Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Data-Driven Optimisation of Supply Chain Management

By

Group 9 - Zihan Zheng, Xi Zhang, Chen Wang, Ning Xia, Yantong Xiang.

BUSINESS DATA MINING

Module Leader: Dr. Nicholas P. Danks

Trinity Business School
TRINITY COLLEGE
UNIVERSITY OF DUBLIN

December 2024

Declaration

I declare that this work has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

Generative AI Declaration

Please choose A or B with regards to your use of ChatGPT & other generative AI tools in this project:

- ■ **A. Nothing to declare.** I did not use ChatGPT or any other generative AI software. (see note)
- ☐ **B. I used ChatGPT** or other generative AI software (see note)

NOTE:

- If you answer A and the corrector/supervisor finds evidence that you have indeed used ChatGPT, this behaviour will be considered as unethical and you will be penalized accordingly with reference to the TCD policy on plagiarism.
- If you answer B, please clearly explain for which chapters or parts of your dissertation you used ChatGPT and how it helped you to improve your learning process within ethical guidelines. You may include your answer – 300 to 600 words approx.- in the appendix.

Signed: Zihan Zheng (print name) ID No. 20308073

Signed: Xi Zhang (print name) ID No. 24337502

Signed: Chen Wang (print name) ID No. 24347542

Signed: Ning Xia (print name) ID No. 24342569

Signed: Yantong Xiang (print name) ID No. 23339874

Date: 13 December 2024

Table of Contents

Executive Summary

This project seeks to enhance demand forecasting and supply chain management by addressing critical factors that influence logistics performance. By employing Random Forest and ARIMA models, the analysis identifies key determinants of order delays and forecasts future demand trends. These insights facilitate proactive strategies aimed at minimising delays, aligning inventory with seasonal demand fluctuations, and improving overall logistics efficiency. The proposed recommendations are designed to support cost reduction initiatives, enhance customer satisfaction, and foster operational resilience.

The complete code and process has been uploaded to GitHub.

Problem description

- **Business goal**

Throughout the research, it was identified that the ability to accurately forecast demand served as a critical factor in the effectiveness of logistics and supply chain management(Real et al, 2008). An accurate demand forecasting system would enhance the organisation's ability to manage supply chain risk and support the maintenance of a high level of operational resilience(Seyedan and Mafakheri, 2020). The primary goal for this project is to optimise its demand forecasting capability along with its production scheduling, inventory, and aggregate planning, and the strategies employed involving identify key factors that impact the ability of its logistic supply chain and optimising a more accurate demand forecasting model through the examination and evaluation of those factors

In this context, the main stakeholders can be defined as customers, supply chain partners, and customers, all of whom are engaged throughout the entire supply chain process. Through accurate demand forecasting and effective supply chain management, they can reduce costs. Additionally, the organisation can identify and mitigate factors that may hinder long-term stability and collaborative efforts, ensuring a more resilient and efficient supply chain(Gunasekaran, 2004).

The main challenges are the associated costs that occurred during the customer group identification, and developing tailored strategies to effectively target these customers. From the short-term view, higher costs would greatly affect the financial situation of the organisation which becomes a key risk factor in implementing a complex prediction model into its logistic system(Abolghasemi et al., 2020). Meanwhile, the identification process includes examining the importance of each factor and the impact it has made on the whole supply chain can be changed On the other hand, the opportunities could be seen as more availabilities for the organisation to not only optimise its supply chain but especially improve the ordering process to minimise delays, More importantly, these opportunities can drive the organisation's digital transformation efforts.

- **Data mining goal**

Our data mining goals are to optimise supply chain management by reducing order delays and enhancing supply chain capacity. We will be focusing on two main areas in order to improve logistics efficiency and meet future demand.

The first area that this project will focus on is analysing the **factors that will result in an order delay.** We decide to make use of the Random Forest model to solve this binary classification problem by predicting whether orders will delay or not. We believe that on-time delivery is crucial in gaining customer satisfaction and maintaining competitive advantage, as delays affect customer experience

and may lead to revenue loss and brand damage. By analysing historical data with "delayed" or "on-time" labels, we can identify key delay factors. This interpretability analysis will reveal specific variables affecting delays, helping optimise order processing and improve on-time delivery rates. Additionally, the Random Forest model's predictive capabilities will help identify high-risk orders early, allowing preventive measures to further enhance logistics efficiency.

The second focus of this project is to determine **how we predict the future order demand in order to optimise supply chain capacity** which involves time series forecasting. We will use an ARIMA model to analyse historical order data, we can capture the trends and seasonality to provide more accurate demand forecasts. This supervised prediction will help businesses allocate resources efficiently, prevent inventory surplus or shortage, and improve supply chain flexibility and responsiveness to better meet market demands.

Data description

1. About the data

The dataset used in this project was obtained from Kaggle, consisting of 15,549 transaction records collected from the logistics supply chain between January 2015 and December 2017. It includes 41 attributes that provide insights into customer, order, and logistics information (*see Appendix 1*).

2. Data preparation

The data in the dataset is relatively complete with no missing values. Only the customer_state field has an outlier, which has little effect on the entire dataset. After discussion, our team selected 11 key features (*see Appendix 2*).

3. Processing process

- Duplicate records were eliminated using the **unique** function.
- State code '91732' represents an outlier in the dataset, therefore, we filtered out the irrelevant records to ensure the analysis remains focused and unbiased.

4. Label transformation

- Values -1 and 0 were set to 0, representing "Not Delayed."
- Value 1 was labeled as "Delayed."

Data mining solutions

1. Methods

1.1 Random Forest Model

For the first data mining problem - predicting order delay - the Random Forest classification algorithm is adopted, which is implemented by the randomForest package in R. Random forests can effectively process multidimensional feature data, reduce overfitting, and are less susceptible to outliers. The dataset was divided into 40% for training, 30% for validation, and 30% for testing to balance training and evaluation. The code can be found in *Data_Mining_Problem_1_Code.Rmd* file.

In this project, the random forest model was chosen for its performance across evaluation metrics*(Appendix 3)* such as accuracy, recall and F1 scores, compared to the decision tree and logistic

regression models. It offers high interpretability and faster runtime when implemented in R, making it well-suited for addressing the problem at hand. The implementation code can be found in *Model_Comparison.Rmd* file.

1.2 ARIMA model

For the second data mining problem - predicting future demand - the ARIMA model is used to predict the order time series. The ARIMA model is highly interpretable and provides a good understanding of how historical data affects current forecasts. The code can be found in *Data_Mining_Problem_2_Code.Rmd* file.

2. Evaluations

For the first problem, since this is a supervised classification task involving multiple factors, we used a Random Forest model to analyse it. The confusion matrix (*Appendix 3*) shows that the model performs moderately in identifying the "Delayed" class compared to the "Not Delayed" class, with a relatively high misclassification rate. The overall accuracy is 66.45% on the validation set and 65.64% on the test set, indicating moderate predictive ability. Lastly, based on the AUC values and ROC curves (*Appendix 4*), it is clear that the model has a moderate ability to distinguish between "Delayed" and "Not Delayed" classes. Overall, the model shows a reasonable performance but leaves room for improvement, especially in accurately identifying "Delayed" cases.

For the second problem, we use the time series ARIMA model to predict future order demand. First, the significance of the coefficients shown by low standard errors, indicates that these terms are important for fitting the model. The negative drift coefficient suggests a downward trend over time. For the model's error metrics (*Appendix 6*) , in relation to the scale of the dataset, the reported error values are comparatively small when viewed against typical order quantities or demand levels. Given that the order volumes likely fall within the range of hundreds or thousands, deviations of RMSE(101.6) or MAE(69.7) constitute a relatively minor proportion of the total values, reflecting a high level of predictive accuracy.

3. Practical implication

For the first problem, based on the variable importance plot (*Appendix 9*), Mean Decrease Accuracy(MDA) and Mean Decrease Gini(MDG), some important features are shown.

Firstly, ‘shipping_mode’ and ‘order_item_total_amount’ are highly ranked in both metrics, indicating they are consistently important for the model, which suggests that these two variables play a crucial role in predicting delays. Secondly, ‘order_region’ , ‘customer_state’ and ‘category_name’also demonstrate notable importance in the context of the model, particularly when examined through the MDG metric. The subsequent analysis will focus on the evaluation of these five variables in greater detail.

3.1 shipping_mode

According to the *Appendix 10*, First Class has the highest delay rate (98.45%), with a total 2390 orders, of which 2353 orders are delayed, while Standard Class has the lowest delay rate (40.58%), with a total 9116 orders, of which only 3699 orders are delayed. This difference may reflect the timeliness and service level of different delivery modes. Even though the First Class is expected to be faster, its highest delay rate indicates that there may be efficiency issues in practical operation.

3.2 order_item_total_amount

Regarding the variable order_item_total_amount, its significance in predicting delays is emphasized. However, this interpretation necessitates a nuanced understanding. The bar chart (*Appendix 11*) indicates that the model attributes substantial weight to this variable primarily due to two high-value orders: one with a total amount of €1939.99 that did not experience delays, and another with a total amount of €1505.51 that was delayed. The contrasting delay statuses of these high-value orders contribute to the variable's importance in the model.

When these two outlier orders are excluded, the analysis reveals that the total order amount has minimal correlation with delay rates across other transactions. Nevertheless, it is essential to recognise that these high-value orders represent genuine and meaningful cases. Their inclusion reflects actual scenarios that the model must account for when making predictions. Consequently, instead of dismissing these values as anomalies, they should be considered valid instances that can influence delay predictions, especially in situations involving exceptionally large transactions. This perspective underscores the necessity of balancing general trends with the potential impact of such exceptional cases in the dataset.

3.3 order_region

The chart (*Appendix 12*) indicates that orders destined for North Africa have the highest delay rate at 63.5%. In contrast, delay rates for orders to other destinations show less variation, remaining within the 58% to 60% range. This suggests a distinct regional disparity in delay rates, with North Africa experiencing significantly higher delays compared to other regions.

3.4 customer_state

The chart (*Appendix 13*) reveals that orders from sellers based in the states of Iowa (IA), Minnesota (MN), and West Virginia (WV) have the highest delay rates, at 75.0%, 72.3%, and 70.4%, respectively. These states warrant particular attention due to their notably elevated delay rates, indicating potential regional challenges impacting order timeliness.

3.5 Category_name

The analysis (*Appendix 14*) shows that product categories with the highest delay rates are concentrated in sporting goods, specifically in items related to strength training, men's golf clubs, soccer, lacrosse, and golf shoes, with delay rates of 83.3%, 83.3%, 79.3%, 75.7%, and 73.1%, respectively. Sellers in the sports equipment sector should pay particular attention to these categories, as the elevated delay rates suggest potential issues that may impact customer satisfaction and operational efficiency.

For the second problem, the line chart (*Appendix 15*) sees the forecasted product volume for 2018. The projection indicates a peak in products from June to August, suggesting that sellers should increase inventory and ensure adequate logistics resources during this period. In contrast, the forecasted product volume from October to December shows a downturn, indicating that sellers should consider reducing inventory levels during these months to optimize storage and reduce potential overstock costs.

Recommendations and Limitations

Based on the modelling analysis, we propose the following recommendations:

For customers, to mitigate high delay rates in First and Second Class shipping, we recommend that customers with urgent needs should opt for Same Day service, while those with less time-sensitive orders should choose Standard Class for cost efficiency.

For companies, there are some measurements that are feasible.

- For delays in North Africa and among sellers in Iowa, Minnesota, and West Virginia, strategies such as partnering with local providers, adopting multimodal transport, establishing nearby warehouses, and expediting customs processes are advised to companies.
- Companies should align inventory levels with seasonal demand, increasing stock from June to August and reducing it from October to December to control costs.
- High-delay items, such as sporting goods, require close monitoring. Collaborating with suppliers to maintain adequate stock and ensure smooth processing is essential for companies.

This work provides a foundation for data-driven supply chain improvements with reasonable model performance, actionable insights for optimisation, and scalability. However, it acknowledges limitations such as moderate accuracy, reliance on historical data, and sensitivity to outliers, which may impact the generalisability of the findings. Furthermore, ethical considerations such as ensuring data privacy, mitigating bias in models, and addressing inequities in resource allocation require careful attention.

Overall, this work lays a solid foundation for improving supply chain operations while emphasizing the need for ongoing refinement and ethical oversight.

References

Abolghasemi, M., Beh, E., Tarr, G. and Gerlach, R. (2020) 'Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion', *Computers & Industrial Engineering*, 142, Article 106380. Available at: https://www.sciencedirect.com/science/article/pii/S0360835220301145 (Accessed: November 2024).

Carbonneau, R., Laframboise, K. and Vahidov, R. (2008) 'Application of machine learning techniques for supply chain demand forecasting', *European Journal of Operational Research*, 184(3), pp. 1140–1154. Available at: https://www.sciencedirect.com/science/article/pii/S0377221706012057 (Accessed: November 2024).

Gunasekaran, A. (2004) . Supply chain management: Theory and applications. European Journal of Operational Research 159 (2), 265–268.

Kwarteng, S.B. and Andreevich, P.A. (2024) 'Comparative Analysis of ARIMA, SARIMA and Prophet Model in Forecasting ', *Research & Development*, 5(4), pp. 110-120. Available at: https://doi.org/10.11648/j.rd.20240504.13 (Accessed: November 2024).

Seyedan, M. and Mafakheri, F. (2020) 'Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities', *Journal of Big Data*, 7(1), Article 52. Available at: https://link.springer.com/article/10.1186/s40537-020-00329-2 (Accessed: November 2024).

Appendices

Appendix 1: Screenshot of the Parts Dataset

shipping_mode	order_region	order_date	category_name	order_item_total_amount	customer_segment	customer_state	department_name	order_status	payment_type	label
Standard Class	Western Europe	2015-08-12 00:00:00+01:00	Cardio Equipment	84.99157	Consumer	PR	Footwear	COMPLETE	DEBIT	-1
Standard Class	South America	2017-02-10 00:00:00+00:00	Water Sports	181.99	Consumer	CA	Fan Shop	PENDING	TRANSFER	-1
Second Class	Western Europe	2015-01-01 00:00:00+00:00	Indoor/Outdoor Games	93.81015	Consumer	PR	Fan Shop	COMPLETE	DEBIT	1
Second Class	Central America	2017-05-31 00:00:00+01:00	Cleats	99.8906	Consumer	PR	Apparel	PROCESSING	TRANSFER	0
Standard Class	Central America	2015-03-28 00:00:00+00:00	Water Sports	171.07587	Consumer	CA	Fan Shop	COMPLETE	DEBIT	1
Standard Class	East of USA	2016-06-06 00:00:00+01:00	Electronics	145.46329	Consumer	PR	Footwear	CLOSED	CASH	1
Standard Class	West of USA	2016-05-17 00:00:00+01:00	Indoor/Outdoor Games	167.99	Corporate	PR	Fan Shop	COMPLETE	DEBIT	1
Standard Class	East of USA	2016-06-09 00:00:00+01:00	Men's Footwear	116.99	Home Office	PR	Apparel	PROCESSING	TRANSFER	-1
Standard Class	Southeast Asia	2016-06-06 00:00:00+01:00	Men's Footwear	113.15623	Consumer	KY	Apparel	ON_HOLD	DEBIT	1
First Class	Central America	2017-08-29 00:00:00+01:00	Shop By Sport	127.39	Corporate	PR	Golf	PENDING_PAYN	PAYMENT	1
Standard Class	Western Europe	2017-08-29 00:00:00+01:00	Men's Footwear	111.575935	Corporate	PR	Apparel	ON_HOLD	DEBIT	1

Dataset Access: [*Logistics Supply chain real world data*]

Appendix 2: Key Features and Explanations

- shipping_mode: Different modes of transportation(categorical, 4 unique categories).
- order_region: Order delivery area (categorical, 24 unique categories).
- order_date: Date on which the order is made (date format)
- category_name: Description of the product category (categorical).
- order_item_total_amount: Total amount per order (numerical).
- customer_segment: Types of Customers Consumer (categorical, 3 unique categories).
- customer_state: State to which the store where the purchase is registered belongs (categorical).
- department_name: Department name of store (categorical).
- order_status: Order Status (categorical, 9 unique categories).
- payment_type: Type of transaction made (categorical, 4 unique categories).
- label: delivery outcomes: -1 early arrival, 0 on time, 1 delayed (categorical).

Appendix 3: Random Forest Model Performance Metrics

Validation

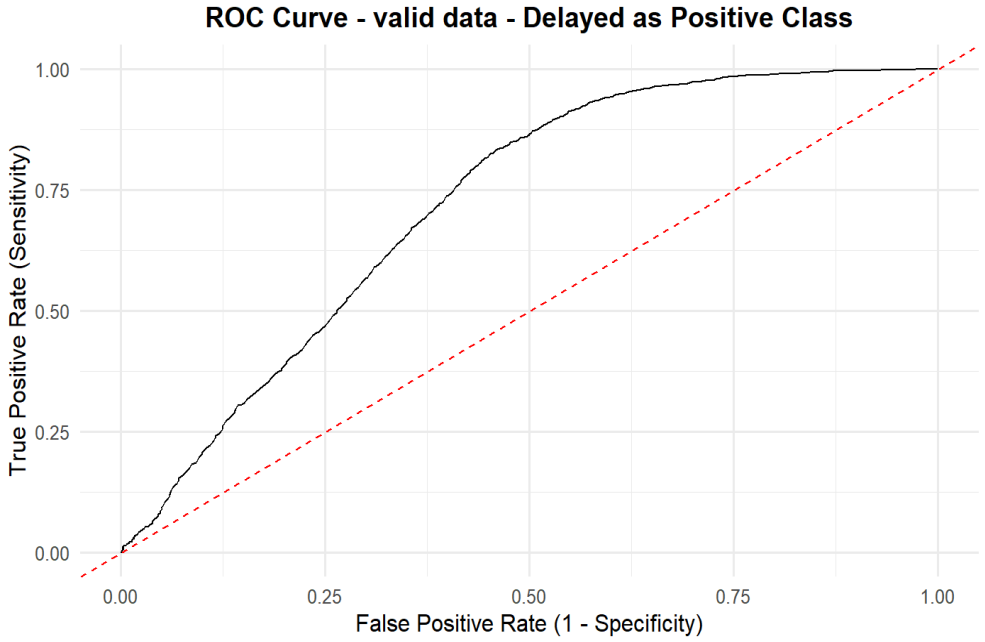
Prediction	Reference: Delayed	Reference: Not Delayed
Delayed	1531	403
Not Delayed	1162	1569

Test

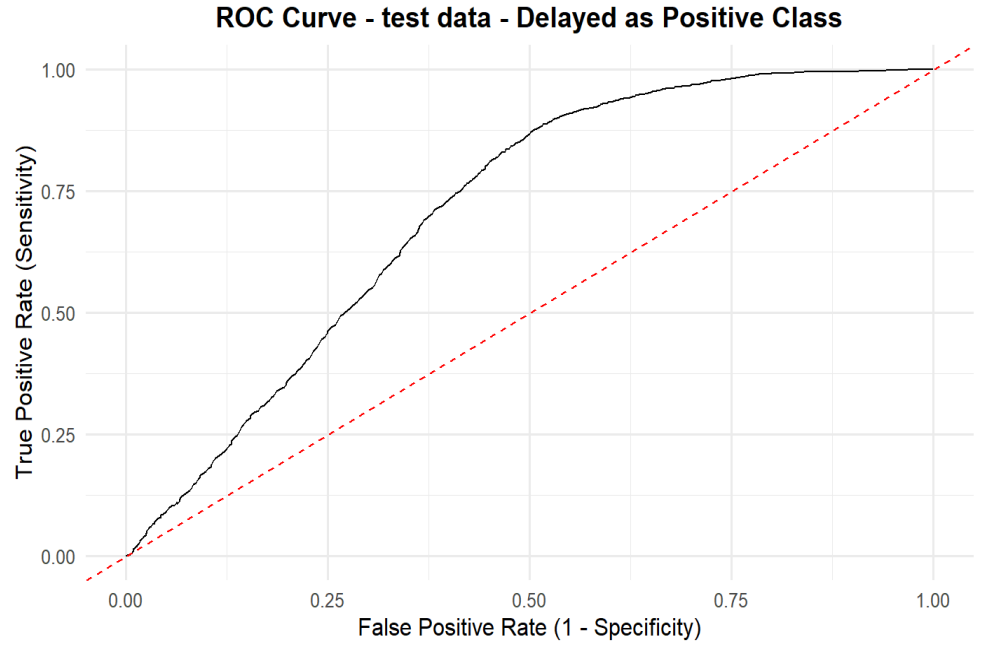
Prediction	Reference: Delayed	Reference: Not Delayed
Delayed	1529	439
Not Delayed	1163	1532

Set	Accuracy	Precision	Recall	F1 Score	AUC
Validation	0.6645	0.7916	0.5685	0.6618	0.7197
Test	0.6564	0.7769	0.5680	0.6562	0.7103

Appendix 4: ROC Curve for Validation Data (Delayed as Positive Class)



Appendix 5: ROC Curve for Test Data (Delayed as Positive Class)



Appendix 6: ARIMA Model Training Set Error Measures

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Value	1.968449	101.5968	69.70962	-4.29687	12.39157	0.448951	0.06160744

Appendix 7: ARIMA Model Coefficient Data

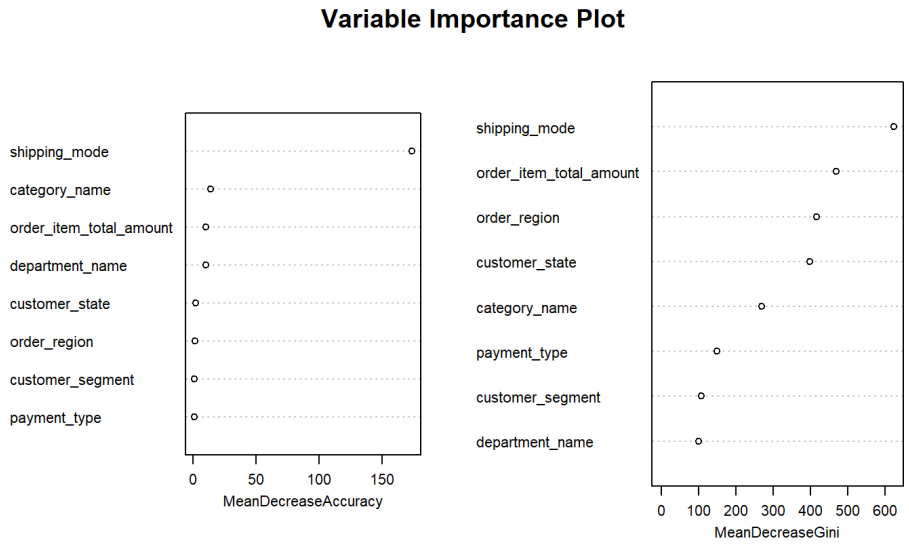
Metric	Value	s.e.
mal	0.3923	0.1623
sar1	-0.5433	0.2074
drift	-8.8379	2.3312

Appendix 8: Logistic regression Model and Decision tree Performance Metrics

Logistic Regression	Accuracy	Precision	Recall	F1 Score	AUC
Validation	0.3262	0.4183	0.4278	0.4229	0.7255
Test	0.3159	0.4097	0.4198	0.4147	0.7339

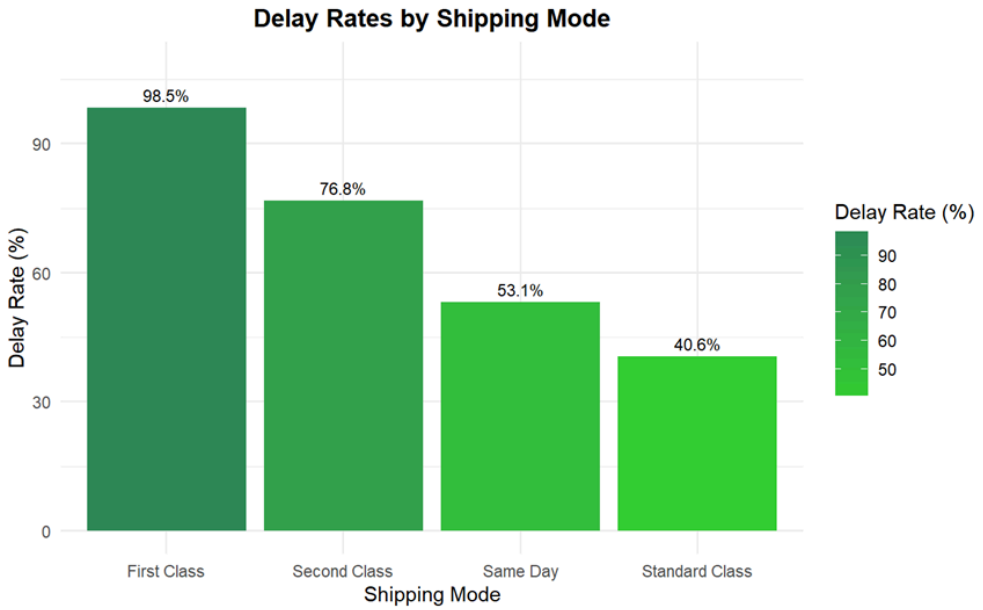
Decision Tree	Accuracy	Precision	Recall	F1 Score	AUC
Validation	0.6888	0.8593	0.5512	0.6716	0.7140
Test	0.6849	0.8611	0.5416	0.6649	0.7111

Appendix 9: Variable Importance Plot

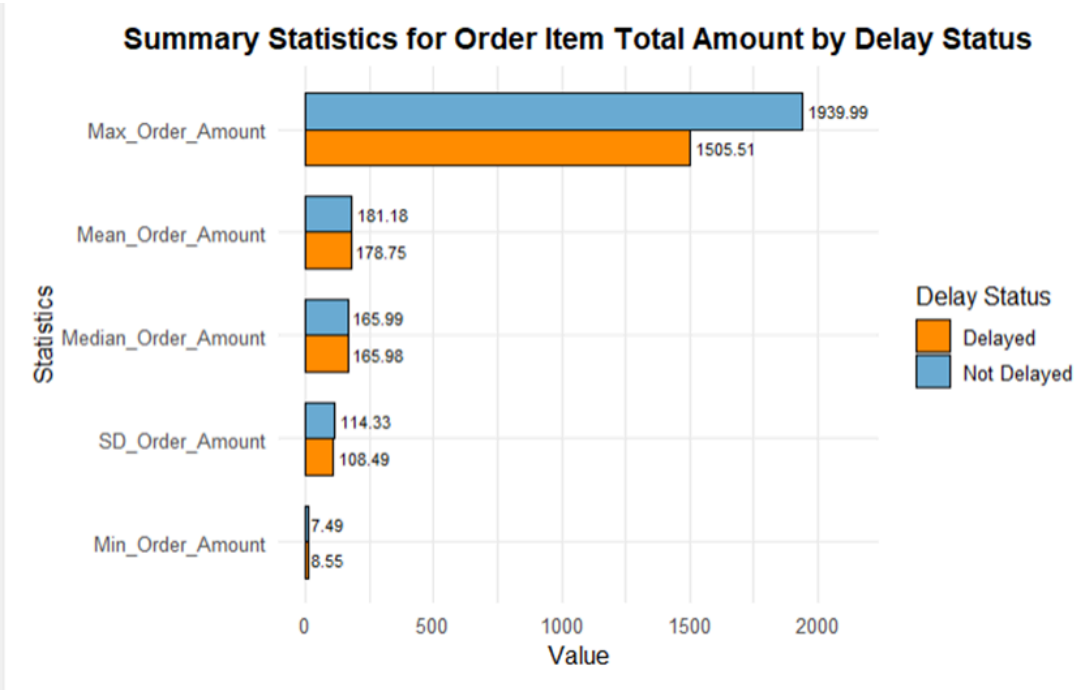


Appendix 10: Figures of shipping_mode

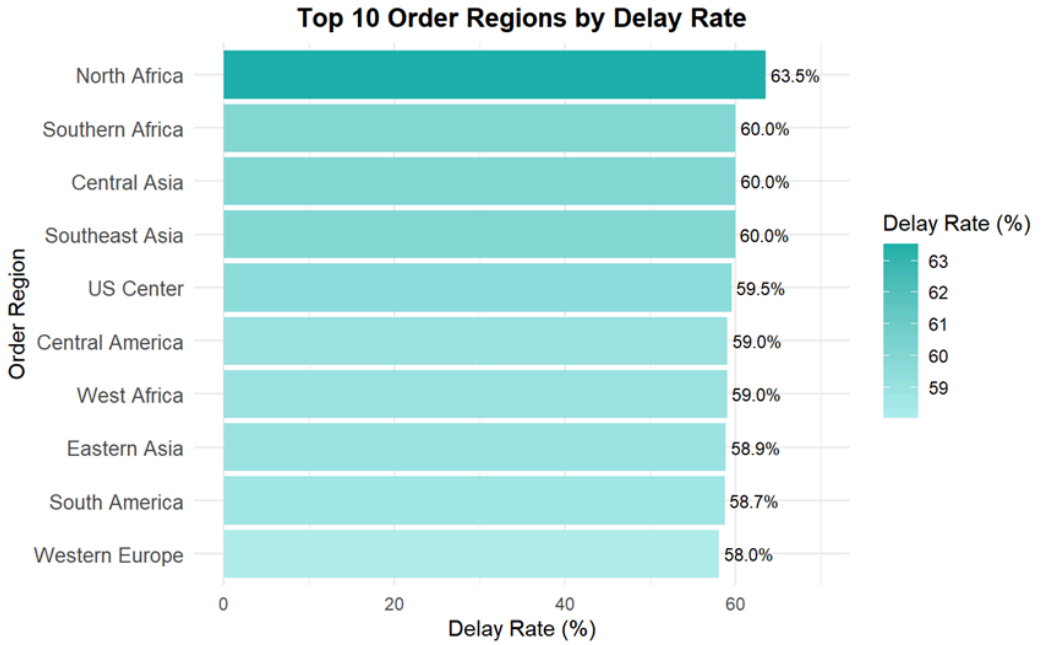
shipping_mode <fctr>	Total_Orders <int>	Delayed_Orders <int>	Delay_Rate <dbl>
First Class	2390	2353	98.45188
Same Day	759	403	53.09618
Second Class	3283	2520	76.75906
Standard Class	9116	3699	40.57701



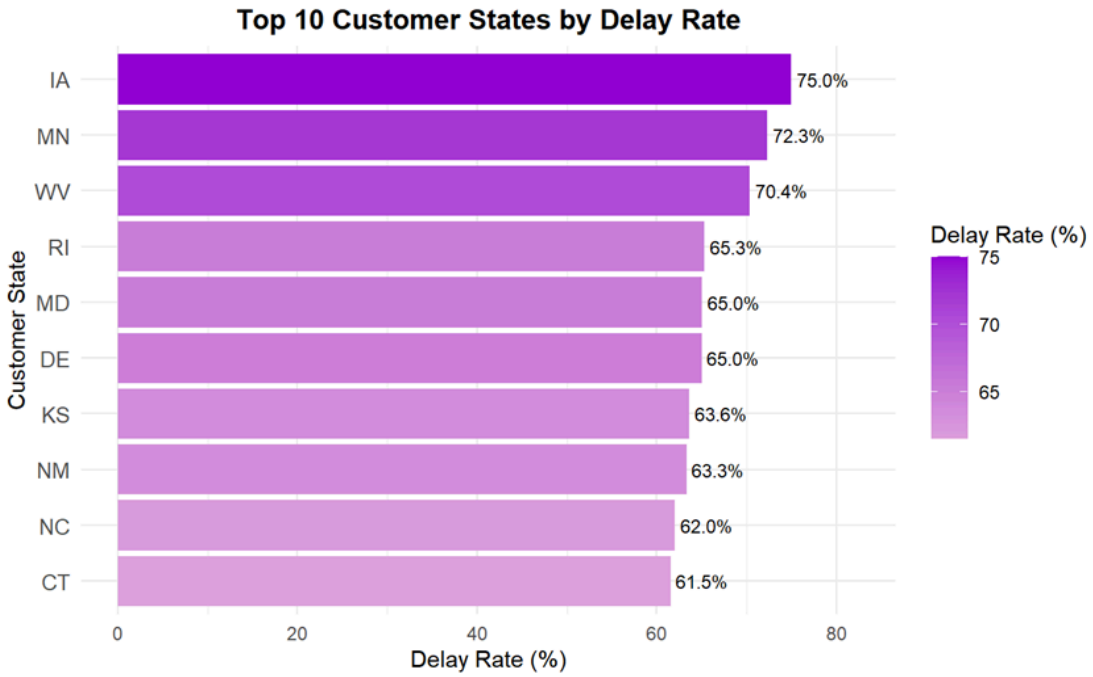
Appendix 11: Figure of order_item_total_amount



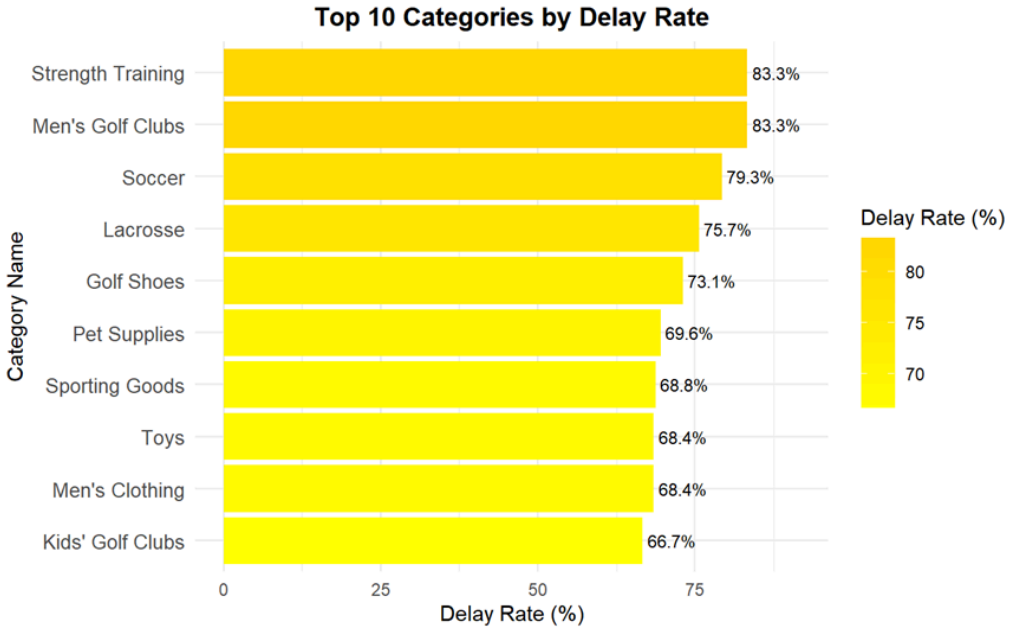
Appendix 12: Figure of order_region



Appendix 13: Figure of customer_state



Appendix 14: Figure of Category_name



Appendix 15: Figure of Comparison of Actual and Forecasted Monthly Product Totals