Investigating Poor Performance Regions of Black Boxes: LIME-based Exploration in Sepsis Detection

Mozhgan Salimiparsa[1], Surajsinh Parmar[1,*], San Lee[1], Choongmin Kim[2], Yonghwan Kim[2] and Jang Yong Kim[3]

[1]SpassMed Inc., Toronto, Canada
[2]Spass Inc., Seoul, Korea
[3]St. Mary's Hospital, Seoul, Korea

Abstract

Interpreting machine learning models remains a challenge, hindering their adoption in clinical settings. This paper proposes leveraging Local Interpretable Model-Agnostic Explanations (LIME) to provide interpretable descriptions of black box classification models in high-stakes sepsis detection. By analyzing misclassified instances, significant features contributing to suboptimal performance are identified. The analysis reveals regions where the classifier performs poorly, allowing the calculation of error rates within these regions. This knowledge is crucial for cautious decision-making in sepsis detection and other critical applications. The proposed approach is demonstrated using the eICU dataset, effectively identifying and visualizing regions where the classifier underperforms. By enhancing interpretability, our method promotes the adoption of machine learning models in clinical practice, empowering informed decision-making and mitigating risks in critical scenarios.

Keywords

Performance Analysis, LIME, Model Explanation, Sepsis Prediction

1. Introduction

Machine learning has exhibited impressive achievements in diverse fields, including healthcare [1]. The complexity of these models, however, creates challenges for their adoption in healthcare [2]. To address this issue, eXplainable AI (XAI) has been introduced, enabling machine learning models to provide explanations for their predictions. Model explainability is essential for gaining a deeper understanding of a model's decision-making process [3]. In critical domains such as sepsis detection [4] in the ICU, where incorrect predictions can result in fatal consequences, the reliability of these models is of utmost significance. This paper aims to tackle a specific aspect of the interpretability challenge associated with these models, specifically the identification

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.
mozhgan.salimiparsa@spassmed.ca (M. Salimiparsa); suraj.parmar@spassmed.ca (S. Parmar); sanlee@spassmed.ca (S. Lee); cmkim@spass.ai (C. Kim); kyh@spass.ai (Y. Kim); vasculakim@catholic.ac.kr (J. Y. Kim)
0000-0002-0162-032X (M. Salimiparsa); 0009-0005-5463-0563 (S. Parmar); 0009-0001-6488-7988 (S. Lee); 0000-0001-5812-889X (C. Kim); 0000-0003-3022-6847 (Y. Kim); 0000-0001-8437-9254 (J. Y. Kim)

and explanation of scenarios in which black box predictive models fail or exhibit unexpected performance.

Examining instances in which machine learning models exhibit deviations from their usual performance holds significant importance. These insights empower decision-makers to exercise caution in deploying models in situations where their predictions are prone to errors, thereby mitigating potential adverse consequences. Previous research endeavors have primarily centered on assessing the overall performance of these models through the adoption of evaluation metrics and methodologies aimed at gauging their reliability [5, 6]. W. Duivesteijn et al. [7] present an evaluation method that assesses the performance of a classifier, highlighting subspaces where the classifier excels or struggles in classification tasks, however, the method's applicability is limited to binary datasets and lacks model agnosticism. L. Torgo et al. [8] propose approaches that aim to offer interpretable descriptions of expected performance; however, the proposed visualization may not be well-suited when dealing with a high number of features. This paper provides an analysis by focusing on the identification of specific regions where the models exhibit significant deviations from their usual performance. The identification of these regions empowers healthcare practitioners to make informed decisions by exercising caution when relying on the model. Additionally, these findings offer valuable insights that can guide the development of potential strategies aimed at improving and refining the model's overall performance [9, 8].

To achieve this, we propose an analytical approach that combines visual techniques to identify regions in the input space where the models' performance significantly diverges from their average performance. This visualization empowers users to grasp how various values of a particular predictor impact the models' performance.

2. Methodology

In this study, we adopted a modified visualization approach inspired by L. Torgo et al. [8] to identify the regions where a black-box model exhibits poor performance. L. Torgo et al. utilized the confusion matrix and cross-validation, and employed error distribution plots for each individual feature to demonstrate areas of inadequate model performance. However, we recognized the challenge of visual clutter arising from a large number of features. To address this limitation, we focused our analysis on identifying recurrent conditions associated with misclassifications, rather than visualizing misclassifications for each individual feature. To achieve this, we employed a rule extraction method, specifically LIME (Local Interpretable Model-agnostic Explanations) [11]. LIME is a model-agnostic explainability technique that assigns importance weights to features, indicating their contribution to individual predictions. We applied LIME to misclassified data samples, allowing us to pinpoint the specific features responsible for incorrect predictions made by the classifier. This process was performed for each misclassified sample, enabling us to accumulate the features with high importance. We then intersected these features, focusing on those that consistently appeared as contributing factors to misclassifications. This allowed us to discern regions or intervals in which the classifier demonstrated poor performance and was prone to misclassification. We calculated the error rates within these regions by examining how often instances with these specific features were

correctly classified versus misclassified.

3. Result

In this study, we employed the publicly available eICU dataset [12] to develop a predictive model for sepsis. The dataset comprises the vital signs of thousands of patients sampled at various rates. The vital signs considered in our experiments included systolic blood pressure, diastolic blood pressure, heart rate, respiration rate, oxygen saturation (SpO2), and gender of the patients. To standardize the sampling rates, all vital signs were resampled at a frequency of 5 minutes. To build a classifier for this time-series data, we employed LightGBM. The time-series data was transformed into a format compatible with the LightGBM classifier [13] by calculating rolling statistical properties such as mean, standard deviation, and lag values from previous timestamps. The model's parameters were optimized using Python library Optuna [14]. The model achieved a recall score of 0.9308 and 0.8125 on in-sample and out-of-sample splits.

To gain a deeper understanding of the model's performance and identify areas where it exhibits suboptimal results, we applied the proposed method. To visualize and communicate the regions of poor model performance, we present Fig. 1 a) demonstrates the error distribution over specific regions where the classifier exhibited suboptimal performance. Additionally, Fig. 1 b) provides a magnified view of the error distribution, offering a clearer resolution and facilitating a more detailed examination of the error rates within these identified regions. These figures serve as visual aids, aiding in the comprehension and interpretation of the model's performance shortcomings.

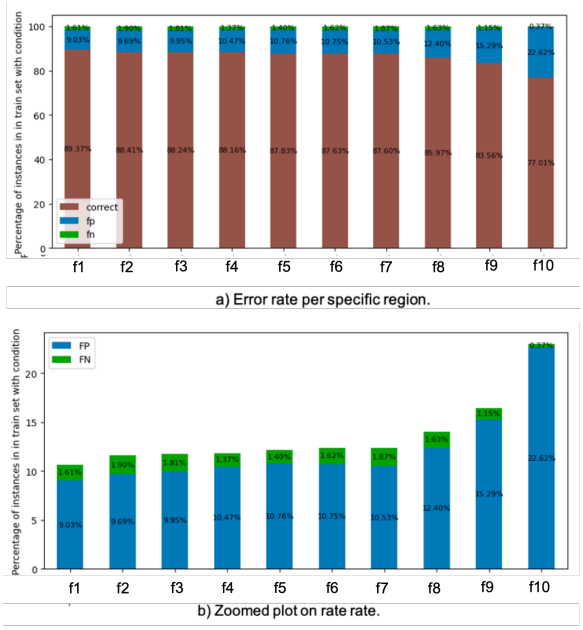

Figure 1: Error Rate Plot for Feature Conditions (Regions) Impacting Poor Performance in Sepsis Detection Using eICU Dataset.

In order to gain insights into the causes of misclassifications, we conducted a detailed analysis to determine which feature regions were most influential in contributing to these errors. Employing the LIME technique, we extracted the most significant features that consistently played a role in misclassification instances.

By identifying and examining these recurring features, we revealed specific regions where the classifier exhibited poor performance. Figure 1 visually illustrates the feature regions that meet this criterion, highlighting the factors associated with the model's suboptimal predictions.

4. Discussion

In this study, we utilized LIME to identify regions where a black-box model exhibits poor performance. This approach allows us to investigate the error distribution across misclassification regions in both training and test data. The proposed method is model agnostic and can be utilized for any classifier. By analyzing the model's fit to the training data, we gain insights into its performance and identify areas where it inadequately represents the underlying patterns in the feature space. This assessment helps us understand the model's limitations in capturing the complexities of the training dataset.

When evaluating the model's generalization error on test data, we pinpoint specific regions within the feature space that contribute to erroneous predictions for unseen data. This knowledge is crucial for important decision-making situations, such as sepsis, where being aware of regions requiring caution is essential when relying on the classifier's predictions. By conducting this analysis, we obtain a comprehensive understanding of the model's limitations and areas of poor performance. This knowledge empowers healthcare professionals and decision-makers to make informed judgments, taking into account the regions in the feature space where the classifier's predictions may be less reliable.

5. Conclusion

Our study contributes to the understanding of machine learning models' performance by introducing a modified visualization approach that identifies regions of poor performance. By leveraging LIME for the rule extraction method, we effectively pinpointed specific features responsible for misclassifications, allowing us to identify recurrent conditions associated with the classifier's suboptimal performance. The application of this methodology to the eICU dataset demonstrated its effectiveness in capturing regions where the classifier exhibits poor performance. These findings enhance interpretability and provide insights for decision-makers, enabling them to make informed choices regarding the deployment of machine learning models in critical domains such as sepsis detection. In light of the study's insights, our future work aims to enhance the model's performance by making specific modifications to the model architecture, feature engineering, and training strategies.

References

[1] Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *New England Journal Of Medicine*. **380**, 1347-1358 (2019)

[2] Antoniadi, A., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. & Mooney, C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*. **11**, 5088 (2021)

[3] Adadi, A. & Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. **6** pp. 52138-52160 (2018)

[4] Fleuren, L., Klausch, T., Zwager, C., Schoonmade, L., Guo, T., Roggeveen, L., Swart, E., Girbes, A., Thoral, P., Ercole, A. & Others Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*. **46** pp. 383-400 (2020)

[5] Cerqueira, V., Torgo, L. & Mozetič, I. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*. **109** pp. 1997-2028 (2020)

[6] Flach, P. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **33**, 9808-9814 (2019)

[7] Duivesteijn, W. & Thaele, J. Understanding where your classifier does (not) work–the SCaPE model class for EMM. *2014 IEEE International Conference On Data Mining*. pp. 809-814 (2014)

[8] Torgo, L., Azevedo, P. & Areosa, I. Beyond Average Performance–exploring regions of deviating performance for black box classification models. *ArXiv Preprint ArXiv:2109.08216*. (2021)

[9] Roshan, K. & Zafar, A. Utilizing XAI technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation (SHAP). *ArXiv Preprint ArXiv:2112.08442*. (2021)

[10] Fryer, D., Strümke, I. & Nguyen, H. Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*. **9** pp. 144352-144360 (2021)

[11] Ribeiro, M., Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining, San Francisco, CA, USA, August 13-17, 2016*. pp. 1135-1144 (2016)

[12] Pollard, T., Johnson, A., Raffa, J., Celi, L., Mark, R. & Badawi, O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*. **5**, 1-13 (2018)

[13] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T. Lightgbm: A highly efficient gradient boosting decision tree. *Advances In Neural Information Processing Systems*. **30** (2017)

[14] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. *Proceedings Of The 25th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining*. pp. 2623-2631 (2019)