Explainable AI in Decision Support Systems

A Case Study: Predicting Hospital Readmission Within 30 Days of Discharge

Alexander Vucenovic[1, 2], Osama Ali-Ozkan[1, 3], Clifford Ekwempe[1], and Ozgur Eren[1]

[1]Erie St. Clair Local Health Integration Network, Chatham, ON N7M 5Z8 Canada,
{alex.vucenovic, osam.ali, clifford.ekwempe, ozgur.eren}@lhins.on.ca
[2]Department of Industrial and Systems Engineering, Wayne State University, Detroit, MI 48202 USA, avucenovic@wayne.edu
[3]Department of Electrical and Computer Engineering, University of Western Ontario, London, ON N6A 5B9 Canada, oali8@uwo.ca

***Abstract*—Explainable models are a critical requirement for predictive analytics applications in the healthcare domain. In this work we develop a hypothetical clinical decision support system for the classification task of predicting hospital readmission within 30 days of discharge. We compare a baseline logistic regression model with an implementation of the coordinate descent algorithm known as lasso. We choose lasso because it inherently performs variable selection during optimization which leads to an explainable model. Using model evaluation data we achieve an area under the ROC curve score of 0.795 improving on the baseline score of 0.683 without inflating the feature space.**

***Index Terms*—clinical decision support systems, cross-validation, explainable AI, feature selection, healthcare, lasso, machine learning, model selection, regularization, shrinkage**

I. Introduction

Readmission within 30 days of discharge is an important metric for evaluating health outcomes [1] including patient satisfaction, safety, and successful transitions of care between health service providers. This metric, also called the readmission rate, is also an indicator of appropriate management of health system resources and cost efficiency. Predictive models can support providers in identifying high-risk patients and providing more appropriate discharge planning and clinical pathways.

A 2013 study by Health Quality Ontario and the Canadian Institute for Health Information, Hospital Admission Risk Prediction (HARP) [2], produced an explainable model for identifying patients at risk of readmission within 30 days of discharge. The binary response variable y_i is defined as follows for discharge $i = 1, ..., N$:

$$y_i = \begin{cases} 1, & \text{readmission within 30 days = Yes} \\ 0, & \text{readmission within 30 days = No.} \end{cases} \tag{1}$$

This model, which we refer to as HARP-Original, is a simple logistic regression classifier that used administrative data only (in other words no EMR or clinical data). The 127 candidate predictor variables were selected based on the advice of clinical experts. These features were reduced to 37 using best subset selection based on p-values of fitted models. HARP-Original produces an area under the Receiver Operating Characteristic curve (AUROC) of 0.678. The predictor variables used in HARP-Original will be used to develop a benchmark for this work.

Centralized Canadian healthcare databases feature data collected from acute care hospitals, emergency departments, outpatient clinics, primary healthcare, as well as home and community care settings [3][4]. The experimental data for this work is sourced from the Discharge Abstract Database (DAD) and National Ambulatory Care Reporting System (NACRS). These resources are available through the integrated internal data warehouse at Erie St. Clair Local Health Integration Network (LHIN) which is used for analysis and research purposes. This Business Intelligence (BI) platform features a range of tools and methods to gather, store, analyze, visualize, and share information related to healthcare services and outcomes. This helps healthcare information users to make better decisions and improve the quality and performance of programs and services. The BI platform consists of many independent and interrelated components to facilitate the management of data as well as provide analytic solutions and data visualization capability as shown in Figure 1 [5].

The main objective of this work is to produce a model with AUROC that beats the benchmark HARP model which we refer to as HARP-Baseline. A supplementary objective of this work is to produce such a model that is also explainable. In AI and machine learning parlance the term *explainable*, when used in the context of describing a predictive model, implies that the model possesses at least one of the following characteristics:

- A reasonably low number of predictor variables,
- little to no transformation of predictor variables,
- $\hat{y}_i$ is a linear combination of the predictor variables.

In this work we present a method for generating a model possessing all three of these characteristics. All experiments are performed in R Studio version 3.6.2 using datasets generated from the BI platform which is based on Microsoft SQL Server infrastructure.

II. Experimental Design

In the first phase of this work we develop the HARP-Baseline model using the HARP-Original predictor variables to establish the performance baseline. In the second phase of

978-1-7281-5442-8/20/$31.00 ©2020 IEEE

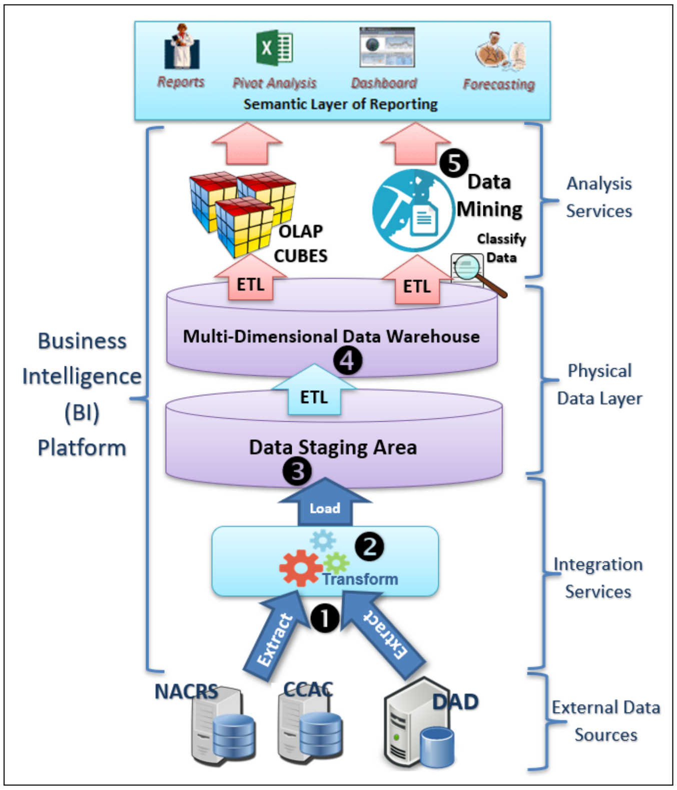

Fig. 1. Business Intelligence-Integrated Decision Support System Platform.

this work we will implement a strategy using K-fold cross-validation to derive a lasso regularized model using a much wider set of candidate predictor variables. We refer to this as the Lasso model.

A. Data Extraction

The dataset for model building is extracted from the BI platform yielding a much smaller number of observations as compared to HARP-Original. This dataset is then split into a subset for model development and a subset for model evaluation. The development subset contains 32,965 discharges and the evaluation subset contains 35,284 discharges.

B. HARP-Baseline Model

Since every predictor variable in HARP-Original is categorical any continuous predictors are transformed by discretization. Only one HARP-Original predictor variable - ‘Paracentesis Yes/No’ - was excluded from HARP-Baseline.

Logistic regression models are developed using maximum likelihood estimation on the conditional distribution $\Pr(G = K|X = x) = p_k(x;\theta)$. For the binary classification problem the log-likelihood is

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{N} \Big\{ y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta)) \Big\} \\ &= \sum_{i=1}^{N} \Big\{ y_i \beta^T x_i - \log(1+e^{\beta^T x_i}) \Big\}.\end{aligned} \tag{2}$$

We use the base R function `glm()` to fit the HARP-Baseline logistic regression model.

C. Lasso Model and K-Fold Cross-Validation

Unlike HARP-Baseline this model will feature a total of 2,468 candidate predictor variables. Lasso regularization is a shrinkage method where the parameter λ imposes a penalty on the coefficients similar to ridge regression. However, unlike ridge regression, lasso shrinks some of the model coefficients to zero as the tuning parameter λ increases [6]. The lasso inherently performs variable selection across p predictor variables (excluding the intercept term β_0) [7].

K-fold cross-validation is a model validation technique for estimating error on the evaluation set by using only the data contained in the development set. K is typically set to a value of 5 or 10 because these values have been shown in statistics research literature to balance the bias-variance problem [8]. In this work the development subset was split into 10 folds where each fold is randomly sampled from the development set without replacement. The bias-variance problem is an important consideration in the development of AI systems and can be represented by decomposing the classification error into two components:

- the error resulting from a difference between $\hat{y}_i$ and y_i,
- the error resulting from a difference between $\hat{y}_i^*$ and $\beta^T x_i$ [9].

The first component represents the bias or difference between the estimated and true response. The second component represents the variance or difference between the optimal estimate of the response and the model parameters given the development data. If a model is too sensitive to noise in the development data then it may *overfit* and not generalize well to unseen data. K-fold cross-validation for binary classification is

$$\mathrm{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{Err}_k \tag{3}$$

where $\mathrm{Err}_k = \Sigma_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$.

The lasso problem is formulated by adding the product of the scalar tuning parameter λ and the sum of the absolute values of the model coefficient weights to the log-likelihood given by (2):

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \Big[y_i(\beta_0+\beta^T x_i) - \log(1+e^{\beta_0+\beta^T x_i}) \Big] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{4}$$

We use the `glmnet` library to fit lasso models across a grid of λ values with K-fold cross-validation as an inner loop.

III. Experimental Results

For HARP-Baseline the AUROC performance on the evaluation set is 0.683 which is representative of the performance capability demonstrated in HARP-Original. The Lasso model produces AUROC on the evaluation set of 0.795.

The regularization path [10] in Figure 2 shows the number of non-zero predictor variables along the upper x-axis in

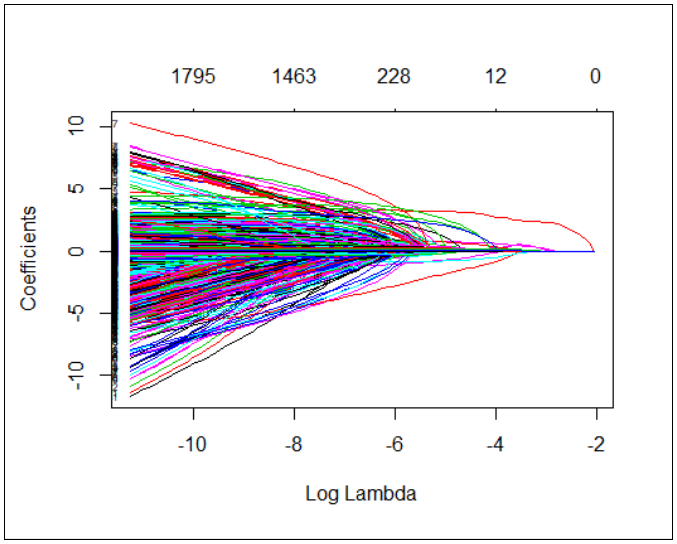

Fig. 2. Regularization Path of Log Lambda.

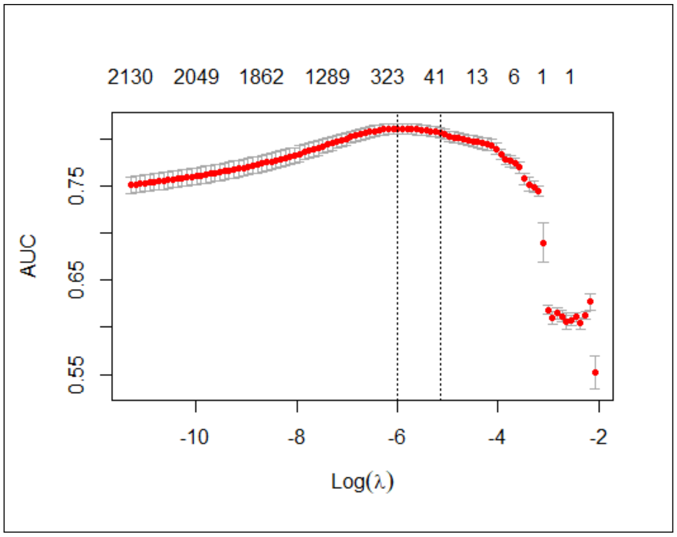

Fig. 3. 10-Fold Cross-Validation.

relation to the changes in λ (along the lower x-axis). The y-axis shows the coefficient value for each predictor. As λ decreases, the coefficients approach the values obtained with (2). This illustrates the variable selection property of the lasso model.

The cross-validation plot in Figure 3 again shows the number of non-zero predictor variables across the upper x-axis and the tuning parameter lambda across the lower x-axis. This plot shows the performance metric AUROC on the y-axis. The cross-validated AUROC is plotted using the red dotted line with error bars shown by the gray whiskers around each dot. The vertical line intersecting with the maximum AUROC is the maximum cross-validated performance on the development set. The vertical line to the right indicates a model within one standard error of the maximum AUROC and we select this parsimonious model as it balances the tradeoff between being both explainable and relatively high performing.

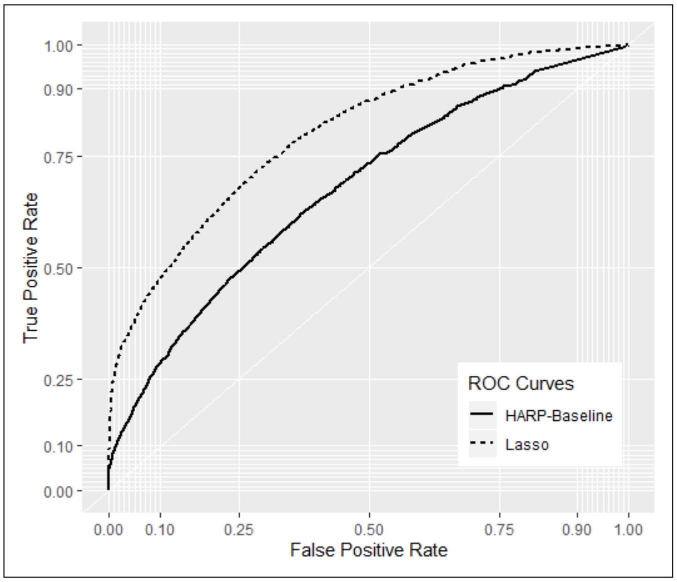

Fig. 4. ROC Curves.

Both the HARP-Baseline and Lasso ROC curves are plotted in Figure 4 [11]. Although the optimal λ is $\lambda^* = 0.002506$, we chose $\lambda_{1SE} = 0.005788$. This model contains only 35 non-zero coefficients whereas the model corresponding to λ^* contains 236 non-zero coefficients. We were able to realize this performance improvement using just a fraction of the development sample size (approximately one sixth) as compared to that which was used in HARP-Original. Table I and Table II illustrate comparative and lasso cross-validation results, respectively.

TABLE I
COMPARATIVE RESULTS

	HARP-Original	HARP-Baseline	Lasso
Performance (AUROC)	0.678	0.683	0.795
Non-zero Predictors (p^*)	37	36	35
Candidate Predictors (p)	127	-	2,468
Development Subset (N)	191,321	32,965	32,965
Evaluation Subset (N)	191,627	35,284	35,284

TABLE II
LASSO CROSS-VALIDATION RESULTS

	λ	Standard Error (SE)	Non-zero Predictors
Minimum	0.002506	0.005575	236
Within 1 SE	0.005788	0.005460	35

IV. DISCUSSION & RECOMMENDATIONS

Unplanned readmissions to hospitals are costly, undesirable to patients, and potentially avoidable [12]. In most investigations into the causes of readmissions, analysis has pointed to predictors relating to diagnosis, age, economic factors, and transitions of care between health service providers. Ability to explain the factors that lead to readmissions, or better yet effectively predict risk of readmissions, could lead to improved actions and outcomes for health care planners, providers, and patients [13].

This work shows that it is possible to develop a clinical decision support system using a business intelligence platform as the foundation and explainable AI for prediction and inference. In comparison to the simple logistic regression classifier with variables selected using the best subset method, implementation of the lasso resulted in an explainable model with improved prediction performance by using a twenty-fold increase in candidate predictor variables. Although the lasso procedure was developed for cases where $p >> N$, in this work we show its effectiveness in reducing a *relatively* high-dimensional feature space (where $N > p$) to one that generates both good performance and explainability.

Although this work focuses on a problem in the acute care sector, AI-driven tools and methods can also support studies in epidemiology and public health-related areas [14]. Opportunities in these areas that would be suitable for AI would be to help gain insight into prevention and protection from disease, infectious disease surveillance, promotion of healthy lifestyles, and population screening. Despite the potential benefits of AI there can be concerns in generating predictions on a data sample that is different from the data used for model development [15]. To reduce the risk of implicit bias in AI models researchers should take a cautious approach to ensure that the model development data sample is representative of the entire population of interest [16]. Another challenge is the potential for researchers to access large sets of data needed to apply and develop AI methods. For health care organizations planning to integrate AI into strategy and operations, finding researchers with the required skills can be challenging and presents another potential barrier for implementation [17].

The importance of variable selection in AI is that an explainable model has the potential to provide insight into focus areas for, in this case, prevention of hospital readmissions. Although the majority of health care organizations and planning bodies in Ontario are not using AI to inform actions, we see opportunities considering the emergence of EMRs (live time data), big data, health risk assessments, socio-economic data, genetic data, and more recently bio-metric sensor data. AI could also be leveraged to evaluate treatment effectiveness, healthcare value, strengths and weaknesses of alternative care models, or potential policy interventions [18].

In future work, this classification task can be extended to exploit non-linearity and interaction in the data using more powerful methods including gradient boosting and deep artificial neural networks. Such methods, however, come with the risk of potentially violating the characteristics of explainable AI. Although these methods may result in a less explainable model, the tradeoff between model performance and explainability should always be considered.

Acknowledgment

We as researchers acknowledge the support of the Decision Support Department at Erie St. Clair LHIN and gratefully thank those who made this work easier to get published.

References

[1] Vest, Joshua R. et al. “Determinants of Preventable Readmissions in the United States: A Systematic Review.” Implementation Science : IS 5 (2010): 88 - 88.

[2] HQ Ontario. Early Identification of People at Risk of Hospitalization: Hospital Admission Risk Prediction (HARP) - A New Tool for Supporting Providers and Patients. Toronto: Health Quality Ontario (2013).

[3] O. Ali, A. Ouda, ”A Classification Module in Data Masking Framework for Business Intelligence Platform in Healthcare”, in 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, Canada (2016).

[4] Lalonde, A. “Canadian Institute for Health Information.” Canadian Journal of Medical Technology 56 1 (1994): 15-6.

[5] O. Ali, H. Johnson, P. Crvenkovski, ”Using a Business Intelligence Data Analytics Solution in Healthcare”, IEEE 7th Annual Conference (IEMCON), Vancouver, Canada, (2016)

[6] Tibshirani, Robert. “Regression Shrinkage and Selection via the Lasso.” (1996).

[7] Hastie, Trevor J. et al. “The Elements Of Statistical Learning.” (2001).

[8] James, Gareth M. et al. “An Introduction to Statistical Learning.” (2013).

[9] Bishop, Christopher M.. “Pattern Recognition and Machine Learning (Information Science and Statistics).” (2006).

[10] Friedman, Jerome H. et al. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” Journal of Statistical Software 33 1 (2010): 1-22 .

[11] M.C. Sachs. plotROC: A Tool for Plotting ROC Curves. Journal of Statistical Software, Code Snippets, 79(2), 1-19. (2017). doi:10.18637/jss.v079.c02

[12] Gohil, Shruti K. et al. “Impact of Hospital Population Case-Mix, Including Poverty, on Hospital All-Cause and Infection-Related 30-Day Readmission Rates.” Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 61 8 (2015): 1235-43 .

[13] Canadian Institute for Health Information, All-Cause Readmission to Acute Care and Return to the Emergency Department (Ottawa, Ont.: CIHI, 2012).

[14] Navarro, Vicente. “What is a National Health Policy?” International Journal of Health Services 37 (2007): 1 - 14.

[15] Thiébaut, Rodolphe and Frantz Thiessard. “Artificial Intelligence in Public Health and Epidemiology.” Yearbook of Medical Informatics 27 (2018): 207 - 210.

[16] Ashrafian, Hutan and Ara Darzi. “Transforming Health Policy Through Machine Learning.” PLoS Medicine 15 (2018).

[17] Canadian Institute for Advanced Research (CIFAR) and Canadian Institutes of Health Research (CIHR). ”Application of Artificial Intelligence Approaches to Tackle Public Health Challenges - Workshop Report”, Toronto, ON (2018).

[18] Crown, William H.. “Potential Application of Machine Learning in Health Outcomes Research and Some Statistical Cautions.” Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research 18 2 (2015): 137-40 .