RESEARCH ARTICLE OPEN ACCESS

Drug Sales Forecasting with Metadata and ACF Supported LSTM

Volkan Demir*, Dogan Tilkici*, Metin Zontul**

*(Caretta Software R&D Center, Istanbul/Turkey)

** (Department of Computer Engineering, Istanbul Arel University,Istanbul/Turkey)

--************************----------------------------------

Abstract:

Data stored based on time in the form of seconds, minutes, hours, days, years is called time series. The process of predicting future values using time series is called time series forecasting. One of the common time series forecasting method is Long short-term memory (LSTM) which is an artificial recurrent neural network (RNN) with feedback connections in the field of deep learning. In this study, he forecasting models are constructed on weekly drug sales time series data by using a combination of Metadata, ACF and LSTM. Metadata is produced based on the time points with an error rate compared with a threshold value in training. ACF is used to determine he stationary status of the data and LSTM is used for drug sales time series forecasting.

Keywords **—LSTM, ACF, Time Series Forecasting, Drug Sales Forecasting**

--************************----------------------------------

I. INTRODUCTION

Thanks to advances in information technology, data collection has become both easy and cost-effective. Data created when data is stored based on time in the form of seconds, minutes, hours, days, years is called time series. If consecutive records of a single variable are kept, the resulting data is expressed as a multivariate time series if the data belonging to more than one variable is sequentially included in the univariate time series. The process of predicting future values using time series is called time series forecasting. Decision support systems based on time series prediction models are important in many areas such as finance, energy, meteorology and health. In these areas, the data are mostly in the form of univariate time series, and linear or nonlinear models are used in time series estimation [1].

One of the common time series forecasting method is Long short-term memory (LSTM) which is an artificial recurrent neural network (RNN) with feedback connections in the field of deep learning. LSTM models are able to store information over a period of time because of their memory capacity. This characteristic is extremely useful for Time-Series data.

There are many studies regarding time series prediction with LSTM models. Sangiorgio and Dercole proposed LSTM networks as general purposes multi-step predictors for nonlinear time series. They compared LSTM nets with the benchmarks set by feed-forward, one step-recursive and multi-output predictors by analysing artificial, noise-free data generated by chaotic oscillators. They also showed that LSTM architectures give good performances in case the number of time lags included in the input are different from the actual embedding dimension of the dataset [2].

Wang et al developed a forecasting model of COVID-19 by an improved LSTM method. Their model is different from traditional epidemical models in a way that it can predict both rising and declining trends for long-term projections. In a similar study [4], the key features to predict the trends and possible stopping time of the current COVID-19 outbreak in Canada and around the world is evaluated with LSTM networks.

Karevan and Suykens proposedTransductive LSTM (T-LSTM) to obtain a data driven forecasting model for an application of weather forecasting by exploiting the local information in time-series prediction. Also, they used a weighted quadratic cost function based on the cosine similarity between the training and the test samples for the regression problem [5]. Xiao et al proposed a machine learning method combining LSTM deep recurrent neural network model and the AdaBoost ensemble learning model (LSTM-AdaBoost) to predict the short and mid-term daily Sea surface temperature. They considered LSTM for modelling long-term dependencies and AdaBoost for avoiding overfitting. Their case study showed that the proposed LSTM-AdaBoost combination model outperformed the LSTM and AdaBoost separately.

There are also other studies regarding demand and sales forecasting with LSTM. Abbasimehr et al proposed a demand forecasting model based on multi-layer LSTM networks. Their model automatically selects the best time series forecasting model using the grid search method. The proposed model outperformed ARIMA, ANN, KNN, SVM and single layer LSTM on demand data of a furniture company.

In another study, after analysing and clustering natural gas daily consumption profiles, a comprehensive LSTM models were built according to load behaviour. LSTM results were compared with seasonal time series with exogenous variables models, MLP neural network approach, and multiple linear regression model. It was indicated that forecasting accuracy especially for days with exceptional customers consumption behaviour change is improved [8]. As a similar study, Su et al proposed a robust hourly gas consumption model based on hybrid structure of Wavelet Transform, Genetic Algorithm, LSTM Network. Wavelet was used for decomposing the original series of gas loads into several sub-components. Forecasting models whose layers were optimized by Genetic Algorithm were built with LSTM [9].

Sarkar and De Bruyn [10] used LSTM models for direct marketing by replacing feature engineering with Deep Learning. They predicted customer behaviours with great accuracy on panel data observed repeatedly over time and along multiple dimensions. Their LSTM model beats 269 of 271 hand-crafted models that use a wide variety of features and modelling approaches.

Muzaffar and Afshari [11] collected an electrical load data with external variables with the inclusion of temperature, humidity and wind speed to train the LSTM network, which was compared with traditional methods for modelling the load time series over the different time periods using RMSE and MAPE metrics. They also used the ACF/PACF plots to determine whether the time series are stationary or not. It was indicated that a further data processing is required for non- stationary time series data.

In another recent study [12], the researchers tried to determine the focus of Chinese consumers with respect to new energy vehicles and to understand how their interests affect the sales volumes by combining NLP and LSTM. They pointed out that demographic factors, national policies, safety awareness and infrastructure are closely connected to the sale volumes.

In this study, the drug sales data is obtained from the warehouses of Santa Farma firm in Turkey, a drug producer. Due to the different forms of drugs with the same effect and the different active ingredients of each, it is decided to consolidate all drug data at the product name level during the data consolidation phase, taking the opinions of experts. As a result, 8 daily sales time series are produced for 8 different drug names from 2015 January to 2018 December. Later, in order to reduce noise in data, the daily data is converted to weekly data. Finally, the forecasting models are constructed on weekly drug sales time series data (Allerset and Dicloflam drug datasets) by using a combination of ACF and LSTM. ACF is used to determine the stationary status of the data and LSTM is used for drug sales time series forecasting.

II. LONG-SHORT TERM MEMORY (LSTM)

Long-Short Term Memory (LSTM), a member of the recurrent neural networks which is a sub-branch of artificial neural networks, is a widely used algorithm in time series prediction, thanks to its ability to store long and short-term contextual information in cells and transfer this contextual information to subsequent cells. The working flow of LSTM cells is illustrated in Fig. 1. An LSTM cell exhibits a four-layer structure. Several steps are performed in order on each of these layers. In the first step, it is determined how much of the contextual information previously remaining in the first layer, the forgetting door layer, is forgotten. In the second step, it is selected which new information will be stored as contextual information. This process is achieved by determining which values will be updated in the second layer, the input gate layer, and generating potential new information in the tanh layer. In the last layer, the previous processes are implemented and the contextual information from the previous cell is updated.

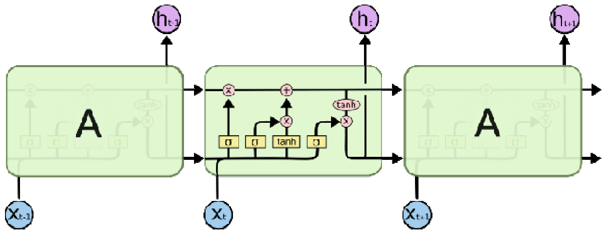


Fig. 1 Working flow of LSTM cells

III. DETERMINATION OF INPUT DATA WITH AUTOCORRELATION FUNCTION

In order to make predictions with LSTM, the data must be suitable for supervised learning. This requirement can be achieved by giving delay values of data such as Xt-1, Xt-2 as input to the model while estimating the moment of Xt. Thus, while the LSTM algorithm calculates the value at Xt, by looking at the history of the data, it can find a correlation between the time Xt and the delayed values of the Xt instant. These delay data, which have a high relationship with the moment Xt, were determined by ACF. The estimation power of the algorithm has been increased by including only the delay values outside the critical area into the model. These critical values were determined as (-0.184) and (0.184), corresponding to a 1% confidence interval. The ACF graph drawn for the Allerset data set is shared in Fig. 2, and the code fragment that chooses delay values that exceed critical thresholds is shared in Fig. 3. Thus, when the Xt-1 delay is called 1, the delays for Allerset data set should be included in the model (2, 5, 7, 9, 51, 52).

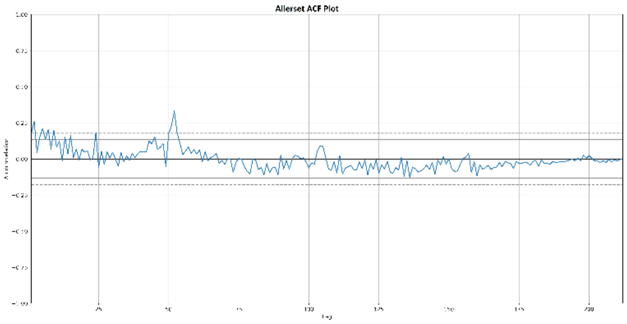
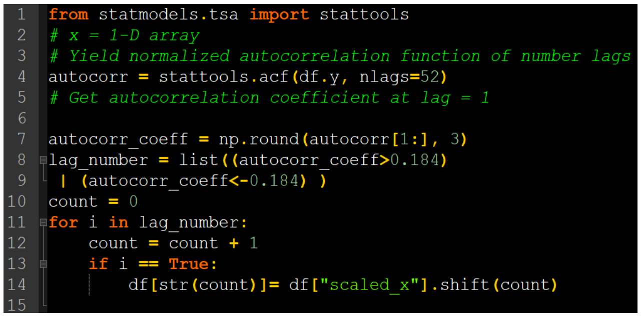


Fig. 2 ACF plot for Allerset dataset



```
from statmodels.tsa import stattools
# x = 1-D array
# Yield normalized autocorrelation function of number lags
autocorr = stattools.acf(df.y, nlags=52)
# Get autocorrelation coefficient at lag = 1

autocorr_coeff = np.round(autocorr[1:], 3)
lag_number = list((autocorr_coeff>0.184)
 | (autocorr_coeff<-0.184) )
count = 0
for i in lag_number:
    count = count + 1
    if i == True:
        df[str(count)]= df["scaled_x"].shift(count)

```

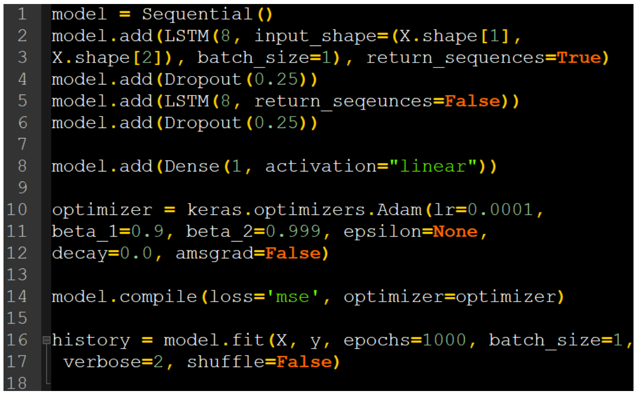
Fig. 3 Code fragment that selects delays that exceed critical thresholds

IV. DETERMINATION OF MODEL PARAMETERS

Major parameters of an LSTM model are as follows: the number of hidden layers, the number of neurons in each hidden layer, batch size, percent dropout, error function, optimizer, learning rate, and epoch number.

The number of hidden layers is limited to a maximum of two to prevent memorization due to the small amount of data. Although there is no method agreed on in the literature for the number of cells in these hidden layers, one of the widely accepted methods is determined to be the power of two closest to the amount of input data entering the model. Since there is relatively little data in the data set, the batch size parameter, which is the parameter that determines the amount of elements of the stack to be included in the model, is determined as one. In order to prevent memorization, dropout ratio, which is the ratio of randomly selected cells to be disabled for each epoch in each hidden layer, was chosen as 0.25.

While the error function was determined as the mean square error among the parameters selected while compiling the model, the optimizer was chosen as the "Adam", which was determined to give the best performance through trial and error. While the learning rate parameter determined while fitting the model was determined as 0.0001, the number of epochs was chosen as 1000 because a low learning rate was preferred. According to these explanations, the code fragment that constitutes the model built for the Allerset dataset is shared below in Fig. 4.



```
model = Sequential()
model.add(LSTM(8, input_shape=(X.shape[1],
X.shape[2]), batch_size=1), return_sequences=True)
model.add(Dropout(0.25))
model.add(LSTM(8, return_seqeunces=False))
model.add(Dropout(0.25))

model.add(Dense(1, activation="linear"))

optimizer = keras.optimizers.Adam(lr=0.0001,
beta_1=0.9, beta_2=0.999, epsilon=None,
decay=0.0, amsgrad=False)

model.compile(loss='mse', optimizer=optimizer)

history = model.fit(X, y, epochs=1000, batch_size=1,
 verbose=2, shuffle=False)
```

Fig. 4 Test-1 on Allerset dataset

V. GENERATING METADATA WITH ERROR FLAGS

The model was trained on the whole data set without being tested. In this training, time points with an error rate greater than 20% are marked as 1, those below -20% are marked as -1, and the remaining time points as 0. Thus, a vector is defined that can take (-1, 0, 1) values indicating the time points where the model makes a much less and much higher estimate than expected. The aim here is to test whether the predictive power improves when this metadata is given to the model as an input, with subsequent tests; If it is improving, it is to find external resources with high potential to improve the model by investigating whether there are external sources that correlate with this metadata created with error flags.

VI. TEST-1 AND TEST-2

After creating metadata with error flags, two experiments were performed on Allerset data set named Test-1 and Test-2 to test whether the predictive power of the model improved when it received this metadata as input.

In Test-1, the model uses only real values as input. Therefore, the estimation can only be made after a time point. In one of the two models created in the same architecture and with the same parameters, the model was trained only with delayed data points determined by ACF, while in the other, it was trained with metadata created with delayed data points and error flags determined by ACF. In the chart shared in Fig. 5, while the X axis is the week numbers for 2018, the y axis is the amount sold or estimated to be sold in milligrams; The data shown in red are real values, while the data shown in blue are estimates made without using metadata, while the data shown in green are estimates made using metadata.

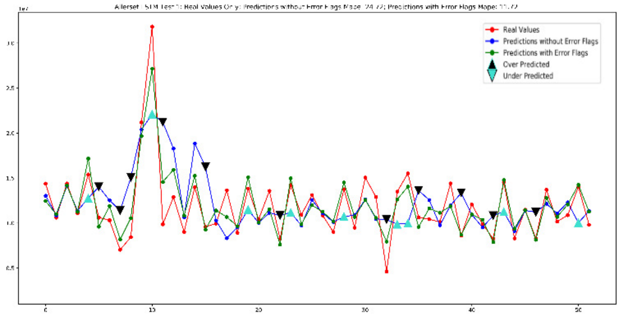


Fig. 5Test-1 on Allerset dataset

In Test-2, on the other hand, the model obtains the lag data it takes as input from its previous estimates. Thus, the estimation is not limited to only estimating a time point, it can offer estimation up to a desired time interval. However, as this time interval gets longer, the predictive power of the model gradually decreases as it gets further away from the real data. In Test-2, one of the two models created in the same architecture and with the same parameters, the model was trained only with delayed data points determined by ACF, while in the other, it was trained with metadata created with delayed data points and error flags determined by ACF. Below, in the chart shared in Fig. 6, while the X axis is the week numbers for 2018, the y axis is the amount sold or estimated to be sold in milligrams; The data shown in red are real values, the data shown in blue are estimates made without using metadata, while the data shown in green are estimates made using metadata.

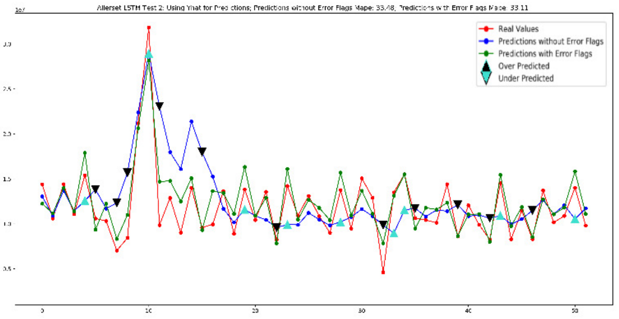


Fig. 6 Test-2 on Allerset dataset

VII. EXPERIMENT RESULTS

In the experiments performed with Test-1 and Test-2, it was observed that the prediction made by using metadata created with error flags gave better results than predictions without metadata created with error flags. In Test-1, while the average absolute percentage error (MAPE) of the estimates of the model without metadata was 24.72%, the estimations of the model made with metadata were 11.72%. In Test-2, although the relevant values are 33.48% and 33.11%, respectively, it was observed that the model with metadata makes a more appropriate estimation to the fluctuation in the data while the model without metadata makes a more horizontal estimation. For this reason, it was decided to search for external sources correlated with the relevant metadata, add these sources and remove the metadata created with error flags and continue the experiments.

When the experiment with Test-2 is examined, the model that uses the predicted values as input gives better results than expected, especially if it is concluded that the deviation from Test-1 is acceptable for the first four predictions and there are external sources correlated with the metadata created with error flags. It was decided to continue the Test-2 experiments.

After these results, Test-1 and Test-2 pioneer experiments were applied on the Dicloflam dataset, which is relatively difficult to predict, without optimizing the model architecture and parameters. The results of the relevant experiments are shared in Fig. 7 and Fig. 8. According to the preliminary results obtained from these pioneering experiments, it was observed that the addition of error flagged metadata increased the estimation power of the model, but according to the experiments performed on the Allerset dataset, these models performed very poorly. In the related experiment with Test-1, the MAPE value of the predictions made with the model without metadata was 100.59%, while the MAPE value of the predictions made with the metadata model was 49.62%. The same values for Test-2 were 94.99% and 106.64%, respectively, but the model without metadata did not make any estimates and returned a value that would subtract the lowest possible mean square error value in all predictions. For this reason, more comprehensive model and optimization studies will be done for the relevant dataset.

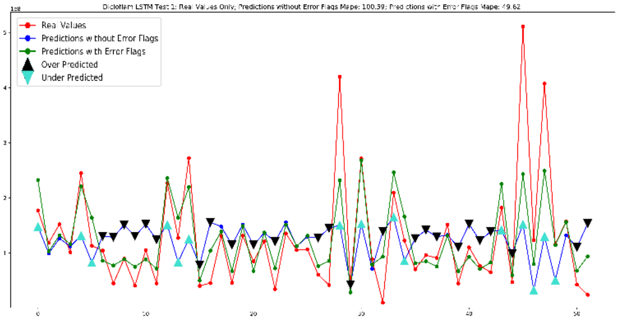


Fig. 7 Test-1 on Dicloflam dataset

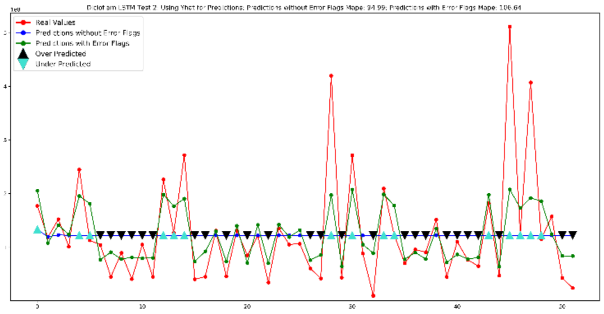


Fig. 8 Test-2 on the Dicloflam dataset

VIII. CONCLUSIONS

With this study, using the weekly consolidated data of drugs as milligrams that being sold by drug stores has been analysed and applied time series algorithms to predict future outcomes. For this, after the data preparation, qualitative analysis of the data was performed, the delay values of the LSTM model were determined by applying ACF and PACF, and it was ensured that it is suitable for supervised learning. The hyperparameters of the LSTM model have tuned however these outputs without metadata did not show good enough result. Thus, the metadata was obtained with error flags and provided as input to the model. This model, considered as a pilot application, is expected to be successful in some other drug time series. As stated in chapter 7, a more comprehensive model and optimization will be done to get stable results regardless of the type of drug.

REFERENCES

[1] Ü. Ç.Büyükşahin, and Ş.Ertekin, “A feature-based hybrid ARIMA-ANN model for univariate time series forecasting,”*Journal of the Faculty of Engineering & Architecture of Gazi University*, vol. 35, 2020.

[2] M.Sangiorgio, and F.Dercole,*Robustness of LSTM neural networks for multi-step forecasting of chaotic time series*, Chaos, Solitons & Fractals, 139, 110045, 2020.

[3] P.Wang, X.Zheng, G.Ai, D.Liu, and B.Zhu, *Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: case studies in Russia, Peru and Iran,* Chaos, Solitons & Fractals, 110214, 2020.

[4] V. K. R.Chimmula, and L.Zhang, *Time series forecasting of COVID-19 transmission in Canada using LSTM networks,* Chaos, Solitons & Fractals, 109864, 2020.

[5] Z.Karevan, and J. A.Suykens, “Transductive LSTM for time-series prediction: An application to weather forecasting,” *Neural Networks*, vol. 125, pp. 1-9, 2020.

[6] C.Xiao, N.Chen, C.Hu, K.Wang, J.Gong, and Z.Chen,“Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach,” *Remote Sensing of Environment*, vol. 233, 2019.

[7] H.Abbasimehr, M.Shabani, and M.Yousefi, “An optimized model using LSTM network for demand forecasting,” *Computers & Industrial Engineering*, 106435, 2020.

[8] O.Laib, M. T.Khadir, and L.Mihaylova, “Toward efficient energy systems based on natural gas consumption prediction with LSTM Recurrent Neural Networks,” *Energy*, vol. 177, pp. 530-542, 2019.

[9] H.Su, E.Zio, J.Zhang, M.Xu, X.Li, and Z.Zhang, “A hybrid hourly natural gas demand forecasting method based on the integration of wavelet transform and enhanced Deep-RNN model,” *Energy*, vol. 178, pp. 585-597, 2019.

[10] M.Sarkar, A. De Bruyn, “LSTM Response Models for Direct Marketing Analytics: Replacing Feature Engineering with Deep Learning,” *Journal of Interactive Marketing*, vol. 53, pp. 80-95, 2021.

[11] S.Muzaffar,A.Afshari, “Short-term load forecasts using LSTM networks,” *Energy Procedia*, vol. 158, pp. 2922-2927, 2019.

[12] L.Wang, Z. L.Fu, W.Guo, R. Y.Liang, and H. Y.Shao, “What influences sales market of new energy vehicles in China? Empirical study based on survey of consumers’ purchase reasons,” *Energy Policy*, vol. 142, 2020.