DELAY PREDICTION ON INVENTORY SHORTAGES IN SPORTS EQUIPMENT SUPPLY CHAIN

LIEW YNG JENG

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes the methodology and procedures used to conduct the research in detail (Kothari, 2004). It includes the research approach, research design, population and sample collection, instruments and models used, data collection procedures, and data analysis techniques. The most important thing is the ethical considerations to ensure the validity and reliability of the research. This chapter outline aims to provide a clear and concise research framework for the problem of delay prediction on inventory shortages in the sports equipment supply chain.

3.2 Research Approach

This study uses the quantitative method to analyze numerical data and evaluate models (Creswell & Creswell, 2017). This method will generate the statistics data or historical data from large-scale survey research related to the sports equipment supply chain, and predict the problems of inventory shortages and transportation delays accurately. The data used is the consumer purchase record data that summarizes the product prices and types, consumer information, and delivery status. This method enables to monitor the measurements and record the results for statistical verification.

In addition, the ontology type that used during analyzing the data in this study is realism (Bhaskar, 1975). It is because of the assumption that delivery delays or inventory shortages, and those factors that may affect these phenomena in the sports equipment supply chain exist independently of human perception or interpretation. At the same time, the predictive model that developed in this study is based on

historical data and measurable variables, which conforms to the realism view. The reality of this study can be understood through observation, measurement, and analysis.

In terms of epistemology, the way of knowledge to acquired, justified, and validated in this study belongs to positivism (Bryman, 2016). It mainly emphasizes that knowledge is conjectural rather than absolute truth and comes from empirical evidence and objective observation. The purpose of this study is to study the causal relationship between supplier reliability, seasonal demand, and delayed delivery. This study relies on historical data on the sports equipment supply chain. It analyzes supply chain behavior in a data-driven manner by adopting quantitative techniques of statistical models and machine learning algorithms to minimize bias and objectively interpret data and predict results to ensure objectivity and unbias of the knowledge.

3.3 Research Design

This study will adopt the modeling method and experiment design as research design. The modeling method is a systematic approach that used to model real-world systems, processes or phenomena to understand, analyze or predict the behavior (Law & Kelton, 2015). It mainly uses mathematical modeling to identify the optimal conditions in the supply chain, predict the delays and inventory shortage problems in complex supply chain systems, view the asymptotic behavior of the system, and propose the effective solutions. The implementation steps are to parameterize the different attributes of the collected historical data, analyze its parameters to establish the prediction model, and finally test the system behavior under different conditions to find the optimal solution for delays and inventory shortages.

The most suitable design under the experiment design is the time series design and non-randomized control group pretest-posttest design or non-equivalent group design in the quasi-experimental design series (Shadish, Cook, & Campbell, 2002). The quasi-experimental design is to generate supply chain data from historical records,

which can be fully utilized and the influence of confounding variables can be controlled as much as possible. The time series design is suitable because it allows to observe the specific interventions to achieve dynamic time analysis. This design focuses on the changes in data collected at multiple time points before and after the intervention. In contrast, the non-randomized control group pretest-posttest design is suitable because it can achieve non-randomized grouping functions according to innate conditions. This design focuses on creating natural groupings and comparing the results of each group based on different conditions in the supply chain environment, such as transportation distance or seasonal demand changes.

3.4 Population and sample

This section will describe the steps on how the data obtained will be processed and analyzed. In this study, this section will focus on the interaction and relationship between the population, sample, and sampling.

3.4.1 Population

The population for this study consists of historical supply chain data from a sports equipment company. This includes a wide range of data related to customer information, order information, product information, payment and shipping information, department information, and category information. Specifically, the population covers all available records over a three-year period from 2015 to 2018, which is crucial for identifying patterns and predicting inventory shortages and shipment delays (Kumar, 2019).

3.4.2 Sample

A sample is a subset of a population to become manageable for selective analysis. The sample for this study covers three years of historical supply chain data. It is selected using a stratified sampling method to ensure the representation of key

variables such as seasonal demand during peak and off-peak sales periods. At the same time, judgmental sampling is also used to focus on the complete part of data to be accurate in analysis (Etikan et al., 2016).

3.4.3 Sampling

Sampling is the process of selecting a subset of individuals or items from a larger population to represent that population. The following are the steps in a sampling plan:

1. Define Population:

   All historical supply chain data, including payment type, shipping days, benefit per order, sales per customer, delivery status, late delivery risk, category id, and name; the customer information like city, country, email, first name, last name, password, segment, state, street, and zipcode; the department information like id and name, latitude, longitude, market, order information like city, country, customer id, date, id, item, item card prod id, item discount rate, item id, item product price, item profit ratio, item quantity, sales, item total, profit per order, region, state, status, and zipcode, the product information like card id, category id, description, image, name, price, and status; the shipping information like date and mode. These data are categorized by customer information, order information, product information, payment and shipping information, department information, and category information.
2. Attain Sample Frame:

   Collect comprehensive records from the company's databases.
3. Design Sample Plan:

   For probability samples known likelihood of selection, use stratified sampling to divide the population into distinct subgroups and select the samples from each subgroup as representative such as seasonal or regional subgroups. Likewise, for non-probability samples known likelihood of selection, adopt judgmental sampling or purposive sampling to select the samples based on the judgment to find the best representative since it allows to make selective

analysis and target the high-impact data such as late delivery risk or customers in a particular segment.

4. Draw Sample:

   Extract a dataset covering from 2015 to 2018, ensuring the low demand period and high demand period are involved. The following figure is the sports equipment supply chain dataset.

Type	Days for shi	Days for shi	Benefit per	Sales per cu	Delivery Sta	Late_delive	Category Id	Category Na	Customer C	Customer C
DEBIT	3	4	91.25	314.64001	Advance shi	0	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	5	4	-249.09	311.35999	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
CASH	4	4	-247.78	309.72	Shipping on	0	73	Sporting Go	San Jose	EE. UU.
DEBIT	3	4	22.860001	304.81	Advance shi	0	73	Sporting Go	Los Angeles	EE. UU.
PAYMENT	2	4	134.21001	298.25	Advance shi	0	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	6	4	18.58	294.98001	Shipping car	0	73	Sporting Go	Tonawanda	EE. UU.
DEBIT	2	1	95.18	288.42001	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	2	1	68.43	285.14001	Late deliver	1	73	Sporting Go	Miami	EE. UU.
CASH	3	2	133.72	278.59	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
CASH	2	1	132.14999	275.31	Late deliver	1	73	Sporting Go	San Ramon	EE. UU.
TRANSFER	6	2	130.58	272.03	Shipping car	0	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	5	2	45.689999	268.76001	Late deliver	1	73	Sporting Go	Freeport	EE. UU.
TRANSFER	4	2	21.76	262.20001	Late deliver	1	73	Sporting Go	Salinas	EE. UU.
DEBIT	2	1	24.58	245.81	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	2	1	16.389999	327.75	Late deliver	1	73	Sporting Go	Peabody	EE. UU.
DEBIT	2	1	-259.58	324.47	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
PAYMENT	5	2	-246.36	321.20001	Late deliver	1	73	Sporting Go	Canovanas	Puerto Rico
CASH	2	1	23.84	317.92001	Late deliver	1	73	Sporting Go	Paramount	EE. UU.
DEBIT	2	1	102.26	314.64001	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
PAYMENT	0	0	87.18	311.35999	Shipping on	0	73	Sporting Go	Mount Pros	EE. UU.
TRANSFER	0	0	154.86	309.72	Shipping on	0	73	Sporting Go	Long Beach	EE. UU.
TRANSFER	5	4	82.300003	304.81	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	4	2	22.370001	298.25	Late deliver	1	73	Sporting Go	Rancho Cor	EE. UU.
TRANSFER	3	2	17.700001	294.98001	Shipping car	0	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	2	2	90.279999	288.42001	Shipping car	0	73	Sporting Go	Billings	EE. UU.
DEBIT	6	2	131.17	285.14001	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	5	2	90.540001	278.59	Late deliver	1	73	Sporting Go	Wilkes Barr	EE. UU.
PAYMENT	4	4	82.589996	275.31	Shipping on	0	73	Sporting Go	Caguas	Puerto Rico
DEBIT	3	4	-17.14	272.03	Advance shi	0	73	Sporting Go	Roseville	EE. UU.

Figure 3.1: Sports Equipment Supply Chain Dataset

5. Assess Sample:

   Check the sample for representativeness and ensure it aligns with research objectives.

6. Resample:

   Refine the sampling method to solve the missing or biased data points if necessary (Cochran, 1977).

3.5 Instrumentation

This section will illustrate the project requirements in this study, including software, hardware, methodology or algorithm, and dataset. The following is introducing the table of project requirements with descriptions:

Requirements	Items	Descriptions
Software	Python	A high-level and general-purpose programming language
	Jupyter Notebook	A web-based interactive compuitng platform
	Tableau	An interactive data visualization software that can help anyone see and understand the data
	PowerBI	A unified, scalable platform for self-service and enterprise business intelligence
Hardware	8GB RAM HP Laptop	A device to run the software
Methodology or Algorithm	Autoregressive Integrated Moving Average (ARIMA)	A time series forecasting statistical model (Gabellini et al., 2024)
	Extreme Gradient Boosting (XGBoost)	A machine learning algorithm for predictive modeling that achieves scalable and accurate forecasting (Keung et al., 2021)
	Seasonal ARIMA (SARIMA)	A statistical model for time series forecasting that can handle periodic changes in data, making up for the inability of the ARIMA to predict seasonal changes (Keung et al., 2021)
	Long Short-Term Memory (LSTM)	A recurrent neural network (RNN) that handles long-term reliance on sequence data, and can handle problems with recurring delays and fluctuating demand (Keung et al., 2021)
Datasets	Customer Information	The information of customers such as city, country, email, first name, last name, password, segment, state, street,

		and zipcode
	Order Information	The information of orders such as city, country, customer id, date, id, item, item card prod id, item discount rate, item id, item product price, item profit ratio, item quantity, sales, item total, profit per order, region, state, status, and zipcode
	Product Information	The information of products such as card id, category id, description, image, name, price, and status
	Payment and Shipping Information	The information of payment and shipping such as payment type, shipping days, benefit per order, sales per customer, delivery status, late delivery risk, category id, name, date, mode
	Department Information	The information of departments such as id and name, latitude, longitude, market

Table 3.1 Project Requirement Table

3.6 Data Collection Methods

The data source for this study was obtained from the Kaggle platform, which is shared data in the secondary data collection method. Since the Kaggle platform allows to access to the datasets without permission and the datasets in a variety of fields are provided (*Kaggle Inc., n.d.*). The sources of these datasets are all public data sets shared by researchers, organizations or individuals (*Johnston, 2017*). The selected datasets include sales records, transportation records, and so on, which are essential for modeling and predicting the supply chain delays and inventory shortages of sports equipment. The secondary data collection method allows researchers to focus on analysis and modeling without spending a lot of time and resources on primary data collection method (*Johnston, 2017*).

3.7 Data Analysis

This section will describe the data anlaysis in this study, including analysis technique, software, and unit of analysis.

3.7.1 Analysis Technique:

Analysis Technique describes the data analysis methods used in the research process. In addition to the previously mentioned XGBoost and LSTM to build predictive modeling for data analysis, descriptive statistics will also help analyze the situation. Descriptive statistics are divided into categorical frequency and grouping frequency in the frequency distribution. The categorical frequency will use the nominal level to analyze the uncalculated part of the data, while ordinal level data or grouped frequency will group and classify the calculable data to read each data level easily (*Smith, 2020*).

3.7.2 Software:

The Software section will describe the software used in the research process. As mentioned before, Python will use pandas and numpy types for data cleaning and processing, so that XGBoost can perform pattern training (*Wang & Liu, 2019*).. Tableau or powerbi is used to display the visualization of sales forecast results and trends, making it easy to read the situation at each level (*Kellen, 2021*).

3.7.3 Unit of Analysis:

Unit of Analysis describes the analysis object of interest in the data analysis process to confirm how to organize data and make inferences. The Unit of Analysis used in this study is Product level and transaction level. The product level is to observe the sales of each product and the transaction level is to confirm the sales, customer information and delivery status of each transaction (*Tan & Lee, 2018*).

3.8 Validity, Reliability, and Ethical Considerations

This section will describe the validity, reliability, and ethical consideration in this study.

- **Validity:** This study uses well-established statistical methods, such as ARIMA, SARIMA, XGBoost, and LSTM prediction methods, to ensure the research is stable and effective (*Jones & Lee, 2020*).
- **Reliability:** This study will test various prediction models on the dataset at different timeframes to ensure that the results obtained from repeated tests are consistent (*Robinson, 2019*).
- **Ethical Considerations:** The dataset used in this study has encrypted the private information of company, suppliers and customers. Therefore, this study adheres to the ethical guidelines of data use and transparency to ensure that all data processing processes comply with privacy regulations (*Smith & Green, 2018*).

3.9 Chapter Summary

This chapter outlines the methodology used to predict delays and inventory shortages in the sports equipment supply chain. It begins with an explanation of the research approach, which adopts a quantitative method to analyze historical supply chain data using statistical models and machine learning techniques. The research is grounded in realism ontology and positivist epistemology, emphasizing objective observation and empirical evidence.

The research design employs modeling and quasi-experimental approaches, including time-series design and non-randomized control group pretest-posttest design. These methods allow for dynamic time analysis and control of confounding variables, ensuring accurate predictions and effective solutions.

The population consists of historical supply chain data from 2015 to 2018, covering customer, order, product, payment, shipping, and department information. Stratified sampling ensures representation of key variables, while judgmental sampling targets high-impact data, such as late delivery risks.

Instrumentation includes software like Python, Jupyter Notebook, Tableau, and Power BI for data processing and visualization, alongside statistical models such as ARIMA, SARIMA, XGBoost, and LSTM for predictive analysis. Hardware requirements include an 8GB RAM laptop to support the computation.

Data collection is based on secondary data from Kaggle, enabling access to publicly shared datasets for analysis. Data analysis combines descriptive statistics with machine learning to identify patterns and predict outcomes. The unit of analysis includes product-level and transaction-level data, focusing on sales, customer details, and delivery statuses.

Validity and reliability are ensured through established statistical methods and repeated testing of prediction models. Ethical considerations include encrypting private information to adhere to privacy regulations and maintain transparency throughout the research process.

In summary, this comprehensive methodology ensures the research's rigor and reliability in addressing supply chain challenges.

3.10 Reference

Cochran, W. G. (1977). Sampling techniques (3rd ed.). Wiley.

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics, 5*(1), 1-4.

Johnston, M. P. (2017). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries (QQML), 3*, 619-626.

Jones, M., & Lee, R. (2020). Statistical methods in predictive modeling. *Journal of Applied Statistics, 25*(4), 30-42.

Kaggle Inc. (n.d.). About Kaggle datasets. Retrieved from https://www.kaggle.com

Kellen, V. (2021). Data visualization tools for business analytics. *Journal of Data Science, 9*(2), 105-116.

Kumar, R. (2019). Research methodology: A step-by-step guide for beginners. Sage Publications.

Law, A. M., & Kelton, W. D. (2015). Simulation modeling and analysis. McGraw-Hill Education.

Robinson, T. (2019). Testing reliability in predictive analytics. *Journal of Data Science, 11*(3), 56-62.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

Smith, J. (2020). Descriptive statistics and its use in business research. *Statistical Review, 8*(3), 45-56.

Smith, L., & Green, P. (2018). Ethical issues in data collection and use. *Journal of Information Ethics, 7*(2), 120-130.

Tan, Y., & Lee, J. (2018). Identifying unit of analysis for e-commerce transactions. *Journal of Business Analytics, 7*(1), 23-30.

Wang, Z., & Liu, S. (2019). Python libraries for machine learning: An overview. *Journal of Machine Learning Research, 15*(4), 112-118.