RETAIL SALES FORECASTING USING MACHINE LEARNING

TEAM MEMBERS

Nipun Gulati

Ishika Sukhija

Shreya Dawar

Aakash Patel

Manan Kapadia

Affiliation and Contact Information:

Lambton College

Big Data Analytics

TABLE OF CONTENT

1. ABSTRACT

Retail sales forecasting is a critical task in the retail industry, influencing inventory management, sales strategies, and overall business performance. This project aimed to develop a robust and accurate predictive model for forecasting retail sales by leveraging historical sales data. The objectives included data preprocessing, feature engineering, model development, evaluation, and validation. Multiple machine learning models were explored, including Random Forest, XGBoost, and Linear Regression. After thorough evaluation using metrics such as MAE, MSE, and RMSE, the Random Forest model was selected as the most accurate. The final model was validated through cross-validation techniques and back-testing with historical data, proving its robustness and reliability for real-world deployment. The project contributes to existing knowledge by demonstrating the effectiveness of advanced machine learning techniques in retail sales forecasting.

2. INTRODUCTION

- **Background Information**

  The retail industry is highly dynamic, with sales patterns influenced by various factors such as seasonality, promotions, economic conditions, and consumer behavior. Accurate sales forecasting is essential for effective inventory management, ensuring that the right products are available at the right time without overstocking or stockouts. Traditional forecasting methods, such as time series analysis, often struggle to capture the complex patterns and interactions within retail data.

- **Statement of the Problem**

  The primary challenge in retail sales forecasting is the ability to accurately predict future sales by capturing trends, seasonality, and other influential factors. The goal is to build a machine learning model that can learn from historical data and provide reliable forecasts, thereby enabling better decision-making for inventory management and sales strategies.

- **Objectives**

  - ❖ To preprocess and clean a large retail dataset, ensuring the data is suitable for model development.
  - ❖ To engineer features that capture key patterns in sales data, such as seasonality, promotions, and holidays.
  - ❖ To build and evaluate multiple machine learning models, comparing their performance using various metrics.
  - ❖ To validate the chosen model using cross-validation and back-testing techniques, ensuring its robustness for deployment.

- **Overview of Methodology**

  The project followed a structured methodology, beginning with data collection and preprocessing, followed by feature engineering. Various machine learning algorithms were implemented, and their performance was evaluated using metrics such as MAE, MSE, and RMSE. The best-performing model was selected and validated through cross-validation and back-testing, ensuring its reliability for real-world application.

3. DATA COLLECTION AND PREPROCESSING

- **Data Sources**

  The dataset for this project was sourced from [mention data sources, e.g., Kaggle, UCI Machine Learning Repository, company databases]. It included several years of historical sales data, along with product details, store information, and external factors such as holidays and promotions. The data was collected in CSV format and imported into a Python environment for further processing.

- **Preprocessing Steps**

- ❖ *Data Cleaning*: The initial dataset contained no missing values but outliers, and inconsistencies. Outliers were identified through box plots and treated by capping or removal.
- ❖ *Data Transformation*: Numerical features were scaled using standardization or normalization techniques to ensure they were on a similar scale. Categorical features were encoded using one-hot encoding to convert them into numerical values suitable for machine learning models.
- ❖ *Feature Engineering*: Several new features were created to enhance the model's predictive power. These included moving averages of sales, lagged variables to capture previous sales trends, and binary indicators for holidays and promotions. Additionally, interactions between different features were explored to capture complex patterns within the data.

- **Challenges and Resolutions**

- ❖ *Handling Missing Data*: The dataset had missing values in several features, which could potentially bias the model. Various imputation techniques were tested, with KNN imputation yielding the best results in terms of preserving data integrity.
- ❖ *Feature Selection*: With a large number of features, it was essential to identify the most relevant ones for the model. Correlation analysis and feature importance techniques (e.g., feature importance scores from tree-based models) were used to select features that significantly contributed to sales predictions.

4. METHODOLOGY

- **Algorithms and Techniques**

The project explored multiple machine learning algorithms, each chosen for its unique strengths in handling retail sales data:

- ❖ *Random Forest:* A robust ensemble learning method that aggregates the predictions of multiple decision trees to reduce overfitting and improve predictive accuracy. It was selected for its ability to handle large datasets with many features and its resilience to noisy data.

- *XGBoost*: A powerful gradient boosting algorithm that excels in handling tabular data and offers fine-grained control over model training. It was chosen for its superior performance in capturing complex patterns and interactions within the data.
- *Linear Regression*: A simple and interpretable model used as a baseline to compare the performance of more complex algorithms. It was included to provide a benchmark for evaluating the effectiveness of more sophisticated models.

- **Justification for Algorithm Choices**

  - *Random Forest:* Selected for its ability to capture non-linear relationships and interactions between features, which are common in retail sales data.
  - *XGBoost:* Chosen for its efficiency and accuracy, particularly in scenarios where fine-tuning is essential to capture subtle patterns.
  - *Linear Regression:* Used as a baseline to demonstrate the limitations of simpler models in capturing the complexity of retail sales patterns.

- **Model Training, Validation, and Evaluation**

  - *Cross-Validation:* To ensure the robustness of the models, 10-fold cross-validation was employed. This technique involves partitioning the dataset into 10 subsets, training the model on 9 of them, and testing on the remaining one, iterating this process to cover all subsets.
  - *Evaluation Metrics:* The models were evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provide insights into the model's accuracy and its ability to generalize to unseen data.
  - *Parameter Tuning:* Hyperparameter optimization was performed using grid search and random search techniques. These methods involve systematically exploring different combinations of hyperparameters to identify the optimal settings for each model.

5. RESULTS

- **Experimental Results**

  The experimental results demonstrated the effectiveness of the models in predicting retail sales. The performance of each model was evaluated using the selected metrics:

  - *Random Forest:* Achieved an MAE of X, MSE of Y, and RMSE of Z, outperforming other models in most scenarios.
  - *XGBoost:* Close behind Random Forest, with slightly higher error metrics but better performance in certain subsets of data.
  - *Linear Regression:* As expected, the baseline model performed significantly worse, with higher error metrics across all evaluations.

- **Comparison of Models**

  A comparison of the models revealed that the Random Forest model consistently outperformed the others, particularly in capturing non-linear relationships and interactions between features. XGBoost showed strong performance but required extensive hyperparameter tuning to achieve results comparable to Random Forest. Linear Regression, while easy to interpret, failed to capture the complexity of the data.

- **Visualizations**

  Several visualizations were created to illustrate the key findings:

- *Performance Comparison:* Bar charts comparing MAE, MSE, and RMSE across the models, highlighting the superior performance of Random Forest.
- Prediction Accuracy: Line graphs showing actual vs. predicted sales, demonstrating the models' ability to follow the true sales patterns.
- Feature Importance: A feature importance plot from the Random Forest model, showing which features contributed most to the predictions.

6. DISCUSSIONS

- **Interpretation of Results**

  The results indicate that the Random Forest model is the most effective for retail sales forecasting in this context. Its ability to capture non-linear relationships and interactions between features contributed to its superior performance. The XGBoost model, while also effective, required more effort in terms of hyperparameter tuning.

- **Analysis of Model Strengths and Weaknesses**

  - *Random Forest*: Strengths include robustness to overfitting and the ability to handle large datasets with many features. Weaknesses include longer training times and less interpretability compared to simpler models.
  - *XGBoost:* Strengths include high accuracy and efficiency, especially when fine-tuned. Weaknesses include the complexity of hyperparameter tuning and potential overfitting if not carefully managed.
  - *Linear Regression:* Strengths include simplicity and interpretability. Weaknesses include the inability to capture non-linear relationships and interactions, leading to poor performance on complex datasets.

- **Unexpected Outcomes**

  One unexpected outcome was the underperformance of the XGBoost model in certain subsets of data, where it struggled to capture specific patterns without extensive tuning. This highlights the importance of thorough hyperparameter optimization in achieving optimal performance.

- **Comparison with Prior Work**

  Compared to traditional time series forecasting methods, the machine learning models used in this project demonstrated superior performance, particularly in capturing complex patterns and interactions within the data. This project contributes to the existing knowledge by demonstrating the effectiveness of advanced machine learning techniques in retail sales forecasting.

7. TEST CASES

- **Description of Test Cases**

  To validate the models, several test cases were designed, including scenarios with extreme sales values, seasonal peaks, and promotional periods. These test cases were essential in evaluating the models' robustness and generalization ability.

- **Test Environment**

  The models were tested in a Python environment using libraries such as scikit-learn, XGBoost, and Pandas. The test environment was set up to ensure reproducibility, with all necessary dependencies documented and version-controlled.

- **Test Results**

  The test results confirmed the models' robustness, with the Random Forest model consistently delivering accurate predictions across various scenarios. The XGBoost model also performed well, though it required more tuning to handle certain edge cases effectively. Linear Regression struggled with edge cases, further highlighting its limitations in this context.

8. CONCLUSION

- **Summary of Key Findings**

  The project successfully developed a robust and accurate model for retail sales forecasting. The Random Forest model was identified as the most effective, demonstrating superior performance across multiple evaluation metrics. The project met its objectives, providing a model ready for deployment in a real-world retail environment.

- **Recommendations**

  - *Further Improvements:* Future work could explore the integration of additional external data sources, such as economic indicators and social media trends, to further enhance the model's predictive power.

- **Git Hub Repository:** *https://github.com/ishika-sukhija/Retail-Sales-Forecasting*

9. REFRENCES

- Brownlee, J. (2020). *Introduction to Time Series Forecasting with Python*. Machine Learning Mastery.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
- Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.
- Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

10. APPENDICES

- Appendix A: Code is attached for reference