Management

FORM FOR THE SUBMISSION OF ASSESSED COURSEWORK

THIS FORM MUST BE COMPLETED AND ATTACHED TO **ALL** ASSESSED WORK.

Student Number (7 digit number)	**2**	**3**	**6**	**6**	**0**	**3**	**6**		
Username (2 letters & 5 digit number ab12345)	**u**	**e**	**2**	**3**	**3**	**0**	**3**		
Unit Code	**M**	**G**	**R**	**C**	**M**	**0**	**0**	**2**	**1**
Unit Title	**Optimisation & Algorithms**								

By submitting this work online using my unique log-in and password, I hereby confirm that this work is entirely my own. I understand that all marks are provisional until ratified by the Faculty Examination Board.

Please remember to save **your work** either as a **Word document (.doc or** as a **PDF (.pdf)** unless another format is specified by the Unit Director with the filename [Unit Code_Student Number] **e.g. MGRCM0000_1234567.docx**.

Please start writing your coursework on the next page.

Predicting the Sales of Rossmann Drug Store by Using Gradient Descent Optimisation Algorithm

Introduction

Optimisation algorithms are now well applied in many industries, we can make production more efficient, utilise limited resources or make predictions. For instance, logistic companies may use Deterministic Dynamic Programming (Shortest Route) to find an efficient path and the companies could save much money and human capital; product companies may conduct Unconstrained Nonlinear Programming (Bisection or Gradient Descent Method) to solve optimisation problems like predicting sales, inventory management, and supply chain management.

My study is mainly to solve the optimisation problem of predicting sales for a German drug company, Rossmann, and I used the dataset containing 1,115 stores' information from the first of January 2013 to the thirty-first of July 2015. To solve this optimisation problem, I used the Gradient Descent Method, Adaptive Moment Estimation (Adam), to train the model, and I drew a graph to see its training process like decreasing the loss, and then made a prediction. Apart from predicting the sales of Rossmann, I would do exploratory data analysis (EDA) to see if there is valuable information, and then I would make some interpretations based on my observation. The management of Rossmann could change their strategies to manage inventory and maximise profit.

Literature Review

Varshney, R.P. and Sharma, D.K. (2024) introduce a novel stacked deep learning framework that integrates bi-directional long-short-term memory networks, 1D convolutional layers, and 1D max-pooling layers within each stack. Additionally, they present an enhanced variant of adaptive moment estimation (Adam) along with an error correction technique aimed at enhancing both the accuracy and convergence speed of the model (time-series forecasting). Arouri, Y. and Sayyafzadeh, M. (2022) suggested employing the adaptive moment estimation (Adam), a first-order gradient-based framework, in combination with a stochastic gradient approximation, to address optimization problems related to well location and trajectory (maximisation of oil production). Li, X. et al. (2023) introduced a novel algorithm called

Double Total Variation (DTV) aimed at mitigating the staircase effect by enhancing the variational information within the two diagonal subbands (fast MR image reconstruction). To optimize the solution algorithm, they leveraged the Improved Adaptive Moment Estimation (IADAM) method in conjunction with the conjugate gradient algorithm. He, J. and Peng, F. (2022) introduced the Adaptive Moment Estimation (Adam) strategy, which effectively utilizes historical data incrementally on an asset portfolio (investment). This approach enhances both performance robustness and maintains linear time complexity. Adam demonstrates strong performance across various evaluation metrics while also being able to sustain reasonable transaction costs. Liu, X. et al. (2019) introduced AdmOAM, an algorithm named Adaptive Moment Estimation (Adam) for Online AUC (Area under ROC curve) Maximization. This method leverages gradient moment estimation to expedite convergence and alleviate the rapid decay of learning rates. We establish the regret bound of the proposed algorithm and conduct extensive experiments to showcase its effectiveness and efficiency. Gayathri Devi, K. et al. (2022) predicted and classified corn leaf disease using Adaptive Moment Estimation (Adam) optimiser in Deep Learning Networks. Their proposed approach achieved an accuracy of 99.43% using an optimal learning rate of 0.0001. This work has potential applications across various agricultural sectors, facilitating the implementation of automated systems such as robotic pesticide sprayers and drone-operated systems. Nhat-Duc Hoang (2020) introduced a computer vision model designed to automatically identify localized spall objects present on the surfaces of reinforced concrete elements. They applied a logistic regression model, trained using the state-of-the-art Adaptive Moment Estimation (Adam), which is employed to establish a decision boundary for predicting the status of "nonlocalized spall" and "localized spall." Experimental results illustrate that the newly developed model achieves high detection accuracy, with a classification accuracy rate of 85.32%, precision of 0.86, recall of 0.79, negative predictive value of 0.85, and F1 score of 0.82. Therefore, the proposed computer vision model can serve as a valuable tool to aid decision-makers in the periodic survey of structural health conditions.

Multiple studies are focusing on optimisation with Adam optimiser. For instance, Adam optimiser could be used on time-series forecasting, well placement optimisation (oil production), MR image reconstruction, asset portfolio optimisation, area under ROC curve optimisation, predicting corn leaf disease and spall object detection, etc. Thus, Adam optimiser is a means that is widely used and could applied to multiple industries for optimisation problems. Due to Adam's advantages of adjusting the learning rate and its momentum-like behaviour, I will implement it in my study on predicting sales for Rossmann drugstore.

Methodology and Data

Data Collection and Cleaning

This study uses the datasets from Kaggle (See Appendix). The datasets contain information about a German drug store, Rossmann, from the first of January 2013 to the thirty-first of July 2015. There are three datasets: ‘store.csv’, ‘train.csv’ and ‘test.csv’. I will use ‘store.csv’ and ‘train.csv’ to analyse data since I will focus on using an optimisation algorithm to get the best result or the lowest loss.

The ‘store.csv’ contains information like ‘Store’, ‘StoreType’, ‘Assortment’, ‘CompetitionDistance’, and ‘Promo’ (Promotion), there is content talking about the competition, the date the store opened, the promotion, etc. The ‘train.csv’ contains information like ‘Sales’, ‘Customers’, and some holidays (Is it open on certain holidays), etc.

The whole dataset contains 1,115 stores and 1,017,209 transactions; categorical data like ‘StoreType’ and ‘Assortment’ exist. First, I merged the dataset ‘train.csv’ and ‘store.csv’ using the LeftJoin function to get all data into one dataset. Second, I tried to figure out the duration of competition and promotion, to do so, I needed to transform the ‘Date’ from ‘object’ to ‘data time’, and then I could calculate the period of competition and promotion. If the value of the addressed data contains a negative number or null value, I will replace them with 0. Finally, I checked if there was a null value on each column, I found out there was a null value on ‘CompetitionDistance’, so I needed to add values to them (There are other columns that contain null values, but I will not use the columns, so I did not tackle them). The average sales of ‘CompetitonDistance’ with null values is 4,536; the mean of sales is 5,774, the first quantile is 3,727, the second quantile is 5,744, the third quantile is 7,856; the mean of ‘CompetitionDistance’ is 5,430, the first quantile is 7,100, the second quantile is 2,330, the third quantile is 6,890. By observing the above information, I choose the 35% quantile to fit into the null values of the ‘CompetitionDistance’ column.

Exploratory Data Analysis

First, I observed that most sales happened at the beginning, medium and late of the month, which means the drug stores need to have enough inventory at that time (Figure 1).

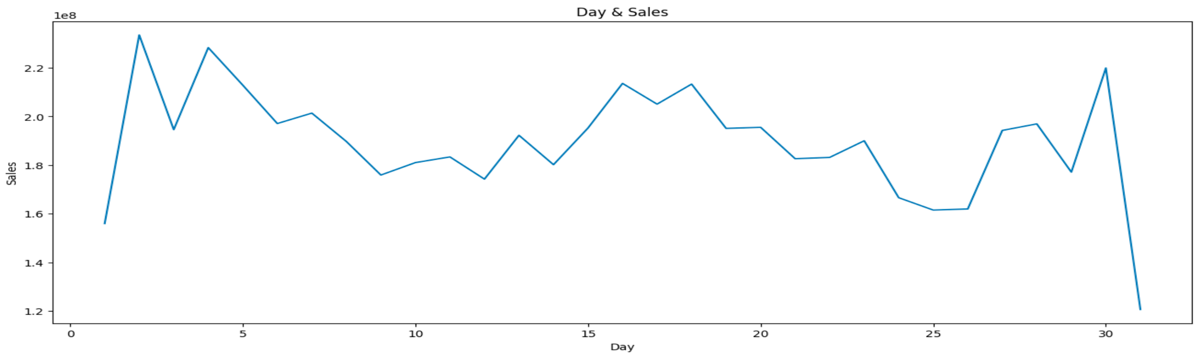

Figure 1: The Day of the month versus the sales

Second, I observed that the peak of sales is from January to July, and there is an upturn in December, maybe it's because there are more holidays or people tend to get cold at that time (Figure 2). The demand for drugs is relatively low from August to November if the production line needs to be repaired annually, then it's better to do so at the time.

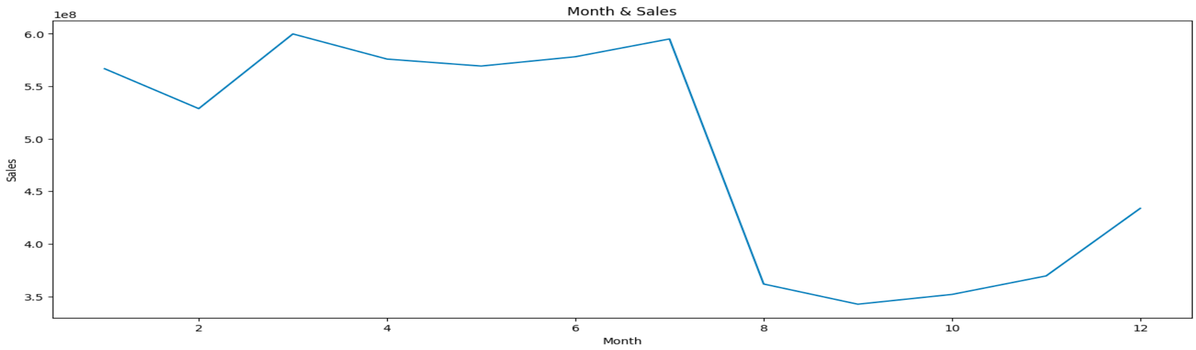

Figure 2: The Month of the year versus the sales

Third, StoreType A is the most successful type on the market, and StoreType B suffers from the lowest sales (Figure 3). When there is limited inventory to distribute, the headquarters needs to fulfil StoreType A first, then D, C, and B.

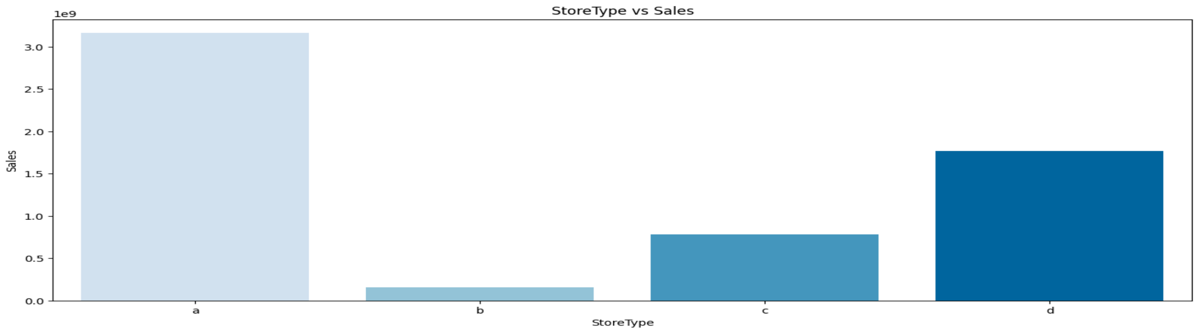

Figure 3: StoreType versus Sales

Fourth, Promotion is more efficient than Promotion 2 (Figure 4&5). Thus, maybe the drugstore can save some money by stopping Promotion 2 and then using the money on Promotion.

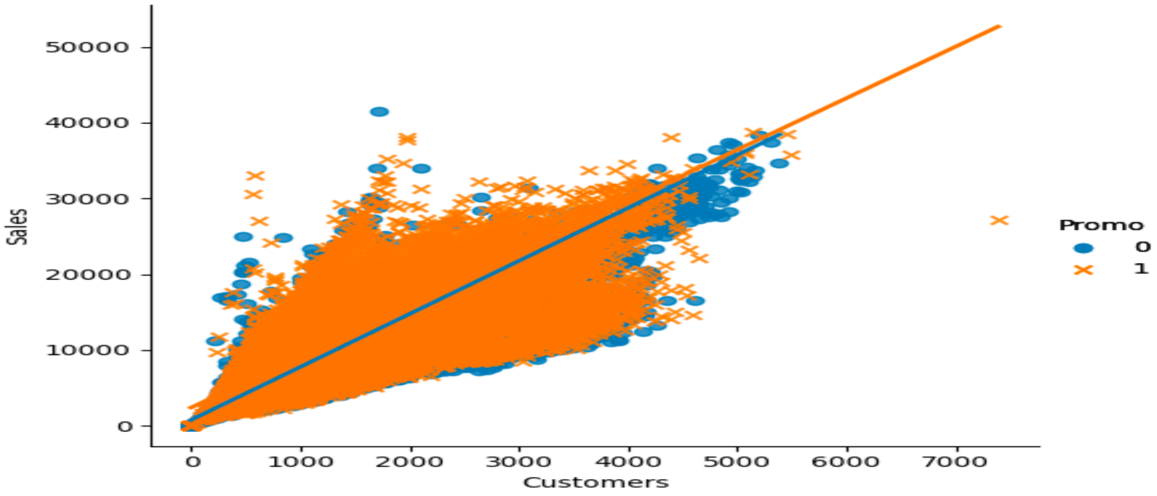

Figure 4: Customers versus Sales on Promotion (0: no promotion, 1: promotion)

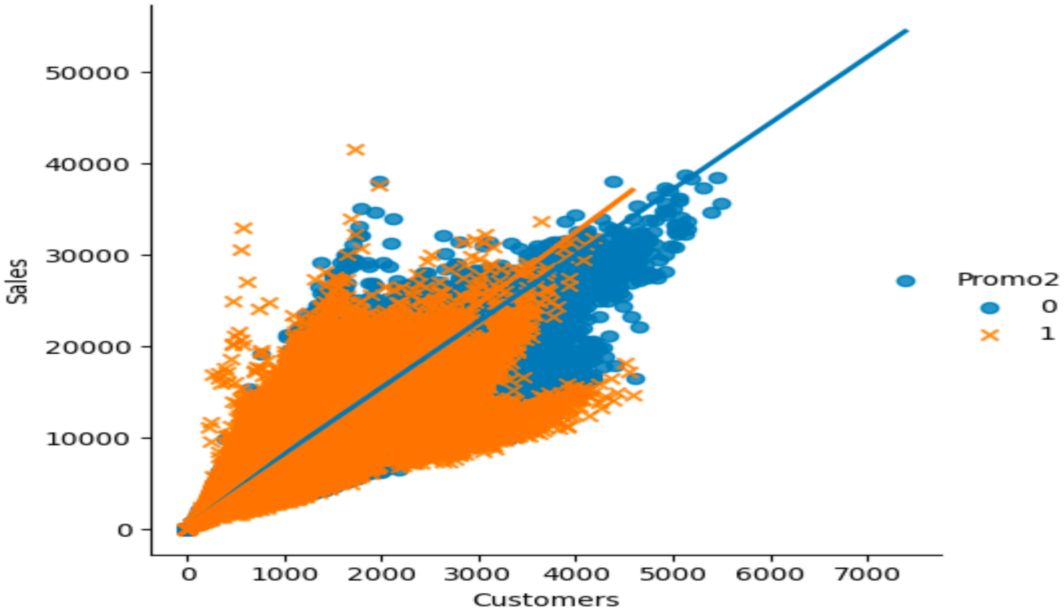

Figure 5: Customers versus Sales on Promotion 2 (0: no promotion, 1: promotion)

Fifth, the sale of Assortments A and C is significantly higher than Assortment B (Figure 6). Assortments A and C need to prepare much more inventory for the potential sale; furthermore, Rossmann can consider if Assortment B should be closed.

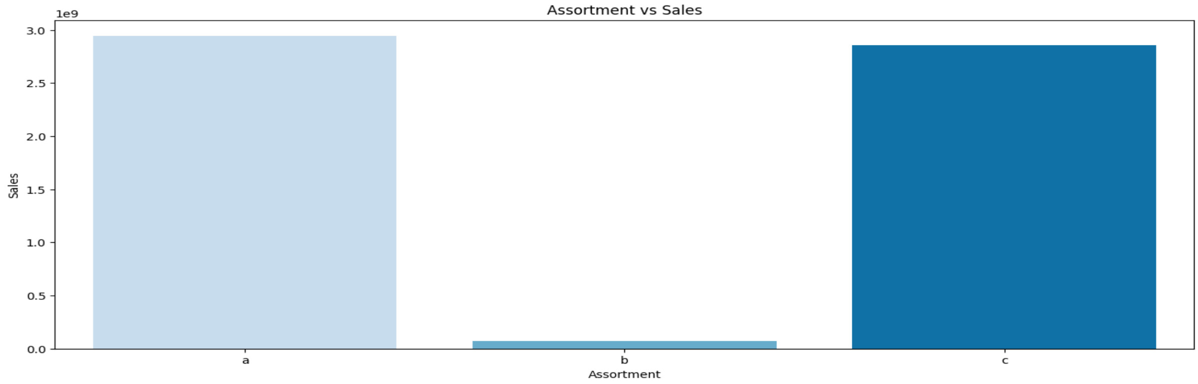

Figure 6: Assortment versus Sales

At last, two factors were calculated from raw data, 'CompetitionTime' (The time the drug store has confronted competition) and 'Promo2Time' (The time the drug store has used promotion 2). Figure 7 shows that 'CompetitionTime' affected sales, but the effect became small and stable after 200 months. Figure 8 shows that the effect became insignificant as time passed. When drug stores face competition in the first place, the store managers should use Promo to boost their sales. The sales are higher without Promo2, and the effect of Promo2 would decrease as time passes, maybe the store managers could abandon this promotion.

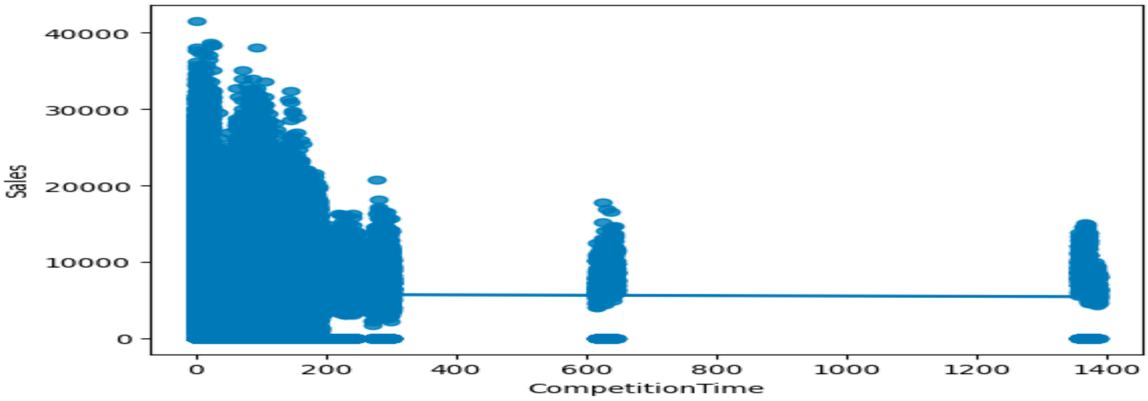

Figure 7: CompetitionTime (Month) versus Sales

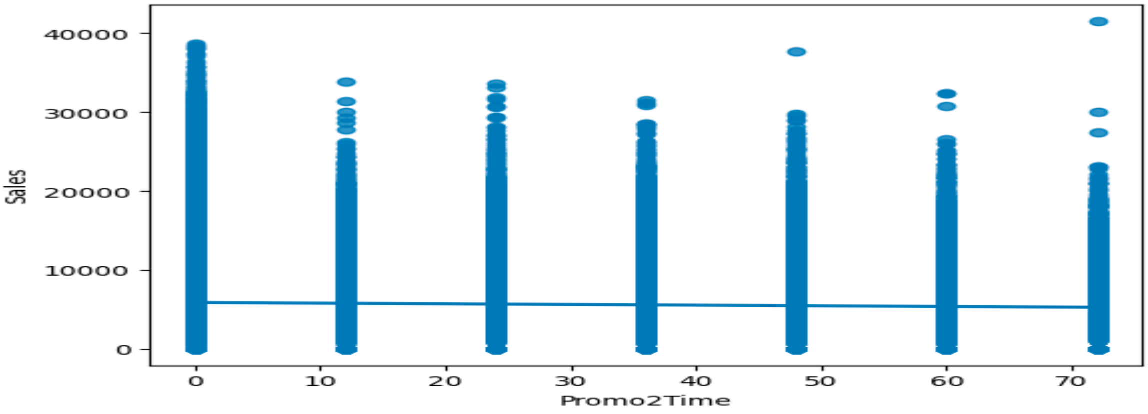

Figure 8: PromoTime (Month) versus Sales

Adaptive Moment Estimation (Adam)

Adaptive Moment Estimation (Adam) is a popular optimization algorithm used for training deep learning models, particularly in the context of neural networks. It combines ideas from two other popular optimization algorithms: RMSprop (Root Mean Square Propagation) and Momentum.

The benefits of Adam:

- Adam adapts the learning rates for each parameter based on the magnitude of the gradients and their moving averages. This can lead to faster convergence and better performance.
- Adam is less sensitive to the choice of hyperparameters compared to other optimization algorithms, making it easier to use in practice.
- By maintaining momentum-like behaviour through the first moment term, Adam can navigate through flat regions and escape saddle points more effectively.

The algorithm computes the first-moment estimate m_t and the second-moment estimate v_t of the gradients as follows:

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_1) \cdot g_t$$

These estimates are biased towards zero at the beginning, so Adam corrects this bias by computing bias-corrected estimates:

$$\widehat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

Finally, Adam updates the parameters using the bias-corrected estimates and a small constant ϵ to prevent division by zero:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \cdot \widehat{m}_t$$

In the formula:

- t is the current time step.
- m_t and v_t are the first and second-moment estimates of the gradients respectively.

- β_1 and β_2 are the exponential decay rates for the moment estimates.
- η is the learning rate.
- ϵ is a small constant to prevent division by zero.

Adaptive Moment Estimation (Adam) combines these moment estimates to adaptively adjust the learning rates for each parameter during training, leading to faster convergence and better performance.

Optimisation and Results

To do optimisation, first, I scaled the numeric columns ‘Store’, ‘DayOfWeek’, ’ ‘Promo’, ‘Schoolholiday’, ‘CompetitionDistance’, ‘Promo2’, ‘CompetitionTime’ and ‘Promo2Time’. I choose MinMaxScaler from Sklearn to do normalisation on these values. After normalisation, the numeric values would be on a scale of 0 to 1. Second, I encode the categorical columns ‘StateHoliday’, ‘StoreType’ and ‘Assortment’ using OneHotEncoder from Sklearn to get numeric values. Third, I would generate training and testing datasets with an 80:20 ratio. Subsequently, we'll utilise the training dataset to create training and validation datasets at a ratio of 90:10. Finally, I applied a neural network model Sequential from Keras; the hidden layer contains 150 neurons and the activation function is ReLU (rectified linear unit); the output layer contains one neuron and the activation function is linear, then I configured the model with Adam optimiser for training and the loss function is set to mean absolute error (MAE). During the training process, the model iterates through the training dataset for 10 epochs, processing the data in batches of 64. After each epoch, the model's performance is evaluated using the validation data to assess its generalization capability and detect potential overfitting.

The result showed that MAE is 1,697 on the training dataset and 1,638 on the testing dataset. Figure 9 shows that the lowest MAE is around 1,650 in this model, and then I tried to get a better result by adding epochs. However, when I tried to increase the epochs to 20 from 10, there was just a little improvement in MAE of 1,643 (See Appendix), maybe there are other ways like different activation functions or different algorithms to tackle this problem. Table 1 shows the predicted and real sales values. The result tells us that the model still predicted sales even when I put the column ‘Open’ into consideration, the ‘Open’ column tells us that if the store is still in operation, the model seems not to get this information well. Besides, the predicted values for operating stores are very similar (around 7,000), which means that the model can not tell which factors are most important for the sales, maybe it’s because I used

ReLU as a hidden-layer function and a linear function as the output layer, or I should increase the complexity of the architecture to make model absorb more information.

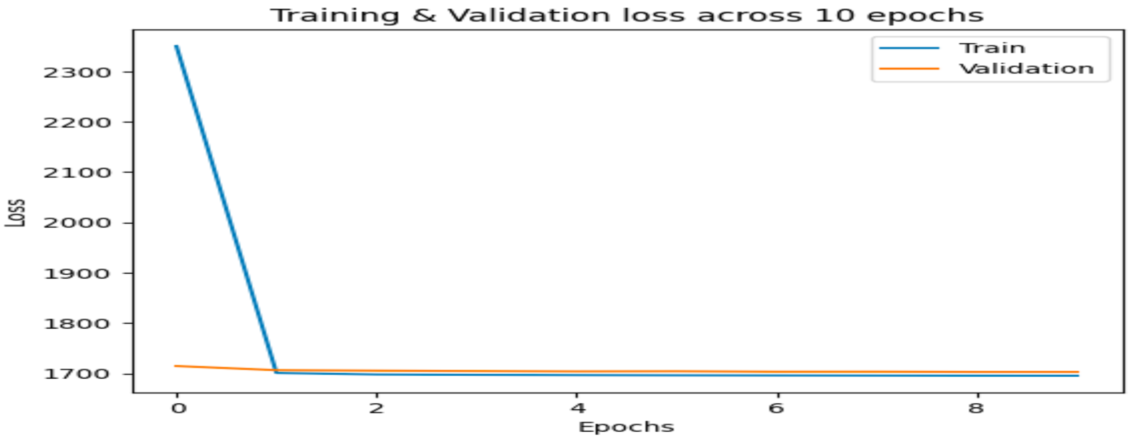

Figure 9: Training and Validation Loss across 10 epochs

Table 1: Predicted Sales versus Real Sales

Predicted Sales	Real Sales
7659	5263
7325	6064
7154	8314
7797	13995
7273	4822
…	…
84	0
222	0

226	0
226	0
94	0

Conclusion and Limitations

My study is to predict the sales of Rossmann drugstore, and I used the Adaptive Moment Estimation (Adam) Gradient Descent Optimisation to make predictions. After training the model, I got the result of 1,697 mean absolute error (MAE) from 3,401 MAE. I found out that there is a chance to lower the MAE if I add columns for training, for instance, I did not use the column ‘DayOfWeek’ for training as I thought it was not that important for predicting the sales; however, the MAE without ‘DayOfWeek’ is above 2,000, which is much higher than the result encompass ‘DayOfWeek’. Thus, if I can divide existing columns into value factors, I may get a much lower MAE after training. I got relatively consistent performance on the testing dataset, the MAE of the testing dataset is 1,638, which is close to the training dataset. My work demonstrated how to optimise predicting problems by using Adam, and there is much room for improvement, maybe people could try to use different parameters to train the model. Besides, the EDA part could give the Rossmann management some insights, they could upgrade their operation based on my research.

The limitation of the model is that there are too many factors into consideration, I could find out certain relationships between a single factor versus sales like ‘StoreType’ or ‘Assortment’; however, when I trained the models, I would like to implement as much as information into training, since the outcome with few factors showed higher MAE, but there is a limit to lower MAE. Thus, to optimise the predicting problems with many factors, there may be other algorithms that are better than Gradient Descent, or I should make more effort to clean data and try to exact the most valuable information. On the other hand, I used a two-layer model, and the output layer is a linear function, it’s possible that I would get better results when I add more layers or the number of neurons.

References

Varshney, R.P. and Sharma, D.K. (2024) “Optimizing Time-Series forecasting using stacked deep learning framework with enhanced adaptive moment estimation and error correction,” Expert Systems With Applications, 249. Available at: https://doi.org/10.1016/j.eswa.2024.123487.

Arouri, Y. and Sayyafzadeh, M. (2022) “An adaptive moment estimation framework for well placement optimization,” Computational Geosciences: Modeling, Simulation and Data Analysis, 26(4), pp. 957–973. Available at: https://doi.org/10.1007/s10596-022-10135-9.

Li, X. et al. (2023) “Double total variation (DTV) regularization and Improved adaptive moment estimation (IADAM) optimization method for fast MR image reconstruction,” Computer Methods and Programs in Biomedicine, 233. Available at: https://doi.org/10.1016/j.cmpb.2023.107463.

He, J. and Peng, F. (2022) “Adaptive moment estimation for universal portfolio selection strategy,” Optimization and Engineering: International Multidisciplinary Journal to Promote Optimization Theory & Applications in Engineering Sciences, 24(4), pp. 2357–2385. Available at: https://doi.org/10.1007/s11081-022-09776-7.

Liu, X. et al. (2019) “An Adaptive Moment estimation method for online AUC maximization,” PloS one, 14(4), p. e0215426. Available at: https://doi.org/10.1371/journal.pone.0215426.

Gayathri Devi, K. et al. (2022) “Accurate Prediction and Classification of Corn Leaf Disease Using Adaptive Moment Estimation Optimizer in Deep Learning Networks,” Journal of Electrical Engineering & Technology, 18(1), pp. 637–649. Available at: https://doi.org/10.1007/s42835-022-01205-0.

Nhat-Duc Hoang (2020) “Image Processing-Based Spall Object Detection Using Gabor Filter, Texture Analysis, and Adaptive Moment Estimation (Adam) Optimized Logistic Regression Models,” Advances in Civil Engineering, 2020. Available at: https://doi.org/10.1155/2020/8829715.

Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Appendix

Dataset: https://www.kaggle.com/competitions/rossmann-store-sales

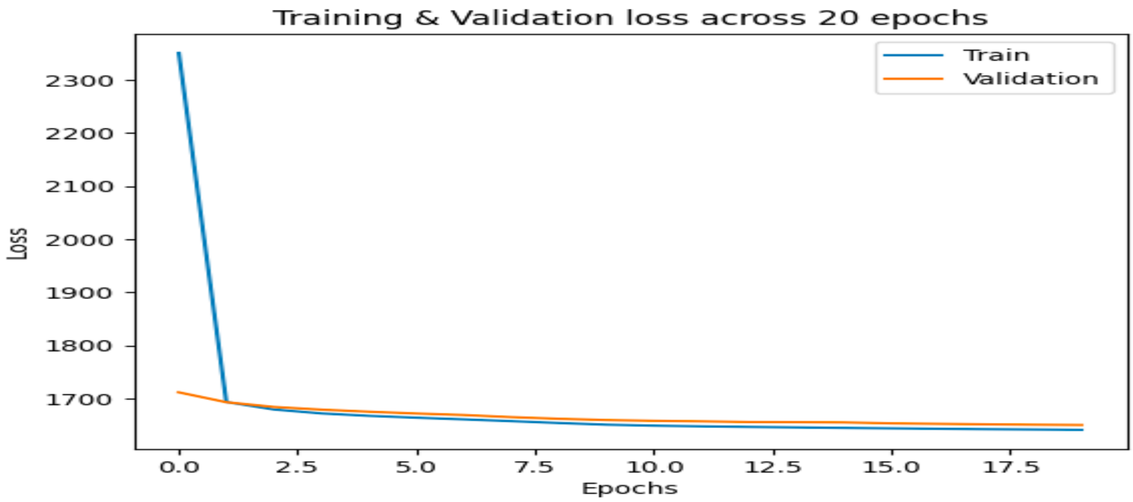

Figure 10: Training and Validation loss across 20 epochs