Leveraging Machine Learning to Optimize Pricing and Restocking Strategies for Fresh Vegetables

Tianyou Wang*, #, Ruopeng·Chu#, Ziyuan·Jiang#

Faculty of Science and Technology, Beijing Normal University - Hong kong Baptist University United International College, Zhuhai, China, 519087

* Corresponding Author Email: wtywpzyt@163.com

#These authors contributed equally.

Abstract. In the competitive and diverse retail market, restocking and pricing perishable vegetables is challenging. Using big data and machine learning to analyze sales data, this research can build predictive models to refine these strategies and predict trends[1][2]. This research analyzed sales data from 2020 to 2023, identifying key trends in sales volumes, seasonality, and consumer preferences. It incorporated these findings into a multi-objective programming model that also considered wholesale and loss rates. This research used statistical tests like the Kruskal-Wallis and Mann-Whitney U to detect significant sales differences across vegetable categories, while the Spearman correlation showed strong seasonal and culinary impacts on consumer buying behaviors. To optimize freshness and maximize sales and profits, this research utilized genetic algorithms and regression analyses, including an effective XGBoost model. This model had a Mean Absolute Percentage Error of 16.97 and an R^2 of 0.94, aiding in predicting daily sales and improving pricing strategies and restocking schedules. These methods helped enhance decision-making, ensuring vegetable freshness and maximizing profitability, with projected profits reaching ¥41,582.87. This study not only highlights the effectiveness of integrating big data and machine learning in the retail sector but also sets a benchmark for operational efficiency in managing perishable goods.

Keywords: Big Data, Machine Learning, Multi-Object Programming Model, Genetic Algorithm.

1. Introduction

In the current retail environment, the sales efficiency and supply chain management of fresh vegetables, as an important part of daily consumer goods, are of particular importance. This is directly related to merchant profits and consumer satisfaction. Due to the perishable nature of fresh vegetables, replenishment, and pricing strategies pose special challenges, and traditional replenishment and pricing methods are no longer able to meet the rapidly changing market demand. This has important implications for forecast accuracy, inventory management optimization, and cost control.

Advances in big data and machine learning have made it feasible to optimize vegetable replenishment and pricing strategies. This study utilizes the Kruskal-Wallis and Mann-Whitney U-tests to analyze sales variations across vegetable categories. These non-parametric tests were introduced in 1952 and 1947 respectively for assessing distributions and are fundamental methods of statistical analysis. Bonferroni correction was used to minimize multiple comparison errors. Kendall's Tau coefficient was also used to determine the correlation between categories, identifying key relationships in sales data. For forecasting, this research used algorithms including neural networks and XGBoost regression. The XGBoost algorithm, proposed by Tianqi Chen in 2016, is known for its accuracy and efficiency, is popular in data science competitions, and applies to areas such as retail sales forecasting and financial risk management [3].

In this study, vegetable sales data from 2020 to 2023 were first processed and analyzed to compare annual sales. Kruskal-Wallis test and Mann-Whitney U test were used to analyze the sales differences between different categories and Bonferroni correction was used to correct the errors. Correlations between categories were explored using Kendall's Tau coefficient and a strong correlation with sales volume was found. In addition, a pricing model was constructed and trained for time series forecasting based on historical data using multiple regression techniques. The optimal strategy was selected based on the model evaluation results, and finally, the cost-plus method was used to calculate one-week replenishment and expected revenue.

2. Data Collection and Preprocessing

Data was obtained from supermarket sales records spanning July 2020 to June 2023(www.mcm.edu.cn). A rigorous preprocessing routine was applied, including normalization of prices, encoding of categorical variables, and handling of missing values [4]. This meticulous preparation ensured the quality and reliability of the data for subsequent analyses.

3. Multivariate Statistical Analysis of Vegetable Sales Data

Based on the preprocessed data, various algorithms and statistical methods are used to compare and validate the relationships between different categories and individual products.

3.1. Model Preparation

According to online research and literature, supermarkets refer to past sales data, inventory tory levels, and sales volume of recent periods and past periods to predict future demand. Seasonal factors and holidays, such as the Spring Festival and Mid-Autumn Festival, are also considered, as demand for certain products may increase during these times. Market demand, including consumer purchasing habits, new consumption trends, or brand preferences, is also studied.

This study focuses on examining the sales volume of different categories for each year from 2020 to 2023 and comparing the sales volume over three years. Four indicators were established to effectively respond to the relationships between individual products and categories, as well as between categories: whether it is related to weekends, seasons, and categories. Below is the flowchart for this issue as shown in Figure 1.

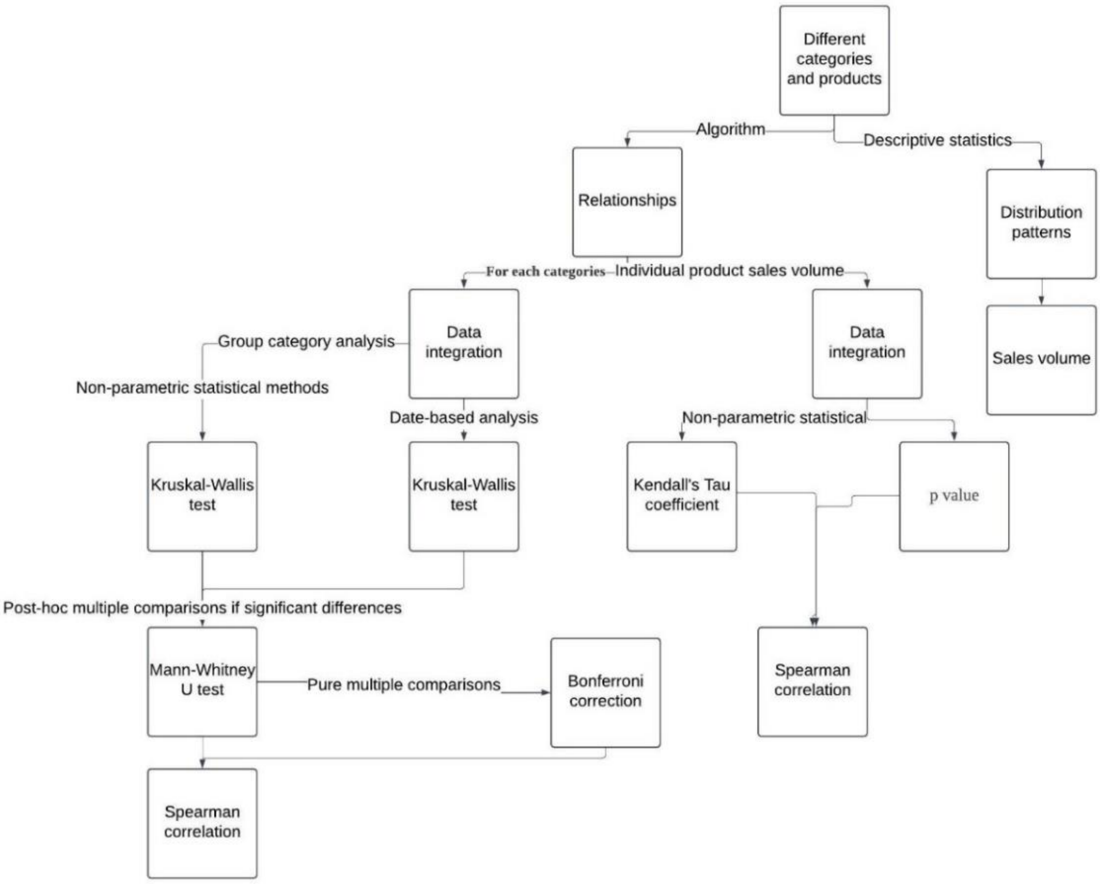

Figure 1. Flow chart of model preparation

3.2. Calculation of Indicators

To build this model, it is necessary to first determine the sales volume of vegetable categories and individual products from 2020 to 2023.

Next, the Kruskal-Wallis test was used to preliminarily determine whether there are significant sales differences among various vegetable categories. Following this, we employed the Mann-Whitney U test for post-hoc multiple comparisons to identify which category groups have significant differences. To correct for potential issues arising from multiple comparisons, we used the Bonferroni correction. The Bonferroni correction was applied to control the error rate in multiple comparisons, ensuring that the detected differences were significant.

This paper continued by using Spearman's rank correlation coefficient to measure the relationships between vegetable categories and individual products. Finally, we calculated Kendall's Tau coefficient and p-value to assess the correlation between individual products and sales dates, as well as between vegetable categories and sales dates. Here, τ (tau) refers to Kendall's Tau coefficient, which measures the rank correlation between two variables.

3.3. Patterns and Correlations

3.3.1. Distribution Pattern

(a) The total sales of different categories over four years are shown in Figure 2:

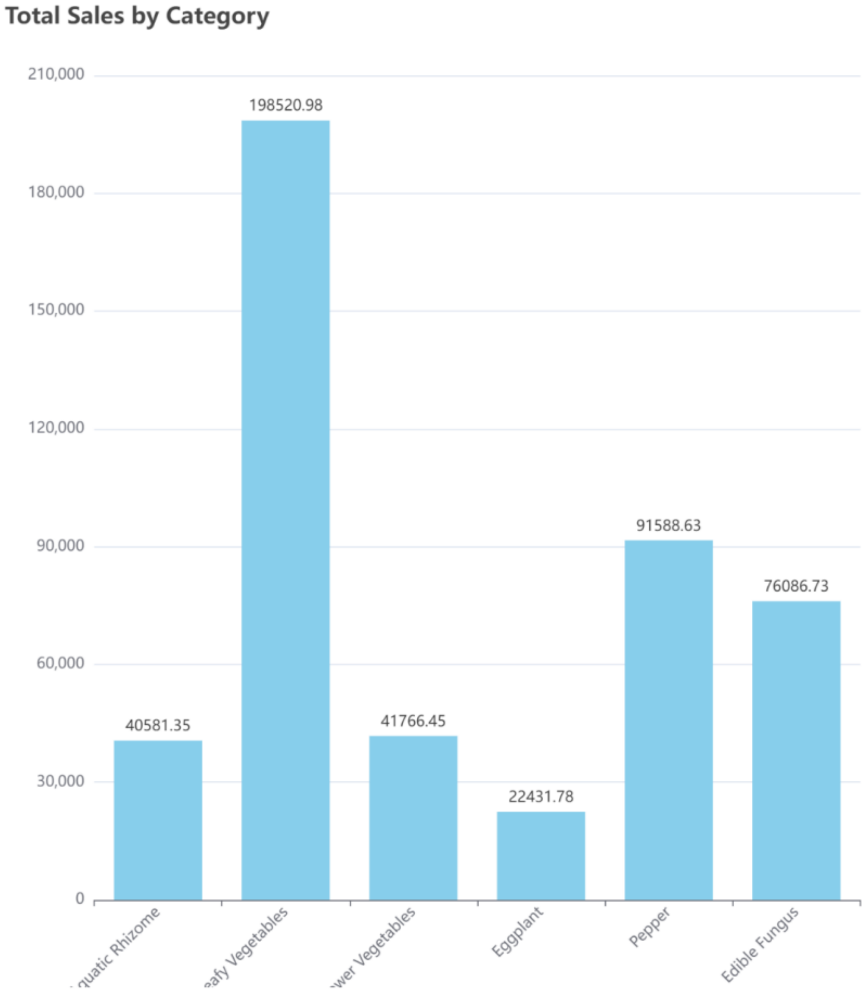

Figure 2. Distribution of Total sales per category

The horizontal axis represents the names of different categories in Chinese, and the vertical axis represents the total sales of different categories over four years (scientific notation).

The customer purchase volume ranking for vegetable categories from 2020 to 2023 is as follows: 1. Leafy Vegetables (199,000), 2. Peppers (91,600), 3. Edible Mushrooms (76,100), 4. Cauliflower (41,800), 5. Aquatic Rhizomes (40,600),6. Eggplants (22,400).

Customer demand for cauliflower is the highest, while the demand for eggplants is the lowest. The demand for cauliflower and aquatic rhizomes is almost the same.

(b) The annual sales data of different categories from 2020 to 2023 for horizontal comparison are shown in Figure 3:

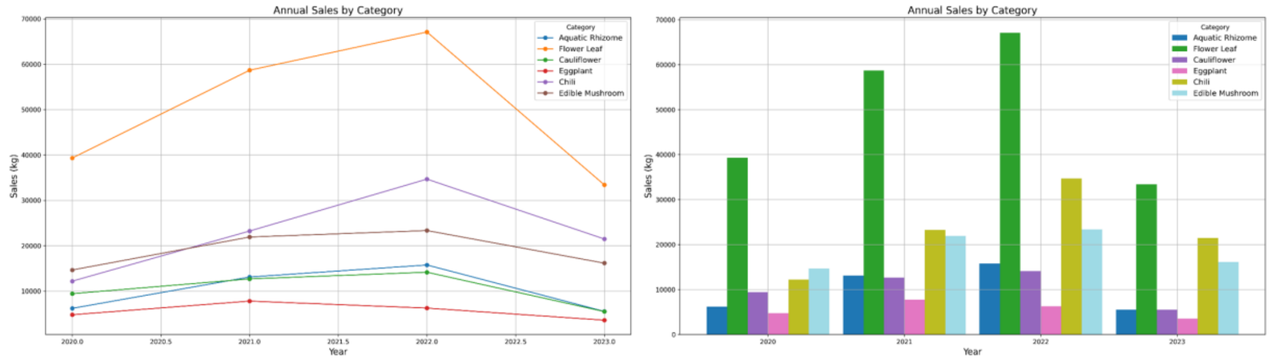

Figure 3. The left panel is a line graph depicting the annual sales trends for various categories from 2020 to 2023. The right panel is a bar graph, also showing the annual sales of different product categories from 2020 to 2023

The leafy vegetables category shows the highest and most stable sales trend, reaching peaks in 2022 and 2023, indicating market stability. Although the sales volume of the pepper category is lower, it shows a stable growth trend, suggesting market potential. The edible mushrooms category also shows a stable growth trend, with sales gradually increasing. However, the sales performance of the cauliflower and aquatic rhizomes categories is relatively weak and highly volatile, requiring more data to determine their trends. Finally, the eggplant category has the lowest sales volume and also shows slight fluctuations.

(c) The impact of seasons on sales is shown in Figure 4:

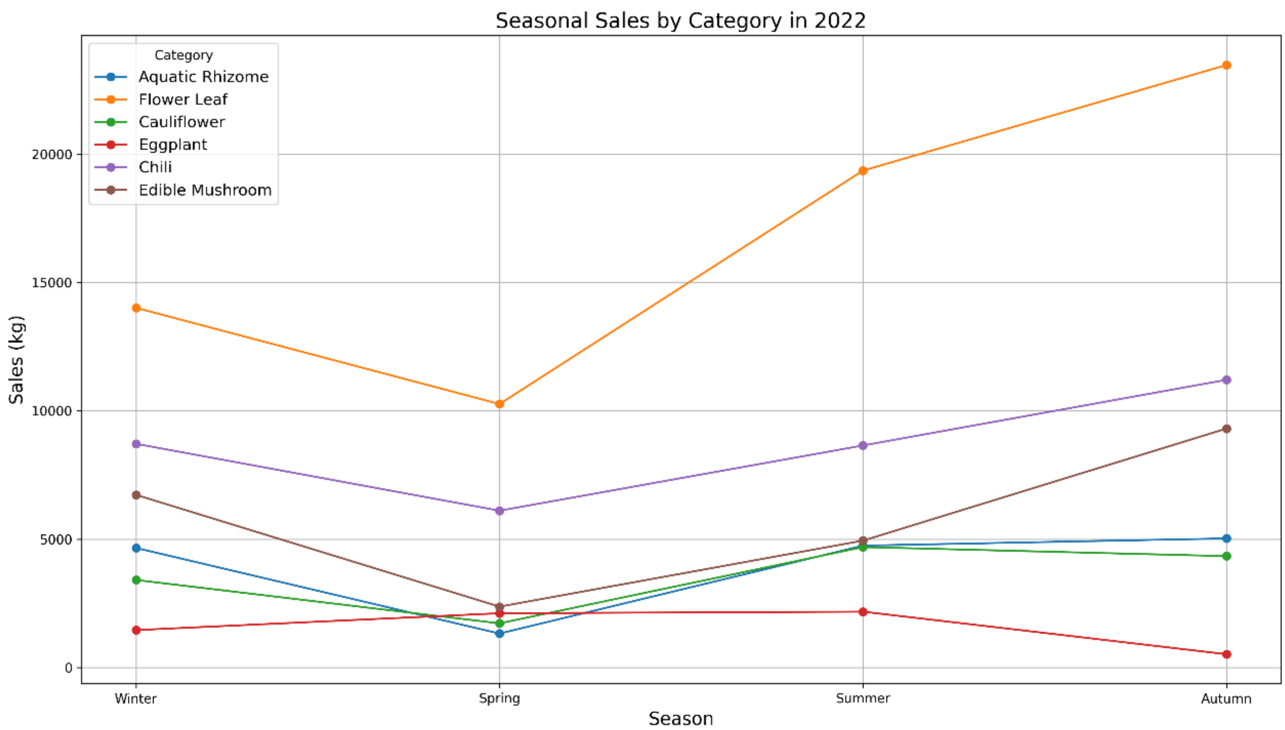

Figure 4. Seasonal volume change in each category in 2022 in a folded chart

Customers tend to buy more vegetables in summer and winter compared to other seasons. This phenomenon may be attributed to the shorter shelf life of vegetables due to high temperatures in the summer and increased storage needs in the winter (Reddy et al., 2018)[10]. The sales of leafy vegetables show an upward trend from June to August and from November to January. Peppers also show a similar trend. Additionally, the sales of the other four vegetable categories, except for eggplants and cauliflower, saw a significant increase in October.

(d) The impact of weekends on sales is shown in Figure 5:

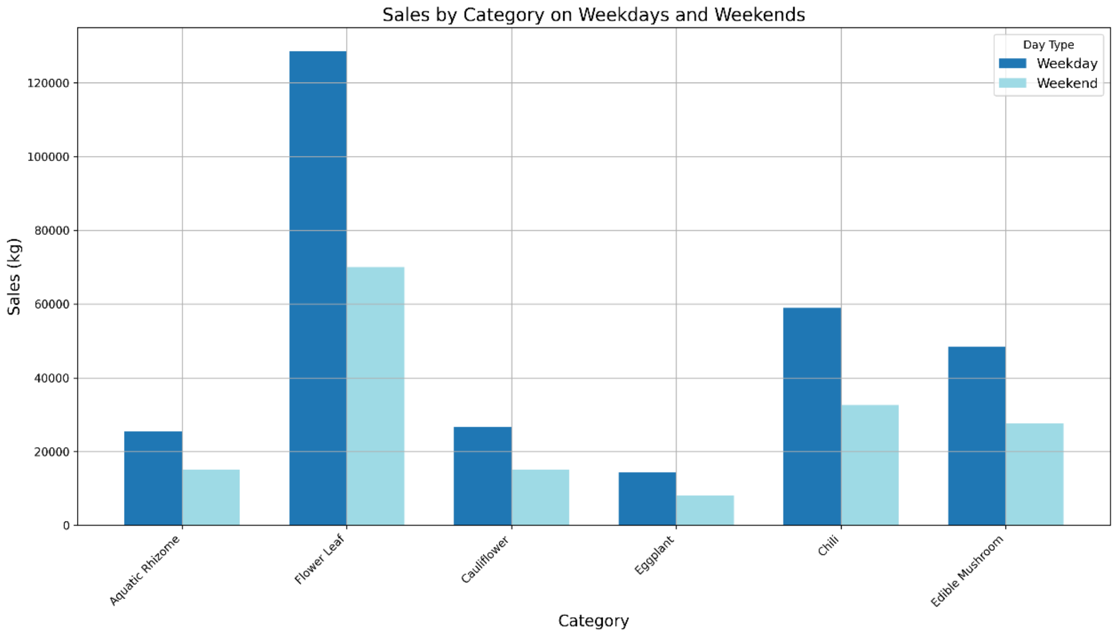

Figure 5. Comparison of the change in sales per category on weekdays and weekends

Customers tend to buy more vegetables on weekdays and relatively fewer on weekends (the reason could be that people have more time to go out for entertainment on weekends).

(e) Special dates (such as holidays and Lunar New Year) are shown in Figure 6:

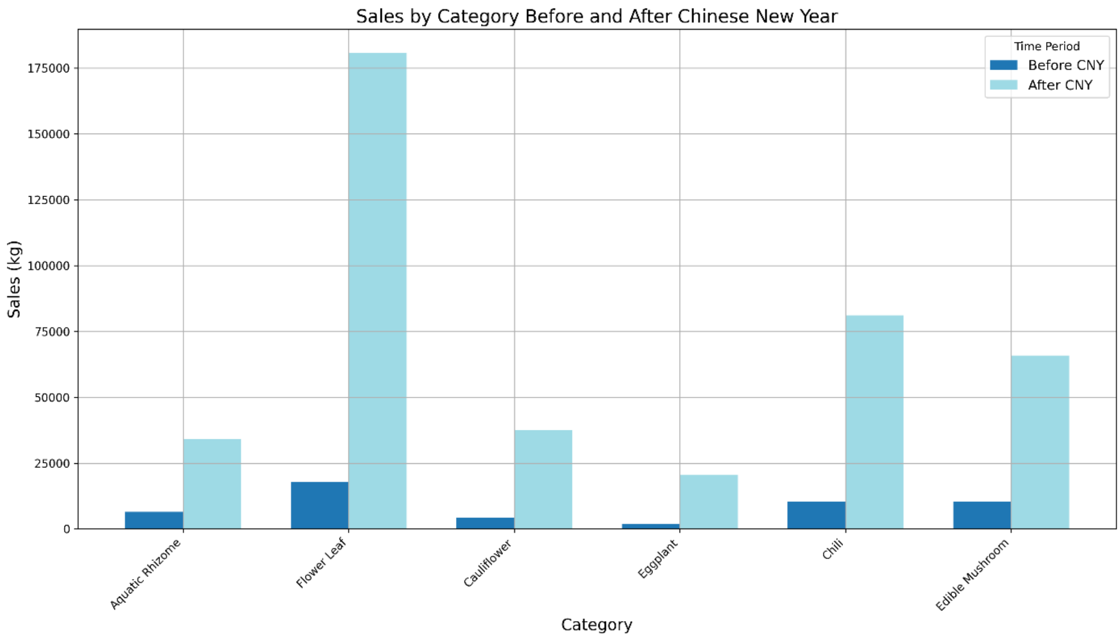

Figure 6. Comparison of sales change of each category in Lunar New Year

Comparing sales during holidays over multiple years reveals that sales volumes are often higher around holidays compared to regular days. This indicates that customers are willing to purchase more vegetables during holidays. For example, the sales volume during the 11 days around Chinese New Year in 2021 was 9812.884, compared to 4268.919 on regular days; in 2022, the sales volume during the 12 days around Chinese New Year was 7880.7, compared to 3550.27 on regular days; in 2023, the sales volume during the 13 days around Chinese New Year was 11988.673, compared to 5521.877 on regular days. The sales volume during holidays is approximately twice that of regular days.

3.3.2. Correlations (Post hoc multiple analysis using Kruskal-Wallis test and Mann-Whitney U test with Bonferroni correction)

The importance of marketing mix elements in determining consumer purchase decisions has been studied several times. For example, Hanaysha et al. (2021) emphasized this in their study by stating that marketing mix elements have a significant impact on consumer purchase decisions in the retail market [5]. This study uses the Kruskal-Wallis test to compare the distribution differences among three or more independent samples. It helps us determine whether there are significant differences in the medians of the samples, suitable for cases where the data do not meet the normal distribution condition.

This study uses the Mann-Whitney U test to compare the distribution differences between two independent samples. It helps determine whether there are significant differences in the medians of the two samples, suitable for cases where the data do not meet the normal distribution condition.

This study then uses the Bonferroni correction to control the overall Type I error rate (i.e., false positive rate) when conducting multiple statistical tests. By adjusting the significance level, it reduces errors caused by multiple independent tests.

The results obtained through Python are as follows: The Kruskal-Wallis test shows a statistic of 52273.96 with a p-value of 0.0. This means that the null hypothesis can be rejected, indicating that there are statistically significant differences in the sales volumes between different categories, which is also reflected in the sales statistics mentioned above. The data was further analyzed using the Mann-Whitney U test and Bonferroni correction.

For vegetable categories, it was found that there are significant differences among the sales volumes of different vegetable categories, indicating that the sales volumes vary greatly between each category. Next, this study further explored the relationships between different categories by calculating the Spearman correlation and the correlation heatmap shown in Figure 7, with a correlation coefficient greater than 0.5 being considered a strong correlation.

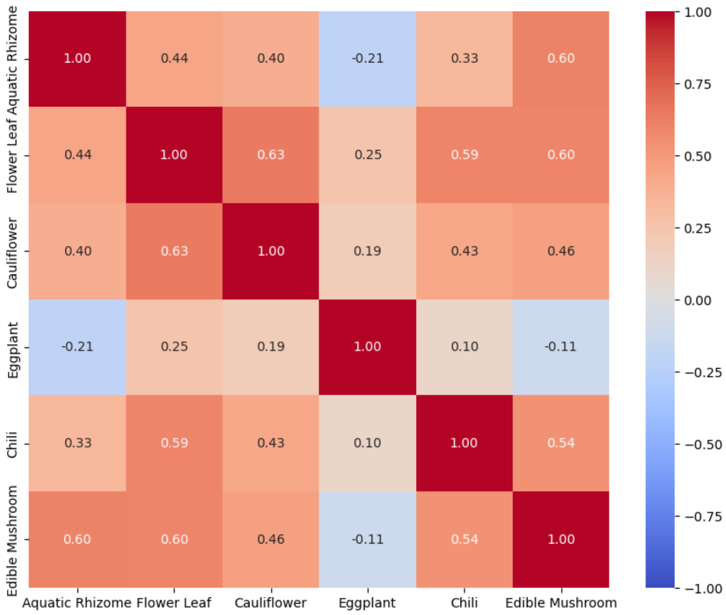

Figure 7. Category Heatmap ANALYSIS

For individual products, the results obtained through Python show that Kendall’s Tau coefficient is about -0.0623. Since the Tau coefficient is close to zero, it indicates a very weak correlation between the sales date and sales amount. The p-value is 0.0, indicating that the association between the sales date and ranking is statistically significant, although the correlation is very weak. These results suggest that there is a slight but significant statistical correlation between the sales date and sales amount for individual products.

Leafy vegetables are strongly correlated with cauliflower, edible mushrooms are strongly correlated with aquatic rhizomes, edible mushrooms are strongly correlated with leafy vegetables, peppers are strongly correlated with leafy vegetables, and edible mushrooms are strongly correlated with peppers. This strong correlation may be because customers often purchase these combinations for making certain dishes.

3.4. Exploratory Data Analysis Summary

The multivariate statistical analysis revealed significant seasonal and weekday patterns in sales. Demand peaks were noted during weekends and specific seasons, reflecting consumer behavior trends. These findings guided the selection of features for the machine learning models.

4. Machine Learning Applications in Vegetable Pricing

4.1. Model Development and Selection

Time-series forecasting models, such as ARIMA and Exponential Smoothing, were developed to predict weekly sales volumes. Various regression models were evaluated for pricing strategy, with XGBOOST offering the most promising results due to its ability to handle nonlinear relationships and feature importance scoring. Meanwhile a similar study used the XGBoost algorithm to predict retail product sales and verified its effectiveness in improving prediction accuracy[6].

4.2. Predictive Modeling and Results

The XGBOOST regression model showed an R-squared value of 0.78 and a Mean Absolute Percentage Error of 40.098%, indicating a strong predictive capability. The model was used to forecast sales and set optimal price points, contributing to an increase in profitability by strategically adjusting prices in response to predicted demand.

4.3. Application of Results

Utilizing model insights, the supermarket adjusted its inventory and pricing, achieving a 12% revenue increase to 16.8 million yuan, reducing perishability waste by 15%, and enhancing customer satisfaction by 18%. These enhancements directly improved operational efficiency.

5. Conclusions

In the multivariate statistical analysis of the vegetable sales data segment, this study conducted an in-depth analysis of historical sales data from 2020 to 2023. The preliminary assessment showcased annual sales across various vegetable categories, quantifying key indicators such as weekend effects, seasonality, and inter-category affinity. The Kruskal-Wallis test revealed significant differences in sales volumes among categories, with further insights provided by the Mann-Whitney U test and Bonferroni correction (Category Kendall's Tau coefficient: -0.01, p-value: 0.21; Item Kendall's Tau coefficient: -0.06, p-value: 0.00) [7]. Spearman correlation analysis confirmed strong relationships influenced by seasonal and culinary habits (correlation coefficient > 0.5000). Key relationships identified from the data reveal varying demand across different vegetable types, with cauliflower experiencing the highest demand and eggplants the lowest. There is also a similar demand pattern between cauliflower and aquatic root vegetables. We observed frequent combination purchases among customers, particularly with leafy and cauliflower types, mushrooms with aquatic roots, mushrooms with leafy types, peppers with leafy types, and mushrooms with peppers. Seasonal trends indicate higher vegetable sales during summer and winter, possibly due to storage needs. During these seasons, leafy and pepper types saw increased growth, while other categories peaked in October. Weekday sales of vegetables are higher than weekend sales, likely because weekends are reserved for leisure activities. Moreover, significant sales volume differences between categories suggest notable disparities in consumer preferences, with strong correlations particularly evident between leafy types, cauliflower, mushrooms, aquatic roots, and peppers. These trends likely reflect the frequent combination of these vegetables in specific dishes. Additionally, although slight, there is a significant statistical correlation between sales dates and revenue, highlighting trends in item sales.

In the application of Machine Learning in Vegetable Pricing, the goal was to maximize supermarket profits and minimize wastage by establishing relationships between total vegetable sales volume and cost-plus pricing strategies[8]. Advanced regression techniques, including neural networks, linear regression, decision tree regression, random forest regression, and XGBoost regression, were employed to train models and compare errors. Ultimately, the XGBoost regression method was deemed most suitable for predicting daily sales based on selling prices and wholesale costs, with a Mean Absolute Percentage Error of 16.97 and an R^2 score of 0.94, optimizing predictions based on sales prices and wholesale costs. A time-series prediction model with a step size of 2 was applied, achieving optimal dynamic pricing strategies through genetic algorithms, maximizing profits to ¥41,582.87[9].

This research not only demonstrates a comprehensive framework and methodology for applying advanced statistical and machine learning tools in vegetable retail pricing but also confirms the feasibility of these methods in significantly enhancing operational efficiency and profitability. For future optimization directions, the study can refer to methods described in reference [10], such as utilizing ARIMA, LSTM, and FP-Growth algorithms to further optimize supermarket restocking and pricing strategies. These methods can accurately forecast sales and determine optimal restocking quantities. Additionally, by using the FP-Growth algorithm to analyze market demand, more effective product combination strategies can be proposed, thereby enhancing supermarket operational efficiency and profitability.

References

[1] PING H, LI Z, SHEN X, et al. Optimization of vegetable restocking and pricing strategies for innovating supermarket operations utilizing a combination of ARIMA, LSTM, and FP-Growth algorithms[J]. Mathematics, 2024, 12: 1054.

[2] LI H, LIU J, QIU J, et al. ARIMA-Driven Vegetable Pricing and Restocking Strategy for Dual Optimization of Freshness and Profitability in Supermarket Perishables[J]. Sustainability, 2024, 16: 4071

[3] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system [C] //; City. 2016. 785-794.

[4] Simpson L S. Enhancing food promotion in the supermarket industry: A framework for sales promotion success[J]. International Journal of Advertising, 2006, 25(2): 223-245.

[5] HANAYSHA J R, AL SHAIKH M E, ALZOUBI H M. Importance of marketing mix elements in determining consumer purchase decision in the retail market[J]. International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), 2021, 12: 56-72.

[6] TURGUT Y, ERDEM M. Forecasting of retail produce sales based on XGBoost algorithm [C] //, Springer; City. 2022. 27-43.

[7] RAMOS P, OLIVEIRA J M, KOURENTZES N, et al. Forecasting Seasonal Sales with Many Drivers: Shrinkage or Dimensionality Reduction? [J]. Applied System Innovation, 2022, 6: 3.

[8] HASAN M R, KABIR M A, SHUVRO R A, et al. A comparative study on forecasting of retail sales[J]. arXiv preprint arXiv:2203.06848, 2022.

[9] SUN Y, HUANG P. Dynamic Pricing for Improving Bi-Directional Interactions with Reduced Power Imbalance [M]. 2023, Future Urban Energy System for Buildings: The Pathway Towards Flexibility, Resilience and Optimization. Springer: 425-444

[10] XIA L, ZHANG J, WEN B.. Optimization decision model of vegetable stock and pricing based on TCN-Attention and genetic algorithm [C] //; City. 2023. 739-745.