IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 12, December 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 12, December 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0912043 |

Comparative Analysis of Regression Techniques for Retail Sales Forecasting: A Case Study on Walmart

Prof. Abhishek Singh, Prof. Zohaib Hasan, Prof. Saurabh Sharma, Prof. Vishal Paranjape

Department of Computer Science and Engineering, Global Nature Care Sangathan's Group of Institutions, Jabalpur, Madhya Pradesh, India

ABSTRACT*:* This research paper presents a comprehensive analysis of various regression techniques for forecasting Walmart sales, focusing on Ridge Regression, Lasso Regression, Decision Tree Regression, and Random Forest Regression. Using historical sales data, the study evaluates the performance of these models based on metrics such as Root Mean Squared Error (RMSE) and R-squared (R^2). The results indicate that while Ridge and Lasso regressions perform moderately, Decision Tree and Random Forest regressors significantly outperform them, with the latter achieving the highest predictive accuracy. This study provides valuable insights into the effectiveness of different regression models for retail sales forecasting, contributing to the optimization of inventory management and strategic planning in the retail sector.

KEYWORDS*:* Sales Forecasting, Regression Analysis, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression

I. INTRODUCTION

Sales forecasting is a vital function for retail businesses, providing critical insights for inventory management, supply chain optimization, and strategic planning. Accurate sales forecasts enable retailers to minimize stockouts and overstock situations, improve cash flow management, and enhance customer satisfaction by ensuring the right products are available at the right time. In the context of large retail chains like Walmart, which operates a vast network of stores with diverse product lines, the complexity of sales forecasting increases significantly.

The advent of big data and machine learning has revolutionized the approach to sales forecasting. Traditional methods, such as moving averages and exponential smoothing, often fail to capture complex patterns and interactions within the data. Machine learning techniques, on the other hand, can model nonlinear relationships and interactions between variables, leading to more accurate and robust forecasts. This study explores the application of various regression techniques for forecasting weekly sales at Walmart, leveraging historical sales data and additional features such as economic indicators and store characteristics.

Walmart's sales data provides a rich and challenging dataset for forecasting, with its high dimensionality and the influence of external factors like holidays and economic conditions. The dataset used in this study includes store-specific weekly sales data along with features such as store size, fuel prices, CPI, and unemployment rates. These features provide valuable context that can enhance the predictive power of the models.

This research focuses on four regression models: Ridge Regression, Lasso Regression, Decision Tree Regression, and Random Forest Regression. Ridge and Lasso regressions are regularized linear models that can handle multicollinearity and feature selection, respectively. Decision Tree Regression offers a non-linear approach by splitting the data into homogeneous subsets, while Random Forest Regression, an ensemble method, combines multiple decision trees to improve generalization and reduce overfitting.

The primary objective of this study is to compare the performance of these models in terms of their predictive accuracy and robustness. By evaluating metrics such as Root Mean Squared Error (RMSE) and R-squared (R^2) on both training and testing datasets, the study aims to identify the most effective model for sales forecasting in a retail context. The findings will provide valuable insights for retailers looking to implement machine learning techniques for more accurate sales predictions, ultimately aiding in better inventory and resource management.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 12, December 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0912043 |

Through this comparative analysis, the paper aims to contribute to the growing body of literature on machine learning applications in retail sales forecasting, offering practical guidance for industry practitioners and researchers alike. The insights gained from this study will help retailers optimize their forecasting processes, leading to improved operational efficiency and competitive advantage in the marketplace.

II. LITERATURE REVIEW

Previous studies have focused on predicting sales for retail industry corporations using relevant historical data. Researchers from Fiji National University and The University of the South Pacific analyzed the Walmart dataset to predict sales, utilizing tools like Hadoop Distributed File Systems (HDFS), the Hadoop MapReduce framework, and Apache Spark, along with high-level programming environments such as Scala, Java, and Python. Their study aimed to determine whether the factors included in the dataset impacted Walmart's sales ("Walmart's Sales Data Analysis - A Big Data Analytics Perspective," 2017).

In 2015, Harsoor and Patil (Harsoor & Patil, 2015) worked on forecasting Walmart Store sales using big data applications such as Hadoop, MapReduce, and Hive to manage resources efficiently. They used the same sales dataset analyzed in this study but forecasted sales for the upcoming 39 weeks using Holt's winter algorithm. The forecasted sales were visually represented in Tableau using bubble charts.

Michael Crown (Crown, 2016), a data scientist, analyzed a similar dataset but focused on time series forecasting and non-seasonal ARIMA models for his predictions. He used ARIMA modeling to create one year of weekly forecasts from 2.75 years of sales data, considering features like store, department, date, weekly sales, and holiday data. The performance was measured using normalized root-mean-square error (NRMSE).

Forecasting techniques have been applied beyond business enhancement. Researchers have used machine learning and statistical analysis to build predictive models for various applications, such as predicting weather, monitoring stock prices, analyzing market trends, and predicting illnesses in patients.

In 2017, Chouskey and Chauhan (Chouksey & Chauhan, 2017) developed a weather forecasting model that accurately predicts the weather and sends out warnings to help people and businesses prepare for unforeseeable conditions. They used MapReduce and Spark to create their models, gathering data from various weather sensors and using parameters like temperature, humidity, pressure, and wind speed for better predictions.

Rajat Panchotia (Panchotia, 2020) highlighted the importance of defining regression techniques and metrics for creating predictive models using linear regression. He emphasized considering the number of independent variables, the type of dependent variables, and determining the best fit based on the nature of the data. His article discussed the use of regression coefficients, p-values, variable selection, and residual analysis to study regression models' performance. While Panchotia focused on the direct relationship between independent and dependent variables, James Jaccard and Robert Turrisi (Jaccard & Turrisi, 2018) explored how the presence of a third variable, called the moderator variable, affects the relationship between an independent and dependent variable.

Kassambara (Kassambara, 2018) discussed implementing interaction effects with multiple linear regression in R. Using a basic multiple regression model to predict sales based on advertising budgets for YouTube and Facebook, he created an additive model. With an R2 score of 0.98, he found an interactive relationship between the two predictor variables, showing that the additive model outperformed the regular regression model.

This study adopts a similar approach to Kassambara (Kassambara, 2018), examining the interaction effects between multiple independent variables like unemployment, fuel prices, and CPI, and exploring their relationship with weekly sales. Additionally, this study extends predictive techniques by implementing random forest algorithms. Researchers at San Diego State University (Lingjun et al., 2018) highlighted the superiority of this tree-based machine learning algorithm over other regression methods for predictive models in the higher education sector. They used a standard classification and regression tree (CART) algorithm along with feature importance in Weka and R, comparing its efficacy with models like lasso regression and logistic regression.

This review aimed to identify similar practices used by other researchers in creating predictive models influenced by multiple independent variables. The research reviewed shows that many authors use a combination of tools and techniques to create efficient models and compare their findings across models to select the best-performing one.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 12, December 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0912043 |

Similar to Harsoor and Patil, as well as Chouskey and Chauhan, who used Hadoop, MapReduce, and Hive for predictions, this study employs algorithms such as linear and lasso regression, random forest, and multiple regression interaction effects for predictions. Performing a comparative analysis with several models ensures accurate predictions and broad applicability, as models perform differently based on data nature and size.

III. METHODOLOGY

DATA COLLECTION AND PREPROCESSING

DATA SOURCES:

Train Data: Historical sales data of Walmart stores, including store number, department number, date, weekly sales, and whether the week included a holiday.

Features Data: Economic indicators and additional features such as temperature, fuel price, CPI (Consumer Price Index), and unemployment rate.

Stores Data: Information about individual stores, including store number and size.

Test Data: Data used for testing the model's performance.

DATA PREPROCESSING STEPS:

Merging Datasets: The features_df and stores_df datasets were merged using the Store column to create a comprehensive dataset with both store-specific and economic features.

Date Conversion: The Date columns in the datasets were converted to datetime format to facilitate time-based analysis.

Feature Engineering: Extracted additional features such as the week number and year from the Date column.

Handling Missing Values: Ensured there were no missing values in the merged dataset as indicated by the info() method outputs.

EXPLORATORY DATA ANALYSIS (EDA)

Visualization of Key Features:

Department Sales Distribution: A pie chart was created to visualize the sales distribution across the top 10 departments.

Store Sales Distribution: A pie chart was created to visualize the sales distribution across the top 10 stores.

Holiday Impact: A pie chart was created to visualize the proportion of weeks that were holidays versus non-holidays.

These visualizations provided insights into the sales patterns across different departments and stores and the impact of holidays on sales.

MODEL SELECTION AND IMPLEMENTATION

Four regression models were selected for comparison: Ridge Regression, Lasso Regression, Decision Tree Regression, and Random Forest Regression.

1. *Ridge Regression:*
   - *Regularization Technique:* Adds a penalty on the size of the coefficients to prevent overfitting.
   - *Parameter Tuning:* The alpha parameter, controlling the regularization strength, was tuned using GridSearchCV with a range of values from 0.0001 to 100000.
   - *Model Evaluation:* The model's performance was evaluated using Root Mean Squared Error (RMSE) and R-squared (R^2) metrics.
2. *Lasso Regression:*
   - *Regularization Technique:* Adds a penalty on the absolute value of the coefficients, promoting sparsity in the model.
   - *Parameter Tuning:* The alpha parameter was tuned using GridSearchCV with the same range of values as Ridge Regression.
   - *Model Evaluation:* The model's performance was evaluated using RMSE and R-squared (R^2) metrics.
3. *Decision Tree Regression:*
   - *Non-linear Technique*: Splits the data into homogeneous subsets based on feature values.
   - *Parameter Tuning:* The max_depth parameter, controlling the depth of the tree, was tuned using GridSearchCV with values ranging from 3 to 30.
   - *Model Evaluation:* The model's performance was evaluated using RMSE and R-squared (R^2) metrics.
4. *Random Forest Regression:*
   - *Ensemble Technique:* Combines multiple decision trees to improve generalization and reduce overfitting.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 12, December 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0912043 |

- *Parameter Tuning:* The parameters n_estimators (number of trees), min_samples_split (minimum samples required to split an internal node), and min_samples_leaf (minimum samples required to be at a leaf node) were tuned using RandomizedSearchCV.
- *Model Evaluation:* The model's performance was evaluated using RMSE and R-squared (R^2) metrics.

MODEL EVALUATION AND COMPARISON

Train-Test Split: The data was split into training and testing sets with a 70-30 ratio to ensure robust evaluation of the models.

Performance Metrics:

Root Mean Squared Error (RMSE): Measures the average magnitude of the errors. It is calculated as the square root of the average squared differences between the actual and predicted values.

R-squared (R^2): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R^2 value indicates a better fit.

MODEL TRAINING AND PREDICTION:

1. Each model was trained on the training set using the selected hyperparameters.
2. Predictions were made on both the training and testing sets.
3. The performance metrics (RMSE and R^2) were calculated for both sets to assess the models' accuracy and generalization capabilities.

IV. RESULTS

The performance of the models was compared based on the RMSE and R^2 values obtained from both training and testing datasets. The following table summarizes the results:

Model	RMSE (Train)	R^2 (Train)	RMSE (Test)	R^2 (Test)
Ridge Regression	21696.19	0.0844	21811.99	0.0845
Lasso Regression	21975.47	0.0607	22092.15	0.0609
Decision Tree Regression	1998.50	0.9922	4889.26	0.9540
Random Forest Regression	2736.12	0.9854	3881.37	0.9710

Table 1 Comparative analysis of different parameters of ML algorithms

Ridge and Lasso Regression: These linear models exhibited moderate performance with low R^2 values, indicating limited ability to capture the complexity of sales data.

Decision Tree Regression: Showed high accuracy on the training set but a noticeable drop in performance on the test set, suggesting potential overfitting.

Random Forest Regression: Achieved the best performance with high R^2 values on both training and test sets, indicating a robust model with excellent predictive capability.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 12, December 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0912043 |

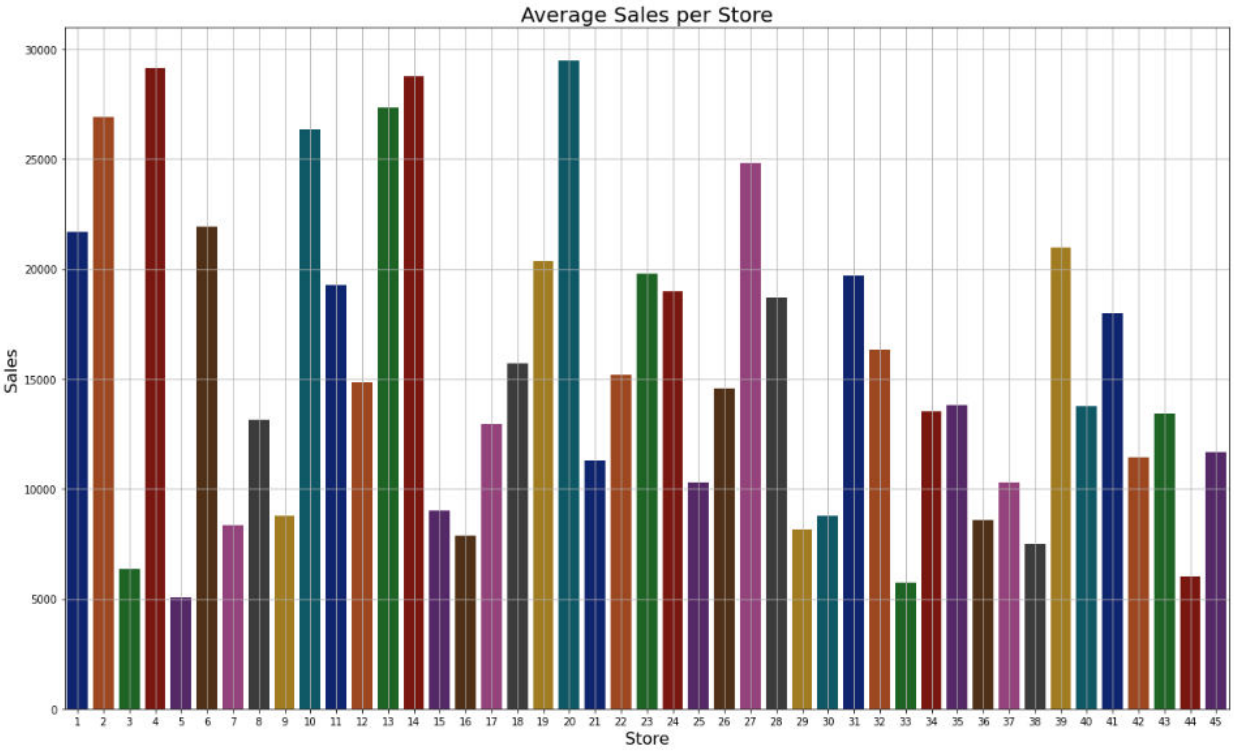

Figure 1 the average weekly sales per store using a bar plot.

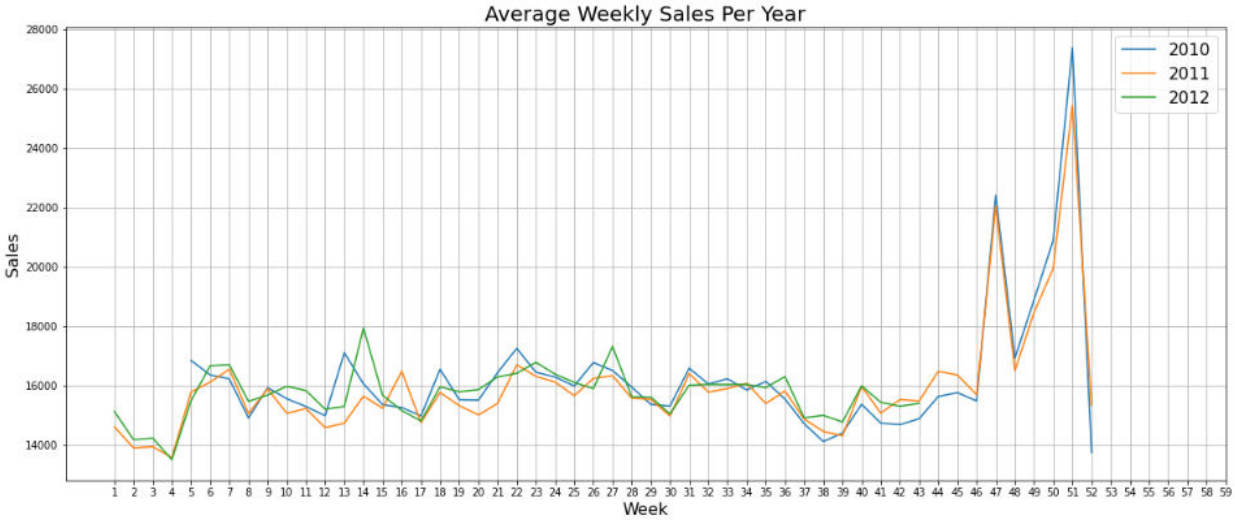

Figure 2 Average weekly sales per year

IV. CONCLUSION

This study demonstrates that while linear models like Ridge and Lasso regression provide a baseline for sales forecasting, advanced techniques such as Decision Tree and Random Forest regressions significantly enhance predictive accuracy. The Random Forest model, in particular, emerged as the most effective for forecasting Walmart sales, offering valuable insights for optimizing inventory and improving strategic planning in the retail industry. Future work could explore additional features and more sophisticated ensemble methods to further improve forecasting accuracy.

By leveraging machine learning techniques, retailers can achieve more accurate sales forecasts, ultimately leading to better inventory management, reduced costs, and improved customer satisfaction.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 12, December 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0912043 |

REFERENCES

[1] Bakshi, C. (2020). Random forest regression. https : / / levelup . gitconnected . com / random-forest-regression-209c0f354c84

[2] Bari, A., Chaouchi, M., & Jung, T. (n.d.). How to utilize linear regressions in predictive analytics. https://www.dummies.com/programming/big-data/data-science/ how-to-utilize-linear-regressions-in-predictive-analytics/

[3] Baum, D. (2011). How higher gas prices affect consumer behavior. https : / / www . sciencedaily.com/releases/2011/05/110512132426.htm

[4] Brownlee, J. (2016). Feature importance and feature selection with xgboost in python. https : / / machinelearningmastery . com / feature - importance - and - feature - selection-with-xgboost-in-python/

[5] Chouksey, P., & Chauhan, A. S. (2017). A review of weather data analytics using big data. International Journal of Advanced Research in Computer and Communication Engineering, 6. https://doi.org/https://ijarcce.com/upload/2017/january17/IJARCCE%2072.pdf

[6] Crown, M. (2016). Weekly sales forecasts using non-seasonal arima models. http : / / mxcrown.com/walmart-sales-forecasting/ Editor,

[7] M. B. (2013). Regression analysis: How do i interpret r-squared and assess the goodness-of-fit? https : / / blog . minitab . com / en / adventures - in - statistics - 2 / regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-offit

[8] Ellis, L. (2019). Simple eda in r with inspectdf. https://www.r-bloggers.com/2019/05/ part-2-simple-eda-in-r-with-inspectdf/ Frost, J. (2021). Regression coefficients- statistics by jim. https://statisticsbyjim.com/ glossary/regression-coefficient/

[9] Glen, S. (2016). Elementary statistics for the rest of us. https://www.statisticshowto. com/correlation-matrix/ 58 Guide, U. B. A. R. P. (n.d.). Gradient boosting machines.

[10] http://uc-r.github.io/gbm_ regression Harsoor, A. S., & Patil, A. (2015). Forecast of sales of walmart store using big data applications. International Journal of Research in Engineering and Technology eIS, 04, 51–59. https : / / doi . org / https : / / ijret . org / volumes / 2015v04 / i06 / IJRET20150406008.pdf

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details