Predicting bank failures using a simple dynamic hazard model

Rebel A. Cole *
College of Business
Florida Atlantic University
Boca Raton, FL 33431 USA
Email: coler@fau.edu
Tel.: 1-561-297-4969

Jon Taylor
College of Business
Florida Atlantic University
Boca Raton, FL 33431 USA
Email: jtaylo98@fau.edu
Tel.: 1-561-297-2607

Qiongbing Wu*
School of Business,
The University of Western Sydney
Penrith South DC, NSW 2751, Australia
Email: q.wu@uws.edu.au
Tel.: 61-2-96859151

Abstract: We compare the out-of-sample accuracy of three methodologies—the time-varying hazard model of Shumway (2001), the static probit model used by Cole and Gunther (1998), and a static logistic regression model similar to Cole and White (2012)—in forecasting U.S. bank failures. When we limit all models to financial data available at the time of prediction, we find that the logistic and probit models outperforms the hazard model, indicating that the superior performance of hazard models documented in previous empirical research is attributable to use of more timely financial data rather than to incorporation of time-varying covariates. We also find that the logistic model outperforms the probit model. Finally, we also find that a parsimonious specification fit to data over 1985-1993 performs well in forecasting bank failures during 2009-2012—evidence that the characteristics of "distressed banks" have experienced little change over the past two decades despite substantial changes in structure and regulation of the U.S. banking industry. Our findings support supervision focusing on banks' traditional CAMELS risk ratios. We also add to the literature finding that simpler models outperform more complex models in out-of-sample forecasting.

Keywords: banking crisis, bank failure, bank supervision, CAMELS, failure prediction, financial crisis, hazard model, logit model, offsite monitoring probit model

JEL Classifications: G17, G21, G28,

* *Corresponding author.* We are grateful to seminar participants at a FDIC's Center for Financial Research and at the University of Newcastle; and to session participants at the 22nd Australasian Finance and Banking Conference in Sydney, Australia. Earlier versions of this paper were distributed under the titles "Is hazard or probit more accurate in predicting financial distress? Evidence from US bank failures" and "Hazard versus probit in predicting US bank failures: a regulatory perspective over two crises," as well as under the current title, which is the original. Wu acknowledges the financial support from the University of Newcastle and the University of Western Sydney.

Predicting bank failures using a simple dynamic hazard model

1. Introduction

Recent empirical research strongly supports the theoretical view that a well-functioning banking system is a critically important determinant of a country's economic growth (Levine 2005). Because the banking system plays such an important role in a country's economic development, a banking crisis can generate an independent negative real effect (Dell' Ariccia *et al* 2008; Campello *et al*. 2010) and cause serious disruptions of a country's economic activities (Hoggarth *et al*. 2002; Boyd *et al*. 2005; Hutchison and Noy 2005; Serwa 2010; and Kupiec and Ramirez 2013).

During the previous decade, a housing boom-and-bust in the U.S. led to massive losses on mortgages and mortgage-backed securities, which were magnified by leverage from derivatives—primarily credit default swaps—tied to those securities. These losses forced regulators to seize a large number of banks and other financial institutions, both in the U.S. and in other countries around the world, and subsequently led to an international freeze-up in credit markets and a serious global recession (see, e.g., Ivashina and Scharfstein, 2010). How to differentiate sound banks from troubled banks in order to ensure that the banking sector continues to provide credit to the private sector is a primary concern of both policy makers and bank regulators around the world. Consequently, development of more effective statistical models for predicting future bank failures, i.e., early warning models, would prove of great value in preventing future crises.

In this study, we develop three statistical models of bank-failure: one based upon the time-varying hazard model as proposed by Shumway (2001); a second based upon the simple

static probit model [1] similar to the one used by Cole and Gunther (1998); and the simple static logit model similar to the one used by Cole and White (2012). We then compare the out-of-sample forecasting accuracy of these three alternatives using a number of different measures in order to identify the best tool for this task. We also compare the early warning indicators of U.S. bank failures based upon the three alternative methodologies, i.e., what variables are statistically significant in a model based upon each methodology. The years 1985 – 1993, during which more than a thousand U.S. banks failed and with more than 100 failures in almost every year, provide a rich sample period for conducting our comparison of forecasting accuracy.[2] We then combine the coefficients estimated from this period with bank financial information as of year-end 2008 to test how accurately these models perform in predicting bank failures over the period 2009 – 2012, a period when more than 400 banks failed. This is an out-of-sample test based upon two waves of bank failures separated by two decades and provides powerful evidence on the robustness of the two methodologies in predicting bank failure.

Most previous studies of bank failures rely upon bank-level accounting data, occasionally augmented with market-price data (e.g., Meyer and Pifer 1970; Martin 1977; Pettway and Sinkey 1980; Cole and Gunther 1995; Cole and White 2012; Shaffer 2012). These studies aim to develop models of an early warning system for individual bank failure. The indicators of these early warning models are closely related to supervisory rating system of banks. The most widely known rating system for banks is the CAMELS system, which is an acronym for the system's six component—*C*apital Adequacy, *A*sset Quality, *M*anagement, *E*arnings, *L*iquidity, and *S*ystemic

[1] King, Nuxoll and Yeager (2006) provide a survey of research on bank-failure early warning models and also provide a summary of early warning models used by the three major bank regulators in the U.S.: the Office of the Comptroller of the Currency ("OCC"), the Federal Deposit Insurance Corporation ("FDIC"), and the Federal Reserve System ("FRS").

[2] Fewer than 25 banks failed in each year from 1994 – 2008, and about 100 failed only during 1994 and 2008.

Risk.[3] However, Cole and Gunther (1998) provide evidence that the information content of the CAMELS ratings decays rapidly as the financial conditions of banks change over time, becoming obsolete in as little as six months; they also report that a static probit model using only publicly available bank-level accounting data almost always provides a more accurate prediction of bank failure than do the actual bank CAMEL ratings.

While the static probit methodology models the health of a bank as a function of its financial condition taken from a single point in time, the time-varying hazard methodology models the health of a bank as a function of its financial condition over a number of different time periods. Shumway (2001) cites three econometric reasons why a time-varying hazard model is preferred compared to static models,[4] he goes on to demonstrate that a simple hazard model provides more consistent in-sample estimations and more accurate out-of-sample predictions for U.S. corporate bankruptcies occurring during the 1962 – 1992 period than do several bankruptcy prediction models.

Distinguished from the existing studies, this paper examines the early warning models from a unique perspective, and our findings have very important implications for policy makers/regulators and individuals involved in bank risk management. Firstly, to our knowledge, this paper is the first empirical study to investigate the relative forecasting accuracy of time-varying hazard model, the static probit model, and the static logit model. We find that both of the

[3] The Uniform Financial Rating System, informally known as the CAMEL ratings system, was introduced by U.S. regulators in the November 1979 to assess the health of individual banks. Following an onsite bank examination, bank examiners assign a score on a scale of one (best) to five (worst) for each of the five CAMEL components; they also assign a single summary measure, known as the "composite" rating. In 1996, CAMEL evolved into CAMELS, with the addition of sixth component ("S") to summarize Sensitivity to market risk.

[4] See pages 102-103 in Shumway (2001).

simpler static models significantly outperform the hazard model, in sharp contrast to the results in Shumway (2001).

Existing empirical studies apply the time-varying hazard model exclusively from an academic perspective. For example, earlier research (e.g., Shumway 2001; Mannasoo and Mayes, 2009) compares the accuracy of the static models and the time-varying hazard model in predicting corporate bankruptcy, more recent empirical studies (DeYoung and Torna 2013; Hong and Wu 2013; Betz *et al.* 2014) apply the time-varying hazard models in bank failure prediction; these studies typically use one-year lagged independent variables for the time-varying hazard model—assuming the company/bank financial data are available from the beginning of each period. Consequently, a time-varying hazard model based upon such data can only predict bank failures one year ahead.

However, bank regulators are most concerned about identifying potential bank failures with sufficient lead time to take supervisory action that might forestall a bank's failure, or, in the event of failure, mitigate the cost of that failure and its impact on the functioning of the banking system, such as contagion. In other words, bank regulators are more concerned about identifying potential bank failures at least two years into the future in order to be able to take appropriate supervisory actions.

In practice, future bank financial data are not available at any given point of time, so regulators must rely upon what data actually are available, without "peaking" at future data, as academics have done. For example, at year-end 2009, bank financial data exist for a one-year-ahead forecast of failures that occur during 2010, but year-end 2010 bank financial data are not yet available for a one-year-ahead prediction of failures that occur during 2011.

We utilize a technique for matching bank financial data with bank failure data so that all models rely upon *the same most recently available bank financial data*; when we do so, we find that the static logit and probit models both *outperform* the time-varying hazard model. This finding is surprising—given the three econometric advantages of the time-varying hazard model over the logit and probit models cited by Shumway (2001). It also suggests that incorporating stale data into the time-varying model actually degrades, rather than improves, out-of-sample forecasting accuracy. Therefore, how to reconcile an econometrically superior model with practical applications opens up a new direction for future research.

In the second major part of our analysis, we find that the parsimonious specifications we document for the 1985-1993 period perform well in predicting bank failures during the recent global financial crisis, even though the two crises are separated by two decades during which the U.S. and international regulatory structures changed drastically, and consolidation eliminated almost half of all U.S. banks. When we apply the coefficients estimated from the 1985-1993 period to bank financial data reported up to the end of 2008 and use this information to forecast bank failures during 2009-2010, we find that both the probit and hazard models work surprisingly well, but the probit model significantly outperforms the hazard model by identifying an average of 66% (57% for hazard model) and 76% (67% for hazard model) failed banks among the top 5% and 10% worst predicted failure probabilities, respectively. This finding is most surprising and has very important policy implications. This suggests that the characteristics of "distressed banks" have experienced little change despite the substantial changes of the U.S. banking industry in terms of consolidation, technology, regulations, and business activities that have taken place over the past two decades. Our research provides direct empirical evidence in support of supervisory/regulatory policies focusing on banks' traditional risk ratios, particularly

those based upon the CAMELS methodology. This finding is also consistent with related recent empirical research. For example, Berger and Bouwman (2013) find capital plays a fundamental role in banks' performance, it helps small banks to increase the chance of survival and market share at all time, and enhances the performance of median to large banks during banking crises. Cole and White (2012) analyze why commercial banks failed during the recent financial crisis and find the traditional proxies for CAMELS components do an excellent job in explaining the U.S. bank failures in 2009. Fahlenbrach *et al*. (2012) find that banks performed poorly during 1998 when Russia defaulted some of its debt are prone to perform poorly during the recent financial crisis, and provide empirical evidence on the persistence of a bank's risk culture.

The remainder of our manuscript is organized as follows. Section 2 provides a review of the literature on failure prediction and distinguishes our paper from the existing literature. Section 3 describes our data and econometric models. Section 4 presents our empirical results, while Section 5 provides a summary and conclusion.

2. Literature review

The literature on forecasting bankruptcy and firm failure dates back to the 1960s. Altman (1993) provides a summary of this research through the early 1990s. The vast majority of these studies rely upon static methodologies, primarily discriminant analysis during the early years and probit/logit models during the later years.

Time-varying hazard analysis (or its variants), combined with the traditional default risk prediction models, have been employed to predict corporate bankruptcy in recent empirical studies. Using the corporate bankruptcy data over the period 1962 – 1992 in the U.S., Shumway (2001) demonstrates that the hazard model outperforms the traditional bankruptcy models

(Altman 1968; Zmijewski 1984), and that a new hazard model combining both accounting and market variables can substantially improve the accuracy in predicting corporate bankruptcy.

Beaver *et al.* (2005) extends Shumway's (2001) research by analyzing the corporate bankruptcy data over the period from 1962 – 2002 and find that the traditional accounting ratios remain robust in predicting corporate bankruptcy, but that a slight decline in the predictive ability of financial ratios can be compensated for by adding market-driven variables into the hazard model estimation. In contrast, Agarwal and Taffler (2008) compare the performance of market-based and accounting-based bankruptcy prediction models and find little difference between the market-based and accounting-based models in terms of predictive accuracy.

Shumway's hazard model has been widely applied to the prediction of corporate bankruptcy in recent empirical studies (Chava and Jarrow 2004; Bharath and Shumway 2008; Campbell *et al.* 2008; Nam *et al.* 2008; Bonfim 2009). Campbell *et al*. (2008) finds that the Shumway (2001) and Chava and Jarrow (2004) specifications appear to behave differently in financial-services industry. The financial-services industry, and especially commercial banking, plays a crucial role in a country's economic development, and is subject to heavy regulations relative to other industries. Moreover, standard financial ratios for commercial banks differ from those for other corporate sectors, so that traditional specifications of corporate bankruptcy prediction models cannot be applied to commercial banks.

Empirical studies applying hazard models to bank-failure prediction primarily rely on variants of Cox (1972) proportional hazard model. Early empirical research uses a static Cox (1972) model by using one set of explanatory variables at a point of time (Lane *et al*. 1986; Whalen 1991). A time-varying Cox model is utilized in the later empirical research by using one-

year lagged explanatory variables (Wheelock and Wilson, 2000; Molina, 2002; Brown and Dinc, 2005; Mannasoo and Mayes, 2009; Brown and Dinc, 2011; Liu and Ngo, 2014).

Our research differs from these previous studies on a number of dimensions. First, previous empirical studies mostly rely on the variants of the Cox proportional hazard model. The Cox hazard model is estimated from a sample of failed banks and non-failed banks, either a peer comparison group (Lane *et al.* 1986) or a selected group of non-failed banks (e.g., Whalen, 1991; Wheelock and Wilson, 2000; Molina, 2002; Brown and Dinc, 2005; Mannasoo and Mayes, 2009; Brown and Dinc 2011), by utilizing partial-likelihood estimation. In contrast, we use the simple discrete-time-varying hazard model proposed by Shumway (2001)[5], and analyze the entire population of U.S. banks over the sample period [6] by utilizing the full-information maximum likelihood hazard model proposed by Shumway (2001), which is much more tractable and more efficient than partial-likelihood Cox models (see Effron, 1977). As demonstrated in a proof by Shumway (2001), the discrete-time hazard model can be estimated using the logistic methodology and can produce more consistent and efficient estimators than alternative estimators. It does require adjustments to test statistics to account for the fact that the appropriate number of degrees of freedom is based upon the number of banks rather than the number of bank-year observations.

Secondly, our primary focus is on the out-of-sample forecasting accuracy, whereas previous empirical research has focused on the *posterior* probability of bank failure for the in-sample estimations. For example, Wheelock and Wilson (2000) use the Cox proportional hazard

[5] Cole and Wu (2009) was the first to apply Shumway(2001)'s model to bank failure analyses; later DeYoung and Torna (2013) use a multi-period logit model similar to Shumway(2001)'s approach in analyzing recent bank failure.
[6] Wheelock and Wilson (2000) use the data over the same sample period by analyzing a non-random selected sample of only 4,022 banks, out of which only 231 failed, while we analyze the entire population of U.S. banks, using data on more than 12,000 banks, out of which 1,277 failed.

model with time-varying covariates, estimated by partial likelihood, to identify specific factors that explain time to bank failures during 1984-93. Utilizing the proportional hazard model to study the large private banks in 21 major emerging markets in the 1990s, Brown and Dinc (2005) find that political concerns play a significant role in delaying government interventions to failing banks; they also find that the government is less likely to take over or close a failing bank if the country's banking system is weak (Brown and Dinc 2011). Similarly, Liu and Ngo (2014) find that bank failure is 45% less likely to occur in the year leading up to an election in the U.S. over the period from 1934 to 2012. Our primary concern is on the *anterior* probability of banking failure by comparing the out-of-sample accuracy of the time-varying hazard model and the static probit model in predicting bank failures, although we also conduct the in-sample prediction for both models.

Finally, and most importantly, we compare the predictive accuracy of hazard model and static logit and probit models from the regulator's perspectives while previous research exclusively examines this issue from the academic perspective. For example, Shumway (2001) combines the coefficients estimated from the in-sample period over 1962-1983 with the subsequent annual data to predict the out-of-sample corporate bankruptcy in the U.S. over the period of 1984-1992. Mannasoo and Mayes (2009) employ the discrete time hazard model estimated over the years 1997-2001 to predict bank distress over 2002-2004 for 19 Eastern European transition economies; similarly, Hong and Wu (2013) examine U.S. bank failures over the out-of-sample period from 2005-2011 by using an in-sample period of 1985 to 2004. These studies use one-year lagged independent variables for the hazard model and technically can only predict failure/bankruptcy one year ahead. In practice, bank regulators are more concerned about bank failures at least two or three years ahead so that they have sufficient time to take

supervisory action to avert the failure of bank. However, future bank financial data are not available at a point of time, e.g., at year-end 2009, bank financial data for year-end 2010 are not available yet for the prediction of failures during 2011. We use an innovative technique to compare the predictive accuracy of both models by restricting both models to the same information set.

When only currently available bank financial data are used to predict bank failures two or three years ahead, the static probit model outperforms the hazard model. We also find that the parsimonious specification fitting the data from 1985 to 1993 performs very well in predicting recent banks failures in the U.S. In particular, the simple probit model is capable of identifying between two-thirds and four-fifths of these more recent bank failures. As we have discussed in last section, the findings open up future research avenue to reconcile an econometrically superior model with practical applications and have very important policy implications for bank supervision.

3. Data and econometric models

3.1. Data

Bank financial data are taken from the year-end Reports of Condition and Income (the "Call Reports") filed by all FDIC-insured commercial banks with the U.S. Federal Financial Institutions Examination Council ("FFIEC"), which collects this information on behalf of the three primary U.S. banking regulators—the Federal Deposit Insurance Corporation ("FDIC"), the Federal Reserve System ("FRS") and the Office of the Comptroller of the Currency ("OCC").[7] Our data include basic balance-sheet and income-statement information of individual banks during the period of 1984-1993. Information on the identity and closure dates of individual failed banks over the period of 1985-1994 obtained from the FDIC website.[8] This sample period, during which more than one thousand banks failed, provides us the best data to compare these two econometric approaches to modeling bank failures. We also obtain the financial information of individual banks during the period of 2007-2010 and bank failure information during the period of 2008-2011. We use the estimated coefficients from 1984-1988 model fits to estimate failures in this later period. [9]

We construct financial variables that measure the capital adequacy, asset quality, profitability, and liquidity of banks. Numerous previous studies (e.g., Martin 1977; Gajewski 1989; Demirguc-Kunt 1989; Whalen 1991; Thomson 1992; Cole and Gunther 1995, 1998; Cole

[7] At the time when this manuscript was written, these datasets, along with supporting documentation for their use, were publicly available for download from the website of the Federal Reserve Bank of Chicago at: https://www.chicagofed.org/banking/financial-institution-reports/commercial-bank-data.
[8] At the time when this manuscript was written, these datasets, along with supporting documentation for their use, were publicly available for download from the website of the FDIC at: https://www.fdic.gov/bank/individual/failed/banklist.html. According to FDIC, bank failure is defined as when a bank is closed by its chartering authority and the FDIC is named as receiver.
[9] There were only 15 failures in 1994 and fewer than 12 failures over 1995-2007, too few for estimating a failure model.

and White 2012; Shaffer 2012; DeYoung and Torna 2013) have found these variables to be statistically significant in predicting bank failures.

Capital adequacy: We measure capital adequacy using the ratio of total equity capital to total assets. Bank capital can absorb unexpected losses and preserve confidence of banks. Thus, capital adequacy is expected to be negatively associated with the probability of bank failure.

Asset quality: We construct four variables that measure bank asset quality, which refers primarily to the payment status of the bank's loan portfolio. Banks are forced by regulators to write down the carrying value of non-current loans, which reduces the value of the bank's capital; Hence, asset quality variables are expected to explain the likelihood of bank failures. The variables on bank asset quality include: *Past due loans* measured by loans 90 days or more past due divided by total assets; *Nonaccrual loans* measured by nonaccrual loans divided by total assets; *Other real estate owned* measured by foreclosed real estate divided by total assets; *Nonperforming loans* measured by the sum of *past due loans, nonaccrual loans and other real estate owned*. Each asset quality variable is expected to have a positive coefficient.

We measure bank profitability using return on assets (*ROA*), which is defined as net income divided by total assets. The more profitable is a bank, the less likely is the failure of the bank. *ROA* is expected to have a negative coefficient.

We measure bank liquidity using two different variables—one on the asset side and one on the liability side of the balance sheet. The asset-side liquidity variable is: *Investment securities*, which is defined as investment securities divided by total liabilities. Banks mostly hold investment securities in the form of U.S. government bonds. Consequently, a bank's portfolio of investment securities is highly liquid, enabling the bank to quickly convert these

assets into cash to meet unexpected withdrawals of deposits, thereby minimizing fire-sale losses. Therefore, *Investment securities* is expected to have a negative coefficient.

The liability-side liquidity variable is *Large CDs* (*certificates of deposit*), which is defined as large certificates of deposit ($100,000 or more) divided by total liabilities. These types of deposits are highly volatile, in that they are the first to be withdrawn when a bank is rumored to have financial difficulties. In addition, banks that rely heavily upon purchased funds, such as large certificates of deposit, rather than core deposits, often have more aggressive investment strategies and face higher funding costs (Cole and Gunther 1995, 1998). Hence, we expect that *Large CDs* will have a positive coefficient.

For robustness, we also construct an alternative pair of bank liquidity variables that were used by Cole and Gunther (1995): *Securities to assets*, which is defined as investment securities divided by total assets, and *Large CDs to assets,* which is defined as large CDs divided by assets.

In addition to the above variables, we construct a proxy variable for *Bank Size,* which we define as the natural logarithm of total assets. We expect that small banks are more vulnerable to failure because, in general, their asset portfolios are less diversified; thus, the probability of failure will be negatively associated with bank size.

We check each of our variables for outliers. To reduce the impact of extreme values on the model fitting process, we winsorize each variable at the 0.01/0.99 level. In our sample sets, the number of failed banks is 1,229 over the period of 1985-1993, and 353 over 2008-2011, with more than one percent of banks failed each year during the periods of 1987 to 1990 and 2009 to 2010.[10]

[10] In a study of large international banks in emerging-market countries over 1993-2000 period, Brown and Dinc (2005) report that about 25 percent of their sample banks failed.

Table 2 reports the means and standard errors of the bank financial ratios of all banks, as well as the difference in means for surviving banks and failed banks, over the years of 1984-1988, 1989-1993 and 2007-2010. Compared with non-failed banks, failed banks are generally smaller, have lower capital ratio, earnings ratio and investment securities (and securities to assets) ratio; and higher non-performing loans and large CDs (and Larger CDs to assets) ratios. As expected, failed banks generally have much lower capital adequacy and profitability, worse asset quality and less liquidity; and this result holds when we examine the data for all of the time periods of our analysis separately.

A number of bank regulatory reforms were introduced subsequent to the first wave of bank failures, focusing on bank capital adequacy.[11] We find that the bank capital ratios for the period of 2007-2010 are significantly higher than the ones over the period of 1984-1988 and 1989-1993, with the mean value of 11.37% for all banks (4.08% for failed banks and 11.47% for non-failed banks, respectively) versus 8.89% (2.20% for failed banks and 8.97% for non-failed banks, respectively), and 9.27% for all banks (1.05% for failed banks and 9.33% for non-failed banks, respectively). The banking industry has experienced substantial changes since 1990s. Whether the coefficients fitting the data from the first wave of bank failures in the 1980s can be used to predict the second wave of bank failures over 2007-2010 is an interesting issue we are going to examine in Section 4.2.

3.2. Econometric models

We use a static probit model similar to the one used by Cole and Gunther (1998), a static logit model similar to the one used by Cole and White (2012), and a simple discrete-time hazard

[11] The Financial Institutions Recovery, Reform and Enforcement Act (FIRREA) was passed in 1989 and the FDIC Improvement Act was passed in 1991.

model as proposed by Shumway (2001). Unlike the static models, which assume bank failure is a function of bank financial variables at a single point of time, the time-varying hazard model assumes bank failure is a function of a time-series of bank financial variables over multiple points in time. These models are briefly explained in the following sections.

3.2.1. Static logit/probit model

We assume that $Failure^*_{i,t}$ is an unobservable index of the probability that bank *i* fails during year *t* and is a function of bank-specific characteristics x_t, so that:

$$Failure^*_{i,t} = \beta_t' \boldsymbol{X}_{i,t-k} + \mu_{i,t} \quad (1)$$

where $\boldsymbol{X}_{i,t-k}$ are a set of financial characteristics of bank *i* as of December 31st in the calendar year that was *k* years before *t*; β_t is a vector of parameter estimates for the explanatory variables measures as of year $t - k$, $\mu_{i,t}$ is a random disturbance term, $i = 1, 2, \ldots, N$, where *N* is the number of banks. Let $FAIL_{i,t}$ be an observable variable that is equal to one if $Failure^*_{i,t} > 0$ and zero if $Failure^*_{i,t} \leq 0$. In this particular application, $FAIL_{,i,t}$ is equal to one if a bank fails during year t and zero otherwise. Since $Failure^*_{i,t}$ is equal to $\beta_t' \boldsymbol{X}_{i,t-k} + \mu_{i,t}$, the probability that $FAIL_{i,t} > 0$ is equal to the probability that $\beta_t' \boldsymbol{X}_{i,t-k} > 0$, or, equivalently, the probability that ($\mu_{i,t} > -\beta_t' \boldsymbol{X}_{i,t-k}$). Therefore, one can write the probability that $FAIL_{i,t}$ is equal to one as the probability that ($\mu_{i,t} > -\beta_t' \boldsymbol{X}_{i,t-k}$), or, equivalently, that Prob($FAIL_{i,t} = 1$) = 1 - Φ ($-\beta_t' \boldsymbol{X}_{i,t-k}$), where Φ is the cumulative distribution function (CDF) of ε. For the logit model, the CDF is assumed to be logistic while for the probit model, the CDF is assumed to be normal. The probability that $FAIL_{i,t}$ is equal to zero is then simply Φ ($-\beta_t' \boldsymbol{X}_{i,t-kt}$). The likelihood function *L* for this model is:

$$L = \prod_{FAIL_i = 0} [\Phi(-\beta_t' \boldsymbol{X}_{i,t-kt})] \prod_{FAIL_i = 1} [1 - \Phi(-\beta_t' \boldsymbol{X}_{i,t-k})], \quad (2)$$

where:

$$\Phi(-\beta_t' \boldsymbol{X}_{i,t-k}) = \exp(-\beta_t' \boldsymbol{X}_{i,t-k}) / [1 - \exp(-\beta_t' \boldsymbol{X}_{i,t-k})]$$

$$= 1 \;/\; [1 + \exp(-\beta_t' \boldsymbol{X}_{i,t-k})] \qquad (3)$$

and

$$1 - \Phi(-\beta_t' \boldsymbol{X}_{i,t-k}) = \exp(-\beta_t' \boldsymbol{X}_{i,t-kt}) \;/\; [1 + \exp(-\beta_t' \boldsymbol{X}_{i,t-k})] \,. \qquad (4)$$

3.2.2. Time-varying hazard model

We assume that a bank can fail only at a discrete point of time, $T_i = 1, 2, 3, \ldots$. We define a dummy variable Y_i that is equal to one if a bank failed at time T_i, and otherwise is equal to zero. Let $F(t_i, X; \gamma)$ be the probability mass function of failure, where γ represents a vector of parameters and X represents a vector of explanatory variables. Following the hazard model conventions, the survivor function that gives the probability of surviving up to time T can be defined as:

$$S(T, X; \gamma) = 1 - \sum_{J<T} F(J, X; \gamma) \qquad (5)$$

The hazard function that gives the probability of failure at T conditional on surviving to T can be expressed as:

$$H(T, X; \gamma) = \frac{F(T, X; \gamma)}{S(T, X; \gamma)} \qquad (6)$$

The likelihood function of the discrete-time hazard model is given by:

$$\boldsymbol{L} = \prod_{i=1}^{n} \boldsymbol{H}(\boldsymbol{T}_i, \boldsymbol{X}_i; \gamma)^{Y_i} \boldsymbol{S}(\boldsymbol{T}_i, \boldsymbol{X}_i; \gamma) \qquad (7)$$

The discrete-time hazard model is equivalent to a multiple-period logit model with the following likelihood function (Cox and Oakes 1984; Shumway 2001):

$$L = \prod_{i=1}^{n} \left\{ F(T_i, X_i; \gamma)^{Y_i} \prod_{J<T_i} [1 - F(J, X_i; \gamma)] \right\} \quad (8)$$

Where *F* is the cumulative density function (CDF) of failure that depends on *T*, which can be interpreted as a hazard function.

Therefore, a discrete-time hazard model can be estimated using a logit model with appropriate adjustment to the test statistics. The test statistics estimated from a logit model assume that the bank-years are independent observations. However, for a discrete-time hazard model, the bank-year observations of a particular bank cannot be independent because a bank cannot fail in period *T* if it failed in period *T-1*; similarly, a bank surviving to period *T* cannot have failed in period *T-1*. Thus, each bank's life span only makes one observation for the hazard model; the correct number of independent observations is the number of banks in the data instead of the number of bank years. As a result, the correct test statistics of the hazard model can be derived by dividing the tests statistics of a logit model by the average number of firm years per bank.

3.3. Measures of Goodness of Fit

We measure how well our models fit the data using a number of different statistics. We assess both in-sample fit and out-of-sample fit.

3.31. Akaike Information Criterion (AIC)

The Akaike Information Criterion ("AIC") (Akaike, 1974) is an estimator of the relative quality of statistical models for a given set of data. As such, AIC can be used to compare models, but AIC does not measure out of sample performance. AIC Formally, AIC is given by:

$$AIC = 2k - 2\ln(\hat{L})$$

where $\hat{L}$ is the maximum value of the likelihood function for the model and k is the number of estimated parameters in the model. When comparing models, the model with the lowest AIC is generally preferable.

3.32. Bayesian Information Criterion (BIC)

BIC is defined as ln(n) x K – 2 x Log-Likelihood. N is the number of observations. Hence, BIC is closely related to AIC but penalizes a model more heavily when it uses more parameters. Smaller values indicate better in-sample fit.

3.33. Precision Score and Average Precision Score,

Precision is the ratio $TP/(TP + FP)$ where TP is the number of true positives and FP is the number of false positives. Precision is the ability of a classifier to label as negative a sample that is negative.

Related to precision is Average Precision (AP), which can be graphed. AP is used to provide a single metric of measure representing a precision-recall curve, similar to the AUC measure in an ROC graph. AP is the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n$$

Where P_n and R_n are the precision and recall at the nth threshold. Figure 4 provides an example of a Precision-Recall Curve graph

3.34. F-beta Score

Because bank failures are rare events, we include a measure that includes class imbalance as a weight in the calculation of the metric. We use the F-beta score, which is the weighted harmonic mean of precision and recall. Precision is calculated as the ratio of correctly predicted positives divided by the total number of predicted positives. Recall is calculated as the number of

true positives divided by the sum of the number of true positives and the number of false negatives.

An F-beta score of 1.0 is a perfect score, and a score of zero represents a perfectly bad model. The beta parameter determines the weight of precision in the combined score, which is calculated by determining precision and recall for surviving banks and failed banks separately. Then the average is found, weighted by the number of true instances for surviving banks and failed banks. In this way, the score is adjusted to account for the imbalance in the data.

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall}$$

3.35. Logarithm-Loss Score

The Logarithmic-Loss Score, simplified to Log Loss, is a classification loss function that quantifies the accuracy of a classifier by penalizing false classifications. Minimizing log loss essentially maximizes the accuracy of the classifier. Log loss relies on the probabilistic output of classifiers rather than the binary class prediction. Log loss is given by:

$$Log\ Loss = -\frac{1}{N}\sum_{i=1}^{N} * \sum_{j=1}^{M} y_{ij} \log p_{ij}$$

where N is the number of observations, M is the number of possible labels, y_{ij} is a binary indicator of whether or not label j is the correct classification for observation i, and p_{ij} is the model probability of assigning label j to observation i.

A perfect classifier would have a log loss of precisely zero. A model no better than random guessing would have a log loss score of 0.693147, equivalent to $-\ln(0.5)$. Anything higher than 0.693147 may not be a reliable predictor.

For this study of bank failures, the equation for log loss above simplifies to:

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1-y_i)\log(1-p_i)]$$

When a prediction algorithm is confident and wrong, the log loss score penalizes the model heavily. A prediction with probability 0% that is wrong has an infinite log loss score. A prediction with probability 100% earns a zero score. Rarely will a model assign full confidence to a classification, so an infinite score would likely come from a model which fails to fit in the first place, which is impractical. Log loss allows for a direct comparison on out of sample prediction accuracy and the quality of predictions given by the different models.

3.36. Matthews Correlation Coefficient

The Matthews Correlation Coefficient (MCC) accounts for true and false positives, and can be used even when class sizes are very different. It can be thought of as the correlation between the actual and predicted binary classifications, and, like Pearson correlations, falls between -1.0 and +1.0.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives.

3.37. ROC-AUC Score

The ROC curve plots the response of TPR(T) against FPR(T) as T varies (Fawcett, 2006). AUC is the "area under the curve." A perfect classifier would receive an AUC of 1.00, a series of random guesses would score 0.50, and a perfectly bad classifier would score 0.00. Figure 3 provides an example of an ROC Curve.

4. Empirical results

4.1. Econometric performance of hazard, probit and logit models

In this section, we assess accuracy of the time-varying hazard model, the traditional static probit model and the traditional static logit model assuming that one-year-lagged bank financial data are available for estimating the time-varying hazard model (e.g., similar to previous empirical research by Shumway 2001, Mannasoo and Mayes 2009, and Hong and Wu 2013). We first perform the in-sample estimations, and then calculate and compare the out-of-sample forecasting accuracy of all models. For example, we use in-sample data for 1984-1988 and failures taking place during 1985-1989 to estimate the hazard model and apply these coefficients to obtain forecasts for 1990-1994 failures. For the probit and logit model, we use three different specifications: 1) in-sample bank financial data for 1984 and failures taking place during 1985, 2) in-sample bank financial data for 1988 and failures taking place during 1989, and 3) in-sample bank financial data for 1988 and bank financial data for banks which failed in 1985, 1986, 1987, and 1988 in addition to the financial data for banks which fail in 1989. The latter model specification allows the logit and probit models to estimate coefficients with the same number of failures as the Hazard model but retains the ability to estimate one year ahead. We use these coefficients on out-of-sample financial data from 1989-1993 and from 2007-2010 to create one-year-ahead forecasts for 1990-1994 and 2008-2011, respectively.

4.1.1. In-sample estimation

Table 3 presents the in-sample model fit statistics for all models. The hazard model presents the highest Akaike Information Criterion (AIC) score of all the models, which is a direct consequence of the model being fit to a greater number of healthy (surviving) banks compared to the other models. Interestingly, the probit model fit to year-end 1988 financial data has the

lowest AIC score, which indicates the best fit. The next best AIC score is the logit model fit on year-end 1988 financial data. Both models indicate a better fit than the hazard model. This is supported by the Bayesian Information Criterion (BIC) score and the pseudo-R squared, where both the probit and the logit models outperform the hazard model. These statistics for the hazard model are not adjusted for the repeated observations of banks in the data, because the model fit statistics are not the relevant metrics for this study; they are presented to represent the performance of the models on the data to which they are trained. In a later section, we will examine the models out-of-sample performance to determine which model performs the best.

Table 4 presents the Average Marginal Effects (AME) for the three models fit to bank financial data as described in section 4.1 above. Panel B provides the estimation results for the probit model for the three in-sample periods identified in section 4.1 above, and Panel C represents the estimation results for the logit model in a similar fashion.

All models produce the expected signs of the coefficients for the financial variables: smaller banks with higher level of non-performing loans and relying more on large CDs for funding are more likely to fail, while banks with higher capital adequacy, greater profitability, and more liquidity are less likely to fail. As with previous research, these results support the robustness of the CAMELS regulatory framework.

Panel A in Table 4 presents the hazard model AMEs. As suggested by Shumway (2001), the test statistics for the hazard model have been adjusted for the average number of firm years per bank. The hazard model identifies three variables that are statistically significant at the 1% level: Capital Adequacy, Non-Performing Loans, and Size. The same is true in regressions

including the alternate specifications of Investment Securities and Large CDs but are unreported[12] to save space.

Panel B in Table 4 presents AMEs for the three single-period probit model. The coefficient for *Earnings Ratio* and *Investment Securities* is statistically significant at the 1% level when fit on year-end 1984 data, but *Earnings Ratio* lacks significance when fit on year-end 1988 data. When we include additional failure observations from the full sample period and fit on year-end 1988 financial data, the coefficient regains statistical significance and achieves significance at the 1% level, and all the other coefficients in the model are significant at the 1% level as well. In each model specification, every bank financial variable except for *Earnings Ratio* estimated using the probit model is statistically significant at better than the 5% level, with the *p*-values close to zero. In Panel A, however, the coefficient of the *Earnings ratio* estimated using the hazard model lacks statistical significance; this indicates that the marginal impact of a temporary higher ratio of earnings to assets is negligible when the failure probability of a bank is higher, and a temporary loss will not lead to bank failure when the failure probability of a bank is relatively low. This result is consistent with Wheelock and Wilson (2000), who find that the earnings ratio is not significant in predicting 1985-1993 in-sample U.S. bank failures.

Panel C in Table 4 presents AMEs for the three single-period logit models. The results are nearly identical to the probit model. Depending on which sample period is used, *Earnings Ratio* is either lacking significance, or the p-values approach zero.

[12] Regression tables including alternate specifications of Investment Securities and Large CDs are available from the authors upon request.

What is evident from Table 4 is that the selection of the representative financial data for bank health is very important for determining the best fit of a model, and the higher the failure rate, the better the model will fit to the data.

4.1.2. Out-of-sample prediction

Table 5 presents the out-of-sample Type I and Type II error rates for each model and each out of sample year and each sample period following 4.1 as well as the full out-of-sample period combined. Table 5 is an important investigation of the bias present in each model and how each model reacts to the different conditions in the two out of sample periods. Considering that the hazard model benefits from observing far more surviving bank observations than any other model, we should expect the hazard model to make the fewest Type II errors. In our context, a Type II error is made when a model predicts that a bank will survive in the next year, but the bank actually fails. To a regulatory agency, this is a bad result; the regulator may have avoided investigating the bank and missed an opportunity to enact corrective actions to save the bank.

Table 5 demonstrates that the hazard model does not have the lowest Type II error rate on the full out of sample period, nor is it the best in the 1989-1993 period. If we select the winning model based on Type II errors alone, then we ignore the cost of forecasting failure too frequently and sending bank regulators on an expansive search that may be fruitless.

An important consideration is the bias-variance tradeoff. The hazard model is fit on more survival observations than any other model, and the probit and logit models, which benefit from additional failure observations proportional to survivors, have a higher Type II error rate when

exposed to more failure observations, but have the lowest Type I error rates. The conclusion from this table is that more investigation is necessary.

Figure 1 represent ROC-AUC Curves for the hazard, probit and logit models fit on the indicated sample periods. An interesting feature of bank failure data is the low incidence of failure relative to the number of observations. Because of the extreme imbalance of the data, common graphing software is unable to accommodate the imbalance and visually capture the effect this has on AUC. As a result, the graphs are inconclusive; the AUC scores are very similar, and no clear winner can be determined.

Rather than continue to confuse the reader, we instead move to a better visual aid in Average Precision-Recall curves. Figure 2 displays the Precision-Recall curves for the hazard, probit and logit models for the sample periods of interest. Precision-Recall better captures the out-of-sample performance of the models since the graph uses the distribution of the data to determine the cut-points used to display the performance of each model. Now, the performance of the static probit and logit models become apparent; both score higher than the hazard model when fit on year-end 1988 financials. Their performance appears to be the same with the additional failure observations in the last row, but still better than the hazard model.

4.1.3 – Additional Scoring Metrics

Up to this point, we have only focused on Type I and Type II error rates, AUC scores and Precision-Recall scores. We have not evaluated enough measures to determine that static models outperform hazard models in our empirical setting. In order to determine the winner, we score

each model on its out-of-sample performance in the two periods of interest as well as on the total out-of-sample period. This follows the methodology of Cole and Taylor (2019). Table 6 presents the Average Precision Score, F1 Score, Log Loss Score, Matthews Correlation Coefficient, Precision Score, ROC-AUC Score and Recall Score. All of these scores have been described in Section 3 above.

Table 6 demonstrates that the logit model fit with 1988 year-end financial data performs the best during the 1989-1993 evaluation period. The hazard model leads in zero categories. The probit and logit models fit with the more recent failure observations perform better than the models fit to year-end 1984 financial data, but the logit model fit to 1988 year-end financial data is the clear winner.

Table 6 also demonstrates that the logit and probit model fit to 1988 year-end financial data generalizes to the 2007-2011 wave of failures, but from the table, a clear winner is not evident. The overall winner from the total out-of-sample period is the static logit model fit to year-end 1988 financial data. This directly contradicts the findings of Shumway (2001) who claims that a hazard model is better than a static model.

4.1.4 – Robustness Test

Claiming that the static model outperforms the time-varying hazard model with qualitative factors may not be satisfactory in an empirical setting. In order to assuage any doubts, we also conduct a McNemar's Test to determine that the models in question are actually different from one another in a statistically valid way. McNemar's Test was described in Section 3 above, and the results are presented in Table 6 and 7.

Panels A, B and C in Table 6 display the McNemar Test contingency tables for the hazard, probit and logit models for the 1989-1993 out-of-sample period. Panel A shows that the hazard model is statistically different from and more accurate than the probit model fit to year-end 1984 financial data. This panel validates the scores presented in 4.1.3 above; the hazard model is more accurate and the chi-square test statistic is statistically significant above the 0.01 level. Panel B compares the probit and logit models fit to 1984 year-end financial data. This table helps to differentiate the similar scores for these models and shows that the logit model is more accurate than and statistically different from the probit model fit to the same data. We can say that the logit model is better than the probit model for this wave of bank failures. Panel C compares the hazard model to the logit model fit on 1984 year-end financial data. This panel demonstrates that the hazard model, which benefits from observing many more failures and financial conditions much later in the failure wave, is more accurate than and statistically different from the logit model. We can say that the hazard model is better than the static logit model fit to stale data for this wave of bank failures.

What happens when we fit the probit and logit models to the later portion of the in-sample observations? Panel D, E and F in Table 6 repeats the analysis above, but compares the hazard model to the probit and logit model which are fit to year-end 1988 financial data. Here, the hazard model is not the better model. In fact, the logit model is more accurate than and statistically different from both the hazard model and the probit model. The probit model is better than the hazard model, and the logit model is better than the probit model. The test statistics indicate that the models are statistically different above the 0.01 level in each panel. We can say that the logit model is better than the hazard model. Our out-of-sample performance

scores indicate that this is the best model overall, and the McNemar's Test demonstrates that the models are different for this wave of bank failures.

Panels G, H and I in Table 6 repeat the analysis above and show that the additional bank failures from 1984-1987 greatly improve the accuracy of the logit and probit models, but not enough to outperform the hazard model. The hazard model is only statistically different from the logit model at the 0.05 level of significance.

Panels A, B and C in Table 7 test each model on the 2007-2010 wave of bank failures. The tables are structured in a similar fashion to Table 6 above. The logit model always beats the probit model, and the logit model fit to 1988 year-end financial data is the better model. The static logit model is better than and statistically different from the hazard model at the 0.01 level of significance.

5. Summary and conclusions

In this paper, we examine two bank early warning models developed from two widely used methodologies -- a simple time-varying hazard model as proposed by Shumway (2001) and a simple static probit model similar to the one used by Cole and Gunther (1998). We contribute to the literature by comparing the predictive accuracy of these two models from both a qualitative and a quantitative perspective while the existing empirical studies examine the performance of models exclusively from a qualitative perspective. We find that, from a quantitative perspective, the hazard model does not outperform the static probit or logit models, in direct contrast to Shumway (2001). Similar with existing empirical studies (e.g., Shumway 2001; Mannasoo and Mayes, 2009; DeYoung and Torna 2013, Hong and Wu 2013, Betz, et al 2014), we use one-year lagged independent variables for the hazard model, therefore technically

the hazard model can only predict bank failures one year ahead as future bank financial data are not available yet at the time of prediction. However in practice, bank regulators are more concerned about identifying potential bank failures two or three years ahead so that they have sufficient lead time to take supervisory action. When we limit the models to the bank financial data available from the failure wave in the 1980's, we find that the hazard model underperforms the logit and probit model, and that the predictive accuracy of all models declines substantially when predicting bank failures so far into the future. Therefore the substantial improvement in predictive accuracy of hazard model compared to the static models fit on stale observations mainly comes from the use of latest bank financial data rather from the time-varying covariates. Our findings support the perception that the health of a bank is the function of its latest financial conditions, and opens up a new direction for future research to bridge an econometrically superior model with practical applications. We used the evaluation methodology proposed in Cole and Taylor (2019) and demonstrated its usefulness, which should be used in further studies of bank failure.

We also find that the simple parsimonious specification fitting the data over 1988 year-end financial data performs very well in predicting bank failures over 2007-2010. This finding is most surprising but has very important policy implications. It suggests that the characteristics of "distressed banks" have experienced little change even though the U.S. banking industry has gone through substantial changes in technology, regulations, and business activities over the past two decades, and provides direct empirical evidence in support of the supervisory/regulatory policy targeting banks' traditional CAMELS risk ratios.

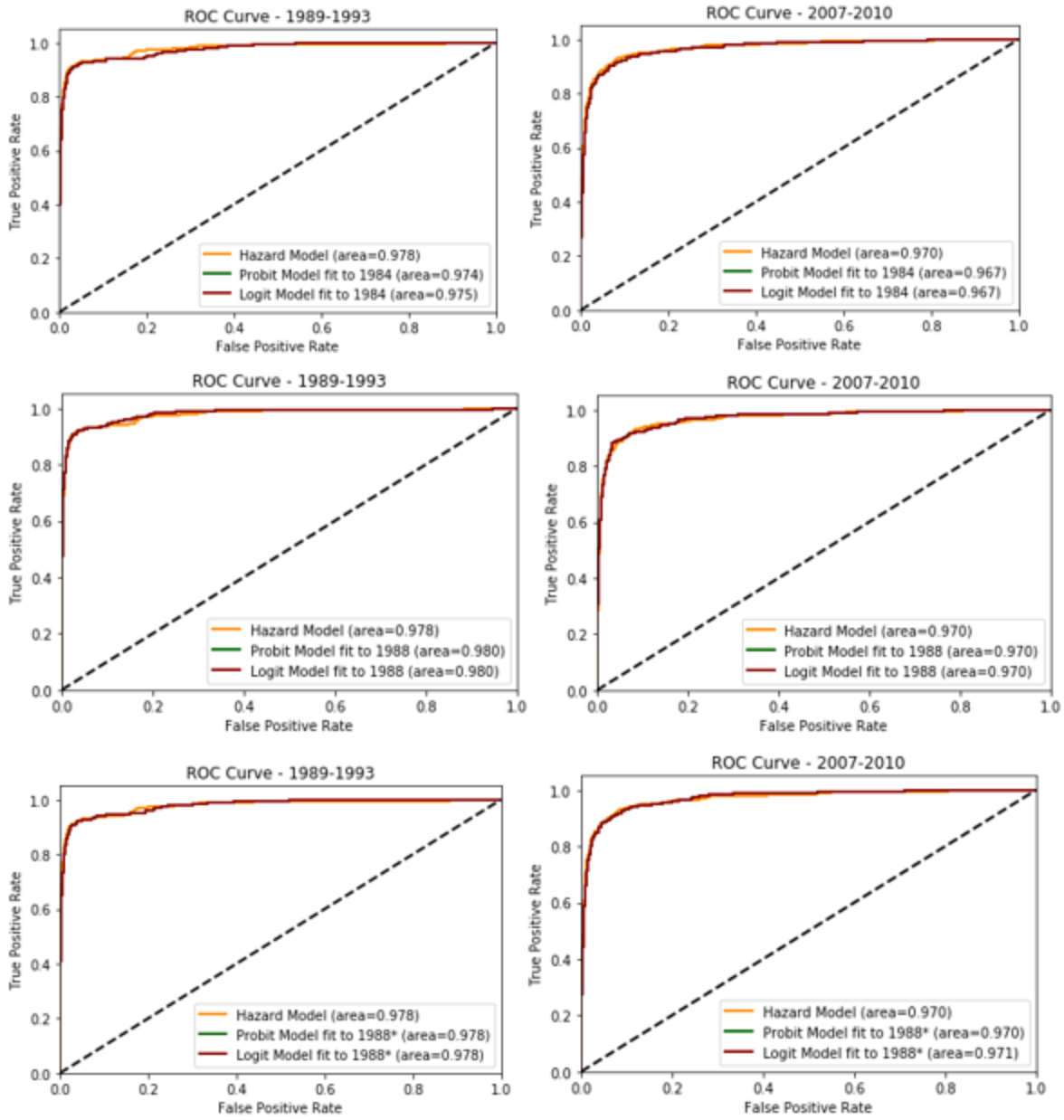

Figure 1: ROC Curves

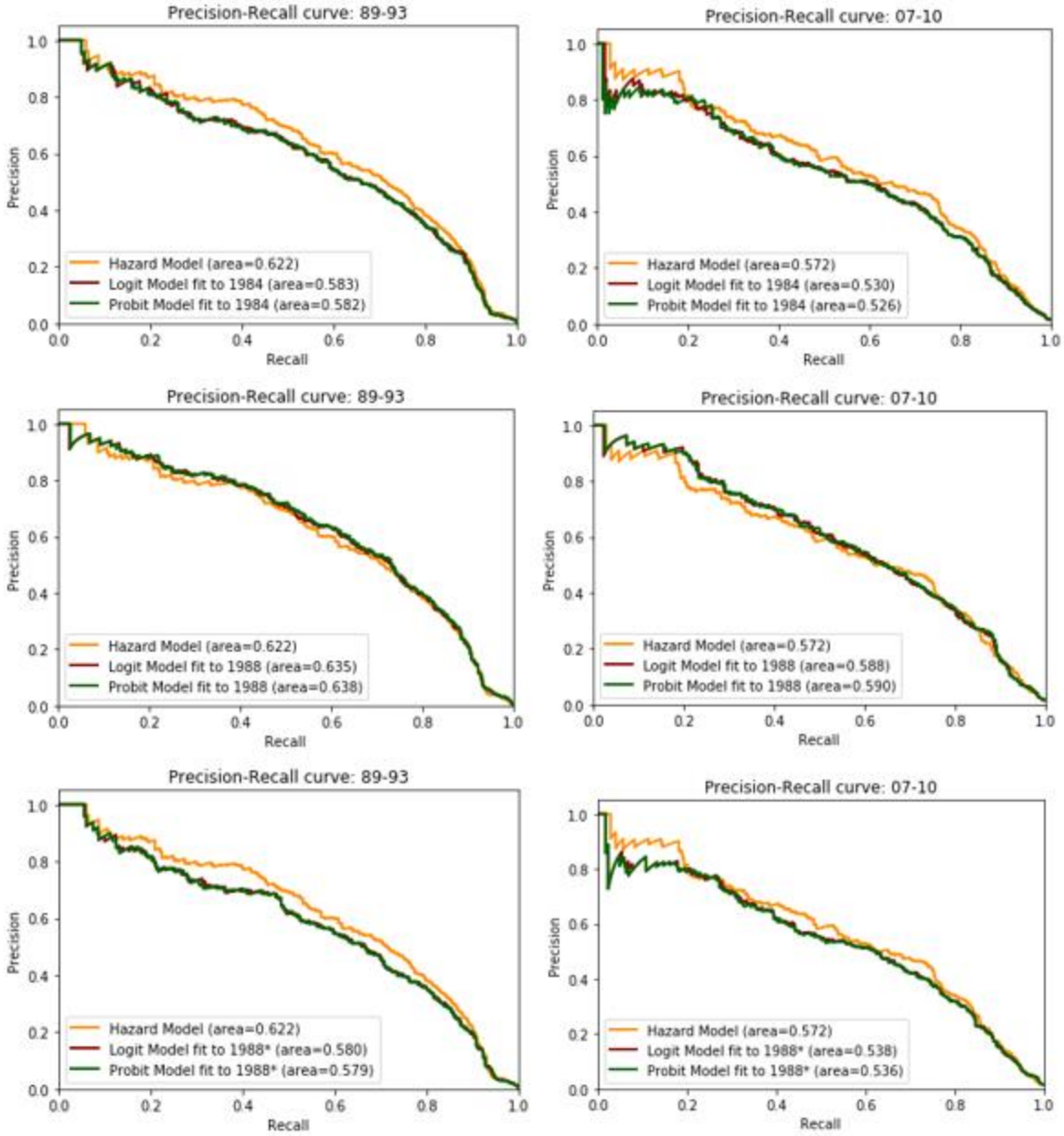

Figure 2: Precision-Recall Curves

Table 1

Variable construction and description. All variables have been winsorized at the 0.001 and 0.999 levels.

Variable	Description
Capital Adequacy	The ratio of a bank's total equity capital to total assets. (Total Equity = RCFD2170, Total Assets = RCFD3210)
Past Due Loans	Loans 90 days or more past due divided by total assets. (Past Due Loans 90+ = RCFD1407)
Nonaccrual Loans	Nonaccrual loans divided by total assets. (Nonaccrual loans = RCFD1403)
Other Real Estate Owned (OREO)	Other Real Estate Owned divided by total assets. (OREO = RCFD2150)
Non-performing Loans	Non-performing loans is the sum of Past Due Loans, Nonaccrual Loans, and OREO.
Earnings Ratio	Net Income divided by total assets. (Net Income = RIAD4340)
Investment Securities	Investment Securities divided by Total Liabilities (Prior to 1994: Investment Securities = RCFD0390; otherwise = RCFD1754 + RCFD1773; Total Liabilities = total assets - total equity capital)
Large CDs	Large Certificates of Deposit ($100,000 or more) divided by total liabilities. (Large CDs = RCFD2604)
Investment Securities*	Investment Securities divided by total assets.
Large CDs*	Large Certificates of Deposit ($100,000 or more) divided by total assets.
Bank Size	natural logarithm of total assets in thousands of same year dollars

Table 2

Univariate comparisons for each variable covering 1984-1988, 1989-1993, and 2007-2010.

	Training Sample (1984-1988)		Difference in Means (survivors and failures)		Out of Sample (1989-1993)		Difference in Means (survivors and failures)		Out of Sample (2007-2010)		Difference in Means (survivors and failures)	
Variable	**Mean**	**S.E.**	**Diff**	**t-statistic**	**Mean**	**S.E.**	**Diff**	**t-statistic**	**Mean**	**S.E.**	**Diff**	**t-statistic**
Capital Adequacy	0.089	0.0002	0.068	50.451 ***	0.093	0.0002	0.083	30.281 ***	0.114	0.0004	0.074	44.930 ***
Earnings Ratio	0.005	0.0001	0.051	36.166 ***	0.008	0.0001	0.055	28.726 ***	0.003	0.0001	0.054	30.985 ***
Past Due Loans	0.006	0.0000	-0.016	-19.961 ***	0.003	0.0000	-0.010	-11.742 ***	0.002	0.0000	-0.002	-4.696 ***
Non Accrual Loans	0.009	0.0001	-0.043	-33.061 ***	0.007	0.0000	-0.041	-23.877 ***	0.013	0.0001	-0.074	-32.995 ***
OREO	0.006	0.0000	-0.028	-25.563 ***	0.006	0.0001	-0.041	-21.810 ***	0.006	0.0001	-0.031	-16.651 ***
NPL	0.021	0.0001	-0.089	-40.513 ***	0.016	0.0001	-0.093	-30.499 ***	0.021	0.0002	-0.114	-30.887 ***
Investment Securities/TL	0.307	0.0007	0.171	47.092 ***	0.340	0.0008	0.199	29.043 ***	0.233	0.0011	0.127	26.975 ***
Large CDs/TL	0.119	0.0004	-0.087	-18.242 ***	0.098	0.0003	-0.031	-8.001 ***	0.186	0.0006	-0.065	-9.602 ***
Investment Securities/TA	0.277	0.0006	0.145	41.076 ***	0.305	0.0006	0.168	28.730 ***	0.204	0.0009	0.102	22.719 ***
Large Cds/TA	0.108	0.0004	-0.094	-20.126 ***	0.089	0.0003	-0.040	-10.255 ***	0.165	0.0005	-0.076	-11.742 ***
LNSIZE	10.668	0.0047	0.374	8.986 ***	10.958	0.0052	0.153	2.402 ***	11.944	0.0077	-0.604	-9.393 ***

Table X - Fit Statistics

Fit statistics from estimating a hazard model, probit models and logistic regression models to explain bank failures, where the dependent variable FAIL takes on a value of one if a bank failed in the next year, and a value of zero otherwise. Explanatory variables are defined in Table 1.

	Hazard	Probit84	Probit88	Probit8488	Logit84	Logit88	Logit8488
Akaike Information Criterion	4594.299	830.909	707.616	2458.044	868.779	726.926	2495.566
Bayesian Information Criterion	4649.175	876.355	752.448	2503.158	914.225	771.758	2540.680
Df Model	5	5	5	5	5	5	5
Df Residuals	69273	14385	12985	13610	14385	12985	13610
Log Likelihood	-2291.150	-409.454	-347.808	-1223.022	-428.390	-357.463	-1241.783
No. Obs.	69279	14391	12991	13616	14391	12991	13616
No. Failures	828	116	203	828	116	203	828
pseudo-R squared	0.490	0.393	0.667	0.608	0.365	0.658	0.602

Table X **Average Marginal Effects**

Average marginal effects (AME) from estimating models to explain bank failures, where the dependent variable FAIL takes on a value of one if a bank failed in the next year, and a value of zero otherwise. Explanatory variables are defined in Table 1. The number of failures and survivors per sample period used for fitting the models is described in Table 2.

Hazard Model Fit on 1984-1988 Data to Predict 1989 Failures

Variables	dy/dx	Std. Err.	t		Pr(>|t|)	Conf. Int. Low	Cont. Int. Hi.
Capital Adequacy	-0.283	0.014	-4.417	***	0.000	-0.306	-0.259
Earnings Ratio	-0.035	0.012	-0.633		0.004	-0.055	-0.015
NPL	0.139	0.007	4.335	***	0.000	0.127	0.151
Investment Securities	-0.021	0.003	-1.548		0.000	-0.026	-0.016
Large CDs	0.018	0.003	1.505		0.000	0.013	0.022
lnSize	-0.002	0.000	-3.705	***	0.000	-0.002	-0.002

Panel B

Probit Model Fit on 1984 Data to Predict 1989 Failures

Variables	dy/dx	Std. Err.	t		Pr(>|t|)	Conf. Int. Low	Cont. Int. Hi.
Capital Adequacy	-0.166	0.028	-5.887	***	0.000	-0.213	-0.120
Earnings Ratio	-0.109	0.029	-3.777	***	0.000	-0.156	-0.061
NPL	0.144	0.019	7.615	***	0.000	0.113	0.175
Investment Securities	-0.016	0.006	-2.608	***	0.009	-0.026	-0.006
Large CDs	0.016	0.005	3.030	***	0.002	0.007	0.024
lnSize	-0.002	0.000	-8.841	***	0.000	-0.003	-0.002

Probit Model Fit on 1988 Data to Predict 1989 Failures

Variables	dy/dx	Std. Err.	t		Pr(>|t|)	Conf. Int. Low	Cont. Int. Hi.
Capital Adequacy	-0.409	0.035	-11.762	***	0.000	-0.467	-0.352
Earnings Ratio	-0.004	0.028	-0.150		0.881	-0.051	0.042
NPL	0.096	0.015	6.325	***	0.000	0.071	0.120
Investment Securities	-0.028	0.007	-4.287	***	0.000	-0.039	-0.017
Large CDs	0.037	0.007	5.511	***	0.000	0.026	0.047
lnSize	-0.001	0.000	-5.353	***	0.000	-0.002	-0.001

Probit Model Fit on 1988 Data with additional failures from 85-87 to Predict 1989 Failures

Variables	dy/dx	Std. Err.	t		Pr(>|t|)	Conf. Int. Low	Cont. Int. Hi.
Capital Adequacy	-0.626	0.052	-12.107	***	0.000	-0.711	-0.541
Earnings Ratio	-0.556	0.061	-9.186	***	0.000	-0.656	-0.456
NPL	0.520	0.032	16.321	***	0.000	0.468	0.572
Investment Securities	-0.090	0.011	-8.175	***	0.000	-0.108	-0.072
Large CDs	0.112	0.012	9.505	***	0.000	0.093	0.132
lnSize	-0.006	0.000	-13.016	***	0.000	-0.006	-0.005

Panel C

Logit Model Fit on 1984 Data to Predict 1989 Failures

Variables	dy/dx	Std. Err.	t		Pr(>|t|)	Conf. Int. Low	Cont. Int. Hi.
Capital Adequacy	-0.159	0.027	-5.857	***	0.000	-0.204	-0.115
Earnings Ratio	-0.091	0.026	-3.542	***	0.000	-0.133	-0.049
NPL	0.134	0.017	7.807	***	0.000	0.106	0.163
Investment Securities	-0.016	0.006	-2.480	**	0.013	-0.026	-0.005
Large CDs	0.017	0.005	3.144	***	0.002	0.008	0.025
lnSize	-0.002	0.000	-8.161	***	0.000	-0.003	-0.002

Logit Model Fit on 1988 Data to Predict 1989 Failures

Variables	dy/dx	Std. Err.	t		Pr(>|t|)	Conf. Int. Low	Cont. Int. Hi.
Capital Adequacy	-0.416	0.034	-12.301	***	0.000	-0.472	-0.360
Earnings Ratio	-0.001	0.027	-0.038		0.970	-0.046	0.044
NPL	0.091	0.015	6.260	***	0.000	0.067	0.115
Investment Securities	-0.030	0.007	-4.416	***	0.000	-0.042	-0.019
Large CDs	0.039	0.007	5.884	***	0.000	0.028	0.049
lnSize	-0.001	0.000	-4.581	***	0.000	-0.001	-0.001

Logit Model Fit on 1988 Data with additional failures from 85-87 to Predict 1989 Failures

Variables	dy/dx	Std. Err.	t		Pr(>|t|)	Conf. Int. Low	Cont. Int. Hi.
Capital Adequacy	-0.629	0.051	-12.226	***	0.000	-0.714	-0.545
Earnings Ratio	-0.543	0.059	-9.211	***	0.000	-0.640	-0.446
NPL	0.506	0.031	16.386	***	0.000	0.455	0.556
Investment Securities	-0.094	0.011	-8.236	***	0.000	-0.112	-0.075
Large CDs	0.115	0.012	9.793	***	0.000	0.096	0.135
lnSize	-0.005	0.000	-11.743	***	0.000	-0.006	-0.004

*, **, *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively

Table X Type I and Type II Errors

Type I and Type II errors (rates) per year per model. A Type I error occurs when a model predicts that a bank will survive in the next period but actually fails. A Type II error occurs when a model predicts that a bank will fail in the next period but actually survives. Error rates are calculated by using the models fit with their listed year-end financial data to predict failures in the out of sample period each year; end-of-year financial results from 1989 are used to predict failure in 1990, end-of-year financial results from 1990 are used to predict failure in 1991, end-of-year financial results from 1991 are used to predict failure in 1992, end-of-year financial results from 1992 are used to predict failure in 1993, and end-of-year financial results from 1993 are used to predict failures in 1994. The last row displays the total Type I and Type II error rates per model specification. Error rates in parenthesis.

	Hazard		**Probit - 1984**		**Probit - 1988**		**Probit - 1984-88**		**Logit - 1984**		**Logit - 1988**		**Logit - 1984-88**	
	Type I	**Type II**	**Type I**	**Type II**	**Type I**	**Type II**	**Type I**	**Type II**	**Type I**	**Type II**	**Type I**	**Type II**	**Type I**	**Type II**
1989	799	3	1256	2	596	7	832	3	1283	2	557	7	742	3
	(6.43%)	(1.90%)	(10.11%)	(1.27%)	(4.80%)	(4.43%)	(6.70%)	(1.90%)	(10.33%)	(1.27%)	(4.48%)	(4.43%)	(5.97%)	(1.90%)
1990	706	4	1144	4	490	4	721	4	1170	4	458	4	635	4
	(5.83%)	(3.96%)	(9.44%)	(3.96%)	(4.05%)	(3.96%)	(5.95%)	(3.96%)	(9.66%)	(3.96%)	(3.78%)	(3.96%)	(5.24%)	(3.96%)
1991	525	18	882	18	313	19	542	19	897	18	291	19	478	19
	(4.48%)	(19.35%)	(7.53%)	(19.35%)	(2.67%)	(20.43%)	(4.63%)	(20.43%)	(7.65%)	(19.35%)	(2.48%)	(20.43%)	(4.08%)	(20.43%)
1992	303	4	523	4	161	4	307	4	526	4	153	4	269	4
	(2.68%)	(10.53%)	(4.62%)	(10.53%)	(1.42%)	(10.53%)	(2.71%)	(10.53%)	(4.64%)	(10.53%)	(1.35%)	(10.53%)	(2.38%)	(10.53%)
1993	156	2	268	1	75	3	165	1	266	1	69	4	146	2
	(1.44%)	(18.18%)	(2.47%)	(9.09%)	(0.69%)	(27.27%)	(1.52%)	(9.09%)	(2.45%)	(9.09%)	(0.63%)	(36.36%)	(1.34%)	(18.18%)
89-93	2489	31	4073	29	1635	37	2567	31	4142	29	1528	38	2270	32
	(4.26%)	(7.73%)	(6.97%)	(7.23%)	(2.80%)	(9.23%)	(4.39%)	(7.73%)	(7.09%)	(7.23%)	(2.61%)	(9.48%)	(3.88%)	(7.98%)
2007	34	11	101	10	23	12	101	10	121	10	24	12	84	10
	(0.47%)	(57.89%)	(1.40%)	(52.63%)	(0.32%)	(63.16%)	(1.40%)	(52.63%)	(1.68%)	(52.63%)	(0.33%)	(63.16%)	(1.17%)	(52.63%)
2008	147	35	391	21	79	46	368	22	415	21	76	48	333	23
	(2.13%)	(29.41%)	(5.66%)	(17.65%)	(1.14%)	(38.66%)	(5.33%)	(18.49%)	(6.01%)	(17.65%)	(1.10%)	(40.34%)	(4.82%)	(19.33%)
2009	371	4	674	3	241	6	662	3	709	3	236	8	611	3
	(5.58%)	(3.05%)	(10.13%)	(2.29%)	(3.62%)	(4.58%)	(9.95%)	(2.29%)	(10.66%)	(2.29%)	(3.55%)	(6.11%)	(9.18%)	(2.29%)
2010	385	1	647	1	276	1	615	1	678	1	272	1	577	1
	(6.02%)	(1.19%)	(10.12%)	(1.19%)	(4.32%)	(1.19%)	(9.62%)	(1.19%)	(10.60%)	(1.19%)	(4.25%)	(1.19%)	(9.02%)	(1.19%)
07-10	937	51	1813	35	619	65	1746	36	1923	35	608	69	1605	37
	(3.45%)	(14.45%)	(6.68%)	(9.92%)	(2.28%)	(18.41%)	(6.43%)	(10.20%)	(7.08%)	(9.92%)	(2.24%)	(19.55%)	(5.91%)	(10.48%)
Total	3426	82	5886	64	2254	102	4313	67	6065	64	2136	107	3875	69
	(4.00%)	(10.88%)	(6.88%)	(8.49%)	(2.63%)	(13.53%)	(5.04%)	(8.89%)	(7.09%)	(8.49%)	(2.50%)	(14.19%)	(4.53%)	(9.15%)

Table X - McNemar's Test Contingency Tables - Main Specification 89-93

Contingency tables provide the discordant errors for each model in a pairwise fashion. The upper left quadrant displays the number of banks incorrectly classified by both models. The lower right quadrant displays the number of banks correctly classified by both models. The discordant errors are displayed in the diagonal, and the more accurate model has the greatest number of banks correctly classified compared to the other model. Chi Square values are provided below, with *** representing statistical significance at the 0.01 level.

Probit fit to 1984
Hazard fit to 1984-1988

Hazard vs. Probit	Probit Wrong	Probit Right
Hazard Wrong	2501	19
Hazard Right	1601	54726
	Chi Square:	1545.939***

b Probit fit to 1984
Logit fit to 1984

Probit vs. Logit	Logit Wrong	Logit Right
Probit Wrong	4024	78
Probit Right	147	54598
	Chi Square:	20.551***

c Logit fit to 1984
Hazard fit to 1984-1988

Hazard vs. Logit	Logit Wrong	Logit Right
Hazard Wrong	2513	7
Hazard Right	1658	54669
	Chi Square:	1635.135***

Probit fit to 1988
Hazard fit to 1984-1988

Hazard vs. Probit	Probit Wrong	Probit Right
Hazard Wrong	1557	963
Hazard Right	115	56212
	Chi Square:	665.5***

e Probit fit to 1988
Logit fit to 1988

Probit vs. Logit	Logit Wrong	Logit Right
Probit Wrong	1554	118
Probit Right	12	57163
	Chi Square:	84.808***

f Logit fit to 1988
Hazard fit to 1984-1988

Hazard vs. Logit	Logit Wrong	Logit Right
Hazard Wrong	1439	1081
Hazard Right	127	56200
	Chi Square:	751.829***

Probit fit to 1988 incl. failures from 84-88
Hazard fit to 1984-1988

Hazard vs. Probit	Probit Wrong	Probit Right
Hazard Wrong	2216	304
Hazard Right	382	55945
	Chi Square:	8.643***

h Probit fit to 1988 incl. failures from 84-88
Logit fit to 1988 incl. failures from 84-88

Probit vs. Logit	Logit Wrong	Logit Right
Probit Wrong	2301	297
Probit Right	1	56248
	Chi Square:	292.03***

i Logit fit to 1988 incl. failures from 84-88
Hazard fit to 1984-1988

Hazard vs. Logit	Logit Wrong	Logit Right
Hazard Wrong	2051	469
Hazard Right	251	56076
	Chi Square:	65.401***

Table X - McNemar's Test Contingency Tables - Main Specification - 07-10

Contingency tables provide the discordant errors for each model in a pairwise fashion. The upper left quadrant displays the number of banks incorrectly classified by both models. The lower right quadrant displays the number of banks correctly classified by both models. The discordant errors are displayed in the diagonal, and the more accurate model has the greatest number of banks correctly classified compared to the other model. Chi Square values are provided below, with *** representing statistical significance at the 0.01 level.

Probit fit to 1984

Hazard fit to 1984-1988

Hazard vs. Probit	Probit Wrong	Probit Right
Hazard Wrong	971	17
Hazard Right	877	25643
	Chi Square:	825.37***

b

Probit fit to 1984

Logit fit to 1984

Probit vs. Logit	Logit Wrong	Logit Right
Probit Wrong	1834	14
Probit Right	124	25536
	Chi Square:	86.094***

c

Logit fit to 1984

Hazard fit to 1984-1988

Hazard vs. Logit	Logit Wrong	Logit Right
Hazard Wrong	971	17
Hazard Right	987	25533
	Chi Square:	935.22***

Probit fit to 1988

Hazard fit to 1984-1988

Hazard vs. Probit	Probit Wrong	Probit Right
Hazard Wrong	642	346
Hazard Right	42	26478
	Chi Square:	236.621***

Probit fit to 1988

Logit fit to 1988

Probit vs. Logit	Logit Wrong	Logit Right
Probit Wrong	667	17
Probit Right	10	26814
	Chi Square:	1.333

Logit fit to 1988

Hazard fit to 1984-1988

Hazard vs. Logit	Logit Wrong	Logit Right
Hazard Wrong	627	361
Hazard Right	50	26470
	Chi Square:	233.820***

Probit fit to 1988 incl. failures from 84-88

Hazard fit to 1984-1988

Hazard vs. Probit	Probit Wrong	Probit Right
Hazard Wrong	970	18
Hazard Right	812	25708
	Chi Square:	757.649***

Probit fit to 1988 incl. failures from 84-88

Logit fit to 1988 incl. failures from 84-88

Probit vs. Logit	Logit Wrong	Logit Right
Probit Wrong	1641	141
Probit Right	1	25725
	Chi Square:	136.063***

Logit fit to 1988 incl. failures from 84-88

Hazard fit to 1984-1988

Hazard vs. Logit	Logit Wrong	Logit Right
Hazard Wrong	970	18
Hazard Right	672	25848
	Chi Square:	617.984***

Table X - Out of Sample Classification Scores

This table presents out-of-sample classification scores for each model on year-end bank financial data from 1989-1993, 2007-2010, and the combination of those periods to predict bank failures in the next year. Each model specification is compared using 7 different classification metrics. The total out-of-sample (OOS) section pools all years to compare general OOS classification scores for each model. Bold text represents the best score for that classification metric in that sample period.

Out-of-Sample Classification Metric	**hazard**	**probit84**	**probit88**	**probit8488**	**logit84**	**logit88**	**logit8488**
89-93 Average Precision Score	0.526	0.506	0.545	0.525	0.505	**0.549**	0.530
89-93 F1 Score	0.973	0.958	0.981	0.972	0.957	**0.982**	0.975
89-93 Log Loss Score	0.017	0.017	**0.016**	0.024	0.017	0.016	0.026
89-93 Matthews Correlation Coefficient	0.337	0.267	0.400	0.332	0.265	**0.410**	0.350
89-93 Precision Score	0.129	0.084	0.182	0.126	0.082	**0.192**	0.140
89-93 ROC-AUC Score	0.940	0.929	0.940	0.939	0.928	0.940	**0.941**
89-93 Recall Score	0.923	0.928	0.908	0.923	**0.928**	0.905	0.920
07-10 Average Precision Score	0.551	0.526	**0.568**	0.526	0.522	0.563	0.531
07-10 F1 Score	0.974	0.956	0.980	0.957	0.953	**0.981**	0.960
07-10 Log Loss Score	0.033	**0.033**	0.035	0.043	0.033	0.034	0.046
07-10 Matthews Correlation Coefficient	0.446	0.351	**0.500**	0.356	0.342	0.497	0.369
07-10 Precision Score	0.244	0.149	0.318	0.154	0.142	**0.318**	0.164
07-10 ROC-AUC Score	0.911	0.917	0.897	0.917	0.915	0.891	**0.918**
07-10 Recall Score	0.856	0.901	0.816	0.898	**0.901**	0.805	0.895
Total OOS Average Precision Score	0.528	0.510	0.545	0.525	0.509	**0.546**	0.530
Total OOS F1 Score	0.973	0.957	0.981	0.967	0.956	**0.981**	0.970
Total OOS Log Loss Score	0.022	0.022	0.022	0.030	0.022	**0.022**	0.032
Total OOS Matthews Correlation Coefficient	0.372	0.297	0.433	0.343	0.293	**0.439**	0.359
Total OOS Precision Score	0.164	0.105	0.224	0.137	0.102	**0.232**	0.150
Total OOS ROC-AUC Score	0.926	0.923	0.919	0.930	0.922	0.917	**0.932**
Total OOS Recall Score	0.891	0.915	0.865	0.911	**0.915**	0.858	0.908