EXPERIMENT REPORT

Student Name	Simon Lim
Project Name	Predictive Model of the Sales Revenue
Date	10/10/2023
Deliverables	Lim_Simon-24661225-Pred_XGBoost.ipynb SimonUTS24661225/ADV_MLA_AT22 (github.com)

---

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 goal of the project is to build potential predictive models using a Machine Learning algorithm to accurately predict the sales revenue for a given item in a specific store at a given date. Also, the project aims to deploy trained models with Heroku and thus models can be used as a service for business uses.
1.b. Hypothesis	1.It is hypothesized that XGBoost Regressor model will be likely to be a potential model for this project, given that the dataset is large and contains a variety of items across state, store and category. XGBoost Regressor can efficiently handle large datasets. Also, XGBoost is one of the optimal machine learning algorithms, which shows very strong predictive performance. As the model will be deployed for real use, the performance of the model is also crucial. 2.Also, RMSE score will be likely to be a suitable performance metric in this project, in order to predict the sales revenue for a specific item at a specific date.

1.c. Experiment Objective	The expected outcome and scenarios of the experiment is that XGBoost models will be trained and used to predict the sales revenue of items at specific dates and stores. It is assumed that the RMSE score of models will be between 0 and 1, indicating strong predictive performance. Subsequently, once the performance of the model is assured, the model will be deployed by using deployment techniques and tools, including FastAPI, Docker and Heroku. Finally, the trained model will be used as a service, in order to predict the sales revenue for a specific item at a specific date and store.

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	1. There were five separate datasets, including training dataset, testing dataset, selling price per week, calendar and calendar of events. The first stage in data preparation was to merge those datasets. Training and testing datasets had many columns representing each day's sales. I switched those columns into rows using "pd.melt()". 2. Calendar and calendar of events were merged on a common column "**date**" and then this combined calendar dataset was merged with training and testing dataset on a common column "**d**" (day). 3.Finally, combined training and testing datasets were merged with selling price per week dataset on common columns **"store_id**" and **"item_id".** 4.There were some null values in **'sell_price', 'event_name'** and '**event_type'** columns. I converted null values to 0, as all values in those columns will be used for feature engineering. 5.Explatory Data Analysis was performed to investigate trends in data.
2.b. Feature Engineering	Six new features were created using existing features. - The sales revenue (**'revenues'**) column was created by multiplying '**sales'** by **'sell_price'.** This column will be used as a target variable when training models. - The event day ("is_event") column was created by using "event_name" column. If there was any event on a row of data, "is_event" was 1 and if there is not event, "is_event" was 0. - The day of week ("day_of_week") column was created from Monday to Sunday by using "date" column. - The "year", "month" and "day", in numeric format, were created by using datetime.

2.c. Modelling	Data Pipeline was a main preprocessing technique used to transform categorical columns to ordinal numeric columns, scale numeric columns and train XGBoost Regressor algorithm with pre-processed data. Categorical values were converted to numeric values using “OrdinalEncoder”, including “store_id”, “cat_id”, “day_of_week”, “dept_id”, and “state_id”. All numeric columns were normalized, using “StandardScaler”. Finally, “ColumnTransformer” technique was used to assemble all transformed data into a preprocessor. There were two models of XGBoost Regerssor used in this experiment. One was XGBoost with the default model. The second model was XGBoost regressor with tuned hyperparameters. Since the dataset was too huge, it was unable to use “RandomSearchCV” or other tuning techniques due to a huge delay of processing. Thus, I manually specified hyperparameters that are particularly optimal for a large dataset.

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	For performance metrics, four different performance metrics were tested, including RMSE, MAE, MSE and R2. However, RMSE is mainly used to assess and compare the performance of trained models. The purpose of the project is to establish a predictive model that can accurately predict the sales revenue. RMSE is sensitive to the difference between predicted and actual values. Also, it is sensitive to large errors and given that the trained model will be used in real-world applications, RMSE is a suitable performance metric in this project.
3.b. Business Impact	As hypothesized, XGBoost Regressor models have shown great performance in predicting the sales revenue of items. RMSE score of default XGBoost model was 1.58, while RMSE score of tuned XGBoost model was 1.53. As the performance of predictive model was outstanding, models can be deployed for real-world application uses. Models will be deployed by using Heroku, which allows an access to online.
3.c. Encountered Issues	1. It was planned that 'RandomSearchCV' would be used to find optimal hyperparameters of models. However, since the final training dataset contained too many data (`47107050`), it was unable to use "RandomSearchCV" or other tuning techniques (unsolved). 2. Also, it was difficult to visualize the predicted data due to its large size (Unsolved).

---

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	1. Explanatory Data Analysis (EDA) is an essential stage prior to training models, to investigate and obtain particular trends of data. 2. Feature engineering is a crucial process in machine learning, which can potentially increase the performance of models. 3. It is also crucial to find the optimal hyperparameters of the model for better results. 4. If conducting a few different experiments, using functions and methods (i.e., custom packages) can be useful to simplify codes and identify patterns. 5. Consider subsampling if the dataset is extremely large. 6. Selection of algorithm for training model is also important depending on the context of dataset. 7. Deploying trained models, using tools such as Fastapi and Heroku for users to access the models

4.b. Suggestions / Recommendations	1.Deciding main performance metric in advance is very helpful in building optimal models. 2. Feature engineering is an essential stage in machine learning, that can influence models and outcomes of projects. 3. Use functions or classes to simplify code and use in other experiments. 4. Use docker containers to smoothly operate and perform productions 5. Use Heroku that allows an access to everyone online.