Supply Chain Analytics Using Machine Learning: An Exploratory and Predictive Study

Houssam Tachihante
Course: Advanced Statistics

Abstract

Supply chain operations generate large volumes of data that can be leveraged to improve efficiency, reduce delays, and mitigate risk. This project applies data analytics and machine learning techniques to a real-world supply chain dataset in order to explore delivery delay patterns and develop a predictive model for delay risk. Inspired by recent literature in supply chain data analytics, the study combines exploratory data analysis with a supervised learning approach to derive actionable business insights.

1 Introduction

Modern supply chains are increasingly complex and data-driven. Organizations collect detailed information on orders, shipments, customers, and delivery performance. However, transforming this data into actionable insights remains a challenge. Data analytics and machine learning have emerged as effective tools to support decision-making in supply chain management.

This project is inspired by the recent literature review by Darbanian et al. (2024), which highlights the growing role of data analytics and machine learning techniques in improving supply chain performance. Following the approaches discussed in the literature, this study applies exploratory and predictive analytics to analyze delivery delays and identify key factors influencing supply chain performance.

2 Research Background and Literature Review

Darbanian et al. (2024) provide a comprehensive review of data analytics applications in supply chain management, emphasizing the importance of descriptive, predictive, and prescriptive analytics. The authors highlight that machine learning methods are particularly effective in identifying complex patterns in large-scale supply chain data, such as demand fluctuations, delivery risks, and operational inefficiencies.

Motivated by these findings, this project focuses on predictive analytics to assess delivery delay risk. The use of a supervised learning model aligns with the literature's emphasis on leveraging historical data to anticipate future operational issues and support proactive decision-making.

3 Research Questions

Based on an initial analysis of the dataset and guidance from the literature, the following research questions were formulated:

1. Which shipment and customer-related factors are most strongly associated with delivery delays?
2. Are certain product categories or customer regions more prone to delayed deliveries?
3. Can a machine learning model accurately predict delivery delay risk using available numerical features?
4. What operational insights can be derived to improve supply chain efficiency and reduce delays?

4 Dataset Description

The dataset used in this project contains detailed information on supply chain transactions, including shipping times, delivery status, customer information, product categories, and financial metrics. Key variables include delivery status, scheduled versus actual shipping days, customer country, and product category.

Due to the large size of the original dataset, a representative subset was used for analysis to ensure computational efficiency and reproducibility. Data preprocessing included handling missing values and selecting relevant numerical features for modeling.

5 Methodology

The analysis follows a two-stage methodology:

5.1 Exploratory Data Analysis

Exploratory analysis was conducted to understand the distribution of delivery delays and identify high-risk segments. Summary statistics and grouped analyses were used to compare delay rates across customer regions and product categories.

5.2 Predictive Modeling

A Random Forest classifier was implemented to predict whether a shipment would be delayed. The target variable was constructed using delivery status and shipping time information. Numerical features were used as predictors, and the model was evaluated using standard classification metrics such as precision, recall, and accuracy.

Random Forest was selected due to its robustness, ability to handle nonlinear relationships, and interpretability through feature importance measures.

6 Results and Discussion

The exploratory analysis revealed that delivery delays are not uniformly distributed across the dataset. Certain customer regions and product categories exhibit higher delay rates, suggesting potential logistical or operational bottlenecks.

The machine learning model demonstrated that delivery delays can be reasonably predicted using historical shipment data. Feature importance analysis indicated that shipping duration variables and order-related metrics play a significant role in determining delay risk. These findings align with the literature, which emphasizes the value of predictive analytics for identifying risk factors in supply chain operations.

7 Managerial Insights and Recommendations

Based on the analysis, several managerial insights can be derived:

- High-risk customer regions and product categories should be monitored more closely to reduce delivery delays.
- Predictive models can be integrated into operational systems to flag shipments with high delay risk.
- Continuous data collection and analytics-driven decision-making can improve overall supply chain performance.

These recommendations are consistent with the conclusions of Darbanian et al. (2024), who stress the strategic importance of analytics capabilities in modern supply chains.

8 Reproducibility

All analysis code and documentation are provided through a GitHub repository. The Streamlit dashboard enables interactive exploration of the dataset and allows users to retrain the predictive model, ensuring full reproducibility of results.

9 Conclusion

This project demonstrates how data analytics and machine learning can be applied to supply chain data to identify delivery delay patterns and support informed decision-making. Inspired by recent academic literature, the study highlights the practical value of predictive analytics in improving supply chain efficiency and resilience.

10 References

References

[1] Darbanian, F., Brandtner, P., Falatouri, T., & Nasseri, M. (2024). Data Analytics in Supply Chain Management: A State-of-the-Art Literature Review. *Operations and Supply Chain Management*, 17(1).