Store Item Demand Forecasting Using Deep Learning Approach

Safwan Muntasir
180104084@aust.edu
Dept. of CSE,
AUST

Mashfiq Rahman
180104087@aust.edu
Dept. of CSE,
AUST

Arifur Rahman Jawad
180104097@aust.edu
Dept. of CSE,
AUST

Abdullah Al Mohaimen
180104098@aust.edu
Dept. of CSE,
AUST

Ashfakur Fahim
160204032@aust.edu
Dept. of CSE,
AUST

***Abstract*—Demand forecasting means analysing present market value of some product and predict it's demand in future after studying pertinent data of the products. It is very important in making profit, specially in business fields, in an efficient way where waste will be minimized and profit will be maximized. Researchers of Today have accomplished fantastic results in time series problems like future demand forecasting using Deep Learning Models like MLP, CNN, LSTM, GRU, BiLSTM, BiGRU as well as combination of some deep learning models like CNN-LSTM, CNN-BiLSTM and CNN-BiGRU. We haved used these deep learning models and tried to accomplish less error rate in predicting future sales of some specific product. Among these models, MLP performed best for our research purpose and showed the minimum error rate of almost 19%.**

I. Introduction

In a business the department responsible for maintaining supply and demand of items in market is management. Management is responsible for business flow and continuity of a product. Analysing the market and sales A kind of forecast is used as a basis to meet the needs. The outcome is more reliable as much as the forecast is accurate. Forecasting helped management for decades, but now with the advancement in computers and AI, forecasting methods that were previously impossible to explore are currently achievable.

The basic goal is to have the least amount of inventory to satisfy the customer's demands and at the same time minimizing the cost of buying and reserving the stock. A grocery shop with an oversupply in inventory causes increase in cost for storage, stock deterioration and expiration of the items. With an under-supply, results in sales failure. Reliable forecasts are essential for a shop to maintain healthy business. So, both over and under supply is a problem that might seem small but plays a vital rule that decreases unseen loss and increases easily gain able profit. A new term recently used is demand forecasting. This demand is the demand of customers who wants a product for immediate use.

Due to many external factors is it hard to build a forecasting model. Some of these factors are store's location, time of the year, change in economy and so on. All these will form a huge data that no human could examine to perform a prediction. But, it is very much possible for a computer and an AI model to process this amount of data and try prediction or forecasting of demand.

We have found out some work has been done on this field earlier so there ware varieties of dataset that we were able to use. We pinpointed some features from a dataset for fulfilling our own data requirement.Then preprocessing the data we made it ready to run through some models. As our work progressed after running some models such as MLP, CNN, LSTM, GRU etc. for our dataset, we have observed performance of MLP was better than other deep learning models. This result was judged on the verdict of their Mean Squared Error (MSE).

II. Related Works

In paper[1] we have found that the main purpose was to forecasting the sale amount of a choen item for a particular week and a distribution warehouse. The method consists of 4 main stages:

(i)Firstly, data was obtained from retailers.

(ii)Moving average value was calculated for every product.

(iii)As group warehouses and sub-distribution has similar sales behaviour so a bipartite clustering algorithm was applied..

(iv)Applying Bayesian Network to get forecasting results.

For generating forecast results the information were taken from a company fromm turkey. It was a national level dried nut and fruit company that provided the sales report of 2011-13. The dataset contains-

• Warehouse related attributes: size, number, location, transport vehicle, weekly sale report, area, employee amd consumer number.

• Product related attributes: category, amount, time.

Error rate was 49% after first trial when Hybrid model with MAPE was used for all warhouses. But in the next trial main warehouses clustered for their sale behaviour. The error rate was 24% which decreased considerately for this method as every main warhouse cluster had same model applied.

For paper[2] forecasting sale is difficult for retail grocery. Here, the complex area of cost for data preparation versus forecasting accuracy to detect efficient model. Benchmark were used to evaluate models. SKUs were introduced to do so. It also supported vector regression. The dataset designed was for one product with four subcategories. And stores were also differentiated for size. Over a span of 76 weeks, the data were divided for training and testing the model. All model got better results considering their benchmarks but the regression tree outperformed with an improvement of 65.17% forecast accuracy.

In paper[3], it is the reinforcement of two research ideas. First, a time series model empowered by machine learning is developed for forecasting demand. Secondly, this research shows how important it is to highlight the magnitude of improvement in performance by employing forecasting method.

Data were obtained from sales of a multinational steel manufacturing company. Though it operated in four different segments , only one was used to collect data. Two traditional forecasting Holt-winter and Damped trend were currently used by corporate. Holt-winter gave an accuracy of 84.1% while damped trend gave 81.3%. later, the dataset were divided for test, train and prediction was for ARIMAX and ANN models. For 2017, ARIMAX got forecast accuracy of 88.9% when ANN was slighlty better with a result of 89.4%. All these test supoorted their hypothesis of supply chain efficiency between traditional and ML-based forecasting models.

In paper[4] they proposed the method of storage management by trying to predict the demand of product sale and service. They used XGBoost regression model that used decision tree to perform predictions.

Amazon Simple Storage Service (Known as S3) is an interface based on web interfaces It is specially developed for the developers and to make web-interface digitized computing easier for them. It gives industrialist flexibility about leading, availability of the data, improve their security and enhance their performance. This technology is invented for industrialists and customers both to protect all kind of knowledge.

In this paper, they obtained the following RMSE values based on their dataset and predicted outputs for every week. The next week's RMSE value is increased because the previous week's RMSE values are appended to it and it made the train data large also.

Week	RMSC
3	0.6758
4	0.6931
5	0.6644
6	0.6824
7	0.7025

Table: RMSE values for each week extrapolation

The author of article[5] said, about the Industry 4.0 (Bartodziej, 2017) which is a new invented model by the researchers that describes and analyses the 4h Industrial Revolution. It overcome the radical change in the previous three industries, represented by intelligent, virtual, and digital performance specially in giant industries which are improved in advanced technologies across their production and distribution. This Industry 4.0 allows giant industries to analyze and upgrade their management system in the advance level which makes them expert in predicting the future sales. Therefore it reduces possibility of product shortage as well as minimize financial cost in certain compounds.

The main aim is to make industrial warehouse management more optimal and easier with the minimization of total financial cost section also in the areas of storing and transportation of the products directly from their suppliers It also helps the industries to relieve from product shortage.

The main aim is to make industrial warehouse management more optimal and easier with the minimization of total financial cost section also in the areas of storing and transportation of the products directly from their suppliers It also helps the industries to relieve from product shortage.

III. Background Study

A. Recurrent Neural Networks (RNN)

Recurrent Neural Network are neural networks that focuses on processing a sequences of data x(t)= x(1),.... x(t) with the time step index t ranging from 1 to t.RNN differs when compared to other neural networks because it uses feedback loops. These loops help RNN in processing the data sequence. This loops enables the data to be distributed among various nodes and generate predictions based on the gathered data. This process can be referred as memory.

B. Long Short-Term Memory(LSTM) The Long Short Term Memory [6] Network is a complicated Recurrent Neural Network (RNN) , that is, a sequential network which permits information persistence.

Simple LSTM has 3 cell that works by the following direction.

The first section determines whether the information from the previous timestamp needs to be remembered or ignored. The cell attempts to learn new information from the input to this cell in the second section. The cell finally transmits the updated data from the current timestamp to the following timestamp in the third section.

An LSTM has a hidden state, just like a straightforward RNN. Additionally, LSTMs have a cell state for the prior and current timestamps.The hidden state is known as Short term memory and the cell state is known as Long term memory.

C. Convolutional Neural Network(CNN)

CNN is a deep learning algorithm that mostly takes image as an input and then process it to differentiate and find variations from others. Though it extracts raw data from images, however the extraction is unnecessary for time series analysis because of its numerical pattern.

CNN has the ability to increase the performance as it can run model twice faster than others. Its weight division helps it to reduce its number of parameters to increase efficiency of model learning. It is more suitable for forecasting time series as it offers dilated convolutions.

D. Bidirectional LSTM (BiLSTM)

BiLSTM is short for bidirectional LSTM, meaning that the signal can propagate both forward and backward in time. BiLSTMs effectively increase the amount of information available to the network, improving the context available to the algorithm. In both directions of the sequence, it is a potent tool for modeling the sequential relationships between words and phrases. In our project accurate prediction is difficult due to the time series' high frequency and high noise levels. BiLSTM is applied in financial time series. The outcomes demonstrate that the BiLSTM model, which can fully capture the past and future data information concurrently, has the highest prediction accuracy.

Bidirectional LSTMs train two LSTMs on the input sequence as opposed to only one. he first on an original copy of the input sequence, and the second on a reversed copy. This can give the network extra context and lead to a quicker and even more thorough learning process for the problem.

E. Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU) is a type of Recurrent Neural Network (RNN) that, in certain cases, has advantages over long short term memory (LSTM). GRU uses less memory and is faster than LSTM. GRUs solve this problem through the use of two gates, the update gate and reset gate. A GRU is a very helpful method for recurrent neural networks to address the vanishing gradient issue. The vanishing gradient problem arises when the gradient gets so small that the weight cannot change its value. When working with smaller datasets, they perform better than LSTM.

F. Bidirectional Gated Recurrent Unit (BiGRU)

The sequence processing paradigm known as a bidirectional GRU, or BiGRU, consists of two GRUs. one processing the information forward and the other processing it backward. Only the input and forget gates of a bidirectional recurrent neural network are present.

the bidirectional GRU is a special form of the bidirectional RNN that divides the regular GRU into two directions: a forward direction associated with historical data and a reverse direction associated with future data. This allows for simultaneous use of the input historical data and future data. This allows the information from both future and past to impact the current states. The bi-GRU is defined as follows:

$$\overrightarrow{h_t}$$
$$= \text{G R L}_{fwd}\left(x_t, \overrightarrow{h_{t-1}}\right)$$

$$\overleftarrow{h_t}$$
$$= \text{G R U}_{bwd}\left(x_{i,}\ \overleftarrow{h_{f+1}}\right)$$
$$h_t = h_t\ \overleftarrow{h_f}$$

where →ht→ is the state of the forward GRU, ←ht← is the state of the backward GRU, indicates the operation of concatenating two vectors.

G. Mean Squared Error(MSE)

The mean squared error tells us how close a regression line is to a set of points. It is done by taking the distances from the points to the regression line (these distances are the "errors") and squaring them.To remove any negative signs the squaring is necessary. It gives more weight to larger differences. It's called the mean squared error as we are finding the average of a set of errors. The forecast gets better with lower value of MSE.

MSE formula = (1/n) * E(actual – forecast)2
Where,
n = number of items,
E = summation notation,
Actual = observed y-value,
Forecast = y-value from regression.

IV. Dataset

A. Dataset Colection

We have collected our dataset from Kaggle[7]. The objective of our research is to predict sale data of a store for the month for some items. These data are separately stored in different files for convenience.

The dataset are divided into six paths. The sales dataset has five features. The features are date_block_num, shop_id, item_id, item_price, item_cnt_day. The features are represented with integer numbers. There are 10,48,576 data in the sales dataset.

Item_categories dataset represents the id of all the items. There are 83 categoris in this dataset. In items dataset, it contains 22,172 items. Sample_submission dataset represents count of items per month and it contains 2,14,199 data. Shops dataset represents 59 shop id and shop names.

The train datas are tested with test dataset, It has three features: id, shop_id and item.

B. Dataset Preprocessing

A raw dataset is processed for being used for another purpose is said to be processing of data. It is a crucially important for performance and accuracy of the result. The

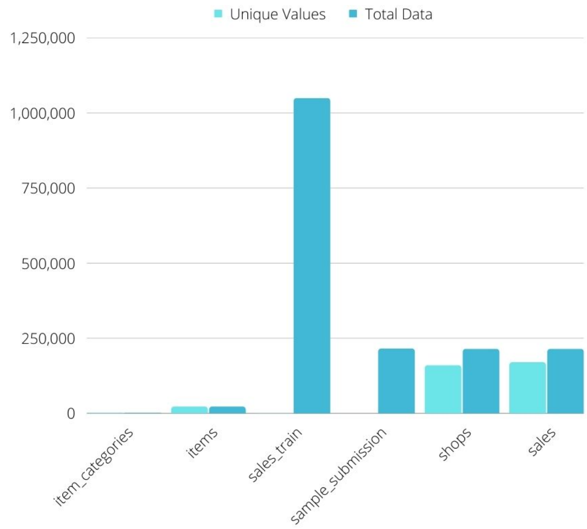

Fig. 1. Statistics of the dataset

processed data are more effective in ensuring better results and saves storage and time.

Handling missing data: NaN values has been handled in our dataset prepossesing. If our dataset contain some missing data, then it may create huge problem for our deep learning models. Hence it is necessary to handle missing values in our dataset. The missing data are replaced with zeros.

Formating Data: The feature 'Date' in the sales dataset was not in the date, month, year seperately. Pandas to_datetime converted the 'Date' into date, month, year format. Then base on the date_block_num and item_cnt_day a pivot_table named 'dataset' is created with shop_id and item_id. Then the test dataset is merged with pivot_table named 'dataset'. A new dataset has been created dropping shop_id, item_id and id.

The dataset is splitted in X_train, X_test and y_train. Dimension of X_train is (214200,33,1), dimension of y_train is (214200,1) and dimension of X_test is (214200,33,1).

V. Methodology

Model Preparation

After completing all the data preprocessing, we trained several deep learning models to preditct next sequence. We used have used Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Bi-directional LSTM, Gated Recurrent Unit (GRU), Bi-directional GRU, Convolutional Neural Netork (CNN) and different combinations of CNN and RNN based networks to run our processed data for the predictions.

All the models were trained with Mean Squared Error loss and Adam optimizer. An Epoch of 10 was used with a batch size of 4096.

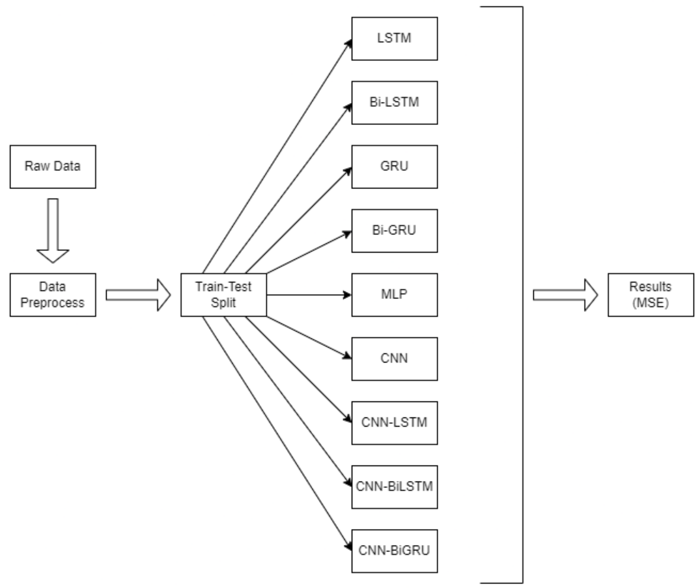

Fig. 2. Loss Curve for MLP

LSTM: This architecture contains 64 unit. A LSTM layer with 64 hidden with a dropout rate of 0.4 is applied. The output layer contains 1 neuron.

BiLSTM: This architecture contains 64 unit. A BILSTM layer with 64 hidden with a dropout rate of 0.4 is applied. The output layer contains 1 neuron.

GRU: This architecture contains 64 unit. A GRU layer with 64 hidden with a dropout rate of 0.4 is applied. The output layer contains 1 neuron.

BiGRU: This architecture contains 64 unit. A BIGRU layer with 64 hidden with a dropout rate of 0.4 is applied. The output layer contains 1 neuron.

MLP: A fully connected layer with 100 neurons was applied with "relu" activation and a dropout rate of 0.4. The output layer contains 1 neurons.

CNN: A convolutional layer with 64 neurons, kernel size = 2 was applied with "relu" activation. A max pooling layer with pool size = 2 is followed by a flatten and a dense layer with 50 neurons and "relu" activation.

CNN-LSTM: The convolution layer is identical to the CNN architecture previously defined. The CNN is follow by a LSTM layer with 50 neurons and "relu" activation.

CNN-BiLSTM: The convolution layer is identical to the CNN architecture previously defined. The CNN is follow by a BiLSTM layer with 50 neurons and "relu" activation.

CNN-BiGRU: The convolution layer is identical to the CNN architecture previously defined. The CNN is follow by a BiGRU layer with 50 neurons and "relu" activation.

VI. Result

After running the data in our selected models, we observed that Multilayer Perceptron (MLP) perfermed better than other models. Mean Squared Error (MSE) for MLP was 19.0906. We know the lower the MSE the better the result. The performance of different models are summarized in the following performance table:

PERFORMANCE TABLE

Model Name	Mean Square Error (MSE)
MLP	19.0906
CNN	21.6758
LSTM	29.6026
GRU	29.4742
BiLSTM	29.2528
BiGRU	29.3320
CNN-LSTM	27.6175
CNN-BiLSTM	26.8782
CNN-BiGRU	26.6248

We measured the performance of the models using Mean Squared Error (MSE). Total 9 models are analyzed in the performance table. The results are summarized as: Multi Layer Perceptron performes the best with minimum error rate of around 19.1% whereas Convolutional Nural Network (CNN) also performed well with error rate of 21.67%. And the other 7 models LSTM, BiLSTM, GRU, BiGRU almost showed the error rate of around 29% which is 9 to 10 percent higher than MLP and CNN models. Besides, combination of CNN-LSTM, CNN-BiLSTM and CNN-BiGRU showed error rate of 26 to 27 percent. So, analyzing the performance table, NLP performed best for demand forcasting whereas GRU performed worst for our research purpose. Higher results can be achieved for every model increasing epochs as 10 epochs has been used for the given models. We know, the higher epochs means the less loss or MSE.

In Fig-2 loss curve for MLP and CNN are represented. For MLP and CNN, the loss value changed for each epoch. The loss continued to show higher and lower values at different points. It did not show continuity of increasing or decreasing.

In Fig-3, for BiGRU the loss function decreased with each epochs. It showed the continuity of decreasing whereas GRU the loss started to decrease but at last epoch it increased little bit. For BiLSTM, the loss showed continuity and decreased with each epoch. For LSTM it did not show continuity and changed to higher or lower with different epochs.

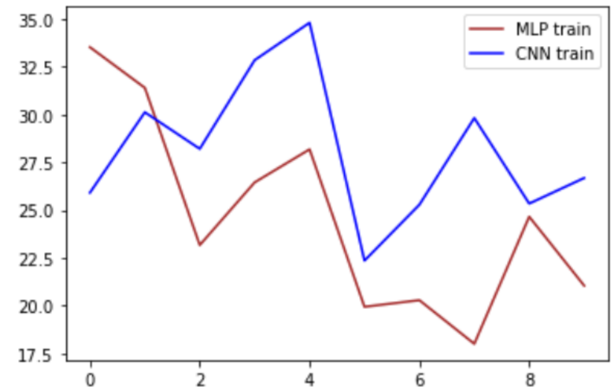

Fig. 3. Loss Curve for MLP & CNN

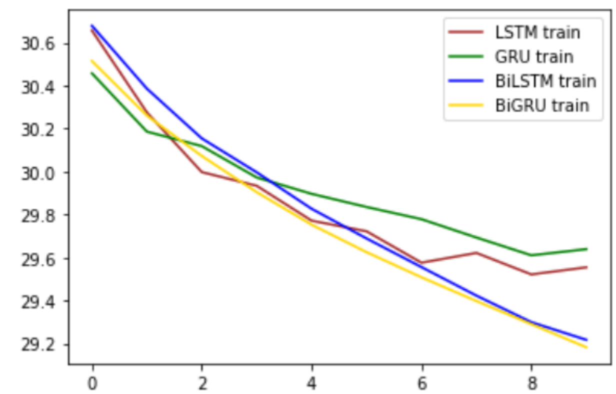

Fig. 4. Loss Curve for LSTM, GRU, BiLSTM & BiGRU

The last figure, Fig-4 represents the loss value for CNN-LSTM, CNN-BiLSTM and CNN-BiGRU. For these 3 models, the loss values changed drastically for each epoch. It is impossible to find continuity and it does not represent the last epoch as the minimum error rate.

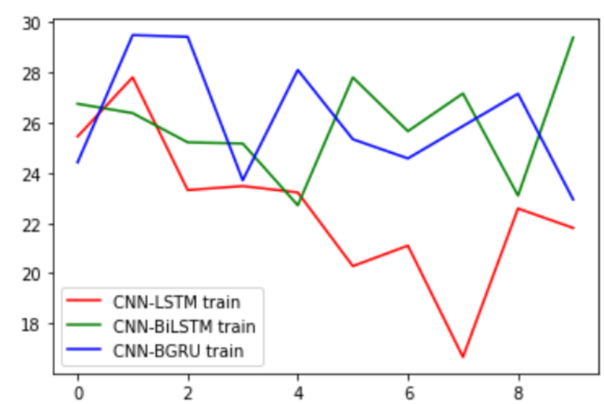

Fig. 5. Loss Curve for CNN-LSTM, CNN-BiLSTM & CNN-BiGRU

VII. Conclusion and Future Work

As management of a business are struggling to maintain the proper supply chain of products to store due lack of knowledge of product demands , there were efforts made to overcome this problem using deep learning. There has been studies related to predicting store item demand. So that , before disbursement of stock from the company, the demand for those product could be known. We have studied works related to this field of store item demand forecasting and have made a dataset of our own by preprocessing various dataset at various level to use it for achieving our target of predicting demand of particular items. After running the data through several models we got result of MSE (Mean Squared Error) for each and have come to the verdict that MLP was the best model to give better output of all for prediction. The minimum error found in MLP which is around 19%. The better result could be found with higher epoch value and different models might give better results than MLP. In future, we will try to implement Recurrent Nural Network (RNN), Bi-directional Recurrent Nural Network (Bi-RNN), ARIMAX and the combination of every model possible for time series recurrence problem.

References

[1] İrem İşlek and Şule Gündüz Öğüdücü, "A Retail Demand Forecasting Model Based on Data Mining Techniques," *2015 IEEE 24th International Symposium on Industrial Electronics (ISIE).*

[2] Özden Gür Ali, Serpil Sayın , Tom van Woensel and Jan Fransoo, "SKU demand forecasting in the presence of promotions", *Science Direct: Expert Systems with Applications*, vol. 36, issue 10, pp. 12340-12348, December 2009.

[3] Javad Feizabadi, " Machine learning demand forecasting and supply chain performance", *International Journal of Logistics Research and Applications*, vol. 25, issue 2, pp. 119-142, 04 August 2020.

[4] Praveen K B, Prateek J, Pradyumna Kumar, Pragathi G and Prof. Madhuri J., "Inventory Management using Machine Learning", *International Journal of Engineering Research and Technology.*, vol. 9, issue 6, issn 2278-0181, June 2020.

[5] V. Lototsky, R. Sabitov, G. Smirnova and B. Sirazetdinov, N. Elizarova, Sh. Sabitov,"Model of the Automated Warehouse Management and Forecasting System in the Conditions of Transition to Industry 4.0", *Science Direct: IFAC-PapersOnLine*, vol. 52, issue 13, pp. 78-82, 2019.

[6] Gers, Felix A. Schmidhuber, J urgen Cummins, Fred. (1997). Learning to Forget: Continual Prediction with LSTM. 10.1162/089976600300015015.

[7] https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data