Developing and improving risk models using machine-learning based algorithms

Yan Wang
ywang63@students.kennesaw.edu

Sherry Ni
sni@kennesaw.edu

Follow this and additional works at: https://digitalcommons.kennesaw.edu/dataphdgreylit

Recommended Citation
Developing and Improving Risk Models using Machine-learning Based Algorithms

Yan Wang
Kennesaw State University
Kennesaw, GA, USA
ywang63@students.kennesaw.edu

Xuelei Sherry Ni
Kennesaw State University
Kennesaw, GA, USA
smi@kennesaw.edu

ABSTRACT
The objective of this study is to develop a good risk model for classifying business delinquency by simultaneously exploring several machine learning based methods including regularization, hyper-parameter optimization, and model ensembling algorithms. The rationale under the analyses is firstly to obtain good base binary classifiers (including Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), and Artificial Neural Networks (ANN)) via regularization and appropriate settings of hyper-parameters. Then two model ensembling algorithms including bagging and boosting are performed on the good base classifiers for further model improvement. The models are evaluated using accuracy, Area Under the Receiver Operating Characteristic Curve (AUC of ROC), recall, and F1 score via repeating 10-fold cross-validation 10 times. The results show the optimal base classifiers along with the hyper-parameter settings are LR without regularization, KNN by using 9 nearest neighbors, DT by setting the maximum level of the tree to be 7, and ANN with three hidden layers. Bagging on KNN with K valued 9 is the optimal model we can get for risk classification as it reaches the average accuracy, AUC, recall, and F1 score valued 0.90, 0.93, 0.82, and 0.89, respectively.

CCS CONCEPTS
• Computer systems organization → Machine learning; Modeling.

KEYWORDS
Improve Risk Model, Machine Learning

ACM Reference Format:

1 PROBLEM AND MOTIVATION
Many studies have demonstrated that the performance of risk models can be improved by using many machine-learning based methods including regularization, hyper-parameter optimization, and ensembling algorithms [4]. In this study, we aim to develop a good risk model for classifying business delinquency by jointly and comprehensively exploring the effect of the above-mentioned model-improving algorithms.

2 BACKGROUND AND RELATED WORK
Logistic Regression (LR) is a widely used technique for binary classification because of its strong interpretability and competitive performance. Moreover, regularized LR, which leads to a substantial decrease in variance and prediction error, outperforms LR in some studies. Two commonly used versions of LR include L1-regularized LR and L2-regularized LR, with the former version penalizes the L1 norm of the coefficients while the latter version penalizes the L2 norm of the coefficients [1]. Decision Tree (DT) is another widely acceptable binary classification technique but different settings of its hyper-parameters could largely affect its performance [2]. Similarly, in an Artificial Neural Network (ANN) and k-Nearest Neighbor (KNN), careful tuning of the hyper-parameters can improve their performance [6] [8]. Moreover, model ensembling is another effective way to improve the model performance [7] [9].

3 APPROACH
The dataset used in this study contains the financial information of 9500 US companies in 2014 with the delinquent rate valued 49.69%. We randomly select 80% of the data as the training set and use the rest as the testing set. The dimensionality of the features is reduced to 100 via hierarchical variable clustering in the data pre-processing stage. Four base classifiers including KNN, LR, DT and ANN are developed by using regularization and hyper-parameter optimization algorithms. To be specific, LR is regularized by using both L1 and L2-regularization. In KNN, the hyper-parameter ‘K’, denoting the number of nearest neighbors, is tuned by taking a series of values ranging from 3 to 13. In DT, we tuned ‘max_depth’, which denotes the maximum level of the tree structure, by using different values ranging from 5 to 15. In ANN, the hyper-parameter ‘layer_size’, representing the number of hidden layers with units in each layer, is tuned by taking a series of values of ‘50’, ‘50_25’, ‘50_25_13’, and ‘50_25_13_6’. For example, the value of ‘50_25_13’ means there are three hidden layers in the ANN and the number of units in each layer is 50, 25, and 13, respectively. Area Under the Receiver Operating Characteristic Curve (AUC of ROC), recall, and F1 score are used as evaluation metrics based on testing set by repeating 10-fold cross validation 10 times [3] [5]. After building the base classifiers as accurate as possible, two model ensembling techniques (including bagging and boosting) are performed to examine whether the performance can be further improved or not.
4 RESULTS

The best base classifiers along with their hyper-parameter setting are \( \text{LR}-0 \) (i.e., \( \text{LR} \) without regularization), \( \text{KNN}-9 \) (i.e., \( K \) valued 9), \( \text{DT}-7 \) (i.e., ‘max_depth’ valued 7), and \( \text{ANN}-4 \) (i.e., contains three hidden layers and having units valued 50, 25, and 13 respectively in each layer). Figure 1 shows the result of the best classifiers as well as their ensemble models. The post-fix bag and boo denote bagging and boosting, respectively. In \( \text{LR} \), bagging and boosting cannot significantly improve the model performance with respect to accuracy, recall, and F1 score. Quite unexpectedly, boosting on \( \text{LR} \) even hurt the AUC by a large extent. In \( \text{KNN} \), both bagging and boosting are beneficial when considering accuracy, recall, and F1 score measures. On the contrary, boosting on \( \text{KNN} \) decrease AUC significantly. Bagging on \( \text{DT} \) outperforms base \( \text{DT} \) as it produces significantly higher AUC and marginally higher accuracy. However, boosting significantly decrease AUC of \( \text{DT} \). Compared with base \( \text{ANN} \), boosting on \( \text{ANN} \) is beneficial in terms of accuracy, recall, and F1 score, indicating that \( \text{ANN} \) is not a good base classifier to be bagged on. By comparing all the aforementioned results, we conclude that \( \text{KNN}-9 \)-bag (i.e., bagging on the base classifier \( \text{KNN} \) with \( K \) valued 9) is the optimal risk model in our study with average accuracy, AUC, recall, and F1 score valued 0.90, 0.93, 0.82, and 0.89, respectively.

5 CONCLUSIONS AND CONTRIBUTIONS

In this paper, we aim at developing and improving the risk modeling via the widely used machine-learning based algorithms including regularization, hyper-parameter optimization, and ensembling simultaneously. The results show that the optimal hyper-parameter settings for the base classifiers are \( \text{LR} \) without regularization, \( \text{KNN} \) by using 9 nearest neighbors, \( \text{DT} \) by setting the level of the tree to be 7, and \( \text{ANN} \) with three hidden layers. The optimal model we get for classifying business delinquency is through bagging on \( \text{KNN} \) with \( K \) valued 9, which reaches the average accuracy, AUC, recall, and F1 score valued 0.90, 0.93, 0.82, and 0.89, respectively. Although different conclusions may be obtained because of various dataset used in the future, the study methodology provided by us is a good reference for studies that aiming to improve risk modeling using machine-learning based algorithms.

REFERENCES