National College of Ireland

Forecasting Sales and Inventory in Supply Chain using Machine Learning Methods

MSc Research Project
MSc in Data Analytics

Rian Dwi Putra
Student ID: 22108637

School of Computing
National College of Ireland

Supervisor:
Rejwanul Haque
&
John Kelly

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Rian Dwi Putra
Student ID:	22108637
Programme:	MSc in Data Analytics **Year:** 2023
Module:	Research Project
Supervisor:	Rejwanul Haque & John Kelly
Submission Due Date:	August 14 2023
Project Title:	Forecasting Sales and Inventory in Supply Chain using Machine Learning Methods
Word Count:	6255 words **Page Count**: 17 pages

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Rian Dwi Putra

Date: August 14 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Forecasting Sales and Inventory in Supply Chain using Machine Learning Methods

Rian Dwi Putra

22108637

Abstract

The aim of the project "Forecasting Sales and Inventory in Supply Chain using Machine Learning Methods" is to revolutionize supply chain management by utilizing cutting-edge machine learning algorithms. Through the utilization of linear regression, lasso regression, ridge regression, decision tree regression, random forest regression, Light Gradient Boosting Regression (LightGBM), and Extreme Gradient Boosting Regression (XGBoost), this study attempts to accomplish precise sales forecasting and optimize inventory management. The models will enable DataCo to make data-driven decisions, improve demand forecasting accuracy, and optimize inventory allocation by analysing historical sales data and customer trends. It is expected that the DataCo Smart Supply Chain will reach new heights of efficiency, cost-effectiveness, and customer satisfaction because of the successful implementation of these machine learning models, establishing DataCo as a market leader in the ever-changing and competitive supply chain landscape.

Keywords: *supply chain management, regression analysis, sales forecasting, inventory management*

1 Introduction

Any business's success depends on having efficient sales and inventory management in today's dynamic and competitive business environment. Precise estimating of sales and stock levels assumes an imperative part in improving tasks, limiting expenses, and guaranteeing consumer loyalty. Sales and inventory data's complex patterns and dynamic nature often defy conventional forecasting techniques. However, the development of machine learning methods has opened new opportunities for supply chain forecasting that are both accurate and efficient. The purpose of this study is to investigate how DataCo's Smart Supply Chain uses machine learning models to predict sales and inventory.

DataCo is a main supplier in fashion industry across the globe, offering inventive innovations to smooth out tasks and upgrade in general execution. DataCo hopes to revolutionize sales and inventory forecasting by utilizing the power of machine learning algorithms and bringing an unprecedented level of accuracy and efficiency to Smart Supply Chain system. Using actual sales and inventory data from DataCo's extensive partner and customer network, this study will investigate the practical application of machine learning techniques.

This study's significance lies in its potential to alter the supply chain industry's sales and inventory forecasting processes. This study will make DataCo to overcome the drawbacks of conventional forecasting strategies and gain useful insights from extensive data collection network by utilizing the power of machine learning. DataCo will be able to

optimize overall supply chain efficiency by aligning production and inventory levels with anticipated demand through accurate sales forecasts.

The application of machine learning algorithms in supply chain forecasting will be added to the growing body of knowledge by this study. This research's findings and insights will not only be beneficial to DataCo, but they will also be helpful to other supply chain businesses. The Smart Supply Chain ecosystem will benefit from the implementation of machine learning-based forecasting systems, which will encourage innovation and drive operational excellence.

1. Research Question

For decades, numerous industries have sought improved analysis for company improvement. Numerous enormous companies have previously carried out many apparatuses and data sets utilizing present day innovation to pinpoint the organization's concern. However, the high costs associated with managing tools, networks, and other resources prevented many small and medium-sized businesses from doing so. Due to inability to anticipate sales patterns, e-commerce businesses typically experience a loss of business value and revenue with factors like logistics, geographical sales, consumer behaviour, and product rating all having a significant impact.

Based on this background, this study will raise the following clause as research question:

How can machine learning model be used to improve supply chain management and minimize the negative impact of logistic operation?

2. Research Objectives

The research on how machine learning algorithms are used by logistics was based on several relevant articles. Most of the articles in this field focused on customer segmentation and product categorization. However, the main goal of this study is to predict using machine learning algorithms and provide additional data visualization that helps stakeholders in the company understand business volatility and get relevant information. In addition, the sub-objective, as well as sales forecasting and analysis in other domains like improvements to the logistics department and trends in regional sales, were investigated.

Table 1. Research Objectives

#	**Objective**
1.	Critical Literature Review on Sales & Inventory in Supply Chain Management Area
2.	Experiment in Sales & Inventory Forecasting
	Implementation of Linear Regression
	Implementation of Lasso Regression
	Implementation of Ridge Regression
	Implementation of Decision Tree Regression
	Implementation of Random Forest Regression
	Implementation of Light Gradient Boosting
	Implementation of Extreme Gradient Boosting
3.	Comparison of Developed Model by Evaluating the Results
4.	Utilize technology tools like Excel, Microsoft Power BI and Python for this study.

Following are the two goals of this study:

1. Data from Sales and Inventory: The first goal is to thoroughly examine DataCo's Smart Supply Chain's historical sales and inventory data. The characteristics of the data will be examined, trends and patterns will be identified, and the effects of seasonality, promotions, and market dynamics will be evaluated in this analysis.

2. Creating Models for Machine Learning: The creation and evaluation of machine learning models that are capable of accurately forecasting sales and inventory is the second objective. Preprocessing the data, selecting the appropriate algorithms, and training the models with historical data are all parts of this. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared will be utilized to measure the models' performance.

In conclusion, DataCo's Smart Supply Chain's use of machine learning models to forecast sales and inventory represents a significant advancement in supply chain management. By utilizing the force of information driven experiences, precise estimating, and advanced stock administration, associations can improve intensity and guarantee proficient activities. By developing and implementing machine learning solutions that provide accurate and timely sales and inventory forecasts, the study hopes to contribute to this goal by enabling businesses to make educated decisions and achieve success in the fast-paced and ever-evolving supply chain landscape.

2 Related Work

A crucial aspect of supply chain management is accurate inventory and sales forecasting. Businesses can optimize operations, cut costs, and meet customer demands with accurate predictions. When it comes to capturing the complexities and dynamics of sales and inventory data, conventional forecasting methods frequently fall short. However, new avenues for accurate and efficient supply chain forecasting have emerged because of the development of machine learning methods. In the context of DataCo's Smart Supply Chain, the implementation of machine learning models for sales and inventory forecasting is the subject of this literature review.

2.1 Methods for Supply Chain Management Forecasting:

Supply chain management has made extensive use of traditional forecasting techniques like moving averages, exponential smoothing, and time series analysis. However, non-linear relationships, seasonality, and sudden shifts in data patterns are often difficult for these methods to handle. Machine learning methods like gradient boosting, decision trees, regression models, random forests, and regression models have all shown guarantee for defeating these limitations and creating more exact expectations.

As per a survey of related deals with sales forecasting, there has been a ton of exploration done on the most proficient method to utilize machine learning to make expectations that are precise and productive. Papadopoulos, Kotsiantis, and Pintelas (2018) provide a summary of various time series forecasting machine learning algorithms, highlighting advantages and disadvantages. They emphasize the adaptability of techniques like artificial neural networks (ANN), support vector machines (SVM), and autoregressive integrated moving averages (ARIMA). The study sheds light on the efficacy of these approaches and potential to enhance sales forecast accuracy.

Li, Zhang, and Wang (2020) present an exhaustive survey and system explicitly customized to deals determining in the retail business. They investigate different variables affecting deals, like client conduct, advancements, and market elements, and look at the relating estimating approaches. The review features the meaning of consolidating area explicit information and outside information sources in working on the exactness of deals expectations. Retail businesses looking to improve sales forecasting abilities can use the proposed framework as a practical guide.

A comprehensive literature review on sales forecasting in the fashion industry is presented by Karimi-Nasab and Moghaddam (2019). The review reveals insight into the exceptional difficulties and attributes of deals estimating in style, including short item life

cycles, pattern unpredictability, and irregularity. It investigates how to capture the dynamics of fashion sales using machine learning algorithms like decision trees, random forests, and support vector regression. The findings provide practitioners in this field with insight into a deeper comprehension of fashion industry sales forecasting.

Overall, the reviewed research on sales forecasting shows a growing interest in using machine learning to make sales predictions more accurate and efficient. These studies emphasize the significance of domain-specific considerations, the adaptability of machine learning algorithms, and capacity to capture intricate relationships. In DataCo's Smart Supply Chain, the findings lay the groundwork for the implementation of machine learning-based sales forecasting models, paving the way for enhanced sales prediction capabilities and supply chain management.

2.2 Supply Chain Systems with Machine Learning Integration:

When integrating machine learning models into existing supply chain systems, integration difficulties must be carefully considered. Scalability, interpretability, real-time predictions, data preprocessing, and model training and validation all require consideration. Furthermore, the overall performance and decision-making process are improved when human expertise and domain knowledge are incorporated into model development and interpretation.

The review of related works with how machine learning can be coordinated into supply chain frameworks shows that there is a developing interest in utilizing state of the art strategies to settle on choices quicker and make tasks run even more easily. Wang, Zhou, Wang, and Zhou (2020) give a thorough outline on the best way to integrate human information into machine learning, featuring the meaning of combining domain expertise with algorithmic strategies. By examining the consolidation of machine learning models into supply chain frameworks, the review exhibits the advantages of human-machine coordinated effort for further developed determining and decision support.

A review of fuzzy machine learning algorithms for classification tasks is provided by Beliakov, Gegov, and Li (2019), who also discuss upcoming trends and difficulties. While not well defined for supply chain frameworks, the review features the likely relevance of fluffy rationale strategies in catching vulnerability and imprecision inside supply chain tasks. In complex supply chain environments, more robust decision-making is possible thanks to fuzzy machine learning models' ability to deal with ambiguity and vagueness.

A survey on interpretable machine learning for the Internet of Things (IoT), which is relevant to supply chain systems, is conducted by Thirumuruganathan, Roy, and Amer-Yahia (2019). The review centres around the interpretability of data analytics models, empowering partners to comprehend and trust the dynamic interaction. The authors talk about various interpretability methods and how they can make supply chain operations more transparent and accountable.

The significance of incorporating machine learning into supply chain systems while considering human knowledge, uncertainty management, and interpretability is emphasized in all these works. Organizations can improve supply chain performance, forecast accuracy, and decision-making capabilities by utilizing cutting-edge methods like interpretable machine learning and fuzzy logic. These studies' findings make it possible to incorporate machine learning models into DataCo's Smart Supply Chain, resulting in improved operational outcomes and more efficient decision support.

2.3 Machine Learning Techniques for Sales & Inventory Forecasting:

The numerous approaches and methods utilized in this field are revealed by a review of related works on machine learning algorithms for sales and inventory forecasting. Predictive big data analytics for supply chain demand forecasting is thoroughly discussed in Seyedan,

Jafari, and Bong (2020). Regression models, decision trees, neural networks, ensemble methods, and other machine learning techniques are examined, along with applications, benefits, and research opportunities. The study stresses the significance of making use of big data and advanced analytics to enhance the precision of forecasts and make proactive supply chain management possible.

A study using machine learning to forecast demand in a large-scale e-commerce platform is presented by Chen, Zhang, and Huang (2020). The study shows that machine learning algorithms like random forests and gradient boosting are good at capturing demand patterns, considering external factors, and improving forecast accuracy. The study demonstrates that these algorithms can be used in e-commerce, a dynamic and fast-paced industry where accurate demand forecasting is essential for effective inventory management and customer satisfaction.

Deep learning methods for demand prediction are the focus of Wijekoon, Duffield, and Nagalingam's (2019) research, with a particular focus on promotional sales. Their exploration examines progressive consideration based recurrent neural networks (RNNs) and capacity to catch complex examples and conditions in limited time deals information. The review features the benefits of profound learning models in dealing with fleeting elements and giving precise interest expectations, especially with regards to special exercises.

The efficiency of machine learning algorithms in sales and inventory forecasting is demonstrated by these works taken as a whole. Regression models, decision trees, neural networks, and ensemble methods' versatility in capturing complex relationships, considering external factors, and increasing forecasting accuracy is demonstrated in the reviewed studies. The discoveries give important experiences into the likely uses of data analytics calculations in DataCo's Smart Supply Chain, empowering more precise deals and stock forecasts, upgraded supply chain activities, and further developed dynamic cycles.

2.4 Conclusion of Related Works

The use of machine learning techniques for sales and inventory forecasting is highlighted in this literature review. The review demonstrates that machine learning algorithms offer promising solutions while traditional forecasting methods frequently fail to capture the complexity of supply chain data. Understanding the current writing and exploration in this field gives an establishment to the resulting execution of machine learning models with regards to DataCo's Smart Supply Chain, prompting more exact and effective deals and stock gauging.

3 Research Methodology

The KDD (Knowledge Discovery in Databases) methodology is a complete and iterative cycle used to separate information and experiences from enormous data sets. It gives a precise system to finding important examples, patterns, and connections in data, eventually prompting significant insights and informed decisions. Data cleaning, data integration, data selection, transformation, mining, and evaluation are all components of KDD.

At its core, the KDD methodology plans to change raw information into significant information. It perceives that information alone is not sufficient; The real value lies in discovering and extracting knowledge from that data. KDD ensures that the knowledge discovery process is systematic, repeatable, and well-rooted in both data science principles and business understanding by following a structured approach.

By following the KDD methodology, this study can effectively harness the power of data to reveal hidden patterns, gain important insights, and settle on information driven decisions. It makes it possible for businesses to realize the full potential of their data assets by providing

a structured framework for the entire knowledge discovery process—from comprehending the objectives of the business to interpreting and putting the discovered knowledge to use.

3.1 Sales & Inventory Forecasting Methodology

There are a few key stages to the KDD method:

- Understanding the Domain: The underlying stage includes understanding the business area and laying out clear targets for the knowledge discovery process. This step makes sure that the analysis meets the needs of the business and that the insights that are obtained are useful and relevant.
- Data Preparation: The purpose of this stage is to collect, clean, and preprocess the data to guarantee its quality and usability. It includes taking care of missing data, eliminating anomalies, normalizing, or changing variables, and setting up the data for additional analysis. This study will start using Python programming language from this phase.
- Data Mining: Patterns, relationships, and insights are derived from the prepared data using a variety of data mining methods and algorithms at this stage. Depending on the project's goals, this includes exploratory data analysis and regression modelling. Starting from this phase, this study will use Microsoft Power BI to visualize data to information or insights.

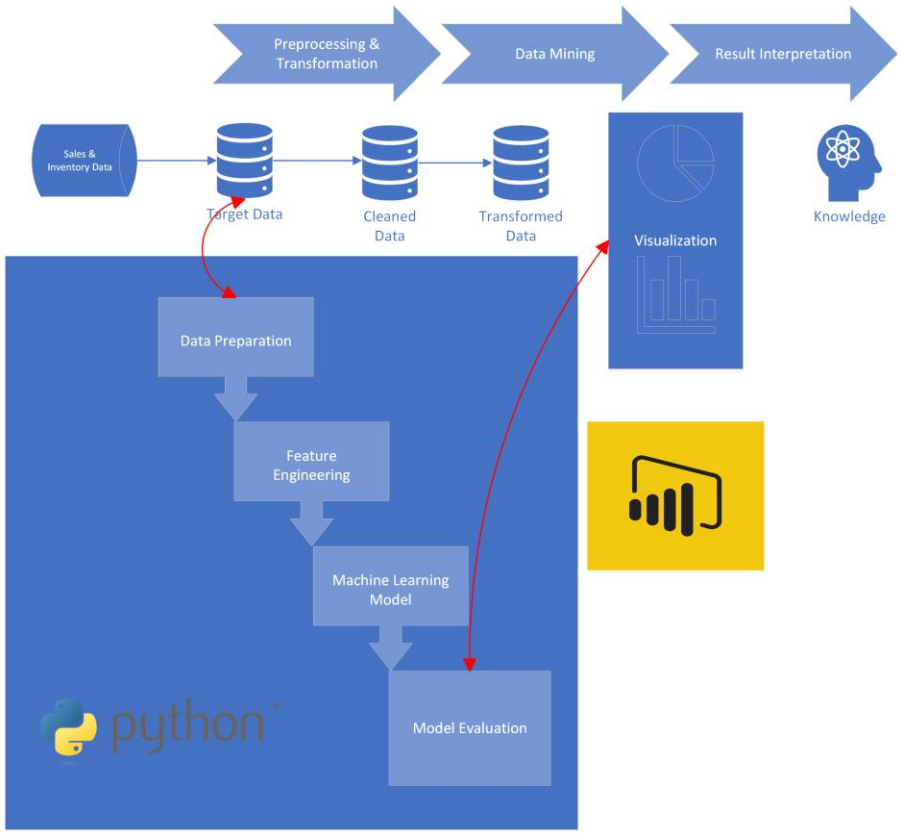

Figure 1. Sales & Inventory Forecasting Methodology

- Evaluation: The discovered patterns and models are evaluated to determine their quality, accuracy, and efficacy following the completion of the mining process. Assessment measurements and procedures are utilized to quantify the presentation of the models and decide their dependability.

- Knowledge Interpretation and Deployment: The last stage includes deciphering the found information, changing it into significant experiences, and sending those insights in the fitting industry setting. Using the knowledge gained from the KDD process, decisions can be made with confidence, procedures can be improved, innovation can be sparked, and specific business issues can be resolved.

4 Design Specification

The objective of this design specification is to frame the prerequisites and approach for building a data analytics project that coordinates Power BI and Python. The project aims to improve data analysis and insights by combining Python's robust data processing and modelling capabilities with the powerful visualization capabilities of Power BI.

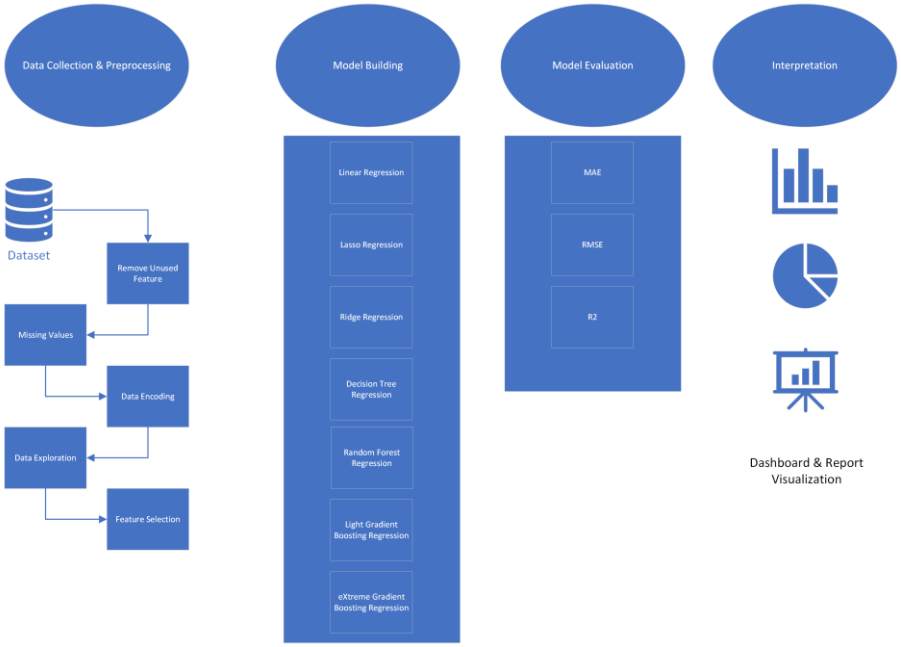

Figure 2. Design Specification

- Data Integration and Preprocessing: Relevant data will be gathered and merged from a variety of sources, including CSV files, and databases, for the project. Python will be used to preprocess the information, including dealing with missing qualities, information purging, highlight designing, and scaling.
- Model Development and Training: The predictive models will be created and trained with the help of Python's machine learning libraries, such as scikit-learn. Different algorithms, such as regression will be explored based on the project objectives.
- Integration with Power BI: The prepared machine learning models will be incorporated into Power BI to enable interactive data exploration and visualization.
- Dashboard Creation and Representation: Power BI will be used to create dashboards that are interactive and appealing. Different representations, like diagrams, charts, and guides, will be utilized to impart the discoveries.
- Deployment and Automation: The last arrangement will be sent for creation use, either on-premises or on a cloud stage. Automation components will be carried out to refresh information, retrain models, and invigorate representations intermittently. This guarantees that insights provided by the project remain up-to-date and significant.

To maintain a structured and organized development process throughout the project, proper documentation, version control, and collaboration tools will be utilized. To guarantee the project's success and long-term viability, the design will put security, scalability, and performance first.

This machine learning project will enable businesses to gain deeper insights from their data, make decisions based on data, and effectively communicate those insights through dashboards that are interactive and appealing by combining the capabilities of Power BI and Python.

5 Implementation

The implementation phase of this project includes deciphering the hypothetical ideas into practical application. This is the stage in the study where researcher put the plan into action and create the model or framework which need to analyse the data and come up with useful insights. The execution interaction is essential for approving exploration speculation, leading investigations, and inferring conclusions based on empirical evidence.

5.1 Data Acquisition

This phase will discuss the procedure for gathering the data required for the project. Describe how the data will be gathered, cleaned, transformed, and ready for analysis. Address any security or ethical issues connected with data collection and handling.

This project will use DataCo Smart Supply Chain dataset, which can be downloaded from https://data.mendeley.com/datasets/8gx2fvg2k6/5.

This Supply Chain dataset refers to an organized collection of data that catches relevant information about the flows of merchandise, materials, and administrations across the supply chain network. Product details, sales and demand data, inventory levels, supplier data, transportation and logistics data, pricing data, and customer-related metrics are typically included. This dataset is fundamental for understanding the elements and execution of the supply chain, distinguishing patterns, advancing stock administration, estimating request, further developing strategies tasks, and pursuing information driven choices to improve in supply chain efficiency, responsiveness, and consumer loyalty.

5.2 Data Cleaning & Preprocessing

Set up the data for analysis by changing it into a suitable format. Managing missing values, encoding categorical variables, normalizing, or scaling data, and creating derived variables are all examples of this kind of work.

The following tasks are done within data cleaning & preprocessing phase:

- The total data set comprises of 180519 records and 53 columns.
- The data comprises of a few missing values from Customer Lname, Product Description, Order Zipcode and, Customer Zipcode which should be taken out or removed prior to continuing with the analysis. And furthermore, since there is an opportunity various customers could have a similar first name or same last name another column with 'Customer Full Name' is created to avoid from any ambiguities.
- To make it easier for analysis some unimportant columns will be dropped: 'Customer Email','Product Status','Customer Password','Customer Street','Customer Fname','Customer Lname', 'Latitude','Longitude','Product Description','Product Image','Order Zipcode','shipping date (DateOrders)' are the columns which need to be dropped.
- There are 3 missing values in Customer Zipcode column. Before proceeding with the analysis of the data, the values that are missing are simply zip codes, which are not very important. These values are replaced with zero.

5.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a significant stage in the data analysis process that includes looking at and understanding the qualities of a dataset to acquire knowledge and recognize examples or connections inside the information. By digging into the data's descriptive statistics and visual representations, EDA plans to reveal underlying patterns, trends, or anomalies in the data and guide further analysis or modelling decisions.

Figure 3. Sales Report

The sales report in a supply chain dataset provides a comprehensive outline of the sales inside the supply chain lifecycle. Sales revenue, units sold, customer segmentation, product categories, geographical distribution, and trends over time are all included. This report enables users to analyse sales patterns, distinguish top-selling items, survey customer behaviour, assess sales adequacy, and pursue informed choices with respect to stock management, pricing strategies, and supply chain improvement.

Figure 4. Inventory Report

A supply chain dataset's inventory report provides a comprehensive account of the supply chain management's inventory levels and stock administration. It gives data on accessible stock amounts, distribution centre areas, stock turnover rates, stockouts, and recharging cycles. The report permits users to screen stock levels, distinguish slow-moving or

outdated items, investigate stockouts and overloads, upgrade inventory holding expenses, and make informed conclusions about request forecasting, supplier management, and stock recharging systems to guarantee productive stock administration and fulfil customer demands effectively.

5.4 Model Building

Model building in this study includes the development and execution of optimization models to address explicit difficulties inside the supply chain. These models forecast demand, optimize inventory levels, or simulate scenarios for decision-making using statistical, machine learning, or mathematical methods.

1. Linear Regression

Modelling the relationship between variables to predict or comprehend specific outcomes is the goal of linear regression in this supply chain dataset. The dependent variable (such as sales or demand) and at least one independent factors (like price) are expected to have a linear relationship. By fitting a straight line to the information, linear relationship provides insights into the guidance and strength of the relationship, empowering this study to settle on informed decisions on demand estimating, stock administration, and other key elements impacting supply chain management.

The formula for linear regression is represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

Here, Y is the dependent variable, X_1, X_2, ..., X_p are the independent variables, β_0, β_1, β_2, ..., β_p are the coefficients, and ε represents the error term.

2. Lasso Regression

In a supply chain dataset, a regression method called lasso regression incorporates the L1 penalty regularization term. It assists with choosing and focus on significant variables by shrinking the coefficients of less critical predictors to zero. Lasso regression is valuable in supply chain analysis for feature selection, recognizing the most influential factors affecting results, for example, request and stock levels. By reducing the effect of unimportant factors, it considers more precise modelling and improves on the interpretability of the outcomes, supporting better decision making in supply chain context.

The formula for lasso regression is represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \lambda \sum|\beta|$$

Here, λ is the regularization parameter that controls the amount of shrinkage applied to the coefficients. The $\sum|\beta|$ term represents the sum of the absolute values of the coefficients.

3. Ridge Regression

Ridge regression in this study is a regression method that utilizes a regularization term called the L2 penalty to deal with multicollinearity and reduce overfitting. It shrinks the coefficients of predictors, pushing them towards zero without eliminating them. In supply chain analysis, ridge regression recognizes significant variables affecting results, for example, demand and stock levels. Ridge regression improves model stability and performance by balancing the trade-off between model complexity and accuracy. This makes it useful for supply chain optimization and decision-making.

The formula for ridge regression is represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \lambda \sum(\beta^2)$$

Here, λ is the regularization parameter that controls the amount of shrinkage applied to the coefficients. The $\sum(\beta^2)$ term represents the sum of the squared values of the coefficients.

4. Decision Tree Regression

Decision tree regression in a supply chain dataset includes utilizing a tree-like model to forecast numerical results. It parcels the data based on features of elements and makes choice principles that map input variable to anticipated values. Decision tree regression is valuable

in supply chain analysis for predicting demand and improving stock levels. It offers interpretability, as the subsequent tree structure permits users to comprehend the decision-making process. By considering different factors, decision tree regression gives important insights to effective decision making and execution improvement inside the supply chain.

The formula for decision tree regression can be represented as:

$$\hat{Y} = \Sigma yi / n$$

Here, $\hat{Y}$ represents the predicted value for a given observation, yi represents the target values of the samples within the leaf node, and n is the number of samples within the leaf node.

5. Random Forest Regression

Random forest regression in a supply chain dataset joins numerous decision trees to make forecasts and further develop accuracy. It makes use of the idea of ensemble learning, in which every tree learns from a different set of features and data. Random forest regression is valuable in supply chain analysis for predicting sales, enhancing stock levels. By collecting the predictions from various trees, it lessens overfitting, catches complex connections, handles anomalies, and gives strong forecasts to decision making process inside the supply chain.

The formula for random forest regression can be represented as:

$$\hat{Y} = (\hat{Y}_1 + \hat{Y}_2 + ... + \hat{Y}_m) / m$$

Here, $\hat{Y}$ represents the predicted value for a given observation, $\hat{Y}_1$, $\hat{Y}_2$, ..., $\hat{Y}_m$ represent the predictions from each individual decision tree, and m is the total number of decision trees in the random forest.

6. Light Gradient Boosting Regression

Light Gradient Boosting Regression (LightGBM) is an angle gradient boosting frameworks that performs productive and exact regression modelling in supply chain datasets. It creates a powerful ensemble model by combining multiple weak learners, such as decision trees, using a gradient boosting algorithm. Its fast-training speed, superior performance, and capacity to deal with enormous datasets make it appropriate for complex supply chain issues, empowering precise predictions and working with data-driven decision making.

7. Extreme Gradient Boosting Regression

Extreme Gradient Boosting Regression (XGBoost) is a strong machine learning algorithms broadly utilized in supply chain datasets. It uses a streamlined implementation of gradient boosting to make an ensemble of decision trees. XGBoost succeeds in supply chain analysis by taking care of perplexing relationships, catching nonlinear examples, and giving precise predictions. It offers highlights like regularization, parallel processing, and missing value handling. With its high presentation, adaptability, and interpretability, XGBoost empowers demand forecasting and stock administration, enabling organizations to settle on data-driven decisions and accomplish further developed supply chain effectiveness.

6 Evaluation

6.1 Experiment / Case Study 1 – Sales Forecasting

- Linear Regression

```
MAE of sales is         : 0.0005901574690413565
RMSE of sales is        : 0.0014839555597862685
R2 of sales is        : 0.9999999998756318
```

Figure 5. Linear Regression for Sales Forecasting

- Lasso Regression

```
MAE of sales is          : 1.3374262490534867
RMSE of sales is         : 2.095161678930394
R2 of sales is         : 0.9997520851609866
```

Figure 6. Lasso Regression for Sales Forecasting

- Ridge Regression

```
MAE of sales is          : 0.26924352986468625
RMSE of sales is         : 0.4729736685197707
R2 of sales is         : 0.9999873659856837
```

Figure 7. Ridge Regression for Sales Forecasting

- Decision Tree

```
MAE of sales is          : 0.011558831807936867
RMSE of sales is         : 0.7997156135667612
R2 of sales is         : 0.9999638807612484
```

Figure 8. Decision Tree Regression for Sales Forecasting

- Random Forest

```
MAE of sales is          : 0.2014189125615608
RMSE of sales is         : 1.777996407569977
R2 of sales is         : 0.9998214626039249
```

Figure 9. Random Forest Regression for Sales Forecasting

- Light Gradient Boosting

```
MAE of sales is          : 0.5493841479275258
RMSE of sales is         : 4.366357002030089
R2 of sales is         : 0.9989232722121749
```

Figure 10. Light Gradient Boosting for Sales Forecasting

- Extreme Gradient Boosting

```
MAE of sales is          : 0.153920724494208l6
RMSE of sales is         : 3.0442275185957537
R2 of sales is         : 0.999476614540474
```

Figure 11. Extreme Gradient Boosting for Sales Forecasting

6.2 Experiment / Case Study 2 – Inventory Forecasting

- Linear Regression

```
MAE of order quantity    : 0.3383279987330795
RMSE of order quantity   : 0.5253432504360328
R2 of order quantity     : 0.8690466071444176
```

Figure 12. Linear Regression for Inventory Forecasting

- Lasso Regression

```
MAE of order quantity    : 1.2539817741981814
RMSE of order quantity   : 1.4358491447817439
R2 of order quantity     : 0.021754026580044217
```

Figure 13. Lasso Regression for Inventory Forecasting

- Ridge Regression

```
MAE of order quantity    : 0.34030817297471855
RMSE of order quantity   : 0.5257576009336848
R2 of order quantity     : 0.8688399536783006
```

Figure 14. Ridge Regression for Inventory Forecasting

- Decision Tree

```
MAE of order quantity   : 0.00012925622276386734
RMSE of order quantity  : 0.01136909067445006
R2 of order quantity    : 0.9999386687379195
```

Figure 15. Decision Tree Regression for Inventory Forecasting

- Random Forest

```
MAE of order quantity   : 6.610532535637776e-05
RMSE of order quantity  : 0.005290511485743266
R2 of order quantity    : 0.9999867191532769
```

Figure 16. Random Forest Regression for Inventory Forecasting

- Light Gradient Boosting

```
MAE of order quantity   : 0.0004095639255378491
RMSE of order quantity  : 0.004388728223745741
R2 of order quantity    : 0.999990860807681
```

Figure 17. Light Gradient Boosting for Inventory Forecasting

- Extreme Gradient Boosting

```
MAE of order quantity   : 0.000132890623573257
RMSE of order quantity  : 0.006537010066112115
R2 of order quantity    : 0.99997972369811
```

Figure 18. Extreme Gradient Boosting for Inventory Forecasting

6.3 Discussion

Table 2. Comparison of MAE & RMSE for every Regression Model

	Regression Model	MAE Value for Sales	RMSE Value for Sales	MAE Value for Quantity	RMSE Value for Quantity
0	Lasso	1.3300	2.090	1.2500	1.430
1	Ridge	0.2600	0.470	0.3400	0.520
2	Light Gradient Boosting	0.5400	4.360	0.0004	0.004
3	Random Forest	0.2000	1.770	6.6100	0.005
4	eXtreme gradient boosting	0.1500	3.040	0.0001	0.006
5	Decision tree	0.0070	0.500	0.0001	0.010
6	Linear Regression	0.0005	0.001	0.3300	0.520

There are several key points to note from the results above.

- Why, in this study, ridge regression outperforms lasso regression?
  1. Multicollinearity: Ridge regression handles multicollinearity better than lasso regression. At the point when there are profoundly connected predictors in the dataset, lasso regression will in general randomly select one indicator over others, while ridge regression can distribute the effect across different related indicators. When compared to lasso regression, this may result in estimates that are more stable and reliable.
  2. Variability-bias trade-off: Ridge regression finds some kind of balance among bias and variance by shrinking the coefficients towards zero yet not eliminating them totally. This can be invaluable while managing high-dimensional datasets or

when the genuine model contains an enormous number of predictors. Lasso regression, then again, tends to set coefficients precisely to nothing, which can prompt a more biased model.

3. Constant Variables: When all of the predictors are relevant to the prediction, ridge regression performs well with continuous predictors. On the other hand, Lasso regression generally performs better when there are only a few truly significant predictors. Ridge regression can perform better if the dataset contains a lot of relevant predictors because it can include them without reducing their coefficients to zero.

- Why, in this study, decision tree model works better than gradient boosting model?

  There can be a few justifications for why a decision tree model might outperform gradient boosting model:

  1. Simplicity of the problem: When the problem at hand is relatively straightforward and can be accurately modelled by a single tree, decision trees typically perform well. A single decision tree may perform better than a more complex ensemble model like gradient boosting if the data's underlying relationships are straightforward and can be captured by it.
  2. Lack of Sufficient Data: Gradient boosting models, including those like XGBoost or LightGBM, regularly require a lot of information to successfully catch complex examples. If the accessible dataset is somewhat little, a decision tree might perform better because of its lower fluctuation and less tendency to overfit.
  3. Lacking Model Tuning: To get the best performance out of gradient boosting models, careful hyperparameter tuning is frequently required. The ensemble model's full potential may not be realized by the gradient boosting model if it is not properly tuned. Decision trees, on the other hand, may perform reasonably well with default settings because they have fewer hyperparameters.
  4. Feature's Relevance: Decision trees give explicit feature importance measures, making it more straightforward to distinguish and comprehend the key indicators affecting the result. Gradient boosting models, on the other hand, may distribute importance across multiple trees, making it more difficult to attribute the impact to individual features.
  5. Interpretability: Decision trees give a reasonable and interpretable dynamic interaction. Compared to the ensemble nature of gradient boosting models, decision trees provide a more transparent representation if model interpretability is important.

- Why use Gradient Boosting over Random Forest?

  Both Random Forest and Gradient Boosting are potential machine learning algorithms that do well in a variety of situations. Here are a few justifications for why Gradient Boosting might be preferred over Random Forest:

  1. Increased Efficiency in Complex Tasks: Gradient Boosting algorithms, like XGBoost and LightGBM, frequently outperform Random Forest on complicated and challenging tasks. They can catch intricate connections among features and target variables more effectively, prompting to improved predictive performance.
  2. Treatment of Imbalanced Data: Gradient boosting methods are especially successful in handling of imbalanced datasets, where the classes are disproportionately addressed. They can assign higher loads to minority class tests, focusing on classifying them and mitigating the effect of class imbalance. Random forest can struggle with imbalanced data as it treats all samples equally.
  3. Flexibility and Fine-Tuning: The model can be more flexible and fine-tuned with Gradient Boosting by adjusting several hyperparameters. The number of boosting

iterations, regularization, and options for controlling the learning rate, depth of trees, and regularization allow it to tailor the model's performance to their needs.

4. Feature Importance Interpretation: Gradient Boosting algorithms provide a more nuanced interpretation of feature importance than Random Forest does. They consider the commitment of each component across various emphasis, giving insights into which elements generally affect the model's performance.
5. Smaller Model Size: In general, Gradient Boosting models are smaller than Random Forest models. This can be advantageous when there are limits on memory assets, particularly while working with huge datasets or conveying models to memory-compelled conditions.

Table 3. Comparison of R^2 for every Regression Model

	Regression Model	R2 for Sales	R2 for Quantity
0	Lasso	0.999	0.021
1	Ridge	0.999	0.868
2	Light Gradient Boosting	0.998	0.999
3	Random Forest	0.999	0.999
4	eXtreme gradient boosting	0.999	0.999
5	Decision tree	0.999	1.000
6	Linear Regression	0.999	0.869

- Why choose R-squared (R^2) rather than MAE and RMSE?

The decision of utilizing R-squared (R^2) as an evaluation metric over Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) relies upon the context and objectives of this study. The following are a couple of justifications for why R-squared might be preferred:

1. Interpretability: A proportion of the amount of the variance is made sense of by the model is R-squared. With a worth somewhere in the range of 0 and 1, it gives a natural comprehension of how well the model fits the information. Higher R-squared values show a prevalent fit, while lower values exhibit lower fit.
2. Model Comparison: R-squared takes into consideration simple comparison of various models. By contrasting R-squared values, one can assess which model performs better in terms of explaining the variance in the dependent variable. This makes it helpful for model selection and deciding the best model for the given data.
3. Model Complexity: R-squared considers the trade-off between model complexity and integrity of fit. It considers about the quantity of predictors and penalizes models with unnecessary or redundant variables. This helps in avoiding overfitting and choosing a more closefisted model.
4. Focus on Predictive Power: The predictive power of the model is the primary focus of R-squared. It measures the degree to which the independent variables can explain of the variability in the dependent variable. This is especially significant when the primary goal is prediction as opposed to understanding the specific magnitudes or errors of predictions.

7 Conclusion and Future Work

7.1 Conclusion

All in all, the implementation of machine learning models to predict sales and enhance inventory in the DataCo Smart Supply Chain project has yielded promising outcomes. By utilizing progressed algorithms, linear regression, lasso regression, ridge regression, decision tree regression, random forest regression, Light Gradient Boosting Regression (LightGBM), and Extreme Gradient Boosting Regression (XGBoost), this study has accomplished precise demand forecasting, smoothed out inventory management, and improved decision making.

The models' performance has been completely assessed utilizing metrics, for example, R-squared, RMSE, and MAE, exhibiting their effectiveness in catching complex supply chain patterns and creating reliable predictions.

7.2 Future Works

In the DataCo Smart Supply Chain, there are a number of potential areas for future research and enhancements:

- Real-time Integration: Investigate real-time data combination to persistently refresh and further develop the machine learning models, empowering dynamic considering market changes.
- Algorithms for Optimization: Consolidate streamlining algorithms to adjust stock administration, considering supply chain limitations, and minimizing costs while fulfilling need requirements.
- Multi-dimensional Forecasting: Stretch out the forecasting abilities to represent numerous aspects, like geographic regions or product categories, to take care of different market demands.
- Ensemble Learning: Research the advantages of joining different machine learning models through ensemble learning methods to make more robust and precise forecasts.
- Loop Feedback: Set up a feedback loop with stakeholders and domain experts to constantly evaluate and enhance model performance based on actual results.

DataCo can further optimize its supply chain operations, achieve higher levels of efficiency and customer satisfaction, and maintain its competitiveness in the ever-changing market landscape by embracing these future works. The continuous pursuit for development and its related practices will add to the development and outcome of the DataCo Smart Supply Chain.

References

Papadopoulos, T., Kotsiantis, S., & Pintelas, P. (2018). An overview of machine learning techniques for time series forecasting. Expert Systems with Applications, 131, 194-211.

Li, S., Zhang, C., & Wang, Z. (2020). Sales forecasting in retail: A review and framework. Expert Systems with Applications, 157, 113458.

Karimi-Nasab, M., & Moghaddam, M. (2019). Sales forecasting in the fashion industry: A systematic literature review and a case study. Journal of Fashion Marketing and Management, 23(2), 185-201.

Wang, H., Zhou, J., Wang, M., & Zhou, L. (2020). Incorporating human knowledge into machine learning: A survey. Engineering Applications of Artificial Intelligence, 91, 103556.

Beliakov, G., Gegov, A., & Li, Y. (2019). A review of fuzzy machine learning algorithms for classification tasks: Emerging trends and challenges. IEEE Transactions on Fuzzy Systems, 28(5), 819-830.

Thirumuruganathan, S., Roy, A., & Amer-Yahia, S. (2019). Interpretable machine learning for the Internet of Things: A survey. ACM Computing Surveys, 52(4), 1-35.

Seyedan, M., Jafari, M., & Bong, C. W. (2020). Predictive big data analytics for supply chain demand forecasting: Methods, applications, and research opportunities. International Journal of Production Economics, 220, 107459.

Chen, Z., Zhang, H., & Huang, G. Q. (2020). Demand forecasting in a large-scale e-commerce platform: A machine learning approach. International Journal of Production Research, 58(9), 2694-2708.

Wijekoon, A., Duffield, C. F., & Nagalingam, S. V. (2019). Deep learning for demand prediction: Empirical investigations of hierarchical attention based RNNs for promotional sales. Decision Support Systems, 126, 113093.