Recurrent Neural Network for Predicting Sequential Supply Chain Delays

by

Anirudh Narula
Bachelor of Arts in Economics, University of California, Berkeley (2018)

and

Yu-Hsin Lin
Bachelor of Business Administration in Business Administration, National Chengchi University (2018)

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Signature of Author: _____
Department of Supply Chain Management
May 10, 2024

Signature of Author: _____
Department of Supply Chain Management
May 10, 2024

Certified by: _____
Tim Russell
Program Engineer, Center for Transportation and Logistics
Capstone Advisor

Accepted by: _____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Recurrent Neural Network for Predicting Sequential Supply Chain Delays

by

Anirudh Narula

and

Yu-Hsin Lin

Submitted to the Program in Supply Chain Management
on May 10, 2024 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

This capstone project addressed the challenges faced by GlaxoSmithKline's (GSK) supply chain in managing sequential delays, essential for ensuring timely delivery in the pharmaceutical industry. The key objectives included pinpointing planned dates within GSK's system and developing a robust machine learning model to predict sequential delays accurately. Through literature review and methodology development, the project focused on utilizing neural network machine learning methods, specifically recurrent neural networks (RNNs). Meanwhile, detailed summary statistics showcasing the delay frequencies and locations within GSK operations for the Benlysta brand revealed that approximately 40% of process orders exhibited delay issues, primarily in primary manufacturing sites. Further examination highlighted specific areas prone to delays, providing GSK with managerial insights for targeted action. The model development involved data acquisition, manipulation, preprocessing, and RNN and explanatory model construction, followed by hyperparameter tuning to optimize performance, resulting in a reduced mean absolute error (MAE) of 4.89 days. Subsequent SHAP value analysis helped identify specific process sequences and the assembly phase as key drivers of delays, enabling GSK to develop strategies and mitigate supply chain risks. Although challenges in linking manufacturing and quality data limited the initial scope, the project provided valuable insights and laid a solid foundation for future enhancements. Leveraging the findings and insights gained from this capstone project, GSK can enhance operational efficiency, mitigate supply chain risks and deliver medications to patients more effectively.

Capstone Advisor: Tim Russell
Title: Program Engineer, Center for Transportation and Logistics

Table of Contents

1 Introduction 6
1.1 Motivation 6
1.2 Problem Description 6
1.3 Objectives 7
1.4 Approach 7
1.5 Benefits 7
2 State of the Practice 8
2.1 The Pharmaceutical Supply Chain 8
2.2 Manufacturing Sequential Delays: Implications for Pharmaceutical Industry 9
2.3 Machine Learning Methods: Neural Networks 10
2.3.1 Feedforward Neural Networks (FNN) 12
2.3.2 Recurrent Neural Networks (RNN) 12
2.3.3 Long Short-Term Memory (LSTM) Networks 12
3 Methodology 14
3.1 Analyzing Summary Statistics 14
3.2 Establishing Plan Rates and Tracking in GSK's System 14
3.3 Development of Machine Learning Models: RNNs and LSTM 14
3.4 Validation of Model Effectiveness 15
4 Summary Statistics 17
4.1 Overall Time Delay 17
4.2 Delay by Primary and Secondary Manufacturing 18
4.3 Delay by Site Locations 19
4.4 Delay by Process Steps 19
4.5 Delay by Workcenter Types 21
5 The Recurrent Neural Network (RNN) Model 23
5.1 Data Acquisition 23
5.2 Data Manipulation 23
5.3 RNN Model Preprocessing 24
5.4 RNN Model Construction 26
5.5 Explanatory Model Construction 28
6 Results 29
6.1 Hyperparameter Tuning 29

6.2 RNN Model Performance Evaluation 29
6.3 Explanatory Model Results 31
7 Recommendations 34
8 Conclusion 36
References 38

List of Figures

Figure 1 Neural Network Layers 11
Figure 2 RNN Simple Cell versus LSTM Cell 13
Figure 3 Percentage of Different Deviation Types 18
Figure 4 Distribution of Delays by Primary and Secondary Manufacturing (with and without Outliers) 18
Figure 5 Distribution of Delays by Site Locations (with and without Outliers) 19
Figure 6 Distribution of Delays by Process Steps (with and without Outliers) 20
Figure 7 Distribution of Delays by Process Steps at Each Site Location (without Outliers) 20
Figure 8 Distribution of Delays by Workcenter Types (with and without Outliers) 21
Figure 9 Distribution of Delays by Workcenter Types at Each Site Location (without Outliers) 22
Figure 10 Distribution of Step Count Frequency for Benlysta 25
Figure 11 Model Loss (MSE) Evolution 30
Figure 12 Distribution of Prediction Errors (MAE) 31
Figure 13 SHAP Summary Plot for All Features 32
Figure 14 SHAP Force Plot for All Features 32
Figure 15 SHAP Values for Assembly Workcenter Type 33

1 Introduction

1.1 Motivation

GlaxoSmithKline (GSK) is an organization deeply committed to the well-being of its end users – the patients. One of the core challenges of supply chain management is the timely delivery of products. Timely delivery is not just about maintaining a business reputation, but more critically, about ensuring that patients get their essential medications when they need them. The pharmaceutical industry has significantly longer lead times than other industries, and these delays can seriously affect the health outcomes of countless individuals. GSK's aspiration for this project emerged from this very concern: to improve the on-time delivery of their products to the patients. This ambition led them to consider the development of an End-to-End (E2E), from active pharmaceutical ingredients (API) site to Local Distribution Centers, early warning system.

1.2 Problem Description

The global supply chain mechanism of GSK is intricate. It is an E2E system spread across multiple global locations, each assigned unique roles and functions. GSK further divides its supply chain into primary and secondary locations. In general, "primary location" indicates the API manufacturing processes, which is the upstream bound in the GSK supply chain, while "secondary location" indicates the stages after API to create finished products for patients. On account of data accessibility, this project focuses primarily on GSK's in-house supply chain silos. When a batch or shipment faces delays at a certain location or when its commencement is postponed, the ripple effects are felt downstream.

The first challenge was to track the planned dates within GSK, which helped identify the specific stages and timings of delay. The second issue arose from the domino effect of supply chain delays. The delay of one batch or shipment can potentially disrupt the timing of multiple subsequent shipments, leading to a compounded delay. As a result, the patient might not receive the necessary medications on time.

Based on preliminary observations and an understanding of the problem, we hypothesized that the solution lay in preempting these delays. The development of an early

warning system flow model could potentially illustrate and predict the cascading effects of a delay on a particular batch or shipment.

1.3 Objectives

The initial objective was to pinpoint the planned data within GSK. Our secondary objective was to furnish GSK with a robust model. This model should be capable of reading system data and subsequently providing insightful predictions on potential compounded delays.

1.4 Approach

Our proposed strategy was comprehensive and involved the following stages:

1. **Data Gathering:** The initial step involved intensive data collection. Our approach was dynamic and remained agile to adapt to challenges and findings.
2. **Data Analysis:** Employed statistical tools to unearth patterns and trends that may be contributing to subsequent delays.
3. **Model Development:** Based on insights from the data analysis phase, we architected a model tailored to read and interpret GSK data, offering accurate predictions on delay implications.
4. **Recommendation Generation:** Rooted in the empirical evidence and insights drawn from the model, we provided data-driven recommendations. The recommendations would guide GSK on the key drivers of delays and allow GSK to take proactive actions.

1.5 Benefits

1. **Proactive Delay Management:** The model serves as a beacon, highlighting potential delay multipliers. Insights from the model will enable GSK's management to realign resources proactively, ensuring on-time delivery.
2. **Strategic Improvements:** Armed with the insights from the model, GSK can embark on both immediate and long-term initiatives to shorten cycle times. These insights will not only assure timely deliveries but also optimize inventory management, supply stability, and demand fulfillment.

2 State of the Practice

In this chapter, we present the landscape of relevant literature on sequential delays within pharmaceutical supply chains. The section covers the overview of the pharmaceutical supply chain, focuses on the sequential delays in manufacturing processes, and delves deeper into the machine learning methods that are suitable to address and alleviate the challenges posed by time delays in the pharmaceutical supply chain.

2.1 The Pharmaceutical Supply Chain

The pharmaceutical industry operates on a complex and finely-tuned supply chain. This chain, crucial for delivering essential medications and vaccines, generated a staggering $1.48 trillion in annual revenue in 2022, as reported by Statista (Mikulic, 2023). Despite its success, the industry faces significant challenges in supply chain management, where inefficiencies can lead to substantial financial losses. According to a report referenced by the Management Center of Europe, based on McKinsey's findings, pharmaceutical companies could save up to $65 billion annually by optimizing their supply chains (Management Centre Europe, 2012).

The supply chain of the Benlysta brand in GSK is a comprehensive five-stage process, each critical to the timely and safe delivery of pharmaceutical products:

1. **Raw Material Procurement:** In this initial stage, the quality of raw materials sets the tone for the entire production process. Any compromise in quality here can lead to significant downstream problems.
2. **Primary Manufacturing:** In this stage, the focus is on producing APIs. Delays or quality problems here can have a domino effect, impacting all subsequent stages.
3. **Secondary Manufacturing:** APIs are converted into consumable forms like tablets or liquids in this stage. Consistency and quality control are paramount, as any deviation can cause delays and compromise product safety.
4. **Multi-Warehouse Distribution Centers:** These centers are the hubs of inventory management and distribution. Inefficiencies or delays here can disrupt the supply chain, leading to shortages or overstocking in various markets (Supply Chain Brain, 2022).

5. **Local Distribution Centers:** The final step involves getting the products to healthcare providers and pharmacies. Timeliness and accuracy in this stage are crucial for meeting the end-user demand.

GSK executives have identified quality assurance at each stage as a potentially significant bottleneck. Quality control is non-negotiable in the pharmaceutical industry, but it requires time and resources. Rigorous testing and validation at each stage, while essential for safety and compliance, can slow down the production and distribution process (Tayyab et al., 2021).

Moreover, delays in the pharmaceutical supply chain often have a cascading effect. If one stage experiences a delay, it can lead to compounded delays in downstream stages (Supply Chain Brain, 2022). For example, an interruption in primary manufacturing not only pushes back secondary manufacturing but also disrupts the scheduling and logistics in distribution centers. These delays can multiply, resulting in a ripple effect throughout the entire supply chain, significantly impacting the availability of medications to end-users.

Optimizing the supply chain in the pharmaceutical industry, therefore, involves a delicate balance between maintaining stringent quality standards and minimizing bottlenecks and delays. Technologies like automation, improved forecasting, and enhanced communication systems can play a crucial role in achieving this balance, thereby ensuring efficient and timely delivery of vital pharmaceutical products.

2.2 Manufacturing Sequential Delays: Implications for Pharmaceutical Industry

Sequential delays have a compounding effect, where disruptions in one stage of the manufacturing process cascade through subsequent stages, magnifying the impact of the initial delay and leading to significant operational disruptions and financial losses (Supply Chain Brain, 2022). These delays in the manufacturing sector, especially in the pharmaceutical industry, significantly impact the on-time and in-full (OTIF) delivery of products.

In the pharmaceutical sector, the repercussions of such delays are profound. They not only affect the company's financial performance but also have critical implications for patient care and public health. For example, delays in the supply chain can lead to shortages of essential medicines, impacting patient treatment plans. In the case of temperature-sensitive drugs, such

as vaccines, delays can result in spoilage, leading to substantial financial losses and a loss of public trust. The COVID-19 pandemic further highlighted the vulnerability of pharmaceutical supply chains to these delays, emphasizing the need for timely delivery of medical supplies in a public health crisis (Supply Chain Brain, 2022).

Developing machine learning (ML) models for early warning systems can help address these challenges. ML algorithms can analyze vast amounts of data to identify patterns and predict potential delays in the supply chain. For example, an ML model could analyze historical data on raw material availability, production timelines, shipping schedules, and external factors like weather or political instability to forecast potential disruptions.

The application of ML in creating predictive models for supply chain management is supported by recent research. According to Kashem et al. (2023), ML algorithms can be used to predict supply chain disruptions, providing valuable insights for decision-makers to proactively address potential issues. Moreover, ML models can enhance real-time monitoring and decision-making. For example, Bodendorf et al. (2023) discussed the use of ML in improving supply chain resilience and responsiveness, particularly in high-stakes industries like pharmaceuticals.

The integration of ML-based early warning systems in the pharmaceutical supply chain can thus lead to significant improvements in forecasting and mitigating delays. By providing advance notice of potential disruptions, these systems enable companies to implement contingency plans, adjust production schedules, and communicate more effectively with stakeholders, thereby maintaining OTIF performance and ensuring the timely delivery of critical healthcare products. Additionally, such systems can play a pivotal role in anticipating and mitigating the effects of sequential delays, ensuring the efficient and reliable delivery of pharmaceutical products, which is essential for patient care and public health.

2.3 **Machine Learning Methods: Neural Networks**

Among various ML models, neural networks aim to emulate the information processing mechanism of human brains. The capability of neural networks to process data through an extensive network of interconnected nodes (Wu & Feng, 2018) positions this machine learning

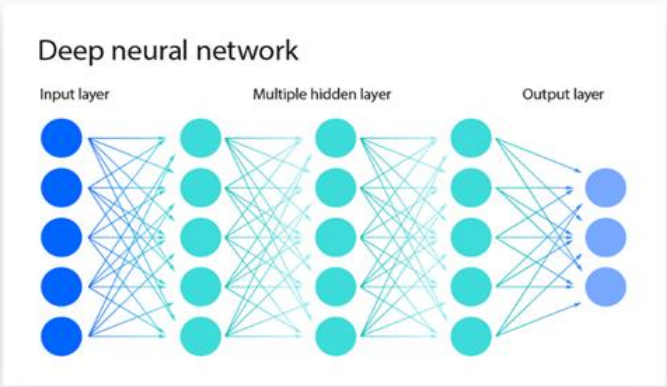
method as well-suited for the intricacies of pharmaceutical supply chains. Within this section, we will delve into the exploration of different types of neural networks.

Pharmaceutical supply chains comprise multiple layers and each layer involves numerous nodes. The characteristics and complexity of these layers closely mirror the structure of neural networks. Figure 1 (IBM, n.d.) shows how a neural network with multiple layers of nodes is further divided into three main sections. Each interconnecting arc between the nodes is assigned a random weight, which is multiplied with the input from the training data and then added with bias to yield the value of each node. This value, subject to an activation function, dictates the information transmitted from the current node to the subsequent one (Great Learning, 2022).

This dynamic empowers neural networks to generate forecasts based on diverse inputs. Given that the project's objective is to establish an early warning signal system for predicting delays within pharmaceutical supply chain, this machine learning approach is worth exploring.

Figure 1

Neural Network Layers



From *What are Neural Networks?, by IBM, n.d.*

This section focuses on three common neural network models – feedforward, recurrent and Long Short-Term Memory (LSTM) – highlighting their shared characteristics to manage multiple layers with diverse nodes. These models stand out for their ability to navigate through complex data streams to predict downstream nodes by leveraging information from upstream sources, which aligns seamlessly with our objective of forecasting sequential delays within pharmaceutical supply chains.

2.3.1 Feedforward Neural Networks (FNN)

FNN is a fixed network as the model does not have loops and the data is only sent one direction (Sharkawy, 2020), which allows easier maintenance and enables faster outputs. Consequently, the simple and straightforward processing of FNN allows its application to face and speech recognition (Great Learning, 2022). However, the downside of FNN lies in the unsuitability for deeper learning, attributed to the fixed weights of each node that do not change.

2.3.2 Recurrent Neural Networks (RNN)

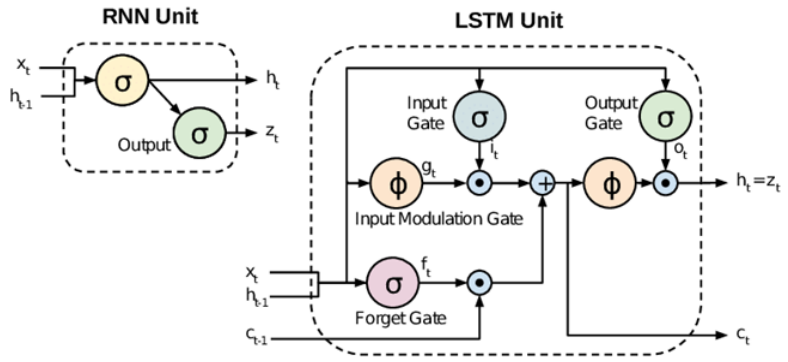
Compared to FNN, RNN offers more flexibility by enabling backpropagation for the model to continuously adjust the weights of each connection over time. Not only does RNN incorporate the value of the preceding node, but RNN also considers all the way back to the beginning of data input. In addition, since RNN restores and relies on historical information to generate more accurate outputs, the model can be used to predict sequential data like translation and text auto suggestion. However, the huge amount of computation can give rise to model explosion issues, amplifying the challenges of model training. An additional challenge associated with RNN is the issue of vanishing gradients. The long-term sequential data could impede the model's ability to effectively capture and retain long-term dependencies, as the gap between current node and the relevant node given previously is so large that the model faces difficulty in establishing robust connections. (Amini, 2023).

2.3.3 Long Short-Term Memory (LSTM) Networks

LSTM, a type of RNN, was introduced by Hochreiter and Schmidhuber (1997) as an enhancement to address the issue of excessive information storage in each node of traditional RNNs. Introducing the concept of "gate cells" within each node (Yu et al., 2019), as illustrated in Figure 2 (Rassem et al., 2017), LSTM incorporates a forget gate mechanism. When new information enters a node, the forget gate is activated, determining whether to retain or discard this information. This innovative addition empowers the model to optimize data utilization, facilitating more effective learning and self-training, particularly in scenarios involving long-term dependencies.

Figure 2

RNN Simple Cell versus LSTM Cell



From *Cross-Country Skiing Gears Classification using Deep Learning*, by Rassem et al., 2017

This chapter established a solid groundwork for the methodology development. As the research unfolded, it became evident that there is a scarcity of prior studies addressing the specific issue of sequential delays in pharmaceutical supply chains, underscoring the significance of this project within the research domain. Additionally, the literature review of the neural network machine learning methods not only provided insights but also led the direction of this project.

3 Methodology

The state of the practice conducted for this project played a crucial role in shaping the methodology by identifying gaps and extracting key insights from prior research. This chapter describes our methodology, including summary statistics analysis, the establishment of plan rates within the system of GSK, the selection of machine learning models, and validation of model effectiveness.

3.1 Analyzing Summary Statistics

The summary statistics analysis was crucial for understanding GSK supply chain dynamics. By examining key metrics such as mean and standard deviation of delays, and the distribution of delay occurrences across different supply chain stages, GSK identified potential inefficiencies and pinpoint areas of risk within its supply chain. The analysis unveiled the frequency and locations of delays within the company, providing a comprehensive overview of the delay issue and valuable insights to inform effective actions moving forward.

3.2 Establishing Plan Rates and Tracking in GSK's System

To enable GSK to achieve timely product delivery, the critical initial step was to accurately determine their plan rates for reliable tracking within GSK’s system. This phase involved a thorough analysis of GSK's historical data on production, quality, and frequency of delays. By understanding the typical rates at which different stages of their supply chain operate, we could identify standard performance benchmarks. These benchmarks served as the foundation for subsequent analysis. We collaborated with GSK to identify key data points within their existing systems that can be leveraged for this purpose. By integrating these data points into our analysis, we could effectively track actual progress against the established plan rates, thereby identifying deviations that may signal potential delays.

3.3 Development of Machine Learning Models: RNNs and LSTM

For the development of a beneficial tool providing early warning signals of sequential time delays, we employed two types of machine learning models: RNNs and LSTM networks. These

models are particularly suitable for GSK's supply chain due to their abilities to process sequential data and recognize patterns over time. The RNN model is adept at handling data where the current state is dependent on the previous state, a common scenario in supply chain operations. This characteristic makes RNNs ideal for predicting potential delays based on historical patterns.

However, RNNs have limitations in capturing long-term dependencies due to the vanishing gradient problem. This is where LSTM models come into play. LSTM, with its unique gating mechanism, can remember information for longer periods, making it more effective in predicting delays that may arise from issues occurring much earlier in the supply chain. This feature is particularly beneficial for GSK's complex, multi-echelon supply chain where early-stage issues can have far-reaching impacts.

3.4 **Validation of Model Effectiveness**

In evaluating the effectiveness of our RNN model, two distinct statistical measures play pivotal roles in validating the goodness of fit: Mean Squared Error (MSE) as the loss function and Mean Absolute Error (MAE) as the error term. MSE measures the average squared difference between the predicted and actual values, providing a comprehensive assessment of the model's overall performance. A lower MSE indicates a better fit between predicted and actual values, signifying higher accuracy. On the other hand, MAE calculates the average absolute difference between predicted and actual values, offering insights into the model's precision in terms of magnitude. While MSE heavily penalizes large errors due to squaring, MAE treats all errors equally, making it particularly useful for assessing outliers.

Besides RNN model evaluation, it's essential to develop an explanatory model that incorporates SHapley Additive exPlanations (SHAP) values for interpreting model outcomes. SHAP values serve as indicators of feature importance within the model. A higher SHAP value indicates a greater positive impact of the feature on the target variable, which helps identify key drivers influencing the predictions.

By rigorously evaluating these models against these statistical measures, we fine-tuned model output to provide reliable and actionable insights for GSK. The model outcome ultimately

aids in the proactive management of their intricate supply chain and ensures timely delivery of essential medications to patients.

4 Summary Statistics

Before embarking on machine learning modeling, we analyzed summary statistics to gain a comprehensive understanding of the frequency and locations of delays within GSK. This analysis provided clarity on the delay problem and will empower GSK to realign its focus and prioritize resources more effectively.

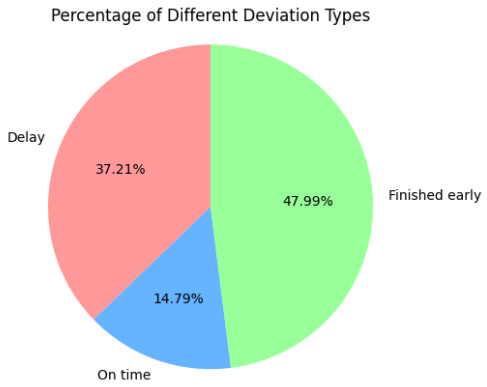
This project's scope centered on the brand Benlysta and encompassed three distinct sites across two manufacturing stages: primary and secondary manufacturing. These sites consisted of one primary manufacturing site (US05) and two secondary manufacturing sites (IT53 and GB57). To ensure data integrity, we removed null entries and utilized process order (PO) numbers as a unique identifier to join manufacturing and quality data. In addition, we defined time delay as the deviation between the actual end date and the planned end date, since (1) a batch that starts late may still get expedited, and (2) simply measuring the duration between actual and planned dates fails to capture the severity of delay. After data manipulation, we performed a multi-layered analysis to extract deeper insights from the dataset. The analysis started by examining delays across all data points before delving into more granular levels, separating primary and secondary sites, site locations, process steps, and workcenter types. Such a multi-layered approach facilitated a thorough understanding of the factors contributing to delays at various operational levels within GSK.

4.1 Overall Time Delay

As shown in Figure 3, the analysis revealed that over 37% of all PO data exhibited delay issues, while the remaining majority either met planned dates or finished ahead of schedule. This finding offered insights for GSK to understand the extent of its delay problem.

Figure 3

Percentage of Different Deviation Types

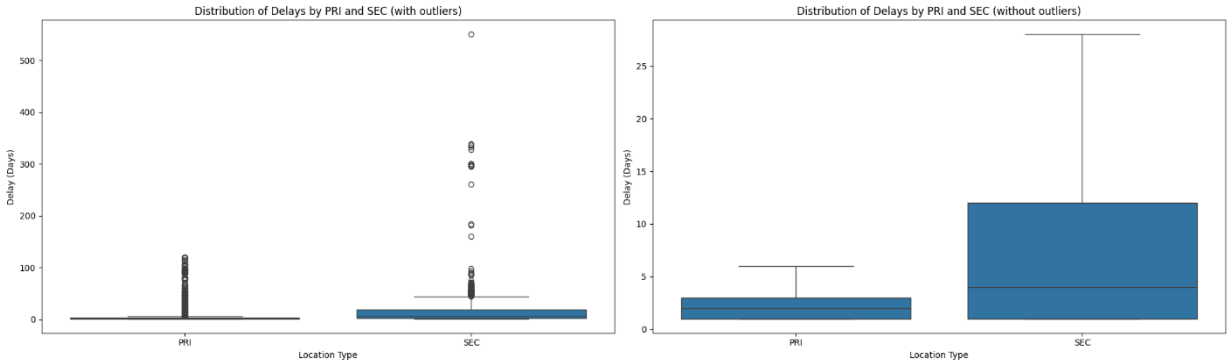


4.2 **Delay by Primary and Secondary Manufacturing**

Upon closer examination of primary and secondary manufacturing, it became apparent that around 90% of all delayed orders originated from the primary manufacturing stage. However, despite this disparity, the average delay at the primary stage was four days, with a standard deviation of 10 days. In contrast, the secondary stage experienced an average delay of 16 days, with a standard deviation of 35 days, indicating a more obvious delay issue at the secondary stage within GSK. The boxplot in Figure 4 illustrates the distribution of delay occurrences across the two different manufacturing stage types. For this project, outliers were excluded based on a criterion of two standard deviations from the mean (Seo, 2006), a method agreed upon with GSK as it retains 95% of the data points.

Figure 4

Distribution of Delays by Primary and Secondary Manufacturing (with and without Outliers)

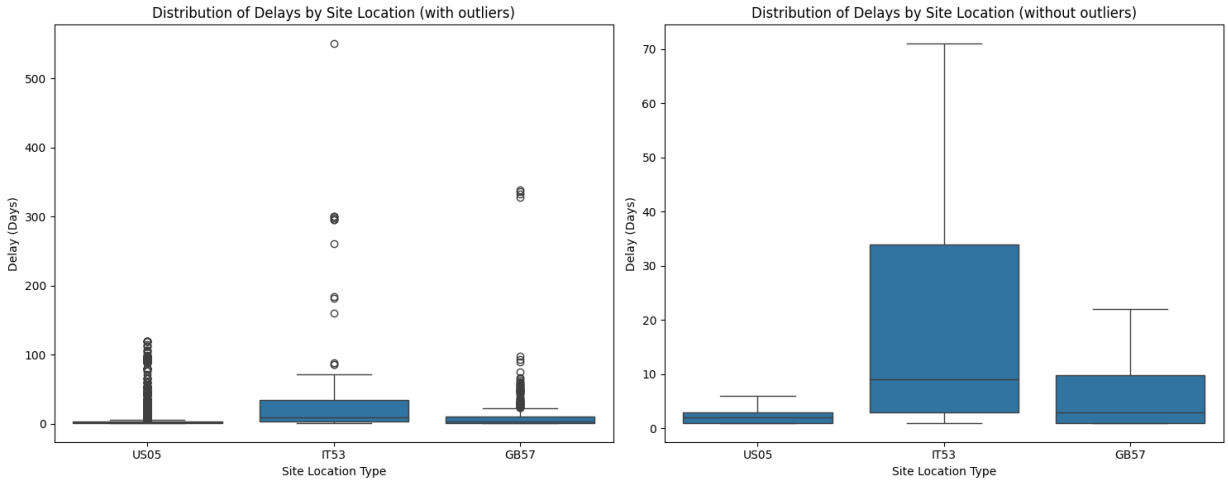


4.3 **Delay by Site Locations**

Continuing with the analysis, we further scrutinized the three site locations, comprising single primary manufacturing site (US05) and two secondary manufacturing sites (IT53 and GB57). Remarkably, nearly 91% of all delay instances were concentrated within US05, but the standard deviation suggested that secondary sites were more profoundly affected by delay. Specifically, the site location with the worst delay issue was IT53, exhibiting an average delay of 22 days and a standard deviation of 41 days. The boxplot in Figure 5 also confirms this observation.

Figure 5

Distribution of Delays by Site Locations (with and without Outliers)

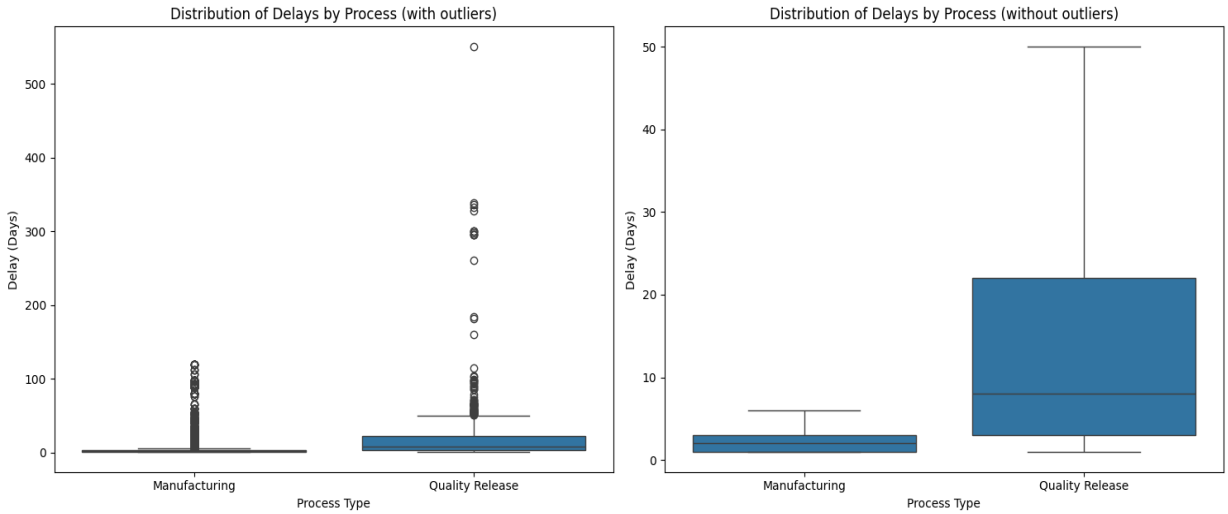


4.4 **Delay by Process Steps**

Given the project's focus solely on primary and secondary stages, the analysis was limited to two process steps: manufacturing and quality release. Among all delayed POs (around 37% out of all POs), 93% occurred during the manufacturing process. However, when considering all POs, it's notable that over half of the quality release processes experienced delay issues, while the manufacturing process accounted for only 36% of delays. Additionally, the quality release process had a longer average delay of 20 days with a standard deviation of 41 days. Figure 6 further underscores the quality release as the primary contributor to delays. The discovery enabled GSK to identify critical processes that were prone to delays within the company.

Figure 6

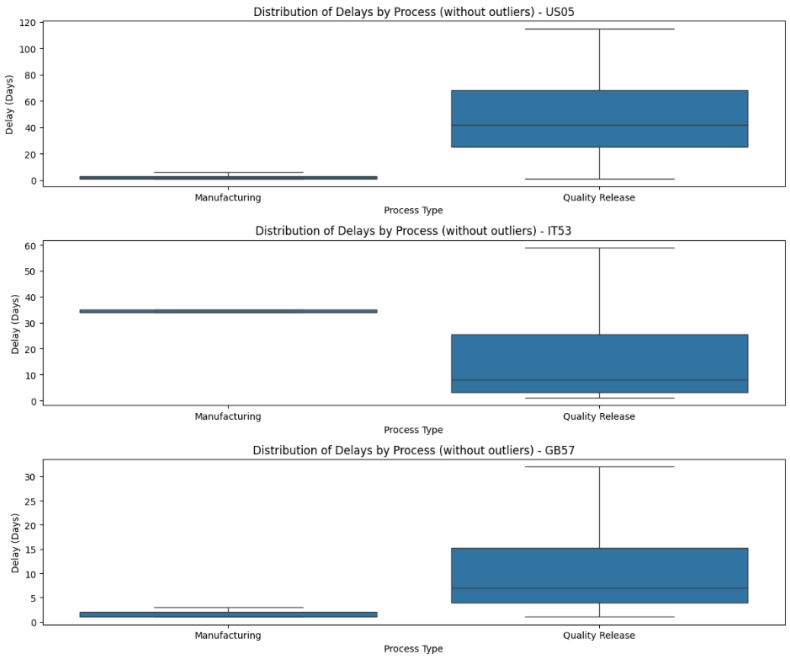
Distribution of Delays by Process Steps (with and without Outliers)



We further analyzed how delays in quality release varied across different site locations. As shown in the boxplot in Figure 7, quality release had longer delays in US05, the primary stage. This discovery served as a pointer for GSK to focus efforts on identifying the root causes of delays within quality release processes at specific locations.

Figure 7

Distribution of Delays by Process Steps at Each Site Location (without Outliers)

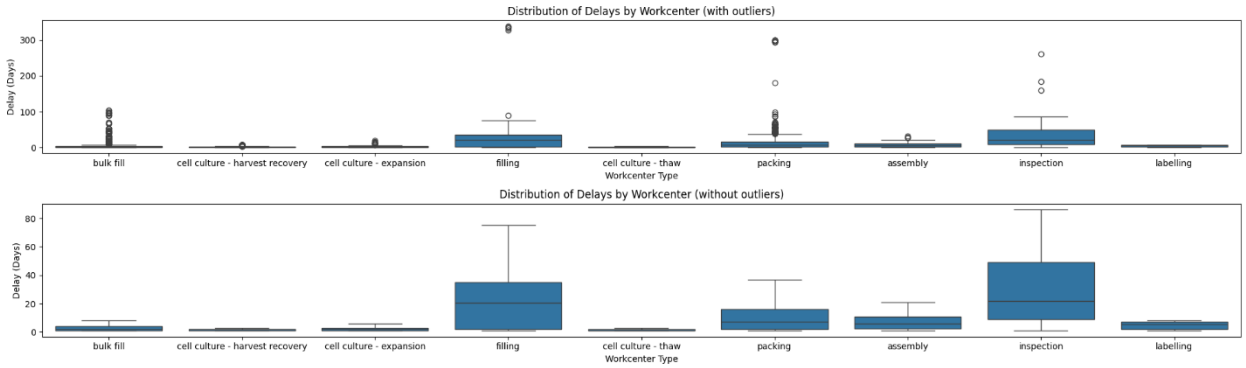


4.5 **Delay by Workcenter Types**

As shown in Figure 8, among the nine different workcenter types within the scope of the project, inspection, filling, and packing emerged as the top three types facing the lengthiest delays, even though they collectively constitute less than 8% of all delayed POs. The analysis revealed that the worst delays took place at the inspection stage, with an average delay of 39 days and a standard deviation of 50 days.

Figure 8

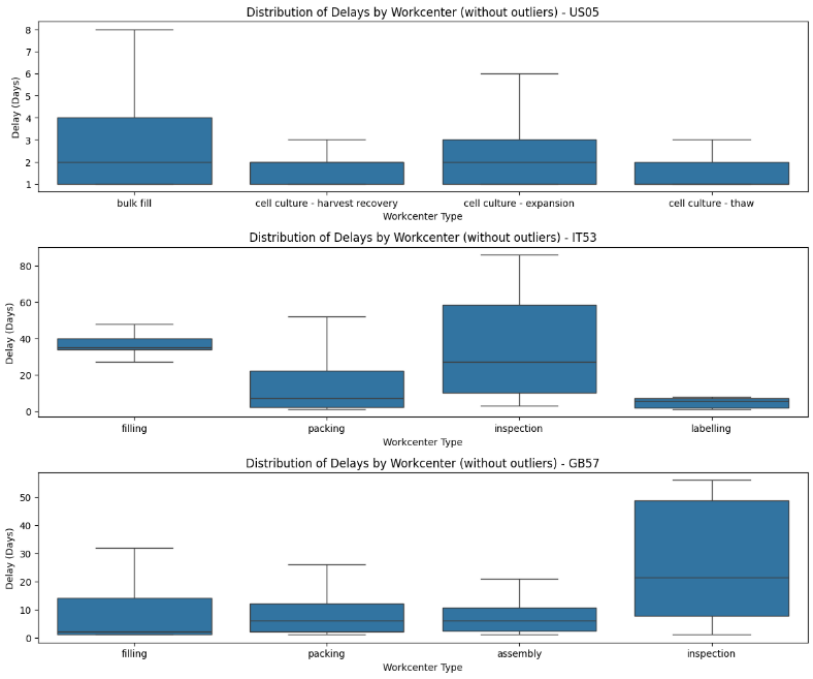
Distribution of Delays by Workcenter Types (with and without Outliers)



Next, we scrutinized delays by workcenter types within each site location. The boxplot in Figure 9 illustrates the specific workcenter types responsible for delays at particular sites. Through this visualization, GSK can swiftly identify bulk fill in US05 and inspection in both IT53 and GB57 as the pivotal workcenter types contributing to delays for the Benlysta brand.

Figure 9

Distribution of Delays by Workcenter Types at Each Site Location (without Outliers)



The analysis of summary statistics furnished GSK with an overview of the delays within their operations, equipping the company with valuable insights to take targeted action in addressing delays more effectively.

5 The Recurrent Neural Network (RNN) Model

The primary goal of the project was to develop a machine learning model capable of predicting sequential delays within GSK supply chain. The creation of the RNN model began with data acquisition, followed by data manipulation, RNN model preprocessing, RNN model construction and ultimately explanatory model construction.

5.1 Data Acquisition

The most crucial dataset required for this project was the planned start and end dates of each step within each process in order to compare them with the actual dates to understand the severity of delays issues within GSK. The planned dates were stored in two distinct datasets: manufacturing data and quality data. Additionally, GSK provided another dataset known as the digital Value Stream Map (dVSM), which offered further granularity, recording more details such as primary or secondary manufacturing, movement types and workcenter types of historical actual goods movements.

The data acquisition phase constituted a significant portion of the project's timeline as the project addressed an unexplored topic within GSK. This project played a pivotal role in facilitating the identification and extraction of the planned data within GSK systems. The planned dataset is now better understood and accessible to GSK.

5.2 Data Manipulation

Given that the required data for the project was dispersed across various datasets, data manipulation served as an essential prerequisite preceding model development. We organized the data, defined project-specific logic, and mapped GSK's supply chain based on the reconfigured data. This process was essential to ensure coherence and accessibility of the data for subsequent model development.

Firstly, as Process Order (PO) number was the key identifier connecting all datasets, we dropped around 36,000 rows of data (41% of all data points) without a PO number to ensure data consistency. Secondly, as each unique PO number may correspond to multiple planned start and

end dates across both manufacturing and quality data, we extracted the minimum scheduled start date and maximum scheduled end date for each PO number, providing a standardized basis for further analysis. Subsequently, to enhance dataset comprehensiveness, we performed an inner join between the dVSM data and the manufacturing and quality data that incorporated the planned dates. Following the data integration, we defined time delay as the deviation between the actual end date and the planned end date to capture the severity of delay within GSK operations.

During the data manipulation phase and through ongoing discussions with GSK, it became evident that the available datasets did not provide a complete mapping of their supply chain. As a result, we leveraged the given datasets to reconstruct a diagraph sequence, showcasing the structure of GSK supply chain specifically for the Benlysta brand.

To establish the path sequence, we concatenated the material number and batch number to create a unique identifier for defining the nodes and edges within the supply chain network. We then identified the end nodes as those lacking outgoing edges and the start nodes as those lacking incoming edge to ensure rationality. Finally, we constructed the path sequence based on the logic that if the previous downstream edge matches the subsequent upstream edge, it forms the same path. This data manipulation phase facilitated the reconstruction of all unique paths across start and end nodes, offering clarity on the structure of GSK's Benlysta supply chain.

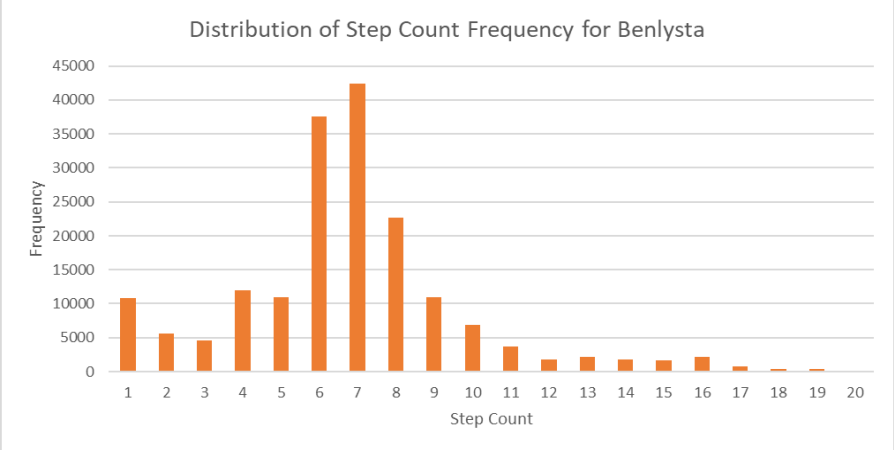
5.3 **RNN Model Preprocessing**

Following the reconstruction of the supply chain network via data manipulation, model preprocessing was indispensable to prepare the data for model construction. This phase began with the creation of a new data frame extracting path-related information, including path ID, step ID, upstream and downstream edges, planned and actual start and end dates, delay days and workcenter types. Each distinct path was labeled with a path ID, while individual steps within each path were labeled with a step ID. Due to self-loop occurrences in certain paths – indicating identical upstream and downstream edges, movement types and site locations in the data – we identified and flagged theses self-loop data rows and assigned their step ID by incrementing the step ID of the preceding non-self-loop step by one to maintain coherence.

The updated data frame allowed us to aggregate and analyze the frequency of each step count within the paths and present its distribution. The analysis unveiled a step count range for the Benlysta brand spanning from 1 to 20 steps. Notably, Figure 10 illustrates that most of the step counts fell within the range of 6 to 8 steps, encompassing nearly 60% of all data points.

Figure 10

Distribution of Step Count Frequency for Benlysta



The next step involved verifying the sequence of planned and actual start and end dates to ensure data rationality. Upon examination, we discovered that 17% of planned dates data and 0.3% of actual dates data exhibited non-conforming date sequences, indicating that the start and end dates of the current step did not align with the previous end date and the subsequent start date. We reported these findings to GSK for further internal investigation. After reaching a consensus to proceed with the analysis using the available data, we recognized the need for data imputation.

For instances of non-conforming planned start and end dates, we opted for simplicity and imputed them with the previous planned end dates, an approach aligned with the company. Similarly, non-conforming actual start dates were imputed with the previous planned end dates for coherence. As for non-conforming actual end dates, we imputed them with the sum of the newly imputed start date and the duration of the previous step, assuming uniform durations between the last step and the current one. Given that actual dates with non-conforming date

sequence issues accounted for less than 0.5% of the dataset, we proceeded under the assumption that imputation would not significantly impact the prediction results.

During the model preprocessing stage, we observed that the manufacturing data and quality data ran parallel to each other. Furthermore, the paths including quality data comprised only one step. Due to the constraints of the timeline and data availability, this discovery required that the project would focus exclusively on predicting sequential delays for manufacturing processes. This phase served as a crucial foundation for a smooth transition to model construction.

5.4 **RNN Model Construction**

The project's objective was to develop an RNN model to predict sequential delays within GSK supply chain. To maintain the integrity of the prediction results and prevent outliers from overly influencing the model, we employed outlier imputation for the delay days. Specifically, we defined upper and lower bounds for the delay days as two standard deviations away from the mean. Outliers exceeding these bounds were imputed with either the upper or lower bound. Subsequently, we filtered out outlier paths with only 1 step or more than 16 steps. One-step paths were excluded since they could not facilitate sequential delay predictions, while paths with more than 16 steps were removed due to their negligible representation, comprising less than 1% of all data points and exerting minimal influence on the model output. Moreover, since the RNN model required uniform step counts across all paths, those with fewer steps than the maximum step count were padded with null values to ensure consistent length for the model. With the data prepared, we conducted a split of 75%-25% (Farias et al., 2020), reserving 75% for training the model and allocating 25% for testing the model performance.

The RNN model adopted a sequence-based approach to predict delays at the final step of a manufacturing process. Inputs of the model comprised sequences of step IDs and workcenter types, representing the product's journey through manufacturing stages.

Initially, a bidirectional LSTM layer was utilized to extract temporal features from the path sequence. This layer processed the sequence in both forward and backward directions, capturing

long-term dependencies between the steps. The output of this bidirectional layer was then passed through a TimeDistributed layer, which applied a dense layer to each time step independently. This layer enabled the model to learn non-linear relationships among features at each step. Finally, an additional LSTM layer was employed to further extract temporal features from the sequence. The output of this layer was again passed through a dense layer with a single neuron to predict delay days at the final step. This RNN model was trained using the Adam optimizer with the mean squared error loss function. Regularization techniques, such as dropout and early stopping, were also implemented to prevent overfitting and enhance the model's generalization capabilities. The following list contains a more detailed explanation of the model's features.

- **Bidirectional LSTM layers:** These layers capture long-term dependencies between steps in the manufacturing process by processing the sequence in both forward and backward directions. This layer is important because the deviation at the final step can be influenced by events that occur earlier in the process.
- **TimeDistributed layer:** This layer allows the model to learn nonlinear relationships among the features at each step. This layer is important because the relationship between the features can vary depending on the specific step in the process.
- **Dropout:** This technique randomly drops out neurons during model training, which helps prevent overfitting.
- **Early stopping:** This technique stops training when the model's performance on a validation set stops improving, which helps prevent overfitting and ensure that the model generalizes well to unseen data.
- **Adam optimizer:** This optimizer has an adaptive learning rate that updates parameter weights during model training.
- **Mean squared error (MSE) loss function:** This function measures the average squared difference between the model's predictions and the actual values, serving as a comprehensive assessment of the model's overall performance. The goal of the model training is to minimize this loss function.

- **Mean Absolute Error (MAE) error term:** This measure calculates the average absolute difference between predicted and actual values, aiding in understanding the model's precision in terms of magnitude.

In summary, the model used a combination of sequence-based processing, nonlinear transformations, and regularization techniques to predict delay days at the final step of a manufacturing process.

5.5 Explanatory Model Construction

For business interpretation, we selected specific features including path ID, step ID, workcenter type, time delay for each path and step, and time delay for the final step of each path. To deepen the analysis, we introduced two new features: step quantity and the combination of workcenter type with step ID. These additions aimed to highlight the most impactful features on the target variable – the time delay of final steps within paths.

Following data preparation, we divided the data frame into 75% training and 25% testing sets and employed Light Gradient Boosting Machine (LightGBM) for evaluation. LightGBM, a type of Gradient Boosting Decision Tree algorithm developed by Microsoft, utilizes a leaf-wise generation strategy to pinpoint the leaf offering the highest gain from splitting within decision trees (Wang et al, 2017). Notably, LightGBM accelerates forecasting speed and minimizes memory usage without compromising prediction accuracy (Ju et al., 2019). Our application of LightGBM yielded an impressive R-squared value of 0.93, indicating a robust fit for this explanatory model. These results allowed us to conduct further analysis using SHAP values to identify features with the greatest impact on the output, which will be discussed in more detail in the subsequent chapter.

This chapter encompasses data acquisition, data manipulation, RNN model preprocessing, RNN model construction and explanatory model construction, providing an overview of the core processes undertaken in this project.

6 Results

The establishment of the RNN model was a pivotal milestone in uncovering crucial insights for the project. This chapter presents hyperparameter tuning, RNN model performance evaluation and explanatory model results, providing a comprehensive overview of the model improvement process and outcomes.

6.1 Hyperparameter Tuning

After model construction, we used hyperparameter tuning to optimize the model performance. Utilizing the “RandomSearch” method within the Keras Tuner library, we explored various combinations of predefined tunable parameters to minimize the validated MAE. During model training, we incorporated essential callback functions, including early stopping and learning rate adjustment, to enhance model generalization.

Hyperparameter tuning determined the optimized model parameters. Early stopping was configured with a patience of 5 epochs, halting training if validation loss failed to improve consecutively for 5 epochs, thus preventing overfitting and conserving computational resources. The optimal learning rate was 0.2, facilitating a 20% reduction in the learning rate whenever validation loss stagnated. A number of 100 epochs and a batch size of 128 were also retrieved.

These fine-tuned parameters led the model to yield its best performance, achieving the lowest MAE of 4.89 days, notably around epoch 23. Considering the inherently long lead times characteristic of pharmaceutical supply chains, this level of precision suggests the model’s efficacy in predicting sequential delays at the final step of the manufacturing process. Consequently, this model equips GSK with the capability to proactively identify potential bottlenecks within the supply chain and take preemptive actions to address them, thereby enhancing overall operational efficiency.

6.2 RNN Model Performance Evaluation

To assess the RNN model’s performance, we employed various visualizations. These graphic representations helped us discern patterns inherent in the model’s predictions.

Figure 11 displays the model loss over epochs during both training and testing phases. The x-axis represents the number of epochs, which are iterations over the entire training dataset, while the y-axis denotes the MSE loss value. Both the training and validation loss decrease as the model learns and enhances its performance. As epochs progress, the validation loss undergoes a transition from large fluctuations to a better fit with the training loss. This graph clearly indicates the model's continuous learning process, ultimately resulting in improved performance.

Figure 11

Model Loss (MSE) Evolution

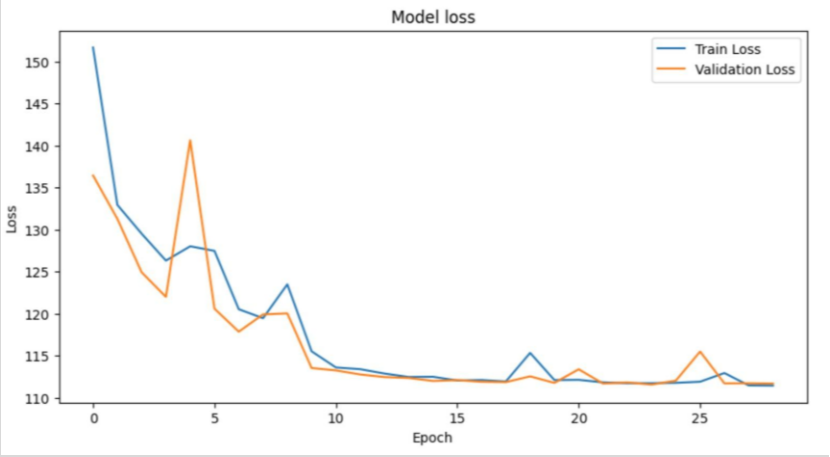
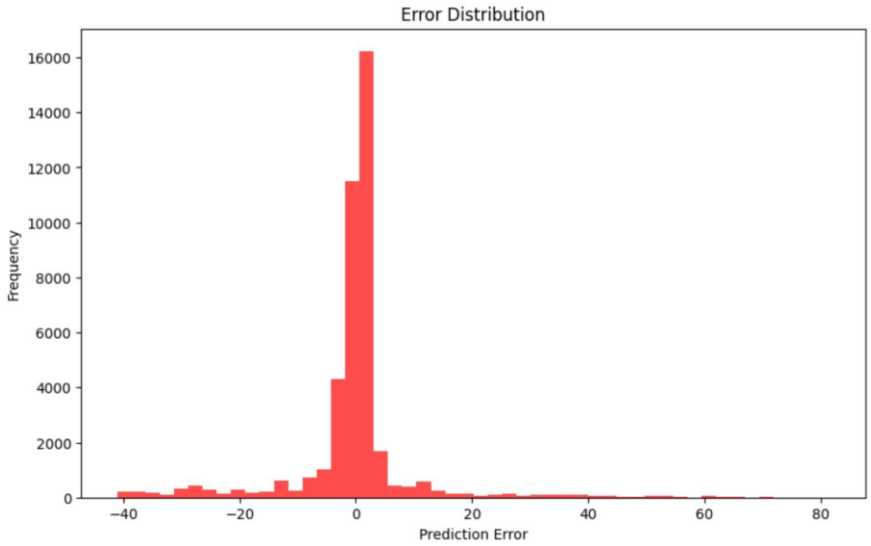


Figure 12 shows the distribution of prediction errors (MAE) of the model. The histogram conveys that the majority of prediction errors are clustered around 0, indicating that the model tends to make predictions close to the actual values. The long tail towards the right-hand side of the histogram suggests that instances of very large prediction errors are relatively infrequent. Despite these deviations, this visualization offers valuable insights into the accuracy and reliability of the model for GSK.

Figure 12

Distribution of Prediction Errors (MAE)



6.3 **Explanatory Model Results**

After the evaluation of RNN model results, we further developed an explanatory model using SHAP values to identify features most influential in causing time delays in final steps within paths. These crucial insights enable GSK to understand the primary drivers of delays within its supply chain.

According to Figure 13, path ID obtained the highest SHAP value and is the most influential feature, indicating that specific path sequences within GSK significantly affect predicted outcomes. Following closely is deviation, representing the delay for each path and step, and then step quantity, indicating the number of steps within each path also plays a role in shaping predictions.

Figure 13

SHAP Summary Plot for All Features

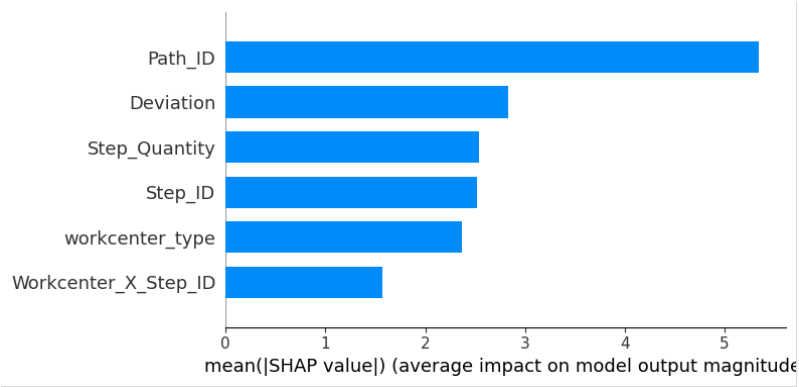


Figure 14 illustrates an overall prediction of finishing 6.62 days earlier in final steps within paths. The force plot provides further insights. For instance, specific path ID 64840 has the most positive impact on time delay predictions, while paths with 7 steps also contribute positively but to a lesser extent.

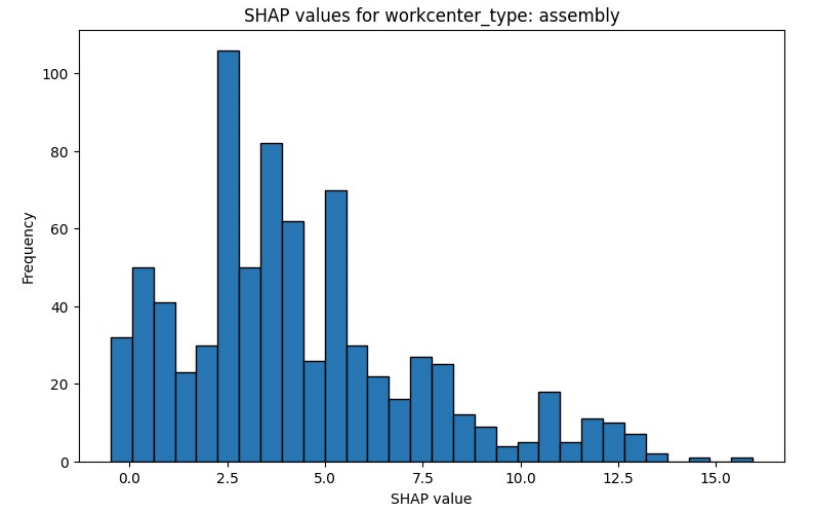
Figure 14

SHAP Force Plot for All Features

For a deeper understanding of feature importance, Figure 15 identifies that assembly has the most positive impact on time delays in final steps within paths, as it exhibits the highest SHAP value among all other workcenter types. Upon sharing this discovery with GSK, the company inferred that the quality stage preceding assembly could be the primary driver of delay impacts.

Figure 15

SHAP Values for Assembly Workcenter Type



Armed with the insights gleaned from the RNN model results, the evaluation of its performance and SHAP value analysis, the next step was to translate these findings into actionable recommendations for GSK. In the next chapter, we will delineate strategic recommendations based on the model's outcomes.

7 Recommendations

The initial objective of the project was to leverage manufacturing and quality data to predict sequential delays for GSK supply chain, specifically for the Benlysta brand. However, due to the parallel nature of the manufacturing and quality data, the project scope underwent modification. Consequently, the RNN model now functions as a proof of concept rather than a scalable tool for GSK. If GSK decides to incorporate the model into its operations in the future, the first crucial step will involve linking manufacturing and quality data.

During data examination, we discovered that 17% of planned date data and 0.3% of actual date data exhibited non-conforming date sequences. To proceed with the project, we aligned with GSK to establish an imputation logic for addressing these discrepancies. If the non-conforming date sequences persist as an ongoing issue for GSK, the company will need to continue implementing the same data imputation technique for subsequent analyses to maintain data consistency.

Several limitations should be acknowledged. Hyperparameter tuning already posed significant computational challenges in the current model, which processed data for only three site locations across primary and secondary manufacturing stages, focusing solely on manufacturing data. To scale the model effectively, GSK will require a robust computational environment capable of handling complex coding, along with a team of experts to fine-tune and debug the model. Furthermore, if Benlysta does not adequately represent the company's generalized supply chain structure, further adjustments to parameters will be necessary to ensure model applicability to the broader context of GSK operations.

If GSK extends the model to encompass other brands or additional sites in the future, we recommend beginning with summary statistics before delving into RNN modeling. This analysis will allow GSK to identify key process and workcenter types contributing to time delays for different brands and sites. Additionally, we highly recommend GSK to integrate summary statistics into its existing frontend dVSM dashboard. This enhancement can transform the dashboard into an early warning system, offering the management team a holistic overview of delay issues to proactively realign action plans and allocate resources.

Given that the project focused solely on primary and secondary manufacturing sites, the code was tailored to address these two specific stages. If GSK intends to expand the analysis to include additional upstream and downstream nodes in the future, we recommend revising and adapting the code to accommodate a broader scope.

To further enhance the value of the RNN model for GSK, we developed a subsequent prediction model for forecasting new data. In addition, we constructed a comprehensive data frame that consolidates all necessary information for identifying specific paths, steps, processes, workcenter types, and their respective upstream and downstream details. This framework empowers GSK to precisely pinpoint bottlenecks and promptly alert relevant departments, enabling proactive actions to be taken.

Within the tight project timeline, spanning less than 10 months, we navigated various challenges, including pinpointing planned date data within the company's system, resolving non-conforming date sequence issues, and ultimately redirecting our focus towards predicting delays only with manufacturing data. Despite these hurdles, this project served as a guidepost for future work within GSK. By facilitating the identification of the planned dates and uncovering underlying data issues, the project provided valuable insights for GSK internal investigations. For instance, while the manufacturing process accounted for the majority of all delayed POs, the quality release process experienced substantially longer average delays. Moreover, the project laid a solid foundation for data manipulation and RNN modeling, providing GSK the opportunity to broaden the project scope. The expansion could involve incorporating additional upstream and downstream stages, additional site locations, and multiple brands to enhance scalability.

8 Conclusion

Timely healthcare delivery is paramount in the pharmaceutical industry, where patient wellbeing relies on prompt medication supply. GSK's intricate supply chain faces challenges in tracking and managing sequential delays, prompting this capstone project.

Our objectives were to pinpoint planned dates within GSK's system and develop a robust machine learning model to predict sequential delays accurately. We began with an extensive review of the literature on sequential delays in pharmaceutical supply chains and explored the application of neural network machine learning methods in predicting delays. The state of the practice underscored the significance of this research question in addressing a critical gap in existing literature and laid the groundwork for subsequent methodology development.

To understand the frequency and locations of delays within GSK operations for the Benlysta brand, summary statistics were calculated across three distinct site locations spanning primary and secondary manufacturing stages. Approximately 40% of POs exhibited delays, predominantly in primary manufacturing sites. Further examination of site locations, process steps, and workcenter types highlighted specific areas prone to delays, equipping GSK with managerial insights for targeted action.

Our methodology employed a RNN due to its suitability for processing sequential data. Through rigorous validation of model effectiveness, using MSE and MAE, we ensured reliable insights for proactive supply chain management, facilitating timely medication delivery to patients.

The RNN model development encompassed data acquisition, manipulation, model preprocessing, and model construction. Data acquisition involved gathering planned start and end dates from multiple datasets, while data manipulation organized and reconstructed the supply chain network. Model preprocessing included outlier imputation and sequence verification, whereas model construction captured temporal dependencies and nonlinear feature learning, incorporating callback functions like dropout and early stopping to prevent overfitting. The creation of an RNN model was followed by hyperparameter tuning to optimize model

performance, reducing MAE to 4.89 days, indicating the model's efficacy in predicting sequential delays within GSK's supply chain. Subsequent SHAP value analysis helped identify key drivers of delays, enabling GSK to strategize and mitigate supply chain risks.

The project initially aimed to predict delays for both manufacturing and quality processes. However, due to challenges in linking manufacturing and quality data, the RNN model serves as a proof of concept rather than a scalable tool. To extend its utility, GSK must address data linkage issues and scale computational resources. Despite these challenges, the project provided valuable insights and laid a foundation for future enhancements.

While our journey may have deviated from the initial trajectory, the insights gained and lessons learned have paved the way for future advancements in GSK supply chain management. The model and the findings of this project allow GSK to proactively reduce delays, potentially resulting in the reduction of safety stock inventory. Building upon this foundation, GSK can further enhance operational efficiency, mitigate supply chain risks, and deliver essential medications to patients more effectively.

References

Amini, A. (2023). MIT 6.S191: Recurrent Neural Networks, Transformers, and Attention. YouTube. https://www.youtube.com/watch?v=ySEx_Bqxvvo

Bodendorf, F., Sauter, M., & Franke, J. (2023). A mixed methods approach to analyze and predict supply disruptions by combining causal inference and deep learning. International Journal of Production Economics, 256(0), 108708. https://www.sciencedirect.com/science/article/pii/S0925527322002900

Farias, F., Ludermir, T., & Bastos-Filho, C. (2020). Similarity Based Stratified Splitting: an approach to train better classifiers. https://www.researchgate.net/profile/Felipe-Farias/publication/344639457_Similarity_Based_Stratified_Splitting_an_approach_to_train_better_classifiers/links/5fa45cb5a6fdcc0624187e24/Similarity-Based-Stratified-Splitting-an-approach-to-train-better-classifiers.pdf

Great Learning. (2022). Types of Neural Networks and Definition of Neural Network. Great Learning. https://www.mygreatlearning.com/blog/types-of-neural-networks/

IBM. (n.d.). What are Neural Networks? IBM. https://www.ibm.com/topics/neural-networks

Ju, Y., Sun, G., Chen, Q., Zhang, M., Zhu, H., & Rehman, M. U. (2019). A Model Combining Convolutional Neural Network and LightGBM Algorithm for Ultra-Short-Term Wind Power Forecasting. https://ieeexplore.ieee.org/abstract/document/8653826

Kashem, M. A., Shamsuddoha, M., Nasir, T., & Chowdhury, A. A. (2023). Supply Chain Disruption versus Optimization: A Review on Artificial Intelligence and Blockchain. MDPI. https://www.mdpi.com/2673-9585/3/1/7

Management Centre Europe. (2012). New Opportunities & Strategies in the Pharmaceutical Industry. MCE CDN. https://cdn.mce.eu/eu/uploads/2016/05/Pharma-Industry-Executive-Issue-38-2012.pdf

Mikulic, M. (2023). Global pharmaceutical industry - statistics & facts. Statista. https://www.statista.com/topics/1764/global-pharmaceutical-industry/#topicOverview

Rassem, A., El-Beltagy, M., & Saleh, M. (2017). Cross-Country Skiing Gears Classification using Deep Learning. ResearchGate https://www.researchgate.net/publication/317954962_Cross-Country_Skiing_Gears_Classification_using_Deep_Learning

Sharkawy, A.-N. (2020). Principle of Neural Network and Its Main Types: Review. Journal of Advances in Applied & Computational Mathematics, 7(0), 8-19. https://www.avantipublishers.com/index.php/jaacm/article/view/851/502

Supply Chain Brain. (2022). Five Critical Challenges Facing Pharma Supply Chains. Supply Chain Brain. https://www.supplychainbrain.com/articles/34798-five-critical-challenges-facing-pharma-supply-chains

Tayyab, M., Awan, M. U., Bukhari, N. I., & Sabet, E. (2021). Key determinants of quality in the pharmaceutical supply chain. International Journal of Quality & Reliability Management, 39(2), 345-366. https://www.emerald.com/insight/content/doi/10.1108/IJQRM-06-2020-0213/full/html

Wang, D., Zhang, Y., & Zhao, Y. (2017). LightGBM: An Effective miRNA Classification Method in Breast Cancer Patients. https://dl.acm.org/doi/10.1145/3155077.3155079

Wu, Y., & Feng, J. (2018). Development and Application of Artificial Neural Network. https://doi.org/10.1007/s11277-017-5224-x

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Computation, 31(7), 1235-1270. https://doi.org/10.1162/neco_a_01199