# **EXPERIMENT REPORT**

| **Student Name** | Tarun Gupta |
| --- | --- |
| **Project Name** | Part 1: Predictive Model |
| **Date** | 10/10/2023 |
| **Deliverables** | <Jupyter Notebook>  <Random Forest Regressor Model>  Git Link: https://github.com/tarungupta293/Sales-Revenue-Prediction |

| 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?  Answer: The goal of this project is to predict the sales revenue on the basis of given dataset. The prediction is in two parts   * 1. Predict the sales revenue for any store id, item id and date.   2. Forecast the sales revenue of all stores for the next 7 days. |
| **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,  Answer: Considering the dataset, the questions arrived in my mind is what are columns named d\_1 to d\_1541 in the train dataset. Also, in the calendar dataset, what represents the feature named ‘wm\_yr\_wk’. |
| **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.  Answer: We need to perform model which can predict the sales revenue. Model should be highly accurate to meet with the business team requirement. Both the models should be able to predict the accurate sales revenue on the basis of input given. |

| 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** | 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  Answer: Steps taken in Data Preparation  Merging the datasets in chunks.  Filter the dataset and remove the records having dates greater than the training dataset dates. |
| **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  Answer:  Label encoding on the Categorical dataset such as store\_id and item\_id  Convert the datetime into date, month and year for the model. |
| **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  Answer: I have performed multiple models on the dataset. All the models comes with the range of accuracies. In all of the models, I have finalised the **Random Forest Regressor** model as my final model. The model is able to achieve the accuracy of 0.46. Also, I have counted the MSE to detect the error in the model and the value was 51.14 which was comparatively very low as compared to other models. |

| 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** | Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.  Answer: The performance of the model was quite satisfactory as compared to the other models performed with the dataset. The Random Forest model comes with the accuracy of 0.46 with the MSE value as 51.14. |
| **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)  Answer: The model is able to predict the sales revenue very accurately on the input parameters provided by the business team. |
| **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.  Answer: I have faced several issues while merging the datasets. As the dataset is much bigger, so while merging it leads to the out of memory error. I had to split the dataset into chunks and run each chunk of data for merging manually. |

| 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** | 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.  Answer: While performing the experiment, I have gained the new insights regarding the data merging in chunks and how to decide the best model from several performed model on the basis of its accuacy and performance. |
| **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.  Answer: In the next step, I would perform the model with more preprocessing of data including any way to overcome outliers. This time, I have removed the outliers, but leads to the low accuracy of model. I will try more way to overcome the outliers which can improve the accuracy of model. |