Journal of Banking & Finance 34 (2010) 2767–2787

Contents lists available at ScienceDirect

Journal of Banking & Finance

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

Consumer credit-risk models via machine-learning algorithms ☆

Amir E. Khandani, Adlar J. Kim, Andrew W. Lo *

MIT Sloan School of Management and Laboratory for Financial Engineering, United States

ARTICLE INFO

Article history:
Received 11 March 2010
Accepted 4 June 2010
Available online 10 June 2010

JEL classification:
G21
G33
G32
G17
G01
D14

Keywords:
Household behavior
Consumer credit risk
Credit card borrowing
Machine learning
Nonparametric estimation

ABSTRACT

We apply machine-learning techniques to construct nonlinear nonparametric forecasting models of consumer credit risk. By combining customer transactions and credit bureau data from January 2005 to April 2009 for a sample of a major commercial bank's customers, we are able to construct out-of-sample forecasts that significantly improve the classification rates of credit-card-holder delinquencies and defaults, with linear regression R^2's of forecasted/realized delinquencies of 85%. Using conservative assumptions for the costs and benefits of cutting credit lines based on machine-learning forecasts, we estimate the cost savings to range from 6% to 25% of total losses. Moreover, the time-series patterns of estimated delinquency rates from this model over the course of the recent financial crisis suggest that aggregated consumer credit-risk analytics may have important applications in forecasting systemic risk.

1. Introduction

One of the most important drivers of macroeconomic conditions and systemic risk is consumer spending, which accounted for over two thirds of US gross domestic product as of October 2008. With $13.63 trillion of consumer credit outstanding as of the fourth quarter of 2008 ($10.47 trillion in mortgages, $2.59 trillion in other consumer debt), the opportunities and risk exposures in consumer lending are equally outsized.[1] For example, as a result of the recent financial crisis, the overall charge-off rate in all revolving consumer credit across all US lending institutions reached 10.1% in the third quarter of 2009, far exceeding the average charge-off rate of 4.72% during 2003–2007. [2] With a total of $874 billion of revolving consumer credit outstanding in the US economy as of November 2009,[3] and with 46.1% of all families carrying a positive credit-card balance in 2007,[4] the potential for further systemic dislocation in this sector has made the economic behavior of consumers a topic of vital national interest.

The large number of decisions involved in the consumer lending business makes it necessary to rely on models and algorithms rather than human discretion, and to base such algorithmic decisions on "hard" information, e.g., characteristics contained in consumer credit files collected by credit bureau agencies. Models are typically used to generate numerical "scores" that summarize the creditworthiness of consumers.[5] In addition, it is common for

☆ The views and opinions expressed in this article are those of the authors only, and do not necessarily represent the views and opinions of AlphaSimplex Group, MIT, any of their affiliates and employees, or any of the individuals acknowledged below. We thank Jayna Cummings, Tanya Giovacchini, Alan Kaplan, Paul Kupiec, Frank Moss, Deb Roy, Ahktarur Siddique, Roger Stein, two referees, the editor Ike Mathur, and seminar participants at Citigroup, the FDIC, the MIT Center for Future Banking, the MIT Media Lab, and Moody's Academic Advisory and Research Committee for many helpful comments and discussion. Research support from the MIT Laboratory for Financial Engineering and the Media Lab's Center for Future Banking is gratefully acknowledged.

* Corresponding author. Tel.: +1 617 253 0920; fax: +1 781 891 9783.
E-mail addresses: khandani@mit.edu (A.E. Khandani), jwkim@csail.mit.edu (A.J. Kim), alo@mit.edu (A.W. Lo).

[1] US Federal Reserve Flow of Funds data, June 11, 2009 release.

[2] Data available from the Federal Reserve Board at http://www.federalreserve.gov/releases/chargeoff/.

[3] See the latest release of Consumer Credit Report published by the Federal Reserve Board, available at http://www.federalreserve.gov/releases/g19/Current/.

[4] See the *Survey of Consumer Finances, 2009* (SCF), released in February 2009 and available at http://www.federalreserve.gov/pubs/bulletin/2009/pdf/scf09.pdf. This report shows that the median balance for those carrying a non-zero balance was $3,000, while the mean was $7,300. These values have risen 25% and 30.4%, respectively, since the previous version of the SCF conducted three year earlier. The SCF also reports that the median balance has risen strongly for most demographic groups, particularly for higher-income groups.

[5] See Hand and Henley (1997) and Thomas (2009) for reviews of traditional and more recent statistical modeling approaches to credit scoring.

0378-4266/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbankfin.2010.06.001

lending institutions and credit bureaus to create their own customized risk models based on private information about borrowers. The type of private information usually consists of both "within-account" as well as "across-account" data regarding customers' past behavior.[6] However, while such models are generally able to produce reasonably accurate ordinal measures, i.e., rankings, of consumer creditworthiness, these measures adjust only slowly over time and are relatively insensitive to changes in market conditions. Given the apparent speed with which consumer credit can deteriorate, there is a clear need for more timely cardinal measures of credit risk by banks and regulators.

In this paper, we propose a cardinal measure of consumer credit risk that combines traditional credit factors such as debt-to-income ratios with consumer banking transactions, which greatly enhances the predictive power of our model. Using a proprietary dataset from a major commercial bank (which we shall refer to as the "Bank" throughout this paper to preserve confidentiality) from January 2005 to April 2009, we show that conditioning on certain changes in a consumer's bank-account activity can lead to considerably more accurate forecasts of credit-card delinquencies in the future. For example, in our sample, the unconditional probability of customers falling 90-days-or-more delinquent on their payments over any given 6-month period is 5.3%, but customers experiencing a recent decline in income—as measured by sharp drops in direct deposits—have a 10.8% probability of 90-days-or-more delinquency over the subsequent 6 months. Such conditioning variables are statistically reliable throughout the entire sample period, and our approach is able to generate many variants of these transactions-based predictors and combine them in nonlinear ways with credit bureau data to yield even more powerful forecasts. By analyzing patterns in consumer expenditures, savings, and debt payments, we are able to identify subtle nonlinear relationships that are difficult to detect in these massive datasets using standard consumer credit-default models such as logit, discriminant analysis, or credit scores.[7]

We use an approach known as "machine learning" in the computer science literature, which refers to a set of algorithms specifically designed to tackle computationally intensive pattern-recognition problems in extremely large datasets. These techniques include radial basis functions, tree-based classifiers, and support-vector machines, and are ideally suited for consumer credit-risk analytics because of the large sample sizes and the complexity of the possible relationships among consumer transactions and characteristics.[8] The extraordinary speed-up in computing in recent years, coupled with significant theoretical advances in machine-learning algorithms, have created a renaissance in computational modeling, of which our consumer credit-risk model is just one of many recent examples.

One measure of the forecast power of our approach is to compare the machine-learning model's forecasted scores of those customers who eventually default during the forecast period with the forecasted scores of those who do not. Significant differences between the forecasts of the two populations is an indication that the forecasts have genuine discriminating power. Over the sample period from May 2008 to April 2009, the average forecasted score among individuals who do become 90-days-or-more delinquent during the 6-month forecast period is 61.9, while the average score across all customers is 2.1. The practical value added of such forecasts can be estimated by summing the cost savings from credit reductions to high-risk borrowers and the lost revenues from "false positives", and under a conservative set of assumptions, we estimate the potential net benefits of these forecasts to be 6–25% of total losses.

More importantly, by aggregating individual forecasts, it is possible to construct a measure of systemic risk in the consumer-lending sector, which accounts for one of the largest components of US economic activity. As Buyukkarabacaka and Valevb (2010) observe, private credit expansions are an early indicator of potential banking crises. By decomposing private credit into household and enterprise credit, they argue that household-credit growth increases debt without much effect on future income, while enterprise-credit expansion typically results in higher future income. Accordingly, they argue that rapid household-credit expansions are more likely to generate vulnerabilities that can precipitate a banking crisis than enterprise-credit expansion. Therefore, a good understanding of consumer choice and early warning signs of over-heating in consumer finance are essential to effective macroprudential risk management policies. We show that the time-series properties of our machine-learning forecasts are highly correlated with realized credit-card delinquency rates (linear regression R^2's of 85%), implying that a considerable portion of the consumer credit cycle can be forecasted 6–12 months in advance.

In Section 2, we describe our dataset, discuss the security issues surrounding it, and document some simple but profound empirical trends. Section 3 outlines our approach to constructing useful variables or feature vectors that will serve as inputs to the machine-learning algorithms we employ. In Section 4, we describe the machine-learning framework for combining multiple predictors to create more powerful forecast models, and present our empirical results. Using these results, we provide two applications in Section 5, one involving model-based credit-line reductions and the other focusing on systemic risk measures. We conclude in Section 6.

2. The data

In this study, we use a unique dataset consisting of transaction-level, credit bureau, and account-balance data for individual consumers. This data is obtained for a subset of the Bank's customer base for the period from January 2005 to April 2009. Integrating transaction, credit bureau, and account-balance data allows us to compute and update measures of consumer credit risk much more frequently than the slower-moving credit-scoring models currently being employed in the industry and by regulators. This data was assembled from a number of different sources which we review in Section 2.1.[9]

Given the sensitive nature of the data and consumer privacy protection laws, all individual identification data elements such as names, addresses, and social security numbers were removed by the Bank before sharing the data with us, and all computations were performed on machines physically located at the Bank's office and within the Bank's electronic "firewalls". We review these security protocols in Section 2.2. In Section 2.3, we document some trends in debt balances and payments over the sample period that clearly point to substantial but gradual deterioration in the financial health of consumers over this period.

2.1. Data sources

The raw data was pre-processed to produce higher-level variables or "feature vectors", and time-aggregated to yield monthly

[6] The impact of such relationship information in facilitating banking engagement has been studied extensively in the area of corporate and small business lending (see, for example, Boot, 2000) and, more recently, in consumer lending (see Agarwal et al., 2009).

[7] As discussed in Avery et al. (2004), local economic circumstances and individual trigger-events must be considered in order to develop powerful predictive models, and including transaction and relationship-specific variables may proxy for such inputs.

[8] See, for example, Foster and Stine (2004), Huang et al. (2006), Li et al. (2006) and Bellotti and Crook (2009) for applications of machine learning based model to consumer credit.

[9] These three different data sources are combined using a unique identifier for each person. More advanced techniques, such as those of Dwyer and Stein (2006), may yield even more powerful predictors by incorporating household information and linking defaults across household members.

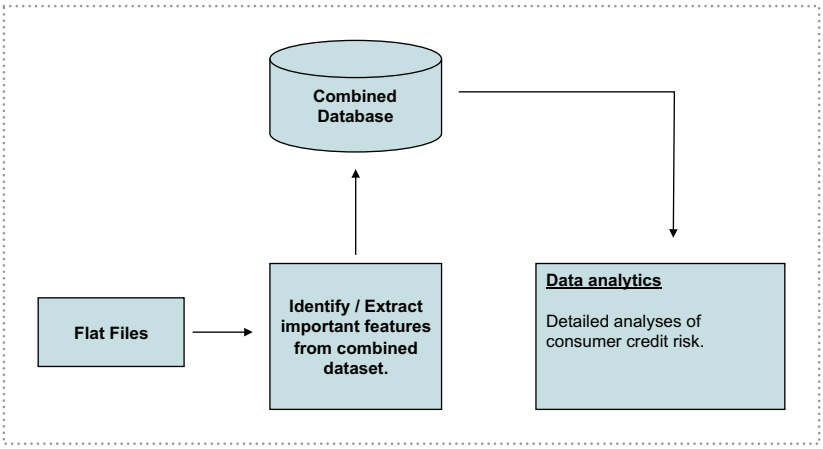


Fig. 1. Construction of an integrated database of transactions and credit bureau statistics used in machine-learning model of consumer credit risk.

observations that were stored in a database (see Fig. 1). The following is a description of each of these components.

2.1.1. Transactions data

The Bank provided us with account-level transactions data for a subset of its customer base. In addition to identifying each transaction's amount and direction (inflow or outflow), this dataset contains two other classification fields with each transaction: *channel* and *category*. The channel refers to the medium through which the transaction took place; a complete list of possible channels are: Automated Clearing House (ACH), Adjustments to Account (ADJ), Automated Teller Machine (ATM), Online Bill Payment (BPY), Credit Card (CC), Check (CHK), Debit Card (DC), Account Fees (FEE), Interest on Account (INT), Wire Transfer (WIR), Internal Transfer (XFR), and Miscellaneous Deposit or Withdrawal (not in another channel) (MSC). There were also very rare occurrences of an Unidentifiable Channel (BAD).

The category classification is much broader and meant to capture the nature of the transaction. Examples of transaction categories include: restaurant expenditure, bar expenditure, grocery, food, etc. In the raw data, we observed 138 categories, which were produced by the Bank's data team based on the information that was associated with each transaction. Naturally, the level of accuracy and coverage varied greatly across channels. In particular, transactions taking place via Credit Card (CC) or Debit Card (DC) channels had much more accurate categories associated with them than those associated with ATM or Check transaction channels.

The sheer size and scope of this dataset proved to be a significant challenge, even for computing basic summary statistics. To develop some basic intuition for the data, we began by computing the total number of transactions per month, total inflows, and total outflows across all transactions for each individual in the database, and then constructed the same metrics for each of the following channels separately: Automated Clearing House (ACH), Automated Teller Machine (ATM), Credit Card (CC), Debit Card (DC), Interest on Account (INT), Online Bill Payment (BPY), and Wire Transfer (WIR). These channels were chosen because they represent the vast majority of the entire sample.

We also attempted to distill the most important information contained in the 138 categories by selecting a smaller set of data elements that broadly represent the key aspects of consumers' daily expenditures. In particular, Table 1 lists the subset of data items we selected from the transactions level data.[10] It should be emphasized that this data represents an incomplete picture of a typical consumer's expenditures. For example, the ultimate purposes of transactions that take place via ATM and Check channels are not typically known. Also, while employment-related payments (e.g., salary, bonus, etc.) are captured if they are directed via ACH channel, these payments are missed and not classified if an individual deposits a check or cash into his or her account. Similarly, any expense paid in cash is not captured in this data. Finally, an individual may have banking relationships with multiple institutions, hence the data from the Bank may be an incomplete view of the individual's financial activities. However, to the extent that such levels of incompleteness are stable through time, the trends observed in the data may still be used to make inferences and predictions about the changing behavior of the population. While a certain degree of caution is

Table 1
Subset of account balance and flow data collected and aggregated from a much larger set of channel and category data collected by the Bank. Due to legal restrictions, not all fields extracted are used, e.g., insurance, healthcare related, social security, collection agencies, unemployment.

Transaction data	
Transaction count	*By category (cont.)*
Total inflow	
Total outflow	Hotel expenses
	Travel expenses
By Channel:	Recreation
	Department store expenses
ACH (count, inflow and outflow)	Retail store expenses
ATM (count, inflow and outflow)	Clothing expenses
BPY (count, inflow and outflow)	Discount store expenses
CC (count, inflow and outflow)	Big box store expenses
DC (count, inflow and outflow)	Education expenses
INT (count, inflow and outflow)	Total food expenses
WIR (count, inflow and outflow)	Grocery expenses
	Restaurant expenses
By Category	Bar expenses
	Fast food expenses
Employment inflow	Total rest/bars/fast-food
Mortgage payment	Healthcare related expenses
Credit-card payment	Fitness expenses
Auto loan payment	Health insurance
Student loan payment	Gas stations expenses
All other types of loan payment	Vehicle expenses
Other line of credit payment	Car and other insurance
Brokerage net flow	Drug stores expenses
Dividends net flow	Government
Utilities payment	Treasury
TV	Pension inflow
Phone	Collection agencies
Internet	Unemployment inflow

[10] Note that not all of these items were used in creating our risk models because of legal restrictions (see Section 2.2 for further discussion). The list of data items used in our analysis is given in Section 4.2 and Table 4.

Table 2
Data items extracted from credit bureaus. Due to legal restrictions not all fields extracted are used, e.g., MSA, zip code, file age, and account-age data.

Individual level	Account level
Credit score	Age of Account
File age	Open/closed flag and date of closure
Bankruptcy (date and code)	Type (CC, MTG, AUT, etc.)
MSA and zip code	Balance
	Limit if applicable
	Payment status
	48-Month payment status history

warranted because of these issues, the apparent success of the statistical models created from this data (see Section 4.4 and 5) suggests that there is considerable incremental value for consumer credit-risk management even in this limited dataset.

2.1.2. Data from the credit bureaus

The Bank complements its account transactions data with credit file data provided by one of the credit bureaus for their customers. This data consists of variables listed under the "Individual Level" data items in Table 2. The generic credit score provided in this data, which we shall refer to as CScore, is similar in terms of statistical properties and predictive power to scores that are standard technique for measuring consumer credit quality.[11] We will use this credit score as a benchmark against which the performance of machine-learning models of credit risk can be compared.

Fig. 2 depicts the relationship between CScore for the months of January of 2005, 2006, 2007 and 2008 and the realized default frequency in the following 6 months. In each case, we consider the initial CScore for individuals and calculate the default frequency in the subsequent 6 months for each level of the score. In addition, we plot second-order-polynomial-fitted curves in each graph, and Fig. 2 shows that the CScore is remarkably successful in rank-ordering future default frequencies. In Fig. 3, we plot the four polynomial approximations to compare the performance across different market conditions. As economic conditions began deteriorating in 2007 and 2008, the curves started to shift to the right, indicating a higher absolute default rate for each level of the CScore. This is to be expected since the distribution of the CScore is static over time as shown in Fig. 4.

In addition, this data contains information regarding all credit or loan facilities, also referred to as "Trade Line", that each customer has across all financial institutions. For example, for a customer with a checking account with the Bank and a mortgage with another lender, this dataset would contain the mortgage balance, age, payment history and other relevant information without specifying the identity of the lender. Trade lines are divided into the following six types: Auto (AUT), Mortgage (MTG), Credit Card (CCA), Home Line of Credit (HLC), Home Loan (HLN), Other Loans (ILA), Other Lines of Credit (LOC), and Speciality type (SPC).

2.1.3. Account balance data

We also match the transaction-level and credit bureau data with information about account balances from checking accounts and CDs that a customer has with the Bank. These items are listed in Table 3.

2.2. Data security

Given the sensitivity of consumer data, a number of steps were taken to preserve the confidentiality of this data. First, the Bank de-identified all records—deleting all identification information such as names, addresses, social security numbers, and customer age—before they were transferred into databases to which we were given access. All computations were performed on dedicated MIT computers installed on site at the Bank's office, and all external network connections for these computers were disabled. At no point during this project was any raw consumer data transferred outside of the Bank's electronic firewalls.

Moreover, several data items were excluded from our models due to legal restrictions such as the Fair Credit Reporting Act. In particular, all healthcare, insurance, unemployment, government, treasury, account and file age information were excluded from our analysis (see Section 4.2 and Table 4 for the list of data items used in our analysis).

2.3. Data trends from 2005 to 2008

To provide some intuition for the empirical characteristics of the data, in this section we document several trends in the data over the period from January 2005 to December 2008. Since the main focus of our study is predictive models for credit-card defaults and delinquencies, we confine our attention in this section to observed trends in credit-card balances and payments.

Fig. 5(a) shows the median, 75th, 90th, and 95th percentiles of credit-card balances, based on the total balance across all credit cards. Because the data used in this calculation are taken from the matched sample obtained from credit bureaus (see Table 2), the balance represents the total balance across credit cards provided by all financial institutions, and is an accurate representation of the total credit-card debt of individuals. This data shows that the top 5% of the sample had credit-card debt of approximately $20,190 at the beginning of the sample, which increased to $27,612 by the end of the sample. The pattern of increasing credit-card indebtedness is clear, and even more extreme for other percentiles of the distribution as shown in Fig. 5(a). For example, the median credit-card balance almost doubled from $647 to $1259 over this period.

While increasing credit-card debt probably does not bode well for systemic risk, it need not represent a decline in the overall financial health of the consumer if the increased debt was accompanied by a comparable increase in income. For example, if increasing credit-card indebtedness is concurrent with a change in the usage among customers such that those consumers with high incomes—and who had not previously carried positive credit-card balances—started borrowing with their credit cards, this scenario may not represent a substantial change in systemic risk. If, on the other hand, customers increase their credit-card borrowing without any change in their income, the result would be an increase in systemic risk. To attempt to distinguish between these two scenarios, we plot the ratio of credit-card balances to monthly income in Fig. 5(b).[12] This figure shows that in the first half of the sample, from January 2005 to January 2006, the increase in credit-card debt documented in Fig. 5(a) was concurrent with increased monthly income or a shift in usage such that there was little or no change in the ratio of credit-card balances to income. However, starting in January 2007, a different pattern emerges—the ratio of the balances to monthly income starts to increase steadily. For example, for the top 5%, this ratio was approximately 11 for the first half of the sample but increased by about 2.5 to reach 13.5 over the period from January 2007 to December 2008.

While the trends shown in Fig. 5(a) and (b) point to an increasing level of credit-card debt, both in absolute terms and relative to income, over this period, this by itself may be the result of a substitu-

[11] See Avery et al. (2003) for an excellent review of consumer data.

[12] We have used the average of monthly income, as measured by the "Employment Inflow" data shown in Table 1, over the preceding 6 months, or as many month as available, as the denominator.

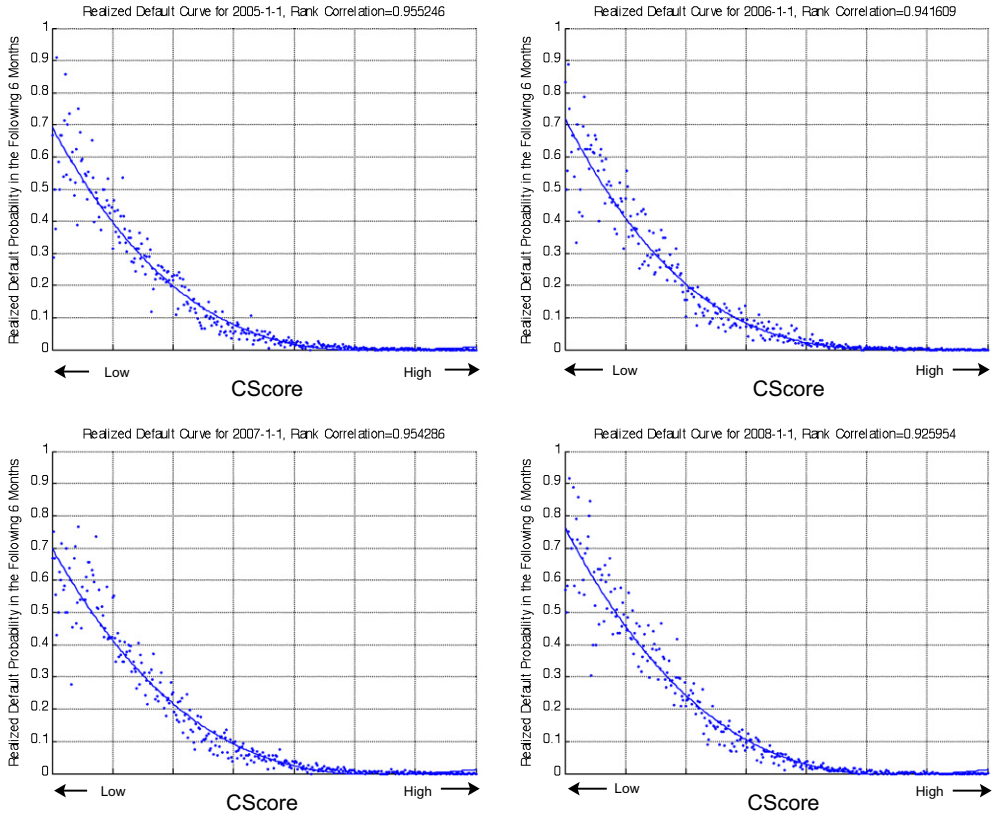


Fig. 2. Relationship between CScore and subsequent default frequency (over the next 6 months) for January of 2005, 2006, 2007, 2008. Second-order polynomial approximations are also shown (solid lines).

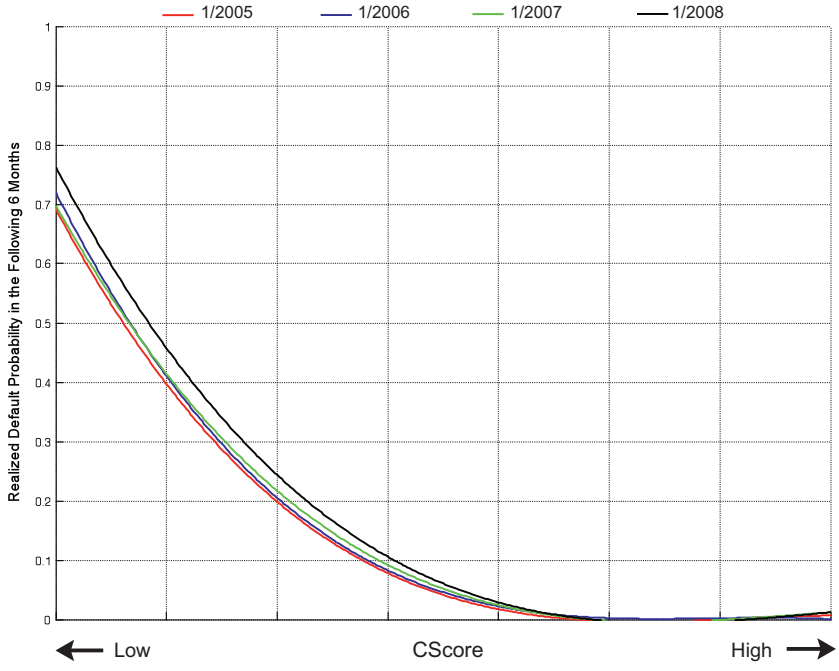


Fig. 3. Second-order polynomial approximation of the relationship between CScore and subsequent default frequency (over the next 6 months) for January of 2005, 2006, 2007, 2008.

tion effect where credit-card debt replaced another form of credit over this period. To explore this possibility, in Fig. 5(c) we depict the same statistics for credit-card payments during this period.[13] We see that credit-card payments exhibited a positive trend from January 2005 to July 2007, but the increase stopped after July 2007. Finally, Fig. 5(d) contains the ratio of credit-card payments to monthly income over this period, which shows that while credit-card payments were increasing as a percentage of monthly income for period of January 2005 to July 2007, this trend broke after July 2007.

[13] In particular, we use average credit-card payments over the preceding 6 months, or as many months as are available.

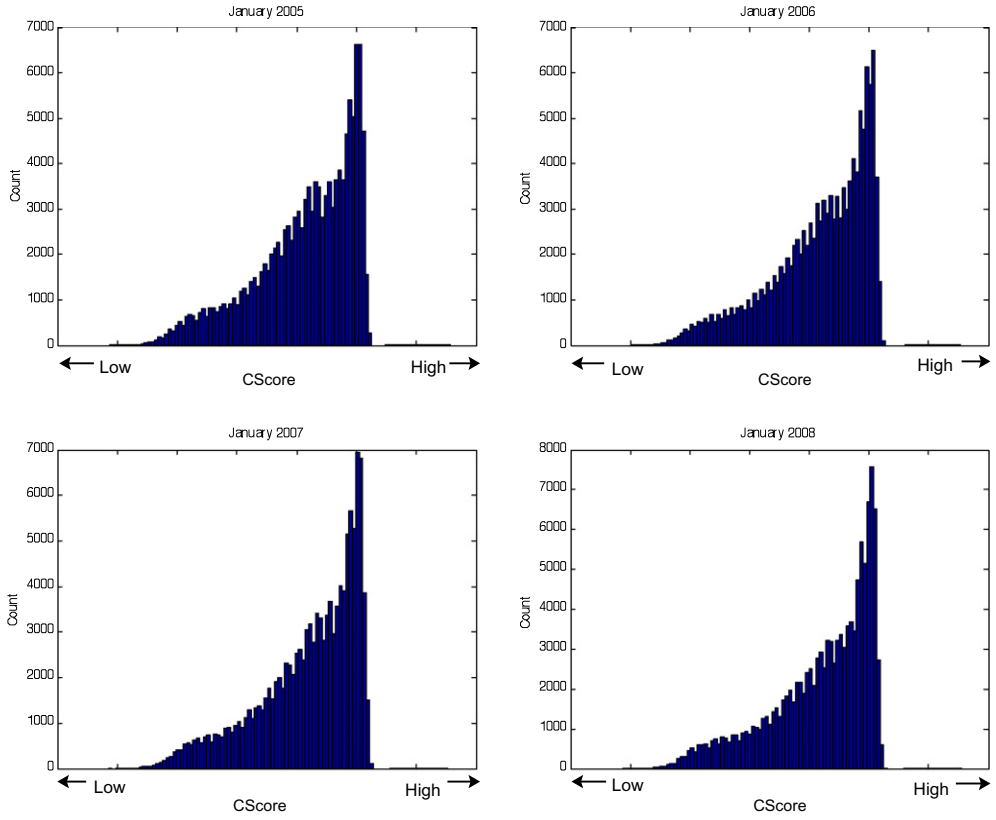


Fig. 4. Histograms of CScore for January of 2005, 2006, 2007, 2008.

Table 3
Balance data for accounts held with the Bank. Due to legal restrictions, not all fields extracted are used.

Deposit data	
Checking account balance	CD account balance
Brokerage account balance	IRA account balance
Savings account balance	

However, this pattern must be interpreted with some care due to certain data issues. In particular, as discussed above, the monthly income data is only captured by the Bank's transaction categorization engine if it comes through ACH (such as direct deposit), and there is no assurance that this system will capture an individual's entire income, even if all of it is deposited with the Bank (for example, income in the form of deposited checks would not be categorized as income). Furthermore, there may be some outliers, as in the case where individuals use their credit cards to purchase big-ticket items such as vacations or major appliances, in which case their credit-card balances can exceed their monthly income. Nevertheless, we believe that the patterns in Fig. 5(a)–(d) are still revealing with respect to the deteriorating financial condition of consumers in the months leading up to the financial crisis.

Fig. 6(a) and (b) contains other trends that confirm the developing problems in consumer credit during this period. Fig. 6(a) shows the ratio of monthly payments for credit cards, auto loans, student loans, and mortgages to monthly income, and the pattern is consistent with Fig. 5(d); monthly payments across all debt instruments were increasing as a percentage of monthly income (but again, we may not have identified all sources of an individual's income). To develop a better sense of these trends, Fig. 6(b) plots the ratio of debt payment to total monthly outflow over this period.[14] The total outflow provides a useful robustness check since the direction of each transaction (inflow or outflow) is known with certainty, so the total monthly outflow is a complete picture of all expenditures from all Bank accounts over the course of the month. The pattern that emerges from this analysis, see Fig. 6(b), supports the trends in the previous figures.

The patterns reported in this section are a very small subset of the many trends that are readily observed in this extremely rich dataset. Even based on this very limited set of metrics, it is clear that credit-card balances were increasing, both in absolute and relative terms, long before the credit crisis hit, and similar trends were apparent among a broader set of debt instruments. Such patterns suggest that predictive systemic risk indicators in the consumer lending business are feasible.

3. Constructing feature vectors

The objective of any machine-learning model is the identification of statistically reliable relationships between certain features of the input data and the target variable or outcome. In the models that we construct in later sections, the features we use include data items such as total inflow, total income, credit-card balance, etc., and the target variable is a binary outcome that indicates whether an account is delinquent by 90 days or more within the subsequent 3-, 6-, or 12-month window. In this section, we provide two illustrative examples of the relationship between certain features and subsequent delinquencies to develop intuition for both the data and our machine-learning algorithms. In Section 3.1, we consider high levels of the ratio of credit-card balances to income, and in Section 3.2, we explore the impact of sharp income declines.

3.1. High balance-to-income ratio

For each month in our sample (January 2005 to December 2008), we estimate the probability of 90-days-or-more delinquen-

[14] Total monthly outflow is measured as the average monthly outflow over the preceding 6 months, or as many months as available.

Table 4
The final inputs used in the machine-learning model of consumer credit risk.

Model inputs	
Credit bureau data	*Transaction data (cont.)*
Total number of trade lines	Total expenses at discount stores
Number of open trade lines	Total expenses at big-box stores
Number of closed trade lines	Total recreation expenses
Number and balance of auto loans	Total clothing store expenses
Number and balance of credit cards	Total department store expenses
Number and balance of home lines of credit	Total other retail store expenses
Number and balance of home loans	
Number and balance of all other loans	Total utilities expenses
Number and balance of all other lines of credit	Total cable TV & Internet expenses
Number and balance of all mortgages	Total telephone expenses
Balance of all auto loans to total debt	Total net flow from brokerage account
Balance of all credit cards to total debt	Total net flow from dividends and annuities
Balance of all home lines of credit to total debt	
Balance of all home loans to total debt	Total gas station expenses
Balance of all other loans to total debt	Total vehicle related expenses
Balance of all other lines of credit to total debt	
Ratio of total mortgage balance to total debt	Total lodging expenses
	Total travel expenses
Total credit-card balance to limits	
Total home line of credit balances to limits	Total credit-card payments
Total balance on all other lines of credit to limits	Total mortgage payments
	Total outflow to car and student loan payments
Transaction data	
	Total education related expenses
Number of Transactions	
Total inflow	*Deposit data*
Total outflow	
Total pay inflow	Savings account balance
	Checking account balance
Total all food related expenses	CD account balance
Total grocery expenses	Brokerage account balance
Total restaurant expenses	
Total fast-food expenses	
Total bar expenses	

cies on any of the credit-card accounts in the following 3 and 6 months for all customers. We then compute the same estimates for customers with credit-card-balance-to-income ratios greater than 10.[15] This threshold is chosen arbitrarily, and is meant to identify consumers experiencing great financial distress. As shown in Fig. 5(b), the top 5% of the sample have credit-card-balance-to-income ratios exceeding 10.

Fig. 7(a) and (b) contains the outcome of this stratification for both 3- and 6-month evaluation periods. Observe that the last month shown on the horizontal axis is June or September 2008 since we need 6 or 3 months, respectively, to calculate subsequent realized delinquency rates, and our sample ends in December 2008. Each axis also begins in June 2005 since we use the average income level over the prior 6 months for the denominator, hence we lose the first 6 months of the sample. This is consistent with the analysis presented in Fig. 5(a)–(d) and Fig. 6(a) and (b).

As shown in Fig. 7(a) and (b), customers with very high credit-card-balance-to-income ratios are much more likely to experience delinquencies in the subsequent 3- and 6-month periods. Furthermore, the separation has become wider in the latter part of the sample.

To evaluate the reliability of the relationship identified in Fig. 7(a) and (b), in Fig. 8(a) and (b) we report the same statistics but on the limited sample of customers who were current on *all* their credit-card accounts as of the conditioning date. The rate of future delinquency is lower (compare the blue lines[16] in Fig. 7(a) and (b) with those in Fig. 8(a) and (b)), but the gap between delinquency rates for typical customers and for highly indebted customers is still clearly apparent.

This example illustrates the presence of statistically stable relationships that can be exploited in creating a statistical model for predicting future delinquencies and defaults. However, since the gap between the future delinquency rates of average customers versus highly indebted customers has become wider in more recent periods, it is clear that the data contain important non-stationarities. Therefore, any machine-learning model of delinquencies and defaults must be re-calibrated frequently enough to capture such changes in a timely manner. We will discuss this issue in more detail in Section 4.

3.2. Negative income shocks

A sharp and sudden drop in income—most likely due to unemployment—is a plausible potential predictor of future delinquency and default, and we proceed in a fashion similar to Section 3.1 to evaluate this possibility. The challenge is to identify income shocks that are "significant", i.e., large relative to his/her historical levels and variability of income. To that end, in each month, we construct an income-shock variable by taking the difference between the current month's income and the 6-month moving-average of the income, and then divide this difference by the standard deviation of income over the same trailing 6-month window. By dividing the income drop by the estimated volatility, we are able to account for those individuals with jobs that produce more volatile income streams. We can stratify the sample according to this income-shock variable, and compare the 3- and 6-month subsequent 90-days-or-more delinquency rates for individuals with and without significant income shocks during each month in our sample (January 2005 to December 2008).

Fig. 9(a) and (b) plots the 90-days-or-more delinquency rates for all customers (blue) and for those customers with income-shock variables less than −2 (red). The gap between the delinquency rates of these two groups is clear. Furthermore, like the balance-to-income ratios of Section 3.1, this gap has also become wider over time, consistent with the patterns in Fig. 7(a) and (b) and Fig. 8(a) and (b).

4. Modeling methodology

In this section, we describe the machine-learning algorithms we use to construct delinquency forecast models for the Bank's consumer credit and transactions data from January 2005 to April 2009.[17] This challenge is well suited to be formulated as a supervised learning problem, which is one of the most widely used techniques in the machine-learning literature. In the supervised learning framework, a learner is presented with input/output pairs from past data, in which the input data represent pre-identified attributes to

[15] We define "credit-card balance" as the total credit-card balances across all financial institutions obtained from credit bureau data. Also, we measure the rate of 90-days-or-more delinquencies on any credit-card account across all financial institutions.

[16] For interpretation of color in Figs. 7–9 and 16, the reader is referred to the web version of this article.

[17] There is an extensive literature on machine learning, and we refer readers to Bishop (2006), Duda et al. (2000), and Vapnik (1998) for excellent overviews of this large and still-growing literature.

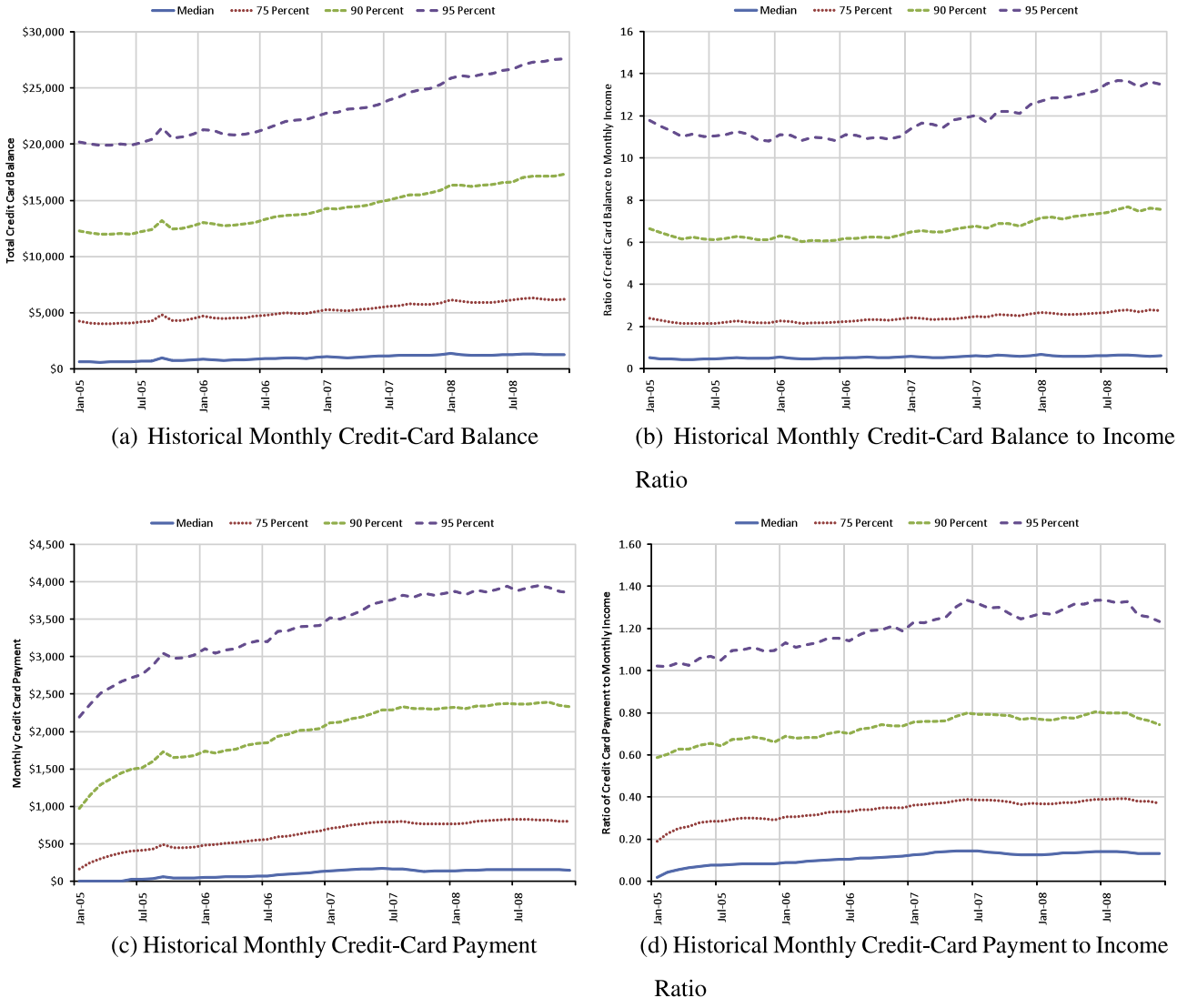


Fig. 5. Characteristics of credit-card balances, payments and their relationship to monthly income, based on data from January 2005 to December 2008. In Panel (a), the credit-card balance, which is the total balance across all credit cards of the customer in the month ending shown on the horizontal axis, is shown. In Panel (b), the ratio of credit-card balance to monthly income is shown. Panel (c) shows credit-card payments over this period while Panel (d) shows the ratio of credit-card payments to monthly income. For this calculation, the average income and credit-card payment over the 6 months preceding the date on the horizontal axis or as many months as available, is used for the calculation. See Section 2 for further details.

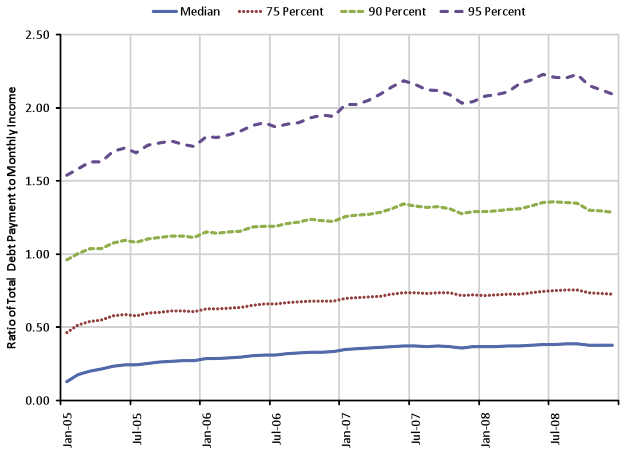
be used to determine the output value. Such input data are commonly represented as a vector and, depending on learning algorithm, can consist of continuous and/or discrete values with or without missing data. The supervised learning problem is referred to as a "regression problem" when the output is continuous, and as a "classification problem" when the output is discrete.

Once such input/output data are presented, the learner's task is to find a function that correctly maps such input vectors to the output values. One brute force way of accomplishing this task is to "memorize" all previous values of input/output pairs. Although this correctly maps the input/output pairs in the training data set, it is unlikely to be successful in forecasting the output values if the input values are different from the ones in the training data set, or when the training dataset contains noise. Therefore, the challenge of supervised learning is to find a function that generalizes beyond the training set, so that the resulting function also accurately maps out-of-sample inputs to out-of-sample outcomes.

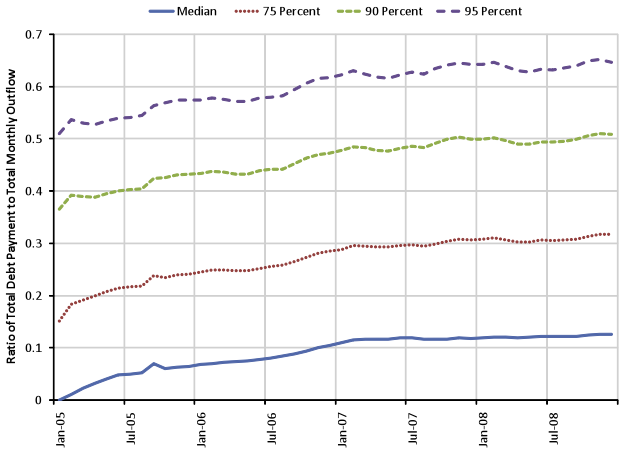
Using this machine-learning approach, we build a credit-risk forecast model that predicts delinquencies and defaults for individual consumers. For example, an output of our model is a continuous variable between 0 and 1 that can be interpreted (under certain conditions) as an estimate of the probability of 90-days-or-more delinquency sometime within the next 3 months of a particular credit-card account, given specific input variables from that account. Using this model, we address two questions:

1. How can this model be used to improve decisions about credit-line cuts?
2. What is the rate of credit-card delinquencies in the aggregate economy?

In addition, we need to take model accuracy into account in addressing these two questions, and more detailed discussion of this and other modeling issues is provided in Section 5. For both questions, selecting a parsimonious set of input variables that are relevant to credit risk is critical, and requires intuition for the underlying drivers of consumer credit, which cannot be resolved by the learning algorithm alone. The approach we described in Sec-

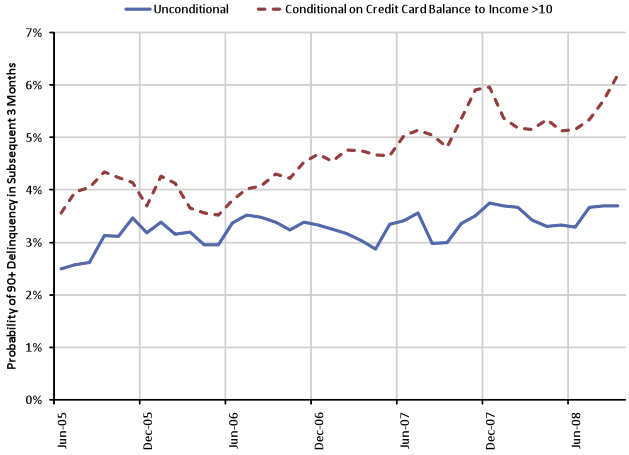


(a) Historical Monthly Ratio of All Debt Payments to Income

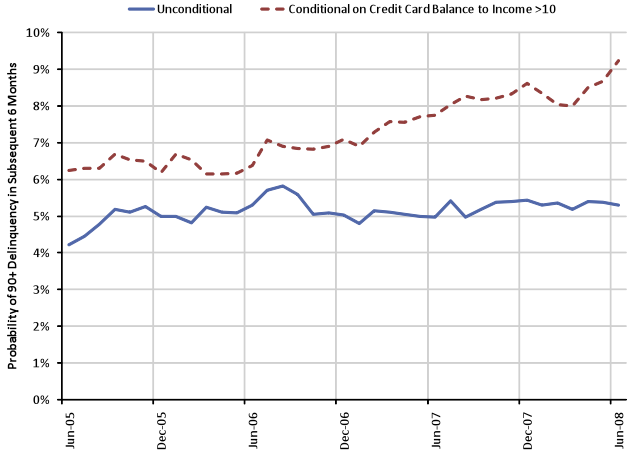


(b) Historical Monthly Ratio of All Debt Payments to Total Outflow

Fig. 6. Characteristics of payments of all forms of debt to monthly income and total monthly outflow based on data from January 2005 to December 2008. Payments for the following debt instruments are included in this calculation: auto loans, student loans, banks loans, mortgages, and credit cards. The average of monthly figures over 6 months preceding the date on the horizontal axis or as many months as available, is used for this calculation. See Section 2 for further details.



(a) Impact of High Balance-to-Income Ratios on Future 3-Month Delinquency Rates



(b) Impact of High Balance-to-Income Ratios on Future 6-Month Delinquency Rates

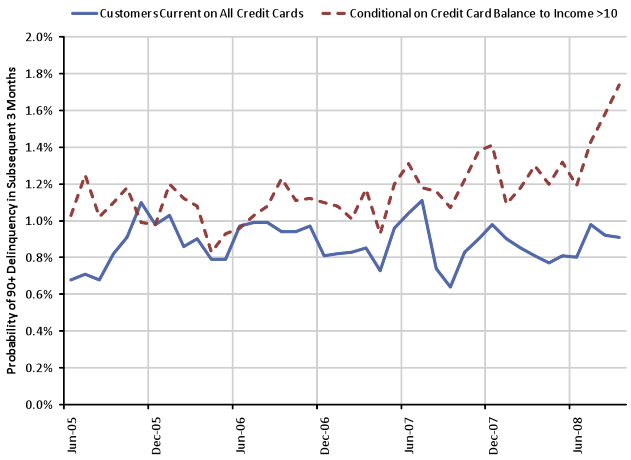
Fig. 7. Delinquency rates of customers with high credit-card-payment-to-income ratios. For each month, the rates of 90-days-or-more delinquencies during the following 3- or 6-month window for all customers as well as customers with total-credit-card-debt-to-income ratios greater than 10 are computed. The data covers the period from January 2005 to December 2008. A 6-month moving-average of income is the denominator used in calculating the ratio of credit-card balance to income, hence the horizontal axis starts in June 2005 and ends in June or September 2008 depending on whether 6- or 3-month forecast windows apply.

tion 3 can be used to identify such variables, and Fig. 10 summarizes our supervised learning framework.

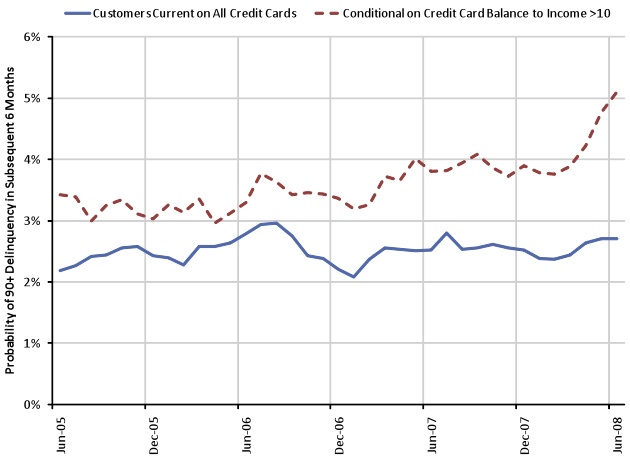
4.1. Machine-learning framework

We use generalized classification and regression trees (CART) (Breiman et al., 1984) to construct our forecast models. CART is a widely used statistical technique in which a dependent or "output" variable (either continuous or discrete) is related to a set of independent or "input" variables through a recursive sequence of simple binary relations (hence the reference to a "tree"). The collection of recursive relations partitions the multi-dimensional space of independent variables into distinct regions in which the dependent variable is typically assumed to be constant (classification tree) or linearly related (regression tree) to the independent variables with parameters unique to each region.

Fig. 11 provides an illustration of a CART model with two non-negative independent variables, $\{x_1,x_2\}$, also known as a "feature vector", and a discrete dependent variable that takes on two possible values, *good* and *bad*. The sequence of recursive binary relations represented by the tree in Fig. 11 partitions the domain of $\{x_1,x_2\}$ into five distinct regions determined by the parameters $L_1,\ldots,L_4$. In particular, this model implies that all realizations of $\{x_1,x_2\}$ in which $x_1 < L_1$ and $x_2 < L_2$ is associated with a *bad* outcome, and all realizations of $\{x_1,x_2\}$ in which $x_1 < L_1$ and $x_2 \geqslant L_2$ is associated with a *good* outcome. The parameters $\{L_j\}$ are chosen to minimize the distance between the dependent variable and the fitted values of the tree using an appropriate distance metric (typically

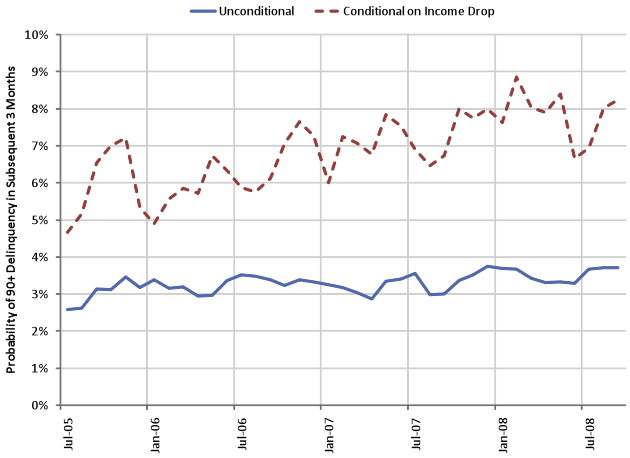


(a) Impact of High Balance-to-Income Ratios on Future 3-Month Delinquency Rates

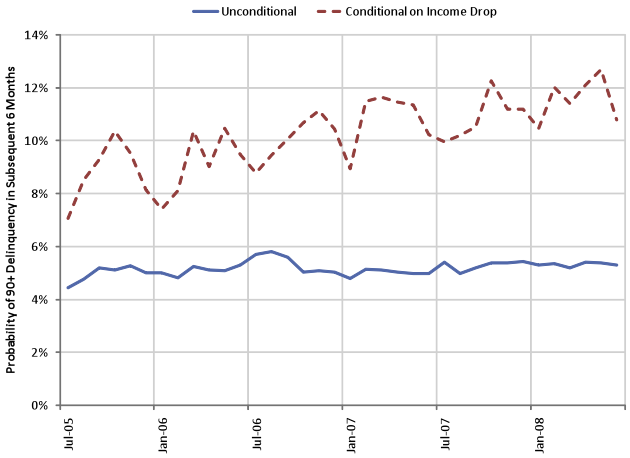


(b) Impact of High Balance-to-Income Ratios on Future 6-Month Delinquency Rates

Fig. 8. Delinquency rates for customers with high credit-card-payment-to-income ratios. For each month, the rates of 90-days-or-more delinquencies during the following 3- or 6-month window for customers who are current on all their credit-card accounts are computed, as well as customers who are current and have total-credit-card-debt-to-income ratios greater than 10. The data covers the period from January 2005 to December 2008. A 6-month moving-average of income is the denominator used in calculating the ratio of credit-card balance to income, hence the horizontal axis starts in June 2005, and ends in June or September 2008 depending on whether 6- or 3-month forecast windows apply.



(a) Impact of Income Ratio Drop on Future 3-Month Default Rates



(b) Impact of Income Ratio Drop on Future 6-Month Default Rates

Fig. 9. Delinquency rates for customers experiencing negative income shocks. For each month and customer, the income-shock variable is defined as the difference between the current month's income and the 6-month moving-average of income over the prior 6 months, divided by the standard deviation of income over that period. For each month, the rates of 90-days-or-more delinquencies during the following 3- or 6-month window for all customers as well as customers with income shocks less than −2 are calculated. The data covers the period from January 2005 to December 2008. Because of the 6-month moving-average for income, the horizontal axis starts in June 2005, and ends in June or September 2008 depending on whether 6- or 3-month forecast windows apply.

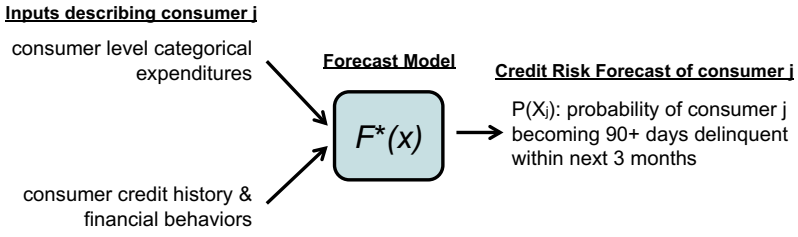


Fig. 10. A summary of the machine-learning algorithm used to construct the consumer credit-risk model.

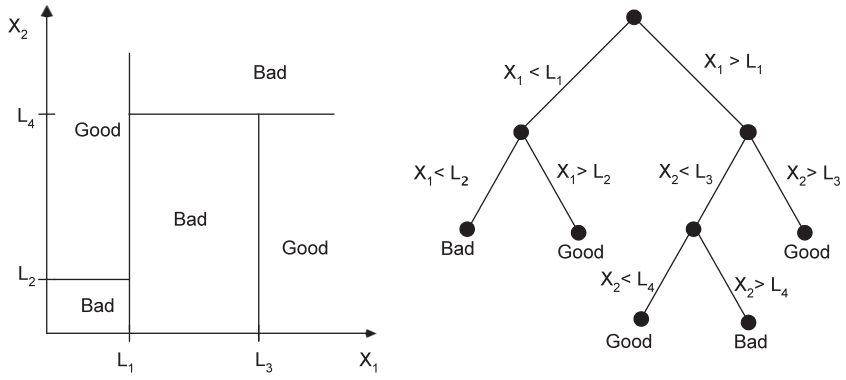


Fig. 11. An example of a CART model for a discrete dependent variable with two outcomes, *good* and *bad*, and two independent variables $\{x_1,x_2\}$.

mean-squared error), and forecasts may be readily constructed from new realizations of the feature vector $\{x'_1,x'_2\}$ by traversing the tree from its top-most node to the bottom.

The growing popularity of the CART model stems from the fact that it overcomes the limitations of standard models such as logit and probit in which the dependent variable is forced to fit a single linear model throughout the entire input space. In contrast, CART is capable of detecting nonlinear interactions between input variables, which dramatically increases the types of relations that can be captured and the number of independent variables that can be used. Moreover, CART models produce easily interpretable decision rules whose logic is clearly laid out in the tree. This aspect is particularly relevant for applications in the banking sector in which "black-box" models are viewed with suspicion and skepticism.

CART models can easily be applied to problems with high-dimensional feature spaces. Suppose we have N observations of the dependent variable $\{y_1,\ldots,y_N\}$ and its corresponding D-dimensional feature vectors $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$. We estimate the parameters of the CART model on the training dataset by recursively selecting features from $\boldsymbol{x}\in\{x_1,\ldots,x_D\}$ and parameters $\{L_j\}$ that minimize the residual sum-of-squared errors. Of course, we must impose a "pruning criterion" for stopping the expansion of the tree so as to avoid overfitting the training data. One of the most widely used measures for pruning is the *Gini* measure:

$$G(\tau)\equiv\sum_{k=1}^{K}P_\tau(k)(1-P_\tau(k)), \tag{1}$$

where τ refers to a leaf node of a CART model and $P_\tau(k)$ refers to the proportion of training data assigned to class k at leaf node τ. Then the pruning criterion for CART model T is defined as,

$$C(T)\equiv\sum_{\tau=1}^{|T|}G(\tau)+\lambda|T|, \tag{2}$$

where $|T|$ refers to a number of leaf nodes in CART model T, and λ refers to a regularization parameter chosen by cross validation. Once the pruning criterion reaches the minimum, the CART algorithm will stop expanding the tree.

An issue that is common among credit-default data is highly skewed proportion of *bad* and *good* realizations (credit defaults make up only 2% of our dataset). There are several ways to improve the forecast power of CART models in these cases, one of which is a technique called "boosting". Instead of equally weighting all the observations in the training set, we can weight the scarcer observations more heavily than the more populous ones (see Freund and Schapire, 1996). This "adaptive boosting" technique trains the model on examples with pre-assigned weights (initially equally weighted), and the weights are adjusted recursively as a function of the goodness-of-fit. In particular, the weight for observation i at the nth iteration is given by:

$$w_i^{(n)}=w_i^{(n-1)}\exp[\alpha_{n-1}I(f_{n-1}(\boldsymbol{x}_i)\neq y_i)] \tag{3}$$

and the data re-weighting coefficient α_{n-1} is defined as,

$$\alpha_{n-1}\equiv\ln\left(\frac{1-\epsilon_{n-1}}{\epsilon_{n-1}}\right), \tag{4}$$

where $I(\cdot)$ is an indicator function that indicates whether the model has correctly predicted the outcome y_i given the input vector $\boldsymbol{x}_i$, and ϵ_{n-1} is the weighted average error of the model from the $(n-1)$th iteration.

Related approaches have been used by Atiya (2001), Shin et al. (2005), and Min and Lee (2005), in which artificial neural networks (ANN) and support-vector machines (SVM) were applied to the problem of predicting corporate bankruptcies. Ong et al. (2005) conclude that genetic-algorithms-based consumer credit-scoring models outperform models based on ANNs, decision trees, rough sets, and logistic regression. Martens et al. (2007) introduce rule-extraction methods for SVMs that can generate consumer credit models with human-interpretable rules and little loss in accuracy. Galindo and Tamayo (2000) test CART decision-tree algorithms on mortgage-loan data to identify defaults, and also compare their results to k-nearest neighbor (KNN), ANNs, and probit models. Huang et al. (2004) provide an excellent survey of these and other related studies.

In the remainder of this section, we describe the attributes we selected for our inputs and discuss the model's performance. To identify the most powerful set of attributes, we combine bank transaction, consumer credit, and account balance databases and process them in parallel to construct our feature vectors (see Fig. 1).

4.2. Model inputs

Table 4 lists the data elements from the three sources discussed in Section 2.1 that were used as inputs to our delinquency prediction model. As discussed in Section 2.2 not all data elements can be used because of legal restrictions. We conducted extensive exploratory data analysis, similar to illustrative examples in Section 3, and finalized the set of input variables to be used in constructing our feature vectors listed in Table 4. For all transaction-related items, their average values over the prior 6 months, or as many months as available, are used.

4.3. Comparison to CScore

The empirical results of Section 2.1 show that the CScore is indeed useful in rank-ordering customers by their delinquency and default rates, hence it should serve as a reasonable benchmark for the machine-learning forecasts.

Fig. 12 compares the machine-learning forecasts of 90-days-or-more delinquencies of the Bank's credit-card accounts to CScore for December 2008. The machine-learning forecasts are constructed by training the model on delinquency data from October 2008 to December 2008 and using input data from September 2008 to construct the feature vectors, and then computing "fitted values" using features in December 2008, which yields December 2008 forecasts of delinquencies in the 3-month period from January to March 2009. CScore scores from December 2008 are plotted on the *x*-axis and the corresponding December 2008 machine-learning forecasts are plotted on the *y*-axis.

This figure shows that machine-learning forecasts differ substantially from CScore scores. In particular, there are a handful of accounts with relatively high CScore scores (note that higher CScore scores indicate higher credit quality or lower credit risk) that have high forecasted delinquency and default risk according to the machine-learning model. To provide further intuition for these conflicts, in Fig. 13 we reproduce the same plot but have included color coding for the status of each account as of the prediction date (December 2008)—green for accounts that are current (no late payments), blue for accounts that are 30-days delinquent, yellow for accounts 60-days delinquent, and red for accounts that are 90-days-or-more delinquent. Individuals with high CScore and high credit-risk forecasts are those already delinquent in their payments as of the prediction date (in some cases 90-days-or-more). However, those with low CScore but low forecasted credit risk are individuals that are current in their payments. While these patterns are not meant to be a conclusive comparison of the predictive ability of the two measures, Fig. 13 does suggest that machine-learning forecasts contain different information than traditional credit scores like CScore. We provide a more quantitative assessment in the next section.

4.4. Model evaluation

We now turn to a more rigorous evaluation of our models. We will focus on the model used to produce forecasts of 90-days-or-more delinquencies over subsequent 3-month windows for the period from January 2008 to April 2009. We choose this period because it includes the most severe deterioration in consumer credit quality in recent history, providing a natural laboratory for gauging the performance of machine-learning models during periods of financial dislocation. To minimize the effects of look-ahead bias, we only train the model based on delinquencies over 3-month windows that were observable at the time of the forecast. For example, the first model was calibrated using input data from January 2008 and realized delinquencies over the 3-month window from February to April 2008, and then the parameter estimates were applied to the input data in April 2008 to produce forecasts of delinquencies over the 3-month window from May to July 2008. This rolling-window approach yields 10 calibration and testing periods listed in Table 5, and sample sizes for each of the 10 prediction periods are reported in Table 6. The percentage of 90-days-or-more delinquent accounts varies from 2.0% to 2.5%.

The output of the machine-learning model is a continuous variable that, under certain conditions, can be interpreted as an estimate of the probability of an account becoming 90-days-or-more delinquent during the subsequent 3-month forecasting window. One measure of the model's success is the difference between the forecasts of those accounts that did become delinquent and those that did not; if accounts that subsequently became delinquent had the same forecasts as those that did not, the forecasts have no value. Table 7 reports the difference between the average forecast associated with accounts that did and did not fall into the 90-days-or-more delinquency category over the 10 evaluation periods. For example, during the period from May to July 2008,

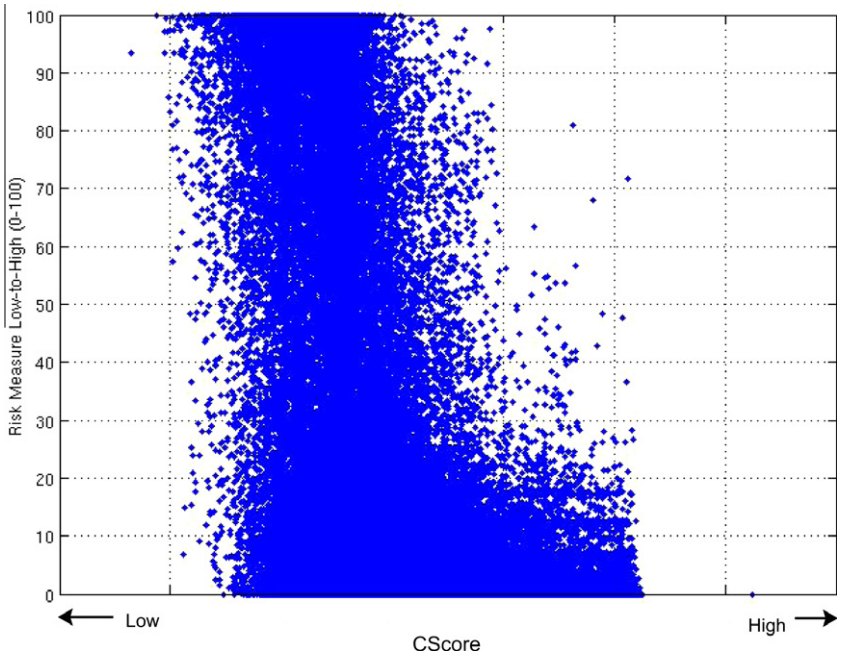


Fig. 12. Comparison of December 2008 CScore (*x*-axis, and where higher values are associated with lower credit risk) and machine-learning forecasts of 90-days-or-more delinquency rates over subsequent 3-month windows using realized delinquency events from October to December 2008 and September 2008 feature vectors. Once calibrated, the machine-learning model is applied to the December 2008 feature vector, and the resulting "fitted values" are plotted (*y*-axis) against the December 2008 CScores.

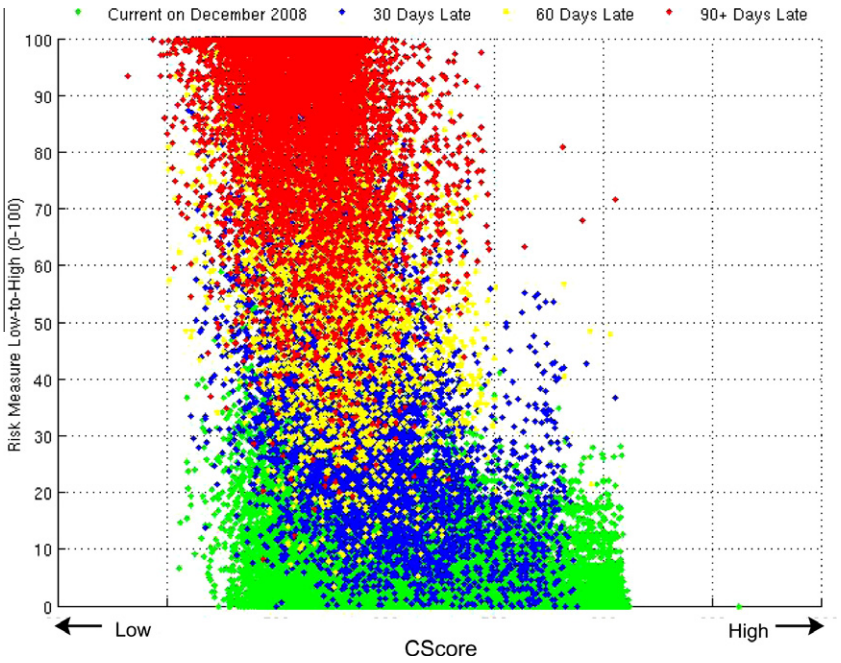


Fig. 13. Color-coded comparison of December 2008 CScore (*x*-axis, and where higher values are associated with lower credit risk) and machine-learning forecasts of 90-days-or-more delinquency rates over subsequent 3-month windows using realized delinquency events from October to December 2008 and September 2008 feature vectors. Once calibrated, the machine-learning model is applied to the December 2008 feature vector, and the resulting "fitted values" are plotted (*y*-axis) against the December 2008 CScores. The color coding indicates whether an account is current (green), 30-days delinquent (blue), 60-days delinquent (yellow), or 90-days-or-more delinquent as of December 2008.

the model's average forecast among the 2.4% of accounts (see Table 6) that did become 90-days-or-more delinquent was 61.2, whereas the average forecast among the 97.6% of accounts that did not fall into this category was only 1.0. We should emphasize that these forecasts are truly out-of-sample, being based on input data available at the end of April 2008, and calibrated using data from February to April 2008 (see Table 5). This implies significant forecast power in distinguishing between accounts that will and will not become delinquent over subsequent 3-month periods. Moreover, this forecast power does not diminish over the 10 calibration and evaluation periods, likely due to the frequent re-calibration of the model that captures some of the changing dynamics of consumer behavior documented in Section 3.

Another interesting challenge for the machine-learning model is the identification of "straight-rollers", accounts that are current as of the forecast date but become 90-days-or-more delinquent within the subsequent 3-month window. Such accounts typically do not become current after short delinquencies, but usually "roll straight" into the most extreme delinquency level that is commonly used as an indication of likely default and a charge-off loss for that bank. Recall from the examples of Section 3 (see, in particular, Fig. 8(a) and (b)) that certain input variables do seem to be related to future delinquencies, even among customers with current accounts as of the forecast date.

To evaluate the machine-learning model's ability to identify straight-rollers, we compare the model's average forecast among customers who were current on their accounts but became 90-days-or-more delinquent with the average forecast among customers who were current and did not become delinquent. Since this is a harder learning problem, we expect the model's performance to be less impressive than the values reported in Table 7. Nevertheless the values reported in Table 8 still indicate that the model is remarkably powerful, and is clearly able to separate the two populations. For example, using input data from April 2008, the average model forecast among customers who were current on their account and did not become 90-days-or-more delinquent from May to July 2008 was 0.7, while the average forecast for straight-rollers

Table 5
The 10 calibration and testing periods used in evaluating the performance of machine-learning forecasts of consumer credit risk.

Input date	Training period		Prediction date	Evaluation period	
	Start date	End date		Start date	End date
Jan-08	Feb-08	Apr-08	Apr-08	May-08	Jul-08
Feb-08	Mar-08	May-08	May-08	Jun-08	Aug-08
Mar-08	Apr-08	Jun-08	Jun-08	Jul-08	Sep-08
Apr-08	May-08	Jul-08	Jul-08	Aug-08	Oct-08
May-08	Jun-08	Aug-08	Aug-08	Sep-08	Nov-08
Jun-08	Jul-08	Sep-08	Sep-08	Oct-08	Dec-08
Jul-08	Aug-08	Oct-08	Oct-08	Nov-08	Jan-09
Aug-08	Sep-08	Nov-08	Nov-08	Dec-08	Feb-09
Sep-08	Oct-08	Dec-08	Dec-08	Jan-09	Mar-09
Oct-08	Nov-08	Jan-09	Jan-09	Feb-09	Apr-09

Table 6
Sample sizes of accounts and 90-days-or-more delinquencies for each of the 10 3-month evaluation windows from May 2008 to April 2009. This data is based on Bank-issued credit cards only.

Three months starting date	Three months ending date	Customers going 90 + days delinquent (%)	Customers NOT going 90 + days delinquent (%)
May-08	Jul-08	2.4	97.6
Jun-08	Aug-08	2.2	97.8
Jul-08	Sep-08	2.1	97.9
Aug-08	Oct-08	2.0	98.0
Sep-08	Nov-08	2.1	97.9
Oct-08	Dec-08	2.1	97.9
Nov-08	Jan-09	2.1	97.9
Dec-08	Feb-09	2.3	97.7
Jan-09	Mar-09	2.6	97.4
Feb-09	Apr-09	2.5	97.5

Table 7
Performance metrics for machine-learning forecasts over 10 3-month evaluation windows from May 2008 to April 2009. For each evaluation period, the model is calibrated on credit-card delinquency data over the 3-month period specified in the *Training Window* columns, and predictions are based on the data available as of the date in the *Prediction Date* column. For example, the first row reports the performance of the model calibrated using input data available in January 2008 and credit-card delinquency data from February to April 2008, and applied to the April 2008 feature vector to generate forecasts of delinquencies from May to July 2008. Average model forecasts over all customers, and customers that (*ex post*) did and did not become 90-days-or-more delinquent over the evaluation period are also reported.

Training window		Prediction date	Evaluation window		Average predicted probability of 90 + delinquency on credit card in the next 3 months		
Start	End		Start	End	Among all customers	Among customers going 90 + days delinquent	Among customers NOT going 90 + days delinquent
Feb-08	Apr-08	Apr-08	May-08	Jul-08	2.5	61.2	1.0
Mar-08	May-08	May-08	Jun-08	Aug-08	2.0	62.1	0.6
Apr-08	Jun-08	Jun-08	Jul-08	Sep-08	1.9	60.4	0.7
May-08	Jul-08	May-08	Aug-08	Oct-08	1.9	62.5	0.6
Jun-08	Aug-08	Aug-08	Sep-08	Nov-08	2.0	62.4	0.7
Jul-08	Sep-08	May-08	Oct-08	Dec-08	2.1	63.6	0.8
Aug-08	Oct-08	Oct-08	Nov-08	Jan-09	2.1	62.5	0.8
Sep-08	Nov-08	May-08	Dec-08	Feb-09	2.2	59.8	0.8
Oct-08	Dec-08	Dec-08	Jan-09	Mar-O9	2.4	60.8	0.8
Nov-08	Jan-09	May-08	Feb-09	Apr-O9	2.4	62.8	0.9

Table 8
Performance metrics for machine-learning forecasts over 10 3-month evaluation windows from May 2008 to April 2009 for "straight-rollers", customers who are current as of the forecast date but who become 90-days-or-more delinquent in following 3-month window. For each evaluation period, the model is calibrated on credit-card delinquency data over the 3-month period specified in the *Training Window* columns and predictions are based on the data available as of the date in the *Prediction Date* column, using only customers who are current as of the forecast date. For example, the first row shows the performance of the model calibrated using input data available in January 2008 and credit-card delinquency data from February to April 2008, and applied to the April 2008 feature vector to generate forecasts about delinquencies from May to July 2008.

Training window		Prediction date	Evaluation window		Average predicted probability of 90 + delinquency on credit card in the next 3 months		
Start	End		Start	End	Among all customers	Among customers going 90 + days delinquent	Among customers NOT going 90 + days delinquent
Feb-08	Apr-08	Apr-08	May-08	Jul-08	0.7	10.3	0.7
Mar-08	May-08	May-08	Jun-08	Aug-08	0.4	8.0	0.4
Apr-08	Jun-08	Jun-08	Jul-08	Sep-08	0.5	9.7	0.4
May-08	Jul-08	May-08	Aug-08	Oct-08	0.4	8.6	0.3
Jun-08	Aug-08	Aug-08	Sep-08	Nov-08	0.4	9.6	0.4
Jul-08	Sep-08	May-08	Oct-08	Dec-08	0.5	9.5	0.4
Aug-08	Oct-08	Oct-08	Nov-08	Jan-09	0.5	10.1	0.4
Sep-08	Nov-08	May-08	Dec-08	Feb-09	0.5	10.6	0.5
Oct-08	Dec-08	Dec-08	Jan-09	Mar-09	0.5	8.6	0.5
Nov-08	Jan-09	May-08	Feb-09	Apr-09	0.5	10.2	0.5

was 10.3. And as before, the degree of separation is remarkably consistent across the 10 evaluation periods listed in Table 8.

5. Applications

In this section we apply the models and methods of Sections 3 and 4 to two specific challenges in consumer credit-risk management: deciding when and how much to cut individual-account credit lines, and forecasting aggregate consumer credit delinquencies for the purpose of enterprise-wide and macroprudential risk management.

The former application is of particular interest from the start of 2008 through early 2009 as banks and other financial institutions responded to the developing crisis with massive risk reductions. In Section 5.1, we use input data from December 2008 to make forecasts about individual-account delinquencies from January to March 2009, which are then compared with realized delinquencies during that period to evaluate the out-of-sample predictive power of our machine-learning model.

In Section 5.2, we investigate the robustness of machine-learning forecasts by comparing their performance across subsets of the data stratified by the number of features available for each account. We find that delinquencies among accounts with more features are significantly easier to predict than those with few features. This difference underscores the potential value of the conditioning information we have constructed from the combined data items in Table 4.

In Section 5.3, we combine individual-account forecasts to produce an aggregate forecast of consumer credit delinquencies, which we propose as a new measure of systemic risk. Despite the fact that our data represents a small subset of the Bank's customer base, the machine-learning model is surprisingly effective in capturing the changing risk characteristics of the population. In fact, our results indicate near-monotonic increases in aggregate consumer credit risk from January 2005 to April 2009, suggesting that consumer-based measures of financial distress may play an important role in macroprudential risk management.

5.1. Credit-line risk management

Once a line of credit is extended to a customer by the Bank, the Bank actively monitors the creditworthiness of the customer. The credit line may be increased or the interest rate reduced in the case of improving creditworthiness, while the opposite may occur if there is a decline in the customer's credit profile. Therefore, the particular application considered in this section is the primary focus of any consumer credit business. For the business to be successful, credit-line decisions must be based on a model that captures changes in a customer's risk profile in a timely and accurate manner. In fact, during the sample period of our dataset, the major decision facing banks and other consumer lenders was identifying high-risk customers and taking the appropriate actions with respect to their credit lines, i.e., lowering credit limits, increasing interest rates, or some combination of both. However, to simplify

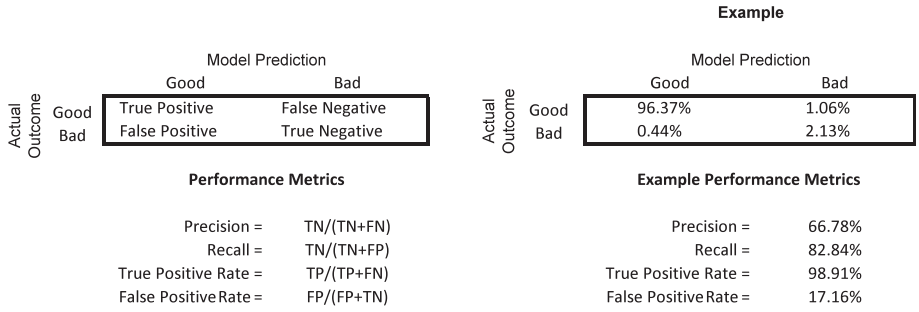


Fig. 14. Confusion matrix framework, model performance metrics, and numerical example based on December 2008 machine-learning forecasts of 90-days-or-more delinquencies over the 3-month forecast horizon from January to March 2009. Rows correspond to actual customer types ("Bad" is defined as 90-days-or more delinquent) and columns correspond to forecasted types. TN: True Negative, FN: False Negative, FP: False Positive; TP: True Positive.

our analysis, we assume that once a customer is identified as high-risk, his/her remaining (unused) credit line will be reduced to zero.

As noted in Section 4, the type of problem in which the target variable or output is binary is known as a *classification* problem in the machine-learning literature. Although the forecasts produced by the model described in Section 4 can be interpreted, under certain assumptions, as estimated delinquency probabilities, these numerical scores can be easily converted to binary decisions by comparing the forecast to a specified threshold and classifying customers with scores exceeding that threshold as high-risk. Setting the appropriate level for this threshold involves a trade-off. A very low threshold results in many customers being classified as high-risk, and while it may correctly identify customers who are genuinely high-risk and about to fall behind on their payments, this choice may also result in many low-risk customers incorrectly classified as high-risk. On the other hand, a high threshold may allow too many high-risk customers to be classified as low-risk, yielding greater subsequent delinquencies and losses.

This type of trade-off is present in any classification problem, and involves the well-known statistical balancing act between Type-I (false positives) and Type-II (false negatives) errors in classical hypothesis testing contexts. In the more practical context of credit-risk management, the trade-off can be made explicitly by a cost/benefit analysis of false positives versus false negatives, and selecting the threshold that will optimize some criterion function in which such costs and benefits are inputs. Before turning to this cost/benefit analysis, we wish to develop further intuition for the statistical properties of this trade-off.

To understand the statistical behavior of any classification algorithm, one commonly used performance metric in the machine-learning and statistics literatures is a 2×2 contingency table often called the "confusion matrix" for reasons that will become obvious once we define its entries (see the left side of Fig. 14).[18] The two rows correspond to the two types of customers in our sample, low-risk or "Good" and high-risk or "Bad", or *ex post* realizations of delinquency. In our application, we define Good customers as those who did not become 90-days-or-more delinquent during the forecast period, and Bad customers as those who did. The two columns correspond *ex ante* classifications of the customers into these same two categories, Good and Bad. When any forecast model is applied to a given set of customers, each customer falls into one of these four cells, hence the performance of the model can be gauged by the relative frequencies of the entries (see Fig. 14). In the traditional Neyman–Pearson hypothesis-testing framework, the lower-left entry is defined as Type-I error and the upper right as Type-II error, and the objective of the classical statistician is to find testing procedures

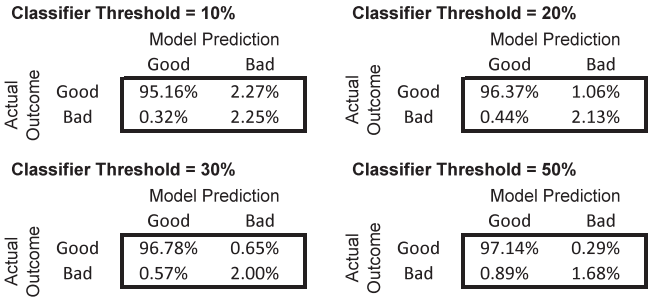


Fig. 15. Confusion matrices of machine-learning forecasts of 90-days-or-more delinquencies for four classification thresholds. Rows correspond to actual customer types ("Bad" is defined as 90-days-or more delinquent). The numerical example is based on the December 2008 model forecast for the 3-month forecast horizon from January to March 2009.

to reduce Type-II error (or maximize "power") subject to a fixed level of Type-I error (or "size").

The right side of Fig. 14 contains the confusion matrix for our machine-learning model for December 2008 using a 3-month forecasting horizon from January to March 2009 and a threshold of 20%, i.e., accounts with estimated delinquency probabilities greater than 20% are classified as "Bad" and 20% or below as "Good". For this month, the machine-learning model classified 96.37% + 0.44% = 96.81% of the accounts as good, of which 96.37% were, in fact, good and 0.44% of which were bad, i.e., were 90-days-or-more delinquent sometime during January to March 2009. Similarly, of the 0.44% + 2.13% = 2.57% that became 90-days-or-more delinquent during the 3-month forecast period, the model correctly identified 2.13% of the accounts, a correct-classification rate of 82.9%. Of course this success comes at the cost of labeling 1.06% of customers who ended up not falling into the 90-days-or-more delinquency state as bad.

Four specific performance metrics are typically computed from the entries of the confusion matrix, and their formulas are given in Fig. 14. "Precision" refers to the model's accuracy in instances that the model classified an account as bad, while "Recall" refers to the number of bad accounts identified by the model divided by the actual number of bad accounts. "True positive rate" and "false positive rate" are self-explanatory. Ideally, we would like our model to have very high precision and recall, or, equivalently, to have a high "true positive rate" and a low "false positive rate". The role of the classification threshold is now clear: if a lender wishes to aggressively identify bad accounts by using a lower threshold, such aggressiveness will result in more incorrect classifications of good accounts as bad, leading to the opportunity cost of lost interest income from cutting the credit lines of low-risk customers. Fig. 15 illustrates such trade-offs for four different threshold levels.

[18] A confusion matrix is typically used in supervised learning to describe actual and predicted classifications by the learning model. See Kohavi and Provost (1998) for a formal definition of this and other machine-learning terminology.

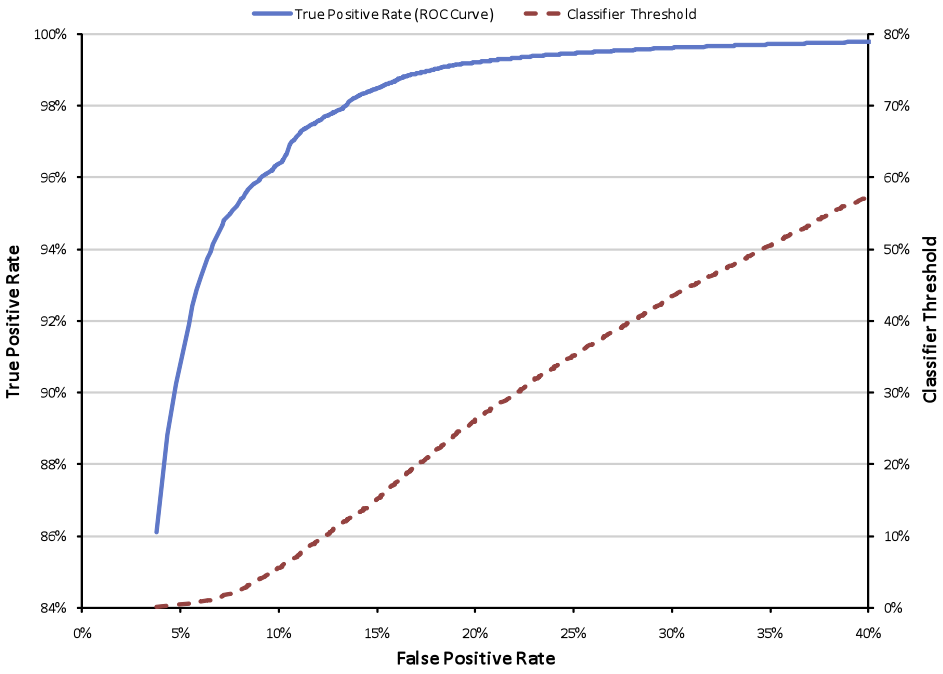


Fig. 16. A Receiver Operating Characteristic (ROC) curve and corresponding classification threshold value of December 2008 machine-learning forecasts of 90-days-or-more delinquencies over the 3-month forecast horizon from January to March 2009.

By applying our forecast model with varying levels of the classification threshold, we can trace out explicitly the trade-off between true and false positives, and this trade-off is plotted in Fig. 16. The blue line, known as the Receiver Operating Characteristic (ROC) curve in the machine-learning literature, is the pairwise plot of true and false positive rates for various classification thresholds (green line), and as the threshold increases, Fig. 16 shows that the true positive rate increases, but the false positive rate increases as well.

The ROC curve shows that the trade-offs are not linear, meaning that increase in true positive rates is not always commensurate with the increase in false positive rates. If the cost of false positives is equal to the gain of true positives, the optimal threshold will correspond to the tangent point of the ROC curve with the 45° line. If the cost/benefit trade-off is not one-for-one, the optimal threshold is different, but can easily be determined from the ROC curve. We have provided a simple optimization framework in the appendix for making such trade-offs.

Table 9 presents a set of standard performance statistics widely used in machine-learning literature (see Appendix A for their definitions) for each of the 12 models discussed in Section 4 (see Table 5).[19] Our models exhibit strong predictive power across the various performance metrics. The two simplest measures reported are the percentage of correctly and incorrectly classified observations, and the reported values show that our models routinely achieve approximately 99% correct-classification rates. However, the correct-classification rate is not necessarily a sufficient measure of a model's predictive power when the distribution of the dependent variable is highly skewed. Therefore, we report several more-sophisticated performance measures to capture other aspects of our model's predictive power.

The kappa statistic is a well-known relative measure of predictive success in classification problems, where success is measured relative to purely random classifications. According to Landis and Koch (1977), a kappa statistic between 0.6 and 0.8 represents "substantial", and a value greater than 0.8 implies "almost perfect" agreement. Table 9 shows that our model's kappa statistic ranges from 0.72 to 0.83, indicating strong agreement between real and forecasted values, even after correcting for agreement due to chance. Both mean-absolute-error and root-mean-squared-error measures also indicate significant predictive power, as do the *F*-measures which are the harmonic means of the model's precision and recall (see Fig. 14), which are 0.73 or greater, implying a good balance between high precision and high recall.

Finally, the area under the ROC curve is a widely used measure in the machine-learning literature for comparing models (see Fig. 16), and can be interpreted as the probability of the classifier assigning a higher score, i.e., a higher probability of being of type-1, to a type-1 observation than to a type-0 observation. The ROC area of our model ranges from 0.89 to 0.95, which again shows that our machine-learning classifiers have strong predictive power in separating the two classes.

5.2. Robustness

Although the dataset used in our analysis is from a large consumer bank in the US, it does not cover the entirety of a consumer's spending and saving activities. Consumers usually have business relationships with multiple banks and credit-card companies, hence it is rare that all elements of the feature vectors we described in Section 4.2 are available for a given account. To check the robustness of our model to missing features, we divide our data set into equally sized samples separated by the availability of features. More sophisticated analysis can be performed—using feature-selection techniques widely used in machine learning—to assign weights on features by their importance, however in this study, we focus on the simpler approach of rank-ordering the data by simply counting the number of missing values, without differentiating among the features that are missing.

After dividing the dataset in this manner, for each group we perform 10-fold cross validation (CV), i.e., stratifying the dataset into 10 bins, using 9 of them for training of the model and the

[19] The reported results are for classifiers based on the ROC-tangency classification threshold. In particular, for each model we constructed the ROC curve, determined the threshold that corresponded to the 45° tangency line, and used that threshold to classify observations into the two possible classes.

Table 9
Standard performance statistics for machine-learning classification models of consumer credit risk. Each model is calibrated based on the prior 3 months of data (*Training Period*) and applied to the input data available on the *Prediction Date*. The results of classifications versus actual outcomes over the following 3 months (*Evaluation Period*) are used to calculate these performance metrics. See also Table 5 for the list of training and evaluation periods.

Model for the prediction date of	Apr-08	May-08	Jun-08	Jul-08	Aug-08	Sep-08
Panel A: performance metrics for April to September 2008						
Correctly classified instances (rate)	0.989	0.991	0.991	0.991	0.991	0.991
Incorrectly classified instances (rate)	0.011	0.009	0.009	0.009	0.009	0.009
Kappa statistic	0.751	0.753	0.735	0.751	0.751	0.749
Mean absolute error	0.006	0.004	0.004	0.004	0.004	0.004
Root mean squared error	0.075	0.061	0.062	0.061	0.062	0.065
TP rate	0.997	0.998	0.998	0.998	0.997	0.997
FP rate	0.312	0.345	0.348	0.322	0.319	0.309
Precision	0.839	0.896	0.856	0.853	0.849	0.828
Recall	0.688	0.656	0.652	0.678	0.682	0.692
F-measure	0.756	0.757	0.740	0.755	0.756	0.754
ROC area	0.952	0.927	0.933	0.937	0.940	0.935
Model for the prediction date of	Oct-08	Nov-08	Dec-08	Jan-09	Feb-09	Mar-09
Panel B: performance metrics for October 2008 to March 2009						
Correctly classified instances (rate)	0.990	0.989	0.988	0.989	0.992	0.992
Incorrectly classified instances (rate)	0.010	0.011	0.012	0.011	0.008	0.008
Kappa statistic	0.741	0.726	0.735	0.745	0.832	0.830
Mean absolute error	0.004	0.005	0.005	0.005	0.006	0.005
Root mean squared error	0.065	0.067	0.069	0.070	0.075	0.073
TP rate	0.997	0.997	0.997	0.997	0.995	0.993
FP rate	0.320	0.356	0.346	0.320	0.129	0.036
Precision	0.827	0.848	0.854	0.839	0.803	0.734
Recall	0.680	0.644	0.654	0.680	0.871	0.964
F-measure	0.746	0.732	0.741	0.751	0.836	0.834
ROC area	0.937	0.921	0.929	0.924	0.944	0.895

remaining one for testing, and repeating this step 9 times using a different set of bins for testing and training. Fig. 17 contains the average confusion matrix (see Section 5.1) across the 10-fold CV results for the 90-days-or-more delinquency forecasts over a 6-month forecast horizon. The panels are labeled Group I through Group IV where Group I contains accounts with the largest number of missing features, and Group IV contains accounts with the fewest.

This analysis shows a significant improvement in both precision and recall values from Group I to Group II. Group I customers are typically single-account holders, i.e., individuals with only a checking account and a low number of transactions, also known as "thin-file" customers because their files contain few records. However, as a customer's file becomes "thicker" with more transactions and accounts, the ability of machine-learning models to forecast future delinquencies improves, presumably because of the additional features available for such customers.

To get a sense to the economic significance of the improvement possible with this model, Fig. 18 plots the Value Added (VA) measure (B.4) that is developed in Appendix B,[20] estimated via the 10-fold CV performed on four different forecasts of 90-days-or-more delinquencies over forecast horizons of 3, 6, 9, and 12 months. As expected, the model's forecast accuracy becomes better as the prediction horizon lengthens because the forecast problem is easier, e.g., predicting 90-days-or-more delinquency over a 12-month horizon is considerably easier than over a 3-month horizon. Also, across all four models, there is a significant improvement in VA from Group I to Group II, whereas the performance of the model in Groups III and IV are comparable to that of Group II, with Group IV yielding slightly lower performance.[21]

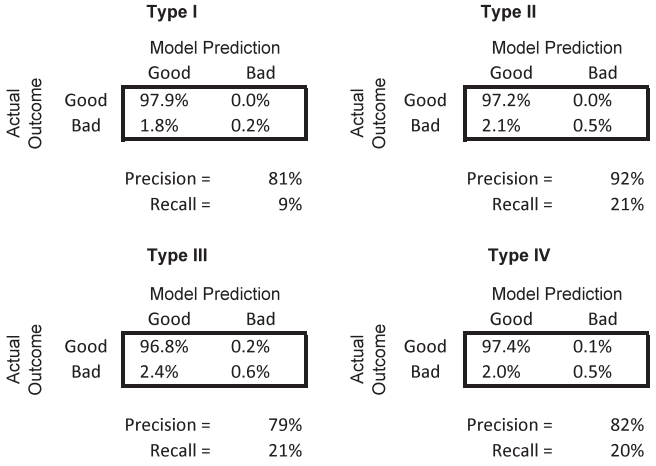


Fig. 17. Average confusion matrices, averaged over 10 trials of machine-learning forecasts of 90-days-or-more delinquency over a 6-month forecast horizon on data from January 2005 to April 2008, with each trial based on a randomly chosen 10% subset of the entire dataset. For each trial, four models are estimated, one for each of four stratified sub-samples of equal size, stratified by on the number of non-missing features. Group I is the sample of accounts with the fewest available features ("thinnest"), and Group IV is the sample with the most ("thickest").

5.3. Macroprudential risk management

The role of the banking sector in the current financial crisis has underscored the need for better risk measures among large financial institutions (see Pérignon and Smith (2010) for a recent critique of traditional measures reported by commercial banks). Machine-learning forecasts for individual accounts can easily be aggregated to generate macroeconomic forecasts of credit risk in the consumer lending business by simply tabulating the number of accounts forecasted to be delinquent over a given forecast period and for a specific classification threshold. The proportion of forecasted delinquencies in the population may then be used as an indicator of systemic risk for consumer lending, and its properties

[20] We assume an amortization period $N = 3$ years, a run-up before default of 30% ($B_d = 1.3B_r$), and an interest rate differential ($r = 5\%$).

[21] One conjecture for Group IV's decline may be self-rationing behavior of borrowers with the "thickest" files, i.e., those with the fewest missing features. These individuals are likely to have the longest relationships with the Bank and, according to Han et al. (2009), may be most easily discouraged from applying for new loans and reduce their borrowing voluntarily.

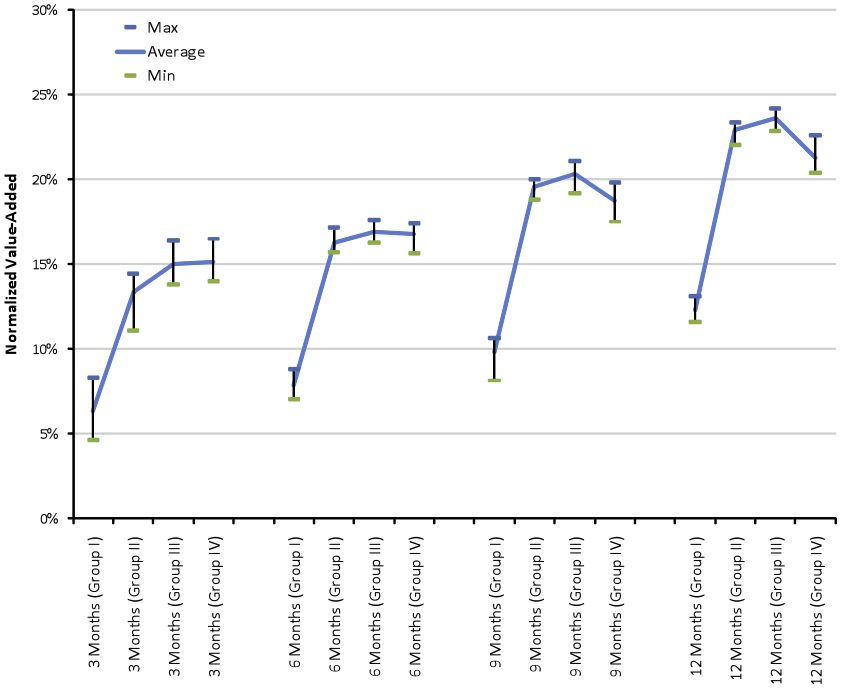


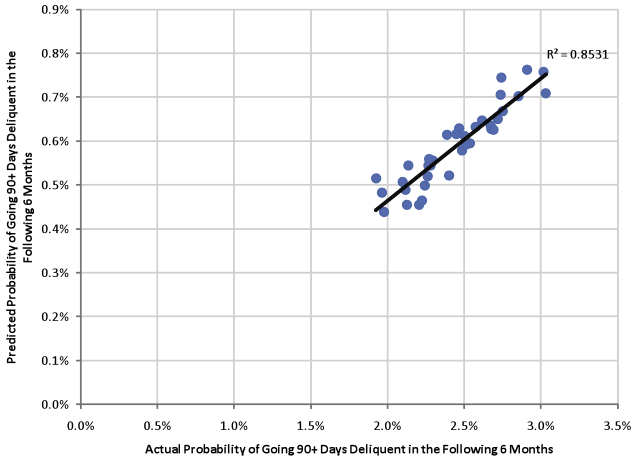
Fig. 18. Value added of machine-learning forecasts of 90-days-or-more delinquency over 3-, 6-, 9-, and 12-month forecast horizons on data from January 2005 to April 2008, for four equally stratified sub-samples of randomly chosen 10%-subsets of the data, stratified by the number of non-missing features. Group I is the sample of accounts with the fewest available features ("thinnest"), and Group IV is the sample with the most ("thickest"). Minimum and maximum VA values are calculated using 10-fold cross validation from the entire dataset using Eq. (B.4) with amortization period $N = 3$ years, run-up before default of 30% ($B_d = 1.3B_r$), $r = 5\%$, and the ROC-tangency classification threshold.

can also be summarized with confusion matrices with different classification thresholds, and incorporated into enterprise risk management protocols such as that of Drehmann et al. (2010).

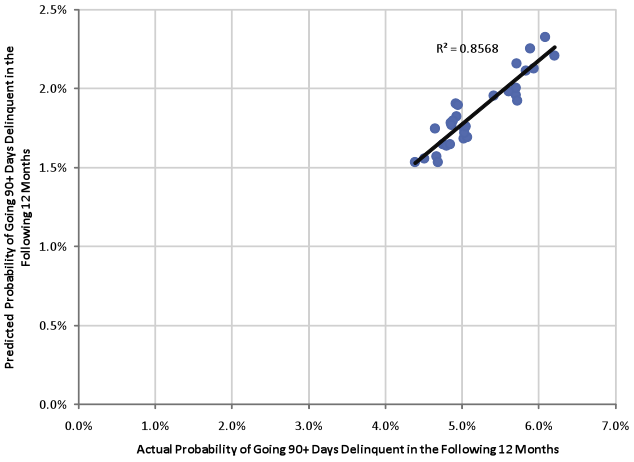
In this section, we construct such aggregate delinquency probabilities in the year prior to the credit crisis of 2007–2009, and show that this approach yielded increasingly higher forecasted delinquency probabilities, even during the boom years. This estimated trend suggests that machine-learning techniques applied to combined transactions and credit scores are capable of generating leading indicators of deterioration in consumer creditworthiness.

For purposes of aggregate risk measurement, a longer forecast horizon may be more appropriate, so we apply the same machine-learning algorithm of Section 4 to data from January 2005 to December 2008 and define the target variable as 90-days-or-more delinquency over a 6- or 12-month forecast horizon. Fig. 19(a) and (b) shows the aggregate results of the machine-learning forecasts as compared to realized delinquencies for 6- and 12-month forecast horizons using the ROC-tangency classification threshold (in which the trade-off between true and false positive rates is one-to-one; see Section 5.1). The forecasted delinquencies are highly correlated with realized delinquencies—with linear regression R^2's of 85% for both horizons—hence our forecasts are capturing the dynamics of the consumer credit cycle over the sample.

However, note that the forecasts consistently under-estimate the absolute level of delinquencies by a scalar multiple, which may be an artifact of the particular classification threshold se-

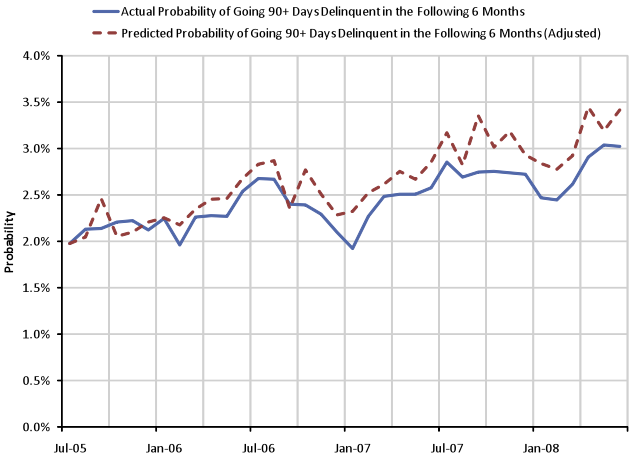


(a) Predicted and actual 90-days-or-more delinquency rates (6-month)

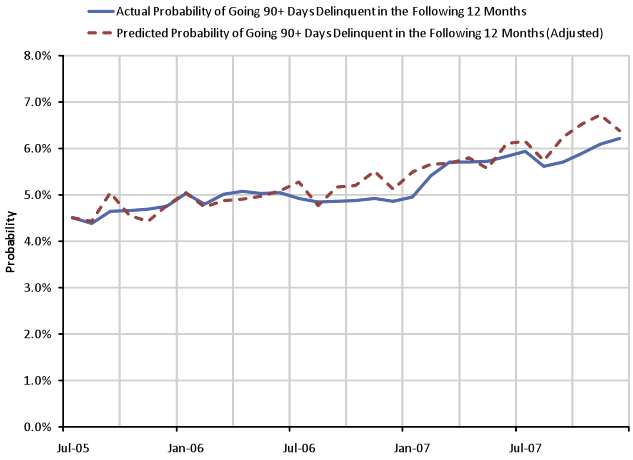


(b) Predicted and actual 90-days-or-more delinquency rates (12-month)

Fig. 19. Forecasted and realized 90-days-or-more delinquency rates for 6- and 12-month forecast horizons, with ROC-tangency classifier threshold, based on data from January 2005 to December 2008. The horizontal axis starts in July 2005 since some features require 6 months of training data, and ends June 2008 or December 2007 since either 6 or 12 months of data are needed to calculate the realized delinquencies events to be compared again the predicted values.



(a) Time series of actual and predicted 90-days-or-more delinquency rates (6-month)



(b)Time series of actual and predicted 90-days-or-more delinquency rates (12-month)

Fig. 20. Time series of predicted and realized 90-days-or-more delinquency rates for 6- and 12-month forecast horizons, with ROC-tangency classifier threshold, based on data from January 2005 to December 2008. The horizontal axis starts in July 2005 since some features require 6 months of training data, and ends June 2008 or December 2007 since either 6 or 12 months of data are needed to calculate the realized delinquencies events to be compared again the predicted values.

lected. To correct for this underestimation, and to reduce the effects of overfitting, we use 10-fold cross validation to estimate the model's recall value (the number of accounts classified as bad as a percentage of the total number of bad accounts). Then we simply multiply the forecasted default rate by the reciprocal of this estimated recall rate. Fig. 20(a) and (b), which contain time series plots of the adjusted forecasts as well as the realized aggregate delinquency rates, shows that the adjustment factor is reasonably successful in correcting the underestimation bias, yielding forecasts that are accurate both in terms of level and dynamics.

6. Conclusion

In the aftermath of one of the worst financial crises in modern history, it has become clear that consumer behavior has played a central role at every stage—in sowing the seeds of crisis, causing cascades of firesale liquidations, and bearing the brunt of the economic consequences. Therefore, any prospective insights regarding consumer credit that can be gleaned from historical data has become a national priority.

In this study, we develop a machine-learning model for consumer credit default and delinquency that is surprisingly accurate in forecasting credit events 3–12 months in advance. Although our sample is only a small percentage of the Bank's total customer base, the results are promising. Our out-of-sample forecasts are highly correlated with realized delinquencies, with linear regression R^2's of 85% for monthly forecasts over 6- and 12-month horizons. Moreover, crude estimates of the value added of these forecasts yield cost savings ranging from 6% to 23% of total losses. Given the scale of industry wide charge-offs during 2008, even a 6% cost savings would amount to hundreds of millions of dollars.

From a macroprudential risk management perspective, the aggregation of machine-learning forecasts of individuals may have much to contribute to the management of enterprise and systemic risk. Current credit bureau analytics such as credit scores are based on slowly varying consumer characteristics, and therefore are not as relevant for tactical risk management decisions by chief risk officers and policymakers. We find that machine-learning forecasts are considerably more adaptive, and are able to pick up the dynamics of changing credit cycles as well as the absolute levels of default rates.

We believe that our results are indicative of considerably more powerful models of consumer behavior that can be developed via machine-learning techniques, and are exploring further refinements and broader datasets in ongoing research.

Appendix A. Definitions of performance metrics

Suppose a binary classifier has been applied to a sample of N observations. For each observation i, the model produces the probability that the observation with feature vector $\boldsymbol{x}_i$ belongs to class 1. This predicted probability, $f(\boldsymbol{x}_i)$ is then compared to a threshold and observations are classified into class 1 or 0 based on the result of that comparison.[22]

For a given level of threshold, let True Positive (TP) be the number of instances that are actually of type 0 that were correctly classified as type 0 by the classifier, False Negative (FN) be the number of instances that are actually of type 0 but incorrectly classified as type 1, False Positive (FP) be the number of instances that are of type 1 but incorrectly classified as type 0 and, finally, True Negative (TN) be the number of instances that are of type 1 and correctly classified as type 1 (see Fig. 14). Then one can define the following metrics to evaluate the accuracy of the classifier:

$$\text{Correctly classified instances (rate)} \equiv (\text{TP}+\text{TN})/N,$$
$$\text{Incorrectly classified instances (rate)} \equiv (\text{FP}+\text{FN})/N,$$
$$\text{kappa statistic} \equiv (P_a - P_e)/(1-P_e), \quad P_a \equiv (\text{TP}+\text{TN})/N,$$
$$P_e \equiv [(\text{TP}+\text{FN})/N] \times [(\text{TP}+\text{FN})/N],$$
$$\text{Mean absolute error} \equiv 1/N \sum_{i=1}^{N} |f(\boldsymbol{x}_i) - y_i|,$$
$$\text{Root mean squared error} \equiv 1/N \sum_{i=1}^{N} (f(\boldsymbol{x}_i) - y_i)^2,$$
$$\text{True positive (TP) Rate} \equiv \text{TP}/(\text{TP}+\text{FN}),$$
$$\text{False positive(FP) Rate} \equiv \text{FP}/(\text{FP}+\text{TN}),$$
$$\text{Precision} \equiv \text{TN}/(\text{TN}+\text{FN}),$$
$$\text{Recall} \equiv \text{TN}/(\text{TN}+\text{FP}),$$
$$F\text{-Measure} \equiv (2 \times \text{Recall} \times \text{Precision})/(\text{Recall}+\text{Precision}),$$

[22] As discussed in Stein (2005), such hard cut-offs are not optimal and a more flexible pricing approach may improve performance. A successful application must also take into account the underlying economic drivers of the credit-card business model (see Scholnick et al., 2008), and incorporate the effects of pricing on consumer incentives and decision-making processes (see Simon et al., 2010).

where $f(\boldsymbol{x}_i)$ is the model's predicted probability that a customer with feature vector $\boldsymbol{x}_i$ is of type 1.

Appendix B. Framework for calculating value added

In this section, we provide a simple framework for analyzing the value added of classification-based algorithms for credit-line reduction. This framework requires some simplifying assumptions regarding the revenues and costs of the consumer lending business. As a starting point, suppose that in the absence of any forecasts, a lender will take no action with respect to credit risk, implying that customers who default or are delinquent will generate losses for the lender, and customers who are current in their payments will generate financing fees on their running balances. To make the analysis simpler, assume that all non-defaulting customers have a running balance of B_r, and defaulting customers also start with a balance of B_r but will increase their balance to B_d between the point at which the lender's credit-risk management decision is made and the point when they default. For example, an empirically plausible set of parameters are B_r = \$1000 and B_d = \$1200. As it will be clear shortly, only the ratio of run-up balance, $B_d - B_r$, and the running balance, B_r, enters into the potential savings calculation. In this particular example, the critical factor is the 20% run-up in the balance before default.

While the lender's loss from default is simply B_d, calculating the profit due to a customer in good-standing requires further discussion. For now, let the present value of the profit from such a customer be given by a fraction of the running balance, denoted by P_m for "Profit Margin Rate", which can be a function of the interest rate and the expected length of time the account holder will use this credit line.[23]

Given these assumptions, we can calculate the profit with and without a forecast model as follows:

$$\text{Profit without forecast} = (\text{TP} + \text{FN}) \cdot B_r \cdot P_m - (\text{FP} + \text{TN}) \cdot B_d,$$
$$\text{Profit with forecast} = \text{TP} \cdot B_r \cdot P_m - \text{FP} \cdot B_d - \text{TN} \cdot B_r.$$

Then the dollar savings is simply given by the difference:

$$\text{Savings} = \text{TN}(B_d - B_r) - \text{FN} \cdot B_r \cdot P_m, \tag{B.1}$$

where

$$\text{TN}\ (B_d - B_r) = \text{Savings due to correct decision},$$
$$\text{FN} \cdot B_r \cdot P_m = \text{Opportunity cost due to incorrect decision}.$$

We can simplify this even further by dividing the savings by the savings that would have been possible under the perfect-foresight case, i.e., the case in which all bad customers, and only bad customers, are identified and their credit was reduced. We will call this ratio the "value added" (VA) of the forecast, which is given by:

$$\text{Value added} \equiv \frac{\text{TN} - \text{FN} \cdot P_m \cdot \frac{B_r}{B_d - B_r}}{\text{TN} + \text{FP}}. \tag{B.2}$$

Note that the final expression depends on the running and default balance only through the ratio of the running balance to the difference between the default and running balance. For example, it is clear from (B.1) that if this difference is zero, no value added can be achieved, which is expected because it is only the difference between B_d and B_r that can potentially be captured by a correct and timely decision by the lender.

Before attempting to evaluate VA, we must first specify how the profit margin rate, P_m, is calibrated. This factor is meant to capture the total financing charge that customers pay to the lender on their outstanding balances. This value depends on the credit-card interest rate as well as the speed with which customers pay off their balances. For simplicity, we assume that the difference between the credit-card interest rate and the lender's funding cost is a fixed rate given by r, and customers pay back their entire credit-card debt over an N-year period. In this simple setting, the financing charge is an annuity hence P_m is given by:[24]

$$P_m = r \cdot \frac{1 - (1 + r)^{-N}}{r} = 1 - (1 + r)^{-N}. \tag{B.3}$$

Combining (B.2) and (B.3) yields the following expression for VA:

$$\text{VA}(r, N, \text{TN}, \text{FN}, \text{FP}) = \frac{\text{TN} - \text{FN} \cdot [1 - (1 + r)^{-N}] \cdot [\frac{B_d}{B_r} - 1]^{-1}}{\text{TN} + \text{FP}}. \tag{B.4}$$

[23] For simplicity, we make the assumption that all "good" customers who are subjected to a line reduction immediately pay off their balances and close their accounts. To the extent that most customers do not behave in this manner (see, for example, Zinman, 2009), our estimates of the potential cost savings are conservative.

[24] In the interest of parsimony and simplicity, we have implicitly assumed that the lender's funding cost is also r. For example, if r = 5% implies that the lender pays 5% for funds and charges an interest rate of 10% on credit-card balances.

References

Agarwal, S., Chomsisengphet, S., Liu, C., Souleles, N., 2009. Benefits of Relationship Banking: Evidence from Consumer Credit Markets? Working Paper No. 1440334, SSRN.

Atiya, A., 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE Transactions on Neural Networks 12, 929–935.

Avery, R., Bostic, R., Calem, P., Canner, G., 2003. An Overview of Consumer Data and Reporting. Federal Reserve Bulletin, Board of Governors of the Federal Reserve System.

Avery, R.B., Calem, P.S., Canner, G.B., 2004. Consumer credit scoring: Do situational circumstances matter? Journal of Banking and Finance 28, 835–856.

Bellotti, T., Crook, J., 2009. Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications 36, 3302–3308.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer-Verlag, New York.

Boot, A., 2000. Relationship banking: What do we know? Journal of Financial Intermediation 9, 7–25.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth and Brooks Cole Advanced Books and Software, Pacific Grove, CA.

Buyukkarabacaka, B., Valevb, N.T., 2010. The role of household and business credit in banking crises. Journal of Banking & Finance 34, 1247–1256.

Drehmann, M., Sorensen, S., Stringa, M., 2010. The integrated impact of credit and interest rate risk on banks: A dynamic framework and stress testing application. Journal of Banking & Finance 34, 713–729.

Dwyer, D.W., Stein, R.M., 2006. Inferring the default rate in a population by comparing two incomplete default databases. Journal of Banking & Finance 30, 797–810.

Duda, R., Hart, P., Stork, D., 2000. Pattern Classification. Wiley-Interscience, New York.

Foster, D., Stine, R., 2004. Variable selection in data mining: Building a predictive model for bankruptcy. Journal of the American Statistical Association 99, 303–313.

Freund, Y., Schapire, R., 1996. Experiments with a new boosting algorithm. Proceedings of the Thirteenth International Conference on Machine Learning, 148–156.

Galindo, J., Tamayo, P., 2000. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. Computational Economics 15, 107–143.

Han, L., Fraser, S., Storey, D.J., 2009. Are good or bad borrowers discouraged from applying for loans? Journal of Banking & Finance 33, 415–424.

Hand, D., Henley, W., 1997. Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society 160, 523–541.

Huang, C., Chen, M., Wang, C., 2006. Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications 33, 847–856.

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

Kohavi, R., Provost, F., 1998. Glossary of terms. Machine Learning 30, 271–274.

Landis, R., Koch, G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Li, S., Shiue, W., Huang, M., 2006. The evaluation of consumer loans using support vector machines. Expert Systems with Applications 30, 772–782.

Martens, D., Baesens, B., Gestel, T., Vanthienen, J., 2007. Comprehensive credit scoring models using rule extraction from support vector machines. European Journal of Operational Research 183, 1466–1476.

Min, J., Lee, Y., 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications 28, 603–614.
Ong, C., Huang, J., Tzeng, G., 2005. Building credit scoring systems using genetic programming. Expert Systems with Applications 29, 41–47.
Pérignon, C., Smith, D., 2010. The level and quality of value-at-risk disclosure by commercial banks. Journal of Banking & Finance 34, 362–377.
Scholnick, B., Massoud, N., Saunders, A., Carbo-Valverde, S., Rodriguez-Fernandez, F., 2008. The economics of credit cards, debit cards and ATMs: A survey and some new evidence. Journal of Banking & Finance 32, 1468–1483.
Shin, K., Lee, T., Kim, H., 2005. An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications 28, 127–135.
Simon, J., Smith, K., West, T., 2010. Price incentives and consumer payment behaviour. Journal of Banking & Finance 34, 1759–1772.
Stein, R.M., 2005. The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. Journal of Banking & Finance 29, 1213–1236.
Thomas, L., 2009. Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press, New York.
Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience, New York.
Zinman, J., 2009. Debit or credit? Journal of Banking & Finance 33, 358–366.