EXPERIMENT REPORT

Student Name	Sahil Kotak
Project Name	AT2 - Machine Learning as a Service
Date	10/10/2023
Deliverables	kotak_sahil_24707592_predictive_xgboost.ipynb XGBoost

1. EXPERIMENT BACKGROUND	
Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.	
1.a. Business Objective	The primary goal of this project was to predict daily sales for specific items in specific stores. By having accurate predictions, the business can optimize inventory, improve sales strategies, and enhance customer satisfaction by ensuring stock availability. Accurate predictions can lead to reduced holding costs, improved turnover rates, and maximized profits. Conversely, incorrect results can result in overstock or stockouts, leading to potential loss of sales, increased costs, and dissatisfied customers.
1.b. Hypothesis	We hypothesized that past sales data, combined with other external factors such as events, can be used to predict future sales of items in stores. Given the cyclical nature of retail sales and the influence of events and promotions on consumer buying behavior, it's worthwhile to consider these factors when predicting sales.
1.c. Experiment Objective	The objective was to build a machine learning model that can predict sales with a high degree of accuracy. We aimed to achieve a model with minimized error metrics (like RMSE) to ensure the predictions are close to the actual sales. Scenarios resulting from this experiment include potential adjustments in stock levels, promotional strategies, and supply chain optimizations based on the predicted sales.

2. EXPERIMENT DETAILS

Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them.

2.a. Data Preparation

The data underwent several preprocessing steps:

- Handling missing values: Filled or dropped based on their significance.
- Conversion of data types: Date columns were converted to datetime format.
- Filtering: Only relevant columns and rows (pertaining to the stores and items of interest) were retained.

2.b. Feature Engineering

Several features were engineered to improve model performance:

- Lagged Features: Sales data from past days (like sales 7 days ago, 28 days ago) were added as features to capture the temporal patterns.
- Rolling Averages: To capture trends, rolling averages of sales were calculated over different window sizes (like 7 days, 28 days).
- Event Encoding: Events were encoded to numerical values to capture their impact on sales.
- Time-based Features: Features like day of the week, month, and other time-based patterns were included to capture seasonality.

2.c. Modelling

- Model Choice: We chose the XGBoost model because of its capability to handle large datasets efficiently and its ability to capture non-linear patterns. XGBoost also provides feature importance, allowing us to understand which features significantly impact predictions.
- Hyperparameter Tuning: Parameters like learning rate, max depth of the trees, number of trees, and regularization were tuned to optimize the model's performance. Grid search or random search methodologies could be used for this purpose.
- Model Evaluation: The model was evaluated using metrics like RMSE on a validation set to understand its predictive accuracy. Other models like linear regression or random forests could be considered in future experiments based on the performance of the XGBoost model.

The model was then used to make sales predictions on unseen data, and the results were evaluated against actual sales to measure the model's performance.

3. EXPERIMENT RESULTS

Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.

3.a. Technical Performance	The XGBoost model achieved a satisfactory **RMSE score of 6.6623** on the validation set, indicating a decent fit to the data. However, there were certain days, especially around special events or promotions, where the model seemed to underpredict or overpredict sales. This could be due to the model not entirely capturing the influence of these external factors or due to variability in the data that wasn't captured during training.
3.b. Business Impact	The predictions provided by the model can lead to better inventory management, reducing the chances of stockouts or overstock. Inaccurate predictions, especially around high-sales days or promotional events, could lead to missed sales opportunities or increased holding costs. For instance, underpredicting sales during a big promotion could mean missed revenue, while overpredicting could lead to excess stock that needs to be cleared later, possibly at discounted prices.
3.c. Encountered Issues	● **Data Quality**: There were missing values in the dataset, and some data inconsistencies were noticed. ● **Feature Engineering**: Initial models without the engineered features did not perform as well, indicating the importance of domain knowledge in feature creation. ● **Computational Constraints**: XGBoost, especially with hyperparameter tuning, can be computationally intensive. ● **Deployment Challenges**: Deploying the model to production posed some challenges, especially due to the complex nature of the model and the extensive feature engineering involved. Initial deployment did not yield as promising results as expected, indicating the need for further optimization and potential simplification for a production environment.

---

4. FUTURE EXPERIMENT

Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective.

4.a. Key Learning	● The importance of domain knowledge in feature engineering was evident. Features like lagged sales and rolling averages significantly improved model performance. ● External factors, like events, have a notable influence on sales, and capturing their effect accurately is crucial for prediction. ● Regularly updating the model with new data, especially after significant events, will be beneficial for maintaining prediction accuracy.

4.b. Suggestions / Recommendations	● **Incorporate More External Factors**: Consider data like local holidays, store-specific promotions, or even macroeconomic indicators. ● **Ensemble Methods**: Combine predictions from multiple models to potentially achieve better accuracy. ● **Regular Model Updates**: As new sales data comes in, regularly retrain the model to capture the most recent trends. ● **Deployment Optimization:** Given the issues faced during deployment, future work should focus on optimizing the model for production. This might involve simplifying some features, making the model lighter, or using model quantization techniques to reduce its size and improve prediction speed.