1-2023

An integrated framework on human-in-the-loop risk analytics

Peng LIU
Singapore Management University, liupeng@smu.edu.sg

Citation

LIU, Peng. An integrated framework on human-in-the-loop risk analytics. (2023). *Journal of Financial Data Science*. 5, (1), 58-64.

Available at: https://ink.library.smu.edu.sg/lkcsb_research/7194

An Integrated Framework on Human-in-the-Loop Risk Analytics

Peng Liu

Peng Liu
is an assistant professor of quantitative finance (practice) at Singapore Management University in Bras Basah, Singapore.
liupeng@smu.edu.sg

KEY FINDINGS

- When developing risk models, the qualitative domain expertise or expert opinions can be expressed as quantitative model constraints to bias the resulting model toward a more interpretable, statistically robust, and regulation-compliant one.
- Proper regularization and constraints are good avenues for expressing domain judgment in the development of credit risk models.
- Instead of overriding the model outputs, the author's framework allows human-in-the-loop modeling, combining domain expertise and statistical robustness during the development and estimation of credit risk models.

ABSTRACT

Risk analytics is an integral component in the overall assessment of the risk profile for potential and existing obligors. For example, credit worthiness is often assessed via the use of scorecards, which are regulatory credit risk models developed based on historical data and domain expertise in banks and financial institutions. A pure statistical model, however, often fails to entertain regulatory requirements on both predictiveness and interpretability at the same time. Instead, practical risk models are developed by incorporating expert opinions within the development process, such as forcing the direction of travel for certain financial factors. In this article, the author proposes a unified framework, termed constrained and partially regularized logistic regression (CPR-LR) model, on how human inputs could be embedded in the statistical estimation procedure when developing credit risk models. By expressing such inputs as model constraints at different levels, the proposed approach serves as an effective solution to developing intuitive, easy-to-interpret, and statistically robust credit risk models, as demonstrated in the author's experiments. This work also contributes to the growing field of human-in-the-loop model development, in which the author shows that domain expertise can be formulated as model constraints, thus biasing the resulting statistical model to be more interpretable and regulation compliant.

Risk analytics concerns the study of risk at different aspects and has been a core function in many financial institutions and banks, big and small. For example, credit risk refers to the risk of a borrower not repaying a loan, credit card, or any other type of loan. To better manage credit risk at both the individual and the aggregate level, banks and financial institutions often use credit risk models, often named scorecards, to perform risk assessment and make lending decisions based on a client's credit worthiness and internal differentiated levels of risk appetite, covering both retail customers and corporate institutions of different sizes. Per regulations from the Basel norms and the International Financial Reporting Standard (IFRS) 9, banks and financial institutions are required to perform quantitative assessment on the minimum risk-weighted assets (RWA) needed to absorb the impact from potential defaults, as well as the expected credit loss (ECL) resulting from such defaults. Both metrics require calculating the probability of default (PD) for each obligor, either on a lifetime or on a periodic basis such as 12 months. By extracting patterns from historical transactional data, a scorecard performs credit scoring and outputs the PD for a new loan application, which is then compared with a threshold to derive a binary prediction (Hand 2003). The process is often accompanied by a calibration process as mentioned in Liu 2021.

Logistic regression, the most widely used baseline classification model, offers PD output along with direct interpretation on feature importance, as compared to the black-box solutions from complex and nonlinear classifiers such as support vector machine (SVM) (Francis 2006). The predictors, in the case of retail credit risk models, can include the size of the loan, as well as other personal information, such as a customer's annual income, occupation, past default records, and credit history. For corporate and institutional credit risk models, a combination of financial factors, such as equity ratio, and qualitative factors, such as number of years for the incumbent chief executive officer, is used in the model development. Instead of adopting a purely data-driven approach, which still deserves its merit and serves as a challenger model, the development process also entails a variety of constraints on certain predictors. The additional layer of constraints originates from downstream model users, who look at the meaning and context of each candidate predictor and express constraints on its presence and weightage in the model, as well as the sign of the coefficient.

The constraints can be categorized as either hard or soft constraints. Examples on hard constraints include keeping or removing certain predictors based on its performance in a univariate analysis against the default outcome. A poorly performing predictor with an un-intuitive sign, when modeled with a default variable alone, is filtered out from the pool of candidate predictors. Even if a candidate predictor is assessed together with other candidate variables in a multivariate regression model, it may still be required to follow a target directional relationship with the default variable. For example, an increase in the net profit- related variable may only reduce the predicted PD, thus the resulting sign should be constrained as nonpositive during estimation. Soft constraints, on the other hand, refer to attributes that users prefer to have in the resulting model but are mostly driven by the statistical procedure itself. For example, a predictor should be encouraged to assume a higher weight in the multivariate estimation if it exhibits statistical significance in the univariate estimation stage and is considered to be an important factor by users based on practical experiences.

Incorporating these hard and soft constraints, however, is a nontrivial task. Although one could develop logistic regression models using different combinations of predictors following a forward or backward selection procedure and selecting those meeting the preset criteria via a postmortem fashion, this would lead to suboptimal solutions compared to a simultaneous feature selection and an estimation scheme using least absolute shrinkage and selection operator (LASSO), especially in cases when the number of candidate predictors is larger than the number of available observations (Tibshirani 1996).

On the other hand, machine learning and deep learning models have demonstrated superior predictive performances in many learning tasks, with specific techniques to explain the relative importance of input features. This line of research falls into explainable artificial intelligence (AI), such as using Shapley values to explain model predictions (Sundararajan and Najmi 2020). In this regard, the set of nonlinear transformations and black-box operations make the model results not directly interpretable and thus need to rely on derived Shapley values. My approach, however, is based on logistic regression and offers direct interpretation of model outputs as a generalized linear model of inputs.

In this article, I propose a unified framework to naturally integrate constraints from users in the model development pipeline. Hard constraints, like the sign and presence of specific features, are modeled as a constrained optimization fashion in the logistic regression setting, and soft constraints on preferences over feature selection are modeled via partial regularization of the parameters in the loss function. I demonstrate the advantages of the proposed model, termed CPR-LR model, in integrating these constraints with a satisfactory predictive performance by running experiments on several benchmark datasets.

BACKGROUND

Involving Expert Opinions in Credit Risk Modeling

The process of scorecard development usually involves the following four stages: portfolio profiling, univariate analysis, multivariate estimation, and calibration. Portfolio profiling refers to descriptive and exploratory analysis of the attributes of the development sample, such as population distribution by different predictors. This step is mainly used to ensure that the training data are well-structured and representative of all categories along a certain dimension. The univariate analysis checks the correlation and predictive power of each individual factor against the default status. Promising candidate factors, as indicated by a small P-value, a high correlation with default outcome, and consistent sign of the estimated coefficient with expert opinions, would enter the multivariate estimation stage. The resulting model is then calibrated using the calibration sample.

Although not being a purely statistical approach on determining feature importance using methods such as LASSO regression, involving preferences or constraints expressed from the model users is necessary, in part because an unconstrained model may give un-intuitive results that make it difficult to explain to the regulators. Therefore, adding an overlay of human inputs in the estimation process is more likely to end up with a statistically robust and operationally acceptable model.

Regularized Logistic Regression

Given that the solution obtained from logistic regression supports the direct interpretation in a linear manner, it is a most popular choice in building explainable credit risk models. A logistic regression model is defined by the following process:

$$y_i = f(x_i^T \beta) + \epsilon_i \tag{1}$$

where y_i takes the value of 1 if the i–th observation is the defaulted case and 0 if not. The predictors x_i is a vector, and f is the sigmoid link function used in the logistic regression. For the binary response, the conditional probability is

$$P(y_i = 1 \mid x_{i1}, \ldots, x_{ip}) = \pi_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} \tag{2}$$

where $\{\pi_i, i = 1\dots n\}$ is the predicted conditional probability of the binary response being 1 given $\{x_i, i = 1\dots n\}$, and β is the unknown coefficient vector. The optimal classifier can be obtained by minimizing the negative log-likelihood function

$$l(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i\log(\pi_i) + (1-y_i)\log(1-\pi_i)\} \tag{3}$$

To make the model more robust and less likely to overfit, a separate term is usually added to regularize the magnitude of the estimated coefficients, with LASSO being the most widely used penalty scheme. Adding LASSO in the loss function gives the following:

$$Q(\beta) = l(\beta + \lambda\sum_{i}^{p}|\beta_i| \tag{4}$$

where λ denotes the tuning parameter that controls the level of regularization in the loss function. A higher value of λ produces a bigger penalty and thus a smaller coefficient and vice versa.

It is observed that all the parameters in β do not need to follow the same level of regularization. As we will illustrate in the next section, λ could be decomposed into different parts, catering for user preference on feature importance and weightage. In addition, expected signs for specific parameters could also be added as constraints in the optimization process.

PROPOSED METHOD

Partially Regularized Logistic Regression

We first look at two common types of user preference in feature selection: those expected to stay in the resulting model, thus assuming no penalty on the coefficient, and those expected to receive less penalty due to good performance in the univariate analysis stage. These preferences could be entertained by adjusting the penalty factor λ. Specifically, the loss function now becomes

$$Q(\beta) = l(\beta) + \alpha\sum_{i=1}^{k}\lambda_i\,|\beta_i| + (1-\alpha)\sum_{j=k+1}^{p}|\beta_j|^2 \tag{5}$$

where β is decomposed into two parts, with hyperparameter α playing a balancing role. The first part, $\{\beta_i, i = 1, \dots, k\}$, represents the first k predictors whose penalty factors are individually adjusted based on, say, the estimated P-value from their respective univariate regression with the default outcome. If a predictor is statistically significant in the univariate logistic regression with the default outcome, then it will assume a low P-value and high correlation, which in turn corresponds to a lower penalty factor λ for its coefficient. The second part, $\{\beta_j, j = k + 1, \dots, p\}$, denotes those to be assessed without any LASSO penalty. Note that β_j assumes a squared form, as in ridge regression, in above loss function. This is to prevent these coefficients from exploding. These are typically hand-picked by users due to their empirical importance and thus more likely to assume a larger value in the resulting coefficient.

Constraining Regression Coefficients

Considering the operational requirement on the direction of travel for certain predictors, it is necessary to constrain their coefficients to be either nonnegative or nonpositive. Without loss of generality, assume that the coefficients $\{\beta_z, z = 1, \ldots, Z, Z \leq q\}$ need to remain as nonnegative. The partially regularized loss function now becomes a constrained one, formulated as follows:

$$\text{minimize } l(\beta) + \alpha \sum_{i=1}^{k} \lambda_i \, |\beta_i| + (1 - \alpha) \sum_{j=k+1}^{p} |\beta_j|^2$$
$$\text{s.t.} \beta_z \geq 0, z = \{1, \ldots, Z\} \tag{6}$$

Since $Q(\beta)$ can be reformulated as a quadratic form via the iteratively reweighted least squares approach, such a quadratic program could then be solved via a host of algorithms, including projected gradient descent, active set methods, and so on.

When it comes to implementation of the proposed CPR-LR model, the *glmnet* package in R is used to easily incorporate the aforementioned constraints. For example, when fitting a logistic regression model, the *penalty.factor* argument allows for a separate penalty factor for each coefficient so that prior knowledge or preference over the variables could be integrated in this step. In addition, the *upper.limits* and *lower.limits* arguments can be used to add constraints on the search range of the coefficients, thus effectively allowing for user preference on the direction of travel against the default outcome.

EXPERIMENTS

To examine the comparative performances of the proposed CPR-LR model, I performed experiments using three public credit scoring datasets. The first dataset, Australian Credit, comes from the Dua and Graff (2019). The second is downloaded from a Kaggle competition, “Give me some credit” (2011), and the third is a credit card dataset from William Greene’s *Econometric Analysis* (2003). Exhibit 1 provides the summary statistics of the datasets used. Note that the datasets are selected with varying class ratios to test the performance of the methods when working with different signal ratios.

During the experiments, 98% of the dataset is split into the training set and the rest into the test set. Missing values are filled with zero, and all tests are run for a total of 20 iterations, each starting with a different random seed. Since this is a probabilistic setting, we use receiver operating characteristic curve (ROC) area under the ROC curve (AUC) as our evaluation metric. Thus, a higher AUC corresponds to a better performing model. In each run, we specify the first five predictors to be nonpositive, that is, the upper limit of the coefficient is zero, and the last five predictors to receive a very small penalty factor (10^{-6} in this case). This serves as a constraint on the direction of travel between the predictor and the default outcome, as well as the variables that should receive little or no penalty based on the operational expertise. We compare two models, LASSO and CPR-LR, the latter of which encodes the user preference on the feature-wise penalty via the correlation with the default outcome. In other words, a predictor highly correlated with the default variable will receive a low

EXHIBIT 1

Summary Statistics of the Credit Scoring Datasets

Dataset	Number of Observations	Number of Predictors	Prior Default Rate
Australian Credit	690	14	0.445
Kaggle Competition	150,000	10	0.067
Econometric Analysis	1319	11	0.224

penalty and vice versa. Data and code will be made available as open source once the paper is accepted.

Exhibit 2 shows the test set ROC AUC statistics on three different credit scoring datasets. Each cell denotes the median AUC across 20 runs, with the mean AUC shown in bracket. The results suggest that the CPR-LR model, other than its human-in-the-loop highlight, can deliver similar or better predictive performance compared with the LASSO-based model.

In addition, we also plot a typical ROC AUC curve for both models across the three datasets. As shown in Exhibit 3, the CPR-LR model clearly dominates the LASSO-based model. We observe its superior performance over the majority of runs with different start seeds, suggesting its potential for incorporating user preferences and constraints and building predictive credit risk models.

EXHIBIT 2

Comparison of the Test Set ROC AUC Statistics on Three Different Credit Scoring Datasets

	Median and Mean ROC AUC in Test Dataset		
Model	**Australian Credit**	**Econometric Analysis**	**Kaggle Competition**
LASSO	0.833 (0.82)	0.5 (0.552)	0.731 (0.731)
CPR-LR	0.8 (0.818)	1 (1)	0.765 (0.765)

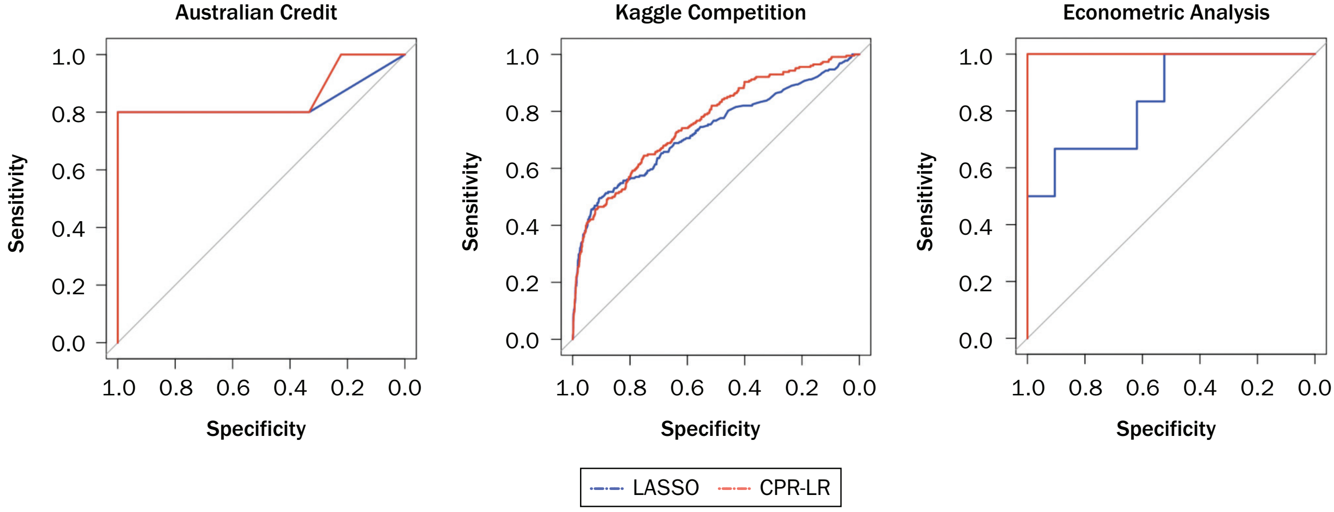
NOTES: Each cell denotes the median AUC across 20 runs, with the mean AUC shown in parentheses. The results suggest that the CPR-LR model, other than its human-in-the-loop estimation process, can deliver similar or better predictive performance compared with the LASSO-based model.

CONCLUSION

In this article, I propose a CPR-LR model in the context of credit scoring. The proposed model is designed to flexibly incorporate user preferences and constraints on coefficient sign and feature importance. Each constraint is explicitly added as either a soft or a hard constraint, giving sufficient transparency and user control in the model development process while ensuring decent predictive performance. I hope that this work contributes to the adoption of regularized and constrained optimization frameworks in the risk management space, where a lot of emphasis is put on building predictive yet intuitive and explainable risk models.

EXHIBIT 3

Sample ROC AUC Curves on Three Different Credit Scoring Datasets



NOTES: This exhibit shows sample ROC AUC curves on three different credit scoring datasets, where the CPR-LR model clearly dominates the LASSO-based model. This is also observed in multiple runs with different starting seeds.

REFERENCES

Dua, D., and C. Graff. 2019. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Francis, L. 2006. "Taming Text: An Introduction to Text Mining." In *Casualty Actuarial Society Forum* 51–88.

Greene, W. H. 2003. *Econometric Analysis*. Pearson Education India.

Hand, D. 2003. "Good Practice in Retail Credit Scorecard Assessment." *Journal of the Operational Research Society* 56 (9): 1109–1117.

Kaggle. "Give Me Some Credit." Kaggle Featured Prediction Competition, 2011, https://www.kaggle.com/c/GiveMeSomeCredit.

Liu, P. 2021. "Improving Credit Scorecard Calibration Using Regularized Logistic Regression and Bayesian Optimization." Paper presented at the Fifth PKU-NUS Annual International Conference on Quantitative Finance and Economics, May 2021.

Sundararajan, M., and A. Najmi. 2020. "The Many Shapley Values for Model Explanation." *Proceedings of the 37*th *International Conference on Machine Learning* 119: 9269–9278.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B* 58 (1): 267–288.

Disclosure

The author reports no conflicts of interest. The author alone is responsible for the content and writing of the article.