Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

Optimization Of Supply Chain System Using ML Algorithms

Shriyash Band
Computer Department
AISSMS IOIT
Pune, India
shriyashband@gmail.com

Aditya Pokharkar
Computer Department
AISSMS IOIT
Pune, India
aditya.d.pokharkar@gmail.com

Anish Dhawalikar
Computer Department
AISSMS IOIT
Pune, India
anishdhawalikar@gmail.com

Himank Tyagi
Computer Department
AISSMS IOIT
Pune, India
tyagihimank28@gmail.com

Prof. Dr. Sarika Zaware
Computer Department
AISSMS IOIT
Pune, India

Abstract - Traditional supply chain management in the present day faces a number of difficulties, including those related to Purchase Probability, Sales Forecasting, and Product Recommendation. The majority of research papers provide solutions to specific sets of problems. To address these issues, a solution is provided that compares the most effective models when used in combination with a supply chain system. The proposed solution consists multitude of ML algorithms such as Random Forest, Gaussian Naive Bayes, Lasso Regressor, XGBoost, Apriori, FP-Growth, KNN, K-Means, and Bayesian Linear Regressor. Real-time Sales Data is used for the prediction and analysis of the supply chain system. Comparison of different algorithms have been evaluated and the most efficient algorithm among them is attained.

Keywords— Sales Forecasting, Product Recommendation, Purchase Probability, Supply Chain System, Machine Learning, Optimization.

I. INTRODUCTION

A supply chain is a network of relationships that enables distribution of various goods and services both locally and globally. The supply chain employs an enumerated structure made up of an effective system and beneficial information based media to validate the delivery of products from raw materials to final consumers. Nowadays, still many of the manufactures and vendors use out dated techniques to communicate with demand and supply problems [1].

A subfield of artificial intelligence (AI) and computer science called machine learning concentrates on using data and algorithms to simulate how humans learn, gradually increasing the accuracy of the system. Machine learning is currently being implemented in the supply-chain industry to facilitate the necessities of modern businesses [2].

Machine learning's significance in supply chain management is as follows:

- Predict sales demand according to previous data. As a consequence, the Inventory can be better maintained in accordance with customer demand.
- Recommending products based on customers' past purchasing patterns which results in enhancing the sales.
- Purchase probability of customers based on the purchasing patterns results in maintaining profitable businesses.

[3, 4]Hence the ML techniques provide smooth and efficient supply chain process in modern systems. Therefore, the paper studies various Machine Learning algorithms which are extensively used for demand and supply communication. The goal of this research is to identify the optimum algorithms when evaluated against the competition and provides resonant solution.

II. RELATED WORK

The applications of machine learning are discussed in [5]. It implies that demand planning and prediction and production are the main uses of machine learning in the supply chain. Compared to clustering, RL, and classification, the use of ML Regression models in demand planning is greatest, at above 45%. With a percentage rate of almost 30%, demand planning and prediction is where classification ML models are used

most frequently. According to the poll, supervised machine-learning algorithms outperform other types of machine-learning algorithms such as unsupervised Learning and Reinforcement Learning by a margin of more than 70%. An analysis of demand forecasting in [6] compares the Naive Bayes, Decision Tree, and KNN algorithms. The study provides an overview of the performance and bias of the models. Naive Bayes Classifier achieves an overall accuracy score of 58.92%, which is the highest compared to the other two algorithms, which earn accuracy scores of 28.57% and 35.71% for DT and KNN, respectively. [7] uses ensemble learning methods to predict the likelihood that consumers will buy retail sales items. It is discovered that combining XGBoost, RF, and CNN increases their effectiveness in comparison to using each model separately. The highest efficiency obtained by combining all three algorithms (CNN+XGB+RFC) was 87.82%. The proposed ensemble approach produced the highest accuracy of all, at 88.84%. The RFC+XGB combination among them produced a score of 78.9%, which is also respectable when you consider regional and small-scale businesses.

[8] The article uses a variety of methods and models from the literature review to categorize the demand sensing requirement. Regressive models for demand forecasting, including ARIMA, NARX, XGBoost, DT, and RF, are also covered, along with how these models can be revised and modified for the best results. The forecasting and the procurement in SCM are both the main topics of the article. The ECNN model produced the best and most precise outcomes out of all the demand management algorithms offered.

[9] sheds light on how machine learning applications can be used to forecast demand and how this could ultimately help to increase the efficiency and effectiveness of the supply chain system. It examines how machine learning and data analytics are used in business sectors, as well as the factors that lead to disruptions.

III. Dataset Description

This dataset was gathered from a food production business with the intention of predicting purchase likelihood. It has 18 columns and 12330 rows. *Table 1*is a description of this dataset's attributes.

Table 1 - Prediction Dataset

Characteristic	Characterization
Administrative	This is the total number of pages the user visited that were of this category (administrative).
Revenue	a Boolean indicating whether the customer finished the transaction or not.
Administrative_ Duration	This reflects the amount of time spent on sites in this category.
Informational	The user viewed these many pages of this kind (informational pages).
Informational_D uration	This represents the time spent on sites in this category.
Weekend	Whether the session is on a vacation is indicated by a Boolean.
VisitorType	A string indicating whether a user is a first-time visitor, a returning visitor, or something
Browser	A number that represents the user's browser when they viewed the website.
Month	Incorporates the pageview's month in string form.
SpecialDay	This value indicates how close the browsing date was to notable occasions or holidays (such as Mother's Day or Valentine's Day), when it was more likely that the transaction would be completed. Below is more information on the formula used to determine this value.
OperatingSyste ms	An integer value that represents the user's operating system at the time the page was viewed.
TrafficType	An integer value that indicates the user's traffic classification.
PageValues	The successful completion of an online purchase.
ExitRates	The proportion of website pageviews that end on that particular page.
ProductRealted	The user visited many pages of this kind (product-related pages).
ProductRelated_ Duration	This represents the time spent on pages in this category.
BounceRates	The proportion of visitors who arrive on that page of the website and leave without performing any further tasks.
Region	An integer indicates the area in which the user is situated.

This dataset was gathered from a website called Kaggle which is used for product recommendation. There are 7 columns and 522065 rows in it. In *Table 2*, attributes are described.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

Table 2 - Recommendation Dataset

Trait	Explanation
BillNo	Each payment is given a 6-digit number.
Country	Country-specific identifier for each client.
Date	When each transaction was created, including the date and time
Price	Product cost.
Quantity	The number of each merchandise in a single transaction.
CustomerID	Each client is given a 5-digit number.
Itemname	Tag line

The dataset is from a nearby producer of edible goods. The training sample is limited to 8523 rows and 12 columns in size. Specifically, the forecasting of sales is done using this info. As a result, no particular data pre-processing model is necessary because the data is already kept in a suitable format. The attributes of the dataset are shown in *Table 3*.

Table 3 - Sales dataset description

Features	Description
Item Identifier	It is the special merchandise ID.
Item Fat Content	It will indicate whether or not the food is minimal in fat.
Item-MRP	The price chart for the item
Outlet-Establishment Year	When the first entrances to the store were unlocked.
Outlet-Size	The entire area that a supermarket takes up.
Outlet-Location	The type of region where the company is located.
Item -Visibility	The portion of the total viewing space allotted to that specific item out of everything in the store.
Item -Type	What category does the product fit into?
Item Weight	It will contain the item's weight.
Outlet-Type	The store is simply a grocery or supermarket.
Item-Outlet-Sales	Sales of the product in the initial store
Outlet-Identifier	An individual spot number

IV. Solution Design

The proposed system architecture is depicted in the image below, with emphasis placed on various algorithms and the dataset used to produce the results. Here, after implementing the algorithms, accuracy, mean squared error, root Mean square error, and maximum residual error are determined. The

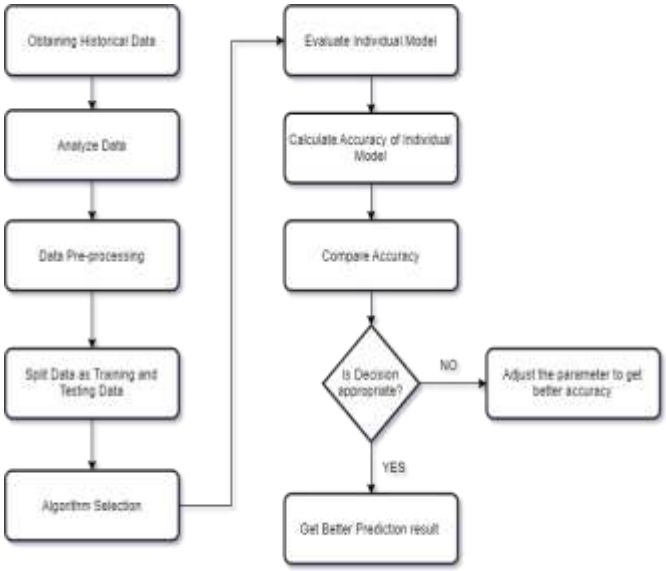

Figure 1 System Architecture

algorithm that delivers the highest result is then used [10, 11].

A. Sales Forecasting

1. XGBoost

We used training data to fit the model using XGB Regressor(objective='req:squarederror,colsample_bytree=1,learning_rate=0.05,max_depth=5,n_estimators=500,subsample=1,min_child_weight=5).

The following list and description of XGBRegressor's parameters:

i. Squared error- squared loss regression
ii. Colsample_bytree- column subsample ratio used to build each tree.
iii. Learning_rate- Following each phase of boosting, we may immediately obtain the weights of new features
iv. Max_depth- The highest part of a tree. The model will become more complex and prone to overfitting if this value is increased.
v. N_estimators- quantity of trees.
vi. Subsample- a ratio of the training instances' subsamples.
vii. Min_Child_Weight- The bare minimum of instance weight required in a child

The dataset's Item_Outlet_Sales attribute is used as a feature for predicting sales. By calculating the values of RMSE, r2 score, ME, and Maximum Residual Error, model performance is shown [12].

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

2. Bayesian Linear Regression

Item_Outlet_Sales is a dependent variable in our sales model that is employed in sales forecasting. Formulas for linear regression can be represented as follows:

$$Y = o1x1 + o2x2 + \cdots + onxn$$

[13]The r2 score, RMSE, MSE, and Maximum Residual Error are used to determine the model performance after the training dataset has been used to fit the model.

3. Random Forest

Here, sales forecasting is carried out using the Random Forest Classifier method. We fit the training data into the Random Forest method with (n_estimators=100, criterion='squared error'). The "squared error" criterion uses the mean of each terminal node to minimize the L2 loss and represents mean squared error, which is equivalent to variance reduction. For predicting a product's sales, we now employ the Item_Outlet_Sales functionality. To evaluate the actual and anticipated results, forecast the sales based on the test data and store the expected values in output_predicted.csv, MSE, RMSE, and Max Residual Error can be used to assess the model's performance [14, 15].

4. Lasso Regressor

L1 penalty formula:

$$l1\_penalty \ = \ sum\ j = 0\ to\ p\ abs(beta\_j)$$

This penalty, known as "Least Absolute Shrinkage and Selection Operator regularization," (LASSO) can be added to the cost function for linear regression.

$$Lasso\_loss \ = \ loss \ + \ (\lambda \ * \ l1\_penalty)$$

The " λ " hyperparameter is used to adjust how much of the penalty is added to the loss function. As a result, Item_Outlet_Sales in our situation immediately contribute to discovering outcomes. The RMSE, MSE, and Maximum Residual Error are used to determine the model performance after the training dataset has been used to fit the model.

B. Purchase Probability

1. K Nearest Neighbor

Here, we utilize the KNN algorithm to forecast purchases. The KNN technique has been used using (n_neighbors=2). The number of neighbors who will vote for the target point's class is indicated here by the variable n_neighbors. Since this method is used to predict the probability that a purchase will be made, the Revenue feature is used to determine the anticipated outcome. Accuracy, precision, recall, and f1-score are used to show this model's performance.

2. Gaussian Naive Bayes

Naive Bayes Formula

$$P(A|B) = \big(P(B|A)P(A)\big)/P(B)$$

Where P(A|B) is Posterior probability: Chance that hypothesis A will be true given the observed occurrence B.

P(B|A) is Likelihood probability: The chance of the information provided that a hypothesis is likely to be correct.

P(A) is Prior Probability: The hypothesis' probability before looking at the evidence.

P(B) is Marginal Probability: Likelihood of the evidence.

In our example, a collection of all the characteristics that do not depend on one another is used as the "evidence" parameter to forecast the outcomes, and the dependent attribute revenue is used as the "labels" parameter dependent attribute to train the algorithm. Once the model has been trained, test the data against the model to determine its correctness. Then, determine the F1 score, true positive rate, and true negative rate.

3. K-Means

The K-Means algorithm for purchase probability is presented here. The K-Means technique has been used using (n_clusters=2,random_state=None). The training data has been fitted with the following conditions. The number of clusters and centroids is represented by n_clusters, while random_state controls the production of random numbers. Here, we have taken advantage of the Revenue function to determine whether or not a particular product would be purchased. Accuracy, precision, recall, and f1-score are used to illustrate this model's performance.

4. Random Forest

For the purpose of generating results for purchase probability, we have also utilized the Random Forest Classifier in this instance.The Random Forest Classifier technique has been employed with (n_estimators=100). The number of trees that will be produced by the algorithm from which we derive the

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

final result is represented by the variable n_estiamtors. Here, we take advantage of the Revenue function to determine whether or not a particular product would be purchased. Accuracy, precision, recall, and f1-score are used to illustrate this model's performance.

C. Product Recommendation

1. Apriori

Apriori mentions:If there is a chance that item X is uncommon then: When X is not frequent, then P(X) < minimum support threshold. If P (X+A) < minimum support threshold, then X+A is not frequent, where A is a part of the set of items.If the value of an itemset set is less than the minimum support, then all of its supersets will also be less than the minimum support and can be disregarded. The Anti Monotone property is the name of this characteristic. Parameters: min_threshold = 1, min_support = 0.02, min_lift = 1.9, min_confidence = 0.2, metric = "lift" [16].

2. FP-Growth

Unlike the Apriori algorithm which uses candidate key generation, the FP-Growth algorithm identifies common sets of items, which increases the overall efficiency of the algorithm. It implements the divide & conquers strategy in many of the several instances. The implementation of the FP-Growth algorithm is accompanied by using FP-Tree (Frequent Pattern Tree) data structure, which stores the information about the associations among the itemsets. In the study performed, below are the parameters set. Parameters: min_threshold = 1, min_support = 0.02, min_lift = 1.9, min_confidence = 0.2, metric = "lift".

V. Results & Discussion

A. Sales Forecasting

Bayesian Linear Regression: The outcomes of Bayesian Linear regression with different parameters are shown in *Table 4*.

Table 4 - Sales Forecasting Results

Model	R2 score	MSE	RMSE	MAX RESIDUAL ERROR
Bayesian Linear Regression	0.8345	275949.0774	525.3085	2452.0920
Random Forest	0.9010	165002.5160	406.2050	3159.2726
XGBoost	0.9149	141930.0915	376.7361	3207.8066
LASSO Regression	0.8349	275408.3043	524.7935	2470.0122

B. Purchase Probability

Precision, recall, F1-score, and support have all been used to assess the model's performance. Various algorithm's accuracy, macro averages, and weighted averages have been compared and examined in order to select the algorithm that predicts purchase probability the best. We find that the accuracy for K-Means is 81.46%, RF is 90.40%, Naive Bayes is 84.71%, and KNN is 82.96%, respectively.

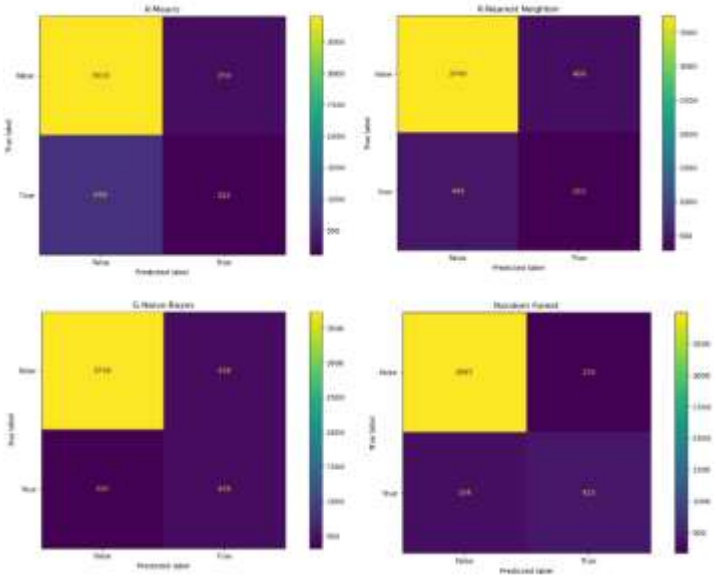

Figure 2 Purchase Probability Results

C. Product Recommendation

Table 5 – FP-Growth v/s Apriori

Parameters	FP-GROWTH	APRIORI
Structure for storage	Tree Centered	Array Centered
Searching Method	Divide and Conquer	BFS
Execution Time (min support = 0.02)	8 s	150 s
Amount of databases scan	2 scans	k+1 scans

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

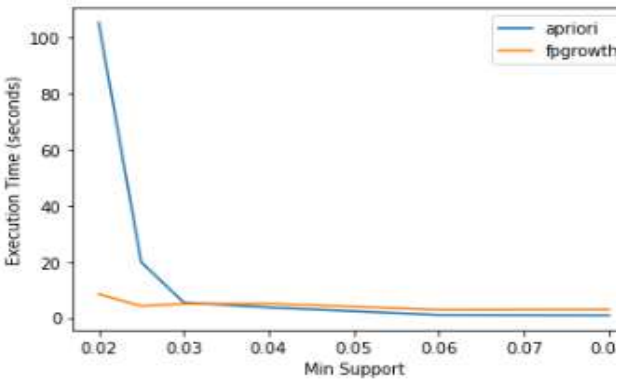

Figure 3 Execution Time Analysis

Similar to the results discussed in [17, 18]. The performance of FP-Growth is evaluated as efficient against Apriori algorithm. This experiment is performed on dataset which has near about 20,210 unique transactions. FP-Growth is more efficient in its execution time when experimented with the transactional database provided. Due to the FP-tree structure and divide & conquer strategy of FP-Growth, the performance measure is increased drastically for a huge amount of data sets. It is seen in the experiment that after a certain threshold value of minimum support the execution time of apriori is exponentially high because of its candidate key generation strategy. Moreover, the memory requirement of the apriori algorithm is also seen to be drastically increased, as it uses the brute force method for frequent pattern itemset mining.

VI. Conclusions

After studying multiple algorithms for each model, we discover that XGBoost, which has the accurate results of all the algorithms, is appropriate for our system in Sales forecasting. Factors like r2 score, RMSE, MSE, and Maximum Residual Error also yield excellent outcomes. We've found Random Forest to be the most useful for the Purchase Probability model, as it delivers 90% accuracy. FP Growth is ideal for Product Recommendation because it implements in 8s and generates the expected outcomes in our case. Based on environmental shifts over time and the dataset being used, the findings above may vary. [19, 20] The study of numerous deep learning and CNN methods for BigData will be the focus of future research because fascinating details may be discovered through multiple hidden layers.

References

[1] S. Ziang and B. Li,, "Research on the Application of Big Data Technology in Electronic Commerce Supply Chain," *International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS),* pp. 691-694., 2020.

[2] Adit Kudtarkar, Danish Shaikh, "Applications of Machine Learning Techniques in Supply Chain Management," *International Journal of Creative Research Thoughts (IJCRT),* 2021.

[3] Sarker, I.H., Kayes, A.S.M. & Watters, P., "Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage.," *Journal of Big Data 6,* 2019.

[4] R. Harikrishnakumar, A. Dand, S. Nannapaneni and K. Krishnan,, "Supervised Machine Learning Approach for Effective Supplier Classification," *18th IEEE International Conference on Machine Learning And Applications (ICMLA),* pp. 240-245, 2019.

[5] Mohamed-Iliasse, Mahraz and Loubna, Benabbou and Abdelaziz, Berrado;, "Is Machine Learning Revolutionizing Supply Chain?," *2020 5th International Conference on Logistics Operations Management (GOL),* pp. 1-10, 2020.

[6] Arif, Md. Ariful and Sany, Saiful and Nahin, Faiza and Rabby, Akm Shahariar Azad, "Comparison Study: Product Demand Forecasting with Machine Learning for Shop," in *8th International Conference on System Modeling & Advancement in Research Trends*, Moradabad, India, 22nd–23rd November, 2019.

[7] Sharma, Archika and Omair Shafiq, M., "Predicting purchase probability of retail items using an ensemble learning approach and historical data," *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) ,* pp. 723-728, 2020.

[8] Pham, Vy and Maag, Angelika and Senthilananthan, Sunthatalingam and Bhuiyan, Moshiur, "Predictive analysis of the supply chain management using Machine learning approaches: Review and Taxonomy," *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA),* pp. 1-9, 2020.

[9] Aamer, Ammar and Eka Yani, Luh Putu and Alan Priyatna, I Made , "Data Analytics in the Supply Chain Management: Review of Machine Learning Applications in Demand Forecasting," *Operations and Supply Chain Management: An International Journal,* vol. 14, pp. 1-13, 2020.

[10] S. Kulshrestha and M. L. Saini, "Study for the Prediction of E-Commerce Business Market Growth using Machine Learning Algorithm," *5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE),* pp. 1-6, 2020.

[11] Haifeng Lin, Ji Lin, Fang Wang, "An innovative machine learning model for supply chain management," *Journal of Innovation & Knowledge,* 2022.

[12] Nikolas Ulrich Moroff, Ersin Kurt, Josef Kamphues, " Machine Learning and Statistics: A Study for assessing innovative Demand Forecasting Models," *Procedia Computer Science,* pp. 40-49, 2021.

[13] Sunitha Cheriyan and Shaniba Ibrahim and Saju Mohanan Mohanan and Susan Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," *2018 International Conference on Computing, Electronics \& Communications Engineering (iCCECE),* pp. 53-58, 2018.

[14] Ghosh, Soumi and Banerjee, Chandan, "A Predictive Analysis Model of Customer Purchase Behavior using Modified Random Forest Algorithm in Cloud Environment," *2020 IEEE 1st International Conference for Convergence in Engineering (ICCE),* pp. 239-244, 2020.

[15] Aarya Pratap Singh Rana, "Supply Chain Optimization using Machine Learning," *International Research Journal of Engineering and Technology (IRJET),* pp. 7110-7114, 2020.

[16] Shah, N., Solanki, M., Tambe, A., & Dhangar, D., "Sales Prediction Using Effective Mining Techniques," *International Journal of Computer Science and Information Technologies (IJCSIT),* pp. 2287-2289, 2015.

[17] Prashasti Kanikar, Twinkle Puri, Binita Shah, Ishaan Bazaz, Binita Parekh., "A Comparison of FP tree and Apriori," *International Journal of Engineering Research and Development,* pp. 78-82, 2014.

[18] Sumit Aggarwal and Vinay Singal,, "A Survey on Frequent pattern mining Algorithms," *International Journal of Engineering Research & Technology (IJERT),* pp. 2606-2608, 2014.

[19] H. Bousqaoui, S. Achchab and K. Tikito,, "Machine learning applications in supply chains: An emphasis on neural network applications," *3rd International Conference of Cloud Computing Technologies and Applications (CloudTech),* pp. 1-7, 2017.

[20] HongJing Liu., " Forecasting Model of Supply Chain Management Based on Neural Network," *Proceedings of the 2015 International Conference on Automation Mechanical Control and Computational Engineering,* pp. 179-183, 2015.