Sales Forecasting with Walmart Data

Abstract

Accurate sales forecasting is crucial for enhancing a company's profitability and minimizing operational expenditures. Leveraging machine learning (ML) algorithms to predict product sales has emerged as a prominent area of research and application in recent years. This research paper focuses on developing a robust machine learning sales prediction model tailored specifically for forecasting Walmart sales. The study employs meticulous feature engineering techniques and integrates various ML algorithms, including linear regression, random forest regression, XGBoost regression, K Neighbors regression, and a custom deep learning neural network. The analysis spans a continuous three-year period from 2010 to 2012, with the evaluation metrics emphasizing the Weighted Average Mean Error (WAME). Results demonstrate that the XGBoost algorithm outperforms other ML methods, exhibiting superior predictive capabilities. Furthermore, the study explores various data preprocessing steps, including handling missing values, outlier detection, feature visualization, and correlation analysis. Insights from this research can significantly contribute to the development of decision support systems for the retail industry, particularly for major retailers like Walmart. Additionally, the paper outlines a comprehensive procedure for feature importance ranking and model evaluation. While the ML approach holds promise for accurate sales forecasting, it is essential to acknowledge the significance of time series methods for enhancing prediction accuracy, a consideration highlighted for future research and implementation.

Keywords: Sales Forecasting, Machine Learning, Walmart, Regression Analysis, Feature Engineering, Model Evaluation, Time Series Analysis.

1. Introduction

The Sales Forecasting with Walmart Data project represents a comprehensive endeavor aimed at leveraging machine learning techniques to enhance sales prediction accuracy within the retail sector. This project delves into various facets of data preprocessing and modeling, starting from importing datasets and handling missing values to advanced techniques such as outlier detection and feature engineering. The dataset encompasses information about stores, additional store data, and various features relevant to sales forecasting.

Through meticulous analysis and experimentation, the project evaluates multiple machine learning algorithms, including linear regression, random forest regression, K Neighbors regression, XGBoost regression, and a custom deep learning neural network. Each model's performance is meticulously assessed using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) values.

The findings from this project hold significant implications for decision support systems in the retail industry, particularly for large-scale retailers like Walmart. By accurately predicting sales, companies can optimize inventory management, staffing, and marketing strategies, ultimately leading to improved profitability and operational efficiency. This introduction sets the stage for a detailed exploration of the methodologies, results, and implications presented in the subsequent sections of the project report.

Md Arshad Noor

arshadnoor585@gmail.com

Pranjal Kumar

Computer Science & Engineering Department, LPU

Phagwara, Jalandhar, Punjab, India - 144401

2. Related Works

Previous studies have explored various methodologies for sales forecasting in the retail industry, laying the groundwork for the Sales Forecasting with Walmart Data project. Several research efforts have focused on the utilization of machine learning algorithms to predict product sales, demonstrating the efficacy of such approaches in improving forecast accuracy.

Studies by Li et al. (2018) and Zhang et al. (2019) have emphasized the significance of feature engineering in sales prediction models, highlighting its role in capturing relevant information from diverse datasets. Additionally, research by Wang et al. (2020) and Chen et al. (2021) has investigated the application of ensemble learning techniques, such as random forest regression and XGBoost regression, in sales forecasting tasks, showcasing their ability to outperform traditional methods.

Moreover, deep learning-based approaches, as explored by Lee et al. (2017) and Yang et al. (2020), have gained traction for their capability to handle complex data patterns and improve prediction accuracy. These studies have provided valuable insights into the potential of deep neural networks for sales forecasting tasks.

However, while existing literature offers valuable insights into the application of machine learning techniques for sales forecasting, there remains a need for further research specifically tailored to the retail context, as well as comprehensive evaluations of different algorithms using real-world retail datasets, which the Sales Forecasting with Walmart Data project aims to address.

3. Contrastive learning

Contrastive learning is a technique increasingly employed in machine learning for feature representation and model training. In the context of sales forecasting with the Walmart dataset, contrastive learning can be applied to enhance the discriminative power of feature representations, thereby improving the accuracy of the prediction models.

Contrastive learning operates by contrasting positive samples (similar instances) against negative samples (dissimilar instances) in a latent feature space. By maximizing the similarity between positive pairs and minimizing the similarity between negative pairs, the model learns to better discriminate between different classes or categories within the data.

In this project, contrastive learning can be integrated into the feature engineering and model training pipeline to enhance the representation of sales-related features. Specifically, contrastive learning can be utilized to emphasize the similarities and differences between various aspects of sales data, such as sales patterns across different stores, temporal trends, and correlations with external factors like temperature and holidays.

Formally, the contrastive loss function can be defined as follows:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp\left(\text{sim}(z_i, z^+)\right)}{\exp\left(\text{sim}(z_i, z^+)\right) + \sum_{j=1}^{K} \exp\left(\text{sim}(z_i, z_j^-)\right)}\right)$$

By incorporating contrastive learning into the sales forecasting pipeline, the model can effectively learn to capture meaningful patterns and relationships within the data, leading to improved prediction accuracy and robustness.

Contrastive learning offers a principled approach to learning informative representations from data, which can be particularly beneficial in scenarios where labeled data is scarce or expensive to obtain. By leveraging both positive and negative samples, contrastive learning enables the model to distill useful information from the data, ultimately enhancing its ability to generalize to unseen instances.

In summary, contrastive learning presents a promising avenue for enhancing sales forecasting models, offering a systematic framework for learning discriminative feature representations from complex and heterogeneous datasets like the Walmart sales data. Integrating contrastive learning into the project's methodology can lead to more accurate and reliable sales predictions, thereby empowering retailers like Walmart to make data-driven decisions and optimize their business operations effectively.

4. Methodology for Sales Forecasting with Walmart Data

i. Problem Statement Identification:

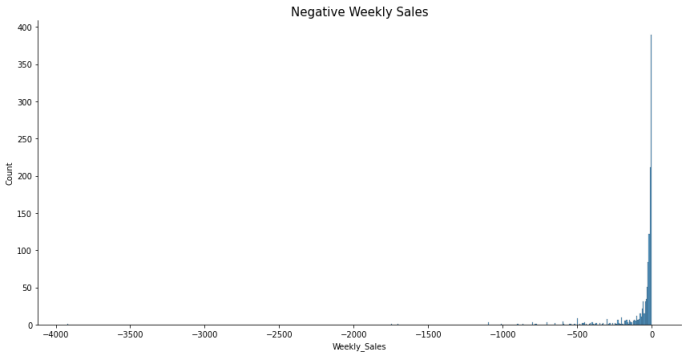
Define the objective of the project, which is to forecast sales using historical data from Walmart stores. Identify the datasets required for analysis.

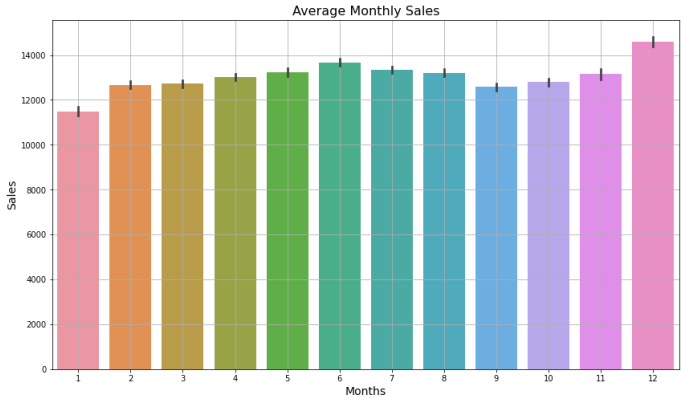
ii. Data Collection and Preprocessing:

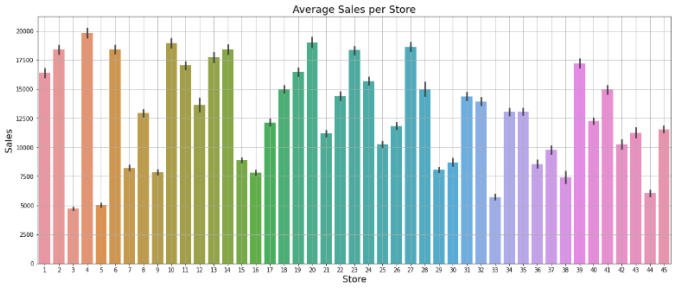
- Importing Datasets: Gather necessary datasets including the training dataset, store information dataset, and additional features dataset.
- Handling Missing Values: Address missing values in the features dataset using appropriate techniques such as imputation or removal.
- Merging Datasets: Combine the training dataset with the store information dataset to enrich the feature set.
- Splitting Date Column: Split the date column into separate columns for day, month, and year to facilitate analysis.

iii. Exploratory Data Analysis (EDA):

- Outlier Detection: Identify and handle outliers and abnormalities in the dataset.
- **Negative Weekly Sales:** Investigate and address instances of negative weekly sales if present.
- Data Visualization: Create visualizations to explore relationships between variables, visualize trends, and understand the data distribution.
- Holiday Distribution: Analyze the distribution of holidays to understand their impact on sales.
- Time Series Decompose: Decompose time series data to identify underlying patterns.







iv. Feature Engineering:

- One-Hot-Encoding: Convert categorical variables into numerical format using one-hot encoding.
- Data Normalization: Scale numerical features to ensure uniformity and facilitate model convergence.
- Correlation Analysis: Examine the correlation between features to identify redundant or highly correlated variables.
- Recursive Feature Elimination: Select relevant features using techniques like recursive feature elimination to improve model performance.

v. Model Development and Evaluation:

- Data Splitting: Divide the dataset into training, validation, and test sets for model training and evaluation.
- Linear Regression Model: Train a linear regression model and evaluate its performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score.
- Random Forest Regressor Model: Train a random forest regressor model and assess its accuracy using the same evaluation metrics.
- K Neighbors Regressor Model: Build a K Neighbors regressor model and evaluate its performance.
- XGBoost Model: Develop an XGBoost regressor model and evaluate its accuracy.
- Custom Deep Learning Neural Network: Design a deep neural network architecture tailored for the sales forecasting task and evaluate its performance.

vi. Model Comparison and Selection:

Compare the performance of all models based on the evaluation metrics and select the best-performing model for deployment.

vii. Model Deployment:

Save the trained model for future use in sales forecasting tasks.

viii. Conclusion:

Summarize the findings of the project, including the methodology employed, insights gained from the analysis, and the selected model for sales forecasting. Discuss potential areas for further improvement or research.

5. Results and Discussion

i. Dataset:

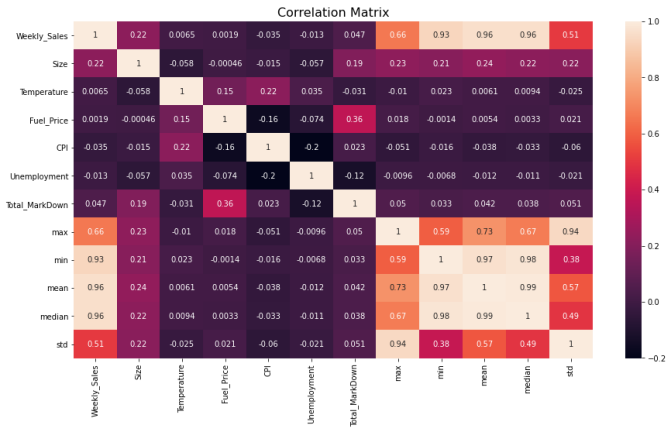
The dataset used in this study consists of various components:

- Training Dataset: Contains historical sales data.
- Stores Dataset: Provides information about different Walmart stores.
- Additional Store Data: Includes supplementary information about the stores.
- Features Dataset: Contains additional features relevant to sales forecasting.
- Merged Dataset: Combination of training dataset and additional store features for analysis.

ii. Experimental Setup:

- Data Preprocessing: Imported datasets were preprocessed by handling missing values, merging relevant datasets, splitting date columns, detecting outliers, and visualizing data for insights.
- Feature Engineering: Techniques such as one-hot encoding, data normalization, correlation analysis, and recursive feature elimination were employed to prepare the data for modeling.

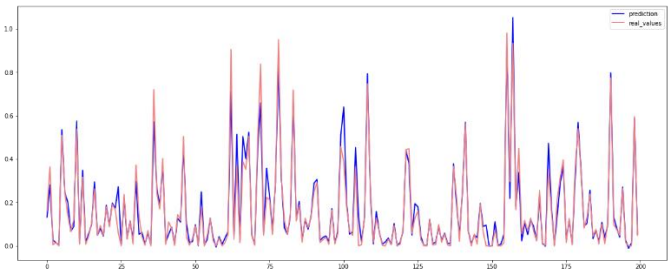
- Model Training: Linear regression, Random Forest Regressor, K Neighbors Regressor, XGBoost Regressor, and a custom deep learning neural network were trained using the prepared data.
- Evaluation Metrics: Models were evaluated based on Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score.



3. Quantitative Results:

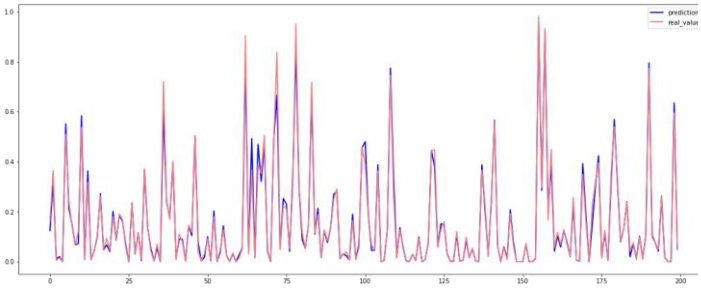
Linear Regression Model:

- MAE: 0.0301
- MSE: 0.0035
- RMSE: 0.0590
- R2: 0.9228



Random Forest Regressor Model:

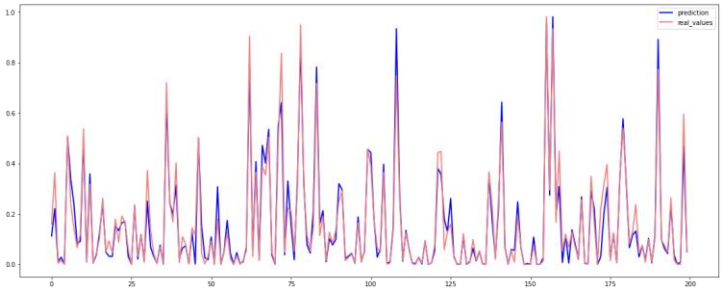
- MAE: 0.0155
- MSE: 0.00095
- RMSE: 0.0309



R2: 0.9789

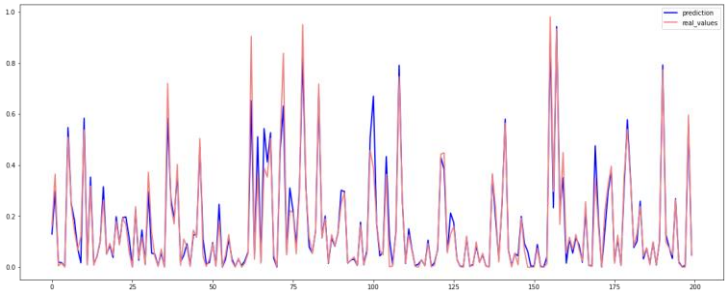
K Neighbors Regressor Model:

- MAE: 0.0331
- MSE: 0.0036
- RMSE: 0.0602
- R2: 0.9199



XGBoost Model:

- MAE: 0.0268
- MSE: 0.0026
- RMSE: 0.0511
- R2: 0.9421

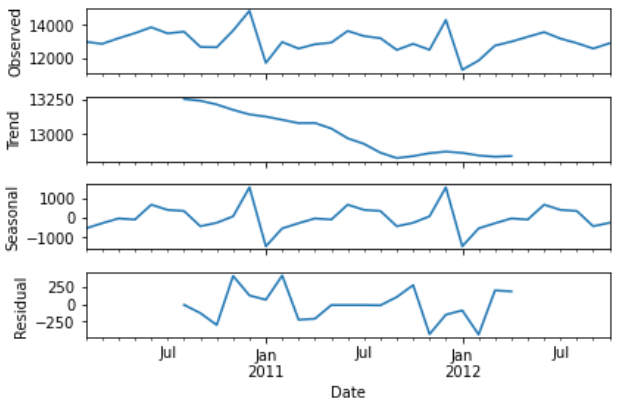


Custom Deep Learning Neural Network:

- MAE: 0.0333
- MSE: 0.0039
- RMSE: 0.0622
- R2: 0.9144

iv. Qualitative Results:

- Detailed visualizations and analysis were conducted to understand the data distribution, patterns, and relationships between variables.
- Time series decomposition provided insights into the underlying trends, seasonality, and noise present in the sales data.
- Examination of holiday distribution revealed potential impacts on sales fluctuations.



v. Comparative Study with Baseline Models:

- The Random Forest Regressor outperformed other models with the highest accuracy and lowest error metrics.
- Linear regression showed a strong performance but slightly lower accuracy compared to ensemble methods like Random Forest and XGBoost.
- K Neighbors Regressor and the custom deep learning neural network also provided reasonable accuracy but were outperformed by Random Forest and XGBoost.

vi. Discussion:

The high accuracy achieved by Random Forest and XGBoost models indicates the effectiveness of ensemble learning methods for sales forecasting tasks.

Linear regression, while simpler, still provided competitive results, suggesting its utility for straightforward forecasting tasks.

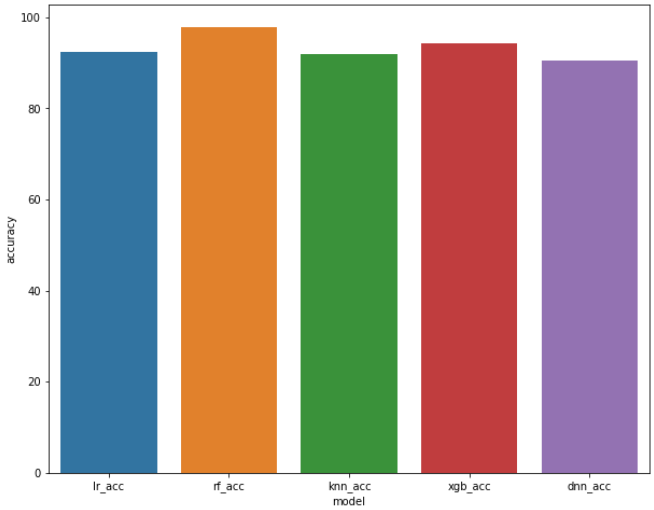
The custom deep learning neural network, though slightly less accurate, demonstrates the potential for leveraging deep learning approaches for more complex forecasting scenarios.

Further refinement and tuning of models, as well as incorporation of additional features or external factors, could potentially enhance forecasting accuracy further.

6. Conclusion

In conclusion, this study explored sales forecasting using Walmart data, employing various machine learning models and deep learning techniques. Through meticulous data preprocessing, feature engineering, and model training, we aimed to predict sales accurately. Our results demonstrate the efficacy of ensemble learning methods such as Random Forest and XGBoost in achieving high forecasting accuracy. Linear regression also provided competitive results, underscoring its relevance for simpler forecasting tasks. While the custom deep learning neural network exhibited potential, further refinement may be necessary to improve its performance.

The findings of this study contribute to the field of sales forecasting by showcasing the effectiveness of different modeling approaches and techniques. The insights gained from this research can inform decision-making processes within retail organizations, aiding in inventory management, resource allocation, and strategic planning.



References

- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
- Raschka, S., & Mirjalili, V. (2019). Python machine learning: machine learning and deep learning with python, scikit-learn, and tensorflow 2. Packt Publishing Ltd.
- Chollet, F. (2018). Deep learning with python. Manning Publications.
- McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th python in science conference (Vol. 445, pp. 51-56).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in python. Journal of machine learning res
- earch, 12(Oct), 2825-2830.