Risk and Risk Management in the Credit Card Industry

Florentin Butaru, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W. Lo and Akhtar Siddique

August 2016

Macro Finance Research Program
of the Becker Friedman Institute

MFM
Macro Financial Modeling

Journal of Banking and Finance 72 (2016) 218–239

Contents lists available at ScienceDirect

Journal of Banking and Finance

journal homepage: www.elsevier.com/locate/jbf

Risk and risk management in the credit card industry☆

Florentin Butaru[a], Qingqing Chen[a], Brian Clark[a,e], Sanmay Das[b], Andrew W. Lo[c,d,*], Akhtar Siddique[a]

[a] U.S. Department of the Treasury, Office of the Comptroller of the Currency, Enterprise Risk Analysis Division, United States
[b] Washington University in St. Louis, Department of Computer Science & Engineering, United States
[c] Massachusetts Institute of Technology, Sloan School of Management, Computer Science and Artificial Intelligence Laboratory, Electrical Engineering and Computer Science, United States
[d] AlphaSimplex Group, LLC, United States
[e] Rensselaer Polytechnic Institute (RPI), Lally School of Management, United States

ARTICLE INFO

Article history:
Received 30 December 2015
Accepted 29 July 2016
Available online 8 August 2016

JEL classification:
G21
G17
D12
C55

Keywords:
Credit risk
Consumer finance
Credit card default model
Machine-learning

ABSTRACT

Using account-level credit card data from six major commercial banks from January 2009 to December 2013, we apply machine-learning techniques to combined consumer tradeline, credit bureau, and macroeconomic variables to predict delinquency. In addition to providing accurate measures of loss probabilities and credit risk, our models can also be used to analyze and compare risk management practices and the drivers of delinquency across banks. We find substantial heterogeneity in risk factors, sensitivities, and predictability of delinquency across banks, implying that no single model applies to all six institutions. We measure the efficacy of a bank's risk management process by the percentage of delinquent accounts that a bank manages effectively, and find that efficacy also varies widely across institutions. These results suggest the need for a more customized approached to the supervision and regulation of financial institutions, in which capital ratios, loss reserves, and other parameters are specified individually for each institution according to its credit risk model exposures and forecasts.

1. Introduction

The financial crisis of 2007–2009 highlighted the importance of risk management within financial institutions. Particular attention has been given to the risk management practices and policies at the mega-sized banks at the center of the crisis in the popular press and the academic literature. Few dispute that risk management at these institutions—or the lack thereof—played a central role in shaping the subsequent economic downturn. Despite this recent focus, however, the risk management policies of individual institutions largely remain black boxes.

In this paper, we examine the practice and implications of risk management at six major U.S. financial institutions, using computationally intensive "machine-learning" techniques applied to an unprecedentedly large sample of account-level credit card data. The consumer credit market is central to understanding risk management at large institutions for two reasons. First, consumer credit in the United States has grown explosively over the past three decades, totaling $3.3 trillion at the end of 2014. From the early 1980s to the Great Recession, U.S. household debt as a percentage of disposable personal income has doubled, although declining interest rates have meant that debt service ratios have grown at a lower rate. Second, algorithmic decision-making tools, including the use of scorecards based on "hard" information, have become increasingly common in consumer lending (Thomas, 2000). Given the larger amount of data, as well as the larger number of decisions compared to commercial credit lending, this new reliance on algorithmic decision-making should not be surprising. However, the implications of these tools for risk management, for individual financial institutions and their investors, and for the economy as a whole, are still unclear.

Credit card accounts are revolving credit lines, and because of this, lenders and investors have more options to actively monitor and manage them compared to other retail loans, such as mortgages. Consequently, managing credit card portfolios is a potential source of significant value to financial institutions. Better risk management could provide financial institutions with savings on the order of hundreds of millions of dollars annually. For example, lenders could cut or freeze credit lines on accounts that are likely to go into default, thereby reducing their exposure. By doing so,

☆ Disclaimer: The statements made and views expressed herein are solely those of the authors and do not necessarily represent official policies, statements, or views of AlphaSimplex Group, the Office of the Comptroller of the Currency, MIT, RPI, Washington University, or their employees and affiliates.

* Corresponding author.
E-mail address: alo-admin@mit.edu (A.W. Lo).

http://dx.doi.org/10.1016/j.jbankfin.2016.07.015

they potentially avoid an increase in the balances of accounts destined to default, known in the industry as "run-up." However, cutting these credit lines to reduce run-up also runs the risk of cutting the credit limits of accounts that will not default, thereby alienating customers and potentially forgoing profitable lending opportunities. More accurate forecasts of delinquencies and defaults reduce the likelihood of such false positives. Issuers and investors of securitized credit card debt would also benefit from such forecasts and tools. Finally, given the size of this part of the industry—$861 billion of revolving credit outstanding at the end of 2014—more accurate forecasts would improve macroprudential policy decisions, and reduce the likelihood of a systemic shock to the financial system.

Our data allow us to observe the actual risk management actions undertaken by each bank at the account level, for example, credit line decreases and realized run-ups over time — and thus determine the possible cost savings to the bank for a given risk management strategy. The cross-sectional nature of our data further allows us to compare risk management practices across institutions, and examine how actively and effectively different firms manage the exposure of their credit card portfolios. We find significant heterogeneity in the credit line management actions across our sample of six institutions.

We compare the efficacy of an institution's risk management process using a simple measure: the ratio of the percentage of credit line decreases on accounts that become delinquent over a forecast horizon, to the percentage of credit line decreases on all accounts over the same period. This measures the extent to which institutions are targeting "bad" accounts, and managing their exposure prior to default.[1] We find that this ratio ranges from less than one, implying that the bank was more likely to cut the lines of good accounts than those that eventually went into default, to over 13, implying the bank was highly accurate in targeting bad accounts. While these ratios vary over time, the cross-sectional ranking of the institutions remains relatively constant, suggesting that certain firms are either better at forecasting delinquent accounts, or view line cuts as a beneficial risk management tool.

Because effective implementation of the above risk management strategies requires banks to be able to identify accounts that are likely to default, we build predictive models to classify accounts as good or bad. The dependent variable is an indicator variable equal to 1 if an account becomes 90 days past due (delinquent) over the next two, three, or four quarters. Independent variables include individual account characteristics such as the current balance, utilization rate, and purchase volume; individual borrower characteristics obtained from a large credit bureau, including the number of accounts an individual has outstanding, the number of other accounts that are delinquent, and the credit score; and macroeconomic variables including home prices, income, and unemployment statistics. In all, we construct 87 distinct variables.

Using these variables, we compare three modeling techniques: logistic regression, decision trees using the C4.5 algorithm, and the random forests method. The models are all tested out of sample as if they were implemented at that point in time, i.e., no future data were used as inputs in these tests. All models perform reasonably well, but the decision tree models tend to perform the best in terms of classification rates. In particular, we compare the models based on the well-known measures of precision and recall, and measures that combine them, the *F*-measure and the kappa statistic.[2] We find that the decision trees and random forest models outperform logistic regression with respect to both sets of measures.

There is a great deal of cross-sectional and temporal heterogeneity in these models. As expected, the performance of all models declines as the forecast horizon increases. However, the performance of the models for each bank remains relatively stable over time.[3] Across banks, we find a great deal of heterogeneity in classification accuracy. For example, at the two-quarter forecast horizon, the mean *F*-measure ranges from 63.8% at the worst performing bank to 81.6% at the best.[4] Kappa statistics show similar variability.

We also estimate the potential cost savings from active risk management using these machine-learning models. The basic estimation strategy is to classify accounts as good or bad using the above models, and then to cut the credit lines of the bad accounts. The cost savings will depend on the model accuracy and how aggressively a bank will cut its credit lines. However, this strategy incurs a potential cost by cutting the credit lines of good accounts, thereby alienating customers and losing future revenue. We follow Khandani et al. (2010) methodology to estimate the "value added" of our models, and report the cost savings for various degrees of line cuts, ranging from no cuts to cutting the account limit to the current balance. To include the cost of alienating customers, we conservatively assume that customers incorrectly classified as bad will pay off their current balances and close their accounts, the bank losing out on all future revenues from such customers.

Ultimately, this measure represents the savings a bank would realize by freezing credit lines of all accounts forecast by our models to default, relative to what the bank would have saved if it had perfect foresight, cutting credit limits on all and only bad accounts. As such, it is representative only of the potential savings from the specific risk management activity we discuss in the paper (i.e., cutting credit lines), and it should not be interpreted as a percentage savings on the entire credit card portfolio, which includes revenues from other sources, including interest and purchase fees.

With respect to this measure, we find that our models perform well. Assuming that cutting the lines of bad accounts would save a run-up of 30% of the current balance, we find that our decision tree models would save about 55% of the potential benefits relative to perfect risk management, compared to taking no action for the two-quarter horizon forecasts (this includes the costs incurred in cutting the lines of good accounts). When we extend the forecast horizon, the models do not perform as well, and the cost savings decline to about 25% and 22% at the three- and four-quarter horizons, respectively. These results vary considerably across banks. The bank with the greatest cost savings had a value added of 76%, 46%, and 35% across the forecast horizons; the bank with the smallest cost savings would only stand to gain 47%, 14%, and 9% by implementing our models across the three horizons. Of course, there are many other aspects of a bank's overall risk management program, so the quality of risk management strategy of these banks cannot be ranked solely on the basis of these results, but the results do suggest that there is substantial heterogeneity in the risk management tools and effective strategies available to banks.

Khandani et al. (2010) is the paper most like ours in applying machine-learning tools to very large financial datasets. Our paper is differentiated from Khandani et al. in two significant ways. The first is that, unlike Khandani et al. (2010) who focus on a

[1] Despite the unintentionally pejorative nature of this terminology, we adopt the industry convention in referring to accounts that default or become delinquent as "bad" and those that remain current as "good."

[2] Precision is defined as the proportion of positives identified by a technique that are truly positive. Recall is the proportion of positives that is correctly identified. The *F*-measure is defined as the harmonic mean of precision and recall, and is meant to describe the balance between precision and recall. The kappa statistic measures performance relative to random classification. See Fig. 1 for further details.

[3] We test the models semi-annually, starting in 2010Q4 through the end of our sample period in 2013Q4.

[4] These *F*-measures represent the mean *F*-measure for a given bank over time.

single bank, we have data on a cross-section of banks. Therefore, we compare models for forecasting defaults across banks, and also compare risk management across the same banks. Another advantage of the cross-section of banks is our ability to compare the drivers of delinquency across the different banks. One set of drivers we look at are macroeconomic variables. On the other hand, Khandani et al. (2010) have a significantly richer dataset for the single bank in that they have account level transactions on credit and debit cards as well as balance information on checking accounts and CDs.

The remainder of the paper is organized as follows. In Section 2, we describe our dataset, and discuss the security issues surrounding it and the sample selection process used. In Section 3, we outline the model specifications and our approach to constructing useful variables that serve as inputs to the algorithms we employ. We also describe the machine-learning framework for creating more powerful forecast models for individual banks, and present our empirical results. We apply these results to analyze bank risk management and the key risk drivers across banks in Section 4. We conclude in Section 5.

2. Data

A major U.S. financial regulator has engaged in a large-scale project to collect detailed credit card data from several large U.S. financial institutions. As detailed below, the data contains internal account-level data from the banks merged with consumer data from a large U.S. credit bureau, comprising over 500 million records of individual accounts over a period of 6 years. It is a unique dataset that combines the detailed data available to individual banks with the benefits of cross-sectional comparisons across banks.

The underlying data contained in this dataset is confidential, and therefore has strict terms and conditions surrounding its usage and dissemination of results to ensure the privacy of the individuals and the institutions involved in the study. A third-party vendor is contracted to act as the intermediary between the reporting financial institutions, the credit bureau, and the regulatory agency, and end-users at the regulatory agency are not able to identify any individual consumers from the data. We are also prohibited from presenting results that would allow the identification of the banks from which the data are collected.

2.1. Unit of analysis

The credit card dataset is aggregated from two subsets we refer to as account-level and credit bureau data. The account-level data is collected from six large U.S. financial institutions. It contains account-level (tradeline) variables for each individual credit card account on the institutions' books, and is reported monthly starting January 2008. The credit bureau data is obtained from a major credit bureau, and contains information on individual consumers reported quarterly starting the first quarter of 2009.

This process results in a merged dataset containing 186 raw data items (106 account-level items and 80 credit bureau items). The account-level data includes items such as month-ending balance, credit limit, borrower income, borrower credit score, payment amount, account activity, delinquency, etc. The credit bureau data includes consumer-level variables such as total credit limit, total outstanding balance on all cards, number of delinquent accounts, etc.[5]

We then augment the credit card data with macroeconomic variables at the county and state level, using data from the Bureau of Labor Statistics (BLS) and Home Price Index (HPI) data from the Federal Housing Finance Agency (FHFA). The BLS data are at the county level, taken from the State and Metro Area Employment, Earnings, and Hours (SM) series and the Local Area Unemployment (LA) series, each of which is collected under the Current Employment Statistics program. The HPI data are at the state level. The BLS data are matched using ZIP codes.

Given the confidentiality restrictions of the data, the unit of analysis in our models is the individual account. Although the data has individual account-level and credit bureau information, we cannot link multiple accounts to a single consumer. That is, we cannot determine if two individual credit card accounts belong to the same individual. However, the credit bureau data does allow us to determine the total number of accounts that the owner of each of the individual accounts has outstanding. Similarly, we cannot determine unique credit bureau records, and thus we have multiple records for some individuals. For example, if individual A has five open credit cards from two financial institutions, we are not able to trace those accounts back to individual A. However, for each of the five account-level records, we would know from the credit bureau data that the owner of each of the accounts has a total of five open credit card accounts.

2.2. Sample selection

The data collection by the financial regulator started in January 2008 for supervisory purposes. For regulatory reasons, the banks from which the data have come have changed over time, although the total number has stayed at eight or less. However, the collection has always covered the bulk of the credit card market. Mergers and acquisitions have also altered its population over this period.

Our final sample consists of six financial institutions, chosen because they have reliable data spanning our sample period. Although data collection commenced in January 2008, our sample starts in 2009Q1 to coincide with the start of the credit bureau data collection. Our sample period runs through the end of 2013.[6]

The very large size of the dataset has forced us to draw a randomized subsample from the entire population of data. For the largest banks in our dataset, we sample 2.5% of the raw data. However, as there is substantial heterogeneity in the size of the credit card portfolios across the institutions, we sample 10%, 20%, and 40% from the smallest three banks in our sample. The reason is simply to render the sample sizes comparable across banks, so that differences in the amount of data available for the machine-learning algorithms are not driving the results. [7]

These subsamples are selected using a simple random sampling method. Starting with the January 2008 data, each of the credit card accounts is given an 18-digit unique identifier based on the encrypted account number. The identifiers are simple sequences starting at some constant and increasing by one for each account. The individual accounts retain their identifiers, and can therefore be tracked over time. As new accounts are added to the sample in subsequent periods, they are assigned unique identifiers that increase by one for each account.[8] As accounts are charged off,

[5] The credit bureau data for individuals is often referred to as "attributes" in the credit risk literature.

[6] We also drew samples at December 2011 and December 2012. Our results using those samples are quite similar. When we test the models, our out-of-time test sample extends to 2014Q2 for our measure of delinquency.

[7] While modern computing can handle increasingly large datasets for machine-learning algorithms, we are limited to a 2.5% sample in the data construction phase. In particular, our raw data (full time horizon of monthly data for all sizes of banks, plus the quarterly credit bureau data) is about 30 TB, which we have to clean, merge, and sort. As such, we are practically limited by the size of the full dataset when building the dataset and creating variables.

[8] For example, if a bank reported 100 credit card accounts in January 2008, the unique identifiers would be {C + 1, C + 2, ..., C + 100}. If the bank then added 20

sold, or closed, they simply drop out of the sample, and the unique identifier is permanently retired. We therefore have a panel dataset that tracks individual accounts through time, a necessary condition for predicting delinquency, and also reflects changes in the financial institutions' portfolios over time.

Once the account-level sample is established, we merge it with the credit bureau data. This process also requires care because the reporting frequency and historical coverage differ between the two datasets. In particular, the account-level data is reported monthly, beginning in January 2008, while the credit bureau data is reported quarterly, beginning in the first quarter of 2009. We merge the data using the link file provided by the vendor at the monthly level to retain the granularity of the account-level data. Because we merge the quarterly credit bureau data with the monthly account-level data, each credit bureau observation is repeated three times in the merged sample. However, we retain only the months at the end of each quarter for our models in this paper.

Finally, we merge the macroeconomic variables to our sample using the five-digit ZIP code associated with each account. While we do not have a long time series in our sample, there is a significant amount of cross-sectional heterogeneity that we use to identify macroeconomic trends. For example, HPI is available at the state level, and several employment and wage variables are available at the county level. Most of the macroeconomic variables are reported quarterly, which allows us to capture short-term trends.

The final merged dataset retains roughly 70% of the credit card accounts. From here, we only retain personal credit cards. The size of the sample across all banks increases steadily over time from about 5.7 million credit card accounts in 2009Q4 to about 6.6 million in 2013Q4.

3. Empirical design and models

In this section, we compare three basic types of credit card delinquency models: decision trees, random forests, and regularized logistic regression. In addition to running a series of "horse races" between the different models, we seek a better understanding of the conditions under which each type of model may be more useful. In particular, we are interested in how the models compare over different time horizons and changing economic conditions, and across banks.

We use the open-source software package Weka to run our machine-learning models. Weka offers a wide collection of machine-learning algorithms for data mining (see http://www.cs.waikato.ac.nz/ml/weka/ for more information). We start by giving a brief overview of the three types of classifiers we use. For the purposes of this discussion, we assume that we are solving a two-class classification problem, so the learning algorithm takes as input a training dataset, consisting of pairs $(\mathbf{x}, y)$, where $\mathbf{x} \in X$ is the feature or attribute vector (and can include categorical- as well as real-valued variables), and $y \in \{0,1\}$. The output of the learning algorithm is a mapping from X to $y \in \{0,1\}$ (or possibly, in the case of logistic regression, to [0, 1] where the output represents $\Pr(y=1)$). We now briefly describe the algorithms underlying these three models.

Decision trees are powerful models that can be viewed as partitions of the space X, with a specific prediction of y (either 0 or 1) for each such partition. If the model partitions the space into k mutually exclusive regions $R_1, \cdots, R_k$, then the model returned by a decision tree can be viewed as $f(\mathbf{x}) = \sum_{m=1}^{k} c_m I[\mathbf{x} \in R_m]$ where $c_m \in \{0,1\}$ and I is an indicator function (see Hastie et al., 2009). The partitioning is typically implemented through a series of hierarchical tests, thus the "tree" nomenclature.

While these models are rich and powerful, the space of decision trees is exponential on the number of features or attributes. It is thus effectively impossible to search the whole tree space to minimize any reasonable criterion on the in-sample training data. Therefore, most decision tree learning algorithms follow a greedy procedure, recursively partitioning the input space on the attribute that most reduces some measure of "impurity" on the examples that have filtered down to that node of the tree. The most commonly used measures of impurity are the Gini index and cross-entropy. We use Weka's J48 classifier, which implements the C4.5 algorithm developed by Quinlan (1993) (see Frank et al., 2011), which uses the reduction in cross-entropy, called the information gain. The other major procedure is that trees are typically restricted in height by some combination of rules to tell the tree when to stop splitting into smaller regions (typically when a region contains some M or fewer training examples), and *post-pruning* the tree after it has been fully constructed, which can be done in a number of different ways. This can be viewed as a form of regularization, reducing model complexity and giving up some in-sample performance, in order to generalize better to out-of-sample data. Since we use a relatively high value of M (see Section 4), we do not use post-pruning.

A major benefit of the decision tree model as a whole is its interpretability. While the greedy algorithm described above is not guaranteed to find the best model in the space of models it searches, greedy decision tree learners have been very successful in practice because of the combination of speed and reasonably good out-of-sample classification performance that they typically achieve. However, this comes as a tradeoff. The major negative of decision trees as a machine-learning algorithm is that they do not achieve state-of-the-art performance in out-of-sample classification (Dietterich, 2000; Hastie et al., 2009). Unfortunately, models that do achieve better performance are typically much harder to interpret, a significant negative for the domain of credit risk analysis. In order to determine how much improvement may be possible, we compare the decision tree models with one of these state-of-the art techniques, namely **random forests** (Breiman, 2001; Breiman and Cutler, 2004).

A random forest classifier is an ensemble method that combines two important ideas in order to improve the performance of decision trees, which are the base learners. The first idea is bagging, or bootstrap aggregation. Instead of learning a single decision tree, bagging resamples the training dataset with replacement T times, and learns a new decision tree model on each of these bootstrapped sample training sets. The classification model is then to allow all these T decision trees to vote on the classification, using majority vote to decide on the predicted class. The big benefit of bagging is that it greatly reduces the variance of decision trees, and typically leads to significant improvements in out-of-sample classification performance. The second key idea of random forests is to further reduce correlation among each of the induced trees by artificially restricting the set of features considered for each recursive split. When learning each tree, as each recursive split is considered, the random forest learner randomly selects some subset of the features (for classification tasks, typically the square root of the total number of features), and only considers those features. Random forests have been enormously successful empirically on many out-of-sample classification benchmarks in the last decade, and are considered among the best "out of the box" learning algorithms available today for general tasks (Caruana and Niculescu-Mezil, 2006; Criminisi et al., 2012).

Our third model is one that is more traditionally used in credit risk modeling and prediction in the finance and economics literature: **logistic regression**. In order to provide a fair comparison to the aforementioned methods, we use a *regularized* logistic regression model, which is known to perform better in

more accounts in February 2008, the unique identifiers of these new accounts would be {C + 101, C + 102, ..., C + 120}.

out-of-sample prediction. In particular, we apply a quadratic penalty function to the weights learned in a logistic regression model (a ridge logistic regression). We use the Weka implementation of logistic regression as per Cessie and van Houwelingen (1992). The log-likelihood is expressed as the following logistic function:

$$l(\beta) = \sum_i [y_i \log p(\mathbf{x_i}) + (1 - y_i) \log (1 - p(\mathbf{x_i}))]$$

where $p(\mathbf{x_i}) = \frac{e^{\mathbf{x_i}\beta}}{1+e^{\mathbf{x_i}\beta}}$. The objective function is then $l(\beta) - \lambda\beta^2$ where λ is the regularization or ridge parameter. The objective function is minimized using a quasi-Newton method.

In all, we have 87 attributes (variables) in our models, composed of account-level, credit bureau, and macroeconomic data.[9] We acknowledge that, in practice, banks tend to segment their portfolios into distinct categories when using logistic regression, and estimate different models on each segment. However, for our analysis, we do not perform any such segmentation. Our rationale is that our performance metric is solely based on classification accuracy. While it may be true that segmentation results in models that are more tailored to individual segments, such as prime versus subprime borrowers, thus potentially increasing forecast accuracy, we relegate this case to future research. For our current purposes, the number of attributes should be sufficient to approach the maximal forecast accuracy using logistic regression. We also note that decision tree models are well suited to aid in the segmentation process, and thus could be used in conjunction with logistic regression, but again leave this for future research.[10]

3.1. Attribute selection

Although there are few papers in the literature that have detailed account-level data to benchmark our features, we believe we have selected a set that adequately represents current industry standards, in part based on our collective experience. Glennon et al. (2008) is one of the few papers with data similar to ours. These authors use industry experience and institutional knowledge to select and develop account-level, credit bureau, and macroeconomic attributes. We start by selecting all possible candidate attributes that can be replicated from Glennon et al. (2008). Although we cannot replicate all of their attributes, we do have the majority of those that are shown to be significant after their selection process.

We also merge macroeconomic variables to our sample using the five-digit ZIP code associated with the account. As mentioned in Section 2, while we do not have a long time series of macroeconomic trends in our sample, there is a significant amount of cross-sectional heterogeneity that we use to pick up macroeconomic trends.

3.2. Dependent variable

Our dependent variable is delinquency status. For the purposes of this study, we define delinquency as a credit card account greater than or equal to 90 days past due. This differs from the standard accounting rule by which banks typically charge off accounts that are 180 days or more past due. However, it is rare for an account that is 90 days past due to be recovered, and therefore it is common practice within the industry to use 90 days past due as a conservative definition of default. This definition is also consistent in the literature (see, e.g., Glennon et al., 2008 and Khandani et al., 2010). We forecast all of our models over three different time horizons—two, three, and four quarters out—to classify whether or not an account becomes delinquent within those horizons.

Table 1
Model timing.
The first column represents the start and end dates of the training data. The test period columns show the quarter in which the models are tested. All models are meant to simulate a bank's actual forecasting problem as if they were at the test period start date.

Training period	Test period start		
Start–end	2Q forecast	3Q forecast	4Q forecast
2009Q4–2010Q4	2011Q2	2011Q3	2011Q4
2010Q2–2011Q2	2011Q4	2012Q1	2012Q2
2010Q4–2011Q4	2012Q2	2012Q3	2012Q4
2011Q2–2012Q2	2012Q4	2013Q1	2013Q2
2011Q4–2012Q4	2013Q2	2013Q3	N/A
2012Q2–2013Q2	2013Q4	N/A	N/A

3.3. Model timing

To predict delinquency, we estimate separate machine-learning models every 6 months, starting with the period ending 2010Q4.[11] We estimate these models at each point in time as if we were in that time period, i.e., no future data is ever used as inputs to a model, and we require a historical training period and a future testing period. For example, a model for 2010Q4 is trained on data up to and including 2010Q4, but no further. Table 1 defines the dates for the training and test samples of each of our models.

The optimal length of the training window involves a tradeoff between increasing the amount of training data available and the stationarity of the training data (hence its relevance for predicting future performance). We use a rolling window of 2 years as the length of the training window to balance these two considerations. In particular, we combine the data from the most recent quarter with the data from 12 months earlier to form a training sample. For example, the model trained on data ending in 2010Q4 contains the monthly credit-card accounts in 2009Q4 and 2010Q4. The average training sample thus contains about two million individual records, depending on the institution and the time period. In fact, these rolling windows incorporate up to 24 months of information each because of the lag structure of some of the variables (e.g., the year over year change in the HPI), and an additional 12-month period over which an account could become 90 days delinquent.

3.4. Measuring performance

The goal of our delinquency prediction models is to classify credit card accounts into two categories: accounts that become 90 days or more past due within the next n quarters ("bad" accounts), and accounts that do not ("good" accounts). Therefore, our measure of performance should reflect the accuracy with which our model classifies the accounts into these two categories.

One common way to measure performance of such binary classification models is to calculate precision and recall. In our model, precision is defined as the number of correctly predicted delinquent accounts divided by the predicted number of delinquent

[9] We refer to our variables as attributes, as is common in the machine-learning literature.

[10] Another reason for not differentiating across segments is that the results might reveal the identity of the banks to knowledgeable industry insiders. The same concern arises with the size of the portfolio.

[11] That is, we build models for the periods ending in 2010Q4, 2011Q2, 2011Q4, 2012Q2, 2012Q4, and 2013Q2. 2013Q2 is our last model because we need an out-of-sample test period to test our forecasts; it is used only for the two-quarter models.

		Model Prediction	
		Good	Bad
Actual Outcome	Good	True Positive (TP)	False Negative (FN)
	Bad	False Positive (FP)	True Negative (TN)

Precision = TN/(TN+FN)
Recall = TN/(TN+FP)
True Positive Rate = TP/(TP+FN)
False Positive Rate = FP/(FP+TN)
F-Measure = (2*Recall*Precision)/(Recall+Precision)
Kappa Statistic = $(P_a - P_e)/(1\text{-}P_e)$,
where P_a = (TP+TN)/N and P_e = [(TP+FN)/N]*[(TP+FN)/N]

Fig. 1. Performance statistics. The figure shows a sample confusion matrix and defines our performance statistics.

accounts, while recall is defined as the number of correctly predicted delinquent accounts divided by the actual number of delinquent accounts. Precision is meant to gauge the number of false positives (accounts predicted to be delinquent that stayed current) while recall gauges the number of false negatives (accounts predicted to stay current that actually went into default).

We also consider two statistics that combine precision and recall, the *F*-measure and the kappa statistic. The *F*-measure is defined as the harmonic mean of precision and recall, and assigns higher values to methods that achieve a reasonable balance between precision and recall. The kappa statistic measures performance relative to random classification, and can be thought of as the improvement over expected accuracy given the distribution of positive and negative examples. According to Khandani et al. (2010) and Landis and Koch (1977), a kappa statistic above 0.6 represents substantial performance. Fig. 1 summarizes the definitions of these classification performance statistics measures in a so-called "confusion matrix."

In the context of credit card portfolio risk management, however, there are account-specific costs and benefits associated with the classification decision that these performance statistics fail to capture. In the management of existing lines of credit, the primary benefit of classifying bad accounts before they become delinquent is to save the lender the run-up that is likely to occur between the current time period and the time at which the borrower goes into default. On the other hand, there are costs associated with incorrectly classifying accounts as well. For example, the bank may alienate customers and lose out on potential future business and profits on future purchases.

To account for these possible gains and losses, we use a cost-sensitive measure of performance to compute the value added of our classifier, as in Khandani et al. (2010), by assigning different costs to false positives and false negatives, and approximating the total savings that our models would have brought if they had been implemented. Our value added approach is able to assign a dollar-per-account savings (or cost) of implementing any classification model. From the lender's perspective, this provides an intuitive and practical method for choosing between models. From a supervisory perspective, we can assign deadweight costs of incorrect classifications by aggregate risk levels to quantify systemic risk levels.

Following Khandani et al. (2010), our value added function is derived from the confusion matrix. Ideally, we would like to achieve 100% true positives and true negatives, implying correct classification of all accounts, delinquent and current. However, any realistic classification will have some false positives and false negatives, which will incur costs.

To quantify the value added of classification, Khandani et al. (2010) define the profit with and without a forecast as follows:

Table 2
Sample description.
The table shows the total number of accounts over time. The six banks' data are combined to show the aggregate each quarter.

Date	Number of accounts (1000's)
2009Q4	5696
2010Q2	5677
2010Q4	5787
2011Q2	5960
2011Q4	5306
2012Q2	6300
2012Q4	6580
2013Q2	6643
2013Q4	6604

$$\Pi_{\text{no forecast}} = (TP + FN)B_C P_M - (FP + TN)B_D \tag{1}$$

$$\Pi_{\text{forecast}} = TP\ B_C P_M - FPB_D - TNB_C \tag{2}$$

$$\Delta\Pi_{\text{no forecast}} = TN(B_D - B_C) - FNB_C P_M \tag{3}$$

where B_C is the current account balance; B_D is the balance at default; P_M is the profitability margin; and *TP, FN, FP*, and *TN* are defined according to the confusion matrix. Note that Eq. (3) is broken down into a savings from lowering balances (the first term) less a cost of misclassification (the second term).

To generate a value added for each model, the authors then compare the savings from the forecast profit ($\Delta P_{\text{forecast}}$) with the benefit of perfect foresight. The savings from perfect foresight can be calculated by multiplying the total number of bad accounts ($TN+FP$) by the run up ($B_D - B_C$). The ratio of the model forecast savings (Eq. (3)) to the perfect foresight case can be written as:

$$\text{Value-Added}\left(\frac{B_D}{B_C}, r, N\right) = \frac{TN - FN\left[1-(1+r)^{-N}\right]\left[\frac{B_D}{B_C}-1\right]^{-1}}{TN+FP} \tag{4}$$

where we substitute $[1-(1+r)^{-N}]$ for the profitability margin, r is the discount rate, and N is the discount period.

4. Classification results

In this section, we report the results of our classification models by bank and by time. There are on average about 6.1 million accounts each month in our sample. Table 2 shows the sample sizes over time. There is a significant amount of heterogeneity in delinquencies across institutions and time (see Fig. 2). Delinquency rates necessarily increase with the forecast horizon, since the longer horizons include the shorter ones. Annual delinquency rates

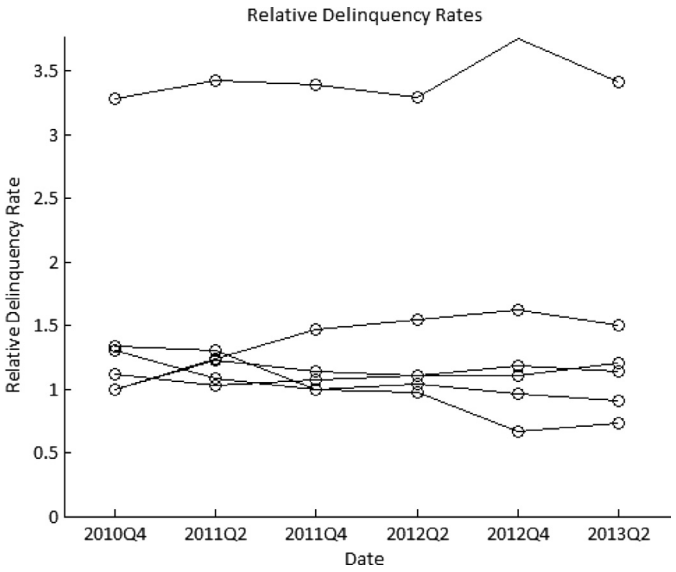

Fig. 2. Relative delinquency rates over time. The figure shows the relative delinquency rates over time. Due to data confidentiality restrictions, we do not report the actual delinquency rates over time. Each line represents an individual bank over time. The delinquency rates are all reported relative to the bank with the lowest two quarter delinquency rate in 2010Q4.

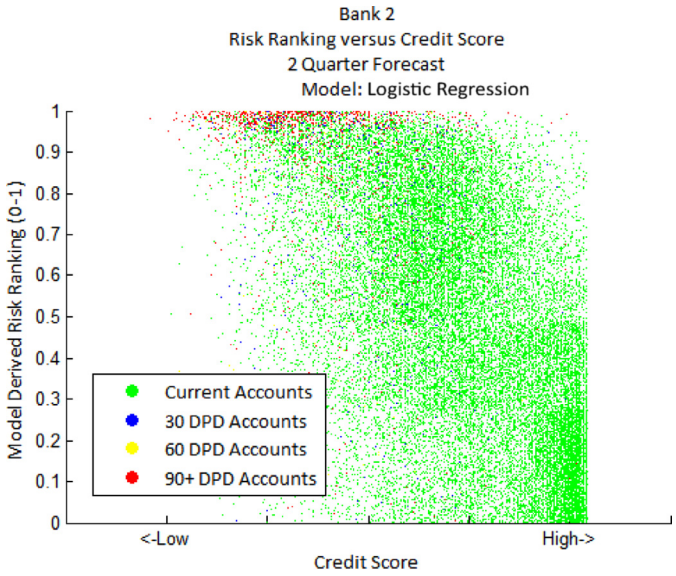

Fig. 3. Model risk ranking versus credit score. The figure plots the model-derived risk ranking versus an account's credit score at the time of the forecast for Bank 2. Accounts are rank-ordered based on a logistic regression model for a two-quarter forecast horizon. Green points are accounts that were current at the end of the forecast horizon; blue points are 30 days past due; yellow points are 60 days past due; and red points are 90+ days past due. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

across banks range from 1.36% to 4.36%, indicating that the institutions we are studying have very different underwriting and/or risk management strategies.

We run individual classification models for each bank over time; separate models are estimated for each forecast horizon for each bank. Because our data ends in 2014Q2, we can only test the three- and four-quarter-horizon models on the training periods ending in 2012Q2 and 2012Q4, respectively.[12]

4.1. Nonstationary environments

A fundamental concern for all prediction algorithms is generalization, i.e., whether models will continue to perform well on out-of-sample data. This is particularly important when the environment that generates the data is itself changing, and therefore the out-of-sample data is almost guaranteed to come from a different distribution than the training data. This concern is particularly relevant for financial forecasting, given the non-stationarity of financial data as well as the macroeconomic and regulatory environments. Our sample period, which starts on the heels of the 2008 financial crisis and the ensuing recession, only heightens these concerns.

We address overfitting primarily by testing out-of-sample. Our decision tree models also allow us to control the degree of in-sample fitting by controlling what is known as the pruning parameter, which we refer to as M. This parameter acts as the stopping criterion for the decision tree algorithm. For example, when $M=2$, the algorithm will continue to attempt to add additional nodes to the leaves of the tree until there are two instances (accounts) or less on each leaf, and an additional node would be statistically significant. As M increases, the in-sample performance will degrade, because the algorithm stops even though there may be potentially statistically significant splits remaining. However, the out-of-sample performance may actually increase for a while because the nodes blocked by an increasing M are overfitting the sample. Eventually, however, even the out-of-sample performance degrades, as M becomes sufficiently high.

To find a suitable value of M for our machine-learning models, we use data from a selected bank for validation. We test the performance for a set of possible M parameters between 2 and 5000 for 15 different "clusters" of parameters used to calculate the value-added (run-up ratios, discount rates, etc.). We found that setting $M=50$ led to the best performance overall across clusters. Further, the results were not very sensitive for values of M between 25 and 250, indicating that the estimates and performance should be robust with respect to this parameter setting. Sensitivity analysis for the other banks around $M=50$ yielded similar results, and in light of these, we use a pruning parameter of $M=50$ in all of our decision tree models.

4.2. Model results

In this section, we show the results of the comparison of our three modeling approaches: decision trees, logistic regression, and random forests. The random forest models are estimated with 20 random trees.[13]

To preview the results, and to help visualize the effectiveness of our models in discriminating between good and bad accounts, we plot the model-derived risk ranking versus an account's credit score at the time of the forecast in Fig. 3 for Bank 2. Accounts are rank-ordered based on a logistic regression model for a two-quarter forecast horizon. Green points represent accounts that were current at the end of the forecast horizon; blue points represent accounts 30 days past due; yellow points represent

[12] For example, for the four-quarter forecast models with training data ending 2012Q2, the dependent variable is defined over the period 2012Q2 through 2013Q2, making the test date 2013Q2. We then need one year of data to test the model out-of-sample, which brings us to our last month of data coverage in 2014Q2.

[13] The C4.5 models produced unreliable results for the 4Q forecast horizon for Bank 5 due to a low delinquency rate combined with accounts that were difficult to classify (the corresponding logistic and random forest forecasts were the worst performing models). The random forest models for the 4Q forecast horizon for Bank 2 failed to converge in a reasonable amount of time (run-time was stopped after 24+ hours at full capacity), so those results are omitted as well. Throughout the paper, those results are indicated with N/A.

Table 3
Precision, recall, true positive rate, and false positive rates.
The table shows the precision, recall, true positive rate, and false positive rate by bank, time, and forecast horizon for each model type. The statistics are defined in Fig. 1. The acceptance threshold is defined as the threshold which maximizes the *F*-measure.

Bank	Test date	C4.5 Decision trees				Logistic regression				Random forests			
		Precision	Recall	True positive rate	False positive rate	Precision	Recall	True positive rate	False positive rate	Precision	Recall	True positive rate	False positive rate
Panel A: Two-quarter forecast horizon													
1	201106	71.3%	63.0%	99.9%	37.0%	17.9%	59.1%	99.0%	40.9%	68.8%	67.8%	99.9%	32.2%
1	201112	62.8%	70.3%	99.8%	29.7%	26.0%	70.2%	98.8%	29.8%	65.0%	68.3%	99.8%	31.7%
1	201206	65.5%	67.8%	99.8%	32.2%	62.7%	60.0%	99.8%	40.0%	64.2%	69.1%	99.8%	30.9%
1	201212	68.0%	65.3%	99.8%	34.7%	62.6%	62.1%	99.8%	37.9%	66.2%	67.3%	99.8%	32.7%
1	201306	68.2%	59.9%	99.9%	40.1%	58.6%	59.3%	99.8%	40.7%	58.3%	70.1%	99.7%	29.9%
1	201312	67.1%	65.6%	99.8%	34.4%	60.6%	64.5%	99.8%	35.5%	64.5%	69.4%	99.8%	30.6%
Average:		**67.2%**	**65.3%**	**99.8%**	**34.7%**	**48.1%**	**62.5%**	**99.5%**	**37.5%**	**64.5%**	**68.7%**	**99.8%**	**31.3%**
2	201106	63.7%	73.0%	99.4%	27.0%	64.2%	71.5%	99.4%	28.5%	65.9%	71.1%	99.4%	28.9%
2	201112	60.5%	75.9%	99.2%	24.1%	61.9%	71.3%	99.3%	28.7%	60.5%	74.2%	99.2%	25.8%
2	201206	64.8%	63.5%	99.4%	36.5%	3.1%	91.8%	53.9%	8.2%	63.4%	71.2%	99.3%	28.8%
2	201212	65.7%	70.7%	99.4%	29.3%	10.0%	67.7%	90.4%	32.3%	62.0%	73.9%	99.3%	26.1%
2	201306	66.5%	66.8%	99.5%	33.2%	63.6%	68.6%	99.4%	31.4%	61.7%	72.3%	99.3%	27.7%
2	201312	63.2%	73.0%	99.3%	27.0%	62.7%	71.2%	99.3%	28.8%	60.8%	72.6%	99.2%	27.4%
Average:		**64.1%**	**70.5%**	**99.4%**	**29.5%**	**44.3%**	**73.7%**	**90.3%**	**26.3%**	**62.4%**	**72.5%**	**99.3%**	**27.5%**
3	201106	79.9%	88.8%	99.9%	11.2%	75.7%	81.2%	99.8%	18.8%	80.0%	87.7%	99.9%	12.3%
3	201112	69.2%	92.6%	99.7%	7.4%	72.5%	82.4%	99.8%	17.6%	80.5%	85.6%	99.9%	14.4%
3	201206	81.1%	84.9%	99.9%	15.1%	73.6%	81.7%	99.9%	18.3%	83.9%	79.0%	99.9%	21.0%
3	201212	79.5%	85.4%	99.9%	14.6%	72.4%	79.3%	99.9%	20.7%	79.0%	85.5%	99.9%	14.5%
3	201306	71.6%	90.2%	99.9%	9.8%	70.8%	80.3%	99.9%	19.7%	70.6%	90.8%	99.9%	9.2%
3	201312	74.8%	88.6%	99.9%	11.4%	70.7%	84.2%	99.9%	15.8%	70.8%	90.3%	99.9%	9.7%
Average:		**76.0%**	**88.4%**	**99.9%**	**11.6%**	**72.6%**	**81.5%**	**99.9%**	**18.5%**	**77.5%**	**86.5%**	**99.9%**	**13.5%**
4	201106	59.4%	64.9%	99.7%	35.1%	57.2%	62.3%	99.7%	37.7%	58.7%	67.2%	99.7%	32.8%
4	201112	61.2%	70.0%	99.8%	30.0%	53.1%	67.1%	99.7%	32.9%	62.4%	67.3%	99.8%	32.7%
4	201206	67.4%	59.0%	99.9%	41.0%	57.6%	59.3%	99.8%	40.7%	59.0%	64.6%	99.8%	35.4%
4	201212	68.6%	60.5%	99.9%	39.5%	59.0%	62.1%	99.8%	37.9%	64.0%	62.1%	99.8%	37.9%
4	201306	62.3%	65.1%	99.8%	34.9%	61.5%	61.3%	99.8%	38.7%	61.3%	66.9%	99.8%	33.1%
4	201312	68.9%	60.7%	99.9%	39.3%	57.5%	67.1%	99.8%	32.9%	64.6%	65.6%	99.9%	34.4%
Average:		**64.6%**	**63.4%**	**99.8%**	**36.6%**	**57.7%**	**63.2%**	**99.8%**	**36.8%**	**61.7%**	**65.6%**	**99.8%**	**34.4%**
5	201106	69.6%	72.8%	99.8%	27.2%	64.5%	71.8%	99.8%	28.2%	67.2%	76.0%	99.8%	24.0%
5	201112	66.1%	72.8%	99.8%	27.2%	65.7%	69.0%	99.8%	31.0%	64.1%	76.4%	99.8%	23.6%
5	201206	70.7%	64.4%	99.9%	35.6%	66.3%	62.2%	99.8%	37.8%	65.6%	72.5%	99.8%	27.5%
5	201212	66.2%	75.4%	99.8%	24.6%	63.5%	72.7%	99.8%	27.3%	66.1%	74.5%	99.8%	25.5%
5	201306	68.4%	71.0%	99.8%	29.0%	68.0%	68.8%	99.8%	31.2%	66.9%	75.4%	99.8%	24.6%
5	201312	63.3%	77.5%	99.7%	22.5%	66.6%	70.4%	99.8%	29.6%	64.3%	75.2%	99.8%	24.8%
Average:		**67.4%**	**72.3%**	**99.8%**	**27.7%**	**65.7%**	**69.1%**	**99.8%**	**30.9%**	**65.7%**	**75.0%**	**99.8%**	**25.0%**
6	201106	69.7%	66.5%	99.9%	33.5%	64.6%	66.4%	99.8%	33.6%	69.9%	65.9%	99.9%	34.1%
6	201112	64.0%	71.1%	99.8%	28.9%	66.0%	66.9%	99.8%	33.1%	64.5%	70.6%	99.8%	29.4%
6	201206	74.7%	67.6%	99.8%	32.4%	69.9%	71.2%	99.8%	28.8%	70.9%	71.4%	99.8%	28.6%
6	201212	42.8%	90.4%	99.1%	9.6%	67.9%	70.2%	99.8%	29.8%	66.4%	72.9%	99.7%	27.1%
6	201306	36.2%	96.2%	98.7%	3.8%	70.6%	69.5%	99.8%	30.5%	71.7%	70.2%	99.8%	29.8%
6	201312	62.8%	72.5%	99.7%	27.5%	60.9%	72.3%	99.7%	27.7%	61.8%	71.7%	99.7%	28.3%
Average:		**58.4%**	**77.4%**	**99.5%**	**22.6%**	**66.7%**	**69.4%**	**99.8%**	**30.6%**	**67.5%**	**70.4%**	**99.8%**	**29.6%**
Panel B: Three-quarter forecast horizon													
1	201109	60.3%	45.8%	99.7%	54.2%	55.8%	42.0%	99.7%	58.0%	56.0%	50.0%	99.7%	50.0%
1	201203	59.0%	44.5%	99.7%	55.5%	54.5%	39.3%	99.7%	60.7%	56.3%	46.1%	99.7%	53.9%
1	201209	53.8%	47.6%	99.6%	52.4%	52.4%	40.1%	99.6%	59.9%	53.4%	47.4%	99.6%	52.6%
1	201303	55.6%	43.3%	99.7%	56.7%	52.0%	37.7%	99.7%	62.3%	49.5%	45.4%	99.6%	54.6%
1	201309	36.9%	54.0%	99.1%	46.0%	54.4%	35.4%	99.7%	64.6%	55.7%	44.1%	99.6%	55.9%
Average:		**53.1%**	**47.0%**	**99.6%**	**53.0%**	**53.8%**	**38.9%**	**99.7%**	**61.1%**	**54.2%**	**46.6%**	**99.6%**	**53.4%**
2	201109	52.3%	51.5%	98.5%	48.5%	54.7%	45.9%	98.8%	54.1%	55.7%	48.1%	98.8%	51.9%
2	201203	55.2%	42.5%	98.9%	57.5%	46.8%	48.8%	98.3%	51.2%	48.9%	47.6%	98.5%	52.4%
2	201209	47.6%	56.0%	98.1%	44.0%	5.0%	80.4%	52.2%	19.6%	50.3%	52.0%	98.4%	48.0%
2	201303	51.1%	45.2%	98.9%	54.8%	9.3%	49.6%	87.8%	50.4%	N/A	N/A	N/A	N/A
2	201309	50.8%	50.8%	98.4%	49.2%	48.3%	50.9%	98.2%	49.1%	N/A	N/A	N/A	N/A
Average:		**51.4%**	**49.2%**	**98.6%**	**50.8%**	**32.8%**	**55.1%**	**87.0%**	**44.9%**	**51.7%**	**49.3%**	**98.6%**	**50.7%**
3	201109	70.1%	56.4%	99.7%	43.6%	64.7%	51.8%	99.6%	48.2%	66.8%	57.8%	99.6%	42.2%
3	201203	70.6%	55.4%	99.8%	44.6%	65.2%	52.9%	99.7%	47.1%	71.2%	55.3%	99.8%	44.7%
3	201209	67.4%	56.8%	99.7%	43.2%	66.3%	53.1%	99.7%	46.9%	70.8%	55.8%	99.8%	44.2%
3	201303	66.7%	60.3%	99.8%	39.7%	64.8%	55.1%	99.8%	44.9%	69.4%	58.1%	99.8%	41.9%
3	201309	72.8%	60.8%	99.8%	39.2%	64.1%	58.9%	99.7%	41.1%	65.7%	63.6%	99.7%	36.4%
Average:		**69.5%**	**58.0%**	**99.8%**	**42.0%**	**65.0%**	**54.4%**	**99.7%**	**45.6%**	**68.8%**	**58.1%**	**99.7%**	**41.9%**

(*continued on next page*)

Table 3 (*continued*)

Bank	Test date	C4.5 Decision trees				Logistic regression				Random forests			
		Precision	Recall	True positive rate	False positive rate	Precision	Recall	True positive rate	False positive rate	Precision	Recall	True positive rate	False positive rate
4	201109	46.1%	48.7%	99.4%	51.3%	46.7%	43.2%	99.5%	56.8%	52.0%	44.3%	99.5%	55.7%
4	201203	25.2%	56.2%	98.5%	43.8%	46.0%	41.0%	99.6%	59.0%	52.9%	42.5%	99.7%	57.5%
4	201209	53.4%	39.6%	99.7%	60.4%	43.8%	43.8%	99.5%	56.2%	47.3%	44.2%	99.5%	55.8%
4	201303	51.3%	38.9%	99.7%	61.1%	48.5%	37.2%	99.7%	62.8%	45.4%	43.4%	99.6%	56.6%
4	201309	46.3%	46.8%	99.5%	53.2%	44.7%	47.4%	99.5%	52.6%	54.4%	43.5%	99.7%	56.5%
Average:		**44.5%**	**46.0%**	**99.4%**	**54.0%**	**46.0%**	**42.5%**	**99.5%**	**57.5%**	**50.4%**	**43.6%**	**99.6%**	**56.4%**
5	201109	30.6%	43.8%	99.2%	56.2%	30.2%	34.6%	99.3%	65.4%	40.0%	36.5%	99.5%	63.5%
5	201203	39.9%	31.2%	99.6%	68.8%	28.8%	32.4%	99.4%	67.6%	36.1%	37.1%	99.5%	62.9%
5	201209	40.4%	33.7%	99.6%	66.3%	22.9%	46.6%	98.7%	53.4%	39.3%	35.8%	99.5%	64.2%
5	201303	41.0%	31.2%	99.7%	68.8%	27.1%	37.4%	99.2%	62.6%	38.9%	34.5%	99.6%	65.5%
5	201309	42.1%	34.6%	99.6%	65.4%	32.6%	31.4%	99.4%	68.6%	42.2%	36.1%	99.6%	63.9%
Average:		**38.8%**	**34.9%**	**99.5%**	**65.1%**	**28.3%**	**36.5%**	**99.2%**	**63.5%**	**39.3%**	**36.0%**	**99.5%**	**64.0%**
6	201109	48.0%	46.0%	99.4%	54.0%	48.3%	39.9%	99.5%	60.1%	56.0%	42.5%	99.6%	57.5%
6	201203	52.9%	43.3%	99.5%	56.7%	47.8%	42.0%	99.4%	58.0%	53.5%	45.3%	99.5%	54.7%
6	201209	42.9%	55.9%	98.9%	44.1%	52.1%	48.4%	99.4%	51.6%	58.2%	51.0%	99.5%	49.0%
6	201303	58.3%	42.8%	99.6%	57.2%	54.2%	43.2%	99.5%	56.8%	59.3%	44.4%	99.6%	55.6%
6	201309	47.7%	51.1%	99.2%	48.9%	48.6%	50.5%	99.3%	49.5%	54.1%	49.1%	99.4%	50.9%
Average:		**50.0%**	**47.8%**	**99.3%**	**52.2%**	**50.2%**	**44.8%**	**99.4%**	**55.2%**	**56.2%**	**46.4%**	**99.5%**	**53.6%**
Panel C: Four-quarter forecast horizon													
1	201112	52.5%	38.9%	99.5%	61.1%	26.6%	38.2%	98.5%	61.8%	48.5%	42.1%	99.4%	57.9%
1	201206	54.5%	36.5%	99.6%	63.5%	44.5%	35.5%	99.4%	64.5%	50.3%	39.2%	99.4%	60.8%
1	201212	49.2%	39.0%	99.5%	61.0%	45.0%	34.8%	99.5%	65.2%	48.9%	40.4%	99.5%	59.6%
1	201306	53.8%	34.4%	99.6%	65.6%	47.5%	29.1%	99.6%	70.9%	48.9%	35.3%	99.5%	64.7%
Average:		**50.5%**	**41.0%**	**99.6%**	**59.0%**	**41.1%**	**36.3%**	**99.3%**	**63.7%**	**49.9%**	**41.3%**	**99.5%**	**58.7%**
2	201112	47.3%	43.1%	98.0%	56.9%	42.1%	47.7%	97.2%	52.3%	N/A	N/A	N/A	N/A
2	201206	53.6%	40.8%	98.5%	59.2%	6.6%	86.9%	46.2%	13.1%	N/A	N/A	N/A	N/A
2	201212	47.1%	43.6%	98.2%	56.4%	5.6%	84.3%	48.4%	15.7%	N/A	N/A	N/A	N/A
2	201306	51.0%	39.6%	98.6%	60.4%	12.9%	51.6%	86.9%	48.4%	N/A	N/A	N/A	N/A
Average:		**48.3%**	**46.2%**	**98.4%**	**53.8%**	**22.5%**	**64.7%**	**75.6%**	**35.3%**	**N/A**	**N/A**	**N/A**	**N/A**
3	201112	63.1%	47.8%	99.6%	52.2%	62.0%	43.9%	99.6%	56.1%	64.2%	47.6%	99.6%	52.4%
3	201206	63.5%	41.9%	99.7%	58.1%	58.2%	41.3%	99.6%	58.7%	68.5%	40.0%	99.7%	60.0%
3	201212	57.9%	44.3%	99.6%	55.7%	49.6%	40.8%	99.5%	59.2%	57.3%	46.0%	99.6%	54.0%
3	201306	60.8%	43.9%	99.7%	56.1%	53.9%	44.3%	99.6%	55.7%	63.5%	42.7%	99.7%	57.3%
Average:		**63.1%**	**47.8%**	**99.7%**	**52.2%**	**56.7%**	**45.8%**	**99.6%**	**54.2%**	**64.8%**	**47.2%**	**99.7%**	**52.8%**
4	201112	38.8%	38.8%	99.2%	61.2%	37.0%	38.8%	99.1%	61.2%	44.3%	36.0%	99.4%	64.0%
4	201206	38.2%	37.8%	99.2%	62.2%	39.3%	33.6%	99.3%	66.4%	44.4%	33.4%	99.4%	66.6%
4	201212	42.9%	36.8%	99.4%	63.2%	40.2%	36.2%	99.3%	63.8%	40.6%	37.4%	99.3%	62.6%
4	201306	26.3%	43.4%	98.4%	56.6%	42.5%	34.7%	99.4%	65.3%	45.3%	36.4%	99.4%	63.6%
Average:		**39.0%**	**39.7%**	**99.2%**	**60.3%**	**40.0%**	**36.5%**	**99.4%**	**63.5%**	**43.9%**	**37.5%**	**99.5%**	**62.5%**
5	201112	N/A	N/A	N/A	N/A	9.1%	31.2%	97.7%	68.8%	9.5%	24.7%	98.3%	75.3%
5	201206	N/A	N/A	N/A	N/A	8.9%	9.8%	99.2%	90.2%	11.8%	16.6%	99.0%	83.4%
5	201212	N/A	N/A	N/A	N/A	9.7%	25.8%	98.2%	74.2%	10.8%	22.0%	98.6%	78.0%
5	201306	N/A	N/A	N/A	N/A	8.9%	33.9%	97.0%	66.1%	10.9%	24.0%	98.3%	76.0%
Average:		**N/A**	**N/A**	**N/A**	**N/A**	**12.8%**	**24.5%**	**98.3%**	**75.5%**	**11.2%**	**24.9%**	**98.5%**	**75.1%**
6	201112	49.9%	36.0%	99.4%	64.0%	48.1%	30.7%	99.4%	69.3%	47.0%	36.8%	99.3%	63.2%
6	201206	55.7%	37.6%	99.4%	62.4%	45.0%	38.8%	99.0%	61.2%	52.3%	40.9%	99.2%	59.1%
6	201212	38.9%	46.0%	98.6%	54.0%	54.0%	37.5%	99.4%	62.5%	49.5%	45.1%	99.1%	54.9%
6	201306	52.9%	40.9%	99.3%	59.1%	54.0%	40.8%	99.3%	59.2%	52.2%	44.2%	99.2%	55.8%
Average:		**48.6%**	**43.7%**	**99.2%**	**56.3%**	**49.8%**	**40.8%**	**99.3%**	**59.2%**	**49.8%**	**44.9%**	**99.3%**	**55.1%**

accounts 60 days past due; and red points represent accounts 90 days or more past due. We plot each account's credit bureau score on the horizontal axis because it is a key variable used in virtually every consumer default prediction model and serves as a useful comparison to the machine-learning forecast.

This plot shows that while credit scores discriminate between good and bad accounts to a certain degree (the red 90+ days past due accounts do tend to cluster to the left region of the horizontal axis with lower credit scores), even the logistic regression model is very effective in rank-ordering accounts in terms of riskiness. In particular, the red 90+ days past due points cluster heavily at the top of the graph, implying that machine-learning forecasts are highly effective in identifying accounts that eventually become delinquent.[14]

Table 3 shows the precision and recall for our models. We also provide the true positive and false positive rates. The results are given by bank, time, and forecast horizon for each model type. The statistics are calculated for the classification threshold that maximizes the respective model's *F*-measure to provide a reasonable balance between good precision and recall.

[14] Analogous plots for our C4.5 decision tree and random forest models look very similar.

Although selecting a modeling threshold based on the test data does introduce some look-ahead bias, we use this approach when presenting the results for two reasons. First, banks are likely to calibrate classification models using an expected delinquency rate to select the acceptance threshold. We do not separately model delinquency rates, and view the primary purpose of our classifiers as the rank-ordering of accounts. To this end, we are less concerned with forecasting the realized delinquency rates than rank-ordering accounts based on risk of delinquency. Therefore, the main role of the acceptance threshold for our purposes is for exposition and to make fair comparisons across models.

Second, the performance statistics we report—the *F*-measure and the kappa statistic—are relatively insensitive to the choice of modeling threshold. Figs. A1–A3 in Appendix A show the sensitivity of these performance statistics to the choice of acceptance threshold for the C4.5 decision tree, logistic regression, and random forest models, respectively. The three plots on the left in each figure show the *F*-measure versus the acceptance threshold, while the plots on the right show the kappa statistic.

There are a few noteworthy points here. First, for each bank, the optimal threshold remains relatively constant over time, which means that it should be easy for a bank to select a threshold based on past results and get an adequate forecast. Second, in the cases where the selected threshold varies over time, the lines are still quite flat. For example, in our C4.5 decision tree models in Fig. A1, the optimal thresholds cluster by bank and the curves are very flat between 20% and 70% of the threshold values for the *F*-measure and the kappa statistics. For the random forest models in Fig. A3, the lines are not quite as flat, but the optimal thresholds tend to cluster tightly for each bank. In sum, it is important to remember that the goal of a bank would not be to maximize the *F*-measure in any case, and as long as the selected threshold is selected using any reasonable strategy, our sensitivity analysis demonstrates that it would, in all likelihood, only have a minimal effect on our main results.

Each of the models achieves a very high true positive rate, which is not surprising given the low default rates. The false positive rates are reasonable, between 11% and 38% for the two-quarter

Table 4
F-measure and kappa statistic by bank and time.
The table shows the *F*-measure and kappa statistic results by bank, time, and forecast horizon for each model type. The statistics are based on the acceptance threshold that maximizes the respective statistic for a given bank-time-model combination. Panel A shows the *F*-measure and Panel B shows the kappa statistic.

Panel A:		2Q forecast			3Q forecast			4Q forecast		
Bank	Test date	C4.5 tree	Logistic regression	Random forest	C4.5 tree	Logistic regression	Random forest	C4.5 tree	Logistic regression	Random forest
1	201106	66.9%	27.5%	68.3%	52.0%	47.9%	52.8%	44.7%	31.4%	45.1%
1	201112	66.3%	37.9%	66.6%	50.8%	45.7%	50.7%	43.7%	39.5%	44.1%
1	201206	66.6%	61.3%	66.5%	50.5%	45.5%	50.2%	43.5%	39.2%	44.3%
1	201212	66.7%	62.3%	66.7%	48.7%	43.7%	47.4%	41.9%	36.1%	41.0%
1	201306	63.8%	58.9%	63.6%	43.8%	42.9%	49.2%	N/A	N/A	N/A
1	201312	66.4%	62.5%	66.9%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**66.1%**	**51.7%**	**66.4%**	**49.2%**	**45.1%**	**50.1%**	**43.5%**	**36.5%**	**43.6%**
2	201106	68.0%	67.7%	68.4%	51.9%	49.9%	51.6%	45.1%	44.7%	N/A
2	201112	67.3%	66.3%	66.7%	48.1%	47.8%	48.3%	46.3%	12.3%	N/A
2	201206	64.2%	6.0%	67.1%	51.5%	9.3%	51.1%	45.3%	10.6%	N/A
2	201212	68.1%	17.4%	67.4%	48.0%	15.6%	N/A	44.6%	20.7%	N/A
2	201306	66.6%	66.0%	66.6%	50.8%	49.6%	N/A	N/A	N/A	N/A
2	201312	67.8%	66.7%	66.2%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**67.0%**	**48.3%**	**67.0%**	**50.0%**	**34.4%**	**50.3%**	**45.3%**	**22.1%**	**N/A**
3	201106	84.1%	78.4%	83.7%	62.5%	57.5%	61.9%	54.4%	51.4%	54.7%
3	201112	79.2%	77.1%	83.0%	62.1%	58.4%	62.2%	50.5%	48.3%	50.5%
3	201206	82.9%	77.5%	81.4%	61.7%	59.0%	62.4%	50.2%	44.8%	51.0%
3	201212	82.3%	75.7%	82.1%	63.4%	59.5%	63.2%	51.0%	48.6%	51.1%
3	201306	79.8%	75.3%	79.4%	66.3%	61.4%	64.6%	N/A	N/A	N/A
3	201312	81.1%	76.9%	79.4%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**81.6%**	**76.8%**	**81.5%**	**63.2%**	**59.2%**	**62.9%**	**51.5%**	**48.3%**	**51.8%**
4	201106	62.1%	59.6%	62.7%	47.3%	44.9%	47.9%	38.8%	37.8%	39.7%
4	201112	65.3%	59.3%	64.7%	34.8%	43.4%	47.1%	38.0%	36.2%	38.1%
4	201206	62.9%	58.4%	61.7%	45.4%	43.8%	45.7%	39.6%	38.1%	38.9%
4	201212	64.3%	60.5%	63.0%	44.2%	42.1%	44.4%	32.7%	38.2%	40.4%
4	201306	63.6%	61.4%	64.0%	46.6%	46.0%	48.3%	N/A	N/A	N/A
4	201312	64.6%	62.0%	65.1%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**63.8%**	**60.2%**	**63.5%**	**43.7%**	**44.0%**	**46.7%**	**37.3%**	**37.6%**	**39.3%**
5	201106	71.2%	67.9%	71.3%	36.0%	32.3%	38.2%	N/A	14.1%	13.8%
5	201112	69.3%	67.3%	69.8%	35.0%	30.5%	36.6%	N/A	9.3%	13.8%
5	201206	67.4%	64.2%	68.9%	36.8%	30.7%	37.5%	N/A	14.1%	14.5%
5	201212	70.5%	67.8%	70.0%	35.5%	31.4%	36.6%	N/A	14.1%	15.0%
5	201306	69.7%	68.4%	70.9%	38.0%	32.0%	38.9%	N/A	N/A	N/A
5	201312	69.7%	68.4%	69.3%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**69.6%**	**67.3%**	**70.0%**	**36.2%**	**31.4%**	**37.6%**	**N/A**	**12.9%**	**14.3%**
6	201106	68.0%	65.5%	67.8%	47.0%	43.7%	48.3%	41.8%	37.5%	41.3%
6	201112	67.4%	66.5%	67.4%	47.6%	44.7%	49.1%	44.9%	41.7%	45.9%
6	201206	71.0%	70.5%	71.1%	48.6%	50.1%	54.3%	42.1%	44.3%	47.2%
6	201212	58.1%	69.0%	69.5%	49.4%	48.1%	50.8%	46.1%	46.5%	47.9%
6	201306	52.6%	70.0%	70.9%	49.3%	49.6%	51.4%	N/A	N/A	N/A
6	201312	67.3%	66.1%	66.4%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**64.1%**	**67.9%**	**68.9%**	**48.4%**	**47.2%**	**50.8%**	**43.7%**	**42.5%**	**45.6%**

(continued on next page)

Table 4
(continued)

Panel B:		2Q forecast			3Q forecast			4Q forecast		
Bank	Test date	C4.5 tree	Logistic regression	Random forest	C4.5 tree	Logistic regression	Random forest	C4.5 tree	Logistic regression	Random forest
1	201106	69.8%	49.9%	70.0%	60.9%	59.1%	61.1%	57.1%	49.6%	57.3%
1	201112	68.1%	49.9%	68.4%	59.8%	57.9%	60.0%	56.8%	55.8%	56.9%
1	201206	68.7%	65.0%	68.8%	59.2%	57.9%	59.9%	56.9%	55.4%	57.4%
1	201212	68.3%	64.4%	68.7%	58.6%	56.1%	58.0%	56.2%	54.0%	56.2%
1	201306	67.3%	61.9%	66.2%	30.5%	56.2%	59.1%	N/A	N/A	N/A
1	201312	68.4%	65.7%	68.7%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**68.4%**	**59.5%**	**68.5%**	**53.8%**	**57.5%**	**59.6%**	**56.7%**	**53.7%**	**56.9%**
2	201106	69.2%	68.7%	69.1%	59.2%	58.7%	59.0%	55.2%	55.0%	N/A
2	201112	68.0%	66.9%	67.2%	57.8%	56.9%	57.0%	53.6%	48.9%	N/A
2	201206	67.9%	50.0%	67.9%	58.9%	49.2%	58.2%	55.5%	49.1%	N/A
2	201212	68.1%	49.6%	67.5%	57.3%	49.4%	N/A	55.5%	49.1%	N/A
2	201306	67.3%	66.2%	66.8%	56.9%	56.9%	N/A	N/A	N/A	N/A
2	201312	67.9%	66.9%	66.3%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**68.1%**	**61.3%**	**67.5%**	**58.0%**	**54.2%**	**58.1%**	**54.9%**	**50.5%**	**N/A**
3	201106	83.5%	78.0%	83.2%	67.4%	64.2%	67.1%	62.7%	61.0%	63.4%
3	201112	75.6%	76.2%	82.4%	67.6%	64.4%	67.7%	61.7%	60.0%	62.1%
3	201206	82.6%	77.0%	81.9%	67.3%	65.0%	67.7%	60.2%	56.7%	60.6%
3	201212	81.8%	75.3%	81.5%	68.1%	62.4%	68.0%	61.7%	59.9%	61.7%
3	201306	77.8%	74.6%	77.6%	69.9%	65.6%	65.5%	N/A	N/A	N/A
3	201312	79.3%	75.6%	77.4%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**80.1%**	**76.1%**	**80.7%**	**68.1%**	**64.3%**	**67.2%**	**61.6%**	**59.4%**	**62.0%**
4	201106	65.3%	62.9%	65.1%	57.3%	55.6%	57.3%	54.8%	53.9%	55.2%
4	201112	66.4%	63.2%	66.6%	−5.9%	55.8%	57.9%	53.9%	52.9%	54.3%
4	201206	66.3%	63.5%	65.2%	56.6%	55.7%	56.8%	54.4%	53.5%	54.4%
4	201212	66.9%	62.8%	66.0%	56.7%	54.9%	56.2%	10.2%	53.7%	54.7%
4	201306	67.7%	63.4%	66.0%	58.1%	56.1%	57.7%	N/A	N/A	N/A
4	201312	67.6%	64.2%	66.3%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**66.7%**	**63.3%**	**65.9%**	**44.6%**	**55.6%**	**57.2%**	**43.3%**	**53.5%**	**54.6%**
5	201106	72.0%	68.0%	71.6%	21.9%	49.8%	53.6%	N/A	49.8%	49.8%
5	201112	70.0%	67.7%	70.6%	53.3%	49.8%	52.9%	N/A	49.8%	49.8%
5	201206	69.6%	67.2%	70.3%	52.2%	49.8%	52.6%	N/A	49.7%	49.8%
5	201212	70.2%	67.6%	70.0%	52.4%	49.8%	52.7%	N/A	49.8%	49.8%
5	201306	70.4%	69.3%	70.7%	52.7%	49.8%	53.4%	N/A	N/A	N/A
5	201312	69.9%	68.7%	70.0%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**70.4%**	**68.1%**	**70.5%**	**46.5%**	**49.8%**	**53.0%**	**N/A**	**49.8%**	**49.8%**
6	201106	68.7%	67.8%	69.7%	47.7%	57.9%	58.1%	55.8%	55.1%	55.7%
6	201112	67.8%	68.3%	68.2%	52.1%	58.1%	59.5%	53.4%	56.9%	57.4%
6	201206	72.3%	72.1%	72.0%	40.4%	60.2%	61.0%	36.3%	56.9%	57.0%
6	201212	34.7%	69.6%	69.9%	59.2%	58.8%	59.4%	51.9%	58.0%	57.6%
6	201306	13.0%	71.1%	72.2%	47.2%	57.5%	57.5%	N/A	N/A	N/A
6	201312	66.4%	64.4%	66.3%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**53.8%**	**68.9%**	**69.7%**	**49.3%**	**58.5%**	**59.1%**	**49.4%**	**56.7%**	**56.9%**

horizon models. However, as the forecast horizon increases, the models become less accurate and the false positive rates increase for each bank.

Table 4 presents the *F*-measure and kappa statistics by bank and by time in Panels A and B, respectively. As mentioned above, the *F*-measure and kappa statistics show that the C4.5 and random forest models outperform the logistic regression models. The performance of the models declines as the forecast horizon increases. The C4.5 and random forest models tend to consistently outperform the logistic regression models, regardless of the forecast horizon, for each statistic.

Table 5 presents the value added for each of the models, which represents the potential gain from employing a given model versus passive risk management. Under this metric, the results are similar in that the C4.5 and random forest models outperform logistic regression. All the value added results assume a run-up of 30% and a profitability margin of about 13.5%.

For the two-quarter forecast horizon, the C4.5 models produce an average per bank cost savings of between 45.2% and 75.5%. The random forest models yield similar values, between 47.0% and 74.4%. The logistic regressions fare much worse based on the bank average values because Banks 1 and 2 show two periods of negative value added—meaning that the models did such a poor job of classifying accounts that the bank would have been better off not managing accounts at all. Even omitting these negative instances, however, the logistic models tend to underperform the others.

Random forests are considered state-of-the-art in terms of out-of-sample prediction performance in classification tasks like the one considered here. It is possible that using more bagged samples would further improve their performance, but given that their economic benefit in performance (in terms of value added) over the more easily interpretable single decision trees seems limited, the single decision tree model may be a preferred alternative for this domain.

There is substantial heterogeneity in value added across banks as well. Fig. 4 plots the value added for all six banks for each model type. All models are based on a two-quarter forecast horizon. Bank 3 is always at the top of the plots, meaning that it performs best with our models. Bank 4 tends to be the lowest (although it still has a positive value added), and the other four banks cluster in between.

Table 5
Value added by bank and time.
The table shows the value added results by bank, time, and forecast horizon for each model type. The statistics are based on the acceptance threshold that maximizes the respective statistic for a given bank-time-model combination. Value added is defined in Eq. (4). Each value added assumes a margin of 5% ($r=5\%$), a run-up of 30% ($(B_d - B_r)/B_d$), and a discount horizon of 3 years ($N=3$). The numbers represent the percentage cost savings of implementing each model versus passive risk management. The profit margin is used to estimate the opportunity cost of a false negative so that misclassifying more profitable accounts is more costly.

Bank	Test date	Value added – 2Q forecast			Value added – 3Q forecast			Value added – 4Q forecast		
		C4.5 tree	Logistic regression	Random forest	C4.5 tree	Logistic regression	Random forest	C4.5 tree	Logistic regression	Random forest
1	201106	51.5%	−63.9%	53.8%	32.1%	26.9%	32.2%	22.9%	−9.6%	21.8%
1	201112	51.4%	−20.5%	51.6%	30.5%	24.4%	29.9%	22.7%	15.4%	21.6%
1	201206	51.6%	43.8%	51.6%	29.0%	23.6%	28.6%	20.7%	15.5%	21.3%
1	201212	51.4%	45.2%	51.7%	27.6%	21.9%	24.4%	21.0%	14.5%	18.6%
1	201306	47.2%	40.2%	47.3%	12.1%	21.9%	28.2%	N/A	N/A	N/A
1	201312	51.0%	45.4%	52.1%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**50.7%**	**15.1%**	**51.4%**	**26.3%**	**23.7%**	**28.6%**	**21.8%**	**8.9%**	**20.8%**
2	201106	54.1%	53.4%	54.4%	30.2%	28.6%	30.8%	21.3%	17.9%	N/A
2	201112	53.4%	51.4%	52.3%	26.9%	23.7%	25.1%	24.7%	−471.2%	N/A
2	201206	47.9%	−1201%	52.5%	28.0%	−618.7%	28.7%	21.3%	−555.0%	N/A
2	201212	53.9%	−209.7%	53.3%	25.6%	−171.0%	N/A	22.3%	−106.4%	N/A
2	201306	51.5%	50.7%	51.9%	28.5%	26.2%	N/A	N/A	N/A	N/A
2	201312	53.8%	51.9%	51.3%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**52.4%**	**−200.5%**	**52.6%**	**27.8%**	**−142.2%**	**28.2%**	**22.4%**	**−278.6%**	**N/A**
3	201106	78.7%	69.4%	77.7%	45.5%	39.0%	44.7%	35.1%	31.7%	35.5%
3	201112	73.9%	68.2%	76.2%	44.9%	40.1%	45.1%	31.0%	27.8%	31.7%
3	201206	75.9%	68.4%	72.1%	44.3%	40.9%	45.3%	29.7%	22.0%	30.4%
3	201212	75.4%	65.6%	75.1%	46.7%	41.5%	46.4%	31.1%	27.1%	31.6%
3	201306	74.0%	65.3%	73.6%	50.5%	44.0%	48.5%	N/A	N/A	N/A
3	201312	75.0%	68.4%	73.4%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**75.5%**	**67.5%**	**74.7%**	**46.4%**	**41.1%**	**46.0%**	**31.7%**	**27.1%**	**32.3%**
4	201106	44.8%	41.1%	45.7%	22.8%	20.8%	25.8%	11.0%	8.7%	15.4%
4	201112	49.8%	40.2%	48.9%	−19.6%	19.2%	25.3%	10.1%	10.0%	14.4%
4	201206	46.0%	39.5%	44.2%	23.9%	18.3%	21.8%	14.6%	11.8%	12.5%
4	201212	47.9%	42.5%	46.2%	22.1%	19.3%	19.7%	−11.9%	13.4%	16.4%
4	201306	47.2%	43.9%	47.7%	22.2%	20.8%	26.9%	N/A	N/A	N/A
4	201312	48.3%	44.6%	49.3%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**47.3%**	**42.0%**	**47.0%**	**14.3%**	**19.7%**	**23.9%**	**6.0%**	**11.0%**	**14.7%**
5	201106	58.4%	53.8%	59.1%	−1.3%	−1.6%	11.6%	N/A	−110.0%	−81.4%
5	201112	55.9%	52.6%	57.0%	9.9%	−3.9%	7.2%	N/A	−36.1%	−40.0%
5	201206	52.3%	47.8%	55.3%	11.2%	−24.5%	10.8%	N/A	−83.2%	−60.3%
5	201212	57.9%	53.7%	57.1%	10.9%	−8.2%	9.9%	N/A	−124.3%	−64.9%
5	201306	56.1%	54.1%	58.4%	13.0%	1.9%	13.7%	N/A	N/A	N/A
5	201312	57.1%	54.4%	56.2%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**56.3%**	**52.7%**	**57.2%**	**8.7%**	**−7.3%**	**10.6%**	**N/A**	**−88.4%**	**−61.7%**
6	201106	53.3%	49.9%	53.0%	23.4%	20.5%	27.3%	19.6%	15.7%	18.0%
6	201112	52.9%	51.3%	52.9%	25.8%	21.2%	27.4%	24.0%	17.3%	23.9%
6	201206	57.2%	57.3%	58.1%	22.2%	28.2%	34.3%	13.2%	23.0%	24.3%
6	201212	35.6%	55.1%	56.1%	28.9%	26.6%	30.6%	24.4%	25.0%	25.8%
6	201306	19.2%	56.3%	57.6%	25.7%	26.3%	30.1%	N/A	N/A	N/A
6	201312	53.1%	51.2%	51.6%	N/A	N/A	N/A	N/A	N/A	N/A
Average:		**45.2%**	**53.5%**	**54.9%**	**25.2%**	**24.6%**	**30.0%**	**20.3%**	**20.2%**	**23.0%**

Moving to three- and four-quarter forecast horizons, the model performance declines, and as a result, the value added declines. However, the C4.5 trees and random forests remain positive, and continue to outperform logistic regression. Although the relative performance degrades somewhat, our machine-learning models still provide positive value at the longest forecast horizons.

Fig. 5 presents the value added versus the assumed run-up. The value added for each model increases with run-up. With the exception of a 10% run-up for Bank 5, all the C4.5 and random forest models generate positive value added for any run-up of at least 10%. The logistic models, however, need to have a run-up of at least 20% for Bank 1 to break even, and they never do so for Bank 2.

4.3. Risk management across institutions

In this section, we examine risk management practices across institutions. First, we compare the credit line management behavior across institutions. Second, we examine how well individual institutions target bad accounts. In credit cards, cutting lines is a very common tool used by banks to manage their risks, and one we can analyze, given our dataset.

As of each test date, we take the accounts predicted to default over a given horizon for a given bank, and analyze whether the bank cut its credit line or not. We use the predicted values from our models to simulate the banks' real problems, and to avoid any look-ahead bias. In Fig. 6 we plot the mean of the ratio of the percent of lines cut for defaulted accounts to the percent of lines cut on all accounts. A ratio greater than 1 implies that the bank is effectively targeting accounts that turn out to be bad and cutting their credit lines at a disproportionately greater rate than they are cutting all accounts, a sign of effective risk management practices. Similarly, a ratio less than 1 implies the opposite.[15] We report the ratio for each quarter between the model prediction and

[15] We plot the natural logarithm of this ratio in Fig. 6, where values above zero should be interpreted as effective risk management.

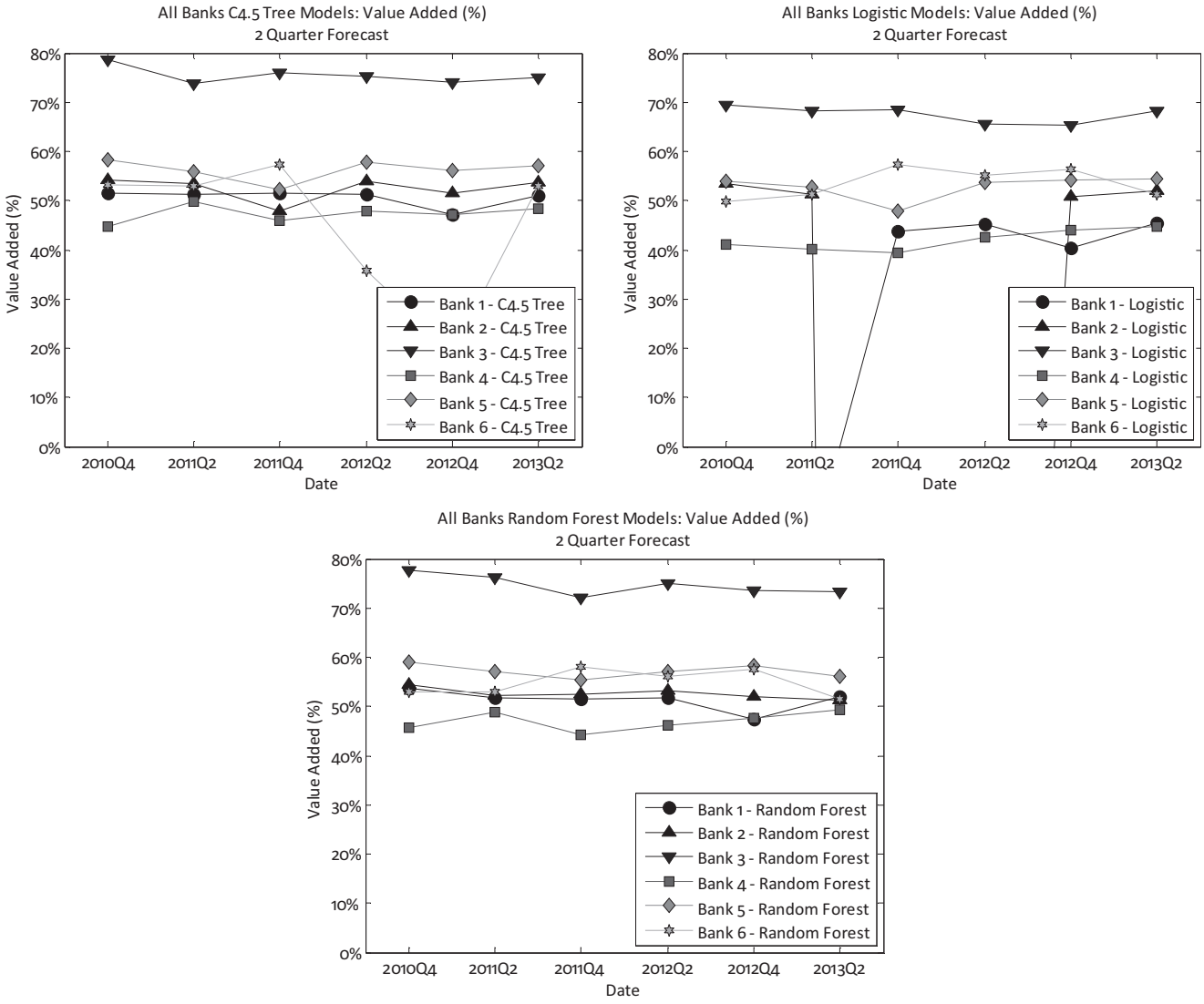

Fig. 4. Value added by model type. These figures plot the value added as defined by Eq. (4) over time. The statistics plotted are for the two-quarter horizon forecasts. Clockwise from the top left, the figures show the value added for C4.5 decision tree, logistic regression, and random forest models. Note the vertical axis is cut off at 0% and the logistic regression models for Bank 1 and Bank 2 are negative for the first two and third and fourth time periods, respectively.

the end of the forecast horizon under the assumption that cutting lines earlier is better if indeed they turn out to become delinquent.

The results show a significant amount of heterogeneity across banks. For example, Fig. 6 shows that three banks (2, 3, and 5) are very effective at cutting lines of accounts predicted to become delinquent—they are between 4.8 and 13.2 times more likely to target accounts predicted to default than the general portfolio. In contrast, Banks 4 and 6 underperform, rarely cutting lines of accounts predicted to default. Bank 1 tends to cut the same number of good and bad accounts. There is no clear pattern to banks' targeting of bad accounts across the forecast horizon.

Of course, these results are not conclusive, not least because banks have other risk management strategies in addition to cutting lines, and our efficacy measurement relies on the accuracy of our models. However, these empirical results show that, at a minimum, risk management policies differ significantly across major credit card issuing financial institutions.

4.4. Attribute analysis

A common criticism of machine-learning algorithms is that they are essentially black boxes, with results that are difficult to interpret. For example, given the chosen pruning and confidence limits of our decision tree models, the estimated decision trees tend to have about 100 leaves. The attributes selected by the algorithm vary across institutions and time, and the complexity of the trees makes it very difficult to compare them. Therefore, the first goal of our attribute analysis is to develop a method for interpreting the results of our machine-learning algorithms. The single decision tree models learned using C4.5 are particularly intuitive.

We propose a relatively straightforward approach for combining the results of the decision tree output, one that captures the results by generating an index based on three principal criteria. We start by constructing the following three metrics for each attribute in each decision tree:

1. *Log of the number of instances classified*: This is meant to capture the importance of the attribute. If attributes appear multiple times in a single model, we sum all the instances classified. This statistic is computed for each tree.
2. *The minimum leaf number*: The minimum leaf number is the highest node on the tree where the attribute sits, and roughly represents the statistical significance of the attribute. The logic

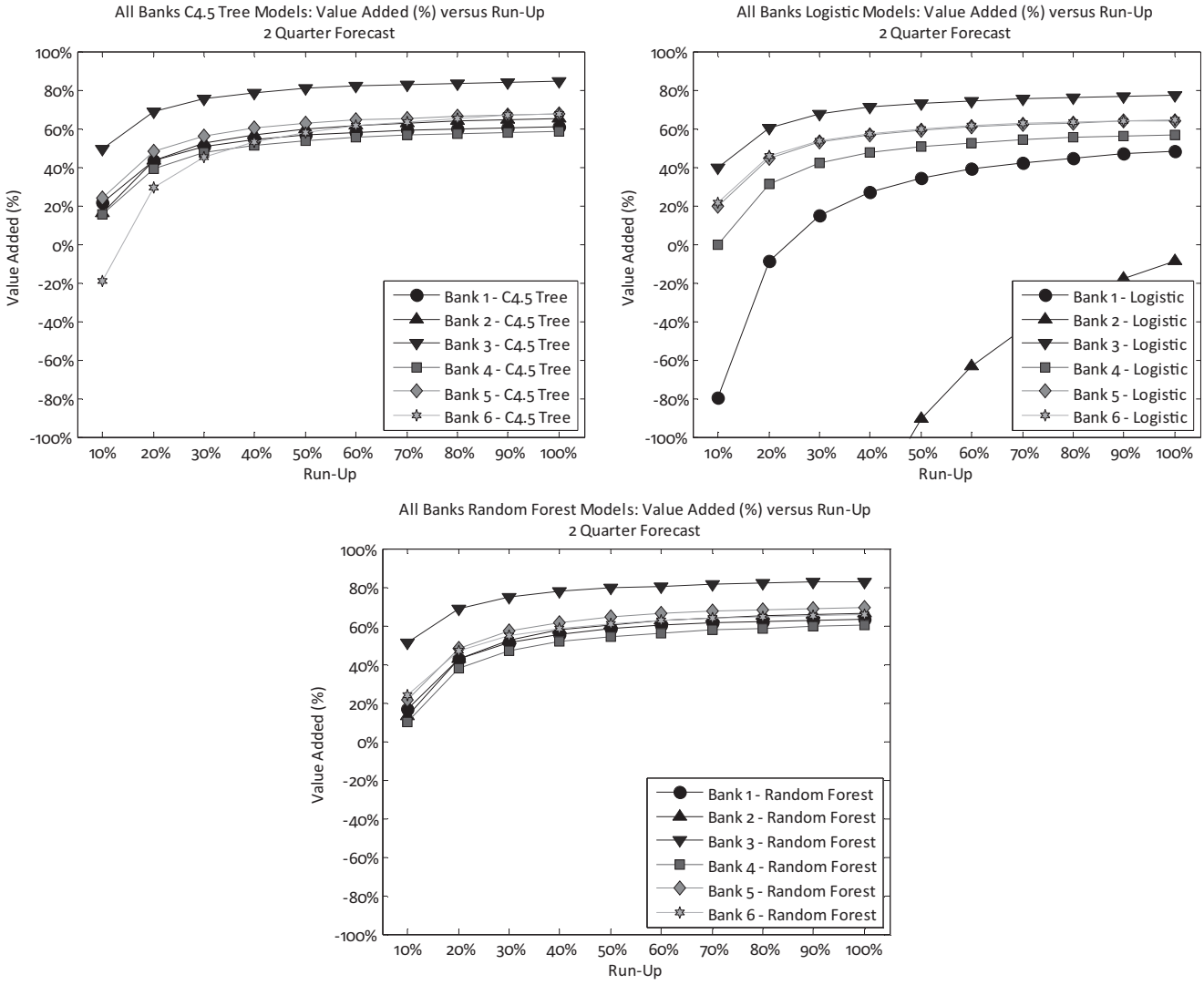

Fig. 5. Value added versus run-up. These figures plot the value added as defined by Eq. (4) versus run-up. The statistics plotted are for the two-quarter horizon forecasts. Clockwise from the top left, the figures show the value added for C4.5 decision tree, logistic regression, and random forest models. Note the vertical axis is cut off at -100% and the logistic regression models for Bank 1, Bank 2, and Bank 3 are negative for low values of run-up.

of the C4.5 classifier is that, in general, the higher up on the tree the attribute is (i.e., the lower the leaf number), the more important it is. Therefore, the attributes will be sorted in reverse order; that is, the variable with the lowest mean minimum leaf number would be ranked first. This statistic is computed for each tree.

3. *Indicator variable equal to 1 if the attribute appears in the tree and 0 otherwise*: We combine the results of multiple models over time to derive a bank-specific attribute ranking based on the number of times attributes are selected in a given model. For example, we run six separate C4.5 models for each bank using a two-quarter forecast horizon. This ranking criterion is the number of times (between zero and six) that a given attribute is selected to a model. This statistic is meant to capture the stability of an attribute over time.

We combine the above statistics into a single ranking measure by standardizing each to have a mean of 0 and a standard deviation of 1, and summing them by attribute. Attributes that do not appear in a model are assigned a score equal to the minimum of the standardized distribution. We then combine the scores for all unique bank-forecast horizon combinations, and rank the attributes. This leaves us with 18 individual scores for each attribute, used to rank them by importance. The most important attributes should have higher scores, appear near the top of the list, and have a lower numerical rank (i.e., attribute 1 is the most important).

In all, 78 of the 87 attributes are selected in at least one model. Table 6 shows the mean attribute rankings across all models, by forecast horizon, and by bank. More important attributes are ranked numerically lower. The table is sorted by the mean ranking for each attribute across all 18 bank-forecast horizon pairs. Columns 2–4 show the mean ranking by forecast horizon and columns 5–10 show the mean ranking by bank.

It is reassuring that the top-ranking variables—days past due, behavioral score, credit score, actual payment over minimum payment, 1 month change in utilization, etc.—are intuitive. For example, accounts that start out delinquent (less than 90 days) are most likely to become 90 days past due, regardless of the forecast horizon or bank.

Looking across forecast horizons, we see little variation. In fact, the pairwise Spearman rank correlations between the attribute rankings (for all 78 attributes that appear in at least one model) are between 89.8% and 94.3%.

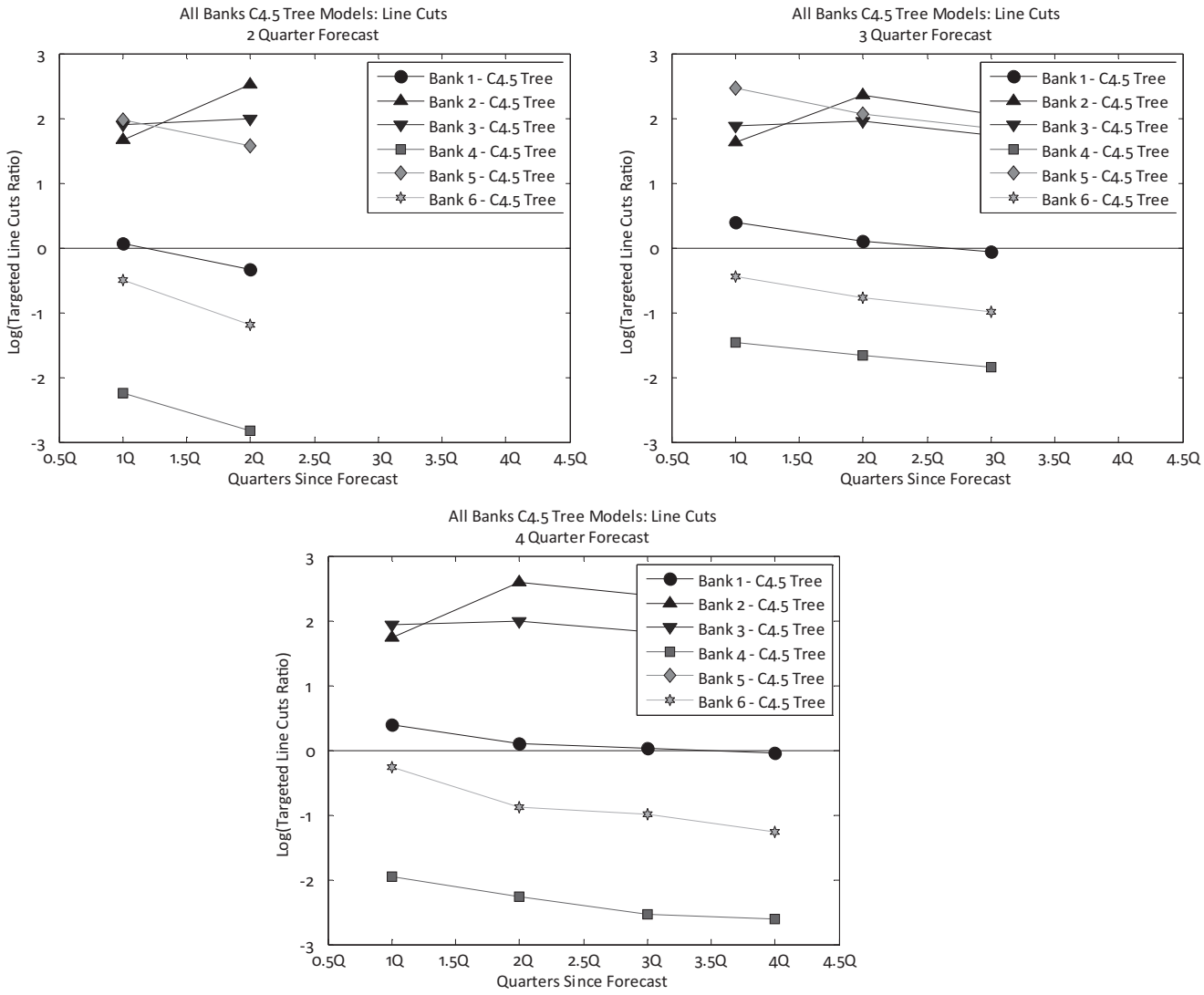

Fig. 6. Credit line cuts. The figures show how well banks target bad accounts and cut their credit lines relative to randomly selecting lines to cut. The targeted line ratio is defined as the percentage of accounts that our models predict to become delinquent whose lines are cut relative to the total percentage of accounts whose lines are cut. A ratio of one (zero on a log scale) means a bank is no more active in cutting credit lines of cards classified as bad than accounts classified as good. Higher ratios signal more active risk management. The ratios for each bank are plotted on a log scale. The plots show the ratios for each quarter following our forecast through the end of the forecast horizon. Clockwise from the top left, the figures show the value added for C4.5 decision tree, logistic regression, and random forest models.

However, there is a substantial amount of heterogeneity across banks, as suggested by the pairwise rank correlations between banks, which range from 46.5% to 80.3%. This suggests that the key risk factors affecting delinquency vary across banks. For example, the change in 1-month utilization (i.e., the percentage change in the drawdown of the credit line) has an average ranking between 2.0 and 4.0 for Banks 1, 2, and 5, but ranks between 10.3 and 15.7 for Banks 3, 4, and 6. For risk managers, this is a key attribute because managing drawdown and preventing run-up prior to default is central to managing credit card risk. Large variation in rank across banks in other attributes, including whether an account has entered into a workout program, the total fees, and whether an account is frozen, further suggests that banks have different risk management strategies.

Overall, the results in Table 6 support the validity of our models and variable ranking criteria, since the most widely used attributes in the industry tend to appear near the top of our rankings. However, looking across institutions, our results suggest that banks face different exposures, likely due to differences in underwriting practices and/or risk management strategies.

There is also substantial heterogeneity across banks in how macroeconomic variables affect their customers. Macroeconomic variables are more predictive (found among the most important 20 attributes) for Banks 2 and 6 in a two-quarter forecast horizon, and for Bank 6, at the 1-year forecast horizon as well. Although they are not the most important attributes, their ranking score is still relatively high, showing that the macroeconomic environment has a significant impact on consumer credit risk.

As mentioned above, we had also drawn the data previously at three other times. Using the data as of 2012Q4 (i.e., with 12 quarters of data, from 2009Q1 to 2012Q4), our results showed a greater sensitivity to the macroeconomic environment. These differences are intuitively consistent, since the macroeconomic environment from the vantage point of 2012Q4 was quite different from the macroeconomic environment of 2014Q2. These results emphasize the dynamic nature of machine-learning models, a particularly important feature for estimating industry relationships in transition.

Table 6
Attribute analysis.
The table shows the mean attribute ranking across all models, by forecast horizon, and by bank. For each unique bank and forecast horizon pair, the time series of C4.5 decision tree models reported in Tables 3–6 are combined, and attributes are assigned a score based on (1) the number of instances classified, (2) the minimum leaf on each tree they appear, and (3) the number of models for which they are selected. The scores are standardized and summed to generate an importance metric for each attribute for each bank-forecast horizon pair. More important attributes have lower numerical rank. The table is sorted by the mean ranking for each attribute across all bank-forecast horizon pairs. Columns 2–4 show the mean ranking by forecast horizon, and columns 5–10 show the mean ranking by bank. In all, 78 of the 87 attributes were selected in at least one model.

Attribute	All models	2Q horizon	3Q horizon	4Q horizon	Bank 1	Bank 2	Bank 3	Bank 4	Bank 5	Bank 6
Days past due	1.4	1.2	1.7	1.5	1.0	1.7	1.0	1.7	2.3	1.0
Behavioral score	3.7	3.2	4.3	3.7	8.3	1.3	2.3	3.0	1.7	5.7
Refreshed credit score	6.3	7.8	6.0	5.0	5.0	8.0	7.0	9.0	4.0	4.7
Actual payment/minimum payment	6.7	5.2	6.3	8.5	9.7	11.3	3.7	5.7	5.0	4.7
1 mo. chg. in monthly utilization	7.8	5.5	7.8	10.0	4.0	3.3	15.7	11.3	2.0	10.3
Payment equal minimum payment in past 3 mo. (0, 1)	8.6	7.8	8.5	9.5	6.3	8.7	6.7	10.3	6.3	13.3
Cycle end balance	9.8	10.8	10.2	8.3	11.0	6.3	12.7	7.7	17.0	4.0
3 mo. chg. in behavioral score	11.9	8.8	16.0	11.0	3.0	13.0	13.0	14.0	12.0	16.7
Cycle utilization	12.1	19.3	8.7	8.3	8.7	21.3	4.7	22.3	9.0	6.7
Number of accounts 30+ days past due	12.6	12.8	12.7	12.2	18.7	5.0	10.3	7.3	13.0	21.0
Total fees	15.9	16.2	12.8	18.8	15.0	21.3	8.3	14.3	9.3	27.3
Workout program flag	16.8	23.5	14.2	12.7	6.7	19.3	10.3	4.0	24.0	36.3
Total number of bank card accounts	17.8	18.5	17.5	17.3	22.0	21.7	19.0	14.3	17.7	12.0
Current credit limit	17.9	18.8	18.2	16.7	21.0	7.7	30.7	16.7	10.0	21.3
Line frozen flag (current mo.)	17.9	17.5	15.7	20.5	9.7	16.3	48.7	1.3	9.0	22.3
Monthly utilization	19.9	21.5	15.3	23.0	16.7	30.0	42.3	12.0	13.7	5.0
Number of accounts 60+ days past due	23.2	22.3	27.2	20.0	21.0	19.0	20.7	18.7	19.3	40.3
3 mo. chg. in credit score	24.4	21.8	24.2	27.2	8.7	27.3	28.3	21.7	32.3	28.0
Number of accounts in charge off status	26.3	26.0	27.7	25.2	27.3	17.0	24.0	18.3	39.0	32.0
1 mo. chg. in cycle utilization	27.0	29.3	26.7	25.0	17.7	38.3	10.7	30.3	28.3	36.7
6 mo. chg. in credit score	27.1	28.8	28.3	24.2	12.7	42.3	25.0	41.3	20.3	21.0
Total number of accounts 60+ days past due	27.9	21.5	32.3	30.0	31.7	24.3	18.0	11.3	41.3	41.0
Total balance on all 60+ days past due accounts	30.2	36.5	30.5	23.7	36.3	28.0	19.7	17.7	32.3	47.3
Total number of accounts verified	30.3	32.3	28.0	30.7	46.7	18.7	42.7	31.0	24.7	18.3
Flag if greater than 0 accounts 60 days past due	30.5	36.2	27.2	28.2	39.3	42.3	16.0	36.0	34.3	15.0
Line frozen flag (1 mo. lag)	30.9	15.5	34.5	42.7	16.3	8.0	33.3	29.0	47.3	51.3
3 mo. chg. in monthly utilization	33.4	30.2	34.8	35.2	19.0	22.7	31.7	42.7	40.0	44.3
Number of accounts 90+ days past due	33.7	43.5	29.8	27.8	34.3	25.0	33.3	31.7	36.0	42.0
6 mo. chg. in behavioral score	34.6	34.5	37.2	32.2	36.0	55.7	22.0	45.3	21.7	27.0
Account exceeded the limit in past 3 mo. (0, 1)	35.3	28.5	46.0	31.3	31.0	23.0	64.7	28.3	34.0	30.7
3 mo. chg. in cycle utilization	35.4	28.8	33.5	44.0	29.7	48.0	29.0	38.7	18.7	48.7
Flag if the card is securitized	36.2	35.5	36.7	36.3	24.0	13.7	30.3	28.7	71.7	48.7
Total number of accounts opened in the past year	36.4	41.7	36.0	31.5	41.0	24.0	38.7	45.0	28.3	41.3
Total number of bank card accounts 60+ days past due	37.4	38.5	32.8	41.0	47.3	25.0	23.7	25.3	40.7	62.7
Total balance of all revolving accounts/total balance on all accounts	39.3	41.0	34.5	42.5	30.0	40.3	43.0	43.3	33.3	46.0
Total number of accounts	41.3	34.2	48.7	41.0	40.7	26.3	35.3	32.3	64.0	49.0
Product type	41.4	38.5	41.7	44.0	20.3	61.0	73.0	71.7	11.3	11.0
Unemployment rate	41.6	41.8	37.2	45.7	42.3	36.7	48.3	54.7	29.3	38.0
Flag if greater than 0 accounts 30 days past due	41.6	47.7	39.7	37.5	55.3	37.3	35.7	44.7	22.0	54.7
Purchase volume/credit limit	43.4	43.5	38.2	48.5	30.3	58.3	32.3	70.3	36.0	33.0
Utilization of all bank card accounts	45.2	53.5	39.5	42.7	39.0	54.0	63.3	52.7	28.3	34.0
Flag if greater than 0 accounts opened in the past year	45.8	49.7	44.0	43.8	64.0	25.7	56.7	58.0	38.7	32.0
Flag if greater than 0 accounts 90 days past due	46.2	47.7	44.8	46.0	42.7	38.3	28.3	54.7	60.0	53.0
Avg. weekly hours worked (private) (12 mo. chg.)	46.2	44.8	49.2	44.5	61.0	37.0	55.7	42.7	52.3	28.3
Avg. hourly wage (private) (3 mo. chg.)	47.7	49.5	43.2	50.3	53.7	56.3	60.0	45.3	36.3	34.3
Avg. weekly hours worked (leisure) (12 mo. chg.)	47.9	49.7	43.0	51.0	53.3	40.0	57.0	60.7	54.3	22.0
Number of total nonfarm (NSA)	48.2	53.2	48.2	43.3	40.7	54.3	52.0	48.7	49.7	44.0
Avg. weekly hours worked (trade and transportation) (12 mo. chg.)	48.6	46.7	51.0	48.2	49.3	49.0	34.3	52.0	51.0	56.0
Avg. weekly hours worked (private) (3 mo. chg.)	49.8	48.2	44.2	57.0	48.7	46.7	53.0	42.3	50.3	57.7
Number of total nonfarm (NSA) (12 mo. chg.)	50.2	49.7	45.2	55.7	45.3	58.0	50.3	44.3	49.0	54.0
Avg. weekly hours worked (trade and transportation) (3 mo. chg.)	50.3	50.8	50.3	49.7	52.7	44.0	55.0	61.0	44.7	44.3
Avg. hourly wage (trade and transportation) (3 mo. chg.)	50.3	48.8	50.0	52.2	55.3	38.0	61.3	38.0	54.3	55.0
Total non-mortgage balance/total limit	50.6	55.0	46.3	50.3	51.7	64.7	55.7	38.7	46.0	46.7
Avg. hourly wage (private) (12 mo. chg.)	51.8	50.3	53.5	51.5	56.0	45.7	59.0	54.0	47.3	48.7
Avg. hourly wage (trade and transportation) (12 mo. chg.)	51.8	57.2	48.8	49.3	52.0	55.0	60.0	47.3	37.3	59.0
Avg. weekly hours worked (leisure) (3 mo. chg.)	51.9	52.5	50.5	52.7	51.3	43.3	39.7	59.7	64.7	52.7
6 mo. chg. in cycle utilization	52.1	46.7	54.7	54.8	33.0	70.3	48.0	64.7	38.3	58.0
Avg. hourly wage (leisure) (12 mo. chg.)	53.2	49.0	53.5	57.2	47.0	48.3	53.3	46.3	62.0	62.3
Avg. hourly wage (leisure) (3 mo. chg.)	53.6	52.7	52.7	55.5	58.7	60.7	62.3	37.3	66.3	36.3
Total credit limit to number of open bank cards	54.0	52.0	52.3	57.7	68.0	56.0	45.3	41.7	49.0	64.0
Number of total nonfarm (NSA) (3 mo. chg.)	54.2	51.3	55.2	56.0	62.3	45.0	54.0	57.3	40.0	66.3

(*continued on next page*)

Table 6 (*continued*)

Attribute	All models	2Q horizon	3Q horizon	4Q horizon	Bank 1	Bank 2	Bank 3	Bank 4	Bank 5	Bank 6
Flag if total limit on all bank cards greater than zero	54.8	50.0	60.2	54.3	59.3	72.0	43.7	33.3	67.3	53.3
Unemployment rate (3 mo. chg.)	55.0	58.3	53.5	53.2	52.3	56.0	68.3	55.3	52.7	45.3
Number of total nonfarm (NSA) (3 mo. chg.)	55.9	59.2	64.2	44.3	58.0	61.7	50.0	55.0	62.0	48.7
Total private (NSA) (12 mo. chg.)	56.0	57.5	53.5	57.0	53.3	47.7	54.3	64.3	56.3	60.0
Percent chg. in credit limit (lagged 1 mo.)	56.5	57.0	52.5	60.0	66.7	74.3	10.7	68.7	58.7	60.0
Unemployment rate (12 mo. chg.)	58.3	53.8	65.3	55.7	42.7	66.7	66.7	61.7	64.3	47.7
Percent chg. in credit limit current 1 mo.	58.4	60.0	59.7	55.5	71.0	74.7	11.7	65.7	68.7	58.7
6 mo. chg. in monthly utilization	58.6	48.7	65.5	61.5	46.0	59.0	50.3	72.7	62.0	61.3
Flag if total limit on all retail cards greater than zero	59.6	55.0	60.3	63.3	62.0	73.0	63.0	31.0	76.0	52.3
Total balance on all accounts/total limit	60.5	55.2	66.2	60.2	72.0	56.7	69.0	57.7	53.3	54.3
Flag if greater than 0 retail cards 60 days past due	60.9	68.0	61.8	53.0	68.7	43.3	55.3	63.7	75.7	59.0
Cash advance volume/credit limit	61.7	64.5	64.5	56.2	72.7	47.0	74.0	63.0	59.0	54.7
Total credit limit to number of open retail accounts	67.0	66.7	67.3	67.0	70.7	72.7	74.0	69.7	67.7	47.3
Line decrease in current mo. flag (0, 1)	67.8	68.7	69.0	65.7	74.7	75.3	57.0	64.7	67.0	68.0
Number of accounts in collection	68.0	66.0	73.3	64.7	69.3	65.3	76.7	65.3	73.7	57.7
Flag if total balance over limit on all open bank cards = 0%	68.1	65.8	67.0	71.3	74.7	76.7	70.0	63.7	66.7	56.7
Number of accounts under wage garnishment	68.7	71.2	68.2	66.7	75.7	70.0	66.3	65.3	67.7	67.0

4.5. Robustness – cross bank model results

The heterogeneity across banks could indicate that fundamental differences exist in the underwriting and/or risk management practices across banks. In particular, the attribute rankings exhibit substantial heterogeneity across banks which could reflect cross-sectional differences in credit card portfolios. Alternatively, it could be a result of poorly fitted models that pick up substantial amounts of noise.[16]

To address this concern, we run our decision tree models across banks. For example, we train the data on Bank 1 and test the data each of Banks 2–6. We repeat this for all pairwise combinations of banks. The idea is that if the true underlying risk drivers across banks are the same and our models are simply picking up noise, then we should not see much degradation in the performance of the models when applied to alternative banks; i.e., using a model trained on Bank 1′s data should perform about as well when tested on Banks' 2–6 data as compared to its own data.

The results of the experiment are given in Table 7. Panels A, B, and C show the results for the two-, three-, and four-quarter forecasts, respectively. The columns represent the bank used to train the data and the rows represent the banks used to test the models. The figures in the table represent the mean value-added of the model forecasts across the time periods so the diagonal values in each panel correspond to the average value-added numbers in Table 5.

The results suggest that the models do pick up differences in the underlying risk drivers across portfolios. This is highlighted by the fact that the diagonal elements of each panel tend to be larger than the off-diagonal terms, implying that the models are best suited for the banks on which they were trained. Note that this is not likely an over-fitting problem as the models are still tested strictly out-of-time meaning there is no look-ahead bias. From a supervisory perspective, these results support our contention that bank-specific models are likely to be better predictors of default as opposed to a single model applied to all banks.[17]

Table 7
Cross model results.
The table shows cross model results. Panels A, B, and C show the results for the two-, three-, and four-quarter forecasts, respectively. The columns represent the bank used to train the data and the rows represent the banks used to test the models. The figures in the table represent the mean value-added of the model forecasts across the time periods so the diagonal values in each panel correspond to the average value-added numbers in Table 5.

Test sample	Model Bank 1	Bank 2	Bank 3	Bank 4	Bank 5	Bank 6	Average
Panel A: Two-quarter forecast							
Bank 1	**50.7%**	17.3%	43.9%	41.0%	37.9%	3.9%	32.4%
Bank 2	39.9%	**52.4%**	45.4%	46.6%	44.8%	40.5%	44.9%
Bank 3	−29.6%	31.8%	**75.5%**	45.6%	41.8%	23.2%	31.4%
Bank 4	41.0%	42.7%	1.8%	**47.3%**	40.0%	8.3%	30.2%
Bank 5	51.0%	28.1%	50.9%	50.7%	**56.3%**	28.2%	44.2%
Bank 6	49.4%	32.7%	22.2%	53.2%	51.5%	**45.2%**	42.4%
Column average	33.7%	34.2%	39.9%	47.4%	45.4%	24.9%	
Panel B: Three-quarter forecast							
Bank 1	**26.3%**	12.8%	24.8%	−8.7%	3.5%	15.2%	12.3%
Bank 2	24.0%	**27.8%**	24.5%	11.5%	14.0%	19.4%	20.2%
Bank 3	19.9%	24.4%	**46.4%**	2.2%	27.2%	21.0%	23.5%
Bank 4	18.1%	3.8%	15.3%	**14.3%**	−21.0%	3.1%	5.6%
Bank 5	−0.5%	−15.5%	−1.7%	−7.0%	**8.7%**	−17.5%	−5.6%
Bank 6	27.1%	20.3%	26.7%	16.3%	15.5%	**25.2%**	21.8%
Column average	19.1%	12.3%	22.7%	4.8%	8.0%	11.1%	
Panel C: Four-quarter forecast:							
Bank 1	**21.8%**	6.2%	16.7%	−44.1%	N/A	12.3%	2.6%
Bank 2	18.6%	**22.4%**	19.2%	−46.7%	N/A	17.9%	6.3%
Bank 3	13.8%	15.9%	**31.7%**	−11.2%	N/A	18.4%	13.7%
Bank 4	10.9%	−5.7%	10.4%	**6.0%**	N/A	2.7%	4.9%
Bank 5	−58.2%	−1540%	−80.5%	−53.5%	**N/A**	−58.5%	−358.3%
Bank 6	21.4%	−627.1%	19.2%	6.2%	N/A	**20.3%**	−112.0%
Column average	4.7%	−354.8%	2.8%	−23.9%	N/A	2.2%	

5. Conclusion

In this study, we employ a unique, very large dataset consisting of anonymized information from six large banks collected by a financial regulator to build and test decision tree, regularized logistic regression, and random forest models for predicting credit card delinquency. The algorithms have access to combined consumer tradeline, credit bureau, and macroeconomic data from January 2009 to December 2013. We find that decision trees and random forests outperform logistic regression in both out-of-sample and out-of-time forecasts of credit card delinquencies. The

[16] We thank an anonymous referee for pointing this out and suggesting this experiment.

[17] We refrain from analyzing any bank specific factors such as business or risk management strategies that could explain the cross-sectional differences to preserve the anonymity of the banks.

advantage of decision trees and random forests over logistic regression is most significant at short time horizons. The success of these models implies that there may be a considerable amount of "money left on the table" by credit card issuers.

We also analyze and compare risk management practices across the banks, and compare drivers of delinquency across institutions. We find that there is substantial heterogeneity across banks in risk factors and sensitivities to those factors. Therefore, no single model is likely to capture the delinquency tendencies across all institutions. The results also suggest that portfolio characteristics alone are not sufficient to identify the drivers of delinquency, since the banks actively manage the portfolios. Even a nominally high-risk portfolio may have fewer volatile delinquencies because of successful active risk management by the bank.

The heterogeneity of credit card risk management practices across financial institutions has systemic implications. Credit card receivables form an important component of modern asset-backed securities. We have found that certain banks are significantly more active and effective at managing the exposure of their credit card portfolios, while credit card delinquency rates across banks are also quite different in their macroeconomic sensitivities. An unexpected macroeconomic shock may thus propagate itself through a greater delinquency rate of credit cards issued by specific financial institutions into the asset-backed securities market.

Our study provides an in-depth illustration of the potential benefits that "Big Data" and machine-learning techniques can bring to consumers, risk managers, shareholders, and regulators, all of whom have a stake in avoiding unexpected losses and reducing the cost of consumer credit. Moreover, when aggregated across a number of financial institutions, the predictive analytics of machine-learning models provide a practical means for measuring systemic risk in one of the most important and vulnerable sectors of the economy. We plan to explore this application in ongoing and future research.

Appendix A. Variables descriptions for tradeline and attributes data

Account level features	Credit bureau features	Macroeconomic features
Cycle end balance	Flag if greater than 0 accounts 90 days past due	Unemployment rate
Refreshed credit score	Flag if greater than 0 accounts 60 days past due	Unemployment rate (3 mo. chg.)
Behavioral score	Flag if greater than 0 accounts 30 days past due	Unemployment rate (12 mo. chg.)
Current credit limit	Flag if greater than 0 bank cards 60 days past due	Number of total nonfarm (NSA)
Line frozen flag (0, 1)	Flag if greater than 0 retail cards 60 days past due	Number of total nonfarm (NSA) (3 mo. chg.)
Line decrease in current mo. flag (0, 1)	Flag if total limit on all bank cards greater than zero	Number of total nonfarm (NSA) (12 mo. chg.)
Line increase in current mo. flag (0, 1)	Flag if total limit on all retail cards greater than zero	Total private (NSA) (3 mo. chg.)
Actual payment/minimum payment	Flag if greater than 0 accounts opened in the past year	Total private (NSA) (12 mo. chg.)
Days past due	Total number of accounts	Avg. weekly hours worked (private) (3 mo. chg.)
Purchase volume/credit limit	Total balance on all accounts/total limit	Avg. weekly hours worked (private) (12 mo. chg.)
Cash advance volume/credit limit	Total non-mortgage balance/total limit	Avg. hourly wage (private) (3 mo. chg.)
Balance transfer volume/credit limit	Total number of accounts 60+ days past due	Avg. hourly wage (private) (12 mo. chg.)
Flag if the card is securitized	Total number of bank card accounts	Avg. weekly hours worked (trade and transportation) (3 mo. chg.)
chg. in securitization status (1 mo.)	Utilization of all bank card accounts	Avg. weekly hours worked (trade and transportation) (12 mo. chg.)
Percent chg. in credit limit (lagged 1 mo.)	Number of accounts 30+ days past due	Avg. hourly wage (trade and transportation) (3 mo. chg.)
Percent chg. in credit limit current 1 mo.)	Number of accounts 60+ days past due	Avg. hourly wage (trade and transportation) (12 mo. chg.)
Total fees	Number of accounts 90+ days past due	Avg. weekly hours worked (leisure) (3 mo. chg.)
Workout program flag	Number of accounts under wage garnishment	Avg. weekly hours worked (leisure) (12 mo. chg.)
Line frozen flag (1 mo. lag)	Number of accounts in collection	Avg. hourly wage (leisure) (3 mo. chg.)
Line frozen flag (current mo.)	Number of accounts in charge off status	Avg. hourly wage (leisure) (12 mo. chg.)
Product type	Total balance on all 60+ days past due accounts	House price index
3 mo. chg. in credit score	Total number of accounts	House price index (3 mo. chg.)
6 mo. chg. in credit score	Total credit limit to number of open bank cards	House price index (12 mo. chg.)
3 mo. chg. in behavioral score	Total credit limit to number of open retail accounts	
6 mo. chg. in behavioral score	Total number of accounts opened in the past year	
Monthly utilization	Total balance of all revolving accounts/total balance on all accounts	
1 mo. chg. in monthly utilization	Flag if total balance over limit on all open bank cards = 0%	
3 mo. chg. in monthly utilization	Flag if total balance over limit on all open bank cards = 100%	
6 mo. chg. in monthly utilization	Flag if total balance over limit on all open bank cards > 100%	
Cycle utilization		
1 mo. chg. in cycle utilization		
3 mo. chg. in cycle utilization		
Account exceeded the limit in past 3 mo. (0, 1)		
Payment equal minimum payment in past 3 mo. (0, 1)		
6 mo. chg. in cycle utilization		

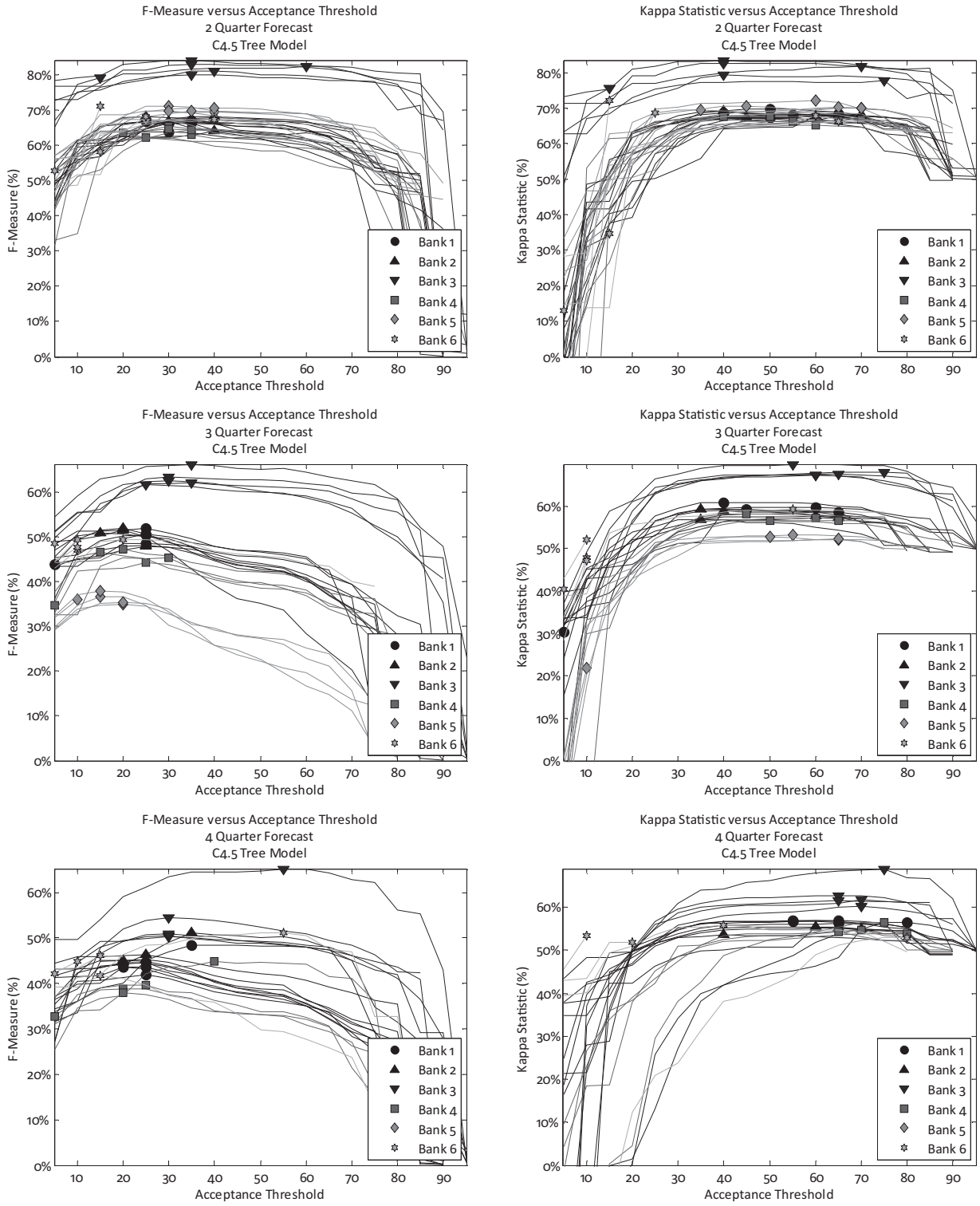

Fig. A1. Sensitivity to choice of acceptance threshold for C4.5 models. The figures on the left show the *F*-measure versus the acceptance threshold for each C4.5 model. The figures on the right show the kappa statistic versus the acceptance threshold. The acceptance threshold is given as a percentage. The dots designate the acceptance threshold that maximizes the respective statistic.

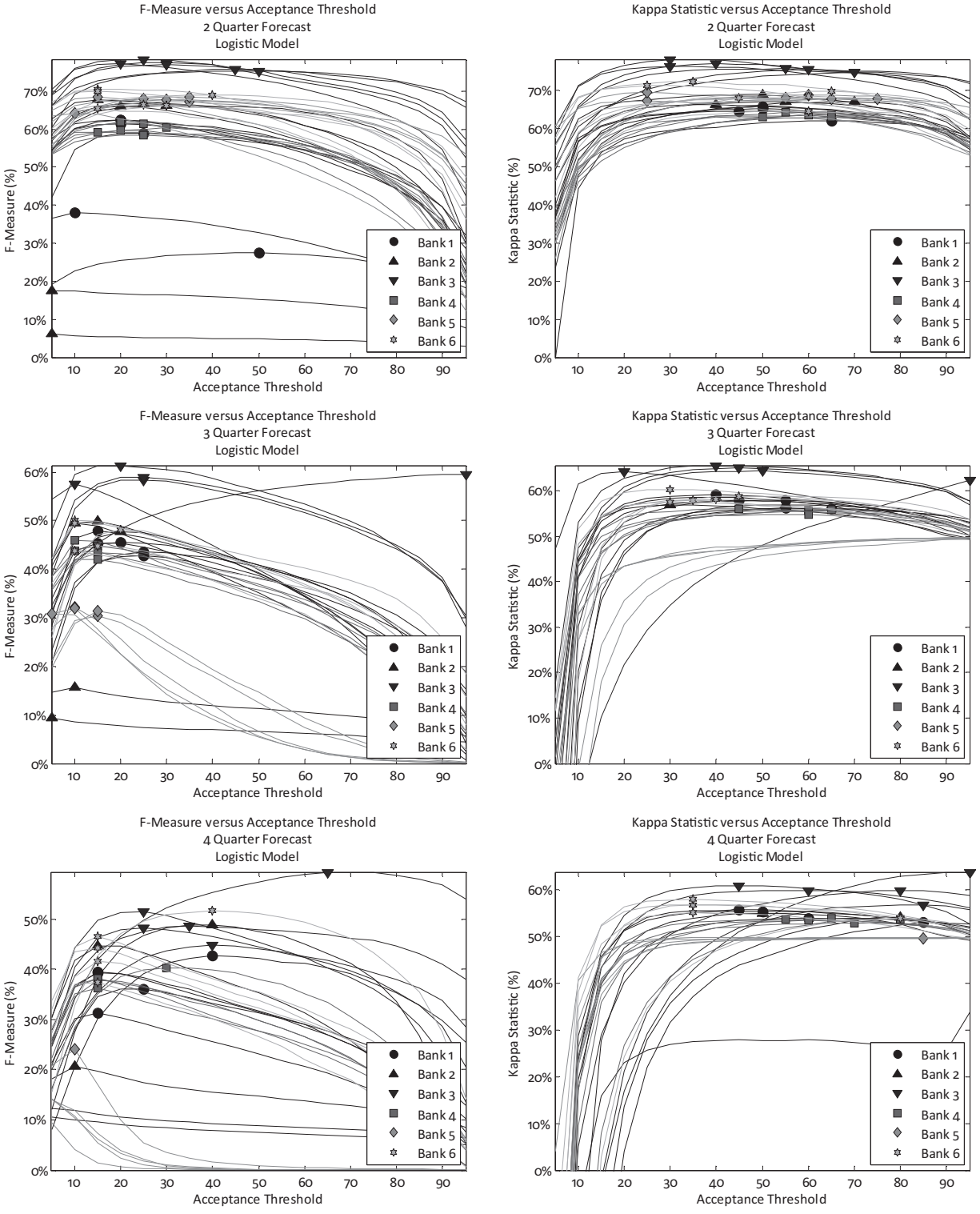

Fig. A2. Sensitivity to choice of acceptance threshold for logistic regression models. The figures on the left show the *F*-measure versus the acceptance threshold for each logistic regression model. The figures on the right show the kappa statistic versus the acceptance threshold. The acceptance threshold is given as a percentage. The dots designate the acceptance threshold that maximizes the respective statistic.

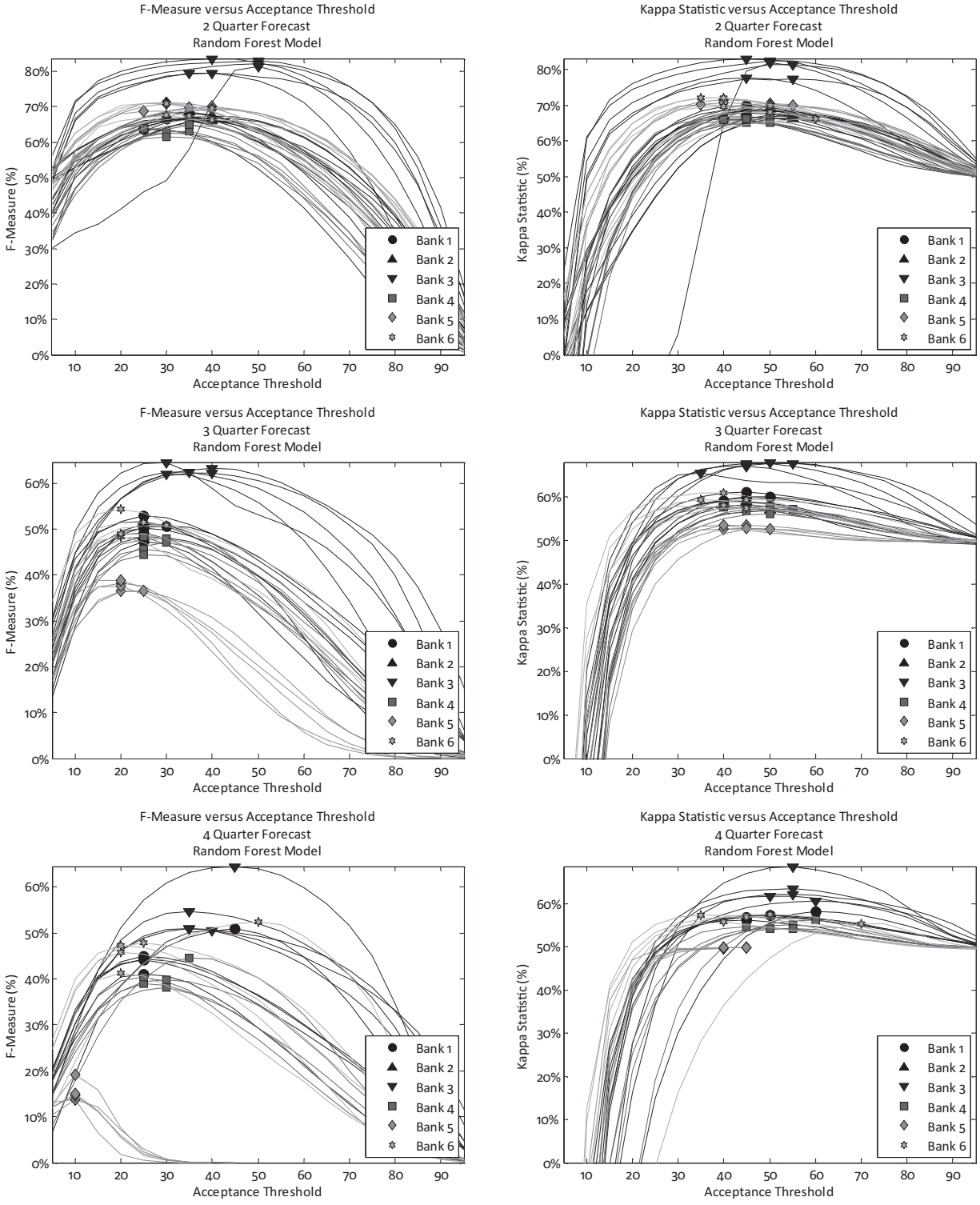

Fig. A3. Sensitivity to choice of acceptance threshold for random forest models. The figures on the left show the *F*-measure versus the acceptance threshold for each random forest model. The figures on the right show the kappa statistic versus the acceptance threshold. The acceptance threshold is given as a percentage. The dots designate the acceptance threshold that maximizes the respective statistic.

References

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

Breiman, L., Cutler, A., 2004. Random forests. Manual http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_papers.htm.

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In: ICML '06 Proceedings of the 23rd International Conference on Machine learning, pp. 161–168.

Cessie, S., van Houwelingen, J.C., 1992. Ridge estimators in logistic regression. Applied Statistics 41 (1), 191–201.

Criminisi, A., Shotton, J., Konukoglu, E., 2012. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations and Trends in Computer Graphics and Vision 7 (2–3), 81–227.

Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning 40 (2), 139–157.

Frank, E., Hall, M.A., Witten, I.H., 2011. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington, MA.

Glennon, D., Kiefer, N.M., Larson, C.E., Choi, H.-s., 2008. Development and validation of credit-scoring models. Journal of Credit Risk 4 (3), 1–61.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning. Springer, New York.

Khandani, A.E., Kim, A.J., Lo, A.W., 2010. Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance 34 (11), 2767–2787.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174. doi:10.2307/2529310.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufman, San Mateo, CA.

Thomas, L.C., 2000. A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumers. International Journal of Forecasting 16 (2), 163–167.