EXPERIMENT REPORT

Student Name	Asmita Sheshrao Kamble
Project Name	Machine Learning as a Service
Date	10-10-2023
Deliverables	Model1 1. Linear regression 2. Random Forest **Git link:** https://github.com/asmitaskamble/AT2_ML-as-a-Service

---

1. EXPERIMENT BACKGROUND	
This experiment aims to develop a Machine Learning algorithm that predicts sales revenue accurately. This project involves predicting sales for specific items in a particular store on specific dates. The purpose of this project is to improve inventory management and sales forecasting. To inform decision-making and optimize stock levels, aim to gain insights into what influences sales.	
1.a. Business Objective	In this project, the business aims to improve inventory management and increase stock levels by accurately predicting sales revenues. It is expected that accurate results will result in improved resource allocation and a reduction in waste. Incorrect results can result in shortages, overstocking, and financial losses.
1.b. Hypothesis	It is hypothesized that implementing a machine learning-based revenue prediction model will significantly improve forecasting accuracy compared to traditional techniques. This study aims to determine whether modern machine learning algorithms are superior to conventional forecasting methods. A successful prediction could provide valuable insights, streamline inventory management, and increase profitability, thus justifying this hypothesis' pursuit.

1.c. Experiment Objective	The ML-based model is expected to achieve significantly higher accuracy rates (e.g., 90%) in sales revenue predictions. There are some possible scenarios: The model may succeed with the most accuracy of the model. Still, It may overfit by getting more testing errors on test data than on train data, and the third scenario is the model may not perform as expected. It might be possible to improve the model by applying some significant engineering techniques if it performs partially well.

2. EXPERIMENT DETAILS	
This approach involves data collection, preprocessing, feature engineering, and model selection. In order to ensure high-quality input data and to leverage machine learning techniques for accurate predictions, I chose this sequence. Rationale: The quality of the data affects the accuracy of predictions, feature engineering optimizes the performance of models, and model selection ensures that the best algorithm is selected for the situation.	
2.a. Data Preparation	To prepare data, some steps had to be taken, including cleaning data, handling missing values, and encoding categorical variables. As a result of these steps, it was ensured that data consistency as well as compatibility with the machine learning algorithms were maintained. I decided not to remove outliers because sales data can frequently provide valuable information about extreme cases, and I did not want to remove them.
2.b. Feature Engineering	I applied the constant features technique to generate features, but it removed id columns, which was not appropriate. Hence, for model 1, the following input features are mentioned: **item_id, store_id, and date**. As far as prediction is concerned, these are the only features set for the model. By creating meaningful variables from raw data, feature engineering improved predictive accuracy. The model captured essential patterns in sales data due to uncovering relationships between attributes, reducing noise, and improving performance. "**Event_name"** and "**Event_type**" could be crucial features in future experiments. A special event or promotion might impact sales based on these indicators. It may be possible to enhance the model's predictive accuracy and make better inventory management decisions as a result of analyzing their influence.

2.c. Modelling	I trained a linear regression model for simplicity and interpretability based on its transparency. In addition, a random forest regression was selected due to its robustness and ability to capture complex nonlinear relationships. # Number of estimators=100, # random_state=21 Regarding limited datasets, SVM (Support Vector Machine) is less suitable due to its computational complexity and sensitivity to hyperparameters. In this experiment, KNN (K-nearest neighbors) is unsuitable since it is computationally expensive, especially when dealing with large-scale or high-dimensional data. It was decided not to use either algorithm due to its potential limitations and complexity.

3. EXPERIMENT RESULTS

There was a significant improvement in performance by using random forest over the linear regression model in the ML model. The error was performed while proving the input variable for analysing prediction. The error was about selecting those features used while training the model. One risk identified is the model's sensitivity to outliers, which led to inaccurate predictions during extreme sales fluctuations.

3.a. Technical Performance

Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.

- 1. Linear Regression

```
Mean Absolute Error (MAE): 6589.5113134022795
Root Mean Squared Error (RMSE): 9137.762321411927
R-squared (R²) Score: 0.011070571115474603
```

- Random Forest

```
Mean Squared Error: 4199.08673635689
R-squared: 0.9999502674839733
```

3.b. Business Impact

It was found that the linear regression model had a high MAE and RMSE, indicating a limited level of accuracy in predicting sales revenue. The result could be overstocking or understocking, resulting in potential financial losses.

As a result of the Random Forest model's exceptional performance, the business can significantly benefit from accurate predictions, leading to a reduction of wastage, an increase in profits, and an optimization of inventory levels.

Linear Regression provides incorrect results with a higher risk than Random Forest, providing accurate forecasts.

3.c. Encountered Issues

List all the issues you faced during the experiments (solved and unsolved)—present solutions or workarounds for overcoming them. Highlight also the problems that may have to be dealt with in future experiments.

1) Issue: Managing and processing large datasets efficiently is one of the most challenging tasks. As a result, the data in this project was merged incompletely; only store_id WI data is reflected in the merged data for all the item IDs and dates.

   Solved: The data was merged by applying chunk size on the CSV files.

2) Issue: The performance of machine learning algorithms was affected by features at different scales.

   Solved: To rescale features within a standard range, use techniques such as Standardization (Z-score scaling) and apply label encoding to treat categorical features.

4. FUTURE EXPERIMENT	
As a result of the experiment, a significant performance gap was highlighted between linear regression and random forest models. It's essential to pick the correct algorithm to get accurate revenue predictions, which has implications for inventory management and profitability.	
4.a. Key Learning	In this experiment, Random Forest outperformed Linear Regression in terms of prediction accuracy. As a result of this insight, it is evident that advanced sales forecasting models are necessary. There is still room for further experimentation with the focus on fine-tuning hyperparameters and refining the feature engineering so that predictive accuracy and business value can be maximized.
4.b. Suggestions / Recommendations	The random forest hyperparameters should be fine-tuned to maximize performance and provide more accurate sales predictions. Refine feature selection and engineering techniques, targeting a reduce MSE and MAE of the model The Random Forest model meets the desired business outcomes, enabling it to be deployed into production and integrated with inventory management systems for real-time sales forecasting and optimization.