Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240728

Predicting Patient No-Shows: Situated Machine Learning with Imperfect Data

Christopher GYLDENKÆRNE[a], Jakob Grue SIMONSEN[b], Gustav FROM[c] and Morten HERTZUM[a,1]
[a] *Department of People and Technology, Roskilde University, Denmark*
[b] *Department of Computer Science, University of Copenhagen, Denmark*
[c] *Digestive Disease Center, Bispebjerg Hospital, Denmark*

Abstract. Patients who do not show up for scheduled appointments are a considerable cost and concern in healthcare. In this study, we predict patient no-shows for outpatient surgery at the endoscopy ward of a hospital in Denmark. The predictions are made by training machine leaning (ML) models on available data, which have been recorded for purposes other than ML, and by doing situated work in the hospital setting to understand local data practices and fine-tune the models. The best performing model (XGBoost with oversampling) predicts no-shows at sensitivity = 0.97, specificity = 0.66, and accuracy = 0.95. Importantly, the situated work engaged local hospital staff in the design process and led to substantial quantitative improvements in the performance of the models. We consider the results promising but acknowledge that they are from a single ward. To transfer the no-show models to other wards and hospitals, the situated work must be redone.

Keywords. Machine learning, participatory design, patient no-shows, healthcare

1. Introduction

Electronic health records (EHRs) store large amounts of data about patients, including their medical history, appointment schedules, and demographics. These data are recorded to enable sharing among healthcare providers to secure the delivery of quality care, but they are also increasingly seen as providing an untapped potential for machine learning (ML) to support healthcare by reducing costs and aiding in decision making [1]. However, typical EHR data are known to be incomplete and suffer from validity issues, for example because data-entry practices vary even within a single hospital or unit and because exceptions to the cases and workflows presumed by EHRs necessitate workarounds [2]. To be applicable in healthcare, ML models must be robust to imperfect data. In this paper, we investigate *the effect of situated work to understand local data practice on the performance of several ML models for predicting patient no-shows*.

Patients who do not show up for their clinical appointments do not receive treatment for their illnesses. In addition, no-shows are a source of wasted time and resources in the healthcare system [3] and of stress and job dissatisfaction among healthcare providers [4]. Reviews of no-show rates find that they vary greatly but average about 23% [5]. Thus, no-shows are a considerable cost and concern. The case considered in the present

[1] Corresponding author: M. Hertzum, Universitetsvej 1, Bldg 44.2, Roskilde, Denmark; E-mail: mhz@ruc.dk.

study involves predicting no-shows among future outpatients at the endoscopy ward of a Danish hospital. We trained off-the-shelf ML algorithms on hospital data and tuned the resulting models to parameters relevant to the prediction of no-shows for endoscopy surgery. The models provide an add-on to EHRs and utilize available EHR data to predict patient no-shows, thereby enabling staff to contact these patients ahead of their appointment to reduce no-show rates. At present, the staff lack guidance about which patients to contact and do not have the resources to contact all patients. The included ML algorithms are random forest (RF), support vector machine (SVM), and XGBoost (XGB).

2. Methods

The study was approved by the case hospital administration. All data used in the study were extracted by the hospital quality assurance department, which also removed information identifying the patients before the data were handed over to us. The ML models were trained, tuned, and tested on the deidentified data. The full process was informed by participatory design [6,7] and spanned a one-year period with five steps.

First, we searched the research literature for factors that were known to cause patients to miss their clinical appointments. Many of these factors could be extracted from the hospital EHR, including no-show history, number of visits to department, weekday of planned visit, time of day, patient age, and patient sex. Factors not in the hospital EHR or excluded due to local data regulations (i.e., GDPR) included ethnicity and substance abuse. The dataset extracted by the hospital quality assurance department gave the available factors for 8840 appointments, with a no-show rate of 8%.

Second, we trained RF, SVM, and XGB models on these data, which we will refer to as the raw data. The RF algorithm [8] combines several classification trees for training and prediction. The SVM algorithm [9] is a discriminate classifier that nonlinearly maps cases to a high-dimensional feature space. The XGB algorithm [10] is based on gradient tree boosting and designed to handle missing values. All models were trained and optimized to the best performance on the selected hyper-parameters through a process of five-fold cross-validation.

Third, we involved staff in the endoscopy ward, EHR data management team, and quality assurance department in the design process. The general aim of these activities was to inform the model building and increase the attention and buy-in to the models and the requirements for realizing their potential. Specifically, the medical secretaries and nurses in the endoscopy ward provided pertinent input. They were the ones who entered the EHR data about appointment scheduling, cancellations, and no-shows. It became apparent that the information was in many cases incomplete because our study was the first initiative where there was a real need for the ward to pay close attention to its registration practice. A particularly important piece of information concerned the status of appointments. This information was crucial to the training of the ML models because we used it for determining whether patients did or did not show up. Due to overlapping and ambivalent status categories, about half of the secretaries simply refrained from registering the status of some appointments. Fourteen days after the appointment date, the EHR automatically set the status of these appointments to “canceled”. Thus, a cancelled appointment in the dataset was not always a real cancellation. In addition, the endoscopy physicians explained that a number of the patients who showed up were unprepared. They had not complied with the three-day cleansing program (i.e., diet restrictions and bowel cleansing medication) necessary before a colonoscopy. These

appointments had to be rebooked, and the physicians requested that appointments with the status "rebooked" be counted as no-shows. The physicians also requested that cancellations made less than five days before the appointment date were counted as no-shows because such late cancelations could not be rebooked for other patients, who would not have sufficient time to complete the three-day cleansing program.

Fourth, the insights from the staff's participation in the design process enabled us to clean the data and contextualize the models. Rebooked appointments and late cancellations were reclassified as no-shows. All appointments with incomplete status information were excluded due to the uncertainty about the actual final appointment status. For the same reason, all unfinished appointments within the last 14 days of the dataset were excluded. With these changes, the final dataset consisted of 2440 appointments, including 197 (8.07%) no-shows.

Fifth, the dataset was imbalanced in that it contained far fewer no-shows than show-ups. We trained models with the imbalanced dataset, but to compensate for the imbalance we also trained models with the synthetic minority oversampling technique (SMOTE) applied [11]. SMOTE oversampling balances a dataset by synthetically increasing the number of cases in the minority class (i.e., the no-shows). In each fold of the five-fold cross-validation, we applied SMOTE on the training set only, not on the test set.

3. Results

Tables 1 and 2 show the performance of the models on the raw (i.e., uncleaned and uncontextualized) data and final (i.e., cleaned and contextualized) data, respectively. All results are averages across the five-fold cross-validation. Overall, the models trained on the final and oversampled data perform substantially better than the models trained on the raw data. In particular, the models trained on the raw data have inferior sensitivity (which is, by definition, the same as recall) for the minority class of patient no-shows.

Table 1. Results for original uncleaned and uncontextualized data (8840 observations).

Model	Class	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's κ
RF	Show-up	1.00	0.96	1.00	0.22	0.98		
	No-show	0.22	1.00	0.22	1.00	0.36	0.96	0.34
SVM	Show-up	0.98	0.96	0.98	0.60	0.97		
	No-show	0.60	0.73	0.60	0.98	0.66	0.92	0.09
XGB	Show-up	1.00	0.95	1.00	0.00	0.97		
	No-show	0.00	0.00	0.00	1.00	N/A	0.95	0.00

N/A: the F-measure is undefined in this case because both recall and precision are exactly zero.

For the no-show class, the XGB model trained on the final data performs best on most performance metrics, both with and without oversampling. Specifically, the high F-measure of the oversampled XGB model on patient no-shows indicates a good balance between precision and recall. The oversampled RF model demonstrates similar performance on sensitivity but at the cost of slightly worse specificity and precision. Conversely, oversampling has severe adverse effects on the SVM model, including performance drops across several metrics for both majority and minority classes, as well as abysmal performance for the no-show class. The XGB model achieves the maximal value of Cohen's kappa (0.64) across both raw and final data, suggesting that this model is better at distinguishing between patient no-shows and patients showing up. However,

the minimal kappa value is also obtained for XGB – on the raw data. This result indicates that the data cleaning and context enrichment benefit the XGB model in particular.

Table 2. Results after data cleaning and contextualization (2440 observations).

Model	Class	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's κ
RF	Show-up	0.96	0.99	0.96	0.79	0.97		
	No-show	0.79	0.48	0.79	0.96	0.60	0.95	0.57
RF *	Show-up	0.56	0.70	0.56	0.97	0.62		
	No-show	0.97	0.95	0.97	0.56	0.96	0.93	0.59
SVM	Show-up	0.97	0.55	0.97	0.13	0.70		
	No-show	0.13	0.78	0.13	0.97	0.23	0.57	0.10
SVM *	Show-up	0.92	1.00	0.92	0.00	0.96		
	No-show	0.00	0.00	0.00	0.92	0.00	0.92	0.00
XGB	Show-up	0.97	0.98	0.97	0.68	0.97		
	No-show	0.68	0.60	0.68	0.97	0.60	0.95	0.61
XGB *	Show-up	0.66	0.68	0.66	0.97	0.67		
	No-show	0.97	0.97	0.97	0.66	0.97	0.95	0.64

* with SMOTE oversampling

To identify the data features that are important for the predictive models, we conducted a leave-one-out feature importance analysis in each fold of the cross-validation. A prior history of no-shows is by far the most important feature in correctly predicting no-shows. Other influential features are the duration of the planned appointment, the number of prior appointments, appointments late in the day, and day of the week. The patient's age and sex are weak no-show indicators in our data.

4. Discussion

The RF and XGB models perform well on all reported metrics, whereas SVM achieves high accuracy but underperforms on most of the other metrics. Specifically, the SVM model performs poorly with SMOTE oversampling. For both RF and XGB, the oversampling of the minority class produces a clear performance increase in the prediction of no-shows. The size of this performance increase is perhaps surprising, given the known ability of these models to handle imbalanced datasets well [10].

For an endoscopy ward, the typical use of one of the developed models would be to contact predicted no-shows ahead of their appointment to nudge them to show up (e.g., phoning them a day in advance). For the non-oversampled XGB model, the sensitivity (0.68) means that two thirds of all actual no-shows would be correctly identified and thus contacted, and the precision (0.60) means that 60% of the patients contacted would be actual no-shows if they were not nudged to show up. For the oversampled XGB model, both sensitivity and precision increase to 0.97. That is, almost all no-shows would be correctly predicted and very few resources would be wasted on contacting patients who would show up anyway. For the concrete ward, if nudging resulted in 50% of the contacted patients either showing up or appropriately canceling their appointment, it would result in $197 \cdot 0.97 \cdot 0.50 = 95$ more patients showing up for appointments, thus decreasing the no-show rate from 8% to 4%. To reduce the no-show rate further, the ward may consider simple workflow changes, such as only planning appointments on Fridays or late in the day for patients with a history of showing up.

In this study, we present results from a single ward at a single hospital. We stress two points regarding generalizability to other wards and hospitals. First, the models

developed in this study use available data features, which will likely differ in both format and registration practice across wards and hospitals. Second, the work performed on site to gain a situated understanding of the local data management and work practices clearly makes a quantitative difference in the performance of the trained models. The models trained on the raw data with all available features display inferior performance (Table 1) compared to the models trained on the final data (Table 2), highlighting the importance of situated work. Apart from leading to model improvements, the situated work also uncovered phenomena that could adversely affect the successful organizational implementation of the no-show models [12], including inconsistent registration practice and limited attention to the importance of data entry to model outputs.

5. Conclusions

The – admittedly tentative – conclusion from the present study is that no-shows can be predicted with high accuracy across different wards and hospitals by off-the-shelf ML models trained on available data. However, the models that perform well for the particular ward on which they are trained are not likely to perform well in other wards but will require situated work and retraining with local data. Given the relative ease with which off-the-shelf ML algorithms can be trained, we anticipate that the main cost of transferring the no-show models to other wards will be the time spent by data scientists and local staff on gaining a situated understanding of the data practice of each ward, rather than the time required to train the ML models themselves.

References

[1] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: Review, opportunities and challenges. Brief. Bioinform. 2018; 19(6):1236–1246, doi:10.1093/bib/bbx044.

[2] Cabitza F, Ciucci D, Rasoini R. A giant with feet of clay: On the validity of the data that feed machine learning in medicine. In: Cabitza F, Batini C, Magni M, editors. Organizing for the Digital World. Cham: Springer; 2019. pp. 121–136, doi:10.1007/978-3-319-90503-7_10.

[3] Berg BP, et al. Estimating the cost of no-shows and evaluating the effects of mitigation strategies. Med. Decis. Mak. 2013; 33(8):976–985, doi:10.1177/0272989X13478194.

[4] Firth-Cozens J. Interventions to improve physicians' well-being and patient care. Soc. Sci. Med. 2001; 52(2):215–222, doi:10.1016/S0277-9536(00)00221-5.

[5] Dantas LF, Fleck JL, Cyrino Oliveira FL, Hamacher S. No-shows in appointment scheduling – A systematic literature review. Health Policy. 2018; 122(4):412–421, doi:10.1016/j.healthpol.2018.02.002.

[6] Bødker K, Kensing F, Simonsen J. Participatory IT design: Designing for business and workplace realities. Cambridge, MA: MIT Press; 2004.

[7] Hertzum M, Simonsen J. Effects-driven IT development: Specifying, realizing, and assessing usage effects. Scand. J. Inf. Syst. 2011; 23(1):3–28, https://aisel.aisnet.org/sjis/vol23/iss1/1.

[8] Breiman L. Random forests. Mach. Learn. 2001; 45(1):5–32, doi:10.1023/A:1010933404324.

[9] Cortes C, Vapnik C. Support-vector networks. Mach. Learn. 1995; 20(3):273–297, doi:10.1007/BF00994018.

[10] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. pp. 785–794, doi:10.1145/2939672.2939785.

[11] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 2002; 16:321–357, doi:10.1613/jair.953.

[12] Gyldenkærne C, Hansen JU, Hertzum M, Mønsted T. Innovation tactics for implementing an ML application in healthcare: A long and winding road. Int. J. Hum. Comput. Stud. 2024; 181:article 103162, doi:10.1016/j.ijhcs.2023.103162.