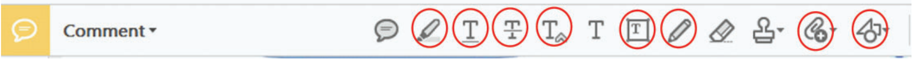
Page Proof Instructions and Queries

Jounal Title: BPR

Article Number: 764505

Greetings, and thank you for publishing with SAGE. We have prepared this page proof for your review. Please respond to each of the below queries by digitally marking this PDF using Adobe Reader (free at http://www.adobe.com/products/reader.html).

Please use ***only*** the circled tools to indicate your requests and responses, as edits via other tools/methods are not compatible with our software. To ask a question or request a formatting change (such as italics), please click the tool and then choose "Text Callout." To access the necessary tools, choose "Comment" from the right-side menu.



No.	Query
	Please confirm that all author information, including names, affiliations, sequence, and contact details, is correct.
	Please review the entire document for typographical errors, mathematical errors, and any other necessary corrections; check headings, tables, and figures.
	Please confirm that the Funding and Conflict of Interest statements are accurate.
	Please confirm you have reviewed this proof to your satisfaction and understand this is your final opportunity for review prior to publication.
AQ: 1	Please provide citations for Tables 1 and 2 and Figures 1 and 3.
AQ: 2	Please provide source information for Tables 1, 2, and 3 and Figures 1, 2, 3, and 4.
AQ: 3	Please provide Figures 1, 2, 3, and 4 in an editable format.
AQ: 4	Please provide issue number for the following reference: • Henley and Hand(1996)
AQ: 5	Please provide text citation for the following references: • Angelini et al. (2008) • Fu et al. (2006) • Huang et al. (2004) • Junqué de Fortuny et al. (2014) • Maher and Maysam (2015) • Mays (1998) • Patil,P., Aghav, J., Sareen, V. (2016).
AQ: 6	Please provide the publication details (full author name, title, name of the book/journal, page, volume, issue, publisher details, etc.) for the same to include in the reference list.: • Page 2, Para 3, Line 7: Angelini (2008) • Page 3, Para 1, Line 2: Hurley (2016) • Page 3, Para 2, Line 7: Zhou and Zhang (2008) • Page 4, Para 5, Line 2: Junque (2014) • Page 5, Para 3, Line 1: Huang (2004) • Page 5, Para 4, Line 2: Fu (2006) • Page 5, Para 5, Line 5: Wayne (2007) • Page 6, Para 2, Line 5: Ala'raj and Abbod (2015) • Page 6, Para 2, Line 9: Zareapoor and Shamsolmoali (2015) • Page 6, Para 2, Line 14: Patil (2016) • Page 9, Para 2, Line 6: Breiman (1996)

Application of Ensemble Models in Credit Scoring Models

Business Perspectives
and Research
6(2) 1–12
© 2018 K.J. Somaiya Institute of
Management Studies and Research
SAGE Publications
sagepub.in/home.nav
DOI: 10.1177/2278533718765531
http://bpr.sagepub.com

Anjali Chopra[1]
Priyanka Bhilare[1]

Abstract

Loan default is a serious problem in banking industries. Banking systems have strong processes in place for identification of customers with poor credit risk scores; however, most of the credit scoring models need to be constantly updated with newer variables and statistical techniques for improved accuracy. While totally eliminating default is almost impossible, loan risk teams, however, minimize the rate of default, thereby protecting banks from the adverse effects of loan default. Credit scoring models have used logistic regression and linear discriminant analysis for identification of potential defaulters. Newer and contemporary machine learning techniques have the ability to outperform classic old age techniques. This article aims to conduct empirical analysis on publically available bank loan dataset to study banking loan default using decision tree as the base learner and comparing it with ensemble tree learning techniques such as bagging, boosting, and random forests. The results of the empirical analysis suggest that the gradient boosting model outperforms the base decision tree learner, indicating that ensemble model works better than individual models. The study recommends that the risk team should adopt newer contemporary techniques to achieve better accuracy resulting in effective loan recovery strategies.

Keywords

Credit scoring model, probability of default, ensemble methods, accuracy, precision

Introduction

The competitive landscape in which retail banks are operating today has laid out a challenging market. Large-scale changes in the banking industry coupled with increasingly intense and diverse competition, and reduced customer trust have led to banks being under immense pressure. The conventional way of face-to-face customer interaction has been replaced by various electronics and digital point of contacts to reduce the cost and time of application process. Hence, it becomes more strenuous for the bank officials to understand each customer, to maintain personal relationship, and to evaluate risk associated with customer profiles.

Loan lending has been observed to be one of the primary businesses for most of the banks worldwide. To avoid the risk of default in loans, it is imperative for banks to find right customers with low credit risks. Bank officials have to play a dual role: On the one hand, they need to understand individual

[1] K.J. Somaiya Institute of Management Studies and Research, Mumbai, Maharashtra, India.

Corresponding author:
Anjali Chopra, K.J. Somaiya Institute of Management Studies and Research, Mumbai, Maharashtra, India.
E-mail: anjali.chopra@somaiya.edu

customers as part of customer relationship initiatives, and on the other hand, they have to evaluate risk associated with different customer profiles.

Large-scale changes in the industry coupled with intense competition from other financial institutions, financial technology startups, and mobile wallets offering customer-friendly options have resulted in deteriorated lending standards. Consequently, banks can expect numerous credit risks ranging from loan defaults and losses, frauds, and other related crimes. Thus, credit risk prediction becomes crucial step in the process of evaluation. Financial institutions use credit scoring as a tool to distinguish among good and bad borrowers. A good borrower is defined as the customer who pays back his loans. Credit scoring models developed by financial institutions assess the non-payment probability of potential borrowers. The credit risk management team provides the lending team with credit score cards of potential borrowers which distinguish creditworthy applicants from likely defaulters. They mainly focuses on maximizing the bank's risk adjusted rate of return by predicting and maintaining the credit risk exposure well within their acceptable parameters. Commercial credit analysts are the most sought-after in the banking industry. They are responsible for determining the ability of loan applicants to repay their loans on time. They perform financial background assessment of an applicant to decide whether to sanction the loan or not. Sometimes depending upon the credit score and repayment history, banks may decide to grant loan with specific conditions. These research insights help financial institutions to provide affordable loans with the best interest rate based on the credit worthiness of the applicants, thus shielding themselves from possibility of defaults.

Traditionally, linear discriminant analysis (LDA) and logistic regression analysis (LRA) have been the core techniques used to construct scoring models. Since credit decisions often involve several billions of dollars, even an improvement in accuracy of a fraction of a percentage point can lead to significant gains. This has prompted both researchers and practitioners to investigate every possibility of improving scoring model accuracy. The use of logistic regression models, nonparametric models, such as k-nearest neighbor, classification trees, and neural network models has been examined by Henley
AQ: 6 (1995), Henley and Hand (1996), Makowski (1985), and Angelini (2008), respectively, in the quest for higher scoring accuracy.

Hurley and Adebayo (2016) discuss credit scoring in the era of big data. According to them, the credit scoring industry has experienced an explosion of startups that take an "all data is credit data" approach, combining conventional credit information with thousands of data points mined from consumers' offline and online activities. Big data scoring tools may now base credit decisions on where people shop, the purchases they make, their online social media networks and various other factors that are not intuitively related to creditworthiness. While the details of many of these products remain closely guarded trade secrets, according to the authors, the proponents of big data credit scoring argue that these tools can reach millions of underserved consumers by using complex algorithms to detect patterns and signals within a vast sea of information. However, Hurley (2016) limits their discussion to how big data tools are transforming the credit scoring industry and the major risks and challenges these new tools pose.

While in-house is working on developing credit scoring models using data spanning different touch points, the credit risk teams of financial institutions are focused on improving the accuracy of existing credit scoring models through various classification techniques. The traditional LDA and logistic regression have low classification accuracy in the credit scoring, as the relationship among variables is linear. To improve the less accuracy of parametric statistical methods, many models based on data mining methods like decision trees proposed by Davis, Edelman and Gammerman (1992), Frydman,
AQ: 6 Altman and Kao (1985), and Zhou and Zhang (2008); artificial neural networks (ANN) by Jensen (1992), West (2000), West, Dellana and Qian (2005); and k-nearest neighbor by Henley and Hand (1996) have become popular.

A novel machine learning technique called ensemble learning is also being used for improving accuracy. A classifier ensemble (also referred to as committee of learners, mixture of experts, multiple classifier system) consists of a set of individually trained classifiers (base classifiers) whose decisions are combined in some way, typically by weighted or unweighted voting, when classifying new examples as stated by Kuncheva (2004). According to Dietterich (1997), it has been found that in most cases, the ensembles produce more accurate predictions than the base classifiers. Researchers have shown that aggregating approach can easily achieve improved accuracies by an aggregation of individual classifiers for credit scoring as well as the classification application.

The objective of this article is to demonstrate the superiority of newer techniques over the traditional data analysis models. This article aims to evaluate and compare a class of machine learning techniques called ensemble learning via decision trees in predicting loan default. The purpose of using ensemble learning is to compare the performance of old-school and contemporary approaches, and recommend models which have better ability to identify potential defaulters. In case of decision tree-based model, user might face practical difficulties like bias and variance. In general with an increase in the complexity of the model, we can expect reduced prediction error due to lower bias but overfitting can cause high variance. Ensemble learning has turned out to be more useful to execute this trade-off analysis of bias–variance errors.

This article will demonstrate the use of specialized class of ensemble models for improved classification. Beginning with traditional decision trees, several ensemble methods such as bagging, random forests, and boosting have been used. Bagging or bootstrap aggregation is a procedure used to reduce the variance of our predictions by combining the output of multiple classifiers modeled on different subsamples with replacement of the same dataset. This can be used for algorithm like classification and regression trees (CART) having high variance. Random forest is an improved version of bagged decision tree, which alters the algorithm for the way that the subtrees are learned, so that the resulting predictions from all of the subtrees have less correlation. Random forest is one of the most accurate learning algorithms and it also provides information of important variables in the classification but overfit is one of the major disadvantages of this technique with noisy classification/regression task. Boosting algorithms are used to combine multiple weak learners together to build a strong learner with exceptionally high predictive power. In case of classification problem, we need to assign each and every observation to a given set of class and binary class encourages the use of AdaBoost algorithm to find its classification boundary. In case of regression problem, where we have continuous variables to predict, the use of gradient boosting algorithm can be used to produce strong regression model having collection of weak predictors.

The precision and recall of classification depends on many aspects, including the selection of a suitable algorithm, the selection of a training dataset and so on. In this article, we have tried to focus on experiments with training dataset samples, with the aim to improve the precision and recall of results. Empirical observations and results draw helpful conclusions and further research directions. The authors have worked with ensemble model on personal loan data of one million customers of a leading private sector bank in India. However, due to data and customer confidentiality, it is not possible to share the results. Hence, we have used publicly available banking dataset for the demonstration of a set of ensemble techniques which were used on actual customer data.

Literature Review

One of the best understood ways to use data to improve decision-making is via predictive analytics. Many researchers have introduced and developed the concept of conventional data analytics in the field of banking and credit risk prediction. Through this research, they tried to utilize the most commonly used statistical techniques such as LDA, LRA, decision tree, and so on. Many financial institutes use these credit scoring models based on traditional statistical theories. It enables them to lower credit risk in credit appraisals, and in granting and supervising credit loans. The earliest application of LDA technique to banking products and services such as credit risk analysis, loan default, and fraud detection was contributed by Durand (1941).

However, these models are less resilient when it comes to large amounts of data input; therefore, some of the assumptions in the classical statistical analysis fail. This influences the accuracy of prediction and model generalizations. West (2000) researched on various neural network credit scoring models for commercial applications using LDA, LRA, and decision tree methods. He proposed that these techniques could be most accurate only in case when relationship between variables is linear and hence it might lack risk prediction accuracy.

An important, open question is as follows: To what extent do larger data lead to better predictive AQ: 6 models? Junque (2014) suggested that banking sector is characterized with larger data assets along with the skill to take advantage of them. Moving beyond traditional data analytics, private sector banks would obtain substantial competitive advantage by adopting machine learning techniques. A lot of literature work can be found in banking analytics and application of ensemble techniques as far as credit risk prediction history is concerned.

Ensemble method is currently thriving in banking industry. Ensemble modeling is the art of combining diverse set of learners (individual models) together to improvise on the stability and predictive power of the model. Dietterich (1997) through his research proved that it is an effective prediction technique which can help bankers to predict the credit risk while extending credit to loan applicants. Various ensemble techniques like bagging and boosting are used to implement predictive model and their accuracy has been compared to achieve better outcome.

AQ: 6 Huang (2004) identified significant difference between traditional statistical techniques and machine learning techniques. He indicated the ability of ensemble methods to learn various structures and patterns of the model from the data itself. Traditional methods are dependent on researchers to impose a particular structure like linearity by estimating parameters to fit the data.

Although most of the above analytics methods can be useful in developing corporate credit scoring AQ: 6 models, Fu (2006) showed that new ensemble techniques, which integrates multiple classifiers into an aggregated result, have shown higher prediction accuracy than any other independent method. Opitz and Maclin (1999) have also recommended ensemble techniques over other analytical methods available for the process of risk evaluation.

Predictive banking analytics can assist credit analysts to optimize functioning of existing processes, identify unexpected areas of opportunities, and anticipate future problems before they even occur. While some organizations have discovered the power of predictive analytics to reduce costs, increase revenues, and optimize business processes, the clear majority are still looking to derive value from their analytical AQ: 6 investments. Wayne (2007) recognized challenges faced by business managers while implementing these various techniques. Even though they understand the improvement that predictive analytics can bring to their organization, most of them are perplexed about how and where to begin. Tsai and Wu (2008) cited the performance of a single classifier as the base learner to compare with multiple classifiers and diversified multiple classifiers by using neural networks. The multiple classifiers exhibit better results in terms of accuracy when compared with the single classifier as the benchmark. Nanni and Lumini (2009) examined the performance of several analysis models based on ensemble methods for credit scoring. The results revealed that ensemble techniques can be used for improving the performance of "stand-alone" classifier.

Abdou and Pointon (2011) suggested the importance of credit scoring applications primarily in finance and banking sector. Suggested techniques can be helpful in identifying the key processes including collection, analysis, and classification of different credit variables while assessing expected risk of customer being bad credit. They strongly advocated that quality of bank loans is the key determinant of survival, competition, and profitability. Hence, accuracy of credit scoring tools is one of the most important criteria during the credit evaluation process. Accurate predictions will not only reduce the present and the future risk of customer being bad credit but also improve the profitability of lenders.

Wang and Ma (2011) revealed the experimental results, highlighting better performance of RS-Boosting algorithm among all other seven techniques, that is, decision tree, LRA, artificial neural network, bagging, boosting, and random forest. They illustrated that RS-Boosting method is more effective and feasible method which can be used as an alternative for credit risk prediction. Ala'raj
AQ: 6 and Abbod (2015) suggested a credit scoring model based on heterogeneous and homogenous classifiers. Ensemble algorithms were based on three classifiers, that is, artificial neural network, logistic regression, and support vector machine. They demonstrated that heterogeneous ensemble classifiers could give improved performance and accurate predictions as compared to homogeneous ensemble
AQ: 6 classifiers. Zareapoor and Shamsolmoali (2015) demonstrated the benefits of the bagging ensemble algorithm to construct detection model that can evaluate the real-life credit card transactions dataset which is highly imbalanced in nature. This model enabled bankers and other financial institutions to keep fraud catching rate high and false alarm rate very low. Proposed ensemble learning model could
AQ: 6 provide promising experimental results to credit risk prediction. Patil (2016) performed experimental analysis in R statistical programming language. The proposed robust data mining model to predict the defaulters using newer ensemble techniques could perform better with higher level of accuracy than individual algorithms such as LDA, LRA, and so on.

While various new techniques have been used to build fraud detection models, the ensemble techniques using logistic regression, artificial neural network, and support vector machine have been used to some extent in loan application models; the authors did not come across any study in the Indian context which compared and evaluated the performance of ensemble models with traditional decision trees, bagging, random forests, and boosting with the objective of improving recall as a performance measure.

Through this research, the authors have proposed superior performance of gradient boosting ensemble method in the context of Indian loan customers. This study recommends credit risk analysts to adopt newer ensemble techniques with higher accuracy in terms of precision and recall which would help banks to reduce risk in Indian market.

Methodology Adopted

The authors used publicly available banking data. The dataset had one special attribute—whether customer has previously defaulted and eight independent variables. The independent variables were as follows:

- Age in years
- Education level
- Number of years employed
- Number of years stayed at same address
- Household income in thousand dollars

- Debt to income ratio
- Credit card debt in thousands
- Other debt in thousands

The dataset consists of 850 customers, of which 700 were labeled (defaulter/non defaulter) while 150 were unlabeled. The research objective was as follows:

- Use training data of 700 customers to build a model which classifies defaulter and non-defaulter
- Evaluate several classifiers such as decision tree, random forest, bagging, ada-boosting, and gradient boosting on the two performance criteria precision and recall
- Identify the best classifier and build model to predict and classify 150 new customers

Predictive analytics platform RapidMiner (version 7) was used for the various methods and iterations. The model was built on 70 percent of the labeled data (490 out of 700 labeled customers), while 30 percent of the labeled data was set aside for evaluation (210 out of 700 labeled customers). After the model was trained, its performance was measured by comparing the predicted values against the labeled ones. Three key measures were used to evaluate the various models: accuracy, recall, and precision.

AQ: 1, 2

Table 1. Explanation of Accuracy, Precision, and Recall

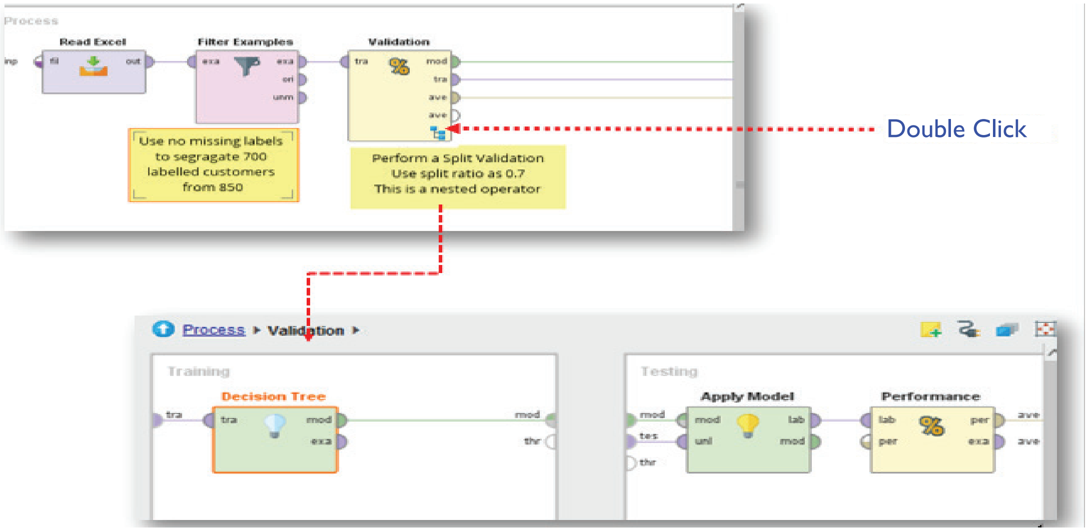
Accuracy = T1P1+T0P0/TOTAL	True 1 (1=Defaulter)	True 0 (0= Non-defaulter)	Precision
Predicted 1	T1P1	T0P1	T1P1/(T1P1+T0P1)
Predicted 0	TIP0	T0P0	T0P0/(T1P0+T0P0)
Recall	TIP1/(T1P1+T1P0)	T0P0(T0P1+T0P0)	

Source:

Accuracy is simply the proportion of correct results that the classifier has achieved. The precision for a class is the number of true positives (i.e., the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e., the sum of true positives and false positives). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e., the sum of true positives and false negatives). For the current dataset, we define precision and recall as follows:

- Precision—Of those predicted as defaulters, how many were correctly classified as defaulters (T1P1/(T1P1+T0P1))
- Recall—Of those who are defaulters, how many were correctly classified as defaulters (TIP1/(T1P1+T1P0))

From a bank's perspective, the business importance for the risk team would be to build a model that detects defaulters correctly. Right detection of defaulters would help the loan team to avoid giving loans to potential defaulters. Thus, it is desirable to have a very high recall, probably at loss of precision, since it is very important that all defaulters are identified or at least suspicions are raised. To begin with, decision tree was used as the base model. The process flow, the decision tree, and the confusion matrix result are given below.



AQ: 4 **Figure 1.** Rapid Miner Process for Decision Tree

Source:

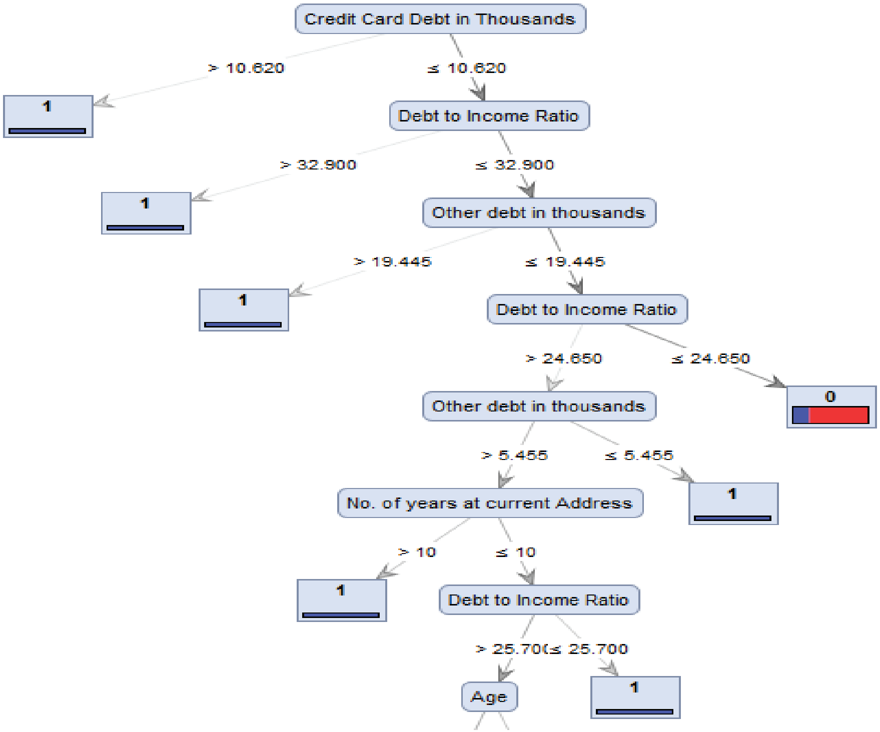
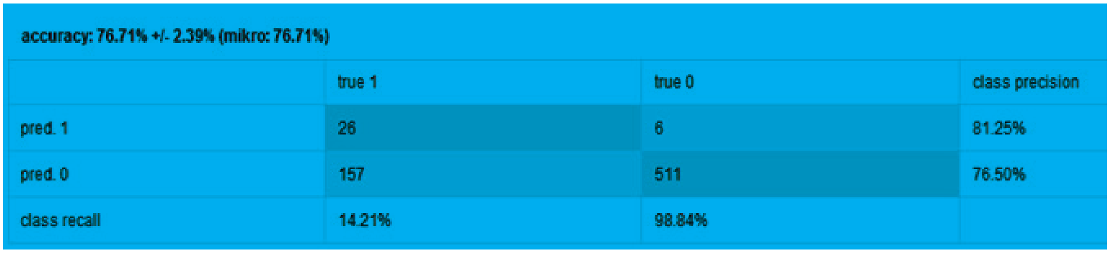


Figure 2. Decision Tree Output

Source:



accuracy: 76.71% +/- 2.39% (mikro: 76.71%)

	true 1	true 0	class precision
pred. 1	26	6	81.25%
pred. 0	157	511	76.50%
class recall	14.21%	98.84%	

Figure 3. Performance Based on Confusion Matrix

Source:

Reviewing the above results, we see that while both accuracy at 77 percent and precision at 81 percent are high, the recall measure at 14 percent is quite poor. As we know that in order to increase recall, precision might reduce but given the business objective of identifying defaulters, sacrificing some precision for increase in recall is desirable.

The next technique used was bootstrap aggregating, also called bagging, which is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging AQ: 6 approach. Bagging leads to "improvements for unstable procedures" (Breiman, 1996) by combining classifications of randomly generated training sets.

The bagging operator has two parameters.

- Sample ratio (N) indicates the fraction of records used for training
- Iterations (m) indicate number of base modes that need to be generated
- Bagging meta model acts as one model with multiple base models inside

Using an iterative process, we get the following results.

Table 2. Summary of Results of Bagging Iterations

For Defaulters	Accuracy	Precision	Recall
Decision Tree—Base Model Default	76	81	14
Bagging—Sample Ratio—0.9; iterations—100	76	79	13
Bagging—Sample Ratio—0.9; iterations—1000	76	79	13
Bagging—Sample Ratio—0.7; iterations—10	76	62	21
Bagging—Sample Ratio—0.7; iterations—100	76	62	21
Bagging—Sample Ratio—0.7; iterations—1000	76	62	21
Bagging—Sample Ratio—0.6, iterations—100	76	60	21
Bagging—Sample Ratio—0.5, iterations—100	76	62	18
Bagging—Sample Ratio—0.4, iterations—100	76	61	22
Bagging—Sample Ratio—0.3, iterations—100	74	52	30
Bagging—Sample Ratio—0.2, iterations—100	73	47	28
Bagging—Sample Ratio—0.1, iterations—100	72	45	34
Bagging—Sample Ratio—0.05, iterations—100	72	47	45

Source:

Compared to decision tree, bagging ensemble model was definitely superior to the base decision tree learner as it could improve the recall measure from 14 percent to 45 percent. It did reduce precision from 81 percent to 47 percent and overall accuracy from 77 percent to 72 percent. However, given that recall is an important criterion over precision, bagging was judged to be superior to decision tree.

The next models used were random forest and boosting (ada-boosting and gradient boosting). Random forest is similar to bagging. Random forest makes a small tweak to bagging and can sometimes result in a very powerful classifier. In both bagging and random forest, many individual decision trees are built on bootstrapped version of the original dataset and are ensemble together. However, the trees produced by different bootstrap samples can be very similar. Random forest overcomes this problem, where only a subset of features are selected at random out of the total and the best split feature from the subset is used to split each node in a tree, unlike in bagging where all features are considered for splitting a node. This results in different or uncorrelated trees in the sample. To reduce the generalization error, the algorithm is randomized into two levels, training record selection and attribute selection, in the inner working of each base classifier. In general, the model works using the following steps:

If there are *n* training records with *m* attributes, and let *k* be the number of trees in the forest, then for each tree:

- An *n* random sample is selected with replacement. This step is similar to bagging.
- A number *D* is selected, where $D << m$. *D* determines the number of attributes to be considered for node splitting.
- A decision tree is started. For each node, instead of considering all *m* attributes for the best split, a random number of *D* attributes are considered. This step is repeated for every node.
- As in any ensemble, the greater the diversity of the base trees, the lower the error of the ensemble.

Once all the trees in the forest are built, for every new record, all the trees predict a class and vote for the class with equal weights. The most predicted class by the base trees is the prediction of the forest. Results of random forest are given in Table 3.

Boosting is another approach for improving the prediction power from a decision tree. Boosting works similarly to bagging except that the trees are grown sequentially. Each tree is grown using information from previously grown trees. Boosting also does not involve bootstrapping; instead, each tree is fit on a modified version of the original data. The boosting process concentrates on the training records that are hard to classify and over-represents them in the training set for the next iteration. In general, boosting can improve upon random forests and is easier to interpret because of the smaller tree structure. The boosting model is built by an iterative and sequential process where a base model is built and tested with all of the training data, and based on the outcome the next base model is developed.

AdaBoost and gradient boosting are two popular boosting techniques. AdaBoost works by weighting the observations, putting more weight on difficult to classify instances and less on those already handled well. New weak learners are added sequentially that focus their training on the more difficult patterns. This means that samples that are difficult to classify receive increasing larger weights until the algorithm identifies a model that correctly classifies these samples (Kuhn, 2013). Gradient boosting is a modified version of boosting where it builds an ensemble of trees one-by-one and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Gradient boosting has three parameters unlike AdaBoost which focuses on only one parameter.

- n.trees: Number of trees (the number of gradient boosting iteration), that is, *N*. Increasing *N* reduces the error on training set, but setting it too high may lead to overfitting.
- interaction.depth (Maximum nodes per tree): Number of splits it has to perform on a tree (starting from a single node); a split of 6—node tree appears to do an excellent job.

- Learning Rate: Reduces the size of incremental steps and thus penalizes the importance of each consecutive iteration. If one of the boosting iterations turns out to be erroneous, its negative impact can be easily corrected in subsequent steps. Conventional use requires very slow learn rates for small datasets and use 0.1 for all datasets with more than 10,000 records.

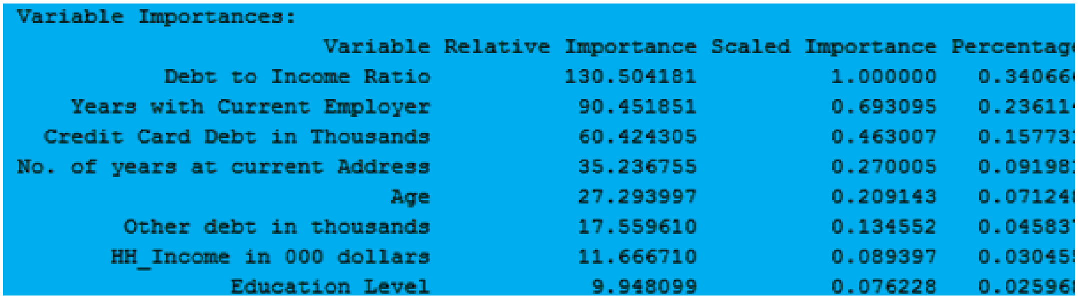
Table 3. Performance Comparisons for the Models Presented Above

For Defaulters	Accuracy	Precision	Recall
Decision Tree—Default	76	81	**14**
Bagging—Sample Ratio—0.05, iterations—100	72	47	45
Random Forest—50 trees, Confidence vote	75	80	4
Random Forest—500 trees, Confidence vote	75	83	5
Ada-boosting—100 iterations	76	81	14
Gradient boosting optimized	81	67	52

Source:

- Decision tree had a poor recall measure but high precision measure and overall accuracy.
- Bagging greatly improved the recall measure (from 14% to 47%) while reducing the precision measure (81% to 47%) and the overall accuracy slightly (72%).
- Random forest gave the worst recall measures.
- Ada-boosting gave very similar results to decision tree.
- Gradient boosting was the best classifier with recall scores of 52 percent and precision at 67 percent. The drop in precision is more than offset by the increase in the recall measures and overall accuracy also improves significantly to 81 percent.
- Thus, gradient boosting was identified as the classifier which would be used to classify the 150 unlabeled customers as opposed to decision tree.

The final model selected based on the empirical tests was gradient boosting. Gradient boosting technique was then applied to the 150 unlabeled customers. The model predicted 38 customers (25.33%) as defaulters and 112 (75%) as non-defaulters. The model identified debt to income ratio as the most important parameter. Referring to Figure 2, it was inferred that the base learner decision tree identified credit card debt in thousands as the most significant variable. The base learner classifier had poor performance measure with only 14 percent recall. The gradient boosting ensemble method had significantly higher performance and was selected as the final model for classifying the unlabeled customers. Summary of the importance of variables as identified by gradient boosting is provided in Figure 4.



```
Variable Importances:
                     Variable Relative Importance Scaled Importance Percentag
           Debt to Income Ratio          130.504181          1.000000   0.34066
      Years with Current Employer           90.451851          0.693095   0.23611
     Credit Card Debt in Thousands           60.424305          0.463007   0.15773
No. of years at current Address           35.236755          0.270005   0.09198
                            Age           27.293997          0.209143   0.07124
          Other debt in thousands           17.559610          0.134552   0.04583
            HH_Income in 000 dollars           11.666710          0.089397   0.03045
                  Education Level            9.948099          0.076228   0.02596
```

Figure 4. Relative Importance of Attributes

Source:

Managerial Implications

Over the last decade, credit scoring has been transforming how traditional financial institutions interact with customers. Initially, credit scores were developed based on qualitative information. This system was called subjective scoring and financial institutions relied on the experience of the risk team to develop subjective scores. This system worked when the customer base was small. However, as financial institutions were facing pressure to increase the scale of customers and competition among lenders, intensified credit scoring models were developed based on quantitative methods. Large volumes of data collected on customers ranging from past credit behavior, payment history, personal factors, and characteristics were used to develop statistical models which determined the weight of each of these factors. The right credit scoring models help banks and financial institutions to increase their risk tolerance and offer loans to a wider segment of creditworthy customers without undermining their profit margins. However, credit scoring also has its drawbacks, with traditional scoring models leaving out some creditworthy customers or giving loans to potential defaulters resulting in higher risk and cost of operations.

Traditional models help the risk team to evaluate potential risks and identify borrowers who are likely to default on loan obligation, by using different assessment criteria to know the borrower's financial condition. Despite having robust models that extensively assess an individual's financial credibility, banks are still exposed to the risk of potential loan default. With the help of advanced data analytics and contemporary prediction techniques, banks are exploring ways in which credit scoring models can be made more robust with higher accuracy level. Historically, traditional models like logistic regression and decision trees have been used, but recent research has indicated that ensemble models work better than individual models with higher accuracy.

Ensemble method is a robust machine learning paradigm which has exhibited more apparent benefits in many applications. This article has presented an empirical investigation of the use of ensemble models for customer loan default prediction under different training algorithms including bagging, random forest, and boosting (AdaBoost and gradient boost). This article has also focused on the comparative study of base decision tree classifier and the ensemble models. The results demonstrated that ensemble model using gradient boosting improved the recall measures substantially (14% to 52%) and was the best performing algorithm. Credit scoring models will greatly benefit by the use of new contemporary techniques. However, in the long run, banks would also need to focus on incorporating non-traditional data like social media data to make credit scoring models more robust.

AQ: 5

References

Abdou, H. A. H., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting*, *Finance & Management*, *18*(2–3), 59–88.

Angelini, E., Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, *48*(4), 733–755.

Davis, R. H., Edelman, D. B., & Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. *IMA Journal of Management Mathematics*, *4*(1), 43–51.

Dietterich, T. G. (1997). Machine-learning research: Four current directions. *AI Magazine*, *18*(4), 96–136.

Durand, D. (1941). Risk elements in consumer instalment financing. *New York: National Bureau of Economic Research*, *8*(1), 22–43.

Eckerson, W. W. (2007). Predictive analytics extending the value of your data warehousing investment. *TDWI Best Practices Report*, *1*(1), 19–27.

Frydman, H., Altman, E. I., & Kao, D. L. (1985). Introducing recursive partitioning for financial classification: The case of financial distress. *The Journal of Finance*, *40*(1), 269–291.

Fu, Z. W., Golden, B. L., Lele, S., Raghavan, S., & Wasil, E. (2006). Diversification for better classification trees. *Computers & Operations Research*, *33*(1), 3185–3202.

Henley, W. E. (1995). *Statistical aspects of credit scoring* (PhD thesis). UK: Open University.

Henley, W. E., & Hand, D. J. (1996). A k-NN classifier for assessing consumer credit risk. *The Statistician*, *65*, 77–95.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, *37*(4), 543–558.

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale Journal of Law and Technology*, *18*(1), 148. Retrieved from http://digitalcommons.law.yale.edu/yjolt/vol18/iss1/5

Jensen, H. L. (1992). Using neural networks for credit scoring. *Managerial Finance*, *18*(6), 15–26.

Junqué de Fortuny, E., Martens, D., & Foster, P. (2014). Predictive modeling with big data: Is bigger really better? *Big Data*, *1*(4), 215–226.

Kuncheva, L. I. (2004). *Combining pattern classifiers: Methods and algorithms*. Hoboken, NJ: Wiley.

Maher, A., & Maysam, A. (2015). A systematic credit scoring model based on heterogeneous classifier ensembles. *Innovations in Intelligent SysTems and Applications (INISTA), 2015 International Symposium*, *10*(3), 471–495.

Makowski, P. (1985). Credit scoring branches out. *The Credit World*, *75*(1), 30–37.

Mays, E. (1998). *Credit risk modeling*. Chicago: Glenlake Publishing.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *36*(2), 3028–3033.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*(1), 169–198.

Patil, P., Aghav, J., & Sareen, V. (2016). An overview of classification algorithms and ensemble methods in personal credit scoring. *IJCST*, *7*(2), 183–187.

Tsai, C-H., & Wu, J-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *34*(1), 2639–2649.

Wang, G., & Ma, J. (2011). Study of corporate credit risk prediction based on integrating boosting and random subspace. *Expert Systems with Applications*, *38*(4), 13871–13878.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(1), 1131–1152.

West, D., Dellana, S., & Qian, J. X. (2005). Neural network ensemble strategies for financial decision applications. *Computers & Operations Research*, *32*(10), 2543–2559.