Spring 2-15-2018

COMPARISON OF BANKRUPTCY PREDICTION MODELS WITH PUBLIC RECORDS AND FIRMOGRAPHICS

Lili Zhang
Kennesaw State University, lzhang18@students.kennesaw.edu

Jennifer Priestley
Kennesaw State University, jpriestl@kennesaw.edu

Xuelei Ni
Kennesaw State University, xni2@kennesaw.edu

COMPARISON OF BANKRUPTCY PREDICTION MODELS WITH PUBLIC RECORDS AND FIRMOGRAPHICS

Lili Zhang[1], Jennifer Priestley[2], and Xuelei Ni[3]

[1]Program in Analytics and Data Science, Kennesaw State University, Georgia, USA
lzhang18@students.kennesaw.edu
[2]Analytics and Data Science Institute, Kennesaw State University, Georgia, USA
jpriestl@kennesaw.edu
[3]Department of Statistics, Kennesaw State University, Georgia, USA
xni2@kennesaw.edu

ABSTRACT

Many business operations and strategies rely on bankruptcy prediction. In this paper, we aim to study the impacts of public records and firmographics and predict the bankruptcy in a 12-month-ahead period with using different classification models and adding values to traditionally used financial ratios. Univariate analysis shows the statistical association and significance of public records and firmographics indicators with the bankruptcy. Further, seven statistical models and machine learning methods were developed, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, Bayesian Network, and Neural Network. The performance of models were evaluated and compared based on classification accuracy, Type I error, Type II error, and ROC curves on the hold-out dataset. Moreover, an experiment was set up to show the importance of oversampling for rare event prediction. The result also shows that Bayesian Network is comparatively more robust than other models without oversampling.

KEYWORDS

Bankruptcy Prediction, Public Records, Firmographics, Classification, Oversampling

1. INTRODUCTION

Bankruptcy prediction has been studied since the 1960s, to improve decision making related to business operations conducted with reliable counterparties [3]. For example, investors want to make investments to organizations that have high potential to succeed. Banks want to lend to the organizations that are less likely to default. Business entities want to do business and build relationships with the ones that can prosper and survive in a long term. Hence, it is valuable to foresee the possibility of the bankruptcy of a business customer or partner.

To improve the accuracy of bankruptcy prediction, researchers and practitioners have pursued two primary paths of study. First, explore important variables for bankruptcy prediction. For example, the predictive ability of financial ratio variables has been thoroughly studied. Second, improve the methodologies used for the bankruptcy prediction, benefiting from the development of both the algorithm theories and computation infrastructure. Besides significant variables and high-performance methods, we observe that appropriate data sampling before modeling is also important for improving bankruptcy prediction, considering that frequently the proportion of bankruptcy cases is substantively lower than the proportion of non-bankruptcies.

In this paper, we aim to make contributions from all above perspectives. First, we explore the impacts of public records and firmographics on bankruptcy prediction to add values to widely used financial ratio variables. Both univariate analysis and multiple variable analysis were

conducted to measure statistical association and significance. With significant variables selected, we comprehensively compare seven classification models from the statistics and machine learning domains, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, Neural Network, and Bayesian Network. The performance of the models are evaluated on the hold-out dataset. The overall classification accuracy, Type I error, Type II error, and ROC curves are evaluated. Finally, we demonstrate the importance of oversampling for the rare event prediction like bankruptcy prediction, and demonstrate the robustness of the Bayesian Network for rare event modeling.

The paper is structured as follows. In Section 2, related work is reviewed. In Section 3, the data processes are described. In Section 4, the univariate analysis between the dependent variable and each individual input variable is performed. In Section 5, the models are developed, diagnosed, evaluated, and compared. In Section 6 and 7, conclusions and future work are discussed.

2. RELATED WORK

Because of its importance in business decisions like investment and loan lending, the bankruptcy prediction problem has been studied through deriving significant predictors and developing novel prediction models. Altman proposed a set of traditional financial ratios, including Working Capital/Total Assets, Retained Earnings/Total Assets, Earnings before Interest and Taxes/Total Assets, Market Value Equity/Book Value of Total Debt, and Sales/Total Assets, and used them in the multiple discriminant analysis for the corporate bankruptcy prediction [2]. Those financial ratios were widely adopted and extended later [13] [4]. Amir came up with some novel financial ratio indicators, including Book Value/Total Assets, Cashflow/Total Assets, Price/Cashflow, Rate of Change of Stock Price, and Rate of Change of Cashflow per Share, in addition to Altman's ones, for a neural network model, and increased the prediction accuracy by 4.04% for a three-year-ahead forecast [4]. Everett et al. studied the impact of external risk factors (i.e. macro-economic factors) on small business bankruptcy prediction and proposed a logistic regression model [7]. Chava et al. demonstrated the statistical significance of industry effects by grouping firms into finance/insurance/real estate, transportation/communications/utilities, manufacturing/mineral, and miscellaneous industries [6].

From the methodology perspective, various statistical methods, machine learning algorithms, and hybrid models have been applied and compared for the bankruptcy prediction problem. Odom et al. proposed the first neural network model for bankruptcy prediction [13]. Zhang et al. showed that the neural network performed better than logistic regression and were robust to sampling variations [17]. Shin et al. found that the support vector machine outperformed the neural network on small training datasets [14]. Min et al. applied support vector machine with optimal kernel function hyperparameters [12]. Zibanezhad showed the acceptable prediction ability of decision tree on the bankruptcy prediction problem and determined the most important financial ratios [8]. Zikeba et al. proposed and evaluated a novel gradient boosting method for learning an ensemble of trees [18]. Sun et al. studied the application of Bayesian network on the bankruptcy prediction problem in respects of the influence of variable selection and variable discretization on the model performance [15]. Ahn et al. presented a hybrid methodology by combining rough set theory and neural network [1]. Huang et al. proposed a hybrid model by incorporating static and trend analysis in the neural network training [9]. Kumar et al. provided a comprehensive review on both the financial ratio variables and methods used for the bankruptcy prediction from 1968 to 2005, discussed merits and demerits of each method, and listed some important directions for future research [11]. Bellovary et al. reviewed 165 existing studies for the bankruptcy prediction and made some suggestions, where one suggestion was that the model accuracy was not guaranteed with the number of factors [5].

Most models proposed for bankruptcy prediction in the literature were directly developed on the dataset with a balanced proportion of bankruptcy and non-bankruptcy observations. However,

data imbalance is a common issue in practice. Kim et al. proposed a geometric mean based boosting algorithm to address the data imbalance problem in the bankruptcy prediction, but only compared it with other boosting algorithms to show its advantage [19]. Zhou studied the effect of sampling methods for five bankruptcy prediction models, but the models were not tuned to their optimal hyperparameters [20].

The models applied to the bankruptcy prediction utilize a variety of algorithms. Logistic Regression formulates a function between the probability of the event ($\hat{p}$) and input variables ($x_1, x_2, \dots, x_n$) defined as:

$$\hat{p} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

The coefficients ($\beta_1, \beta_2, \dots, \beta_n$) in the function are estimated by optimizing the maximum likelihood function defined as below, where y is the actual value with the event denoted as 1 and the nonevent denoted as 0.

$$\max y log\hat{p} + (1 - y)\log(1 - \hat{p})$$

Decision Tree defines hierarchical rules by searching for optimal splits on input variables based on the Entropy or Gini index. The Entropy and Gini index of an input variable are defined below, where x is a given input variable, $1, \dots, k$ are levels in the dependent variable, and $p(i|k)$ is the conditional probability for the dependent variable taking value i given x [16].

$$Entropy(x) = -\sum_{i=1}^{k} p(i|k) log_2\big(p(i|k)\big)$$

$$Gini(x) = 1 - \sum_{i=1}^{k} [p(i|k)]^2$$

Random Forest and Gradient Boosting are an ensemble of multiple decision tree models through bagging and boosting, respectively. In Random Forest, each tree is trained independently on a bootstrap dataset created from the original training dataset and then combined to a single prediction model by taking the average of all trees. In Gradient Boosting, each tree is trained sequentially based on a modified version of the original training dataset by utilizing the information of previously trained trees [10]. In tree-based models, a summary of variable importance can be obtained. The importance of each input variable is measured based on the Entropy or Gini reduction by splitting a given input variable. The larger the value is, the more important an input variable is.

Support Vector Machine defines a hyperplane for two-class classification by maximizing the marginal distance. To handle the nonlinear relationship, a kernel function can be first applied to project the input variables to a higher feature space. Neural Network learns the relationship between the dependent variable and input variables by first transforming input variables with an activation function (Tanh, Sigmoid, etc.) through each hidden unit in one or more hidden layers and then adjusting the weights through backpropagation iteratively to minimize a loss function. Bayesian Network represents the probability relationship and conditional dependencies between the dependent variable and input variables via a directed acyclic graph.

3. DATA

The bankruptcy indicator, public records and firmographics information of 11,787,287 U.S. companies in the 4th Quarter of 2012 and 2013 was collected by a national credit reporting agency, and were approved for use in this study. From the data, a bankruptcy flag indicates whether a corporate is in bankruptcy or in business at the capture time point. Firmographics in the data include industry, location, size, and status and structure. Each corporate is identified by

its unique Market Participant Identifier (MPID). Public Records include judgements and liens reported.

From the dataset provided, we aim to answer the following question explicitly, which can provide decision makers with insights into improved bankruptcy prediction.

Given the public records and firmographics indicators of an organization in one quarter, can we predict its operation status one year in the future?

To answer the question above, the dependent variable Bankruptcy Indicator Change (i.e. BrtIndChg) was created and is provided in Table 1. Originally, Bankruptcy Indicator (i.e. BrtInd) has two levels, 0 and 1, where 0 indicates that the organization is operating and 1 indicates a bankruptcy. If an organization in business in 2012 went to bankruptcy in 2013, then BrtIndChg was assigned to 1. If the organization was still in business in 2013, then BrtIndChg was assigned to 0.

The raw data had to be cleaned and transformed prior to modeling, to address missing values, abnormal/incorrect values, and correlated variables. The following steps were applied to the data.

(1) Only keep observations with the level value 0 in the original 2012 BrtInd.
(2) Create the dependent variable BrtIndChg by comparing BrtInd in the dataset of 2012 and 2013 as shown in Table 1.
(3) Drop interval variables if the percentage of coded values or missing values is greater than 30%. A value of 30% was selected to optimize the percent of variance explained in the dataset.
(4) Drop observations in an interval variable or a categorical variable if the percentage of the abnormal/incorrect values in that variable is less than 5%.
(5) Continuous variables were binned into nominal variables. For example, the variable Number of Current Liens or Judgment was binned into Current Liens or Judgment Indicator (i.e. curLiensJudInd) with two levels, 0 and 1, where 0 means an organization does not have a lien or judgment currently and 1 means an organization has one or more liens or judgments currently.
(6) Retain the variable with the best predictive ability among several correlated variables. For example, based on both the variable definition and the Chi-Square value, the following variables are correlated: Current Liens/Judgment Indicator, Number of Current Liens/Judgment and Total Current Dollar Amounts on All Liens/Judgments. After comparing their performance, only the variable Current Liens/Judgment Indicator was kept.

Table 1. Creation of Dependent Variables

BrtInd 2012	BrtInd 2013	BrtIndChg
0	1	1
0	0	0

After the data was cleaned, the variables in Table 2 were prepared for further analysis and modeling. As described above, the bankruptcy is a rare event, which can be further confirmed by the distribution of the dependent variable BrtIndChg, as shown in Table 3. In our dataset, there are 0.12% of observations going into bankruptcy from 2012 to 2013 and 99.88% of observations staying in business from 2012 to 2013. Because the proportion of event cases is much less than the proportion of nonevent cases, we need to consider oversampling to have sufficient event cases to train the model and achieve better performance, which will be discussed in detail in Section 5.

Table 2. Variables for Analysis and Modelling.

Variable	Type	Description
MPID	Nominal	Market Participant Identifier
BrtIndChg	Binary	Bankruptcy Indicator Change
curLiensJudInd	Nominal	Current Liens/Judgment Indicator
histLiensJudInd	Nominal	Historical Liens/Judgment Indicator
Industry	Nominal	Industry
LargeBusinessInd	Nominal	Large Business Indicator
Region	Nominal	Geographical Region
PublicCompanyFlag	Nominal	Public Company Flag
SubsidiaryInd	Nominal	Subsidiary Indicator
MonLstRptDatePlcRec	Interval	Number of Months Since Last Report Date on Public Records

Table 3. Frequency of Dependent Variable.

BrtIndChg	Frequency	Percent (%)
1	1031	0.12
0	843330	99.88

4. EXPLORATORY ANALYSIS

To examine the statistical association and significance between each individual input variable and the dependent variable, bivariate analysis was performed. The results of odds ratio and Chi-square test can be found in Table 4. Based on the Chi-Square results, all the variables are significantly associated with the dependent variable except the variable PublicCompanyFlag. Based on the odds ratio, we have the following observations regarding their relationship:

- Current Lien/Judgment Indicator: The organizations which currently do not have any lien/judgment is about 47.1% less likely to go into bankruptcy in the following year than those which currently have liens or judgments.
- Historical Lien/Judgment Indicator: The organizations which did not have any lien/judgment is about 32% less likely to go into bankruptcy in the following year than the ones which historically had liens or judgments.
- Large Business Indicator: The organizations which are not large are about 45.8% less likely to go into bankruptcy in the following year than the ones which are large.
- Subsidiary Indicator: The organizations which are not subsidiaries are 74.5% more likely to go into bankruptcy in the following year than those organizations which are subsidiaries.
- Industry: By using the industry group 8 as the reference level, the organizations in the industry group 3 is about 2 times more likely going to the bankruptcy in the following year than the ones in the industry group 8.
- Region: By using the region group 9 as the reference level, the organizations in the region group 2 are about 55.7% less likely to go into bankruptcy in the following year than the ones in the region group 9.
- Number of Months Since Last Report Date on Public Records (i.e. MonLstDatePlcRec): Figure 1 shows that the distribution of MonLstDatePlcRec is very different in different levels of BrtIndChg, indicating their strong relationship.

Table 4. Univariate Odds Ratio and Chi-Square p-value.

Effect	Odds Ratio	95% Confidence Interval	Chi-Square p-value
curLiensJudInd 0 vs 1	0.529	[0.447, 0.627]	<.0001
histLiensJudInd 0 vs 1	0.680	[0.601, 0.768]	<.0001
LargeBusinessInd N vs Y	0.542	[0.474, 0.620]	<.0001
LargeBusinessInd U vs Y	0.202	[0.165, 0.249]	
PublicCompanyFlag N vs Y	0.295	[0.104, 0.838]	0.065
PublicCompanyFlag U vs Y	0.370	[0.138, 0.989]	
SubsidiaryInd N vs Y	1.745	[0.997, 3.053]	<.0001
SubsidiaryInd U vs Y	0.411	[0.261, 0.648]	
Industry 1 vs 8	1.538	[0.947, 2.496]	<.0001
Industry 2 vs 8	3.085	[1.118, 8.514]	
Industry 3 vs 8	2.079	[1.545, 2.797]	
Industry 4 vs 8	1.971	[1.365, 2.847]	
Industry 5 vs 8	1.648	[1.136, 2.392]	
Industry 6 vs 8	2.421	[1.704, 3.439]	
Industry 7 vs 8	1.386	[1.033, 1.859]	
Industry 9 vs 8	1.348	[1.012, 1.795]	
Industry 10 vs 8	0.885	[0.216, 3.629]	
Industry U vs 8	0.473	[0.343, 0.651]	
Region 1 vs 9	0.699	[0.479, 1.019]	<.0001
Region 2 vs 9	0.443	[0.358, 0.549]	
Region 3 vs 9	0.627	[0.505, 0.779]	
Region 4 vs 9	0.913	[0.686, 1.215]	
Region 5 vs 9	0.636	[0.525, 0.772]	
Region 6 vs 9	1.203	[0.928, 1.558]	
Region 7 vs 9	1.084	[0.875, 1.343]	
Region 8 vs 9	1.194	[0.920, 1.549]	
MonLstRptDatePlcRec	0.971	[0.969, 0.973]	<.0001

5. METHODOLOGY

To better train and evaluate the models, the dataset was first oversampled and then split into training dataset and validation dataset, where the training dataset was used for training the models and the validation dataset was used as the hold-out dataset for evaluating the performance of models. Seven different models were developed, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, Bayesian Network, and Neural Network. Their respective performances were then evaluated by overall accuracy, Type I error, Type II error, and ROC curve.

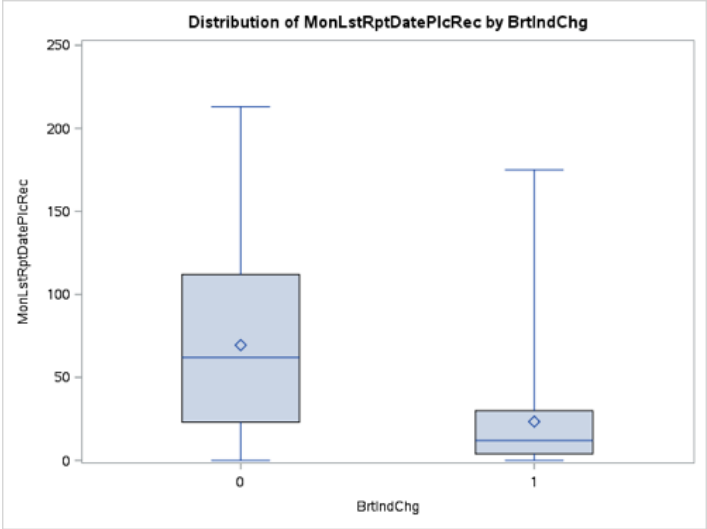

Figure 1. Boxplot of MonLstRptDatePlcRec by BrtIndChg

5.1. Sampling

Sampling was done in three steps.

(1) Oversampling: The proportion of the events is 0.12%, as indicated in Table 3. To have sufficient event cases to train the model and achieve better performance, the oversampling technique is used to adjust the proportion of event observations and non-event observations to 50% versus 50%, which include all the event observations and an equal number of randomly selected non-event observations. That ends up with 1031 bankruptcy observations and 1031 non-bankruptcy observations.

(2) Training Dataset and Validation Dataset Split: The out-of-sample test is used by evaluating the models on the hold-out dataset. Hence the dataset is split into training and validation by 70% versus 30%, respectively.

(3) Oversampling Adjustment: Prior probability and inverse prior weights are applied to the results to adjust oversampling.

5.2. Model Development and Evaluation

The models were developed using SAS Enterprise Miner. All variables in Table 4 are specified as initial inputs for all models. Every model is tuned to their best performance by trying different hyperparameter values.

In Logistic Regression, backwards selection is used to select significant variables with the significance level set to 0.05. The multivariate odds ratio and Chi-Square p-value of the resulting model can be found in Table 5. The significant variables include curLiensJudInd, histLiensJudInd, LargeBusinessInd, Region, and MonLstDatePlcRec. Their multivariate odds ratio is consistent with their univariate odds ratio. For example, univariate odds ratio shows that curLiensJudInd is negatively associated with the dependent variable, which is the same as indicated by the multivariate odds ratio of curLiensJudInd.

Decision Tree, Gradient Boosting, and Random Forest are all tree-based models. Entropy is used as the criteria of searching and evaluating candidate splitting rules for Decision Tree, while Gini index is used for Gradient Boosting and Random Forest. The important variables selected by these models include MonLstDatePlcRec, Region, Industry, curLiensJudInd, histLiensJudInd, and LargeBusinessInd. Their importance measure can be found in Table 6. Note that for Decision Tree and Gradient Boosting, the importance measure presented here is the total Entropy or Gini reduction, while for Random Forest, the importance measure is the marginal Gini reduction.

Table 5. Multivariate Odds Ratio and Chi-Square p-value.

Effect	Odds Ratio	Chi-Square p-value
curLiensJudInd 0 vs 1	0.573	0.0046
histLiensJudInd 0 vs 1	0.508	<.0001
LargeBusinessInd N vs Y	0.796	<.0001
LargeBusinessInd U vs Y	0.332	
Region 1 vs 9	1.067	0.0002
Region 2 vs 9	0.411	
Region 3 vs 9	0.583	
Region 4 vs 9	0.839	
Region 5 vs 9	0.558	
Region 6 vs 9	0.858	
Region 7 vs 9	0.881	
Region 8 vs 9	1.261	
MonLstRptDatePlcRec	0.976	<.0001

Table 6. Variable Importance.

Variable	Decision Tree	Gradient Boosting	Random Forest
MonLstRptDatePlcRec	1.0000	1.0000	0.0911
Region	0.2423	0.2880	0.0048
Industry	0.1663	0.3516	0.0110
curLiensJudInd	0.1550	0.0820	0.0024
histLiensJudInd	0.1192	0.1205	0.0038
LargeBusinessInd	0.0308	0.2752	0.0100

In Support Vector Machine, linear kernel function performs better than polynomial kernel function. In Neural Network, Tanh is used as the activation function in the hidden layer while Sigmoid is used in the output layer. Its architecture can be found in Figure 2. In Bayesian Network, the significant variables selected by G-Square with the significance level 0.2 include MonLstDatePlcRec, Region, Industry, curLiensJudInd, histLiensJudInd, LargeBusinessInd, and SubsidiaryInd. The resulting Bayesian Network can be found in Figure 3.

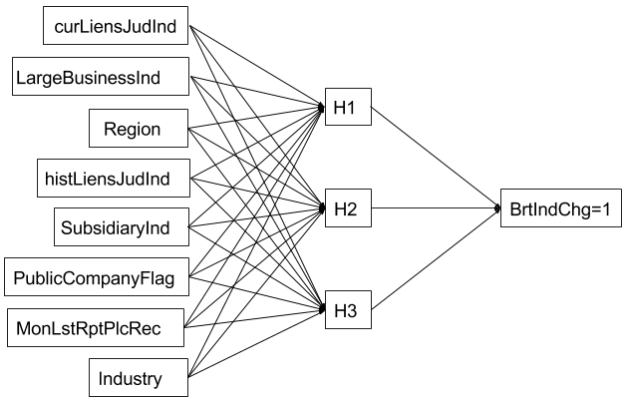

Figure 2. Neural Network Architecture

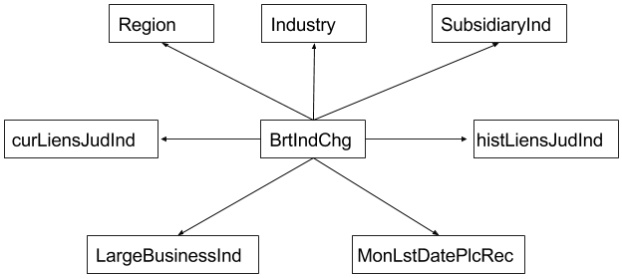

Figure 3. Bayesian Network Diagram

The accuracy, Type I error, and Type II error of all models are summarized in Table 7. Overall speaking, there is always a tradeoff between Type I error and Type II error, where a large Type I error causes less profits by classifying organizations with low bankruptcy risk into high risky ones and a large Type II error brings more losses by classifying organizations with high bankruptcy risk into low risky ones. For example, according to Type II error, Support Vector Machine performs the best, but it gives the worst overall accuracy and Type I error in the meantime. Practitioners are suggested to select the model by making a balance among model characteristics (accuracy, Type I error, Type II error, interpretability, etc.) based on their expectations. For example, Neural Network, Bayesian Network, and Logistic Regression give the same Type II Error in this case, Bayesian Network and Logistic Regression may be favored than Neural Network because of their high interpretability.

Table 7. Performance of Models.

Model	**Accuracy**	**Type I Error**	**Type II Error**
Support Vector Machine	73.39%	40.32%	12.90%
Decision Tree	75.16%	36.45%	13.87%
Gradient Boosting	74.51%	31.29%	17.42%
Random Forest	75.80%	29.35%	19.03%
Neural Network	74.52%	29.35%	19.35%
Bayesian Network	74.35%	31.93%	19.35%
Logistic Regression	74.35%	31.61%	19.35%

To more comprehensively compare these models, ROC curves on both the training and validation dataset are provided in Figure 4 and Figure 5, respectively. For all models, there is no large difference between training ROC and validation ROC, so there is no overfitting. Moreover, all models overall perform similar, because there is no large gap among their ROCs.

5.3. Experiment without Oversampling

To show the influence of oversampling in the rare event prediction, the dataset and models were fitted without oversampling. The resulting performance measures on the validation dataset can be found in Table 8. All models, except Bayesian Network, classify all bankruptcy observations to non-bankruptcy, as indicated by Type II error, although they have high overall accuracy 99.88% which is exactly the proportion of non-bankruptcy observations in the original data. For Bayesian Network without, its overall accuracy and Type II error decreased by 9.92% and 2.03%, respectively, while its Type I error increased by 3.66%. The ROC curve in Figure 6 further shows that all other models perform no better than random selection, except Bayesian Network.

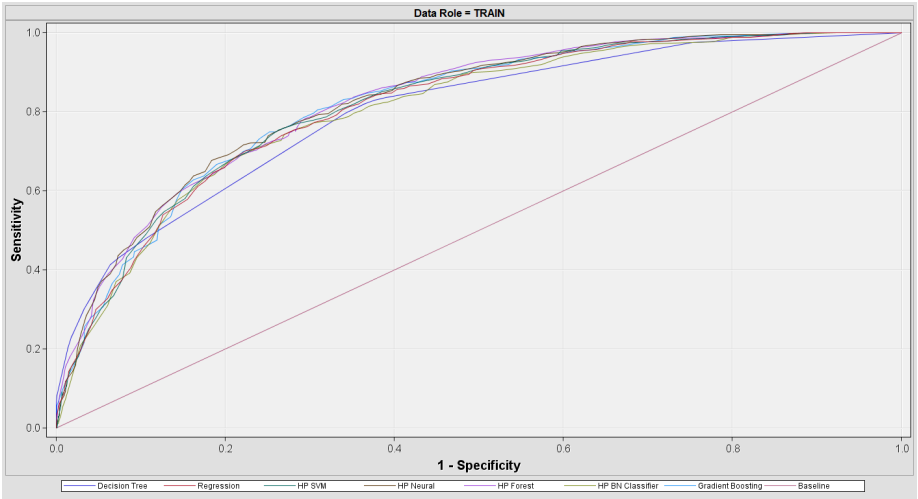

Figure 4. ROC Curve on Training Dataset

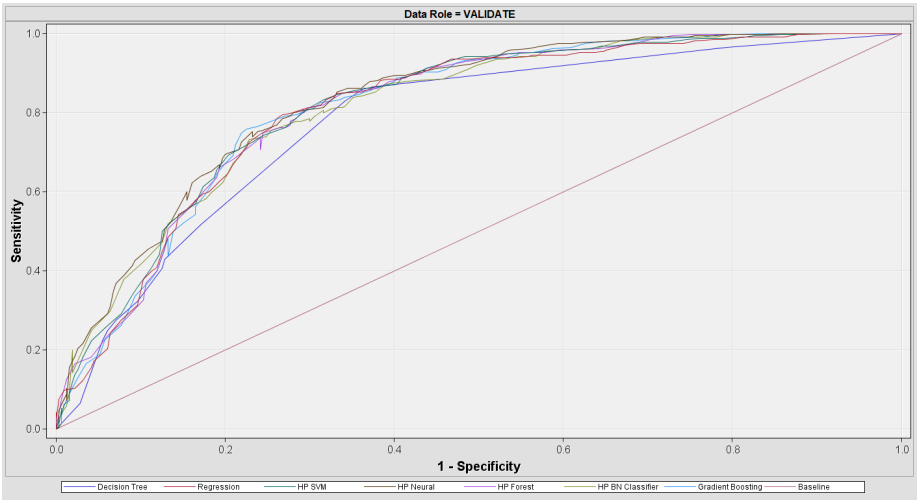

Figure 5. ROC Curve on Validation Dataset

Table 8. Performance of Models without Oversampling.

Model	**Accuracy**	**Type I Error**	**Type II Error**
Support Vector Machine	99.88%	0%	100%
Decision Tree	99.88%	0%	100%
Gradient Boosting	99.88%	0%	100%
Random Forest	99.88%	0%	100%
Neural Network	99.88%	0%	100%
Bayesian Network	64.43%	35.59%	17.32%
Logistic Regression	99.88%	0%	100%

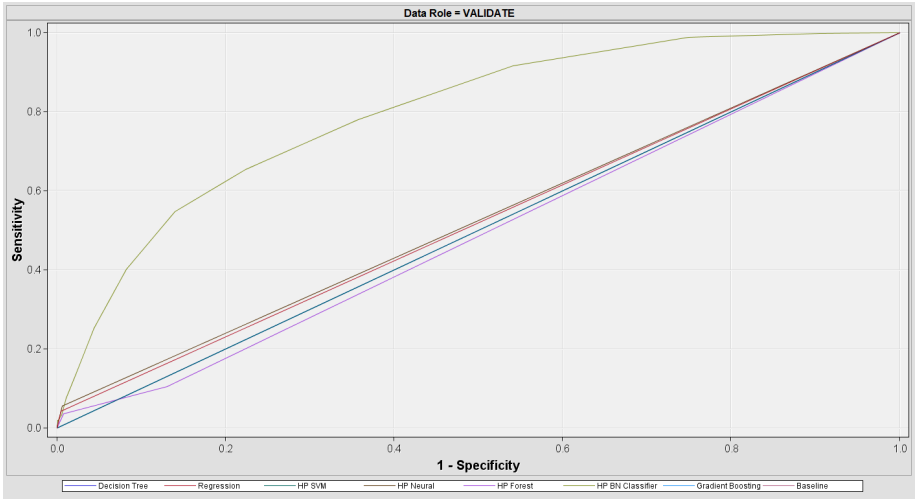

Figure 6. ROC Curve on Validation Dataset without Oversampling

6. Discussions and Conclusions

Based on the univariate analysis and multivariate analysis, the impacts of public records and firmographics indicators were comprehensively studied. With them as input variables of different classification models, the model results show that public records and firmographics indicators play an important role in the bankruptcy prediction. This may serve as a reference for practitioners and researchers to include these information in the bankruptcy prediction model.

Different classification models generate quite different Type I/II error, although their overall accuracy is similar. Support Vector Machine gives the lowest Type II error, and Logistic Regression gives the lowest Type I error. Regarding the interpretability, Logistic Regression, Decision Tree and Bayesian Network might be favorable choices. We also find that on the dataset with small/medium size, simple models may outperform complicated models like ensemble models and Neural Network. Practitioners may handle the tradeoff between Type I error and Type II error as well as the model interpretability and accuracy based on their expectations.

For rare event prediction, the oversampling is necessary before modeling to achieve better performance. Bayesian Network is quite robust for rare event prediction without oversampling, compared to other classification models.

7. Future Work

In this study, we only focused on the public records and firmographics indicators. In the future, we may collect other information like financial ratios to further reduce Type I/II error, and test the model performance in a wider time spanning.

References

[1] Ahn, B. S., S. S. Cho, and C. Y. Kim. "The integrated methodology of rough set theory and artificial neural network for business failure prediction." *Expert systems with applications* 18.2 (2000): 65-74.

[2] Altman, Edward I. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." *The journal of finance*23.4 (1968): 589-609.

[3] Altman, Edward I., and Paul Narayanan. "An international survey of business failure classification models." *Financial Markets, Institutions & Instruments* 6.2 (1997): 1-57.

[4] Atiya, Amir F. "Bankruptcy prediction for credit risk using neural networks: A survey and new results." *IEEE Transactions on neural networks* 12.4 (2001): 929-935.

[5] Bellovary, Jodi L., Don E. Giacomino, and Michael D. Akers. "A review of bankruptcy prediction studies: 1930 to present." *Journal of Financial education* (2007): 1-42.

[6] Chava, Sudheer, and Robert A. Jarrow. "Bankruptcy prediction with industry effects." *Financial Derivatives Pricing: Selected Works of Robert Jarrow*. 2008. 517-549.

[7] Everett, Jim, and John Watson. "Small business failure and external risk factors." *Small Business Economics* 11.4 (1998): 371-390.

[8] Foroghi, Daryush, and Amirhassan Monadjemi. "Applying decision tree to predict bankruptcy." *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on*. Vol. 4. IEEE, 2011.

[9] Huang, Shi-Ming, et al. "A hybrid financial analysis model for business failure prediction." *Expert Systems with Applications*35.3 (2008): 1034-1040.

[10] James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

[11] Kumar, P. Ravi, and Vadlamani Ravi. "Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review." *European journal of operational research* 180.1 (2007): 1-28.

[12] Min, Jae H., and Young-Chan Lee. "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters." *Expert systems with applications* 28.4 (2005): 603-614.

[13] Odom, Marcus D., and Ramesh Sharda. "A neural network model for bankruptcy prediction." *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*. IEEE, 1990.

[14] Shin, Kyung-Shik, Taik Soo Lee, and Hyun-jung Kim. "An application of support vector machines in bankruptcy prediction model." *Expert Systems with Applications* 28.1 (2005): 127-135.

[15] Sun, Lili, and Prakash P. Shenoy. "Using Bayesian networks for bankruptcy prediction: Some methodological issues." *European Journal of Operational Research* 180.2 (2007): 738-753.

[16] Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.

[17] Zhang, Guoqiang, et al. "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis." *European journal of operational research* 116.1 (1999): 16-32.

[18] Zięba, Maciej, Sebastian K. Tomczak, and Jakub M. Tomczak. "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction." *Expert Systems with Applications* 58 (2016): 93-101.

[19] Kim, Myoung-Jong, Dae-Ki Kang, and Hong Bae Kim. "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction." Expert Systems with Applications 42.3 (2015): 1074-1082.

[20] Zhou, Ligang. "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods." Knowledge-Based Systems 41 (2013): 16-25.

Authors

Lili Zhang is currently a Ph.D. candidate in Analytics and Data Science at Kennesaw State University, working on the data mining, machine learning, and data management.