A comparative study of predictive models for pharmaceutical sales data

Jérémy Renaud
FEMTO-ST Institute
Univ. Bourgogne Franche-Comté
Belfort, France
Caduciel informatique
Voray-sur-l'ognon, France
jeremy.renaud@univ-fcomte.fr

Raphaël Couturier
FEMTO-ST Institute
Univ. Bourgogne Franche-Comté
Belfort, France

Christophe Guyeux
FEMTO-ST Institute
Univ. Bourgogne Franche-Comté
Belfort, France

Benoit COURJAL
Caduciel Informatique
Voray-sur-l'ognon, France
benoit.courjal@caduciel.com

Christine GIOT
Caduciel Informatique
Voray-sur-l'ognon, France
christine.giot@caduciel.com

***Abstract*—To provide their patients with the care they need as quickly as possible, pharmacies are supplied by wholesaler-distributors who provide them with a half-day delivery guarantee for most product references. For this purpose, they have set up an efficient and complex supply chain. To further improve the efficiency of their delivery services, some of them want to use machine learning tools to predict future orders and anticipate their inventory needs. This paper investigates different machine learning models for the prediction of sales on molecules of a French wholesaler-distributor. This paper focuses on four molecules and compares the results of the models predictions on these molecules.**

***Index Terms*—Neural network, Gradient Boosting, Transformers, Forecasting, Pharmaceutical products**

I. Introduction

Pharmacies are a key part of the French health care system. To allow their patients to have access to their treatments quickly, French pharmacies are supplied by different wholesaler-distributors. Their main mission is to provide their pharmacist customers with fast delivery of their orders; currently, the standard is two deliveries per day. Wholesaler-distributors must set up an extremely efficient logistic chain to have a maximum of product references as close as possible to each of their customers.

Knowing in advance what pharmacists are going to order can help improve supply chain efficiency, including matching the right product references to potential customers, and ensuring that each warehouse is stocked enough to avoid stock-outs for example. In the world of logistics, being helped by artificial intelligence algorithms has been in place for many years, either by statistical and business intelligence systems or more recently by machine learning algorithms.

On the thousands of different products references on the French pharmaceutical market, predicting with precision which will be sold next, can be very complicated. This problem can be simplified by first studying the sales grouped by molecules.

This paper investigates different machine learning models for the prediction of sales of molecules of a French wholesaler-distributor. After tests and investigations, the choice was made to focus on some molecules, which will be explained in the Pharmaceutical sales data section. Therefore, this study focuses on four molecules and compares the results of the model's predictions for these four molecules.

II. Related Works

Decision support systems are already used in many fields, such as in the world of logistics, whether for optimization [1] or for crisis management through exceptional situations that can put a strain on the supply chain [2]. More recently, the use of machine learning technologies such as neural networks has emerged, for demand prediction in the field of sales for instance [3], notably thanks to the good results of architectures such as LSTM (Long Short-Term Memory).

These architectures have also been used in various domains such as firefighter intervention prediction [4], which is a more abstract subject than simple sales data. In the health domain, the use of decision-making tools to manage hospital logistics platforms has already been observed [5], or the use of deep learning architecture to detect and predict the evolution of diseases in the population [6]. To take a closer look at the pharmaceutical field, the trend is still towards the use of statistical models [7], as machine learning solutions do not yet seem to have been adopted in this type of market.

III. Pharmaceutical sales data

The data correspond to the total sales of a French wholesaler-distributor between 2017 and the end of 2021, more precisely they correspond initially to each product purchased by each customer of this wholesaler-distributor. They are recovered and grouped by molecule. Then the choice of the temporal granularity has been made. It was decided to re-aggregate them by week, which is for the moment the best compromise between the number of available data and the smoothing of the random orders. Then, the choice of which molecule to study has been prioritized, based on one criterion: make focus on molecules with seasonality, because it will help the machine learning algorithms to understand the market.

Four molecules were selected, because they are among the most sold molecules by the wholesaler-distributor, and three of them have a strong seasonality : Amoxicillin which is a very commonly used antibiotic, Cetirizine which is an antihistamine used in the treatment of allergies, Colecalciferol which is a vitamin D, and finally Paracetamol which is one of the most used molecules as a painkiller and anti-fever. The sales data are ordered by week date, so they are considered as time series. They are given as input to the Machine Learning algorithms in the following format: Date, average between n-1 and n-4, minimum between n-1 and n-4, maximum between n-1 and n-4, seasonality function, and trend. The algorithms do not have the previous data as input, to avoid them simply copying those data as a result. The data set is divided into 80% for training, 10% for the validation set, and 10% for the test set.

IV. Studied machine learning models

In this work, several machine learning architectures have been tested and selected.

A. Stacked Long Short-Term Memory (Stacked-LSTM)

The LSTM architecture is a particular type of recurrent neural network based on memory blocks [8]. Previous architectures, such as vanilla recurrent neural networks, can also be used for sequential data with short-term dependencies. The advantage of the LSTM architecture is its ability to use both short and long-term dependencies to make predictions. LSTM has a hidden state as well as a cell state which acts as a memory. The contents of the memory, as well as the output of the network, are controlled using point operators called gates (forget gate, input gate, and output one).

Note that this study did not use a GRU (Gated Recurrent Unit) network, another architecture that often provides comparable performance while being faster to train, because no satisfactory predictions were obtained in preliminary experiments. A possible explanation is that in theory, LSTMs can store longer sequences than GRU networks.

B. Gradient Boosting (LightGBM)

LightGBM is a popular gradient boosting framework created by Microsoft. It is a tree-based learning algorithm designed to have faster-learning speed, reduced memory usage, and better accuracy [9]. It supports execution with GPU and parallelization, has a test function and keeps the best set of parameters. Compared to other boosting algorithms, LightGBM differs in the approach used for tree growth, namely leaves, which are leaf-based (best first) and not level-based (depth-first).

C. Transformers

The architecture of the Transformers models was introduced in 2017 [10] and has allowed great advances in natural language processing. The main feature of these models is an attention system that allows sequential data to be processed without them necessarily being processed in a particular order. The Transformers allow a better parallelization of data, which allows the processing of larger quantities, surpassing old architectures such as GRU or LSTM in the field of natural language processing. Transformers models have also shown their performances on other problems such as image, video, or time series processing.

D. Dumb model - baseline (mean)

The ”dumb” model is a simplistic model that serves as a baseline for comparing the effectiveness of others: always predicts the average value of the training set data.

V. Experimental results

A. Setup of Machine Learning Architectures

1) Stacked-LSTM: The Stacked-LSTM architecture is composed of the first layer with 300 units with a reLU activation function, then a second LSTM layer of 100 units still with a reLU activation, to finish on a ”Dense” layer of size 1. Between the two LSTM layers, a 20% dropout has been added to avoid potential overfitting. The RMSprop optimizer was used with a learning rate of 1e-4 and the look back was set to 5.

2) Gradient Boosting (LightGBM): For the settings of the LightGBM model, a grid search was used on the following parameters: learning rate with a variation of 0.01 between 0.01 and 0.08, number of estimators between 100 and 2000 with a step of 100, number of leaves between 2 and 10 with a step of 1 and the max depth between 2 and 16 with also a step of 1. The early stopping was set to 300 epochs, because the model had a tendency to stop at the beginning of the training with a shorter early stopping.

3) Transformers: The Transformers architecture applied in this paper is the one implemented in Keras for classification [11], but adapted to regression with an output ”dense” layer of size 1 and a ”linear” activation function. The architecture has the following parameters: 256 head size and 4 heads for the multi-head attention, with a dropout of 0.25. The Adam optimizer has been used with a learning rate of 1e-4. The training is 1000 epochs with an early stopping after 20 epochs and a batch size of 64.

B. Global Sales Predictions

1) Root Mean Square Error (RMSE) benchmark: The following tables show a comparison of the different RMSE of each model on the sales prediction of the four molecules: Amoxicilline (Table I), Cétirizine (Table II), Colécalciférol (Table III) and Paracétamol (Table IV). As the scales of the number of products sold are not the same for the four molecules, the column ”/mean” represents the error weighted

Amoxicilline				
model	RMSE (train)	/mean	RMSE (test)	/mean
Baseline	36020	0,36	39244	0,39
Stacked-LSTM	**29092**	**0.29**	26947	0.27
LightGBM	48934	0.49	27977	0.28
Transformers	29986	0.3	**23566**	**0.23**

TABLE I
COMPARATIVE OF RMSE FORECASTING SALES FOR AMOXICILLINE

Cétirizine				
model	RMSE (train)	/mean	RMSE (test)	/mean
Baseline	4006	0,30	3662	0,27
Stacked-LSTM	**3040**	**0.23**	6200	0.47
LightGBM	4470	0.49	3373	0.25
Transformers	3299	0.25	**2103**	**0.16**

TABLE II
COMPARATIVE OF RMSE FORECASTING SALES FOR CÉTIRIZINE

Colécalciférol				
model	RMSE (train)	/mean	RMSE (test)	/mean
Baseline	21397	0,22	25603	0,27
Stacked-LSTM	**13308**	**0.14**	29700	0.31
LightGBM	26358	0.28	27480	0.29
Transformers	17100	0.18	**23433**	**0.24**

TABLE III
COMPARATIVE OF RMSE FORECASTING SALES FOR COLÉCALCIFÉROL

Paracétamol				
model	RMSE (train)	/mean	RMSE (test)	/mean
Baseline	90713	0,27	106626	0,32
Stacked-LSTM	88230	0,27	63182	0,19
LightGBM	**61733**	**0,19**	102648	0,31
Transformers	106431	0,32	**53491**	**0,16**

TABLE IV
COMPARATIVE OF RMSE FORECASTING SALES FOR PARACÉTAMOL

by the average of sales of each molecule, to allow a comparison between them.

The LSTM model can approximate the curve better on training data. However the Transformers model is more efficient with training data: it is able to generalize the shape of the curve, and therefore is more accurate than the LSTM model on data it has never seen. This can be observed on the curves hereafter. It is also interesting to note that the LightGBM model does not learn the curves, but gets better results on paracetamol by mimicking the baseline model and predicting values around the mean.

2) Forecasting curves: For all the figures contained in this article, the purple curves represent the real data, the green curves the predictions on the training data, the red curves (if they exist), are the predictions on the data of the validation set, and the blue curves are the predictions on the test set. The x-axis represents the weeks and the y-axis is for the number of sales.

The Transformers model tends to follow the trend line without worrying about noise spikes (Figure 1, Figure 4, Figure 7, and Figure 10). On the contrary, the Stacked-LSTM model tries to approximate the curve as much as possible (Figure 2, Figure 5, Figure 8, and Figure 11). This makes a difference in the test set which does not have the same proportions as the other years: the Stacked-LSTM predicts a curve similar to the previous years while the Transformers will follow the downward trend in sales. The LightGBM model has very bad predictions skills on the training data but approximates the curve quite well on the test data (Figure 3 and Figure 9). It also tends to approximate close to the mean. This shows that it does not understand seasonality, but enables one to obtain good results on the data of paracetamol which has a more linear curve (Figure 6 and Figure 12).

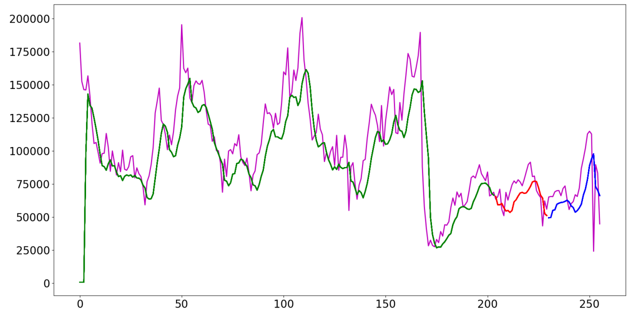

Fig. 1. Forecasting curves of Transformers model for Amoxicilline

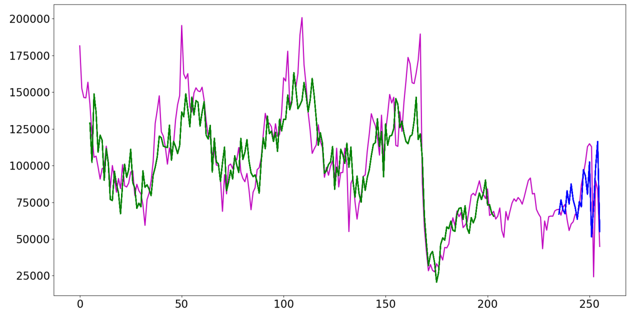

Fig. 2. Forecasting curves of LSTM model for Amoxicilline

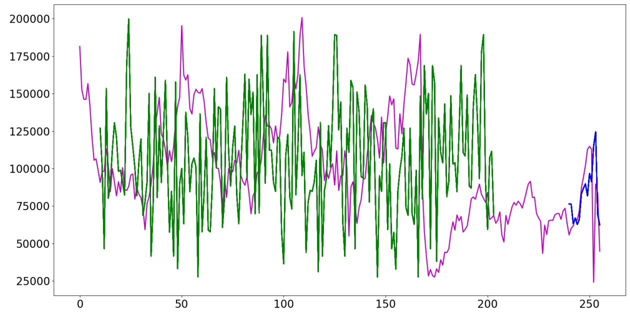

Fig. 3. Forecasting curves of LightGBM model for Amoxicilline

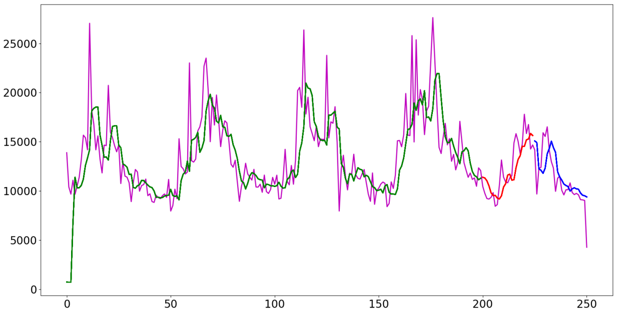

Fig. 4. Forecasting curves of Transformers model for Cetirizine

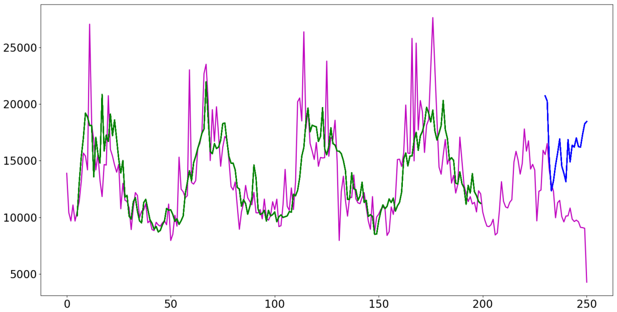

Fig. 5. Forecasting curves of LSTM model for Cetirizine

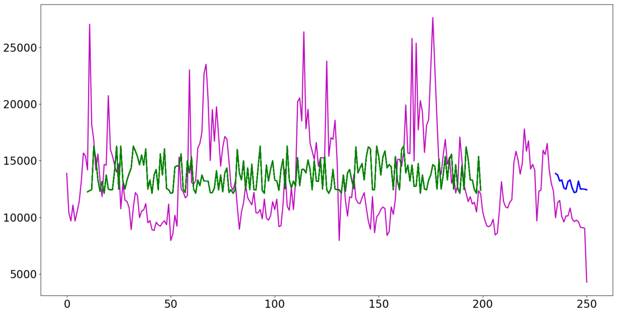

Fig. 6. Forecasting curves of LightGBM model for Cetirizine

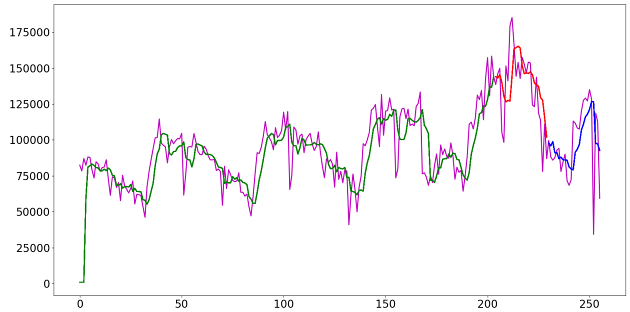

Fig. 7. Forecasting curves of Transformers model for Colecalciferol

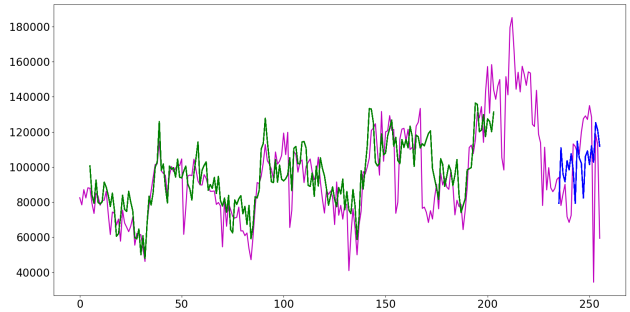

Fig. 8. Forecasting curves of LSTM model for Colecalciferol

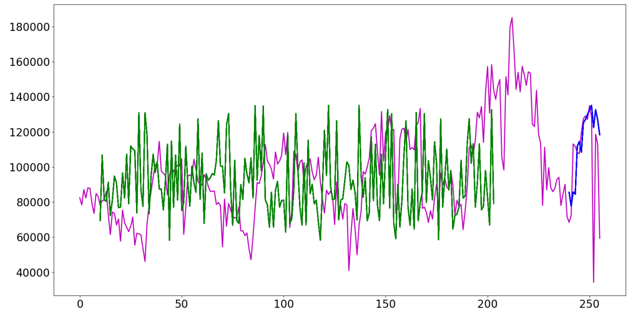

Fig. 9. Forecasting curves of LightGBM model for Colecalciferol

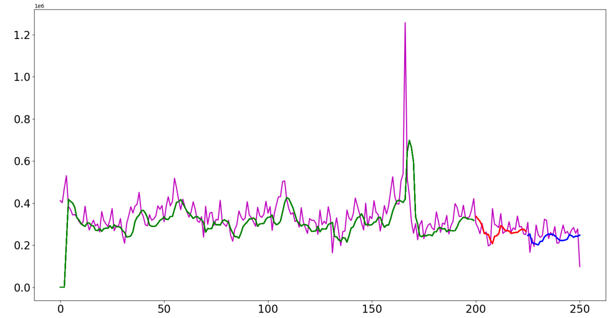

Fig. 10. Forecasting curves of Transformers model for Paracetamol

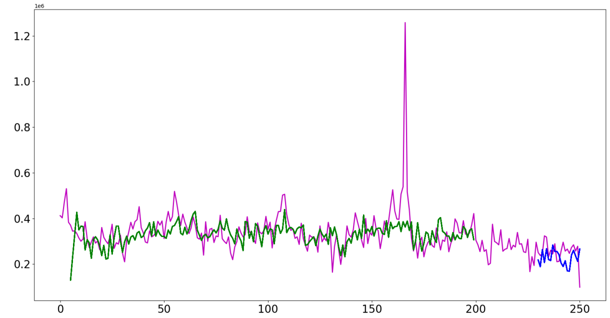

Fig. 11. Forecasting curves of LSTM model for Paracetamol

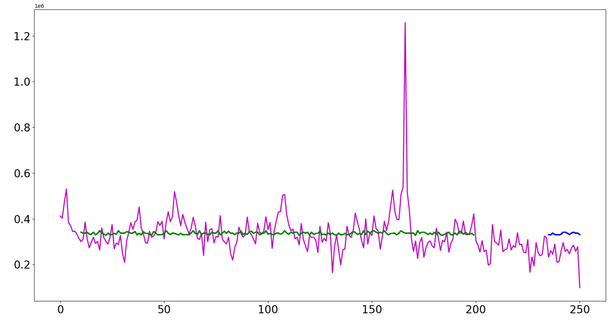

Fig. 12. Forecasting curves of LightGBM model for Paracetamol

VI. Conclusion

In this comparative study, the efficiency of different machine learning architectures for the prediction of pharmaceutical sales related to four molecules was observed. Out of these four molecules, three have a strong seasonality and the 4th is more linear. The results indicate that the Transformer-based model has the best results on the test data because it generalizes better during the training. The Stacked-LSTM model, on the other hand, obtains the best results on the training data, because it tries to approximate the data as well as possible and therefore loses generalization when it sees new data. These two models have difficulties with the molecule which does not have a marked seasonality, and cannot perform better than the basic model which only predicts the mean. In that case, the LightGBM model, which does not succeed to learn correctly for the molecules with seasonality, manages to do better for the 4th by imitating the principle of the basic model and predicting a value close to the average. The next step of our work will be to improve these predictions, especially by adding external data to help the models understand the pharmaceutical market.

References

[1] Sébastien Thomassey, Développement de systèmes d'aide à la décision pour l'optimisation des systèmes de production et de la chaîne approvisionnement de la filière textile habillement distribution Habilitation à Diriger des Recherches, 2018

[2] Vérane HUMEZ, Proposition d'un outil d'aide à la décision pour la gestion des commandes en cas de pénurie : Une approche par la performance, Rapport de thèse 2008.

[3] Bandara K., Shi P., Bergmeir C., Hewamalage H., Tran Q., Seaman B. (2019) Sales Demand Forecast in E-commerce Using à Long Short-Term Memory Neural Network Methodology. In: Gedeon T., Wong K., Lee M. (eds) Neural Information Processing. ICONIP 2019. Lecture Notes in Computer Science, vol 11955. Springer, Cham. https://doi.org/10.1007/978-3-030-36718-3_39

[4] S. L. C. Ñahuis, C. Guyeux, H. H. Arcolezi, R. Couturier, G. Royer and A. D. P. Lotufo, "Long Short-Term Memory for Predicting Firemen Interventions," 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), Paris, France, 2019, pp. 1132-1137, doi: 10.1109/CoDIT.2019.8820671.

[5] Dhia Jomaa. "Outil d'aide à la décision dans le pilotage de plateforme logistique hospitalière : Mise enplace d'un module de préconisation de commandes." Génie logiciel [cs.SE]. Université jean Monnet de Saint-étienne, 2013. Français. tel-01758331 https://tel.archives-ouvertes.fr/tel-01758331/document

[6] Chae S, Kwon S, Lee D. Predicting Infectious Disease Using Deep Learning and Big Data. Int J Environ Res Public Health. 2018 Jul 27 ;15(8) :1596. doi : 10.3390/ijerph15081596. PMID : 30060525 ; PMCID : PMC6121625.

[7] Nawaz, Ahamd,Fouzia Nasir,Usman Aleem,"Sale Forecasting of Merck Pharma Company using ARMA Model", Research Journal of Finance and Accounting Vol 6 Num 21, 2015 https://www.academia.edu/19192280/Sale_Forecasting_of_Merck_Pharma_Company_using_ARMA_Model

[8] F. A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: continual prediction with lstm, in: 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), Vol. 2, 1999, pp. 850–855 vol.2.

[9] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: Advances in neural information processing systems, 2017, pp. 3146–3154.

[10] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin, Attention Is All You Need, 2017 , arXiv:1706.03762

[11] Theodoros Ntakouris,Timeseries classification with a Transformer model, 2021,https://keras.io/examples/timeseries/timeseries_classification_transformer