Performative Prediction in Time Series: A Case Study

Rupali Bhati
Mila & Laval University
rupali.bhati.1@ulaval.ca

Jennifer Jones
Princess Margaret Cancer Centre,
University Health Network

Kristin Campbell
University of British Columbia

David Langelier
University of Toronto

Anthony Reiman
Dalhousie University

Jonathan Greenland
Memorial University of Newfoundland

Audrey Durand[*]
Mila & Laval University

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 prediction 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 patient-reported symptom values collected during their oncology appointments as a time series. We use a decision-tree based model to predict a patient's future symptom values. Based on these predictions, clinicians decide which symptom values will be observed in the future. We propose methods to provide robustness against the problem of performative prediction in time series. Our results characterise how performative prediction may lead to a 29.4% to 40.7% higher error across different symptoms.

1 Introduction

Typically time series prediction (TSP) 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 values for cancer patients, focusing medical care on patients in critical states. We, therefore, aim to predict future symptom values using historical ones. These predictions aid in deciding whether a patient should visit a cancer 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 [Perdomo et al., 2020]. 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 cancer 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 may affect the accuracy of the system.

[*] CIFAR AI Chair

Workshop on Learning from Time Series for Health, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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 prediction*. 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.

2 Related Work

The problem of predicting future symptoms for oncology patients has been tackled with machine learning in many forms, e.g., predicting severe symptoms in cancer patients Seow et al. [2021], Papachristou et al. [2018], Vaz-Luis et al. [2022] and using patient-reported symptom values to improve the performance of models for emergency department visits among cancer patients Sutradhar et al. [2019]. 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. [2020] discuss a typical healthcare setting where post-intervention model updating leads to bias. Lenert et al. [2019] highlight the problem of when users respond to model predictions, downstream characteristics of the data, including the distribution of the outcome, may change and how these problems will need to be mitigated by systematically incorporating interventions into prognostic models to achieve robust performance surveillance of models in clinical use. However, these settings are not modelled as TSP like our setting. TSP can often suffer from a distribution shift. Duan et al. [2022] 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 [2022] deal with TSP under distribution shift using differentiable forgetting. However these solutions may not be used in our setting because the distribution shift is not induced by the decisions made by the model but are rather by the environment.

3 Time Series Performative Prediction

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.

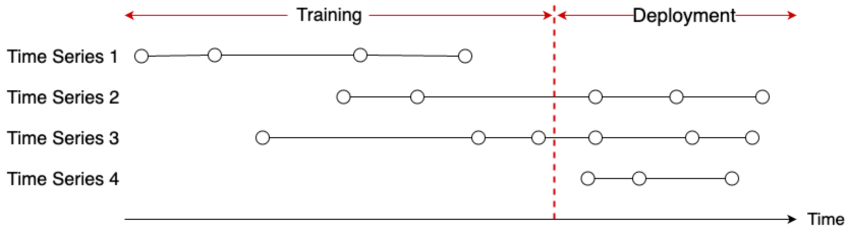


Figure 1: Time series data of symptom values of patients

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 $\mathcal{H}(t_k) = \{(t_1, x_{t_1}), (t_2, x_{t_2}), \dots, (t_{k-1}, x_{t_{k-1}})\}$ denote the history of previous observations available at time $k \geq 1$ for a given time series, where t_i denotes the time of the i-th observation and $1 \leq i < k$. 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_k, a predictive model outputs the prediction $\hat{x}_{t_k} \in \mathcal{X}$ using $\mathcal{H}(t_k)$ and x_{t_k} is observed if $g(\hat{x}_{t_k})$ is true, in which case $\mathcal{H}(t_{k+1}) = \mathcal{H}(t_k) \cup \{(t_k, x_{t_k})\}$. Otherwise, x_{t_k} is not observed and $\mathcal{H}(t_{k+1}) = \mathcal{H}(t_k)$.

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 $k = N + 1$ with the history $\mathcal{H}(t_{N+1}) = \{(t_1, x_{t_1}), (t_2, x_{t_2}), \ldots, (t_N, x_{t_N})\}$.

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 values for cancer patients:

- We split the dataset into the training set and deployment set as shown in Figure 1.
- We train a model on the training set using different training strategies.
- We evaluate the model in the deployment setting simulated using the deployment set. 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 as described below.

We consider a decision function g that acquires the observation when the predicted symptom value exceeds a threshold τ:

$$g(t_k) = \begin{cases} 1, & \text{if} \quad I(\hat{x}_{t_k} \geq \tau) \\ 0, & \text{otherwise} \end{cases}$$

When $\tau = 0$, all the symptom values are acquired, therefore there is full observability. On the other extreme, when $\tau = 10$, we observe only those symptom values that are equal to 10. This would mean we would never observe the other symptom values, making their prediction very difficult. Therefore, the higher the τ, the more difficult it is to track the problem. We change the value of τ to gauge the robustness of the model. In our setting, we use $\tau = 5$ as a medium-level difficulty task.

4.1 Data

The data used in this study has been collected at the Princess Margaret Cancer Centre from 2013 to 2019 for 14586 patients with breast, colorectal, lymphoma, and head and neck cancer types. Post cancer diagnosis, a patient visits the doctor or clinician either when they are undergoing initial diagnostic testing and treatment planning, undergoing treatments, or coming for follow-up surveillance appointments. During a patient's visit to a cancer center, they are asked to fill an electronic survey (filled on iPads) documenting their symptoms using DART (Distress Assessment and Response Tool) [Li et al., 2016]. To predict future symptom values, we use demographic information of patients like sex and age from the cancer registry data, along with surgery and historical symptom value data from the surveys. We leverage the historical symptom values to calculate their gradient of change and use it as an input feature. As the average number of observations per patient is 4.76 and the minimum number of observations required to calculate the gradient of change in symptom value is two, we use the gradient between the two previous symptom values 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 which were measured on the Edmonton Symptom Assessment Scale-Revised [Watanabe et al., 2012]. 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 in the dataset, 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. This class imbalance problem makes the prediction problem more complex.

4.2 Predictive model

Decision trees (DTs) are widely used in healthcare [Podgorelec et al., 2002] and cancer-related applications [Valdes et al., 2016, Deist et al., 2018, Coroller et al., 2017, Gennatas et al., 2018], making them a suitable predictive model for this problem. Among several DT based algorithms, we consider LightGBM [Ke et al., 2017] because it uses gradient descent to minimize the loss when adding new models which is more beneficial as compared to other DT algorithms that either train the different models separately before aggregating them or do not use multiple models at all.

Inspired by previous works [Yang Zhao and Tsui, 2018, Khushi et al., 2021], we rely on Synthetic Minority Oversampling TEchnique (SMOTE) [Chawla et al., 2002] to deal with the class imbalance problem which generates synthetic samples of the minority classes so that the trained model sees the same number of data points of these classes as of 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 [Tomek, 1976] 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 set the threshold $\tau = 0$ in the 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 values. This may not be the best strategy because it does not use recent observations. At deployment, we use the g function.

Random Here, we train using two randomly chosen previous symptom values 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:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{x}_i| \,, \tag{1}$$

where x_i is the true value of the i^{th} 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) as described in eqn. 3. The weight of a class w_c as described in eqn 2 gives value to the MAE$_c$ inversely proportional to the number of samples in that class (N_c). The weight will 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.

$$w_c = \frac{\max_{c' \in \mathcal{C}} N_{c'}}{N_c}, \tag{2}$$

$$\text{WMAE} = \frac{\sum_{c \in \mathcal{C}} w_c \text{MAE}_c}{\sum_{c \in \mathcal{C}} w_c} \tag{3}$$

At deployment, the performance is measured on all symptom values, whether observed or not. However, only the observed symptom values are used as inputs to the model for future predictions.

5 Results

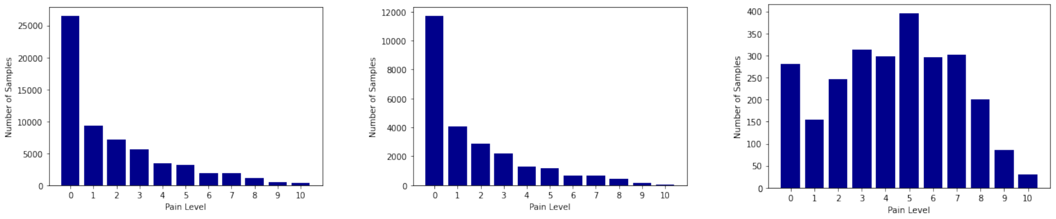
Our results are summarised in table 1. As expected, for both symptoms, the 'Ideal' model has the lowest WMAE. The other three models that consider performative prediction where some symptom values may not be observed at deployment, perform worse overall than the 'Ideal' model as indicated by higher WMAE. When predicting pain, WMAE is 29.4% to 32% higher than the 'Ideal' model. When 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 values, not observing recent values does not affect the prediction. However, for higher symptom values, 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. Moreover, achieving robustness to the performative dynamics is not trivial.

	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

Visual Representation of the Distribution shift Fig. 2 shows the change in distribution due to performative prediction. Fig. 2a and 2b show the distribution of the train and test set, respectively. Their distributions are quite similar, making this problem a good candidate for a prediction model. Fig. 2b is the deployment set used for the 'Ideal' model. However, the distribution of the observed values using the 'Oblivious' training strategy shown in fig. 2c has a different distribution from the train set which makes the prediction non-trivial.



(a) Train set

(b) Test set

(c) Observed values in test set

Figure 2: Distribution of samples over classes for pain levels in the (a) train set, (b) test set, and (c) observed pain levels in the test set

6 Conclusion

In this study, we introduced an important healthcare application where we exhibit performative prediction in time series. We described baseline methods to provide robustness against the problem of performative prediction in time series for cancer symptom prediction. We observe that for some symptom values, the model is not affected by performative prediction. We characterise the extent that performative prediction can impact evaluation can be up to 40.7% increase in error values. Indeed, performative prediction is an important feature of the problem setting that needs to be dealt with. 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.

Acknowledgments and Disclosure of Funding

We would like to thank Mehrnaz Mofakhami and Joey Bose for our insightful discussions. We would also like to thank CIFAR for the CCAI Chair funding and CIHR CCS #706699 and CIHR CAS #168602.

References

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

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.

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.

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.

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.

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.

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.

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.

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.

M. Li, A. Macedo, S. Crawford, S. Bagha, Y. W. Leung, C. Zimmermann, B. Fitzgerald, M. Wyatt, T. Stuart-McEwan, and G. Rodin. Easier Said Than Done: Keys to Successful Implementation of the Distress Assessment and Response Tool (DART) Program. *Journal of Oncology Practice*, 12 (5):1–e526, Apr 2016. ISSN 1554-7477. doi: 10.1200/JOP.2015.010066.

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.

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.

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

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.

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.

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.

I. Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, 1976.

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.

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.

S. M. Watanabe, C. L. Nekolaichuk, and C. Beaumont. The edmonton symptom assessment system, a proposed tool for distress screening in cancer patients: development and refinement. *Psycho-Oncology*, 21(9):977–985, 2012.

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.