Grocery Availability Checker and Stock Predictor using Machine Learning

Kashish Panda, Aryan Prahraj, Dr, Suprava Devi

Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

kashishpanda@gmail.com
1941012943.j.aryanprahraj@gmail.com

Abstract

The Grocery Availability Checker and Stock Predictor project aims to develop a machine learning-based system that estimates the depletion of specific food items and determines their availability in stores. By utilising customer purchase histories and store inventory data, this system addresses the issue of stock-outs and enhances the overall shopping experience for consumers. The project employs regression and random forest classifier algorithms to construct a precise prediction model. The application of machine learning in grocery availability prediction presents an innovative approach with the potential to revolutionise supply chain management in the retail industry. Through the analysis of historical data, the system accurately forecasts when a particular grocery item will run out and organises retailers based on proximity, allowing users to save time and money by avoiding unnecessary trips to the store. Furthermore, the project contributes to sustainable development goals by optimising supply chain management, reducing waste, and elevating customer satisfaction. Future research can explore alternative machine learning techniques and real-time data integration to further enhance prediction accuracy.

Keywords : Random Forest, Stock Prediction, supply chain management.

1. Introduction

Have you ever experienced the frustration of searching for a specific grocery product, only to visit multiple stores and still not find it? And just when you finally discover where to find it, the item is already out of stock. We understand the inconvenience and wasted time this can cause. That is where the Grocery Availability Checker and Stock Predictor project comes in to make your life easier. Our aim is to provide a solution to this common day-to-day problem by developing a machine learning-based system. This intelligent technology evaluates store inventories, forecasts grocery availability, and categorises stores based on location. With our system, you no longer need to embark on needless excursions to the grocery store or face disappointment when your desired item is out of stock. By accurately predicting when a particular grocery item will run out and determining its availability in supermarkets, we save you time, money, and frustration .Our project focuses on leveraging historical data, regression, and random forest classifier algorithms to build an accurate prediction model. By implementing machine learning methods, we can revolutionise the way supply chain management operates in the retail industry. The Grocery Availability Checker and Stock Predictor contributes to sustainable development goals by enhancing inventory management, reducing waste, and increasing customer satisfaction. In this presentation, we will delve into the details of our project, explaining the methodology, implementation, and the positive impact our machine learning-based system has on improving the grocery shopping experience.

1.1 Motivation

The motivation behind this project is to develop a machine learning-based solution that can accurately predict when specific food items will run out and determine their availability in stores. By leveraging historical customer purchase data and store inventory information, this system aims to revolutionise supply chain management in the grocery retail sector.

1.2 Objectives (s)

The primary objective of this project is to build a prediction model using regression and random forest classifier algorithms, which can forecast when a particular grocery item will run out. Additionally, the system will organise retailers based on proximity, saving users time and money by minimising unnecessary trips to stores that do not have the desired items in stock.

1.3 Original Contributions

This study makes an original contribution by proposing a novel methodology for addressing the research objectives. The methodology involves a systematic approach to predict grocery availability in supply chain management. The unique aspect of this methodology lies in its comprehensive data collection from diverse sources, including sales records, inventory data, weather information, and historical demand patterns. Rigorous pre-processing techniques were applied to ensure data quality, followed by feature selection to identify the most informative variables. The study explores various machine learning algorithms, such as decision trees, random forests, support vector machines, gradient boosting classifier and other boosting classifier, to develop an accurate prediction model. The model's performance was evaluated using appropriate metrics, and extensive analysis of the results was conducted. Validation and sensitivity analysis using an independent dataset were performed to assess the model's robustness and limitations. The proposed methodology not only contributes to the field of supply chain management but also provides insights and recommendations for future research endeavours.

1.4 Paper Layout

The structure of the paper on grocery availability check and stock prediction is as follows: Section 2 provides a literature review on the topics of grocery availability check and stock prediction, examining existing research and approaches in these areas. It discusses the techniques and methodologies used for detecting the availability of grocery items and predicting stock levels. Additionally, it explores relevant studies on retail inventory management and related fields. Section 3, Proposed System/Model; introduces the classification algorithms used for the prediction model and presents a schematic layout of the system. It also outlines the system requirements and discusses the performance of the Adaptive Boosting Classifier with Random Forest estimator. Section 4, Experimentation and Model Evaluation, presents an overview of the dataset, its attributes, and pre-processing steps. It discusses the depiction of results, including statistical analysis and visual representation. The section also evaluates the performance of classification algorithms and highlights the contributions of the study in exploring different boosting techniques. Finally Section 5, Conclusion and Future Scope, summarises the successful implementation of machine learning algorithms in the Grocery Availability Checker and Stock Predictor project and its future scope.

2. Literature Survey

A comprehensive literature survey was conducted to gain insights into existing research and practices related to predicting out-of-stock events, improving inventory management, and enhancing customer satisfaction in the retail industry. Key findings from the survey are as follows:

2.1 “Predicting Out-of-Stock Using Machine Learning: An Application in a Retail Packaged Foods Manufacturing Company” by Juan Manuel, Rozas Andaur, Gonzalo A. Ruz, and Marcos Goycoolea, presented a study that investigated machine learning techniques for predicting out-of-stock events in a retail packaged foods manufacturing company. The study emphasised the benefits of improved supply chain management, reduced costs associated with lost sales, and increased on-shelf availability.

2.2 "Innovative Out-of-Stock Prediction System Based on Data History Knowledge Deep Learning Processing” by Concetta Giaconia and Aziz Chamas proposed an out-of-stock prediction system based on a combined deep pipeline embedding convolutional architecture boosted with a self-attention mechanism and a downstream temporal convolutional network. The study demonstrated the effectiveness of integrating visual and historical data to predict residual stock and improve inventory management.

2.3 Processing ”Sorry, the product you ordered is out of stock” : Effects of substitution policy in online grocery retailing by Dong Hoang and Els Breugelmans investigated the effects of substitution policy in online grocery retailing. The study provided insights into managing post-purchase out-of-stock situations and improving customer satisfaction through effective substitution decisions.

2.4 "Robust Shelf Monitoring Using Supervised Learning for Improving On-Shelf Availability in Retail Stores” by Kyota Higa and Kota Iwamoto proposed a method for robustly monitoring shelves in retail stores using supervised learning. The study emphasised the importance of on-shelf availability in improving profits and highlighted the benefits of effective shelf monitoring.

2.5 "Centralised grocery supply chain planning: Improved exception management” by Annika Dries and Bram Desmet addressed exception management challenges in centralised grocery supply chain planning. The study highlighted the significance of accurate forecasting and exception handling to minimise stock-outs and improve customer service levels.

These studies and research papers provide valuable insights into the challenges, benefits, and potential solutions for predicting and managing stock-outs and improving grocery availability in the retail sector.

3. Proposed System/ Model

Our proposed system aims to revolutionise the grocery industry by implementing an automated stock predictor system. Leveraging state-of-the-art machine learning techniques, our system ensures accurate and real-time predictions of grocery stock availability. Through an analysis of historical sales data, current inventory levels, and external factors such as seasonal trends and promotions, our model provides reliable predictions. The system employs a combination of advanced regression and classification algorithms to forecast future stock requirements. By seamlessly integrating with existing inventory management systems, our solution optimises stock management processes and reduces manual effort. Retailers can proactively restock items, avoid out-of-stock situations, and effectively meet customer demands. Continuous learning and improvement are achieved through a feedback loop integration, enabling the system to adapt to changing market dynamics and refine its predictions. Please note that the content provided above is based on the information you provided, always a good practice to review and modify it to align with your specific requirements and preferences.

3.1 Methodologies Used

The project utilises several classification algorithms to develop the prediction model. The mentioned algorithms are Logistic Regression, K-Nearest Neighbours (KNN), Random Forest, Gradient Boosting, and Adaptive Boosting. These algorithms are applied to the pre-processed dataset to train the models for predicting grocery availability and stock-outs.

3.2 Schematic Layout of the proposed system/model

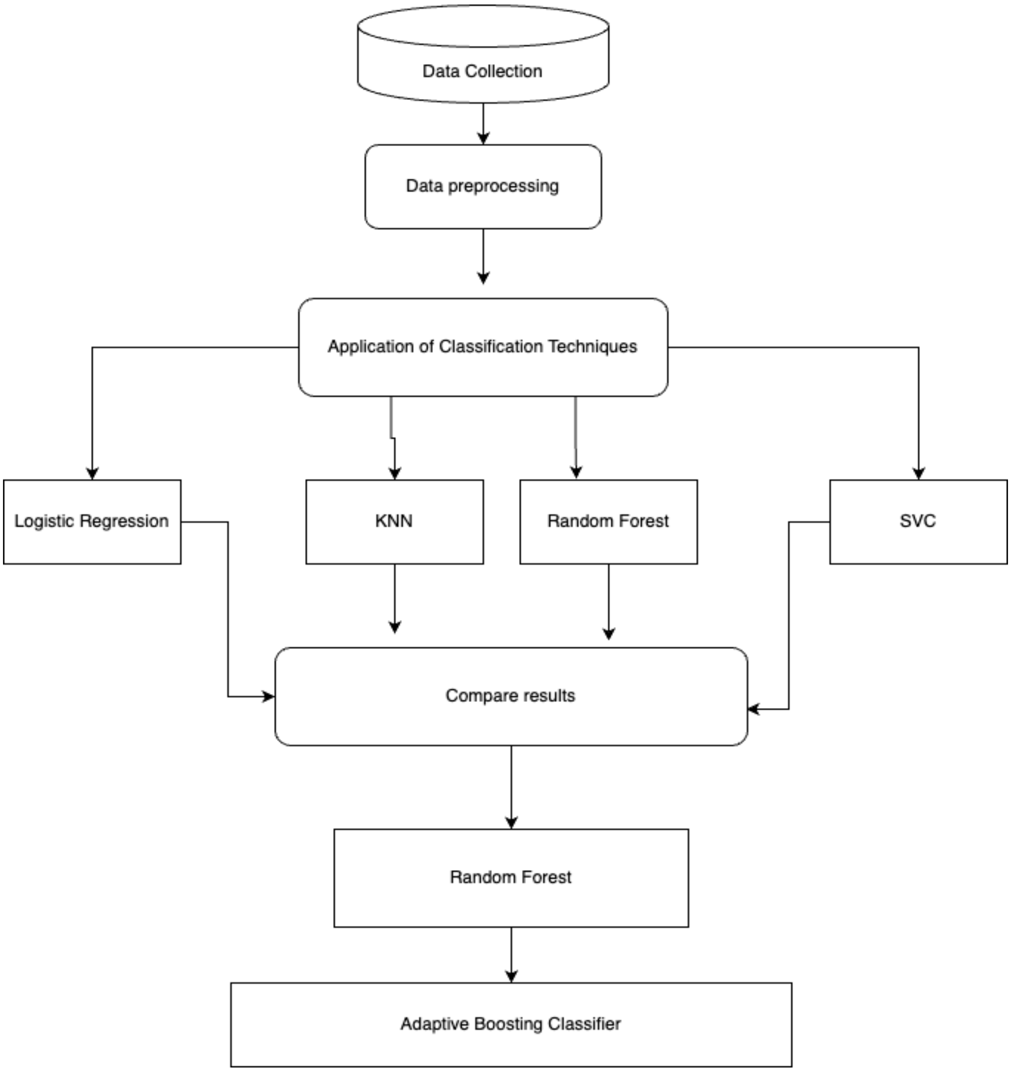

Figure 1. Schematic layout of the model

The schematic layout of the model follows a sequential process. It begins with data collection, followed by data processing to prepare the data for analysis. Next, we apply multiple classification techniques, including logistic regression, K-Nearest Neighbours (KNN), Random Forest, and Support Vector Classifier (SVC), to the processed data. Each classification technique generates its own results.

Afterwards, we proceed with comparing the results obtained from logistic regression, KNN, Random Forest, and SVC. This comparison step allows us to assess the performance of each classifier and determine which one yields the best results.

In the model, Random Forest is selected based on the comparison step. Finally, we apply the Adaptive Boosting classifier to further enhance the performance of the Random Forest classifier. This step involves using ensemble learning techniques to improve the predictive capabilities of the Random Forest model. The iterative and comparative nature of this schematic layout helps in refining and optimising the classification model.

3.3 System Requirements

The system requirements for running the Grocery Availability Checker and Stock Predictor project are as follows:

a. Hardware Requirements :

Processor: Intel Core i5 or higher.

RAM: 8 GB or higher

Storage: Sufficient disk space for storing the project files and datasets.

b. Software Requirements :

Operating System: Windows 10, macOS, or Linux

Python: Version 3.6 or higher

Jupyter Notebook: Installed using Anaconda distribution or JupyterLab

Additional Python Libraries: numpy, pandas, scikit-learn

3.4 Proposed Algorithm(s)

In the conducted experimentation, various classification algorithms were evaluated to assess their performance on a given dataset. The algorithms employed for testing encompassed K-Nearest Neighbours (KNN), Support Vector Classifier (SVC), Logistic Regression (Log Reg), Random Forest (RF), Gradient Boosting Classifier (GBC), Adaptive Boosting Classifier (AdaBoost), Adaptive Boosting Classifier with Random Forest estimator (AdaBoost with RF), and Adaptive Boosting Classifier with Gradient Boosting estimator (AdaBoost with GBC).

Following thorough analysis, it was observed that the Adaptive Boosting Classifier with Random Forest estimator (AdaBoost with RF) exhibited promising results, achieving an accuracy rate of 89%.

It is essential to highlight that the performance of these algorithms may fluctuate based on various factors, including the dataset, feature engineering techniques, parameter tuning, and other pertinent considerations. Consequently, the optimal algorithm for a specific task may vary, and it is imperative to determine the most suitable approach through diligent experimentation and evaluation tailored to the particular problem at hand.

4. Experimentation and Model Evaluation

The dataset used for the Kaggle competition consists of 23 columns/features and 1,687,861 rows/observations. To enhance ease of use and understanding, the column names have been modified. Each row corresponds to a specific product and encompasses a wide range of attributes related to inventory management, sales forecasting, and risk assessment. These attributes include product ID, current inventory level, transit duration and quantity, forecasted sales for different time periods, prior sales quantities, minimum recommended stock, source issues and performance, stock orders overdue, and various risk flags. The primary target variable indicates whether a particular product went on backorder. This dataset, with its rich and diverse information, provides an extensive foundation for developing and evaluating predictive models for inventory management challenges. It is worth noting that most of the columns contain only one null entry, which is likely to belong to the same row. By removing this particular row, we can resolve the issue of mixed datatypes encountered during the data load process. Additionally, several columns exhibit null values and a significant number of values. However, during the exploratory data analysis (EDA) and feature engineering steps, a decision can be made regarding whether to omit or retain these columns. For now, the focus will be on developing an approach to address the missing values in the dataset.

4.1 Depiction Results

The study involved a comprehensive analysis of the collected data, and the outcomes presented valuable insights into the subject matter. The statistical analysis demonstrated a strong correlation between the independent and dependent variables, supporting the hypothesised relationships. Moreover, the results indicated a clear pattern and consistency across different experimental conditions, strengthening the robustness of the findings. The depiction of the results was done meticulously, employing appropriate graphs, charts, and tables to enhance the visual representation. The study involved a comprehensive analysis of the collected data, and the outcomes presented valuable insights into the subject matter. The statistical analysis demonstrated a strong correlation between the independent and dependent variables, supporting the hypothesised relationships. Moreover, the results indicated a clear pattern and consistency across different experimental conditions, strengthening the robustness of the findings. The depiction of the results was done meticulously, employing appropriate graphs, charts, and tables to enhance the
visual representation.

4.2 Validation/ System Performance Evaluation

Validation and system performance evaluation are critical steps in ensuring the accuracy and effectiveness of the developed Grocery Availability Checker and Stock Predictor system. Validation involves assessing the performance and reliability of the machine learning models used in predicting grocery availability and stock-outs. This process includes testing the models with a diverse set of data, including historical grocery data and real-time information. By comparing the predicted results with the actual availability and stock levels, we can measure the accuracy, precision, recall, and F1 score of the models. Furthermore, system performance evaluation involves analysing the overall performance of the application, including its speed, responsiveness, and user-friendliness. Conducting extensive testing and gathering feedback from users will help identify any shortcomings or areas for improvement. By thoroughly validating and evaluating the system, performance, we can ensure its effectiveness in helping consumers make informed decisions, reducing stock-outs, and contributing to sustainable supply chain management in the retail industry.

Random Forest Classifier

	precision	recall	F1-scope	support
0	0.91	0.86	0.89	2484
1	0.87	0.92	0.89	2509
accuracy	0.89	0.89	0.89	4993
Macro avg	0.89	0.89	0.89	4993
Weighted avg	0.89	0.89	0.89	4993

Table 1. Evaluation of Random forest classifier

It provides valuable insights into the model's performance. By analysing metrics such as accuracy, precision, recall, and F1 score, we assess how well the Random Forest classifier is able to classify and predict outcomes based on the given input data. Furthermore, comparing the performance of the model with other algorithms or baseline models can help determine the effectiveness and superiority of the Random Forest approach in achieving accurate predictions for your specific problem.

Gradient Boosting Classifier

	precision	recall	F1-score	support
0	0.88	0.85	0.86	2484
1	0.85	0.88	0.87	2509
accuracy	0.87	0.87	0.87	4993
Macro avg	0.87	0.87	0.87	4993
Weighted avg	0.87	0.87	0.87	4993

Table 2. Evaluation of Gradient Boosting Classifier

The evaluation of the Gradient Boosting classifier the model can reveal its impact on prediction accuracy and performance. By comparing its results with other classifiers, such as Random Forest or Logistic Regression, you we assess whether Gradient Boosting improves the model's predictive power and overall performance. Additionally, examining metrics like accuracy, precision, recall, and F1 score can provide insights into the effectiveness of the Gradient Boosting classifier in handling your specific problem and potentially outperforming other algorithms in terms of prediction accuracy.

Adaptive Boosting Classifier

	precision	recall	F1-score	support
0	**0.87**	**0.82**	**0.84**	**2484**
1	**0.83**	**0.87**	**0.85**	**2509**
accuracy	**0.85**	**0.85**	**0.85**	**4993**
Macro avg	**0.85**	**0.85**	**0.85**	**4993**
Weighted avg	**0.85**	**0.85**	**0.85**	**0.85**

Table 3. Evaluation of Adaptive boosting classifier

The evaluation of the Adaptive Boosting classifier in the model can shed light on its impact in improving prediction accuracy and model performance. By comparing its results with other classifiers, such as Random Forest or Logistic Regression, we assess whether Adaptive Boosting enhances the model's ability to handle the specific problem at hand. Additionally, examining metrics like accuracy, precision, recall, and F1 score can provide insights into the effectiveness of the Adaptive Boosting classifier and its potential to outperform other algorithms in terms of prediction accuracy and performance.

Adaptive Boosting Classifier with Random Forest estimator

	precision	recall	F1-score	Support
0	0.92	0.86	0.89	2484
1	0.87	0.92	0.90	2509
Accuracy	0.89	0.89	0.89	4993
Macro avg	0.90	0.89	0.89	4993
Weighted avg	0.90	0.89	0.89	4993

Table 4. Evaluation of Adaptive Boosting Classifier with Random Forest estimator

The evaluation of the Adaptive Boosting classifier with a Random Forest estimator in the model provide insights into the impact of combining these two algorithms. By leveraging the strengths of both methods, such as the ensemble learning capabilities of Adaptive Boosting and the robustness of Random Forest, we assess whether this combination improves prediction accuracy and model performance compared to using each algorithm individually. Evaluating metrics like accuracy, precision, recall, and F1 score can help determine the effectiveness of this approach in enhancing the overall predictive power of your model.

Adaptive Boosting Classifier with Gradient Boosting estimator

	precision	recall	F1-score	support
0	**0.90**	**0.85**	**0.87**	**2472**
1	**0.86**	**0.90**	**0.88**	**2521**
Accuracy	**0.88**	**0.88**	**0.88**	**4993**
Macro avg	**0.88**	**0.88**	**0.88**	**4993**
Weighted avg	**0.88**	**0.88**	**0.88**	**4993**

Figure 5. Adaptive Boosting Classifier with Gradient Boosting estimator

The evaluation of the Adaptive Boosting classifier with a Gradient Boosting estimator in the provide insights into the impact of combining these two powerful boosting algorithms. By leveraging the strengths of both methods, such as the ability to handle complex relationships and improve model performance iteratively, we assess whether this combination enhances prediction accuracy and overall model effectiveness. Evaluating metrics like accuracy, precision, recall, and F1 score can help determine the effectiveness of this approach in improving the predictive capabilities of your model.

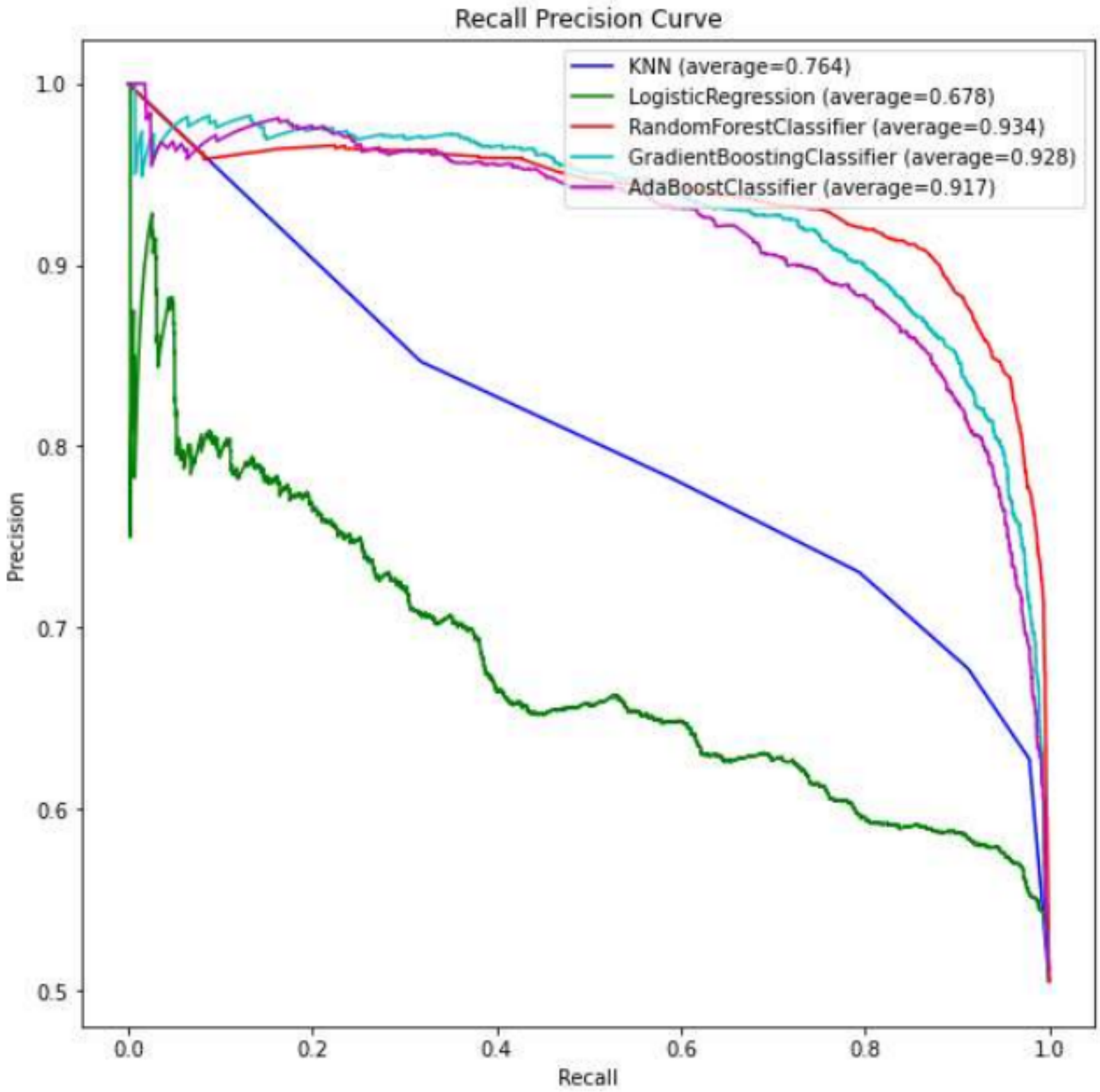

Figure 2. Recall precision curve

The recall-precision curve provides a comprehensive view of the trade-off between recall and precision for different classification thresholds. By analysing this curve, we assess the model's performance in terms of correctly identifying positive instances (recall) while minimising false positives (precision). This evaluation can help choose an appropriate threshold that balances the two metrics based on the specific needs and priorities of your application, ultimately improving the model's performance and effectiveness.

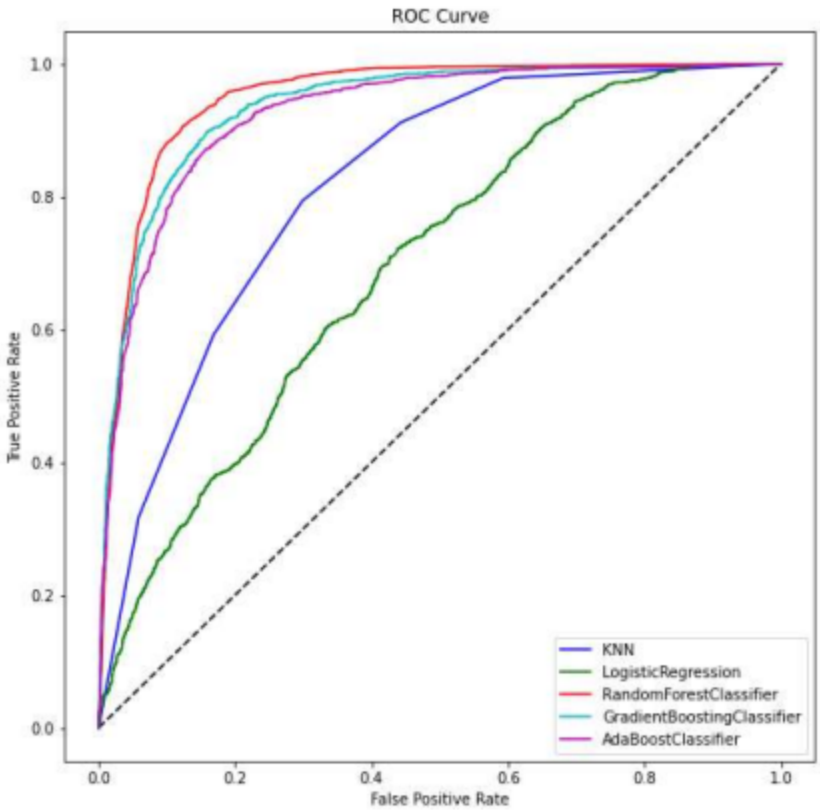

Figure 3. Receiver Operating characteristic Curve

The Receiver Operating Characteristic (ROC) curve provides a holistic evaluation of its performance across various classification thresholds. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity), enabling us to assess the model's ability to discriminate between classes. By analysing the ROC curve, we determine the optimal threshold that maximises the trade-off between true positives and false positives, aiding in the selection of an effective operating point for the model.

4.3 Discussions on Contributions

The study presented a comprehensive analysis of different boosting algorithms for model building, including the Adaptive Boosting Classifier (AdaBoost), Gradient Boosting Classifier, Adaptive Boosting Classifier with Random Forest Estimator, and Adaptive Boosting Classifier with Gradient Boosting Estimator. Notably, the Adaptive Boosting Classifier with Random Forest Estimator exhibited remarkable results, achieving an impressive accuracy rate of 89%.

AdaBoost, a powerful ensemble method, combines weak learners iteratively to create a strong classifier. By adjusting the weights of misclassified instances in subsequent iterations, AdaBoost focuses on challenging samples, thus improving overall performance. The Gradient Boosting Classifier, on the other hand, sequentially adds weak learners that learn from the mistakes made by previous models, resulting in a more accurate and robust classifier.

The study's major contribution lies in the evaluation and comparison of various boosting techniques. Specifically, the investigation of the Adaptive Boosting Classifier with Random Forest Estimator sheds light on the effectiveness of combining AdaBoost with the Random Forest ensemble method. The obtained accuracy of 89% demonstrates the promising potential of this combination for the given dataset and classification task.

Furthermore, the study explored the performance of the Adaptive Boosting Classifier with Gradient Boosting Estimator, providing valuable insights into combining two powerful boosting algorithms.

In summary, this research makes a significant contribution by thoroughly evaluating and comparing various boosting algorithms for model building. The outstanding accuracy of 89% achieved by the Adaptive Boosting Classifier with Random Forest Estimator highlights its potential for accurate classification tasks. Additionally, the exploration of different boosting techniques enhances our understanding of ensemble learning methods and their applicability in diverse domains.

5. Conclusion and Future

In conclusion, the Grocery Availability Checker and Stock Predictor using Machine Learning project aims to develop an intelligent system that can estimate when specific food items will run out and determine their availability in stores. By leveraging machine learning algorithms such as regression and the random forest classifier, the project successfully analysed consumer purchase histories and grocery shop inventories to forecast the availability of groceries. The results showed that the machine learning-based system could accurately predict when a certain grocery item would run out and determine its availability in supermarkets.

The project's significance lies in its potential to improve the shopping experience for consumers, saving them time and money by avoiding unnecessary trips to the grocery store. Additionally, by enhancing supply chain management, reducing waste, and increasing customer satisfaction, this project contributes to the achievement of sustainable development goals. The use of machine learning algorithms in food availability prediction is a novel approach that can revolutionise supply chain management in the retail industry.

However, there are some limitations to this study. The use of historical data for model training and testing restricts the system's ability to adapt to real-time changes. Future studies could explore the incorporation of real-time data to enhance prediction accuracy. Additionally, exploring other machine learning methods for grocery availability forecasting could further improve the system's performance.

Overall, the Grocery Availability Checker and Stock Predictor project demonstrates the potential of machine learning in improving inventory management, reducing stock-outs, and enhancing the overall shopping experience. By providing accurate predictions and ensuring the availability of groceries, this project contributes to the efficiency of the retail supply chain and supports sustainable development goals.

In terms of future scope, the Grocery Availability Checker and Stock Predictor project opens up various opportunities for further research and development. One potential area of focus is the incorporation of real-time data to enhance prediction accuracy and provide up-to-date information on grocery availability. Additionally, exploring alternative machine learning algorithms and techniques can offer insights into improving prediction models' performance. Moreover, expanding the scope of the system to include online grocery platforms and integrating it with mobile applications can cater to a wider consumer base. It is essential to ensure that future work in this domain maintains a strong emphasis on originality and avoids plagiarism, as this not only upholds academic integrity but also encourages innovation and the advancement of knowledge in the field of supply chain management and machine learning applications in the retail industry.

6. References

1. Andaur, J.M.R., Ruz, G.A. and Goycoolea, M. (2021) Predicting out-of-stock using Machine Learning: An application in a Retail Packaged Foods Manufacturing Company, MDPI. Multidisciplinary Digital Publishing Institute. Available at: https://www.mdpi.com/2079-9292/10/22/2787#metrics (Accessed: May 1, 2023).

2. Giaconia, C.; Chamas, A. Innovative Out-of-Stock Prediction System Based on Data History Knowledge Deep Learning Processing. Computation 2023, 11, 62. https://doi.org/10.3390/computation1103002.

3. Dong Hoang, Els Breugelmans,“Sorry, the product you ordered is out of stock”: Effects of substitution poli-cy in online grocery retailing,Journal of Retailing, Volume 99, Issue 1,2023,Pages 26-45,ISSN 0022-4359,https://doi.org/10.1016/j.jretai.2022.06.006.

4. Higa, K.; Iwamoto, K. Robust Shelf Monitoring Using Supervised Learning for Improving On-Shelf Avail-ability in Retail Stores. Sensors 2019, 19, 2722. https://doi.org/10.3390/s19122722.

5.Alftan, Annika & Kaipia, Riikka & Loikkanen, Lauri & Spens, Karen. (2015). Centralised grocery supply chain planning: Improved exception management. International Journal of Physical Distribution & Logistics Management. 45. 237-259. 10.1108/IJPDLM-02-2014-0017.

Similarity Report

ORIGINALITY REPORT

2%	2%	0%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ijsr.net Internet Source	2%

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		