Stockout prediction for Inventory Management using Machine Learning algorithms

Divyam Kumar
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Pune, India
divyam.kumar.btech2021@sitpune.edu.in

Harsh Ratna
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Pune, India
harsh.ratna.btech2021@sitpune.edu.in

Manvi Kalia
Department of Engineering and Supply Chain Management
Aston University
Birmingham, United Kingdom
220379012@aston.ac.uk

Sina Seyfadini Kuhbanani
Department of Engineering and Physical Sciences
Aston University
Birmingham , United Kingdom
220379012@aston.ac.uk

Dr. Kalyani Kadam
Assistant Professor
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed University)
Pune, India
kalyanik@sitpune.edu.in

Dr. Liu, Kurt Y.
Department of Engineering and Physical Sciences
Aston University
Birmingham, United Kingdom
y.liu56@aston.ac.uk

***Abstract*—This research presents a study conducted for a large warehousing firm located in the West Midlands, UK, with over 1.6 million stock-keeping units (SKUs) in inventory nationally. Inventory stockout is a situation where a company faces a shortage of a particular item in its inventory, leading to the inability to meet the increasing demands of customers for that product. This condition leads to financial losses, decreased customer loyalty, heightened customer grievances, and the risk of losing market share to competitors. This research aims to investigate the primary factors contributing to stockouts in the organization and suggest strategies to mitigate or minimize their frequency. We leverage an unprocessed dataset obtained from Kaggle to predict stockout events in inventory management systems. To prepare the dataset for analysis, we performed data imputation for null values. Additionally, we addressed the extreme data imbalance by applying oversampling techniques. Categorical data were label-encoded for enhanced processing. Feature scaling was then applied to minimize bias caused by varying magnitudes of variables. The pre-processed dataset was split into train and test sets using K-fold CV for model development and evaluation. To gain comprehensive insights into the dataset, we conducted Exploratory Data Analysis (EDA), uncovering valuable patterns and trends. For training our predictive model, we employed various classifiers, including SVM and Logistic Regression, along with ensemble approaches. Techniques like CV, Elbow method, etc were implemented for hyperparameter tuning in classifiers. The performance of each classifier was evaluated using metrics such as the ROC curve and accuracy. Our trained model gave 99% accuracy which indicate that our predictive model has the potential to accurately forecast stockout events in inventory management systems. These findings are significant for businesses aiming to prevent stockouts, which can result in sales loss and customer dissatisfaction. Effective prediction of stockouts allows businesses to optimize inventory and enhance operational efficiency.**

Keywords

Machine Learnings, Classifiers , Pre-processing, Evaluation

I. Introduction

Backorders are common in automobile industries as demand unpredictability cannot be denied. Over stocking is avoided primarily to prevent blocked inventory and the accompanying holding costs. Additionally, recent disruptions in the global supply chain due to the Covid-19 pandemic, Suez Canal blockage and the Russia-Ukraine war have further aggravated the problems of inventory management. The lead time of raw materials have now become more uncertain than ever and companies are finding it difficult to establish the exact quantity of buffer stock they shall maintain. However, with growing market competition the wait-time due to backorders can be a loss making affair as customers want their demand to be met at the earliest. Efficient inventory management is a necessity to prevent customer turnover and yet avoid blocked inventory. This is only possible with the use of an effective demand prediction mechanism. Pioneered by Industry 4.0 in automotive industry, Machine Learning can be used as an effective tool for backorder prediction. It enables creating models to analyse the existing scenario and generate means to develop optimum inventory management plans. Thus, preventing supply delays and subsequently retain the customer base for an organization.

II. Aim and Objective

Through this study an attempt has been made to design and implement appropriate Machine Learning algorithms for efficient demand analysis and timely backorder prediction.

III. Literature Review

Pillai and Shukla investigated various supervised machine learning classifiers to select the best model for stockout prediction in a Multi Echelon Supply Chain. They proposed a meta-learning-based stacked ensemble model using XG boost, Ada boost, and random forest classifiers as the base models. This model aims to improve the stockout prediction performance compared to individual classifiers. Finally, they used a five-fold cross-validation technique to validate the models. The performance of the machine learning models has been evaluated on the test dataset using assessment metrics. The auteurs have compared the performance of different classifiers and identified that boosting algorithms perform better [1].

Kurian and Maneesh conducted a study on supply chain inventory management, specifically focusing on predicting stockouts using machine learning classifiers. To assess the impact of stockout prediction, they compared two scenarios: one where the supply chain operated solely based on the "OUT" policy (likely meaning orders are placed when the stock reaches a certain threshold), and another where the supply chain incorporated stockout prediction alongside the "OUT" policy. The results of their study showed that the supply chain with stockout prediction performed better than the one without it. This suggests that incorporating predictive analytics into inventory management can offer significant advantages [2].

Lauras and Humez present a multi-criteria decision support system that aims to manage stockouts effectively while simultaneously preserving customer satisfaction. The chosen evaluation criteria of responsiveness, flexibility, and efficiency help ensure the system's ability to meet these objectives [3].

Schreiber and Lydon's study provides valuable insights into the occurrence and causes of vaccine stockouts. By utilizing data from the JRF and UNICEF's Vaccine Forecasting Tool, the authors shed light on the ongoing challenges in maintaining the availability of essential vaccines. The study emphasizes the necessity for prompt and effective interventions to address stockouts and ensure uninterrupted access to vaccines, calling for collaborative efforts from the global community and individual countries [4].

Sheikh-Zadeha and Rossetti's paper introduces the PBIC method as a performance-based inventory classification approach for a multi-item, multi-echelon inventory system. Their evaluation and comparison demonstrate that the PBIC method outperforms other methods and achieves results comparable to the optimal grouping solution. This research contributes to enhancing inventory management practices by providing a more effective approach to inventory classification [5].

Manuel and Ruz's study focus on developing machine learning-based systems to automatically detect OOS events in a retail packaged foods manufacturing company. The Random Forest classifier showed the highest performance, achieving a satisfactory detection precision and successfully identifying a significant proportion of OOS events. The inclusion of new predictor variables further improved the classifier's performance. This research contributes to improving OOS detection and addressing this long-standing problem for manufacturers in the grocery retailing industry [6].

IV. Proposed Work

The real-life problem that is addressed in the paper is the detection of out of stocks in the stock industry. This problem affects both the retailer and the manufacturer. Our suggested technique seeks to offer a structured and efficient strategy for creating reliable prediction models that can foresee stockouts and help with proactive inventory management. The research methodology consists of several key steps that were essential for the programming of our stockout model. They are listed below :

A. Study of the dataset

Studying the dataset is a vital part of any ml building process since it analyzes the features of the dataset to understand the insights about it.In this study, we examine a dataset that was obtained via Kaggle and focuses on forecasting inventory stockouts. The dataset includes a number of columns, including 'sku', 'national_inv', 'lead_time', 'in_transit_qty', 'forecast_3_month', 'forecast_6_month', 'forecast_9_month', 'sales_1_month', 'sales_3_month', 'sales_6_month', 'sales_9_month', 'min_bank', 'potential_issue', 'pieces_past_due', 'perf_6_month_avg', 'perf_12_month_avg', 'local_bo_qty', 'deck_risk', 'oe_constraint', 'ppap_risk', 'stop_auto_buy', 'rev_stop', 'went_on_backorder'. Our study intends to examine the dataset, find the causes of stockouts, and create a prediction model. The dataset offers details on stock levels, demand projections, past sales, risk considerations, and performance indicators. We learn more about the dataset, evaluate variable distributions, look for missing or incorrect data, and investigate correlations between variables using exploratory data analysis. These studies help us comprehend the dataset and direct the creation of a reliable stockout prediction model.

B. Pre-Processing

Upon deep exploration of the dataset, we were confronted with the significant challenge of preparing the raw data to ensure its suitability for comprehensive exploratory data analysis (EDA) and subsequent model training. This crucial phase involves a meticulous process of data transformation and refinement, enabling us to extract meaningful insights and unlock the predictive power of our models. The preprocessing methods that we used with our dataset are :

1) Handling Null Values: Raw dataset consists of various null values in between them.It becomes important for us to handle those null values either by removing them or by replacing them with the appropriate statistical value. If not imputed , they can cause huge error in training the model.

2) Feature selection: Features of the dataset play a major role during the training of the ml model. So , instead of training our model with all features, we should only select the important features that are affecting the results. In this study, we used correlation analysis and multicollinearity measure by VIF Factor for selecting 12 prominent features out of 21.

3) Dealing with Categorical features: Qualitative variables without a natural numerical representation are represented by categorical features.We transformed categorical features into numerical representations for our dataset so that algorithms might recognize patterns and relationships and be used with a larger variety of methods. Overall, addressing categorical features helps algorithms provide precise predictions while ensuring that vital information is kept.

4) Balancing the dataset: The dataset under investigation was characterized by binary categories, exhibiting a substantial class imbalance. Such a class disparity can introduce bias into our model, causing it to place an excessive amount of emphasis on the majority class and possibly jeopardizing the precision of forecasts for the minority class. We used the Synthetic Minority Over-sampling Technique (SMOTE) to address this important problem. By generating synthetic data points, SMOTE effectively equalizes the number of instances in both classes, mitigating the impact of class imbalance and fostering a more balanced training set.

5) Feature scaling: Feature scaling is an essential preprocessing technique that aims to transform numerical features in a dataset to a comparable scale. Its significance lies in several key aspects. Feature scaling promotes faster convergence of machine learning algorithms by ensuring that all features have similar ranges. This helps algorithms to optimize their objective functions more efficiently.

C. Exploratory Data Analysis (EDA)

Understanding the dataset and obtaining knowledge about the field of stockout prediction requires the use of exploratory data analysis (EDA). We used a thorough EDA approach to unearth the main conclusions and ramifications of our research. Inventory management is a crucial area for businesses since it affects customer happiness and the effectiveness of the supply chain. Our goal is to advance this field by creating a reliable stockout prediction model utilizing machine learning techniques. During EDA, we investigated the correlations between characteristics and stockout events as well as the distribution of stockout occurrences. To find patterns and relationships, we used visualizations, descriptive statistics, and domain-specific analyses. Our correlation graph in Figure 1, highlighted the intensity and direction of feature correlations, offering insightful information about potential indicators of stockout occurrences. The EDA process, coupled with domain expertise, guided our subsequent preprocessing steps and facilitated the selection of appropriate machine learning algorithms.

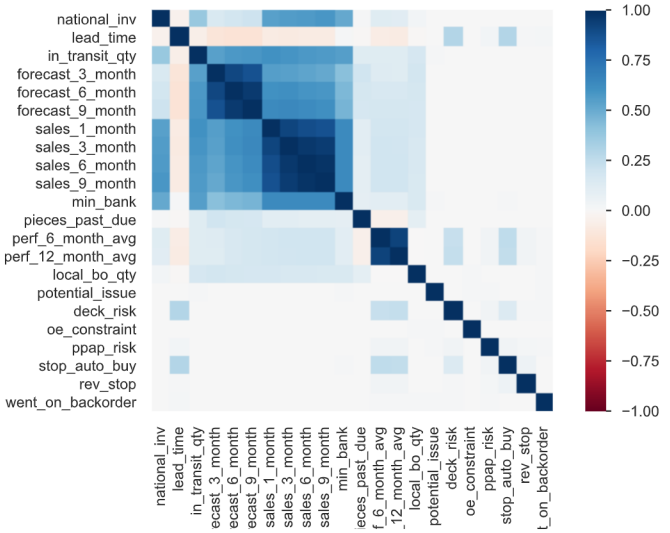

Fig. 1. Correlation Graph

Overall, our EDA process enabled a comprehensive understanding of the dataset and the domain of inventory stockout prediction. By analyzing feature distributions, correlations, and leveraging domain knowledge, we identified significant predictors and gained insights into the factors influencing inventory stockouts. These findings played a crucial role in guiding our subsequent preprocessing steps and selecting suitable machine learning algorithms for stockout prediction. The integration of EDA and domain expertise has contributed to enhancing the accuracy and effectiveness of our research, addressing the challenges of inventory management, and aiding businesses in optimizing their inventory strategies to minimize stockouts and improve customer satisfaction.

D. Training the model

The target variable ['went_on_backorder'] contains only 2 types of values (Yes and No), hence it is a binary classification problem. We have used various popular algorithms for classification like logistic regression, Gaussian Naive Bayes, Decision Tree, and an ensemble technique using random forest, to train our model, and select the best one for accurate and reliable predictions.

- **Logistic Regression algorithm** is used in binary classification tasks where it outputs the probability of a test case lying between 0 and 1 by using sigmoid function. If the predicted probability output is above threshold (generally 0.5) it is assigned to "1" class otherwise '0' class.
- **Gaussian Naive Bayes**- Naive Bayes algorithm makes a 'naive' assumption that the predictor features are independent of each other. It uses the concept of bayes theorem and conditional probability to predict the output of the classification problem. Gaussian Naive Bayes further extends this concept by assuming that the continuous predictor variables are distributed normally (follow a gaussian distribution)

- **Decision Tree** - It is a non-probabilistic approach which uses the concept of tree-like structure and partitions the data recursively, based on its predictor features. It can handle both categorical and continuous variables, and make no assumption about the dataset. Decision tree models are highly interpretable as they form a tree like structure, making it easy to understand the decision-making process, but they are highly prone to overfitting.
- **Random Forest** is an ensemble learning technique based on the concept of "Bagging" or "Bootstrap Aggregation". Individual Decision Trees are trained on small samples of data with replacement, and finally, the result is obtained using aggregation or majority voting. Random Forest combines several decision trees to make a more accurate and robust model less prone to overfitting.

V. Results and Disscussion

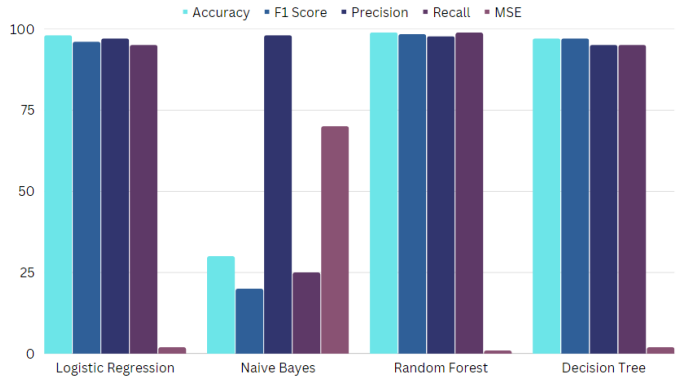

Fig. 2. Comparative analysis of different classifiers

The findings of our research demonstrate the remarkable performance of the Random Forest classifier for inventory stock-out prediction. Among the four classifiers evaluated, including Logistic Regression, Naive Bayes, Decision Tree, and Random Forest, the latter exhibited the highest accuracy rate, reaching an impressive 99%. This exceptional accuracy underscores the effectiveness of Random Forest in capturing complex relationships and interactions within the dataset, making it a powerful tool for stockout prediction.

Figure 2 presents a visual representation of the evaluation metrics comparison for the different classifiers we implemented which is also depicting the comparative analysis based on the performance of the model with each classifier used. We can clearly infer Random Forest, showcasing its robust performance across multiple metrics. The steep increase in accuracy, precision, recall, and F1 score validates the effectiveness of the Random Forest classifier in accurately predicting inventory stockouts.

Table 1 shows the Confusion metrics that we obtained from Random Forest Classifier.

TABLE I
Confusion Matrix

		Prediction	
		Positive	Negative
Actual		2.5×10^5	1568
		2.3×10^3	154

The results obtained from our research highlight the superiority of the Random Forest classifier in predicting inventory stockouts, as evidenced by its exceptional accuracy, precision, recall, and F1 score. The incorporation of preprocessing techniques ensures the reliability and generalizability of our model. These findings have significant implications for inventory management, enabling proactive decision-making to optimize inventory levels, minimize stockouts, and enhance operational efficiency. The success of our model opens avenues for further research, including the exploration of additional features and the validation of the model's performance on diverse datasets.

VI. Conclusion

Our research focused on developing a stockout prediction model for inventory management using machine learning techniques. After thorough preprocessing and analysis, we found that random forest emerged as the most effective classifier, achieving an impressive accuracy of 99%. The superiority of random forest can be attributed to its capability to capture complex relationships and non-linear interactions within the data, enabling it to make accurate predictions. On the other hand, naive Bayes performed comparatively poorly due to its assumption of feature independence, which did not align with the characteristics of our dataset.

The results of our study highlight the critical importance of careful algorithm selection and preprocessing steps in constructing accurate stockout prediction models. By leveraging the power of random forest and other suitable algorithms, businesses can optimize their inventory management strategies, minimize stockouts, and enhance customer satisfaction. These findings contribute to the field by demonstrating the effectiveness of machine learning in predicting stockouts, and emphasize the value of data-driven decision-making in inventory management.

References

[1] S. Shukla and V. M. Pillai, "Stockout Prediction in Multi Echelon Supply Chain using Machine Learning Algorithms," in Proceedings of the 2nd Indian International Conference on Industrial Engineering and Operations Management, Warangal, Telangana, 2022.

[2] Kurian, D.S., Maneesh, C.R. and Pillai, V.M. (2020) 'Supply chain inventory stockout prediction using machine learning classifiers', Int. J. Business and Data Analytics, Vol. 1, No. 3, pp.218–231.

[3] M. Lauras, V. Humez, U. Okongwu, and L. Dupont , "An Advanced ATP Decision Support System in Stockout Situations," in Proceedings of the 17th World Congress, The International Federation of Automatic Control, Seoul, Korea, July 6-11, 2008.

[4] P. Lydon, B. Schreiber, A. Gasca, L. Dumolard, D. Urfer, and K. Senouci, "Vaccine stockouts around the world: Are essential vaccines always available when needed?," Vaccine, vol. 35, pp. 2121-2126, 2017.

[5] A. Sheikh-Zadeha, M. D. Rossetti, and M. A. Scott, "Performance-based inventory classification methods for large-scale multi-echelon replenishment systems," Omega, vol. 101, article no. 102276, 2021.

[6] Electronics — Free Full-Text — Predicting Out-of-Stock Using Machine Learning: An Application in a Retail Packaged Foods Manufacturing Company (mdpi.com)