Logistics Data Analytics and Delay Prediction

Chotmanee Boonma

Logistics Management Program, Faculty of Science and Technology, Bangkok Suvarnabhumi University, Thailand
boonmachotmanee@gmail.com

Abstract

With transportation delays, the logistics sector has many challenges in ensuring the timely delivery of goods. In order to predict transportation delays, this study investigates the use of predictive modelling and logistics analytics. A dataset consisting of 10,000 randomly selected records from a larger Kaggle dataset which use for study intends to create efficient models for forecasting delays based on previous logistical data by applying machine learning methods such as Naïve Bayes, Decision Trees, and Logistic Regression. The CRISP-DM framework is used in the study to prepare and analyze data in a methodical manner, create prediction models, and assess how well they operate. The Decision Tree model achieved the highest accuracy (74.75%) otherwise Logistic Regression had the highest recall (83.31%) which making it the best at detecting delays. The Naïve Bayes classifier achieved 71.58% accuracy, 38.89% precision, 13.39% recall, and an F1-score of 0.63. The Decision Tree model had 74.75% accuracy, 40.00% precision, 0.33% recall, and an F1-score of 0.68, whereas Logistic Regression had 54.37% accuracy, 33.64% precision, 83.31% recall, and an F1-score of 0.68. The findings suggest that Logistic Regression is the best at identifying delays with a higher recall, even though the Decision Tree model has the maximum accuracy. By implementing predictive models and advanced analytics, logistics providers can reduce delays, optimize resources, and improve customer satisfaction. In order to increase operational effectiveness and proactively control delays, this paper offers suggestions for combining real-time tracking systems with predictive analytics.
Keywords: Logistics data analytics, Data Mining, Machine Learning, CRISP-DM

Background and Statement of the problem

In today's highly competitive logistics and shipping world on-time delivery of goods is critical to customer satisfaction and business success. However, shipping delays continue to be a major problem affecting both service providers and recipients. There are so many factors such as Traffic conditions, weather, vehicle error, human error and other unknown factors all affect the efficiency and timeliness of freight transportation. Therefore, it is essential to develop a system that can analyze and predict transportation delays so addressing this issue is crucial as proactive risk management can significantly enhance operational efficiency and mitigate potential disruptions. (Rezki & Mansouri, 2024). According to World Bank data (2023), Southeast Asian countries continue to experience transit delays as a result of traffic congestion, severe weather changes, and inadequate infrastructure. This causes an average delivery delay of 4.8-6.1 days per shipment in some countries. (Arvis et al., 2023).

Bhattacharjee et al., (2024) discovered that inefficient logistics operations, particularly in the mid to late stages of the supply chain, can account for 13-19% of overall logistics expenditures, resulting in a global revenue loss of up to 95 billion USD each year. To understand how important it is to address this issue as researchers have launched the Prediction Logistics Analytics and Delay project which aims to apply current data analytics technologies and develop models to predict distribution delays. The problem of transportation delays is the result of several interrelated factors, both external and internal, including rising customer expectations, all of which play an important role in the overall logistics process. In terms of external factors, traffic congestion, especially in major cities, inevitably leads to transportation delays, particularly during rush hours or festivals.

Al-Selwi et al., (2022) studied Unpredictable weather conditions such as storms, flooding and smog are also significant obstacles that disrupt transportation plans. Construction or repair work on infrastructure such as roads and bridges can lead to partial road closures and therefore unexpected route changes. In addition,

changes in legal regulations, such as anti-pollution measures or weight restrictions, require operators to constantly adapt. (Chawla et al., 2022). Finally, the lack of a system for systematically storing data and analyzing the causes of delays leads to recurring problems because we do not learn from past experience. In addition, customers have higher expectations of delivery, particularly due to the increase in e-commerce business, which demands a faster service. There is also a need for transparency in tracking products through real-time notification systems. If logistics providers are unable to meet these expectations, it will inevitably impact on customer satisfaction and repeat purchase rates. Moreover, given the complexity of these factors may help reducing transportation delays require advanced data analysis and the development of predictive models that can fully account for all factors. This will help to make more accurate estimates of delivery times, reduce errors and truly improve the efficiency of the logistics system.

Objective

1. To analyze the factors affecting transportation delays using order data, transportation methods, and geographical information.
2. To develop and compare models for predicting transportation delays using data analysis and machine learning techniques.
3. To propose solutions for reducing transportation delays and improving the efficiency of the logistics system.

Expected benefits

1. Reduce transportation delays by improving planning and decision-making processes so the company can choose the right routes and avoid potential risks.
2. Reduce the costs caused by delays and increase the efficiency of resource deployment, e.g. of vehicles and personnel, to maximize resource utilization.
3. Increase transparency in tracking product status by using a real-time notification system so that stakeholders can receive accurate and timely information.

Conceptual Framework

In this study was applied CRISP-DM (Cross Industry Standard Process for Data Mining) it is a standard framework for managing structured data to develop models. Forecasting consists of six main steps: business understanding, data understanding, data preparation, modeling, evaluation and deployment. This framework helps to systematically manage the data pipeline to ensure accurate and practical predictions. (Narendra, 2025).

To apply the research systematically and efficiently which CRISP-DM framework is used as a guide for data analysis as follows:

1. Business Understanding: the first step is to clearly define the business problem and the objectives of the research. This includes creating objectives to improve actual delays and analyzing the value, issues, and impact of deliveries.
2. Understanding the data: In this step, all available data is analyzed by starting with a descriptive statistical analysis to understand the basic characteristics of the different variables then explore the dataset on Kaggle check descriptive statistics and perform preliminary data analysis.
3. Data preparation: The data obtained often needs to be improved and prepared before it can be used to create models. This includes dealing with selecting the important attributes, filling in the missing information and converting the numerical values into labels that serve as the step names.
4. Modeling: In this phase, the performance of different model types is compared to finding the most suitable model for predicting delays. These can create a prediction model using Naïve Bayes, Decision Tree, and Logistic Regression.
5. Evaluation: Once the model has been created, we will evaluate the model by using the confusion matrix and the scoring matrix. In addition, a feature importance analysis is performed to determine which variables have the greatest influence on the delays.
6. Deployment: The final step is to implement the model evaluated for effectiveness and evaluate the model using the confusion matrix and the scoring matrix. Then, creating dashboards to track and analyze the prediction results and setting up processes to continuously improve the model to maintain effectiveness over time.

Predictive analytics & forecasting models are a group of techniques used to analyze trends and behaviors of past data to predict future delays including Decision Tree, Naïve Bayes and Logistic Regression which help companies analyze and improve transportation routes to become more efficient. (Can et al., 2023)

In this article, the concept of Machine Learning has been applied to create a model that can analyze and predict problems from existing datasets. Machine Learning is a process that enables computers to learn from data and develop models to predict outcomes or classify data without the need to write specific programs for every situation. (Aljohani, 2023) In this operation, there are three types of machine learning models that were chosen:

- Naïve Bayes is a technique that relies on the principles of probability according to Bayes' Theorem, assuming that the various features of the data are independent of each other. This allows for the efficient calculation of the probability of each class. Even though it is a simple method Naïve Bayes often yields accurate results in many contexts, especially with data that is not very complex. (Azeraf et al., 2020)
- Decision Tree is a model that uses tree structure for decision-making. The data is split based on features that most influence the outcome at each level, making it easy to understand and track the forecasting process. It is also suitable for diverse data. (Priyanka & Kumar, 2020)
- Logistic Regression is a statistical method used for classification, especially in cases where the outcome is a binary value, such as yes/no or delay/not delay. It uses the logistic function to transform the predicted values into a range between 0 and 1 reflecting the probability of each target group. (Abramovich et al., 2020)

Evaluation Methods

Agrawal (2021) studied that the performance of the classifier can be calculated from the confusion matrix. After being compared to the actual result the classifier results can generate four values:

- True Positive (TP): the predicted value is positive; the actual value is positive.
- True Negative (TN): the predicted value is negative; the actual value is negative.
- False Positive (FP): the predicted value is positive; the actual value is negative.
- False Negative (FN): the predicted value is negative; the actual value is positive.
- Four measures were used to measure the performance of selected algorithms: accuracy, precision, recall, and F1 score. These measures are all positively related to the quality of algorithms. Consequently, the higher values of these measures are, the better their performances are. The value of the four measures can be obtained by calculations using these parameters

Table 1. Measures of Evaluation

Metric	Formula and Description
Accuracy	$Accuracy$ = (TP + TN)/ (TP + FP + TN + FN)
Precision	$Precision$ =TP/ (TP + FP)
Recall	$Recall$ =TP/(TP+FN)
F1 Score	$F1$ = (2 x Precision x Recall)/ (Precision + Recall)

Research Methodology

This study examines and develops predictive models for transport delays using quantitative analysis techniques and machine learning. The study follows the framework of data science and statistical validation techniques. (Garg et al., 2025). This essay was written with a focus on data mining, specifically the prediction problem of accurately categorizing and identifying transactions that may be delayed in distribution. In general, predictive procedures employ various supervised learning methods to examine historical data in order to construct predictive models, which are then applied to new data to forecast the chance of a logistics delay. The paper focusses on the prediction problem for identifying logistical delays, where the goal variable is defined as a binomial, with 1 indicating delays and 0 indicating non-delays. Given the binary nature of our target variable, this problem becomes a classification problem, with the purpose of estimating the probability connected to each class or category of the target variable.

Data Collection

This study utilizes a publicly available logistics dataset from Kaggle called “Logistics Data Analysis and Delay Prediction” from https://www.kaggle.com/code/devraai/logistics-data-analysis-and-

delay-prediction/input. The dataset includes a variety of information related to logistics, such as customer state, market, number of items ordered, sales, order region, order status, processing time, shipping method, and more. The prediction of delivery delays was carried out using binary classification, focusing on the variable "IS_DELAY." A detailed description of the dataset's features is shown in Table 1.

Due to the limitations of the RapidMiner software used for analysis, this study extracted a subset of 10,000 rows from the entire data set. The dataset is secondary data contributed by Kaggle contributors and is freely available for educational and research purposes. As a result, the researchers did not collect the data directly but rather received it from the Kaggle website.

Table 2. Attribute Description For Data Set

Attribute Name	Type	Description
Label	Binominal	Is_delay/Not_delay
Profit per order	Real	Earnings per order placed
Category name	Text	Description of the product category
Customer country	Categorical	Country where the customer made the purchase
Customer segment	Categorical	Types of Customers Consumer, Corporate, Home Office
Customer state	Categorical	State to which the store where the purchase is registered
Market	Categorical	Market to where the order is delivered
Order item quantity	Numerical	Number of products per order
Sales	Numerical	Value in sales
Order region	Categorical	Region of the world where the order is delivered
Order status	Categorical	Complete, pending, closed, pending_payment, canceled, processing, on_hold, payment_review
Order date	Datetime	Date on which the order is made
Shipping date	Datetime	Exact date and time of shipment
Shipping mode	Binomial	Standard Class, First Class, Second Class, Same Day

Data Processing

The data processing pipeline was implemented using RapidMiner incorporating essential preprocessing steps to prepare the dataset for machine learning algorithms. The dataset was retrieved from the local repository and the Multiply operator was applied to allow parallel preprocessing without altering the original dataset. Two instances of the Select Attributes operator were used to refine the dataset. The first selection included attributes relevant to prediction tasks such as *label, customer_state, customer_segment, customer_country, profit_per_order, sales, order_status, order_item_quantity, order_processing_time, shipping_mode, order_region, and category_name.* The second selection further refined the dataset based on relevance to customer order patterns and delivery delays, keeping attributes like *label, market, order_item_quantity, order_region, order_status, sales, shipping_mode, and customer_segment.* Missing values were handled using the Replace Missing Values operator by applying average values for numerical attributes and most frequent value replacement for categorical attributes. Data transformation steps included converting nominal variables to numerical format through integer encoding, ensuring that each unique category was mapped to a distinct integer value for algorithm compatibility. Finally, the Set Role operator was used to designate the label attribute as the target variable, likely to represent whether a delivery delay occurred or not. As a result, the dataset was cleaned, transformed, and streamlined, retaining only meaningful features, and is now structured for supervised machine learning, ensuring reliable model training and evaluation.

Data Splitting Process

The effective machine learning model training and evaluation which the dataset was divided into two parts using a 70:30 split ratio 70% of the data was allocated for training while 30% was reserved for testing.

- Training Data (70%): Used to teach model patterns, relationships, and decision-making strategies.
- Testing Data (30%): Kept separate to evaluate how well the model generalizes to unseen data, preventing overfitting.

Moreover, the three Split Data modules were applied individually across different model lines. Each module ensures that data is divided in a consistent manner, maintaining the integrity of the training/testing process for distinct machine learning models. By structuring the data split this way the models can learn effectively while also being tested on scenarios leading to better accuracy and reliability.

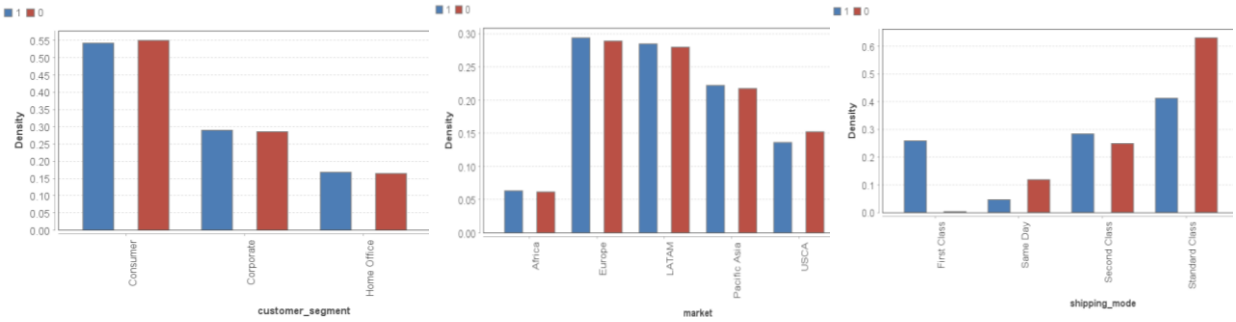

Figure 1. *Data Visualization*

Models Comparison

Table 3 shows the evaluation metrics for each algorithm used in the study. The table highlights the highest value for each metric in bold font, while the lowest value for each metric is denoted by an underscore. The bold values represent the best performance for each evaluation metric across the three models, signifying which algorithm excels in that particular aspect. Also, the underscore values represent the worst performance for each evaluation metric, indicating which algorithm struggled most in that specific aspect. For instance, Logistic Regression achieved the highest recall (83.31%), meaning it was particularly effective at detecting actual delays. In contrast, Decision Tree performed best in accuracy (74.75%) and precision (40.00%), demonstrating its overall ability to classify instances correctly, though with a slightly less effective recall.

Table 3. Evaluation Metrics for Each Algorithm

Algorithm	Evaluation			
	Accuracy	Precision	Recall	F1-Score
1. Naïve Bayes	71.58	38.89	*13.39*	*0.63*
2. Decision Tree	**74.75**	**40.00**	*0.33*	**0.68**
3. Logistic Regression	54.37	33.64	**83.31**	**0.68**

Research Results

The study applied various machine learning models and data visualization techniques to analyze transportation delays using order data, shipping methods, and geographical information. The key findings showed that First Class and Second-Class shipping modes were the most reliable, while Same Day and Standard Class shipping were more prone to delays. The analysis of order processing time revealed that although the distributions for successful and failed deliveries overlapped significantly, there were notable differences in their density patterns which suggest that delays often occur within certain time windows.

From a modeling perspective, three models were tested: Decision Tree, Naïve Bayes, and Logistic Regression. The Decision Tree achieved the highest accuracy (74.75%) but had very low recall for detecting delays. Naïve Bayes balanced better between accuracy and recall, while Logistic Regression had a low overall accuracy but was excellent at detecting potential delays (recall of 83.31%). These results highlight the trade-offs between precision, recall, and accuracy in different modeling approaches.

Summary of the Study

Transportation delays are influenced by multiple factors, including shipping methods, order processing time, and geographical attributes. The analysis indicates that First Class shipping is the most reliable, while Same Day and Standard Class shipments face higher risks of delay. Additionally, orders processed quickly with low sales values are more prone to failures, demonstrating the complex interaction between logistics, sales, and regional handling efficiency. To predict delays, three machine learning models Decision Tree, Naïve Bayes, and Logistic Regression were compared. The Decision Tree model achieved the highest accuracy (74.75%) but struggled to detect actual delays. Naïve Bayes had lower accuracy (71.58%) but performed better in identifying delays, while Logistic Regression demonstrated the highest recall (83.31%) yet suffered from lower overall accuracy. The choice of model depends on application

needs: Decision Trees offer better general classification, while Logistic Regression is more effective at minimizing undetected delays. Based on these insights, several solutions are proposed to improve logistics efficiency. Promoting reliable shipping methods, setting minimum order thresholds, and carefully managing Same Day deliveries can significantly reduce delays. Predictive modeling can flag high-risk orders, allowing proactive intervention, while tailoring logistics strategies based on region and customer segments increase efficiency. These data-driven strategies offer a comprehensive approach to optimizing transportation reliability and ensuring smooth operations.

The models' sensitivity to data imbalance and how they handle missing data are two of their main drawbacks. When there is an imbalance in the data, such as when there are significantly more late submissions than successful submissions, models like Decision Trees and Logistic Regression may tend to skew the predictions in favor of the larger group. Additionally, each model has a different capacity to deal with missing data. In contrast to Naïve Bayes, which requires complete data or employs a method to fill in missing values, Decision Trees are able to handle missing values effectively through surrogate splits. Logistic regression necessitates a great deal of data preparation prior to use and is more impacted by incomplete data. Consequently, it is crucial to take these factors into account when choosing a model in order to get the most accurate results given the data context.

Recommendations

Integrating Logistic Regression into the system to indicate high-risk shipments provides a data-driven strategy to proactively identify and manage delays. This enables the logistics team to intervene before problems get out of hand, improving overall efficiency and customer happiness. Predictive analytics also aids decision-making by ensuring that resources are deployed efficiently, and high-priority shipments are handled with greater attention. Moreover, implement real-time tracking and monitoring systems using IoT and AI. It can identify future delays and send alerts to logistics management as it provides them to respond quickly before the problem goes up. This may assist with route optimized performance, adaptive rescheduling, and altering delivery expectations based on current conditions. This technology-driven strategy improves decision-making with real-time data, reducing delays and increasing overall logistics efficiency. Furthermore, this system might be linked to consumer-facing applications, allowing end users to receive precise shipment updates, increasing customer satisfaction and decreasing uncertainty.

Reference

Abramovich, F., Grinshtein, V., & Levy, T. (2020). *Multiclass classification by sparse multinomial logistic regression*. ArXiv.org. https://arxiv.org/abs/2003.01951

Agrawal, S. K. (2021, July 20). *Evaluation Metrics for Classification Model| Classification Model Metrics*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/

Aljohani, A. (2023). Predictive Analytics and Machine Learning for Real-Time Supply Chain Risk Mitigation and Agility. *Sustainability*, *15*(20), 15088. mdpi.

Arvis, J.-F., Ojala, L., Shepherd, B., Ulybina, D., & Wiederer, C. (2023). Connecting to Compete 2023 Trade Logistics in the Global Economy The Logistics Performance Index and Its Indicators. https://lpi.worldbank.org/sites/default/files/2023-04/LPI_2023_report.pdf

Azeraf, E., Monfrini, E., & Pieczynski, W. (2020). *Using the Naive Bayes as a discriminative classifier*. ArXiv.org. https://arxiv.org/abs/2012.13572

Bhattacharjee, D., Kamil, A., Lukasiewicz, M., & Melnikov, L. (2024, January 5). Digitizing mid- and last-mile logistics handovers to reduce waste. McKinsey & Company. https://www.mckinsey.com/industries/logistics/our-insights/digitizing-mid-and-last-mile-logistics-handovers-to-reduce-waste

Can, C. M., Ravi, S. K., & Saleh, S. (2023). *Enhancing Supply Chain Resilience: A Machine Learning Approach for Predicting Product Availability Dates Under Disruption*. ArXiv.org. https://arxiv.org/abs/2304.14902?utm_source=chatgpt.com

Chawla, P., Hasurkar, R., Bogadi, C. R., Korlapati, N. S., Rajendran, R., Ravichandran, S., Tolem, S. C., & Gao, J. Z. (2022). Real-time traffic congestion prediction using big data and machine learning techniques. *World Journal of Engineering*. https://doi.org/10.1108/WJE-07-2021-0428

Garg, A., Mohmmad Ayaan, Parekh, S., & Vikranth Udandarao. (2025, March 19). *Food Delivery Time Prediction in Indian Cities Using Machine Learning Models*. Https://www.researchgate.net/. https://doi.org/10.48550/arXiv.2503.15177

Narendra, M. F. (2025). Forecasting Manpower Planning Using the CRISP-DM Method and Machine Learning Algorithm: A Case Study of Tiki Jalur Nugraha Ekakurir (JNE) Company. *Journal of Information Systems Engineering and Management*, *10*(19s), 371–378. https://doi.org/10.52783/jisem.v10i19s.3040

Priyanka, N. A., & Kumar, D. (2020). Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*, *12*(3), 246. https://doi.org/10.1504/ijids.2020.108141

Rezki, N., & Mansouri, M. (2024). Machine Learning for Proactive Supply Chain Risk Management: Predicting Delays and Enhancing Operational Efficiency. *Management Systems in Production Engineering*, *32*(3), 345–356. https://doi.org/10.2478/mspe-2024-0033