Learning from failure: Explaining clinical ML model prediction errors

Anthony Li, MD[1], Zhilian Huang, MPH, PhD[2], Chien Wei Oei, BS[1], Wei Xiang Lian, BS[1], Xiaojin Zhang BEc, MS, PhD[1], Hwee Pin Phua, MS[1], Wei Xiang Ng, MBBS, MMED, FAMS[1], Seow-Yian Tay, MBBS, FRCS(A&E), FAMS, MS [1,3,4]

[1] Tan Tock Seng Hospital [2]National Center for Infectious Diseases [3]NTU, LKC School of Medicine [4]National Healthcare Group

Background. Machine Learning (ML) models, applied on complex tabular medical datasets for risk prediction, have shown superior predictive performance over traditional statistical learning methods, such as Logistic Regression[1]. However, prediction errors are inevitable in every ML model. The error etiologies of key concern in ML models for clinical risk stratification are: (1) Origin of False Negatives (FN): In a model used to predict preventable clinical events of high mortality and morbidity, we need to understand how false negatives could be reduced, thus increasing sensitivity. (2) Origin of False Positives (FP): In a model used for screening of diseases and to avoid complications of resulting investigations, we need to understand how false positives could be reduced, thus increasing specificity. If these error causes can be explained, clinicians can comprehend the limits of the model in clinical application. ML researchers can work to reduce prediction error in subsequent model revisions. We address this need by presenting a software tool to explain model prediction errors.

Methodology. This tool requires SHapley Additive exPlanations (SHAP[2]) to be applied to trained ML models. SHAP is the expected marginal contribution by a feature to the performance of the model after consideration of all possible combinations. Features with positive SHAP values are expected to contribute positively to the prediction and vice versa for negative SHAP values. It can thus be inferred that any FN row's feature, with an associated negative SHAP value, is making an erroneous model contribution in the wrong direction. The same heuristic could be applied to FP rows' features. With this approach, we can filter subsets of features and their values contributing to prediction errors. Mean SHAP values, mean feature values and error counts for each of these subsets can thus be calculated. These aggregates are used to rank error features, determine erroneous feature values and find the number of predictions with these errors.

Study Design. We extracted data of 7,690 adult patients (≥18 years old) who presented with sepsis-related ICD-9 codes to the emergency department from the MIMIC-IV[3] dataset. Features included patient demographics, comorbids, vitals recordings, laboratory test results, chronic medications and chest imaging data. Outcome of interest was ICU admission. Logistic regression (LR), support vector machines (SVM), gradient boosted ensemble decision trees (GBoost), multilayer perceptron neural network (NN) models were trained on 5,383 patients (70% of dataset) and tested on held out test set of 2,307 patients (30% of dataset). Finally, error explanation tool was applied on model with the best held out test predictive performance.

Results. The tool was applied on GBoost, which had the best held out test predictive performance (GBoost's AUROC was 0.84 vs LR 0.79 vs SVM 0.78 vs NN 0.80). GBoost made 262 prediction errors (11.3% of test set). Of which, 143 are FN and 119 are FP. Figures 1 and 2 show the top 10 features contributing errors to FN and FP respectively, ranked by their mean SHAP value contributions. Top features contributing to FN error predictions are the presence of a chronic endocrine prescription (67 errors, 46.9% of FN), no blood gas taken on admission (52 errors, 35.3% of FN) and an average minimum systolic blood pressure of 121.2 mmHg (129 errors, 90.2% of FN). Top features contributing to FP error predictions are blood gas taken on admission (55 errors, 46.2% of FP), missing ethnicity in records (88 errors, 73.9% of FP) and average minimum systolic blood pressure of 82.11 mmHg (17 errors, 14.3% of FP).

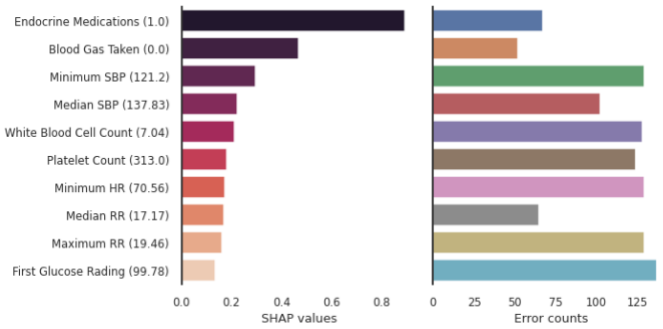

*Figure 1: FN error explanations**

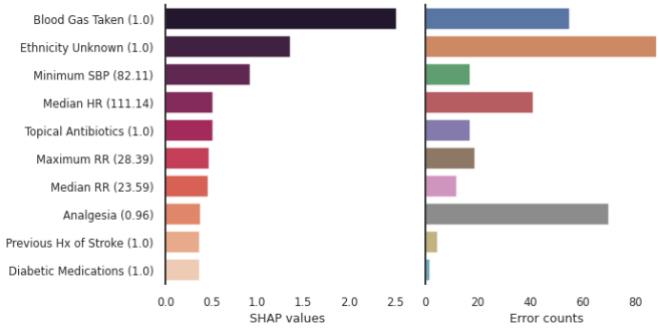

*Figure 2: FP error explanations**

**Number in parentheses is the mean value of the error feature. Binary features will show proportions instead.*

Conclusion and future work. We have demonstrated a SHAP based tool to explain clinical ML model errors. The insights from this tool helps to caution clinicians of ML model predictions driven by features which might cause inadvertent errors. Additionally, this work highlights the need for ML model failure evaluation. As future work, we aim to extend global explanations to local explanations of errors and derive strategies to correct them accordingly.

Link to Data and Software. Sepsis dataset link as provided here. Software link as provided here.

Supplementary Material

References

[1] Forte, J. C., Wiering, M. A., Bouma, H. R., Geus, F., & Epema, A. H. (2017, November). Predicting long-term mortality with first week post-operative data after Coronary Artery Bypass Grafting using Machine Learning models. In Machine Learning for Healthcare Conference (pp. 39-58). PMLR.

[2] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[3] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. circulation, 101(23), e215-e220.