EXPERIMENT REPORT

Student Name	Sareem Sandeed
Project Name	Part A: AT2 - Predictive model using a Machine Learning algorithm
Date	10-October-2023
Deliverables	**Notebook**: sandeed_sareem_24622829_predictive_lightgbm.ipynb **Best Model:** Light Gradient Boosting Machine Regressor **GitHub link:** https://github.com/ssandeed/adv_mla_at2

---

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	**Project Goal:** The goal of this project is to develop a predictive Machine Learning algorithm that accurately predicts the sales revenue for a given item in a specific store on a given date. **Usage of Results:** The results of this predictive model will be used to optimize inventory management, demand planning, pricing strategies, marketing efforts, and financial planning. **Impact of Accurate Results:** Accurate results from the predictive model will enable the business to make data-driven decisions, leading to improved profitability, efficient operations, and enhanced customer satisfaction. **Impact of Incorrect Results:** Incorrect results from the predictive model can lead to bad decision-making, ineffective pricing and marketing strategies, and potential customer dissatisfaction, ultimately affecting the business's success.
1.b. Hypothesis	**Hypothesis:** The hypothesis to find sales revenue for the retail giant provided that have a given date, store, and items. **Question:** Can we accurately predict the sales revenue for a given date, store, and items for the retail giant using a Machine Learning algorithm? **Reason:** It is worthwhile to consider this hypothesis because accurate sales revenue predictions can significantly impact inventory management, demand planning, pricing, and marketing strategies, leading to improved operational efficiency and financial performance for the retail giant.

1.c. Experiment Objective	**Expected Outcome:** The expected outcome of the experiment is to successfully develop a predictive model that accurately determines the sales revenue for a given store, item, and date. The primary evaluation metric for the model will be the Root Mean Squared Error (RMSE), with the goal of minimizing it. **Possible Scenarios:** 1. **Poor Model Performance:** The model's predictions have a high RMSE, indicating inaccurate sales revenue forecasts. In this scenario, further model refinement and feature engineering will be necessary to enhance its accuracy. 2. **Overfitting or Underfitting:** The model may exhibit signs of overfitting (performing well on the training data but poorly on new data) or underfitting (performing poorly on both training and new data). These scenarios would require adjustments to the model architecture and hyperparameters. 3. **Moderate Model Performance: T**he model performs reasonably well but may have a moderate RMSE. While not achieving the ideal accuracy, it still provides valuable insights for the business and can be considered a good starting point for further improvements.

---

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 datasets didn't have major issues. There were four datasets that needed to be joined to get relevant information like revenues, items prices, sales quantity, and date. • **calendar_events**: Major events along with date • **calendar**: Dates with week id and date id • **items_weekly_sell_prices:** Unit sale price of each item • **sales_train:** Sales quantity on dates **The following steps were taken for data preparation:** 1. Checking the null values, duplicates in each of the datasets. 2. Checking data types of all columns. 3. Merging all the datasets. 4. Dropping columns that are correlated and already have the same information in different columns can lead to overfitting and inaccurate predictions. 5. Splitting data store wise as I wanted to model at store level. 6. Final store level data was split into train and validation data.

2.b. Feature Engineering	**In the final dataset before splitting at store level, following features were used for modelling:** • item_id: 3049 unique items • store_id: 10 stores • date • revenue The ask of the project was to predict sales revenue for a given item, store, date. So, it must have these features only for the model to work. Features engineering was done on date to get year, month, day and days of the week. One hot endoing was done on ['year','month','day','day_of_week'] Target encoder was used in item_id as there were too many features to consider for one hot encoding.
2.c. Modelling	I chose to use LightGBM as the predictive model due to its advantages in handling complex tabular data and efficient gradient boosting capabilities. **Reasons for Choosing LightGBM:** 1. **Efficiency:** LightGBM is known for its speed and efficiency, making it suitable for handling large datasets, which is common in retail sales forecasting. 2. **Gradient Boosting:** It uses gradient boosting, an ensemble learning technique known for its ability to capture complex patterns in data and produce accurate predictions. 3. **Categorical Feature Support:** LightGBM has native support for categorical features, which is essential in this retail scenario where store locations, item categories, and dates may be categorical variables. **Hyperparameter -:** No hyperparameter tuning was done as given the large datasets, it require high computational power and hence limited my capability to explore further. Other models, such as Random Forest, SDG were trained but the LightGBM ran faster on large datasets. While Random Forest were giving better results but it was taking too much time. **Potential Model Importance for Future Experiments:** In future experiments, it would be valuable to explore the performance of different models, especially those known for handling complex relationships in data. Hyperparameter tuning should also be systematically performed to optimize model performance. Additionally, different stratified sampling can be done so that it can be used with other algorithms.

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 metric used for assessing model performance is the Root Mean Squared Error (RMSE) score. RMSE is in the same unit as the target variable, making it easy to interpret. This makes it more understandable to stakeholders. Also, RMSE allows for straightforward model comparison for all 10 stores consistently.

	baseline_rmse	train_rmse	val_rmse
store_CA_1	7.242093	7.233376	7.276859
store_CA_2	5.735208	5.721615	5.789264
store_CA_3	9.695821	9.700983	9.675144
store_CA_4	4.415324	4.409242	4.439569
store_TX_1	5.785319	5.787674	5.775888
store_TX_2	7.192938	7.199415	7.166971
store_TX_3	6.388319	6.383801	6.406360
store_WI_1	4.808941	4.805001	4.824669
store_WI_2	6.594577	6.585164	6.632098
store_WI_3	6.373680	6.362419	6.418527

Consistency: Across most stores, the RMSE values are relatively consistent, indicating that the model's performance is similar across different store locations.

Generalization: RSME is similar across baseline train and validation, therefore model has been able to general very well.

Next Steps: The specific actions to be taken should align with the business objectives and the desired level of accuracy. Potential actions include hyperparameter tuning, feature engineering, and exploring alternative modeling techniques to enhance model performance.

3.b. Business Impact

The experiment involved training a LightGBM model to predict sales revenue. The model achieved reasonable RSME on train and validation data. These scores indicate that the model can predict revenue on a given day for an item at a store with good confidence.

Impact of Incorrect Results:
1. Bad pricing decisions resulting from incorrect prediction may lead to missed opportunities for profit during high-demand periods or unnecessary discounts during low-demand periods.
2. Misallocated marketing budgets based on inaccurate predictions can result in ineffective campaigns and missed chances to boost brand visibility and ROI.
3. Inaccurate sales prediction can disrupt financial planning, budgeting, and revenue target setting, potentially leading to financial instability.
4. Inadequate stock levels due to inaccurate predictions can result in customer dissatisfaction, leading to lost sales and reduced loyalty.
5. Inaccurate prediction can damage the credibility of the business, especially when widely communicated to stakeholders, investors, or partners.

3.c. Encountered Issues	**Solved Issues:** **1. Big Datasets:** Handling large datasets can be computationally expensive and may lead to memory constraints during model training and evaluation. **2. Feature Engineering:** Feature engineering was performed, enhancing the predictive power of the model. **3. Datasets merging:** The data preparation required merging multiple datasets and handling them without the memory issue. **Unsolved Issues:** **1. Computational Resources:** Training complex models with extensive hyperparameter tuning can require substantial computational resources, which may be expensive or time-consuming. **2. Hyperparameter Tuning:** Tuning hyperparameters manually can be time-consuming and may not lead to optimal results. **Solutions or Workarounds:** 1. **Hyperparameter Tuning:** Perform hyperparameter tuning using techniques like grid search or random search. Optimize hyperparameters to achieve better model performance. 2. **Model Diversity:** Explore other models like Random Forest, SVM, or Gradient Boosting. Ensemble methods can combine multiple models for better performance. 3. **Regularization:** Apply regularization techniques (L1, L2) to prevent overfitting and improve model generalization. **Issues for Future Experiments:** 1. **Model Evaluation:** While gradient boosting (e.g., LightGBM) was used successfully in this experiment, future experiments may explore other ensemble methods like Random Forest or Neural Network to determine if they offer improved performance. But must have to think with respect to computational capability. 2. **Domain Knowledge Integration:** Involve domain experts to validate model predictions and provide insights on potentially overlooked factors. 3. **Overfitting:** Overfitting occurred in some cases, where the model performed well on the training data but poorly on validation or test data. In summary, while several issues have been addressed, hyperparameter tuning remains a gap, and other aspects such as model diversity and overfitting should be considered in future experiments to enhance the model's overall performance and reliability.

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 choice of LightGBM as the predictive model demonstrated efficiency and effectiveness in handling large, complex tabular data, aligning with the project's goal. LightGBM proved to be an efficient choice for handling large datasets, outperforming other models like Random Forest and SGD in terms of computational speed. This insight is valuable for practical implementation in a business context. The use of RMSE as the performance metric provided a consistent and interpretable measure of model performance across all 10 stores. RMSE's interpretability and consistency make it a suitable choice for assessing model accuracy in a retail sales forecasting context.

	Given the success of LightGBM in terms of efficiency and reasonable model performance, further experimentation is required. However, there are opportunities for improvement and exploration: 1. **Hyperparameter Tuning:** Future experiments should explore hyperparameter tuning to optimize model performance. While computational constraints limited tuning in this experiment, leveraging advanced optimization techniques can potentially enhance accuracy. 2. **Other Models:** Evaluating other ensemble methods like Random Forest and SVM could be beneficial to assess if they provide even better predictive accuracy compared to LightGBM. 3. **Experiment with Sampled Data:** As datasets continue to grow, experimenting with sampled data could help address the challenges of handling even larger datasets efficiently. 4. **Model Robustness:** Investigate strategies to ensure that the model remains robust over time and can adapt to changing market conditions. Continual monitoring and retraining processes may be necessary. 5. **Ethical Considerations:** Future experiments should consider ethical aspects, including bias in predictions and potential consequences of inaccurate forecasts on stakeholders and customers.
4.b. Suggestions / Recommendations	Based on the achieved results and the overall objective of the project, here are some potential next steps and experiments: **Hyperparameter Tuning (High Priority):** • **Expected Uplift:** Hyperparameter tuning can significantly enhance model performance by optimizing model parameters. • **Rank:** High priority. • **Recommendation:** Perform an exhaustive hyperparameter search using techniques like grid search, random search, or Bayesian optimization to fine-tune the model. Consider leveraging cloud resources to expedite this process. **Scaling for Larger Datasets (Medium Priority):** • **Expected Uplift:** Scaling the solution to handle even larger datasets can future proof the system. • **Rank:** Medium priority. • **Recommendation:** Explore distributed computing and big data frameworks but have to consider the trade-offs in terms of complexity and cost. **Feature Engineering (Low Priority):** • **Expected Uplift:** Carefully crafted features can enhance model performance by capturing more relevant information from the data. • **Rank:** Low priority. • **Recommendation:** Explore feature engineering techniques tailored to retail sales data, such as lag features, rolling statistics, and holiday/event indicators. **Deployment Steps for a Successful Solution:** If the experiment achieves the required outcome for the business, follow these steps to deploy the solution into production: **Model Evaluation:** Perform a final evaluation of the model's performance using an independent test dataset to ensure it meets the desired accuracy and reliability. **Scalability Assessment:** Verify that the model deployment can handle the anticipated workload and data volumes in a production environment. **Feedback Loop:** Establish a feedback loop to collect user feedback and monitor the model's accuracy and impact on business operations. **Business Integration:** Integrate the predictive model's outputs into key business processes, such as inventory management, pricing, marketing, and demand planning.