Human-Machine Teaming for Intelligent Demand Planning

by

Ye Ma
Bachelor of Science, Computing and Software System

University of Melbourne, 2015

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ENGINEERING IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Signature of Author: _____
Department of Supply Chain Management
May 8, 2020

Certified by: _____
Dr. Maria Jesus Saenz Gil De Gomez
Executive Director, MIT Supply Chain Management Blended Program
Thesis Advisor

Accepted by: _____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Human-Machine Teaming for Intelligent Demand-planning

by
Ye Ma
Submitted to the Program in Supply Chain Management
on May 8, 2020 in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering in Supply Chain Management

ABSTRACT

The second machine age is reshaping the way we work, do business, and collaborate. Today collaboration is switching from just among humans to between humans and machines. Mundane and repetitive tasks will be done by machines automatically, while humans can develop insights and make wise decisions supported by data streaming from intelligent machines. If and how different human-machine teaming decision-making structures would influence the organization's performance is important to understand, so that human-machine teaming capabilities could contribute the most to business outcomes.

By using the augmented inverse propensity weight estimator method, this research empirically analyzes the average treatment effects of three different human-machine decision-making structures: Full human to AI delegation, Hybrid AI-Human with adequate human intervention, and Hybrid AI-Human with all steps of demand planning overrides. These three decision-making structures are defined as treatment groups, and the traditional manual demand-adjustment process is defined as the control group. Effects of switching human-machine teaming decision-making structures from one to another are also analyzed. The performance of each treatment and control group is measured by the long-term forecast accuracy, short-term forecast accuracy, and customer inventory level. The project is based on an IT collaboration project between a large fast-moving consumer goods company and one of its largest e-commerce customers. The project implemented an AI-enabled demand-adjustment process to incorporate the external e-commerce customer demand signals into existing demand-planning process. Demand planners engage in the demand-adjustment process via web-based interfaces, to apply human judgment-based decisions. All the stock keeping units are randomly assigned to treatment and control groups.

The results show that after the implementation of human-machine teaming decision-making structures, both demand-forecast accuracy and inventory level are strongly improved by at least 47%. Overall, the Hybrid AI-Human with adequate human intervention model is the optimal decision-making structures between human and machine, which improves the short-term forecast accuracy by 53%, long-term forecast accuracy by 64%, and inventory level by 70%. The Hybrid AI-Human with all steps of demand planning overrides model performed worse than the previous model, because of the heavy human overrides. Additionally, those AI enabled decision-making structures works better for low-turnover products than high-turnover ones.

Thesis Advisor: Dr. Maria Jesus Saenz Gil De Gomez
Title: Executive Director, MIT Supply Chain Management Blended Program

ACKNOWLEDGMENTS

My deepest appreciation goes first to Dr. Maria Jesus Saenz Gil De Gomez, who guides me, motivates me, encourages me to move on and forward on the journey, which finally made this work possible. I really appreciate her patience and care. I would also like to express my thanks to Prof. Elena Revilla for her help and suggestions to guide on the methodologies of this work. Thank you so much, Toby Gooley, for your professional writing assistance and editing this work time and time again.

To Prof. Yossi Sheffi, Dr. Chris Caplice, Dr. Eva Ponce, Dr. Maria Jesus Saenz Gil De Gomez (again), Dr. Josué C. Velázquez Martínez, Robert R. Cummings, Leonard Morrison, Justin Snow, Bonnie Borthwick, Aren Ghazarians, Arthur Grau, and the entire SCx, SCMr, SCMb staff, thank you for your dedication to SCM MicroMasters and blended program, which provided me the opportunity to attend MIT and realize my dream. The experience is extraordinary!

To people who helped and made me here: Dr. Michael Fang, Prof. George Huang, Alan Zhao, Jeffrey Chen, Peter He, Krystal Hu, Nia Lu, Qiuyu Ye, Pat Peng, Lin Ge, Stewart Wu.

To friends in MIT, we really experienced a lot, and made it, together. Long live our friendship.

To my parents, who always love me, support me, anything, anywhere.

Finally, to the time, to myself.

谨以此篇纪念在 美利坚合众国 波士顿 剑桥 麻省理工学院 の 经学岁月 人机共舞

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Background

1.1.1 Demand Forecasting and Supply Chain

The demand forecast is an essential part of the supply chain management system for companies operating their business within cut-throat competition (Tai, Ho & Wu, 2010). However, it has been characterized by heavily manual work and ineffective information system handling, which leads to inadequate quality control, trend forecasting, and financial efficiency (Wang, Gunasekaran, Ngai & Papadopoulos, 2016). Demand forecasting is also strongly connected with inventory management due to its impact on the replenishment schedules, production arrangements, delivery plans, and need to process perishable products in the fast moving consumer goods (FMCG) industry (Liu, Sun, Wang & Zhao, 2011).

In large-scale FMCG companies in the e-commerce domain, the replenishment team is overloaded with the demand-planning and replenishment process (Barngetuny & Kimutai, 2015). They need to consider many factors to determine how to adjust the demand plan in accordance with the company's national plan and the specific key customers' demand signals. Those procedures are repeated each time for hundreds of stock-keeping units (SKU) in many online stores. Within supply chain management, integrated demand-planning processes and decision-making structures with advanced tools are urgently needed to improve demand forecast accuracy and inventory efficiency.

1.1.2 AI and Humans in Demand Forecasting

Together with artificial intelligence (AI) algorithms and automation tools, digitalization with either robotic process automation (RPA) or business process management (BPM) software implementation is a proven way to improve business process accuracy and efficiency in many use cases (Anagnoste, 2017). Previous research (Saenz, Revilla and Simon, 2020) shows that human and machine teaming models could determine whether varied AI system capabilities and implementation would be successful or not. However, there has been too little investigation on the AI-Human teaming decision-making structures in the supply chain, specifically in demand signal selection and adjustment, to formulate effective demand forecasting processes.

1.2 Motivation and Research Questions

The project sponsor company supplier P wants to integrate a key customer's demand signals into the existing demand process, by implementing AI-enabled automation process and taking effective human interventions. It is believed that AI algorithms could provide a robust demand forecast automatically and efficiently, while humans could input their expertise and farsighted information to further calibrate the results. Therefore, the human-machine teaming capabilities could contribute to a more accurate demand forecast result, so that the company could further improve the customer-specific order service level, inventory efficiency, turnover rate, and even customer sales.

Because of the importance of the human-machine teaming capabilities in the digitalized supply chain demand-planning systems, there is a need for a better understanding of how different human-

machine interface (HMI)-based decision-making structures influence demand forecasting in e-commerce. However, there is little investigation on how human-machine teaming works together in supply chain demand-planning, especially relative to how to allocate the national demand forecast among specific customers or regions. Through this thesis research, we will formulate a more effective demand forecasting adjustment based on external customer-provided demand signals.

The research questions are:

- If and how could different human-machine teaming decision-making structures improve demand forecast accuracy and inventory level?

- Which of the structures would provide an optimal approach for demand forecasting and inventory level: Full AI delegation, or hybrid (AI-to-Human) with different levels of human intervention?

1.3 Relevance and Company Overview

Supplier P is one of the largest fast-moving consumer goods company, who has an overall supply chain synchronization strategy from end to end. Digitalization in demand-planning is the key part of their blueprint for business transformation. E-commerce company J, as one of the most important e-commerce customers of supplier P, has significant business needs for customer-specific demand forecasting adjustments from supplier P due to its large business scale. In this case, customer J provides its own demand forecast to supplier P, so that supplier P will be able to

consider customer J's demand signals 13 weeks in advance to improve its product-supply service level. This project leveraged the results of the system that translates the customer shipment forecast into demand signals. The intelligent demand adjustment (IDA) system is part of the supplier P Greater China smart customer-collaboration program, which is attempting to integrate the customer demand signal into the supplier P demand forecast process for better planning performance. It is believed that if we can bring the external demand signals into the demand-planning process, assuming good demand signal accuracy and information quality, the overall accuracy of demand-planning and service level (to customers) can be improved, so as to bring business benefits.

1.4 Problem Setting and Overall Process

The overall, newly designed demand-adjustment process flow is shown in Figure 1. The project scope is shown under the "incremental conversion" box area. Detailed adjusted planning steps inside the incremental conversion will be discussed in Methodology Section 3.2.

As Figure 1 shows, customer J provides a rolling 13-week order forecast (based on their sales forecast and considered inventory factors) at the SKU level as customer demand signals to supplier P for implementing necessary demand-adjustment. Before this project, supplier P only considered their demand forecast based on their existing systems, such as a detail assumption tool (DAT) and integrated demand-planning (IDP), to provide a national-level forecast, without any specific customer demand-adjustment.

This project creates an incremental conversion bridge between customer J and supplier P, where demand signals are received, selected, and processed efficiently. Multiple AI-Human teaming decision models are tested in the incremental conversion segment in order to find the optimal roles for machines and humans teaming on demand forecasting in e-commerce. Ultimately, the adjusted demand could influence the production and distribution plan, to better fulfill customer J's actual orders and maintain an efficient inventory level.

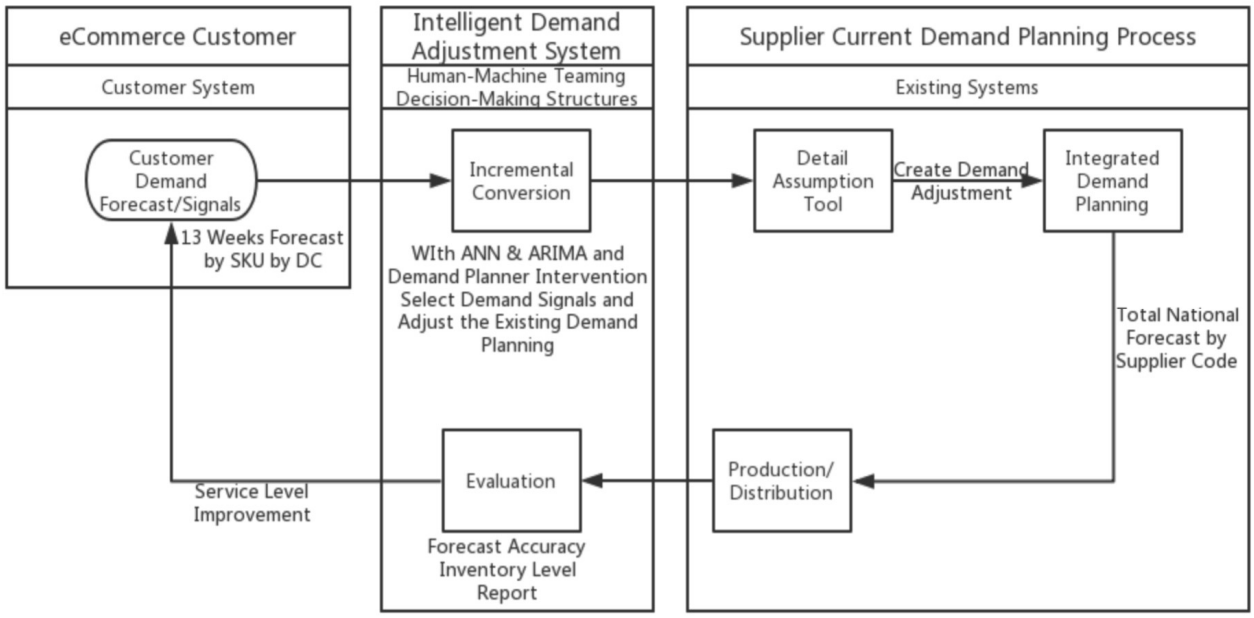

Figure 1. Flow of customer J demand signal to supplier P demand-planning process

1.5 Summary and Thesis Structure

This paper includes 6 sections. In Section 2, we go through a comprehensive review of e-commerce demand forecasting, AI-Human decision-making, teaming capabilities models, and their characteristics. Section 3 illustrates the methodology, and it introduces how different human-machine teaming models' experiment scenarios were designed, and what analysis methods were applied. Section 4 presents and discusses the results of implementing different human-machine

teaming decision-making structure models in a real-world experiment. In Section 5, we discuss those different human-machine cooperation models, justify the methods, and illustrate the insights based on the results. We also identify the optimal teaming model in this e-commerce demand-planning process, and the potential reasons behind it. In Section 6, we conclude the experiment and analysis, and give our demand-planning recommendations in system implementation and human-machine interaction.

2 LITERATURE REVIEW

This section provides an overview of project-related studies. There are five subsections in this literature review section. The review starts by introducing the relation between e-commerce and supply chain, stressing the fast growth rate and rich data availability of e-commerce, as well as the importance and difficulties of data sharing in e-commerce. The second part introduces demand-planning in supply chain management, discussing how demand-planning influences the supply chain, the role the demand planner plays in the process, and how an IT system changes the ways of demand-planning. The third part discusses demand-planning and AI forecast applications. This part introduces AI and machine learning techniques used in the forecast applications and one popular framework in machine learning implementation. The fourth part discusses human-machine teaming features and capabilities. In this part, the decision-making structures this experiment is designed are described. The last section explains how this thesis contributes to the body of literature on human-machine decision-making structures influence on demand forecast and inventory management.

2.1 E-commerce and Supply Chain

E-commerce is the execution of business over the Internet; its success and efficiency highly depend on technology-supported digital commercialism. Electronic systems provide service to all key business units of e-commerce, which include purchasing, sales, marketing, and customer

service (Wong, 2010). E-commerce plays a particularly important role in the current business world, and its importance will continuously increase (Reinsch, 2005). Consumers' shopping habits have changed rapidly with the fast-growing e-commerce and technologies (Klein & Rai, 2009). As evidenced by the downward trend of offline retail stores, in 2019, 638 million Chinese Internet users shopped online and contributed to around USD 1.3 trillion gross merchandise volume (China Internet Network Information Center, 2019). Due to the large scale of the e-business and unpredictable consumer behaviors, e-commerce demand is dramatically fluctuating, which leads to a significant supply chain bullwhip effect. Bullwhip effect is the phenomenon that the supply chain inefficiency due to demand fluctuation yielding from end to end supply chain, which would lead to high customer inventory level (Zhao, Zhu & Zheng, 2018).

Qian (2016) shows that the information sharing would reduce the bullwhip effect and contribute to customer collaboration. Due to the electronic nature of e-commerce, the rich information among the electronic networks could be leveraged to enhance supply chain management (Zhao et al., 2018). Because of the large-scale business and multiple supply chain players, who all want to maximize their own profits, it is hard to synchronize information from end to end (Iyer, Narasimhan & Niraj, 2007). Rached, Bahroun and Campagne (2015) shows that it is important to share information smoothly from suppliers to retailers to enable coordination, so that the whole supply chain could be efficient and profitable. This is also the reason why customer J would like to share its own demand forecast information with supplier P —to help improve the overall supply chain performance. However, Chen & Lee (2009) found that if sharing information among supply chain nodes costs a lot while contributing little value, companies are much less willing to share their supply chain data, such as their cost, demand, order, and inventory

information. Both supplier P and customer J have strong IT capabilities to exchange data with electronic data interchange (EDI) smoothly and cheaply. To answer the further question of how to leverage the external demand signals from retailers effectively to improve the overall supply chain, the next section will introduce the significance of the demand-planning process.

2.2 Demand-planning in Supply Chain Management

Demand-planning is the first master planning task that defines the operation plan, which is a crucial part of supply chain management, where human knowledge particularly matters (Hauke, Lorscheid & Meyer, 2018). To balance demand and supply, demand-planning processes leverage internal and external information to do forecasting, so that company can coordinate material sourcing, product manufacturing, and customer delivery accordingly (Zhou, Benton, Schilling, and Milligan, 2011). Therefore, demand-planning accuracy influences the whole supply chain to a remarkable degree (Chopra & Meindl, 2010), especially in inventory efficiency, production schedule, and customer service levels (Moon, Mentzer, & Smith, 2003). The demand-planning decision-making structure also significantly influences product inventory levels (Wang & Petropoulos, 2016); for that reason, this research takes the absolute inventory amount as one of the outcome variables to measure the business impacts.

Demand planners are responsible for the demand forecast accuracy, based on their expertise, knowledge, and the necessary internal and external information they collect across the whole company as well as among colleagues and systems (Jonsson, Kjellsdotter & Rudberg, 2007). Interpersonal connection and collaboration among cross-function staff such as finance, sales,

marketing, IT, and even external customers are required for demand planners to do the forecast (Oliva and Watson, 2011). In addition, Kaipia, Holmström, Småros and Rajala (2017) show that customer collaboration and information sharing are very important in demand-planning; the stronger the cooperation among supply chain players, the better the demand forecast accuracy. Therefore, the demand planners' comprehensive business awareness and decision-making structures are crucial in the demand-planning process (Barnes and Y., 2012). However, both Fildes, Goodwin, Lawrence and Nikolopoulos (2009) and Moritz, Siemsen & Kremer (2014) found that, because of humans' cognitive nature, a pure human decision-making structure might influence the forecast result by overacting, biased perspective, or lack of information.

Nowadays, the complex demand-planning process includes not only the overall information integration and interpersonal communications, but also the IT systems interaction and technologies (Zoryk-Schalla, Fransoo, and de Kok 2004). For example, web-based software can connect the demand-planning workflow and stakeholders, and process forecasting tasks automatically. The improved demand forecast process might contribute to increase forecast accuracy and saving costs (Chybalski, 2017). Supplier P implemented an intelligent demand-adjustment (IDA) system to handle the demand signals they received from key customers, so that they can adjust their internal demand-planning by SKU level efficiently. Otherwise, it is impossible for demand planners to adjust demand-planning by SKU manually. This research set up the experiment to assess the effects of the redesigned decision-making structure in the demand-forecasting process. In the next section, the demand-planning, related systems, and AI forecasting applications are introduced.

2.3 Demand-planning and AI Forecast Applications

Artificial intelligence, also called machine intelligence, is a technology that mimics the way natural human intelligence functions, and is a science that simulates, expands, and extends human intelligence by integrating theories, methodologies, and systems (Ren & Bao, 2020). The advanced development of AI has solved many difficult tasks, such as image recognition by computer vision and language translation by natural language processing, as well as the Chinese chess Go by AlphaGo with deep learning (Brynjolfsson and McAfee, 2014).

One sub-area of AI that is widely used in supply chain management domain, is machine learning, which is the study about designing computer algorithms that making a machine have the ability to improve itself automatically from data and experience (Wenzel, Smit & Sardesai, 2019). There are three well-known machine learning types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is the process of model training based on labeled data; the model learns to connect the input and output, and then use the trained model to predict the future. Classification and regression are the main objectives for supervised learning (Russell and Norvig, 2010). On the other hand, unsupervised learning is the process to find out patterns from data without labeled data as a training set. Reinforcement learning trains the model by interacting with the environment with reward and punishment. In practice, machine learning is usually developed as the data mining, analytics part of an IT system. One widely used framework (Figure 2) was developed by Wirth and Hipp (2000) to implement the machine

learning project, the Cross-industry standard process for data mining model (CRISP-DM), which is still very popular today across industries.

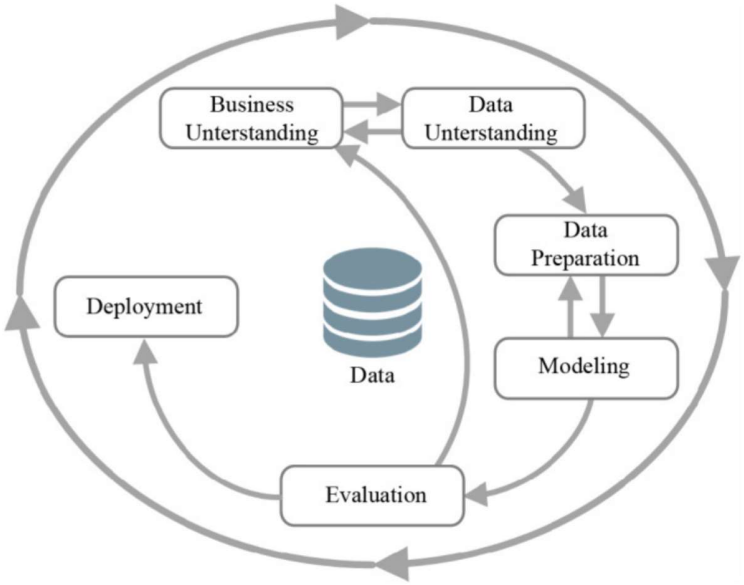

Figure 2.The CRISP-DM process (Shearer, C., 2000)

In forecast domains, for both demand forecasting and sales forecasting applications, machine learning techniques such as artificial neural network (ANN) are often used together with multiple statistical techniques, including time series. A group of time series techniques includes moving average, autoregressive model, autoregressive moving average (ARMA), and autoregressive integrate moving average (ARIMA). ARIMA is used to predict future values efficiently in nonstationary data patterns based on past data and errors (Kapila, Seneviratna, Jianguo and Arumawadu, 2015). Recurrent neural network (RNN) can be used to identify features and data patterns in sequential time ordered, and it is also a data-driven self-adaptive model (Adebiyi, Adewumi & Ayo, 2014), which would correct and improve its own performance based on the evaluation it received. Among the ANN and time series hybrid models, RNN and ARIMA performed with high forecasting accuracy in recent testing (Hiranya, Karunathilake, Achira & Ganegoda, 2018). Therefore, supplier P selected the RNN and ARIMA as the AI algorithm

model embedded in the IDA demand-planning process to take the demand-adjustment forecast, as the research AI model.

2.4 Human-Machine Teaming Capabilities

Decision-making structures engaged with humans and AI algorithms would influence an organization's performance significantly (Georg, 2018). It is a challenge for organizations to introduce an appropriate human-machine decision-making structure to successfully leverage the full human-machine teaming capabilities (Saenz et al., 2020). Professional managers need to clearly understand the strengths and weaknesses of human-AI decision-making structures, because they are still the owners of the relevant business results.

Shrestha & Ben-Menahem (2019) give a human-AI decision-making structure framework to categorize the different human and machine teaming models: Full human-to-AI delegation, hybrid AI-to-human or human to AI sequential decision making, and aggregated human-AI decision making. This research selects the first two models to design the experiment, then examines their treatment effects on treatment groups and a traditional manual control group. Randomly selecting SKUs for the experiment with different levels of human intervention and machine automation in the demand-planning process will reveal the treatment effects of different decision-making structures for human-machine teaming.

2.5 Literature Gap

The current literature illustrates the importance of information sharing in supply chain management and demand-planning process in an e-commerce environment. As part of the large supply chain demand-planning system, research shows that both AI/ML model and human engagement significantly influence forecast accuracy and inventory level. ML algorithms combined with statistical techniques improve the forecast accuracy. On the other hand, human over-intervention decreases the value-added forecast performance. The existing studies also empirically examine the effects on forecast accuracy and inventory level, by applying different planning strategies: statistical models, human judgment-based decisions, and combinations of these strategies. The research has defined models of different human-machine decision-making structures, as well as varied human-machine teaming capacities. However, there are few studies on how different human-machine decision-making structures' implementation would impact the organizational performance, such as forecast accuracy and inventory level. Further, the question of which is the optimal human-machine decision-making structure in the demand-planning process has also not been answered. This thesis designed an experiment to empirically evaluate different human-machine decision-making structures to determine their impact on forecast accuracy and inventory level.

3 DATA AND METHODOLOGY

This chapter introduces the data collected for the research and methods this research project applied. Section 3.1 Data includes the overall information flow, the data sources, and how the data are processed. Section 3.2 Experiment Design and Methodology includes the demand-adjustment process of each step, the empirical methods used to determine the treatment effects, and the pretreatment variables, treatment variables, and outcome variables among those treatment groups.

3.1 Data

The sponsor company supplier P provided data that the research needed from the IDA system. The data contains demand forecasting-related data from both supplier P internally and their customer J externally. Those datasets include demand forecast data (from both supplier P & customer J), actual order shipment data, and system auto-generated accuracy results related to the demand-planning.

3.1.1 Data Flow

Figure 3 shows the overall intelligent demand-adjustment (IDA) system information flow, including four data sources in rhomboids: national-level historical shipments, customer-J-specific historical shipments, supplier P integrated demand-planning demand-forecast, and customer J's order forecast. The rectangles are IDA in-process data-conversion and data-processing nodes; those data would be hidden within the system. Human-machine decision-making and intervention points are listed in circles with step number; those steps of human-machine interaction points are further explained in Section 3.2.1. This research utilizes the four data sources and the performance evaluation report for analysis purposes.

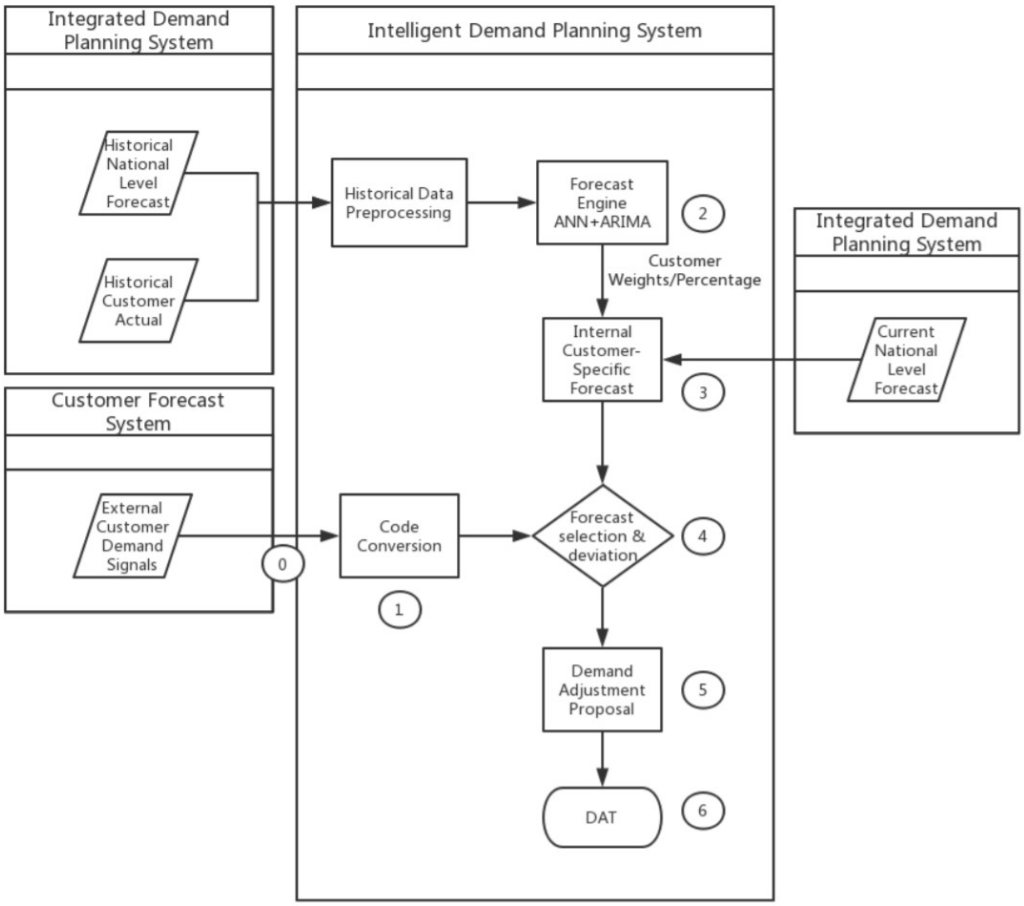

Figure 3. Data flow of the demand-adjustment

3.1.2 Data Overview

All the research data are directly exported from the IDA system data-exportation function and converted to flat data files for use. The datasets include Demand Forecast (supplier P, customer J), Product Master Data, Actual Order Data, and Performance Evaluation. The total datasets contain over 6 million entries, from the project initialization and preparation stage to the time of this writing, which is from January 2019 to April 2020, or about 50 weeks' transaction records. Each entry in the record represents a line item of a demand forecast SKU during a specific week for a specific distribution center (DC). The information included in the flat files is shown in Table 1. Explanations of the data types follow the table.

Table 1. Overview of data collected by types and key fields

Type of Data	Customer J Demand Signals	Supplier P Internal Demand Forecast	Product Master Data	Actual Order Data	Performance Evaluation Report
Key Fields	RPC (Customer J Product Code)	SFU (Supplier P Product Code)	Item Codes	SFU	Item Codes
	Customer J DC Code	Supplier P DC Code	Product Name	Ordered Quantity	Layout Accuracy
	Predict Sales Volume	Predict Demand Volume	Category	Ordered Value	MAPE Accuracy
	Sales Amount	Date	Brand	Date	Date
	Inventory Amount		Product Price		
	Turnover		DC Codes		
	Product Segmentation				
	Date				

Customer J Demand Signals are provided by customer J weekly from their enterprise resource planning (ERP) system and received by supplier P's IDA system through an application programming interface (API) automatically. There is rich information in the customer J forecast file, such as forecast order amount, in-stock inventory quantity, turnover rate, sales amount, and product segmentation. This research is based on the inventory and sales amount provided by this data file to evaluate the impact of different treatment groups. Each file provides the future 13-week forecast by customer J SKU code as the identifiers. On average, each file contains around 1,000 customer J SKUs, and 100,000 records because of the multiple forecasting period and DCs.

Supplier P Internal Demand Forecast is originally provided by supplier P weekly from their advance planning system (APS) model and received by supplier P's intelligent demand-adjustment system. This is the national-level demand forecast information for the coming 13 weeks by supplier P SKU code, which is the traditional default demand forecast processed by demand planners. On average, each file contains around 3,000 supplier P SKUs, and 50,000 records because of the multiple DCs.

Production Master Data includes supplier P and customer J product code mapping relationships, the warehouses and distribution center mapping relationships, and product information such as product brand, category, and names. This data is imported from supplier P's internal master data management system through the API; however, demand planners might modify it to update information. In total it maintains around 4,000 supplier P SKUs. This

translating dictionary conducts the code conversion and smoothly matches codes from customer J and supplier P.

Actual Order Data represent the actual orders that are placed by customer J with supplier P. It includes each supplier P SKU ordered quantity and value for each week by DC. Those data are used as the actual value to calculate the forecast accuracy and train the AI models.

Performance Evaluation Report is generated by IDA automatically through the comparison between forecast data and actual order data. This research leverages this data source to measure demand forecast accuracy. It includes field names such as MAPE formula (1) and layout accuracy formula (2) by SKU, which are the KPI measurements for the models' performance. MAPE, is a traditional forecast accuracy measurement to check the quality of the forecast result and find out how much the forecast is different from the actual results, in a percentage of overall actual demand. The value of MAPE would be 0% if demand forecast is the same as actual demand; the bigger gap between demand forecast and actual demand, the bigger the MAPE value would be.

Layout accuracy is a measurement that is used to measure the matching situation between the actual demand and the demand forecast. Therefore, the value of layout accuracy would be 100% if the actual demand and demand forecast result are the same; larger than 100% if the demand forecast is over actual demand; less than 100% if the demand forecast is less than actual demand.

Those two KPI are defined by supplier P demand planners to fit their measurements' needs; the calculation formulas are shown below. For MAPE accuracy, i = 2, which means taking the most

recent 2 weeks' results as the calculation input, so MAPE can be used as an indicator to check the short-term demand-forecast accuracy. For layout accuracy, i = 13, which means taking the most recent 13 weeks' results as the calculation input, so layout accuracy can be used as an indicator to check the long-term demand forecast accuracy.

$$MAPE = \frac{\sum_{i=1}^{n} \left| \frac{Actual\ Demand_i - Demand\ Forecast_i}{Actual\ Demand_i} \right|}{n} \quad (1)$$

$$Layout\ Accuracy = \sum_{i=1}^{n} \left| \frac{Demand\ Forecast_i}{Actual\ Demand_i} \right| \quad (2)$$

3.1.3 Data Processing

Data Mapping: Because of the different product and warehouse coding systems used by supplier P and customer J, it is necessary to map the flat data files together for the result analysis. This match and mapping are based on the master data file that includes both sides' codes. By joining the customer J demand signals and performance evaluation report, we get all the pretreatment, treatment, and outcome variables within the same table, which is the basis for the effect evaluation.

Data Cleaning and Selection: The overall data quality is good enough to use directly, because the IDA system has already cleaned and converted the raw source data into standard format. Though there are still occasional missing values or mismatched items, the analysis has already removed those records. We also selected data from the beginning (Week 1, 2020) and the end (Week 12, 2020) of this business quarter, because the company reviews their performance for

each business quarter (3 months). Only the SKUs that appear in both the beginning week and the ending week are counted. In total, there are 3,188 records selected into the final analysis.

3.2 Experiment Design and Methodology

In this section, we first describe the overall demand-planning process before and after implementing the IDA system, and then introduce the human-machine teaming interactions and decisions in the adjusted demand-planning process for the experiment. Second, we introduce the empirical methods this research applied, and how the effects of different human-machine teaming decision-making structures were analyzed. Finally, we introduce the setting up of pretreatments, treatments, and outcome variables for the experiment. The overall experiment stages and blueprint is shown in Figure 4. The Figure 4 introduced the experiment design and implementation phases. Before the project implemented, demand planner could only rely on the internal system generated national wide demand forecast, they were not able to deal with the external customer demand signals due to limited efforts. The first stage of the experiment set up the smooth information flow internally and externally (refer to Figure 3). Ample data is available to the demand planners. At stage 2, the project set up the newly designed intelligent demand adjustment process (Figure 6,7,8,9), with embedded ANN and ARIMA AI algorithms, the machine begins to do the demand forecasting tasks. The research experiment designs the empirical analysis of pretreatment variables, and treatment variables, and outcome variables, which are introduced on Section 3.2.3 Treatments by detail.

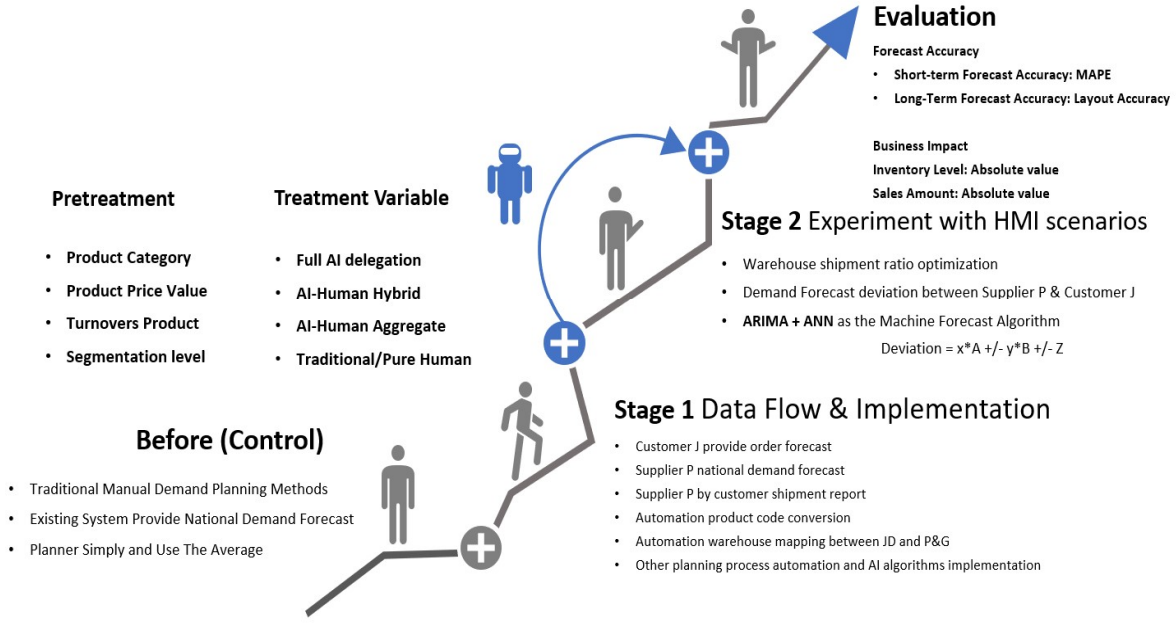

Figure 4. Demand-adjustment human-machine interaction experiment design blueprint

3.2.1 Demand-planning Process

3.2.1.1 Traditional Manual Process

Figure 5 shows the three steps of the human demand-adjusted process before the IDA project implementation, which is the baseline control group demand-adjustment process of this research experiment. In this scenario, it is not possible to take the external specific customer's demand signals, and then adjust the demand forecast by specific customer, geography, and SKU, because of the huge manual effort required. This research takes the results of SKUs with their demand forecast under this traditional manual process as the control group objects.

Description of the traditional process:

Supplier P e-commerce demand planners take only supplier P's internal APS-provided national-level demand-forecast as the demand forecast basis (Step 0).

They convert the internal national level forecast to customer level (Step 2) by the manually calculated (Step 1) average customer sales share based on historical records.

Finally, demand planners aggregate all incoming 13 weeks' adjusted demand forecast, then manually upload them to the detailed assumption tools (DAT) system (Step 3) for further arrangement of production and distribution scheduling.

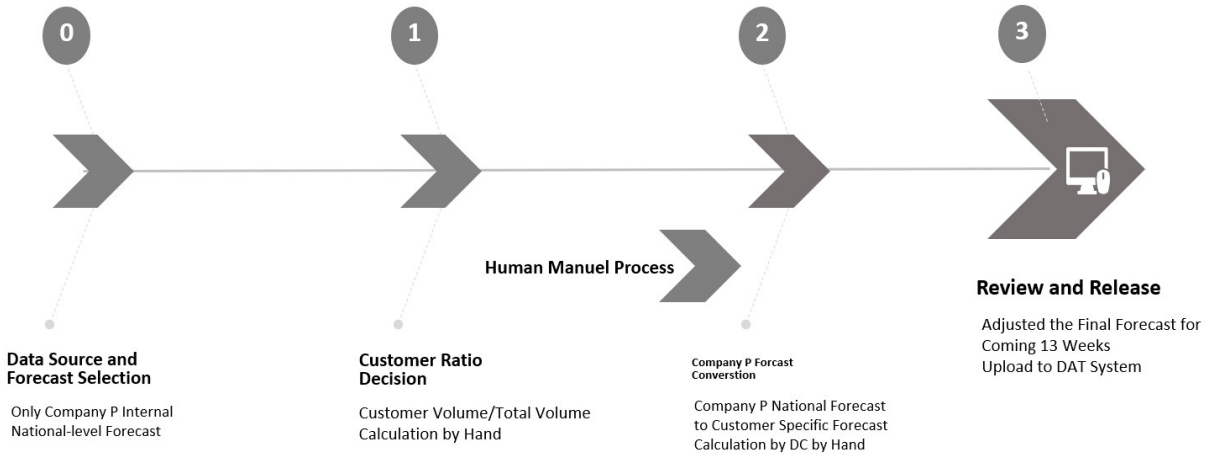

Figure 5. The adjusted demand-planning process for pure human manual process

3.2.1.2 IDA Planning Process – Full AI to Human Delegation

After the IDA system implementation, the demand-adjustment planning steps are increased to 6 steps (with an extra step 0 data selection), shown in Figure 6. The default mode is that demand planners trigger the demand-adjustment process each week, and then all demand-adjustment planning steps could be done automatically by machine with AI algorithms. This research takes

the results of SKUs with a demand forecast under this full AI delegation mode as one of the treatment group objects.

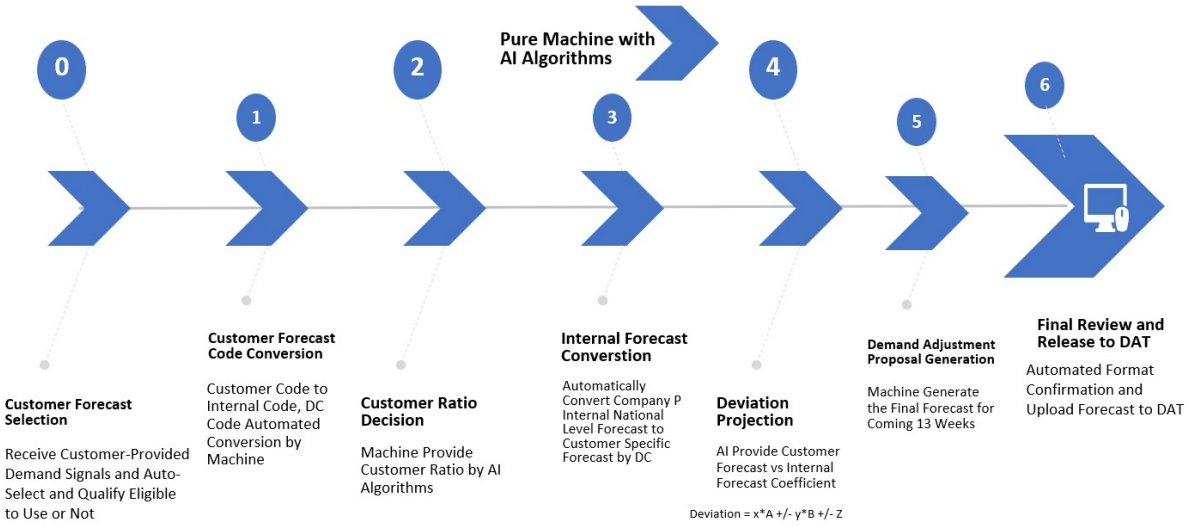

Figure 6. The adjusted demand-planning process by full machine delegation

Description of IDA planning automation process:

Step 0 Customer Forecast Selection: The customer sends incoming 13-week forecast data to IDA. IDA evaluates the forecast according to previous customer J forecast accuracy by layout accuracy formula (2). If layout accuracy is between 50% (under predict) and 150% (over predict), the current week forecast of this SKU record is qualified and selected to the next step of planning. Otherwise (accuracy below 50% or above 150%), the current week forecast of this SKU record will not be used in this round of the IDA forecast process. This selection is done by machine automatically with rules-based algorithms that are defined by demand planners as the default, while the ANN would dynamically adjust the qualified range based on the historical results.

Step 1 Customer Forecast Code Conversion: The machine automatically converts the customer SKU ID (customer J product code RPC) and DC ID to internal demand-planning product ID (supplier P SFU) and DC ID. If there are any missing master data or mismatches, the system will remove them and alert the demand planners.

Step 2 Customer Ratio Decision: The machine calculates the customer demand's percentage of total internal demand by SKU and at the DC level. ARIMA and ANN algorithms take the past actual transactions for the specific customer and the national-level actual transactions as historical data to train the AI model, and then provide the future customer ratio.

Step 3 Internal Forecast Conversion: The machine converts the internal national-level forecast to a customer-specific forecast by multiplying results from Step 2 (customer ratio) and internal national demand forecast. The results are an internal forecast for a specific customer, by SKU at the DC level.

Step 4 Deviation Projection: The machine provides the coefficients of A and B and intercept Z, which decide the weights between the customer forecast and the internal forecast selection for the final forecast, respectively. It is enabled by the ARIMA methods and ANN algorithms that compare a long period of historical results from both the customer side and from internal results. The formula is shown below:

$$Deviation = x * A \pm y * B \pm z \quad (3)$$

A = customer forecast

B = internal forecast

x = coefficient of customer forecast

y = coefficient of internal forecast

z = configurable value

Step 5 Demand-adjustment Proposal Generation: The machine generates formatted, full 13-week demand-adjustment forecast data based on the results of Step 4.

Step 6 Final Review and Release to DAT: In the final format and review, the adjusted demand forecast results are sent to the DAT system by API for further production and distribution scheduling.

3.2.1.3 IDA Planning Process – Machine-Human Hybrid

Based on the implemented IDA planning process, humans can intervene in each step of the default mode described in Section 3.2.1.1. Based on demand planners' knowledge and information, they can make adjustment or change the results that the machine provided in each

step of planning or in specific steps. This research takes the results of SKUs that are under two different levels of interactions between human and machine as the other two treatment groups: Hybrid 1 and Hybrid 2.

Hybrid 1 is the group where demand planners only get involved in the decision-making process of Step 2 and Step 4, shown in Figure 7.

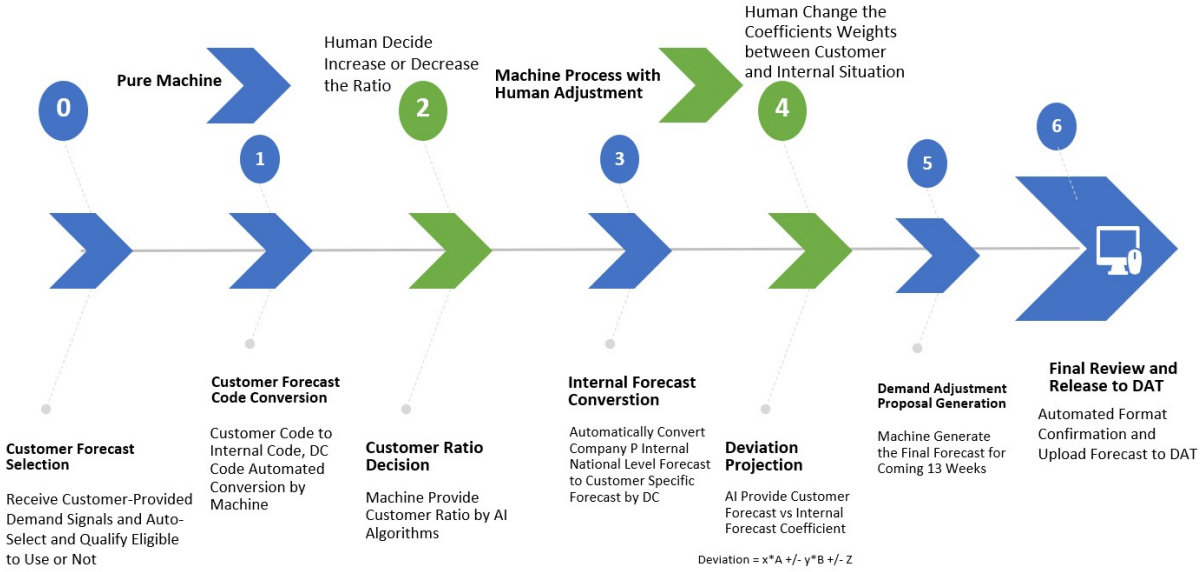

Figure 7. The adjusted demand-planning process by human-machine hybrid (main steps)

Hybrid 2 is the group where, in addition to Step 2 and Step 4, demand planners get involved in the decision-making process of all steps of interaction, as shown in Figure 8.

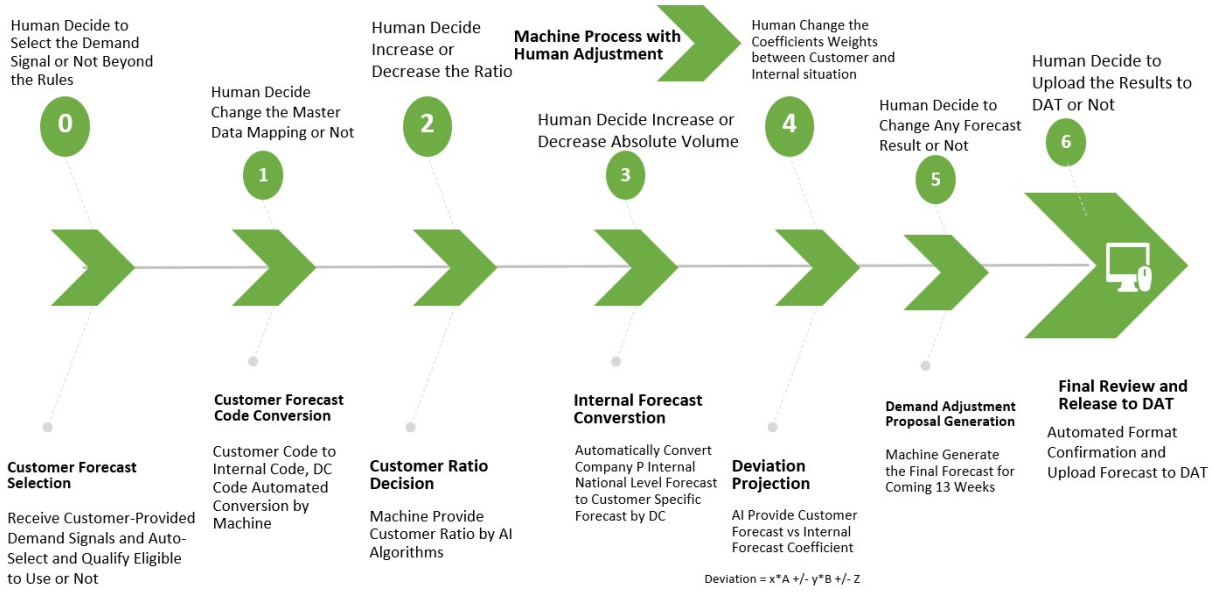

Figure 8. The adjusted demand-planning process by human-machine hybrid (all steps)

The following describes the IDA planning human-engaged process:

Step 0 Customer Forecast Selection: Based on experience and promotion status, demand planners might decide to select or change some SKU demand signal data from what customer J provided, beyond the default rules that limit the accuracy range.

Step 1 Customer Forecast Code Conversion: Demand planners might update/change the master data mapping relations if the master data is out of date. This situation is rare, because most of the master data are updated from the master data management (MDM) system.

Step 2 Customer Ratio Decision: The planners might consider the impact of temporary regional promotion status for this customer or for a specific regional DC, then increase or decrease the target demand.

Step 3 Internal Forecast Conversion: Besides the promotion situation in Step 2, planners might adjust the ratio according to the warehouse operation situation. For example, if a customer or internal warehouse capacity is overloaded, they would decrease the ratio for the overloaded DC. This situation is rare and only happens on a mega-promotion day, once or twice a year.

Step 4 Deviation Projection: Planners balance the customer's needs and supplier P production/distribution situation to adjust the deviation. If there is enough stock, they could increase the customer ratio, and if they are short of inventory, they could increase supplier P ratio.

Step 5 Demand-adjustment Proposal Generation: Demand planners confirm the overall demand forecast for the coming 13 weeks. They might adjust the forecast if any other information they receive from sales, marketing, and business planners indicate that they should decrease or increase the demand forecast for specific SKUs.

Step 6 Final Review and Release to DAT: Demand planners check the data format, then either confirm and release it to the system for demand-adjustment or withdraw this version of the forecast for this week.

3.2.2 Empirical Methodology

This research's empirical objective is to determine the causal effect of different human-AI teaming decision-making structures on forecast accuracy and their business impact. The treatments are different human-AI decision-making models: traditional manual process group as

a control group; full machine delegation group; and AI-Human groups Hybrid 1 and Hybrid 2, respectively. Those treatments and control group are applied at product SKU levels. The outcomes include two parts: forecast accuracy—MAPE accuracy as formula (1), and layout accuracy as formula (2); and business impacts – customer inventory amount and sales volume to end consumers. The SKUs are randomly selected to different treatment groups. All treatments are in binary form. For example, one SKU's demand forecast process could only be decided by one human-AI decision-making structure, so the specific SKU is treated by the Hybrid 1 model, or not by the Hybrid 1 model. To find out the treatment effect, the results of SKUs from different groups are measured through multivalued treatment effects in different treatment status. However, each SKU in the dataset can only have one of the potential outcomes, because only one decision-making structure would be applied to this specific SKU. Nichols (2007) discussed this missing-data problem, and we must estimate other potential treatment parameters to get the causal effect. The structure of the analysis is shown as Figure 9.

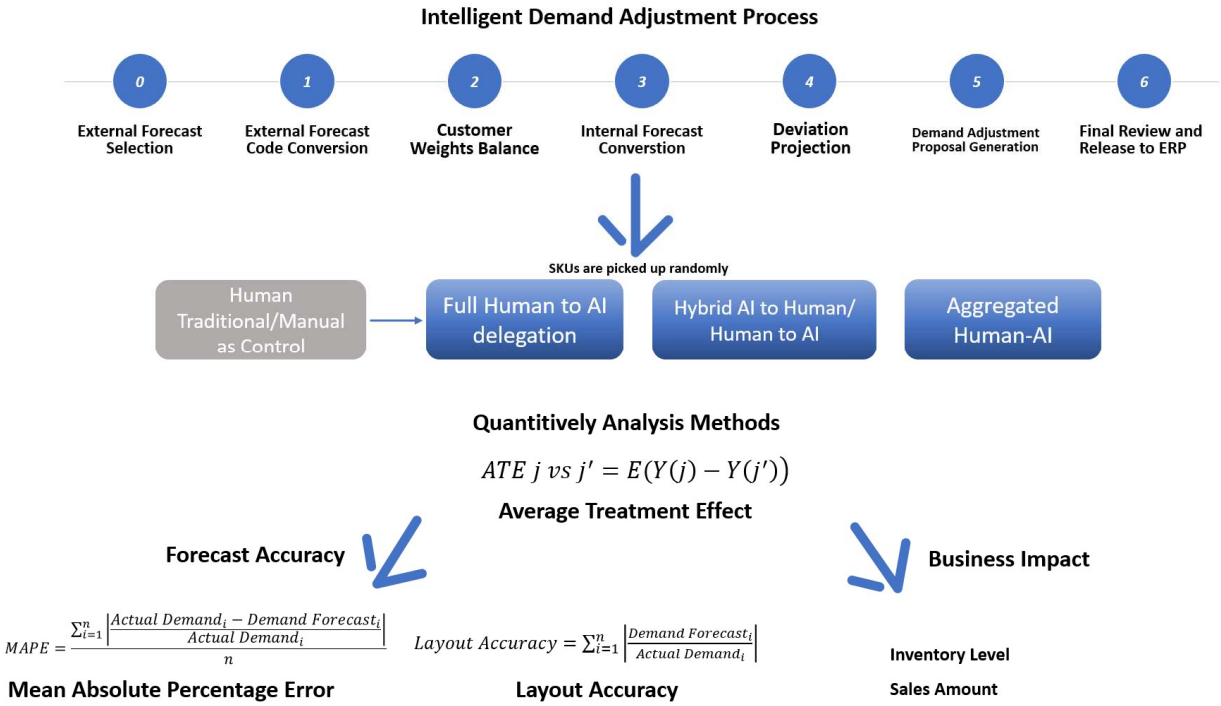

Figure 9. Experiment empirical analysis structure

We follow the analysis methods of Revilla & Rodríguez-Prado (2018) and Cattaneo, Drukker & Holland (2013), by setting up a standard experiment in cross sections, where one of the four possible treatment/control levels would be applied to each specific SKU. The treatment level is represented by j ($j = 1,2,3, \dots J$). For each SKU ($i = 1, 2, 3, \dots n$) in a specific DC, a random vector observation result:

$$z_i = (y_i, w_i, x_i')' \quad (4)$$

y_i = the outcome variables – forecast accuracy, business impacts.

w_i = the treatment level made of types of human-AI decision making structures.

x_i = the $k_x \times 1$ vector of variates, such as the turnover, price, and product segmentations, product categories.

To find out the causal effect, we need to construct a counterfactual model to measure the multivalued treatment effects. Because we only observed the actual outcome y_i, other potential outcomes $y_i(J)$ with different level of treatment j ($j = 1,2,3, \dots J$) are impossible to be observed from experiment. The outcome variable that can be directly observed (observed forecast accuracy, business impacts) could be represented by:

$$y_i = d_i(0)y_i(0) + d_i(1)y_i(1) + \cdots + d_i(J)y_i(J)$$

$\{y_i(1), y_i(2), \ldots, y_i(\mathrm{J})\}$: The independent and equally distributed outcome variables selected from $\{y(1), y(2), \ldots y(J)\}$ for every SKU ($i = 1, 2, 3, \ldots n$) by DC in the sample, where $d_i(j) = 1(w_i = j)$.

d_i = The indicator variables $\mathrm{d_i(j)} = 1(w_i = \mathrm{j})$, the value 1 represents SKU (i) received treatment (j), otherwise the value of $\mathrm{d_i(j)}$ is 0.

It is not possible to find the forecast accuracy and business impacts that the SKU would have had if another human-AI decision making structure were applied. Thus, the individual SKU-level treatment effect is impossible to calculate directly. According to Morgan & Winship (2007), we could determine the aggregated treatment effects. In this case, we want to find out the mean of potential outcomes for every treatment level (POMj) with potential outcome distribution $E(Y(j))$, where $POMj = E(Y(j))$. Average treatment effect (ATE) is the aggregated experiment treatment effect, written as:

$$ATE\ j\ vs\ j' = E(Y(j) - Y(j'))$$

This represent the effects that occur when an SKU receives one treatment instead of another; that is, the average difference in terms of forecast accuracy and business results of applying j type of human-AI decision-making structure instead of j'.

To estimate the ATE, we apply the augmented inverse propensity weight (AIPW) as the estimator in empirical analysis. With AIPW, we only need to decide a regression model to predict the propensity score, then specify a regression model for outcomes. One theoretical advantage of using AIPW is its double robust property (Tan, 2010), which would be consistent for the ATE calculation if either the propensity score estimation model or the outcome regression model is set up correctly.

There are also two assumptions that must be satisfied for the AIPW estimation: conditional independence and common support. Conditional independence could be achieved by adding observed covariates that cover factors that contribute to treatment and outcome results. The common support assumptions require us to have the results of all treatment levels, which is solved by the predicted generalized propensity scores.

By using AIPW, there are three tasks for ATE estimation according to Cattaneo et al. (2013) and Glynn and Quinn (2010) including:

Task 1: Estimate the generalized propensity scores of treatment models by a multinomial logit method, to get the inverse propensity weights.

Task 2: Use a regression model to estimate every treatment j, and calculate the POM of each SKU in each treatment j.

Task 3: Calculate the mean difference of POMs as weighted means (inverse-propensity weights in Task 1) among different treatment groups' predicted outcomes.

We make the analysis based on Stata command group "teffects aipw" to calculate Task 1 through 3. StataCorp. (2019) illustrates the multivalued treatment effects calculation task by task in their reference manual. The sample of Stata codes we use to analyze the results can be found in the Appendix B. The statistical analysis was implemented using a PC Stata package (Stata/IC) on a PC laptop with Microsoft Win 10 operating system, Intel i5-8256 CPU, 8GB RAM.

3.2.3 Treatments

As described in Section 3.2.1, SKUs are randomly selected to different demand-planning processes with different human-AI decision-making structures: traditional/manual group, full AI delegation group, human-AI Hybrid 1 and human-AI Hybrid 2. We formulate one treatment variable that is multivalued for different human-AI decision-making structures in the adjusted demand-planning process. Divergent options were coded with Value 0, Value 1, Value 2, and Value 3:

- Value 0 indicates whether the SKU's demand forecast is under traditional pure manual methods, as shown in Section 3.2.2.1.
- Value 1 indicates whether the SKU's demand forecast is under full AI delegation process, as shown in Section 3.2.2.2.

- Value 2 indicates whether the SKU's demand forecast is under human-AI Hybrid 1 process, in which humans intervene in the main Steps 2 and 4 of demand planning, as shown in Section 3.2.2.3.
- Value 3 indicates whether the SKU's demand forecast is under human-AI Hybrid 2 process, in which humans intervene in the all steps of demand planning from 0 to 6, as shown in Section 3.2.2.4.

As this is a multi-valued treatment experiment, either Value 0, 1, 2, 3 could be treated as the control group to be compared with anyone of the other groups. In this research, we study analyze the average treatment effects in two themes: 1. human-machine teaming structures vs. traditional manual process; 2. comparison among different human-machine teaming structures. The concept of treatment and control varied in the two themes. In the first theme, which is the main theme of the project, Value 0 would be the control group, Value 1, 2, 3 would be treatments compared with Value 0 (1 vs. 0, 2 vs. 0, 3 vs. 0). In the second theme, we study the effects when switching from one of the human-machine teaming decision-making structure to another, so the control group could be Value 1, Value 2, and the treatment group would be Value 2, Value 3 for Value 1 (2 vs. 1, or 3 vs. 1), and Valve 3 for Value 2 (3 vs. 2).

3.2.3.1 Outcome Variables

Four outcome variables are analyzed as the effect measurements of different human-machine decision-making structures applied to SKUs. The first variable is the layout accuracy (2), which is measured by the absolute percentage between forecast quantity and actual quantity during the most recent 13 weeks. This outcome variable stands for the long-term forecast accuracy. The

second variable is the MAPE accuracy (1), which is measured by the percentage difference between forecast quantity and the actual quantity during the most recent 2 weeks. This outcome variable stands for the near short-term forecast accuracy. The third variable is the inventory amount, which is measured by the customer SKU inventory amount by DC. The fourth variable is the sales amount, which is measured by the customer SKU sales volume to end customers by DC.

3.2.3.2 Pretreatments

In AIPW, we need to specify a regression treatment to calculate the estimated generalized propensity score, and then specify a regression outcome model for the conditional mean outcomes of every treatment level. To fulfill the conditional independence assumption, many pretreatment variables are selected to control the potential influence of SKU-specific demand forecast features.

First, we set the turnover rate by assigning a binary variable equal to 1 if the turnover rate is high, and 0 if the turnover rate is low. In the customer's company, turnover of less than 40 days stands for a high-turnover rate, and turnover rate greater than or equal to 40 days is classified as low. The turnover reflects if the product is moving fast or slowly. Second, we set the product price by assigning a binary variable equal to 1 if the product price is high, and 0 if the product price is low. In the customer company, a product purchase price without tax of less than 40 RMB per item stands for a low price for a FMCG product, and a product purchase price without tax that is greater than or equal to 40 RMB per item is classified as high price. Third, we control for the product segmentation based on the SKU, which is defined by the customer, by adding binary

(1,0) variables from segment A to segment F. The product segmentation stands for the importance of the product and how much priority the customer puts on the product. Finally, we control for the product category the SKU belongs to by adding binary (1,0) variables for each category, because different product categories might perform in varied patterns according to different product category management teams and resources.

3.2.3.3 Moderator Variables

The moderator variable is also known as the contextual factor, which is a third variable that might influence the strength of the relationship between the treatment variables and the outcome variables. Some research has shown that the correlation strength between human intervention in statistical forecast model and the forecast performance would be influenced by the moving speed of products (Syntetos et al., 2009). Therefore, we choose the turnover as one of the moderator variables, to study the performance of different treatments under slow or fast product moving speed.

4 RESULTS AND ANALYSIS

In this chapter, the result of all the treatment and control groups are presented and compared. The first part shows the statistical descriptions of all the variables. The second part illustrates the results of the average treatment effects for all treatment and control groups. The results indicate that all the human-machine teaming treatments improved the forecast accuracy and reduced the inventory level. The third part compares the treatment effects between the high-turnover group and the low-turnover group.

4.1 Statistics Summary

Table 2 shows the size of the experiment samples and descriptive statistics of all variables in the experiment. In total, we have 3,188 records by SKU by DC in the analysis. In the sample data,

40% (1,281) of the SKUs are in the traditional process as the control group treatment; 0.27% (873) of the SKUs are in the full AI delegation process as the Treatment 1 group; 23% (746) of the SKUs are in the human-machine Hybrid 1 process as the Treatment 2 group; and 9% (288) of the SKUs are in the human-machine Hybrid 2 process as the Treatment 3 group. The average sales amount per SKU is 85 units per week per region with a high standard deviation (SD) (294); the average inventory is 839 units per DC with a high SD (1796); The turnover day is 71 days on average, while the mean price of the product is 58 with a relatively small SD(61), very close to half product high price and the other half product low price.

Table 2. Sample Size and Descriptive Statistics

			N
Total Observation			3188
Control Groups			
0. traditional Manual			1281
Treatment Groups			
1. full machine AI automation			873
2. human-machine hybrid in main steps 2,4			746
3. human-machine hybrid all steps			288
	Mean	Std. Dev.	Obs
Outcome Variables			
Sales amount	85.65	294.61	3188
Inventory amount	839.00	1796.12	3188
Layout accuracy	1.99	2.57	3188
Mape accuracy	1.10	2.75	3188
Pretreatment Variables			
turnover	71.72	161.99	3188
turnover_dummy	0.48	0.50	3188
Basic piece price tx	58.03	61.55	3188
price_dummy	0.47	0.50	3188
product segmentation			
segmentation_a	0.02	0.14	3188

segmentation_b	0.03	0.18	3188
segmentation_c	0.06	0.24	3188
segmentation_d	0.13	0.34	3188
segmentation_e	0.32	0.47	3188
segmentation_f	0.44	0.50	3188

First, we observe that the more human intervention is involved in the decision-making structures, the fewer records there are—a much smaller proportion of the Hybrid 2 model compared with the Hybrid 1 or full AI delegation model. Second, we can see that all the outcome variables have very high standard deviation. Third, the pretreatment variables of price and turnover are evenly divided around 1:1, while the segmentation distribution is increasing and showing the long tail effect.

4.2 Overall Average Treatment Effect Estimates

Table 3 reflects the potential-outcome mean (POM) for all level of treatments (j) and comparisons between them (ATE). The POMs include four outcome variables: layout accuracy, MAPE, inventory amount, and sales amount. The treatment levels js include: 0 traditional manual; 1 full machine delegation; 2 human-machine Hybrid 1 on main steps of demand planning; and 3 human-machine Hybrid 2 on all steps of demand planning. The ATEs among those treatment groups are the comparison results among samples-averaged treatment effects, which are the results under one treatment instead of another, and a comparison of the POM ratio between them None of the results of the Outcome 4 sales amount are significant, so the below discussion excludes this variable. The Figure 9 shows the ATE among treatment groups.

The research is interested in comparisons of the ATEs among different human-machine decision-making structures in the demand-planning process to find out the impacts of human and machine intervention, rather than using single structures. As mentioned in section 3.2.3.2 Treatments and Controls, we select Treatment 0, 1, 2 as the control groups to be compared with anyone of the other groups (Treatment 1, 2, 3). The ATE results are presented in two themes: 1. The upper part of results: human-machine teaming structures vs. traditional manual process; 2. The lower part of results: comparison among different human-machine teaming structures. In the first theme, Treatment 0 would be the control group, Treatment 1, 2, 3 would be treatments compared with Treatment 0 (1 vs. 0, 2 vs. 0, 3 vs. 0). In the second theme, we study the effects when switching from one of the human-machine teaming decision-making structure to another, so the control group could be Treatment 1 or 2, and the treatment group would be Treatment 2 or 3 for Treatment 1 (2 vs. 1, or 3 vs. 1), and Treatment 3 for Treatment 2 (3 vs. 2).

Table 3. Average Treatment Effects Estimates.

	Outcome 1 Layout Accuracy				Outcome 2 MAPE Accuracy		
Human-Machine Decision Making Structures	Potential Mean		s.e		Potential Mean		s.e
0 Traditional Manual	3.04		0.08		1.65		0.09
1 Full Machine-AI Delegation	1.44		0.07		0.66		0.04
2 Human-Machine Hybrid 1 on Main Steps	1.44		0.09		0.60		0.07
3 Human-Machine Hybrid 2 on All Steps	1.35		0.11		0.82		0.09
	Average Treatment Effect	Significance	s.e		Average Treatment Effect	Significance	s.e
1 vs 0	-0.53	***	0.03		-0.60	***	0.03
2 vs 0	-0.53	***	0.03		-0.64	***	0.04
3 vs 0	-0.56	***	0.04		-0.50	***	0.06
2 vs 1	-0.003		0.08		-0.10		0.11
3 vs 1	-0.06		0.09		0.24		0.16
3 vs 2	-0.06		0.95		0.04	*	0.22

	Outcome 3 Inventory Amount				Outcome 4 Sales Amount		
Human-Machine Decision Making Structures	Potential Mean		s.e		Potential Mean		s.e
0 Traditional Manual	1276.95		64.12		81.91		10.51
1 Full Machine-AI Delegation	511.32		25.57		72.95		4.30
2 Human-Machine Hybrid 1 on Main Steps	381.36		33.72		88.01		9.51
3 Human-Machine Hybrid 2 on All Steps	680.00		61.01		82.74		6.13
	Average Treatment Effect	Significance	s.e		Average Treatment Effect	Significance	s.e
1 vs 0	-0.60	***	0.03		-0.11		0.12
2 vs 0	-0.70	***	0.03		0.07		0.18
3 vs 0	-0.47	***	0.05		0.01		0.15
2 vs 1	-0.25	***	0.07		0.21		0.15
3 vs 1	0.33	**	0.14		0.13		0.11
3 vs 2	0.78	***	0.22		-0.06		0.12

AIPW estimators controlling for products' difference in price, turnover, and segmentation category.
s.e: robust standard errors.
*Significant at 10%.
**Significant at 5%.
***Significant at 1%.

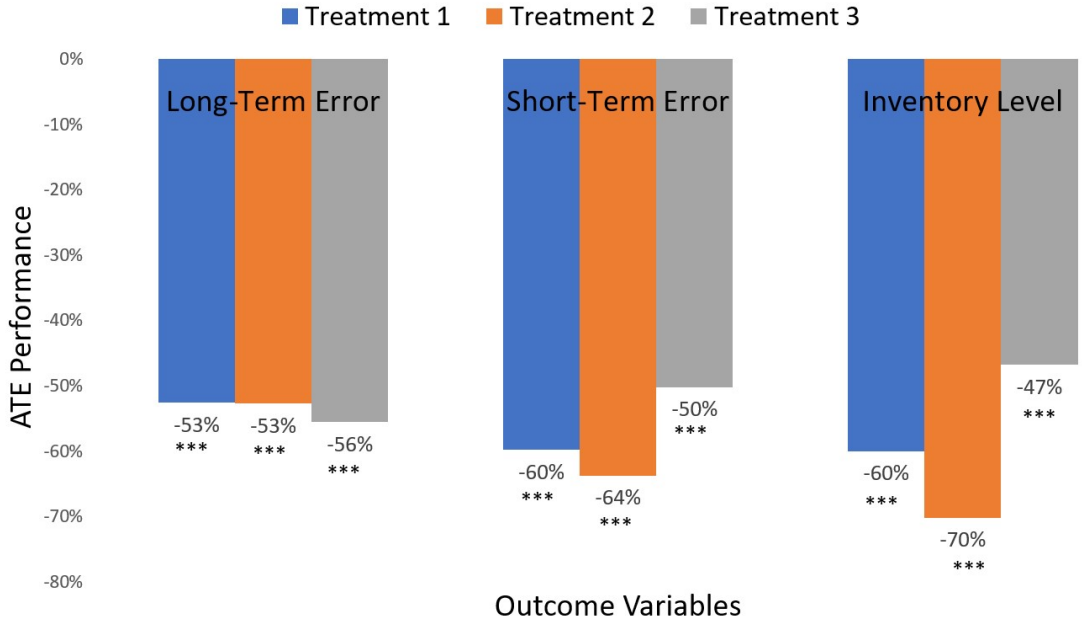

Figure 10. The overall ATEs comparison among treatments

4.2.1 Theme 1 Human-Machine Teaming vs. Traditional Manuel

In this main theme of the research, we are interested in the comparison between the control group (Treatment 0) and the Treatment groups (Treatment 1, 2, 3), as these results would reflect whether the IDA process with human-machine interaction improves the demand-planning process or not. Results from the upper part of Table 3 and Figure 10 indicate that, once applied,

the new IDA demand-planning process, in terms of both forecast accuracy and business results, has been improved significantly on all treatment decision-making structures (j=1,2,3) compared with the control group (j=0). More precisely, the product SKUs' layout accuracy was 53 percentage points more accurate when the SKUs are under a full machine delegation decision-making structure in the demand-planning process, instead of the traditional/manual demand-planning process. In other words, ATE is negative with the forecast errors, and rejects the null effects statistically [Outcome 1: (1 vs 0): ATE -0.53, $p < 0.01$]. Similar results were obtained when we looked at Treatment 2 [Outcome 1: (2 vs 0): ATE -0.53, $p < 0.01$] and Treatment 3 [Outcome 1: (3 vs 0): ATE -0.56, $p < 0.01$].

Except for Outcome 4, we could find a similar result from outcome variable 2 MAPE accuracy and outcome variable 3 inventory amount, that all redesigned human-machine engaged decision-making structures compared with traditional pure human manual process, improved the demand-planning results, both the forecast accuracies and the business results. In other words, the ATE of treatment groups 1, 2, 3 of all Outcomes 1, 2, 3 are less than 0 (improved), and $p < 0.01$, which is statistically significant.

Thus, with respect to Outcomes 1, 2, 3 shown above, we answer the first research question: the human-machine teaming decision-making structures improve demand forecast accuracy in all engagement levels (Full machine delegation, Hybrid 1 and Hybrid 2) compared with the traditional process.

4.2.2 Theme 2 Comparison of Human-Machine Teaming Groups

In this theme, there are the comparisons among treatment groups (Treatments 1, 2, 3). These results illustrate how different levels of human intervention in the AI-machine automation demand-planning process change the demand forecast results. Results for the lower part of ATE indicate that different levels of human intervention in the IDA process only influence the short-term forecast accuracy if humans are engaged in all steps of the decisions; the results are worse by 4 percentage points compared with Hybrid 1 group [Outcome 2: (3 vs. 2): ATE 0.04, $p < 0.1$]. All other forecast accuracy groups (Outcomes 1, 2) are not influenced by human engagement.

In contrast, the inventory amount is significantly influenced by human engagement in the planning process. From Outcome 3 we see that if moderate human intervention is added (Treatment 2, Hybrid 1 group), the inventory amount is further improved by 25% [Outcome 2: (2 vs. 1): ATE -0.25, $p < 0.01$]. However, if humans engage in all steps of planning (Treatment 3, Hybrid 2 group), the inventory amount increase is statistically significant compared to both the full AI delegation group (+33%) and the Hybrid 1 group (+78%). Compare this to [Outcome 2: (3 vs, 1): ATE 0.33, $p < 0.05$] and [Outcome 2: (3 vs. 2): ATE 0.78, $p < 0.01$].

Overall, according to Outcomes 1, 2, 3 shown above, we answer the second research question: The hybrid human-machine teaming decision-making structure with appropriate human intervention provides a better approach for demand forecasting and business results.

4.3 Treatment Effect Estimates by Moderator

Table 4 reflects the results that take the contextual factor into account, in this case, turnover level. The mean equality is under a two-sided test, to show if the high-turnover product is statistically different with the low-turnover product in the experiment. The results reflect that the POMs of the high-turnover group and the low-turnover group are statistically significantly different in regard to MAPE accuracy, inventory amount, and sales amount. The low-turnover products tend to have more inventory amount, lower sales amount, and better MAPE accuracy compared to the high-turnover products.

Table 4. Treatment Effect Estimates by Turnover

	Outcome 1 Layout Accuracy						
	High Turnover Products				Low Turnover Products		
Human-Machine Decision Making Structures	Potential Mean		s.e	Significance of two-side test equality	Potential Mean		s.e
0 Traditional Manual	2.80		0.11		3.28		0.13
1 Full Machine-AI Delegation	1.08		0.07		1.79		0.12
2 Human-Machine Hybrid 1 on Main Steps	1.06		0.07		1.80		0.16
3 Human-Machine Hybrid 2 on All Steps	0.92		0.15		1.69		0.15
	Average Treatment Effect	Significance	s.e		Average Treatment Effect	Significance	s.e
1 vs 0	-0.61	***	0.03		-0.46	***	0.04
2 vs 0	-0.62	***	0.03		-0.45	***	0.05
3 vs 0	-0.67	***	0.06		-0.49	***	0.05
2 vs 1	-0.02		0.09		0.01		0.11
3 vs 1	-0.15		0.15		-0.55		0.11
3 vs 2	-0.13		0.15		-0.06		0.12

	Outcome 2 MAPE Accuracy						
	High Turnover Products				Low Turnover Products		
Human-Machine Decision Making Structures	Potential Mean		s.e	Significance of two-side test equality	Potential Mean		s.e
0 Traditional Manual	2.37		0.16	***	0.96		0.07
1 Full Machine-AI Delegation	0.95		0.08	***	0.39		0.03
2 Human-Machine Hybrid 1 on Main Steps	1.07		0.14	***	0.22		0.04
3 Human-Machine Hybrid 2 on All Steps	1.31		0.19	***	0.36		0.05
	Average Treatment Effect	Significance	s.e		Average Treatment Effect	Significance	s.e
1 vs 0	-0.60	***	0.04		-0.59	***	0.04
2 vs 0	-0.55	***	0.06	***	-0.77	***	0.04
3 vs 0	-0.45	***	0.09	*	-0.63	***	0.06
2 vs 1	0.12		0.17	***	-0.44	***	0.10
3 vs 1	0.37		0.22	*	-0.08		0.14
3 vs 2	0.22		0.24		0.63	*	0.34

	Outcome 3 Inventory Amount						
	High Turnover Products				Low Turnover Products		
Human-Machine Decision Making Structures	Potential Mean		s.e	Significance of two-side test equality	Potential Mean		s.e
0 Traditional Manual	1075.44		76.58	***	1453.16		98.74
1 Full Machine-AI Delegation	480.62		37.06	*	535.42		33.30
2 Human-Machine Hybrid 1 on Main Steps	309.12		41.60	**	449.40		52.29
3 Human-Machine Hybrid 2 on All Steps	552.34		70.42	*	762.80		88.08
	Average Treatment Effect	Significance	s.e		Average Treatment Effect	Significance	s.e
1 vs 0	-0.55	***	0.05		-0.63	***	0.03
2 vs 0	-0.71	***	0.04		-0.69	***	0.04
3 vs 0	-0.49	***	0.07		-0.48	***	0.07
2 vs 1	-0.36	***	0.10		-0.16	**	0.11
3 vs 1	0.15		0.17		0.42		0.19
3 vs 2	0.79	**	0.33		0.67	**	0.28

	Outcome 4 Sales Amount						
	High Turnover Products				Low Turnover Products		
Human-Machine Decision Making Structures	Potential Mean		s.e	Significance of two-side test equality	Potential Mean		s.e
0 Traditional Manual	111.49		18.32	***	52.65		10.35
1 Full Machine-AI Delegation	88.81		6.55	***	57.92		5.52
2 Human-Machine Hybrid 1 on Main Steps	112.15		16.51	**	65.14		10.03
3 Human-Machine Hybrid 2 on All Steps	101.77		9.49	***	64.64		7.14
	Average Treatment Effect	Significance	s.e		Average Treatment Effect	Significance	s.e
1 vs 0	-0.20		0.14		0.10		0.24
2 vs 0	0.01		0.22		0.24		0.31
3 vs 0	-0.09		0.17		0.23		0.27
2 vs 1	0.26		0.21		0.12		0.20
3 vs 1	0.15		0.13		0.12		0.16
3 vs 2	-0.09		0.16		-0.01		0.19

AIPW estimators controlling for products' difference in price, turnover, and segmentation category.
s.e: robust standard errors.
*Significant at 10%.
**Significant at 5%.
***Significant at 1%.

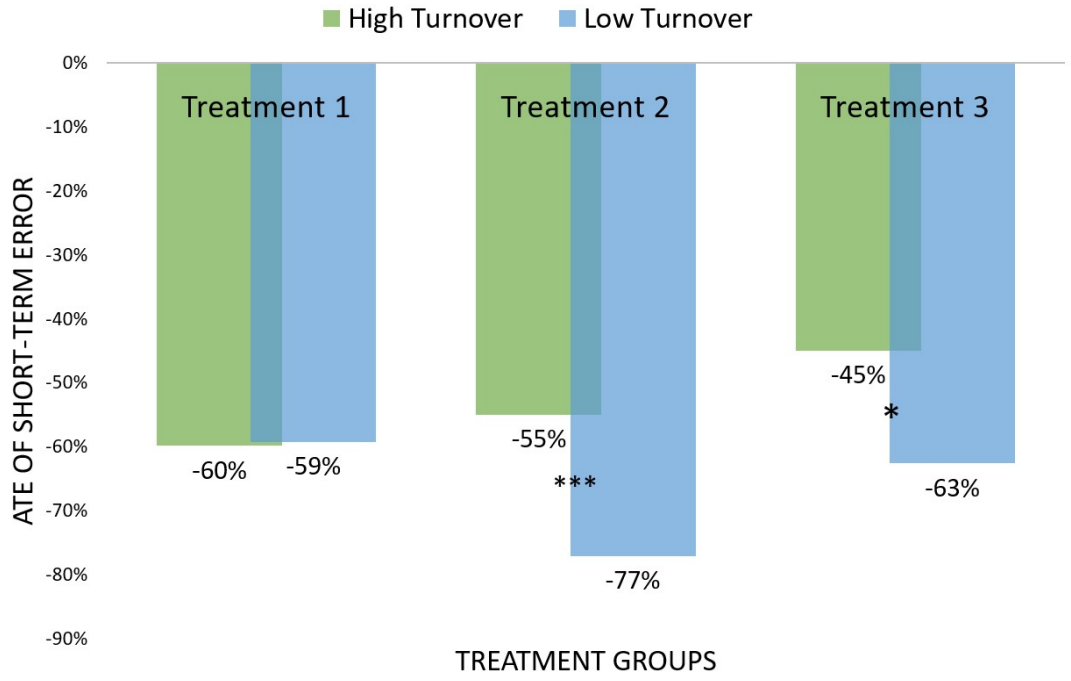

Figure 11. The ATEs of short-term forecast comparison among treatments by turnover

The average treatment effect (ATE) is statistically significantly different in regard to treatment variables in the Outcome 2 MAPE accuracy group. The ATE for the low-turnover product, human-machine decision-making structure engaged with human interventions is stronger than that of high-turnover groups, although the full AI delegation group has the same ATE between high and low-turnover groups. The ATE of MAPE accuracy improvement is increased by 22 percentage points in Hybrid 1 group [Outcome 2 high-turnover: (2 vs. 0): ATE -0.55, $p < 0.01$] compared with [Outcome 2 low-turnover: (2 vs. 0): ATE -0.77, $p < 0.01$], and 18 percentage points in Hybrid 2 group [Outcome 2 high-turnover: (2 vs. 0): ATE -0.45, $p < 0.01$] compared with [Outcome 2 low-turnover: (2 vs. 0): ATE -0.63, $p < 0.1$]. Other comparisons are not fully significant, so they are not considered in Section 5. Figure 11 summarized the ATE of three treatments compared between the high-turnover and low-turnover products in terms of MAPE, the short-term accuracy.

5 DISCUSSION

The human-machine teaming decision-making structures in demand-planning processes are very important in the success of supply chain digitalization in an e-commerce business environment. To help fully realize the efficiency and effectiveness of potential human-machine teaming in demand-planning capabilities, this research examines whether and how different human-machine decision-making structures improve the demand forecast accuracy and inventory level in the demand-planning process. Further, this research determines which of the structures would provide an optimal approach for demand forecast accuracy and inventory level through a random experiment and empirical analysis of treatment effects. The experiment set up four groups of SKUs randomly under those different decision-making structures: Control 0: traditional manual process; Treatment 1: full AI delegation; Treatment 2: human-machine hybrid decision-making structure in the main demand-planning steps; and Treatment 3: human-machine hybrid decision-making structure in all demand-planning steps. In this section, we discuss the results, provide insights drawn from the four key findings, and offer managerial implications for practitioners.

5.1 Results Discussion

Firstly, our findings show that adopting all human-machine teaming decision-making structures in the demand-planning process significantly improves both the forecast accuracy and the inventory level, compared with the pure traditional manual control group.

The results (Treatment 1, 2, and 3 vs. 0, Table 3 and Figure 10) show that all treatment groups significantly improve their long-term forecast accuracy (layout accuracy 55-60% error reduced), short-term forecast accuracy (MAPE 50-64% error reduced), and business results (inventory level 47-60% inventory stock reduced). The only difference between the treatment groups and control group is the introduction of an AI-enabled decision-making structure (IDA system) as the demand-adjustment process. This finding answers our first research question: According to the results of this experiment, human-machine decision-making structures do improve both the demand forecast accuracy and the inventory level.

There are two potential explanations for why the treatment improves the outcomes: the treatment enabled effective information sharing, and it provided new AI forecast capabilities. Comparing the control group's demand-adjustment process, the treatment groups' process receives external demand signals and makes adjustments accordingly, with or without human intervention. However, in the control group, all the procedures are done manually by demand planners, and it is not possible for the demand planners to consider thousands of external demand signals from customers. Therefore, the addition of information sharing between customer and supplier, and a redesigned demand-adjustment process leads to much better demand forecast accuracy and inventory level. Our study provides further evidence that demand information sharing by online retailers would reduce the bullwhip effect and the suppliers' inventory level (Zhao et al., 2018). Additionally, through the comparison between the full AI delegation group and the control group, the AI algorithm shows its strong demand forecast capability to achieve improvements in both long-term and short-term accuracy. As expected, the machine learning ANN model combined with statistical technique ARIMA has a strong performance in forecasting, which has

also been demonstrated in other research in many industries (Vhatkar and Dias, 2016; Adebiyi et al., 2014; Bhadouria and Jayant, 2017; Hiranya Pemathilake et al., 2018).

Secondly, the results show that average treatment effects (ATEs) of inventory level are varied in different human-machine teaming decision-making structures, which depend on the level of human intervention and which steps of demand planning humans get involved in.

The ATE of inventory level performed very differently from demand forecast accuracy. There is an obvious pattern that the adequate human intervention (Treatment 2) in the process would improve the inventory level (-25%) compared to the full AI delegation group (Treatment 1). In Treatment 2, demand planners would adjust the AI-provided results according to their expertise and updated promotion information or warehouse information, in two main steps. This meets the expectation that humans and machines can be partnered, and that human expertise and machines' robust capabilities will contribute to the performance (Saenz, Revilla & Simon, 2020). Similar results were shown in a previous study, which found that human revision of the process of selecting the right combination of forecasts by human experts and forecasts by statistical model would improve inventory efficiency (Wang & Petropoulos, 2016).

However, if there are too many human overrides (Treatment 3), it would drastically reduce the human-machine teaming advantages, leading to an increase in the forecast error of +78% with Treatment 2 and +33% with Treatment 1.Treatment 3 means beyond the main steps, demand planners would engage in all other steps of the demand-adjustment processes as well, such as manual demand signal selection, final proposal adjustment, and code conversion. Research

shows that when many unrequired judgmental adjustments are made, the forecasting result would be sub-optimal, which is the case we faced in this scenario (Goodwin et al., 2011).

On the other hand, long-term forecast accuracy ATEs are almost the same among Treatments 1, 2, and 3 vs. 0, and the ATEs of switching among treatment groups are insignificant. Short-term forecast accuracy has a similar situation, except one outlier in the low-turnover products, which is described in the moderating analysis, as shown in Table 4.

Thirdly, the findings also illustrate that, overall (according to Table 3), the optimal human-machine teaming decision-making structure for different outcome variables varies, but Treatment 2 performs very well for all outcome variables. For long-term forecast accuracy, the ATE of Treatment 3 is the highest, though not significantly different from the other two treatment groups, and the gap is small (3%). For both the short-term forecast accuracy and the inventory level, Treatment 2 is the optimal decision-making structure, and it is significantly better than other groups in the inventory level improvement. Therefore, this addresses our second research question: Overall, the hybrid human-machine decision-making structure with adequate human engagement in the main steps of the demand-adjustment process is the optimal model. Similar rationales are discussed in the second discussion point, above.

Finally, the moderating analysis (according to Table 4 and Figure 11) on the impacts of product turnover provides insights into how to better leverage human-machine teaming forecasting capabilities according to product properties.

The short-term forecast accuracy is statistically different between high-turnover products and low-turnover products, both in potential-outcome means (POMs) and the ATEs. The low-turnover products tend to have a much better short-term forecast accuracy of POMs and ATEs than the high-turnover ones in all control and treatment groups. This means the human-machine teaming decision-making structures work better to improve the short-term forecast for low-turnover products. And as mentioned in discussion point 2, the only exception of forecast accuracy among treatment groups occurs when switching treatments in the low-turnover products from Treatment 1 to Treatment 2, which would result in a huge improvement in short-term accuracy (44%). The low-turnover product forecast would achieve better accuracy if humans engaged in the human-machine decision-making structure only in the main steps of the demand-adjustment process. Both the above points match the previous research, which shows that forecast accuracy performs better in slow-moving category products if human judgmental adjustment is added to the statistical forecast model (Syntetos et al., 2009). The reason might be that demand planners have most update information about the promotion activities in the near future to handle the tail goods, while this is an information asymmetry to AI model because the current statistical model ARIMA and ANN only trains the prediction model by historical data, which could not capture the temporary fluctuation. On the other hand, the long-term forecast accuracy is not influenced by this contextual factor, which further indicate the above hypothesis.

We can also naturally find out that the high-turnover products have statistically significant lower inventory levels and higher sales amounts compared with the low-turnover products in the form of POMs. However, the long-term forecast accuracy for the high-turnover products is not significantly different than the forecast accuracy for low-turnover products. Performance of other

treatment groups in the moderate analysis follows the same trend as the above first three discussion points.

5.2 Managerial Recommendations and Practical Implications

Supply chain and IT managers should be aware of the relationship between humans and machines in the age of supply chain digitalization and AI application, especially in regard to the demand-planning decision-making structure. Information sharing would help the supply chain smooth the demand fluctuation and alleviate the bullwhip effect, especially in an e-commerce business environment. Traditional pure human manual process could not meet the collaboration requirements with high-volume interchange of data. Our research shows significant improvement in long-term and short-term demand forecast accuracy and inventory level, by switching the demand-planning process from a pure manual traditional group to any of the human-machine teaming decision-making structure models: Full AI delegation, Hybrid AI-Human in the main demand-planning steps, and Hybrid AI-Human in all demand-planning steps. The new human-machine teaming structure would help the organization incorporate external demand signals into the existing internal demand forecast efficiently. This supply chain digitalization would help the demand-planning process evolve to a much more intelligent phase, while humans could partner with machines to deliver better inventory-level results based on their expertise, especially for low-turnover, slow-moving products. The demand planner should focus on the slow-moving products to maximize the performance of the human-machine teaming capacities, which would also help handle the long tail products.

When designing, implementing, and operating digital supply chain demand-planning using AI capabilities, a flexible interface between humans and machines on each key step of demand planning would keep the AI-enabled demand-planning process more robust, open, and interoperable. Neither pure AI delegation nor heavy human intervention would bring optimal results to the demand-planning and forecasting process. On the one hand, the balance between AI and human interaction should be carefully calibrated. The machine-AI approach is good at data qualification, selection, modeling process, conversions, and trend-pattern prediction. On the other hand, demand planners monitor the near-future changes and any emergencies not defined in the model or history. The optimal engagement level that demand planners should apply when the company implement AI-enabled demand-planning process, which is only focus on the main steps revision, such as demand forecasts weights selection from different source and/or warehouse weights adjustment based on updated warehouse capacity, too many overrides or breaking the system's predefined rules during the demand-planning process might be detrimental to the human-machine teaming capabilities and business results.

6 CONCLUSION

6.1 Limitations

While the results of the experiment are conclusive, it is important to note the limitations of this research. First, the project focuses on one large-scale FMCG company with one of their leading e-commerce customers. A broader experiment with more data and industries might be needed to further generalize or expand the results across other industries or companies. Additionally, the generalizability might also be limited in companies without as much IT or supply chain resources as the project companies have available. Second, because all data comes from Supplier P's system directly, there are not enough information or data to describe the demand planners' features or other product-related data. Additionally, because of the limitation of the current multivalued effect-estimation tools, the endogenous effects could not be ruled out, and we could only assume that all the pretreatment variables are independent. The current research only focuses on the intervention between humans and machines, and the machines only train their models from the historical data. The potential effects that may result when humans and machines learn from each other might be further developed. Due to the scope and resources, two additional human-machine decision-making structures were not addressed in this research: the aggregated human-machine model and the hybrid human-to-AI model, which can be further studied.

6.2 Contribution

In response to calls in the literature about the implications of organization performance when different decision-making structures are applied (Shrestha & Ben-Menahem, 2019), this research reveals different human-machine AI decision-making structures' impacts and insights for demand-planning, by empirical analysis that quantifies the average treatment effects of different models on forecast accuracy and inventory level. This research also provides support and evidence about the forecast selection based on forecast variability (Wang & Petropoulos, 2016), and judgmental adjustment improvement (Syntetos et al., 2009) with a human-machine teaming decision-making structure and studied its impacts on forecast accuracy and inventory level. It also provides a demand-planning process framework to intake external customer-provided demand signals and incorporate them into existing internal systems, which is a new model for supply chain digitalization and customer collaboration.

6.3 Conclusion

This research is carried out based on an intelligent demand-adjustment system of a large FMCG company and one of its key e-commerce customers. Based on their IT collaboration program, the research wanted to learn if and how the different human-machine teaming decision-making structures would influence the demand forecast accuracy and customer inventory level. The three human-machine teaming decision-making structures are: Full AI delegation model, Hybrid AI-Human in the main demand-planning steps model, and Hybrid AI-Human in all demand-planning steps model. We conducted an empirical study by using an augmented inverse propensity weighted estimator to find out the treatment effects of the different decision-making structures on demand forecast accuracy and inventory level.

The results of this study show that after the implementation of human-machine teaming decision-making structures, both demand forecast accuracy and inventory level are strongly improved by at least 47% compared to the traditional manual process. Based on the significant improvement, it is a great opportunity for company to implement a human-machine teaming decision making structure to improve their demand forecast accuracy and customer inventory level, if the demand-planning and adjustment tasks of the company are still by traditional pure manual process. The Hybrid AI-Human with adequate human intervention model is the optimal decision-making structures between human and machine, which improves the short-term forecast accuracy by 53%, long-term forecast accuracy by 64%, and inventory level by 70%. The Hybrid AI-Human with all steps overrides model performed worse than the previous model, because of the heavy human overrides. This guides the optimal engagement level that demand planners should apply when the company implement AI-enabled demand-planning process, which is only

focus on the main steps of planning revision, such as demand forecasts weights selection from different source and/or warehouse weights adjustment based on updated warehouse capacity, further overrides in all steps of demand planning lead to diminishing of forecast accuracy and inventory efficiency improvement. Additionally, for low-turnover products, the AI-enabled decision-making structure with adequate human revision works better than for the high-turnover ones, in terms of the short-term forecast accuracy (77% vs. 55%). This result clearly indicates that, the demand planner should focus on the slow-moving products to maximize the performance of the human-machine teaming capacities, which would also help handle the long tail products.

This research shows the strong power of human-machine teaming decision-making structures in demand-planning domain, and it opens further research questions and leads to next phase. For example, besides the decision-making structures tested in this paper, other two models are worth testing: hybrid human-to-AI and aggregated human-AI decision making structures. How to measure and assess human performance when the decision is made by human-machine teaming model? And who should own the responsibilities for the business results, data scientists, IT managers or business function leaders? Solving these and further questions would make human live better and work happier in the second machine age.

REFERENCES

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction. *Journal of Applied Mathematics, 2014,* 1–7.

Barnes, J., and Y. Liao. (2012). "The Effect of Individual, Network, and Collaborative Competencies on the Supply Chain Management System." *International Journal of Production Economics 140 (2):* 888–899.

Barngetuny, D. C., & Kimutai, G. (2015). Effects of e-procurement on supply chain management performance in Elgeyo-Marakwet County. *International Academic Journal of Procurement and Supply Chain Management | Volume 1, Issue 5,* pp. 99-120

Bhadouria, S., & Jayant, A. (2017). Development of ANN Models for Demand Forecasting. *American Journal of Engineering Research,* 6.

Brynjolfsson E, McAfee A. (2014). *The Second Machine Age : Work, Progress, and Prosperity in a Time of Brilliant Technologies. New York : W. W. Norton & Company,* [2014];

Cattaneo, M.D., Drukker, D.M., Holland, A.D., (2013). Estimation of multivalued treatment effects under conditional independence. *Stata J. 13 (3),* 407–450.

China Internet Network Information Center (2019). China's online shopping market research report. *The 44th Statistical Report on Internet Development in China.* http://cnnic.com.cn/IDR/ReportDownloads/. Accessed 20 Sep 2019

Chopra, S., & Meindl, P. (2010). *Supply chain management: strategy, planning, and operation*. Boston: Prentice Hall, c2010.

Chybalski, F. (2017). Forecast value added (FVA) analysis as a means to improve the efficiency of a forecasting process. *OPERATIONS RESEARCH AND DECISIONS; ISSN 2081-8858.*

Fildes, R., P. Goodwin, M. Lawrence, and K. Nikolopoulos. (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.

Georg von Krogh, "Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing," *Academy of Management Discoveries, 4/4 (2018):* 404-409;

Glynn, A. N., & Quinn, K. M. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis, 18(1),* 36–56.

Goodwin, P., Fildes, R., Lawrence, M., & Stephens, G. (2011). Restrictiveness and guidance in support *systems. Omega, 39(3),* 242–253.

Halldórsson, Á., G. Stefánsson, P. Jonsson, L. Kjellsdotter, and M. Rudberg. (2007). "Applying Advanced Planning Systems for Supply Chain Planning: Three Case Studies." *International Journal of Physical Distribution & Logistics Management 37 (10):* 816–834.

Hauke, J., Lorscheid, I., & Meyer, M. (2018). Individuals and their interactions in demand-planning processes: An agent-based, computational testbed. *International Journal of Production Research, 56(13),* 4644–4658.

Hiranya Pemathilake, R. G., Karunathilake, S. P., Achira Jeewaka Shamal, J. L., & Ganegoda, G. U. (2018). Sales Forecasting Based on AutoRegressive Integrated Moving Average and Recurrent Neural Network Hybrid Model. *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD),* 27–33.

Iyer G, Narasimhan C, Niraj R. (2007). Information and inventory in distribution channels. *Manage Sci 53(10):*1551–1561

Kaipia, R., J. Holmström, J. Småros, and R. Rajala. (2017). "Information Sharing for Sales and Operations Planning: Contextualized Solutions and Mechanisms." *Journal of Operations Management 52: 15–29.*

Klein R, Rai A. (2009) "Interfirm strategic information flows in logistics supply chain relationships". *MIS Quarterly, pp. 735-762*, 2009.

Liu, Q., Sun, S. X., Wang, H., & Zhao, J. (2011). A multi-agent based system for e-procurement exception management. *Knowledge-Based Systems, 24(1),* 49–57.

Moon, M. A., Mentzer, J. T., & Smith, C. D. (2003). Conducting a sales forecasting audit. *International Journal of Forecasting, 19(1),* 5–25.

Moritz, B., E. Siemsen, and M. Kremer. (2014). "Judgmental Forecasting: Cognitive Reflection and Decision Speed." *Production and Operations Management 23 (7)*: 1146–1160.

Nichols, A., (2007). Causal inference with observational data. *Stata J. 7 (4),* 507–541.

Oliva, R., and N. Watson. (2011). "Cross-functional Alignment in Supply Chain Planning: A Case Study of Sales and Operations Planning." *Journal of Operations Management 29 (5):* 434–448.

Qian, L. (2016). A Study on Information Sharing in Logistics Supply Chains within E-commerce Environments. *International Journal of Simulation -- Systems, Science & Technology. 2016;17(18):17.*1-17.6.

Rached M, Bahroun Z, Campagne JP. (2015). Assessing the value of information sharing and its impact on the performance of the various partners in supply chains. *Comput Ind Eng 88(22)*:237–253

Reinsch, R. (2005), "E-commerce: managing the legal risks", *Managerial Law, Vol. 47 No. 1/2,* pp. 168-196.

Ren, F., & Bao, Y. (2020). A Review on Human-Computer Interaction and Intelligent Robots. *International Journal of Information Technology & Decision Making, 19(01),* 5–47.

Revilla, E., & Rodríguez-Prado, B. (2018). Building ambidexterity through creativity mechanisms: Contextual drivers of innovation success. *Research Policy, 47(9),* 1611–1625.

R. M. Kapila Tharanga Rathnayaka, D. M. K. N. Seneviratna, W. Jianguo and H. I. Arumawadu, (2015). "A hybrid statistical approach for stock market forecasting based on Artificial Neural Network and ARIMA time series models," *2015 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC), Nanjing, 2015,* pp. 54-60.

Saenz, M. J., Revilla, E., & Simón, C. (2020). Designing AI Systems With Human-Machine Teams. *MIT SLOAN MANAGEMENT REVIEW*, 7.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *J Data Warehouse, 5.*

StataCorp. (2019). Stata treatment-effects reference manual: potential outcomes/counterfactual outcomes. Stata: Release 16. *Statistical Software*, 216-224

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.

Stuart J. Russell, Peter Norvig. (2010). *Artificial Intelligence: A Modern Approach, Third Edition, Prentice Hall ISBN 9780136042594.*

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R., & Goodwin, P. (2009). The effects of integrating management judgement into intermittent demand forecasts. *International Journal of Production Economics, 118(1),* 72–81.

Tai, Y. M., Ho, C. F., & Wu, W. H. (2010). The performance impact of implementing web-based e-procurement systems. *International Journal of Production Research, 48(18),* 5397-5414.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika 97*: 661–682.

Vhatkar, S., & Dias, J. (2016). Oral-Care Goods Sales Forecasting Using Artificial Neural Network Model. *Procedia Computer Science*, *79,* 238–243.

Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics, 176*, 98-110.

Wang, X., & Petropoulos, F. (2016). To select or to combine? The inventory performance of model and expert forecasts. *International Journal of Production Research, 54(17),* 5271–5282.

Wenzel, H., Smit, D., & Sardesai, S. (2019). A literature review on machine learning in supply chain management. In C. M. Kersten Wolfgang Blecker, Thorsten Ringle (Ed.), *Artificial Intelligence and Digital Transformation in Supply Chain Management: Innovative Approaches for Supply Chains. Proceedings of the Hamburg International Conference of Logistics (HICL), Vol. 27* (pp. 413–441).

Wong, J. (2010*). Internet Marketing for Beginners, Elex Media Komputindo, Jakarta*, 2010.

Zhao, J., Zhu, H., & Zheng, S. (2018). What is the value of an online retailer sharing demand forecast information. *Soft Computing, 22(16),* 5419–5428.

Zhou, H., W. Benton, D. A. Schilling, and G. W. Milligan. (2011). "Supply Chain Integration and the SCOR Model." *Journal of Business Logistics 32 (4):* 332–344.

Zoryk-Schalla, A. J., J. C. Fransoo, and T. G. de Kok. (2004). "Modeling the Planning Process in Advanced Planning Systems." *Information & Management 42 (1):* 75–87.

APPENDIX

Appendix A Glossary

Terms	Definition
AI	Artificial Intelligence.
AIPW	Augmented Inverse Propensity Weight. An estimator for average treatment effects.
ANN	Artificial Neural Network.
API	Application Programming Interface.
APS	Advance Planning System.
ARIMA	Autoregressive Integrate Moving Average.
ATE	Average Treatment Effect. ATE is a measure used to compare treatments (or interventions) in randomized experiments, by find the difference in mean (average) outcomes between units assigned to the treatment and units assigned to the control.
BPM	Business Process Management
CRISP-DM	Cross-Industry Standard Process for Data Mining model.
Customer J	One of the largest e-commerce platforms in China.
DAT	Detail Assumption Tool. A system of Supplier P that take the demand adjustment in certain format, then convert the adjustment to IDP system for next process.
DC	Distribution Center.
e-commerce	Electronic Commerce.
EDI	Electronic Data Interchange.
ERP	Enterprise Resource Planning.
FMCG	Fast-Moving Consumer Goods
HMI	Human-Machine Interface.
IDA	intelligent demand-adjustment
IDP	Integrated Demand Planning. A system of Supplier P that take the demand forecast then plan the future production and distribution.
MAPE	Mean Absolute Percentage Error, a measure of prediction accuracy of forecasting methods in statistics
MDM	Master Data Management.
ML	Machine Learning.
POM	Potential Outcome Mean. The mean of the outcome that would be realized if the individual received a specific value of the treatment.
RMB	Ren Min Bi. Chinese currency unit.
RNN	Recurrent Neural Network.
RPA	Robotic Process Automation
RPC	Retail Product Code, the product code ecommerce customer used to manager their product.
SFU	Supply Fulfillment Unit. The internal product code of Supplier P use as demand forecast purpose.
SKU	Stock Keeping Unit. A unit that used to manage inventory
Supplier P	One of the largest consumer goods companies, as the sponsor of this research.

Appendix B Sample Stata Analysis Code

```
* thesis analysis code

* Calculate the overall data ATE

* MAPE ATE calculation

* Read dataset
use data, clear

*Calculate the descriptive statistics
. mean mapexaccuracy

. mean mapexaccuracy, over(treatment)

*Regression adjustment
. teffects ra (mapexaccuracy price_dummy turnover_dummy segmentation_a
segmentation_b segmentation_c
segmentation_d segmentation_e segmentation_f, poisson) (treatment)

*AIPW single treatment effects
. teffects aipw (mapexaccuracy price_dummy turnover_dummy segmentation_a
segmentation_b
segmentation_c segmentation_d segmentation_e segmentation_f, poisson)(treatment
price_dummy
turnover_dummy segmentation_a segmentation_b segmentation_c segmentation_d
segmentation_e
segmentation_f), coeflegend

*ATE calculation compared treatment groups with control group
. nlcom (_b[ATE:r1vs0.treatment] / _b[POmean:0.treatment])(_b[ATE:r2vs0.treatment]
/ _b[POmean:
0.treatment])(_b[ATE:r3vs0.treatment] / _b[POmean:0.treatment])

*ATE calculation compared among treatments group
. teffects aipw (mapexaccuracy price_dummy turnover_dummy segmentation_a
segmentation_b
segmentation_c segmentation_d segmentation_e segmentation_f, poisson)(treatment
price_dummy
turnover_dummy segmentation_a segmentation_b segmentation_c segmentation_d
segmentation_e
segmentation_f), control(1) coeflegend

. nlcom _b[ATE:r2vs1.treatment] / _b[POmean:1.treatment], noheader
. nlcom _b[ATE:r3vs1.treatment] / _b[POmean:1.treatment], noheader

. teffects aipw (mapexaccuracy price_dummy turnover_dummy segmentation_a
segmentation_b
segmentation_c segmentation_d segmentation_e segmentation_f, poisson)(treatment
price_dummy
turnover_dummy segmentation_a segmentation_b segmentation_c segmentation_d
segmentation_e
segmentation_f), control(2) coeflegend

```

```
. nlcom _b[ATE:r1vs2.treatment] / _b[POmean:2.treatment], noheader
. nlcom _b[ATE:r3vs2.treatment] / _b[POmean:2.treatment], noheader

. teffects aipw (mapexaccuracy price_dummy turnover_dummy segmentation_a
segmentation_b
segmentation_c segmentation_d segmentation_e segmentation_f, poisson)(treatment
price_dummy
turnover_dummy segmentation_a segmentation_b segmentation_c segmentation_d
segmentation_e
segmentation_f), control(3) coeflegend

. nlcom _b[ATE:r1vs3.treatment] / _b[POmean:3.treatment], noheader
. nlcom _b[ATE:r2vs3.treatment] / _b[POmean:3.treatment], noheader

*Get all POM
. teffects aipw (mapexaccuracy price_dummy turnover_dummy segmentation_a
segmentation_b
segmentation_c segmentation_d segmentation_e segmentation_f, poisson)(treatment
price_dummy
turnover_dummy segmentation_a segmentation_b segmentation_c segmentation_d
segmentation_e
segmentation_f), pom

*count each treatment number
count if treatment == 0
count if treatment == 1
count if treatment == 3
```