Predicting Patient No-Shows Using Machine Learning to Optimize Clinic Scheduling

Lucas De Smet[1]*, Marie Dubois[2], Johan Vermeulen[3], Sofie Claes[4], Thomas Lambert[5]

[1]Department of Electrical Engineering, KU Leuven, Belgium
Email: lucas.desmet@kuleuven.be
[2]Department of Mechanical Engineering, Ghent University, Belgium
Email: marie.dubois@ugent.be
[3]Department of Computer Science, Université catholique de Louvain, Belgium
Email: johan.vermeulen@uclouvain.be
[4]Department of Aerospace Engineering, Vrije Universiteit Brussel, Belgium
Email: sofie.claes@vub.be
[5]Department of Biomedical Engineering, University of Liège, Belgium
Email: thomas.lambert@uliege.be

Abstract

Patient no-shows for scheduled appointments represent a critical operational and financial challenge for healthcare providers, leading to wasted resources, reduced access to care, and increased wait times. This paper investigates the application of machine learning (ML) to predict the risk of a patient missing their appointment. By analyzing historical electronic health record (EHR) and scheduling data, predictive models can identify patients with a high probability of no-show, enabling clinics to implement targeted interventions such as reminder calls, overbooking strategies, or offering slots to waitlisted patients. This research details the development of several classification algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, to forecast no-show events. The results demonstrate that ensemble methods can accurately identify high-risk patients, providing a data-driven foundation for improving clinic efficiency, maximizing resource utilization, and enhancing patient access to care.

Keywords
Patient No-Show, Machine Learning, Healthcare Operations, Scheduling Efficiency, Predictive Analytics, Resource Optimization.

1. Introduction

The efficiency of outpatient clinic operations is severely hampered by patient no-shows—appointments where the patient fails to arrive without prior cancellation. The ramifications are multifaceted: valuable clinical time and resources are wasted, revenue is lost, and other patients face longer wait times for appointments, potentially exacerbating health disparities [1]. Traditional approaches to mitigating no-shows, such as blanket reminder calls or 短信, are inefficient and fail to address the underlying risk factors that vary significantly across a patient population [2].

Predictive analytics offers a sophisticated solution to this persistent problem. By leveraging machine learning algorithms on historical administrative and clinical data, healthcare systems can move from a one-size-fits-all approach to a targeted, risk-based strategy [3]. EHRs contain a wealth of information that can be indicative of no-show risk, including demographic details, previous appointment history, geographic factors, and clinical characteristics [4]. Modeling these complex, non-linear relationships allows clinics to proactively identify which patients are most likely to miss their appointment.

This enables precise interventions. For instance, a patient predicted to be high-risk could receive more intensive reminder protocols (e.g., phone calls plus SMS plus mail), while a low-risk patient might only require a single automated reminder [5]. This optimizes staff effort and intervention costs. This paper will explore the development and validation of a predictive model for patient no-shows. It will review existing literature, detail the methodology for feature engineering and model selection, present performance results, and discuss the practical implications of deploying such a model to improve operational outcomes in a real-world clinical setting.

2. Literature Review

The challenge of patient no-shows has been extensively studied, with early research focusing on identifying correlative factors using traditional statistical methods like logistic regression. These studies consistently found associations between no-show rates and variables such as age, travel distance, lead time (the time between scheduling and the appointment), and a history of previous no-shows [6][7]. While informative, these models often had limited predictive power and could not easily capture complex interaction effects within the data.

The adoption of machine learning has marked a significant advancement in the field. More sophisticated algorithms capable of handling high-dimensional data have been employed to improve prediction accuracy. Studies using decision trees and Random Forests have demonstrated the ability to rank feature importance, often identifying a patient's prior appointment behavior as the most potent predictor of future no-shows [8][9]. This aligns with the behavioral principle that past behavior is a strong indicator of future actions.

Research has expanded to include a wider array of features, including weather conditions on the day of the appointment [10], seasonality [11], and specific socioeconomic proxies derived from patient zip codes [12]. The integration of natural language processing (NLP) to analyze the content of reminder communications is also an emerging area of interest [13]. The consensus in recent literature is that ensemble methods like Gradient Boosting Machines (e.g., XGBoost) consistently outperform traditional regression models for this classification task [14].

However, the literature also highlights critical challenges. A major concern is model fairness and algorithmic bias; a model trained on historical data may perpetuate existing disparities by unfairly targeting vulnerable populations (e.g., those from lower socioeconomic backgrounds) if not carefully designed and audited [15]. Furthermore, the operational integration of these models into existing workflow systems presents a technical hurdle [16]. The ultimate goal is not just to predict no-shows

DOI: https://doi.org/10.5281/zenodo.17075883

but to translate these predictions into actionable strategies that improve efficiency without compromising patient trust or access to care [17].

3. Methodology

This study employed a standard machine learning workflow to develop a binary classification model for predicting patient no-shows.

3.1 Data Source and Collection

De-identified data was extracted from the scheduling and EHR systems of a large multi-specialty outpatient clinic over a 24-month period. The dataset included all completed appointments across various specialties. The target variable was a binary label indicating whether an appointment was a "show" or "no-show."

3.2 Feature Engineering and Preprocessing

A wide range of features was engineered from the raw data:

Patient History: Number of previous no-shows, number of previous shows, show rate.
Appointment Characteristics: Lead time, day of the week, time of day, clinic specialty, type of appointment (e.g., new vs. follow-up).
Patient Demographics: Age, gender, insurance type.
Environmental: Season (derived from date).
Missing data for categorical variables was imputed with a new "missing" category, while numerical used the median. All categorical variables were one-hot encoded.

3.3 Feature Selection

To avoid overfitting, feature importance was calculated using a Random Forest algorithm. The top 20 most important features were selected for the final model training. Correlation analysis was also performed to remove highly correlated redundant features.

3.4 Model Development and Training

Three algorithms were selected for comparison: Logistic Regression (baseline), Random Forest, and Gradient Boosting (XGBoost). The dataset was split temporally, using the first 18 months for training and the subsequent 6 months for testing. This ensures the model is evaluated on future data, simulating a real-world deployment. Class imbalance was addressed using SMOTE (Synthetic Minority Over-sampling Technique) within the training folds during cross-validation.

3.5 Model Evaluation

The final models were evaluated on the held-out test set. Performance was assessed using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), precision, recall, and F1-score. The cost-effectiveness of interventions based on model predictions was also analyzed.

4. Results

The machine learning models demonstrated a strong ability to predict patient no-shows, significantly outperforming a baseline model that assumed all patients would show. The dataset contained 245,780 appointments with a no-show rate of 12.8%.

DOI: https://doi.org/10.5281/zenodo.17075883

The Gradient Boosting (XGBoost) model achieved the highest performance on the temporal test set, with an AUC-ROC of 0.87. The Random Forest model also performed well (AUC-ROC = 0.84), while the Logistic Regression baseline was the least accurate (AUC-ROC = 0.72). The key evaluation metrics for the XGBoost model at a threshold chosen to maximize the F1-score are summarized in Table 1.

Table 1: Performance Metrics of the XGBoost Model on the Test Set.

Metric	Score
AUC-ROC	0.87
Precision	0.73
Recall (Sensitivity)	0.68
F1-Score	0.70
Accuracy	0.91

Analysis of feature importance revealed that the most powerful predictors were related to past behavior: the patient's number of previous no-shows and their historical show rate. Other significant factors included lead time (longer lead times correlated with higher risk), patient age (younger patients were higher risk), day of the week (Mondays and Fridays had higher no-show rates), and insurance type.

5. Discussion

The high AUC-ROC score confirms that machine learning can effectively stratify patients based on their no-show risk. This allows clinic managers to move from inefficient blanket policies to targeted, cost-effective interventions aimed at the ~20% of patients who constitute the highest risk group.

5.1 Interpretation of Findings

The feature importance analysis is highly intuitive and aligns with existing literature and operational experience. The strong predictive power of historical attendance behavior suggests that patient habits are a primary driver. By identifying these high-risk patients, clinics can tailor their engagement strategies, potentially improving overall show rates and operational throughput.

5.2 Limitations and Challenges

This study has limitations. The model was trained on data from a single healthcare network, which may limit its generalizability to other populations with different demographics and behaviors. Furthermore, the model does not include real-time contextual data like day-of-appointment weather or traffic conditions, which could improve accuracy. The most significant challenge is the ethical

DOI: https://doi.org/10.5281/zenodo.17075883

consideration of fairness; the model must be audited to ensure it does not systematically bias interventions against vulnerable groups.

5.3 Conclusion and Future Work

In conclusion, this research provides a robust framework for using predictive analytics to tackle the operational problem of patient no-shows. By enabling proactive and personalized patient engagement, these models can significantly enhance clinic efficiency and patient access. Future work will focus on implementing a fairness audit of the model, integrating real-time data feeds, and conducting a prospective study to measure the actual impact on show rates and resource utilization when the model is deployed in a live clinical environment.

6. Conclusion

The persistent challenge of patient no-shows necessitates a shift from reactive to proactive strategies. This research demonstrates that machine learning models, particularly Gradient Boosting, can accurately predict no-show risk by identifying complex patterns in historical EHR and scheduling data. This capability empowers healthcare providers to optimize their resource allocation by implementing targeted interventions for high-risk patients, thereby improving clinic efficiency, reducing financial waste, and enhancing overall patient access to care. Successful implementation must be coupled with careful attention to ethical considerations to ensure that these data-driven tools promote equity and fairness within the healthcare system.

References

[1] Dantas, L. F., Fleck, J. L., Oliveira, F. L. C., & Hamacher, S. (2018). No-show rate in the outpatient clinic: a systematic review. Health Services Research, 53(1), 1-18.

[2] Hasvold, P. E., & Wootton, R. (2011). Use of telephone and SMS reminders to improve attendance at hospital appointments: a systematic review. Journal of Telemedicine and Telecare, 17(7), 358-364.

[3] Daggy, J., Lawley, M., Willis, D., et al. (2010). Using no-show modeling to improve clinic performance. Health Informatics Journal, 16(4), 246-259.

[4] Huang, Y., & Hanauer, D. A. (2014). Patient no-show predictive model development. AMIA Annual Symposium Proceedings, 2014, –.

[5] Norris, J. B., Kumar, C., & Chand, S. (2014). The effectiveness of automated appointment reminders: a systematic review. Journal of Medical Systems, 38(2), 1-9.

[6] Bean, A. G., & Talaga, J. (1995). Predicting appointment breaking. Journal of Health Care Marketing, 15(1), 29-34.

[7] George, A., & Rubin, G. (2003). Non-attendance in general practice: a systematic review. Family Practice, 20(2), 210-218.

[8] Alaeddini, A., Yang, K., Reddy, C., & Yu, S. (2011). A probabilistic model for predicting the probability of no-show in hospital appointments. Health Care Management Science, 14(2), 146-157.

[9] Mieloszyk, R. J., Hall, C. S., & Badawi, O. (2018). Predicting no-shows in healthcare: a systematic review. IEEE International Conference on Healthcare Informatics (ICHI), –.

DOI: https://doi.org/10.5281/zenodo.17075883

[10] Abdullah, M., Alshami, A., Alhajjaj, H., & Al-Mulla, F. (2021). The impact of weather on patient no-show rate: a machine learning approach. Informatics in Medicine Unlocked, 24, 100565.
[11] Cosgrove, M. (2018). Seasonal variation in patient no-show rates. Journal of Medical Practice Management, 33(5), 315-318.
[12] Moein, S., & Ahmadi, M. (2019). Using machine learning to predict patient no-shows: a case study of a private clinic. Health and Technology, 9(5), 749-759.
[13] Sunny, Md Nagib Mahfuz, et al. "Optimizing healthcare outcomes through data-driven predictive modeling." *Journal of Intelligent Learning Systems and Applications* 16.4 (2024): 384-402.
[14] Munsell, M., Velasco, C., & Gao, J. (2019). Predicting patient no-shows using machine learning: a comparative study. Journal of Data Science, 17(3), 567-584.
[15] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
[16] Muse, S., & Omer, T. (2020). Operationalizing machine learning models for no-show prediction: a framework for integration. Journal of the American Medical Informatics Association, 27(9), 1441-1445.
[17] Mera, D. M., & Ramirez, A. (2020). Ethical implications of predictive analytics for patient no-shows. Cambridge Quarterly of Healthcare Ethics, 29(4), 1-12.

DOI: https://doi.org/10.5281/zenodo.17075883