ORIGINAL RESEARCH

Default definition selection for credit scoring

Terry Harris

Credit Research Unit, Department of Management Studies, The University of the West Indies, Cave Hill Campus, W.I., Barbados

Correspondence: Terry Harris. Address: Credit Research Unit, Department of Management Studies, The University of the West Indies, Cave Hill Campus, W.I, Barbados. Email: terry.harris@cavehill.uwi.edu

Received: July 18, 2012 **Accepted:** August 19, 2013 **Online Published:** September 17, 2013
DOI: 10.5430/air.v2n4p49 **URL:** http://dx.doi.org/10.5430/air.v2n4p49

Abstract

In this paper some of the main causes of the recent financial crisis are briefly discussed. Specific attention is paid to the accuracy of credit-scoring models used to assess consumer credit risk. As a result, the optimal default definition selection (ODDS) algorithm is proposed to improve credit-scoring for credit risk assessment. This simple algorithm selects the best default definition for use when building credit scorecards. To assess ODDS, the algorithm was used to select the default definition for the random forest tree algorithm. The resulting classification models were compared to other models built using the unselected default definitions. The results suggest that the models developed using the default definition selected by the ODDS algorithm were statistically superior to the models developed using the unselected default indicators.

Key words

Credit Risk, Credit Risk Assessment, Credit Scoring, Default Definition Selection, Default Definitions, Random Forest

1 Introduction

In the wake of the global financial crisis, developing more accurate techniques to assess credit risk has become an important pursuit of managers at financial institutions across the developed and developing world. To understand why this is so, one simply needs to examine what is widely considered to be the root cause of the recent financial malaise. That is the granting of subprime mortgages in the United States of America (USA) and the subsequent collapse of the market for securities backed by these assets. This caused financial institutions to cut-back on lending as they faced increased risk of losses associated with (i) the granting of consumer credit and (ii) the provision of corporate loans.

Historically, subjective methods based on staff experience and judgment, have been used to assess credit risk. However, due to the increased demand for credit, quantitative credit risk assessment techniques have received increased attention from decision makers in the financial services industry as they attempt to automate and standardise credit risk assessment. Results have demonstrated that using such models can lead to more accurate and timely credit approval decisions [1].

One popular operations research (OR) application applied to quantitative credit risk assessment is credit-scoring. To help managers at financial institutions make credit granting decisions, this method involves evaluating the likelihood that a credit applicant will default according to some statistical model built from past clients' socio-demographic data (e.g. income, expenses, employment status, etc) and their credit behaviour (whether they defaulted or not). According to this, potential clients are classified as either "good" or "bad" for credit [2].

The development of credit-scoring models based on past client behaviour involves the selection of certain criteria that is used to specify when a default has occurred. Over the years, many models have used the 90 days past due rule as the default definition. This rule is one of two criterion proposed by the Basle Committee on Banking Supervision that considers a default to have occurred when (i) the financial institution considers that the obligor is unlikely to pay/repay their obligations in full, and the institution is unable to sell assets held as security in order to satisfy the obligor's debt, or (ii) the obligor is more than 90 days past due on any material amount owing to the financial institution [3].

Harris [4] investigated whether the performance of credit-scoring models built using the support vector machine (SVM) classification algorithm—a machine learning (ML) technique—could be improved if different default definitions are used (e.g. 30 days past due and 60 days past due). He referred to these models as "broad" and "narrow" models. His results suggest that the default definition has an impact on model performance. This paper takes a closer look at these findings and tests whether they are generalisable to another sample dataset provided by a Barbados based financial institution and a different classification algorithm. Specifically, models will be developed using another popular ML algorithm applied in the fields of OR and management science (MS)—the random forest tree algorithm. In addition, to help find and select the optimal default definition within a search space of possible default definitions the optimal default definition selection (ODDS) algorithm is proposed. To the best knowledge of the author, this work represents the first proposal of such an algorithm for default definition selection in the credit-scoring literature.

This paper is outlined as follows. Section 2 presents a brief background to the global economic crisis, its origins in the USA and its impact on the Barbados financial services industry. In Section 3 the practice of credit-scoring is briefly discussed. The decision tree and the random forest tree algorithms are presented in Section 4. The dataset is presented in Section 5. In Section 6 the ODDS algorithm is presented. Section 7 describes the study methodology. Section 8 highlights the statistical tests used to in this paper. Section 9 discusses the results of the study, while Section 10 presents the conclusions and the author's future research interests.

2 The roots of the crisis and the rationale for credit scoring

2.1 The cause of the crisis

In the USA successive governments have promoted the availability and use of consumer credit. Accordingly, the Congress has passed legislation such as the Equal Credit Opportunity Act (ECOA), the Fair and Accurate Credit Transaction Act (FACTA), the Fair Housing Act (FHA), and the Community Reinvestment Act (CRA). These laws were designed to reduce discriminatory lending practices by banks and other financial institutions and promote lending. To achieve the congressional objectives, these Acts gave federal authorities the power to prosecute financial entities that did not comply. As a result, USA based financial institutions began to give loans to a number of individuals who were once not eligible for credit. The techniques used to assess the creditworthiness of these persons determined some of them to be high-risk but acceptable for credit. Hence, mortgages were given and these loans were referred to as subprime mortgages.

Traditionally, financial institution such as banks practiced an "originate and hold" business model where banks used deposits to fund loans. These loans were kept on the institution's balance sheet until maturity. This practice meant that (i) the financial institution was exposed to all defaults resulting from its lending activities and (ii) the loans held on the balance sheet froze-up capital that was required by regulators to offset any potential loan defaults, thereby, costing the institution the opportunity of granting new loans from this capital. Subsequently, financial institutions sought to address these constraints by moving to an "originate and distribute" business model. In addition to allowing the institution to un-freeze capital, this permitted the credit risk associated with the granting of prime and sub-prime mortgages to be mitigated and transferred through securitisation. Thus individual financial institutions were able to remove the risk attached to the future cash receipts of interest payments and principal repayments by selling the mortgage receivables in the present. Historically, this was achieved via the use of specifically created companies called special purpose entities

(SPEs) that purchase the financial institution's mortgage receivables. The SPEs raised the capital to support these purchases through the issuance of bonds. These bonds were held as assets (investment instruments) by investors in SPEs and are known generally as mortgage backed securities (MBS).

The subprime mortgages issued to high risk individuals required additional measures to deal with the underlying riskiness of the future cash flows in order to make them attractive to potential investors with varying risk appetites. Securitisation was achieved by the creation of derivative-style instruments known as collateralised debt obligations (CDOs). These instruments concentrated the risk into different investment layers or tranches, so that some investors took proportionately more of the underlying default (credit) risk for a larger return while others took-on less risk, which was better suited to their risk preference. Thus a series of notes of different seniorities given ratings ranging from AAA to high yielding by credit rating agencies were issued. Having received the "blessings" of the credit rating agencies and the protection provided by other banks and insurance agencies through credit default swaps, investors around the world were willing to purchase CDO bonds [5]. Figure 1 illustrates the cash flows from a mortgaged backed CDO transaction.

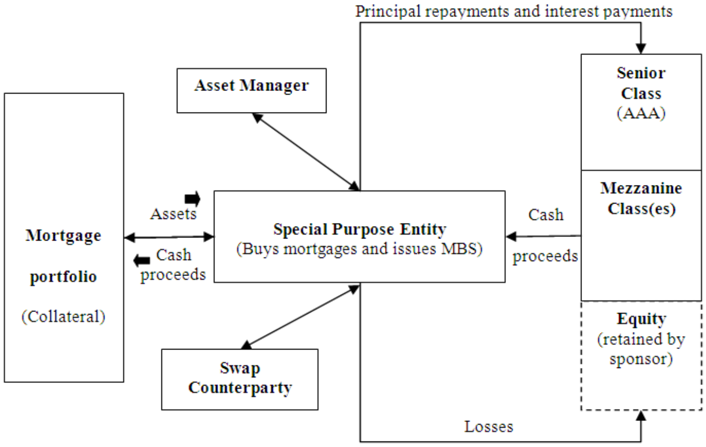

Principal repayments and interest payments
Asset Manager
Senior Class (AAA)
Mezzanine Class(es)
Mortgage portfolio (Collateral)
Assets
Cash proceeds
Special Purpose Entity (Buys mortgages and issues MBS)
Cash
proceeds
Equity (retained by sponsor)
Swap Counterparty
Losses

Figure 1. Showing example of mortgage backed CDO structure and cash flow

Unfortunately, the CDO instruments did not perform as expected and senior investment grade tranches, given AAA ratings by credit rating agencies, began to experience defaults on an unimaginable scale. The declining economic fortunes of the USA economy compounded this problem as rising unemployment and falling property values led to an even greater number of defaults. This meant that some financial institutions were left holding CDO instruments worth much less than their recorded book values. As financial institutions tried to offload these instruments a market (demand) did not exist. Hence, the market values of CDO bonds continued to fall. Given that CDO instruments were actively traded globally this became a worldwide phenomenon. In the absence of being able (or willing) to sell these investments, prevailing accounting rules meant that these financial assets had to be reported at fair value through the firm's profit and loss (income statement). Hence, markdowns in these portfolios were debited (expensed) to the income statement and credited to the asset's section of the balance sheet (i.e. CDO assets were written off). These fair value rules were pro-cyclical as a downturn in the USA economy and housing market led to; losses due to asset markdowns, these losses triggered a deterioration of capital base, leading to lower creditworthiness ratings from rating agencies, and higher costs of capital [6]. A global financial crisis had begun.

Furthermore, because the risk associated with these defaulting subprime loans, now viewed as so called "toxic assets", was distributed through securitisation, the institutions most under threat could not be pinpointed and this led to a contagion of risk to all investors in CDOs as these subprime loans were a poison pill to any securitised loan portfolio. As a result, financial institutions became reluctant to lend to other financial entities out of the fear that the other institution may be insolvent. Simultaneously, the practice of securitisation tapered-off as investors became more hesitant to invest in the resulting bond issues. This meant that financial institutions had to keep more capital in order to satisfy regulatory requirements. Collectively, these occurrences led to a global credit crunch which helped cause the worldwide recession [7].

Looking back, irrespective of the virtue of public policy decisions such as the ECOA, FACTA, FHA, and CRA, the role of fair value accounting rules, the alleged malpractice by credit rating agencies, or the myopia of managers at various financial institutions, it is clear that one of the underlying causes of the credit crunch was a lack of accurate consumer credit risk assessment that would have classified high risk individuals as un-creditworthy without being seen as discriminatory on the grounds of race, religion or ethnicity. Developing these type of models is a real challenge for OR practitioners as the law, personal, and professional ethics dictate that care be taken when designing classification models so as not to unduly disadvantage any specific group [8]. Indeed, had the initial decision been to deny credit to individuals who were seen as high risk as oppose to accepting and then transferring the risk, it is reasonable to believe that the credit crunch and the resulting global economic recession may not have occurred in the way in which it did. Thus, if financial institutions are to successfully leverage risk, abide by public policy directives, and not damage the economy, better means of credit risk assessment need to be pursued. In this regard, the development of more accurate credit scorecards through the use of OR and ML techniques present one possible means of reducing the exposure of financial institution to losses associated with client default.

2.2 The impact of the crisis on the Barbados financial services industry

The financial services industry in Barbados—a Small Island Developing State (SIDS)—has been facing significant challenges as a result of the global economic crisis. This has been indicated by the Central Bank of Barbados which in its annual “Financial Stability Report” disclosed that local financial institutions were experiencing increased delinquency rates since 2007/2008 [9, 10].

This situation has been compounded by the fact that Barbados based financial institutions have been slow to adopt more accurate quantitative techniques for credit risk assessment. Preliminary investigation comprising interviews with senior officials at many of the country’s leading financial institutions have suggested that a lot of them still use traditional judgmental approaches to assess applicants’ credit risk. This ossification has resulted in some institutions being over exposed to the risk of default, and this has made the island’s economy more susceptible to external economic events. As a result, the financial crisis has led to a significant deterioration of credit quality and diminution of profit margins.

The credit-scoring literature has chronicled the improvements in classification accuracy possible if modern approaches to credit risk assessment are adopted. Thus, these methods need to be developed in order to help protect the Barbados financial services sector by shielding local firms from losses resulting from an over exposure to credit risk. This is important in the Barbados case because the financial services sector is a major pillar of the Barbados economy. Furthermore, as a small open economy Barbados is characterised by its high susceptibility to exogenous financial shocks. As a result, all efforts to safeguard its institutions must be pursued as it represents a critical area to the success of Barbados’ national development strategy.

3 Credit scoring

Fisher [11] in his seminal work proposed the use of discriminant analysis to differentiate between two or more populations (classes) in a dataset. Since this time, a number of other studies have outlined further statistical methods aimed this task. Many of these techniques have been applied to build quantitative credit scorecards for financial institutions, and over the years credit-scoring has been demonstrated to be an efficient and effective means of assessing a loan applicant’s credit risk. One reason for this is that credit-scoring allows the credit approval decision process to be automated. Hence, credit-scoring models typically reduce the processing time of applications, and this helps to cut costs. In addition, basing credit approval decisions on the outputs of quantitative credit scorecards has been shown to increase the accuracy and consistency of the credit risk assessment exercise, as credit granting decisions are based on statistical models derived from data rather than the subjectivity of human decision-makers.

Durant [12] and Altman [13] both applied Fisher's [11] discriminant analysis in credit-scoring. Durant used discriminant analysis to assess the creditworthiness of car loan applicants, while Altman used it to explore corporate bankruptcy proposing his popular Z-scores. To build this type of classifier, let A represent the population of all possible credit applicants who are characterised by the features $X = (x_1, x_2, x_3, \cdots, x_n)$, where the variable n represents the number of applicant attributes that are collected by the financial institution (e.g. employment status, salary, expenses, etc). Accordingly, A can be divided into subset A_g and A_b, where A_g represents the set of applicants who are creditworthy and A_b the set of applicants who are un-creditworthy. Similarly, let the proportion of applicants from the population who are creditworthy be denoted P_g and un-creditworthy P_b. When making the credit assessment decision the decision maker wants to accept the members of A_g while rejecting the members of A_b. However, there is a chance of rejecting a member of A_g and accepting a member of A_b; allow the misclassification costs associated with making these errors to be C_g and C_b, respectively. Assuming normality, the conditional density function of both subsets A_g and A_b can be represented by the formulas $f_g(X)$ and $f_b(X)$ which follow the multivariate Gaussian distribution with means $\mu_g$and μ_b, (where $\mu_g \in \mathbb{R}^n$ and $\mu_b \in \mathbb{R}^n$) and covariance matrix $\Sigma \in \mathbb{R}^{n:n}$, which is positive semi-definite. Thus the conditional density functions $f_g(X)$ and $f_b(X)$ can be represented as follows;

$$f_g(X) = (2\pi)^{-\frac{n}{2}}(\det\Sigma)^{-\frac{1}{2}}\exp((-(X.\mu_g)^T\,\Sigma^{-1}(X.\mu_g))/2) \tag{1}$$

and

$$f_b(X) = (2\pi)^{-\frac{n}{2}}(\det\Sigma)^{-\frac{1}{2}}\exp((-(X.\mu_b)^T\,\Sigma^{-1}(X.\mu_b))/2) \tag{2}$$

Unfortunately, the true values for the population means μ_g and μ_b, and the covariance matrix Σ are unknown. However, these values can be estimated from a representative sample of past applicants. Therefore, let the vector $\bar{x}_g$ ($\bar{x}_g \in \mathbb{R}^n$) represent the sample mean of past applicants belonging to set A_g, $\bar{x}_b$ ($\bar{x}_b \in \mathbb{R}^n$) the sample mean of past applicants belonging to set A_b, and S the n by n sample covariance matrix. The classification rule base on the sample values can be derived where a new applicant (described by the vector of features X) would be assigned to A_g if the inequality as in (3) is satisfied and assigned to A_b otherwise;

$$\left[X - (\bar{x}_g + \bar{x}_b)\frac{1}{2}\right]S^{-1}(\bar{x}_g - \bar{x}_b) > \ln((C_bP_b)/(C_gP_g)) \tag{3}$$

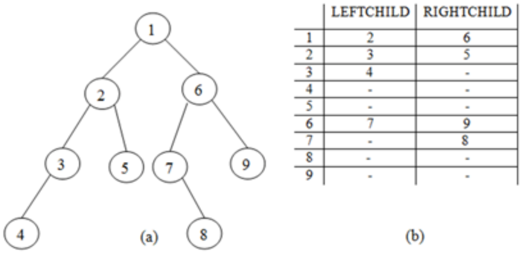
In addition to Fisher's discriminant analysis, other statistical techniques such as linear regression and logistic regression have been used to build credit scorecards [14]. Like disciminant analysis these classical methods have the advantage of being human interpretable (i.e. a human expert can determine how and why the model derived its classification decision) and work well when the data is linearly separable. Unfortunately, linear separability is not always the case in practical credit-scoring as data-frame curvature and interactions among variables often need to be considered when building models. Despite the best efforts of data modelers, there are some limitations to the extent these problems (particularly curvature) can be solved using classical methods. In recent times, more complex techniques such as artificial neural networks (ANNs), genetic algorithms (GAs), SVMs, decision trees, and random forest trees, have been develop to deal with problems characterised by nonlinearly separable data with unknown distributions (e.g. computer vision, digital character recognition, etc). These methods, although not offering the same level of understandability as conventional statistical techniques, have been applied in the credit-scoring space with notable success [15-20]. The works of Rosenberg and Gleit [21], Crook et al. [2], and Yu [22] provide a more comprehensive review of these and other contemporary classification methods.

4 Decision trees and random forest

4.1 Decision trees

A decision tree is a type of directed acyclic graphical model that is frequently used for binary classification. As directed (rooted) trees (see Figure 2(a)), decision trees satisfy the following properties:

- There is exactly one vertex (node) that has no edges entering it. This vertex is known as the root.
- Every node except the root has exactly one edge entering it.
- There is a unique path from the root to each vertex.



	LEFTCHILD	RIGHTCHILD
1	2	6
2	3	5
3	4	-
4	-	-
5	-	-
6	7	9
7	-	8
8	-	-
9	-	-

Figure 2. A binary decision tree (a) and its representation in table (2 parallel arrays) form (b)

Decision trees have been successfully implemented in many practical credit-scoring systems to separate “good” applicants from “bad” [23]. Decision trees can be implemented as Abstract Data Types (ADTs) which are easily represented in memory using parallel arrays (see Figure 2(b)) or linked data structures.

To achieve binary classification, a decision tree is made-up of a set of sequential binary splits on the dataset. When constructing a decision tree classifier the financial institution uses historic data collected from and on past credit applicants/clients. This data contains the past clients’ characteristics (attributes/features) and their resulting credit behaviour (whether they were creditworthy or un-creditworthy). Letting the variable, m, represent the number of instances (past clients) in the dataset, the variable, n, denote the number of client features, and the, $x_i \in \{x_1, x_2, x_3, \dots, x_n\}$, the individual attributes. When a financial institution builds a decision tree classifier, the algorithm begins at the root node and iterates over all possible binary splits in order to find the clients’ feature, x_i, and corresponding cut-off values, c, that achieve the optimal separation between creditworthy and un-creditworthy classes. This process is then repeated for the resulting children vertices until some stopping condition is met. Allowing the purity, p, of a vertex to be defined as the fraction of creditworthy instances it contains, the splitting attributes and cut-off values are those that minimise the sum of the Gini indices $p(1 - p)$ of the children nodes. This results in children nodes that are more homogeneous than their parent nodes. This is because for any attribute or cut-off value, if the sum of the Gini indices of the children vertices is higher than the Gini index of the parent vertex the parent vertex is not split.

The un-split children vertices, now referred to as “leafs” (see Figure 2(a)) are classified based on their most common class. As a result, a leaf node will be classified as “good” if it contains a majority of creditworthy instances and “bad” otherwise. Thus, when the financial institution is presented with a new credit applicant the institution uses the decision tree model to classify the applicant as either “good” or “bad”. This is determined based on the classification of the resulting leaf node that is arrived at after traversing the tree model using the applicant’s features. Hence, an applicant who lands on a “good” leaf is classified as creditworthy, while an applicant whose attributes traverse to a “bad” leaf is classified as uncredit-worthy [24].

Despite their popularity there are some well documented limitations when using decision trees for credit-scoring. First is the tendency for decision trees to over-fit the training dataset when growing large trees. Over-fitting is a problem because it can reduce the generalisability of the model and hence its usefulness in practice. To remedy this, insignificant branches (sub-trees) could be cut (this is known as "pruning"). Another issue that arises when using decision trees for credit-scoring is that small shifts in the data sample used to build the model can result in large variations in the classifications assigned to particular cases. This issue can be dealt with by growing a number of decision tree classifiers—a "forest"—and basing the new applicant's classification on the majority vote of the individual trees. Furthermore, to reduce the probability that any one tree is identical to another in the ensemble of tree classifiers, the historic dataset which contains the records of past applicants can be shuffled and the data instances chosen for the training samples selected at random. The aggregated decision tree approach is taken when building random forest of decision trees for credit-scoring.

4.2 Random forest

Random forest is an extension on the decision tree algorithm, and was first introduced by Breiman [25] for regression and classification problems. As an aggregated tree technique, the random forest algorithm determines the class label of an unknown input vector *x*, through the majority vote of component decision trees. Random forest has been determined to give comparable performance to other types of ensemble tree techniques such as boosting and bagging. Random forest trees have been observed to give state-of-the-art performance when used to build quantitative credit risk assessment models.

Random forest for credit scoring

To build a random forest predictor, as proposed by Breiman [25], for credit-scoring, let the classifier consists of a combination/ensemble of K decision tree predictors ($h_1(x, \Theta_1)$, $h_2(x, \Theta_2)$, $h_3(x, \Theta_3)$,…, $h_K(x, \Theta_K)$), where x represents an input vector corresponding to the attributes of a credit applicant, and Θ_i, a random vector generated to govern the growth of each tree. The random vector Θ_i is a random selection (with replacement) made from the examples in the existing past applicants/clients dataset. Hence, each tree is produced from a random training set of historical clients of size s, which has been independently identically drawn from the dataset of historical clients, of size m. The training examples that are not drawn for use by each tree (the out-of-bag examples) are used to estimate the error, strength, and correlation of the forest.

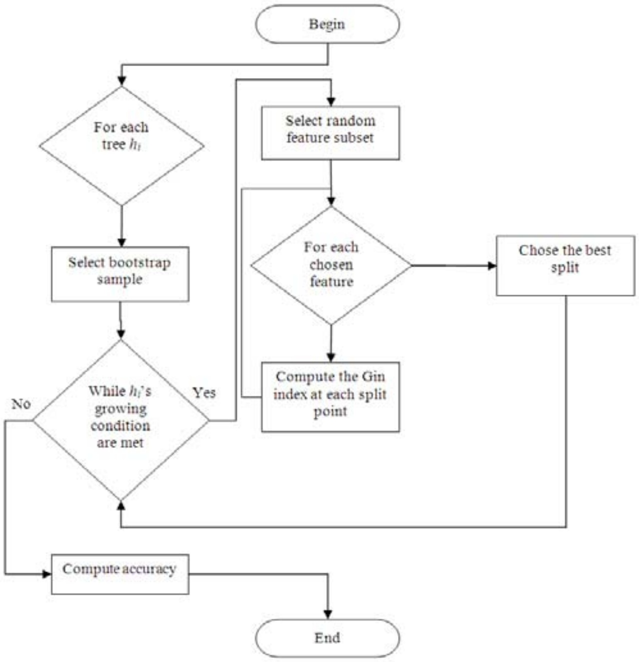


Figure 3. Random forest algorithm flow chart

Each tree is constructed such that at each node a random subset/subspace of the clients' feature space n is chosen. Splitting is only considered using the clients' features from this subset. After constructing each tree classifier, the binary class label ("good" or "bad") for a new applicant is determined by counting the votes of each tree classifier. The ultimate class label (creditworthy or un-creditworthy) is the class that has attracted the majority of votes. Figure 3 illustrates the random forest algorithm.

5 Default definition selection

Previous work has shown that credit-scoring models are sensitive to the default definition used to build them [4]. As a result to select the optimal default definition, the optimal default definition selection (ODDS) search algorithm is posited. The use of this algorithm ensures that the optimal default definition will be found for a given search space. However, this approach can be computationally expensive (time) when the search space gets large. Hence, to test the suitability of this technique while limiting the computing time required, the default definition search space was limited to three popular default definitions—30 days past due, 60 days past due, and 90 days past due.

The optimal default definition selection algorithm

A description of the ODDS algorithm is presented in Figure 4. Here, the variable, D represents the default definition used to build the final credit-scoring model. In addition, F and A are two vectors, in $\mathbb{R}^n$ where n represents the number of default definitions in the search space. The variable F is initilised to the empty set and A the set of all default definitions in the default definition search space.

```
Begin
        for $i$ = 1 to $n$,
                if $A^i \notin F$,

                        let $F = F \cup \{A^i\}$,
                        using the training dataset build a credit-scoring model using $F_i$,
                        record the model's performance using the CV dataset,
                        endif

        set $D$ to be the $F_i$ with the best performance.
        endfor
End
```

Figure 4. The optimal default definition selection (ODDS) algorithm

6 Data

A loan dataset was provided by a Barbados based financial institution. This dataset measured 18 client attributes and contained records of 47,407 instances dating from 2007 to 2010 before cleaning. The client features measured by the dataset included: applicant's marital status, the number of months the applicant has been living at the current address, the number of dependents, the ages of the up to eight of the applicant's dependents, the number of years the applicant has been employed with their current employer, the applicant's employment status, the loan amount, the loan purpose, the loan type, the applicant's monthly income, and the applicant's monthly expenditure.

This first sample is referred to as Sample$_A$. The sample's class distribution was as follows. Taking default to mean that the clients were delinquent 30 days or greater, about 92 per cent of the clients represented in this historic sample dataset were creditworthy and 8 per cent un-creditworthy, an imbalanced ratio of about 11.5:1 in favour of the creditworthy class. Taking a stricter approach and allowing default to be defined as 60 days or greater, approximately 94 percent of the dataset represented creditworthy clients and 6 per cent un-creditworthy (imbalanced ratio 17.6:1). Finally, when default is defined as clients being overdue on material obligations 90 days or greater, approximately 96 percent of the data represented creditworthy applicants while the remaining 4 per cent were un-creditworthy customers (imbalanced ratio 24.5:1).

This sample, Sample$_A$, was pre-processed by transforming all categorical data into numerical data before analysis. In addition, missing values were substituted using the variable median. Lastly, the sample dataset was normalised so as to improve the performance of the random forest tree algorithm.

7 Methodology

Empirical testing commenced with the random sorting of the processed dataset Sample$_A$. To eliminate the class imbalance problem and reduce the computational expense, Sample$_A$ was under-sampled so as to produce a sample dataset with an approx. 1:1 class ratio. This under-sampled data-file is referred to as Sample$_B$. Sample$_B$ was then split into three data-files for analysis—test (20%), cross-validation (20%), and training (60%). Following this, three copies of the training data-file were made —Train$_{30\ day}$, Train$_{60\ day}$, and Train$_{90\ day}$. For each copy of the training dataset a different definition was used to indicate default. The test and cross-validation data-files used a 90 days past due default definition. The withheld test dataset was exclusively used to test the performance of the classification models developed. This approach gives some intuition as to the performance of the models in real world settings. The cross-validation and the training datasets were used to develop the models.

Using the OCTAVE 3.2.4 programming language, the ODDS algorithm was implemented and used as a wrapper function for the random forest tree algorithm. Here, the cross-validation data-file was used to assess the performance of the training models built. This performance was logged and the training dataset that achieved the highest performance on the cross-validation data-file was used to build the ultimate classifier.

To determine the superiority of the chosen default definition, comparative models were built using the training datasets that were not chosen by the ODDS algorithm. The significance of the difference between the mean AUC scores of the models built using the optimal definition as indicated by ODDS, and the comparative models was computed using a popular statistical test—the one-way analysis of variances (ANOVA) test.

8 The ANOVA and Bonferroni tests

8.1 The ANOVA test

To investigate the significance of any differences between the AUC scores of the models the One-way analysis of variances (ANOVA) test was used. The null hypothesis of the ANOVA test was that there was no significant difference between the mean AUC scores of the three types of models, stated mathematically, H$_0$: $\mu_{30day} = \mu_{60day} = \mu_{90day}$. The alternative hypothesis stated that there was at least one group of models whose mean AUC score was significantly different (greater or less) from the other two means, H$_A$: $\exists\ i, j$: $\mu_i \neq \mu_j$. The F-statistic was computed using (4),

$$F = \frac{MS_{between}}{MS_{within}} \tag{4}$$

Here, the variable MS_{within} is used to denote the mean squares within the groups of classifiers while the variable $MS_{between}$ represents the mean squares between the groups. In order to find the value of MS_{within} the sum of squared deviation from the mean within the groups of models, SS_{within}, is calculated and divided by the degrees of freedom within the groups, df_{within}, as,

$$MS_{within} = \frac{SS_{within}}{df_{within}} \qquad (5)$$

For an individual group of models (i.e. 30 day, 60 day, or 90 day), here denoted using the variable i, the sum of squared deviations from the mean can be calculated by the equation $SS_i = \sum_{j=1}^{n_i}\left(Y_{i,j} - \bar{Y}_i\right)^2$, and degrees of freedom using the formula $df_i = n_i - 1$, where the variable n represents the number of models of type i, the variable Y represents an individual model's AUC score and $\bar{Y}$ the mean AUC score for the entire group. Accordingly, the sum of squared deviation from the mean within the groups can be calculated via the equation $SS_{within} = \sum_{i=1}^{k} SS_i$ and the degrees of freedom within the groups $df_{within} = \sum_{i=1}^{k}\left(n_{i,} - 1\right) = N - k$, where N is the total number of random forest classifiers built and k the number of groups of models (in this case 3). Similarly to find the value of $MS_{between}$ the sum of squared deviations from the mean between the groups of models, $SS_{between}$, is divided by the degrees of freedom between the groups, $df_{between}$, as $MS_{between} = \frac{SS_{between}}{df_{between}}$. Here the sum of squared deviations from the mean between the groups of models can be calculated as $SS_{between} = \sum_{i=1}^{k} n_i\left(\bar{Y}_{i,} - \bar{\bar{Y}}\right)^2$, where $\bar{\bar{Y}}$ represents the grand mean and $df_{between} = k - 1$.

The F-statistic shows the ratio of the variance calculated among the means to the variance within the samples. Therefore, if the group means are equal, this value should be lower than the critical value (referenced from the F distribution table); however, if the alternative hypothesis is true this value should be larger.

8.2 The Bonferroni post hoc test

In addition, to determine which groups of models were performing significantly different in terms of AUC, a multiple comparisons post hoc test, the Bonferroni method, was conducted as equal variances were assumed based on Levene's test for the homogeneity of variances. In its formal form the Bonferroni inequality is represented as shown in,

$$P\left(\bigcap_{i=1}^{g} A_i\right) \geq 1 - \sum_{i=1}^{g} P[\bar{A}_i] \qquad (6)$$

In (6) the variable g denotes the number of comparisons, and the variables $A_i$and $\bar{A}_i$ represents the occurrence of complementary events. The event A_i occurs when the calculated confidence interval of a linear combination of treatments includes the true value of that combination. As a result, the right hand side of the inequality represents one minus the sum of the probabilities that the calculated confidence intervals of the linear combination of treatments do not include their true values.

Since simultaneous multiple interval estimates have a confidence of 1 – α (where α = 0.05), each interval estimate has confidence of (1 – α/g). Hence, this method can be used to ensure that the group-wise comparison achieve an overall confidence coefficient of at least 1 – α.

9 Results and analysis

Quantitative credit-scoring is a notoriously difficult task as credit data is very often not easily separable. This is in-part due to the nature of the credit assessment exercise. To begin with, in many cases there is an asynchrony of information between the applicant (who may guard this information) and the firm, as loan applicants often have more knowledge of their own creditworthiness than credit providers. The financial institution is therefore charged with the task of gathering this information about the applicant. This is often done by having the applicant complete standardised application forms and taking statements concerning the applicant from third-parties (e.g. banks, suppliers, employers, credit-bureaus, etc). Nevertheless, despite the best efforts of the firm it is almost impossible to record every aspect of a client's life that may result in default. As a result, credit scorecards often produce higher misclassification rates than would normally be acceptable for other classification problems [26]. The results presented in this section should be interpreted with this in mind.

9.1 The optimal default definition

The ODDS algorithm selected the 60 days past due definition as optimal. Accordingly, the 60 day training dataset was used to build the primary classifiers while the 30 day past due and the 90 days past due training datasets were used to build the comparative models. In total thirty-six credit-scoring models were built. These models were divided equally across the three default definitions used; as a result, the thirty-six models comprised of twelve 30 day models, twelve 60 day models, and twelve 90 day models. Using the withheld test dataset means and standard deviations values for each performance metric were calculated for the three groups of models. These performances are shown in Table 1 below.

Table 1. Showing mean and standard deviation performances of the models using the test dataset

Test dataset	Models		Test accuracy	Precision	Recall	F-score	BAC	AUC	Sensitivity	Specificity
$\text{Test}_{90\ day}$	30 day models	mean	65.44	65.11	66.51	65.80	65.44	65.38	66.51	64.36
		s.d.	0.48	0.47	0.72	0.52	0.48	0.45	0.72	0.64
	60 day models	mean	66.28	66.17	66.63	66.40	66.28	66.27	66.63	65.69
		s.d.	0.52	0.52	0.94	0.60	0.52	0.54	0.94	1.08
	90 day models	mean	65.57	65.79	64.90	65.34	65.57	65.53	64.90	66.25
		s.d.	0.57	0.71	0.68	0.52	0.57	0.62	0.68	1.08

The results presented in Table 1 reveal that on average the models built were reasonably predictive, as indicated by AUC, of creditworthiness using the withheld test dataset. Here, the 30 day models reported a mean AUC score of 65.38 per cent, the 60 day models reported a mean AUC score of 66.27 per cent, and the 90 day models reported a mean AUC score of 65.53 percent.

9.2 Significances of AUC scores on test dataset

Highlighted in Table 2 are summary descriptive statistics and ANOVA analysis for the models on the test dataset. The results show that there was a significant difference between two or more mean AUC scores for the three groups of classifiers (F = 9.22 > F critical-value = 3.28, $p < .05$).

Table 2. Showing summary statistics and the ANOVA computed for the 3 groups of models

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
30 day models	12	784.56	65.38	0.21		
60 day models	12	795.25	66.27	0.29		
90 day models	12	786.40	65.53	0.39		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F critical*
Between Groups	5.44	2	2.72	9.22	6.59E-04	3.28
Within Groups	9.74	33	0.30			
Total	15.19	35				

The Bonferroni test was computed to determine which models were performing significantly different from each other (Levene's statistic = 0.695; p = .506). Table 3 illustrates the results of this testing and shows that there was a significant difference between the mean AUC scores of the 60 day models and both the 30 day and 90 day models, on the test dataset. An inspection of the mean AUC scores (see Tables 1 and 2) indicated that on average the "broad" 60 day models would outperform both the "broad" 30 day models and the "regular" 90 day models when predicting the creditworthiness of individuals ($\mu_{30day} = 65.38$; $\mu_{60day} = 66.27$; $\mu_{90day} = 65.53$). This finding seems to suggest that models developed using a “broader” 60 day past due definition of default were statistically better in terms of AUC score. There was no significant difference between the 30 day and the 90 day models as measured by AUC.

Table 3. Showing comparisons of the models using Bonferroni’s method

(I) Group	(J) Group	Mean Difference (I-J)	Sig.	95% Confidence Interval	
				Lower Bound	Upper Bound
30 day models	60 day models	-0.89083*	0.001	-1.4503	-0.3313
	90 day models	-0.15333	1.000	-0.7128	0.4062
60 day models	30 day models	0.89083*	0.001	0.3313	1.4503
	90 day models	0.73750*	0.007	0.1780	1.2970
90 day models	30 day models	0.15333	1.000	-0.4062	0.7128
	60 day models	-0.73750*	0.007	-1.2970	-0.1780

The * is used to indicate that the difference between the groups of models is significant at the 0.05 level.

10 Conclusion

This paper explored the use of random forest trees when developing “broad” credit-scoring models for a Barbados based financial institution. Random forest is an ensemble learning technique that combines the predictions of several “weak” decision trees in order to form a “strong” classifier. The use of the random forest algorithm and other similar artificial intelligent and machine learning algorithms at financial institutions is becoming increasingly popular but is currently an under-utilised practice in Barbados. The results of this paper suggest that the historic clients’ database of the study institution could be data-mined to unearth patterns which indicate client creditworthiness. These models of client behaviour could be used to assist in the credit evaluation process of the entity.

This work also studied whether models built using more relaxed default definitions (i.e. “broad” models) can have an impact on classifier performance. The findings of this study revealed that models built using a “broad” 60 day past due definition of default were statistically superior to 30 day and 90 day models when used to predict whether a withheld test example would experience 90 days delinquency. The value of any improvement in classification can be enormous to a financial institution when the net-present-value of the misclassification cost savings over the life of the firm are calculated. Thus in practice a mere one per cent improvement in classifier performance can be worth millions of dollars. Thus, the use of the ODDS algorithm to select the best default definition from a given search space represents an improvement to the art.

The future work of this author will seek to improve the ODDS algorithm and will consider other strategies for selecting the optimal default definition. Furthermore, future studies will consider the impact of class distribution shifts on broad models. In addition, incorporating economic time series data into these models will also be explored.

Appendix

Credit scoring metrics

In this paper the Area under the Receiver Operating Characteristic (ROC) curve (AUC) is used as the primary performance metric for analysis. This metric is one of the most promising and widely used during classifier development, as it makes use of the ROC curve. This curve is a two dimensional measure of classification performance where the sensitivity (1), which is the proportion of actual positives predicted as positive, and the specificity (2), this is the proportion of actual negatives that are predicted as negative, are plotted on the Y and X axis, respectively. Equation (3) presents the AUC metric and measures the area under the resulting curve. Here, the variable S_1 represents the sum of the ranks of the creditworthy clients. As a result, a score of 100% means that the classifier is able to perfectly discriminate between the classes; while a score of 50% means that the classifier is poor with insignificant discriminatory power.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{1}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{False Positive} + \text{True Negative}} \tag{2}$$

$$\text{AUC} = \frac{(s_1 - \text{Sensitivity}) * [(\text{Sensitivity} + 1) * 0.5]}{\text{Sensitivity} * \text{Specificity}} \tag{3}$$

In addition to the AUC, a number of other metrics are used to report on the performance of credit-scoring models presented in this paper. These metrics are introduced in the following paragraphs.

Test accuracy, as in (4) is a popular metric with which the performance of credit-scoring models can be assessed. This metric is the measure of how accurately the model classifies credit applicants on a withheld test dataset. When interpreting this metric one should always be cautious as it can lead to performance assessment issues when the dataset is skewed, and skewed data is a common occurrence with real world credit-scoring datasets.

$$\text{Test accuracy} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} + \frac{\text{True Negative}}{\text{False Negative} + \text{True Negative}} \tag{4}$$

Arguably a more useful measure is the balanced accuracy (BAC) as in (5). This measure avoids the misleading affects on accuracy caused by imbalanced datasets by showing the arithmetic mean of sensitivity and specificity.

$$\text{BAC} = \frac{\text{Specificity} + \text{Sensitivity}}{2} \tag{5}$$

Two other metrics which can be used to give an indication of the quality of a classifier when the dataset is skewed are precision, as in (6), and recall, as in (7). Precision measures of how accurately the positive predictions have been classified, while recall assesses how accurately actually positives were predicted as positive (i.e. this is the same as sensitivity).

$$\text{Precision} = \frac{\text{True Positive}}{\text{True positive} + \text{False Positive}} \tag{6}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{7}$$

The F_1 score, as in (8), metric will also be used to report the performance of the models discussed in this paper. This metric can be viewed as a type of average for precision and recall.

$$F_1\,Score = 2\frac{Recall * Precision}{Recall + Precision} \tag{8}$$

References

[1] L. Thomas, R. Oliver, and D. Hand, "A survey of the issues in consumer credit modelling research," Journal of the Operational Research Society. 2005; 56: 1006-1015. http://dx.doi.org/10.1057/palgrave.jors.2602018

[2] J. N. Crook, D. B. Edelman, and L. C. Thomas, "Recent developments in consumer credit risk assessment," European journal of operational research. 2007; 183: 1447-1465. http://dx.doi.org/10.1016/j.ejor.2006.09.100

[3] C. Basel, "International Convergence of Capital Measurement and Capital Standards: A Revised Framework, Comprehensive Version," Switzerland: Bank for International Settlements, 2006.

[4] T. Harris, "Quantitative Credit Risk Assessment using Support Vector Machines: Broad Versus Narrow Default Definitions," Expert Systems with Applications. 2013; 40: 4404-4413. http://dx.doi.org/10.1016/j.eswa.2013.01.044

[5] P. Jorion and G. Zhang, "Good and bad credit contagion: Evidence from credit default swaps," Journal of Financial Economics. 2007; 84: 860-883. http://dx.doi.org/10.1016/j.jfineco.2006.06.001

[6] S. G. Ryan, "Accounting in and for the Subprime Crisis," The accounting review. 2008; 83: 1605-1638. http://dx.doi.org/10.2308/accr.2008.83.6.1605

[7] M. K. Brunnermeier, "Deciphering the liquidity and credit crunch 2007-08," National Bureau of Economic Research. 2008. http://dx.doi.org/10.3386/w14612

[8] R. Ormerod and W. Ulrich, "Operational research and ethics: a literature review," European journal of operational research. 2013; 228: 291-307. http://dx.doi.org/10.1016/j.ejor.2012.11.048

[9] CBB, "Financial Stability Report," The Central Bank of Barbados, Bridgetown, Barbados 2011.

[10] CBB, "Financial Stability Update August 2012". The Central Bank of Barbados, Bridgetown, Barbados 2012.

[11] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Human Genetics. 1936; 7: 179-188. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

[12] D. Durand, Risk Elements in Consumer Instalment Financing. NY: National Bureau of Economic Research, 1941.

[13] E. I. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," The Journal of Finance.1986; 23: 589-609, 1986. http://dx.doi.org/10.1111/j.1540-6261.1968.tb00843.x

[14] Y. E. Orgler, "A credit scoring model for commercial loans," Journal of Money, Credit and Banking. 1970; 2: 435-445. http://dx.doi.org/10.2307/1991095

[15] T. Bellotti and J. Crook, "Support vector machines for credit scoring and discovery of significant features," Expert Systems with Applications. 2009; 36: 3302-3308. http://dx.doi.org/10.1016/j.eswa.2008.01.005

[16] Z. Huang, H. Chen, C. J. Hsu, W. H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: a market comparative study," Decision support systems. 2004; 37: 543-558. http://dx.doi.org/10.1016/S0167-9236(03)00086-1

[17] C.-S. Ong, J.-J. Huang, and G.-H. Tzeng, "Building credit scoring models using genetic programming," Expert Systems with Applications. 2005; 29: 41-47. http://dx.doi.org/10.1016/j.eswa.2005.01.003

[18] Y. Wang, S. Wang, and K. Lai, "A new fuzzy support vector machine to evaluate credit risk," Fuzzy Systems, IEEE Transactions. 2005; 13: 820-831. http://dx.doi.org/10.1109/TFUZZ.2005.859320

[19] D. West, "Neural network credit scoring models," Computers & Operations Research. 2000; 27: 1131-1152. http://dx.doi.org/10.1016/S0305-0548(99)00149-5

[20] D. West, S. Dellana, and J. Qian, "Neural network ensemble strategies for financial decision applications," Computers & Operations Research. 2005; 32: 2543-2559. http://dx.doi.org/10.1016/j.cor.2004.03.017

[21] E. Rosenberg and A. Gleit, "Quantitative methods in credit management: a survey," Operations research. 1994; 42: 589-613. http://dx.doi.org/10.1287/opre.42.4.589

[22] L. Yu, Bio-inspired credit risk analysis: computational intelligence with support vector machines: Springer, 2008. http://dx.doi.org/10.1007/978-3-540-77803-5

[23] H. Frydman, E. I. Altman, and D. L. Kao, "Introducing recursive partitioning for financial classification: the case of financial distress," The journal of finance. 1985; 40: 269-291. http://dx.doi.org/10.1111/j.1540-6261.1985.tb04949.x

[24] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and regression trees. Belmont, CA: Wadworth International Group, 1984.

[25] L. Breiman, "Random forests," Machine Learning. 2001; 45: 5-32. http://dx.doi.org/10.1023/A:1010933404324

[26] B. Baesens, T. van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," Journal of the Operational Research Society. 2003; 54: 1082-1088. http://dx.doi.org/10.1057/palgrave.jors.2601545