Performative Prediction in Time Series: A Case Study

Anonymous Author(s)
Affiliation
Address
email

Abstract

Performative prediction is a phenomenon where a model's predictions, or the decisions based on these predictions, may influence the outcomes of the model. This is especially conspicuous in a time series forecasting setting where interventions occur before outcomes are observed. These interventions dictate which data points in the time series can be used as inputs for future predictions. In this paper, we represent a patient's symptom levels along their cancer rehabilitation plight as a time series. We use a decision-tree based model to predict the future symptom values of a patient. Based on these predictions, clinicians decide which symptom levels will be observed in the future. We propose methods to mitigate the problem of performative prediction in time series. Our results show how performative prediction may lead to a 29.4% to 40.7% higher error across different symptoms.

1 Introduction

Typically time series forecasting (TSF) uses observations of a time series (and perhaps other information) to forecast future values of that series. As time passes, new data of the time series becomes available and is used to predict more future values. However, in some settings, the decisions based on these predictions may influence future observations. These predictions are called performative.

In this work, we consider a concrete application setting where the goal is to develop a monitoring system of symptom levels for cancer patients during their rehabilitation journey focusing medical care on patients in critical states. We, therefore, aim to predict future symptom values using historical symptom levels. These predictions aid in deciding whether a patient should visit a rehabilitation center (where the patient gets their state assessed by a medical professional) or not. Therefore, the prediction model will impact which symptom values will be observed and may be used as inputs for future predictions. This induces a distribution shift in the data, which corresponds to a performative prediction problem [13]. Moreover, such systems introduce a partial observability problem when deployed in the real world, making their evaluation especially tricky at deployment. Indeed, once the system is deployed, the true patient state is only available when the monitoring system sends the patient to a rehabilitation center. Otherwise, the true state remains unknown. It is therefore crucial to conduct an offline evaluation of the system that can simulate these effects. Our aim is to understand the impact of this partial observability and study how this affects the accuracy of the system.

The paper is organized as follows. Section 2 provides an overview of the related work. Section 3 formulates the general problem which we refer to as time series performative predictions. Section 4 presents our proposed methodology for investigating the impact of the performative prediction effect on model performance along with solutions to train models with increased robustness to this effect and the data used in this study. Section 5 finally shows results that highlight the need to consider these aspects when training and evaluating time series predictive models that may induce performative prediction behaviours at deployment.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Related Work

The problem of predicting future symptoms for oncology patients has been tackled with machine learning in many forms, e.g., predicting the severe symptoms in cancer patients [16, 12, 20] and using patient-reported symptom levels to improve the performance of models for emergency department visits among cancer patients [17]. However, these works do not consider the performative prediction dynamics that can occur at deployment but rather study the typical supervised learning problem.

Performative prediction has been observed in healthcare applications in the literature. Liley et al. [11] discuss a typical healthcare setting where interventions lead to bias. Lenert et al. [10] show how prognostic models can become victims of their own success. However, these settings are not modeled as TSF as our setting. TSF can often suffer from a distribution shift. Duan et al. [5] resolve this by using a Hyper TimeSeries Forecasting model that jointly learns the time-varying distributions and the corresponding forecasting models in an end-to-end fashion. Bennett and Clarkson [1] deal with TSF under distribution shift using differentiable forgetting. However, in these settings, the distribution shift is not induced by the decisions made by the model but is rather by the environment.

3 Time Series Performative Predictions

Time series prediction essentially consists of predicting the future value of some process based on values observed for this process until now:

$$\hat{x}_{t+\Delta} = f(x_{t-a}, x_{t-b}, x_{x-c}, \dots),$$

where t denotes the current time. We consider the generic setting where observations are not assumed to be evenly spaced in time nor aligned for multiple time series data. Moreover, we do not make any assumptions regarding the (in-)dependence between time series used for training a model and time series encountered at deployment. Figure 1 illustrates time series data in the considered problem.

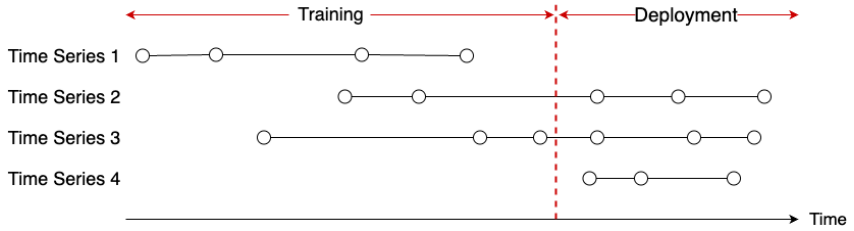

Figure 1: Time series data in the considered problem.

Such a problem will therefore require considering two cases at evaluation: time series that had observations in the training set (e.g. time series 2 and 3 in Fig. 1) and time series entirely encountered at deployment (e.g. time series 4 in Fig. 1).

Let $\mathcal{H}(t) = \{(t_1, x_{t_1}), (t_2, x_{t_2}), \dots\}$ denote the history of previous observations available at time $t \geq 1$ for a given time series, where t_i denotes the time of the i-th observation and $1 \leq t_i < t$. The problem is characterized by a decision function $g : \mathcal{X} \mapsto [0, 1]$ indicating whether an observation should be acquired (1) or not (0) based on the current prediction. At time t, a predictive model outputs the prediction $\hat{x}_t \in \mathcal{X}$ using $\mathcal{H}(t)$ and x_t is observed if $g(\hat{x}_t)$ is true, in which case $\mathcal{H}(t+1) = \mathcal{H}(t) \cup \{(t, x_t\}$. Otherwise, x_t is not observed and $\mathcal{H}(t+1) = \mathcal{H}(t)$.

In practice, one may allow some warm-up when starting to predict on a new time series. In this case, we define warm-up of duration N as systematically observing the first N observations for a time series. Therefore, decisions on a new time series in a system with warm-up of N always begin at time $t = N + 1$ with the history $\mathcal{H}(N+1) = \{(1, x_1), (2, x_2), \dots, (N, x_N)\}$.

4 Methodology

To investigate the impact of the performative prediction dynamics of the problem, we conduct experiments where we simulate the training and deployment of monitoring systems of symptom levels for cancer patients during their rehabilitation journey:

- We split the dataset in two as shown in Figure 1 (training and deployment).
- We train a model on the training set using different training strategies.
- We evaluate the model in the deployment setting simulated using the deployment data. For each time series contained in the deployment set, for each data point in the time series: if the history for that series contains less than $N = 2$ observations, then acquire observation; otherwise, predict and acquire (or not) observation based on the decision function g.

We consider a decision function that acquires the observation when the prediction exceeds a threshold:

$$g = \begin{cases} 1 & \text{if } I(\hat{x}_t \geq \tau) \\ 0 & \text{otherwise,} \end{cases} \qquad \text{with threshold value } \tau = 5.$$

4.1 Data

The data used in this study has been collected from 2013 to 2019 for 14586 patients with cancer types breast, colorectal, lymphoma, and head and neck. We use demographic information of patients like sex and age, along with surgery and historical symptom level data for prediction. Specifically, we use the two previous symptom levels as input to the model. Therefore, we consider only those patients in the dataset that have at least three observations of a symptom. We consider two symptoms to predict, namely, pain and fatigue. Each symptom level can be one of 11 possible values: 0 to 10, where 0 means the lowest level of pain or fatigue and 10 means the highest level. We observe that there are class imbalance problems, where approximately 43% and 21% of the outputs belong to one class (level 0) for pain and fatigue, respectively. Moreover, the number of samples in a class is inversely proportional to the symptom level.

4.2 Predictive model

Decision trees (DTs) are widely used in healthcare [14] and cancer-related applications [15, 6, 19, 4, 3, 7]. In this work, we consider LightGBM [8], a gradient boosting DT approach where new models are created to predict the residuals of prior models and are then combined to make the final prediction.

Inspired by previous works [21, 9], we rely on Synthetic Minority Oversampling TEchnique (SMOTE) [2] to deal with minority classes by generating synthetic samples so that the trained model sees the same number of data points of these classes as for the majority class. A minority class is over-sampled by taking each of its samples and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours. For our experiments, we use the default value $k = 5$. To reduce the impact of the majority classes in the training, we also perform undersampling with Tomek Links [18] after oversampling.

4.3 Training strategies

We evaluate different training strategies to investigate whether the models can be made more robust to the performative prediction dynamic:

Ideal Training the model using the two most recent observations. At deployment, we do not use g function. Therefore, everything is observed. This strategy ignores the performative prediction problem and can be seen as the upper bound in terms of observability.

Oblivious This model is trained exactly like the 'Ideal' model, using the two most recent observations. However, at deployment, values are observed (or not) based on the g function.

Warm-up One way to avoid the problem of performative prediction is to ignore all the data points affected by the distribution shift. In this training strategy, we train the model using only the warm-up history to predict all future symptom levels. This may not be the best strategy because it does not use recent observations. At deployment, we use the g function.

Random Training using two randomly chosen previous symptom levels as inputs. This is an attempt in making the resulting model robust to unobserved symptom values. When deploying this model, we use the g function.

4.4 Evaluation Metrics

We measure performance using the Mean Absolute Error (MAE), which corresponds to the mean absolute deviation in the symptom level making it interpretable for clinicians to infer and analyse. Let N denote the number of samples used for evaluating the performance, the MAE is given by: $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|$, where x_i is the true value of the ith target observation, and $\hat{x}_i$ is the value predicted by the model for this observation.

In the case of imbalanced classes, mistakes on a minority class can be hidden by good performance on a dominant class. To this end, we focus on achieving good performance on individual classes as well as overall. Let $\mathcal{C}$ denote the set of classes. For each class $c \in \mathcal{C}$, we measure the MAE per class (MAE_c) and combine these MAE_c values using the weighted-MAE (WMAE). The weight of a class gives value to the MAE_c of that class inversely proportional to the number of samples in that class (N_c): $w_c = \frac{\max_{c' \in \mathcal{C}} N_{c'}}{N_c}$. The weight will therefore be of value 1 for the class with the largest amount of data and would be 10 times larger for a class with 10 times fewer data. We can then compute the global error as $\text{WMAE} = \frac{\sum_{c\in\mathcal{C}} w_c \text{MAE}_c}{\sum_{c\in\mathcal{C}} w_c}$.

5 Results

Our results are summarised in table 1. For both symptoms, the 'Ideal' model has the lowest WMAE. The other three models that consider performative prediction where some symptom values may not be observed at deployment, perform worse overall than the 'Ideal' model as indicated by higher WMAE. For predicting pain, WMAE is 29.4% to 32% higher than the 'Ideal' model. For predicting fatigue, WMAE is 38.3% to 40.7% higher than the 'Ideal' model. It is interesting to note that on lower classes ($c < 4$), the models considering performative prediction perform almost equivalent to and in some cases better than the 'Ideal' model. This shows that for some symptom levels, not observing recent values does not affect the prediction. However, for higher symptom levels, the 'Ideal' models clearly perform better. The discrepancy between the WMAE of the 'Ideal' and other models highlights the necessity to account for performative prediction for accurate prediction.

Pain				
	Ideal	Oblivious	Warm-up	Random
MAE_0	0.64	0.59	0.68	0.59
MAE_1	1.1	0.97	1.07	1.08
MAE_2	1.68	1.4	1.62	1.57
MAE_3	1.94	1.72	1.91	1.85
MAE_4	2.14	2.25	2.48	2.34
MAE_5	2.41	2.86	2.99	2.96
MAE_6	2.51	3.1	3.05	3.16
MAE_7	2.68	3.5	3.5	3.57
MAE_8	2.51	3.8	3.83	3.82
MAE_9	2.45	3.78	3.58	3.81
MAE_{10}	4.21	5.26	5.35	5.02
WMAE	**3.09**	4.08	4.07	4.00

Fatigue				
	Ideal	Oblivious	Warm-up	Random
MAE_0	0.67	0.7	0.71	0.7
MAE_1	0.97	0.94	1.01	0.97
MAE_2	1.41	1.3	1.41	1.35
MAE_3	1.66	1.53	1.51	1.58
MAE_4	1.82	1.82	1.96	1.88
MAE_5	2.02	2.32	2.49	2.39
MAE_6	2.12	2.62	2.79	2.7
MAE_7	2.07	2.76	2.96	2.88
MAE_8	2.15	2.98	3.03	3.09
MAE_9	1.97	3.04	2.96	3.03
MAE_{10}	2.38	3.46	3.4	3.49
WMAE	**2.11**	2.92	2.93	2.97

Table 1: Summary of results

6 Conclusion

In this study, we introduced an important healthcare application where we exhibit performative prediction in time series. We described baseline methods to mitigate the problem of performative prediction in time series for cancer symptom prediction. We observe that for some symptom levels, the model is not affected by performative prediction. However, overall performative prediction is an important feature of a problem setting that needs to be dealt with, otherwise, it can lead to higher levels of error. Future work may be required to formulate strategies that are more robust to this problem. We hope that our work will elicit discussion of this interesting problem, as machine learning models become more ubiquitously used to guide decisions by policymakers.

References

[1] S. Bennett and J. Clarkson. Time Series Prediction under Distribution Shift using Differentiable Forgetting. *arXiv*, July 2022. doi: 10.48550/arXiv.2207.11486.

[2] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.*, 2002.

[3] T. P. Coroller, W. L. Bi, E. Huynh, M. Abedalthagafi, A. A. Aizer, N. F. Greenwald, C. Parmar, V. Narayan, W. W. Wu, S. M. de Moura, S. Gupta, R. Beroukhim, P. Y. Wen, O. Al-Mefty, I. F. Dunn, S. Santagata, B. M. Alexander, R. Y. Huang, and H. J. W. L. Aerts. Radiographic prediction of meningioma grade by semantic and radiomic features. *PLoS One*, 12(11):e0187908, Nov 2017.

[4] T. M. Deist, F. J. W. M. Dankers, G. Valdes, R. Wijsman, I.-C. Hsu, C. Oberije, T. Lustberg, J. van Soest, F. Hoebers, A. Jochems, I. E. Naqa, L. Wee, O. Morin, D. R. Raleigh, W. Bots, J. H. Kaanders, J. Belderbos, M. Kwint, T. Solberg, R. Monshouwer, J. Bussink, A. Dekker, and P. Lambin. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Med. Phys.*, 45(7):3449–3459, Jul 2018.

[5] W. Duan, X. He, L. Zhou, L. Thiele, and H. Rao. Combating Distribution Shift for Accurate Time Series Forecasting via Hypernetworks. *arXiv*, Feb. 2022. doi: 10.48550/arXiv.2202.10808.

[6] R. Elshawi, M. H. Al-Mallah, and S. Sakr. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med. Inf. Decis. Making*, 19(1):1–32, Dec 2019.

[7] E. D. Gennatas, A. Wu, S. E. Braunstein, O. Morin, W. C. Chen, S. T. Magill, C. Gopinath, J. E. Villaneueva-Meyer, A. Perry, M. W. McDermott, T. D. Solberg, G. Valdes, and D. R. Raleigh. Preoperative and postoperative prediction of long-term meningioma outcomes. *PLoS One*, 13 (9):e0204161, Sep 2018.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

[9] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access*, 9:109960–109975, Aug 2021.

[10] M. C. Lenert, M. E. Matheny, and C. G. Walsh. Prognostic models will be victims of their own success, unless.... *J. Am. Med. Inform. Assoc.*, 26(12):1645–1650, Dec. 2019. ISSN 1527-974X. doi: 10.1093/jamia/ocz145.

[11] J. Liley, S. R. Emerson, B. A. Mateen, C. A. Vallejos, L. J. M. Aslett, and S. J. Vollmer. Model updating after interventions paradoxically introduces bias. *arXiv*, Oct. 2020. doi: 10.48550/arXiv.2010.11530.

[12] N. Papachristou, D. Puschmann, P. Barnaghi, B. Cooper, X. Hu, R. Maguire, K. Apostolidis, Y. P. Conley, M. Hammer, S. Katsaragakis, K. M. Kober, J. D. Levine, L. McCann, E. Patiraki, E. P. Furlong, P. A. Fox, S. M. Paul, E. Ream, F. Wright, and C. Miaskowski. Learning from data to predict future symptoms of oncology patients. *PLoS One*, 13(12), 2018.

[13] J. C. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative Prediction. *arXiv*, Feb. 2020. doi: 10.48550/arXiv.2002.06673.

[14] V. Podgorelec, P. Kokol, and S. B. et al. Decision trees: An overview and their use in medicine. *In: Journal of Medical Systems 26, 445–463 (2002).*, 2002.

[15] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera. Diagnosis of diabetes mellitus using gradient boosting machine (lightgbm). *Diagnostics*, 11(9), 2021.

[16] H. Seow, P. Tanuseputro, L. Barbera, C. C. Earle, D. M. Guthrie, S. R. Isenberg, R. A. Juergens, J. Myers, M. Brouwers, S. Tibebu, and R. Sutradhar. Development and validation of a prediction model of poor performance status and severe symptoms over time in cancer patients (PROVIEW+). *Palliat. Med.*, 35(9):1713–1723, Oct 2021.

[17] R. Sutradhar, M. Rostami, and L. Barbera. Patient-Reported Symptoms Improve Performance of Risk Prediction Models for Emergency Department Visits Among Patients With Cancer: A Population-Wide Study in Ontario Using Administrative Data. *J. Pain Symptom Manage.*, 58 (5):745–755, Nov 2019.

[18] I. Tomek. Two modifications of cnn. 1976.

[19] G. Valdes, T. D. Solberg, M. Heskel, L. Ungar, and C. B. Simone. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. *Phys. Med. Biol.*, 61(16):6105–6120, Jul 2016.

[20] I. Vaz-Luis, A. Di Meglio, J. Havas, M. El-Mouhebb, P. Lapidari, D. Presti, D. Soldato, B. Pistilli, A. Dumas, G. Menvielle, C. Charles, S. Everhard, A.-L. Martin, P. H. Cottu, F. Lerebours, C. Coutant, S. Dauchy, S. Delaloge, N. U. Lin, P. A. Ganz, A. H. Partridge, F. André, and S. Michiels. Long-Term Longitudinal Patterns of Patient-Reported Fatigue After Breast Cancer: A Group-Based Trajectory Analysis. *J. Clin. Oncol.*, page 1, Mar 2022.

[21] Z. S.-Y. W. Yang Zhao and K. L. Tsui. A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection. *In: Journal of Healthcare Engineering*, 2018.

A Appendix

Visual Representation of the Distribution shift Fig. 2 shows the change in distribution due to performative prediction. The distribution of the train and deployment set are quite similar, making this problem a good candidate for a prediction model. This is the deployment set used for the ''Ideal' model. However, the distribution of the observed values using the 'Oblivious' training strategy shown in the right-most column of the figure has a different distribution from the train set.

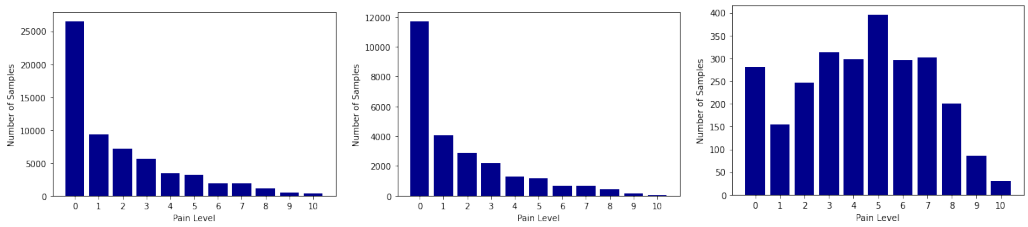

Figure 2: Distribution of samples over classes for pain levels in the train set (left column), test set (middle column) and observed pain levels in the test set (right column)