# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Tahmidul Islam |
| **Project Name** | Machine Learning as a Service |
| **Date** | 16 October 2023 |
| **Deliverables** | Islam\_Tahmidul-24587139-predictive\_xgboost.ipynb   XGBoost  Islam\_Tahmidul-24587139-predictive\_prophet.ipynb  Prophet |

|  |  |
| --- | --- |
| 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 main target of this project was to upgrade an American retail company and increase its operational capacity and effectiveness by building strong and un-fragile predictive and forecasting models. By using these models, the project targeted to predict item-specific sales in individual stores for the next 7 days (because it had 10 stores across different states in the US). Using the models, the store would be better able to manage its supply chain, thereby doing better business. |
| **1.b. Hypothesis** | The Target is to predict sales revenue given an item, store and date. To achieve this, we plan to develop a predictive model by using historical data from different stores across several years. |
| **1.c. Experiment Objective** | The project is divided into two segments and thus has two main goals. The first is to predict sales revenue given an item, store and date. To achieve this, we plan to develop a predictive model by using historical data from different stores across several years. The model must be robust and accurate. This can help stakeholders improve the way they manage inventory. It will do so by accurately predicting the demands, thereby helping order the right amount of items and prevent over-stocking.  In the second segment, we build a forecasting model that forecasts total revenue across all the stores and items for the next 7 days. The forecasting would be done by a time-series analysis and provide the store (the primary stakeholder) to fine-tune its strategies by predicting demand and revenue trends. |

|  |  |
| --- | --- |
| 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 Preparation for Predictive Modelling:**  In order to prepare the data for predictive modeling, a series of preprocessing and feature engineering steps were implemented:  The date column served as the basis for engineering new features, including year\_month (e.g., 2011-01), day\_name (e.g., Monday), and year (e.g., 2011).  The sales\_revenue target column was generated by calculating the product of items\_sold and sell\_price.  To handle missing values in categorical columns, a new category, MISSING, was introduced. This strategy allowed the model to learn how to manage missing data effectively.  Categorical columns with a high number of distinct categories (item\_id, date, event\_name, year\_month) were subjected to mean target encoding, a technique suited for dealing with large category counts in each column.  Categorical columns with low cardinality (dept\_id, cat\_id, store\_id, state\_id, event\_type, day\_name, year) underwent one-hot encoding.  The categorical columns mentioned in the previous two points were subsequently employed for modeling.  In addition to these data preprocessing and feature engineering steps, the dataset was partitioned into distinct train, validation, and test sets. This division was based on sorting the data by the date column. The first 80% of the data constituted the train set, the next 10% was allocated to the validation set, and the remaining 10% formed the test set. This approach was adopted because the data followed a time-series structure, and it was essential to prevent any form of data leakage between the dataset splits.  **Data Preparation for Forecasting Modelling:**  In preparation for forecast modeling, the following steps were executed:  Within the combined dataset, a new target column, denoted as sales\_revenue, was created by multiplying the values in the items\_sold column with those in the sell\_price column.  Subsequently, the combined dataset underwent a grouping process based on the date column. This entailed the calculation of the total sales revenue across all stores for each specific date.  The resulting data frame designed for forecast modeling was structured with two key columns: date (renamed as ds) and total sales revenue (renamed as y).  Moreover, to facilitate the forecast modeling process, the dataset was divided into distinct train, validation, and test sets. This division was carried out following the sorting of the data by the date column. The initial 80% of the data comprised the train set, the subsequent 10% was earmarked for the validation set, and the final 10% was allocated to the test set. This approach was crucial in ensuring the integrity of the modeling process, especially given the time-series nature of the data. |
| **2.b. Feature Engineering** | To prepare the data, we initiated the process by merging the individual datasets into a unified dataset.  We performed a left join between the calendar data frame and the calendar\_events data frame, using the date column as the key, resulting in a combined data frame named merged\_calendar.  Following this we transformed the sales\_train data frame from a wide format to a long format, introducing a day\_num column that encompassed all unique day numbers, along with an items\_sold column representing the number of items sold.  Then, the sales\_train data frame in its long format was merged with the merged\_calendar data frame, using the day\_num column as the linking key, leading to the creation of a data frame known as sales\_train\_calendar.  Further, we incorporated the items\_weekly\_prices data into the sales\_train\_calendar data frame, thus forming the consolidated dataset.  Following the creation of the unified dataset, a comprehensive data analysis was conducted to identify and address any invalid or missing values. It was observed that only the event\_name and event\_type columns contained missing values, and it was logical because not every day, in a year of 365 days, would have an event. |
| **2.c. Modelling** | XGBoost Hyper Parameters used:   * n\_estimator or weak learners: 150   Any more than 150 would overfit the training data   * learning\_rate: 0.05   A slow learning rate was used to have proper convergence   * max\_depth: 5   A value less than 5 would miss out the complicated nuances in the data. Value more than 5 would overfit.   * min\_child\_weight: 3   Having few samples would reduce overfitting and give better partitioning.   * gamma: 3 * subsample: 0.7 * colsample\_bytree: 0.8  Prophet Prophet was chosen for Time-Series analysis for forecasting. It has the ability to model by incorporating several factors (holiday and special events, in this case) for which it was chosen over other Time-Series Algorithms. |

|  |  |
| --- | --- |
| 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** | From the RMSE and MAE values, we can see that the error is lower on the Test when compared to Train. However, it is overfitting when compared to the validation score.    From this result, it can be seen that Prophet is overfitting on the training data because of having high error scores on the test data. |
| **3.b. Business Impact** | The business can focus on these factors to improve on their business by improving them or investing more on them because they have high importance on sales. |
| **3.c. Encountered Issues** | The usage of user-defined libraries created problems in deploying the models. It recommended to not use user-defined libraries in future experiments. |

|  |  |
| --- | --- |
| 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** | Although the RMSE and MAE show mediocre results, this feature importance table can provide good business insights to the store. It shows that CA\_3, TX and HOUSEHOLD have the top 3 highest feature importance. This means that branch 3 in California plays an important role, branches in Texas are significant Team players and HOUSEHOLD category 2 is also important for the business. |
| **4.b. Suggestions / Recommendations** | While the predictive and forecasting models performed mediocre, it did provide certain insights to the business that can greatly improve their service long term.  As future works, the project may be enhanced by incorporating more factors like economy and environment, and the Machine Learning Models may be fine-tuned further for better performance. |