EXPERIMENT REPORT

Student Name	Monali Patil
Project Name	AT2 - Machine Learning as a Service
Date	10/10/2023
Deliverables	Notebook name: Patil_Monali-14370946-EDA.ipynb Patil_Monali-14370946-Predictive_XGBoost.ipynb Model name: XGBoost Regressor for Revenue Prediction Project Repo: https://github.com/MonaliPatil19/adv_mla_assignment2.git API Servicing Repo: https://github.com/MonaliPatil19/assignment2_api.git

---

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

Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?

Goal: The project involves working with an American retail chain operating across California (CA), Texas (TX), and Wisconsin (WI) having 10 stores to develop two distinct machine learning predictive and forecasting models to assist in business decisions related to sales revenue.

Application: A predictive model would be utilized to accurately predict sales revenue for individual items in specific stores for any given date crucial for managing inventory levels, optimizing pricing strategies, and ensuring product availability. The forecasting model would be employed to forecast the total sales revenue across all stores and items for the next 7 days allowing for better resource allocation, staffing decisions, and financial planning.

Impact: The accuracy of these predictive and forecasting models is crucial for the retailer's operational efficiency, profitability, and customer satisfaction. Correct predictions and forecasts empower the retailer to make informed decisions that positively impact its business revenue growth, while inaccurate results can lead to financial losses with unsold inventory by overstocking, stockouts by understocking, and inadequate staffing, leading to poor customer service and lost revenue.

1.b. Hypothesis	Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it, Hypothesis: Can the deployment of machine learning models for sales revenue prediction in a retail environment result in enhanced inventory management, more effective pricing strategies, and increased overall business profitability? Rationale: Accurate sales predictions at the item and store levels can assist in pricing strategies to optimize inventory management. By stocking the right quantity of products, the retailer can minimize holding costs and reduce the risk of understocking or overstocking. Moreover, the hypothesis aligns with a prevalent objective within the retail industry, which involves harnessing data-driven insights to make well-informed decisions that drive business success.
1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment. The retailer's expected outcome of this experiment would be development of a machine learning predictive model utilizing regressor algorithm that can accurately predict the sales revenue for individual items in specific stores for any given date. The goal is to ensure that the predictive model provides accurate sales revenue predictions and model to be deployed as a service, ultimately benefiting the retail business. Possible scenarios resulting from this experiment include the following situations: • Success: The predictive model accurately predicts sales revenue, leading to improved inventory management, optimized pricing strategies, and increased profitability. • Moderate Success: The model performs reasonably well but may require further optimization to achieve the desired accuracy. • Failure: The model fails to provide accurate predictions, which may lead to suboptimal inventory management and pricing decisions, potentially impacting profitability.

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

Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments

The retail data underwent exploration, cleaning, and preparation as part of the subsequent steps required for the utilization of the regression algorithm.

• Data Understanding

1] Loading Data: Imported below given data in separate CSV files into the panda's data frames to use and create a predictive model.

- Training Data
- Evaluation Data
- Calendar
- Events
- Items Price per Week

2] Exploring Data: Examined and studied information using different panda functions to comprehend and uncover patterns, aiming to identify prospective features for revenue prediction. The below tasks are performed to ensure the quality of the data to be utilized by the model analysis.

- Handing missing/null values.
- Eliminating identifiers.
- Combining the individual datasets into a single dataset
- Computing the total revenue based on the items sold and the weekly price of an item
- Processing the 'date' feature to derive additional date related information
- Selecting appropriate features
- Identifying and eliminating identifier
- Splitting data into different datasets
- Handling and imputing missing values

df.head(): Checking initial records of the dataset.
df.shape(): Verifying the dimension of the dataset.
df.columns: Identifying attributes name.
df.info(): Assessing the summary information of the attributes of the dataset.
df.describe(include='all'): Examining statistical summary information for all variables of the dataset across different data types.
df.isnull().sum(): Examining whether there are any null values in the dataset.

• Data Preparation

3] Reshaping the training dataset and renaming the attributes

	id	item_id	dept_id	cat_id	store_id	state_id	d_1	d_2	d_3	d_4	...	d_1532	d_1533	d_1534	d_1535	d_1536	d_1537	d_1538
0	HOBBIES_1_001_CA_1_evaluation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	1	1	1	0	1	0	1
1	HOBBIES_1_002_CA_1_evaluation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	0	0	0	0	0	0	0
2	HOBBIES_1_003_CA_1_evaluation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	0	0	1	0	0	0	0
3	HOBBIES_1_004_CA_1_evaluation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	8	2	0	8	2	3	1
4	HOBBIES_1_005_CA_1_evaluation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	2	0	1	3	2	1	1

* The above training dataset had columns named from d_1 to d_1541 representing the sold item volume for each day for distinct items from 10 stores across 3 separate states. So, melt functionality was utilized to reshape the data to improve attribute organization and for more convenient and suitable data analysis and modelling task.
Additionally, the 'd_*' column was relabeled as 'dayofsale,' and the 'sold item numbers' were renamed as 'volume.

4] Combining the individual datasets into a single dataset

id	item_id	dept_id	cat_id	store_id	state_id	dayofsale	volume	date	wm_yr_wk	d	event_name	event_type	sell_price
BBIES_2_050_CA_4_evaluation	HOBBIES_2_050	HOBBIES_2	HOBBIES	CA_4	CA	d_741	0	2013-02-07	11302	d_741	NaN	NaN	1.97
OODS_3_782_WI_2_evaluation	FOODS_3_782	FOODS_3	FOODS	WI_2	WI	d_114	0	2011-05-22	11117	d_114	NaN	NaN	2.48
EHOLD_2_505_CA_1_evaluation	HOUSEHOLD_2_505	HOUSEHOLD_2	HOUSEHOLD	CA_1	CA	d_686	1	2012-12-14	11246	d_686	NaN	NaN	4.97

* Combining data from various sources into a single data frame is a fundamental step in data preprocessing for modelling. It ensures data consistency, enables feature engineering, and simplifies the overall analysis process, ultimately contributing to better decision-making and model performance.

5] Feature Engineering.

Please refer to the 2.b. Feature Engineering section.

6] Selecting appropriate features

Features selected: 'item_id', 'store_id', 'date', 'event_name', 'event_type', 'revenue'

* To build a machine learning predictive model, above specific features were chosen due to the presence of duplicate columns after aggregating the datasets. This duplication resulted in the large dataset leading to memory related issues. For instance, ‘store_id’ field values (CA_1, WI_2, TX_9) already incorporates the ‘state_id’ attribute values (CA, WI, TX). Therefore, it's logical to choose pertinent features that have generalized unique values and avoid duplicates.

* Additionally, to address the memory challenges associated with analyzing the extensive dataset, the aforementioned features were chosen based on their relevance from a business perspective.

7] Identifying and eliminating identifier

* The 'id' attributes serve as unique identifiers for individual items sold within a specific category, store, and state on a particular date. As a result, during the feature selection process, this identifier was omitted.

* Including it in the analysis could potentially result in overfitting, where the model becomes too focused to these specific values, rather than capturing the underlying, generalized patterns present in the retailer's data.

8] Splitting data into different datasets

To split the dataset into training and validation below strategy was employed.

- Training data: From 2011-01-01 to 2014-12-31
- Validation data: from 2015-01-01 to 2016-12-31

* The retail data covers multiple years, and by splitting it into training (historical) and validation (future) sets, would maintain the temporal consistency of the dataset. This ensures that the model is trained on past data and tested on more recent data, mimicking a real-world scenario.

* Furthermore, this approach aided in handling the relatively compact dataset, mitigating memory-related challenges.

9] Handling and imputing missing values

After combining the datasets, missing values were identified in the attributes.

Attribute Name	Total missing values
event_name	43143350
event_type	43143350
revenue	12291876

* To handle missing values in the 'event_name' and 'event_type' features, 'None' was used as a replacement for the missing values.

* For address missing values in the 'revenue' target feature, 0 was used for imputation. This signifies days with no items sold and, consequently, zero income.

The following measures are crucial and may hold significance for any future regression experiments.

1. When working with date-related datasets and business scenarios, it's vital to split the data into training and validation sets based on the date. This practice helps prevent data leakage, where information from the validation period unintentionally influences the training process. This ensures a fair assessment of the model's predictive performance.
2. To prevent the model from becoming too specialized on particular data points and to enable it to learn generalized patterns, it's essential to eliminate any identifying attributes.
3. Carefully evaluating and handling missing values is crucial to build machine learning models and avoid introducing biases into the model.

2.b. Feature Engineering

Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments

- Computing the total revenue based on the items sold and the weekly price of an item

* Calculating the overall revenue by considering the quantity of items sold and their corresponding weekly prices was essential for developing the machine learning model to predict target variable revenue.

- Processing the 'date' feature to derive additional date related information

* To enhance the dataset's date related information by extracting additional date related features from the 'date' attribute such as day of a week, month, year, and week of the year for model to learn generalized features.

* Therefore, set the datetime type to the date feature and utilized the datetime functionality to extract the day of the week, month, year, and week of the year information.

2.c. Modelling

Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments

Regression Algorithm: The target variable 'revenue' contains continuous values that represent sales income. Since our goal is to predict these continuous incomes, using a regression algorithm is the appropriate choice, aligning with the objective of learning and prediction.

Algorithm Name	Rationale
1. XGBoost Regression	* XGBoost algorithm is an ensemble learning method, which means it combines the predictions of multiple weak learners (typically decision trees) to create a stronger, more accurate model. It builds trees sequentially, each one correcting the errors of the previous trees, leading to improved predictive performance. * The algorithm has demonstrated exceptional predictive performance in various machine learning competitions and real-world applications. It is known for its ability to handle complex, non-linear relationships in the data, making it suitable for regression tasks. * Additionally, it is optimized for speed and efficiency, making it suitable for large datasets. It can handle a substantial number of features and samples efficiently.

2. Linear Regression	* Linear regression algorithm is computationally efficient and does not require the extensive computational resources that some complex machine learning algorithms demand. This efficiency makes it suitable for large datasets and real-time applications. * It serve as a valuable starting point for regression tasks, allowing data scientists to establish a base model before exploring more complex alternatives.

Pipeline

* The pipeline automates the steps required to preprocess input data, make predictions as the same preprocessing and prediction steps are consistently applied to incoming data, ensuring that results are reproducible and dependable.

* Building a pipeline for model deployment for servicing with an API enhances the model's usability and reliability in using the predictive model in a operational environment.

Because of time limitations, could not dedicate time to employ hyperparameters and fine tune the model to optimise the 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

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

The table below informs the Mean Absolute Percentage Error (MAS) and Root Mean Square Error (RMSE) performance scores for two algorithms on both training and validation datasets.

Algorithm	Datasets	MAS Score	RMSE Score
	Baseline Performance	4.3372	9.0484
Experiment 1: XGBoost Regressor	Training Dataset	4.0605	8.7152
	Validation Dataset	4.3389	10.4643
Experiment 2: Linear Regression	Training Dataset	4.2788	9.0114
	Validation Dataset	4.4691	10.6893

The baseline model has a MAS score of 4.3372 and an RMSE score of 9.0484. This serves as a reference point for evaluating the other models.

Experiment 1: XGBoost Regressor

* The XGBoost Regressor algorithm achieves a lower MAS score (4.0605) and RMSE score (8.7152) compared to the baseline on the training data. This indicates that the model provides a better fit to the training dataset.

* However, the model's performance on the validation dataset, deteriorates slightly, with a slightly higher MAS score (4.3389) and considerably higher RMSE score (10.4643), suggesting that the model is overfitting to the training dataset, as it struggles to generalize to unseen data.

Experiment 2: Linear Regression

* The Linear Regression model has a MAS score of 4.2788 and an RMSE score of 9.0114 on the training dataset. It performs slightly worse than the XGBoost Regressor but is still better than the baseline.

* On the validation dataset, the model has a higher MAS score (4.4691) and RMSE score (10.6893) compared to the XGBoost Regressor. Similar to XGBoost model, it also overfitts on the training dataset.

Both the XGBoost Regressor and Linear Regression models show reduced performance on the validation dataset, suggesting overfitting. Therefore, regularization techniques or hyperparameter tuning should be considered for both models to address the overfitting issue.

3.b. Business Impact	Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others) * The XGBoost Regressor outperforms the Linear Regression model in terms of both Mean Absolute Percentage Error (MAS) and Root Mean Square Error (RMSE) scores. This indicates that the XGBoost model provides more accurate predictions of sales revenue. * Both the XGBoost and Linear Regression models show signs of overfitting, particularly on the validation dataset. This means that the models may have learned specific patterns in the training dataset that do not generalize well to new, unseen data. The impact of overfitting is that the models may make less accurate predictions while servicing in real-world scenarios. * The root causes of model performance issues, such as overfitting, need further investigation. It could be due to insufficient data, or data transformation choices like ordinal encoder. Identifying and addressing these root causes is critical for improving model performance and business outcomes. * While the machine learning models show promise in improving sales revenue predictions, they also exhibit overfitting issues. * Additionally, incorrect predictions can have various impacts on the business such as overstocking or understocking of products, affecting inventory management efficiency, suboptimal pricing decisions, affecting profitability and inefficiencies in staffing and supply chain management.
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 issues that may have to be dealt with in future experiments. * Memory constraints was a challenge, particularly during EDA tasks involving the large dataset. Nonetheless, this issue was successfully addressed by splitting the dataset by date and selecting pertinent features. * During the pipeline construction, could not employ one-hot encoding for categorical feature transformation with a large number of unique values in categorical features. This approach significantly increased the dimensionality of the features, leading to excessively long processing times before encountering system crashes. Therefore, ordinal encoder was utilised.

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	Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end. * The experiment outcomes provided valuable insights into the challenges and potential improvements for predictive modelling in the retail sector. One key takeaway is the need for sophisticated feature engineering and data preprocessing techniques to handle large datasets efficiently and address issues like memory constraints. * Additionally, the overfitting observed in both the XGBoost and Linear Regression models suggests the importance of hyperparameter tuning to optimize model performance. * Continuing with the current approach is worthwhile, but it requires further experimentation with techniques such as dimensionality reduction, regularization, and hyperparameter tuning to improve model generalization and reduce overfitting. * Exploring alternative algorithms and ensemble methods could also lead to better predictive accuracy. This experiment highlights the path forward for refining the predictive model, but it requires additional iterations and adjustments to achieve the desired results.
4.b. Suggestions / Recommendations	Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production. Primarily would be performing these two tasks. * Experiment and train the XGBoost regressor algorithm with hyperparameters like max_depth, subsample, max_leaves etc. to improve their predictive performance for better model generalization. * Investigate dimensionality reduction techniques like Principal Component Analysis (PCA) or feature selection to reduce the number of features without sacrificing predictive power. mitigate memory issues.