EMPIRICAL EVALUATION OF THE SALES PREDICTION USING DATA FROM DATACO'S SUPPLY CHAIN MANAGEMENT AND THE ADDITION OF ECONOMIC INDICATORS

By

Md Mozahidur Rahman, CSC, Independent University of Bangladesh, 2007

A Major Research Project
presented to Ryerson University
(renaming to Toronto Metropolitan University in progress)*
in partial fulfillment of the
requirements for the degree of

Master of Science
in the Program of
Data Science and Analytics

Toronto, Ontario, Canada, year of convocation, i.e. 2022

* In April 2022, the university announced the new name of Toronto Metropolitan University, which will be implemented in a phased approach.

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Md Mozahidur Rahman

Abstract

Sales forecasts determine how much product is needed to meet targets. Economic indicator is a macroeconomic metric that researchers use to assess current and future economic activity and opportunities, which is also a key factor of ‘systematic risk’ of the local or global organizations. This research had constructed to establish an empirical comparison between future sales forecast considering by economic indicators. Population, unemployment rate and GDP per capita of customer country have been selected as representative economic indicators. After the deployment of the data model and its examination, some intriguing facts emerged. There is no significant improvement on sales forecast after incorporating economic indicator. In addition, the most accurate predictor was Xgboost, which provided mean absolute error (MAE) score of 0.014. Even after accounting for economic indicators, the Random Forest Predictor still produced the same result.

ACKNOWLEDGEMENTS

I am deeply indebted to Dr. Saman Hassanzadeh Amin for his support and assistance in making this project a reality. I also could not have undertaken this journey without his valuable patients and feedback. My MRP supervisor was Dr. Saman, who has been a wonderful help throughout the semester in directing my research and giving me insightful criticism. The manual approach to improve the dataset took time and effort, and I had his full support throughout.

I really value Dr. Ceni Babauoglo's second reader help for this project, who provided thorough explanations of the growth of research skills, project requirements, and moral support.

Table of Contents Page

List of Figures Page

1.0 Introduction:

The supply chain is made up of a network of businesses that are connected through upstream and downstream links to the various processes and activities that result in value being manufactured in the form of goods and services that are then delivered to the final consumer. (Christopher 2005, p. 17) Supply chain operations are responsible for transforming raw materials, components, and natural resources into finished goods.(Kozlenkova & Irina et al. 2015). In the regular tasks of many professional and personal activities, supply chains and industrial logistics processes are vital. Additionally, they are essential to the development of the global economy. (Min *et al.*, 2019). A supply chain sales forecast is a projection of demand. A forecast can be established using quantitative, qualitative, or a combination of approaches, and it can be based on extrinsic or intrinsic factors. If the sales forecast sniffles, the supply chain as a whole gets sick. Sales forecasts are used to calculate the quantity of products needed to achieve accomplishment. It outlines the quantity of raw and packing materials which should be ordered, the use of existing equipment, and whether new machinery are required.(Boone et al.,2019). Key performance indicators are quantitative matrices built from historical data assembled over time inside an organization and used to track and forecast business performance efficacy. (Mishra *et al.*, 2018). On other words, to anticipate future sales, conventional statistical forecasting techniques merely extrapolate historical trends and seasonal impacts. As a result, these approaches are unable to predict business environment macroeconomic changes which are statistical data readings that depict the state of the economy in a specific nation, area, or industry. Both methods, nevertheless, are infamously biassed, as

Humans often struggle to make these adjustments and take a long time. (Verstraete *et al.*, 2020) A DataSet of Supply Chains was employed in this research of the business DataCo Global. This study will be designed to produce an empirical comparison between prospective sales forecasts that are based on supply chain management's key performance indicators for demand planning and same prophecy with additional bias of economic indicators.

2.0 Literature Review:

Because it aids professionals in achieving better financial budgeting, operational planning, and inventory strategies, sales forecasting is significant in the retail business. (Ma & Fildes, 2021). While the expert panel established an interactive sales forecasting process based on their knowledge of demand forecast, the Delphi method is a common form of qualitative systematic analysis. (Dalkey & Helmer, 1963). Comparatively, quantitative research methods (Ord et al., 2017) might be based on historical data like moving average, exponential smoothing, or time series, as well as on contemporary predictors (statistical regression, machine learning, deep learning).

The most challenging issues in practice relate to the identification of trustworthy predictors and their associated temporal causalities (time lags). Additionally, territories, yearly events, and climate fluctuations would have an impact on certain product categories (Martnez-de-Albeniz & Belkaid, 2021; Ulrich et al., 2021). The consequences of dynamic interactions and economic indicators are both covered by this study. The study evaluated the influence of the weather and consumer behaviour when examining the sales of soft drinks or apparel at department stores. The impact of weather on sales has been seen to differ by category. For instance, experts found that street mall sales of football memorabilia were higher than those at street shops. Seasonal merchandise also typically sells better. Researchers found that the weather had a negligibly little impact on sales.

It is less obvious how much can be saved in terms of actual supply chain performance results, but a recent study suggests that utilizing machine learning-based enhanced demand forecasting could lessen the impact and boost performance. In this study, hybrid approaches for demand forecasting, such as ARIMAX and neural networks, are developed. The presented technique incorporates both time series and explanatory elements. The approach was used and assessed in relation to a functional product and a steel manufacturer. It was demonstrated that there are statistically significant differences between traditional and ML-based approaches to demand forecasting for supply chain performance improvement. (Feizabadi, 2022)

However, for macroeconomic forecasting, researchers created lasso-type penalized regression models. Researchers discovered that penalized regression techniques are more robust to misspecification than factor models, even when the underlying DGP has a factor structure. Furthermore, it has been shown that, despite a decline in their selectivity, penalized regression methods outperform conventional methods in forecasting for non-stationary data containing co integrated variables. (Smeekes et al., 2018)

Moreover, using the quarterly profit data gathered on the three biggest airlines in China along with additional recent historical data on external influencing factors, researchers develop profit models to forecast the performance of airlines in the short run. They specifically suggest the use of the LASSO estimation method for this issue and evaluate its performance against a number of other cutting-edge methods, such as neural networks, ridge regression, support vector regression, tree regression, and support vector regression. It is demonstrated that LASSO performs better than the other techniques in this study overall. They came to a number of conclusions regarding the profitability of Chinese airlines in relation to the oil price and other important variables. (Xu et al., 2021)

Using economic indicators and Google online search data, a novel multivariate model was created to estimate monthly auto sales data in Germany. For the majority of the automotive manufacturers and forecast timeframes, models that used Google search data performed statistically better than competing models. (Fantazzini et al., 2015). Similarly, to determine the long-term effects of economic indicators on sales, a vector error correction model (VECM) of multi-segment vehicle sales was developed based on impulse response functions. In this instance, economic data included the consumer price index (CPI), the unemployment rate, petrol prices, and housing starts. According to the empirical findings, VECM, as opposed to traditional time series approaches, can greatly increase prediction accuracy of car sales for 12-month ahead projection in terms of RMSE (42.73 %) and MAPE (42.25 %).(Sa-Ngasoongsong et al., 2012)

In addition, vegetable prices were decomposed using STL decomposition by Xiong, Li, and Bao (2018). The authors then make individual predictions for each component, utilising seasonal naive method for the seasonality component and extreme learning machines (ELMs) for the trend and error components. The decomposed components are assumed in all of these contributions to be autonomous and subject to a variety of effects. Macroeconomic variables are one of the variables for the tactical window in sales forecasting.

Over a long period of time, researchers have projected sales forecasts in numerous industries while taking economic indicators into account. Sales projections for various industry items have shown varying bias dependence on external economic indicators. Our goal is to add external economic indicators and implement sales estimation on Dataco's smart dataset while assessing their dependencies.

3.0 Data Analysis:

The Data Set of Supply Chains used by the company ‘DataCo Global’ has used for research analysis .The dataset is available at both DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS | Kaggle and DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS - Mendeley Data as well. Provisioning, Production, Sales, Commercial Distribution are the important activities that allows correlation with structured data with unstructured data. The primary dataset is composed structurally of 51 columns and 180520 rows. The goal variable is to predict numeric attribute "sales" in the 37th column. The Days for shipment (actual), Benefit per order, Sales per customer, Category Name, Customer Country, Department Name, order date (DateOrders), Sales, Order Item Total, and Product Name are the data set's most crucial attributes.

A few biases will be incorporated from the world bank indicator collection dataset, which is accessible at World Bank Indicators Collection | Kaggle. The aim is to add additional features based on business insights to help the model for gain more sales regression performance and conduct comparative analysis.

The data type, missing values, and duplicate values have been identified through initial dataset analysis. The dataset now only has 24 numeric columns and 18 object columns after removing extraneous attributes like "Customer Email," "Product Status," and "Customer Password," etc. However, the further analysis of numeric columns revealed that there were 9 more features that were not necessary and were identical to other significant features. For instances "Order Item Cardprod Id" and "Product Category Id," which are both equivalent to "Category Id".
The distribution of "Sales" and "Product Price" across the dataset is shown in the graph. The majority of sales were seen to be focused between "0$+" and "300$," while the product price was slightly lower. Sales in this instance represent each transaction. The apparent explanation of this relationship is that sales value is an addition to product price and profit. Similarly, the following figure will be generated if the benefits per order distribution and order item profit ratio are plotted. ‘Benefit per order’ is the profit amount of each order. Here, one thing stands out: most businesses make a tiny profit, but the largest loss was over $4,000 and the highest profit was just over $1,000.

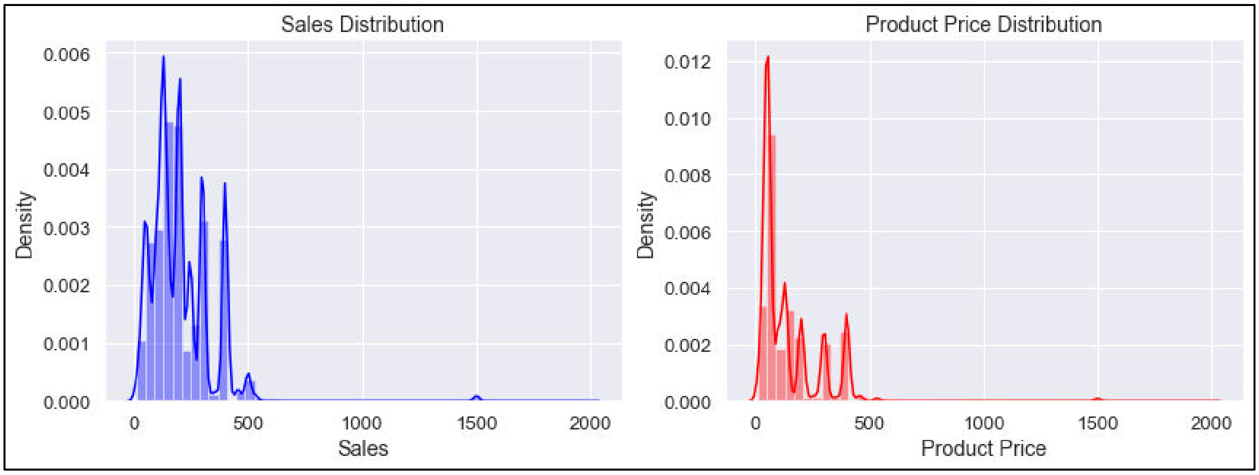

Figure 1: Distribution Plot of Sales and Product Price

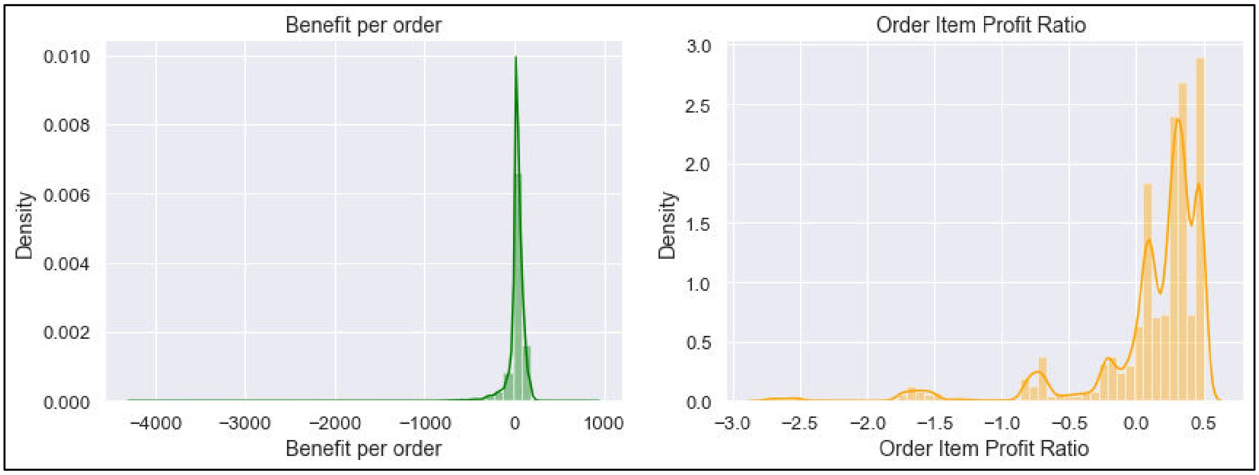

Figure 2: Distribution Plot of ‘Benefit per order’ and ‘Order Item Profit Ratio’

‘Order Item Profit Ratio’ are the ratios of ‘Benefits per order’ and ‘Sales per customer’. The value ranges mostly between "0" and "0.5," indicating that the highest profit threshold is a margin of 100% on any goods. The values of the "Order Item Profit Ratio" fluctuate rather than have a smooth distribution, which indicates that the profit margins of the various products are not comparable. Based on category, demand and place, and product price differ.

Finally, after removing same value numerical features, a correlation heatmap is plotted to display the pair wise relationship among columns.

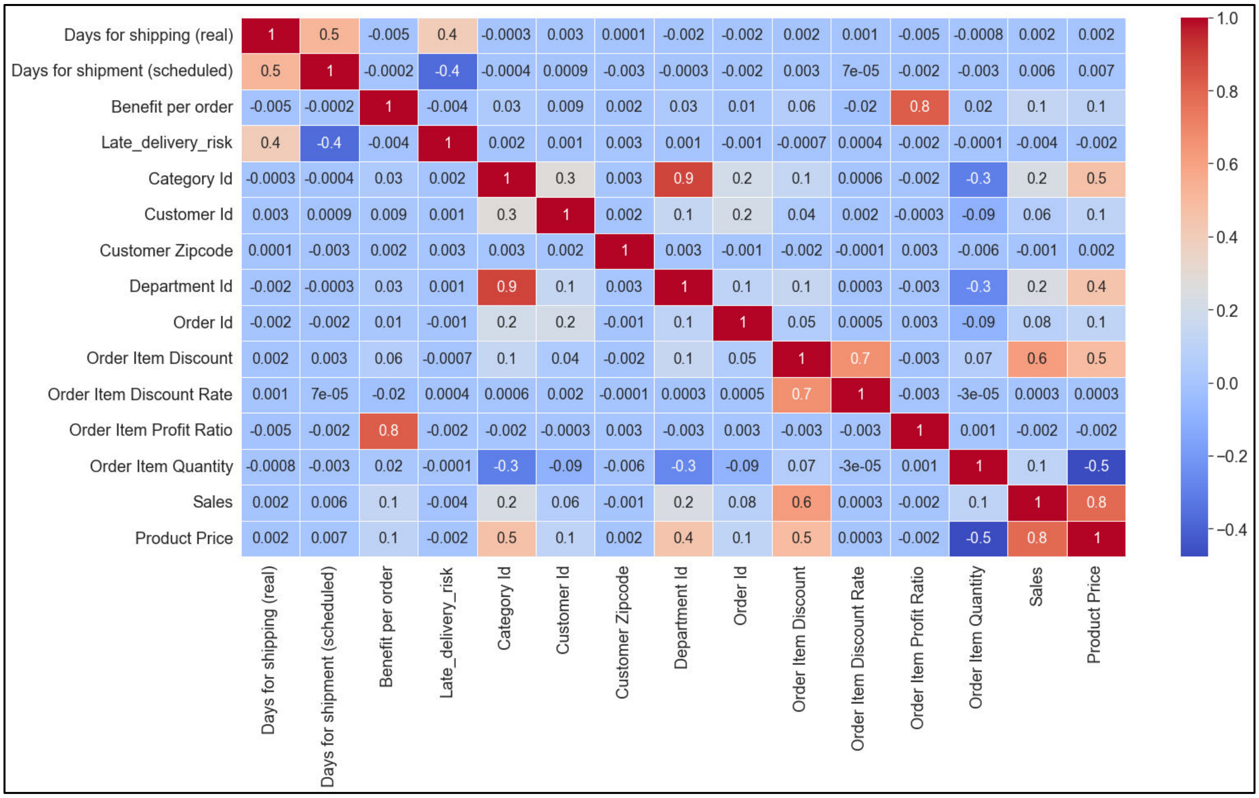

Figure 3: Heatmap of supply chain dataset after initial analysis

However, the greatest documented positive correlation is 0.9, showing that the values of "Category Id," "Department Id," and "Customer Id" are comparable. But according to the heatmap, two pairs had the most meaningful association. Benefits per order and the order item profit ratio come first, followed by sales and product price. However, there is also a positive correlation between "Sales" and "Order Item Discount," which is "0.6". Thus, there are more sales of things that are on sale. Sales and other features have not been found to be negatively correlated. Although there is a negative correlation between "Late delivery risk" and "Days for shipment," this correlation has no discernible impact on studies on regression models.

The next figure appears when to look more closely at how values between "Sales" and "Product Price" were distributed. There are numerous items in the small range with modest price values but relatively high sale values. Other than that, sale values are formulated with a uniform profit ratio. The graphical portrayal of these two features, however, clearly demonstrates the positive association.

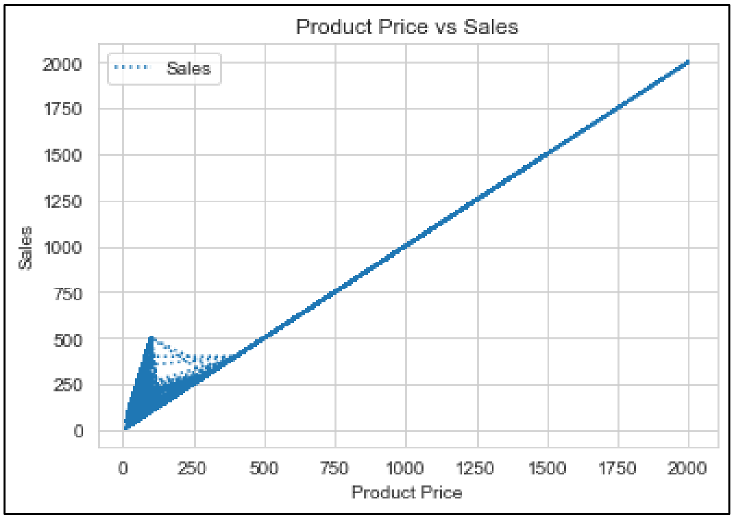

Figure 4: Scatter Plot of 'Sales' and 'Product Price'

Total sales for each market place have been mapped to learn more about sales based on market continents. Five markets have been divided up in the data. They are Africa, Europe, Latin America, Pacific Asia, the USCA (USA and Canada). Africa represented the lowest sales, which are estimated to be almost five times lower than those in Europe. Latin America has the second-highest sales volume. It was pretty intriguing that, although having similar financial strength, USA and Canada only generate less than half the sales of Europe. Africa has the harshest economic situation and the lowest sales, it can be said.

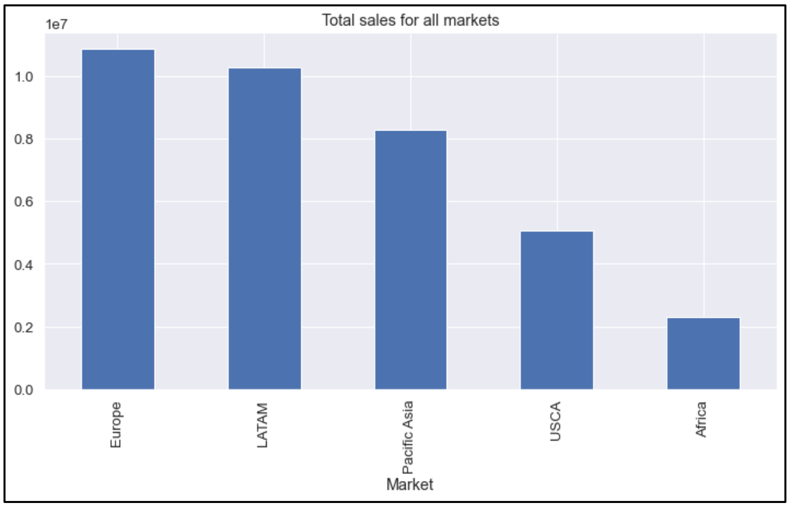

Figure 5: Bar Plot of 'Sales' and 'Market'

There are some intriguing insights revealed by the dataset if we look into sales more thoroughly in the context of the market by region. As an illustration, sales in western Europe are higher than in the north and south of the continent combined. Similar to western Europe, certain regions of central America have sales that are virtually as high. Contrarily, the majority of Asia and Africa both have comparable sales.

The graph that followed shows the total sales over time. The sales were fairly uniform from 2015 to 2017, then eventually steadily decreased in the following year.

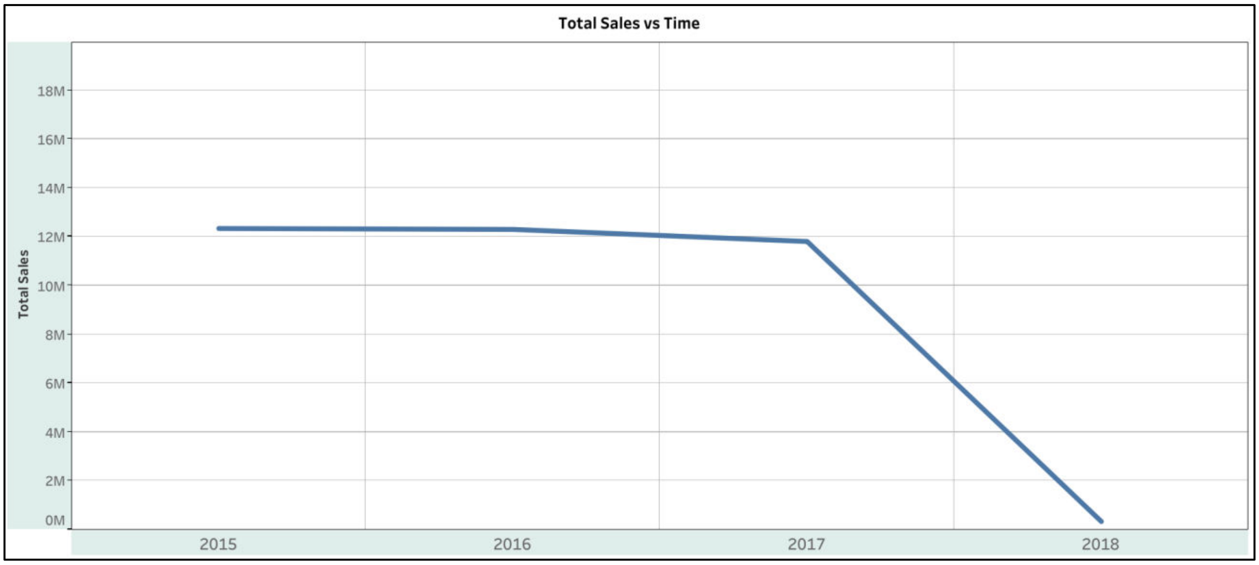

Figure 6: Total sales over time

Total sales from 2015 to 2017 have been created taking into account five different market continents in order to more thoroughly study total sales over time. The graph's five colours represent five different markets, and the line's width indicates the overall benefit per order for those sales during that time period.

This visualisation has revealed some intriguing data. In 2017, sales in the USCA, Pacific Asia, and Africa were incredibly modest. Thus, total sales began to decline in 2017 gradually. which the previous graph correlatingly shows. Nevertheless, USCA (USA & Canada) saw no sales in 2015 and a sharp decline starting in August 2016. The business strategy or the economic situation may be to blame. The pattern for Pacific Asia is also very intriguing as well.

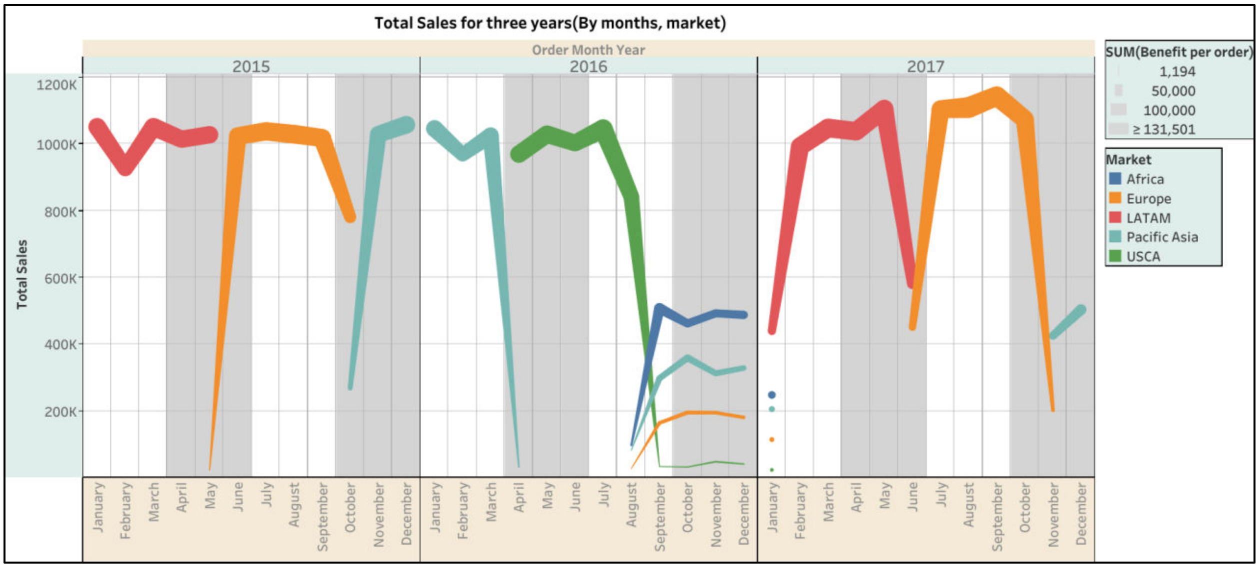

Figure 7: Total sales, Benefit per order and market over monthly time

In October 2015, there was a significant increase, but it decreased in April 2016 in a similar but opposite pattern. Similar to the United States, sales in Europe climbed from the middle of 2015 to the middle of 2016 before stopping. USCA and Africa, on the other hand, recorded no sales in 2016. However, Consumer was successful in launching a lucrative business at Latam in 2017. (Latin America). 2017's third quarter saw the biggest sales ever in Europe. However, benefits per order decrease if sales start to fall or rise.

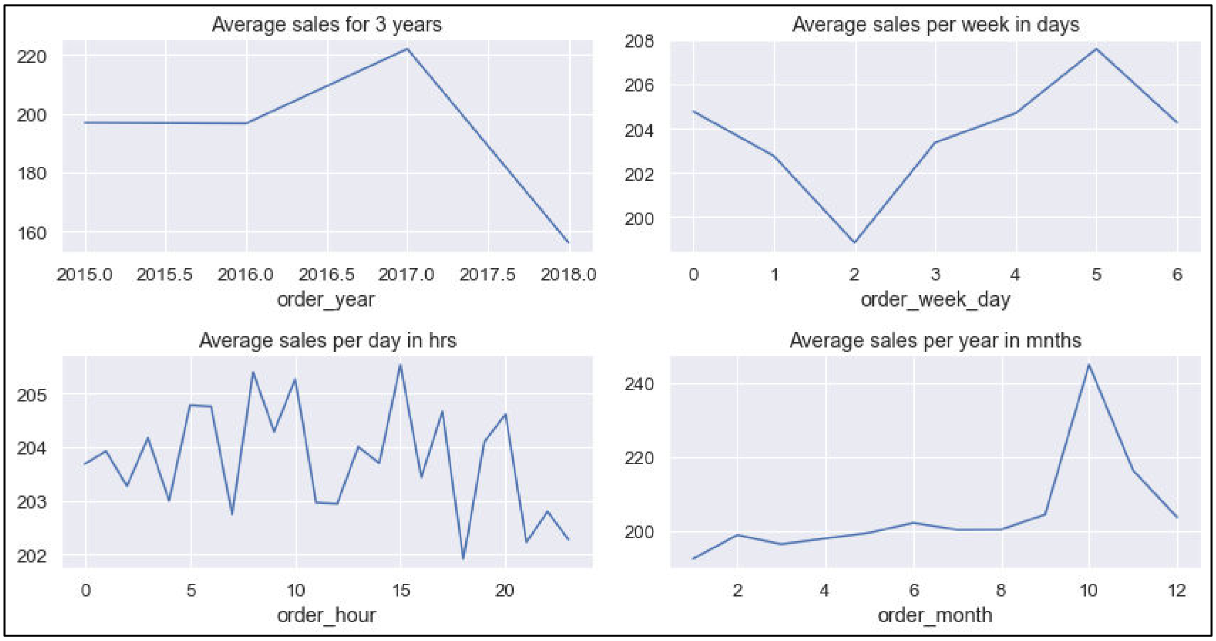

Figure 8: Average sales transaction over years, monthly, weekly and daily

To explore seasonal sales patterns, the average sale transaction count has been plotted by all years, months, weekdays, and daily hours. Beginning in 2017, average sales began to build up and then began to drop progressively in keeping with previously noted pattern. The day with the highest average sales was Saturday, while sales were at their lowest on Wednesday. Regardless of the time of day, the average sales kept consistent. When compared to the all three years, October had high sales.

3.0 Methodology & Implementation:

The dataset was carefully inspected after collection to look for valuable properties such as null values, duplicate values, repeating columns, and so on. The target attribute "Sales" was assessed along with other independent features. The attributes that don't correlate with the goal variable, "Sales," are identified and removed. There were two versions of the dataset developed since the research intended to empirically compare the predictions of "Sales" on the original dataset and an enhanced version of the dataset where an economic indicator was added. The attributes which contain categorical values has been decocted by lebel encoder for numeric representation.Two datasets were separated into training and testing sets once they were ready for usage in the data model.

During the preparation of the dataset, it was found that the object datatype attributes comprised a specific group of repeating categorical data. As an illustration of label encoding, we may say that the numerals "0," "1," and "2" could stand for Canada, Spain, or India in the phrase "Order Country." As a result, they used label encoding technology to transform the numerical value from "0" to "class number - 1". On the other hand, the scalling function 'minmax scaler' has been used to normalises the numeric features. Attributes value has been scaled from "0" to "1". The following formula will be used to calculate the final value, "X," which represents the normalized value of each feature.

$$X_{\text{Scaled}} = \frac{(X - X\text{min})}{(X\text{max} - X\text{min})}$$

The basic data for adding external economic indicators has been chosen as "Customer Country" feature. For the construction of the second edition of the dataset, the country's population, GDP, and unemployment rate in that particular timeframe have been added.

Two dataset has been splitted into training and testing data. Both of the training dataset has used to train data model. Data model has employed a variety of machine learning regression models.

Regression is a statistical method for simulating a target value using multiple variables. The main applications of this technique are forecasting and determining the causal connections between variables. The number of independent variables and the nature of the relationship between the independent and dependent variables are the main determinants of how regression algorithms differ.

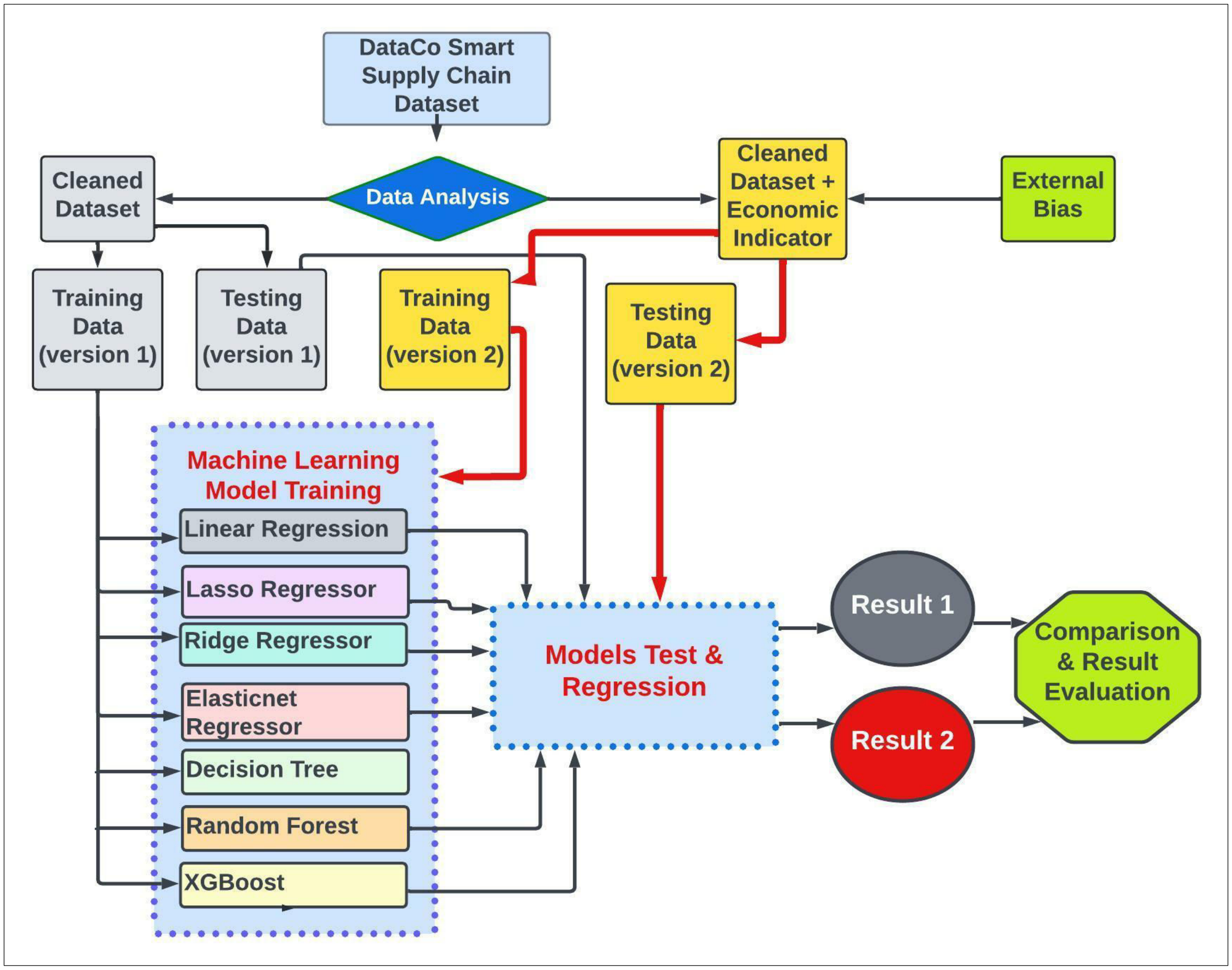

Figure 9: Data Model Architecture

For a given dataset, $X = \{X_1, X_2, \dots X_n, Y\}$

Whereas $X_1, X_2, \dots X_n$ each are individual features and Y is target. The regressor predicted value will be

$Y' = mX + C$

Here, ‘m’ is slope and C is unobserved noise. The aim is to be predict $Y^{/}$ values more closer to target variable Y in the training and testing data. Basic linear model has used at first.

Secondly, data model has used ridge regression whereas, a shrinkage amount from ‘0’ to any positive number denoted as ‘λ' used with square value of slope in the cost function. The purpose of this is to prevent overfitting and obtain the optimal fit line by penalising a certain amount for the change of steep slope to shallow slope in order to forecast test data more accurately. In order to determine the best fitting line taking into account of all features, gradient decsent methods were employed to obtain the minimal cost function.

Mathmatically, the cost function for ridge regression could be denoted as

$$\frac{1}{2n}\sum_{i=1}^{n}(\mathrm{y_i} - \sum \mathrm{x_{ij}}\beta_\mathrm{i})^2 + \lambda \sum_{j=1}^{p} {\beta_\mathrm{i}}^2$$

In this case, the forecasted target is $\mathrm{x_{ij}}\beta_\mathrm{i}$ and the target variable is $\mathrm{y_i}$ However, in this instance, incorrect noise has not been taken into account, and the slope of the best fit line is β_i. Since the dataset contains numerous features, graphical representation is not viable.

Lasso Regression, which is a regularisation technique and very similar to ridge regression has been used in data model also.It uses shrinkage whereas data points shrinks towards to mean value. Additionally, lasso regression is better for use when there are more features because it automatically selects features for better prediction. Regression using the lasso or L1 method is very helpful in preventing overfitting of the data. In this instance, the slope's absulate value was employed. The insignificant features would be eliminated because the target variable has no dependence on them because some of the feature co-efficients will get to "0" in this scenario.

Mathmatically it can be denoted as,

$$\frac{1}{2n}\sum_{i=1}^{n}(\mathrm{y_i} - \sum \mathrm{x_{ij}}\beta_\mathrm{i})^2 + \lambda \sum_{j=1}^{p} |\beta_\mathrm{i}|$$

Additionally, elastic net regression—a combination of lasso and ridge regression—was applied in the data model also. The aim is to have more better predicted value by avoiding overfitting and feature selection. In mathematics, the elastic net regression's mean square error has been employed as follows: $\frac{1}{2n}\sum_{i=1}^{n}(\mathrm{y_i} - \sum \mathrm{x_{ij}}\beta_\mathrm{i})^2 + \lambda_1 \sum_{j=1}^{p} \beta_\mathrm{i}^2 + \lambda_2 \sum_{j=1}^{p} |\beta_\mathrm{i}|$

λ_1 and λ_2 both are different hyperparameter used with squared and absulate values of slopes.

The data model also utilised a decision tree regressor for prediction. The decision tree has a root node, internal nodes, and leaf nodes that branch off of it. The leaf nodes include the final categorization categories or regressed values. A decision tree has served its purpose if it successfully "splits" data into homogeneous groups. A single splitting point is determined by a single attribute among all candidates by a decision tree algorithm. Data is split into two groups when it is partitioned at a split point: the groups to the left of the split point and the groups to the right of the split point.

Gini index cost function has calculated by $G = 1 - (p_1^2 + p_2^2)$, whereas p_1 and p_2 are proportion of instances considering by all instances of parent node.

For regression, root attributes and best conditional split point has determined by calculating lowest sum of residuals. The predicted value is an average value of instances of any pure leaf node.

The data model also includes a random forest regressor. It is a method of supervised learning that makes use of the ensemble learning strategy for regression. In order to produce predictions that are more accurate than those from a single model, the ensemble learning method integrates predictions from several machine learning algorithms. Bagged decision trees are inferior to random forests. The greed of decision trees like CART is a drawback. They pick the variable to divide on greedily in order to minimise error. Random forest generates predictions from all of the subtrees that are less correlated with one another as a result of altering the method used to learn the subtrees. The random forest technique, which changes this process, limits the learning approach to a random sampling of features to seek for. The number (m) of features that can be searched at each split point must be one of the algorithm's parameters.

Mathmatically it can be denoted as, m = $\frac{p}{3}$ whereas, 'm' is the number of selected features and 'p' is the value of input variables.

Lastly, data model has used extreme gradient boosting technique called 'Xgboost' for regression. It is an ensemble boosting technique, whereas, residuls which is average value from mean to target variable used initially for determine base model tree.

Similarity score and information gain has calculated for every single value of specific attribute.

Similarity Score = $\frac{(Sum\ of\ Residuals)^2}{Number\ of\ Residuals + \lambda}$

Gain = $\text{Left}_{\text{siilarity}}$ +$\text{Right}_{\text{similarity}}$ - $\text{Root}_{\text{similarity}}$

$\text{Root}_{\text{similarity}}$ for first variable will remain unchange for any feature. By selecting the greatest information gain threshold, the tree structure will be built. The subsequent attribute values of that specific record will determine the pure leaf threshold. All numerical features and basic models will be created using the same process.

The final output will be

Mean + $\lambda_1 T_1$ + $\lambda_2 T_2$....... $\lambda_n T_n$

Where as mean is on target variable, $\lambda_1, \lambda_2 .. \lambda_n$ are the learning rate of that particular feature and $T_1, T_2 .. T_n$ are the average value of the leaf node of that particular feature.

Finally, the data model visualises the assessment. After testing, mean absolute error (MAE) and root mean square error (RMSE) for both versions of the dataset were determined for evaluation. The model's accuracy is considered to be greater the lower the value.

Mathmatically they can denoted as

RMSE = $\frac{\sqrt{(y_i - y_p)^2}}{n}$

MAE = $\frac{|(y_i - y_p)|}{n}$

Whereas y_i is actual value and y_p is predicted value and n is number of instances. The best prediction will be their median target value if there are multiple examples with the same input feature values. MAE is not sensitive to outliers.On the contrary, The distance between the data points and the regression line is measured by residuals, and the spread of these residuals is measured by RMSE. In other words, it provides information on how tightly the data is clustered around the line of best fit.

4.0 Results & Evaluation:

The data model forecasted "Sales" using seven distinct machine learning algorithms. By taking into account both versions of the dataset, the mean absolute error (MAE) and root mean square error (RMSE) have been utilised to evaluate the results for the data model. It is apparent that for each of the evaluation matrices, linear regression produced the largest difference between test data and projected data. Four types of linear regression analysis have been taken into account by the data model. The four types are lasso regression, ridge regression, elasticnet regression, and linear regression.

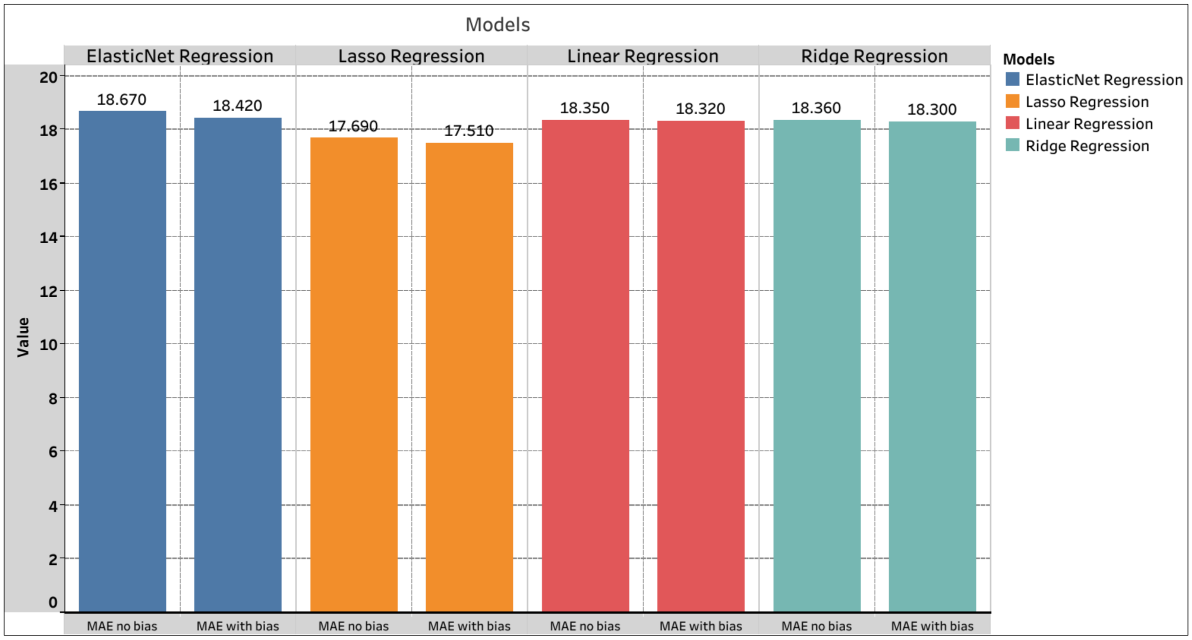

Figure 10: MAE score for elasticnet regression, lasso regression, ridge regression, and linear regression

All types of linear regression provide remarkably comparable results. For MAE, the range has been between "17.51" and "18.67" for both the core dataset and the modified dataset. While '0.0001' is utilised in elastic net, '0.1' is used as the learning rate for lasso and ridge regression. And, L1 ratio for elasticnet is set to "0.0001". On the other hand, the RMSE values for all four of these algorithms hover around 30.

Additionally, the data for population, unemployment rate, and GDP per capita were combined in the second data set. For MAE, it has been evidently demonstrated that incorporating economic variables resulted in slightly lower anticipated values than actual values. Even though there isn't much of a difference, the increased zeal is something to be taken seriously. The unique instance of the MAE score found via elastic net regression analysis.

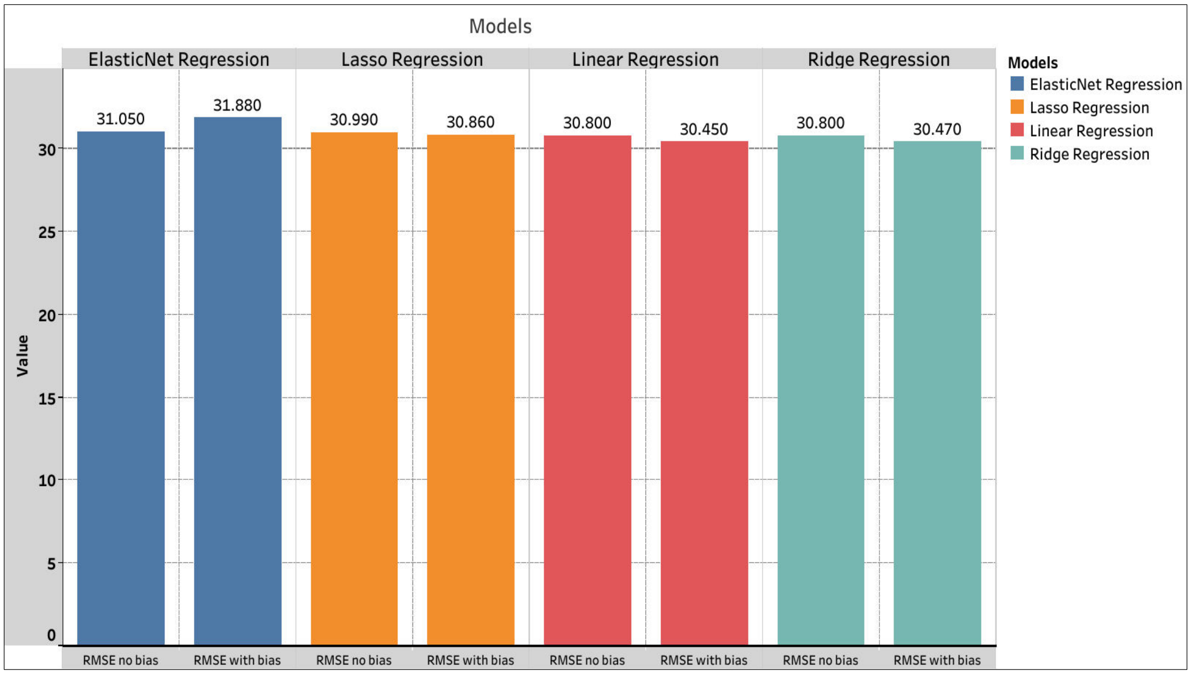

Figure 11: RMSE score for elasticnet regression, lasso regression, ridge regression, and linear regression

The root mean square error(RMSE) score has ranged from "30" to "32" when using the same techniques. Since the test and projected values in these two circumstances are identical, the mathematical criteria for inaccuracy have an impact on the score. For Both dataset versions', RMSE values are similarly quite close. In light of this, it may be said that the regression model based on linearity did not meet expectations.

Furthermore, when deploying data models, hyperparameter tuning was adhered to. In this instance, learning rate values ranging from "1" to "0.001" have been applied. When it reaches "0.1," the score has been optimised from "1" and is almost unaltered. Because the results from linear machine learning models were unsatisfactory, three additional algorithms have been chosen for regression analysis.

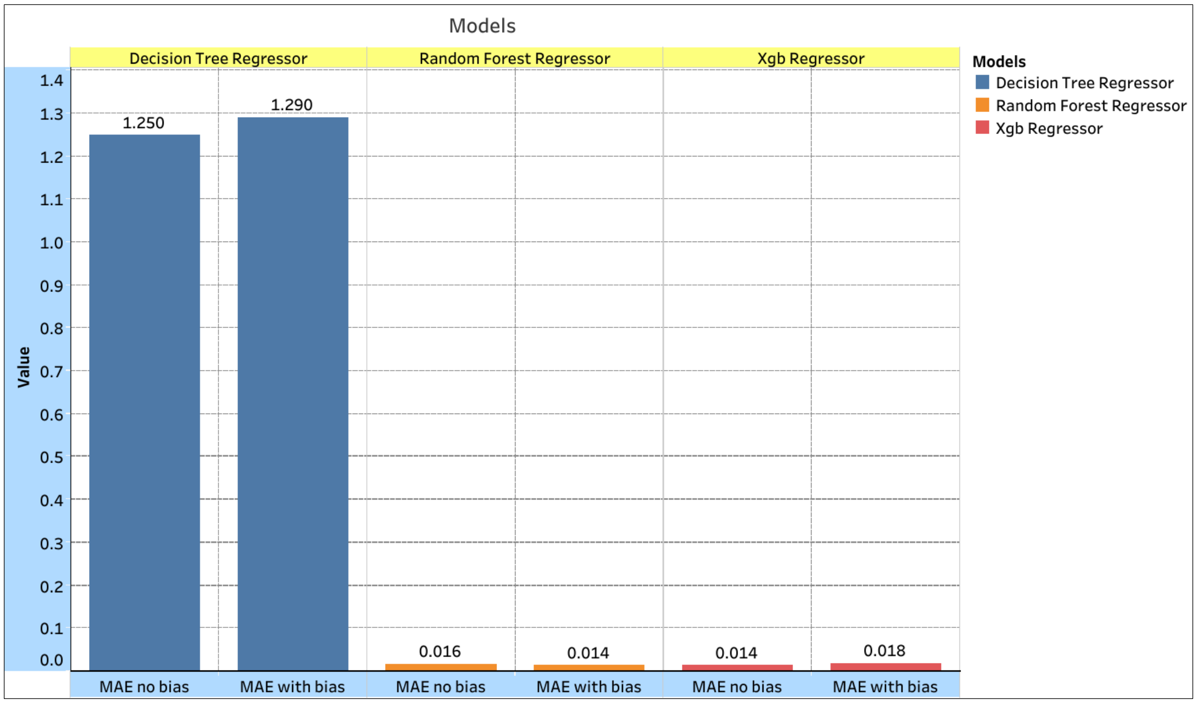

Figure 12: MAE score for Decision Tree regression, Random Forest regression and Xgboost regression

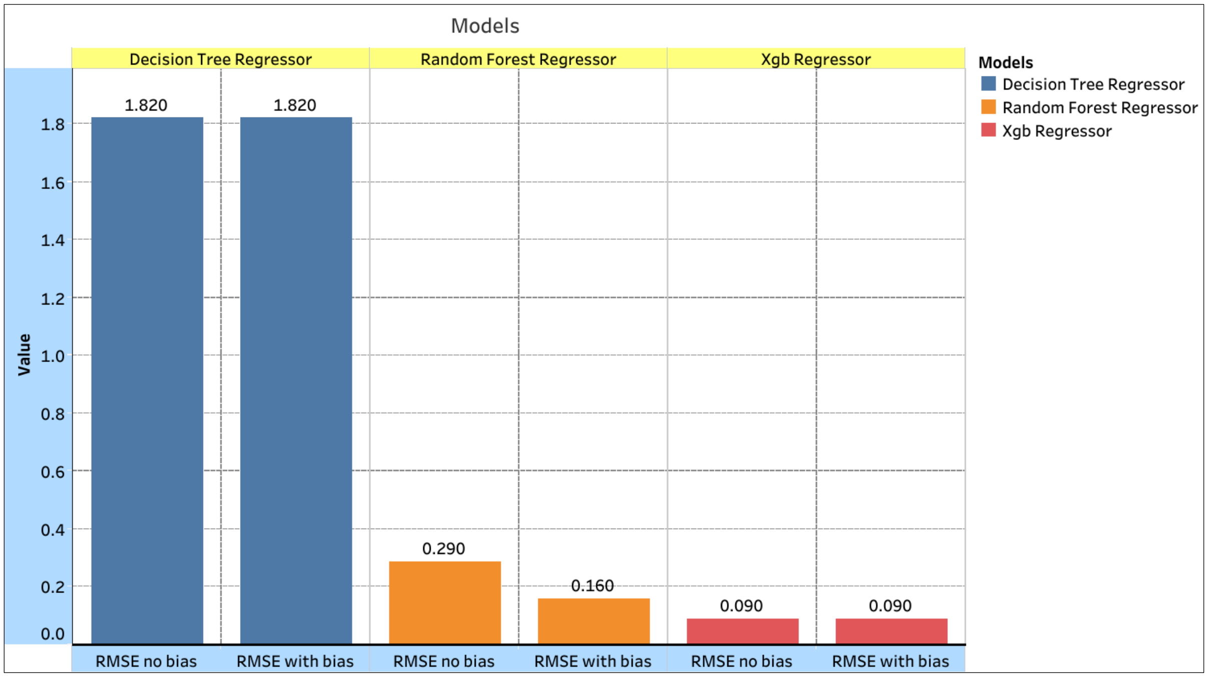

Figure 13: RMSE score for Decision Tree regression, Random Forest regression and Xgboost regression

A decision tree, which is a non-linear algorithm split the dataset into a tree-based structure and divide data into smaller subset. The projection's accuracy is rather high because it was based on a more limited similar subset. Regression's effect on this data model was evident. The main dataset's mean absolute error (MAE) is only projected as "1.25", however, the second version of the dataset has a little higher MAE of "1.29". Comparatively, it produced the same root mean square error (RMSE) score of 1.82.

Another machine learning approach that uses many decision trees to make projection judgments is called random forest. It combines all of the individual results from each decision tree for the same problem formulation to get the final outcome. As a result, random forest is by its very nature a powerful regression projector, as demonstrated by the data model projection. The precision is quite great in this case as indicated by the second version of the dataset's least mean absolute error (MAE), which is 0.014. It has been slightly higher for the main dataset, at "0.016". Additionally, it has been shown that the root mean square error (RMSE) is 0.29 and 0.16, respectively.

The data model has also used the extreme gradient boosting (XgBoost) algorithm, which provided the best accurate projection of all models. The mean absolute error (MAE) score was found to be '0.014' and '0.018' for the two data versions, respectively. Additionally, both values for RMSE are the same and are "0.09".

Therefore, based on observation, it is possible to draw the conclusion that XgBoost provided the most accurate forecast for "Sales" and that the addition of economic indicators to the primary dataset had little to no impact on that forecast.

6.0 Conclusion:

Sales forecasting is traditional yet essential for managers to create price plans and inventory policies since it helps them understand consumer preferences in retail industries. (Martínez-de-Alb´eniz & Belkaid, 2021; Sagaert et al., 2018; Sun et al., 2008). In this significant study, the open source smart supply chain dataset of DataCo was used to forecast sales by incorporating economic indicators. The following summarizes the contributions and results of the research:

-Population, unemployment rate and GDP per capita of customer country have been selected as representative economic indicators. These indicators have been added to the dataset, and the data model has been implemented for both the original dataset and the augmented dataset.

-There is no significant improvement on sales forecast after incorporating economic indicator. In addition, the most accurate predictor was Xgboost, which provided mean absolute error (MAE) score of 0.014. Even after accounting for economic indicators, the Random Forest Predictor still produced the same result.

In contrast to four linear models, seven regression models were used in the research, which did not yield the desired precision. As an alternative, tree-based and boosting machine learning model, on the other hand, provided striking accuracy in this situation. Other strong machine learning models, such as catboost, adaboost, lightgbm, etc., could also be used for further research. Since deep learning models were not tested in this study, alternative neural network models such ANN, CNN, LSTM, and GRU could represent a fresh line of inquiry into the same issue. Before deploying the data model, some of the unnecessary attributes were selected to be removed during data wrangling and processing. Their presence would allow for investigation using both permutations and combinations.

The research used three economic indicators, but it could have chosen any number of combinations of economic indicators based on customer and order data instead. The classic regression model was used to project "sales," however it may be improved by combining ensemble machine learning and deep learning models. The study concentrated on in-depth sales analysis when it should have concentrated on overall sales, allowing time series analysis to be used as a forecasting guarantee. In order to attain greater accuracy in the sales projection, it can be concluded that the data model and dataset might be constructed using a variety of structures and methodologies.

Reference:

1. Boone T., Ganeshan R., Jain A., Sanders N.R. (2019), "Forecasting sales in the supply chain: Consumer analytics in the big data era", *International Journal of Forecasting*, *35*(1), 170-180.
2. Chih-Hsuan Wang C.-H.,2022, "Considering economic indicators and dynamic channel interactions to conduct sales forecasting for retail sectors",*Computers & Industrial Engineering*,Vol. 165,pp. 107965,
ISSN 0360-8352, https://doi.org/10.1016/j.cie.2022.107965.
3. Christopher, M. (2005) *Logistics and supply chain management, creating value-adding networks*, Financial Times Prentice Hall, Harlow, 3rd ed.
4. Dalkey, Norman; Helmer, Olaf (1963). "An Experimental Application of the Delphi Method to the use of experts". *Management Science*. 9 (3): 458–467. doi:10.1287/mnsc.9.3.458. hdl:2027/inu.30000029301680
5. Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics*, *170*, 97-135.
6. Feizabadi, J. (2022). Machine learning demand forecasting and supply chain performance. *International Journal of Logistics Research and Applications*, *25*(2), 119-142.
7. Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*.
8. Kozlenkova, Irina; et al. (2015). "The Role of Marketing Channels in Supply Chain Management". *Journal of Retailing*. 91 (4): 586–609. doi:10.1016/j.jretai.2015.03.003
9. Martínez-de-Albéniz, V. and Belkaid,A. (2021),"Here comes the sun: Fashion goods retailing under weather fluctuations",*European Journal of Operational Research*, Vol. 294, No. 3, pp. 820-830, ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2020.01.064.

10. Min, S., Zacharia, Z.G. and Smith, C.D. (2019), "Defining supply chain management: in the past, present, and future", Journal of Business Logistics, Vol. 40 No. 1, pp. 44-55.

11. Mishra, D., Gunasekaran, A., Papadopoulos, T. and Dubey, R. (2018), "Supply chain performance measures and metrics: a bibliometric study", *Benchmarking: An International Journal*, Vol. 25 No. 3, pp. 932-967. https://doi.org/10.1108/BIJ-08-2017-0224

12. Ord, K., Fildes, R. A., & Kourentzes, N. (2017). Principles of business forecasting.

13. Sagaert, Y. R., Aghezzaf, E. H., Kourentzes, N., & Desmet, B. (2018). Tactical sales forecasting using a very large set of macroeconomic indicators. *European Journal of Operational Research*, *264*(2), 558-569.

14. Sa-Ngasoongsong, A., Bukkapatnam, S. T., Kim, J., Iyer, P. S., & Suresh, R. P. (2012). Multi-step sales forecasting in automotive industry based on structural relationship identification. *International Journal of Production Economics*, *140*(2), 875-887.

15. Smeekes, S., & Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International journal of forecasting*, *34*(3), 408-430.

16. Sun, Z. L., Choi, T. M., Au, K. F., & Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems*, *46*(1), 411-419.

17. Verstraete,G., Aghezzaf, E. and Desmet,B. (2020), "A leading macroeconomic indicators' based framework to automatically generate tactical sales forecasts",*Computers & Industrial Engineering*, Vol. 139, pp. 106169. https://doi.org/10.1016/j.cie.2019.106169.

18. Xu, X., McGrory, C. A., Wang, Y. G., & Wu, J. (2021). Influential factors on Chinese airlines' profitability and forecasting methods. *Journal of Air Transport Management*, *91*, 101969.

19. Xiong, T., Li, C., & Bao, Y. (2018). Seasonal forecasting of agricultural commodity price using a hybrid STL and ELM method: Evidence from the vegetable market in China. *Neurocomputing*, *275*, 2831-2844.