**"Forecasting Late Arrivals: Delivery Estimation"**

**Machine Learning 2**

**(BIA – 5402)**

**Submitted to:**

**Dr. Ameera Al-Karkhi**

**Submitted by:**

**Vrajkumar Patel (N01581006)**

**Param Panchal (N01579822)**

**Pravina Prajapati (N01579926)**

**Dhairya Dangi (N01580705)**

**Prappan Batra (N01579150)**

**Submitted on:**

**June 4, 2024**

Table of Contents

[**1.** **ABSTRACT** 3](#_Toc168436357)

[**2.** **INTRODUCTION** 3](#_Toc168436358)

[**3.** **LITERATURE REVIEW** 4](#_Toc168436359)

[**4.** **ABOUT DATA** 5](#_Toc168436360)

[**5.** **EXPLORATORY DATA ANALYSIS (EDA)** 6](#_Toc168436361)

[**6.** **MODEL IMPLEMENTATION (NEURAL NETWORKS)** 8](#_Toc168436362)

[**7.** **CONCLUSION** 10](#_Toc168436363)

[**8.** **REFERENCES** 11](#_Toc168436364)

# **ABSTRACT**

As recognition grows for effective supply chain management in manufacturing firms, delivery performance becomes crucial. Global supply chains require sophisticated systems to navigate cultural, technical, and regulatory differences. When delivery problems arise, they quickly affect the entire supply chain. Traditionally, management responds with increased buffering, like extra inventory, which raises costs and hampers responsiveness. (Neural networks for financial market prediction, n.d.)

Although delivery performance is acknowledged, few studies empirically assess how supply chain factors influence it. This gap may be due to delivery being influenced by upstream and downstream operations. Thus, a holistic view of the supply chain is essential. To understand this linkage, a conceptual model is needed. This paper offers such a model, exploring the relationship between supply chain complexity and delivery performance.(Neural networks for financial market prediction, n.d.)

# **INTRODUCTION**

In today's competitive e-commerce landscape, ensuring timely delivery of products is crucial for maintaining customer satisfaction and loyalty. However, navigating the complexities of the supply chain often leads to challenges such as transportation delays and inventory management issues, resulting in late deliveries and potential revenue losses. Our business objective is to develop a predictive model that accurately forecasts the risk of late deliveries, allowing us to proactively mitigate risks and optimize delivery performance. By leveraging historical transaction data and machine learning algorithms, we aim to enhance customer satisfaction, optimize operations, and drive financial performance, ultimately gaining a competitive edge in the market.

For this project, the team utilizes the “DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS” dataset from Mendeley Data to build predictive models for latency in delivery. The dataset consists of approximately 180519 records and includes 53 variables, that capture a variety of information like transaction details, customer information, product details, order details, and the target variable ‘Late Delivery Risk’.

# **LITERATURE REVIEW**

Recent studies have explored diverse approaches to Supply Chain Risk Management (SCRM) using Artificial Intelligence (AI) techniques. Stochastic programming methods, exemplified by (SpringerLink, n.d.) and (Transportation Research Part E: Logistics and Transportation Review, n.d.), offer effective strategies for mitigating risks across supply chain operations. Similarly, fuzzy programming-based methodologies, as demonstrated by (Omega, n.d.) and (Taylor & Francis, n.d.), introduce flexibility to accommodate uncertainties in supply chain parameters. Additionally, network-based models, such as Bayesian Belief Networks and Color-Trans-Nets, provide valuable insights into the dynamic nature of SCRM decision-making processes (Econstor, n.d.); (Expert Systems with Applications, n.d.).

Machine Learning (ML) techniques have gained prominence in addressing supply chain challenges by analyzing large datasets and predicting future demands. Notable contributions include the work of (Computers & Operations Research, n.d.), who proposed a neural network-based ML approach for optimizing blood unit transshipment across hospital networks. Similarly, (Springer Link, n.d.) applied ML models to improve product backorder forecasting accuracy, even in the presence of biased data. Additionally, Deep Learning (DL) techniques, such as deep belief networks (DBN) and Long Short-Term Memory (LSTM) neural networks, have shown promise in forecasting future risks and demand fluctuations during the COVID-19 pandemic (IEEE Xplore, n.d.); (European Journal of Operational Research, n.d.).

# **ABOUT DATA**

The dataset includes multiple fields, each offering unique insights into various fields of the supply chain. Below is a detailed description of the key variables in our dataset:

|  |  |
| --- | --- |
| **Transaction Details**: | Type, Order Date (DateOrders), Order Id, Order Item Id, Order Customer Id, Product Card Id, Product Name, Product Description, Product Price, Sales per Customer, Sales, Order Status |
| **Customer Information**: | Customer Id, Customer Fname, Customer Lname, Customer Email, Customer City, Customer Country, Customer Segment, Customer State, Customer Street, Customer Zipcode |
| **Product Details**: | Category Id, Category Name, Department Id, Department Name, Market, Product Status, Product Image |
| **Order Details**: | Order Region, Order City, Order Country, Order State, Order Item Discount, Order Item Discount Rate, Order Item Product Price, Order Item Profit Ratio, Order Item Quantity, Order Item Total, Shipping Mode, Late Days, Order Year (Order Yr), Order Month, Order Day, Order Hour |
| **Target Variable:** | **Late Delivery Risk** |

**Table 4.1** (DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS, n.d.)

The extensive scope of the dataset facilitates a profound comprehension of the supply chain, providing valuable insights that have the potential to yield substantial enhancements in productivity and effectiveness.

# **EXPLORATORY DATA ANALYSIS (EDA)**

Our supply chain optimization project greatly benefits from exploratory data analysis (EDA), which reveals important patterns and insights hidden in the dataset. By employing diverse visualization methodologies like heatmaps, scatter plots, and histograms, our team is able to obtain a thorough comprehension of the correlations and distribution of the predictor variables. EDA is the cornerstone for developing strong predictive models for late delivery risk, which in turn improves the effectiveness and dependability of our supply chain operations by identifying patterns and abnormalities early in the analysis.

The figure below shows the correlation between all the variables in the dataset. As we can see, the variables having identical correlation values are grouped together. For further analysis, we can select only 1 of the group of variables showing identical correlations.

A screenshot of a computer

Description automatically generated

**Fig 5.1** Correlation Confusion Matrix

Below are the pie charts describing the delivery status. Fig. 5.2 below shows all the different possibilities associated with delivery status. Fig. 5.3 shows the distribution of late delivery risk. When there is no late delivery, it includes both the proportions, i.e. on time shipping and advanced shipping. Cancelled shipments are not required to be taken into consideration when looking at the risk of late delivery.

A pie chart with different colored circles

Description automatically generated

**Fig. 5.2** Delivery Status Proportion

A pie chart with numbers and text

Description automatically generated

**Fig 5.3** Late Delivery Risk Proportion

The below figure shows how the risk of late delivery varies with the payment type. As seen in the image, and being the most common type of payment, DEBIT incurs the highest risk for late delivery. Similarly, with cash being the least used method of payment, it incurs the least risk for late delivery.

A graph with blue and orange bars

Description automatically generated

**Fig 5.4** Late Delivery Risk vs Payment Type

The below image shows the distribution of late delivery risk over the frequency of the order status. As expected, the orders that are suspected fraud, and the cancelled orders have no risk of late delivery. And the completed orders have the highest risk of late delivery.

A graph of a bar chart

Description automatically generated with medium confidence

**Fig. 5.5** Late Delivery Risk vs Order Status

# **MODEL IMPLEMENTATION (NEURAL NETWORKS)**

To determine the most pertinent features for forecasting the risk of late delivery (late\_delivery\_risk), we carried out feature selection in this stage of our study. The ANOVA F-value (f\_classif) was particularly employed as the scoring function in the SelectKBest method from the sklearn.feature\_selection module. The dataset was first divided into two parts: the features (X\_s) and the goal variable (y\_s), with late\_delivery\_risk acting as a binary indicator of whether a shipment is delayed. The SelectKBest function was then instantiated, with k=10 specified to choose the top 10 features. Through this method, we were able to narrow down on the most important features, which may have improved the interpretability and performance of the model.

Subsequently, we split the dataset into training and test sets, reserving 30% of the data for testing and using the remaining 70% for training. We then applied the StandardScaler() function from scikit-learn to fit and scale the training data, standardizing the features to prepare the data for modeling. This meticulous preparation facilitated the building of a machine learning model that demonstrated high accuracy in both training and test datasets.

**6.1 MLP Classifier:**

To predict the risk of late delivery, we trained a neural network model using the MLPClassifier from the scikit-learn library. The network was configured with a single hidden layer containing 4 nodes and used the logistic activation function to capture non-linear patterns in the data. We employed the lbfgs solver for optimization, suitable for smaller datasets. The model was trained on the standardized training data (X\_train\_s and y\_train), ensuring consistent feature scaling. The neural network demonstrated an impressive accuracy of 99.47% and a precision of 99.04% on the training data, indicating that it effectively learned the patterns and relationships necessary for predicting the risk of late deliveries.

A blue squares with white text

Description automatically generated

**Fig. 6.1** Confusion Matrix for Training Data

**6.2 Testing:**

The confusion matrix visually represents the model's predictions compared to the actual outcomes for the testing data. It indicates that the model performed exceptionally well, with an accuracy of 99.50% and a precision of 99.10%. These high-performance metrics suggest that the model accurately classified both 'Late' and 'Not Late' deliveries, demonstrating its effectiveness in predicting late delivery risks.

**A blue squares with white text

Description automatically generated**

**Fig. 6.2** Confusion Matrix for Testing Data

# **CONCLUSION**

This project explored the application of machine learning techniques, specifically neural networks, to predict late delivery risks in DataCo Global’s supply chain. Through meticulous data preparation, feature selection using the ANOVA F-value method, and the development of a neural network model with the MLPClassifier, we achieved impressive results. Our model demonstrated high accuracy and precision, both in training and testing phases, indicating its robustness and generalization capability.

The confusion matrix analysis confirmed the model’s effectiveness, showing an accuracy of 99.50% and a precision of 99.10% on the testing data. These results highlight the potential of neural networks in identifying and mitigating risks associated with late deliveries, ultimately enhancing operational efficiency and customer satisfaction.

This study underlines the relevance of employing powerful machine learning algorithms in supply chain management to address complicated issues such as delivery delays. The insights gathered from this research can guide future efforts in optimizing supply chain performance, decreasing costs, and sustaining competitive edge in the marketplace. Future studies may focus on incorporating more complicated neural network designs and researching additional data sources to further improve forecast accuracy and operational outcomes.

# **REFERENCES**

*A new model to mitigating random disruption risks of facility and transportation in supply chain network design*. (n.d.). From Springer Link: https://link.springer.com/article/10.1007/s00170-013-5404-0

*Computers & Operations Research*. (n.d.). From Science Direct: https://www.sciencedirect.com/science/article/abs/pii/S0305054820300587

*DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS*. (n.d.). From Mendeley Data: https://data.mendeley.com/datasets/8gx2fvg2k6/5

*Econstor*. (n.d.). From A new approach for supply chain risk management: Mapping SCOR into Bayesian network: https://www.econstor.eu/handle/10419/188682

*European Journal of Operational Research*. (n.d.). From Science Direct: https://www.sciencedirect.com/science/article/pii/S0377221720306913

*Expert Systems with Applications*. (n.d.). From Science Direct: https://www.sciencedirect.com/science/article/abs/pii/S0957417411011109

*IEEE Xplore*. (n.d.). From Application of Deep Learning Neural Network in Online Supply Chain Financial Credit Risk Assessment: https://ieeexplore.ieee.org/abstract/document/9148203

*Neural networks for financial market prediction*. (n.d.). From IEEE Xplore: https://ieeexplore.ieee.org/abstract/document/374354

*Omega*. (n.d.). From Science Direct: https://www.sciencedirect.com/science/article/abs/pii/S0305048316000177

*Springer Link*. (n.d.). From Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques: https://link.springer.com/article/10.1186/s40537-020-00345-2

*SpringerLink*. (n.d.). From SpringerLink: https://link.springer.com/article/10.1007/s00170-013-5404-0

*Taylor & Francis*. (n.d.). From How to choose mitigation measures for supply chain risks: https://www.tandfonline.com/doi/abs/10.1080/00207543.2013.828170

*Transportation Research Part E: Logistics and Transportation Review*. (n.d.). From Science Direct: https://www.sciencedirect.com/science/article/abs/pii/S1366554514001033