Identifying Arbitrage Opportunities in Retail Markets with Artificial Intelligence

Jitsama Tanlamai
Corresponding author
Ph.D. Student
HEC Montreal, 3000, Chemin de la Cote-Sainte-Catherine, Montreal, H3T 2A7, QC, Canada.
jitsama.tanlamai@hec.ca

Warut Khern-am-nuai
Associate Professor of Information Systems
Desautels Faculty of Management, McGill University, 1001 Rue Sherbrooke O., Montreal, H3A 1G5, QC, Canada.

Yossiri Adulyasak
Associate Professor of Operations Management
HEC Montreal, 3000, Chemin de la Cote-Sainte-Catherine, Montreal, H3T 2A7, QC, Canada.

Abstract

This study uses an artificial intelligence (AI) model to identify arbitrage opportunities in the retail marketplace. Specifically, we develop an AI model to predict the optimal purchasing point based on the price movement of products in the market. Our model is trained on a large dataset collected from an online marketplace in the United States. Our model is enhanced by incorporating user-generated content (UGC), which is empirically proven to be significantly informative. Overall, the AI model attains more than 90% precision rate, while the recall rate is higher than 80% in an out-of-sample test. In addition, we conduct a field experiment to verify the external validity of the AI model in a real-life setting. Our model identifies 293 arbitrage opportunities during a one-year field experiment and generates a profit of $7.06 per arbitrage opportunity. The result demonstrates that AI performs exceptionally well in identifying arbitrage opportunities in retail markets with tangible economic values. Our results also yield important implications regarding the role of AI in the society, both from the consumer and firm perspectives.

Keywords: arbitrage; field experiment; artifical intellegence; user-generated content; retail markets

Declarations

Funding details: This work is financially supported by the Social Science and Humanities Research Council of Canada (SSHRC) grant number 430-2019-00520.

Conflict of interest: The authors declare no competing interests

Data Availability Statement: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Author contributions

Conceptualization: Y. Adulyasak, W. Khern-am-nuai
Methodology: J. Tanlamai, Y. Adulyasak, W. Khern-am-nuai
Investigation: J. Tanlamai, Y. Adulyasak, W. Khern-am-nuai
Visualization: J. Tanlamai, Y. Adulyasak, W. Khern-am-nuai
Writing: J. Tanlamai, Y. Adulyasak, W. Khern-am-nuai

1. Introduction

Market efficiency in electronic commerce has been one of the core research topics that yield important implications to our society in the past decade (Bapna et al. 2018). As the efficiency of the market is generally driven by activities between buyers and sellers (McMillan 2003), prior works in this area have studied how information technologies enhance buyer–seller interactions and the overall impact on market efficiency (Ghose and Yao 2011; Wang et al. 2022). In particular, artificial intelligence (AI) agents, which are commonly utilized on electronic commerce platforms, can reduce buyer-seller interactions and provide information that benefits customers' interests (Serenko et al. 2007). Nevertheless, arbitrage activities, in which people buy products to resell for profits, are still consistently observed (Overby and Forman 2015; Subramanian and Overby 2017; Yao and Alexiou 2022).[1] These arbitrage opportunities are considered anomalies that challenge the core assumption of the efficient market hypothesis (Ardeni 1989; Malkiel and Fama 1970) and are generally difficult to identify (Shleifer and Vishny 1997).

There exist prior works on arbitrage opportunities identification. The majority of these work are in the area of financial markets where researchers examine how investors take advantage of certain external shocks to exploit short-term price dispersions (Avellaneda and Lee 2010; Birău 2015; Huy et al. 2020). Previous works have also examined arbitrage activities in commodity markets, including commodity futures (Lee et al. 1985), live cattle and wheat (Kohzadi et al. 1996), real estate markets (Limsombunchai 2004), and palladium and crude oil (Yao and Alexiou 2022). However, little research has been done in the context of retail markets and hence our understanding on arbitrage opportunities in retail markets is particularly limited. This knowledge gap is important because retail markets play an important role in our society. On the one hand, participants in large online marketplaces, such as eBay and Amazon, may behave similarly to participants in stock and commodity markets because there are millions of products available in marketplaces where multiple sellers and buyers transact with real-time price changes (Jopson 2012). On the other hand, these participants in online marketplaces are inherently different from those in other markets in multiple ways. First, there are price differences across marketplaces for the same product (Cavallo 2017). In addition, the market itself has a low barrier of entry rather than being exclusive to limited groups, as is the case with the financial and energy markets. It also usually provides advanced information technologies facilitating buyers' or sellers' transactions. Furthermore, product sales in retail markets are also associated with several product-related factors (Ebina and Kinjo 2019; Ehrenthal et al. 2014; Elmaghraby and Keskinocak 2003) and user-generated content (Hu et al. 2008; Khern-am-nuai et al. 2023), which potentially allow retail arbitrageurs to leverage such information to detect arbitrage opportunities. In this study, we aim to fulfill this research gap and investigate the following research question:

RQ: Can we develop an AI model that can consistently identify arbitrage opportunities in a retail market with tangible economic values?

In this work, we utilize data collected from Amazon Marketplace between February 8, 2015 and June 30, 2016. Using this dataset, we develop an AI-based predictive model to identify arbitrage opportunities by leveraging the combination of traditional price-related features and user-generated content (UGC) including online reviews and questions and answers. We demonstrate the performance of the model using a common cross-validation technique based on the out-of-sample dataset. In addition, we conduct a field experiment to empirically verify the external validity and to demonstrate the economic value of our model. The remainder of the paper is organized as follows. In Section 2, we survey past literature that is closely related to our study. Specifically, we review prior works that study arbitrage opportunities, machine learning algorithms and predictive tasks, and social impact of AI in e-commerce markets. Following that, in Section 3, we describe our research methodology in detail. Then, in section 4, we present our empirical results. Finally, in Section 5, we conclude the paper and discuss its theoretical and practical implications, limitations, and future research opportunities.

2. Literature Review

In this section, we review the previous literature that pertains to our study. More specifically, we survey prior work on arbitrage, particularly the factors of its emergence. We then review past management studies that adopt machine

[1]The definition of arbitrage can vary based on the context and discipline. For example, in finance, arbitrage is usually defined as the act of simultaneously buying and selling an asset to take advantage of differing prices. In this paper, we adopt a looser definition of arbitrage that is commonly employed in e-commerce literature. Specifically, the term arbitrage in our study describes a scenario in which a product is purchased and subsequently resold to exploit a price discrepancy with a minimal level of risk (Subramanian and Overby 2017).

learning, a subfield of artificial intelligence, as a core research methodology. Finally, we review prior works on the social impact of AI usage in the context of electronic commerce.

2.1. Arbitrage Opportunities

Past studies on arbitrage have been conducted across various areas. For example, in economics, Anson et al. demonstrate that currency shock tends to lead to customer arbitrage opportunities in cross-border e-commerce (Anson et al. 2019). Gębarowski et al. focus on triangular arbitrage in the Forex market and conclude that such opportunities are caused by abrupt changes among currencies and significant events, such as Brexit (Gębarowski et al. 2019). Fisch and Schmeisser explore global arbitrage from an operational perspective. They suggest that multinational corporations can better leverage arbitrage opportunities by improving management practices in host countries. For instance, improved efficiency in sale operations can lead to tax arbitrage (Fisch and Schmeisser 2019). Avdjiev and Aysun study the regulatory arbitrage of internationally active banks. They find that global financial situations cause the banks' arbitrage activity. In particular, the banks rapidly expand their claims to less-regulated countries when the global financial risk is higher because they face higher compliance costs (Avdjiev et al. 2022).

Studies related to arbitrage opportunities are also common in management literature. For example, Goncalves-Pinto at el. (Goncalves-Pinto et al. 2020) analyze trading data from three US stock exchanges and find that the option prices add pressure to fundamental stock prices, which leads to the predictability of the trading price of stocks and triggers trading arbitrage opportunities. However, the opportunities are limited when the gap between the latest option-implied stock price and the actual stock trade price is wider than previously recorded. Relatedly, Kozhan and Tham study high-frequency arbitrage in the spot foreign exchange market. They find that while arbitrage opportunities are spurred by high competition among arbitrageurs for scarce assets and uncertainty regarding the arbitrage portfolios' profitability, these opportunities can be eliminated by market orders under high market liquidity, short arbitrage duration, and significant arbitrage deviation (Kozhan and Tham 2012).

Earlier e-commerce studies have also focused on arbitrage in non-financial markets. For example, Roach's study reveals that economic differences between countries (e.g., cost of living and lowest wage) lead to global labor arbitrage (Roach 2004). Overby and Clarke investigate the effect of e-commerce on spatial arbitrage in the context of the US wholesale used vehicle market and conclude that using a webcast channel can greatly limit arbitrage opportunities (Overby and Clarke 2012). Subramanian and Overby extend Overby and Clarke's work by conducting a quasi-natural experiment to examine arbitrage opportunities where standalone electronic markets are included in the study. They find that while e-commerce channels reduce spatial arbitrage, standalone electronic markets facilitate the capture of remaining opportunities by arbitrageurs (Subramanian and Overby 2017). Zhang and Feng empirically demonstrate that cross-channel effects trigger arbitrage opportunities in the gray market, as manufacturers sell an identical product through multiple markets at different price points (Zhang and Feng 2017). Fedoseeva and Irek investigate the price dispersion within one chain online food retailer during the COVID-19 pandemic. They find that the differences in economic indicators, competitive pressure, and the number of new COVID-19 cases lead to geological arbitrages, but the magnitude differs by product type (Fedoseeva and Irek 2022). The current research contributes to this stream of research by providing a predictive analytics framework to identify arbitrage opportunities in retail markets. We leverage a machine-learning algorithm that is commonly used in the literature for predictive tasks. Next, we survey prior works that use a similar approach.

2.2. Machine-Learning Algorithms and Predictive Tasks

Machine-learning is a subgroup of artificial intelligence (AI) whose algorithms enhance AI applications. As big data has grown significantly, recent arbitrage research has leveraged machine-learning algorithms as a core research methodology. This practice is particularly seen in the business and management literature. For instance, Krauss et al. use deep neural networks, gradient-boosted trees, and random forests (RFs) to predict statistical arbitrage opportunities in the S&P 500. They find that RFs produce the most accurate results (Krauss et al. 2017). Huck predicts statistical arbitrage in the US market using three classification models, deep belief networks, RFs, and elastic net regression based on high-frequency trading data. The results show that RFs yield the best performance (i.e., highest portfolio return) (Huck 2019). Relatedly, Fischer et al. develop an RF model to predict statistical arbitrage opportunities in cryptocurrency markets in the timespan of two hours and demonstrate that the model performs reasonably well in capturing these arbitrage opportunities (Fischer et al. 2019). Zhang M et al. focus on statistical arbitrage in China stock market. They employ five classification models, RF, deep neural net, extreme gradient-boosted trees, support vector machine, and long short-term memory, to predict the stock price. The result shows that RF outperforms the other models (Zhang et al. 2021). Khan develops an RF model to predict real-time electricity locational marginal price and approximate subsequent profit for the corresponding hour of the next day. The model can successfully capture the

arbitrage opportunities and incur a net profit for the market participators (Khan 2022).

In the management literature, the growing importance of incorporating predictive analytics into research has been highlighted (Shmueli and Koppius 2011). The authors provide six roles that predictive analytics can play in business management research and argue that predictive analytics is useful not only for developing practical, useful models but also for assisting with theory building and testing. Indeed, over the last few years, numerous works have applied machine-learning algorithms for predictive tasks in different contexts. For example, Zheng et al. use fast causal inference and really fast causal inference models to assess the usage spillover effect of one instant messaging application in China on the others (Zheng et al. 2019). Nam and Seong propose a transfer entropy model to predict stock price direction in the Korean stock exchange using financial and economic news articles (Nam and Seong 2019). Liu et al. evaluate products' competitive advantage in the automotive market based on UGC shared on social media. They use bagging logistic regression as a competing product classification model (Liu et al. 2019).

The current work connects to this stream of literature as we leverage a machine-learning algorithm—specifically the RF model—to identify arbitrage opportunities in retail markets. Later in the methodology section, we discuss the data we used in this study and the proposed predictive analytics framework in more details.

2.3. Social Impact of Artificial Intelligence in E-Commerce Markets

Online retail is one of the markets where artificial intelligence (AI) tools are actively and seeminglessly integrated in their operations and buyers and sellers regularly engage in the AI-based technologies. As such, the use of AI in these market yields various significant social implications. While ethical issues have raised significant concerns when technologies are used (Cohen et al. 2023; White 2013), data privacy has also emerged as a major concern due to the vast amount of personal data required for AI development (Ouchchy et al. 2020). In that regard, AI creditability, transparency, and trustworthiness are also considered holistically. As a result, several works have called for AI literacy to be a common knowledge for end-users in such markets (Kim and Lee 2019; Shin 2022; Shin et al. 2022). However, the benefits of using AI tools tend to exceed the potential risks, and, as a result, end-users continue to utilize AI tools even when they are well aware of ethical and privacy concerns. Ultimately, the benefits of using AI applications are not limited to the end-users but also intertwined with firms or implimenters who provide the systems. A recommendation system is an example of a frequently used AI in e-commerce. While it improves the efficiency of customers' information search, sellers who provide the system also increase their profits (Cao 2021; Hinz and Eckert 2010; Kannan et al. 2022). Another example is Chatbot, which can timely provide information as per customers' inquiries while decreasing sellers' operational costs (Luo et al. 2019b; Oosthuizen et al. 2021). Several firms also have utilized AI in their operations, which not only benefits the firm itself but can also improve consumer surplus and social welfare as well (e.g., Adulyasak et al. 2023). This study contributes to this research stream, demonstrating the benefit of an AI model for optimal purchasing. Customers would pay for a product at a reasonable cost, representing an AI application that potentially improves customers' surplus and welfare.

3. Methodology

In this section, we discuss the data we collected and describe the predictive analytics framework employed in this study.

3.1. Data Source

We collected the primary data for this study from Amazon marketplace. We chose Amazon as our data source because it is the largest online retailer in North America with millions of buyers and sellers transacting in real time. Furthermore, the product prices on Amazon change frequently and are set by multiple sellers. Therefore, the current selling price for a given product on Amazon can generally represent the notion of market price in the context of retail markets. Moreover, Amazon marketplace also incorporates AI tools (e.g., product recommendation system and chatbots) to elevate their customers' shopping experiences, which reflect the power of advanced information technologies on market efficiency. Regarding the context of the study, we selected the market of video game software for the following reasons. First, the average lifecycle of products in this market is generally long, unlike technology products such as laptops or mobile phones, which typically become obsolete in less than one year. This characteristic allows us to have data about products in different life-cycle stages. Therefore, the insights developed are not limited to only newly released products. Second, as we aim to incorporate UGC as a candidate set of predictors, it is well known that video game console software actively garners a significant amount of related UGC due to its rich characteristics (Khern-am-nuai et al. 2023). Third, many prior studies have used the video game console market as the market of interest because it generally represents the market for experience goods (Kim et al. 2021; Nair 2007;

Zhu and Zhang 2010).

To collect the data, we first obtained a list of video game console software available on Amazon.com that was released within the last two years of the study period. The dataset consists of the stock-keeping unit identification, the product names, the gaming platforms of the product, the publisher that officially published the products, and the products' official market release date. In addition, for each product we collected its pricing data and sales rank on a daily basis using an automated script. The data collection task was performed between February 8, 2015 and June 30, 2016. The data consist of each item's lowest price (including shipping cost), the number of sellers who offer that product, and the sales rank. In addition, we also collected UGC, including online reviews and questions and answers related to each product. Our final dataset has 2,626 products.

3.2. Predictive Task Formulation and Target variable

The primary objective of this study is to identify arbitrage opportunities. Conceptually, a given price point would represent a good buying opportunity (i.e., an arbitrage opportunity) if, later on, the price of that product sufficiently increases within a reasonable timeframe. It is important to note that we need both conditions to hold true when defining an arbitrage opportunity because there are costs associated with every transaction and with the length of the product holding period. Without loss of generality, we consider a sufficient increase when the original price rises at least 20% plus two dollars. For example, if an original price is 100 dollars, the required minimum margin is 122 dollars. We choose 20% increase plus two dollars as the minimum margin because online marketplaces, such as eBay and Amazon, tend to charge end users who sell products on their platform about 15%–18% of the selling price as commission, while a two-dollar charge tends to cover fixed costs such as closing and listing fees (Godmanis 2019). We choose 30 days as the reasonable timeframe because such a length generally represents the risk-free period in retail markets in the US. For instance, most stores in the US allow buyers to return products for a refund within 30 days (with the exception of final clearance items with no returns). Therefore, if necessary, the arbitrageur can return the product and obtain a refund if she cannot sell it within 30 days; thus, she only suffers from an opportunity cost. In addition, many premium credit cards in the US also offer "purchase protection" and "return production" for 30 days after the purchase date, which provide an additional layer of protection for arbitrageurs during this period. In the field experiment discussed later, we neither return any products nor utilize any credit card protection.

To operationalize our definition of arbitrage opportunities in the predictive framework, we follow the strategy commonly used in the literature to construct a single target variable that captures both the variation of price and time period such that the final target variable is binary (which is used to indicate a purchase opportunity). As such, the predictive task at hand can be formulated as a simple classification problem (Torgo 2016). Our target variable is a binary variable that takes the value one (or positive class) if a price point has an arbitrage opportunity and zero (or negative class) otherwise. In order to classify an arbitrage opportunity, the opportunity should be sufficiently large to trigger the recommendation. We count the number of days a focal price point meets our condition of arbitrage opportunity (i.e., the price increase of more than 20% plus two dollars within 30 days), and the higher number indicates a better arbitrage opportunity. The sufficiency of the opportunity depends on a threshold, and it can be set based on a pre-defined criterion or as a hyperparameter optimized through cross-validation. In this paper, we set the threshold equal to or greater than 4 days. The detailed calculation for our target variable is described in Appendix A.

3.3. Predictors

Our study includes two sets of predictors, both of which are calculated on a daily basis. The first set of predictors consists of product- and price- related information including: (1) sales rank, (2) market price of a brand-new version, (3) market price of a used version, (4) trade-in value, (5) total number of units available for sales for brand new version, (6) total number of units available for sale for a used version. These predictors are equivalent to the stock-related variables, such as market price, market volume, and amount of bids and offers, that are commonly used in the literature in finance and other related areas that perform arbitrage opportunities predictions (Cao et al. 2011; Iqbal et al. 2013).

The second set of predictors consists of UGC available on Amazon. This set of predictors includes online product reviews and questions and answers related to the product from the Amazon Answer System (Khern-am-nuai et al. 2023). From here, we derive the predictors following the information systems and marketing literature that studies UGC. In particular, prior works in this area have empirically demonstrated the economic value of UGC volume (Chevalier and Mayzlin 2006; Zhu and Zhang 2010), content length (Khern-am-nuai et al. 2023), and textual content characteristics (Ghose and Ipeirotis 2010). Following these prior works, the predictors in this set include (7) total number of reviews, (8) total number of questions, (9) total number of answers, (10) average star rating, (11) average review length, (12) average question length, (13) average answer length, (14) average review sentiment, (15) average question sentiment, (16) average answer sentiment, (17) average review subjectivity, (18) average question

subjectivity, and (19) average answer subjectivity. The sentiment and subjectivity of UGC are calculated using the pre-defined score from the TextBlob package in Python (Loria 2018), which is a commonly used tool to analyze textual content. Note that both sets of predictors of each observation are calculated up to day i of the target variable to eliminate the potential issue of look-ahead bias.

3.4. Data Pre-processing

In this subsection, we describe the data pre-processing procedures and the development process of our predictive models. First, we pre-process the data by examining data points with a missing value. For each data point that contains missing values, we employ a common approach in predictive analytics literature (Larose 2015) to fill the missing value of a predictor with the latest value of the same product (e.g., if a trade-in value is missing on day *i*, we fill it with the latest trade-in value of the same product). After this process, if it still has any missing values (e.g., records with a predictor that has a missing value for the entire period), we remove the entire record. Our final dataset consists of 87,949 records. Following that, we randomly separate 70% of the data into a training dataset, which is used to develop our predictive model. Meanwhile, the rest of the data (30%) is used as a test dataset to validate model performance. Such a separation allows us to avoid potential overfitting issues when training/evaluating the model.

3.5. Model Development and Predictive Framework

We develop the AI predictive models using Python 3, a popular programming language in predictive analytics. In particular, we use the *sklearn* package (Pedregosa et al. 2011), one of the most popular data analytics and machine-learning packages in Python, as the primary tool. We use an RF machine-learning algorithm (Breiman 2001) as the predictive algorithm because it is commonly used in the literature for predictive tasks, given its overall efficiency and effectiveness (Luo et al. 2019a; Müller et al. 2016). Note that we let the *RandomForest* algorithm balance the class weights when it performs classification tasks in this study because the target variable in our dataset is highly imbalanced (i.e., the buying opportunities where the target variable takes the value one are relatively rare in the dataset). In addition, we tune the hyperparameter *n_estimators*, which determines the number of trees built in the classification process, using the *RandomSearchCV* method. We choose the value that yields the highest average F1 score from 10-fold cross-validations during the hyperparameter tuning process. All other hyperparameters are at the default value.

Our evaluations consist of two iterations. In the first iteration, we focus on the scope of the prediction tasks and develop three models based on different scopes:

(1) The generic model where all products are included within a single model
(2) The per-category model where we separately develop a prediction model per product category. Recall that Amazon categorizes all video game software based on its associated gaming console. In our dataset, there are eight product categories in total: Nintendo 3DS, Nintendo Wii, Nintendo Wii U, PlayStation 3, PlayStation 2, PlayStation Vita, Xbox 360, and Xbox One.
(3) The per-product model where we separately develop a prediction model per product. In our context, a product is a game that has a unique Amazon Standard Identification Number (ASIN). In other words, we develop one prediction model for each game.

The second iteration focuses on the selection of predictors. In this regard, we develop three models with different sets of predictors as follows:

(1) The first model only includes product information. This model represents predictive models that are commonly used in the finance literature to identify arbitrage opportunities or to predict price changes.
(2) The second model includes both product information and UGC. The results from this model demonstrate the predictive power of UGC in the context of arbitrage opportunity identifications in retail markets.
(3) The third model applies a common feature selection criterion used in the literature (Torgo 2016). Specifically, only predictors that increase the classification accuracy by at least 5% are included in the model. The performance of this model allows us to observe whether a feature selection process can improve the prediction performance in our study context.

3.6. Model Performance Evaluation

In this section, we report on the performance of our AI model, which is calculated based on the 30% out-of-sample dataset that we separated before the model training process. We evaluate the performance of the model using the common measures used in the predictive analytic literature. In particular, for each predictive model, we calculate the classification accuracy score, the precision score, the recall score, and the F1 score.

Recall that our target variable is imbalanced (i.e., there are more observations where the target variable is 0 than where the target variable is 1). Since the accuracy score considers a model's predictive performance for all classes at once, the minority class could be ignored and not be well represented by the score. As such, the accuracy score may not accurately indicate our models' overall performance. For this reason, although the accuracy score shows a model's overall correctness, we report the other three performance measures along with it. The precision score indicates the degree to which our model's predicted opportunities become real, the recall score determines our model's ability to capture the actual opportunities, and the F1 score combines the precision and the recall scores into a single measure. Appendix B describes the calculation for each of them.

3.7. Field Experiment Setup

We further strengthen our model evaluation by performing a field experiment where we deploy our AI model in a field setting and observe its economic performance using the monetary value. It is important to note that the primary objective of our experiment is to empirically confirm the external validity of our model rather than to optimize our model in a real-life setting. Therefore, we implement our model in the field in an especially conservative manner. First, we do not perform model retraining even when more training data are available during the experiment period.[2] Second, we do not re-optimize the hyperparameters of our AI model with RF algorithm during the experiment period. As a result, the performance of our model during the experiment period is naturally expected to degrade compared to its performance from the out-of-sample cross-validation.

We conducted our field experiment between July 1, 2016 and June 30, 2017 using the following procedures. We trained our model using the specification that yields the best F1 score from the model evaluation (i.e., the generic model with all predictors). The training data for the field experiment was a combination of the training and test dataset we previously used to develop the model in the earlier section (i.e., the entire dataset between February 8, 2015 and June 30, 2016).

After we obtained the trained model, we performed the following actions at 8 a.m. Eastern Time on a daily basis. First, we collected the data from a major online marketplace in the US with the same product category (i.e., video game software). Second, we executed the trained model with the data collected on that day, along with past data if applicable. We purchased each product that the model identified as an arbitrage opportunity from the platform. Third, for each product in our inventory, we observed the lowest price listed on the platform. If the lowest price was 20% plus two dollars higher than the purchase price, we listed that product on the platform with the same lowest listed price available. Note that we used a completely new seller profile in the marketplace to conduct the field experiment. The profile was created on July 1, 2016 (i.e., the first day of the experiment). We continued to use the same profile throughout the experiment period.

In addition, to account for the impact of false positive prediction, instead of returning the product to the seller if the desired profit margin (i.e., 20 percent plus 2 dollars) could not be obtained during the risk-free period (i.e., 30 days), all products older than 30 days were listed with the lowest listed price available on the platform regardless of their purchase price. As a result, these products were usually sold at a loss during the experiment. For each product sold on the platform, we paid the commission fee to the platform and shipped it using a standard shipping method offered by the platform. Note that we did not purchase or keep more than one copy of the same product at a given time (i.e., if the product was in the inventory, we did not purchase it until the existing copy was sold, even if the model indicated an additional arbitrage opportunity).

4. Findings

In this section, we present the results of our developed AI models. First, model performances based on test sets with three prediction scopes and three sets of predictors are shown. Then the most promising model is used to conduct the file experiment, thus evaluating its economic value.

4.1. Models with Different Prediction Scopes

We next report the performance of the three AI models with RF algorithm developed with different prediction scopes

[2] In practice, it is very common for predictive models in retail markets to perform model retraining when the data become more available. For example, a model that is used to predict the target variable on January 15 would be trained on the data up to January 14. Meanwhile, when the model is used the next day (i.e., January 16), it would be retrained on the data up to January 15 before performing the prediction task. The retraining can be daily, weekly, or monthly depending on the context and computational resources required. Nevertheless, in this study, we do not perform any model retaining throughout the experiment period to ensure that our validation is as conservative as possible.

while all predictors were used[3]. Recall that the models were developed by tuning one hyperparameter, *n_estimators*; because the algorithm is an ensemble method, this hyperparameter tended to have a major effect on model performance. We also performed additional analyses where we tuned additional hyperparameters of the RF. The corresponding results, which are qualitatively similar to our main results, are reported in Appendix D[4].

The performance measures of the three models with different prediction scopes are summarized in Table 1. Note that boldface letters indicate the best result obtained under each measure when comparing the different models. The scores of the per-category model and those of the per-product model are the average of the scores of the model of each category and the average of the scores of the model of each product, respectively.

Table 1. performance Measures with Different Prediction Scopes

	Accuracy	Precision	Recall	F1
Generic	**0.9813**	0.9536	**0.8230**	**0.8835**
Per Category	0.9809	**0.9538**	0.8173	0.8803
Per Product	0.9726	0.9306	0.8171	0.8701

Interestingly, we find that the generic model significantly outperformed the per-category and per-product models for most performance measures, including accuracy, recall, and F1 score. Meanwhile, the per-category model yielded the best precision score, but the difference between the precision score of the per-category model and that of the generic model is marginal. Considering that the generic model also benefits from its superior generalizability, we used the generic model where all products are pooled together to develop the predictive model as the main one in this study. Next, we explore the performance of the model with respect to different sets of predictors.

4.2. Models with Different Sets of Predictors

We observed that our AI model performed exceptionally well when all products were included together in the model development (i.e., the generic model). Next, we examine the relative performance of the models with different sets of predictors. Table 2 below reports the performance measurement of the generic model that utilizes only the product-related and price-related information features, both product-related information and user-generated information, and the two information but only the predictors that increase the classification accuracy by at least 5%. Additional details regarding our feature selection method and corresponding feature importance scores are reported in Appendix F.

Table 2. Performance measures with different feature sets

	Accuracy	Precision	Recall	F1
Product Info Only	0.9599	0.9543	0.5605	0.7062
Product Info & UGC	**0.9813**	0.9536	**0.8230**	**0.8835**
Significant Features Only	0.9797	**0.9564**	0.8010	0.8718

Interestingly, the model in which only the first set of predictors (i.e., product information) was included performed reasonably well. The precision score is 0.9543, indicating that the positive predictions (i.e., arbitrage opportunities identified) are very accurate. However, the recall of 0.5605 indicates that only roughly half of the existing arbitrage opportunities in the dataset were captured by the model. Meanwhile, the results from the model with both set of predictors demonstrate that adding UGC can significantly enhance the performance of the prediction model. Specifically, the recall score (0.8230) and the F-1 score (0.8835) improved significantly, while the precision score decreased slightly (0.9536), although it was already high in the base model with only product-related information. This finding confirms that UGC indeed possesses significant informative power in identifying arbitrage opportunities in retail markets. Finally, the results from the last model where a common feature selection technique was applied show that a feature selection procedure does not appear to improve the classification performance in the context of this study. Although this model yields the best precision score, the difference between the precision score of the model with feature selection and that of other models is only marginal. Meanwhile, other performance measures were weaker

[3] This research utilizes the random forest algorithm as the primary model because it is commonly used in both research and practice and has a reasonably good predictive performance. We also develop alternative models that utilize other machine-learning algorithms to ensure that the results are not solely driven by random forest. The performance of these alternative models is reported in Appendix C.

[4] The threshold is set at 4 in the paper because sensitivity analyses show that this value is the most conservative option. We report results with different threshold values of in Appendix E.

than the model with the full set of predictors. This finding suggests that feature selection processes may have a limited impact in our AI model.[5]

Although our out-of-sample cross-validations robustly demonstrate that the AI predictive model we developed could consistently identify the arbitrage opportunities that exist in the out-of-sample dataset, there could be several concerns regarding the validity and practicability of our results.

(1) First, because we constructed our dataset to be cross-sectional, all our predictors and the target variable do not have a time component. As such, our test dataset was separated based on the out-of-sample principle rather than the out-of-time principle (i.e., the training and test datasets are from the same time period), which differs from the predictive model test that used time series-based data. However, when the model is used in reality, it would only be used for future data. Therefore, even though our validation is scientifically valid, it might be practically irrelevant depending on how the model is used. This is especially important given that the product category of our choice is video game software where the nature of older products may not necessarily reflect that of newer ones.

(2) Second, we measured the performance of our model using accuracy, precision, recall, and the F1 score, which are common performance metrics for prediction models. However, although these scores represent the performance of the model in terms of its prediction accuracy, they do not capture the economic impacts that the model generates (e.g., the model might accurately capture arbitrage opportunities, but those arbitrage opportunities may only generate negligible economic values).

(3) Third, one may argue that the price movement in the retail market alone may not necessarily constitute arbitrage opportunities. For example, Amazon or eBay buyers may be especially conscious of a seller's credibility. As such, even when the model can consistently identify good purchasing points, it does not necessarily mean that these products can be consistently sold to obtain arbitrage profits.

For these reasons, we went the last research mile (Nunamaker Jr et al. 2015) to validate the economic value and practicability of our model by conducting a field experiment in a real-life setting.

4.3. Field Experiment

Table 3 lists the results of our field experiment in terms of the model's performance. The total number of arbitrage opportunities identified during our one-year study period is 293, which is significantly lower than the number of arbitrage opportunities identified in our modeling phase, as expected. This decrease supports our hypothesis that the nature of the video game software market significantly changes over time, resulting in a significant decrease in the recall rate of the model (i.e., the ability for the model to identify positive predictions). Nevertheless, the model still performed exceptionally well in terms of precision. Out of 293 arbitrage opportunities identified, 268 (91.47%) were true arbitrage opportunities, while only 25 (8.53%) were false positives, which is consistent with the precision rate from the modeling phase. Next, we report the economic value of our arbitrage identification model.

Table 3. Field experiment results (model performance)

	Results
Total number of arbitrage opportunities	293
Total number of products sold at a profit	268
Total profit from true arbitrage opportunities	$2,126.40
Total number of products sold at a loss	25
Total loss from false arbitrage opportunities	$58.32

Table 4 lists the results of our field experiment in terms of the predictive model's economic value. As noted before, the model identified 293 arbitrage opportunities within the one-year period of our experiment. The total amount of purchasing cost that covers all arbitrage opportunities identified was $1,825.13, while the total revenue after selling those products (both the case of true arbitrage opportunities where a profit is realized and the case of false arbitrage opportunities where the product has to be potentially sold at a loss after 30 days) is $5,200.72. The total profit during the experiment was $2,068.08 or about $7.06 per arbitrage opportunity. A student t-test with unequal variance

[5] The feature selection procedure here uses feature importance scores that are calculated during the evaluation. We also calculate the feature importance score based on the best model and revisit the performance of the best model if only predictors selected by the feature selection procedure are used. Details of this exercise are available in Appendix G.

confirmed that the arbitrage opportunities identified by our AI model had an economic value higher than zero with a p-value < 0.001. Our field experiment reaffirmed the practicability and the economic value of the AI predictive model we developed to identify arbitrage opportunities in retail markets.

Table 4. Field experiment results (economic value)

	Results
Total number of arbitrage opportunities identified	293
Total costs	$1,825.13
Total revenue	$5,200.72
Total profit (after commission and shipping costs)	$2,068.08
Average profit per arbitrage opportunity	**$7.06**

5. Discussion

In this section, we conclude our study, discuss the implications for research and practice, and describe limitations and potential future research directions.

5.1. Conclusion

Recent studies have examined the arbitrage phenomenon as an anomaly in overall market efficiency that is affected by the use of emerging information technologies. However, most existing studies focus on arbitrage in financial, commodity, and real-estate markets. Meanwhile, insights on arbitrage opportunities in the retail context, especially at the places where AI tools are available, that are relevant to both researchers and practitioners have been limited in prior works.

Motivated by this gap in the literature, this paper investigated arbitrage opportunities in retail markets. In particular, we studied the possibility of systematically identifying arbitrage opportunities in retail markets using an AI model. To operationalize our research agenda, we collected a dataset for our analysis from Amazon marketplace, one of the largest online marketplaces in the US with advanced AI applications. This particular market has millions of available products and thousands of buyers and sellers who transact with each other in real-time. After collecting the data, we utilized one of the most popular machine-learning algorithms to develop an AI model to identify arbitrage opportunities in retail markets. Specifically, we developed several AI predictive models, each with a different modeling scope—a generic model where all products were included, a per-category model where one model per product category was developed, and a per-product model where one model was developed per product. We found that while the performance of all three models did not drastically differ, the generic model not only performed best in terms of most performance measures but also benefited from better model generalizability. In addition, we developed two models that used the generic modeling scope, one that used only product- and price-related information and another that used both product- and price-related information and UGC information but that also leveraged a common feature selection framework to reduce the model complexity. We found that UGC, such as product reviews and questions and answers, possesses significant predictive power in terms of identifying arbitrage opportunities. The model that includes these variables, along with the product- and price-related variables commonly used in financial prediction models, yielded a significantly better predictive performance. In contrast, we found that the common feature selection framework did not improve the overall performance of our predictive model. Overall, the AI model with the best performance could identify arbitrage opportunities with 95.36% precision and could capture about 82.30% of arbitrage opportunities that existed in the market.

We further validated the performance of our model on its economic values by conducting a field experiment that applied our model to a real-life setting in a large online marketplace in the US over a one-year period. We found that, as expected from our conservative approach in the model implementation, the model can identify significantly fewer arbitrage opportunities during the field experiment compared with the model performance in the out-of-sample test. Nevertheless, the model still performed well in terms of precision and generated $7.06 in profit on average for each identified arbitrage opportunity.

5.2. Implications for Research and Practice

Insights generated from this study contribute to both research and practice. From a research perspective, while prior studies focus on arbitrages in financial and commodity markets (Anson et al. 2019; Birău 2015; Goncalves-Pinto

et al. 2020; Huy et al. 2020), this study empirically demonstrate the existence of arbitrage opportunities in retail markets, especially in the contexts where AI tools are available to facilitate customers' decision-making. In addition, our work demonstrates that an AI model based on a machine-learning algorithm can consistently and efficiently identify these opportunities. Furthermore, we show the predictive power of UGC, which is typically available on online platforms, in enhancing the identification of arbitrage opportunities. These results imply that advanced technological tools, like AI systems, play an important role in moderating the customer's welfare with respect to arbitrage opportunities. Meanwhile, our findings inform practitioners across multiple levels. For end users, our model, when enhanced with contextual data, could be considered an alternative to the commonly used AI predictive models for arbitrage identification in the finance literature. For firm managers, this framework could be integrated as a part of the procurement decision process for common parts, materials, and services. Specifically, our arbitrage identification model could be incorporated into a procurement system that is traditionally used to facilitate the procurement process and manage inventory through pre-defined inventory policies. Our AI model could also introduce the optimal purchasing time into the overall objective function of the tool, which considers the arbitrage opportunities and jointly optimizes the savings from a price reduction (over the speculated future market price) in conjunction with ordering and inventory holding costs. Lastly, for managers of large retailers, our predictive model empirically demonstrates the existence of arbitrage opportunities and the overall framework to identify them, which could be integrated into the price optimization or anomaly detection engine that platforms utilize to optimize their benefits.

5.3. Limitations and Future Directions

Our study is not without limitations, which also represents an opportunity for future research. First, this study intentionally used a common machine-learning algorithm (i.e., RF) to demonstrate that arbitrage opportunities in retail markets can be identified using price movement and UGC data. Nevertheless, future research should explore more sophisticated machine-learning algorithms that improve the performance of arbitrage opportunity identification. Given the sheer amount of transactional data in retail markets, advanced deep-learning models may be suitable for this research avenue. Second, our current model is a black box (i.e., does not allow any interpretation of the processes we employ). Future research that focuses on the interpretability of an arbitrage opportunity identification process can investigate alternative machine-learning algorithms that allow for interpretability at the cost of model performance. Third, the objective of our field experiment was to enhance the validity and economic values of our AI model. Therefore, the experiment was designed without considering the scalability of the process. As such, even though the profit margin we attain may appear to be high, it is plausible that such a number is not sustainable when the scale of the arbitrage opportunity identification is significantly increased. Future research should examine the scalability of arbitrage opportunity predictions in retail markets by considering alternative specifications of the AI model and/or the experimental settings. Last, as past research has empirically demonstrated that the impact of UGC is moderated by product types (e.g., search goods vs. experience goods), future research could examine whether the proposed framework for arbitrage opportunity identification that we propose can be consistently used for search goods.

Data Availability Statement

The data that support our findings are available from the corresponding author, J.T., upon reasonable request.

References

Adulyasak Y, Benomar O, Chaouachi A, Cohen MC, and Khern-Am-Nuai W (2023) Using Ai to Detect Panic Buying and Improve Products Distribution Amid Pandemic. Ai & Society Forthcoming:1-30

Anson J, Boffa M, and Helble M (2019) Consumer Arbitrage in Cross-Border E-Commerce. Review of International Economics 27:1234-1251

Ardeni PG (1989) Does the Law of One Price Really Hold for Commodity Prices? American Journal of Agricultural Economics 71:661-669

Avdjiev S, Aysun U, and Tseng MC (2022) Regulatory Arbitrage Behavior of Internationally Active Banks and Global Financial Market Conditions. Economic Modelling 112:105857

Avellaneda M, and Lee J-H (2010) Statistical Arbitrage in the Us Equities Market. Quantitative Finance 10:761-782

Bapna R, Bichler M, Day B, and Ketter W (2018) Call for Papers—Special Issue of Information Systems Research—Market Design and Analytics. Information Systems Research 29:1067-1068

Birău FR (2015) Emerging Capital Market Efficiency: A Comparative Analysis of Weak-Form Efficiency in Romania and Hungary in the Context of the Global Financial Crisis. AI & SOCIETY 30:223-233

Breiman L (2001) Random Forests. Machine learning 45:5-32

Cao L (2021) Artificial Intelligence in Retail: Applications and Value Creation Logics. International Journal of Retail & Distribution Management 49:958-976
Cao Q, Parry ME, and Leggio KB (2011) The Three-Factor Model and Artificial Neural Networks: Predicting Stock Price Movement in China. Annals of Operations Research 185:25-44
Cavallo A (2017) Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers. American Economic Review 107:283-303
Chen T, and Guestrin C (2016). Xgboost: A Scalable Tree Boosting System. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794
Chevalier JA, and Mayzlin D (2006) The Effect of Word of Mouth on Sales: Online Book Reviews. Journal of marketing research 43:345-354
Cohen MC, Dahan S, Khern-Am-Nuai W, Shimao H, and Touboul J (2023) The Use of Ai in Legal Systems: Determining Independent Contractor Vs. Employee Status. Artificial intelligence and law Forthcoming:1-30
Ebina T, and Kinjo K (2019) Consumer Confusion from Price Competition and Excessive Product Attributes under the Curse of Dimensionality. AI & SOCIETY 34:615-624
Ehrenthal J, Honhon D, and Van Woensel T (2014) Demand Seasonality in Retail Inventory Management. European Journal of Operational Research 238:527-539
Elmaghraby W, and Keskinocak P (2003) Dynamic Pricing in the Presence of Inventory Considerations: Research Overview, Current Practices, and Future Directions. Management science 49:1287-1309
Fedoseeva S, and Irek J (2022) Within-Retailer Price Dispersion in E-Commerce: Prevalence, Magnitude, and Determinants. Q Open 2:1-20
Fisch JH, and Schmeisser B (2019) Upgrading Local Operations for Global Arbitrage. Long Range Planning 52:101845
Fischer TG, Krauss C, and Deinert A (2019) Statistical Arbitrage in Cryptocurrency Markets. Journal of Risk and Financial Management 12:31
Gębarowski R, Oświęcimka P, Wątorek M, and Drożdż S (2019) Detecting Correlations and Triangular Arbitrage Opportunities in the Forex by Means of Multifractal Detrended Cross-Correlations Analysis. Nonlinear Dynamics 98:2349-2364
Ghose A, and Ipeirotis PG (2010) Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. IEEE transactions on knowledge and data engineering 23:1498-1512
Ghose A, and Yao Y (2011) Using Transaction Prices to Re-Examine Price Dispersion in Electronic Markets. Information Systems Research 22:269-288
Godmanis I (2019) Economic Studies in Business Platforms: Completely New Approach in Microeconomics. Economics 6:59-66
Goncalves-Pinto L, Grundy BD, Hameed A, van der Heijden T, and Zhu Y (2020) Why Do Option Prices Predict Stock Returns? The Role of Price Pressure in the Stock Market. Management Science 66:3903-3926
Hinz O, and Eckert J (2010) The Impact of Search and Recommendation Systems on Sales in Electronic Commerce. Business & Information Systems Engineering 2:67-77
Hu N, Liu L, and Zhang JJ (2008) Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects. Information Technology and management 9:201-214
Huck N (2019) Large Data Sets and Machine Learning: Applications to Statistical Arbitrage. European Journal of Operational Research 278:330-342
Huy DTN, Loan BTT, and Pham TA (2020) Impact of Selected Factors on Stock Price: A Case Study of Vietcombank in Vietnam. Entrepreneurship and Sustainability Issues 7:2715
Iqbal Z, Ilyas R, Shahzad W, Mahmood Z, and Anjum J (2013) Efficient Machine Learning Techniques for Stock Market Prediction. International Journal of Engineering Research and Applications 3:855-867
Jopson B (2012) The Amazon Economy. Penguin Canada
Kannan K, Saha RL, and Khern-am-nuai W (2022) Identifying Perverse Incentives in Buyer Profiling on Online Trading Platforms. Information Systems Research 33:464-475
Khan I (2022). Efficiency Analysis and a Random Forest Based Trading Strategy for Heteroscedastic Electricity Market Data. 2022 IEEE Power & Energy Society General Meeting (PESGM): IEEE, pp. 1-5
Khern-am-nuai W, Ghasemkhani H, Qiao D, and Kannan K (2023) The Impact of Online Q&as on Product Sales: The Case of Amazon Answer. Information Systems Research Forthcoming:
Kim A, Saha RL, and Khern-am-nuai W (2021) Manufacturer's "1-up" from Used Games: Insights from the Secondhand Market for Video Games. Information Systems Research 32:1173-1191
Kim D, and Lee J (2019) Designing an Algorithm-Driven Text Generation System for Personalized and Interactive News Reading. International Journal of Human–Computer Interaction 35:109-122

Kohzadi N, Boyd MS, Kermanshahi B, and Kaastra I (1996) A Comparison of Artificial Neural Network and Time Series Models for Forecasting Commodity Prices. Neurocomputing 10:169-181
Kozhan R, and Tham WW (2012) Execution Risk in High-Frequency Arbitrage. Management Science 58:2131-2149
Krauss C, Do XA, and Huck N (2017) Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500. European Journal of Operational Research 259:689-702
Larose DT (2015) Data Mining and Predictive Analytics. John Wiley & Sons
Lee CF, Leuthold RM, and Cordier JE (1985) The Stock Market and the Commodity Futures Market: Diversification and Arbitrage Potential. Financial Analysts Journal 41:53-60
Limsombunchai V (2004). House Price Prediction: Hedonic Price Model Vs. Artificial Neural Network. New Zealand agricultural and resource economics society conference, pp. 25-26
Liu Y, Jiang C, and Zhao H (2019) Assessing Product Competitive Advantages from the Perspective of Customers by Mining User-Generated Content on Social Media. Decision Support Systems 123:113079
Loria S (2018) Textblob Documentation. Release 0.15 2:
Luo X, Lu X, and Li J (2019a) When and How to Leverage E-Commerce Cart Targeting: The Relative and Moderated Effects of Scarcity and Price Incentives with a Two-Stage Field Experiment and Causal Forest Optimization. Information Systems Research 30:1203-1227
Luo X, Tong S, Fang Z, and Qu Z (2019b) Frontiers: Machines Vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. Marketing Science 38:937-947
Malkiel BG, and Fama EF (1970) Efficient Capital Markets: A Review of Theory and Empirical Work. The journal of Finance 25:383-417
McMillan J (2003) Reinventing the Bazaar: A Natural History of Markets. WW Norton & Company
Müller O, Junglas I, Brocke Jv, and Debortoli S (2016) Utilizing Big Data Analytics for Information Systems Research: Challenges, Promises and Guidelines. European Journal of Information Systems 25:289-302
Nair H (2007) Intertemporal Price Discrimination with Forward-Looking Consumers: Application to the Us Market for Console Video-Games. Quantitative Marketing and Economics 5:239-292
Nam K, and Seong N (2019) Financial News-Based Stock Movement Prediction Using Causality Analysis of Influence in the Korean Stock Market. Decision Support Systems 117:100-112
Neely CJ, Rapach DE, Tu J, and Zhou G (2014) Forecasting the Equity Risk Premium: The Role of Technical Indicators. Management science 60:1772-1791
Nunamaker Jr JF, Briggs RO, Derrick DC, and Schwabe G (2015) The Last Research Mile: Achieving Both Rigor and Relevance in Information Systems Research. Journal of management information systems 32:10-47
Oosthuizen K, Botha E, Robertson J, and Montecchi M (2021) Artificial Intelligence in Retail: The Ai-Enabled Value Chain. Australasian Marketing Journal 29:264-273
Ouchchy L, Coin A, and Dubljević V (2020) Ai in the Headlines: The Portrayal of the Ethical Issues of Artificial Intelligence in the Media. AI & SOCIETY 35:927-936
Overby E, and Clarke J (2012) A Transaction-Level Analysis of Spatial Arbitrage: The Role of Habit, Attention, and Electronic Trading. Management science 58:394-412
Overby E, and Forman C (2015) The Effect of Electronic Commerce on Geographic Purchasing Patterns and Price Dispersion. Management Science 61:431-453
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, and Dubourg V (2011) Scikit-Learn: Machine Learning in Python. Journal of machine learning research 12:2825-2830
Roach S (2004). How Global Labour Arbitrage Will Shape the World Economy. from http://ecocritique.free.fr/roachglo.pdf
Serenko A, Ruhi U, and Cocosila M (2007) Unplanned Effects of Intelligent Agents on Internet Use: A Social Informatics Approach. AI & SOCIETY 21:141-166
Shin D (2022) How Do People Judge the Credibility of Algorithmic Sources? Ai & Society 1-16
Shin D, Lim JS, Ahmad N, and Ibahrine M (2022) Understanding User Sensemaking in Fairness and Transparency in Algorithms: Algorithmic Sensemaking in over-the-Top Platform. AI & SOCIETY 1-14
Shleifer A, and Vishny RW (1997) The Limits of Arbitrage. The Journal of finance 52:35-55
Shmueli G, and Koppius OR (2011) Predictive Analytics in Information Systems Research. MIS Quarterly 35:553-572
Subramanian H, and Overby E (2017) Electronic Commerce, Spatial Arbitrage, and Market Efficiency. Information Systems Research 28:97-116
Torgo L (2016) Data Mining with R: Learning with Case Studies. CRC press

Wang J, Cai S, Xie Q, and Chen L (2022) The Influence of Community Engagement on Seller Opportunistic Behaviors in E-Commerce Platform. Electronic Commerce Research 22:1377–1405
White JB (2013) Infosphere to Ethosphere: Moral Mediators in the Nonviolent Transformation of Self and World. Moral, Ethical, and Social Dilemmas in the Age of Technology: Theories and Practice 215-233
Yao W, and Alexiou C (2022) Exploring the Transmission Mechanism of Speculative and Inventory Arbitrage Activity to Commodity Price Volatility. Novel Evidence for the Us Economy. International Review of Financial Analysis 80:102027
Zhai Y, Hsu A, and Halgamuge SK (2007). Combining News and Technical Indicators in Daily Stock Price Trends Prediction. International symposium on neural networks: Springer, pp. 1087-1096
Zhang M, Tang X, Zhao S, Wang W, and Zhao Y (2021). Statistical Arbitrage with Momentum Using Machine Learning. International Conference on Identification, Information and Knowledge in the internet of
Things, pp. 194-202
Zhang Z, and Feng J (2017) Price of Identical Product with Gray Market Sales: An Analytical Model and Empirical Analysis. Information Systems Research 28:397-412
Zheng J, Qi Z, Dou Y, and Tan Y (2019) How Mega Is the Mega? Exploring the Spillover Effects of Wechat Using Graphical Model. Information Systems Research 30:1343-1362
Zhu F, and Zhang X (2010) Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. Journal of marketing 74:133-148

Appendix A

In this appendix, we show the calculation of our target variable, an arbitrage opportunity. First, for each product, we define an arbitrage opportunity when the future market price over the next k days is expected to be at the margin m above the current spot price P on day i. The margin is calculated as:

$$m_i = (\alpha P_i + \beta)/P_i, \tag{1}$$

where α is the predefined spread threshold (as a percentage) of the future market price above the spot price, and β is the fixed cost (in dollars) to cover fixed costs. In our study, we consider an opportunity to arbitrage when the selling price (i.e., market price) of a product is expected to be 20% (α) higher than the current buying price (i.e., spot price) plus two dollars (β) within 30 days (k).

Next, for our model binary prediction, we initially define a set of intermediate variables for each product on day i, denoted by V_i , as:

$$V_i = \begin{Bmatrix} 1 & if\ \frac{P_{i+j}-P_i}{P_i} \geq m_i \\ 0 & else \end{Bmatrix}_{j=1}^{k}, \tag{2}$$

where P_i is the product price on day i and k is the number of days following day i. On each of the following k days (i.e., 30 days) after day i, the corresponding element in the set V_i indicates if the price on day $i + j$ is at least equal to the defined margin m_i (i.e., 20% plus two dollars). Following that, we define an indicator T_i such that:

$$T_i = \sum_{v} \{v \in V_i\}. \tag{3}$$

Here, the indicator T_i represents the number of days during the arbitrage period (over the next k days) where the actual margin is at least equal to the required margin m. Note that the use of technical indicators to detect and predict price trends is commonly employed by practitioners to identify investment opportunities (Neely et al. 2014; Torgo 2016; Zhai et al. 2007). Once the variable T_i is derived, we transform it into a target binary variable by checking if $T_i \geq \gamma$, where γ is a threshold indicating sufficiently large arbitrage opportunities. Our study set the threshold γ at 4. As such, the final target variable in our study values one if T_i of the focal product on day i is higher than four and zero otherwise.

Appendix B

We evaluate each model's performance with four matrices, consisting of accuracy score, precision score, recall score, and F1 score. The classification accuracy score measures the overall accuracy of the classification task with the following specification:

$$Accuracy\ (Acc.) = \frac{TP+TN}{TP+TN+FP+FN}, \tag{4}$$

where TP is the number of true positive predictions, cases a model correctly predicts as arbitrage opportunities. TN is the number of true negative predictions, cases a model correctly predicts as non-arbitrage opportunities. FP is the number of false positive predictions, cases a model predicts as arbitrage opportunities but are not. FN is the number of false negative predictions, cases a model predicts as non-arbitrage opportunities but actually are.

The precision score shows how well the model can predict the positive class and is calculated with the following specification:

$$Precision\ (Prec.) = \frac{TP}{TP+FP}, \tag{5}$$

The recall score, which represents the opportunities captured by the positive prediction, is calculated with the following specification:

$$Recall\ (Rec.) = \frac{TP}{TP+FN}. \quad (6)$$

The F1 score combines the precision score and the recall score into a single measure. It is calculated based on the harmonic mean of the precision score and the recall score. Specifically,

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}. \quad (7)$$

Appendix C

In the paper, we use random forest (RF) as the primary algorithm because it is commonly use in both research and practice. In this appendix, we report the performance of alternative AI models that use other machine learning algorithms. Specifically, the alternative algorithms include (1) K-nearest neighbors (KNN), (2) artificial neural network (ANN), and (3) extreme gradient boosting (XGB). The development of these models is similar to that of the main model (i.e., random forest) presented in section 3.5.

We implement the KNN and NN using *sklearn* package (Pedregosa et al. 2011), and XGB using *DMLC-XGBoost* package (Chen and Guestrin 2016). Each model is developed under three prediction scopes (i.e., generic, per-category, and per-product), and all predictors are included. All models are set to balance class weights and their hyperparameters are tuned with a similar method presented in the paper (i.e., *RandomSearcCV*). Specifically, for KNN, we tuned the following hyperparameters: the number of neighbors, *n_neighbors*, distance function, *p*, and weight function, *weights*. For ANN, we tuned the number of neurons in the hidden layer, *hidden_layer_sizes*, strength of the regularization, *alpha*, learning rate schedule for weight updates, *learning_rate*, activation of the hidden layer, *activation*, and solver for weight optimization, *solver*. Lastly, for XGB models, we tuned their learning rate, *eta*, minimum loss reduction for leave node partition, *gamma*, tree growing policy, *grow_policy*, and *booster*. Since KNN and ANN models were sensitive to predictor ranges, all predictors were standardized. Table 5 presents the results of the additional models, and the boldface letters show the result from the main model used in this study. Since our data have an imbalanced class, we focus on the performance comparison based on F1 scores. Overall, the additional models' predictive performance is comparable with the main model. To evaluate whether the performance of these alternative models is better than that of the main model, we perform the two-sided permutation test with 10,000 iterations to get the p-value with up to four decimal places We find that none of the p-value of the differences is higher than 0.10, indicating that the performance of alternative models is statistically higher than that of the main model.

Table 5. Performance measures with different machine learning models

	Model	Accuracy	Precision	Recall	F1
Generic	**RF**	**0.9813**	**0.9536**	**0.8230**	**0.8835**
	KNN	0.9737	0.8659	0.8217	0.8432
	ANN	0.9655	0.7428	0.9168	0.8207
	XGB	0.9791	0.8589	0.9058	0.8817
Per-Category	**RF**	**0.9809**	**0.9538**	**0.8173**	**0.8803**
	KNN	0.9743	0.8703	0.8239	0.8465
	ANN	0.9738	0.8175	0.8952	0.8546
	XGB	0.9849	0.9168	0.9071	0.9119
Per-Product	**RF**	**0.9726**	**0.9306**	**0.8171**	**0.8701**
	KNN	0.9695	0.8797	0.8689	0.8742
	ANN	0.9730	0.8831	0.8975	0.8902
	XGB	0.9760	0.8712	0.9426	0.9055

Appendix D

In the paper, our best AI model is based on the optimized value of the hyperparameter *n_estimator*. We also tune other hyperparameters but do not observe significant differences in model performance after hyperparameter tuning. In this appendix, we report the comparison of model performance when additional three hyperparameters are tuned for illustration. Specifically, in addition to *n_estimator*, we tune the function for quality measurement (*criterion*),

minimum number of samples in the leaf node, (*min_samples_leaf*), and minimum number of samples for splitting (*min_samples_split*). Table 6 shows the results for the models with optimized hyperparameters, and the boldface letters present those of the main model. We observe that tuning 4 hyperparameters improve the performance by less than 1%, and the two-sided permutation test shows that the differences are not statistically significant at p<0.10.

Table 6. Performance measures with different prediction scopes and number of tuned hyperparameters

	No of hyperparam	Accuracy	Precision	Recall	F1
All Features	**One**	**0.9813**	**0.9536**	**0.8230**	**0.8835**
	Four	0.9818	0.9567	0.8265	0.8868
Prod. Info Only	**One**	**0.9809**	**0.9538**	**0.8173**	**0.8803**
	Four	0.9623	0.8886	0.6424	0.7457
Sig. Features Only	**One**	**0.9726**	**0.9306**	**0.8171**	**0.8701**
	Four	0.9802	0.9194	0.8441	0.8802

Appendix E

In the paper, we choose the threshold value for observations to be considered as an arbitrage opportunity (γ) to be 4. In this appendix, Table 7 and 8, we report the number of arbitrage opportunities and the results from various sensitivity analyses in which we vary the value of γ.

Table 7. Number of arbitrage opportunities with respect to γ

Scope	Data Set	N of Arbitrage Opportunities (% to Total N)						
		γ						
		2	3	4	5	6	7	8
Generic	Training	6,265	5,743	5,300	4,971	4,633	4,297	4,006
	Test	2,685	2,461	2,271	2,130	1,986	1,842	1,717
Total N = 87,949	Total	8,950 (10.18%)	8,204 (9.33%)	**7,571** (8.61%)	7,101 (8.07%)	6,619 (7.53%)	6,139 (6.98%)	5,723 (6.51%)
Per-cat.	Training	6,265	5,743	5,299	4,971	4,631	4,298	4,004
	Test	2,685	2,461	2,272	2,130	1,988	1,841	1,719
Total N = 87,949	Total	8,950 (10.18%)	8,204 (9.33%)	**7,571** (8.61%)	7,101 (8.07%)	6,619 (7.53%)	6,139 (6.98%)	5,723 (6.51%)
Per-item[a]	Training	1,432	1,263	1,144	1,034	921	804	768
	Test	620	542	492	449	394	353	330
	Total	2,052 (13.09%)	1,805 (11.99%)	**1,636** (11.27%)	1,483 (11.03%)	1,315 (10.48%)	1,157 (10.31%)	1,098 (9.97%)
	Total N =	15,680	15,048	14,513	13,450	12,547	11,218	11,010

[a] The models include only items whose arbitrage opportunities occur at least 5% and more than 4 times.
Note: the values of the main model are marked in bold.

Table 8. N sensitivity analysis with respect to γ

Scope		Feature	γ						
			2	3	4	5	6	7	8
F1 score	Generic	All	0.903	0.895	**0.883**	0.888	0.895	0.878	0.885
		Feature Selection	0.887	0.882	0.872	0.873	0.882	0.866	0.874
		Product Info Only	0.715	0.706	0.706	0.698	0.704	0.676	0.672
	Per-cat.	all	0.899	0.896	0.880	0.885	0.891	0.882	0.875
	Per-item	all	0.870	0.872	0.870	0.866	0.857	0.888	0.870
Precision	Generic	All	0.965	0.961	**0.954**	0.958	0.968	0.965	0.957
		Feature Selection	0.968	0.961	0.956	0.958	0.970	0.960	0.963
		Product Info Only	0.962	0.947	0.954	0.943	0.953	0.951	0.953
	Per-cat.	all	0.955	0.952	0.954	0.954	0.956	0.957	0.961
	Per-item	all	0.954	0.950	0.931	0.938	0.945	0.954	0.940
Recall	Generic	All	0.848	0.838	**0.823**	0.827	0.832	0.806	0.822
		Feature Selection	0.819	0.815	0.801	0.802	0.809	0.788	0.800
		Product Info Only	0.569	0.563	0.561	0.554	0.558	0.524	0.520
	Per-cat.	all	0.849	0.845	0.817	0.825	0.834	0.818	0.803
	Per-item	all	0.800	0.806	0.817	0.804	0.784	0.830	0.809

Note: the values of the main model are marked in bold.

Appendix F

In our AI model development, we evaluate whether a feature selection technique would improve the performance of our model. There, we leverage feature importance scores reported by random forests. Particularly, we develop a set of trees using the random forest algorithm and measure the decrease in the classification accuracy of the trees when each predictor is removed from the overall random forest model. Table 9 reports the predictive power of each predictor, and the predictors with boldface letters (i.e., those that improve model accuracy by at least 5%) are selected in the evaluation procedure. Four out of the five most important features are extracted from user-generated content (UGC), and information from reviews has higher predictive power compared with the information from other types of UGC (e.g., questions and answers). These results clearly present the role of UGC in arbitrage opportunity predictions.

Table 9. Predictive powers of each predictor

Group of Information	Predictor	% Inc in Acc.
Product information	**Price (brand new)**	**0.1647**
UGC	**Total number of reviews**	**0.0695**
UGC	**Average review sentiment**	**0.0633**
UGC	**Average review length**	**0.0643**
UGC	**Average star ratings**	**0.0663**
Product information	**Total available unit (brand new)**	**0.0605**
Product information	**Sales rank**	**0.0599**
Product information	**Price (used)**	**0.0569**
UGC	**Average review subjectivity**	**0.0571**
Product information	**Trade-in value**	**0.0537**
Product information	Total available unit (used)	0.0470
UGC	Average answer length	0.0335
UGC	Average answer sentiment	0.0315
UGC	Average question subjectivity	0.0327
UGC	Average question sentiment	0.0307
UGC	Average answer subjectivity	0.0301
UGC	Total number of answers	0.0279
UGC	Average question length	0.0259
UGC	Total number of questions	0.0245

Appendix G

Initially, we perform calculate the feature importance scores using random forest during the evaluation phase. As a result, the feature importance score is not calculated based on the best model (i.e., the model with optimal scope and hyperparameters). In this appendix, we calculate the feature importance scores from the best model and report the scores in Table 10. Observe that the scores reported here are remarkably consistent with the scores we obtain during the evaluation phase.

Table 10. Predictive powers of each predictor

Group of Information	Predictor	% Inc in Acc.
Product information	Price (brand new)	0.1612
UGC	Average review length	0.0651
UGC	Average star ratings	0.0621
UGC	Total number of reviews	0.0602
UGC	Average review sentiment	0.0597
Product information	Price (used)	0.0567
UGC	Average review subjectivity	0.0558
Product information	Total available unit (brand new)	0.0543
Product information	Sales rank	0.0487
Product information	Trade-in value	0.0467
UGC	Average answer length	0.0421
Product information	Total available unit (used)	0.0419
UGC	Average question sentiment	0.0415
UGC	Average answer sentiment	0.0388
UGC	Average answer subjectivity	0.0371
UGC	Average question subjectivity	0.0371
UGC	Average question length	0.0349
UGC	Total number of questions	0.0283
UGC	Total number of answers	0.0278

Based on these feature importance scores, we also perform an additional analysis where we refit the best model with only predictors that improve the model accuracy by at least 5%. The performance of this alternative model is qualitatively similar to the performance of our main model.