EXPERIMENT REPORT

Student Name: Yasaman Mohammadi
Project Name: Machine Learning as a Service

Date:
06/10/2023
Deliverables:Mohammadi_Yasaman-24612626-LGBMregressor.ipynb

1.EXPERIMENT BACKGROUND

1.a. Business Objective

The goal of this project for the business is to develop predictive models that accurately forecast sales revenue for specific items in each of its ten stores across three states (California, Texas, and Wisconsin) on a given date. The primary use of these predictive models is to inform inventory management, pricing strategies, and overall store operations.

The impact of incorrect results:

- **Increased Profitability:** Accurate predictions will lead to optimized inventory, reduced waste, and improved pricing strategies, ultimately increasing profitability.
- **Improved Customer Satisfaction:** By having the right products in stock when customers want them, the retailer can enhance the shopping experience and customer satisfaction.
- **Better Resource Allocation:** Accurate predictions allow for the efficient allocation of resources, such as labour and storage space, reducing costs and improving operational efficiency.
- **Competitive Advantage:** The retailer can gain a competitive edge by offering better inventory availability and pricing, attracting more customers.

The impact of incorrect results:

- **Overstocking or Stockouts:** Incorrect predictions can lead to overstocking, tying up capital and storage space, or stockouts, leading to lost sales and customer dissatisfaction.
- **Loss of Profit:** Inaccurate predictions can result in poor pricing decisions, causing potential revenue loss due to underpricing or lack of sales due to overpricing.
- **Inefficient Resource Allocation:** Resources may need to be better allocated, leading to higher operational costs and reduced efficiency.
- **Customer Dissatisfaction:** Customers may be disappointed if they cannot find the desired items, leading to a negative shopping experience.

1.b. Hypothesis

Hypothesis: Implementing machine learning models to predict sales revenue for specific items in our retail stores can significantly improve inventory management, pricing strategies, and overall store operations, ultimately leading to increased profitability.

In summary, considering the hypothesis that machine learning models for sales revenue predictions can improve inventory management, pricing strategies, and store operations is worthwhile because it has the potential to lead to a wide range of benefits, including increased profitability, better resource allocation, improved customer satisfaction, and a competitive advantage in the retail market. This hypothesis aligns with the broader trend of using data-driven insights to enhance decision-making in the retail sector.

1.c. Experiment Objective

Implementing machine learning models for sales predictions in the retail business is expected to yield several positive outcomes, including:

1. **Improved Inventory Management:** Accurate sales predictions will enable optimal inventory levels, reducing overstocking and understocking issues.
2. **Effective Pricing Strategies:** Accurate forecasts will guide pricing decisions, leading to increased revenue and improved profitability.
3. **Enhanced Store Operations:** Knowledge of high-demand items will lead to better store operations, resulting in a better shopping experience and increased sales.
4. **Data-Driven Decision-Making:** The organization will adopt data-driven decision-making processes, leading to more informed and effective decisions.
5. **Competitive Advantage:** Accurate sales predictions will provide a competitive edge, increasing market share and customer loyalty.
6. **Cost Reduction:** Efficient inventory management based on accurate predictions will reduce operational costs and improve profitability.

Possible Scenarios:

1. **Positive Outcome (Goal Achieved):** The machine learning models accurately predict sales, leading to improved inventory management, pricing strategies, and store operations. The retailer achieves its goals, such as a specific percentage reduction in overstocking, a revenue increase target, or improved customer satisfaction.
2. **Neutral Outcome:** The machine learning models provide some improvement in certain areas, but the impact is not as significant as expected. The retailer may achieve partial goals or make modest inventory management, pricing, and operations improvements.
3. **Negative Outcome:** The machine learning models do not perform as expected, leading to inaccurate predictions and potential disruptions in inventory management, pricing, and operations. The retailer may miss its goals or even face operational challenges.

2.EXPERIMENT DETAILS

2.a. Data Preparation

	the steps taken for preparing the data	Explanation
1	Data collection	The process of constructing a machine learning model commences with gathering data from diverse sources. In this instance, the data was obtained from Canvas UTS, consisting of four separate CSV files. Then they merged.
2	Memory optimization	Memory optimization is a crucial technique for effectively managing and preserving the memory resources of a data frame, offering significant benefits when working with large datasets. It boosts performance while keeping memory usage to a minimum.
3	Explore Dataset	A general understanding is derived from the data in this step.
4	Data cleaning	Data was preprocessed to remove any irrelevant, missing, or corrupted data points and any discrepancies in the data.
5	Data visualization	Visual understanding of data distribution is gained by plotting different charts in both Python and Tableau.
6	Feature engineering	New features added in this step and some features were removed.
7	Memory-Friendly Machine Learning Subset	Given the substantial dataset resulting from extensive feature engineering, we opted to employ machine learning on a more manageable subset, explicitly focusing on the data from the last two years to

		avoid kernel crashes or performance issues.
8	Splitting the dataset	The dataset was divided into two subsets: a training set, utilized to train the model, and a test set, employed to assess the model's final performance. The test size parameter was set to 20%, indicating that the last 20% of the dataset was designated as the test data.

Table 1- the steps taken for preparation of the data for previous experiment.

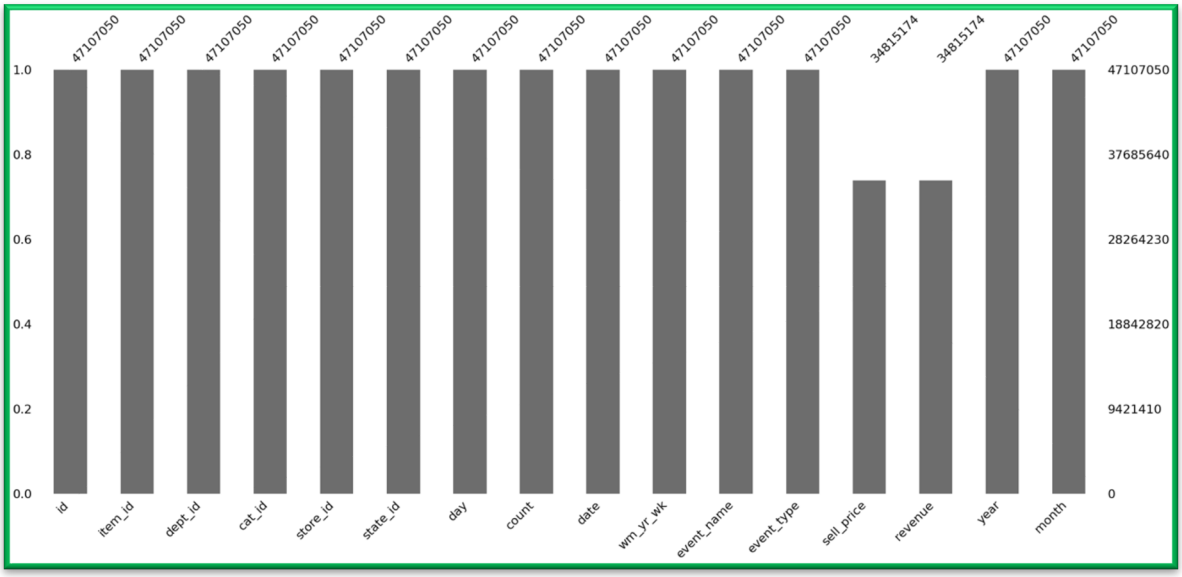

Figure2-The mnso bar chart for fiding missing values

2.b. Feature Engineering

Binary encoding:

Given the multitude of unique attributes in the event type and name categories, we employed binary encoding as an alternative to one-hot encoding. Binary encoding, a technique that converts categorical data into a binary format of 0s and 1s, offers distinct advantages in machine learning by reducing data dimensionality while retaining critical information. It is worth noting that label encoding, another option for numerical representation, was considered but deemed unsuitable due to the non-ordinal nature of these columns.

Lags:
Lag calculation is a method that involves analysing historical sales data to determine if it exerts any impact on current-day sales. This approach allows for an investigation into whether there exists a connection between the sales figures from the recent past and those of today in a retail store.

2.c. Modelling

Due to the importance of specific categorical features, namely item_id and store_id, in our dataset, we strategically chose to utilize the LightGBM model. LightGBM is particularly well-suited for handling categorical variables efficiently and effectively. Its gradient boosting framework, designed to handle categorical data, provides speed and predictive accuracy advantages.
This decision enables us to capitalize on the strengths of LightGBM, ensuring that we can leverage the categorical information in our dataset to improve the model's overall performance and predictive power. By doing so, we aim to obtain more accurate and insightful results considering the nuances of item and store characteristics.

- **N estimator**

  This is the number of boosting rounds or trees trained in the LightGBM model. It is a crucial hyperparameter because it determines how many iterations the model will go through to learn from the data. Increasing the value of "n_estimators" generally allows the model to learn more complex patterns in the data, but it also makes the training process slower and may risk overfitting if set too high.

  With grid search, a n estimator of 250 to 450 was selected for this experiment, and 450 was determined to be the best value.

K-fold cross-validation with five folds is employed in the code to evaluate and fine-tune the LightGBM regressor's hyperparameters robustly. This technique ensures that the model's performance is thoroughly assessed across multiple data subsets, reducing the risk of overfitting and providing a more reliable estimate of its effectiveness. GridSearchCV systematically explores hyperparameter combinations, such as the number of estimators ("n_estimators"), using the "neg_mean_absolute_error" metric, resulting in a well-tuned and generalized model.

3. EXPERIMENT RESULTS

3.a. Technical Performance

An RMSE value of 0.77 was achieved on the test set. It is worth mentioning that in this experiment, a subset of the dataset was used for machine learning to prevent kernel crashing. For future experiments, it is worth considering other hyperparameters and exploring different models like CatBoost. Additionally, incorporating some carefully engineered feature columns may enhance predictive performance further.

3.b. Business Impact

Accurate predictive models are pivotal for the business's goal of forecasting sales revenue across its ten stores in California, Texas, and Wisconsin. Correct predictions lead to increased profitability through optimized inventory, reduced waste, and enhanced pricing strategies, fostering better customer satisfaction and resource allocation. Conversely, incorrect results can cause overstocking or stockouts, loss of profit, inefficient resource allocation, and customer dissatisfaction, potentially impacting revenue and operational efficiency. Accurate predictions are financially advantageous and essential for maintaining a competitive edge and ensuring a positive shopping experience for customers.

3.c. Encountered Issues

The primary challenge was the dataset's substantial size, which consistently led to code crashes during processing. To mitigate this issue, we initially reduced the dataset's size significantly. However, even with this reduction, we still encountered crashes during the machine-learning phase. As a solution, we resorted to working with a smaller subset of the data to ensure smooth processing and analysis.

4.FUTURE EXPERIMENT

4.a. Key Learning

An interactive visualization in Tableau has been created to visualize data and were uploaded to GitHub repo.

Given the high level of accuracy achieved by the model, it is worth considering its deployment for practical use. Exploring alternative models like CatBoost or fine-tuning other hyperparameters to enhance predictive performance in future experiments may also be beneficial.

4.b. Suggestions / Recommendations

In future experiments, different machine learning models with hyperparameter tuning can be used.
To deploy the final solution into production, the user interface must be refined so that predictions are easy to access, and ongoing model adjustments should be monitored in real-time.