Accept or Reject Integration in the Final Forecast.

We analyze the performance of Module 2, which operationalizes our adjustment-level classification approach, relative to popular context-based classification rules that were studied in past work. Specifically, we benchmark our approach against the following five easy-to-implement classification rules: ($\mathcal{C}_1$) always accept AF; ($\mathcal{C}_2$) always accept HF; ($\mathcal{C}_3$) two heuristicss that always-reject small-sized adjustments (Fildes et al., 2009); ($\mathcal{C}_4$) an always-reject rule for adjustments made to products in low-volatility demand environments (Kremer et al., 2010); and ($\mathcal{C}_5$) selective application of always-accept or -reject rule at the product-store level based on the historical performance of human forecasters.

Compared to strategies $\mathcal{C}_1$ and $\mathcal{C}_2$ that apply blanket rules, which ignore any trade-off between human vulnerability to biases and their private information advantage, the other strategies take always-accept or -reject positions based on humans' average performance in applying beneficial adjustments. For example, Fildes et al. (2009) find, on average, small-size adjustments do not improve HF accuracy. Likewise, Kremer et al.

[7] *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Level	MAD	RAD2OD	SOI
psw	$\text{AD}_{psw}(S,\ f)= \lvert S_{psw}\text{-}f_{psw}\rvert$	$\frac{AD_{psw}(S,f)}{AD_{psw}(S,f^{\mathcal{O}})}$	$\mathbb{I}(f = f^{\mathcal{O}})$
ps	$\frac{1}{W_{ps}}\sum_{w=1}^{W_{ps}} \text{AD}_{psw}(S,\ f)$	$\frac{1}{W_{ps}}\sum_{w=1}^{W_{ps}} \frac{AD_{psw}(S,f)}{AD_{psw}(S,f^{\mathcal{O}})}$	$\frac{1}{W_{ps}}\sum_{w=1}^{W_{ps}} \mathbb{I}(f=f^{\mathcal{O}})$
Overall	$\sum \text{MAD}_{ps}/N_{ps}$	$\sum \text{RAD2OD}_{ps}/N_{ps}$	$\sum \text{SOI}_{ps}/N_{ps}$

1. W_{ps} is the count of weeks in which product p was sold in the store s among the out-of-sample observations
2. $\text{AD}_{psw}(S,\ f) = |S_{psw} - f_{psw}|$
2. $f^{\mathcal{O}}$ is the oracle's forecast. And $AD(S,\ f^{\mathcal{O}})$ is the AD of the oracle: $\min(S - f^{h}_{psw},\ S - f^{m}_{psw})$
3. N_{ps} is the number of all product-store combinations.
4. $\mathbb{I}$ is the *Indicator* function.

Table 2. Forecast Accuracy Metrics

(2010) show that human vulnerability to biases becomes prominent in stable environments and, thus, automated decision-making should be emphasized in such environments. Finally, in a recent work, Kesavan and Kushwaha (2020) document human adjustments perform much better for growth-stage products than for mature-stage products.

Theoretically, our adjustment-level classification approach renders greater flexibility in integrating human adjustments and should outperform the always-accept or always-reject rules that are applied at a broader level. However, whether our proposed two-module framework, which leverages predictability in human biases to deliver adjustment-level classification, leads to a substantial improvement in forecast accuracy is an open question. To answer it, we implement the following out-of-sample analysis.

We evaluate the performance of each of the strategies using the same out-of-sample observations ($N = 143,830$) that were used to compare the two prediction model performances (see discussion in previous section). Recall that Module 2 does not involve any "model training". Its implementation, however, requires us to set an ex-ante (of sales) threshold for classifying an adjustment as accept or reject for integration with the final forecast (see eq(1) on page 9). We set $\tau_{ps} = \min(\overline{binv}^{8}_{\text{product-store}}, 0.5)$. Here, $\overline{binv}^{8}_{\text{product-store}}$ is the adjustment's beneficial rate in the immediate eight weeks before the first test week. Effectively, this threshold definition forces our classification procedure to reject an adjustment with poor predicted quality if a human forecaster has shown better application of her/his private information advantage in the past.

Likewise, for $\mathcal{C}_5$, we execute the following rule: for a product-store combination, always accept adjustments only if its $\overline{binv}^{8}$ is higher than the 50^{th} percentile (0.5), otherwise always reject adjustments for it. For $\mathcal{C}_3$, we follow Fildes et al. (2009) in labeling an adjustment as a 'small-sized' one if it changes the AF by less than 20%. We also test an alternate version of this rule that reflects the human forecasters' adjustment patterns in our sample. Specifically, we alternatively label an adjustment as 'small-sized' if the resultant change in AF is below the 50^{th} percentile (5.12%). Finally, to implement $\mathcal{C}_4$, for a forecasting instance psw, we capture demand volatility using the standard deviation in sales $\text{SD}(S)_{psw}$ and the sales rolling-average $\text{RAVG}(S)_{psw}$ over the preceding four weeks ($w-1, \ldots, w-4$). We classify demand volatility for forecast f_{psw} as low if $\text{SD}(S)_{psw}/\text{RAVG}(S)_{psw}$ is below the 50^{th} percentile (0.23).

We evaluate forecast accuracy under a classification rule using three metrics: (i) Mean Absolute Deviation (MAD), (ii) Ratio of Absolute Deviation to Oracle Deviation (RAD2OD), and (iii) Share of Oracle-like Instances (SOI). We use the same bootstrap method, as in the previous section, to infer the statistical difference in forecast accuracy under our adjustment-level classification rule (referred to as $\mathcal{C}_0$) against the five benchmark rules $\mathcal{C}_1$-$\mathcal{C}_5$. Next, we provide some intuition for these metrics. Formal definitions are shown in Table 2.

MAD is the average of absolute deviations (AD) between the realized sales and the final forecast across

weeks. AD-based measures are commonly used to measure and compare forecast accuracy. Theoretically, the optimal forecast is one that attains zero AD – desirable but rarely achievable. Note that under a classification rule that either accepts or rejects an adjustment, the best achievable AD in forecasting instance psw is $\min(|S_{psw} - HF|, |S_{psw} - AF|)$. We refer to this best-possible AD as the absolute deviation under the oracle forecast, $AD(S, f^{\mathcal{O}})$, where $f^{\mathcal{O}}$ refers to the oracle's forecast. Even though the lowest possible $AD(S, f^{\mathcal{O}})$ is zero, this is not achievable unless either the HF or AF matches exactly with observed sales. To provide a more realistic accuracy measure, we use RAD2OD and SOI to evaluate a classification rule's success in attaining forecast accuracy vis-à-vis that of the optimal accuracy under the oracle's final forecast (Kremer et al., 2011).

RAD2OD measures a classification rule $\mathcal{C}$'s ability to emulate the oracle's absolute deviation. The value of RAD2OD ranges from 1 to ∞. The higher the RAD2OD value the poorer the $\mathcal{C}'s$ forecast accuracy. RAD2OD of 1 implies accuracy as good as an oracle. The metric SOI evaluates $\mathcal{C}$'s success rate in achieving the oracle's forecast over time. SOI's value ranges between 0 and 1, with 1 indicating perfect success in achieving the oracle's forecasts. In Table 2, we begin by defining each of our performance metrics for each instance of judgemental adjustment (psw-level). These measures are then aggregated to the product-store level (ps-level), and finally to the bootstrap iteration level for statistical comparison between the classification rules.

Table 3 provides a comparison of forecast accuracy with our adjustment-level classification approach $\mathcal{C}_0$ and the benchmark rules $\mathcal{C}_1$-$\mathcal{C}_5$. The top row denotes the absolute performance of $\mathcal{C}_0$ with the Module 1 prediction model that includes both PI and BR predictors. We find that our focal approach $\mathcal{C}_0$(PI+BR) outperforms *all* other benchmarks to provide statistically significant improvement in forecast accuracy, as measured by the three metrics. Focussing on comparison with $\mathcal{C}_5$ on SOI metric, we find that $\mathcal{C}_0$(PI+BR) and $\mathcal{C}_0$(PI) approaches improve forecast accuracy by nearly 12.5% (∼(0.57/0.51-1)×100) and 6.8% (∼(0.54/0.51-1)×100), respectively. This, in turn, suggest that about 46% (∼(1- 6.8/12.5)×100) of the improvement seen in $\mathcal{C}_0$(PI+BR) is associated with the addition of BR predictors. Further, we find that the forecast accuracy improvement with $\mathcal{C}_0$(PI+BR) is not attributable to select products, stores, or time periods.

Discussion and Implications for Future Research

We showcase a novel adjustment-level classification framework for integrating judgemental adjustments in the final forecast. It uses an algorithm to leverage the predictors of (a) human vulnerability to systematic biases and (b) their private information advantage to predict the quality of their judgments applied to data-driven algorithm-generated forecasts. Using data from an industry collaborator, we show that a simple threshold-based heuristic that classifies each adjustment as accept or reject for integration into the final forecast can substantially improve forecast accuracy compared to popular extant context-based classification rules. Our framework expands the practitioners' toolkit to integrate judgemental adjustments in the final forecast.

Our exploration into human-machine collaboration opens the door to several new lines of inquiry that merit academic attention. First, an extension of Module 1 could consider the search for behavioral response (BR) predictors that can enhance classification model performance. The current binary assessment of adjustment quality could be further refined. For example, a continuous measure of the forecast's deviation from actual sales could offer more nuanced insights into the performance of the two-module framework proposed in this study.

Second, there's an opportunity to employ more advanced machine learning models to improve the predictability of adjustment quality. This dovetails with Module 2's focus on developing effective strategies for using predicted adjustment quality to enhance forecast accuracy. As we bridge these two modules, the question arises: Can a classification-based approach improve an ensemble of algorithmic and human-generated direct forecasts (DFs), especially given the inconsistency in human application of private information as indicated by Ibrahim et al. (2021)?

Additionally, as data-driven algorithms gain traction in practitioner settings (Berk et al., 2018), identifying new contexts for their application becomes vital. Does the documented bias in demand forecasting prevail in these new areas? If so, context-specific BR predictors could be developed, thereby further unlocking the

#	Rule	MAD[1]	RAD2OD	SOI	MSE[1]	SMAPE	Classification Rule
$\mathcal{C}_0$	Classification: PI+PR	7.45	1.93	0.57	194.61	0.146	Adjustment-level classification, PI+BR model
$\mathcal{C}_0$	Classification: PI	-0.08 (-45.2)	-0.03 (-19.02)	0.03 (45.85)	-5.94 (6.35)	-0.001 (14.4)	Adjustment-level classification, PI-only model
$\mathcal{C}_1$	Always AF	-0.13 (-26.7)	-0.04 (-15.67)	0.04 (39.47)	-5.28 (5.80)	- 0.006 (69.9)	Always reject adjustments
$\mathcal{C}_2$	Always HF	-0.40 (-65.7)	-0.18 (-33.16)	0.09 (58.28)	-20.0 (21.70)	- 0.003 (38.8)	Always accept adjustments
$\mathcal{C}_3$	Small-size Adjustments	-0.07 (-23.6)	-0.04 (-18.2)	0.04 (42.8)	-5.54 (5.92)	-0.001 (17.3)	Reject adjustments changing AF < 20%.[2]
$\mathcal{C}_3$	Small-size Adjustments	-0.30 (-52.0)	-0.16 (-32.8)	0.06 (49.8)	-17.9 (19.4)	- 0.003 (36.2)	Reject adjustments changing AF < 50^{th} percentile.
$\mathcal{C}_4$	High-Low Volatile	-0.07 (-20.5)	-0.05 (-28.16)	0.03 (32.42)	-1.45 (1.65)	-0.001 (11.0)	Reject in low-volatile forecasting
$\mathcal{C}_5$	Product-Store	-0.30 (-47.1)	-0.12 (-29.9)	0.06 (47.8)	-16.3 (17.5)	- 0.002 (29.5)	Product-store combinations with $\overline{binv}^8 > (\leq) 50^{th}$ percentile

1. The number in brackets provides the *t*-statistic of the pairwise equal-mean test. Lower values indicate larger errors. 2. Same threshold as in Fildes et al. (2009)

Table 3. Forecast Accuracy Under Different Classification Rules

potential of human-algorithm collaboration.

Another pressing issue is human resistance to machine errors, which is often greater than their tolerance for their own mistakes (Donohue et al., 2018). Could feedback mechanisms on forecast accuracy, incorporating BR predictors, help alleviate this algorithm aversion? Additionally, it's worth noting that the scalability of our current method across different geographies remains an open question. A trans-national study would offer valuable insights into the framework's robustness and adaptability, further validating its effectiveness beyond the initial industrial setting.

Moreover, the manner in which humans' private information is utilized could significantly influence the effectiveness of our classification approach. Specifically, research could explore whether this effectiveness varies depending on whether human information is integrated at the algorithm input stage or during the output (Brau et al., 2023).

Lastly, two avenues of research focus on the long-term implications of our framework. One explores how humans might adapt to adjustment regulation over time. Do they become more risk-seeking, given the framework's filtering of potentially harmful adjustments? The second pertains to the design of incentives for forecasting teams who lack full control over the final forecasts. Here, the literature on combining separate direct forecasts from humans and machines could offer invaluable insights.

By addressing these questions, future research can make substantive contributions to the optimization of human-machine collaboration in decision-making, thus advancing the field of Information Systems.

References

Akkas, A., Gaur, V., & Simchi-Levi, D. (2019). Drivers of product expiration in consumer packaged goods retailing. *Management Science*, *65*(5), 2179–2195.

Arvan, M., Fahimnia, B., Reisi, M., & Siemsen, E. (2019). Integrating human judgement into quantitative forecasting methods: A review. *Omega*, *86*, 237–252.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44.

Berrar, D., & Dubitzky, W. (2013). Information gain. *Encyclopedia of Systems Biology; Dubitzky, W., Wolkenhauer, O., Yokota, H., Cho, K.-H., Eds*, 1022–1023.

Brau, R., Aloysius, J., & Siemsen, E. (2023). Demand planning for the digital supply chain: How to integrate human judgment and predictive analytics. *Journal of Operations Management*, *69*(6), 965–982.

Caro, F., & de Tejada Cuenca, A. S. (2023). Believing in analytics: Managers' adherence to price recommendations from a DSS. *Manufacturing & Service Operations Management*, *25*(2), 524–542.

Chen, T., He, M., Tand Benesty, & Khotilovich, V. (2019). Package 'xgboost'. *R version*, *90*, 1–66.

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

Donohue, K., Katok, E., & Leider, S. (Eds.). (2018). *The handbook of behavioral operations*. John Wiley & Sons, Inc.

Evans, J. S. (1982). Psychological pitfalls in forecasting. *Futures*, *14*(4), 258–265.

Fawcett, T. (2004). Roc graphs: Notes and practical considerations for researchers. *Machine learning*, *31*(1), 1–38.

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.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting*, *16*(1), 85–99.

Ibrahim, R., & Kim, S.-H. (2019). Is expert input valuable? the case of predicting surgery duration. *Seoul Journal of Business, Forthcoming*.

Ibrahim, R., Kim, S.-H., & Tong, J. (2021). Eliciting human judgment for prediction algorithms. *Management Science*, *67*(4), 2314–2325.

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.

Kremer, M., Minner, S., & Wassenhove, L. N. V. (2010). Do random errors explain newsvendor behavior? *Manufacturing & Service Operations Management*, *12*(4), 673–681.

Kremer, M., Moritz, B., & Siemsen, E. (2011). Demand forecasting behavior: System neglect and change detection. *Management Science*, *57*(10), 1827–1843.

Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., & Lawrence, M. (2007). Providing support for the use of analogies in demand forecasting tasks. *International Journal of Forecasting*, *23*(3), 377–390.

Moritz, B. B., Narayanan, A., & Parker, C. (2022). Unraveling behavioral ordering: Relative costs and the bullwhip effect. *Manufacturing & Service Operations Management*, *24*(3), 1733–1750.

Schweitzer, M. E., & Cachon, G. P. (2000). Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science*, *46*(3), 404–420.

Sun, J., Zhang, D. J., Hu, H., & Mieghem, J. A. V. (2022). Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*, *68*(2), 846–865.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*(4), 437–450.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

van Donselaar, K. H., Gaur, V., van Woensel, T., Broekmeulen, R. A. C. M., & Fransoo, J. C. (2010). Ordering behavior in retail stores and implications for automated replenishment. *Management Science*, *56*(5), 766–784.

Appendix

Variable Definitions

Humans are prone to a range of biases that may negatively affect the quality of their judgemental adjustments. Interestingly, these biases often occur in systematically repeatable patterns (Tversky & Kahneman, 1974). We leverage this predictability by including bias-related (BR) predictors of intervention quality. There are many examples of systematic biases in the context of demand forecasting. We give a non-comprehensive list below:

- Humans are known to engage in demand-chasing behavior, where they over-emphasize the previous period's demand realization and the error in their forecast when making forecasts in the current period (Schweitzer & Cachon, 2000).
- Humans tend to over-emphasize recent performance (i.e., recency bias – see Evans (1982))
- Humans exhibit anchoring-and-adjusting bias, where they anchor on an initial value (e.g., AF in our case) and insufficiently adjust it to yield the final answer (Tversky & Kahneman, 1974)
- Humans erroneously avoid algorithms after seeing them make mistakes (Dietvorst et al., 2015).
- Human judgements are more biased in stable environments (Kremer et al., 2011).

Rows (1) to (11) in Table 4 list the BR predictors that we include in our model to capture human vulnerability to systematic biases. We note that the same predictors can potentially capture multiple underlying biases. For example, using current and first-lag values of algorithm-generated forecasts and judgemental adjustments (rows 1, 2, 4, 5, and 7) in conjunction with indicators of accuracy in the previous period (rows 8, 9, and 11), we intend to capture demand chasing behavior. At the same time, the current period algorithm-generated forecasts (row 1) and judgemental adjustments (rows 2, 4, 5, and 7) can act as indicators of bias caused by employing the anchoring and adjusting heuristic. We include the standard deviation of sales over a rolling window of 4 weeks (row 6) to capture "stability of environment" – which is known to influence bias in human judgments (Kremer et al., 2011). Further, to capture humans' relative trust in algorithms versus their own judgment and potential for algorithm aversion, we include first-lag of accuracy metrics for humans and algorithms (rows 8, 9, 10, and 11). Finally, the first-lag variables (rows 1, 2, 3, 4, 5, 8, 9, 10, and 11) can help to capture potential recency bias. We acknowledge that the proposed list of BR predictors is not a comprehensive list that can predict all kinds of systematic biases to which a human is vulnerable while making a judgmental adjustment.

Rows (12) to (27) list the PI predictors included in the model to capture human forecasters' private information advantage. Motivated by the findings of Kesavan and Kushwaha (2020) and van Donselaar et al. (2010), we add categorical variables that capture time-invariant product-level (rows 17 - 22) and store-level (rows 23 - 26) characteristics. Here, we wish to highlight our use of assortment width metrics, at various levels of product category hierarchy, as PI predictors. These metrics capture humans' potential private information advantage in understanding customer preferences for variety and, consequently, the effects of stockouts within a product subgroup (van Donselaar et al., 2010). As the retailer selectively ranges products across stores and over time to match customer preferences, these metrics are time-varying. However, assortment width is also associated with inattention bias (Akkas et al., 2019), since the effects of customer preferences grow exponentially with the increase in the number of products. Thus, it can be argued that these assortment-width metrics also capture cognitive load for humans and, hence, are potential BR predictors. We take a conservative approach and term these predictors as PI type. This choice could potentially under-represent the incremental predictive power of BR-type predictors compared to PI-type predictors.

Implementation of Module 1 using XGB model

We use the `xgboost` (XGB) package from CRAN to train and test our classification model (Chen et al., 2019). At a high level, XGB models build upon gradient boosting machines (GBM), combining estimates from multiple decision trees where each tree is built iteratively based on the residuals of a prior tree. The algorithm implements approximations for wbuilding decision trees that significantly improves speed and scalability, making it ideal for machine-learning problems with large datasets (like sales forecasting). For a detailed

#	Variable {periods}	Unit	Definition	Type
1	AF, $\{w,\ w-1\}$	psw	Algorithm-generated forecast for demand of product p in store s for week w, current period and first lag. Intended to capture demand chasing, anchoring bias, rece.	BR
2	HF, $\{w,\ w-1\}$	psw	Human-adjusted forecast, current period and first lag	BR
3	S, $\{w-1\}$	psw	Quantity sold, only first lag	BR
4	AdjDir, $\{w,\ w-1\}$	psw	A categorical variable that captures an adjustment's direction. It is set to U if HF > AF and D if HF < AF. Current period and first lag	BR
5	AbsAdjSize, $\{w,\ w-1\}$	psw	Absolute value of adjustment. Current period and first lag	BR
6	StdDevS, $\{w\}$	psw	Standard deviation of the S over the preceding four weeks rolling window (i.e. periods $w-1,\ w-2,\ w-3,$ and $w-4$)	BR
7	AdjScaled, $\{w\}$	psw	Adjustment size scaled by the preceding four weeks running average of sales (i.e. periods $w-1,\ w-2,\ w-3,$ and $w-4$).	BR
8	DevHum, $\{w-1\}$	psw	Deviation of HF compared to sales (= HF - S). First lag only.	BR
9	DevAlgo, $\{w-1\}$	psw	Deviation of AF compared to sales (= AF - S). First lag only.	BR
10	DevAligned, $\{w-1\}$	psw	Set to 1 if (S-AF)×(S-HF) > 0 and to 0 otherwise. Captures whether algorithm and human deviated in the same direction in week w. First lag only.	BR
11	$binv$, $\{w-1\}$	psw	Set to 1 if |(S-HF)| ≤ |(S-AF)| and to 0 otherwise. Captures if HF was a more accurate forecast than AF. First lag only.	BR
12	NP-Store, $\{w\}$	psw	Number of products in store	PI
13	NP-L2, $\{w\}$	psw	Number of products for sale in a level 2 product category	PI
14	NP-L3, $\{w\}$	psw	Number of products for sale in a level 3 product category	PI
15	NP-L4, $\{w\}$	psw	Number of products for sale in a level 4 product category	PI
16	NP-L5, $\{w\}$	psw	Number of products for sale in a level 5 product category	PI
17	IDProd	p	Unique product identifier	PI
18	IDL2	p	Categorical variable for level 2 product category	PI
19	IDL3	p	Categorical variable for level 3 product category	PI
20	IDL4	p	Categorical variable for level 4 product category	PI
21	IDL5	p	Categorical variable for level 5 product category	PI
22	Life	p	Shelf life of a product, which can range from 1 to 14 days	PI
23	IDStore	s	Unique store identifier	PI
24	StoreAttr1[1]	s	A store-level categorical variable capturing primary customer target profile	PI
25	StoreAttr2[1]	s	A store-level categorical variable capturing primary customer target profile	PI
26	StoreAttr3[1]	s	A store-level categorical variable capturing its format and size.	PI
27	StoreAttr4[1]	s	A store-level categorical variable capturing its location type.	PI

[1] Limited information presented on our collaborator's request.

Table 4. Variable Definitions

description of XGB, see Chen et al. (2019) and Chen and Guestrin (2016).

We use a fixed-size rolling window procedure (also known as fixed-size rolling sample procedure) for out-of-sample testing of our models (Tashman, 2000). This is a popular procedure for testing out-of-sample model performance when working with time-series data.

In line with the context of our collaborator, we set our *forecasting horizon* to one week. Our fixed-size rolling *training period* is 8 weeks long and our *test period* is the subsequent walk-forward week. As the retailer selectively ranges products across stores and over time to match customer preferences, not every $product \times store$ is seen throughout the 8 weeks in every training period. In each *training period*, we drop product-store combinations that do not have at least 5 (more than half the window size) observations. We work with 18 weeks of overall data. Due to the selective ranging policies, each $product \times store$ features in our test set 4.5 times on average and up to a maximum of 10 times. The total number of test period observations is 143,830.

We train and test our models using the subsample of observations in which humans made a non-zero adjustment to the algorithm-generated forecast ($N = 409k$). We perform a grid search to tune the hyperparameters of our XGB model. Our grid search yielded the following hyperparameters for our XGB model: depth = 9, eta rate = 0.05, number of rounds: 100, objective to binary:logistics, and default values for the remaining model parameters.

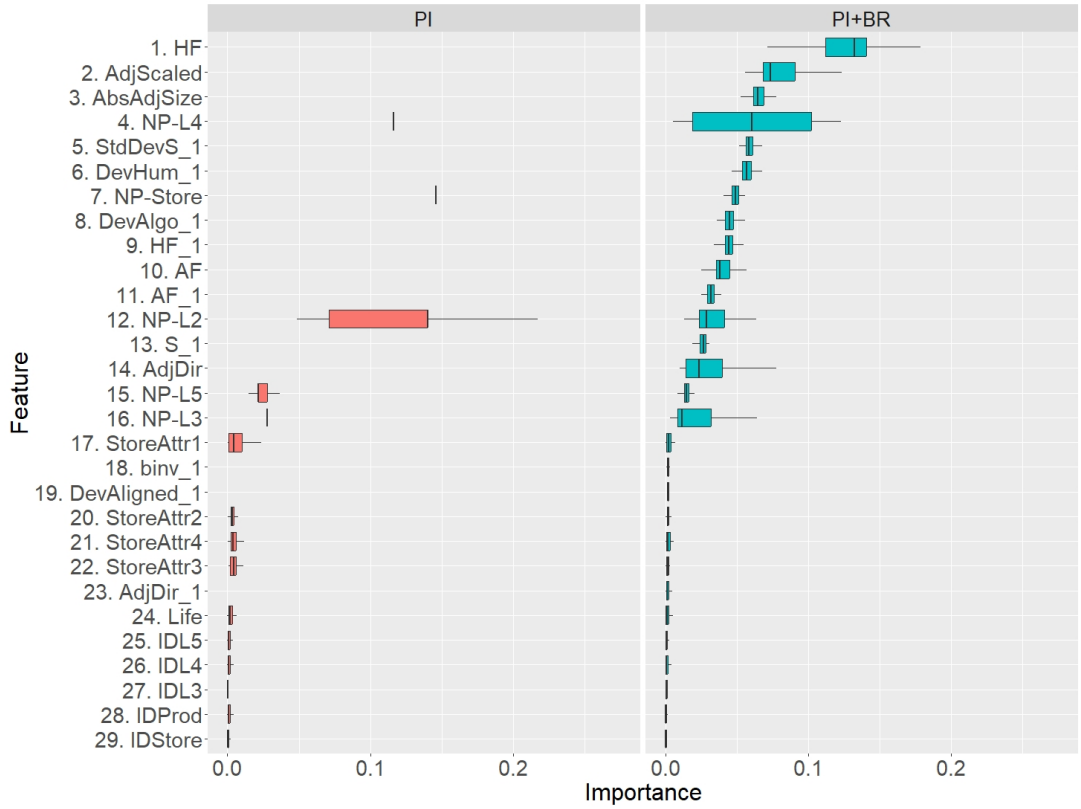
Bootstrap Implementation Details.

Our sample consists of 27 level-4 product categories. We chose level-4 as our sampling level because human forecasters are responsible for multiple similar products. Therefore, we sought to find a balance between (i) groupings that are too broad such that products which are unlikely to be forecasted by the same human end up in the same group, and (ii) groupings that are too narrow such that a human forecaster might be responsible for products across multiple groupings. Sampling at the balance of these two allows us to preserve cross-product effects in each draw.

For each bootstrap iteration, we make random draws, with replacement, of ten level-4 product categories. Next, we train and test both the PI-only and PI+BR models using all the $product \times store \times week$ observations of these ten drawn categories. We capture the corresponding AUC values. To generate an empirical distribution of the performance of the two models, we repeat the iteration procedure a hundred times. Summary statistics on the bootstrap sample, compared to the full sample, are shown in Table 5. We conduct statistical inference using paired t-tests where performance is calculated across models based on the same individual samples.

	Sample			Bootstrap		
Statistic	Mean	St. Dev.	50^{th}	Mean	St. Dev.	50^{th}
Sales $\widehat{S}$	0	1	-0.31	0	0.98	-0.08
Algorithm-generated forecast $\widehat{f}^{a}_{psw}$	1.23	1.09	1.04	1.22	1.16	1.03
$ScaledAbsRevSize > 0$	7.64	7.11	5.07	7.84	8.65	5.03
Algorithm-generated forecast $\widehat{\text{MAD}}_{\text{AF}}$	0.36	0.59	0.29	0.32	0.21	0.28
Human-adjusted forecast $\widehat{\text{MAD}}_{\text{HF}}$	0.34	0.37	0.28	0.32	0.19	0.28

Table 5. Summary statistics



Note: A predictor's importance is measured using *information gain* – a metric that measures how information contained in the predictor lowers the uncertainty in the prediction task (Berrar & Dubitzky, 2013). A higher percentage means a more important predictor. The box plot shows the redictors' importance distribution, across the 100 iterations, using the summary-tuple of (mean, 25^{th} percentile and 75^{th} percentile).

Figure 4. Predictors Relative Importance Summary

Relative importance of Features

To understand the relative importance of each predictor, we use *information gain* – a metric that measures how the information contained in a particular predictor lowers the uncertainty in the prediction task (Berrar & Dubitzky, 2013).

Figure 4 in the appendix shows the relative importance of the predictors in both models. A few insights from this analysis are noteworthy: first, while the PI predictor *NP-L4*, which measures the number of products in a level-4 category, is top ranked in the PI-only model, it is replaced by a BR predictor (HF_w) in the PI+BR model. As we discuss in an earlier section within the appendix, HF_w predictor could contain indicators for multiple systematic biases when used in combination with other predictors. For instance, in combination with last period sales (S_{w-1}), and other variables (see appendix) it proxies for any demand chasing behavior. Second, eight of the top ten ranked predictors in the PI+BR model are of the BR type. Third, the PI+BR model list includes both the current and past week BR predictors. Collectively, these insights corroborate the relevance of BR predictors in predicting human adjustment quality.