Forecasting High-Risk Composite CAMELS Ratings

Gaul, Lewis, Jonathan Jones, and Pinar Uysal

Please cite paper as:
Gaul, Lewis, Jonathan Jones, and Pinar Uysal (2019). Forecasting High-Risk Composite CAMELS Ratings. International Finance Discussion Papers 1252.

https://doi.org/10.17016/IFDP.2019.1252

International Finance Discussion Papers
Board of Governors of the Federal Reserve System

Number 1252
June 2019

Board of Governors of the Federal Reserve System

International Finance Discussion Papers

Number 1252

June 2019

Forecasting High-Risk Composite CAMELS Ratings

Lewis Gaul, Jonathan Jones, and Pinar Uysal

Forecasting High-Risk Composite CAMELS Ratings*

Lewis Gaul† Jonathan Jones‡ Pinar Uysal§

June 7, 2019

Abstract: We investigate whether statistical learning models can contribute to supervisors' off-site monitoring of banks' overall condition. We use five statistical learning and two forecast combination models to forecast high-risk composite CAMELS ratings over time (1984-2015), where a high-risk composite CAMELS rating is defined as a CAMELS rating of 3, 4, or 5. Our results indicate that the standard logit model, which is already widely used to forecast CAMELS ratings, comes close enough to be an adequate model for predicting high-risk ratings. We also find that the overall accuracy of the individual forecasts could be modestly improved upon by using forecast combination methods.

Keywords: Bank supervision and regulation, early warning models, CAMELS ratings, machine learning.

JEL classifications: G21, G28, C53.

*The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Office of the Comptroller of the Currency, of the U.S. Department of the Treasury, of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.

†Senior Financial Economist, Policy Analysis Division, Office of the Comptroller of the Currency, 400 7th St. SW, Washington, DC 20219, e-mail: Lewis.Gaul@occ.treas.gov

‡Lead Modeling Expert, Enterprise Risk Analysis Division, Office of the Comptroller of the Currency, 400 7th St. SW, Washington, DC 20219, e-mail: Jonathan.Jones@occ.treas.gov

§Senior Economist, Federal Reserve Board of Governors, 20th and Constitution Ave, N.W. Washington, D.C. 20551, e-mail: pinar.uysal@frb.gov

I. Introduction

Bank supervisors have various tools to evaluate the health of banks at their disposal and in their evaluation, supervisors use a combination of on-site examinations and off-site monitoring systems. Current supervisory practice based on FDIC Improvement Act of 1991 mandates that regulators complete annual on-site examinations for all banks, unless banks meet specific requirements which allow on-site exams to take place every 18 months.[1] After the examination, supervisors assign the bank a CAMELS rating that summarizes its financial condition and performance. Between on-site examinations, supervisors use off-site monitoring tools which typically analyze several bank risk factors such as balance sheet ratios and past examination ratings. Subsequently, the examination frequency is higher for banks that supervisors perceive to be in unsatisfactory condition based on off-site monitoring tools.

In this paper, we study whether several statistical learning models can contribute to supervisors' off-site monitoring of banks' overall condition in between on-site examinations. Specifically, we examine whether statistical learning models can aid in off-site monitoring by improving forecasts of commercial banks' CAMELS ratings. The CAMELS rating system is intended to classify the quality of a bank's financial condition, risk profile, and overall performance. The CAMELS name is an acronym for the six types of risk components assessed in the rating, which are (C) capital adequacy risk, (A) asset quality risk, (M) management risk, (E) earnings risk, (L) liquidity risk, and (S) sensitivity to market risk.[2] The CAMELS ratings range from 1, the lowest risk classification to 5, the highest risk classification. Most banks have ratings of 1 or 2 and are considered to be in satisfactory condition. Banks with a rating of 3, 4, or 5 are generally considered to be in unsatisfactory condition and are required to take actions to improve their conditions.

In our analysis, we investigate whether statistical learning models can improve forecasts of commercial bank risk ratings and provide evidence as to whether an investment in employing these models could be worthwhile for bank regulators. Specifically, we use bank risk factors to forecast whether a bank will have a high-risk CAMELS rating within the next four quarters. We use CAMELS ratings for about 6000 banks with a national charter over the period 1984-2015 along with information on several risk factors that predict CAMELS ratings to analyze the out-of-sample forecast power of statistical learning models.

Why should we try to use statistical learning models to forecast high risk CAMELS ratings? Statistical learning models have been shown to have substantial predictive power for other applications. For example, these models have been used to forecast credit scores, emergency room visits in hospitals, and sort spam email from email inboxes. Despite their impressive results, statistical learning models have not been frequently used in banking supervision literature to forecast important indicators or risk factors.

Why haven't statistical learning models already been put to use in economics and banking?

[1] Banks are allowed to defer their examinations for up to 18 months if the bank has assets of less than $500 million, the bank is well capitalized, the bank has both a management and composite CAMELS rating of 1 or 2, the bank is not subjected to an enforcement procedure or order by the Office of the Comptroller of the Currency (OCC), Federal Deposit Insurance Corporation (FDIC), or the Federal Reserve Board (FRB), and the bank has not had a change in control of the bank in the preceding 12 months.

[2] The sixth component (S) reflecting a bank's sensitivity to market risk was added in 1997.

To explain such scarcity, one could argue that the majority of researchers in banking supervision come from finance and economics backgrounds where researchers are generally trained to focus on tests of causal hypotheses in empirical research. As a result, researchers generally attempt to identify unbiased and/or consistent estimates of causal effects. In contrast, a main characteristic of several effective statistical learning models is that they often improve forecast performance by providing uninterpretable or biased estimates of the relation between variables. Therefore, statistical learning models are generally not well suited for economic research and have not made their way into mainstream and financial economists' toolkits. However, given these models' purported usefulness in several non-economic applications, we suggest that it is important to further examine the usefulness of statistical learning models in economic research applications, such as generating early-warning models of commercial bank risk.

We directly use five statistical learning models in our analysis. We use a logit model to reflect the standard statistical model that bank regulators and researchers would typically use to forecast CAMELS ratings. We also use a linear discriminant analysis model (LDA), a quadratic discriminant analysis (QDA) model, a mixture discriminant analysis model (MDA) and a flexible discriminant analysis model (FDA). In addition, we use two forecasts that are combinations of the individual five statistical learning models. The forecast combinations include a simple average of the five statistical learning models forecasts, and a logit combination model suggested by Kamstra and Kennedy (1998). We use the two forecast combination models to infer whether statistical learning models can be useful in conjunction with the standard logit model, even if any statistical learning model does not provide clearly superior performance over the standard logit model or any other particular statistical learning model we use.

Our results indicate that the standard logit model performs as well as or better than the other individual statistical learning models. We base this assertion on the result that the logit model correctly forecasts a larger fraction of banks with high-risk CAMELS ratings for a fraction of banks that are incorrectly forecasted to have a high-risk CAMELS ratings. That is, the logit model tends to have a higher true positive rate for a given false positive rate.

We also note that, the basic logit model performs similar to the other statistical learning models in classifying high-risk ratings and it performs better than the other individual models in classifying low-risk ratings. Additionally, the average and logit forecast combination models have better performance than individual models in general. At this point, given the simplicity of the logit models, we suggest that it is not clear from our results whether the small increase in performance from the forecast combination models justifies the cost of investing additional resources in implementing statistical learning models beyond the logit model. Further ongoing research will be need to better understand whether other statistical learning models can effectively contribute to early-warning model forecasts.

We also find that forecast accuracy from any model declined from the beginning of the sample period until the 2008 financial crisis period, and then substantially dropped in the immediate aftermath of the crisis, and is currently rising again. Therefore, our results suggest that just prior to the 2008 financial crisis, the regulators were relatively more likely to downgrade banks for reasons unrelated to bank risk factors we use in our models. However, during the stressful post-crisis period,

regulators began to downgrade banks more frequently for reasons correlated with our set of risk factors.

Finally, we point out that we only use a relatively small set of statistical learning models out of the total available universe of models, and that future work generating new forecasts could potentially identify other models with superior performance to the logit model or our forecast combination models. We also note that although the LDA, QDA, MDA, and FDA models are varieties of the same class of discriminant analysis models, these models do allow us to analyze a broad range of statistical modeling features that differ from standard logit and econometric techniques. For example, all of the discriminant analysis models are forms of Bayes classifiers, the MDA model uses flexible statistical mixture distributions, and our FDA implementation uses a highly flexible Multivariate Adaptive Regression Spline (MARS) model. Therefore, we emphasize that while we have implemented a limited number of models thus far, these models share modeling features with a wider range of statistical learning models.

II. Statistical Learning Models

In this section, we provide brief and simplified descriptions of the statistical learning models that we use to identify banks which would be assigned high-risk composite CAMELS ratings within the next four quarters. In all of our models we forecast a dummy variable, which we denote as $d_{i,t}$, equal to one if a bank, i, in time period t, has a high risk CAMELS rating of 3, 4, or 5 and is equal to zero otherwise. We attempt to forecast the CAMELS dummy with a matrix of variables which we denote as X, where $x_{j,i,t}$ denotes an observation of variable j for bank i at time t. In our forecast exercises, we predict the dummy variable between quarters $t+1$ and $t+k$ using information on predictor variables at date t.

We use five statistical learning models: Logit, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Mixture Discriminant Analysis (MDA), Flexible Discriminant Analysis (FDA), and two forecast combinations of these five models. All of our model descriptions are based in part on James, Witten, Hastie, and Tibshirani (2013), Hastie, Tibshirani, and Friedman (2016), and Kuhn and Johnson (2013).[3]

A. Logit

The base model that we use is a standard logit model that is commonly used in economics and finance research. First, we denote probability of high-risk and low-risk ratings conditional on our matrix of predictors as $P[d_{i,t+1,t+k} = 1|X]$ and $P[d_{i,t+1,t+k} = 0|X]$ respectively. The logit model assumes the following functional forms for the conditional high-risk and low-risk rating probabilities:

$$P\left[d_{i,t+1,t+k} = 1|X\right] = \frac{e^{X\beta}}{1 - e^{X\beta}}, \tag{1}$$

$$P\left[d_{i,t+1,t+k} = 0|X\right] = \frac{1}{1 - e^{X\beta}}. \tag{2}$$

[3]All three references discuss the logit, LDA, and QDA models, but the MDA and FDA models are only discussed in Hastie et al. (2016), and Kuhn and Johnson (2013).

To implement the logit model we estimate the vector of coefficients β that multiplies the matrix of predictor variables and denote the estimate as $\widehat{\beta}$. Given the estimate $\widehat{\beta}$, we predict the probability of high-risk and low-risk ratings as:

$$\widehat{P\left[d_{i,t+1,t+k}=1|X\right]} = \frac{e^{X\widehat{\beta}}}{1-e^{X\widehat{\beta}}}, \tag{3}$$

$$\widehat{P\left[d_{i,t+1,t+k}=0|X\right]} = \frac{1}{1-e^{X\widehat{\beta}}}, \tag{4}$$

where $\widehat{P\left[d_{i,t+1,t+k}=1|X\right]}$ and $\widehat{P\left[d_{i,t+1,t+k}=0|X\right]}$ denote the estimated high-risk and low-risk rating probabilities. Logit classifies a bank's rating as high-risk if the predicted probability of a high-risk rating is greater than some threshold τ:

$$\widehat{P\left[d_{i,t+1,t+k}=1|X\right]} > \tau. \tag{5}$$

B. LDA and QDA

LDA and QDA use straightforward applications of Bayes rule to estimate the probability of a high-risk rating conditional on the data in the X matrix. The formulas for Bayes Rule is given by:

$$P\left[d_{i,t+1,t+k}=r|X\right] = \frac{P\left[X|d_{i,t+1,t+k}=r\right]P\left[d_{i,t+1,t+k}=r\right]}{\sum_{r=0}^{1}P\left[X|d_{i,t+1,t+k}=r\right]P\left[d_{i,t+1,t+k}=r\right]}, \tag{6}$$

where r is equal to one for high-risk rating and zero for low-risk rating. LDA classifies a bank's rating as high-risk if the estimated probability of a high-risk rating is greater than some threshold τ, similar to the equation (5).

To implement LDA, we must calculate an estimate of equation (6). To do so, we first choose prior values for $P\left[d_{i,t+1,t+k}=r|X\right]$ and assume normal distribution for each r given by:

$$P\left[d_{i,t+1,t+k}=r|X\right] = \frac{1}{\sqrt{2\pi}|\mathbf{\Sigma}|^{1/2}}e^{-\frac{1}{2}(X-\mu_r)^T\mathbf{\Sigma}^{-1}(X-\mu_r)}. \tag{7}$$

Equation (7) assumes that the variance and covariance matrix of X, $\mathbf{\Sigma}$ is the same for all r and that the vector of means, μ_r, can differ for each r. The final step in estimating the high-risk rating with LDA is to estimate the the parameters of $P\left[d_{i,t+1,t+k}=r|X\right]$ which include the means and variances of equation (7). The LDA model uses the maximum likelihood estimates (MLEs) of these means and variances. The MLEs for mean of each variable, $j=1\ldots K$, in the μ_r vectors are:

$$\widehat{\mu_{j,r}} = \frac{\sum_{i=1}^{N_r}x_{i,j}}{N_r} \tag{8}$$

and, the MLE of each element of $\mathbf{\Sigma}$, which we denote as $\widehat{\sigma_{j,s}}$, where variables are denoted as $j=1\ldots K$ and $s=1\ldots K$ is given by:

$$\widehat{\sigma_{j,s}} = \frac{\sum_{i=1}^{N}(x_{i,j}-\widehat{\mu_j})(x_{i,s}-\widehat{\mu_s})}{N-1}. \tag{9}$$

LDA assumes variance and covariance between risk factors do not vary with whether banks have a high-risk rating or not (i.e. common $\mathbf{\Sigma}$ for high-risk and low-risk ratings). However, QDA assumes the variance and covariance between risk factors vary with whether banks have a high-risk rating or not (i.e. different $\mathbf{\Sigma}$ for high-risk and low-risk ratings).

C. MDA

A single normal distribution to model a class, as in LDA and QDA, could be too restrictive. MDA is an extension of LDA that models $P\left[d_{i,t+1,t+k} = r|X\right]$ as mixtures of one or more normal distributions separately for both classes. We denote these distributions as:

$$P\left[X|d_{i,t+1,t+k} = r\right] = \sum_{h=1}^{H_r} \alpha_{r,h} \frac{1}{\sqrt{2\pi}\mathbf{\Sigma}^{1/2}} e^{-\frac{1}{2}\left(X-\mu_{r,h}\right)'\mathbf{\Sigma}^{-1}\left(X-\mu_{r,h}\right)}. \tag{10}$$

In equation (10), h represents number of the sub-distribution within each $P\left[X|d_{i,t+1,t+k} = r\right]$ which runs $h = 1 \ldots H_r$, and $\alpha_{r,h}$ represents the weight applied to each sub-distribution where $\sum_{h=1}^{H_r} \alpha_{r,h} = 1$. In MDA, the probabilities of interest, $P\left[d_{i,t+1,t+k} = r|X\right]$, can be calculated simply by plugging the probabilities in equation (10) in equation (6).

To implement MDA, one needs to estimate each the set of parameters in equation (10) given by $\{\alpha_{r,h}, \mu_{r,h}, \mathbf{\Sigma} : r \in (0; 1)\}$. As in LDA, we use the maximum likelihood estimates each of these parameters. However, because of the summation in equation (10), the MLEs are complex and obtained with the Expectation Maximization (EM) algorithm. Because of the complexity of this algorithm, we relegate explanation of the details to Hastie et al. (2016). However, in summary, the main point is that to implement MDA, we simply compute estimates of the parameters in equation (10) and calculate $P\left[X|d_{i,t+1,t+k} = r\right]$, and then use these estimates in equation (6) along with the assumed values for the prior probabilities of failure to obtain estimates of $P\left[d_{i,t+1,t+k} = r|X\right]$.

D. FDA

In this section we give a brief description of the FDA model. Because of the technical complexity of the FDA model, we only provide a brief qualitative overview of the model and refer the reader for technical details to Hastie et al. (2016).

The main motivation for the FDA model is the realization that a linear regression of a dummy variable defining two classes onto a set of predictor variables provides coefficient estimates for the predictor variables that are proportional to a particular classification rule for the LDA model. If the estimated probability of a particular classification is greater than the threshold τ, the particular classification rule classifies observations to a particular class. It can be shown that the classification rule is a linear relation between the predictor variables and the odds ratio between the two classes.

FDA is a non-parametric and flexible alternative to LDA model and the ability to derive a classification rule from a linear regression is the basis of the FDA model. The main virtue of the FDA model is that the linear regression that motivates the FDA rule, can replaced by several linear non-parametric regression models. In this paper, we use a variant of the FDA model that replaces the linear regression model with the Multivariate Adaptive Regression Spline (MARS) model. This provides "flexibility" in defining models, and motivates the term Flexible Discriminant Analysis.

For our purposes, it can also be show that the estimated probability of a particular classification can be solved by solving for the probability of a classification from the odds ratio in the aforementioned classification rule. In our analysis, it is this probability that we will use to classify a high-risk rating.

III. Forecast Combination

Next, we generate two forecast combinations. Our forecast combination models are taken from Kamstra and Kennedy (1998). Our first forecast combination model is a simple majority vote of the existing high-risk rating forecasts based on whether high-risk rating probability estimates exceed the high-risk rating probability thresholds τ . In this model, we classify a CAMELS rating as an expected high-risk rating if 3 or more models classify the rating as an expected high-risk rating.

Our second forecast model attempts to combine the individual high-risk rating probability estimates into a joint high-risk rating probability estimate. Kamstra and Kennedy (1998) suggest that analysts can combine individual forecast probabilities by estimating a logit model with a high-risk rating dummy variable as the dependent variable and the high-risk rating probability estimates from the individual statistical learning models as the independent variables. Therefore, if we would like to estimate a model for quarter t, then the high-risk rating dummy variable is dated up to the period $t-4$ through $t-1$ and the high-risk rating probability estimates are dated up to $t-5$. We use these coefficient estimates to create the joint high-risk rating probability estimate for period t.

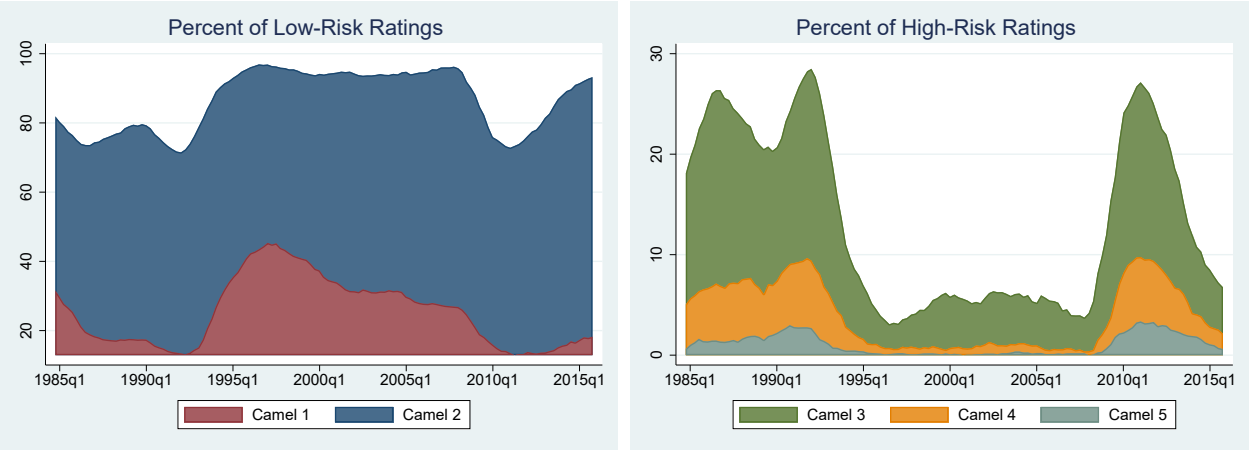
IV. Data

We gather data from the bank Call Report forms and confidential supervisory data on CAMELS ratings from the Office of the Comptroller of the Currency (OCC). We use data beginning in the fourth quarter of 1984 through the last quarter of 2015.

From the OCC, we gather quarterly information on OCC rated banks' current CAMELS rating at a given point in time. The OCC periodically examines banks and releases an up-to-date CAMELS rating upon completion of the exam. Figures 1a and 1b plot the fractions of banks with high- and low-risk ratings over time.

We construct eight different predictor variables from the Call Report data that have been used by previous researchers in forecast models to predict CAMELS ratings.[4] Supervisors have many variables and/or models to consider for assigning supervisory ratings, but we selected a more parsimonious specification to avoid over fitting. The variables we construct are a measure of balance sheet equity to assets (RCFD 3210 divided by RCFD 2170), return on assets (ROA) (the sum of the current and previous four quarters of quarterly RIAD 4300 divided by RCFD 2170), non-performing loans (NPL) to assets (the sum of the current and previous four quarters of quarterly RIAD 4230 divided by RCFD 2170), allowance for loan and lease losses (ALLL) to assets (RCFD 3123 divided by RCFD2170), loan loss provisions (LLP) to assets (RIAD4230 divided by RCFD2170), bank size

[4]Please see Cole, Gunther, and Cornyn (1995), Hirtle and Lopez (1999), and Bassett, Lee, and Spiller (2015).



(a) Low-Risk Ratings

(b) High-Risk Ratings

Figure 1. CAMELS Ratings over time

(log of RCFD 2170), total deposits (RCON 2200 divided by RCFD 2170), a measure of commercial and industrial lending exposure (RCON 1766 divided by RCFD 2170), and a measure of real estate lending exposure (RCFD 1410 divided by RCFD 2170). The summary statistics are provided in Table I.

Table I. Summary statistics

Variable	**Mean**	**Std. Dev.**
log(Assets)	11.534	1.398
Return on Assets	0.008	0.011
Equity/Assets	0.097	0.042
Loans/Assets	0.563	0.160
RE Loans/Assets	0.303	0.162
C&I Loans/Assets	0.107	0.078
Deposits/Assets	0.853	0.091
ALLL/Assets	0.009	0.006
LLP/Assets	0.004	0.008
N	315,771	

V. Estimation Results and ROC Analysis

Next, we discuss how we use Receiver Operating Curve (ROC) analysis to choose the rules we will use for forecasting high-risk CAMELS ratings with our probability estimates. In the ROC analysis, we attempt to infer an optimal rule to classify banks as having either high- or low-risk ratings. This rule will be based on basic probability theory and the expected costs of incorrectly predicting whether a bank has a high-risk CAMELS rating.

The ROC analysis makes reference to several basic definitions of forecast error rates which we explain in the next paragraphs. In this analysis, we refer to the hypothesis that a bank has a high-risk CAMELS rating in the future as the null hypothesis and define error rates in terms of this

null hypothesis. We classify a bank to have a high-risk CAMELS rating if the probability estimate of the bank's high-risk rating, denoted as p, is greater than or equal to a threshold, τ.

Table II presents the error matrix. In standard statistical language, a Type I error consists of false negatives and refers to rejecting the null hypothesis when it is true. In this paper, a Type I error refers to misclassifying a bank to have low-risk CAMELS rating, when a bank actually does receive a high-risk CAMELS rating. Therefore, the Type I error rate is the percentage of banks that have high-risk ratings but are predicted to have low-risk ratings. We denote a high-risk rating indicator variable, D, equal to 1 if a bank receives a high-risk rating and equal to 0 otherwise. Given this notation, the Type I error is defined as the case where $D = 1$ and $p \leqslant \tau$, and the Type I error rate is denoted as $P[p \leqslant \tau | D = 1]$. As a result, the number of banks correctly predicted to have high-risk ratings, where $D = 1$ and $p > \tau$, is often referred to as the true positive rate, and could be defined as $y(t) = P[p > \tau | D = 1]$.

Table II. Type I vs Type II Error

		Predicted Risk	
		High-Risk	Low-Risk
True Risk	High-Risk	True Positive	False Negative Type-I error
	Low-Risk	False Positive Type II error	True Negative

True Positive Rate (TPR) = $\frac{TP}{TP+FN}$
False Positive Rate (FPR) = $\frac{FP}{FP+TN}$

Similarly, a Type II error consists of false positives and refers to accepting the null hypothesis when it is not true. A Type II error refers to misclassifying a bank to have a high-risk rating, when it subsequently receives a low-risk rating. The Type II error is the case where $D = 0$ and $p > \tau$, and the Type II error rate is $P[p > \tau | D = 0]$. The number of banks incorrectly predicted to have high-risk ratings is often referred to as the false positive rate, and could be defined as $x(t) = P[p > \tau | D = 0]$.

A standard empirical ROC curve plots the true positive rate (1 minus the Type I error rate) against the false positive rate (Type II error rate) for a granular set of threshold τ values varying from 0 to 1. Each corresponding true positive rate and false positive rate is plotted as a point in a graph with the true positive rate on the vertical axis and the false positive rate on the horizontal axis. The scatter of points creates an approximation to a ROC curve that continuously varies with τ. Therefore, in terms of our notation, the ROC curve is a plot of $y(t) = P[p > \tau | D = 1]$ versus $x(t) = P[p > \tau | D = 0]$.

The main use of ROC analysis here is to better understand the expected error rates of our forecast model and to help pin down an optimal probability threshold τ, in order to determine which banks will be predicted to have a high-risk CAMELS rating. We will discuss how the ROC analysis can be used to assess the costs of forecast errors which will largely determine our optimal τ, and describe summary measures of forecast accuracy derived from the ROC analysis.

The ROC curve also provides overall summary measures of model fit. One common statistic is

the Area Under the Curve (AUC) which is defined as:

$$AUC = \int_0^1 y\,(x)\,dx. \tag{11}$$

The AUC is the integral over the domain of the true positive rate when the true positive rate, $y(t)$, is written as a function of the false positive rate, $x(t)$.

The AUC has two useful interpretations. The first interpretation is that it is the average value of the true positive rate where the true positive rate has a uniform distribution between zero and one. The second interpretation is that AUC measures the probability that a bank with a high-risk CAMELS rating will have a greater high-risk rating probability estimate than a bank with a low-risk CAMELS rating. For a proof of this second interpretation see Krzanowski and Hand (2009).

These interpretations of the AUC can be used to compare the performance of models. First, we could interpret models with higher AUC as superior because these models have a higher average true positive rate under a uniform distribution, hence perform better at correctly identifying high-risk ratings. However, one should temper this interpretation if users of the model care about segments of the true positive distribution conditioned on certain values of the false positive rates. For example, there are generally many more banks with low-risk ratings than banks with high-risk ratings. Hence, even moderately high false positive rate could lead to significant supervisory costs and make models relatively uninformative if hundreds of banks are erroneously predicted to have high-risk ratings. Furthermore, available models may not have uniform or even approximately uniform distributions and would significantly misstate the average true positive rates.

Figure 2 presents the ROC curve plots and figure 3 presents the true positive rate and the false positive rate against threshold values. The AUC ranges from 0.85 *(QDA model)* to 0.89 *(FDA model)*, which suggests that overall, the average available model can identify the majority of high-risk ratings. However, this result relies on models with probability thresholds that likely have false positive rates that are well above the level to be considered of practical use to the regulators. For example, choosing a model with over a 50 percent or greater false positive rate could lead analysts searching through more than 400 false leads to find about 3 high-risk ratings. The average and logit forecast combinations perform very similar when we compare AUC values, and their performance is slightly worse than the FDA model.

Figure 3 shows that setting the probability threshold for high-risk rating can be set fairly low (high) to increase (decrease) the true positive rate while not greatly increasing (decreasing) the false positive rate. The true positive rate of FDA and QDA models rise quickly if the probability threshold is lowered slightly below one and then rises gradually until the threshold is close to zero. The true positive rate of LDA and the MDA models rise slower than the FDA and QDA models when the threshold is lowered below one. Finally, the behavior of the logit model's true positive rate is somewhere in between these two sets of models and the average forecast combination model's true positive rate mimics the logit model's true positive rate. The logit forecast combination model's true positive rate appears to have a slow response when the cutoff is lowered below one, yet the true positive rate rises quickly once the threshold is around 0.90. Additionally, the true positive

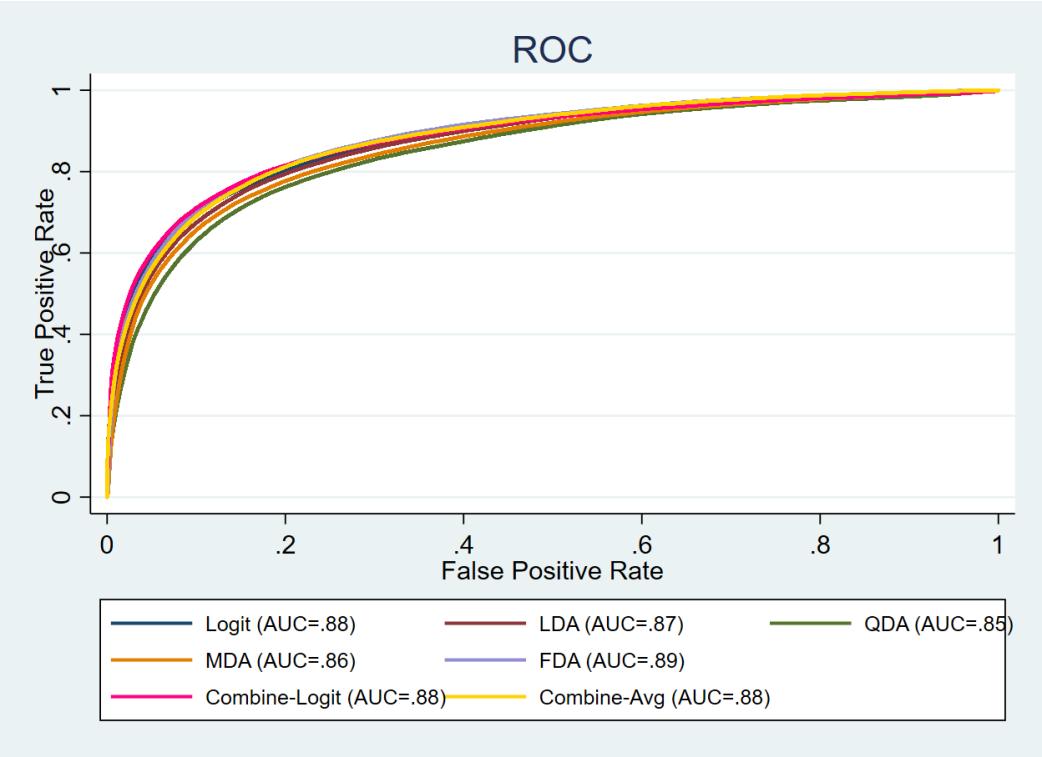


Figure 2. ROC and AUC

rate of the logit forecast combination model reaches one before any other model, once the cutoff value is lowered enough (around 0.05). Overall, Figure 3 shows that models' true positive rates respond somewhat differently to the varying threshold values, however the models' false positive rates respond to the varying threshold values in a very similar manner.

One remedy to the limitations of the AUC is to examine the integral of the true positive rate over a constrained range of the false positive rates. This is referred to the Partial Area Under the Curve (PAUC) and is given by:

$$PAUC = \int_a^b y\,(x)\,dx. \tag{12}$$

The PAUC measures the area of the ROC curve for a restricted range of false positive rates. While there is no simple interpretation of the PAUC, if we divide PAUC estimates by $b-a$, the resulting values across models can can be used to compare different models. The PAUC normalized by $b-a$ provides us with an estimate of the average true positive rate within the range of false positive rates from b to a under the assumption that the set of available models in the range from b to a are uniformly distributed.

We calculate normalized PAUC values for several non-overlapping partitions between 0 and 100 percent. This will provide information on average true positive rates for local and cumulative sets of the total ROC curve and allow analysts to better infer the ability of a subset of more practical useful models to detect high-risk ratings. Table III presents the PAUC values which indicate that the average true positive rate for the false positive rate in the 0.01-0.10 range may be much less than the values suggested by the overall ROC curve.

For example, the logit model has PAUC values of 0.40 to 0.65 in the 0.01-0.10 false positive

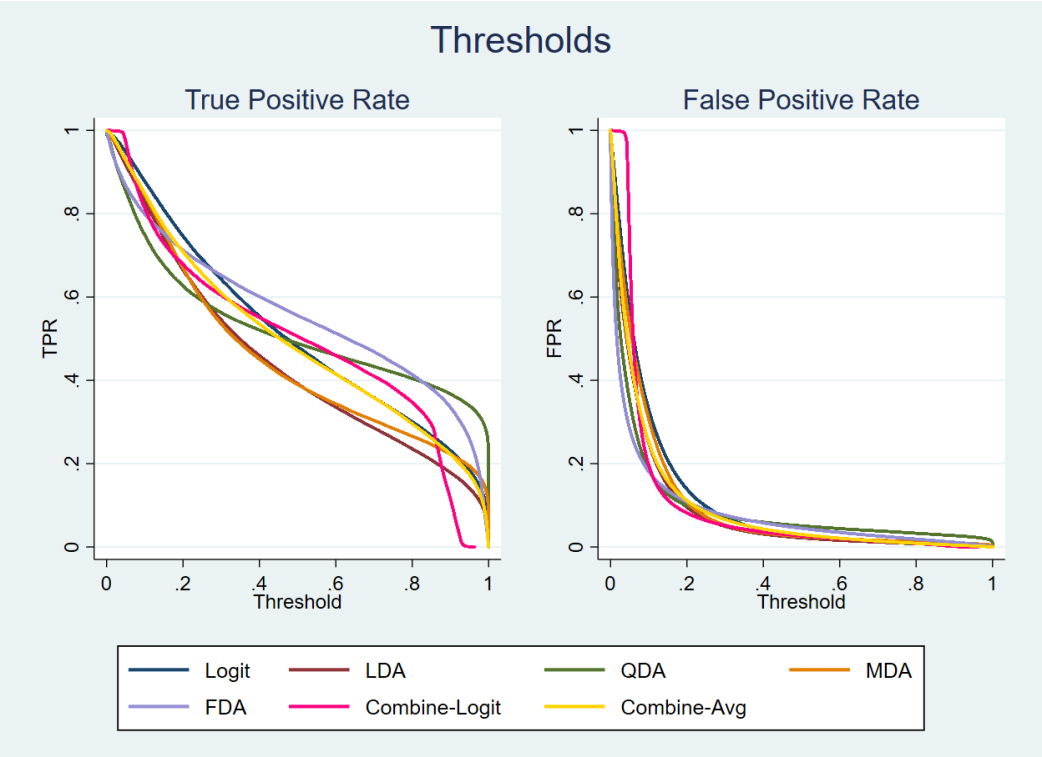


Figure 3. Thresholds vs True Positive Rate and False Positive Rate

rate range which suggests Type I error rates of 0.35 to 0.60. This suggests that the probability of misclassifying a high-risk rating as a low-risk rating could be much higher that the Type I error rate of 0.13 implied by the average true positive rate of about 0.87 suggested by the overall AUC estimates in Figure 2.

The results for the LDA, QDA, MDA, FDA, and forecast combination models are similar to those from the logit model and also imply that the feasible false negative rate is likely vastly higher than the 11 to 15 percent suggested by the overall AUC estimates. Therefore, focusing on the PAUC results would provide a much more realistic and different expected true positive rate than the overall AUC results.

A. Choosing the Classification Threshold τ

To pin down an optimal threshold, we derive the formula for the expected cost of forecast errors for an individual bank. We define the expected cost of forecast errors as a weighted average of the cost of Type I and Type II errors. We denote the cost of a Type I error as $C_1 = C\left[p \leqslant \tau | \, D = 1\right]$ and the cost of a Type II error as $C_2 = C\left[p > \tau | \, D = 0\right]$. The expected cost is given by:

$$C(t) = P(High\ Risk) \times TPR(\tau) \times C_1 + P(Low\ Risk) \times FPR(\tau) \times C_2, \tag{13}$$

where TPR is the true positive rate and FPR is the false positive rate. To derive the optimal threshold, τ, we maximize equation (13) with respect to τ. At the optimum, the slope of the ROC curve is given by:

Table III. PAUC for the Year-Ahead Forecasts

FPR Range	Logit (1)	LDA (2)	QDA (3)	MDA (4)	FDA (5)	Forecast Combination Logit (6)	Forecast Combination Avg (7)
.01-.02	0.390	0.322	0.261	0.290	0.376	0.419	0.364
.02-.03	0.473	0.416	0.349	0.388	0.458	0.495	0.441
.03-.04	0.527	0.480	0.415	0.458	0.514	0.546	0.498
.04-.05	0.567	0.527	0.464	0.506	0.556	0.583	0.543
.05 -.1	0.647	0.618	0.565	0.598	0.645	0.661	0.632
.1 -.2	0.755	0.742	0.705	0.724	0.765	0.769	0.758
.2 -.3	0.836	0.829	0.798	0.811	0.848	0.844	0.845
.3 -.4	0.885	0.880	0.853	0.865	0.897	0.889	0.892
.4 -.5	0.920	0.916	0.894	0.904	0.929	0.920	0.924
.5 -.6	0.945	0.945	0.928	0.935	0.952	0.944	0.950
.6 -.7	0.963	0.967	0.952	0.956	0.970	0.961	0.969
.7 -.8	0.977	0.981	0.968	0.972	0.982	0.975	0.983
.8 -.9	0.987	0.990	0.979	0.984	0.991	0.985	0.991
.9 -1.0	0.995	0.997	0.991	0.994	0.998	0.993	0.998

Note: FPR Range is the range of false positive rates.

$$\underbrace{\frac{\partial TPR(\tau)}{\partial FPR(\tau)}}_{Slope\ of\ ROC} = \frac{P(Low\ Risk) \times C_2}{P(High\ Risk) \times C_1}. \tag{14}$$

Therefore, we can find the optimal threshold, τ, by estimating where the slope of the ROC curve equals the value on the right hand side of Equation (14).

We try to decide on a value of τ by examining the cost minimizing values of τ for different assumptions about the relative costs of Type I and Type II errors. In this analysis, we normalize the cost of a false positive rate to 1, allow the cost of false negative rate to run from 1 to 20, and calculate the cost minimizing value of τ for each set of costs. Because we do not have enough supervisory experience to definitively determine the relative cost of Type I and Type II errors, in order to classify banks as having high-risk ratings for our analysis, we arbitrarily assume that the relative cost of a Type I versus Type II error is 10 to 1. As a result of this assumption, the Table IV presents the various thresholds and the average costs associated with them. The logit model has the largest threshold at 0.115 and the FDA model has the smallest threshold at 0.047. Next, using these optimal thresholds for each model, we estimate the forecast combination models. The average and logit forecast combination models have optimal threshold values of 0.081 and 0.086 respectively, which are within the range of the largest and smallest thresholds.

VI. Historical Performance of Models

We present time series information on each model's performance using the models based on the optimal thresholds from the previous section. Figure 4 plots the out-of-sample Type I error rates

Table IV. Cost minimizing probability thresholds

FN Cost	Logit τ	Logit Avg Cost	LDA τ	LDA Avg Cost	QDA τ	QDA Avg Cost	MDA τ	MDA Avg Cost	FDA τ	FDA Avg Cost	Forecast Combination: Logit τ	Forecast Combination: Logit Avg Cost	Forecast Combination: Average τ	Forecast Combination: Average Avg Cost
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	0.438	0.11	0.359	0.12	0.614	0.13	0.363	0.12	0.568	0.11	0.463	0.11	0.464	0.12
2	0.314	0.18	0.227	0.19	0.252	0.21	0.241	0.20	0.305	0.18	0.233	0.18	0.256	0.19
3	0.247	0.24	0.180	0.25	0.161	0.27	0.198	0.26	0.177	0.24	0.168	0.23	0.202	0.24
4	0.203	0.29	0.148	0.30	0.118	0.32	0.169	0.31	0.151	0.28	0.138	0.28	0.164	0.29
5	0.177	0.33	0.131	0.34	0.103	0.37	0.158	0.36	0.111	0.32	0.109	0.32	0.136	0.33
6	0.165	0.37	0.118	0.38	0.091	0.41	0.140	0.40	0.082	0.36	0.108	0.36	0.115	0.36
7	0.145	0.40	0.110	0.41	0.078	0.45	0.126	0.43	0.075	0.39	0.104	0.39	0.109	0.39
8	0.132	0.43	0.094	0.44	0.064	0.48	0.116	0.47	0.059	0.41	0.086	0.42	0.099	0.42
9	0.116	0.46	0.085	0.47	0.060	0.51	0.104	0.50	0.056	0.44	0.086	0.45	0.099	0.44
10	**0.115**	**0.48**	**0.080**	**0.49**	**0.054**	**0.54**	**0.090**	**0.52**	**0.047**	**0.46**	**0.081**	**0.47**	**0.086**	**0.47**
11	0.099	0.51	0.076	0.52	0.053	0.57	0.086	0.54	0.038	0.48	0.077	0.50	0.083	0.49
12	0.094	0.53	0.069	0.54	0.038	0.59	0.075	0.57	0.036	0.50	0.074	0.52	0.083	0.51
13	0.083	0.54	0.059	0.56	0.034	0.61	0.074	0.58	0.036	0.52	0.073	0.54	0.074	0.53
14	0.083	0.56	0.059	0.57	0.026	0.63	0.062	0.60	0.03	0.53	0.068	0.56	0.067	0.55
15	0.074	0.58	0.045	0.59	0.026	0.64	0.061	0.62	0.028	0.55	0.068	0.57	0.067	0.57
16	0.074	0.59	0.045	0.60	0.022	0.65	0.053	0.63	0.026	0.56	0.065	0.59	0.051	0.58
17	0.063	0.61	0.038	0.61	0.021	0.67	0.049	0.64	0.026	0.58	0.062	0.61	0.051	0.59
18	0.062	0.62	0.038	0.62	0.019	0.68	0.049	0.66	0.022	0.59	0.062	0.62	0.044	0.60
19	0.062	0.63	0.038	0.63	0.019	0.69	0.047	0.67	0.022	0.60	0.062	0.63	0.044	0.61
20	0.056	0.64	0.037	0.64	0.019	0.70	0.046	0.68	0.022	0.62	0.061	0.64	0.044	0.63

Note: FN Cost is the relative cost of false negative to false positive.

over time and figure 5 plots the out-of-sample Type II error rates over time.

Overall, all of the models tend to show relatively high Type I error rates leading up to the 2008 financial crisis and a sharp drop in the Type I error rates during the financial crisis. We also note that the type I error rates are currently rising again. Relatively high false negative rates just prior to the financial crisis would suggest that the regulators were more likely to assign high-risk ratings for reasons unrelated to bank risk factors that we use in our models. That said, the sharp drop in the false negative rate during the stressful financial crisis period would imply that the regulators were relatively more likely to assign a high-risk rating for reasons correlated with the bank risk factors we used in our models.

In contrast, all of the models tend to show relatively low Type II error rates leading up to the 2008 financial crisis and a sharp increase in the Type II error rates during the financial crisis. There is a similar increase in the false positive rates after the 1990s recession. The steep increase in the false positive rates during periods with economic stress would imply that the regulators were relatively less likely to assign a low-risk rating for reasons uncorrelated with the bank risk factors we used in our models. One potential explanation might be that the supervisory standards used in the assignment of CAMELS ratings were unnecessarily tight during these periods and supervisory were more likely to assign high-risk ratings. This is in line with Bassett, Lee, and Spiller (2015)'s finding that even though the supervisory standards appear to have been fairly constant over time (1991-2013), the standards have been somewhat tighter than average during the early 1990s and in 2008.

The largest Type I error rates are observed right before the 2008 financial crisis and are about

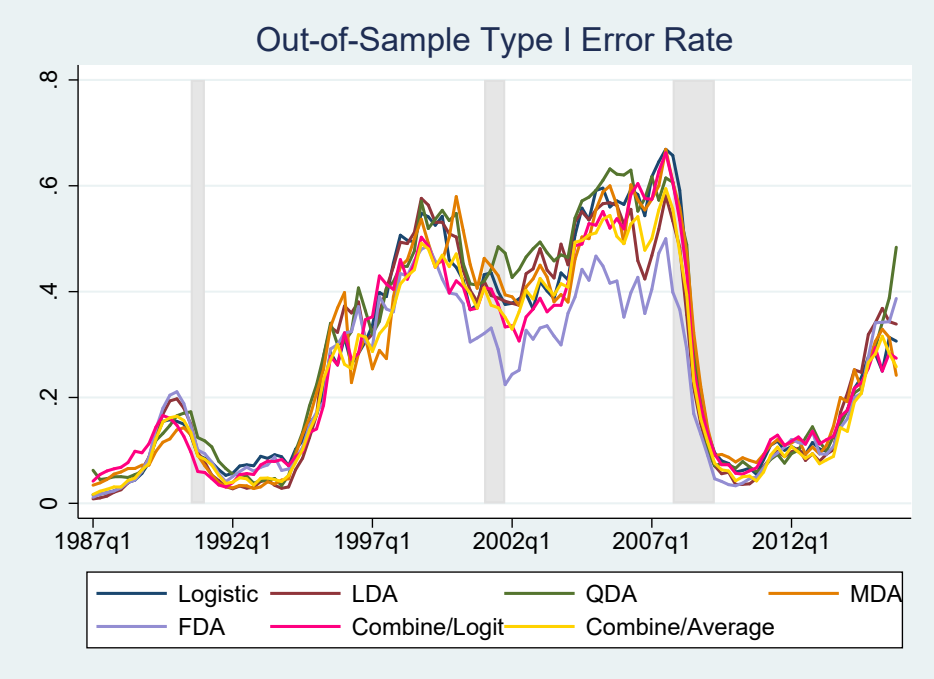


Figure 4. Type I Error

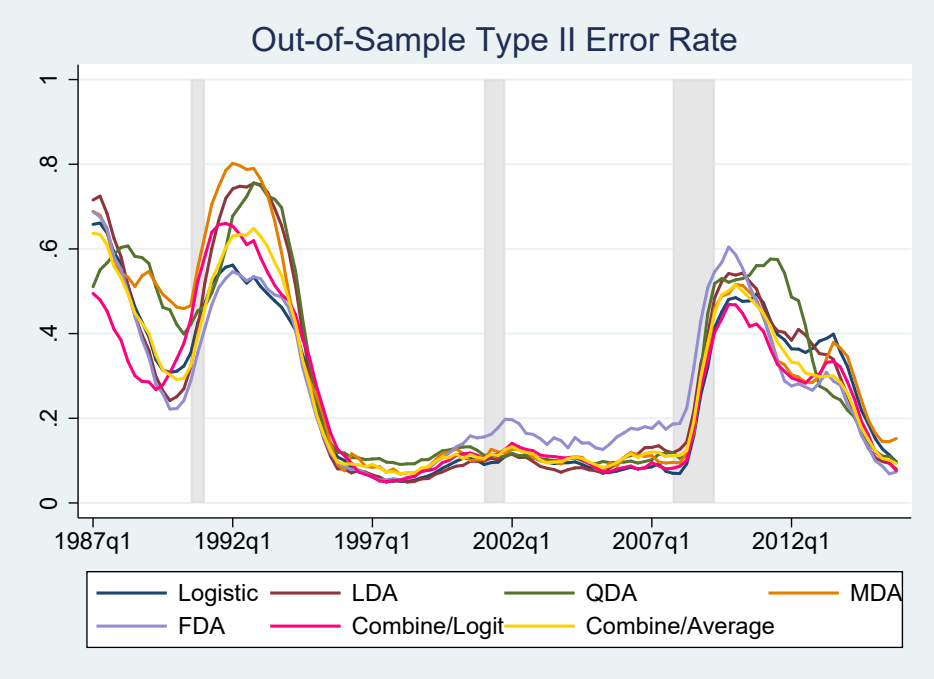


Figure 5. Type II Error

0.45 to 0.65. Table V shows that Type I error rates are about 0.23 to 0.28 on average and are generally higher during recession periods compared to periods without a recession. Type I error rates imply that the models correctly identify about 72 to 77 percent of high-risk rating banks. Additionally, Type I error rates are well above the 0.13 Type I error rate suggested by the AUC, but are more consistent with the Type I error rates of about 0.30 to 0.65 suggested by the PAUC as discussed earlier.

The highest Type II error rates are observed in the late 1980s, and after the early 1990s and 2000s recessions. Table V shows that Type II error rates are about 0.27 to 0.31 and are only slightly higher than Type I error rates on average. This implies our models misclassify a low-risk rating as a high-risk rating slightly more often than they misclassify a high-risk rating as a low-risk rating. This is the outcome of our assumption that the cost of false negatives is higher than the cost of false positives. Therefore, we should be certain that false positives truly are less costly than false negatives, as it is apparent that the models will indeed generate noise. This would imply that if regulators wanted to identify all potential high-risk ratings, they would have to search through a large number of false leads.

Table V. Type I and Type II Error Statistics

	Type I			Type II			TPR/FPR		
		No			No			No	
	Overall	Recession	Recession	Overall	Recession	Recession	Overall	Recession	Recession
Logit	0.268	0.262	0.313	0.265	0.272	0.216	2.760	2.715	3.181
LDA	0.262	0.260	0.273	0.288	0.294	0.249	2.561	2.520	2.925
QDA	0.284	0.279	0.323	0.304	0.311	0.254	2.355	2.321	2.666
MDA	0.266	0.259	0.317	0.306	0.312	0.263	2.398	2.376	2.594
FDA	0.225	0.226	0.214	0.278	0.275	0.296	2.794	2.814	2.651
Comb/Logit	0.256	0.253	0.284	0.257	0.258	0.247	2.895	2.895	2.901
Comb/Avg	0.243	0.240	0.273	0.272	0.276	0.235	2.785	2.751	3.089
N	118	104	14	118	104	14	118	104	14

Note: TPR/FPR is the ratio of true positive rate to false positive rate.

We also note that the logit model has similar performance to the other statistical learning models—except for the FDA model—in identifying high-risk ratings and it performs better than the other individual models in identifying low-risk ratings. FDA model has the best performance in identifying high-risk ratings, but performs worse than the logit model in identfying low-risk ratings. FDA's superior performance in identifying high-risk ratings is mainly driven by the fact that it has the lowest classification threshold, τ, among all the other models.

Last three columns of Table V reports the ratio of true positive rate to false positive rate. Overall, the logit model performs better than the other individual statistical learning models, except for the FDA model. We base this assertion on the result that the logit model correctly forecasts a larger fraction of banks with high-risk CAMELS ratings for a fraction of banks that are incorrectly forecasted to have a high-risk CAMELS ratings. That is, the logit model tends to have a higher true positive rate for a given false positive rate. Also, the logit model outperforms the other models during recessionary periods.

Additionally, the average and logit forecast combination models have better performance than individual models in general. They can classify high-risk banks better than the individual models,

except for the FDA model, and they can classify low-risk banks better than the individual models, except for the logit model. The forecast combination models also have the highest true positive rate for a given false positive rate overall.

Taken together, given the simplicity of the logit models and economists' familiarity with it, we suggest at this point, there is not much reason to consistently favor any other model over the standard logit model for forecasting high-risk CAMELS ratings. However, we recommend that we should continue to search through other models which might be an improvement over the logit model.

VII. Conclusion

In this paper, we study potential early-warning models that could help supervisors identify which banks will be assigned a high-risk CAMELS rating within a year. We use five statistical learning and two forecast combination models to investigate whether statistical learning models can improve forecasts of high-risk CAMELS ratings. We use information for over 6000 banks' CAMELS ratings over the period 1984-2015 along with information on several bank risk factors that help predict CAMELS ratings to analyze the out-of-sample forecast power of several statistical learning models. Our results indicate that the standard logit model, which is already widely used to forecast CAMELS ratings, comes close enough to be an adequate model for predicting high-risk ratings. We also find that the overall accuracy of the forecasts could be modestly improved upon by using forecast combination methods.

References

Bassett, W. F., S. J. Lee, and T. P. Spiller (2015). Estimating Changes in Supervisory Standards and Their Economic Effects. *Journal of Banking & Finance 60*, 21–43.

Cole, R., J. Gunther, and B. Cornyn (1995). FIMS: A New Financial Institutions Monitoring System for Banking Organizations. *Federal Reserve Bulletin 81*, 1–15.

Hastie, T., R. Tibshirani, and J. Friedman (2016). *The Elements of Statistical Learning*. Springer.

Hirtle, B. and J. Lopez (1999). Supervisory Information and the Frequency of Bank Examinations. *Economic Policy Review 5*, 1–19.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning*. Springer.

Kamstra, M. and P. Kennedy (1998). Combining Qualitative Forecasts Using Logit. *International Journal of Forecasting 14*, 83–93.

Krzanowski, W. J. and D. J. Hand (2009). *ROC Curves for Continuous Data*. Taylor and Francis.

Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling*. Springer.