Title: Advanced Model Evaluation and Selection for Retail Data
Name: Satyaprakash challagulla
Date: 02/03/2024

Abstract

This project aims to identify the most effective predictive model for forecasting total sales from transactional retail data. We evaluated three models: Linear Regression, Random Forest, and Gradient Boosting, using RMSE as the performance metric. The Random Forest model showed superior performance and was recommended for operational deployment.

Introduction

The objective of this project is to develop a predictive model that accurately forecasts total sales based on transactional retail data. Accurate sales forecasting helps in better inventory management and customer satisfaction.

Data Description

The dataset includes transaction records from a retail business, comprising features such as product descriptions, stock codes, invoice details, and total price. The data was preprocessed to handle categorical and numerical variables appropriately, setting the stage for reliable model training.

Methodology

We implemented three predictive models:

- **Linear Regression:** A baseline model for comparison.
- **Random Forest:** An ensemble method known for handling non-linear data effectively.
- **Gradient Boosting:** A robust method that sequentially corrects errors of the prior models.

Each model was trained on the preprocessed dataset, with RMSE as the key metric to evaluate model performance.

Results

The performance of the models was as follows:

- **Linear Regression:** RMSE = 175.83
- **Random Forest:** RMSE = 90.04 (Best Performance)
- **Gradient Boosting:** RMSE = 280.32

Visualizations were created to compare the predictions versus actual outcomes, highlighting the superior accuracy of the Random Forest model.

Discussion

The Random Forest model outperformed other models, suggesting its effectiveness in capturing complex patterns in the data without overfitting. This superiority could be attributed to its ability to manage diverse data types and its robustness against the noisy retail environment.

Recommendations

1. **Deploy the Random Forest Model:** Use it for real-time sales forecasting to enhance inventory decisions.
2. **Continuous Model Training:** Regularly update the model with new transaction data to maintain its accuracy.
3. **Feature Engineering:** Experiment with additional features, such as time-based variables, to potentially improve model performance.

Conclusion

The project demonstrates the effectiveness of using advanced machine learning techniques for sales forecasting in retail. The Random Forest model, with the lowest RMSE, is recommended for further development and operational use.