Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/53/20241461

Walmart sales prediction based on random forest model and application of feature importance

Deli Chen

Whittier Christian High School, La Habra CA90631, United States

delichen@wchs.com

Abstract. Sales forecasting is crucial for efficient resource allocation and inventory management in retail. This study employs Random Forest to predict weekly sales for 45 Walmart stores, leveraging a diverse dataset with store-specific sales and external factors. Through meticulous preprocessing and model application, one achieves outstanding accuracy, with a Weighted Mean Absolute Error (WMAE) as low as 1.2030 and an impressive accuracy rate of 98.8%. Additionally, integrating feature importance ranking sheds light on influential variables in sales forecasting. This study provides a blueprint for developing precise and adaptable sales forecasting models, offering profound significance for the retail industry. It underscores the effectiveness of machine learning techniques, e.g., Random Forest and insightful feature engineering in achieving highly accurate predictions. By enhancing the industry's understanding of intricate sales dynamics, this research contributes to optimizing resource allocation, inventory management, and strategic planning. Ultimately, it drives operational efficiency and success in the dynamic landscape of the retail sector.

Keywords: Random forest, feature importance, WMAE.

1. Introduction

In the dynamic and highly competitive landscape of modern business, the ability to accurately forecast sales have become a critical strategic imperative. Sales forecasting, a pivotal component of business management, plays a prominent role in resource allocation, marketing strategies, and financial planning [1]. It also serves a helpful function for managing diversified processes in the business [2]. Not all of businesses can easily utilize sales forecasting, however, as the rapid change of customers or other environmental factors influence the sales, such as in fashion retailing that has a short product's life cycle [3]. Undoubtedly, using the method of sales forecasting and maximally control the budget and properly adjust sales in certain periods [4].

Sales forecasting is a continuous development, which can be traced back to more than 50 years ago [5]. Many real-world industries apply sales forecasting and gradually develops various systems of sales prediction. Time series sales-forecasting model, which can be divided to linear models and nonlinear models, have distinct effective applications. Especially for linear models, the autoregressive integrated moving averages (ARIMA) model contributes to effective applications about sales trend, capturing the long memory of the series [6]. Besides, some researchers have established the univariate forecasting model, a model directly utilizes the input data extracted from historical sales data, typically under the fundamental premise that the underlying data-generating process of the time series remains constant [5].

DOI: 10.54254/2755-2721/53/20241461

This assumption contradicts with a great number of influencing factors like dynamic business environment, which leads to the failure of using univariate forecasting model in a unstable business [5].

Multivariate forecasting model seems to be a potential model for a fluctuate environment, which concerns with the various influencing factors, such as a holiday promotion and Consumer Price Index (CPI). It is a hard task to distinguish the actual relationship between the various influencing factors and final sales, and usually, the insufficient historical data can cause the inaccuracy of the forecasting [5]. The advantage of using multivariate forecasting model, nevertheless, outweighs the its deficiencies. Multivariate forecasting model is potentially having a higher accuracy because it tests multiple variables, compared to univariate models. It is also a great optimizing tool for resource allocation based on the impact of multiple variables on resource needs. Attempting a multivariate forecasting model shows to be a useful tool in managing the resources and observing indicators of the data, which leads us to explore further relationship between sales and other variables [6, 7].

Utilizing a Random Forest model for sales prediction presents several compelling advantages and motivations in the context of a multivariate forecasting model within a fluctuating environment. Random Forest, as an ensemble learning method, combines predictions from multiple decision trees, leading to enhanced accuracy and resilience against overfitting, a valuable trait when dealing with multiple influencing factors [8]. This model excels in handling multivariate data, effectively capturing non-linear relationships, and remaining robust to outliers as it has the outlier detection techniques [9]. Moreover, it offers insight into feature importance, aiding in resource allocation and decision-making [10]. Importantly, its reduced susceptibility to data overfitting, adaptability to missing data, and parallelization capabilities make it a dependable choice, even when historical data is limited. Overall, Random Forest stands out as a potent tool for improving the precision and reliability of sales forecasts, particularly in complex, changing environments. This paper will show how to use random forest to forecast the sales in business. This paper will explore dataset information, data preprocessing, data analysis, sales prediction using random forest, feature engineering, model performance and comparison, and limitations and future outlooks.

2. Data and method

Walmart, one of the world's largest and most renowned retail giants, operates an extensive network of stores across the United States. Understanding and predicting sales performance within this vast retail ecosystem is a fundamental imperative for both operational efficiency and strategic planning. In kaggle, this study picks Walmart datasets that includes 45 stores, each identified by a unique number [7]. Given train data with variables, the main objective is to predict week sales of store in test data. This dataset encompasses various crucial variables, including store-specific sales, temporal data, and external factors like a bool value of whether or not it is a holiday, temperature, fuel prices, promotional markdowns, the Consumer Price Index (CPI), store's size, and local unemployment rates.

The meaning and the significance of each variable in the Walmart sales dataset:

- Store: An identifier for the specific Walmart store, allowing store-specific analysis. It counts from 1 to 45.
- Date: The temporal variable that represents the date in day-month-year format.
- Weekly Sales ($): The target variable, representing the weekly sales figures for each store.
- Markdowns 1-5: Anonymized Walmart promotional markdown data, post-November 2011, with occasional missing values marked as Null.
- Holiday Flag: A binary indicator that identifies weeks coinciding with specific holidays. In this dataset, Walmart have four critical holidays: Super Bowl, Labour Day, Thanksgiving, and Christmas.
- Temperature (°C): The temperature in Celsius on the day of sale.
- Fuel Price ($): Cost of fuel in the region.
- CPI: Prevailing consumer price index.
- Unemployment: Prevailing unemployment rate.

There are four separate Comma Separate Value files, so one merges them together to have a better view to analyse. Then one finds out the shape of the data, eliminating weekly sales that has 0 or negative

DOI: 10.54254/2755-2721/53/20241461

value, which is unreasonable for a store. Moreover, this study checks if there is any NaN value in the dataset. Since Markdowns 1-5 have many NaN value, this study simply changes all of their NaN values to 0. Morevier, since the date cannot be easily interpreted by a model, one converts date to the type of datetime, which includes week, month, and year, and add these three columns to the dataframe. Furthermore, to avoid multicollinearity problem, this study makes a heatmap to show correlations and try to drop the column that may influence the model prediction. In this case, since Markdown 4 and 5 are highly correlated with Markdown 1, one drops Markdown 4 and 5 (seen from Figure 1).

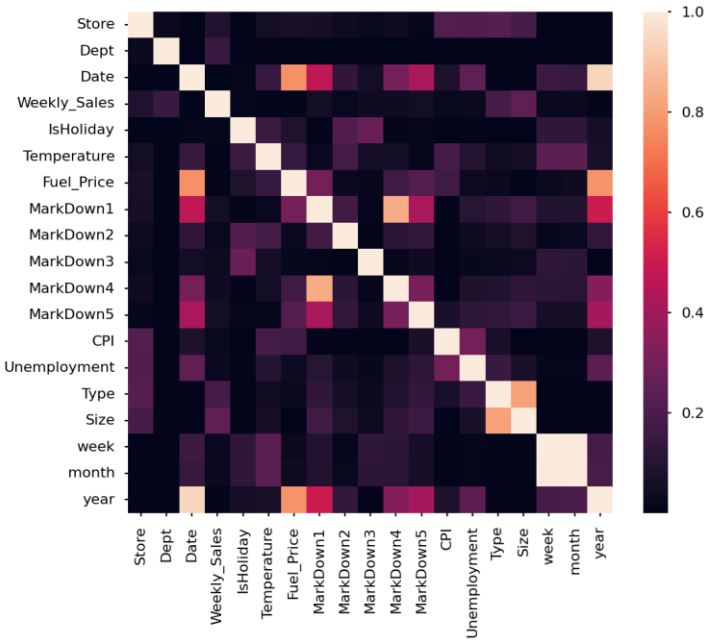


Figure 1. Heatmap that shows the correlations (Photo/Picture credit: Original).

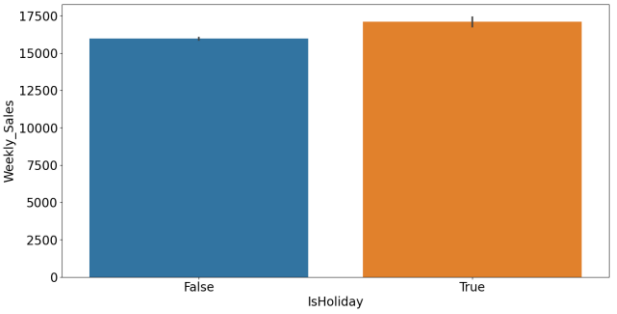


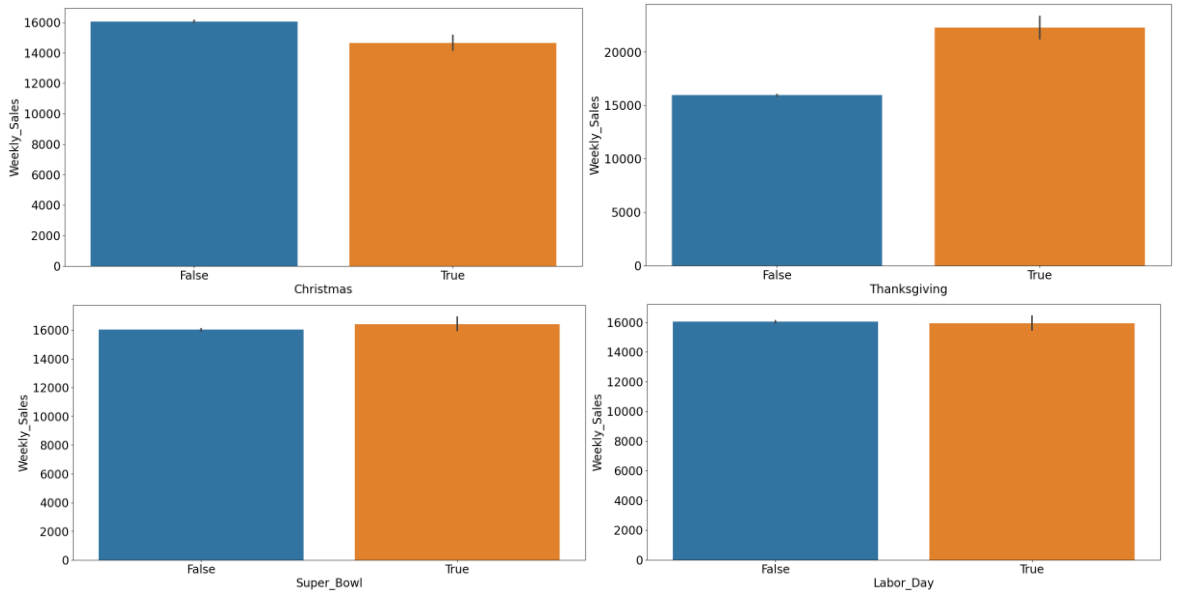
Figure 2. Holiday Weekly Sales vs. Non-Holiday Weekly Sales (Photo/Picture credit: Original).

3. Results and discussion

In order to have a clear view, this study draws bar plots to see the weekly sales of holiday and non-holiday (Seen from Figure 2 and Figure 3). This study provides a pie chart to see the percentage of type A, type B, and type C for the stores (Figures 4). The side by side bar graph shown below provide the average weekly sales according to holidays by types as given in Figure 5. The type for each type are distributed by its size, as shown in Figure 6. For store distribution, one finds out that Store 10 has the highest individual weekly sales, and Store 20 has the highest weekly sales average as shown in Figure 7 and Figure 8. For department distribution, Department 72 has highest individual weekly sales, but its average weekly sales is lower than the average weekly sales of Department 92, which can lead to an assumption that Department 72 is a seasonal department as illustrated in Figure 9 and Figure 10.

DOI: 10.54254/2755-2721/53/20241461

Compared each month's average weekly sale in 2010, 2011, 2012, it looks like 2012 average weekly sales are lowest because it misses Novemember and December, but it is actually because that 2012 misses average weekly sales of November and December; 2012 is close to the averages of 2010 and 2011, so adding November and December will absolutely make 2012 the highest average weekly sales (shown in Figure 11). The weekly sales in 2010, 2011, 2012 reveals that 51th week and 47th weeks have significantly higher averages as Christmas, Thankgiving, and Black Friday effects (shown in Figure 12). This study also plots fuel price, temperature, CPI, unemployment rate against the weekly sales respectively, finding that they have no apparent correlation and fluactuate randomly.



Figures 3. Christmas, Thanksgiving, Super Bowl, Labor Day Weekly Sales vs non-Holiday Weekly Sales (Photo/Picture credit: Original).

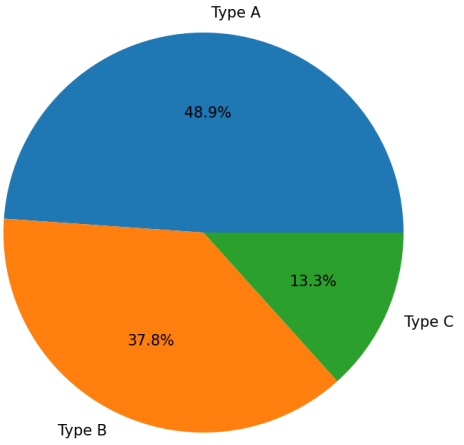


Figure 4. Proportion for each type of store (Photo/Picture credit: Original).

DOI: 10.54254/2755-2721/53/20241461

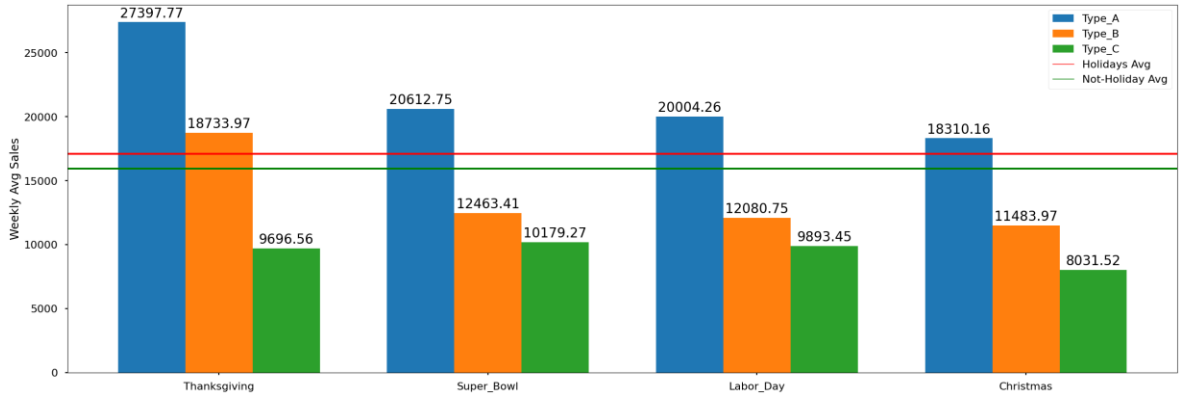


Figure 5. Average weekly sales for Thanksgiving, Super Bowl, Labor Day, and Christmas in each type of store (Photo/Picture credit: Original).

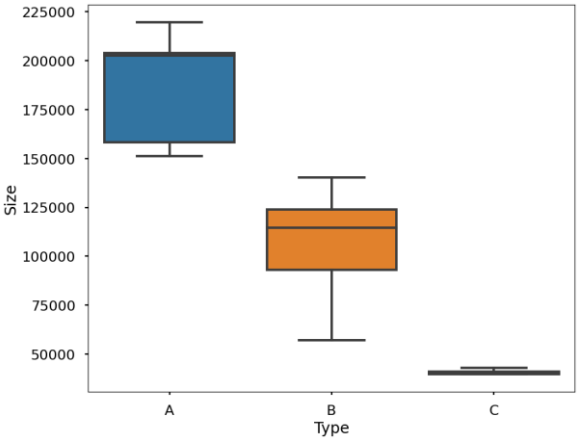


Figure 6. Size distribution for Type A, Type B, and Type C (Photo/Picture credit: Original).

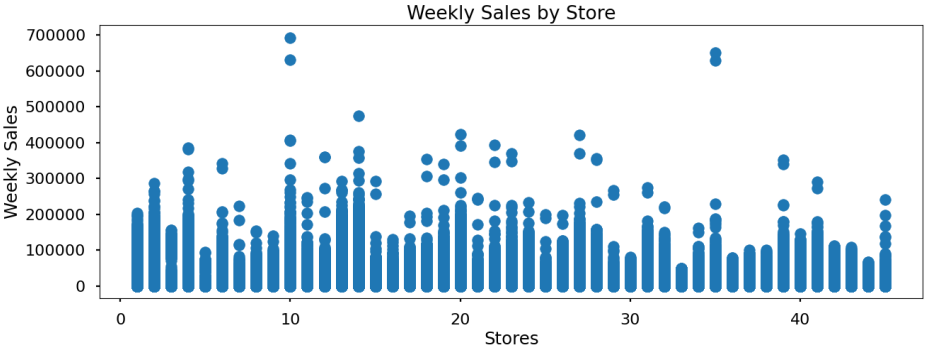


Figure 7. Every sale in each store (Photo/Picture credit: Original).

DOI: 10.54254/2755-2721/53/20241461

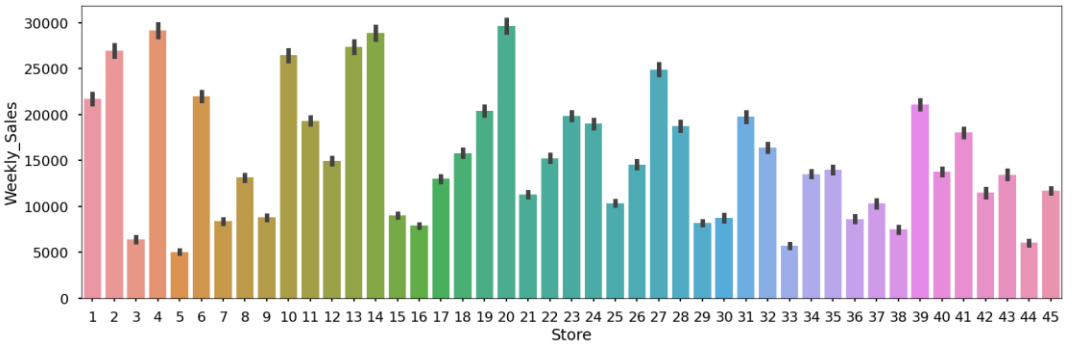


Figure 8. Each store's average sale (Photo/Picture credit: Original).

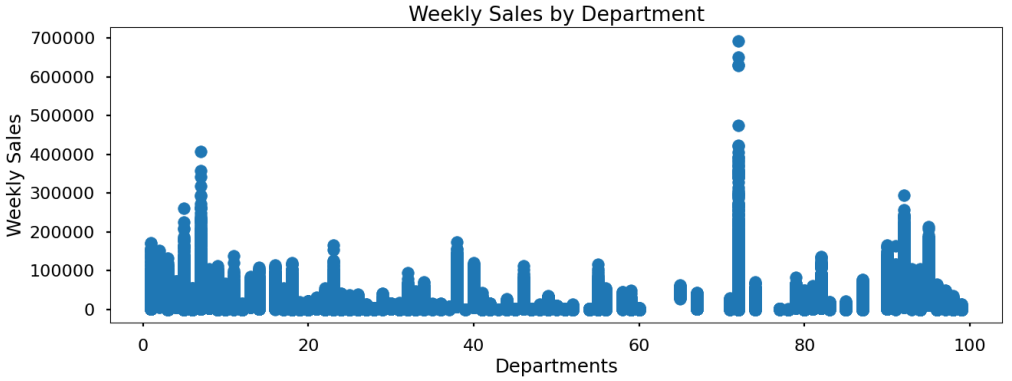


Figure 9. Every sale in each department (Photo/Picture credit: Original).

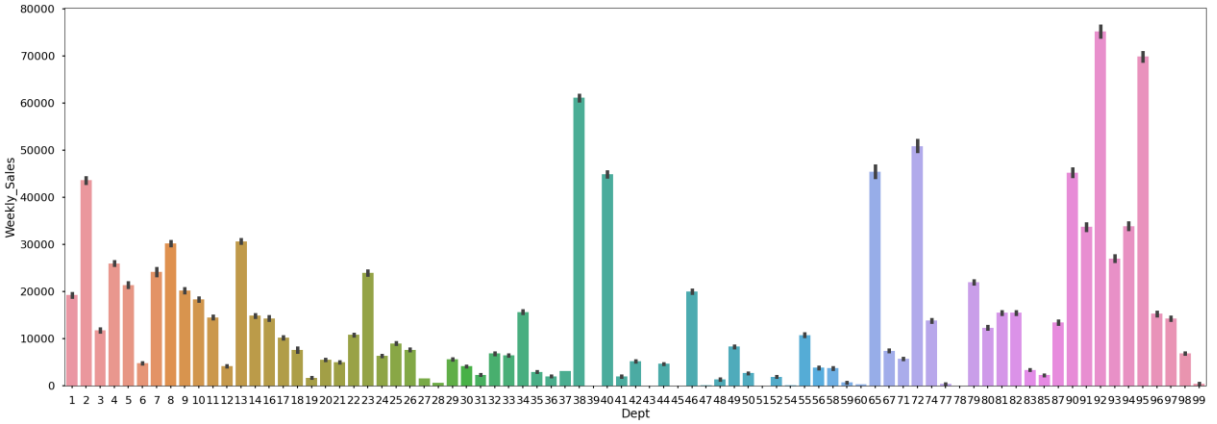


Figure 10. Each department's average sale (Photo/Picture credit: Original).

DOI: 10.54254/2755-2721/53/20241461

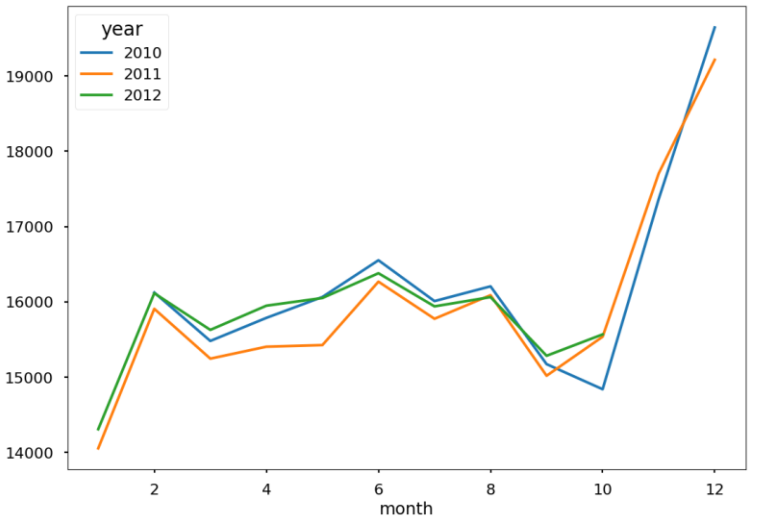


Figure 11. Each month's average weekly sale in 2010, 2011, and 2012 (Photo/Picture credit: Original).

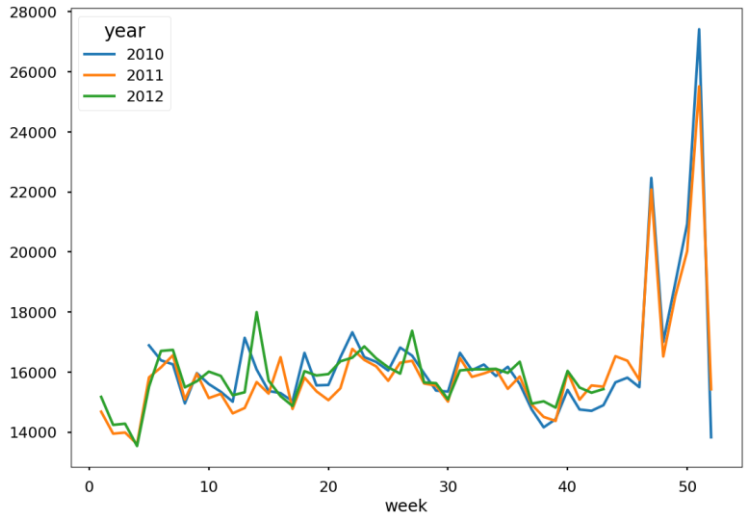


Figure 12. Weekly sales in 2010, 2011, and 2012 (Photo/Picture credit: Original).

To comprehensively evaluate the performance of the Random Forest model for sales forecasting, this study meticulously splits the training dataset into a 70/30 proportion, designating 70% of the data for training and reserving the remaining 30% for testing. This approach was chosen for a specific reason: one wanted to ensure the continuity of the date features, which provide valuable insights into sales patterns. By manually splitting the data, one maintained a coherent timeline of sales data, allowing the model to recognize and capture the temporal dynamics inherent in the sales process. The Random Forest model was meticulously configured with a careful selection of hyperparameters to ensure optimal performance. The model was set to include 50 trees within the forest, offering a balance between accuracy and computational efficiency. One also defined a random state of 42 to enable reproducibility, a critical aspect in machine learning experiments, ensuring that results can be consistently replicated. The maximum depth of 35 was specified to prevent overfitting, as overly deep trees can fit the training data too closely, resulting in reduced generalization to unseen data. Additionally, one sets the minimum number of samples required to split a node to 10, which served to control the granularity of splits in the decision trees. To measure the performance of the sales forecasting model, one employed the Weighted Mean Absolute Error (WMAE) as the primary evaluation metric. The WMAE metric is highly valuable

DOI: 10.54254/2755-2721/53/20241461

in scenarios where the significance of predictions may vary across different data points. In the case, the factors of importance originate from both the user and the seller's perspectives. Each prediction is assigned a weight, signifying its subjective importance in the context of the specific use case [11]. The WMAE formula is constructed to address the uniqueness of each prediction, accounting for the heterogeneity in the importance attached to it:

$$WMAE = \frac{\Sigma_i^U \Sigma_j^{N_i} \omega_{i,j} |p_{i,j} - r_{i,j}|}{\Sigma_i^U \Sigma_j^{N_i} \omega_{i,j}} \quad (1)$$

Here, U represents the number of users, Ni represents the number of items predicted for the i th-user, $r_{i,j}$ represents the rating given by the i th-user to the item I_j, $p_{i,j}$ represents the rating predicted by the model, $w_{i,j}$ represents the weight associated to this prediction [11].

In the realm of the Walmart sales forecasting, it's paramount to consider the weight associated with each prediction. This weight signifies the varying significance attached to different predictions, reflecting the importance from the perspectives of both the customers (users) and the retailer (seller). Specifically, in the case, one considers the holiday factor as a significant contributor to the weight of predictions. During holiday weeks, customers are more likely to increase their purchases, and the retailer is keen on ensuring optimal stocking and resource allocation during these periods. Hence, one sets a weight of 5 for holiday weeks and 1 for normal weeks. This differential weighting reflects the higher significance attributed to predicting weekly sales during holiday weeks, compared to non-holiday weeks. In addition to the WMAE metric, one incorporated a feature importance ranking (FIR) mechanism within the Random Forest model. FIR is a pivotal tool in evaluating the impact of each input feature or variable on the overall performance of a supervised learning model [12]. By ranking features based on their importance, one gains valuable insights into which variables have the most influence on the model's predictions as presented in Figure 13.

The rigorous approach to evaluating the Random Forest model for sales forecasting underscores the commitment to precision and robustness. The manual data splitting strategy preserves the temporal continuity of date features, ensuring the model captures essential time-based patterns. The selection of hyperparameters in the Random Forest model demonstrates the dedication to fine-tuning for optimal performance. One has adopted the WMAE metric, which aligns with the varying significance of predictions and emphasizes the holiday factor. the weighting scheme considers the perspectives of both customers and the retailer, effectively reflecting the relative importance of different predictions. Additionally, the integration of feature importance ranking (FIR) enriches the analysis by revealing the influential variables, thereby guiding the decision-making processes. The results of the Random Forest Regressor model in the analysis reveal intriguing insights into the significance of specific feature columns and the overall performance of the model in predicting sales. The four distinct scenarios, each with varying sets of input features, provided a comprehensive perspective on the influence of these features on predictive accuracy.

In the first scenario, where the model was trained without the divided holiday columns, one observed a relatively high Weighted Mean Absolute Error (WMAE) of 5850. This suggests that omitting information related to holiday promotions in the dataset has a substantial negative impact on the model's accuracy. However, the model still achieved an impressive 95% accuracy, emphasizing the robustness of Random Forest in handling other variables. Similarly, the second scenario, where the month column was excluded, yielded a WMAE of 5494. While the accuracy slightly improved to 95.3%, this scenario reaffirmed that seasonality, represented by the month column, plays a significant role in sales predictions. The third scenario, using the entire dataset, substantially improved model performance with a considerably lower WMAE of 2450. The accuracy also saw a noteworthy boost, reaching 97.3%. This underscores the importance of comprehensive data in making more accurate sales predictions. The fourth scenario, employing the whole dataset with feature selection, demonstrated the potential for further refinement. With a significantly reduced WMAE of 1801, this approach reached an impressive accuracy of 98.8%. Feature selection effectively identified the most influential variables, enhancing the model's precision. In general, the findings underscore the value of feature selection and the critical role

DOI: 10.54254/2755-2721/53/20241461

of holiday promotions and seasonality in sales forecasting. The Random Forest Regressor showcased its adaptability and robustness, consistently providing high levels of accuracy in various scenarios. By leveraging these insights, businesses can optimize resource allocation and decision-making, ultimately leading to more accurate sales predictions and improved operational efficiency. The choice of feature selection and careful consideration of data variables are pivotal in achieving highly accurate predictive models.

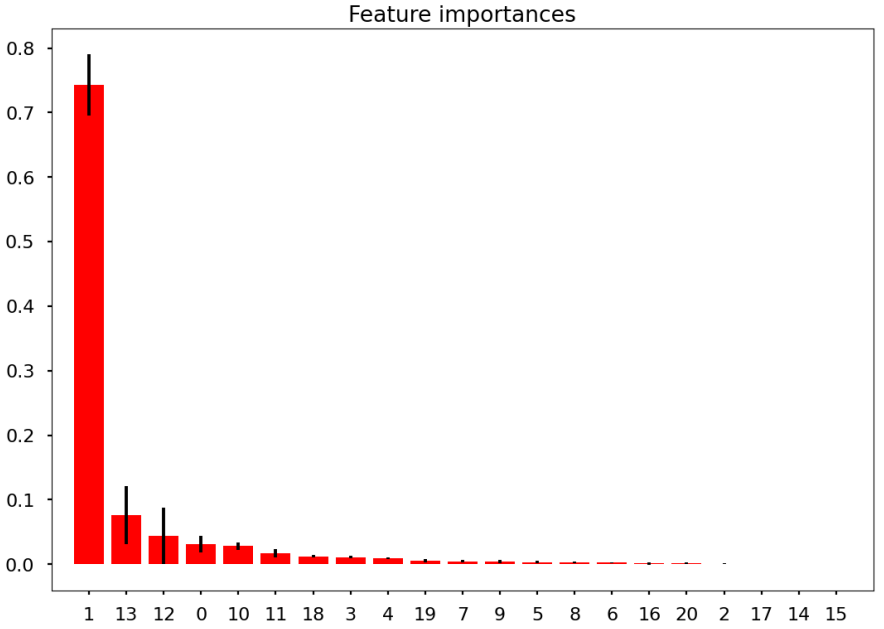


Figure 13. feature importance ranking (FIR) for whole encoded dataset (Photo/Picture credit: Original).

4. Conclusion

To sum up, this study provides a blueprint for developing precise and adaptable sales forecasting models, offering profound significance for the retail industry. It underscores the effectiveness of machine learning techniques like Random Forest and insightful feature engineering in achieving highly accurate predictions. By enhancing the industry's understanding of intricate sales dynamics, this research contributes to optimizing resource allocation, inventory management, and strategic planning. The limitations of the sales forecasting using the features in Random Forest Regressor include data dependency, feature engineering, data distribution changes, and overfitting risks, underscoring the need for better data quality, advanced feature engineering, ongoing monitoring, and hyperparameter fine-tuning. Furthermore, sales forecasting encounters inherent challenges, including the unpredictable nature of consumer behaviour. One key limitation is that customers' purchase decisions are influenced by a multitude of factors, often varying with personal preferences and immediate needs. While historical data provides insights, it may not capture the subtle shifts in buying patterns, making accurate predictions challenging. Additionally, external factors such as economic changes, unforeseen events, or even cultural trends can significantly impact sales, introducing further uncertainty. The future outlook for the sales forecasting model is focused on enhanced adaptability. This includes integrating advanced data sources, embracing deep learning techniques, enabling real-time predictions, and implementing explainable AI for transparency. All these techniques will refine the predictions to meet dynamic business needs.

References

[1] Gustriansyah R, Ermatita E and Rini D P 2022 Expert Systems with Applications vol 207 p 118043.

DOI: 10.54254/2755-2721/53/20241461

[2] Kahn K B and Adams M E 2001 The Journal of Business Forecasting Methods and Systems vol 19 p 19.
[3] Liu N, Ren S, Choi T M, Hui C L and Ng S F 2013 Mathematical Problems in Engineering vol 2013 pp 1–9.
[4] Xia Z, Xue S, Wu L, Sun J, Chen Y and Zhang R 2020 Distributed and Parallel Databases vol 38(3) p 713–738.
[5] Edling E 2016 Intelligent Decision-Making Models for Production and Retail Operations Createspace Independent Publishing Platform (Berlin Heidelberg: Springer)
[6] Aras S Deveci K İ and Polay C 2017 Journal of Business Economics and Management vol 18(5) pp 803–832.
[7] Walmart Sales Prediction (nd) Wwwkagglecom Retrieved from: https://wwwkagglecom/datasets/divyajeetthakur/walmart-sales-prediction/data
[8] Segal M R 2004 Machine learning benchmarks and random forest regression (New York: Escholar).
[9] Zhang J, Zulkernine M and Haque A 2008 IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews) vol 38(5) pp 649-659.
[10] Cassidy A P and Deviney F A (2014 October) Calculating feature importance in data streams with concept drift using Online Random Forest In 2014 IEEE International Conference on Big Data (Big Data) pp 23-28.
[11] Cleger-Tamayo S, Fernández-Luna J M and Huete J F 2012 RUE@ RecSys pp 24-26.
[12] Wojtas M and Chen K 2020 Advances in Neural Information Processing Systems vol 33 pp 5105-5114.