# **EXPERIMENT REPORT**

| **Student Name** | Trung Kien Hoang |
| --- | --- |
| **Project Name** | Machine Learning as a Service |
| **Date** | 06/10/2023 |
| **Deliverables** | <Experiment 1>  <XGboost>  <Predictive models>  <https://github.com/KenUTS/adv\_mla\_assignment\_2/tree/master/reports> |

| 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 objective of this business project is to develop a prediction model utilizing a Machine Learning algorithm in order to effectively forecast the sales income of a particular item within a designated retailer on a specific date. The findings will be utilized by an American retailer to inform their sales strategies, such as implementing discount promotions on days with lower revenues. Additionally, this tool can assist in the identification of stores that are experiencing poor sales performance for a given item. This enables retailers to implement appropriate solutions, such as optimizing item delivery and implementing effective marketing strategies. Similarly, in the context of a store generating substantial money from a particular item, it is advantageous for the store to prioritize the delivery of a larger quantity of its top-selling items in order to optimize its overall earnings. The attainment of precise outcomes contributes to the optimisation of sales and the formulation of effective marketing strategies for an American store. However, the presence of inaccurate outcomes may result in significant financial losses pertaining to the transportation and storage of goods, particularly in the case of perishable food items. | |
| **1.b. Hypothesis** | There exists a correlation between sales income and date-related factors, including the day of the month, month of the year, and day of the week. Additionally, there exists a correlation between the quantity of sales on days with activities and those without events. It is imperative to take into account the significance of this matter, as some commodities, such as food products, are frequently purchased during events. Furthermore, the consideration of date features holds significance in examining this project, as individuals tend to engage in shopping activities during weekends or designated shopping days. Customers are aware of the recurring nature of discount dates, which are consistently observed on specified days or months. These factors merit consideration in the context of this task. | |
| **1.c. Experiment Objective** | The objective of this experiment is to minimize the root mean square error (RMSE) score for each of the selected models in order to identify the optimal model for the business. Additionally, the selected models are expected to exhibit robustness against overfitting when presented with new data. | |

| 1. **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** | Data cleaning:   1. Combining calendar and events datasets. 2. Combine the main dataset vs combined datasets above. 3. Combine the dataset above with items' weekly price dataset. 4. Fill null values in weekly sell price with average prices of the same item in different stores. 5. Remove unused columns such as sale, day, wm\_yr\_wk,sell\_price and event\_type.   Data splitting:  As a result of the substantial size of the final dataset, it has been partitioned into seven smaller subsets, taking into consideration the various departments across all stores. The subsets are partitioned into training and validation datasets, with a ratio of 8:2 for the purpose of training and evaluating models. Finally, a stratify technique is performed to ensure that the distributions of items in training and validating data. | |
| **2.b. Feature Engineering** | In the present study, a novel variable denoted as "revenue" is generated through the multiplication of the sales figure by the price of each item. A newly added column, titled "event," has been developed to verify the presence of an event on a certain day.  A pipeline is built to transform features such as scaler all numerical columns, one hot encoding for store identification and label encoding for all item identifications. Furthermore, date records are converted from timestamp to new columns as day of month, month of year and day of week. | |
| **2.c. Modelling** | In this experiment, three machine learning methods, including linear regression, decision tree, and XGboost, were taught. The ideal hyperparameter for the decision tree is determined to be a maximum tree depth of 10 to reduce data overfitting. In addition, the random\_state parameter is consistently assigned a value of 42 for all procedures, including data partitioning and data modeling. | |

| 1. **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** | RMSE scores for the linear regression model for training and validating sets:    RMSE scores for the decision tree model for training and validating sets:    RMSE scores for the XGboost model for training and validating sets:    It is evident that the RMSE scores demonstrate superior performance of XGboost models in terms of having the lowest scores and exhibiting less overfitting across all departments, as compared to the other two models. Moreover, the departments of Hobbies 2 and Household 2 are more suitable for this model. | |
| **3.b. Business Impact** | The model utilizes the difference of mean square values, around 7, to make predictions about sales revenue. One potential approach for identifying and predicting the sales revenue of an item across several stores is to employ a suitable methodology or model. The potential inaccuracies in the results may have a detrimental effect on the appropriate allocation of stocks across stores, resulting in increased financial losses in terms of delivery and storage. | |
| **3.c. Encountered Issues** | As a consequence of the substantial size of the dataset, my laptop is unable to train models on the entire dataset due to memory limitations. Hence, a proposed approach involves partitioning the dataset into seven distinct subsets, each corresponding to a specific item department. | |

| 1. **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 XGboost model was chosen as the optimal method for predicting the target variable, as determined by the Root Mean Square Error (RMSE) score. In addition, the XGboost algorithm is characterized by its lightweight nature, resulting in significant reductions in memory usage and processing time during deployment. This experiment has reached a dead end as no more models have been selected for training, and the observed performance of the chosen model appears to be satisfactory. | |
| **4.b. Suggestions / Recommendations** | Rank 1: Subsequently, it is imperative to evaluate the performance of this model with data that has not been before encountered.  Rank 2: Once the model demonstrates satisfactory results, the subsequent course of action involves deploying the model. | |