Sales Forecasting

Introduction

The objective of this project was to forecast retail sales using a combination of machine learning (Random Forest and XGBoost) and deep learning (LSTM) models. We enriched the data with features like calendar-based flags, lag features, rolling statistics, and external factors (such as oil prices and holidays). After training the models, we evaluated them using key metrics: RMSE, MAPE, and R^2 score, which provide insights into prediction accuracy, error, and model fit.

Model Comparison Results

The following table summarizes the performance of the models evaluated on the validation dataset.

Model	RMSE	MAPE	R^2 Score
Naive Forecast	933.30	643,236,750,449,569,792.00%	0.727
Random Forest	629.44	262,192,615,915,253,094.00%	0.876
XGBoost	600.76	181,587,149,708,048,076.00%	0.887
LSTM	118.10	10,958,593,713,181,294,592.00%	0.346

Key Observations

1. **Naive Forecast (7-Day Lag)**:
    - **RMSE**: 933.30
    - **MAPE**: Extremely high (indicating large percentage errors), likely due to small actual values or zeros in the validation set.
    - **R^2 Score**: 0.727
    - **Insights**: Although the Naive forecast provides a decent baseline model (with an R^2 score of 0.727), it is unable to capture complex patterns in the data. The high MAPE suggests that the model struggles to adapt to outliers and extreme sales values.
2. **Random Forest Regressor**:
    - **RMSE**: 629.44
    - **MAPE**: 262,192,615,915,253,0944.00% (relatively lower than Naive forecast).
    - **R^2 Score**: 0.876

- **Insights**: Random Forest performed well overall, with an R^2 score of 0.876, indicating that it effectively captured the non-linear patterns and interactions within the data. The model provides a solid alternative when interpretability and training speed are more important. However, the MAPE is still relatively high, and it may not be the best at capturing outliers.

3. **XGBoost**:
   - **RMSE**: 600.76
   - **MAPE**: 181,587,149,708,048,076.00% (lower than Random Forest).
   - **R^2 Score**: 0.887
   - **Insights**: XGBoost emerged as the best model overall, providing the highest R^2 score (0.887) and lowest RMSE. This indicates that it was able to model the non-linear interactions in the data and handle external factors (like oil prices and holidays) more effectively. It is particularly well-suited for structured tabular data.
4. **LSTM (Long Short-Term Memory)**:
   - **RMSE**: 118.10
   - **MAPE**: Extremely high (due to issues with very small values).
   - **R^2 Score**: 0.346
   - **Insights**: While the LSTM model achieved an impressively low RMSE, its R^2 score was significantly lower than the tree-based models, suggesting that the model struggled to capture the long-term dependencies in the data effectively. Additionally, the very high MAPE points to problems with model overfitting and poor prediction of small sales values.

Model Performance Evaluation

1. RMSE (Root Mean Squared Error)

- **Best Model**: LSTM (118.10)
- **Worst Model**: Naive Forecast (933.30)

RMSE provides a quantitative measure of how far the predicted values are from the actual sales. The LSTM model performed exceptionally well in terms of RMSE, indicating that its predictions are closer to the actual sales, especially in terms of numerical error.

2. MAPE (Mean Absolute Percentage Error)

- **Best Model**: XGBoost (181,587,149,708,048,076)
- **Worst Model**: Naive Forecast (643,236,750,449,569,792.00%)

MAPE measures the percentage error in predictions. Although LSTM had a low RMSE, its MAPE was extremely high, suggesting that the model failed to predict very small sales values accurately. XGBoost showed lower MAPE, indicating its ability to better predict sales values across the board.

3. R^2 Score (Coefficient of Determination)

- **Best Model**: XGBoost (0.887)
- **Worst Model**: LSTM (0.346)

The R^2 score reveals how well the model fits the data. XGBoost achieved the highest R^2 score, meaning it captured the underlying patterns in the data more effectively than the other models. LSTM, despite its low RMSE, had a much lower R^2 score, indicating that it was not as effective at capturing the variance in the data.

Business Insights and Recommendations

1. External Factors Influence on Sales

- **Oil Prices**: The fluctuation in oil prices has a significant impact on consumer behavior. For instance, a sudden drop in oil prices can increase consumer spending, indirectly boosting sales. Models like XGBoost, which can capture this external influence, performed better as they accounted for oil price changes.
- **Holidays**: Sales during holidays often differ from regular days. The presence of holidays was used as a feature in all models, and XGBoost and Random Forest managed to capture the patterns related to sales spikes or dips due to holidays more effectively.
- **Event-Based Influences**: The project also analyzed events like earthquakes, which can affect sales significantly. These special events should be included as features in any model to improve forecasting accuracy.

2. Inventory Management

Using the top-performing models (XGBoost or Random Forest), businesses can forecast demand more accurately. This leads to optimized inventory management, ensuring that stockouts or overstocks are minimized, reducing costs, and improving customer satisfaction.

3. Targeted Promotions

Sales forecasts can also help design targeted promotions around key dates such as holidays, when sales are expected to spike. XGBoost's ability to predict sales fluctuations during holidays allows businesses to plan promotions effectively.

4. Dynamic Pricing

With accurate sales predictions, businesses can adopt dynamic pricing strategies. For instance, when oil prices drop, consumer purchasing power may increase, justifying price increases in certain product categories. Predicting demand accurately using models like XGBoost can assist businesses in adjusting prices based on expected demand.

5. Localized Strategies

Using store-specific and cluster-based features in the model, businesses can tailor strategies for different regions or store types. Understanding the local sales trends allows businesses to craft region-specific marketing campaigns and stocking strategies, ensuring that the right products are available at the right time.

Conclusion

The XGBoost model outperformed all other models in terms of accuracy, making it the recommended model for deployment. It provides a robust framework for capturing complex patterns and interactions in retail sales data, including external factors like holidays and oil prices.

- **Recommendation**: **XGBoost** is the final model for deployment, as it consistently provides the best performance across RMSE, MAPE, and R^2 scores.
- **Future Enhancements**: The LSTM model, although it showed potential in terms of RMSE, could benefit from better sequence handling (e.g., more time steps) and improved feature engineering. Moreover, tuning hyperparameters and investigating better scaling strategies for time-series data might help boost its performance.

In summary, retail businesses can leverage these forecasting models to optimize inventory management, plan promotions, and implement dynamic pricing strategies, ultimately leading to better business decision-making and improved profitability.