# **EXPERIMENT REPORT**

| **Student Name** | Trung Kien Hoang |
| --- | --- |
| **Project Name** | Machine Learning as a Service |
| **Date** | 06/10/2023 |
| **Deliverables** | <Experiment 2>  <SARIMA model with grid search>  <forecasting>  <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 project is to develop a forecasting model utilizing a time-series analysis technique to predict the aggregate sales income for all retailers and items during the upcoming 7-day period. The outcomes of this research will be utilized by an American shop to optimize their revenue generation by identifying days with high sales revenues and formulating an effective marketing strategy. In order to enhance sales performance during periods of anticipated low overall income, retailers may opt to implement discount promotions. The significance of obtaining precise outcomes lies in the possession of an optimal tool that has the potential to enhance one's income. However, the company would incur financial losses in its marketing efforts if the model being utilized is inaccurate. | |
| **1.b. Hypothesis** | A correlation exists between the total revenue generated from sales and the corresponding dates. It is worth consideration due to the observed upward trend in overall sales income over time. | |
| **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 all columns except date and revenue. 6. Removing outliers that total revenues is very small.   Data splitting:  Dataset is divided into training and validation datasets, with a ratio of 7:3 for the purpose of training and evaluating models. | |
| **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. | |
| **2.c. Modelling** | In this experiment, two machine learning methods, including ARIMA, SARIMA model with grid search and XGboost, were trained. The ideal p, d and q for the ARIMA is (1,1,1) while the optima, hyperparameter for XGboost: n\_estimators=100, learning\_rate=0.04 and max\_depth=2 In addition, the random\_state parameter is consistently assigned a value of 42 for all procedures, including data partitioning and data modeling. For the Seasonal Autoregressive Integrated Moving Average, it can be seen that there is a season for each every 7 days, so the parameter s is set to 7. | |

| 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** | The root mean square error (RMSE) values obtained for the validation data are 21485 for the ARIMA model, 13996 for the SARIMA model and 23164 for the XGboost model. This implies that the SARIMA model exhibits higher levels of predicting accuracy compared to the ARIMA and XGboost model. | |
| **3.b. Business Impact** | The findings obtained from this experiment can be utilized to provide empirical evidence in favor of the practice of predicting future sales revenues across all retail establishments. However, the presence of inaccurate findings has a significant impact on the efficacy of the firm sales strategy. | |
| **3.c. Encountered Issues** | Unsolved: The XGBoost algorithm was executed with 34 lag variables, resulting in a highly satisfactory performance outcome. Nevertheless, evaluating the model's performance with novel data is a challenge due to its reliance on lagged variables for predicting sales revenues. | |

| 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 accuracy of model forecasts decreases as the duration of time increases. However, this might be considered a dead end as there are no further models selected to be trained just with date as input. | |
| **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.  Rank 1: Once the model demonstrates satisfactory results, the subsequent course of action involves deploying the model. | |