UNIVERSITY OF CALIFORNIA,
IRVINE

Essays in Value Creation with Digital Technologies: The Role of Machine Learning and Human Experts

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Management

by

Jin Sik Kim

Dissertation Committee:
Professor Vijay Gurbaxani, Chair
Professor Vibhanshu Abhishek
Professor Tingting Nian

2021

DEDICATION

To

my parents, professors, and friends

in recognition of their worth

And To my 30s

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven..."

(Charles Dickens
A Tale of Two Cities)

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my most profound appreciation to my committee chair, Professor Vijay Gurbaxani, who has been a tremendous advisor throughout my doctoral program. Dr. Gurbaxani continually and convincingly conveyed a spirit of adventure regarding research and new knowledge. Without his guidance, encouragement, and persistent help, this dissertation would not have been possible.

I would also like to sincerely thank my committee members, Professor Vibhanshu Abhishek and Professor Tingting Nian, for their insightful and constructive feedback. Dr. Abhishek sets an excellent example for me to expand my ability and grow up as a better researcher.

I thank Professor Yi-Jen (Ian) Ho at Pennsylvania State University, who provided me sincere brothership and substantial help in this Ph.D. journey. And I express my gratitude to Dr. Mingdi Xin, who provided significant support in my job search and helpful suggestions for my studies. I would also like to acknowledge beneficial suggestions on the research design from Dr. Vidyanand Choudhary and Dr. Jiawei Chen as members of my proposal committee.

My experience in the Ph.D. program study at the University of California, Irvine, benefited from the friendship and support of my fellow doctoral students. In particular, I would like to thank Shengjun Mao, Jooho Kim, Aruhn Venkat, Yoojin Lee, and Hyewon Park. In addition, I would also like to thank a special friend, Noel Negrete. Noel carefully supported various administrative tasks to help me concentrate on my studies and research during my Ph.D. program. Finally, I would like to acknowledge Hwan Heo's helps. I will not forget that he is the number one contributor to my Ph.D. journey.

Lastly, I want to thank my dearest parents - Jong Wook Kim and Hee Soon Kim, my younger sister and brother-in-law – Seo Yun Kim and Choulki Park, and my two litter angel nieces – Yuha Park and Yuna Park. Thank you for encouraging me to pursue a doctorate degree and for supporting me during this long journey. I love you so much.

VITA

Jin Sik Kim

2004	Bachelor of Arts in Journalism & in Management Information Systems, Yonsei University, South Korea (Double Major)
2006	Master of Science in Business Administration, The School of Business, Yonsei University, South Korea (Management Information Systems)
2010	MPIA, The School of Global Policy & Strategy, The University of California, San Diego (International Management)
2013	Master of Arts in Economics, The Maxwell School of Citizenship & Public Affairs, Syracuse University
2021	Ph.D. in Management, The Paul Merage School of Business, The University of California, Irvine (Information Systems)

FIELD OF STUDY

Information Systems

ABSTRACT OF THE DISSERTATION

Essays in Value Creation with Machine Learning
by

Jin Sik Kim

Doctor of Philosophy in Management

University of California, Irvine, 2021

Professor Vijay Gurbaxani, Chair

One of the most exciting tools that has entered our life in recent years is machine learning. The family of cognitive learning algorithms has already proved to bring convenience in many aspects considerably in our daily life. However, it is still unclear how machine learning can be utilized and create incremental value over existing information technologies and algorithms. This dissertation examines value creation with machine learning and compares its value to other competing systems. In the first two essays, we begin with an examination of the forecasting accuracy of machine learning algorithms compared to other traditional econometric models in predicting daily demand cases at a theme park. In the first essay (Chapter 2), we find improvements in machine learning algorithms' prediction accuracy in nonlinear and non-stationary daily demand forecasting. In the second essay (Chapter 3), we create a deep-learning neural network using a Human-in-the-Loop (HITL) approach. This chapter shows that the HITL approach outperforms prior approaches in prediction accuracy & consistency and in (in)direct economic benefits over other benchmarks. These results indicate that human expertise complements, rather than substitutes, machine learning. Also, the HITL approach performs better in demand prediction with sudden changes, bringing considerable advantages to business practices. In Chapter 4, we show the existence of price dispersion online. We leverage a machine learning application to measure information asymmetry in this research context. The analysis finds robust evidence of value creation with machine learning. The last chapter discusses implications and limitations for academia and practice for future machine learning applications, investments, and value creation research.

Keywords: Machine learning algorithms, HITL AI, DNN, value creation, human experts

CHAPTER1
Introduction

Advances in computing power, coupled with large digital datasets, have made it practical to apply artificial intelligence (AI) methods to solve problems in real-time and at scale. Not surprisingly, venture capitalists invested over $19 billion in AI startups in 2019 alone and established firms invested several times that amount in AI initiatives.[1] Software vendors, such as IBM, Amazon, and SAS, have built the off-the-shelf, "general-purpose AI (GPAI)" systems that are deployed at many companies.[2]

In reality, these AI initiatives remain early-stage applications in business practice, and their economic value is still unclear. A recent study by IDC shows that the failure rate of AI investments is significantly high.[3] About 80% of enterprises currently engaged in ML/ AI projects have stalled.[4] Hosanagar & Saxena (2017) suggest that the frequent failures are due to the complexity of tasks being targeted by these systems. The perception of AI/ ML algorithms as a plug-and-play technology with immediate returns is a significant factor in these failures (Fountaine et al. 2019). Therefore, important questions arise in academia and business practice about the real economic value of advanced machine learning algorithms,

[1] Forbes, January 5, 2020. Is Venture Capital Investment In AI Excessive? https://www.forbes.com/sites/cognitiveworld/2020/01/05/is-venture-capital-investment-for-ai-companies-getting-out-of-control/#79307a07e050

[2] Forbes, February 20, 2020. Gartner's 2020 Magic Quadrant for Data Science And Machine Learning Platforms Has Many Surprises https://www.forbes.com/sites/janakirammsv/2020/02/20/gartners-2020-magic-quadrant-for-data-science-and-machine-learning-platforms-has-many-surprises/#654e3fa53f55

[3] Forbes, July 19, 2019. This week In AI Stats: Up To 50% Failure Rate In 25% Of Enterprises Deploying AI https://www.forbes.com/sites/gilpress/2019/07/19/this-week-in-ai-stats-up-to-50-failure-rate-in-25-of-enterprises-deploying-ai/#6e8f074872ce

[4] Readwrite, August 30, 2020. Challenge of Adopting AI in Businesses https://readwrite.com/2020/08/30/challenges-of-adopting-ai-in-businesses/

how these algorithms can be utilized, and what human roles should be in the adoption and use of machine learning applications.

This dissertation will investigate machine learning applications in a theme park and an online store to answer the research questions. In the second chapter, we compare the forecasting accuracy of machine learning algorithms to approaches from traditional econometrics. By doing so, we try to explain the forecasting value of machine learning algorithms in cases where there is nonlinear and non-stationary demand. The third chapter investigates the human role in machine learning applications and their economic and non-economic values. The third chapter focuses on studying questions about the role of humans due to the surge of machine learning applications. Then, we would like to highlight the complementary relationship between human experts and machine algorithms for demand forecasting. In the fourth chapter, we use machine learning applications to understand market mechanisms. We show the existence of market inefficiency using machine learning algorithms and market structure. The rest of this chapter presents the research contexts.

1.1 Machine Learning System for Non-Parametric Demand Prediction

Chapter 2 compares non-parametric demand prediction using machine learning is better to predictions using traditional econometrics. We have witnessed a proliferation of prediction models using machine learning methods (e.g., Chen & Wang 2007; Claveria et al. 2015; Hu & Song 2019; Law et al. 2019). Previous literature has revealed the value of machine learning algorithms, which by reshaping the underlying economic mechanisms, has many dimensions: increase in international trade (Brynjolfsson et al. 2019); a company's market value (Rock 2019); productivity (Tambe 2014); profitability (Cui et al. 2006); and

consumer surplus (Miklós-Thal & Tucker 2019). In the same vein, we can expect the reduced cost of prediction and increased prediction accuracy from the machine learning algorithms (Loebbecke & Picot 2015). However, there is no consensus on identifying the best forecasting model (e.g., Peng et al. 2014; Sinha & May 2004; Zhou et al. 2004).

Among machine learning models, neural network models generally report the best predictions for nonlinear and non-stationary prediction problems (Tam & Kiang 1992). However, variable selection and the size of the data are the essential criteria in the accuracy of the model (Peng et al. 2014; Walczak 2001; Zhou et al. 2004). Moreover, neural network models do not identify as the best prediction model in the complex demand forecasting problems (Cui et al. 2006; Lin et al. 2017).

Building better learning algorithms in IS has studied the interdependence of data and algorithms (Grover et al. 2018; Ransbotham et al. 2017), the optimal amount of training data (Walczak 2001; Yan et al. 2017), the effect of noise or biased data (Cowgill 2018), characteristics of data (Grover et al. 2018; Zhang & Hu 1998), the optimal number of parameters (Yan et al. 2017; Zhang & Hu 1998), and computational ability along with hardware capability (Monroe 2018; Lu et al. 2018).

Given these points, we compared machine learning algorithms and twelve traditional econometric models heavily used in demand forecasting. For this chapter, we partnered with one of the biggest theme parks in the world and the biggest in South Korea. We find the forecasting performance using our machine learning approach in demand prediction outperforms the twelve traditional econometric models. Non-parametric methods such as deep-learning neural networks, Bayesian Structural Time Series, and OLS outperform other

parametric econometrics for the more non-stationary and nonlinear demand trends. In contrast, parametric time series models outperform non-parametric models, including the machine learning approach, for the more linear and stationery demand trends.

1.2 Demand Prediction using ML: A Human-in-the-Loop Approach

Chapter 3 explores the question: do machine algorithms substitute for or complement human expertise? With the growth of machine learning algorithms, human roles are expected to change in business practice. Human decision-makers in complex environments are constrained by their computation limitations. The family of cognitive intelligence algorithms coupled with high-performance computation power can process information similarly to the processes used by humans with lower cost and less time (Lu et al. 2018). Moreover, since some forms of data are not scarce, can be inexpensive to obtain, and available in digital format, machine learning algorithms should provide a cheaper and better substitute for human predictions and automation (Acemoglu & Restrepo 2018; Agrawal et al. 2016; Agrawal & Dhar 2014; Davenport et al. 2018). Hence, the demand for laborers who deal with unskilled labor tasks (Sachs & Kotlikoff 2012) and data-driven managerial decision-making processes can decrease (Huang & Rust 2018; Loebbecke & Picot 2015; McAfee & Brynjolfsson 2012). However, we focus on applying the Human-in-the-Loop (HITL) approach, which exploits the complementary relationship between human experts and machine algorithms. More specifically, the HITL approach uses humans to provide continuous feedback to the ML system in an iterative manner to improve its performance, making these systems more resilient to sudden changes in the data generating process.

In the third chapter, we compared prediction accuracy, prediction consistency, direct economic benefits, and indirect economic benefits of HITL AI to general-purpose AI and human experts. As in chapter 2, we partner with one of the biggest theme parks in the world and the biggest in South Korea. Subsequently, we ran a field experiment for six months to test the performance of the various models. Then, we compared the forecasting results and their consequences and provide possible explanations of the superior performance of the complementarity of human experts and machine algorithms.

To summarize our results, we find robust evidence for the complementarity of human experts and machine algorithms. The field experiment shows that HITL AI forecasting performance is significantly better than both human experts and GPAI in accuracy and consistency for predicting daily demand. Also, the forecasting performance of HITL AI is superior to the traditional econometric models. From the economic value perspective, HITL AI improves the theme park's average daily revenues by around 5~6% compared to experts and GPAI. Finally, we found a substantial improvement of HITL AI in indirect economic benefits. For example, HITL AI reduces the data processing time in demand prediction, giving the park an advantage to react in dynamic environments. This can lead to better park management and customer satisfaction.

1.3 Revisiting Price Dispersion in Electronic Markets

Theory predicts that the prices of homogenous products will display relative convergence in electronic markets (Bakos 1997, Brynjolfsson and Smith 2000a). Yet, as any online shopper will have observed, there is considerable dispersion in the prices of homogeneous products in electronic markets. The fourth chapter aims to provide

explanations for the observed price dispersion for homogeneous products in electronic markets.

Previous literature predicts that the nature of electronic markets will promote price competition (Bakos 1997; Brynjolfsson & Smith 2000a; Kauffman & Wood 2000; Smith 2001). This is because travel costs are no longer a constraint in electronic markets (Bakos 1997), consumers can search for alternatives easily, and retailers can track competitors' pricing with low costs (Brynjolfsson & Smith 2000b). However, scholars also provide substantial evidence to the contrary - the existence of price dispersion (e.g., Lynch & Ariely 2000; Degeratu et al. 2000; Clay et al. 2001; Ancarani & Shankar 2004; Dimoka et al. 2012).

Price dispersion in electronic markets is evidence of the existence of inefficiency. For example, consumers are not acquainted with complete information on sellers and products in electronic markets (e.g., Dimoka et al. 2012; Kim & Krishnan 2015). Also, the price dispersion can be explained by market competition (Haynes & Thompson 2008). For instance, we observed that competition due to the number of sellers leads to an empirically converged price (e.g., Barron et al. 2004).

We study a single online marketplace with 12,409 retailers. We compare four different products rather than focus on a single product category like some prior studies. This allows us to control for the product uncertainty inherent in different product categories. We capture price dispersion by examining multiple sellers who sell the same product on a single platform, thereby controlling for platform heterogeneity. We use text-mining, a popular machine learning technique, to capture accurate quality information on products and sellers from other customers to measure information asymmetry. By doing so, we directly measure

the other customers' evaluation of products and sellers. Finally, we use the Herfindahl Hirschman Index to understand the market structure and competition on price dispersion in an integrated manner. We use a well-controlled empirical setting, Hierarchical Linear Modeling (HLM).

We find that product uncertainty and seller certainty judged by other customers and market competition are significant factors correlated to price dispersion. Our study adds value to the existing literature on price dispersion in electronic market by carefully addressing the limitations of previous studies. For example, most previous studies do not effectively handle the product uncertainty inherent in different product categories. This it makes it easier to generalize the empirical results.

From a machine learning application perspective, we introduce a new way to measure information asymmetry (uncertainty and certainty). We first use text-mining and sentiment analysis using a natural language process, and then we calculate the level of uncertainty (or certainty) using entropy and an information gain function. The new way of measuring information asymmetry works better to understand the existence of market inefficiency (price dispersion) than the previous methods (e.g., utilizing star ratings, certifications, or third-party evaluations). We argue that text-mining is an essential application for sellers in the electronic market which sells homogenous products.

In sum, machine learning approaches can bring superior results in demand prediction, especially in nonlinear and non-stationary cases. Also, the machine learning application can be beneficial to the sellers in online markets to set their pricing strategy. Finally, with the

widespread use of machine learning applications, It is important to consider human experts as complements to machine systems, not substitutes.

References

Acemoglu, D., & Restrepo, P. (2018). *Artificial intelligence, automation and work* (No. w24196). National Bureau of Economic Research. Available at NBER: https://www.nber.org/papers/w24196

Agrawal, A., Gans, J., & Goldfarb, A. (2016). The simple economics of machine intelligence. *Harvard Business Review*, *17*, 2-5. https://hbr.org/2016/11/the-simple-economics-of-machine-intelligence

Agarwal, R., & Dhar, V. (2014). Editorial - Big data, data science, and analytics: The opportunity and challenge for IS research, Information Systems Research, 25(3), 443-448.

Ancarani, F., and Shankar, V. (2004). Price Levels and Price Dispersion Within and Across Multiple Retailer Types: Further Evidence and Extension. *Journal of Academy of Marketing Science*, 32(2), pp. 176-187.

Bakos, Y. (1997). Reducing Buyer Search Costs: implications for Electronic Marketplaces, *Management Science,* 43(12), pp. 1679-1692.

Barron, J. M., Taylor, B. A., & Umbeck, J. R. (2004). Number of sellers, average prices, and price dispersion. *International Journal of Industrial Organization*, *22*(8-9), 1041-1066.

Brynjolfsson, E., Hui, X., & Liu, M. (2019). Does machine translation affect international trade? Evidence from a large digital platform. Management Science, Article in Advance, 1-12.

Brynjolfsson, E. and Smith, M. (2000a). The Great Equalizer? Consumer Behavior at Internet Shopbots, *Working Paper*, MIT, Cambridge, MA.

Brynjolfsson, E. and Smith, M., (2000b). Frictionless Commerce? A Comparison of Internet and Conventional Retailers, *Management Science*, 46,(4), pp.563-585.

Chen, K. Y., & Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*, *28*(1), 215-226.

Claveria, O., Monte, E., & Torra, S. (2015). Tourism demand forecasting with neural network models: different ways of treating information. *International Journal of Tourism Research*, *17*(5), 492-500.

Clay, K., Krishnan, R., and Wolff, E. (2001). Prices and Price Dispersion on the Web: Evidence from the online book industry, *Journal of Industrial Economics*, 49(4), pp. 521-540.

Cowgill, B. (2018). Bias and productivity in humans and algorithms: Theory and evidence from resume screening. Columbia Business School, Columbia University, August 29.

Cui, G., Wong, M. L., & Lui, H. K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. Management Science, 52(4), 597-612.

Davenport, T. H., Libert, B. & Beck, M. (2018). Robo-Advisers Are Coming to Consulting and Corporate Strategy. Harvard Business Review, January 12. Available at: https://hbr.org/2018/01/robo-advisers-are-coming-to-consulting-and-corporate-strategy

Degeratu, A., Rangaswamy, A., and Wu, J. (2000). Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes, *International Journal of Research in Marketing*, 17(1), pp. 55-78.

Dimoka, A., Hong, Y., and Pavlou, P.A. (2012). On Product Uncertainty in Online Markets: Theory and Evidence, *MIS Quarterly*, 36(2), pp. 395-426.

Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review*, *97*(4), 62-73.

Grover, V., Chiang, R. H., Liang, T. P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. Journal of Management Information Systems, 35(2), 388-423.

Haynes, M., & Thompson, S. (2008). Price, price dispersion and number of sellers at a low entry cost shopbot. *International Journal of Industrial Organization*, *26*(2), 459-472.

Hosanagar, K., & Saxena, A. (2017). The First Wave of Corporate AI Is Doomed to Fail. Harvard Business Review. April 18, Available at: https://hbr.org/2017/04/the-first-wave-of-corporate-ai-is-doomed-to-fail

Hu, M., & Song, H. (2020). Data source combination for tourism demand forecasting. *Tourism Economics*, 26(7), 1248-1265.

Huang, M. H., & Rust, R. T. (2018). Artificial intelligence in service. Journal of Service Research, 21(2), 155-172.

Kauffman, R. and Wood, C. (2000). Follow the Leader? Strategic Pricing in E-Commerce, *ICIS 2000 Proceeding Paper*, 14. Available at: http://aisnet.org/icis2000/14

Kim, Y. and Krishnan, R. (2015). On Product-Level Uncertainty and Online Purchase Behavior: An Empirical Analysis, *Management Science*, Published online in Articles in Advance: 02 April 2015.

Law, R., Li, G., Fong, D. K. C., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of tourism research*, *75*, 410-423.

Lin, Y. K., Chen, H., Brown, R. A., Li, S. H., & Yang, H. J. (2017). Healthcare predictive analytics for risk profiling in chronic care: A Bayesian multitask learning approach. MIS Quarterly, 41(2). 473-496 & A1-A3.

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. Journal of Strategic Information Systems, 24(3), 149-157.

Lu, Y., Qian, D., Fu, H., & Chen, W. (2018). Will supercomputers be super-data and super-AI machines?. Communications of the ACM, 61(11), 82-87.

Lynch, J., and Ariely, D., (2000). Wine Online: Search Costs Affect Competition on Price, Quality, and Distribution, *Marketing Science*, 19(1), pp. 83-103.

McAfee, A. & Brynjolfsson, E. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Miklós-Thal, J., & Tucker, C. (2019). Collusion by Algorithm: Does Better Demand Prediction Facilitate Coordination Between Sellers?. Management Science, 65(4), 1552-1561.

Monroe, D. (2018). Chips for artificial intelligence. Communications of the ACM, 61(4), 15-17.

Peng, B., Song, H., & Crouch, G. I. (2014). A meta-analysis of international tourism demand forecasting and implications for practice. Tourism Management, 45, 181-193.

Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2017). Reshaping business with artificial intelligence: Closing the gap between ambition and action. MIT Sloan Management Review, 59(1).

Rock, D. (2019). Engineering Value: The Returns to Technological Talent and Investments in Artificial Intelligence. Working paper, May. 1. Available at SSN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3427412

Sinha, A. P., & May, J. H. (2004). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. Journal of Management Information Systems, 21(3), 249-280.

Sachs, J. D., & Kotlikoff, L. J. (2012). Smart machines and long-term misery (No. w18629). National Bureau of Economic Research. Available at NBER: https://www.nber.org/papers/w18629

Smith, M.D. (2001). The Law of One Price? The Impact of IT-Enabled Markets on Consumer Search and Retailer Pricing. *Working Paper*, Carnegie Mellon University, PA.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. Management science, 38(7), 926-947.

Tambe, P. (2014). Big data investment, skills, and firm value. Management Science, 60(6), 1452-1469.

Walczak, S. (2001). An empirical analysis of data requirements for financial forecasting with neural networks. Journal of management information systems, 17(4), 203-222.

Yan, D., Zhou, Q., Wang, J., & Zhang, N. (2017). Bayesian regularisation neural network based on artificial intelligence optimisation. International Journal of Production Research, 55(8), 2266-2287.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker Jr, J. F. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, *20*(4), 139-166.

CHAPTER2
Machine Learning System for Non-Parametric Demand Prediction

2.1 Introduction

Mayer-Schönberger and Ramge (2018) have argued that the source of innovation is shifting from human ingenuity to data-driven machine learning. Researchers have suggested many measures of the economic value of machine learning algorithms: (1) an increase in international trade (Brynjolfsson et al. 2019), (2) the market value of a company (Rock 2019), (3) productivity, profitability, and consumer surplus by reshaping of underlying economic mechanisms and by reducing the cost of prediction (Tambe 2014), and (4) the accuracy of prediction (Cui et al. 2006; Miklós-Thal & Tucker 2019). Therefore, advanced machine learning algorithms can be seen as a source of economic value creation and competitive advantage.

Even though expectations for sophisticated machine learning algorithms are high, the reality is that machine learning initiatives remain early-stage applications in business practice. Most of the projects using the family of cognitive learning-driven algorithm projects fail (Flam 2018; Hosanagar & Saxsena 2017; Gartner Sep 2015; Feb 2018; Pumplun et al. 2019). Moreover, the success rate of machine-based learning algorithms is also unclear (Grover et al. 2018). For example, nearly 25% of 2,500 business organizations worldwide have implemented artificial intelligence initiatives; yet report a 50% failure rate (IDC Global survey 2019). Also, 85% of machine-based learning algorithm projects do not deliver value (Gartner 2015).

This chapter explores the performance of non-parametric demand prediction using machine learning in a dynamic/ non-linear environment. To determine its performance, we compare it to the conventional econometric models.

To fulfill this objective, we compare the accuracy of the daily demand forecast of machine learning algorithms and traditional econometrics models in the context of a theme park. The accuracy of daily park demand prediction is a core task that affects the provision of services and staffing and the operation of related leisure facilities, such as ordering related food, beverage, and goods. The daily demand for a theme park is non-linear and non-stationary, requiring considerable effort and cost to predict. In tourism demand forecasting, we have witnessed a growth of prediction models using machine learning algorithms (Burger et al. 2001; Chen & Wang 2007; Claveria et al. 2015; Goh et al. 2008; Hu & Song 2019; Law et al. 2019). However, there is no consensus on identifying the best forecasting model.

2.2 Literature Review

To further understand the role and related issues in the business environment, we identify three streams of machine learning research that are relevant to our study. The three selected streams are (1) machine learning and economic value, (2) machine learning and prediction accuracy, and (3) developing better machine learning systems.

The first stream estimates the economic value of machine learning algorithms (Table 2.1). Mayer-Schönberger and Ramge (2018) argue that the source of innovation is shifting from human ingenuity to data-driven machine learning. Therefore, the new sources of competitive advantage result from access to advanced data-driven machine learning systems, not traditional factors such as market concentration and scale benefits. However,

there is still a need to better understand how machine algorithms increase business value. For example, Davenport and Ronanki (2018) argue that the economic benefits of early artificial intelligence/ machine learning adoption with cumulative data include improved products, decisions, and operations. However, Agrawal et al. (2018) argue that the accumulated data is limited to a one-time benefit from training. In addition, the contents of data collected by one business practice are imitable from other organizations in similar business practice (Agrawal et al. 2018; Grover et al. 2018), which means that advantages are difficult to sustain.

Table 2.1 Summary of Selected Research on ML and Economic Value

Study	Source	Findings
Stream: Artificial Intelligence and Economic Value		
Cui, Wong, & Lui (2006)	Management Science	Machine learning is an innovative method that can potentially improve forecasting models and assist management decision-making. Using the Bayesian networks learning model, the authors show that accurate prediction, transparent procedures, interpretable results, and explanatory power are higher by using the Bayesian network learning model, which can lead to sales, reduce costs, and improve profitability direct marketing decisions.
Meyer, Adomavicius, Johnson, Elidrisi, Rush, Sperl-Hillen, & O'Connor (2014)	Information Systems Research	The authors use machine learning techniques to support complex, ill-structured dynamic decision-making contexts to identify the conditions predicting failures in a dynamic decision-making context. Through anticipating conditions in failure cases, machine learning techniques can lead to a cost-effective method to achieve the desired goal.
Tambe (2014)	Management Science	The paper analyzes the investment of big-data analysis systems, the productivity of investment, and labor market factors. The

		productivity growth related to the investment is associated with a 3% faster growth rate when the organization that invests in the extensive data analysis system is a data-intensive organization or has enough technical skills experts.
Loebbecke & Picot (2015)	Journal of Strategy Information Systems	Reshaping with digitization and (machine-based) bid-data analytics impacts underlying economic mechanisms (centralized production, increased harmonization of demand, & erosion of property rights) and business models.
Agrawal, Gans, & Goldfarb (2016)	Harvard Business Review	Machine intelligence system drops in the cost of prediction.
Agrawal, Gans, & Goldfarb (2017)	Harvard Business Review	Artificial intelligence algorithm drops in the cost of prediction. As the cost of machine prediction falls, machines will make more and more predictions.
Agrawal, Gans, & Goldfarb (2018)	Harvard Business Review	The value of the accumulated data is limited to a one-time benefit from training. At the same time, the unique operational data have economic value because it can enhance the prediction of the machine.
Davenport & Ronanki (2018)	Harvard Business Review	AI and cognitive technologies bring economic benefits to early adopters. As the number of AI deployments increases, the cumulative benefits grow. Business organizations view AI as a way to improve products, decisions, and operations.
Mayer-Schönberger & Ramge (2018)	Harvard Business Review	The authors argue that the source of innovation is shifting from human ingenuity to data-driven machine learning. Hence, the firms that benefit are the ones that have access to the most data, not the ones that have market concentration and scale.
Grover, Chiang, Liang, & Zhang (2018)	Journal of Management Information Systems	Big-data analytics can provide novel and valuable insights to exploit new business opportunities or defend against competitive threats; however, data contents may be imitable. The things which make big-data analytics unique are analytics knowledge, human talent,

		expertise, and experience with advanced big-data management.
Brynjolfsson, Hui, & Liu (2019)	Management Science	A new artificial intelligence system (machine translation system) has significantly increased international trade on the online platform by 10.9%, substantially reducing translation costs and improving economic efficiency.
Miklós-Thal & Tucker (2019)	Management Science	Using a game-theoretic model, the authors found that better demand prediction by learning algorithms such as AI/ML allows colluding firms to adjust prices better to meet demand conditions. But each firm has incentives to deviate from collusive prices because it increases their temptation to undercut prices. Hence, better demand prediction leads to lower prices (set their price below the monopoly price) and higher consumer surplus.

Other studies have revealed the value of machine learning algorithms, which by reshaping the underlying economic mechanisms, has many dimensions: increase in 10.9% international trade with a substantial reduction in translation costs (Brynjolfsson et al. 2019); increases in a company's market value by 4-7% (Rock 2019); and grow productivity 3% faster (Tambe 2014).

From a microeconomics perspective, machine learning enhances products, decisions, and operations (Davenport & Ronanki 2018). Significantly, machine learning algorithms lead to an increase in sales, reduce costs, and improve profitability (Cui et al. 2006) by reducing prediction costs (Agrawal et al. 2016; 2017; Meyer et al. 2014) and improve prediction accuracy (Cui et al. 2006; Miklós-Thal & Tucker 2019).

Finally, with the low variable cost of data storage, processing, transmission, and machine-based interpretation, machine learning is expected to reshape underlying economic mechanisms, favoring centralized production, increased harmonization of demand, and

erosion of property rights, leading to changes in the role of employment in the market strategy perspective (Loebbecke & Picot 2015).

The second stream in machine algorithms examines whether machine learning algorithms perform better than traditional forecasting models (Table 2.2). Studies in Information Systems have tested various machine learning algorithms such as neural networks, Bayesian network learning, Bayesian Multitask learning model, support-vector machine, CART algorithm, decision trees, ID3, fuzzy model, and grey model. However, the literature in IS fails to reach a consensus on identifying the most precise machine learning algorithm for prediction (Sinha & May 2004; Zhou et al. 2004). Also, IS studies have not found a consensus on whether machine learning algorithms are better forecasting models than other econometric models (Peng et al. 2014; Sinha & May 2004; Yu & Schwartz 2006).

Table 2.2 Summary of Selected Research on ML and Prediction

Study	Source	Findings
Stream: Artificial Intelligence Systems and Prediction		
Tam & Kiang (1992)	Management Science	The authors show a neural network outperforms to predict a nonlinear discriminant pattern recognition compared to a linear classifier, logistic regression, kNN, and ID3 in predictive accuracy, adaptability, and robustness with bank data.
Bansal, Kauffman, & Weitz (1993)	Journal of Management Information Systems	By comparing two prediction models, regression and neural network analysis, the authors conclude that linear regression outperformed neural net when linear relationships are present in terms of r-square; however, neural network analysis forecasts better than linear regression as data accuracy degraded.
Walczak (2001)	Journal of Management Information Systems	As noted, neural network systems incur costs (not only financial but also impact development time and effort) from training data. Only one to two years of

		training data is required to produce the "best" performing backpropagation trained neural network forecasting model.
Sinha & May (2004)	Journal of Management Information Systems	The authors compare five models – neural networks, logistic regression, linear discriminant analysis, decision trees, and nearest-neighbor - on two binary classification problems from the credit evaluation domain. The authors focus on minimizing misclassification costs using ROC and AUC instead of accuracy, finding logistic regression and neural network models were the best performings in expenses.
Zhou, Burgoon, Twichell, Qin, & Nunamaker Jr. (2004)	Journal of Management Information Systems	Decrease the deception in human communication; the authors test prediction performance of the four classification methods: discriminant analysis, logistic regression, decision trees, and neural network. The authors find all of the four methods are potentially good alternatives for predicting deception. Neural networks generally represent a high degree of nonlinear relationships between input and output and are reliable across all testing settings. However, it is highlighted that selecting inputs are significant for the classification.
Cui, Wong, & Lui (2006)	Management Science	The authors compare the accurate prediction, transparent procedures, interpretable results, and explanatory power of the Bayesian network learning model to neural networks, CART, and latent class regression with a large dataset, concluding that Bayesian network models have distinctive advantages.
Yu & Schwartz (2006)	Journal of Travel Research	The study examines two artificial intelligence models (Fuzzy and Grey model) with a simple time series model to predict short time-series tourism demand and finds that AI forecasting methods do not generate a better prediction model.
Chen and Wang (2007)	Tourism Management	The support vector machine's tourism forecasting methods outperform other traditional models, such as the

		autoregressive integrated moving average (ARIMA) model in prediction accuracy (NMSE and MAPE). To build stable and reliable models based on AI, the parameters must be specified carefully.
Guiterrez, Solis, & Mukhopadhyay (2008)	International Journal of Production Economics	The authors compare neural networks to traditional econometrics models such as SES, CM, and SB approximation to forecast lumpy demand. The results reveal that neural networks perform better in measuring lumpy demand forecasting.
Song & Li (2008)	Tourism Management	To forecast tourism demand, new artificial intelligence methods have been identified along with time-series and econometric models. Those are heuristic methods, notably genetic algorithms, fuzzy logic, neural network, and support vector machines.
Peng, Song, & Crouch (2014)	Tourism Management	The authors conduct a meta-analysis in travel demand forecasting between 1980-2011 studies. As no consensus has been reached to find the best forecasting model, the results could not identify a single best model. Moreover, prediction models based on AI are not the best accurate prediction models (lowest MAPE) in overall forecasting ranking most times.
Lin, Chen, Brown, Li, & Yang (2017)	MIS Quarterly	The authors perform experimental evaluations using the Bayesian multitask learning model (BMTL) as an intelligent clinical system. The BMTL attains a better predictive performance on clinical risk profiling than with an alternative prediction model (predict multi-events separately). Also, the BMTL reduces the failure and delays in preventive intervention in clinical practice.
Cowgill (2018)	Working Paper Columbia Business School	The author reports that the employees hired by algorithms outperformed the prediction of hiring and productivity than the employees hired by human interviewers, even with noisy history data.
Dressel & Farid (2018)	Science Advances	Algorithms for predicting recidivism used in general-purpose software are not more accurate or fair than the prediction made

	by people with little or no criminal justice expertise. In addition, even though the algorithm uses more variables to predict, the same accuracy can be achieved with a simple linear predictor with only two features.

For example, studies find that neural networks generally outperform to predict a high degree of nonlinear prediction problems (Tam & Kiang 1992). Compared to other traditional econometrics models such as single exponential smoothing, exponential smoothing, Syntetos-Boylan approximation, the forecasting results reveal that neural networks statistically outperform in measuring lumpy demand forecasting (Guiterrez et al. 2008). Also, the forecasting accuracy of neural networks is more precise than regression models as data accuracy is degraded (Bansal et al. 1993). However, without carefully selecting input variables and considering the size of the training dataset, other models can outperform neural network models even in non-linear relationship problems (Peng et al. 2014; Walczak 2001; Zhou et al. 2004).

Neural networks do not perform better in R-squared performance than regression models in linear problems (Bansal et al. 1993). Also, the algorithms are not significantly superior in minimizing misclassification costs compared to logistic regression in binary classification problems (Sinha & May 2004). Furthermore, Neural nets do not show significant differences in detecting human deception in communication compared to other econometrics and machine learning models (Zhou et al. 2004).

Finally, Bayesian (multitask) learning algorithms show distinctive advantages in prediction accuracy, transparent procedures, interpretable results, and explanatory power

to neural network models with a large dataset, models with latent variables, and multi-event forecasting cases (Cui et al. 2006; Lin et al. 2017).

In our research context of nonlinear and non-stationary demand forecasting - tourism demand forecasting - we have witnessed a proliferation of prediction models using machine learning methods (Burger et al. 2001; Chen & Wang 2007; Claveria et al. 2015; Goh et al. 2008; Hu & Song 2019; Law et al. 2019). However, it is also reasonable to argue that a machine learning-based predictive model is superior in tourism demand forecasting.

The third stream of work discusses how to create better machine learning algorithms (Table 2.3). Since Turing (1950; 2009) first proposed the conceptualization of machine learning, and thinking, the research on cognitive intelligence algorithms has been of interest. Simon (1995) and Schaeffer & Simon (1992) see intelligence algorithms as computer-coded problem solvers with Information Processing Systems (IPS) like human beings.

Table 2.3 Summary of Selected Research on Developing a Better ML Systems

Study	Source	Findings
Stream: Developing better artificial intelligence systems		
Zhang & Hu (1998)	Omega	The authors identify the selection of the optimal number of input and hidden nodes. The size of the training sample on the in-sample and out-of-sample will be the key to building a good forecasting neural network model. The authors find that neural network algorithms perform better in forecasting an exchange rate when the forecast horizon is short. The number of input nodes significantly impacts performance compared to the number of hidden nodes, while a more significant number of observations for training reduces prediction errors.
Walczak (2001)	Journal of Management	Producing high-quality neural network systems one-to-two years of training data

	Information Systems	set would give a better performance in accuracy than a more extensive training data set.
De Leon, Soo, & Williamson (2011)	Journal of Applied Statistics	The authors develop a new algorithm for classifying by the joint distribution of the mixed discrete and continuous variables. The authors find that the approach yield nearly unbiased estimates for large samples. The sample size requirements for the training and estimating are essential to get unbiased results.
Ransbotham, Kiron, Gerbert, & Reeves (2017)	MIT Sloan Management Review	Understanding the interdependence between data and AI algorithms is vital to building successful AI projects. Generating value from an AI algorithm is more complex than simply making or buying AI for a business process. Hence, training AI algorithms with appropriate data is required to have a better result.
Yan, Zhou, Wang, & Zhang (2017)	International Journal of Production Research	The authors propose a Bayesian-regularized artificial neural network (BR-ANN) to solve high noise and a strongly nonlinear problem - stock price prediction. By optimizing the number of parameters and number of training, the authors insist BR-ANN can perform the best because it reduces the potential over-fitting and over-training problem.
Cowgill (2018)	Working Paper Columbia Business School	The author tests heterogeneous predictions about when machine learning algorithms can improve human biases, finding machine algorithms remove human tendencies exhibited in historical training data, but only if the human training decisions are sufficiently noisy; otherwise, the algorithms will codify or exacerbate existing biases.
Davenport, Libert, & Beck (2018)	Harvard Business Review	Developing a specific artificial intelligence system may be more difficult and expensive than many companies willing to venture into. Hence, By either partner with the startup or acquire one would be a good solution.

Di Fiore, Schneider, & Farri (2018)	Harvard Business Review	As more companies invest in AI units, many managers face problems in understanding what AI can do. The authors suggest scouting AI technology, applications, and partners, testing small AI pilots, building the integration and application of AI tools and solutions, educating the organization on the power of AI, and attracting new talents.
Kleber (2018)	Harvard Business Review	The prediction will be wrong when the data is defective and has perpetuated bias or opinion on the subject. It is because of the value-free learning algorithms extracting patterns from historical data, finding patterns, and applying, repeatedly, these patterns to forecast.
Monroe (2018)	Communi-cations of the ACM	As deep-learning algorithms have been increasingly and successfully applied to many business practices, these algorithms' computational and power requirements need to be carefully considered to meet the rapidly growing demand of data processing through customizing hardware, neuromorphic hardware, etc.
Lu, Qian, Fu, & Chen (2018)	Communi-cations of the ACM	High-performance computing and application development based on high computational power makes it possible to promote lots of problems using big data and artificial intelligence.

Scholars in information systems have tried several different approaches to create better machine learning algorithms than previous economics models. Creating business value from machine learning algorithms is linked to effectively training those algorithms with valid data sets (Grover et al. 2018). Ransbotham et al. (2017) point out a common misconception that sophisticated AI algorithms can provide valuable business solutions without sufficient data. Also, compared to organizations with no adoption of AI, pioneering and leading-edge organizations in AI adoption are eight times more likely to understand the data that is needed for training algorithms and twelve times more likely to understand the training

process (Ransbotham et al. 2017). Hence, we need to understand the interdependence between big data and machine learning algorithms to build a superior forecasting algorithm.

Considering data quality, the forecasting results are not expected to be accurate when the data is defective and has perpetuated prejudices or opinions on the subject (Kleber 2018). Conversely, those machine learning algorithms only remove human biases embedded in training data if human training decisions are sufficiently noisy (Cowgill 2018). Moreover, some scholars insist that having unique and large data as a training set is required for the algorithms to perform better (Grover et al. 2018; Zhang and Hu 1998), while others argue that one-to-two years of well-structured and a limited training data set can lead to better performance in accuracy (Walczak 2001; Yan et al. 2017).

In developing algorithms, optimizing the number of parameters has been identified as another key factor in improving algorithms because it reduces over-fitting problems (Yan et al. 2017; Zhang & Hu 1998). To find accurate patterns, forecasting, and solution from algorithms, computational power is another essential factor to consider. Especially, scholars note that the recent explosion in artificial intelligence is less related to the development of new algorithms than the availability of high-performance hardware.[5] Hence, scholars promote meeting rapidly growing demand for data processing through customizing hardware in organizations (Monroe 2018; Lu et al. 2018).

[5] Joel Hruska (2017), "Google's dedicated TensorFlow Processor, or TPU, crushes Intel, Nvidia in Inference workloads," April 6, Available at:https://www.extremetech.com/computing/247199-googles-dedicated-tensorflow-processor-tpu-makes-hash-intel-nvidia-inference-workloads

Given these considerations, we compare the forecasting accuracy of 13 different models such as generic artificial intelligence systems and 12 standard econometrics models. In the following section, we will discuss models, data, and empirical strategies.

2.3 Data Description

In this research, we partner with one of the biggest theme parks in the world and the biggest in South Korea, which receives around 5 million annual visitors. The theme park has made considerable effort to understand the impact of customer visit data and has utilized its demand forecasts to plan human resources, supplies of food & beverage, merchandise, and park operations. The theme park has identified precise prediction of daily visitors as one of its most critical business goals.

The data spans from January 1, 2015, to October 31, 2018, and comprises 267 variables that include proprietary and non-proprietary information such as customer-specific characteristics, weather conditions, day-specific characteristics, events, discounts, reservation information, and so on. Park attendance from January 1, 2015, to December 31, 2017, is shown graphically in Figure 2.1. A total of 14.46M customers visited the theme park in the three years 2015 to 2017. Of these, individual visitors (solo tourism) accounted for 65.96% (9.54M), with 32.75% (3.12M) visiting in the spring and 30.99% (2.96M) in the fall, the two most preferred seasons. Meanwhile, 22.35% (2.13M) and 13.92% (1.33M) of individual visitors visited during the summer and winter, due to the severe seasonal characteristics of these seasons in Korea.

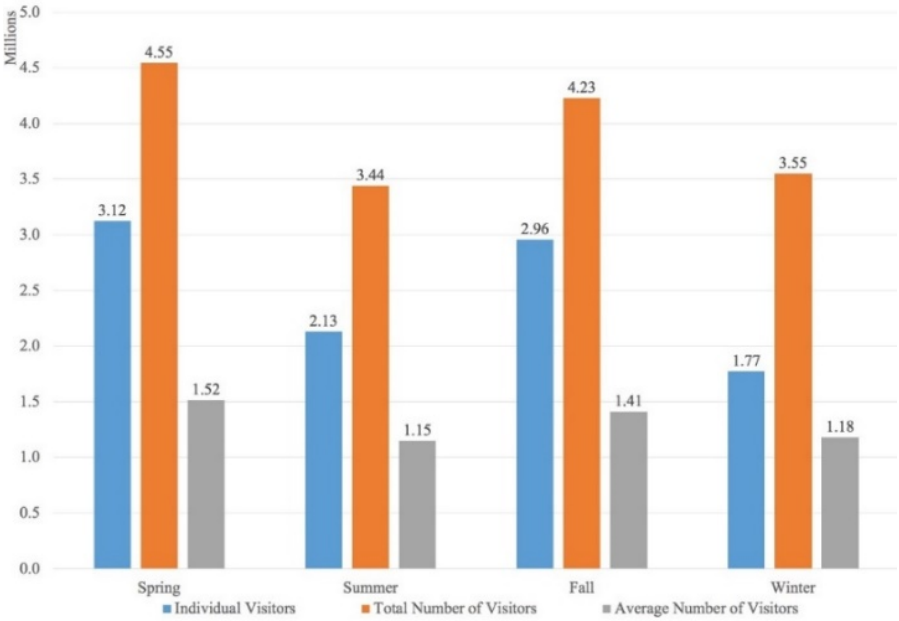

Figure 2.1 Total Number of (Individual) Visitors by Seasons for 3 Years (Jan. 2015- Dec. 2017) and Average Number of Visitors by Season & per Year (in Millions)

The data from national data centers and third-party databases (popular search engine and market analysis engine) include air quality, traffic, alerts, buzz in social media, and local competitors (see Table 2.4). Subsequently, we ran a field experiment to test the performance of the models for roughly six months, from November 1, 2018, to April 28, 2019.

Table 2.4 Variables, Source of Variables, Number of Variables, Format of Variables

Source of Variables/ Variables	# of Variables	The initial format of Variables
Theme Park private data		
Day-Specific variables	55 variables	Text/ Numeric
Event-related variables	5 variables	Numeric
Discount related variables	2 variables	Numeric
Discount/ Package variables	121 variables	Numeric
Customer-related variables	19 variables	Numeric
Reservation related variables	2 variables	Numeric
Weather-related variables	42 variables	Numeric
National Air Quality Monitoring Center		
Air Quality variables	5 variables	Numeric/ Geographic
National Transport Information Center		

Traffic-related variables	2 variables	Numeric/ Geographic
3rd Party Private data (Daum/ Cheil Inc.)		
Buzz (Blog + Search) related variables	14 variables	Text/ Numeric

To provide a more comprehensive summary, we combined training data and some of the testing data. The results are shown in Table 2.5. First, the total number of weekends in each season from 2015 to 2018 is very similar (around 104 days per season). The number of holidays in summer was slightly lower than in other seasons. Similar to the trend of visitors in Figure 2.1, the total number of individual visitors in spring and fall was 32.47% (3.99M) and 32.16% (3.95M), respectively; while the summer and winter were 21.99% (2.70M) and 13.38% (1.64M), respectively (Table 2.5). There is a significant amount of inter-day variation in every season, which is significantly higher than the variation in visits across seasons. For example, the average number of visitors per day in spring is 15,736, while the standard deviation was 12,109.8. This huge variance of the daily mean for visitors makes it difficult for the theme park to accurately predict the number of daily visitors. The ratio of smart reservations, similar to the fast-track tickets of Disney and Universal Studios, was around 10 percent overall. Word-of-mouth on Instagram totaled 1.5M in the past four years. The most mentioned season is fall at 35.37% and immediately following respectively, in spring, summer, and winter at 29.89%, 22.15%, and 12.59%.

Table 2.5 Summary Analysis for Selected Variables

	Spring 2015-2018	Summer 2015-2018	Fall 2015-2018	Winter 2015-2018
# of Days (2015-2018)	368 days	368 days	364 days	361 days
# of Weekends (2015-2018)	105 days	104 days	104 days	105 days
# of Holidays (2015-2018)	16 days	8 days	18 days	17 days

Total Individual Visitors (2015-2018)	3,991,205	2,702,930	3,953,329	1,644,210
Total Annual Member Visitors (2015-2018)	1,799,593	1,703,569	1,627,905	1,137,496
Average Daily # of Visitors (Std. deviation)	15,736 (12,109.8)	11,974 (7,261.3)	15,333 (10,749.8)	7,706 (6,613.4)
Avg. Daily % Smart Reservation (Fast Track)	7.96%	11.59%	7.48%	15.38%
Avg. High Temp. (°C)	19.49	30.49	20.34	4.81
Avg. Low Temp. (°C)	7.17	21.54	9.90	-4.98
Total Rainfall / Season & Year (mm)	237.50	627.08	183.35	80.30
Avg. PM10 (mg/m3)	63.81	37.04	39.73	68.15
Avg. daily # of Blogs about Air Quality	16,653	2,700	3,159	3,894
Avg. daily # of Theme Park Buzz on Instagram	1,234	914	1,476	530

2.4 Methodology

This section first discusses the deep neural network algorithms, the generic artificial intelligent systems, we mainly focus on. Then we outline other traditional econometric models for demand prediction.

2.4.1 Deep-learning Neural Network

We employ deep learning methods (LeCun et al. 2015) to develop a demand prediction model. A deep-learning neural network algorithm represents a learning architecture composed of multiple layers of neurons, which generate nonlinear data transformations (Guo et al. 2018; Liu et al. 2019). Each input variable $(x_1, x_2, \dots\dots, x_{n-1}, x_n)$ in the previous layer (output of the previous layer or raw data in the first layer) passes its information into neurons in the next layer. Hence, each module abstracts higher-level representations from lower-level information or raw data. For example, a piece of voice information can be converted to higher-level features representing tone, speed, or interval from lower-levels of

amplitude and frequency between neuron layers. Each neuron processing input values with the mathematical functions (f) throughout the layers, as shown in Figure 2.2. For example, linear function, sigmoidal function, hyperbolic tangent function, softmax function, or ReLURectified linear unit function may be applied in each neuron (Equation 2.1 to 2.6), and the processed information (weighted information) in the neurons of the hidden layers pass onto the output layer neurons (Nwankpa et al. 2018).

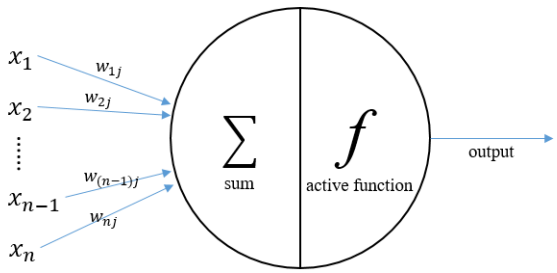

Figure 2.2 Input, Neuron, & Output

$$f(u) = \sum_{i=1}^{n} w_{ij}x_j + b_j \quad (2.1)$$

Sigmoid function: $f(u) = \frac{1}{1+\exp(-cu)}$, where c is constant value (2.2)

Linear function: $f(u) = u$ (2.3)

Hyperbolic Tangent function: $f(u) = \tanh(cu)$, where c is constant value (2.4)

Softmax function: $f(u) = \frac{\exp(u/T)}{\sum_k \exp(u/T)}$, where k is # of class (2.5)

ReLuRectified linear unit: $f(u) = \max(0, u)$ (2.6)

The weights in the neural network are learned through backpropagation that works in the following manner. Backpropagation starts with randomly selected weights from the first variables to the neurons of the hidden layer. It then adjusts the weights to gradually reduce the error (using gradient descent) between the predicted results calculated through each layer and the results to be obtained (Lillicrap et al. 2016). In this study, for instance, we utilize a normalized level of particulate matter concentrations ($PM_{2.5}$ or PM_{10}) as one of the inputs. The information may be transformed with one of the activation functions (e.g., softmax function). Backpropagation starts from the last layer's last parameter (error) of the neuron(s). With the reduced error (bias) and adjusted weights, each activation function's outputs pass information, leading to the last neuron, daily demand.

We decide to use three hidden layers and 30 to 45 neurons over the three hidden layers. In academia, scholars agree that deep-learning neural network algorithms should be able to extract the most efficient informative features and patterns or stack up with sufficient statistics with a minimal number of layers and a minimal number of neurons within each layer (Tishby & Zaslavsky 2015; Zhang et al. 2014). But, finding a minimal number of layers and neurons within the neurons can be problematic because a deep-learning neural network may perform badly even if the training results return negligible errors. For example, using more neurons per layer to detect more patterns in the raw data; however, the more neurons use in each layer may increase the risk of over-fitting by fitting noise relationships as patterns. Also, an extravagantly large number of neurons in the hidden layers can increase training time in each layer. And that leads to the point that it is impracticable to train the data with algorithms suitably.

To reduce such problems, prior studies in computer science have been suggesting three solutions. The first solution is the most frequently used model selection technique – trial and error (Panchal et al. 2011). Based on the findings from Panchal et al. (2011) study, we compared AIC values with the one layer to five layers cases with one to thirty neurons in each layer. And select the optimal number of neurons and layers which provide the best AIC values.

Second, a dropout from designated layers during training or dropping a particular connection is a popular methodology for finding the optimal number of neurons and layers in computer science (Gal & Ghahramani 2016; Kubo et al. 2016). In dropout, the hyperparameter p (the probability of retaining a hidden unit in the network) is one of the critical variables to select. Prior research shows that the test errors were obtained as a function of the hyperparameter p. Few neurons may survive with small retaining probability, leading to high training errors (Srivastava 2013). A general dropout rate suggested is a probability of 0.5 for retaining with the p range of 0.5 and 0.8 (Srivastava et al. 2014).

Finally, batch normalization is one of the popular methods applied to any layer within the algorithm for solving the number of neurons and layers problem (Ioffe & Szegedy 2015; Ma & Klabjan 2017). Batch normalization is a technique to normalize the input variables and hidden neurons by adjusting and scaling the activations even as the mean and variance change over time during training. By applying batch normalization, studies show that the deep-learning neural network computes the results much faster and more stable with prediction improvement (Ioffe & Szegedy 2015; Santurkar et al. 2018). Hence, we applied batch normalization in each hidden layer.

General-purpose AI (Deep-learning neural network):

The past few years have witnessed a proliferation of "turnkey" AI solutions provided by many software companies (Carvalho et al. 2019). For example, cloud-based off-the-shelf AI (including deep learning) technologies developed by Google, IBM, Amazon, and Microsoft have been positioned as "ready-to-use" technologies for decision-makers as universal approximators in many different fields (Dwivedi et al. 2019). These products offer an easy way to enable AI capabilities within organizations by providing boxed software, typically driven by deep learning, with various input variables and a target variable. The software trains the neural net to learn the mapping from the inputs to the output variable. For example, cloud AI platforms have been used to solve various problems such as object recognition, forecasting, sentiment extraction, simulation, and optimization (Carvalho et al. 2019). GPAI is often marketed as solutions that can be deployed in lieu of data scientists. The dearth of good AI/ML talent has gained significant adoption in the industry (Dwivedi et al. 2019).

In this study, our partner organization also used an off-the-shelf deep learning-based solution that works on all the theme park data. The benchmark GPAI system used at the theme park is an off-the-shelf deep-learning system provided by a major US solution company. This GPAI usually performs well on big data with many training sets. The GPAI was trained using cumulative data from January 2012, and the first prediction was made in January 2015.

2.4.2 Commonly used Econometric Approaches for Demand Prediction

Scholars have tested various econometric models to find a superior forecasting model in nonlinear and non-stationary situations like tourism demand (Peng et al. 2014; Song et al. 2010; Yu & Schwartz 2006). So, we examine whether the machine learning algorithms perform better compared to other econometric forecasting models. We initially estimate commonly used twelve econometric models[6]. Of the twelve, we select the best five models.

A. Seasonal Naïve model (Naïve 2): The number of visitors to the theme park varies from day to day as well as across the season. Hence, we used the Seasonal Naïve model (1) with a drift method (2). The seasonal Naïve model is widely used, but a simple model that is employed when there is a highly seasonal trend present in the data (Chan et al. 1999).

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}, \quad (2.7)$$

$$\hat{y}_{T+h|T} = y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1}), \quad (2.8)$$

where T is the last observed time, and h is the period for prediction, m is the seasonal period, and k is the number of complete years in the forecast period prior to time T+h, and $y_1, y_2, \ldots, y_T$ are the number of visitors to the theme park in historical data.

B. Simple Exponential Smoothing (SES) model: Simple Exponential Smoothing (SES) is suitable for forecasting data with no clear trend or seasonal pattern (Chen et al. 2008). We have seasonal patterns; however, we will utilize SES forecasting to compare with other forecasting models. The philosophy of the Seasonal Naïve forecasting model shows that the last observation is the most vital variable to predict future prediction (h = 1, 2, ……), while SES is a model of estimating by weighting more or less from most recent to

[6] 12 Models: Naïve 1, Seasonal Naïve, SMA, MAMA, SES, HDES, BDES, H-WDES, ARIMA, BSTS, OLS, & GARCH

the oldest observations (3). After applying our data in this model, we optimize the prediction using the SIMPLEX algorithm by finding one smoothing constant (α), minimizing the mean squared errors of the forecasting model.

$$\hat{y}_{T+h|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots, \quad (2.9)$$

$$\hat{y}_{T+h|T} = \alpha y_T + (1-\alpha)\left[\hat{y}_{T|T}\right], \quad (2.10)$$

where α is a smoothing constant ($0 \leq \alpha \leq 1$).

C. Holt-Winter's Double Exponential Smoothing (H-WDES) model: The core of H-WDES is the same as Single Exponential Smoothing models (SES); however, they have better prediction data with a seasonal trend. In our case, we are forecasting the everyday demand of a theme park and observing a seasonality within a week, such as high demand on Friday, Saturday, and Sundays. Hence, we apply our data in these H-W DES models. After applying our data in this model, we optimize the prediction using the GRG Nonlinear algorithm by finding two smoothing constants (α, β, and γ), minimizing mean squared errors of the forecasting model.

$$L_8 = \frac{y_8}{S_1} \text{ and } Tr_8 = \frac{y_8}{S_1} - \frac{y_7}{S_7}, \quad (2.11)$$

when subscript 1 represents Sunday of the first week, 7 represent Saturday of the first week, and 8 represent Sunday of the second week.

$$S_i = \frac{y_i}{\left(\sum_{j=1}^{7} y_j / 7\right)}, \quad (2.12)$$

where i = (1, 2, 3, … , 7) represents Sunday to Saturday of first week.

$$L_T = \alpha\left(\frac{y_T}{S_{T-7}}\right) + (1-\alpha)(L_{T-1} + Tr_{T-1}), \quad (2.13)$$

$$Tr_T = \beta(L_T - L_{T-1}) + (1-\beta)\, Tr_{T-1}, \quad (2.14)$$

$$S_T = \gamma\left(\frac{y_T}{L_T}\right) + (1-\gamma)S_{T-7}, \quad (2.15)$$

$$\hat{y}_{T+h|T} = (L_T + h\, Tr_T)S_{T-7+h}. \quad (2.16)$$

D. Bayesian Structural Time Series (BSTS) model: Bayesian Structural Time Series (BSTS) model is a popular model for feature selection, time series forecasting, and inferring causal relationships (Brodersen et al. 2015). The previous three models make it possible to extrapolate previous data in the series to predict future trends. Although time series approaches are useful tools in tourism demand forecasting, their major limitation is that their construction is not based on any economic theory that underlines tourists' decision-making processes. In forecasting demand for theme parks, there are important variables such as weather, promotion, air quality, and other issues and the existing historical trends. BSTS can take into account historical trends, seasonality, regression, and other common time series factors (Jammalamadaka et al. 2019). The Bayesian Structural Time Series we used here is based on Scott & Varian Model (2013a; 2013b)

$$\hat{y}_t = \mu_t + \tau_t + \beta^T x_t + \varepsilon_t \,, \qquad \varepsilon_t \sim N(0, H_t) \tag{2.17}$$

$$\mu_{t+1} = \mu_t + \delta_t + \eta_{0t} \,, \qquad \eta_t \sim N(0, Q_t) \tag{2.18}$$

$$\delta_{t+1} = \delta_t + \eta_{1t}, \tag{2.19}$$

$$\tau_{t+1} = -\sum_{S=1}^{S-1}(\tau_t + \eta_{2t}), \tag{2.20}$$

with three state components: a trend μ_t, a seasonal pattern τ_t, & regression component $\beta^T x_t$, and δ_t is the amount of extra trend. ε_t and η_t are idiosyncratic shocks independent of everything else.

The basic Bayesian Structural Time Series model has two equations: the observation equation and the Transition equation, which is very similar to the equation we used.

Observation (15) and Transition (16) equations:

$$y_t = Z_t^T \alpha_t + \varepsilon_t \,, \qquad \varepsilon_t \sim N(0, H_t) \tag{2.21}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \qquad \eta_t \sim N(0, Q_t) \tag{2.22}$$

where α_t is a latent variable known as state, and Z_t, T_t, & R_t are Structural parameters. ε_t and η_t are idiosyncratic shocks independent of everything else.

E. Ordinal Least Square (OLS) forecasting model: The Ordinal Least Squares (OLS) forecasting model provides the minimum value for the sum of squared errors, finding the best-estimated coefficient and model for the given data. Like BSTS, the OLS prediction model seeks to find a dependent relationship between tourism demand and a set of explanatory variables rather than extrapolation. Hence, Tourism forecasts can then be produced as a function of the value taken by these explanatory variables. However, the OLS model that contains a trend series tends to create a spurious correlation between dependent and independent variables, thus invalidating the model's diagnostic statistics.

$$\hat{y}_t = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_{it}, \quad (2.23)$$

where t is time, i is the number of independent variables considered for prediction, and k are all the independent variables.

2.5 Results

2.5.1 Error and Accuracy Comparison of Different Prediction Methodologies

In this section, we compare the predictive performance of the different techniques presented in Section 4 using measures of MAPE, MSE, and accuracy. The theme park can easily accommodate up to 20% deviations from predictions, and the errors within this range do not lead to any meaningful financial loss for the park. Hence, we also present the accuracy by considering a contingency 20% error threshold. The predictive performance of all the models is shown in Table 2.6.

Table 2.6 Prediction Results Comparisons

Prediction Model	MAD	MSE	MAPE	Accuracy w/ Threshold ± 20% error
General-purpose AI	2253.348	10293993	46.54%	73.74% (132/179)
Naïve 2 Model	2912.905	18882751	49.34%	70.39%

Simple Exponential Smoothing (SES)	2776.425	19812413	51.00%	(126/179) 68.72%
Holt-Winter's DES	2930.535	19137843	62.95%	(123/179) 64.80%
Bayesian Structural Time Series	2284.227	12152605	42.09%	(116/179) 75.42%
OLS Regression	2067.369	7199708	45.21%	(135/179) 75.42% (135/179)
Other Econometric Models				
Naïve 1 Model	277133	20426189	50.22%	70.39% (126/179)
Simple Moving Average	3207.534	23585817	65.28%	61.45% (110/179)
Moving Average of Moving Average	3589.566	27143186	78.71%	58.10% (104/179)
Holt's Double Exponential Smoothing (DES)	3133.723	19979172	72.20%	59.22% (106/179)
Brown's Double Exponential Smoothing (DES)	2932.729	18735881	63.94%	67.04% (120/179)
ARIMA	3243.981	20803800	70.72%	62.01% (111/179)
Generalized Autoregressive Conditional Heteroscedasticity	3070.391	25432765	54.97%	70.95% (127/179)

Of the twelve, we select the best five models (we still report all twelve results). The machine learning algorithm's performance (MAPE: .4654 & accuracy: 73.74%) is comparable to or better than most forecasting models. More specifically, GPAI does not produce better forecasting results in MAPE and accuracy than BSTS (.4209 & 75.42%)[7] and OLS (.4521 & 75.42%)[8]. It does show better predictability as measured by MSE than all other benchmark models except OLS (2683.23). For example, the HW-DES has a higher MAPE of 62.95% among the selected five models, while also being slightly more inaccurate from a business decision-making perspective (64.80%).

[7] t-statistics for the mean difference of APE for GPAI and BSTS is .7866 with p-value: .4321

[8] t-statistics for the mean difference of APE for GPAI and OLS is .2459 with p-value: .8059

The worst econometric models in the daily demand forecasting for the theme park are moving average of moving average (MAMA) and Holt's Double Exponential Smoothing (DES). For example, the MAPE of MAMA and DES are 78.71% and 72.20%, respectively. And, the accuracy of each model is 58.10% and 59.22%, respectively. These results mean that daily demand is reasonably scattered and nonlinear. Therefore, the machine learning approach is a better way to predict these nonlinear and non-stationary forecasting. However, demand forecasting using machine learning did not excel at forecasting using other traditional econometric models, but marginally better. Significantly, the performance of demand prediction of GPAI is still not particularly far off from that of BSTS and OLS. These results suggest that the number of visitors to the theme park is not well-predicted through past time series trends. But, a comprehensive consideration of diverse variables and past visitor trends enables more accurate demand forecasting for machine learning algorithms. Also, non-parametric approaches such as DNN, BSTS, and OLS better represent nonlinear and non-stationary future trends.

2.5.2 Logistic Regression Comparison of Different Prediction Methodologies

In this section, we compare the predictive performance of the different techniques presented in Section 4 using measures of logistic regression. Again, the theme park can easily accommodate up to 20% deviations from predictions, and the errors within this range do not lead to any significant financial loss for the park. Hence, we also present the accuracy by considering a contingency 20% error threshold. The equations we used for this logistic regression is presented from the equation 2.24 to 2.26

Since the dependent variable is dichotomous, representing "whether each model correctly forecasts or not in a given situation," we decide to use logistic regression to investigate the hypothesized relationship. The mapping between each model's latent utility (μ_i) and the observed predictability y_i is:

$$y_i = \begin{cases} 1, |Visitor_i - Forcasting_i| \leq 20\% * Visitor_i \\ 0, Elsewhere \end{cases} \quad (2.24)$$

where i is specific date.

$$y_i \sim Bernoulli(\mu_i), \quad (2.25)$$

We assume that latent utility (μ_i) is formed by a linear combination of our determinants which measure the visitors with less than 5,000, visitors with more than 16,000, and twelve models with the visitors' situations. Hence, we can specify our latent model as following:

$$\mu_i = logistic(\beta_0 + \sum_k \sum_i \beta_k x_{ik}), \quad (2.26)$$

, where $\beta_{k\,or\,0} \sim N(M_i, S_i)$ and k is each determinant

and $logistic(\beta_0 + \sum_k \sum_i \beta_k x_{ik}) = \frac{1}{(1+e^{-(\beta_0+\sum_k \sum_i \beta_k x_{ik})})}$

The predictive performance of all the models is shown in Table 2.7.

Table 2.7 Bayesian Logistic Regression Results

	Bayesian Logistic regression Accuracy w/ Threshold ± 20% error					
	Mean	ESS	Odds Ratio	95% HDI (Low, High)	P-value	Sig.
Main Variables						
Below 5,000 Dummy	.700	2703.9	2.0146	(.4189, .9744)	.0001	***
Over 16,000 Dummy	-1.173	4566.9	.3095	(-1.512, -.8415)	.0000	***
Below 5 X Naïve 1 Dummy	.553	4464.1	1.7390	(.1840, .9307)	.0039	***
Below 5 X Naïve 2 Dummy	.400	4181.8	1.4911	(.0432, .7621)	.0303	**
Below 5 X SMA Dummy	.074	5180.3	1.0772	(-.3287, .4952)	.7250	

Below 5 X MAMA Dummy	-.026	2072.4	.9746	(-.4087, .3598)	.9085	
Below 5 X SES Dummy	.521	3046.7	1.6840	(.1569, .8859)	.0058	***
Below 5 X HDES Dummy	-.521	3818.2	.5918	(-.8575, -1956)	.0020	***
Below 5 X BDES Dummy	-.034	4189.2	.9668	(-.3455, .2911)	.8380	
Below 5 X HWDES Dummy	.093	5166.9	1.0973	(-.2254, .4129)	.5692	
Below 5 X ARIMA Dummy	-1.062	6496.5	.3457	(-1.368, -.7425)	.0000	***
Below 5 X BSTS Dummy	1.446	8317.7	4.2478	(1.1242, 1.775)	.0000	***
Below 5 X OLS Dummy	-.125	12513.6	.8827	(-.4632, .2218)	.4752	
Below 5 X GARCH Dummy	1.330	26587.3	3.7812	(.9713, 1.705)	.0000	***
Over 16 X Naïve 1 Dummy	-.942	11123.2	.3897	(-1.509, -.3734)	.0013	***
Over 16 X Naïve 2 Dummy	-.776	11951.2	.4603	(-1.384, -.1764)	.0139	**
Over 16 X SMA Dummy	-1.328	9034.3	.2651	(-1.829, -.8147)	.0000	***
Over 16 X MAMA Dummy	-1.206	4116.4	.2994	(-1.715, -.6927)	.0004	***
Over 16 X SES Dummy	.430	5161.7	1.5370	(-.1596, 1.0141)	.1432	
Over 16 X HDES Dummy	-.797	3590.9	.4506	(-1.618, -.0177)	.0581	*
Over 16 X BDES Dummy	.265	3339.5	1.3029	(-.6075, 1.1247)	.5687	
Over 16 X HWDES Dummy	.407	5025.0	1.5023	(-.3397, 1.1309)	.2920	
Over 16 X ARIMA Dummy	-.534	5702.1	.5861	(-1.300, .23302)	.1757	
Over 16 X BSTS Dummy	1.87	6968.1	6.5124	(1.1998, 2.552)	.0000	***
Over 16 X OLS Dummy	.143	21847.9	1.1537	(-.3262, .6195)	.5564	
Over 16 X GARCH Dummy	-3.225	66431.7	.0398	(-4.565, -2.023)	.0000	***

| Bayesian HDI (Highest Density Interval): the distribution by specifying an interval that spans most of the distribution, 95%, such that every point inside the interval has a higher

|| Runed MCMC methods using a Metropolis-Hastings algorithm with a random walk chain for 100,000 iterations.

||| ***, **, and *: Significance at 1%, 5%, and 10% level, respectively in frequentist approach

The machine learning approach has greater odds of forecast right for the visitors less than 5,000 but lower odds of forecasting right for the visitors less than 16,000. Interestingly, with less than 5,000 visitors – mostly weekdays, we can observe that other time-series econometric models outperform better odds ratio than the machine learning approach statistically. It seems to be because most of the days with less than 5,000 visitors have stationary trends. On the other hand, in the cases of more than 16,000 visitors – often weekends and holidays, non-parametric demand predictions such as machine learning and BSTS are statistically significant than time-series models.

2.6 Conclusion

This chapter compares forecasting results of daily demand for a theme park by machine learning algorithms, DNN, and other econometric models. The daily demand for theme parks has nonlinear and non-stationary characteristics. So, it has been argued which predictive model best predicts demand with such nonlinear and non-stationary features.

Overall, the demand prediction with non-parametric methods such as machine learning, Bayesian structural time series, and ordinal least-square are superior to forecast nonlinear and non-stationary characteristics. For example, the MAPEs of machine learning, BSTS, and OLS are 46.54%, 42.09%, and 45.21%, respectively. Compared to the MAPEs, those of parametric methods – Naïve 2, SES, Naïve 1, H-W DES are 49.34%, 51.00%, 50.22% and 62.95%, respectively. These results suggest that the number of visitors to the theme park is not well-predicted through past time series trends. Yet, a non-parametric approach utilizing the function of more data and parameters might better predict the nonlinear and non-stationary future trends.

To be more specific, we investigated further. In the cases of more stationary and linear visitors (5,000 or fewer), parametric modeling of time series is statistically significant at predicting demand than the machine learning approach, even though the machine learning approach is superior overall. Conversely, the machine learning methodology was statistically superior to other time-series analyses in cases where the number of visitors changed suddenly (more than 16,000) and non-linear are.

Surprisingly, demand forecasting using machine learning did not outperform other traditional econometric models huge but marginally better. It might be because the

underlying neural network needs to learn thousands of parameters. The training data in this situation might not be sufficient to train the model, even though it extends several years. Also, because we have tremendous data; thus, the search space is too large for the training algorithm to perform well without any form of human supervision (and training). Therefore, we will study in-depth in the next chapter how humans, as complementary to machine algorithms, can help in nonlinear and non-stationary demand predictions along with machine learning.

In conclusion, the machine learning approach for nonlinear and non-stationary demand forecasting outperforms traditional econometric models. Non-parametric econometric models such as BSTS and OLS also exhibit statistically similar performance to machine learning in nonlinear and unstructured demand predictions. However, in more linear and stationary demand forecasting cases, parametric time series models perform better in demand forecasting than machine learning.

Reference:

Acemoglu, D., & Restrepo, P. (2018). Artificial intelligence, automation and work (No. w24196). National Bureau of Economic Research. Available at NBER: https://www.nber.org/papers/w24196

Aghion, P., Jones, B. F., & Jones, C. I. (2017). Artificial intelligence and economic growth (No. w23928). National Bureau of Economic Research. Available at NBER: https://www.nber.org/papers/w23928

Agrawal, A., Gans, J., & Goldfarb, A. (2016). The simple economics of machine intelligence. Harvard Business Review, 17, 2-5.

Agrawal, A., Gans, J., & Goldfarb, A. (2017). How AI will change the way we make decisions. Harvard Business Review. July 26, Available at: https://hbr.org/2017/07/how-ai-will-change-the-way-we-make-decisions

Agrawal, A., Gans, J., & Goldfarb, A. (2018). Is Your Company's Data Actually Valuable in the AI Era. Harvard Business Review. January 17. Available at: https://hbr.org/2018/01/is-your-companys-data-actually-valuable-in-the-ai-era

Agarwal, R., & Dhar, V. (2014). Editorial - Big data, data science, and analytics: The opportunity and challenge for IS research, Information Systems Research, 25(3), 443-448.

Aleksander, I. (2017). Partners of humans: a realistic assessment of the role of robots in the foreseeable future. Journal of Information Technology, 32(1), 1-9.

Bansal, A., Kauffman, R. J., & Weitz, R. R. (1993). Comparing the modeling performance of regression and neural networks as data quality varies: A business value approach. *Journal of Management Information Systems*, *10*(1), 11-32.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, *9*(1), 247-274.

Brynjolfsson, E., Hui, X., & Liu, M. (2019). Does machine translation affect international trade? Evidence from a large digital platform. Management Science, Article in Advance, 1-12.

Burger, C. J. S. C., Dohnal, M., Kathrada, M., & Law, R. (2001). A practitioners guide to time-series methods for tourism demand forecasting—a case study of Durban, South Africa. Tourism management, 22(4), 403-409.

Carvalho, A., Levitt, A., Levitt, S., Khaddam, E., & Benamati, J. (2019). Off-the-shelf artificial intelligence technologies for sentiment and emotion analysis: a tutorial on using IBM

natural language processing. *Communications of the Association for Information Systems*, *44*(1), article 43.

Chan, Y. M., Hui, T. K., & Yuen, E. (1999). Modeling the impact of sudden environmental changes on visitor arrival forecasts: The case of the Gulf War. Journal of Travel Research, 37(4), 391-394.

Chen, K. Y., & Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. Tourism Management, 28(1), 215-226.

Chen, R. J., Bloomfield, P., & Cubbage, F. W. (2008). Comparing forecasting models in tourism. Journal of Hospitality & Tourism Research, 32(1), 3-21.

Cowgill, B. (2018). Bias and productivity in humans and algorithms: Theory and evidence from resume screening. Columbia Business School, Columbia University, August 29.

Cui, G., Wong, M. L., & Lui, H. K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. Management Science, 52(4), 597-612.

Davenport, T. H., Libert, B. & Beck, M. (2018). Robo-Advisers Are Coming to Consulting and Corporate Strategy. Harvard Business Review, January 12. Available at: https://hbr.org/2018/01/robo-advisers-are-coming-to-consulting-and-corporate-strategy

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard Business Review, 96(1), 108-116.

De Leon, A. R., Soo, A., & Williamson, T. (2011). Classification with discrete and continuous variables via general mixed-data models. Journal of Applied Statistics, 38(5), 1021-1032.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1), eaao5580.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 101994.

Flam, F. (2018). IBM's Watson Hasn't Beaten Cancer, But AI Still Has Promise. Bloomberg Opinion. Aug 24, Available at: https://www.bloomberg.com/opinion/articles/2018-08-24/ibm-s-watson-failed-against-cancer-but-a-i-still-has-promise

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In Advances in neural information processing systems (pp. 1019-1027).

Gartner. (2015, September 15). Gartner Says Business Intelligence and Analytics Leaders Must Focus on Mindsets and Culture to Kick Start Advanced Analytics. Available at: https://www.gartner.com/en/newsroom/press-releases/2015-09-15-gartner-says-business-intelligence-and-analytics-leaders-must-focus-on-mindsets-and-culture-to-kick-start-advanced-analytics

Gartner. (2018, February 13). Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence. Available at: https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence

Goh, C., Law, R., & Mok, H. M. (2008). Analyzing and forecasting tourism demand: A rough sets approach. Journal of Travel Research, 46(3), 327-338.

Grover, V., Chiang, R. H., Liang, T. P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. Journal of Management Information Systems, 35(2), 388-423.

Gutierrez, R. S., Solis, A. O., & Mukhopadhyay, S. (2008). Lumpy demand forecasting using neural networks. *International journal of production economics*, *111*(2), 409-420.

Guo, J., Zhang, W., Fan, W., & Li, W. (2018). Combining geographical and social influences with deep learning for personalized point-of-interest recommendation. *Journal of Management Information Systems*, *35*(4), 1121-1153.

Hosanagar, K., & Saxena, A. (2017). The First Wave of Corporate AI Is Doomed to Fail. Harvard Business Review. April 18, Available at: https://hbr.org/2017/04/the-first-wave-of-corporate-ai-is-doomed-to-fail

Hu, M., & Song, H. (2019). Data source combination for tourism demand forecasting. Tourism Economics, 1354816619872592.

Huang, M. H., & Rust, R. T. (2018). Artificial intelligence in service. Journal of Service Research, 21(2), 155-172.

IDC (2019, July 19). This Week In AI Stats: Up To 50% Failure Rate in 25% of Enterprises Deploying AI, Forbes, Available at: https://www.forbes.com/sites/gilpress/2019/07/19/this-week-in-ai-stats-up-to-50-failure-rate-in-25-of-enterprises-deploying-ai/#c02d09e72ce4 /IDC (2019, July 8): https://www.idc.com/getdoc.jsp?containerId=prUS45344519&utm_medium=rss_feed&utm_source=Alert&utm_campaign=rss_syndication

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Jammalamadaka, S. R., Qiu, J., & Ning, N. (2019). Predicting a stock portfolio with the multivariate Bayesian structural time series model: do news or emotions matter?. *International Journal of Artificial Intelligence*, *17*(2), 81-104.

Kleber, S. (2018). As AI Meets the Reputation Economy, We're All Being Silently Judged. Harvard Business Review. January 29. Available at: https://hbr.org/2018/01/as-ai-meets-the-reputation-economy-were-all-being-silently-judged

Kubo, Y., Tucker, G., & Wiesler, S. (2016). Compacting neural network classifiers via dropout training. arXiv preprint arXiv:1611.06148.

Law, R., Li, G., Fong, D. K. C., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of tourism research*, *75*, 410-423.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. Nature communications, 7(1), 1-10.

Lin, Y. K., Chen, H., Brown, R. A., Li, S. H., & Yang, H. J. (2017). Healthcare predictive analytics for risk profiling in chronic care: A Bayesian multitask learning approach. MIS Quarterly, 41(2). 473-496 & A1-A3.

Liu, X., Zhang, B., Susarla, A., & Padman, R. (2019). Go to YouTube and call me in the morning: use of social media for chronic conditions. *Forthcoming, MIS Quarterly*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3061149

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. Journal of Strategic Information Systems, 24(3), 149-157.

Lu, Y., Qian, D., Fu, H., & Chen, W. (2018). Will supercomputers be super-data and super-AI machines?. Communications of the ACM, 61(11), 82-87.

Ma, Y., & Klabjan, D. (2017). Convergence analysis of batch normalization for deep neural nets. arXiv preprint arXiv:1705.08011.

Mayer-Schönberger, V. & Ramge, T. (2018). Are the Most innovative Companies Just the Ones With the Most Data?. Harvard Business Review. February 7, Available at: https://hbr.org/2018/02/are-the-most-innovative-companies-just-the-ones-with-the-most-data

McAfee, A. & Brynjolfsson, E. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Meyer, G., Adomavicius, G., Johnson, P. E., Elidrisi, M., Rush, W. A., Sperl-Hillen, J. M., & O'Connor, P. J. (2014). A machine learning approach to improving dynamic decision making. Information Systems Research, 25(2), 239-263.

Miklós-Thal, J., & Tucker, C. (2019). Collusion by Algorithm: Does Better Demand Prediction Facilitate Coordination Between Sellers?. Management Science, 65(4), 1552-1561.

Monroe, D. (2018). Chips for artificial intelligence. Communications of the ACM, 61(4), 15-17.

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*.

Panchal, G., Ganatra, A., Kosta, Y. P., & Panchal, D. (2011). Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. International Journal of Computer Theory and Engineering, 3(2), 332-337.

Peng, B., Song, H., & Crouch, G. I. (2014). A meta-analysis of international tourism demand forecasting and implications for practice. Tourism Management, 45, 181-193.

Pistrui, J. (2018). The future of human work is imagination, creativity, and strategy. Harvard Business Review, 18. January 18. Available at: https://hbr.org/2018/01/the-future-of-human-work-is-imagination-creativity-and-strategy

Pumplun, L., Tauchert, C., & Heidt, M. (2019). A New Organizational Chassis for Artificial Intelligence-Exploring Organizational Readiness Factors (No. 112582). Darmstadt Technical University, Department of Business Administration, Economics and Law, Institute for Business Studies (BWL), Available at: https://aisel.aisnet.org/ecis2019_rp/106/

Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2017). Reshaping business with artificial intelligence: Closing the gap between ambition and action. MIT Sloan Management Review, 59(1).

Rock, D. (2019). Engineering Value: The Returns to Technological Talent and Investments in Artificial Intelligence. Working paper, May. 1. Available at SSN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3427412

Sachs, J. D., & Kotlikoff, L. J. (2012). Smart machines and long-term misery (No. w18629). National Bureau of Economic Research. Available at NBER: https://www.nber.org/papers/w18629

Santurkar, S., Tsipras, D., Ilyas, A., & Madry, A. (2018). How does batch normalization help optimization?. In Advances in Neural Information Processing Systems (pp. 2483-2493).

Schaeffer, J., & Simon, H. (1992). The Game of Chess. In R. J. Aumann & S. Hart (Eds.), Handbook of Garne Theory with economic applications, (Vol 1, pp. 1-17). Amsterdam: Elsevier.

Scott, S. L., & Varian, H. R. (2013a). Predicting the present with Bayesian Structural Time Series. Available at SSRN 2304426.

Scott, S. L., & Varian, H. R. (2013b). Bayesian variable selection for nowcasting economic time series (No. w19567). National Bureau of Economic Research.

Simon, H. A. (1995). Artificial intelligence: an empirical science. Artificial Intelligence, 77(1), 95-127.

Sinha, A. P., & May, J. H. (2004). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. Journal of Management Information Systems, 21(3), 249-280.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. Tourism management, 29(2), 203-220.

Song, H., Li, G., Witt, S. F., & Fei, B. (2010). Tourism demand modelling and forecasting: how should demand be measured?. Tourism economics, 16(1), 63-81.

Srivastava, N. (2013). Improving neural networks with dropout. University of Toronto, 182(566), 7.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. Management science, 38(7), 926-947.

Tambe, P. (2014). Big data investment, skills, and firm value. Management Science, 60(6), 1452-1469.

Tishby, N., & Zaslavsky, N. (2015, April). Deep learning and the information bottleneck principle. In 2015 IEEE Information Theory Workshop (ITW) (pp. 1-5). IEEE.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433-460.

Turing, A. M. (2009). Computing machinery and intelligence. In Parsing the Turing Test (pp. 23-65). Springer, Dordrecht.

Walczak, S. (2001). An empirical analysis of data requirements for financial forecasting with neural networks. Journal of management information systems, 17(4), 203-222.

Wolcott, R. C. (2018). How Automation Will Change Work, Purpose, and Meaning. Harvard Business Review, 1-1. January 11. Available at: https://hbr.org/2018/01/how-automation-will-change-work-purpose-and-meaning

Yan, D., Zhou, Q., Wang, J., & Zhang, N. (2017). Bayesian regularisation neural network based on artificial intelligence optimisation. International Journal of Production Research, 55(8), 2266-2287.

Yu, G., & Schwartz, Z. (2006). Forecasting short time-series tourism demand with artificial intelligence models. Journal of travel Research, 45(2), 194-203.

Zhang, G., & Hu, M. Y. (1998). Neural network forecasting of the British pound/US dollar exchange rate. Omega, 26(4), 495-506.

Zhang, X., Trmal, J., Povey, D., & Khudanpur, S. (2014, May). Improving deep neural network acoustic models using generalized maxout networks. In 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 215-219). IEEE.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker Jr, J. F. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. Journal of Management Information Systems, 20(4), 139-166.

CHAPTER3
Demand Prediction using Machine Learning: A Human-in-the-Loop Approach

3.1 Introduction

This chapter explores the question: do machine algorithms substitute or complement human expertise? Advances in computing power, coupled with large digital datasets, have made it practical to apply artificial intelligence (AI) methods to solve problems in real-time and at scale. Agrawal et al. (2016) argue that reducing the cost and accuracy of predictions has significant implications for productivity, consumer surplus, and firm value. Not surprisingly, venture capitalists invested over $19 billion in AI startups in 2019 alone and established firms invested several times that amount in AI initiatives.[9] Software vendors, such as IBM, Amazon, and SAS, have built the off-the-shelf, "general-purpose AI (GPAI)" systems deployed at many companies.[10]

We would like to draw a distinction between GPAI and General AI. GPAI, in the context of this chapter, represents tools like AutoML by Google and other black box software promoted by companies such as SAS, TIBCO, and Alteryx for enterprise applications of AI (that can be deployed without building custom models). GAI, on the other hand, is AI that thinks and behaves like human beings. Our chapter focuses on GPAI. Researchers have argued that

[9] Forbes, January 5, 2020. Is Venture Capital Investment In AI Excessive? https://www.forbes.com/sites/cognitiveworld/2020/01/05/is-venture-capital-investment-for-ai-companies-getting-out-of-control/#79307a07e050

[10] Forbes, February 20, 2020. Gartner's 2020 Magic Quadrant for Data Science And Machine Learning Platforms Has Many Surprises https://www.forbes.com/sites/janakirammsv/2020/02/20/gartners-2020-magic-quadrant-for-data-science-and-machine-learning-platforms-has-many-surprises/#654e3fa53f55

humans, especially data-driven managers, can be wholly substituted for machine learning (ML) systems in the future (Huang & Rust 2018; McAfee & Brynjolfsson 2012). In reality, these AI initiatives remain early-stage applications in business practice, and their economic value is still unclear. A recent study by IDC shows that the failure rate of AI investments is significantly high.[11] About 80% of enterprises currently engaged in ML/ AI projects have stalled.[12] Hosanagar & Saxena (2017) suggest that frequent failures are due to the complexity of tasks being targeted by these systems. The view of AI/ ML algorithms as a plug-and-play technology with immediate returns is a significant factor in these failures (Fountaine et al. 2019). To address the disappointing results of AI-based systems (particularly ML systems), an approach gaining prominence is Human-in-the-Loop (HITL) design.

The HITL approach uses humans to provide continuous feedback to the ML system in an iterative manner to improve its performance, making these systems more resilient to sudden changes in the data generating process. In this approach, human expertise complements the ML algorithm, rather than substitutes it, as has been argued in the literature (Lake et al. 2015; 2016). Since human experts have domain-specific knowhow, while machines have superior computational power, HITL promises to bring the best of both worlds to improve the prediction accuracy of ML systems. Cognitively, humans find it difficult to process vast

[11] Forbes, July 19, 2019. This week In AI Stats: Up To 50% Failure Rate In 25% Of Enterprises Deploying AI https://www.forbes.com/sites/gilpress/2019/07/19/this-week-in-ai-stats-up-to-50-failure-rate-in-25-of-enterprises-deploying-ai/#6e8f074872ce

[12] Readwrite, August 30, 2020. Challenge of Adopting AI in Businesses https://readwrite.com/2020/08/30/challenges-of-adopting-ai-in-businesses/

amounts of data to identify underlying patterns. Stereotyping or prejudgment by humans can affect data processing and interpretation, increasing uncertainty in finding solutions.

Conversely, in the absence of human intuition and experience, the search space for ML algorithms can be extremely large. The inability to handle bias due to over or under-fitting data can also lead ML algorithms to unclear or erroneous conclusions. Finally, ML algorithms can be quite brittle to rapid changes in the environment. The relevance of the distribution of a future test set to that of the training set increases the predictability of ML algorithms (Brynjolfsson & Mitchell 2017). Interestingly, many ML systems that predict consumer behavior were caught unaware by COVID-19, a shortcoming that HITL might be able to overcome.[13] Hence, exploiting the complementary relationship between human experts and machine algorithms might lead to improvement in business performance.

In this chapter, we focus on apply HITL to demand prediction. Demand prediction is an essential but extremely challenging issue in business and has received considerable attention (Farias et al. 2013; Peng et al. 2014). In our context of theme parks, accurate demand prediction is significant to balance revenues with the costs of operating attractions and customer dissatisfaction from long wait times. In this paper, we propose a deep-learning-based HITL system for predicting demand at one of the largest theme parks in Asia, and compare its performance with experts and an off-the-shelf deep-learning-based GPAI that was deployed at the park. In our study, demand prediction is performed by experts, off-

[13] MIT Technology Review, May 11, 2020. Our weird behavior during the pandemic is messing with AI models https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/

the-shelf GPAI, and HITL. Moreover, we evaluate the demand forecasting results, economic benefits, and indirect economic benefits to the theme park for each model.

We train and compare the models using slightly less than four years of data from January 2015 to Oct. 2018. Subsequently, we ran a field experiment for six months to test the performance of the various models. Then, we assessed the direct and indirect economic benefits of different approaches in 13 restaurants and 21 snack bars. The field experiment shows that HITL AI performs significantly better than both experts and GPAI in predicting the number of daily park visitors, improving prediction accuracy by 14.52% over experts and 20.11% over GPAI. In terms of prediction error, HITL AI has 18.07% and 27.93% lower mean absolute percentage error (MAPE) compared to experts and GPAI, respectively. We also selectively present the best-performing five econometrics models (in the previous chapter) for daily prediction. It is interesting to note that GPAI performs worse than experts during the field experiment and was considered a failure by the management, consistent with the poor performance observed in industry surveys. From a financial perspective, HITL AI improves the theme park's average daily revenues by 4.98% and 5.87% as compared to experts and GPAI, respectively. It also marginally reduces the operating cost of the theme park as compared to the other alternatives. These differences are not only economically significant but also statistically significant.

This chapter makes several contributions. We believe this study is distinctive in demonstrating a complementary relationship between human experts and ML algorithms rather than merely seeing human and machine learning algorithms as exchangeable economic factors. It adds to the nascent but growing literature on the use of ML techniques

to solve business problems and measure the business value of IT for a specific technology. From a practical perspective, it provides an explanation for the high observed failure rate in unmediated AI/ ML, and provides direction for better AI systems design. HITL AI can be valuable in solving complex problems where the solution space is ample or the data are sparse.

The rest of this chapter is organized as follows. The following section reviews related IS literature. Then, we present our research design. This is followed by a section on research methodology and the results. Finally, we discuss implications and provide suggestions for future research.

3.2 Literature Review

The primary objective of our chapter is to explore how machines and humans work together, focusing on how they act as substitutes or complements in our context of demand prediction. While a large number of scholars have argued that ML/ AI algorithms are a "general-purpose technology (GPT)" that substitutes for human performance (Agrawal et al. 2016), researchers have also indicated that these approaches can complement human decision making, though there has been less systematic work from this perspective.

General-purpose technologies are characterized by pervasiveness, potential for improvement over time, and foster complementary innovation (Brynjolfsson et al. 2018). Machine learning exhibits these characteristics. ML/ AI systems have been adopted in a number of industries and a variety of business contexts (Chui et al. 2018), such as predicting customer churn (De Caigny et al. 2018), credit risk (Baesens et al. 2003), and search personalization (Yoganarasimhan 2020). Davenport et al. (2018) point out that ML/AI tools

have become more powerful over time, and are simultaneously both less scarce and cheaper to obtain. Agrawal et al. (2016) argue that reducing prediction cost and improving accuracy contribute to productivity, consumer surplus, and firm value. Finally, scholars have unveiled evidence of ML systems surpassing human-level performance in specific tasks such as image recognition (Krizhevsky et al. 2017), speech recognition (Graves et al. 2013), and gaming (Guo et al. 2014). Ultimately, some argue that AI/ML algorithms will eventually display human-like intelligence, using processes similar to those deployed in humans (Simon 1995).

Acemoglu & Restrepo (2018) have argued that automation by ML/ AI reduces the demand for labor and wages. Research has also shown that ML/ AI may serve to substitute for knowledge workers in addition to production workers. (Loebbecke & Picot 2015; McAfee & Brynjolfsson 2012). Moreover, Frey and Osborne (2013) find that 47% of U.S. employment related to non-routine tasks, high levels of perception or manipulation, creative intelligence, and social intelligence is at high risk of computerization because of machine learning and its application to mobile robotics. Moreover, Mayer-Schönberger & Ramge (2018) argue that the source of innovation is shifting from human ingenuity to data-driven ML.

On the other hand, some scholars have argued that even smart machines will not eliminate all human labor (Davenport & Kirby 2016) because ML/ AI systems do not "think" but instead look for patterns in the data (Bergstein 2017). Specifically, these algorithms are incredibly narrow in their focus on data and instructions fed to the system (Brooks 2017), while human intelligence is more general and intuitive (Frantz 2003). For these reasons, these researchers predict that as the costs of these systems decrease, the demand for labor

with judgment, a high degree of imagination, creative analysis, and strategic thinking will increase (Pistrui 2018), consistent with complementarity.

Peeters et al. (2020) argue that true intelligence lies in the collective approach of intelligent humans and machines working together. Brynjolfsson and Mitchell (2017) consider the "Human-in-the-Loop (HITL)" approach as one where AI acts as an apprentice to assist the human, while also learning by observing and capturing the human's decisions as additional training examples. Schirner et al. (2013) point out that "HITL" is not the classic supervised ML approach. Rather, it is different because it puts humans into continuous physical feedback loops.

Our study adds to the existing literature on the complementary relationship of humans and machines, which have mainly been analyzed either at the level of industry or, more relevant to this paper, at the job level (Huang & Rust 2018; Pistrui 2018). In a recent working paper, de Silva and Thesmar (2021) focus on the complementarity of humans and ML/ AI in predicting corporate earnings. Specifically, they compare the predictions of human forecasters, humans working alongside ML systems, and exclusively data-driven machine (statistics) forecasts. They find that humans outperform machine predictions in a one-year horizon but this result is reversed in longer prediction horizons. Like us, they find that humans and ML together perform significantly better.

We study the complementarity of the two input factors, human experts and machine algorithms, in very different settings allowing us to gain new insights. Corporate earnings are predicted for relatively long horizons of one to two years and are reasonably static. In contrast, demand prediction at a theme park is conducted daily under frequently and widely

changing conditions. Moreover, the dynamic nature of daily forecasting in our study allows us to more deeply examine the complementary relationship between humans and machines. For example, humans are more efficient than algorithms in learning the complex rules of their dynamic environments and can self-adjust (Lake et al. 2015; 2016). On the other hand, machines possess high-performance computational power (Lu et al. 2018). Taken together, as we shall see later, this leads to comparative advantages for both input factors in the underlying steps of developing a forecast. Finally, our work compares the temporal prediction consistency and direct/ indirect economic benefits in addition to the accuracy of demand prediction.

Turning next specifically to prior literature in demand prediction, several studies have examined the usefulness of ML/AI in different settings (Cui et al. 2006; Lin et al. 2017; Sinha & May 2004; Walczak 2001; Zhou et al. 2004;). In particular, demand forecasting in tourism has received considerable attention (Claveria et al. 2015). Taken as a whole, the predictive performance of exclusively ML/ AI approaches relative to other econometric benchmarks based on time-series analysis is mixed (Chen & Wang 2007; Peng et al. 2014; Yu & Schwartz 2006). These results suggest that past time-series trends are not strongly predictive of future demand. This literature suggests the importance of incorporating additional variables (e.g., search engine dataset) in the analysis that go beyond past visitor trends (Hu & Song 2020). We now turn to our analysis, but we begin by describing our dataset.

3.3 Data Description

In this research, we partner with one of the biggest theme parks in the world and the biggest in South Korea, which receives around 5 million annual visitors. The theme park has

made considerable effort to understand the impact of customer visit data, and has utilized its demand forecasts to plan human resources, supplies of food & beverage, merchandise, and park operations. The theme park has identified precise prediction of the number of daily visitors as one of its most critical business goals.

The data spans from January 1, 2015 to October 31, 2018, and comprises 267 variables that include proprietary and non-proprietary information such as customer-specific characteristics, weather conditions, day-specific characteristics, events, discounts, reservation information, and so on. Park attendance from January 1, 2015 to December 31, 2017 is shown graphically in Figure 3.1. A total of 14.46M customers visited the theme park in the three years 2015 to 2017. Of these, individual visitors (solo tourism) accounted for 65.96% (9.54M), with 32.75% (3.12M) visiting in the spring and 30.99% (2.96M) in the fall, the two most preferred seasons. Meanwhile, 22.35% (2.13M) and 13.92% (1.33M) of individual visitors visited during the summer and winter, due to the severe seasonal characteristics of these seasons in Korea.

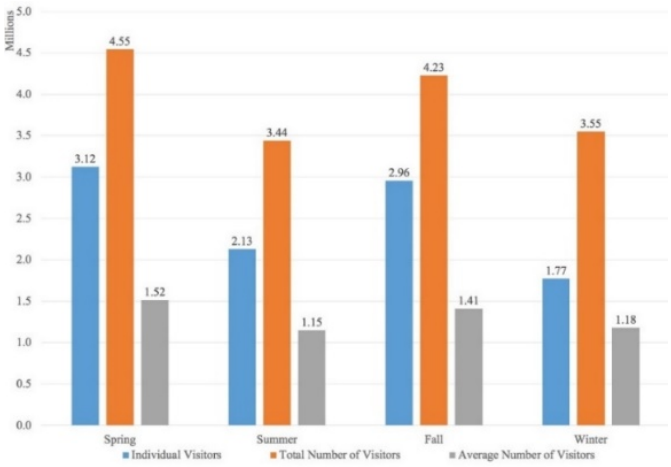

Figure 3.1 Total Number of (Individual) Visitors by Seasons for 3 Years (Jan. 2015- Dec. 2017) and Average Number of Visitors by Season & per Year (in Millions)

The data from national data centers and third-party databases (popular search engine and market analysis engine) include information such as air quality, traffic, alerts, buzz in social media, and local competitors (see Table 3.1). Subsequently, we ran a field experiment to test the performance of the models for roughly six months, from November 1, 2018, to April 28, 2019. In addition to the real-time daily data (179 days) for testing prediction accuracy for daily demand, the park also provided financial performance data (323 days) for 13 restaurants and 21 snack bars from January 1, 2018, to Nov. 18, 2018. Later, the park provided additional attendance data for 31 days from March 1 to March 31, 2020. The timeline is illustrated in Figure 3.2.

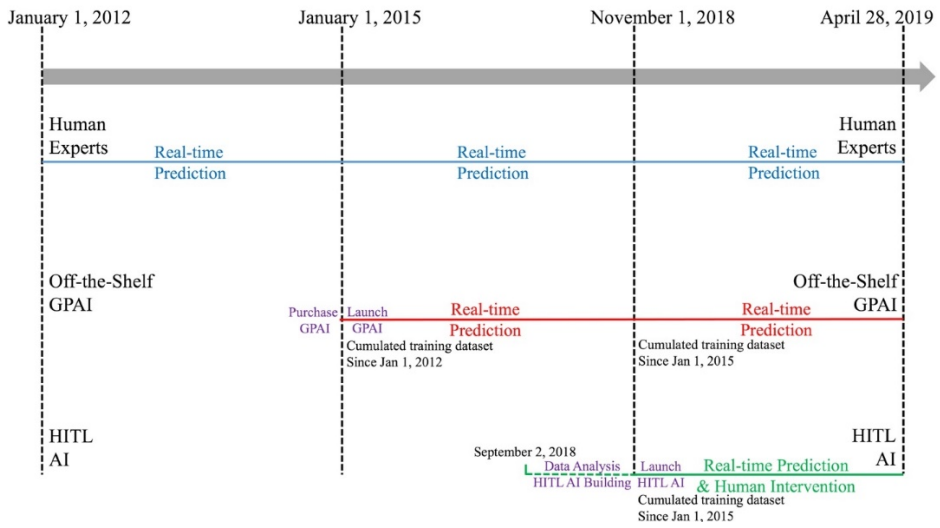

Figure 3.2 Timeline for prediction for Human Experts, GPAI, HITL AI – Data & Prediction

Table 3.1 Variables, Source of Variables, Number of Variables, Format of Variables

Source of Variables/ Variables	# of Variables	The initial format of Variables
Theme Park private data		
Day-Specific variables	55 variables	Text/ Numeric
Event-related variables	5 variables	Numeric
Discount related variables	2 variables	Numeric

Discount/ Package variables	121 variables	Numeric
Customer-related variables	19 variables	Numeric
Reservation related variables	2 variables	Numeric
Weather-related variables	42 variables	Numeric
National Air Quality Monitoring Center		
Air Quality variables	5 variables	Numeric/ Geographic
National Transport Information Center		
Traffic-related variables	2 variables	Numeric/ Geographic
3rd Party Private data (Daum/ Cheil Inc.)		
Buzz (Blog + Search) related variables	14 variables	Text/ Numeric

To provide a more comprehensive summary, we combined training data and some of the testing data. The results are shown in Table 3.2. First, the total number of weekends in each season from 2015 to 2018 is very similar (around 104 days per season). The number of holidays in summer was slightly lower than in other seasons. Similar to the trend of visitors in Figure 3.1, the total number of individual visitors in spring and fall was 32.47% (3.99M) and 32.16% (3.95M), respectively; while the summer and winter were 21.99% (2.70M) and 13.38% (1.64M), respectively. There is a significant amount of inter-day variation in every season, which is significantly higher than the variation in visits across seasons. For example, the average number of visitors per day in spring is 15,736, while the standard deviation was 12,109.8. This huge variance of the daily mean for visitors makes it difficult for the theme park to accurately predict the number of daily visitors. The ratio of smart reservations, similar to the fast-track tickets of Disney and Universal Studios, was around 10 percent overall. Word-of-mouth on Instagram totaled 1.5M in the past four years. The most mentioned season is fall at 35.37% and immediately following respectively, in spring, summer, and winter at 29.89%, 22.15%, and 12.59%.

Table 3.2 Summary Analysis for Selected Variables

	Spring 2015-2018	Summer 2015-2018	Fall 2015-2018	Winter 2015-2018
# of Days (2015-2018)	368 days	368 days	364 days	361 days
# of Weekends (2015-2018)	105 days	104 days	104 days	105 days
# of Holidays (2015-2018)	16 days	8 days	18 days	17 days
Total Individual Visitors (2015-2018)	3,991,205	2,702,930	3,953,329	1,644,210
Total Annual Member Visitors (2015-2018)	1,799,593	1,703,569	1,627,905	1,137,496
Average Daily # of Visitors	15,736	11,974	15,333	7,706
(Std. deviation)	(12,109.8)	(7,261.3)	(10,749.8)	(6,613.4)
Avg. Daily % Smart Reservation (Fast Track)	7.96%	11.59%	7.48%	15.38%
Avg. High Temp. (°C)	19.49	30.49	20.34	4.81
Avg. Low Temp. (°C)	7.17	21.54	9.90	-4.98
Total Rainfall / Season & Year (mm)	237.50	627.08	183.35	80.30
Avg. PM10 (mg/m3)	63.81	37.04	39.73	68.15
Avg. daily # of Blogs about Air Quality	16,653	2,700	3,159	3,894
Avg. daily # of Theme Park Buzz on Instagram	1,234	914	1,476	530

3.4 Methodology

This section first discusses the deep neural network, which forms the basis of our demand prediction section. Then we outline how our HITL approach differs from the GPAI approach.

3.4.1 Deep-learning Neural Network

We employ deep learning methods (LeCun et al. 2015) to develop a demand prediction model. A deep-learning neural network algorithm represents a learning architecture composed of multiple layers of neurons, which generate nonlinear data transformations (Guo et al. 2018; Liu et al. 2019). Each input variable $(x_1, x_2, \dots\dots, x_{n-1}, x_n)$ in the previous layer (output of the previous layer or raw data in the first layer) passes its information into

neurons in the next layer. Hence, each module abstracts higher-level representations from lower-level information or raw data. For example, a piece of voice information can be converted to higher-level features representing tone, speed, or interval from lower-levels of amplitude and frequency between neuron layers. Each neuron processing input values with the mathematical functions (f) throughout the layers, as shown in Figure 3.3. For example, linear function, sigmoidal function, hyperbolic tangent function, softmax function, or ReLURectified linear unit function may be applied in each neuron, and the processed information (weighted information) in the neurons of the hidden layers pass onto the output layer neurons (Nwankpa et al. 2018).

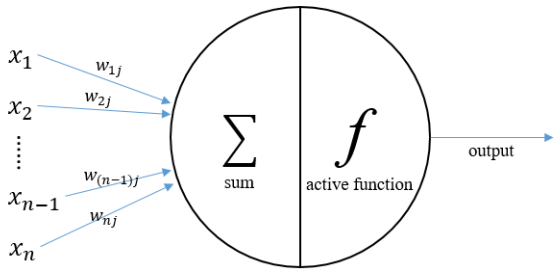

Figure 3.3 Input, Neuron, & Output

The weights in the neural network are learnt through backpropagation that works in the following manner. Backpropagation starts with randomly selected weights from the first variables to the neurons of the hidden layer. It then adjusts the weights to gradually reduce the error (using gradient descent) between the predicted results calculated through each layer and the results to be obtained (Lillicrap et al. 2016). In this study, for instance, we utilize a normalized level of particulate matter concentrations ($PM_{2.5}$ *or* PM_{10}) as one of the inputs. The information may be transformed with one of the activation functions (e.g.,

softmax function). Backpropagation starts from the last layer's last parameter (error) of the neuron(s). With the reduced error (bias) and adjusted weights, each activation function's outputs pass information, leading to the last neuron, daily demand.

3.4.2 Commonly used Approaches for Demand Prediction

Human Expert driven decision-making:

Human experts in the marketing department of the theme park have at least two years of experience and predict the number of daily visitors two days ahead. These experts pay attention to different sets of variables that vary by season. Based on domain knowledge and experience, these human experts (HE) predict daily demand using three or four important variables.

General-purpose AI (Deep-learning neural network):

The past few years have witnessed a proliferation of “turnkey” AI solutions provided by many software companies (Carvalho et al. 2019). For example, cloud-based off-the-shelf AI (including deep learning) technologies developed by Google, IBM, Amazon, and Microsoft have been positioned as “ready-to-use” technologies for decision-makers as universal approximators in many different fields (Dwivedi et al. 2019). These products offer an easy way to enable AI capabilities within organizations by providing boxed software, typically driven by deep-learning, with various input variables and a target variable. The software trains the neural net to learn the mapping from the inputs to the output variable. For example, cloud AI platforms have been used to solve a variety of problems such as object recognition, forecasting, sentiment extraction, simulation, and optimization (Carvalho et al. 2019). GPAI is often marketed as solutions that can be deployed in lieu of data scientist and

given the dearth of good AI/ML talent has gained significant adoption in the industry (Dwivedi et al. 2019).

In this study, our partner organization also used an off-the-shelf deep-learning-based solution that works on all the theme park data. The benchmark GPAI system used at the theme park is an off-the-shelf deep-learning system provided by a major US solution company. This GPAI usually performs well on big data with many training sets. The GPAI was trained using cumulative data from January 2012 and the first prediction was made in January 2015. Note that the GPAI and HITL use the same underlying deep learning algorithm, with the difference being that the GPAI was automated and did not receive iterative human inputs.

3.4.3 Human-in-the-Loop AI

Following the HITL literature in computer science (Licklider 1960; Licklider & Clark 1962), we define a human-in-the-loop design for ML as (a) handling, collecting, storing, analyzing, and interpreting of complex unstructured and structured data by human experts, (b) selecting variables which are domain-specific, statistically relevant, and unique to a business by human experts, (c) training selected variables for the ML algorithm for superior results, and (d) giving feedback for improvement through trial and error as well as reviewing of results with optimizing the ML algorithms, iteratively. Finally, if there are sudden immediate issues to consider, human experts provide feedback for the HITL AI.

In our study, human experts, with their experience and knowledge, not only interact with the ML algorithms by reducing the size and scope of the dataset, selecting features, and optimizing the neurons in the procedure, but also continuously change, tune, and improve

algorithms based on feedback from the actual results <Figure 3.4>. In our context, when using GPAI, we observe that employees were trained on how to use the GPAI tool. One observation of interest is that they show a tendency to use all the data they can generate.

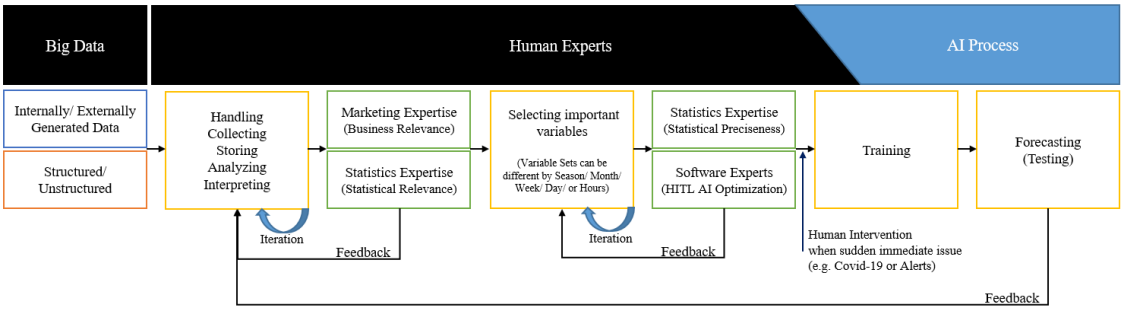

Figure 3.4 Human-in-the-Loop approach

The continuous interaction of the system with human beings in the prediction process distinguishes the HITL approach from a typical ML prediction model. More formally, the model parameters in the standard application of ML are fixed after the training is complete. On the other hand, these parameters of the HITL approach can change over time based on human input.

In our approach, we use a deep-learning neural network developed with continuous feedback from experts.[14] Human experts first use their domain know-how to identify business-relevant data from the vast amount of data (76 variables/ 267 variables: 28.36%). For example, marketing experts emphasized the "seasonality" effect. Previously, for GPAI, the seasonality variable was included with other variables in one big piece, increasing noise in the data. The noise can lead to under/ over-estimation of the effect of the "season." However, with human experts, a different set of variables was considered to be important in

[14] It is important to note that the HITL approach can be used with any ML model and is not restricted to a deep learning model.

each season. Human experts tweaked the variables using their expertise, resulting in around 12-15 variables in the best models.[15]

Conversely, GPAI requires a specific, well-defined task based on massive but specific data in a controlled environment. However, in this research context of a theme park that predicts future demand, the environment is not well controlled, and the data comes from diverse sources with a great deal of noise. Therefore, GPAI requires judgment and verification by humans to capture and process quality data. As human experts interact with ML algorithms, the HITL AI approach has the advantage of securing more situation-specific and controlled data in a dynamic and noise-prone environment relative to GPAI.

We iteratively study several different deep-learning neural network models that use the same set of data for each season, and determine the optimal number of neurons and layers for the deep-learning architecture. The number of layers and neurons can change over time as the theme park uses the HITL approach due to continuous training as the theme park uses the HITL approach. Finally, dropped variables, usually identified as irrelevant by human experience or statistical analysis, can be reconsidered as inputs in HITL AI by human experts in immediate or sudden events (weather alerts, national tragedy, pandemic, or other unexpected events).

[15] We considered (1) Principal Component Analysis & (2) Information Gain Function as the linear dimensionality reductions as well. However, principal component analysis is heavily dependent on a correlation (or a covariance) matrix. Thus, the principal component analysis may be unable to capture complex non-linear correlations, leading to overestimate the true dimensionality of the data (Bishop 1995). Lastly, the principal components similar to latent variables are difficult to interpret. Second model we consider is that information gain function. Information gain function is a mathematical algorithm to calculate the reduction in entropy with additional information (post status) from the prior status, used in information theory (Shannon 1948). Moreover, we do not use the information gain function approach because our outcome variable, the number of daily visitors, can be hard to set a threshold for creating multiple bins.

We expect the HITL AI approach to surpass GPAI's forecasting results in that HITL has extracted more sophisticated and relevant data from vast and noisy data in a dynamic environment using human intuition and experience. We also expected that HITL AI, which can reflect significant immediate events compared to GPAI, which relies on existing data patterns and is not sensitive to sudden changes, would be superior to predicting daily demand for the theme park. Compared to demand prediction by human experts, we claim that the forecasting results by HITL AI would present a better fit to the actual demand, given that it provides more diverse and relevant data analysis. Human experts predict demand for the theme park by comparing patterns and meaning based on their experiences with a few variables in each season. We expect that the forecasts of HITL AI will be more accurate than those of human experts by finding various patterns in the data that humans cannot recognize.

3.5 Results

3.5.1 Comparison of Different Prediction Methodologies

This section compares the predictive performance of the different techniques presented in Section 4 using measures of MAPE, RMSE, and accuracy. The theme park is able to easily accommodate up to 20% deviations from predictions, and the errors within this range do not lead to any meaningful financial loss for the park. Hence, we also present the accuracy by considering a contingency 20% error threshold. The predictive performance of all the models is shown in Table 3.3.

Table 3.3 Model Comparisons: Human Experts, GPAI, HITL AI, and other benchmarks

	MAPE	RMSE	Accuracy w/ Threshold ± 20% error
(1) Human Experts	.3668	2961.276	79.33%
(2) General-purpose AI	.4654	3208.425	73.74%
(3) HITL AI	.1861	1541.994	93.85%
Selected 5 Benchmarks from previous chapter			
- Seasonal Naïve	.4934	4345.429	70.39%
- SES	.5100	4451.109	68.72%
- H-W Des	.6295	4374.680	64.80%
- BSTS	.4209	3486.059	75.42%
- OLS	.4521	2683.227	75.42%

The baseline prediction of human experts has a MAPE of 36.68%, while being "accurate enough" for business decision-making at 79.33%. On the other hand, the GPAI has a higher MAPE of 46.54%, while also being slightly more inaccurate from a business decision-making perspective (73.74%). Clearly, human experts were able to make better decisions compared to an off-the-shelf GPAI. This is surprising because the experts only use a few variables to make predictions, while the GPAI uses a much larger number of predictor variables. The GPAI might not perform well in this situation because the underlying neural network needs to learn thousands of parameters, and the training data in this situation might not be sufficient to train the model, even though it extends several years. The GPAI internally tries to optimize the model to improve prediction accuracy, but the search space is too large for the training algorithm to perform well without any form of human supervision (and training). Given the relatively poor performance of GPAI as compared to experts, the theme park had written off the investments in AI. While this might seem to be an exception, off-the-shelf GPAI software often does not perform well in practice (Strickland 2019). This finding is supported in our setting where even with an extremely well-defined task, and rich, high-quality data, the GPAI does not perform well in predicting daily theme-park demand.

The HITL approach, on the other hand, performs exceptionally well as compared to both human experts and GPAI. The MAPE of HITL is 18.61%, whereas the model makes predictions within a 20% range of error 93.85% of the time. This demonstrates that human expertise and the learning models are complements in our context. We infer that HITL AI replaces human experts when computers perform routine, codifiable, and repetitive tasks via the specific dataset, while amplifying the comparative advantage of human experts in problem-solving, adaptability, and creativity (Autor 2015). Cognitively, humans find it difficult to process vast amounts of data to identify underlying patterns. Also, human stereotypes or prejudgment can affect the processing and interpretation of data, increasing uncertainty in finding solutions. On the other hand, the search space is extremely large for machine learning algorithms to substitute for human intuition and experience. Hence, exploiting the complementary relationship between human experts and machine algorithms leads to performance improvement.

Next, we examine whether the artificial intelligence/ machine learning algorithms perform better compared to other econometric forecasting models in previous chapter 2. We initially estimate twelve econometric benchmarks. Of the twelve, we select the best five models. Although GPAI is not better than human experts or HITL, its performance (MAPE: .4654, RMSE: 3208.43, & Accuracy: 73.74%) is comparable to or better than most forecasting models. Interestingly, human experts utilize only a small number of variables to forecast daily demand and show superior performance in MAPE (.3668), RMSE (2961.28), and Accuracy (79.33%) than all of the other benchmarks, other than the RMSE of OLS. Finally, the demand forecast performance of HITL AI (MAPE: .1861, RMSE: 1541.99; and Accuracy: 93.85%) is far superior to other benchmarks. The outcome of the demand forecast of HITL

AI and human experts is evidence of the complementarity of human experience and intuition and deep-learning algorithms in nonlinear forecasting.

3.5.2 Predictive Performance during the Field Experiment

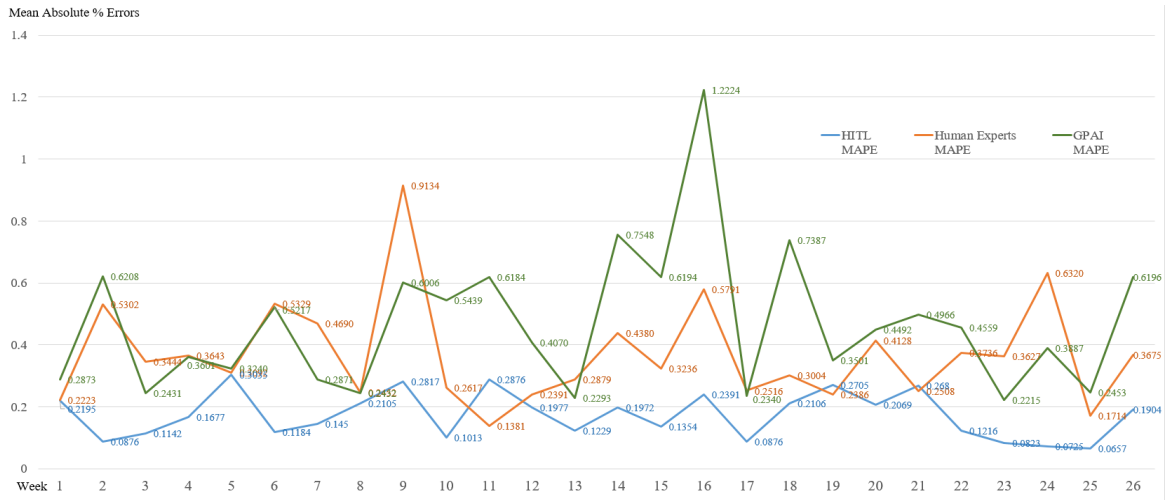

Figure 3.5 Comparison of the prediction performance (MAPE) per Week for (i) experts, (ii) GPAI, and (iii) HITL during the field experiment

In order to test the real-world validity of the HITL approach, we ran a field experiment for 26 weeks from November 1, 2018 to April 28, 2019. The temporal performance of the three approaches is shown in Figure 3.5. The largest MAPE in this period is 91.34% and 122.24% for human experts and GPAI, respectively, while that of HITL AI is 30.35%. The lowest MAPE for human experts and GPAI is 13.81% and 22.15%, respectively, while HITL AI is 6.57%. The average standard deviation of MAPE in the weekly forecasts was .1672 for the human experts and .2270 for GPAI, compared to .0737 for HITL AI. Overall, we observe that HITL performs far more consistent predictions.

The GPAI does considerably worse when there is a change in season. In the absence of input from human experts, it is difficult for GPAI to adapt to sudden changes in the environment quickly. Humans are more efficient than algorithms in recognizing shifts in

their environments and are able to self-adjust (Lake et al. 2015; 2016). As a result, our context allows us to identify and understand humans and machines' relative strengths and weaknesses. To assess the performance of GPAI compared to HE and HITL AI in dynamic environments, we perform three additional analyses, with the results shown in Tables 3.4 – 3.9.

3.5.3 Predictive Performance under Dynamic Environments

3.5.3.1 Seasonal Changes

In Korea, the winter season begins in December and the spring season starts in March. During the course of our field experiment, seasons change from fall to winter and then to spring. The prediction results for these changes are shown in Table 3.4. From November 28 to December 4, 2018, in the last three days of fall season and the first four days of the winter season, the park hosts a total of 29,922 visitors. In this period, the MAPE for HITL AI is 24.00%, with the MAD of 979. During the same period, MAPEs of HE and GPAI are 27.26% and 34.86%, respectively. As we can see, GPAI underperforms the other two approaches.

We observe similar trends in the shift to spring. From February 26 to March 4, 2019, in the last three days of winter season and the first four days of the spring season, 69,294 customers visit the theme park. While the MAPEs of GPAI and HE are 50.20% and 24.61%, respectively, the MAPE of HITL AI is 14.45% with the MAD of 1,164. HITL AI outperforms its benchmarks. Importantly, we observe that the prediction performance of HITL is statistically better than the prediction performance of both human experts and GPAI.

Table 3.4 Prediction Error (Absolute Difference & Absolute % Error) & Season changes

	Date	Abs. Diff. of HE	Abs. Diff. of GPAI	Abs. Diff. of HITL	Abs. % Error of HE	Abs. % Error of GPAI	Abs. % Error of HITL
	11/28	88	637	197	2.85%	20.63%	6.38%
	11/29	744	325	853	19.87%	8.68%	22.79%
	11/30	70	136	994	1.78%	3.46%	25.28%
Winter Starts	12/01	597	4,056	1,034	6.94%	47.18%	12.02%
	12/02	294	6,890	2,112	5.15%	120.75%	37.02%
	12/03	2,373	591	430	145.85%	36.32%	26.41%
	12/04	270	226	1,230	8.36%	7.00%	38.08%
Average w/ $\pm$3 days		634	1,837	1,045	27.26%	34.86%	24.00%
	02/26	1,521	2,884	190	17.85%	33.84%	2.22%
	02/27	1,774	906	121	20.22%	10.32%	1.38%
	02/28	1,620	1,266	1,386	13.94%	10.90%	11.93%
Spring Starts	03/01	6,969	5,604	3,967	36.74%	29.54%	20.92%
	03/02	1,033	1,411	409	7.36%	10.05%	2.91%
	03/03	1,986	2,908	1,161	39.61%	58.00%	23.16%
	03/04	863	4,696	913	36.52%	198.73%	38.64%
Average w/ $\pm$3 days		2,252	2,811	1,163	24.61%	50.20%	14.45%

3.5.3.2 Air-quality Alerts

In South Korea, air quality is an essential consideration for outdoor activity. The National Air Quality Agency of South Korea provides four major air-quality alerts (PM10, PM2.5, NO2, and CO to the public). Of these four, PM10 and PM2.5 receive the most attention. HE and HITL AI reconsider demand via a feedback loop when PM10 and PM2.5 alerts are issued, while GPAI does not. On the other hand, NO2 and CO levels are less critical in the choice to engage in outdoor activity and are not considered by all of the model. The comparison of the models presents in Table 3.5.

Table 3.5 Prediction Error (MAD & MAPE) & Air-quality alerts

Alerts	Error	Human Feedback	Human Experts	General-purpose AI	HITL AI
PM 10 Alerts	MAD	Yes	1560.800	2352.769	1210.802
(40 days)	MAPE	Yes	.2617	.4968	.2212
- t-statistics for MAPE to GPAI			-1.670*	N/A	-2.683***

(p-value)			(.0992)		(.0093)
PM 2.5 Alerts	MAD	Yes	1528.548	2440.027	1176.388
(42 days)	MAPE	Yes	.2686	.5266	.2225
- t-statistics for MAPE to GPAI			-2.014**	N/A	-3.087***
(p-value)			(.0475)		(.0030)
NO2 Alerts	MAD	No	1803.250	1984.736	1541.968
(36 days)	MAPE	No	.2810	.4214	.2321
- t-statistics for MAPE to GPAI			- .416	N/A	-1.124
(p-value)			(.6646)		(.2652)
CO Alerts	MAD	No	1458.773	2136.805	1482.603
(22 days)	MAPE	No	.2545	.5398	.2753
- t-statistics for MAPE to GPAI			-1.144	N/A	-1.306
(p-value)			(.2595)		(.1988)

P-values are in parentheses (the benchmark is GPAI): *** p <.01 / ** p <.05 / * p <.10

Interestingly, for sudden PM10 and PM2.5 alerts, the demand prediction errors (MAD and MAPE) of human experts and HITL AI are statistically significantly lower than those of GPAI. Moreover, in all of these cases, the demand prediction errors (MAD and MAPE) of HITL AI are lower than both human experts and GPAI. Finally, since all three methods do not consider NO2 and CO alerts, the differences between the demand prediction errors (MAPE) of the three methods are not statistically significant.

3.5.3.3 Unusual Number of Visitors

We investigate the prediction errors on extreme days which are defined by an unusual number of visitors in Table 3.6. Extreme days include: (1) those with less than 2,000 visitors, (2) more than 13,000 visitors, (3) weekends, (4) national holidays, and (5) considerable variation from the previous day. In all cases, the MAPE of HITL AI ranges from 11.41% to 18.04%, while human experts are between 21.87%, and 60.97%, and GPAI is between 24.55% and 78.98%.

Among these five cases, we further examine the three cases (1), (2), and (5). The two instances of less than 2,000 & more than 13,000 visitors' cases are specified as extreme days

because of their low frequency. For low attendance days (1), the prediction error (MAPE) of HITL AI was 16.32% compared to 60.97% for Human experts and 78.98% for GPA. For high attendance days (2), the prediction error (MAPE) of HITL AI was 11.41% compared to 21.87% for Human experts and 25.23% for GPAI.

For case (5), shown in table 3.6, there are 60 days with considerable variation from the previous day. To classify these days, we calculate the absolute difference between the number of visitors of the previous day and the target day (eq. 3.1) and calculate the average variation of the absolute differences for 178 days (eq. 3.2). Finally, we specify the 60 days as extreme days because the absolute difference of the target day is more than the average variation of the absolute differences (eq. 3.3). For those 60 days with large divergence from the previous day, HITL AI prediction error (MAPE) is 14.39%, much lower than those of HE and GPAI are 28.18% and 32.28%, respectively.

$$Abs.Diff_i = \left|\#\ of\ Visitors_{i-1\ day} - \#\ of\ Visitors_i\right| \tag{3.1}$$

$$Avg.Abs.Diff = \frac{\sum_{i=2}^{179}\left|\#\ of\ Visitors_{i-1\ day} - \#\ of\ Visitors_i\right|}{178\ days} \tag{3.2}$$

$$High\ Variation\ Day_i = \begin{cases} 1, & Abs.Diff_i > Avg.Abs.Diff \\ 0, & Elsewhere \end{cases} \tag{3.3}$$

Table 3.6 Prediction Error (MAD & MAPE) & Extreme Days

Model	Errors	Visitors Less than 2,000	Visitors More than 13,000	Weekends	National Holidays	Considerable Variation From Previous Day
# of Day (out of 179 days)		17 days	15 days	52 days	6 days	60 days
Human Experts	MAD	940.941	4125.467	3093.000	3019.833	2970.383
	MAPE	.6097	.2187	.3393	.3109	.2818
GPAI	MAD	1218.648	4758.333	4298.546	2384.424	3403.836
	MAPE	.7898	.2523	.4715	.2455	.3228
HITL AI	MAD	251.846	2151.945	1606.667	1752.551	1517.403

MAPE	.1632	.1141	.1762	.1804	.1439

3.5.3.4 Sudden Events

The Covid-19 situation in South Korea became severe after a highly infectious individual was identified on Feb. 18, 2020. Six days later, there were three-digit confirmed cases per day, being a heightened awareness of Covid-19. Thus, the theme park faced difficulties in forecasting demand in March 2020. In Table 3.7, the number of visitors in March 2020 decreased by up to 63.73% compared to the last five years of March.

Table 3.7 Total number of Individual Visitors in March 2020 & Comparison to the Other Years

	2015	2016	2017	2018	2019	2020
Individual Visitors in March	356,750	218,023	195,632	195,841	191,967	129,383
% Decrease:	- 63.73%	- 40.66%	- 33.86%	- 33.93%	- 32.60%	-

The similarity between the trends inherent in the training set and the distribution of future test sets increases the predictability of machine learning algorithms (Brynjolfsson & Mitchell 2017). Therefore, we create a new (binary) variable called MERS (Middle East Respiratory Syndrome) to better forecast demand using HITL in the Covid-19 context.

In South Korea, MERS patients were first reported on May 20, 2015. One hundred eighty-six people were infected, and 38 of the infected patients died in June 2015. Then, there was no MERS from late June to July 4, 2015, and the MERS situation ended. MERS is the same strain of coronavirus as Covid-19. Also, MERS has a rapid spread rate and high fatality rate, affecting the demand of the theme park and the entire economy of Korea. Thankfully, since our dataset started on January 1, 2015, it can reflect trends in daily demand for the theme park affected by MERS. The MERS variable might not be utilized important in any other machine learning application. The main reason is that it is easy for a human to infer the

coronavirus from the previous MERS, but not to machine algorithms. In addition, the MERS variable is statistically insignificant in linear approximation on the daily demand of the theme park (Table 3.8). Therefore, it will not be considered one of several other variables for machine learning. Even if considered, it will not be treated as an essential variable to find an underlying pattern in machine learning applications.

Table 3.8 Statistically Insignificant MERS Variable & Other Significant Variables

	Estimate	Std. Error	t-statistics	p-value	Sig.
Major Variables					
Intercept	4682.237	1091.884	4.288	.0000	***
National Holiday	5075.470	594.459	8.538	.0000	***
Weekednds	2834.435	245.716	11.535	.0000	***
Long weekends	3745.873	813.761	4.603	.0000	***
The day b/w Holiday & Weekends	2549.392	1009.146	2.526	.0116	**
Introd. New Attraction or Event	416.921	252.860	1.649	.0994	*
Night Ticket	1004.758	298.676	3.364	.0008	***
Ten Dollar Ticket	1680.425	746.093	2.252	.0245	**
Web Reservation	9.446	.2969	31.815	.0000	***
Web Promotion	-3821.24	368.186	-10.379	.0000	***
Rainfall (mm)	-26.957	9.330	-2.889	.0039	***
Rain Before opening the park	-923.784	297.529	-3.105	.0019	***
Rain a Day Before	449.485	219.153	2.051	.0405	**
Daily Temperature Range	365.844	75.033	4.876	.0000	***
Lowest Temperature (°C)	-140.914	67.964	-2.082	.0375	**
Average Humid (%)	-25.492	9.5311	-2.675	.0076	***
PM 10 Alerts	-2889.65	1243.231	-2.324	.0203	**
# of Posts in Weather Alerts Buzz	-.0234	.0062	-3.763	.0002	***
# of SNS Posts about competitor 1	-.1104	.0224	-4.931	.0000	***
# of SNS Posts about competitor 2	1.941	1.174	1.653	.0985	*
# of SNS Posts about us	7.623	2.640	2.888	.0039	***
# of Blogs about us	63.366	17.298	3.663	.0003	***
# of SNS Posts about Foreign Travel	-14.373	4.635	-3.101	.0020	***
The variable we are interested in					

MERS (Yes/ No)	-1111.765	1080.601	-1.029 .3037
With Other 12 variables Selected as Not Significant			
Residual Standard Error: 3341 on 1364 Df Multiple R-squared: .8018 Adjusted R-Squared: .7967 F-statistics: 157.7 on 35 and 1364 Df, p-value: < .0000			

P-values are in parentheses (the benchmark is GPAI): *** p <.01 / ** p <.05 / * p <.10

We use the MERS variable as one of the primary considerations in the feedback loop of the HITL approach. The daily demand forecasting results for the theme park in early Covid-19 are shown in Table 3.9. Table 3.9 only compares the GPAI and HITL AI because the theme park no longer utilizes human experts' forecasting.

Table 3.9 Prediction Error (MAD & MAPE) on March 2020 under Covid-19

	GPAI	HITL AI
MAD	2,737.625	1,649.584
MAPE	85.30%	50.43%
Accuracy w/ ± 20% Errors	64.52%	83.87%
# of Over-Predict Days	29 days	24 days
# of Under-Predict Days	2 days	7 days
Type 1 Error Costs / Day	$4,658.00	$2,697.19
Total Type 1 Error Costs	$135,082	$64,732.50

The overall MAPE, MAD, and accuracy have deteriorated considerably compared to the 179-days results earlier in Table 3.3. This shows that predictions using ML are vulnerable to sudden changes. However, HITL AI still shows much better prediction results than GPAI. For example, the accuracy in the ordinal days is 93.85% (Table 3.3) and the accuracy in the Covid-19 becomes 83.87% for HITL AI (Table 3.9). Also, the MAPE of HITL AI in early Covid-19 is 50.43% compared to 85.30% for GPAI.

3.5.3.5 Statistical comparison of the three models

Finally, we check whether the mean absolute differences between the observed daily number of visitors and the predicted number of visitors in each prediction model are

statistically significant with the mean comparisons <Table 3.10>. Table 3.10 shows that the mean absolute differences in the daily number of visitors for the theme park predicted in GPAI and that experts are statistically meaningless or better indicated by HE. More importantly, HITL AI produces statistically significantly better predictions compared to that of GPAI and HE. Therefore, we stress that HITL AI can bring the accuracy of the prediction and consistency of prediction compared to its alternatives.

Table 3.10 Comparison of Mean Absolute Difference between Models

Effective Size Between	Cohen's D	99% C.I.	Effect Size
Human Experts & GPAI	.1627	(-.112, .437)	Small Effect
Human Experts & HITL AI	.4839	(.206, .761)	Small Effect
GPAI & HITL AI	.6892	(.407, .971)	Medium Effect

3.6 Managerial Benefits

In the previous section, we have demonstrated the superior performance of HITL AI in terms of prediction accuracy and consistency compared to human experts and GPAI. This section explores the economic benefits that HITL AI can achieve compared to Human Experts and GPAI. We explore the benefits by dividing them into direct and indirect economic values.

3.6.1 Economic Benefit

More accurate predictions are beneficial only if the business can exploit them. Next, we examine the economic benefits of the various predictions to the theme park. To measure the economic benefit of each model, we present the costs that are incurred when a model makes incorrect predictions (Table 3.11). For example, when a model under-predicts the daily number of visitors, the lack of sufficient support for visitors to the restaurants and snack bars can increase bottlenecks, and may lead consumers to forgo spending. In accordance with the ML literature, we call this Type 2 error. On the other hand, when the predictions are

higher than the actual count, the park is spending unnecessarily on labor, attractions and other associated costs, or Type 1 error.

Table 3.11 Type 1 and Type 2 error costs

	Visitors: large # of visitors	Visitors: small # of visitors
Theme Park: Predicted large # of visitors		**Over-Forecasting:** *Theme Park: Type 1 Error Costs* - Overspending on Staff (Wasted Labor wages or other resources)
Theme Park: Predicted small # of visitors	**Under-Forecasting:** *Theme Park: Type 2 Error Costs* - Visitor Support Problem due to lack of Staff: Long line-up, Longer Waiting Time, Inefficient restaurant/ snack bar turnover, etc. *Visitors* - Give up their consumptions in restaurants/ snack bars → Spend less per visitors (Opportunity Cost)	

3.6.1.1 Estimating Type 2 error: To compute the cost of Type 2 error, we add the difference between the mean of the daily consumption per visitor when the park was sufficiently staffed to when it was under-staffed.

$$OvStf_i = \begin{cases} 1, & Real\ number\ of\ visitors \leq Forecasted\ number\ of\ visitors \\ 0, & Else \end{cases}, \quad (3.4)$$

$$ResSp/Visitor_i = \frac{Total\ \$\ Spent\ in\ Res_i}{Real\ Number\ of\ Visitors_i}, \quad (3.5)$$

$$SnackSp/Visitor_i = \frac{Total\ \$\ Spent\ in\ Snack_i}{Real\ Number\ of\ Visitors_i}, \quad (3.6)$$

$$ResSp/Visitor_i = \alpha + \beta OvStf_i, \quad (3.7)$$

$$SnackSp/Visitor_i = \gamma + \delta OvStf_i, \quad (3.8)$$

where $OvStf_i$ is the dummy variable that captures when the theme park was under (or over) staffed in a specific day i. $ResSp/Visitor_i$ and $SnackSp/Visitor_i$ represents the different types of dependent variables, such as the consumer spending in restaurants, and snack bars. Since the prediction error is completely random, we do not face any endogeneity concerns. These estimates are presented in Table 3.12.

As we can see in Table 3.12, the difference is estimated at $3.88 and $1.43 for restaurants and snack bars, respectively, and are statistically significant. Based on the average differences in expenditure per visitor, we can calculate the average Type 2 cost per day for restaurants and snack bars. The mean Type 2 error costs per day by HITL AI totals $8,692.14 and are lower than those of GPAI and HE as shown in Table 2.6. The mean Type 2 error costs per day for HITL AI account for 5.59% of the average daily revenue, while those of GPAI account for 11.33% and HE for 10.46%.

Table 3.12 Mean $ Spent Per Visitors in Restaurants and Snack Bars

	Estimate	Std. Error	t-statistics	p-value	Sig.
Restaurants					
Intercept	13.693	1.011	13.545	.0000	****
Over Staff	3.873	1.387	2.792	.0056	**
Snack Bars					
Intercept	10.275	.436	23.545	.0000	****
Over Staff	1.436	.569	2.398	.0170	*

3.6.1.2 Estimating Type 1 error: We calculate the type 1 error cost based on Korea's legal minimum wage. The average daily costs of type 1 errors are slightly over 1% (1.02%~1.31%) of the daily average revenue, regardless of which demand forecast model was used, as shown in Table 3.13. More specifically, the average daily Type 1 error costs equals $2,229.75 (= $1,014.71 for restaurant + $1,215.04 for snack bars), which are lower than Type 1 error costs for GPAI ($2,523.84) and HE ($2,861.86). Moreover, these differences between the average daily Type 1 error costs are statistically significant.

Table 3.13 Type 1 and 2 Error Costs and Comparisons (Currency Rate: $1/₩1,100)

Type 2 Error Costs	HE	GPAI	HITL AI
13 Restaurants / Day	$11,875.70	$12,866.58	$ 6,343.85
21 Snack Bars / Day	$ 4,396.00	$ 4,762.79	$ 2,348.29
Total Type 2 Error Costs / Day	$16,271.70	$17,629.37	$8,692.14
% Compared to Avg. Daily Revenue	10.46%	11.33%	5.59%
Total # of Under-predicted Days	179 days	282 days	171 days
Type 1 Error Costs	HE	GPAI	HITL AI
13 Restaurants / Day	$1,270.36	$1,200.12	$ 1,014.71
21 Snack Bars / Day	$1,591.50	$1,323.42	$ 1,215.04
Total Type 2 Error Costs / Day	$2,861.86	$2,523.54	$2,229.75
% Compared to Avg. Daily Revenue	1.31%	1.15%	1.02%
Total # of Over-predicted Days	143 days	40 days	151 days
Total Error Costs	$3.22M	$3.73M	$1.42M
% Compared to Total Revenue	5.40%	6.25%	2.38%

In addition, we further investigate Type 1 error costs during early Covid-19. Interestingly, both GPAI and HITL AI tend to over-predict during this early Covid-19 situation (Table 3.9). Like other results, HITL approach is superior to GPAI in Type 1 costs perspectives. For example, the type 1 error costs in HITL AI is $2,697.19 per day compared to $4,658 per day for GPAI. This means that the type 1 error costs for the entire March of 2020 become around $64,733 for HITL AI compared to $135,082 for GPAI. Therefore, it can be seen that using HITL AI to make more sophisticated demand predictions, even when in the Covid-19, can significantly help reduce the cost of prediction errors compared to its alternatives.

3.6.1.3 Overall Economic Benefit

Putting these estimates together as shown in the last two rows of the Table 3.13, we conclude that the economic benefits of HITL AI are considerably higher than the other two methods. HITL AI improves the theme park's revenues by 3.02% and 3.87% as compared to experts and GPAI, respectively. It also marginally reduces the operating cost of the theme

park as compared to the other alternatives. In summary, the HITL leads to additional profits of $1.8-$2.3M during the duration of the field experiment.

3.6.2 Intangible Benefits

GPAI, the traditional predictive model used by the park took 2 days to compute its forecasts. In contrast, HITL AI takes only 15 minutes to process the variables to predict daily demand. The importance of the shorter lead time cannot be overstated in terms of having enough time to respond to an emergency. For instance, in the event of a sudden weather situation, the use of HITL AI gives a business time to respond. In the case of GPAI, this agility is lost. HITL AI also makes it easier to collect, maintain, and refine small amounts of vital information. More importantly, better and reliable planning for the theme park is of substantial indirect economic value, which can also lead to higher customer satisfaction. We conclude that human experience or human experts complement ML algorithms even from an indirect-economic benefits perspective.

3.7 Conclusion

In this paper, we have tested a new approach to deploying ML in practice, a Human-in-the-loop AI system that combines the expertise of humans with the processing power of ML algorithms. And, we examined why HITL approach outperforms alternative approaches. Using data from a large theme park, we show that not only does the HITL outperform human experts and GPAI, it also leads to significant economic benefits. In addition, we show that un-mediated GPAI underperforms experts. Given that the prediction accuracy of HITL AI is statistically superior to the other two methods, we conclude that in this setting, humans and the system are complementary inputs. There are also indirect benefits associated with HITL

AI. First, shorter lead times allow the theme park to respond more quickly to changes in the market conditions. Second, due to fewer features required by HITL AI, it makes the process of collecting, maintaining, and repairing small amounts of vital information for the theme park easier.

This study has a number of implications for both academia and practice. For practice, we hope this study allows managers to reduce AI systems' failure rate and improve their performance. We see HITL as valuable for solving complex problems where the solution space is extremely large or the data are sparse. What sets HITL AI apart from GPAI and human experts is that human experts and machine algorithms can be combined to offset the weaknesses of each alone. With the machine learning algorithms/ AI investment decision, companies must invest considerable agony over how to utilize, develop, update, and maintain the ML/ AI algorithms with their human experts.

For academia, our research has shown that HITL AI, designed according to a complementary relationship between human experts and ML algorithms, can accrue significant economic value. Many studies have discussed human role changes and replacement in machine learning algorithms. In particular, the cost of data processes and forecasts was reduced with the machine learning/ AI application, expecting a replacement of experts represented by experts. On the contrary, it is not desirable to replace human experts on the cost side. Given our result, the machine learning algorithms of HITL design have shown more precise forecasting and less costly in terms of type 1 and 2 error costs than GPAI and HE. However, we are aware that this is one study in a burgeoning field. There is a

need for further studies to examine further the nature of the complementary relationship between these two inputs.

In sum, this study introduces and evaluates a novel method for deploying AI, i.e., Human-in-the-Loop. Furthermore, we hope this paper encourages more research on the complementary role of human expertise and AI, which we perceive as a potent new opportunity to design scalable and impactful AI applications.

References

Acemoglu, D., & Restrepo, P. (2018). *Artificial intelligence, automation and work* (No. w24196). National Bureau of Economic Research. Available at NBER: https://www.nber.org/papers/w24196

Agrawal, A., Gans, J., & Goldfarb, A. (2016). The simple economics of machine intelligence. *Harvard Business Review*, *17*, 2-5. https://hbr.org/2016/11/the-simple-economics-of-machine-intelligence

Autor, D. H.. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of economic perspectives*, *29*(3), 3-30.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, *54*(6), 627-635.

Bergstein, B. (2017). The Great AI Paradox. *MIT Technology Review, Dec. 15. 2017.*

Brooks, R. (2017). The Seven Deadly Sins of AI Predictions. *MIT Technology Review, Oct. 6. 2017.*

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530-1534.

Brynjolfsson, E., Rock, D., & Syverson, C. (2018). *Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics* (No. w24001). National Bureau of Economic Research.

Carvalho, A., Levitt, A., Levitt, S., Khaddam, E., & Benamati, J. (2019). Off-the-shelf artificial intelligence technologies for sentiment and emotion analysis: a tutorial on using IBM natural language processing. *Communications of the Association for Information Systems*, *44*(1), article 43.

Cui, G., Wong, M. L., & Lui, H. K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, *52*(4), 597-612.

Chen, K. Y., & Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*, *28*(1), 215-226.

Chui, M., Manyika, J., & Miremadi, M. (2018). What AI can and can't do (yet) for your business. *McKinsey Quarterly*, *1*, 97-108.

Claveria, O., Monte, E., & Torra, S. (2015). Tourism demand forecasting with neural network models: different ways of treating information. *International Journal of Tourism Research*, *17*(5), 492-500.

Davenport, T. H., & Kirby, J. (2016). Just how smart are smart machines?. *MIT Sloan Management Review*, *57*(3), 21.

Davenport, T. H., Libert, B., & Beck, M. (2018). Robo-Advisers are coming to consulting and corporate strategy. In The latest research: AI and machine building. Harvard Business School Publishing.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760-772.

de Silva, T., & Thesmar, D. (2021). The Complementarity Between Man and Machine in Forecasting. *Available at SSRN 3762348*.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 101994.

Farias, V. F., Jagabathula, S., & Shah, D. (2013). A nonparametric approach to modeling choice with limited data. *Management Science*, *59*(2), 305-322.

Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review*, *97*(4), 62-73.

Frantz, R. (2003). Herbert Simon. Artificial intelligence as a framework for understanding intuition. *Journal of Economic Psychology*, *24*(2), 265-277.

Frey, C. B., & Osborne, M. (2013). The future of employment. *Working Paper. Oxford Martin School.*

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). IEEE.

Guo, J., Zhang, W., Fan, W., & Li, W. (2018). Combining geographical and social influences with deep learning for personalized point-of-interest recommendation. *Journal of Management Information Systems*, *35*(4), 1121-1153.

Guo, X., Singh, S., Lee, H., Lewis, R. L., & Wang, X. (2014). Deep learning for real-time Atari game play using offline Monte-Carlo tree search planning. *Advances in neural information processing systems*, *27*, 3338-3346.

Hosanagar, K., & Saxena, A. (2017). The First Wave of Corporate AI Is Doomed to Fail. *Harvard Business Review*. April 18.

Hu, M., & Song, H. (2020). Data source combination for tourism demand forecasting. *Tourism Economics*, 26(7), 1248-1265.

Huang, M. H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research, 21*(2), 155-172.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332-1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *Behavioral and brain sciences*, *24,* 1-101.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE transactions on human factors in electronics*, (1), 4-11.

Licklider, J. C. R., & Clark, W. E. (1962, May). On-line man-computer communication. In *Proceedings of the May 1-3, 1962, Spring Joint Computer Conference* (pp. 113-128). ACM.

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, *7*(1), 1-10.

Lin, Y. K., Chen, H., Brown, R. A., Li, S. H., & Yang, H. J. (2017). Healthcare predictive analytics for risk profiling in chronic care: A Bayesian multitask learning approach. *MIS Quarterly*, *41*(2). 473-496.

Liu, X., Zhang, B., Susarla, A., & Padman, R. (2019). Go to YouTube and call me in the morning: use of social media for chronic conditions. *Forthcoming, MIS Quarterly*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3061149

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *Journal of Strategic Information Systems*, *24*(3), 149-157.

Lu, Y., Qian, D., Fu, H., & Chen, W. (2018). Will supercomputers be super-data and super-AI machines?. *Communications of the ACM*, *61*(11), 82-87.

Mayer-Schönberger, V. & Ramge, T. (2018). Are the Most innovative Companies Just the Ones With the Most Data?. *Harvard Business Review*. February 7. Available at: https://hbr.org/2018/02/are-the-most-innovative-companies-just-the-ones-with-the-most-data

McAfee, A. & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60-68. Available at: https://hbr.org/2012/10/big-data-the-management-revolution

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*.

Peeters, M.M., van Diggelen, J., Van Den Bosch, K., Bronkhorst, A., Neerincx, M.A., Schraagen, J.M., & Raaijmakers, S. (2020). Hybrid collective intelligence in a human–AI society. *AI & Society*, 1-22.

Peng, B., Song, H., & Crouch, G. I. 2014. A meta-analysis of international tourism demand forecasting and implications for practice. *Tourism Management*, (45), pp. 181-193.

Pistrui, J. (2018). The future of human work is imagination, creativity, and strategy. *Harvard Business Review*, *18*. January 18.

Schirner, G., Erdogmus, D., Chowdhury, K., & Padir, T. (2013). The future of human-in-the-loop cyber-physical systems. *Computer*, *46*(1), 36-45.Scott, S. L., & Varian, H. R. (2013). *Bayesian variable selection for nowcasting economic time series* (No. w19567). National Bureau of Economic Research. Available at: https://www.nber.org/system/files/working_papers/w19567/w19567.pdf

Simon, H. A. (1995). Artificial intelligence: an empirical science. *Artificial Intelligence*, *77*(1), 95-127.

Sinha, A. P., & May, J. H. (2004). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems*, *21*(3), 249-280.

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, *56*(4), 24-31.

Walczak, S. (2001). An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of management information systems*, *17*(4), 203-222.

Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science*, *66*(3), 1045-1070.

Yu, G., & Schwartz, Z. (2006). Forecasting short time-series tourism demand with artificial intelligence models. *Journal of travel Research*, *45*(2), 194-203.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker Jr, J. F. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, *20*(4), 139-166.

CHAPTER4
Revisiting Price Dispersion in Electronic Markets

4.1 Introduction

Theory predicts that the prices of homogenous products will display relative convergence in electronic markets (Bakos 1997, Brynjolfsson and Smith 2000a). This chapter investigates whether electronic market price dispersion exists for homogeneous products, meaning that electronic markets are still inefficient. Furthermore, it will examine possible factors influencing electronic market price dispersions.

A considerable variation in prices for homogeneous products from brick-and-mortar retailers can be explained by the variety of ways in which these retailers offer products, including their location, ambiance, layout, and extra services (Baker et al. 1992, Berry 2000). Some forms of ex-ante heterogeneity increase customers' search costs and result in excess profits for sellers (Stigler 1961; Pratt et al. 1979; Salop & Stiglitz 1977; Burdett & Zudd 1983). This is because such heterogeneity can influence consumers' willingness to pay. Additionally, market concentration, which is a function of the number of competitors and their respective shares, leads to price dispersions of homogeneous products in brick-and-mortar retail stores (Cui et al. 2019).

Contrary to those in brick-and-mortar markets, the ex-ante heterogeneity features in electronic markets are diminished. The distance between sellers and buyers is not a primary factor to consider in electronic markets due to low search costs (Bakos 1997). Furthermore, electronic markets can convey complex information with low communication costs (Lee &

Clark 1996). Hence, consumers can more easily search for alternatives with a few clicks. Retailers can track their competitors' prices and modify their menus with low costs. Such features of electronic markets should lead to price competition among sellers (Bakos 1997; Brynjolfsson & Smith 2000b; Kauffman & Wood 2000; Smith 2001). Similar to the Bertrand price competition model, scholars have hypothesized that sellers with a homogeneous product in electronic markets will set lower and more converged prices than conventional retailers (Lynch & Ariely 2000; Brynjofsson et al. 2003). However, as online shoppers will have observed, there is considerable dispersion in the prices of homogeneous products in electronic markets.

This study aims to provide explanations for the observed price dispersions for homogeneous products in electronic markets. In particular, it aims to answer whether the three factors of price dispersion (product information asymmetry from customer reviews, seller certainty of ex-ante heterogeneity evaluated by customer reviews, and market structure assessed by the number of sellers) help explain the persistent price dispersion in electronic markets. We collect information from a single online marketplace (Amazon.com) on 12,409 retailers. We then analyze hypothesized relationships between price dispersion and (1) product uncertainty from customer reviews, (2) seller certainty from customer reviews, and (3) market competition.

4.2 Literature Review

Since travel costs are no longer a constraint in electronic markets (Bakos 1997), consumers can easily search for alternatives. Additionally, retailers can track competitors' pricing with low costs. More consumers and retailers are willing to participate in online

markets due to low search costs and entry costs, respectively (Brynjolfsson & Smith 2000b). This nature of electronic markets should promote price competition (Bakos 1997; Brynjolfsson & Smith 2000a; Kauffman & Wood 2000; Smith 2001). Therefore, previous researchers have viewed electronic markets as having high competition, leading to lower and converged prices (Bakos 1997; Alba et al. 1997; Brynjolfsson & Smith 2000b; Kauffman & Wood, 2000; Tang & Xing 2001; Brown & Goolsbee 2002; Pan et al. 2002; Ancarani & Shankar 2004).

Scholars have also provided substantial evidence to the contrary, arguing that consumers in electronic markets experience higher prices and greater price dispersion than those in offline markets (Lee 1998; Bailey 1998a; 1998b; Lal & Sarvary 1999; Lynch & Ariely 2000; Degeratu et al. 2000; Clay et al. 2001; Ancarani & Shankar 2004; Dimoka et al. 2012). In line with this observation of high price variation, a rich body of research has investigated the underlying reasons for the existence of high price levels and price dispersion in electronic markets. Previous research on online demands has identified information asymmetry as a significant reason for high prices and high price dispersion in markets (Dimoka et al. 2012; Kim & Krinan 2015).

First, consumers are not acquainted with complete information about sellers and products in electronic markets. Suppose that several sellers are selling the same product on an online platform. In that case, each seller has much more information (the cost of warehousing, the total supply of the product, the cost of providing the product, the number of competitors selling the same product, delivery situations, etc.) about their competitors than the customers know about the sellers. This difference in information can be used as

leverage for sellers, creating incentives for opportunistic behavior (Dimoka et al. 2012; Kim & Krishnan 2015; Jarvenpaa et al. 2000; Ba & Pavlou 2002; Pavlou 2003; Dellarocas 2003; Pavlou & Gefen 2004, 2005; Pavlou & Dimoka 2006). Therefore, sellers can utilize their attributes and signal customers (e.g., delivery time guarantee when the seller know other sellers do not have the products in their inventory) to reduce the risk of information asymmetry.

Because the attributes of sellers can greatly influence consumers' selections, the characteristics of sellers may lead to price dispersion. Nelson et al. (2007), Pan et al. (2002), and Ratchford et al. (2003) tried to measure seller attributes and how the attributes lead to price dispersion online. They found that web-quality ranking, product selections, product representation, and service quality are significant attributes that customers consider while buying online. Therefore, sellers can exercise pricing power due to information asymmetry, leading to price dispersion.

Second, imperfect information about products has been identified as a factor in high prices and high price dispersion. Product uncertainty due to insufficient product information is defined as the difference between the perceived and realized qualities of a product. Because a large variation in qualities can increase online purchase risk, customers are reluctant to buy products or prefer to purchase products from sellers who have better descriptions, better reputations, or expertise in a product. Customers tend to hedge the risk of buying highly uncertain products with seller certainty. This is an important reason why new entrants or low-profile retailers set low prices (Baye & Morgan 2001). This phenomenon is observed often, especially when customers cannot easily tell the quality of a

product. Therefore, sellers can utilize their heterogeneity attributes (e.g., insurance and star ratings) and obtain better seller evaluations (seller certainty) from other consumers, engaging in strategic behavior. The strategic behavior of price setting can lead to price dispersion (Wang & Li 2020).

Finally, some studies have examined the relationship between market competition and price dispersion; however, no agreement has been reached on market competition and price dispersion in online retail. Industrial organization literature has tried to explain price dispersion in the context of market competitiveness (i.e., the number of sellers). In offline markets, competition due to the number of sellers leads to an empirically converged price (e.g., Barron et al. 2004). Haynes and Thompson (2008) showed that a larger number of e-tailers is associated with low price dispersion online. Compared to this evidence, some studies have concluded that competition in the online market instead increases price dispersion. Chandra and Tappata (2011) showed that price dispersion is positively correlated to the number of gas stations in the offline market. Wang and Li (2020) showed that price dispersion arises online because more competition for sellers would make sellers engage in strategic pricing behaviors such as loss-leader pricing.

Our study adds value to the existing literature on price dispersion in the electronic market in a way that carefully addresses the weaknesses of previous studies. First, we measured the price dispersion of multiple sellers who sell the same product on one platform, thereby controlling platform heterogeneity. By doing so, we could purely compare the heterogeneity of sellers, contrary to previous studies. However, previous studies did not separate the features of retailers from those of products.

Additionally, previous studies captured the platform (ranking or service) differences rather than the differences of embedded retailer attributes. Second, we were not focusing on a single product category; we stretched our scale to compare four different products. Those four products were either search or experience products with well-controlled empirical settings. Therefore, we could provide a complete picture of online price dispersion. Third, we used one of the popular machine learning techniques, text-mining, to capture accurate quality evaluations of products and sellers from other customers. This was notably different from previous studies, since the studies used quality evaluations from the third-party or sales platforms (Dimoka et al. 2012). Additionally, our methods could directly measure the other consumers' evaluations of products and sellers.

Forth, we considered that product uncertainty is inherent in different product categories, as Kim and Krishnan (2015) mentioned. Nevertheless, our research was more sophisticated in that we controlled the product information asymmetry inherent to different product categories even though we collected information from multiple products in different product categories. In order to do that, we used a well-controlled empirical setting, hierarchical linear modeling (HLM), because the information asymmetry of product–seller–market structure (lower level) can be subordinate to the product category (higher level: product information asymmetry inherent to different product categories). Finally, we used the Herfindahl-Hirschman Index (HHI) to understand the market structure and competition on price dispersion in an integrated manner.

In this carefully designed study, we examined the reason for price dispersion for products online. Furthermore, we may have solved the problems of previous studies, which

struggled to generalize conclusions on price dispersion issues in online markets. Finally, we included multiple products in different product categories; we could thus provide a complete picture of online price dispersion.

4.3 Model

We considered three main factors that can influence the standardized price dispersion of each seller. The three main factors were (1) product uncertainty from customer reviews, (2) seller certainty from customer reviews, and (3) market competition.

First, we hypothesized that greater uncertainty about a product from customer reviews will lead to higher risk of the product, leading to greater price dispersion. Second, we hypothesized that the larger a retailer's reputation from customer reviews is realized, the lower the risk is of the seller being recognized by customers, for retailers having their bargaining power. Therefore, when sellers have more certainty, they can set higher prices than their competitors, leading to greater dispersion. Finally, we hypothesized that the higher the market competition, the more likely retailers would be to set lower prices to compete with their competitors, leading to lower price dispersion.

Customers would face purchasing risks when product uncertainty existed. Because of the product uncertainty from other reviews, consumers may have varying perceptions of the actual quality, attributes, and future performance of products (Dimoka & Pavlou 2008; Dimoka et al. 2012; Kim & Krishnan 2015). In particular, product attributes such as non-digital attributes may not be provided by retailers or other buyers (Koppius et al. 2004; Suh & Lee 2005; Suh & Chang 2006). Not knowing those non-digital attributes may influence their product description, which may increase uncertainty about the product (Dimoka et al.

2012). Furthermore, sellers may not be willing to present the full description of product attributes because this may attenuate their bargaining power (Arrow, 1985).

Hypothesis 4.1: *Products about which there is greater uncertainty will have greater price dispersion, because consumers may vary greatly in their evaluations of the true product attributes.*

Retailers with more information or expertise can take opportunistic action (Akerlof 1970). If customers are not fully informed about the attributes and qualities of retailers, their purchasing risks increase. Therefore, we assume that consumers in electronic markets prefer to buy products from highly reputable sellers or experts (Lal & Sarvary, 1999) to reduce purchase risks. Therefore, when there is high certainty about a seller from their customer reviews, the seller would set the price higher, leading to price dispersion.

Hypothesis 4.2: *The other customers' review increasing seller certainty will cause greater price dispersion.*

We utilized text-mining and sentiment analysis of customer reviews on products. Using text-mining, we first obtained fragments of sentences from customer reviews. For sentiment analysis, we applied the natural language process of the fragmented words. Then, we used sentiment analysis to determine whether other customers positively, neutrally, or negatively evaluated the products and retailers. Finally, with the information from the text-mining, we calculated information asymmetry of sellers and products by leveraging entropy and the information gain function.

The previous literature on industry organizations has shown empirical evidence on converged prices with many offline and online sellers (Barron et al. 2004; Haynes &

Thompson 2008). Interestingly, however, some recent studies have argued that competition in the market positively relates to price dispersion (Chandra & Tappata 2011; Wang & Li 2020). Therefore, we wanted to test the hypothesis about the negative relationship between market competition and price dispersion in the online market in terms of integration between various products and two product categories.

Hypothesis 4.3: *Greater competition between retailers with a homogenous product will result in greater price dispersion.*

The HHI was applied to measure market competition, because the index is well-defined and widely used (Rhoades 1993). Equation 4.2 represents the HHI. In previous studies, market competition was defing ed as high when the HHI was over 2,500. A HHI between 2,500 and 1,500 was defined as a moderately concentrated market, while a HHI of less than 1,500 was defined was a low-competition market (Matsumoto et al. 2012: Owen et al. 2007).

$$HHI_j = \sum_{i=1}^{n} S_{ij}^2, \quad (4.2)$$

where i represents individual retailer i, j represents product j, n represents the total number of retailers, and S represents the seller (retailer)

4.4 Data Description

This study aimed to provide explanations for the observed price dispersion for homogeneous products in electronic markets. In particular, we examined the role and impact of (1) product uncertainty from customer reviews, (2) seller certainty from customer reviews, and (3) market competition. We collected information from a single online marketplace (Amazon.com) on 12,409 retailers. The attributes we included were retailers' price, minimum price, product rating, seller rating, seller positive feedback, and customer reviews, among others (August 8th and 18th).

Before we collected the information, we drew a total of eight products to represent search and experience goods (Vaidyananthan & Aggarwal 2003; Leahy 2005; Weathers et al. 2007; Huang et al. 2009). The framework used to finalize the products representing each product category was based upon the previous literature (Nelson 1970; Alba et al. 1997; Peterson et al. 1997; Klein 1998; Girard & Dion 2010). After conducting several pretests, five of the eight products (two products for search and three products for experience goods) were identified as statistically significant products based on a preselection survey through Amazon Mechanical Turk between August 1 and 7 (100 subjects). Table 4.1 shows the pretest results.

Table 4.1 Pretest results

t-value (p-value)	(1)	(2)	(3)	(4)	(5)
Estimating function before use					
(1) SD Cards	-				
(2) Digital Cameras	0.77 (0.44)	-			
(3) Golf Clubs	11.93 ***	12.16 ***	-		
(4) Outdoor Tents	10.13 ***	10.18 ***	-0.64 (0.52)	-	
(5) Skincare Products	10.97 ***	11.16 ***	-1.18 (0.24)	-0.45 (0.65)	-
Estimating quality before use					
(1) SD Cards	-				
(2) Digital Cameras	0.74 (0.46)	-			
(3) Golf Clubs	12.00 ***	11.88 ***	-		
(4) Outdoor Tents	10.66 ***	10.48 ***	-0.62 (0.53)	-	
(5) Skincare Products	12.67 ***	12.58 ***	0.23 (0.82)	0.87 (0.39)	-
Familiarity of the products in online markets					
(1) SD Cards	-				
(2) Digital Cameras	-1.31 (0.19)	-			
(3) Golf Clubs	0.41 (0.69)	1.82 **	-		
(4) Outdoor Tents	1.22 (0.22)	2.75 **	0.83 (0.41)	-	
(5) Skincare Products	1.35 (0.18)	2.82 **	0.97 (0.33)	0.19 (0.85)	-

Based on the pretest results (Table 4.1), SD cards and digital cameras were selected as search products, while golf clubs, outdoor tents, and skincare products were chosen as experience products. We collected information about 12,409 retailers from a single online

marketplace (Amazon.com) to reduce platform heterogeneity. Each of the sellers could sell one of the five products or all the products. Among these 12,409 retailers, we excluded 2,003 subjects because those subjects were either the only sellers who sold a particular product or sellers who sold used products. Among the 10,406 retailers, 5,756 retailers sold search goods, and 4,649 retailers sold experience goods. Table 4.2 shows the descriptive statistics for each product and product category. Since there was significant variation in the average price and standard deviation for each product in Table 4.2, we utilized z-score transformation to make them comparable to each other.

Table 4.2 Descriptive Statistics of Search and Experience Products

	Search Product		Experience Product		
	SD Card	Camera	Golf Club	Outdoor Tents	Skincare Products
Average Retail Price	$58.85	$350.63	$91.69	$131.24	$18.56
Std. Deviation of the Price	$54.51	$624.83	$127.58	$86.97	$10.60
Total # of Products	371	249	415	123	105
Total # of Sellers	3564	2192	2201	1407	1041
Average # of Sellers/Product	9.6	8.8	5.3	11.4	9.9

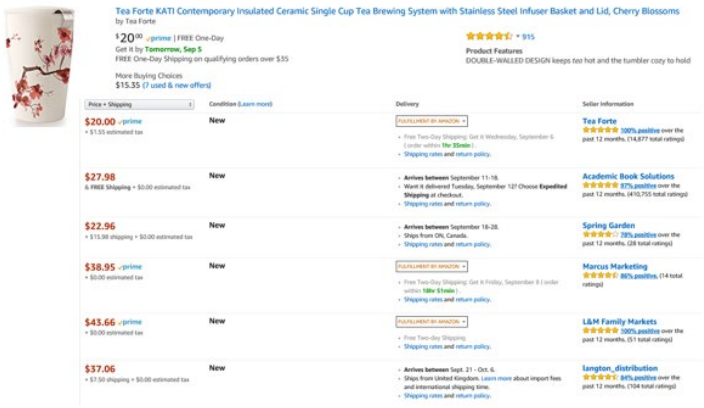

Figure 4.1 Search result of Amazon

First, the minimum price is the lowest price for a product among other retailers' prices when customers searched for a particular product. The lowest price is presented at the top of the retailer list by Amazon.com. Hence, the lowest price is crucial information to evaluate the value of a specific product by customers. Retailer price is the second piece of information that shows the particular price customers need to pay when they decide to buy from one particular retailer. We assumed that these two prices were vital to discuss price dispersion, since price dispersion is defined as the range of prices.

$$ZPrice_{ij} = \frac{Price_{ij} - Price_{minj}}{Standard\ Deviation_j}, \quad (4.1)$$

where *i* represents individual retailer *i* & *j* represents product *j*

Equation 4.1 represents how we determined the extent to which a price of the product *j* of retailer *i* was dispersed and comparable with that of other products. We first calculated the distance difference between the specific retailers' prices and the minimum price of a particular product *j*. Furthermore, each distance difference was divided by the standard deviation of the specific product *j*, called the standardized price distance of the product *j* of retailer *i*. Thus, if there are *n* number of retailers who sell the product *j*, there will be *n* standardized price distance for product *j*. By standardizing so, one can compare the price dispersion between retailers selling the same product. Additionally, one can compare the price differences between retailers selling different products in the same and other categories.

4.5 Methodology

To test our research hypotheses between three factors (information asymmetry of sellers, products, and market structure) and price dispersion, we decided to use HLM. HLM is a complex form of the ordinary least squares model used to analyze variance in outcome

variables when predictor variables are at varying hierarchical levels (Woltman et al. 2012). In order words, HLM explains the overlapped variance in hierarchically structured data. Thus, the technique accurately estimates lower-level coefficients and their implementation in evaluating higher-level outcomes (Hofmann 1997). For example, in our case, the highest level of the hierarchy was the product category variable, which differentiated whether the product was in search or experience goods. The lower level of the order was information asymmetry of products.

Nelson (1970) classified types of goods into search goods and experience goods. Search goods are defined as products for which accurate attributes can be obtained ahead of purchasing. Hence, the non-digital characteristics can generally be discoverable without the consumer interacting with the product. Therefore, if a product is categorized as a search good, the level of information asymmetry for the product would be low. On the contrary, experience goods in electronic markets have more significant non-digital features that cannot be known without direct experience (Klein 1998). Hence, the level of information asymmetry for the products is high for experience goods.

Similarly, because experience goods have unclear non-digital attributes captured beforehand, consumers may face opportunisms from sellers (e.g., the moral hazard problem; Akerlof 1970; Dellarocas 2005). Hence, sellers' attributes in experience goods may play a more significant role (Lal & Sarvary 1999). Additionally, because there is possible room for retailers to be opportunistic, more retailers might sell experience goods, leading to price competition. HLM in equations 4.3 to 4.10 is a relevant methodology to test our hypotheses and control such hierarchical impact.

$$ZPrice_{ijk} = \beta_{ok} + \beta_{1k}X_{1ijk} + \beta_{2k}X_{2ijk} + \beta_{3k}X_{3jk} + \beta_{4k}X_{4jk} + \beta_{5k}X_k + \beta_{6k}X_{1ijk}X_k + \beta_{7k}X_{2ijk}X_k + \beta_{8k}X_{3jk}X_k + \beta_{10k}X_{4jk}X_k + \gamma_{ijk}, \quad (4.3)$$

$X_{1ijk}: Short -$

$term\ (30\ days) reputation\ of\ a\ seller\ i\ who\ sells\ product\ j\ within\ product\ category\ k$

,where $X_{1ijk} = MaxEntropy\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right) -$

$$\left(\frac{Total\ Negative\ Review\ for\ 30\ days_i}{Total\ Review_i}\log_2\frac{Total\ Negative\ Review\ for\ 30\ days_i}{Total\ Review_i}\right) +$$

$$\frac{Total\ Neutral\ Review\ for\ 30\ days_i}{Total\ Review_i}\log_2\frac{Total\ Neutral\ Review\ for\ 30\ days_i}{Total\ Review_i} +$$

$$\frac{Total\ Positive\ Review\ for\ 30 days_i}{Total\ Review_i}\log_2\frac{Total\ Positive\ Review\ for\ 30\ days_i}{Total\ Review_i}) \quad (4.4)$$

$X_{2ijk}: Long -$

$term\ (Lifetime) reputation\ of\ a\ seller\ i\ who\ sells\ product\ j\ within\ product\ category\ k$

,where $X_{2ijk} = MaxEntropy\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right) -$

$$\left(\frac{Total\ Negative\ Review\ for\ Lifetime_i}{Total\ Review_i}\log_2\frac{Total\ Negative\ Review\ for\ Lifetime_i}{Total\ Review_i} +\right.$$

$$\frac{Total\ Neutral\ Review\ for\ Lifetime_i}{Total\ Review_i}\log_2\frac{Total\ Neutral\ Review\ for\ Lifetime_i}{Total\ Review_i} +$$

$$\left.\frac{Total\ Positive\ Review\ for\ Lifetime_i}{Total\ Review_i}\log_2\frac{Total\ Positive\ Review\ for\ Lifetime_i}{Total\ Review_i}\right) \quad (4.5)$$

$X_{3jk}: Inoformation\ Asymmetry\ of\ a\ product\ j\ within\ product\ category\ k$

,where $X_{3jk} = MaxEntropy\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right) - \frac{Total\ Negative\ Review_j}{Total\ Review_j}\log_2\frac{Total\ Negative\ Review_j}{Total\ Review_j} +$

$$\frac{Total\ Neutral\ Review_j}{Total\ Review_j}\log_2\frac{Total\ Neutral\ Review_j}{Total\ Review_j} + \frac{Total\ Positive\ Review_j}{Total\ Review_j}\log_2\frac{Total\ Positive\ Review_j}{Total\ Review_j}$$

(4.6)

$$X_{4jk} = \begin{cases} High, where\ \ HHI_j = \sum_{i=1}^{n} S_{ij}^2 \geq 2500 \\ Moderate, where\ \ HHI_j = \sum_{i=1}^{n} S_{ij}^2 \geq 1500 \\ Low, else \end{cases} \quad (4.7)$$

$$X_{\mathrm{k}} = \begin{cases} 1, \ when\ Product\ j\ is\ Experience\ product \\ 0, when\ Product\ j\ is\ Search\ Product \end{cases} \quad (4.8)$$

$\gamma_{ijk} =$

$random\ error\ associated\ with\ seller\ i\ \&\ product\ j\ within\ the\ product\ category\ k$

, where:

$$\beta_{ok} = \delta_{00} + \delta_{01} X_{\mathrm{k}} + \epsilon_{OJ} \quad (4.9)$$

$$\beta_{Nk} = \delta_{N0} + \delta_{N1} X_{\mathrm{k}} + \epsilon_{NJ} \text{ with N is 1~10} \quad (4.10)$$

To measure the information asymmetry of a product, we used the entropy formula in equation 4.6. Entropy is a scientific concept and a physical property most commonly associated with a state of disorder (Segal 1960). In information theory, entropy is a measure of information uncertainty (Anand & Bianconi 2009). Hence, the greater the entropy of information, the higher the uncertainty in concluding. In our example of customer reviews on a product, one can analyze the reviews in positive, neutral, and negative tones. Therefore, if there are equal numbers of positive, neutral, and negative words about a product in the reviews, the information about that product is assumed to be highly uncertain. In the same logic, if there is a dominant tone of reviews, the product has less uncertainty of information. Equation 4.11 represents the entropy formula to measure product uncertainty.

$$ProEnt_j = H\left(\frac{Positive\ Review_j}{Total\ Review_j}, \frac{Neutral\ Review_j}{Total\ Review_j}, \frac{Negative\ Review_j}{Total\ Review_j}\right)$$

$$= -\sum_{All\ Cases}\left(\frac{Each\ Case_j}{Total\ Review_j}\log_2\frac{Each\ Case_j}{Total\ Review_j}\right) \quad (4.11)$$

, where j is a product j

To measure a seller's information asymmetry, we used the same entropy formula used in equation 4.11 and applied the same logic. However, the two differences were as follows:

(1) The reviews about a seller were collected over a short (30 days) and long term (lifetime). Therefore, the information asymmetry for a seller was measured by two variables: 30 days and lifetime.

(2) We captured the information certainty instead of the information uncertainty of sellers. Therefore, we calculated the information gain function, or the difference between the maximum entropy and the entropy with reviews: X_{2jk} and X_{3jk} (equations 4.4 and 4.5).

4.6 Results

This section presents the statistical results using HLM of price dispersion on product uncertainty, seller certainty, and market structure introduced in Section 5. The price dispersion results are presented in Table 4.3.

Table 4.3 Price Dispersion and predictors

Random Effects:	Variance	Std. Dev.	2.5% Lower CI	97.5% Upper CI		
Experience Products	0.000	0.000	.000	.038		
Residual	1.312	1.146	1.129	1.161		
Fixed Effects:						
Variables	Coeff.	Std. Errors	2.5% Lower CI	97.5% Upper CI	t-value	Sig.
-Intercept	.439	.278	-.106	.985	1.578	*

-Product Entropy	.723	.116	.496	.951	6.240	***
-Information Gains of Sellers						
Short-term	-.205	.042	-.287	-.123	-4.914	***
Long-term:	.177	.052	.075	.278	3.419	***
-HHI:						
High Competition	-.502	.254	-1.000	-.003	-1.972	**
Moderate Competition	-.155	.257	.658	.348	-.605	
Low Competition	-.015	.251	-.507	.477	-.061	
-Product Type:	.526	.224	.088	.964	2.353	***
1. Experience Product						
2. Experience Product & Short-term IG of Seller	.237	.071	.097	.376	3.328	***
3. Experience Product & Long-term IG of Seller	-.050	.081	-.209	.108	-.619	
4. Experience Product & HHI: Moderate	-.106	.101	-.305	.092	-1.049	
5. Experience Product & HHI: Low	.087	.079	-.067	.241	1.108	
6. Experience Product & Product Entropy	.276	.178	-.073	.626	1.551	*

P-value: <.01: *** | <.05: ** | <.1: *
AIC: 32380.6 | BIC: 32489.4

When controlling product categories, we found that product uncertainty, seller certainty (either short-term or lifetime), and high competition are the significant factors correlated with price dispersion. When a customer reads product reviews from other customers and the product uncertainty (product entropy) increased by one unit, the price dispersion increased by .723. Meanwhile, our second hypothesis was partially accepted. The hypothesis—price dispersion increases when seller certainty increases—was accepted when the seller certainty was consistently good for a lifetime. In this case, the price was dispersed by .177 units, and it was statistically significant. However, if the seller certainty increased by one unit for 30 days, the price dispersion instead decreased by .205 units. Finally, our third hypothesis that price dispersion will be reduced in competitive markets

was accepted. When HHI indicated that the market was highly competitive, the price dispersion decreased by .502 and was statistically significant. However, when the market had moderate or low competition, price dispersion was statistically insignificant.

Furthermore, when the product category was considered, the price dispersion of the experience product category was as high as .526 units. In addition, the price dispersion of .238 was statistically significant, even for the one unit increases in seller certainty for 30 days. Lastly, in the experience product category, where product uncertainty was one unit high due to customer reviews, the price dispersion was .276 units.

Table 4.4 Price Dispersion Scenarios

Price Dispersion Scenarios	Experience Goods	Search Goods
Basis: Price Dispersion	.965	.439
Scenario1: Price Dispersion + Short-term IG of Seller	.997	.234
Scenario2: Price Dispersion + Long-term IG of Seller	1.142	.616
Scenario3: Price Dispersion + Short-term IG of Seller + High Competition	.495	-.268
Scenario4: Price Dispersion + Long-term IG of Seller + High Competition	.640	.114
Scenario5: Price Dispersion + Product Entropy	1.964	1.162
Scenario6: Price Dispersion + Short-term IG of Seller + Product Entropy	1.996	.957
Scenario7: Price Dispersion + Long-term IG of Seller + Product Entropy	2.141	1.339
Scenario8: Price Dispersion + Short-term IG of Seller + High Competition + Product Entropy	1.494	.455
Scenario9: Price Dispersion + Long-term IG of Seller + High Competition + Product Entropy	1.639	.837

The results of our analysis of several scenarios considering statistical significance are shown in Table 4.4. Overall, one can observe that the experience product category was more

price dispersed in all respects than the search product category. Interestingly, when the seller certainty increased by one unit for 30 days and the seller sold products in the search product category in a highly competitive environment, the price dispersion decreased by .268 units. However, in all other cases, we found the existence of price dispersion.

4.7 Conclusion

This study aimed to better understand and explain the market inefficiency in price dispersion for homogeneous products in electronic markets. In particular, we aimed to answer whether the three factors (information asymmetry, ex-ante heterogeneity, and market structure) of price dispersion would help explain the persistent price dispersion in electronic markets.

Using HLM, we found that price dispersion still exists in electronic markets. In addition, we found that all three factors were the source of the market inefficiency. More specifically, uncertainty in how a product was valued by other customers was a major source of price dispersion in the electronic market.

Second, customer reviews about a seller (seller heterogeneity or certainty) resulted in different price dispersion directions. When measured over a lifetime, seller certainty caused sellers behave opportunistically, leading to price dispersion. However, the short-term (30 days) increment of seller certainty reduced price dispersion. Since we did not measure causality but rather measured correlation, we might need to conduct further research about the relationship. In other words, if one were to analyze this data in a time-series manner, perhaps sellers would offer a homogeneous product cheaper than their competitors to

increase their short-term certainty. Moreover, the lower price of a homogeneous product may reduce the risk of purchasing for customers, increasing the sellers' certainty.

Third, the market competition was high, and in that case one can expect price dispersion to decrease. However, if market competition was moderate or low, the market competition would be statistically insignificant to price dispersion.

Finally, sellers in the experience product category could expect greater price dispersion than those in the search product category. Since experience products have more non-digital attributes than search products, consumers may rely on a seller's heterogeneity to lower uncertainty. Therefore, sellers can price the homogeneous product higher than their competitors when they have higher certainty and sell the experience products. Sellers can thus price the homogeneous product higher than their competitors by utilizing seller certainty and market structure.

The current study provided several implications for both academic areas and business practice. First, this study provided profound explanations on the micro-level such as why price dispersion exists in electronic markets and how price dispersion increases depending on uncertainty, heterogeneity, and market structure. Hence, it is possible to further investigate price dispersion when consumers change their risk preferences in different marketplaces. Moreover, researchers can extend this research to other product categories (e.g., information goods). Additionally, we showed that text-mining and sentiment analysis with the natural language process are an excellent method to measure the information asymmetry of products and sellers.

In practice, sellers could have a price-setter strategy based on their products and reputations. For example, a seller with the same attributes as sellers could set a higher price to sell one unit of an experience good than to sell one unit of a search product. Additionally, sellers could bribe consumers to obtain cumulative lifetime certainty, allowing them to set higher prices later. Moreover, text-mining is an essential application for electronic market sellers who sell homogenous products, because text-mining increases understanding of the uncertainty of products, sellers, and competitors.

As do all studies, this study had several limitations. One of the limitations was that this study only included five products representing search and experience goods. If we had ample product samples, we would be able to explain the results by comparing standardized price dispersion in different conditions in multiple product categories. Overcoming these limitations can provide fruitful directions for future research.

In summary, this study gave a micro-level of explanations for the uncertainty and price dispersion in electronic markets. Furthermore, the study showed that different product classifications impact levels of price dispersions.

References

Akerlof, G.A. (1970). The Market for "Lemons": Quality Uncertainty and Market Mechanism, *The Quarterly Journal of Economics*, 84(3), pp. 488-500.

Alba, J., Lynch, J., Weitz, B, Jainszewski, C., Lutz, R., Sawyer, A, and Wood, S. (1997). Interactive Home Shopping: Consumer, Retailer, and Manufacturer Incentives to Participate in Electronic Market-places, *Journal of Marketing*, 61(3), pp. 38-53.

Anand, K., & Bianconi, G. (2009). Entropy measures for networks: Toward an information theory of complex topologies. *Physical Review E*, *80*(4), 045102.

Ancarani, F., and Shankar, V. (2004). Price Levels and Price Dispersion Within and Across Multiple Retailer Types: Further Evidence and Extension. *Journal of Academy of Marketing Science*, 32(2), pp. 176-187.

Arrow, K. J. (1985). Informational structure of the firm. *The American Economic Review*, *75*(2), 303-307.

Ba, S. and Pavlou, P.A. (2002). Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior, *MIS Quarterly*, 26(3), pp. 243-268.

Bailey, J.P. (1998a). Electronic Commerce: Prices and Consumer Issues for three Products: Books, Compact Discs, and Software, *Report OCDE/GP, 98(4).* Organization for Economic Co-Operation and Development.

Bailey, J.P. (1998b). Intermediation and Electronic Markets: Aggregation and Pricing in Internet Commerce. Ph.D. Dissertation. Program in Technology, Management, and Policy. MIT, Cambridge, MA.

Baker, J., Levy, M., and Grewal, D. (1992). An Experimental Approach to Making Retail Store Environmental Decisions, Journal of Retailing, 68(4). pp. 445-460.

Bakos, Y. (1997). Reducing Buyer Search Costs: implications for Electronic Marketplaces, *Management Science,* 43(12), pp. 1679-1692.

Barron, J. M., Taylor, B. A., & Umbeck, J. R. (2004). Number of sellers, average prices, and price dispersion. *International Journal of Industrial Organization*, *22*(8-9), 1041-1066.

Baye, M. R., & Morgan, J. (2001). Information gatekeepers on the internet and the competitiveness of homogeneous product markets. *American Economic Review*, *91*(3), 454-474.

Berry, L.L. (2000). Cultivating service brand equity, *Journal of the Academy of Marketing Science,* 28(1), pp. 128-137.

Brown, J. R., & Goolsbee, A. (2002). Does the Internet make markets more competitive? Evidence from the life insurance industry. *Journal of political economy*, *110*(3), 481-507.

Brynjolfsson, E., Hu, Y., and Smith, M., (2003). Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product variety at Online Booksellers, *Management Science*, 49(11), pp. 1580-1596.

Brynjolfsson, E. and Smith, M. (2000a). The Great Equalizer? Consumer Behavior at Internet Shopbots, *Working Paper*, MIT, Cambridge, MA.

Brynjolfsson, E. and Smith, M., (2000b). Frictionless Commerce? A Comparison of Internet and Conventional Retailers, *Management Science*, 46,(4), pp.563-585.

Burdett, K. and Judd, K.L. (1983). Equilibrium Price Dispersion, *Econometrica,* 51(4). pp. 955-969.

Chandra, A., & Tappata, M. (2011). Consumer search and dynamic price dispersion: an application to gasoline markets. *The RAND Journal of Economics*, *42*(4), 681-704.

Clay, K., Krishnan, R., and Wolff, E. (2001). Prices and Price Dispersion on the Web: Evidence from the online book industry, *Journal of Industrial Economics*, 49(4), pp. 521-540.

Cui, Y., Orhun, A. Y., & Duenyas, I. (2019). How price dispersion changes when upgrades are introduced: Theory and empirical evidence from the airline industry. *Management Science*, *65*(8), 3835-3852.

Degeratu, A., Rangaswamy, A., and Wu, J. (2000). Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes, *International Journal of Research in Marketing*, 17(1), pp. 55-78.

Dellarocas, C. (2003). The Digitization of Word-of-Mouth: Promise and Challenges of Online Feedback Mechanisms, *Management Science*, 49(10), pp. 1407-1424.

Dellarocas, C. (2005). Reputation Mechanism Design in Online Trading Environments with Pure Moral Hazard, *Information Systems Research*, 16(2), pp. 209-230.

Dimoka, A. and Pavlou, P.A. (2008). Understanding and Mitigating Product Uncertainty in Online Auction Marketplaces, *Industry Studies Conference Paper*. Available at SSRN: http://ssrn.com/abstract=1135006 or http://dx.doi.org/10.2139/ssrn.1135006

Dimoka, A., Hong, Y., and Pavlou, P.A. (2012). On Product Uncertainty in Online Markets: Theory and Evidence, *MIS Quarterly*, 36(2), pp. 395-426.

Girard, T. and Dion, P. (2010). Validating the Search, Experience, and Credence product classification framework, *Journal of Business Research*, 63(9 & 10), pp.1079-1087.

Haynes, M., & Thompson, S. (2008). Price, price dispersion and number of sellers at a low entry cost shopbot. *International Journal of Industrial Organization*, *26*(2), 459-472.

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of management*, *23*(6), 723-744.

Huang, P., Lurie, N.H., and Mitra, S. (2009). Search for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods, *Journal of Marketing*, 73(2), pp. 55-69.

Jarvenpaa, S.L., Tractinsky, N., and Saarinen. (2000). Consumer Trust in an Internet Store: A Cross-Cultural Validation, *Journal of Computer-Mediated Communication*, 5(2), pp. 1-33. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1999.tb00337.x/full

Kauffman, R. and Wood, C. (2000). Follow the Leader? Strategic Pricing in E-Commerce, *ICIS 2000 Proceeding Paper*, 14. Available at: http://aisnet.org/icis2000/14

Kim, Y. and Krishnan, R. (2015). On Product-Level Uncertainty and Online Purchase Behavior: An Empirical Analysis, *Management Science*, Published online in Articles in Advance: 02 April 2015.

Klein, L.R. (1998). Evaluating the Potential of Interactive Media Through a New Lens: Search Versus Experience Goods, *Journal of Business Research*, 41(3), pp. 195-203.

Koppius, O.R., Van Heck, E., and Wolters, M. (2004). The Importance of Product Representation Online: Empirical Results and Implications for Electronic Markets, *Decision Support Systems*, 38(2), pp. 161-169.

Lal, R. and Sarvary, M. (1999). When and How is the Internet Likely to decrease Price Competition, *Marketing Science*, 18(4), pp. 485-503.

Leahy, A. S. (2005). Search and Experience Goods: Evidence from the 1960's and 70's. *Journal of Applied Business Research (JABR)*, *21*(1).

Lee, H.G. (1998). Do Electronic Marketplaces Lower the Price of Goods? *Communications of the ACM*, 41(1), pp. 73-80.

Lee, H.G. and Clark, T.H. (1996). Market prcess reengineering through electronic market systems: opportunities and challenges, Journal of Management Information Systems, 13(3), 1996, pp. 113-136.

Lynch, J., and Ariely, D., (2000). Wine Online: Search Costs Affect Competition on Price, Quality, and Distribution, *Marketing Science*, 19(1), pp. 83-103.

Matsumoto, A., Merlone, U., & Szidarovszky, F. (2012). Some notes on applying the Herfindahl–Hirschman Index. *Applied Economics Letters*, *19*(2), 181-184.

Nelson, P. (1970). Information and Consumer Behavior, *Journal of Political Economy*, 78(2), pp. 311-329.

Nelson, R. A., Cohen, R., & Rasmussen, F. R. (2007). An analysis of pricing strategy and price dispersion on the internet. *Eastern Economic Journal*, *33*(1), 95-110.

Owen, P. D., Ryan, M., & Weatherston, C. R. (2007). Measuring competitive balance in professional team sports using the Herfindahl-Hirschman index. *Review of Industrial Organization*, *31*(4), 289-302.

Pan, X., Shankar, V., and Ratchford, B.T. (2002). Price Competition Between Pure Play vs. Bricks-and-clicks e-tailers: Analytical Model and Empirical Analysis, July, Available at SSRN: http://ssrn.com/abstract=328840 or http://dx.doi.org/10.2139/ssrn.328840

Pavlou, P.A. (2003). Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model, *International Journal of Electronic Commerce*, 7(3), pp. 101-134.

Pavlou, P.A. and Dimoka, A. (2006). The Nature and Role of Feedback Text Comments in Online Marketplace: Implications for Trust Building, Price Premiums, and Seller Differentiation, *Information Systems Research*, 17(4), pp. 392-414.

Pavlou, P.A. and Gefan, D. (2004). Building Effective Online Marketplaces with Institution-Based Trust, *Information Systems Research*, 15(1), pp. 37-59.

Pavlou, P.A. and Gefan, D. (2005). Psychological Contract Violation in Online Marketplaces: Antecedents, Consequences, and Moderating Role, *Information Systems Research*, 16(4), pp. 372-399.

Pratt, J.W., Wise, D., and Zeckhauser, R. (1979). Price Differences in Almost Competitive Markets, *The Quarterly Journal of Economics*, 93(2), pp. 189-211.

Ratchford, B. T., Pan, X., & Shankar, V. (2003). On the efficiency of internet markets for consumer goods. *Journal of Public Policy & Marketing*, *22*(1), 4-16.

Peterson, R.A., Balsubramanian, S., and Bronenberg, B.J. (1997). Exploring the Implications of the Internet for Consumer Marketing, *Journal of the Academy of Marketing Science*, 25(4), pp. 329-346.

Rhoades, S. A. (1993). The herfindahl-hirschman index. *Fed. Res. Bull.*, *79*, 188.

Salop, S. and Stiglitz, J. (1977). Bargains and Ripoffs: A model of Monopolistically Competitive Price Dispersion, *The Review of Economic Studies*, 44(3), pp. 493-510.

Segal, I. E. (1960). A note on the concept of entropy. *Journal of Mathematics and Mechanics*, 623-629.

Smith, M.D. (2001). The Law of One Price? The Impact of IT-Enabled Markets on Consumer Search and Retailer Pricing. *Working Paper*, Carnegie Mellon University, PA.

Stigler, G.J. (1961). The Economics of Information, *Journal of Political Economy*, 69(3). Pp. 213-225.

Suh, K.S. and Lee, Y.E. (2005).The Effects of Virtual Reality on Consumer Learning: An Empirical Investigation, *MIS Quarterly*, 29(4), pp. 673-697.

Suh, K.S. and Chang, S.H. (2006). User Interfaces and Consumer Perceptions of Online Stores: The Role of Telepresence, *Behaviour & Information Technology*, 25(2), pp. 99-113.

Tang, F. and Xing, X. (2001). Will the Growth of Multi-Channel Retailing Diminish the Pricing Efficiency of the Web?, *Journal of Retailing*, 77(3), pp. 319-333.

Vaidyanathan, R., & Aggarwal, P. (2003). Who is the fairest of them all? An attributional approach to price fairness perceptions. *Journal of Business Research*, *56*(6), 453-463.

Wang, W., & Li, F. (2020). What determines online transaction price dispersion? Evidence from the largest online platform in China. *Electronic Commerce Research and Applications*, *42*, 100968.

Weathers, D., Sharma, S., & Wood, S. L. (2007). Effects of online communication practices on consumer perceptions of performance uncertainty for search and experience goods. *Journal of retailing*, *83*(4), 393-401.

CHAPTER5
Conclusion

In this dissertation, we studied the economic value of machine learning algorithms. In this work, we have attempted to go beyond the current state of knowledge. Through rigorous analysis with rich data from multiple sources, we develop more insights into the applications of machine learning and the process of value creation., To this end, we examine different application cases: two for demand prediction and one for sentiment analysis. In addition, we present a robust methodology to evaluate its use in demand forecasting by comparing machine learning algorithms with a variety of traditional econometrics models and human experts. For the price dispersion study, we also include a variety of products in different product categories to facilitate generalization. Finally, we have considered human roles and contributions to the economic value creation process with machine learning algorithms. This last chapter summarizes the main conclusions of this dissertation and gives some insights for future research.

5.1 Machine Learning System for Non-Parametric Demand Prediction

We have compared daily demand forecasting results for a theme park from machine learning algorithms and twelve traditional econometric models. We find considerable improvement in forecasting performance in demand prediction using our machine learning approach compared to twelve conventional econometric models. For example, the machine learning algorithm's performance has a MAPE of 46.54% and accuracy of 73.74%, comparable to or better than most forecasting models. More specifically, the GPAI does not produce better forecasting results in MAPE and accuracy than Bayesian Structural Time Series (MAPE as 42.09% & accuracy as 75.42%) and OLS (MAPE as 45.24% & accuracy as

75.42%). The differences in the prediction performance between GPAI and the other two models are statistically insignificant.

The logistic analysis shows that non-parametric methods such as machine learning (DNN), BSTS, and OLS outperform other parametric econometrics for the more non-stationary and nonlinear future demand trends. For example, the best performing parametric econometric model – Seasonal Naïve (Naïve 2) - has a .46 odds ratio compared to GPAI. In contrast, parametric time series models outperform non-parametric models, including the machine learning approach, for the more linear and stationery demand trends. For example, the best performing parametric econometric model – Seasonal Naïve (Naïve 2) - has a 1.49 odds ratio compared to GPAI.

Unlike prior studies, we can conclude that machine learning algorithms outperform daily demand prediction compared to other econometric models. Also, we conclude that non-parametric models are better performing when we have nonlinear and non-stationary demand prediction cases. Yet, parametric econometric models fit better when we have linear and stationary demand prediction cases.

These results have important managerial implications. First, business practice must check the trends in daily demand before they invest in new prediction systems. If demand is more nonlinear and non-stationary, leveraging machine learning or other non-parametric econometric methods is a good way to increase prediction accuracy. However, when demand is more stable and linear, other parametric econometrics are better in prediction accuracy. Second, business practices must comprehensively consider diverse variables and past visitor trends to enable accurate demand forecasting from machine learning algorithms.

This essay is not without limitations. We consider twelve econometric benchmarks. However, it is possible that other traditional econometric models may outperform the machine learning prediction accuracy. Also, we tear out our demand prediction data into two cases: one that shows more linear and stable cases and another that shows more transformative and nonlinear. However, if there were two significantly different training and testing data sets for demand prediction, the two situations could be studied more closely.

5.2 Demand Prediction using ML: A Human-in-the-Loop Approach

Chapter 3 compares prediction accuracy, prediction consistency, direct economic benefits, and indirect economic benefits of HITL AI to general-purpose AI and human experts. We do so because we wish to study whether machine algorithms substitute or complement human experts.

Based on six months of data from field experiments, we find robust evidence for the complementarity of human experts and machine algorithms. The field experiment shows that HITL AI forecasting performance is significantly better than both human experts and GPAI in accuracy and consistency for daily demand. For example, human experts have a MAPE of 36.68% and an accuracy of 79.33%. On the other hand, the GPAI has an even higher MAPE of 46.54%, while also being slightly more inaccurate from a business decision-making perspective (73.74%). Clearly, human experts were able to make better decisions compared to an off-the-shelf GPAI. The HITL approach, on the other hand, performs exceptionally well compared to both human experts and GPAI. The MAPE of HITL is 18.61%, whereas the model makes predictions accuracy of 93.85%. From a consistency perspective, the temporal performance of the three approaches shows that HITL AI has a minor average standard

deviation of MAPE compared to human experts and GPAI. Similar results can be found for the prediction accuracy for dynamic environments.

We calculate type 1 and type 2 error costs from the economic benefit perspective because we expect all three models to produce under- or over-forecasting cases. When the models are over-forecasted, the mis-prediction cost per day due to type 1 errors for HITL AI is $2,229 compared to $2,861 for human experts and $2,523 for GPAI. In addition, when the models are under-forecasted, the mis-prediction cost per day due to type 2 errors for HITL AI is $8,692 compared to $16,271 for human experts and $17,629 for GPAI. Therefore, the total error costs to total revenue are 2.38% for HITL AI compared to 5.40% for human experts and 6.25% for GPAI.

Finally, from the indirect economic benefit perspective, we find that HITL AI requires only 15 minutes to process the variables for predictions while GPAI takes two days to compute. This shorter lead time of the HITL approach cannot be overstated because the theme park will have enough time to respond to an emergency. More importantly, better and reliable planning for the theme park is related to customer satisfaction.

All of these results have several implications for both academic and business practice. First, our results show that human expertise complements, rather than substitutes for, the ML algorithm. Since human experts have domain-specific knowhow, while machines have superior computational power, HITL promises to bring the best of both worlds to improve the prediction accuracy of ML systems.

For practice, we hope this study allows managers to reduce failure rates of AI systems and improve their performance. We see HITL as valuable for solving complex problems

where the solution space is enormous or the data are sparse. What sets HITL AI apart from GPAI and human experts is that human experts and machine algorithms can be combined to offset the weaknesses of each alone. With the machine learning algorithms/ AI investment decision, companies must invest considerable agony over utilizing, developing, updating, and maintaining the ML/ AI algorithms with their human experts.

For academia, our research has shown that HITL AI, designed according to a complementary relationship between human experts and ML algorithms, can accrue significant economic value. Many studies have discussed human role changes and replacement in machine learning algorithms. In particular, the cost of data processes and forecasts was reduced with the machine learning/ AI application, expecting a replacement of experts represented by experts. On the contrary, according to our research, it is not desirable to replace human experts on the cost side.

Turning to limitations, while we have a well-designed field experiment and promising results, we are limited by studying just one setting, making it hard to generalize. We encourage other scholars to test the HITL approach in different business contexts.

5.3 Revisiting Price Dispersion in Electronic Markets

In Chapter 4, we add value to the existing literature on price dispersion in electronic markets in a way that carefully addresses the weaknesses of previous studies. We collect information from a single online marketplace on 12,409 retailers to study the reason for price dispersion in electronic markets. We use a well-controlled empirical setting, Hierarchical Linear Modeling (HLM). After controlling for product category, we find that the increases in product uncertainty are correlated to price dispersion by .723. Also, with these

control for product categories, lifetime seller certainty increases are associated with price dispersion by .177. Finally, the highly concentrated market structure decreases price dispersion by .502.

In this study, we leverage a category of machine learning algorithms – text-mining. We utilize text-mining in a sentiment analysis of customer reviews on a product. Using text-mining, we first have fragments of the sentences of customer reviews. For sentiment analysis, we apply a natural language process to the fragmented words. Finally, we use sentiment analysis to determine whether other customers have positively, neutrally, or negatively evaluated the products and retailers. Text-mining helps understand the uncertainty of products and sellers in the reviews of different customers. Text-mining is an essential application for directly measuring information asymmetry.

This chapter contributes to the literature in two ways. First, we introduce a new way to measure information asymmetry (uncertainty and certainty). We first use text-mining and sentiment analysis using the natural language process, and then we calculate the level of uncertainty (or certainty) using entropy and an information gain function. Our new way of measuring information asymmetry works better to understand the existence of market inefficiency (price dispersion) than previous methods (e.g., utilizing star ratings, certifications, or third-party evaluations). Second, we provide a generalizable explanation of the existence of price dispersion in the electronic market by minimizing the limitations of the prior literature. To minimize the limitations of the previous study, we carefully collect the data from one online shop, include multiple products in different product categories and retailers, and apply rigorous empirical methods.

This study has several limitations. First, we are unable to collect actual customer purchase data, which can provide a complete picture of price dispersion in electric markets. Also, we do not have time-series data of the price-changing history of sellers. With the time-series data on price from sellers and actual purchase data from consumers, the explanations of price dispersion online could be somewhat different. Second, this study utilizes 10,406 retailers who sell 5,756 products in search product category and 4,649 products in experience product category. However, when we sort this out into the upper product categories, there are only 5 product categories of SD cards, Digital Cameras, Golf Clubs, Outdoor Tents, and Skincare Products. If we could include ample product samples, we would provide an intensive explanation of the results by comparing the standardized price dispersion in different conditions of multiple product categories. Overcoming these limitations can provide fruitful directions for future research.

5.4 Summary

This dissertation has shown that machine learning is a powerful tool and set of approaches, encompassing algorithms and data. While we have shown its value in demand prediction, especially when complemented by human experts, we believe the scope of application is extensive. Moreover, we also demonstrate its use in expanding our understanding of other economic phenomena like price dispersion.

Some have argued that AI is a general-purpose technology, with the potential to transform the economy and society. We hope our study is one small, but significant, step towards furthering our understanding of transformative technology.