Insurers' Insolvency Prediction using Random Forest Classification

Anastasia V. Kartasheva*

Mikhail Traskin†

March 15, 2011

*Department of Insurance and Risk Management, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302

†Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340

Insurers' Insolvency Prediction using Random Forest Classification

Abstract

This paper uses a modification of the Random Forest classification algorithm to predict insolvency of insurers. RF orders companies according to their propensity to default. We show that RF methodology delivers higher quality of prediction compared to other existing methods. In addition, RF classification can be used to gather further insights about the fragile companies. It ranks the explanatory variables in the order of their ability to predict insolvency. Also it is used to describe the relationship between the propensity to default and the individual characteristics of an insurer. We show that many of these relationships are highly non-linear.

Keywords: property-casualty insurance, insolvency prediction, Random Forest Classification.

JEL Codes: C14, C44, G17, G22, G32

The financial strength of an insurer is one of the key qualities of an insurance policy. The complexity of insurers' operations and asymmetric information can restrict the ability of insurance buyers to evaluate insurer's insolvency risk. As a consequence, insurance solvency regulation is common in every developed economy. However, the efficiency of the regulation depends on the ability to identify vulnerable insurance companies and on understanding the drivers of insolvencies. This paper contributes to the literature on insurer insolvency prediction. It describes a new method to predict insolvency of insurers based on the Random Forest (RF) classification algorithm.

Random Forest is a non-parametric machine learning algorithm that orders companies according to their propensity to default. We compare the performance of RF to logistic regression commonly used for insurers' insolvency prediction. The results show that RF methodology delivers higher quality of prediction for a wide range of misclassification costs. In addition, RF classification can be used to gather further insights about the fragile companies. It ranks the explanatory variables in order of their ability to predict insolvency and provides a practical guidance for insolvency monitoring to regulators and market participants. Also it can be used to describe how the propensity to default depends on the individual characteristics of an insurance company such as capital, investment strategy and others that are commonly used by regulators. We show that many of these relationships are highly non-linear.

The advantage of the Random Forest algorithm is its ability to handle effectively several specific features of the insurers insolvency prediction.

First, insolvencies are rare events which result in a highly unbalanced dataset and which reduce the set of parameters that can be estimated efficiently. As Table 1 shows, only about 1% of companies fail each year. Therefore, a passive prediction that all companies are solvent leads to 99% accuracy. Clearly this solution cannot be used for solvency monitoring.

The second challenge of insolvency prediction is the temporal aspect of company's performance. Typically a failure in a given year is preceded by poor management decisions

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
NC	1903	1897	1968	2012	2061	2065	2084	2100	2096	2096	2042	1952
NF	25	23	33	21	22	7	8	25	6	10	18	16
% failed	1.31	1.21	1.68	1.04	1.07	0.34	0.38	1.19	0.29	0.48	0.88	0.82

Table 1: Number of property/causality insurance companies and number of bankruptcies among them in 1989–2000. NC—number of companies, NF—number of failures.

or/and asset returns. Including past performance can substantially enhance the quality of prediction but it also increases the number of parameters to be estimated. Combined with a low number of insolvencies, estimation by logistic regression becomes numerically unstable.

The third challenge lies in the importance of interactions among different company characteristics. The amount of the economic capital can substantially vary for companies with different degrees of geographic and line of business diversification. Similarly, companies entering new business lines are likely to need more capital than those operating in the familiar business areas. Handling interactions effectively is one of the key features of RF classification.

The Random Forests algorithm was developed by Breiman (2001). A number of other works preceded the Random Forests algorithm in using randomization to grow a forest (Dietterich, 2000; Breiman, 2000; Ho, 1998; Amit and Geman, 1997). The algorithm has been successfully applied in many areas. Some appealing examples include predicting customer retention and profitability (Larivière and Van den Poel, 2005), forecasting murder within a population of probationers and parolees (Berk et al., 2009), protein-protein interaction prediction (Qi et al., 2005) and classification of microarray data (Díaz-Uriarte and Alvarez de Andrés, 2006). Khandani et al. (2010) employ the machine learning algorithms to construct consumer credit risk models.

In our experiments we used `R` (R Development Core Team, 2009), a free software environment for statistical computing and graphics. We used package `randomForest` (Liaw and Wiener, 2002), a readily available implementation of the Random Forest algorithm which can be downloaded from the R-project website.[1] This implementation is based in the code

[1] `http://cran.cnr.berkeley.edu/web/packages/randomForest/index.html`. Algorithm is also avail-

written by Leo Breiman and Adele Cutler.

The rest of the paper is organized as follows. The next section explains the main features of the Random Forrest algorithm. Section 2 reviews the models used to predict insurers' insolvencies and relates RF methodology to these studies. Sections 3 describes the data. Section 4 presents the prediction accuracy results for the logistic regression. Our main results, RF estimation and comparison of RF prediction accuracy with other models, are presented and discussed in Section 5. Section 6 applies RF classification to identify the main variables that predict insolvency and to study the individual effect of predictors. Section 7 concludes. The technical details are explained in the appendices. Figures and tables that do not appear in the paper are in Appendix D.

1 The Random Forest Algorithm

In this section we present a simple example that describes the essential features of Random Forest algorithm. Pseudocode for the algorithm is given in Appendix A.

Consider a sample of 10 companies that contains two subsamples of solvent and insolvent companies. Each company has two characteristics, x and y, with $x, y \in [0,1]$. Thus all companies are located on a square $U = [0,1] \times [0,1]$. The initial sample is illustrated on Figure 1(a), where solvent companies are depicted as points and insolvent ones as crosses. The classification objective is to divide the square into two subsets S and N, $S \cup N = U$ such that a company i is classified as solvent if $(x_i, y_i) \in S$ and insolvent if $(x_i, y_i) \in N$.

RF classification consists of the following basic steps.

1. Select a random subsample D from the initial sample of companies. Suppose that the outcome of the random draw is 5 companies illustrated as circled points on Figure 1(b).

2. On subsample D grow a classification tree. This involves the following steps.

able in some commercial software packages and in standalone commercial implementations.

(a) Consider a rectangle containing a subset of points in D. At the first step, this rectangle coincides with the unit square (the support of the distribution).

(b) Randomly choose between characteristics x or y to split the rectangle. Suppose the outcome is x.

(c) On the x axis choose a split point x^* to maximize the homogeneity of observations in rectangles to the left and to the right of x^*. Divide D according to the criterion $x > x^*$ and $x < x^*$. On Figure 1(c), the left rectangles contains two solvent companies, while the right rectangle contains three insolvent companies. Thus the left rectangle is classified as solvent class and the right rectangle is classified as insolvent class. In general, a rectangle is classified as an insolvent class when it contains a majority of insolvent companies.

(d) Repeat the steps 2.a–2.c for each of the two resulting rectangles until each subsample contains exactly one point or the rectangle is homogeneous. On Figure 1(c), the rectangles to the right and to the left of x^* are homogeneous after the first iteration. In RF terminology, a rectangle that contains only one point or is homogeneous is referred to as a leaf of a tree. The partition of U in leafs constitutes a tree. Figure 1(d) is an illustration of a partition determined by this procedure for a subsample D. On Figure 1(d), the tree classifies all companies that belong to the clear area as solvent and those that belong to the shaded area as insolvent.

3. Repeat steps 1 and 2 independently many times. Figures 1(e) and 1(f) illustrate another two examples of partitions of the unit square, or trees. Random forest consists of a large number of trees grown using these steps.

4. Classification of an out-of-sample company A in is done by majority vote of individual trees. For each tree, A belongs to a leaf that is classified as either solvent or insolvent. A company A is classified as solvent (insolvent) if the majority of trees classify it as solvent (insolvent). In the example on Figures 1(d)-(f), a company A denoted by a

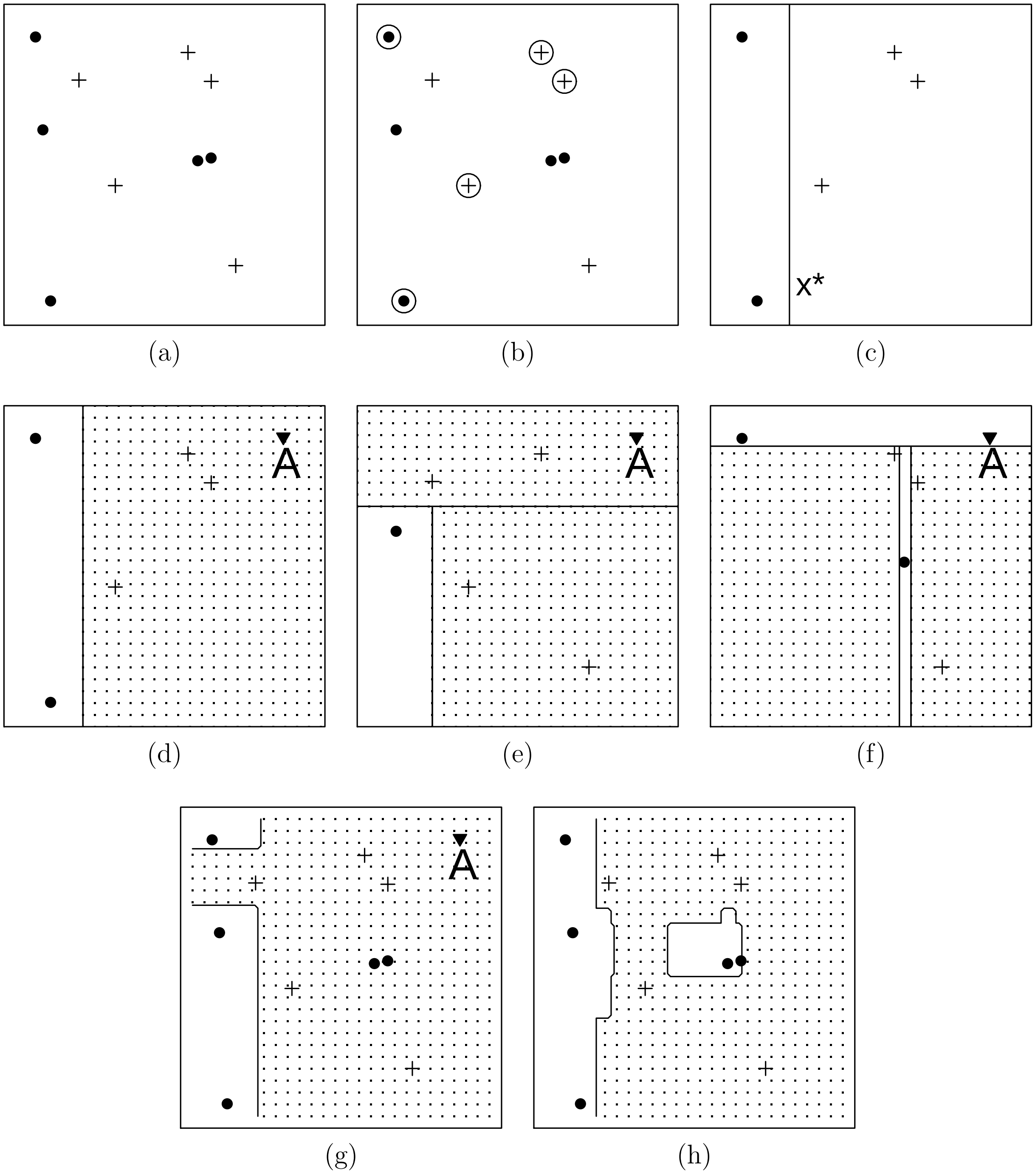
black triangle is classified as insolvent by trees 1(d) and 1(e), and as solvent by tree 1(f). Thus the majority voting of these three trees prescribes that company A is insolvent. Figure 1(g) depicts the classification boundaries generated by the forest consisting of three trees in Figures 1(d)–(f). In general, as the number of classification trees increases, classification boundaries become more refined. For example, Figure 1(h) gives an example of classification boundary produced by the forest consisting of seven trees.

RF has several characteristics that make it a suitable method for insolvency prediction. RF adapts to non-linear decision boundaries between solvent and insolvent classes. For example, consider the decision boundaries illustrated on Figures 2(a,b). In neither of these cases a linear decision boundary produced by classical methods like logistic regression, is suitable. However, RF will approximate these boundaries effectively.

Growing multiple trees on random subsets and then averaging over them reduces a sample-to-sample variation in the prediction (Breiman, 1996). Also the randomized splitting procedure allows us to account better for the local properties of the data. Altogether, RF leads to a highly flexible procedure that does not put any restrictions on the predictor variables (it equally well works with continuous and categorical variables), is insensitive to monotonic transformations of the variables, and does not make any restrictive model assumptions on the relationship between the insolvency event and the predictor variables. We discuss further advantages of RF in the next section.

2 Comparison of random forest algorithm to other methods of insolvency prediction

This section reviews the common methods of insurers' insolvency prediction. Also it discusses different approaches to deal with unbalanced datasets and large number of parameters, and compares them to Random Forest.



(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 1: Example illustrating growth of a single tree in the Random Forest and final classification by the forest. Round dots have labels −1, star dots have labels +1 and shaded regions in (d)-(f) should be classified as +1 with the unshaded regions as −1.

2.1 Insolvency prediction in insurance

Regulatory solvency prediction methods. The National Association of Insurers Commissioners (NAIC) has developed property-liability Risk-Based Capital System (RBC) to calculate the

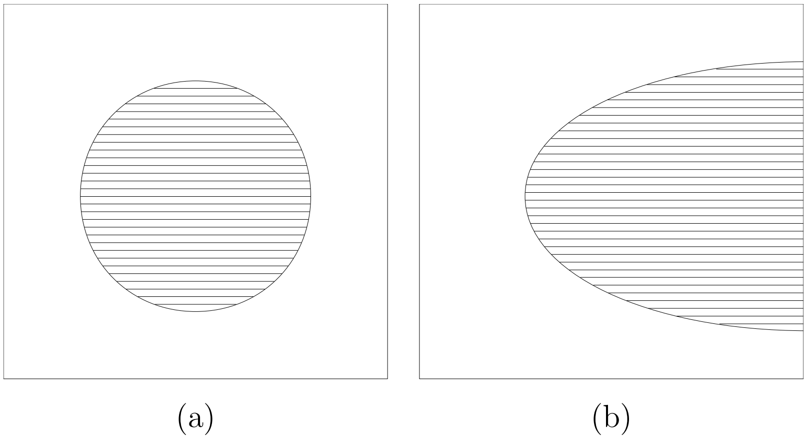


Figure 2: Two examples of non-linear decision boundaries, that are hard for linear classification methods, but easy for Random Forest algorithm.

amount of risk-based capital that an insurer needs to hold. RBC system computes capital charges for four major types of risk—asset risk, credit risk, underwriting risk and growth and off-balance sheet risk. The sum of the charges, less the covariance adjustment, results in the regulatory risk-based capital. Then the actual capital that the insurer has is divided by its risk-based capital. Companies with the RBC ratio below certain threshold, called "Authorized Control Level", are subject to different degrees of regulatory intervention.

The FAST (Financial Analysis and Surveillance Tracking) system is used to prioritize insurers for further regulatory action. It consists of twenty-nine audit ratios and corresponding scores for each ratio. The resulting overall FAST score is a weighted sum of audit ratios. Companies are ranked according to FAST score, and the extent of regulatory intervention is determined according to the score. The ability of the FAST system to predict insolvencies has been questioned by Grace et al. (1995). They tested FAST against a logistic regression model based on FAST ratios and other firm characteristic variables and concluded that FAST did not lead to better predictions. Cummins et al. (1999) note that Grace et al. (1995) results are subject to potential bias because the scores and audit ratios used by NAIC were modified to maximize their predictive power and thus may have even lower ex-ante ability to predict insolvency.

Ratings and rating changes. Major credit rating agencies, A.M. Best, Fitch, Moody's and Standard and Poor's "provide an opinion about the insurer's ability to meet its contractual obligations to policyholders" (quote from A.M. Best). Compared to a large number of potential financial ratios, and their changes over time, ratings and rating changes are simple summary measures that may be preferred by consumers and regulators. Pottier (1998) analyzes a sample of forty-eight insolvent life insurers in 1990-1992 rated by A.M. Best. The comparison among the predictive ability of (i) ratings, rating changes and total assets, (ii) financial ratios, and (iii) financial ratios combined with ratings and rating changes reveals that (i) and (iii) have comparable predictive ability that is higher than (ii). Ambrose and Carrol (1994) consider a sample of life insurers and show that financial ratios, combined with an indicator that an insurer is rated A or higher, are better insolvency predictors than either ratios or indicators alone. These finding suggest that A.M. Best ratings and financial ratios contain complementary information about insurers insolvency. At the same time, the number of companies that major rating agencies classify as vulnerable, is large and thus is not practical for identifying those that need regulatory attention. Furthermore, several studies have shown that the rating agencies delay the information release, so relying on ratings alone can increase the cost of insolvencies.

Multiple discriminant analysis (MDA). A large number of studies have used MDA since Altman (1968). MDA assumes that groups of solvent and insolvent insurers are drawn from a normal distribution with different means but a common variance-covariance matrix. Then each observation is assigned a score that is developed to maximize the ratio of between-group to within-group variance. An observation is assigned to the solvent or the insolvent group according to the score. Trieschmann and Pinches (1973); Pinches and Trieschmann (1974, 1977) analyze distressed property-liability insurers using MDA approach. These studies for the most part ignore that MDA relies on strong distributional assumptions and that the actual number of insolvent insurance is significantly lower than the solvent ones.

Logit, Probit and Hazard models. Logit models estimate the probability of insurer insolvency where the dependent variable is a binary variable of one when the insurer is insolvent and zero otherwise. The hazard model estimates a hazard function that is a probability of solvency in year t given that a firm has survived till t. Lee and Urrutia (1996) compare the performance of Logit and Hazard models in predicting insolvency and selecting variables that have a statistically significant impact on insolvencies. They find that in logit model four variables are significant at 5% level of better: ratio of net premiums written to surplus, return to policyholders' surplus, proportion of premiums written in long-tailed lines, and market value of invested bonds as a proportion of total admitted assets. The hazard model has four additional statistically significant variables; operating margin, current liquidity ratio, growth rate of statutory surplus and growth rate of the net premium written. Tests to forecast accuracy indicate that the logit model has a somewhat lower misclassification rate than the hazard model, but the logit model has higher misclassification costs for Type I errors. Lee and Urritia conclude that the two models have comparable forecasting accuracy.

Cash flow simulation. Cummins et al. (1999) develop a cash flow simulation model that uses pro-forma cash flow statements of property-liability insurers and simulates the cash inflows and outflows implied by the insurer's initial position over a twenty-year horizon. A company is classified to be solvent if it has positive net resources in twenty years. The model relies on several (non-stochastic) scenarios and contains separate modules, including the premium and underwriting expense module, the loss modules and the investment module. The accuracy of the simulation model is compared to the regulatory RBC and FAST systems using the logit regression. The independent variables are drawn from the three solvency prediction models. The authors confirm earlier findings that RBC and FAST systems have low predictive power, though FAST performs better than RBC. They show that cash flow simulation variables add significant explanatory power to RBC and FAST. The model also estimates the predicted time to failure that is not available with RBC and FAST systems.

2.2 Advantages of the Random Forest

The Random Forest algorithm has a number of advantages over logistic regression and other classification methods when it comes to predicting insurer insolvencies.

First, it is fairly insensitive to the values of tuning parameters such as a number of variables used to split a leaf and a subsample size (Breiman, 2001). Even with the default settings results are usually good and further tuning is often not necessary. This is in contrast to some other machine learning methods, such as Support Vector Machines, which require accurate choice of tuning parameters in order to obtain good results.

Second, as a byproduct Random Forest produces a measure of variable importance that is useful for variable selection. Unlike in linear or logistic regression, Random Forest does not estimate a coefficient associated with an individual variable. However, it orders variables according to the ability to obtain accurate prediction. In many cases, variables at the top of this list are sufficient to obtain classification which is nearly as accurate as classification based on all the variables. Moreover, these are often the variables that work well in other models. The ranking of variables importance is perfomed by evaluating the decrease of prediction accuracy when one of the variables is eliminated from the classification procedure. A high drop in accuracy due to elimination of a particular variable indicates that the variable contains valuable information about propensity to default. We present the results of variable importance analysis in Section 6.

Third, RF suits well for the analysis of large unbalanced data sets. All classification methods have difficulties with datasets where one of the classes dominates the other and tend to classify everything according to the majority class. One of the standard method of dealing with such data is to use subsampling of the majority class. In particular, the prior literature (Lee and Urrutia, 1996; Cummins et al., 1999) has addressed the issue of the small number of insolvencies by using subsampling of the overrepresented class. Subsampling builds a new sample where the number of insolvent and solvent companies is either the same (Lee and Urrutia, 1996), or at least is more balanced (Cummins et al., 1999). The main advantage of subsampling, aside from simplifying the problem considerably by reducing total sample size, is that it improves accuracy of quantile estimation. Usual estimates of quantiles that are close either to 0 or 1 are not very accurate, while it was shown in (Mease et al., 2007) that subsampling improves the accuracy of quantile estimation. The obvious deficiency of subsampling approach is that we lose power due to decreased sample size. At the same time subsampling is an integral part of the Random Forest algorithm. Individual trees are grown on *different* subsamples, and hence Random Forest looks at *all* the data.

Accuracy of quantile estimation can also be improved through introduction of weights. However, these methods may not be suitable for large data sets. In our study we use four years of data to predict insolvency in the next year. Weighted logistic regression becomes numerically unstable when used with such a large data set due to the large number of variables relative to the sample size. The appropriate regularization (shrinkage) performed by ridge regression (Hoerl and Kennard, 1970; Hastie et al., 2001), lasso regression (Tibshirani, 1996; Hastie et al., 2001) or elastic net (Zou and Hastie, 2005) could be used to remedy the insufficient sample size relative to the number of variables in the model. The latter two approaches also allow doing variable selection. However, these methods are not sufficient to estimate the non-linear boundary between solvent and insolvent insurers. All the methods based on the logistic regression assume a linear decision boundary between solvent and insolvent companies. This restriction can be compensated to a certain degree by including interactions into the model. However, introducing only first order interactions in a model with 100 variables increases the number of parameters to be estimated from 101 to 5051 (we are not mentioning higher order interactions here!). Even shrinkage methods, such as lasso and elastic net, face computational difficulties in this setting. This leads us to the fourth advantage of the Random Forest: it produces highly non-linear decision boundary and is capable of dealing with large number of predictor variables. This is because when splitting a leaf Random Forest looks only at a small subset of all available variables to choose the best one to split on, it can grow trees rapidly even if number of predictors is large.

In general, Random Forest often outperforms other classifiers (Breiman, 2001; Berk, 2006).

3 Data

We analyze the sample of property-casualty insurers during the period between 1993 and 2000 which represents a period of relative stability in the insurance sector. We employ two sets of data: one including the characteristics of insurance companies; the other including the information on insolvencies during the period. Information on insurers' financial strength, business strategy and organizational characteristics comes from the annual regulatory statements filed with the National Association of Insurance Commissioners (NAIC). The list of insolvent insurers was compiled from the various years of A.M. Best publications "A.M. Best Insolvency Review". Our definition of insolvency is common in the literature. An insurer is considered insolvent if the domiciliary state insurance commissioner declared the insurer insolvent or placed it in receivership, conservatorship, or liquidation.[2]

We consider a wide set of predictors. It includes 25 FAST financial ratios that summarize the individual characteristics of insurers; the information about the product mix along two dimensions, commercial/personal and short/long tail lines; insurer's organizational form that distinguishes between mutual and stock companies; the information about the investment strategy and liquidity of assets; the characteristics of the competitive environment. As a result, each company has 37 characteristics in each year of the panel. Table 6 contains the complete variables list. Table 7 provides the summary statistics for the solvent and insolvent insurers. It indicates that the two groups are statistically different in terms of size, leverage, profitability, liquidity, etc.

We focus on the prediction that an insurance company becomes insolvent within one year and use the current year's data and the data for the past three years to predict insolvency in

[2] This definition excludes insurers that merged into another company, were completely reinsured or liquidated voluntarily. The exclusion of these companies from the analysis is because the data is not available.

the subsequent year. Thus we recognize that the financial strength can deteriorate over time, even though the insurer may not declare insolvency for several years. Predicting insolvency within a longer time period, say two or three years, can be done with minor modification of the data preprocessing step.

The following data preprocessing procedure is used to obtain the working data set from the NAIC data. In the original data set we have 37 variables and an indicator if a company is insolvent in a particular year t. Then we construct a new vector of predictors that corresponds to a vector of predictors for the same company for years $t - i$, $i = 0, 1, 2, 3$, and an indicator if a company is insolvent in year t. We account for missing values by replacing them with mean value and creating an extra variable $\tilde{x}_i^m \in \{0, 1\}$ which is set to 1 if value of $\tilde{x}_i$ is missing and to 0 otherwise.

After the above transformations, the resulting data set contains 200 predictors and 13648 records. Total number of defaults is 83, or 0.608% of all the records. The 83 insolvent companies belong to 70 unique insurance groups. Thus our results are not driven by the fact that insolvent companies are subsidiaries of a limited number of national insurance groups that become insolvent. The focus of our analysis is on individual characteristics of an insurer that affect its probability of default.

In the constructed dataset, every company has several records that correspond to different years. We treat these records as independent samples from the underlying distribution. This brings up the problem of censoring since an insolvent company disappears from the data set. However, the assumption that, given the four years of data, the event of insolvency is independent of other events, alleviates this problem.

Ideally, the evaluation of performance of any classification algorithm would be the validation on an independent data set. Small number of insolvencies makes this approach unreliable. Hence we resort to the second-best option: a cross validation. This does not allow us to evaluate the performance given the training data, but it is sufficient to evaluate the *comparative* performance of various insolvency prediction methods. Therefore, our results should be interpreted as comparisons of methods, and not the actual classifiers generated by methods from the data.

4 Prediction accuracy

We use logistic regression as a base line for accuracy comparison. Logistic regression finds a linear boundary between two classes. It usually works poorly with highly unbalanced datasets similar to the one we have. On the other hand, in high dimensional spaces like the one we consider, a linear decision boundary is often a good choice. For our data set, fitting a logistic regression model that uses all the available predictors (no interactions) results in the confusion table shown in Table 2.[3] These are very good results for a problem where training dataset contains so few insolvent cases. In fact, usually one should expect to see at most 1 or 2 predictions of insolvency on the training data set. Here we have 39, and 10 of these are correct. Therefore one may conclude that logistic regression model is sufficient for this problem and that even simpler model may be enough. For comparison, we report confusion table for the model of Lee and Urrutia (1996)[4] in Table 3. These results are more like what we expected *a priory*, but they hint at usefulness of having more predictors. Finally, Table 4 reports results of fitting a logistic regression model to part of the data and then computing a confusion table on the remaining part. There we see that model is making too many errors (error rate is .775), though majority of these errors are false positives. If we take

	predict solvency	predict insolvency	model error
solvency	13536	29	0.002
insolvency	73	10	0.880
use error	0.005	0.744	Total error = 0.007

Table 2: Logistic regression classification table for forecasts of insolvency using the training sample. A linear model is based on all the predictors available in the data set.

[3] Notice that all these errors are *in-sample* errors and hence are likely to be overly optimistic.

[4] We also tried to use the same predictor variables as those used by Lee and Urrutia over the 4 year span, but the results were essentially the same.

	predict solvency	predict insolvency	model error
solvency	13565	0	0
insolvency	83	0	1
use error	0.006	N/A	Total error = 0.006

Table 3: Logistic regression classification table for the model of Lee and Urrutia for forecasts of insolvency using the training sample. The Linear model uses a subset of predictors available in the data set.

	predict solvency	predict insolvency	model error
solvency	1507	5278	0.778
insolvency	10	29	0.256
use error	0.007	0.995	Total error = 0.775

Table 4: Logistic regression classification table for forecasts of insolvency using the test sample.

misclassification cost into account, the results may be not that bad. But as we show next, Random Forest produces results superior to those of logistic regression even when accounted for asymmetric loss function.

5 Simulation analysis

We compare predictive qualities of several models, including the model of Lee and Urrutia (1996), logistic regression model that uses a larger set of variable than the one used by Lee and Urrutia (1996) and model based on the Random Forest algorithms using our dataset. The models performance is ranked by comparing the misclassification costs. It involves setting the cost of making a decision that a company will fail when it does not (type I error) and the cost of making a decision that a company will not fail when it does (type II error).

We present the comparison that directly accounts for asymmetric loss function. Since the objective is to compare the relative performance of the methods for a fixed loss function, it is without loss of generality to assume that type I error costs and type II error costs sum to one. For example, suppose that the cost ratio is 4. It implies that the cost of type I error

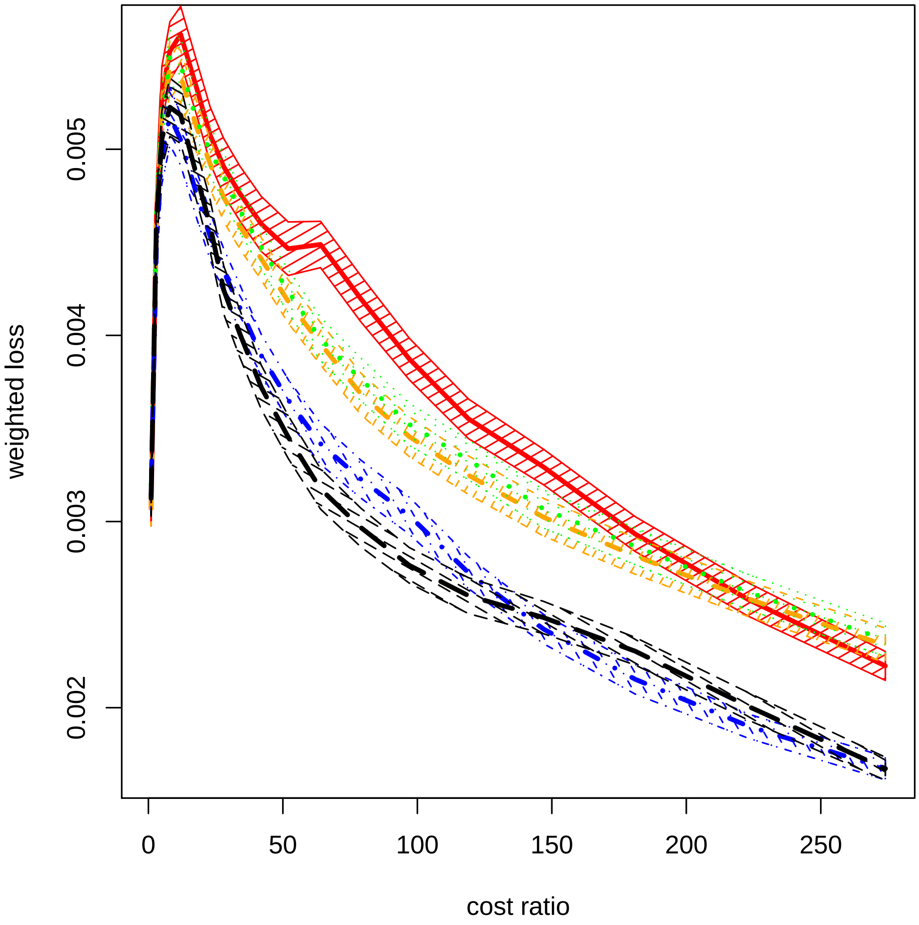


Figure 3: Weighted misclassification cost as a function of cost ratio for different models. Red solid line corresponds to Model 1, orange dashed to Model 2, green dotted to Model 3, blue dash-dotted to Model 4 and black long dashed to Model 5. The dashed polygons around the lines correspond to pointwise 95% confidence intervals.

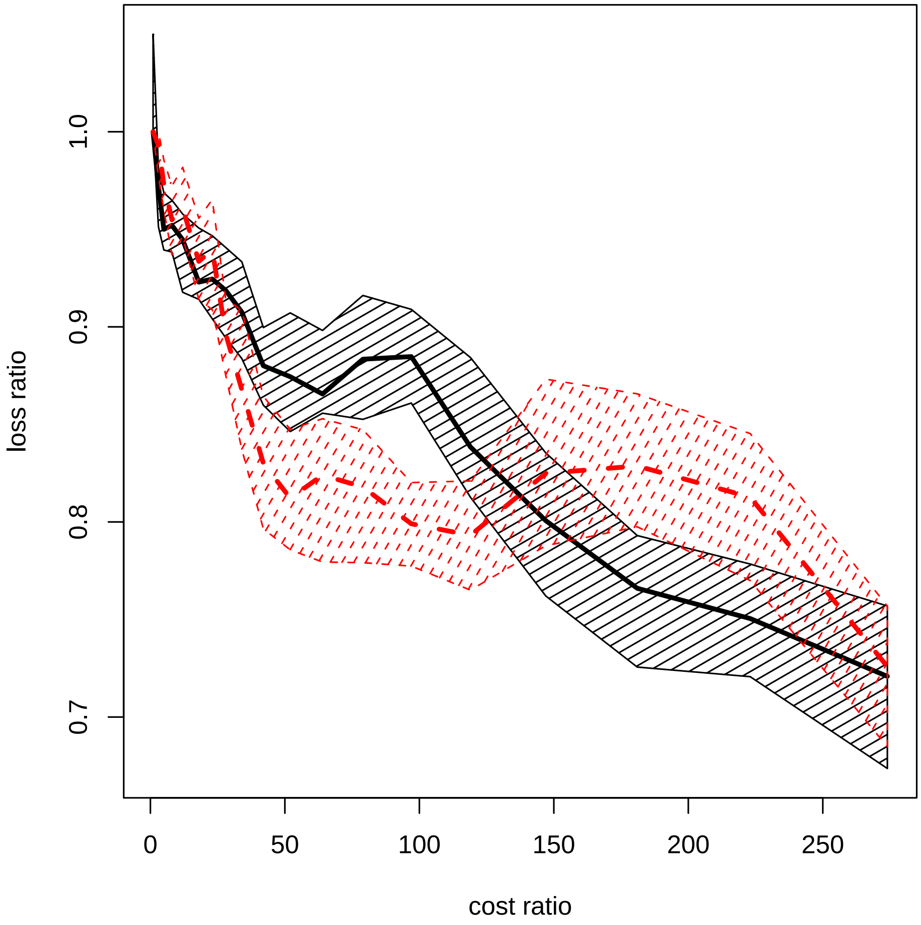


Figure 4: Median ratio of the weighted misclassification loss. Black solid line corresponds to the ratio of Model 4 to Model 2, red dashed line corresponds to the ratio of Model 5 to Model 2. Bands indicate non-parametric 95% confidence bands.

is 0.2 and the cost of type II error is 0.8. In the context of Table 4, the total weighted cost is going to be

$$C = 0.2 * 5278 + 0.8 * 10 = 1063.6.$$

In practice, different regulators may have different ideas about the appropriate ratio of the false negative to false positive costs. If the regulators fail to respond with appropriate regulatory actions against a financially vulnerable insurer (type II error), the delay may increase the guaranty fund assessments. These costs will eventually be passed on to policyholders, taxpayers and financially viable insurers. However, regulatory intervention in operations of a financially solid insurer (type I error) can result in inefficiencies in the operation of insurance markets. Also the cost ratios are likely to change over time even for the same regulator[5]. To account for these features, we consider a wide range of cost ratios ranging from 1 to 274.

We compare the performance of five different models.

- **Model 1.** The first method is the usual logistic regression that uses the set of variables similar to the one used by Lee and Urrutia (1996)[6]. To account for the asymmetric loss function, we first fit the logistic regression model and then use the data and loss function to determine the threshold level for classification. This last step is a deviation from the approach used by Lee and Urrutia who fitted a standard logistic regression model that uses 0.5 threshold to classify into solvent and insolvent companies. Since they had equal number of solvent and insolvent companies in the training set, they did not have to adjust the classification threshold. In our case, because of highly unbalance training set, this would result in classifying every company as solvent. Hence, we have to adjust the classification threshold. This model provides a benchmark method for comparison.

- **Model 2.** The second method is the logistic regression that uses all the available

[5] See Lamm-Tennant et al. (1996) and references therein for further discussion.

[6] We were not able to perfectly match the set of variables we had to the one used by Lee and Urrutia (1996), therefore we came up with a set that is expected to provide a similar behavior.

variables but employs only one year of data to make prediction for the next year. We use a theoretically determined threshold level for classification. The reason to include this method is to verify that there are benefits in looking back for large time spans and using non-parametric methods, like the Random Forest algorithm. Otherwise, a simple increase in the number of variables in the logistic regression model provides a similar payback.

- **Model 3.** This model is the same as Model 2 with the only difference that it uses data to determine the threshold level for classification.
- **Model 4.** Random Forest that uses subsampling to obtain quasi-probability estimates and then uses calibration to determine optimal thresholding level for classification. We use the same set of variables as in Models 2 and 3. In particular, this model uses only one year of data to predict insolvency next year.
- **Model 5.** This model is the same as Model 4, but it employs four years worth of data to predict next year insolvency.

As can be seen in Figure 3, all three logistic regression models give roughly similar results, while Random Forest algorithm gives lower weighted misclassification error. Notice that calibration (Model 3) does not help in case of logistic regression: it already performs as well as it can with theoretically determined threshold level (Model 2). Random Forest algorithm (Model 4) does better than logistic regression for cost ratio greater than 40 but does not seem to benefit from looking back several years (Model 5).

Notice that we should not compare values of the weighted loss curve at different cost ratios as these are not really comparable since we restricted our costs to sum to one. To simplify the comparison across different cost ratios, we used a ratio of the losses produced by Random Forest models to the loss of Model 2. Results of such comparison are shown in Figure 4. Overall, for cost ratio greater than 50 we get 10–25% improvement. For smaller cost ratios Random Forest still performs better than logistic regression.

6 What drives insolvencies?

We present two applications of RF classification. First, we identify the variables that contain the most information about an insurer's propensity to default. Then we describe the results on the individual effect of the most important variables and show that these are often highly non-linear.

6.1 Variables selection

What are the key financial variables that predict that an insurance company has a high risk of insolvency? One of the advantages of the Random Forest classification is that it can be used to rank the characteristics of the insurance company according to their ability to predict insolvency. The basic idea of the ranking procedure is to order the explanatory variables according to the reduction in the accuracy of prediction when the variable is excluded from the set of explanatory variables.

In general, the accuracy of RF is assessed using a bootstrap sample. In the process of growing the forest the algorithm repeats many times the same two steps. First, it generates a bootstrap sample from the original sample. Second, it grows a tree on a bootstrap sample. Hence, every time RF grows a tree using only a subset of the original sample. The data that did not go into the bootstrap sample (referred to as Out-Of-Bag (OOB) sample) can be used as an independent test sample for a particular tree, where the tree predictions are compared to the actual outcomes in OOB.

In order to measure the information of a particular variable, RF uses the following procedure. Given the data table in which each column corresponds to one variable, RF randomly shuffles values in the corresponding column to remove dependence of classification on this variable. Then classification accuracy with shuffled data is compared to classification with unshuffled data. The decrease in accuracy in the shuffled sample is the measure of variable importance. Thus the importance measure of a given company characteristic should be interpreted relative to the set of variables used for prediction.

Figure 5 shows the relative importance of the 30 most informative variables in the set of over 150 variables that include the 37 characteristics for the period of 4 years. It presents the results for the ranking estimated for the cost ratio 1:150, but the other cost ratios produce very similar results. Since the variable importance estimates are obtained from a random algorithm, the estimates are scaled by standard errors. The set of the top 30 variables has 12 distinct characteristics and their lagged values. For several variables, the current and the lagged value of the variable appears on the top of the list (for example, SB.1 and SB.2). The interpretation of this result is that the variables are important both individually and as a group. In particular, RF would not put on the top of the list the variables that are simply highly correlated but at the same time do not provide additional information as a group.

Figure 6 shows the unstandardized variable importance estimates. The order of variables is very similar to Figure 5. Unscaled versions in Figure 6 give very similar *relative* importance and allow to assess the effect of excluding the variable on the overall weighted variable. For example, if we exclude one year growth in equity capital (FAST8.1), then weighted error increases by approximately 0.00018, or, given that for cost ratio of 1:150 weighted loss is approximately 0.0025, the prediction loss increases by about 7%. For investment in stocks and bonds (SB.1) it is about 8%.

These results show variables importance for the case where we are concerned both with missing those companies that will default and mispredicting defaults. In contract, Figure 7 shows the ranking of variables importance in the class of insolvent companies. For variables SB1 and FAST8.1 this increase is about 3%. Since we put much larger weight on missing insolvent companies, the relative importance is essentially the same as in Figure 6 (at least at the top of the list).

The results summarized in Figure 5 show that the top ten most informative insurer's characteristics include the measures of the capital structure and its dynamics, composition of the investment portfolio, size and profitability. The ratio of the net premiums written to equity capital (FAST8), aide to equity capital due to reinsurance (FAST6) and the ratio of the adverse reserve development to equity capital (FAST9) are the most significant indicators that explain insolvencies. These characteristics are followed by another two usual measures of insurer's capital, the ratio of net premiums written to equity capital (FAST1) and the ratio of the insurance reserves to equity capital (FAST3). The top 10 list also includes the share of assets invested in stocks and bonds (SB). The operating cash flow (OCF) and the amount of net premiums written (NPW) reflect the importance of insurer's ability to generate funds from insurance operations and insurer's size for predicting insolvencies.

The next twenty variables list contains the lagged values of the top 10 list, along with the characteristics of assets liquidity, such as the one year change in liquid assets (FAST13) and the percentage of cash and short term investment in the insurer's investment portfolio (CSTI).

Table 6.1 summarizes how the propensity to default depends on the past years variables. The percentage invested in stocks and bonds and the reliance on reinsurance (FAST6) have the longest history of four years. That is, all four years of data on these characteristics are relevant to predict insolvency next year. It suggests that the composition of the investment portfolio and persistent reliance on reinsurance as a source of capital are important indicators of insurer's financial quality. The majority of the other variables have a history of two to three years. The percentage of cash and short term investment matters only for the current year.

Our analysis indicates that the insurer's propensity of insolvency can be summarized by a small number of characteristics that reflect the capital structure, allocation and liquidity of assets, size and profitability measures. Interestingly, while for some variables it is important to consider the historic performance (like the consistent use of reinsurance) for other variables the near past values are sufficient (like the operating cash flow). These results have both positive and normative implications. From the positive perspective, the variables lists of Figures 5–7 reveal the key drivers of insolvencies. From the normative perspective, our

Number of years of past performance	Variables
4	Stocks and Bonds (%) (SB) Aid to equity capital due to reinsurance (%) (FAST6)
3	Net premiums written to equity capital (FAST1) Insurance reserves to equity capital (FAST3) Aid to equity capital due to reinsurance (FAST6) Adverse reserve development to equity capital (FAST9) 1 year change in liquid assets (FAST13)
2	Gross premiums written to equity capital (FAST2) Net premiums written (NPW) Operating cash flow (OCF)
1 (current year)	Misc recoverable to equity capital (FAST20) Cash and short term investment (%) (CSTI) Premiums written in commercial short-tail lines (perpwcst)

Table 5: Summary of influence of past data on the propensity to default.

results can be used to justify the variable selection in other models of solvency prediction that are applied in the literature.

6.2 Individual effects of predictors

In this section, we describe the partial dependence relations for the most important variables. The partial dependence function measures how the propensity to default depends on each insurer's characteristic after averaging the propensity to default over other characteristics. The vertical axis is given on the logit scale defined as

$$f(x) = \log[p(x)] - \frac{\log[1 - p(x)] + \log[p(x)]}{2},$$

where $p(x)$ is the probability of default at point x. Thus, a linear dependence implies that the functions can be described in terms of the logit model. However, our results reveal that many of the dependencies are highly non-linear.

In the previous section, we identified the set of capital structure variables that are important to predict insolvency. The partial dependence analysis indicates that these variables have an asymmetric impact on the probability of default. Figure 8 shows the partial dependence plot for the one year growth in the equity capital in the current year (FAST8). It illustrates that the negative growth has a drastic impact on the probability of default. At the same time, the positive growth does not substantially reduce it. Thus while negative development in FAST8 is a strong signal that a company has high risk, positive FAST8 contains almost no information. Similarly, the value of adverse reserve development to equity capital (FAST9), plotted on Figure 9 is an important factor to predict insolvency when the value is slightly negative or positive. It does not decrease the propensity to default for the negative values. Unlike FAST8 and FAST9, the aid to equity capital due to reinsurance (FAST6), plotted on Figure 10, has almost linear relationship with the propensity to default (on a logit scale).

The value of the operating cash flow (OCF) and net premiums written (NPW) show how the impact of variation in a given characteristics on the propensity to default may depend on its current value. [7] On Figure 11, the response to the change of OCF is negligent for the high values of OCF. As the value decreases below 7, each marginal reduction of OCF causes a high increase in the probability of default. For the values of OCF that are around zero or slightly negative, the insolvency risk stabilizes. It increases again for more pronounced negative values of the cash flow.

Figure 12 illustrates the response to the change of NPW. It indicates that the response is almost absent for the values above 20 and below 5. However, the propensity to default increases substantially when the value of NPW reduces from 18 to 5.

The last two Figures 13 and 14 show the impact of the stock and bonds (SB) in the portfolio. When the company invests 85% or higher of its assets in stocks and bonds in the current year (Figure 13), this variable does not affect the probability of default. However, SB has an important impact on the default probability for values between 65% and 85%. When the SB falls below 65%, further decrease in investment increases the probability of default,

[7] The values of OCF and NPW are shown on logarithmic scale. Since the values can be positive or negative, each value x was converted according to the formula $sign(x)log(abs(x) + e)$.

but at a lower rate. Figure 14 shows a similar result for the investment in the previous year.

The results of this section suggest that impact of the most important financial strength variables depends on the initial conditions of the insurance company. The analysis identified the "critical zones" of the variables where the marginal changes in insurer's strategy have a substantial effect on the insolvency risk. The results have important policy implications. They imply that the efficiency of identification and regulatory actions for high risk companies can be substantially improved when the policy differentiates between the marginal effects of different company characteristics on its financial strength.

7 Conclusion

In this paper we analyzed the insolvency prediction for the U.S. property-casualty insurance companies using the Random Forrest algorithm, a non-parametric machine learning method. Our results show that Random Forrest produces higher quality of prediction compared to other common methods, in particular, the logistics regression. The superiority of the prediction quality is due to the ability of Random Forrest to handle effectively highly unbalanced data with a wide pool of explanatory variables and their interactions. We show that the superiority of prediction arises over a wide range of cost ratios for type I/ type II errors.

We also show how Random Forrest can be used to identify the key variables responsible for company's insolvency. The analysis of partial dependencies indicates that the effect of the variables is highly non-linear. The importance of the past years variables varies from the long history of four years for the reliance on reinsurance to the short history of the current year for the percentage of cash and short term investment.

An important policy implication of our results is that the solvency requirements have to recognize that the marginal effect of the variable on the propensity of default is heterogeneous across different values of the company's characteristics.

A The Random Forest algorithm

Pseudocode of the algorithm is as follows.

1. Generate a bootstrap sample from the original sample. Sampling can be done either with or without replacement and size of the bootstrap sample does not have to be the same as of the original sample.

2. Grow a decision tree on the bootstrap sample. When splitting a node randomly select k variables out of d (where k is fixed for the given forest) and choose a best split with respect to these k variables (this is different from regular CART trees (Breiman et al., 1984) where split is done by searching over all variables), selection of variables is done independently in each node.

3. Repeat steps 1 and 2 m times, where m is the number of trees we want to grow. We would prefer to use $m = \infty$, but since this is not feasible $m = 500$ works well.

4. To classify a point x let each tree in the forest vote and then average votes over the trees.

B Comparison procedure

Since we want to compare performance of the Random Forest to the performance of the logistic regression model across different costs of type I and type II errors, we cannot use the value of the error, since it will depend on these costs. Instead we look at the ratio of the misclassification costs for the same pair of training and test data sets for different classification methods. The cross-validation algorithm we used to compare the methods can be described as follows.

1. Decide on the number of cross-validation iterations. We used $B = 200$.

2. For $i = 1$ to B do

   (a) Randomly split the data set into training and test parts.

   (b) For each cost on the list do

      i. Obtain three classifiers using the methods under consideration.

      ii. Validate their cost-sensitive error on the test data set.

      iii. Compute the ratio of the error for the logistic regression model that use all the variables under consideration to the error of the logistic regression model of Lee and Urrutia (1996). Also compute the ratio of the error of the Random Forest Model to the error of the logistic regression model of Lee and Urrutia (1996).

3. For the two $12 \times B$ matrix of ratios we obtain (each row of each matrix corresponds to one of the costs and each column corresponds to one of the splits of the data into train and test parts) for each row compute the median and the 95% confidence interval for the median.

4. Plot the corresponding graphs.

C Variable importance algorithm

The variable importance algorithm, implemented by `randomForest` package (Liaw and Wiener, 2002), does not handle the case of the unbalanced data properly: it does a good job at estimating the in-class variable importance, but a poor job when estimating the average variable importance when taking into account the prior class weights. Here we provide an algorithm that handles the misclassification costs properly.

1. For every tree in the forest do

   (a) Generate a test sample from the out-of-bag (OOB) data (that is, the part of the original training sample that was not selected into the subsample used to build this particular tree.) The subsample is generated in such a manner, that we take all the examples of the underrepresented class and subsample examples from the overrepresented class, so that proportion of examples from the underrepresented class in the subsample is the same as in the original training sample.

   (b) Determine the *weighted* tree accuracy on the obtained OOB sample.

   (c) For every variable do

      i. Permute values in the OOB sample for this variable.

      ii. Compute the *weighted* tree accuracy on the permuted OOB dataset.

      iii. Compute the decrease in accuracy between the permuted and unpermuted OOB datasets.

2. Compute average accuracy decrease for each variable across all the trees.

3. Compute the SD of accuracy decrease for each variable.

D Figures and Tables

Variable name	Description
FAST1	Kenney Ratio: Net Premiums Written to Equity Capital
FAST2	Gross Premiums Written to Equity Capital
FAST3	Insurance Reserves to Equity Capital
FAST4	1 Yr. Growth in Net Premiums Written (%)
FAST5	1 Yr. Growth in Gross Premiums Written (%)
FAST6	Aid to Equity Capital due to Reinsurance
FAST7	Investment Yield (%)
FAST8	1 Yr. Growth in Equity Capital (%)
FAST9	Adverse Reserve Development to Equity Capital (%)
FAST10	D (Losses Incurred + Und. Exp's to Net Premiums Written Ratio)
FAST11	Gross Expenses to Gross Premiums Written
FAST12	1 yr. Change in Gross Expenses (%)
FAST13	1 yr. Change in Liquid Assets (%)
FAST14	1 Yr. Change in Agents Balances (%)
FAST15	Reins. Recoverable on Paid Losses to Equity Capital
FAST16	Reins. Recoverable on Unpaid Losses to Equity Capital
FAST17	Net Premiums Written in Liability Lines to Total Net Premiums
FAST18	Investments in Affiliates to Equity Capital
FAST19	Receivables from Affiliates to Equity Capital
FAST20	Misc. Recoverables to Equity Capital
FAST21	Non-investment Grade Bonds to Equity Capital
FAST22	Other Invested Assets to Equity Capital
FAST23	Dummy = 1 if insurer has a large single agent
FAST24	Dummy = 1 if insurer has a large single agent they control
FAST25	Losses, Exp's, Div's and Taxes Paid to Premiums Collected
NPW	Net Premiums Written
PWHERF_51	State of Business of Herfindahl
LINEHERF_PW	Line of Business Herfindahl
Ind_Agent	Dummy=1 if Insurer has an independent agent
perpwcst	Percentage of Premiums Written in Commercial Short-Tail Lines
perpwclt	Percentage of Premiums Written in Commercial Long-Tail Lines
perpwpst	Percentage of Premiums Written in Personal Short-Tail Lines
mutgrp	Dummy=1 if Insurer is Part of a Mutual Group
lta	Logarithm of total assets
smallco	Dummy=1 if Insurer is a Small Company
CSTI	Cash and Short-term Investment (%)
OCF	Operating Cash Flow, ($000)
QL	Quick Liquidity, (%)
SB	Stocks and Bonds (%)

Table 6: Variables and their definitions.

	Solvent Insurers		Insolvent Insurers		Wilcoxon
Variable	Mean	SD	Mean	SD	Ranked P-value
FAST1	1.13	0.85	1.87	1.12	0.0000
FAST2	2.2	1.92	4.19	2.66	0.0000
FAST3	1.03	0.94	1.65	1.25	0.0000
FAST4	11.87	41.62	11.69	61.21	0.0013
FAST5	11.94	37.63	11.06	52.93	0.0003
FAST6	2.05	4.34	6.07	7.52	0.0000
FAST7	5.71	1.38	5.41	1.55	0.0042
FAST8	8.82	16.3	-8.5	19.87	0.0000
FAST9	-2.73	10.8	4	11.62	0.0000
FAST10	0.01	0.23	0.06	0.37	0.0759
FAST11	0.58	0.76	0.55	0.64	0.8134
FAST12	0.05	0.47	0.09	0.58	0.9005
FAST13	1.17	2.66	0.37	1.79	0.0008
FAST14	-0.03	0.99	0.04	1.29	0.0777
FAST16	0.12	0.3	0.24	0.42	0.0000
FAST17	0.59	0.33	0.57	0.36	0.9385
FAST18	0.58	1.32	0.94	1.74	0.0285
FAST19	0.02	0.04	0.04	0.06	0.0000
FAST20	0.03	0.05	0.07	0.08	0.0000
FAST21	0.65	2.37	0.68	2.49	0.2558
FAST22	0.01	0.03	0.02	0.04	0.0375
FAST23	0.12	0.33	0.22	0.42	0.0000
FAST24	0.08	0.28	0.12	0.32	0.0816
FAST25	1.29	0.73	1.59	0.84	0.0000
NPW	15.84	4.80	15.42	3.99	0.0000
PWHERF_51	0.61	0.38	0.69	0.37	0.0022
LINEHERF_PW	4.41	506.14	1.07	6.39	0.0026
Ind_Agent	0.76	0.43	0.88	0.33	0.0021
perpwcst	0.3	0.34	0.2	0.28	0.0000
perpwclt	0.37	0.38	0.41	0.39	0.2536
perpwpst	0.09	0.15	0.11	0.19	0.0644
mutgrp	0.26	0.44	0.08	0.28	0.0000
lta	17.7	2.03	16.73	1.6	0.0000
smallco	0.77	0.42	0.93	0.25	0.0000
CSTI	13.59	19	13.11	16.26	0.8437
OCF	4.01	7.32	-3.58	7.31	0.0000
QL	99.6	174.41	41.52	71.29	0.0000
SB	73.01	20.83	51.32	26.7	0.0000

Table 7: Summary statistics for each variable across solvent and insolvent companies and p-values for the Wilcoxon rank-sum test of equality of means of two groups.

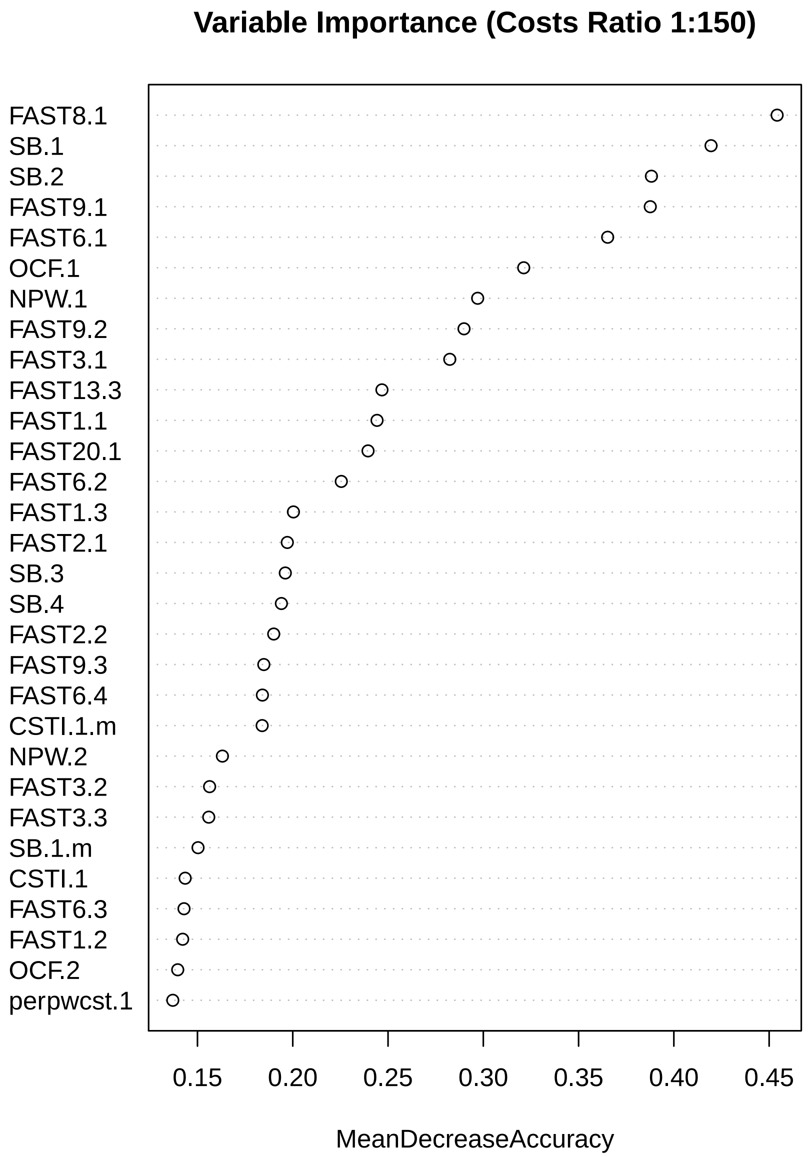


Figure 5: Estimated weighted mean decrease accuracy scaled by its standard error. Variable.t denotes the variable value at year t, and $t = 1$ stands for the current year. Variable.t.m is a dummy equal to one if the value of the variable is missing in year t.

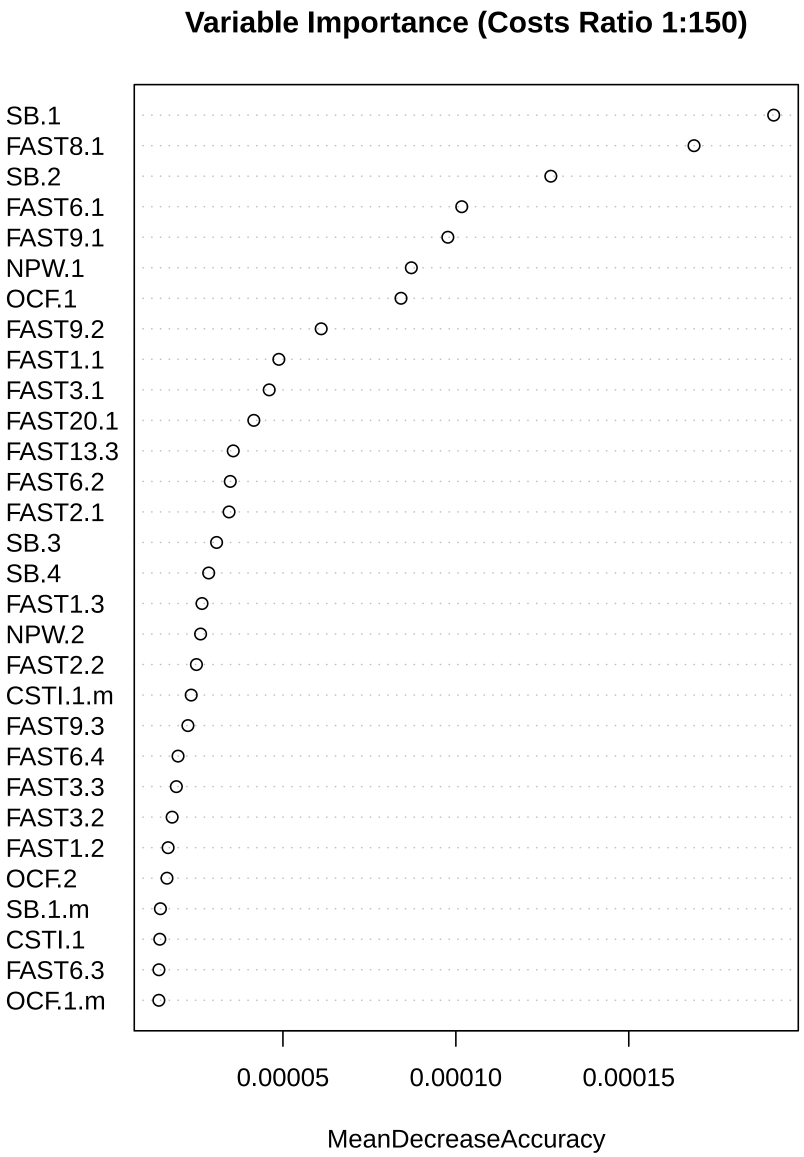


Figure 6: Estimated weighted loss increase. Variable.t denotes the variable value at year t, and $t = 1$ stands for the current year. Variable.t.m is a dummy equal to one if the value of the variable is missing in year t.

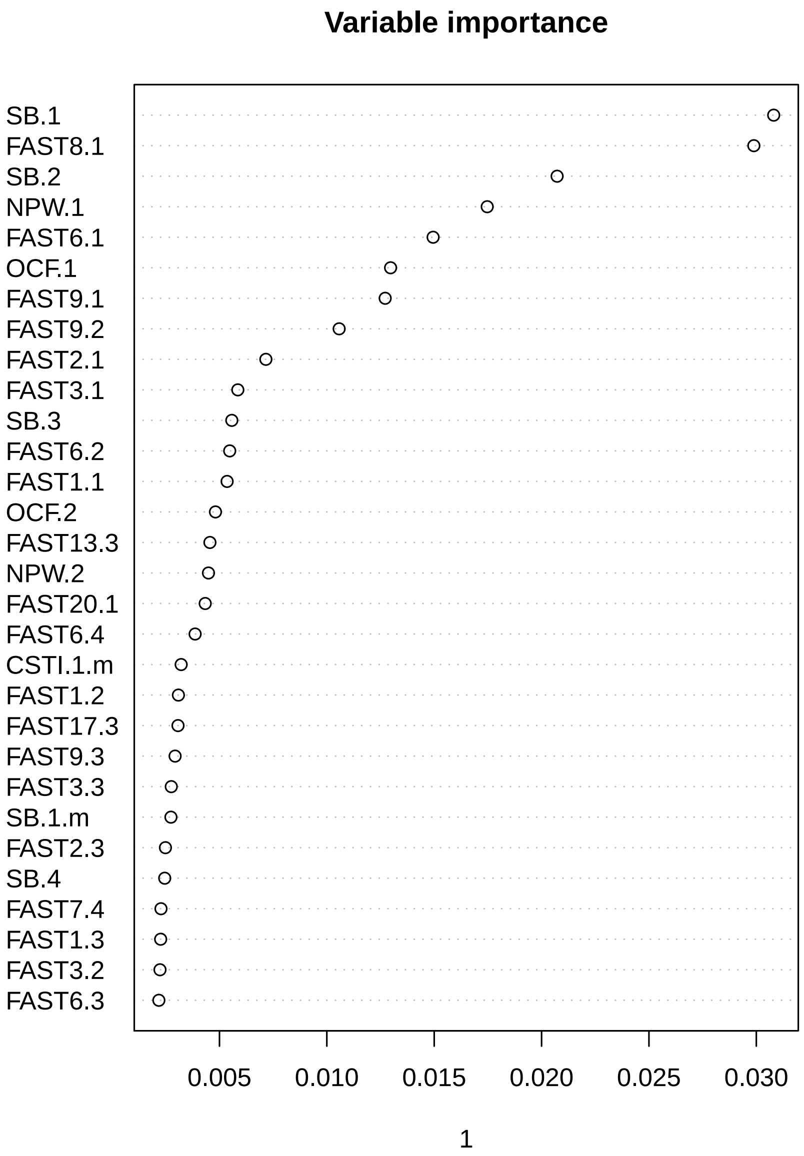


Figure 7: Predictor importance when predicting among defaulting companies. Variable.t denotes the variable value at year t, and $t = 1$ stands for the current year. Variable.t.m is a dummy equal to one if the value of the variable is missing in year t.

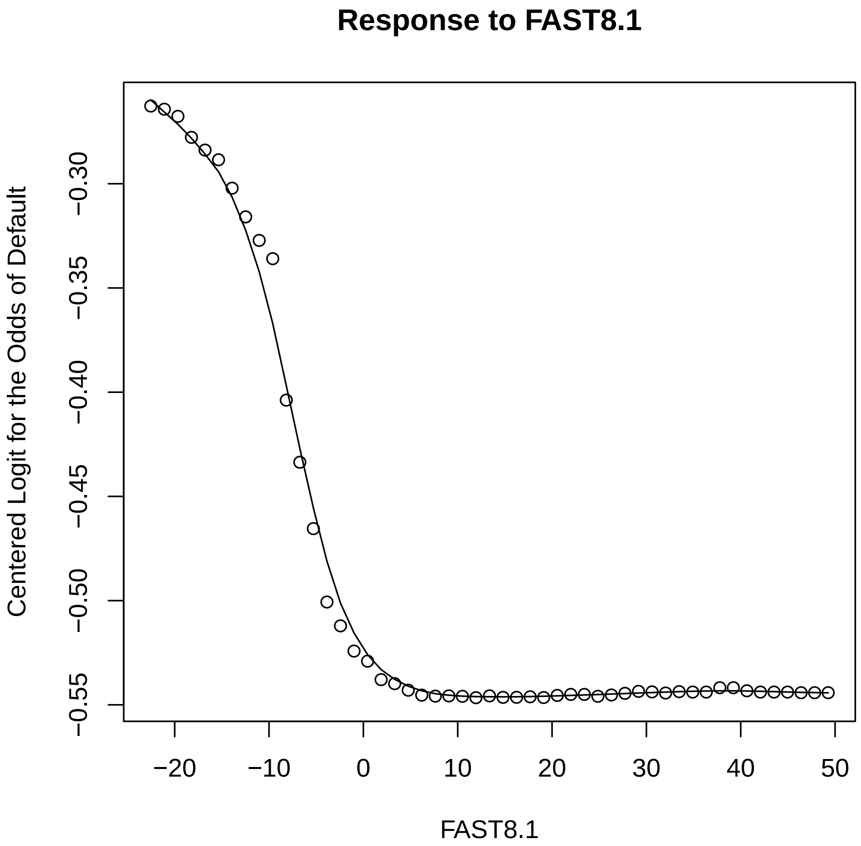


Figure 8: Partial dependence plot for one year growth in equity capital (FAST8.1)

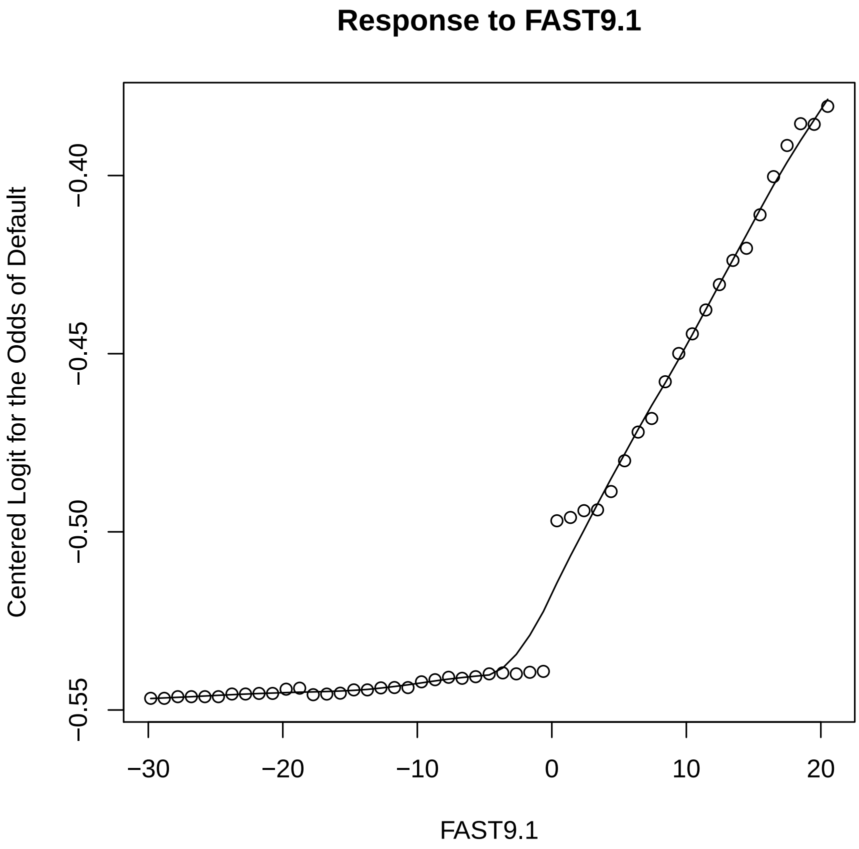


Figure 9: Partial dependence plot for adverse reserve development to equity capital (FAST9.1)

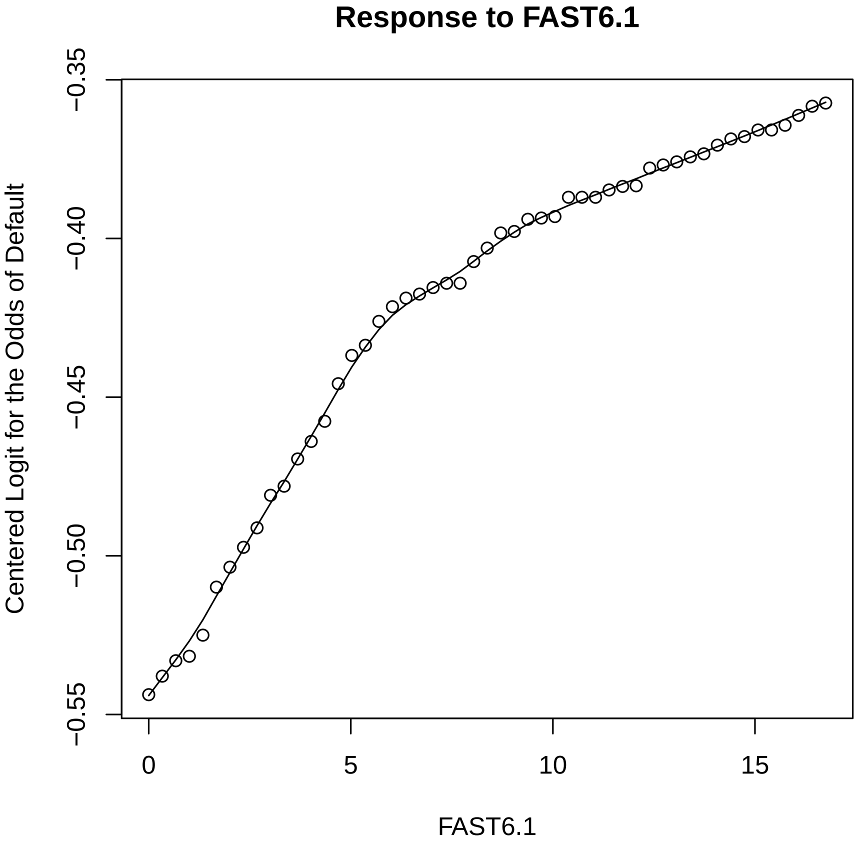


Figure 10: Partial dependence plot for aid to equity capital due to reinsurance (FAST6.1)

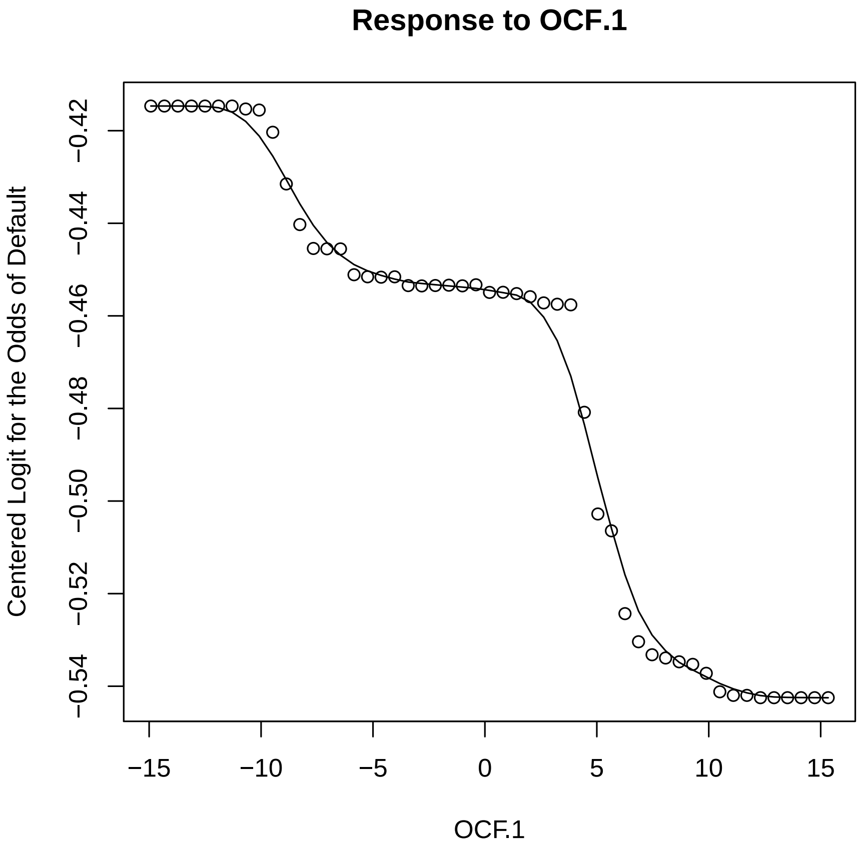


Figure 11: Partial dependence plot for operating cash flow (OCF.1)

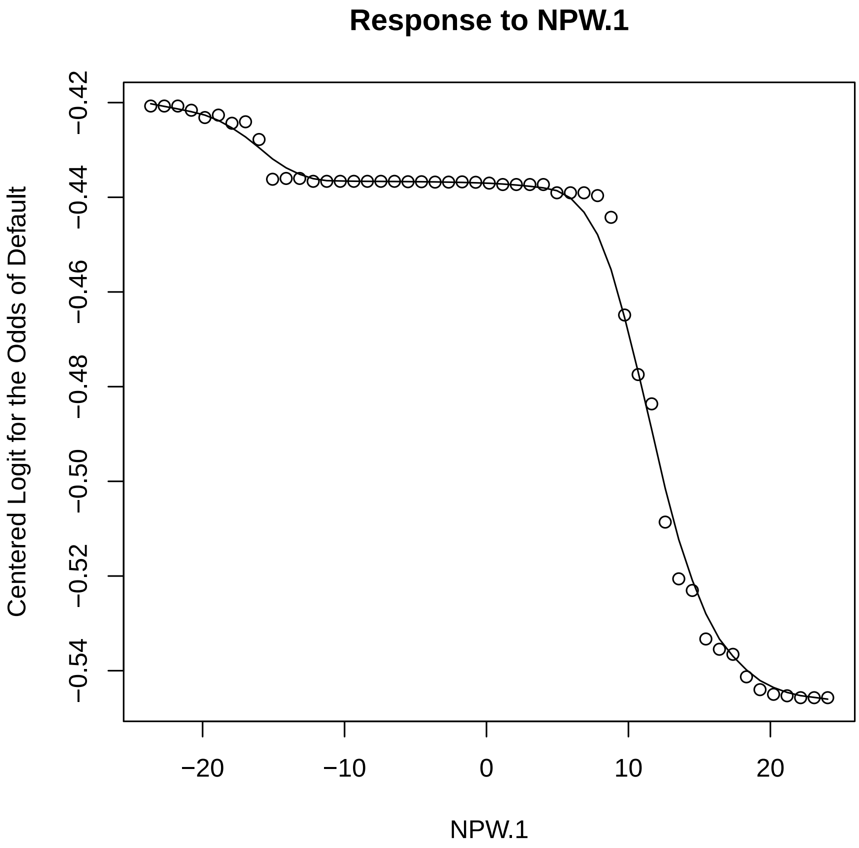


Figure 12: Partial dependence plot for net premiums written (NPW.1)

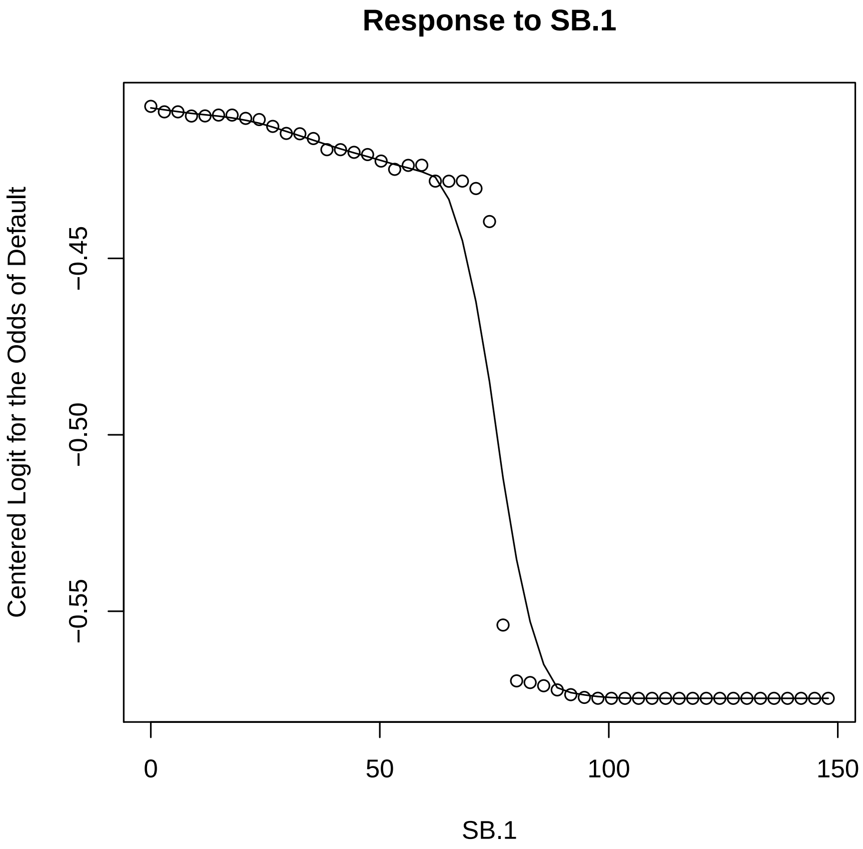


Figure 13: Partial dependence plot for the percentage invested in stocks and bonds in the current year (SB.1)

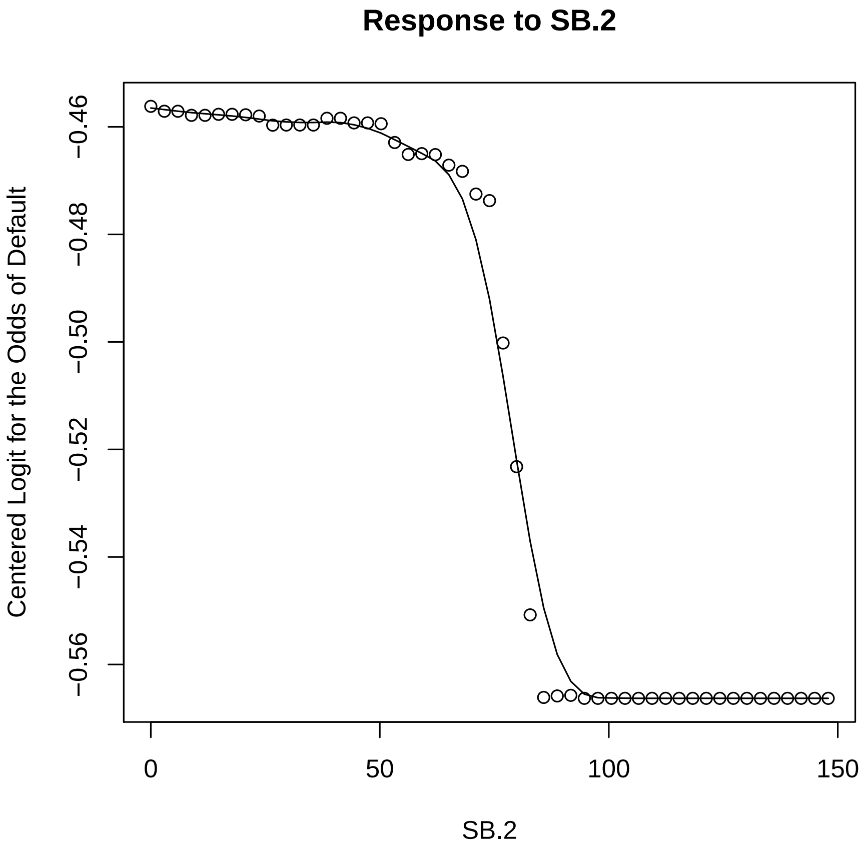


Figure 14: Partial dependence plot for the percentage invested in stocks and bonds in the previous year (SB.2)

References

Edward I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609, 1968. ISSN 00221082. URL `http://www.jstor.org/stable/2978933`.

Jan M. Ambrose and Anne M. Carrol. Using best's ratings in life insurer insolvency prediction. *Journal of Risk and Insurance*, 61(2):317–327, 1994.

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

R.A. Berk. An introduction to ensemble methods for data analysis. *Sociological Methods and Research*, 34(3):263–295, 2006.

Richard Berk, Lawrence Sherman, Geoffrey Barnes, Ellen Kurtz, and Lindsay Ahlman. Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):191–211, 2009. doi: 10.1111/j.1467-985X.2008.00556.x. URL `http://dx.doi.org/10.1111/j.1467-985X.2008.00556.x`.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40 (3):229–242, 2000.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen. *Classification and Regression Trees.* Chapman & Hall/CRC, 1984.

J. David Cummins, Martin F. Grace, and Richard D. Phillips. Regulatory solvency prediction in property-liability insurance: Risk-based capital, audit ratios, and cash flow simulation. *The Journal of Risk and Insurance*, 66(3):417–458, Sep. 1999.

Ramón Díaz-Uriarte and Sara Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-3. URL `http://www.biomedcentral.com/1471-2105/7/3`.

Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 40 (2):139–158, 2000.

Martin F. Grace, S. Harrington, and R. Klein. An analysis of the fast monitoring system. Technical report, NAIC, December 2 1995. A Report Presented to the NAIC's Financial Analysis Research and Development Working Group.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2001.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

A. E. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit risk models via machine-learning algorithms. *Journal of Banking and Finance*, forthcoming, 2010.

Joan Lamm-Tennant, Laura Starks, and Lynne Stokes. Considerations of cost trade-offs in insurance solvency surveillance policy. *Journal of Banking and Finance*, 20:835–852, 1996.

Bart Larivière and Dirk Van den Poel. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29(2):472–484, 2005.

Suk Hun Lee and Jorge L. Urrutia. Analysis and prediction of insolvency in the property-liability insurance industry: A comparison of logit and hazard models. *The Journal of Risk and Insurance*, 63(1):121–130, Mar. 1996.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2 (3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

David Mease, Abraham J. Wyner, and Andreas Buja. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*, 8:409–439, May 2007.

George E. Pinches and James S. Trieschmann. The efficiency of alternative models for solvency surveillance in the insurance industry. *The Journal of Risk and Insurance*, 41(4): 563–577, 1974. ISSN 00224367. URL `http://www.jstor.org/stable/251955`.

George E. Pinches and James S. Trieschmann. Discriminant analysis, classification results, and financially distressed p-l insurers. *The Journal of Risk and Insurance*, 44(2):289–298, 1977. ISSN 00224367. URL `http://www.jstor.org/stable/252140`.

Steven W. Pottier. Life insurer financial distress, best's ratings and financial ratios. *Journal of Risk and Insurance*, 65(2):275–288, 1998.

Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph. Random forest similarity for protein-protein interaction prediction from multiple sources. In *Pacific Symposium on Biocomputing 10*, pages 531–542, 2005.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58:267–288, 1996.

James S. Trieschmann and George E. Pinches. A multivariate model for predicting financially distressed p-l insurers. *The Journal of Risk and Insurance*, 40(3):327–338, 1973. ISSN 00224367. URL `http://www.jstor.org/stable/252222`.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Methodological)*, 67:301–320, 2005.