European Journal of Operational Research 290 (2021) 358–372

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Interfaces with Other Disciplines

Predicting mortgage early delinquency with machine learning methods

Shunqin Chen[a], Zhengfeng Guo[b], Xinlei Zhao[c,*]

[a] Shunqin Chen is a Senior Economist at Fannie Mae, United States
[b] Zhengfeng Guo is a Principal Economist at Fannie Mae, 1100 15th St NW, Washington, DC 20005, United States
[c] Xinlei Zhao is a Deputy Director of the Commercial Credit Risk Analysis Division of the Office of the Comptroller of the Currency, 400 7th Street SW, Mail Stop 6E-3, Washington, DC 20219, United States

ARTICLE INFO

Article history:
Received 29 April 2019
Accepted 29 July 2020
Available online 5 August 2020

Keywords:
Credit scoring model
Mortgage early delinquency
Machine learning
Gradient boosting
Random forest
Neural network, Ensemble

ABSTRACT

This paper investigates the performance of thirteen methods for modelling and predicting mortgage early delinquency probabilities. These models include variants of logit models, some commonly used machine learning methods, and variants of ensemble models. We find that heterogenous ensemble methods lead other methods in the training, out-of-sample, and out-of-time datasets in terms of risk classification. Nonetheless, various predictive accuracy performance measures yield different rankings among the thirteen methods and no method consistently dominates in this performance dimension in the training, out-of-sample, and out-of-time data. Lastly, predictive accuracy is a major challenge facing all mortgage early delinquency models, even in the training data.

Published by Elsevier B.V.

1. Introduction

Credit risk (i.e., the failure to make required payments to a debt) is the primary risk facing financial institutions, and a sudden spike in credit risk was at the center of all banking crises in the past. As a result, managing credit risk is a critical component of risk management operations among financial institutions. In the recent decades, risk management tools for consumer loans have become more and more model driven, and financial institutions typically use credit scoring models to assess the delinquency or default probabilities of consumer loans, such as mortgages, auto loans, credit cards, or personal loans.[1] Today, most lending decisions on consumer loans in the financial industry are automated based on model outputs. Financial institutions use these models to assess borrowers' creditworthiness, and to determine whether the borrower will be granted a loan and if so, how to form loan contracts. From this perspective, credit scoring models are closely related to a bank's lending decisions and have a significant impact on the operating profitability of the portfolios the models are applied to. Consequently, developing good credit scoring models and fine-tuning such models has become a central function of risk management operations in all financial institutions. Even a small improvement in the predictive power of such models can potentially significantly improve a financial institution's profitability, and a small change in the scoring models might be associated with a major shift in a financial institution's business strategy.

Given this background, not surprisingly, many researchers have shown tremendous interests in such models. A widely established technique for this type of modeling is the logistic regression. Starting from the 1990s, academic researchers started to apply machine learning (ML) methods to consumer credit risk probability modeling.[2] Financial institutions have also shown an increase in interest in ML methods; they have been testing, and in some cases, have started implementing ML methods for operational decisions.[3]

Although the outstanding balance in the mortgage segment is multiple times higher than all other lines of consumer debt combined, and mortgage lending is a major line of business operations in financial institutions, there are fewer papers on credit scoring models for mortgages than for credit cards or personal loans

* Corresponding author.
E-mail address: Xinlei.Zhao@occ.treas.gov (X. Zhao).

[1] See for example, the summary in Thomas, Crook, and Edelman (2017).

[2] For instance, David, Edelman and Gammerman (1992), Desai, Crook, and Overstreet (1996), Galindo and Tamayo (2000), Baesens et al. (2003), Crook, et al. (2007), Martens et al. (2007), Khandani, Kim, and Lo (2010), Kennedy, Namee, and Delany (2013), Verbraken, et al. (2014), Lessmannn et al. (2015), Fitzpatrick and Mues (2016), Butaru et al. (2016), Li, Bellotti, and Adams (2017), Addo, Guegan, and Hassan (2018), Kvamme et al. (2018), and Sirignano, Sadhwani, Giesecke (2018), and Li et al. (2020).

[3] See for example, https://www.americanbanker.com/news/the-banks-warming-to-ai-based-lending?tag=0000015a-1e76-d1c9-adfa-1e7e0cf70000.

https://doi.org/10.1016/j.ejor.2020.07.058
0377-2217/Published by Elsevier B.V.

in the credit scoring literature. This paper aims to move the literature on mortgage credit scoring models forward by exploring ML methods for mortgage early delinquency prediction. Some earlier papers have investigated application of ML methods on mortgage models (for instance, Lessmann, Baesens, Seow, and Thomas (2015) and Fitzpatrick and Mues (2016)). We add to this literature by applying the ML methods to a much larger sample representing the US prime mortgage market over many post-crisis periods, and we conduct a comprehensive analysis along the full range of the operational issues. We contribute to the literature in two ways. First, most of the papers in the literature (for example, Lessmann et al. (2015) and Fitzpatrick and Mues (2016)) largely focus on risk classification or rank ordering of credit risk probabilities, although a few recent papers (such as Sirignano, Sadhwani, and Giesecke (2018)) have started to investigate predictive accuracy at portfolio level. Risk classification or rank ordering the credit risk probability is a dimension that typically does not overlap much with the assessment of the level of credit risk probability, with the latter more difficult to capture. Such a distinction has been pointed out in the prior literature, such as Qi, Zhang, and Zhao (2014), but is still commonly overlooked in the literature.

Model predictive accuracy is important for credit scoring models because the relative riskiness of a borrower (risk classification or rank ordering) cannot be directly used in risk management. One way scoring models are used in financial institutions is that the model output is a key input downstream in the net-present-value (NPV) models for underwriting purpose, and the *absolute* levels of predicted delinquency or default rate are directly used in NPV calculation. For lenders not using NPV models for underwriting, they typically apply cut-off points to the rank orders of delinquency/default probabilities when making the decision on which borrowers can obtain loans. In this case, simply knowing that borrower A is riskier than borrower B does not mean that borrower A will be granted a loan while borrower B will be denied. It is possible that both borrowers might be denied, or they might both be granted the loan. The decision of whether a borrower can be granted a loan depends on cut-off points applied to the models, and the cut-off points depend on the prediction of the *absolute* level of credit risk of the future. In addition, when setting the various features of loan contracts (such as the size of the mortgage, interest rate, and the design of payment schedule, etc.), lenders also rely on the *absolute* predicted level of riskiness, but not the relative riskiness of borrowers. For example, when setting interest rates, lenders need to assess the absolute level of predicted credit risk of borrowers so that the interest rates charged on the loans or other features of the loan contracts can offset the expected levels of potential credit losses. Accordingly, the trend in the financial industry over the recent decade is the increasing emphasis on predictive accuracy of credit scoring models.

Second, we place a high emphasis on out-of-time analysis, as credit scoring models in financial institutions are always implemented out-of-time in real business operations. Driven by the lifecycle of a model from data collection, model development, model validation to final model production, models are typically applied out-of-time at least one-year after the ending period of the development data. As such, for real operational purposes, a model that performs well in-time but poorly out-of-time would not be particularly useful. However, this out-of-time investigation has not been well understood in the financial industry. Based on conversations with practitioners, it is far from uncommon for credit scoring models to be put in production in financial institutions before sufficient out-of-time evidence is collected. Out-of-time investigation does not receive much attention in the existing literature either, as the findings in the prior studies on ML methods are typically based on in-sample and out-of-sample evidence (see for example, the discussions in Varian (2014)). However, conducting out-of-time investigation for ML methods is particularly important-(see Hand (2009a) for a comprehensive explanation). A statistical model is built on consumer behavior observed in historical data, but such behavior could change in the future and the historical pattern observed at certain periods for a sub-population may not be generalized over time. Therefore, ML methods could be more sensitive to over-time disturbances because of their complexity and flexibility in the training process, which might be more subject to the overfitting problem and become more unstable over time. Additionally, a lack of interpretability of a complex model may pose new challenges and uncertainties in generalizing the ML models over time. From this perspective, the performance of a ML model that is well trained in-sample and performs well out-of-sample may deteriorate more significantly than that of traditional statistical models out-of-time. Therefore, it is important to examine the relative performance of ML models out-of-time.[4]

ML methods are primarily built for prediction, not for parameter estimation - i.e., trying to identify the underlying causality relationship between the explanatory variables and outcome variables (see the discussions in Mullainathan and Spiess (2017)). As a result, we focus on the prediction of mortgage early delinquency probabilities in this paper instead of investigating how various factors drive early delinquencies in mortgages.

We explore the performance of thirteen methods in this paper. In addition to the general logit regressions with linear terms (GLM1) and piece-wise linear splines (GLM2), we also include a logit regression with P-splines added to some explanatory variables (GLM3). In addition, we investigate the regularized logistic regression (RLGR), generalized additive model (GAM) (Hastie & Tibshirani, 1990), Neural Networks (NN) (Bishop, 1995), Support Vector Machine (SVM) (Cortes and Vapnik (1995)), and K-Nearest Neighbors (KNN) (Altman (1992)). For ensembled classifiers, we investigate three homogenous ensemble methods–AdaBoost (ADA) (Freund, Schapire, and Abe (1999); Nikolaou, Edakunni, Kull, Flach, & Brown, 2016), Random Forest (RF) (Breiman (2000, 2004), Ernst and Wehenkel (2006), and Genuer, Poggi, and Tuleau (2008)), and Extreme Gradient Boosting (XGB) (Chen & Guestrin, 2016),[5] and two heterogenous ensemble methods–Stack (Opitz & Maclin, 1999) and Hill-climbing (HC) ensemble selection (Caruana, Niculescu-Mizil, Crew and Ksikes (2004)). Our analysis is based on the Fannie Mae Public dataset (LPPUB), and we conduct analysis based on multiple estimation and forecast periods.[6]

Our findings can be summarized as follows. First, the two heterogenous ensemble methods can out-perform other methods in terms of rank ordering mortgage early delinquency risk in the training, out-of-sample, and out-of-time datasets. AdaBoost, Support Vector Machine, and K-Nearest Neighbors often underperform even the linear Logit Regression (GLM1) in risk classification.

Second, ranking of various methods along different predictive accuracy measures shows a high level of variation, and no method clearly stands out in terms of predictive accuracy. Although Stack leads other methods in Brier score in the training, out-of-sample, and out-of-time data, the literature has shown weaknesses in the Brier score (Jewson (2018)), and thus results from Brier score alone cannot be conclusive. We also examine the confusion matrix and investigate how well the models can correctly capture the true

[4] Out-of-time testing has been undertaken in some earlier studies (for example, Butaru et al. (2016)). However, we focus on a different type of lending from Butaru et al. (2016).

[5] Extreme Gradient boost is one type of gradient tree boosting machine methods (GBM) originally proposed by Friedman (2001) and Friedman et al. (2001).

[6] In addition, because of the very low early delinquency rates, we also conduct robustness tests by varying the proportions of the "good" population sampled. Such results are mostly discussed but not reported because of space limitations.

positives and true negatives when we assign costs to false positives and false negatives. We find that the rankings of different methods vary depending on the costs we assign to false positives or false negatives. No method consistently leads in the ability to 1) correctly identify both true positives and true negatives or 2) capture true positives or true negatives reliably with different cost function assumptions.

When we compare actual early delinquency rates with the predicted ones in different vintages, we do not find any method dominates. Across the refined FICO groups, we find that ML models typically are not able to achieve satisfactory accuracy even in the training data, and the predictive accuracy at the FICO level is by no means better in the training data than in the out-of-time data. These results suggest that predictive accuracy for mortgage early delinquency models is a challenge for all statistical models even in the training data, especially at the refined subgroup level. This finding is likely due to the inability to incorporate important risk drivers (such as borrower income,[7] job stability and expenditure patterns) among the list of explanatory variables.

The above results show the potential benefit of disentangling mortgage risk classification from predictive accuracy in production, possibly relying on heterogenous ensemble methods for risk classification. Meanwhile, given the challenge to incorporate critical risk drivers, it is very difficult even for highly sophisticated models to capture early delinquency rates well even in the training sample.

The studies that are closely related to this paper are Fitzpatrick and Mues (2016), Li, Bellotti, and Adams (2017) and Sirignano et al. (2018) and they all apply ML methods for mortgage credit models. Fitzpatrick and Mues (2016) examine one-year default rates, while Li et al. (2017) investigate early delinquency models, and both papers focus on risk classification or rank ordering without considering predictive accuracy. Sirignano et al. (2018) explore mortgage credit risk management models for a seasoned mortgage portfolio, and such models can be used for reserve forecast, capital calculation, stress testing, etc. Among the risk drivers in the models in Sirignano et al. (2018) there is information on past loan performance, and such information can significantly enhance the prediction of loan status in the following months. However, past loan performance information is not available at certain stages of loan risk management, for example at loan origination. Past loan performance information is not available for mortgage early delinquency models, and as a result, the questions we examine in this paper differ significantly from those examined in Sirignano et al. (2018).

The rest of the paper is organized as follows. Section 2 describes the data and Section 3 discusses research design. Section 4 presents empirical results and we draw a brief conclusion in Section 5.

2. Data description

The Fannie Mae Public dataset (LPPUB) provides both origination and monthly performance information at the loan level. Variables at origination include consumer credit bureau score, combined loan-to-value ratio (CLTV), debt-to-income ratios (DTI), the number of borrowers, loan term, loan purpose, occupancy status, property type, the number of units, first time home buyer indicator. Spread at origination (SATO) is defined as the difference between mortgage interest rate and market mortgage rate for the same loan term; we construct this variable using mortgage interest rate provided in the data and market mortgage rate obtained from Freddie Mac website.[8]

We constructed a cross-sectional dataset, with one observation for each loan.[9] Our dependent variable is equal to 1 (i.e., "bad") if the loan becomes 60 days delinquent (or days past due, DPD) in the first 12 months since origination,[10] and equal to 0 otherwise (i.e., "good"). We include both 30-year and 15-year fixed-rate mortgages in this study. LPPUB has data from 2000; we do not use loans originated before 2009 in the main results reported in the paper, as one can see from Appendix A that mortgages originated before 2009 are very different from those originated post 2009. The origination FICO scores are much higher, and the early delinquency rates are much lower among mortgages originated post 2009. Therefore, mortgages originated post 2009 might be more homogeneous and the models built on these loans might be less contaminated by model instability. In the tables of the paper, we report results from the 2009 to 2015 period as in-time and 2016 as out-of-time.

The percentage of loans being delinquent in the first 12 months since loan origination is very low, and the impact of such imbalanced datasets on model estimation has been a research topic in the extent literature (for example, Fitzpatrick and Mues (2016)). To test the robustness of our results, we implement different sampling schemes, always sampling 100% of the "bad" and 0.5%, 1%, and 2.5% of the "good", and we investigate how our conclusions might be affected by different sampling schemes. The results reported in the tables of the paper are based on the 1% "good" sample. Because of space limitations, we only discuss results from the alternative "good" sample in robustness tests but choose not to report such results in the tables. These results are available upon request.

We have also conducted all the analyses in the tables using four-year rolling windows to estimate the models and then test the models out-of-time in the subsequent one year; we conduct this exercise using data from the entire period of 2009–2016 and using different sampling schemes. We largely describe results from such an exercise and summarize the out-of-time results via box-plots.

Macroeconomic variables considered in this paper include four-quarter growth in housing price indices (HPIs), four-quarter unemployment rate changes, and the credit spread between the interest rates on corporate BAA bond and 10-year T-bond.[11] HPIs are at 3-digit ZIP level obtained from Federal Housing Finance Agency (FHFA), and unemployment rates are at MSA level obtained from Bureau of Labor Statistics (BLS). Bond yields are obtained from the St Louis Federal Reserve. Each macro is merged with the loan level data by the date of the first scheduled mortgage loan payment to be made by the borrower under the terms of the mortgage loan.

Table 1 presents the summary statistics based on the full data from 2009 to 2016. We can see that most of the loans in our study are used by borrowers with reasonably high credit bureau scores, with only 14 percent of them first-time home buyers. The origination LTVs are overwhelmingly at or below 80 percent and the

[7] Borrower income is observable for mortgage applications. However, lenders typically do not use income information directly in mortgage scoring models because of potential fair lending concerns in the U.S.

[8] Technically, SATO is not available at the underwriting stage. We include this variable among the list of explanatory variables because only a limited number of explanatory variables are available in the dataset, and some important variables used in underwriting could be missing, which in turn can explain the lack of predictive accuracy that we observe in the paper. We use SATO to address such missing information problem.

[9] We investigate cross-sectional models in this paper because they are more commonly used as business tools in practice than survival models.

[10] We count the loan as "bad" even if the loan is cured after falling 60 dpd.

[11] We have investigated all macro variables specified in the Federal Reserve stress tests as well as additional variables such as existing home sales, new home sales, and new construction permit. We narrow down to the three macro variables used in the paper as these variables show significant and intuitive coefficients. Signs of the other variables are either counter-intuitive or non-significant in the logit regression, especially when used together with other macro variables.

Table 1
Summary statistics based on the full data.

Panel A Continuous variables Variable	Mean	25th Pctl	50th Pctl	75th Pctl	Std Dev
Proportion of loans being 60-day delinquent in the first 12 months of loan origination	0.25%				4.95%
Credit bureau score (FICO)	758	730	769	792	44
Origination combined LTV (%) (OCLTV)	71.21	61	75	80	17.27
Debt-to-income rate (DTI)	32.24	25	33	40	9.92
Difference between mortgage interest rate and market mortgage rate for the same loan term (SATO)	0.21	−0.04	0.21	0.47	0.44
Four-quarter HPI change (3-digit zip level)	0.69%	−2.51%	0.81%	4.92%	6.96%
Four-quarter unemployment rate change (MSA level)	−0.06	−1.10	−0.60	0.10	1.70
BAA Spread	3.08	2.66	2.94	3.24	0.77
Panel B Proportions of loan types					
15-year fixed rate loans	25.96%		Occupancy Status (%)		
First-time home buyers	13.53%		Investor		8.05
Single borrower	44.58%		Second		3.63
Single unit property	98.11%		Principal owner		88.32
Property Type (%)			Loan purpose (%)		
Single-Family	63.58		Cash-out refinance		24.33
Planned unit development (PUD)	27.18		No Cash-out refinance		37.45
Others (condo, co-op, manufactured housing)	9.24		Purchase		38.22

spread at origination is on average 21 basis points. About one-quarter of our sample are 15-year mortgages, with the rest 30-year mortgages. Most loans result from refinancing, with roughly one quarter of them cashing out equity in the residential properties. The overwhelming majority of the mortgages are used for principal residence, and roughly two-thirds of the sample have single-family homes as the collateral. Our sample shows that a wild range of economic conditions across various regions, with the mean unemployment rate growth at −0.06% but median at −0.60%. The HPI changes also shows much variation in the sample, with the interquartile ranging from negative growth to positive growth. The BAA spread, on the other hand, demonstrates a narrow range during this sample period. Finally, unreported results show that the distributions of the explanatory variables of the various "good" samples are quite similar among themselves and they are close to the summary statistics reported in Table 1.

3. Research design

3.1. Model description

Even though nonparametric methods are more flexible than parametric methods in terms of explanatory variable functional forms, they are still built on a pre-specified set of variables and the transformation of these variables are determined by the researchers. The advantage of ML methods over traditional nonparametric methods is their ability to discover complex structures and interactions that were not specified in advance.

We present results from the three GLMs in this paper. In GLM1, we include only linear terms. GLM2 and GLM3 aim to address the non-linear relationships between the dependent variable and risk drivers, by specifying piece-wise linear splines as well as smoothing splines for important explanatory variables. We rely on tree-based importance measures (available in Scikit-learn RandomForestClassifier() and XGBClassifier() functions) to evaluate the importance of the explanatory variables in relation with the dependent variable. In unreported results, we find that FICO is overwhelmingly the dominant factor determining the event of mortgage early delinquencies. SATO and the macro factors (unemployment or HPI changes) come next, followed by DTI and OCLTV.[12] There also exist some non-linear relationships between the dependent variable and these important risk drivers. To account for such non-linearity identified from the univariate relationship between the dependent variable and the risk drivers, we include in GLM2 knots of FICO (knots defined around 680, 720, and 760), OCLTV (defined around 70, 80, and 90), DTI (defined around 35, 40, and 45), four-quarter HPI growth (defined around −0.05 and 0.05), four-quarter unemployment rate change (defined around −0.5) and SATO (defined around 0.25), and BAA spread (defined around 3). To allow more flexibility to capture the nonlinearity, GLM3 adopts penalized B-splines (i.e., p-spline, see Eilers and Marx (1996)) with degree of two for the four loan-level variables (DTI, FICO, CLTV, SATO) and the three macroeconomic variables. GAM is another method of smoothing splines that adds more flexibility than GLM2 in accounting for potential non-linearities. In contrast top-splines in GLM3, smoothing functions in GAM are estimated from a local polynomial regression in an iterative procedure. The smoothing functions are approximated using local polynomial functions in a neighborhood around each value of the explanatory variables. We use the glm() function in R to estimate GLM1, GLM2 (bs() function for linear b-spline) and GLM3 (pspline() function for p-spline), and the gam() function to estimate GAM.

We use the Python package Scikit-learn (Pedregosa et al., 2011) to train most of the ML methods under investigation in this paper. For regularized logistic regression (RLGR), we use LogisticRegression() function with ridge regularization, with the key tuning parameter being the regularization strength. We use the MLPClassifier() function for neural network, and the hyper-parameters are number of layers, node size, and learning rate. We use LinearSVC() function to train the linear SVM model, and the key tuning parameter is the penalty term. For K-Nearest Neighbors, we use KNeighborsClassifier() function which requires to find optimal number of neighbors and distance metric. For Adaboosting, we use AdaBoostClassifier() function and tune the number of trees as well as learning rate. We use the RandomForestClassifier() function for random forest, and the key hyper-parameters are the number of trees and max depth. For extreme gradient boosting, we use the XGBClassifier() function. There are many hyper-parameters that can be tuned

[12] The definition of DTI is debt/combined income for mortgages with multiple borrowers.

in this classifier, and the key tuning parameters include the number of trees, max depth, learning rate, min split loss (gamma), min child weight and regularization (alpha). We rely on AUC as the evaluation metric in order to find the best set of hyper-parameters through grid search.

3.2. Hyper-parameter search

We split each of the in-time sample into 50% for training (we call this sample training sample in the rest of the paper), 30% for in-time out-of-sample testing (we call this sample out-of-sample in the rest of the paper), and 20% as validation set. The validation set is used for probability calibration (i.e., Platt Scaling (Platt (1999))) as well as for training heterogenous ensemble models.[13] In the 50% training data, we use 3-fold cross validation for parameter tuning in ML methods.[14]

We apply the grid search method in an iterative process. Take Random Forest (RF) as an example. As a first step, we start with a broad range of parameter values (i.e., the number of trees in the range of (300, 500, 700, 900, 1100, 1300, 1500, 1700, 2000), and the maximum depth in the range of (3, 5, 7, 9, 11, 13, 15)). Then we repeat the process by narrowing down the value ranges, until no significant gain in AUC in cross validation can be obtained. Similar process is done to choose optimal parameters for the other ML methods. The hyper-parameters behind the results presented in the tables are reported in Appendix B.

The two heterogeneous ensemble methods are constructed as follows. We use a stacking method which combines the probabilities from all methods via a logistic regression. We implement a forward selection algorithm for hill-climbing (HC) and select the combination of methods that has the largest AUC in the validation dataset.

3.3. Performance metrics

3.3.1. Risk classification

We apply two performance metrics for risk classification or rank ordering: AUC and the H-measure. AUC is the area under the receiver operating characteristic (ROC) curve. The maximum AUC value is 1, and a higher AUC value indicates better rank ordering ability. However, there are weaknesses in the AUC measure. In particular, our sample is very imbalanced, and the AUC is thus naturally quite high. To gain more insight on the models' rank ordering ability, we use the H-measure as the second measure of risk classification. The H-measure is proposed by Hand (2009b), and detailed discussions on this measure can be seen in Adams, Anagnostopoulos, and Hand (2012). We follow Adams et al. (2012) and use a Beta distribution with its mode equal to the proportion of 60 DPD in the dataset. A higher H-measure suggests better rank ordering ability.

3.3.2. Predictive accuracy

We gauge predictive accuracy via several measures. The first measure is the Brier score, which compares the predicted and actual outcome at the individual loan level. However, Brier score can yield counter-intuitive results, as it is calculated on a straight difference between the predicted and actual probabilities, thus encouraging the prediction of very low or zero probabilities for a low default portfolio (see for example, Jewson (2018)).

The next a few measures are derived from the confusion matrix, accounting for non-negative costs of false positives (C_FP) and false negatives (C_FN), and we assume true positives and true negatives to have zero costs. The threshold to classify positives or negatives is defined as $threshold = \frac{C\_FP \times Neg}{C\_FN \times Pos + C\_FP \times Neg}$, where *Neg* and *Pos* are the actual number of negatives and positives from the training sample. In the same sample, as the cost of false negative (C_FN) becomes higher relative to the cost of false positives (C_FP), the threshold decreases. We then define two measures as follows,

$$Precision = \frac{\#\ true\ positives}{\#\ true\ positives + \#\ false\ positives}$$

$$Recall = \frac{\#\ true\ positives}{\#\ true\ positives + \#\ false\ negatives}$$

A low number of false positives (negatives) is associated with a higher precision (recall). In other words, the measure 'precision' gauges how well the model can capture the true positives, and the measure 'recall' assesses how well the model captures the true negatives. The third measure from the confusion matrix is total false classification, which is defined as TFC = FP +FN*c, where c = C_FN/C_FP. This measure examines how many wrong classifications the method generates, given a particular set of C_FP and C_FN values. The higher this measure is, the more wrong classifications a method yields.

The last set of predictive accuracy measures compare the actual early delinquency rate and the model-predicted early delinquency rate at portfolio level. We conduct this analysis for various vintages. We further examine predictive accuracy by various FICO groups and then calculate the mean absolute error (MAE) and root mean squared error (RMSE) across the FICO groups (weighted by loan accounts of different FICO groups to account for the changes in FICO composition through time).

4. Empirical results

4.1. Layout of the tables and figures

We report in the tables results from the training and out-of-sample datasets with 100% of the "bad" and 1% of the "good" from the 2009 to 2015 period,[15] and out-of-time results from 2016 (consisting of 100% of the "bad" and 1% of the "good"). Appendix C summarizes the loan counts of "good" and "bad" by year in the training, out-of-sample and out-of-time samples. We have also conducted all the analyses on two alternative sampling schemes: sampling 100% of the "bad" and 0.5% and 2.5% of the "good." Because of space limitations, we do not report, and instead discuss, the results from the alternative sampling schemes in the tables; those results are qualitatively similar to those reported in the tables based on the 1% of the "good."

Besides the different sampling schemes, we have also tried estimating the models using four-year rolling windows from the 2009 to 2015 period and then tested the models out-of-time using data from the subsequent one year. We employ three different sampling schemes in this exercise as well, always sampling 100% of the "bad" and 1%, 0.5%, and 2.5% of the "good." So, in total, we have 9 sets of results, and we re-tune the hyper-parameters for each sample. For each set of results, we compare the rank ordering and predictive accuracy for the thirteen models and we rank these methods based on the various performance measures. Because of space

[13] The scores of some methods under investigation (for example, Adaboosting and SVM) do not have the appropriate scales in probability. Therefore, an additional probability calibration is required. To have a fair comparison, we apply the Platt Scaling to all ML methods, except for Regularized logistic regression and GAM which inherently fit a logistic distribution in probability.

[14] We have also tried Nx2-fold cross-validation (Dietterich (1998)) when choosing the hyper-parameters. This approach has four steps: (1) randomly splitting a data set in half, (2) using the first and second half for model building and evaluation, respectively, (3) swapping the two datasets, and (4) repeating steps (1) – (3) N times. We find that hyper-parameters from repeated cross-validation tend to overfit in our data.

[15] Loans originated between 2009 and 2015 were pooled in model estimation.

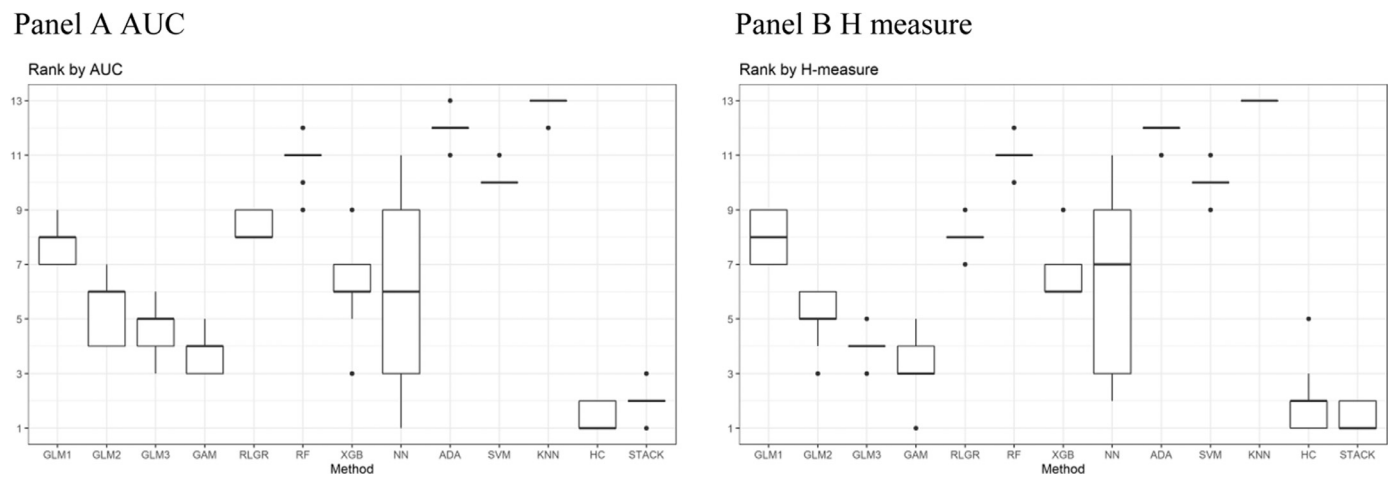

Fig. 1. Out-of-time Rankings of AUC and H measure.

Table 2
Rank Ordering Results in Training, Out-of-sample, and Out-of-time Datasets.

Panel A AUC	GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
Training (2009–2015)	0.828	0.829	0.830	0.830	0.827	0.830	0.830	0.832	0.822	0.826	0.820	0.833	0.832
Out-of-sample (2009–2015)	0.829	0.831	0.831	0.831	0.829	0.826	0.830	0.832	0.823	0.828	0.822	0.832	0.832
Out-of-time (2016)	0.805	0.807	0.807	0.807	0.805	0.803	0.807	0.807	0.801	0.804	0.799	0.808	0.808
Panel B H measure	**GLM1**	**GLM2**	**GLM3**	**GAM**	**RLGR**	**RF**	**XGB**	**NN**	**ADA**	**SVM**	**KNN**	**HC**	**STACK**
Training (2009–2015)	0.364	0.368	0.370	0.370	0.364	0.368	0.371	0.374	0.351	0.362	0.347	0.376	0.374
Out-of-sample (2009–2015)	0.369	0.372	0.373	0.373	0.369	0.362	0.370	0.376	0.354	0.366	0.354	0.375	0.375
Out-of-time (2016)	0.317	0.323	0.323	0.323	0.317	0.313	0.321	0.320	0.309	0.315	0.303	0.322	0.323

limitations, we only summarize the distributions of their respective rankings for the nine sets of out-of-time results in the boxplot figures in the paper.

In the boxplots, the middle line is median, and upper and below edges are 25 and 75 quantiles. The end of the whiskers represents the values that are located within 1.5 times the inter-quartiles (IQR) beyond the upper and lower quartiles and the dots depict the values that are more extreme. In other words, the upper whisker typically represents the maximum value. However, if the maximum value is greater than 1.5 times the IQR, the upper whisker represents 1.5 times the IQR, and the maximum value is instead presented by a dot (for example, Stack in Panel A of Fig. 1). Similarly, the lower whisker usually represents the minimum value. However, if the minimum value is lower than 1.5 times the IQR, the lower whisker represents 1.5 times the IQR, and the minimum value is represented by a dot. Sometimes, there is no whisker for a box plot, for example, for HC in Panel A of Fig. 1. So, the minimum value for HC is 1 (i.e., the best ranking possible) in this graph and it is equal to the 50th percentile. The same graph shows that the lowest ranking of HC is 2, which is also its 75 percentage.

We report risk classification results in Section 4.2 and predictive accuracy results in Section 4.3. In Section 4.4, we discuss results from additional investigations we have conducted.

4.2. Risk classification

Table 2 reports the risk classification or rank ordering results from the training and out-of-sample data from the period 2009 to 2015 and out-of-time data from 2016. We can see that, for both the AUC and H-measure in the training sample, the two heterogeneous ensemble methods lead other methods, followed by NN, XGB, GAM, GLM3, and GLM2. ADA and KNN under-perform even GLM1.

For out-of-sample, the two heterogeneous ensemble methods and NN lead other methods based on the AUC, while NN outperforms other methods gauged by the H-measure. However, the performance difference between HC, Stack, and NN in the out-of-sample data is rather minor. ADA and KNN still under-perform GLM1 in the out-of-sample data.

From the result using out-of-time data from 2016, we can see that both the AUCs and H-measures decline noticeably from the training and out-of-sample data, but the two heterogeneous methods still lead other methods based on the AUC performance measure, while GLM2, GLM3, GAM and Stack lead based on the H-measure. However, the differences in either AUC or the H-measure among GLM2, GLM3, GAM, XGB, NN, and the two heterogeneous methods are not substantial, and all seven methods lead the remaining methods in the out-of-time data. ADA and KNN again clearly trail the simple GLM1 based on both measures.

For both alternative sampling schemes with 0.5% and 2.5% of the "goods," the two heterogeneous ensemble methods tend to lead in risk classification in the training, out-of-sample, and out-of-time data, followed by NN, XGB, GAM, and GLM3, while ADA and KNN under-perform even GLM1. Such results are not reported to save space and are available upon request.

Fig. 1 reports the out-of-time AUC and H-measure results on the 9 sets of rolling windows and different sampling schemes via boxplots.

We can see from Fig. 1 Panel A that HC and Stack lead other methods based on the AUC measure. Among the two, HC has a clear advantage with its median at 1 and its lowest ranking being 2. By contrast, the median ranking of Stack is 2, and in one case,

Table 3
Brier Score in Training, Out-of-sample, and Out-of-time Samples.

	GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
Training (2009–2015)	0.113	0.112	0.112	0.112	0.113	0.112	0.112	0.112	0.114	0.113	0.115	0.111	0.111
Out-of-sample (2009–2015)	0.112	0.111	0.111	0.111	0.112	0.114	0.113	0.112	0.114	0.112	0.114	0.111	0.111
Out-of-time (2016)	0.139	0.139	0.139	0.138	0.139	0.141	0.140	0.140	0.141	0.140	0.143	0.139	0.138

Stack is ranked the third. The ranking of NN has a much larger range, and RF, ADA, SVM, and KNN largely trail other methods. The performance of GLM2, GLM3, and GAM is quite stable, and they do not underperform the commonly used ML methods such as RF, XGB and NN. Based on AUC, RLGR slightly underperforms GLM1. ADA and KNN are clearly the laggers in Panel A.

Fig. 1 Panel B depicts risk classification results based on the H-measure. In this figure, HC and Stack still lead other methods, and Stack seems to have an edge over HC in this graph. NN still has a wide range, and RF, ADA, SVM, and KNN still lag. The performance of GLM2, GLM3, and GAM again does not show much variation, and they performance very decently even relative to XGB and NN. Based on the H-measure, RLGR slightly outperforms GLM1, showing less variation in risk classification ranking, while ADA and KNN again clearly underperform other methods.

Combining the results from Table 2 and Fig. 1, as well as the unreported results from alternative sampling methods, we conclude that the two heterogeneous ensemble methods can produce superior risk classification in the training, out-of-sample, and the out-of-time datasets. By contrast, some ML methods, especially RF, ADA, SVM, and KNN, do not outperform better than some simpler methods, such as GLM2, GLM3, and GAM in risk classification.

4.3. Predictive accuracy

Tables 3–6 reports results on different measures of predictive accuracy in various samples.

4.3.1. Brier score

In Table 3, we present results on the Brier score of the training sample and the out-of-sample datasets from 2009 to 2015 and for the out-of-time dataset from 2016. We can see that HC and Stack perform rather similarly and lead in the training sample, and they tie with GLM2, GLM3, and GAM in the out-of-sample data in Table 3. Stack and GAM outperform other methods in out-of-time dataset in 2016. By contrast, KNN demonstrates the worst performance in Table 3, and its sub-par performance shows up in all three samples in Table 3. ADA shows the second worst performance in Table 3.

Results from the alternative sampling methods and the rolling windows are quite comparable to those reported in Table 3. To illustrate, we depict in Fig. 2 the boxplots of the out-of-time Brier score from the nine sets of results from the rolling window exercise and the three alternative sampling methods. We can see that Stack clearly leads other methods in this figure, while KNN is clearly the worst.

However, the differences of the magnitude of the Brier scores in Table 3 are small, and it is not clear whether a Brier score of 0.143 points towards substantial under-performance relative to a Brier score of 0.138. In addition, because of the issues with the Brier score as discussed earlier, the superior performance of Stack in Table 3 is far from conclusive concerning models' predictive accuracy. We next investigate other predictive accuracy metrics.

4.3.2. Confusion matrix results

We report in Table 4 measures based on the confusion matrices. In our codes, we set C_FP=1, and C_FN=c, and we report in the table results from three values of c: 1, 5, 10, corresponding to the three sets of values of C_FN and C_FP: 1) C_FN$_=$ C_FP, 2) C_FN$_=$ 5*C_FP, and 3) C_FN$_=$ 10*C_FP.[16] We assume c to be greater than 1 because a misclassified 'bad' would cost a bank far more in lost funds than a rejected applicant that turned out to be 'good'.

We can make the following observations from Table 4. First, the precision measure falls with c and the recall measure rises with c. This result is intuitive, because as c increases (i.e., the cost of false negatives becomes higher relative to the cost of false positives), there are lower costs for false positives, and as a result, more observations are classified as positives. Second, the three measures (precision, recall, and TFC) do not provide consistent rankings with the same c. For c=1 in the training sample, RLGR ranks the first for precision, but 9th for recall, and 10th for TFC. By contrast, NN ranks 11th in precision, but ranks first under both recall and TFC for c=1 in the training sample. Third, the ranking of the methods changes when we alter the value of c. For example, XGB and NN lead in recall in the training sample for c=1, but they trail all other methods in recall, except for RF, when c = 10. Therefore, Table 4 suggests that no method can clearly lead in different predictive accuracy measures from the confusion matrix and when different cost-sensitive functions are assumed.

Results from the alternative sampling methods and the rolling windows are quite similar to those reported in Table 4. To illustrate, we depict in Fig. 3 the boxplots of the out-of-time ranking of precision, recall, and TC from the nine sets of such results. We can see again that no method can clearly lead in these graphs. Furthermore, the ranking of different methods varies in all panels of Fig. 3 when we change c (i.e. when different cost-sensitive functions are assumed), and there is no clear pattern which method can capture the true positives or true negatives better. For example, in Panel B (recall), XGB leads for c=1; XGB is in the middle of the pack for c=5; and it trails most methods for c=10. In addition, no method can consistently capture either true positives or true negatives better than other methods. Even though the heterogeneous ensemble methods tend to outperform other methods in rank ordering, they are largely in the middle of the pack in Table 4 and Fig 3.

4.3.3. Predictive accuracy at different portfolio levels

We first examine the predictive accuracy for different vintages, and we report in Table 5 the actual early delinquency rates and the difference between the model-predicted and actual early delinquency rates across all loans originated in that year in the training, out-of-sample, and out-of-time 2016 datasets. In the last two columns of the table, we test the difference between the predicted and actual delinquency rates and we report the 95% confidence intervals around the mean of zero to assess whether the predicted numbers are significantly different from the actual numbers. We underline the predictions that are outside the 95% confidence intervals.

We can see from this table that the predictive errors for the entire training sample are close to zero for all methods.

[16] We have tried many different sets of C_FP and C_FN values. The results reported in Table 4 are representative of the different cost-sensitive functions we have explored. Additional results from the different cost-sensitive functions are available upon request.

Table 4
Confusion Matrix with cost-sensitive cutoffs.

Training (2009–2015)		GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
$c = 1$	Precision	0.87	0.86	0.86	0.86	0.88	0.87	0.82	0.83	0.88	0.87	0.82	0.87	0.87
	Recall	0.07	0.08	0.09	0.08	0.07	0.11	0.14	0.14	0.04	0.07	0.09	0.11	0.11
	TC	12,399	12,272	12,227	12,251	12,428	11,953	11,730	11,691	12,685	12,336	12,200	11,981	11,951
$c = 5$	Precision	0.65	0.66	0.66	0.66	0.66	0.65	0.66	0.66	0.63	0.65	0.64	0.66	0.66
	Recall	0.42	0.43	0.43	0.43	0.42	0.45	0.44	0.44	0.43	0.43	0.41	0.44	0.45
	TC	40,917	40,547	40,653	40,482	40,915	39,472	40,089	39,843	40,581	40,655	41,848	39,859	39,562
$c = 10$	Precision	0.53	0.54	0.54	0.54	0.53	0.58	0.58	0.58	0.52	0.53	0.54	0.57	0.55
	Recall	0.62	0.61	0.61	0.61	0.62	0.55	0.56	0.57	0.60	0.62	0.58	0.58	0.60
	TC	57,259	57,818	57,664	57,647	57,219	64,880	63,179	62,571	59,345	57,460	62,019	61,147	58,895
Out-of-sample (2009–2015)		GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
$c = 1$	Precision	0.90	0.88	0.89	0.89	0.90	0.86	0.83	0.83	0.87	0.89	0.86	0.87	0.87
	Recall	0.07	0.08	0.09	0.09	0.07	0.10	0.14	0.14	0.04	0.08	0.10	0.11	0.11
	TC	7564	7481	7445	7470	7578	7380	7161	7148	7775	7517	7382	7338	7339
$c = 5$	Precision	0.66	0.66	0.66	0.66	0.66	0.64	0.65	0.66	0.63	0.65	0.65	0.66	0.66
	Recall	0.42	0.43	0.43	0.42	0.42	0.44	0.43	0.44	0.43	0.43	0.41	0.44	0.44
	TC	25,061	24,947	24,880	24,963	25,100	24,511	24,845	24,420	25,099	24,949	25,766	24,566	24,283
$c = 10$	Precision	0.53	0.54	0.54	0.54	0.53	0.58	0.58	0.58	0.53	0.53	0.55	0.57	0.56
	Recall	0.62	0.61	0.61	0.61	0.62	0.54	0.56	0.56	0.60	0.62	0.58	0.57	0.60
	TC	35,261	35,913	35,654	35,698	35,312	40,233	38,983	38,484	36,773	35,435	37,856	37,921	36,441
Out-of-time (2016)		GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
$c = 1$	Precision	0.84	0.82	0.85	0.84	0.83	0.83	0.78	0.82	0.80	0.82	0.72	0.84	0.82
	Recall	0.03	0.04	0.05	0.05	0.03	0.01	0.11	0.09	0.02	0.03	0.07	0.04	0.06
	TC	3921	3872	3862	3858	3925	3975	3706	3730	3945	3909	3837	3869	3836
$c = 5$	Precision	0.65	0.66	0.66	0.66	0.66	0.65	0.66	0.66	0.63	0.65	0.64	0.66	0.66
	Recall	0.42	0.43	0.43	0.43	0.42	0.45	0.44	0.44	0.43	0.43	0.41	0.44	0.45
	TC	40,917	40,547	40,653	40,482	40,915	39,472	40,089	39,843	40,581	40,655	41,848	39,859	39,562
$c = 10$	Precision	0.53	0.54	0.54	0.54	0.53	0.58	0.58	0.58	0.52	0.53	0.54	0.57	0.55
	Recall	0.62	0.61	0.61	0.61	0.62	0.55	0.56	0.57	0.60	0.62	0.58	0.58	0.60
	TC	57,259	57,818	57,664	57,647	57,219	64,880	63,179	62,571	59,345	57,460	62,019	61,147	58,895

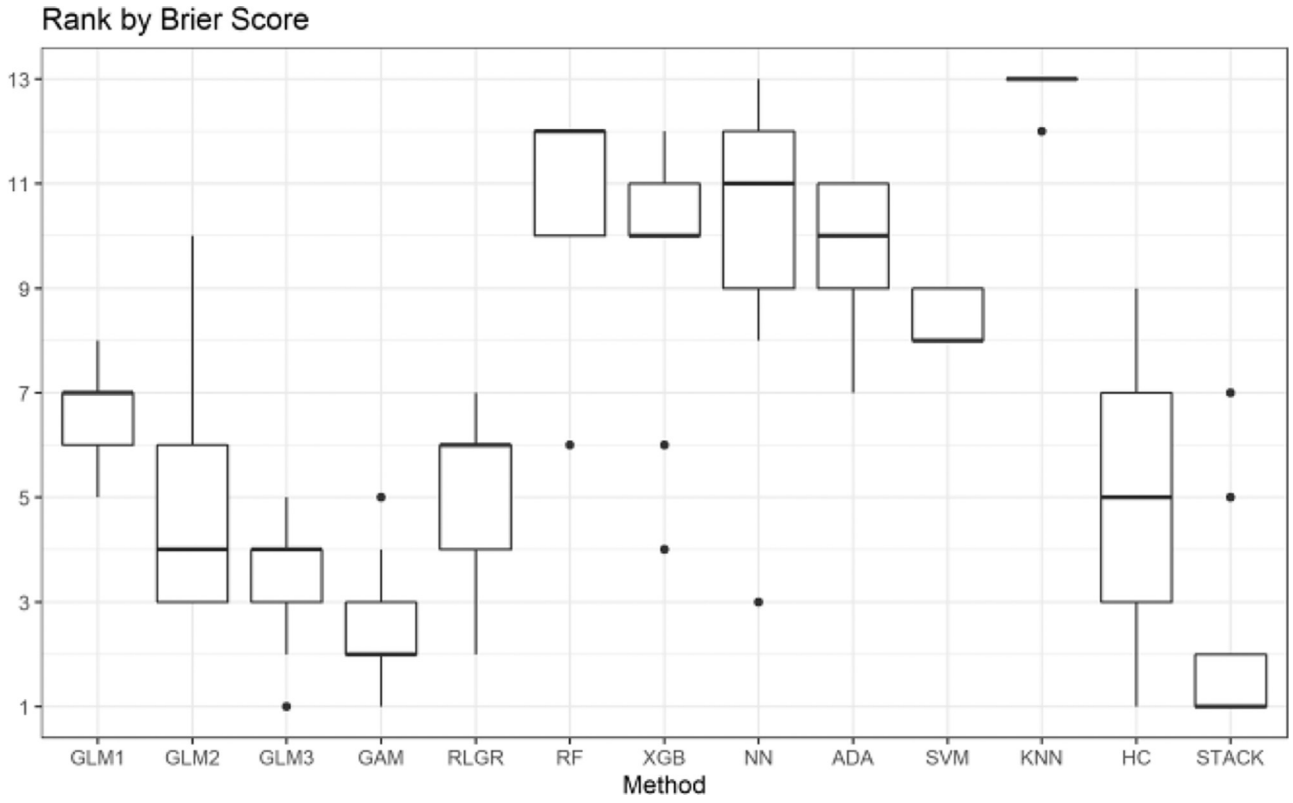

Fig. 2. Out-of-time Ranking of Brier Score.

Table 5
Predictive accuracy in training, out-of-sample and out-of-time by vintage.

Panel A Training (2009–2015)															
Orig. year	Actual avg delq rate (%)	Model-predicted delinquency rate - actual delinquency rate (%)													95% C.I.
		GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK	Lower bound	Upper bound
2009	30.27	−0.68	−0.50	−0.23	−0.32	−0.68	−0.77	−0.67	−0.42	−4.04	−0.34	−1.35	−0.48	−0.33	−0.81	0.81
2010	16.28	0.17	0.00	0.14	0.14	0.06	−0.12	0.04	0.11	0.21	0.21	−0.98	0.04	−0.07	−0.81	0.81
2011	14.03	0.34	0.43	0.01	0.11	0.16	0.49	0.40	0.59	0.73	0.40	−0.08	0.37	0.02	−0.78	0.78
2012	13.16	0.92	0.81	0.85	0.81	0.82	0.36	0.45	1.05	0.65	1.17	0.24	0.75	0.65	−0.62	0.62
2013	13.34	0.72	0.59	0.30	0.34	0.83	1.10	0.74	0.90	1.92	0.42	0.94	0.77	0.52	−0.63	0.63
2014	22.74	−1.20	−0.61	−0.28	−0.41	−1.09	−0.05	0.01	−1.44	1.25	−1.53	1.35	−0.59	−0.10	−0.94	0.94
2015	22.16	−0.48	−0.82	−0.90	−0.77	−0.36	−0.90	−0.85	−1.10	0.39	−0.67	0.13	−0.97	−0.78	−0.80	0.80
All	19.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	−0.29	0.29
Panel B Out-of-sample (2009–2015)															
Orig. year	Actual avg delq rate (%)	Model-predicted delinquency rate - actual delinquency rate (%)													95% C.I.
		GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK	Lower bound	Upper bound
2009	30.27	−0.87	−0.68	−0.42	−0.50	−0.85	−1.12	−1.07	−0.71	−4.24	−0.54	−1.61	−0.76	−0.55	−1.03	1.03
2010	16.30	0.10	−0.04	0.05	0.06	−0.01	−0.04	0.00	0.23	0.16	0.17	−1.01	0.08	−0.06	−1.03	1.03
2011	14.03	0.28	0.36	−0.07	0.04	0.10	0.40	0.35	0.41	0.73	0.34	−0.13	0.25	−0.11	−1.00	1.00
2012	13.16	1.01	0.82	0.86	0.83	0.90	0.38	0.53	1.19	0.67	1.25	0.21	0.81	0.67	−0.79	0.79
2013	13.35	0.35	0.23	−0.07	−0.02	0.45	0.92	0.57	0.63	1.62	0.06	0.56	0.50	0.16	−0.80	0.80
2014	22.77	−1.50	−0.96	−0.63	−0.76	−1.39	−0.40	−0.33	−1.67	0.89	−1.85	1.11	−0.90	−0.41	−1.20	1.20
2015	22.17	−0.47	−0.75	−0.82	−0.71	−0.35	−0.93	−0.82	−1.01	0.30	−0.68	0.10	−0.92	−0.70	−1.03	1.03
All	19.16	−0.13	−0.13	−0.14	−0.14	−0.13	−0.13	−0.13	−0.09	−0.14	−0.13	−0.16	−0.12	−0.13	−0.38	0.38
Panel C Out-of-time (2016)															
Orig. year	Actual avg delq rate (%)	Model-predicted delinquency rate - actual delinquency rate (%)													95% C.I.
		GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK	Lower bound	Upper bound
2016	24.32	−0.25	−1.00	−1.14	−0.93	−0.07	−1.52	−1.14	−1.43	−0.24	−0.17	−0.31	−1.36	−1.29	−0.65	0.65

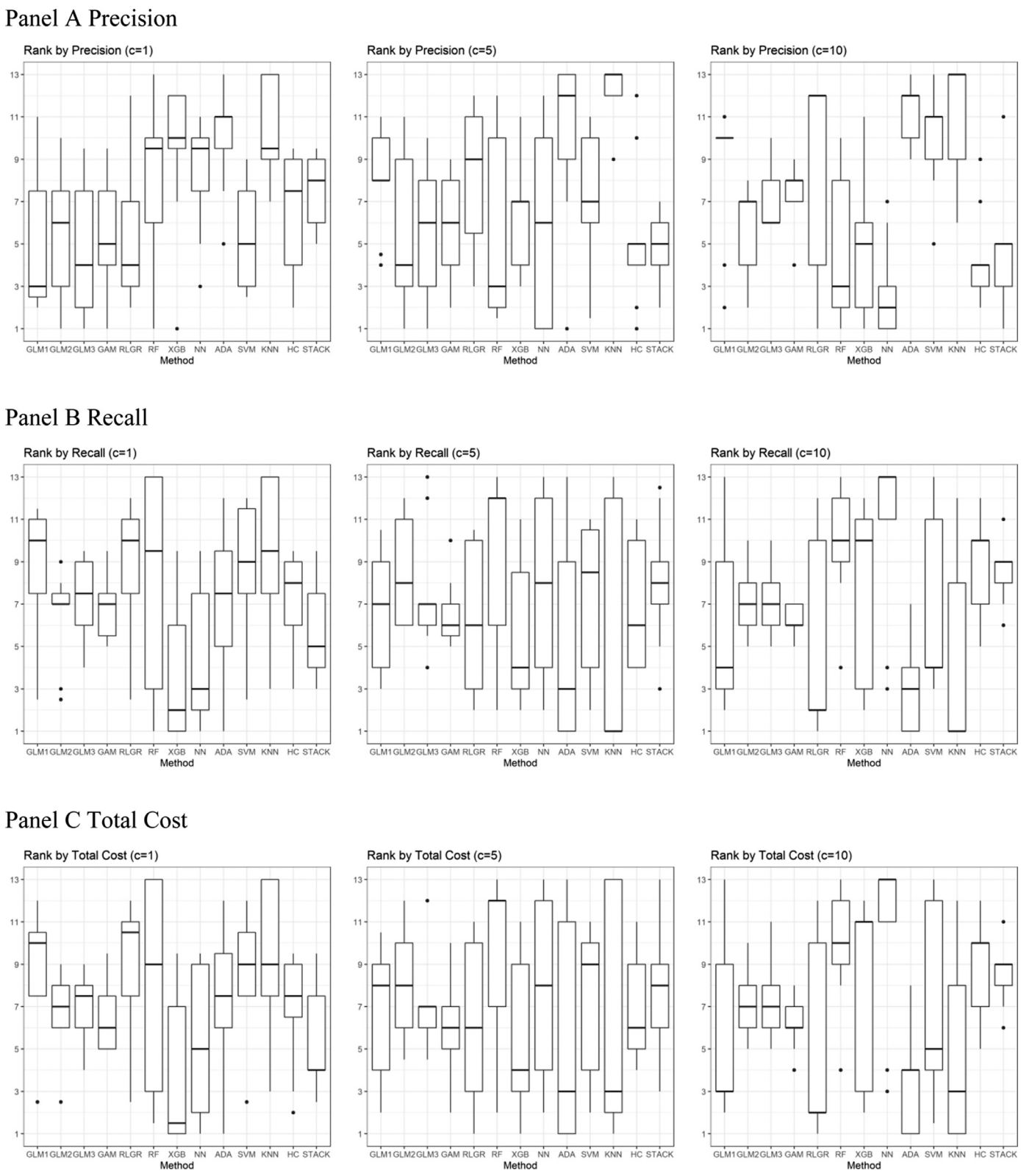

Fig. 3. Out-of-time Rankings of Precision, Recall and Total Cost.

The predictive errors for the out-of-sample data are also very low, and these results are intuitive. Since the models are estimated on the pooled 2009–2015 data, the predicted delinquency rate can be different from the actual delinquency rates for each vintage. Indeed, we find some over-predictions for the vintages 2012 and 2013, and some under-predictions for other vintages for the training and out-of-sample data. The cells outside the confidence interval are scattered across all methods with no concentration among any methods. In addition, the underlined numbers show up in every column of Table 5.

Most methods, including the two heterogeneous ensemble methods, significantly under-predict the early delinquency rates for

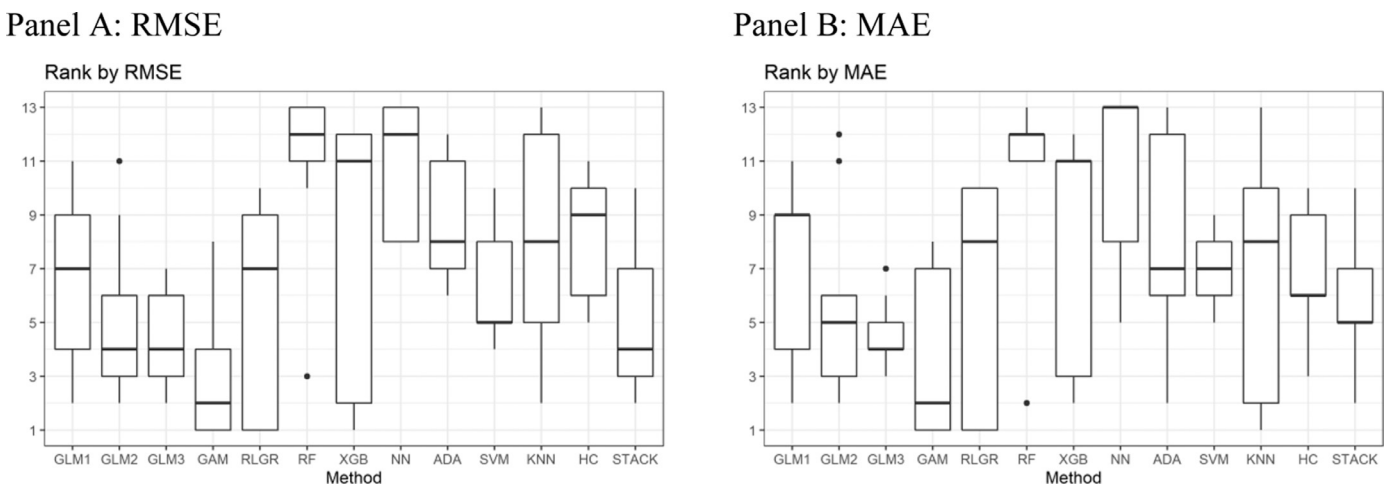

Fig. 4. Out-of-time Ranking of RMSE and MAE by FICO buckets.

the 2016 vintage. By contrast, KNN, ADA, and SVM, the three methods that lag in risk classification, predict early delinquency rates for the 2016 vintage within the 95% confidence interval. Note that KNN is the worst performer gauged by the Brier score. This finding points toward ranking inconsistency from different predictive accuracy measures.

In addition, it seems hard to gauge which method will yield better predictive accuracy out-of-time based on their performance in the training and out-of-sample data. For example, Stack only slightly exceeds the 95% upper confidence interval for one year in the training sample and it does not show any breach in any year in Panel B of table 5. Nevertheless, Stack under-predicts significantly out-of-time for 2016. By contrast, RLGR does not demonstrate better predictive accuracy than other methods in Panels A and B, but RLGR shows the smallest predict error for 2016. Overall, no method clearly stands out when we investigate the predictive errors for vintages.

During our exercise, we find that it is important to include all three macro variables in the model. When we only include two of the three, we observe overwhelming under-prediction in 2009 and severe over-prediction from 2011 to 2013, and the prediction errors are large for all methods. The predictive accuracy for vintages improves significantly when we include all three macro variables. Thus, the insight we derive from this exercise is that including the proper explanatory variables is more important than the statistical models used. However, even after we control for many additional macro factors in the model beyond the three macro factors, results from such exercises do not lead to patterns substantially different from those in Table 5. The reason behind the patterns in Table 5 is that the coefficients might change in magnitude from vintage to vintage. Although models estimated by each vintage or models incorporating vintage dummy variables can achieve vintage predictive accuracy, such models cannot be used in real time prediction.[17] As a result, capturing the actual early delinquency rates by vintages remains a difficult task for all methods.

There is one catch in the model comparison based on the numbers in Table 5. That is, the overall numbers in Table 5 might be contaminated by the offsetting effect among different loans. For example, the errors might be large in sub-samples. If, however, the errors are in different directions, they can cancel out at the vintage level. To address this issue, we next calculate predictive accuracy by FICO sub-groups and then calculate the mean absolute error (MAE) and root mean squared error (RMSE) across the FICO groups, weighted by loan accounts of different FICO groups. We use relatively granular FICO groups in this calculation: below or equal to 660, (660, 680], (680, 700], (700, 720], (720, 740], (740, 760], (760, 780], (780, 800], and above 800. We present such results in two panels in Table 6. For each row in Table 6, we underline the lowest numbers (i.e., the method with the lowest predictive errors).

Examination of the different panels of this table suggests that the underlined numbers are quite scattered across different methods. Although no method clearly out-performs other methods, GAM seems to have a slight edge over other methods in both panels of Table 6. Even though the two heterogeneous ensemble methods lead in rank ordering and Stack leads in Brier scores, their performance in Table 6 by no means stands out.

Results from the alternative sampling methods and the rolling windows are quite similar to those reported in Table 6. To illustrate, we depict in Fig. 4 the boxplots of the out-of-time ranking of RMSE and MAE from the rolling window exercise. We see again that no method can clearly lead in these figures, and GAM again appears to have a minor lead in Fig. 4.

Furthermore, we can compare the RMSEs and MAEs from the training, out-of-sample, and out-of-time data by each method. We can see that the RMSEs and MAEs are not necessarily lower in the training than the out-of-sample data. In addition, the RMSEs and MAEs from the out-of-time sample are by no means larger than those in the training sample. Therefore, the predictive accuracy in the training sample might be unsatisfactory to begin with.

Even though the RMSEs and MAEs reported in Table 6 can be used to compare various models, they do not provide enough intuition as to the absolute predictive errors at the sub-group level. To gain a better understanding of the predictive accuracy of various models at different FICO groups, we investigate the actual early delinquency rates and the differences between the actual and predicted early delinquency rates at the FICO groups and present such results in Appendix D. We can observe from Appendix D many breaches at the 95% confidence level; even for the training sample – roughly 50% of the FICO bucket-method observations have breaches. Neither GLM3 nor GAM has any breaches in the training sample, while GAM performs slightly better than GLM3 out-of-sample and out-of-time. In the tables in Appendix D, GAM has the

[17] We have tried adding vintage dummies and tried modeling the coefficients of the vintage dummies based on the time series of the past vintage dummies. However, we do not find much success from such an exercise. This exercise cures under- or over-prediction for some years but creates new under- or over-prediction in other years.

Table 6
RMSE and MAE by FICO buckets: training, out-of-sample and out-of-time by vintage.

Panel A Root Mean Squared Error (RMSE)													
Training	GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
2009	0.014	0.014	0.012	0.013	0.014	0.038	0.016	0.020	0.047	0.010	0.038	0.020	0.017
2010	0.025	0.016	0.015	0.015	0.025	0.031	0.016	0.016	0.027	0.024	0.031	0.018	0.015
2011	0.018	0.012	0.010	0.010	0.018	0.028	0.018	0.017	0.019	0.016	0.025	0.016	0.011
2012	0.016	0.013	0.013	0.013	0.015	0.032	0.024	0.022	0.023	0.019	0.013	0.020	0.016
2013	0.017	0.010	0.008	0.009	0.017	0.026	0.021	0.018	0.028	0.013	0.020	0.014	0.009
2014	0.022	0.015	0.016	0.016	0.021	0.038	0.033	0.030	0.032	0.023	0.022	0.023	0.018
2015	0.022	0.017	0.017	0.017	0.022	0.036	0.027	0.021	0.031	0.021	0.013	0.022	0.020
Out-of-sample	GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
2009	0.021	0.023	0.019	0.021	0.022	0.047	0.024	0.026	0.053	0.018	0.043	0.028	0.025
2010	0.026	0.021	0.021	0.020	0.027	0.033	0.026	0.025	0.022	0.024	0.043	0.024	0.020
2011	0.022	0.016	0.016	0.015	0.022	0.027	0.021	0.021	0.022	0.020	0.026	0.018	0.015
2012	0.021	0.015	0.015	0.015	0.020	0.027	0.020	0.019	0.024	0.024	0.009	0.017	0.015
2013	0.019	0.012	0.012	0.011	0.019	0.023	0.020	0.021	0.023	0.016	0.024	0.015	0.009
2014	0.019	0.014	0.012	0.013	0.018	0.042	0.032	0.026	0.033	0.021	0.015	0.023	0.016
2015	0.024	0.021	0.021	0.021	0.024	0.037	0.032	0.027	0.033	0.024	0.023	0.026	0.023
Out-of-time	GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
2016	0.021	0.013	0.014	0.011	0.021	0.034	0.023	0.019	0.021	0.019	0.016	0.019	0.015
Panel B mean absolute error (MAE)													
Training	GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
2009	0.011	0.012	0.010	0.011	0.011	0.034	0.012	0.017	0.041	0.009	0.031	0.018	0.016
2010	0.016	0.008	0.008	0.008	0.015	0.025	0.014	0.014	0.018	0.015	0.015	0.014	0.010
2011	0.013	0.010	0.007	0.007	0.012	0.025	0.017	0.016	0.009	0.012	0.017	0.015	0.010
2012	0.011	0.009	0.010	0.009	0.011	0.027	0.021	0.019	0.015	0.013	0.012	0.017	0.012
2013	0.012	0.007	0.005	0.006	0.013	0.023	0.019	0.017	0.020	0.010	0.017	0.012	0.007
2014	0.017	0.009	0.011	0.010	0.017	0.033	0.028	0.022	0.023	0.018	0.020	0.020	0.015
2015	0.020	0.015	0.017	0.016	0.019	0.028	0.020	0.016	0.026	0.019	0.009	0.018	0.018
Out-of-sample	GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
2009	0.015	0.018	0.016	0.017	0.015	0.036	0.019	0.018	0.042	0.013	0.035	0.021	0.019
2010	0.017	0.017	0.017	0.016	0.017	0.029	0.022	0.023	0.017	0.016	0.029	0.021	0.018
2011	0.016	0.014	0.014	0.013	0.016	0.023	0.016	0.016	0.015	0.015	0.018	0.014	0.012
2012	0.015	0.011	0.011	0.010	0.014	0.021	0.014	0.014	0.015	0.017	0.007	0.012	0.008
2013	0.014	0.009	0.007	0.007	0.015	0.021	0.019	0.019	0.016	0.011	0.018	0.014	0.008
2014	0.015	0.012	0.011	0.012	0.014	0.033	0.024	0.018	0.023	0.018	0.013	0.017	0.014
2015	0.019	0.018	0.017	0.017	0.018	0.030	0.026	0.022	0.025	0.020	0.018	0.020	0.020
Out-of-time	GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK
2016	0.018	0.011	0.011	0.010	0.018	0.026	0.019	0.016	0.019	0.016	0.010	0.014	0.014

fewest number of breaches, consistent with its weak leading status in Fig. 4.[18] Further, for the same FICO groups, we can see from Appendix D that while some methods significantly over-predict, other methods significantly under-predict. These findings are likely driven by some important factors that are not incorporated in the models, such as borrower income, job stability, expenditure patterns, etc. As a result, Appendix D suggests that even achieving in-sample predictive accuracy is very difficult for cross-sectional mortgage scoring models, especially at the sub-portfolio level.

4.4. Further investigations

4.4.1. Further investigation along the dimension of predictive accuracy

It might be argued that the poor performance along the predictive accuracy dimension could be due to the use of AUC when we select the hyper-parameters. To address this concern, we have conducted all analysis using the log losses as the performance metrics when selecting the tuning parameters. In unreported results,[19] we find that conclusions on 1) no method clearly leads in predictive accuracy and 2) the inability to accurately capture the early delinquency rates at the vintage level and the FICO sub-group level is robust, while not surprisingly, the rank ordering results deteriorate from those reported in the paper. Therefore, even if we use log losses to select the hyper-parameters, the ML methods (including the two heterogeneous ensemble methods) are still not able to consistently out-perform simpler methods along the different measures of predictive accuracy.

It can also be argued that the inability to capture the early delinquency rates well in the main results might be due to the inclusion of the 2009 data, since the early delinquency rates are obviously higher for loans originated in 2009 than post 2009, as can be seen from Appendix A. When we use the 2010–2016 data and repeat all the exercises in the paper, we find more severe under-prediction of the early delinquency rates from 2014 to 2016.[20] Thus the inability to capture the early delinquency rates is not due to inclusion of the 2009 data.

4.4.2. Including all data from 2000 to 2016

We do not use loans originated before 2009 in the tables and figures presented in the paper because loans originated before

[18] These results are not reported because of space limitations. They are available upon request.

[19] These results are not reported because of space limitations. They are available upon request.

[20] These results are not reported because of space limitations. They are available upon request.

2009 are quite different from those after 2009. However, we have conducted all the analysis using data from 2000 to 2016, and the conclusions we draw from such analysis do not change materially from the conclusions stated in this paper. The only difference is that it is even more difficult to capture predictive accuracy if we include the data before 2009.

4.4.3. Additional right-hand-side variables

In addition to the variables listed in Section 3, we also included in the ML methods geographic information on the location of the property by including the 3-digit zip code among the potential explanatory variables. It is impossible to add such variables on the RHS of either GLM or GAM. We find that adding such additional variables does not improve either the rank ordering ability or predictive accuracy of the ML methods.

5. Conclusions

We investigate the performance of thirteen methods to model mortgage early delinquency. Compared to papers in the existing literature that apply ML methods on mortgage cross-sectional delinquency models, such as Fitzpatrick and Mues (2016), Lessmann et al. (2015), and Li et al. (2017), our investigation is more comprehensive, covering not only risk classification but also predictive accuracy, on a much larger sample representing the U.S. prime mortgage market over post financial crisis periods.

In this paper, we find that, in terms of risk classification, heterogeneous ensemble methods out-perform the other methods in the training, out-of-sample, and out-of-time datasets in our study. However, predictive accuracy is a major challenge facing all mortgage early delinquency models. Ranking of various methods differ based on different predictive accuracy measures. No method clearly and consistently stands out in various metrics of this performance dimension in the training, out-of-sample, and out-of-time datasets. In fact, none of the models can accurately capture predictive accuracy at the refined FICO groups even in the training sample. Such results could be driven by the inability to incorporate some critical risk drivers (such as income) among the list of explanatory variables, and predictive accuracy is thus a major challenge facing all mortgage early delinquency models.

Appendices

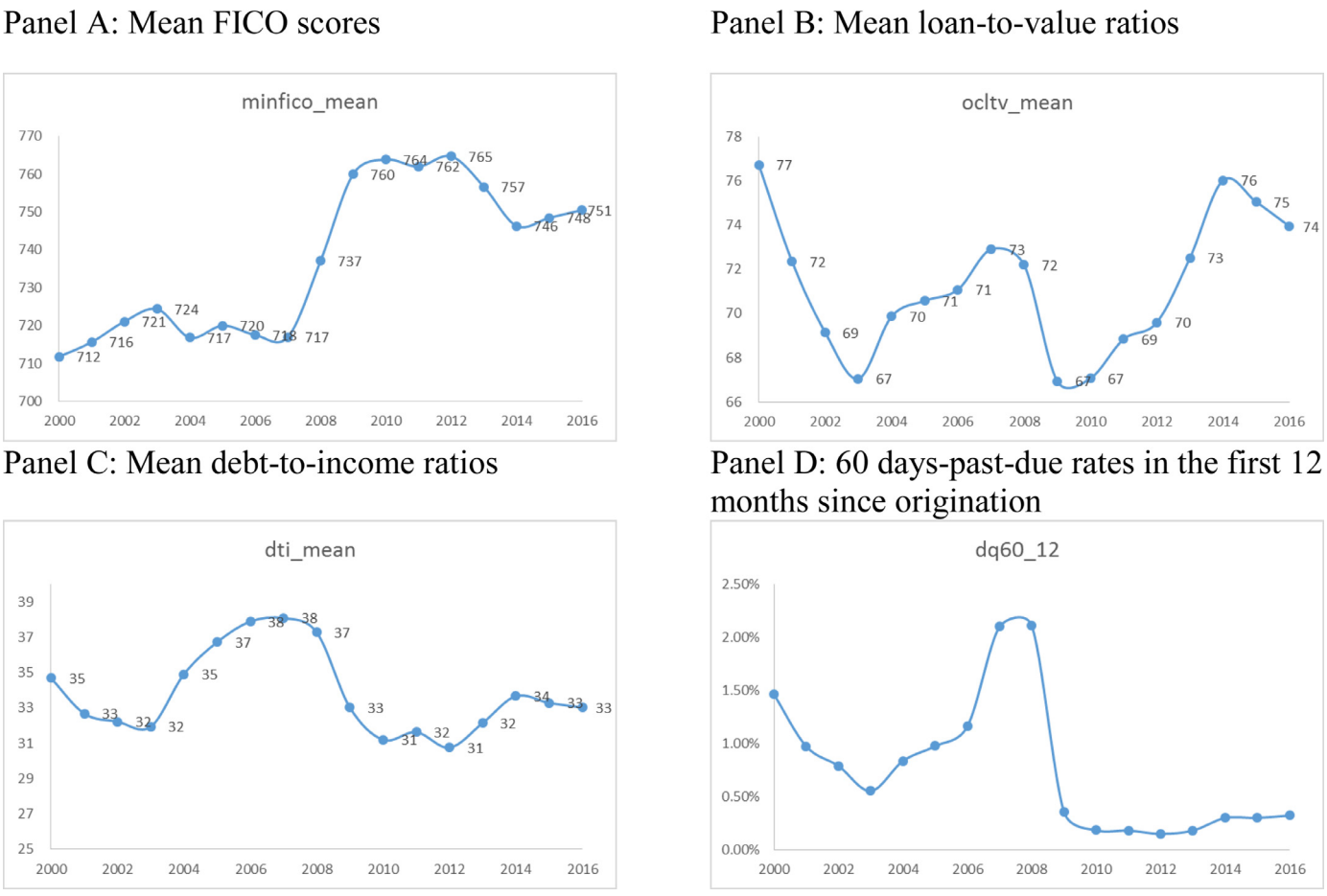

Fig. Appendix A: Loan characteristics and 60 DPD rates in the first 12 months since origination by vintage.

Appendix B. Hyper-parameters used in ML algorithms.*

Regularized Logistic Regression(RLGR)	
C	0.02
solver	newton-cg

Random Forest (RF)	
max_depth	6
n_estimators	1500

Extreme Gradient Boosting (XGB)	
learning_rate	0.008
max_depth	6
n_estimators	500
gamma	0.3
min_child_weight	0.8
reg_alpha	20

Neural Networks (NN)	
hidden_layer_sizes	(50 20)
learning_rate_init	0.04

AdaBoost (ADA)	
learning_rate	0.06
n_estimators	100

Support Vector Machine (SVM)	
C	0.08

K-Nearest Neighbors (KNN)	
n_neighbors	3000
p	2

* We only report hyperparameters that we optimized for the training period 2009-2015 in 1% sample. Different sampling schemes and training periods result in slightly different hyperparameters. Other hyperparameters are set at default values in the scikit-learn package in python. Details can be found in: https://scikit-learn.org/stable/.

Appendix C. Loan count by vintage in training, out-of-sample and out-of-time samples.

Training				Out-of-sample				Out-of-time			
Year	# Good	# Bad	Total	Year	# Good	# Bad	Total	Year	# Good	# Bad	Total
2009	8652	3755	12,407	2009	5298	2300	7598	2016	12,485	4013	16,498
2010	6690	1301	7991	2010	4097	798	4895				
2011	6552	1069	7621	2011	4013	655	4668				
2012	9902	1500	11,402	2012	6064	919	6983				
2013	9836	1514	11,350	2013	6023	928	6951				
2014	5928	1745	7673	2014	3630	1070	4700				
2015	8017	2282	10,299	2015	4909	1398	6307				
Total	55,577	13,166	68,743	Total	34,034	8068	42,102				

Appendix D. Predictive accuracy in training, out-of-sample and out-of-time by FICO bucket.

Panel A: Training (2009–2015)

FICO bucket	Actual avg delq rate (%)	Model-predicted delinquency rate - actual delinquency rate (%)													95% C.I.	
		GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK	Lower bound	Upper bound
(0,620]	81.10	1.03	4.36	2.01	1.64	1.02	0.39	−0.87	4.51	−7.77	2.57	−2.88	2.30	3.10	−3.54	3.54
(620,680]	56.24	−2.12	−0.49	−0.25	−0.04	−2.12	3.66	2.54	0.60	3.23	−1.20	−3.56	1.33	1.29	−1.10	1.10
(680,700]	35.97	1.37	0.66	0.04	0.12	1.36	2.37	−1.04	−0.27	2.70	1.48	−1.22	0.71	0.90	−1.34	1.34
(700,720]	26.56	1.74	−0.22	0.08	−0.36	1.73	−5.97	−2.57	−1.48	0.50	1.59	−0.67	−2.47	−1.38	−1.12	1.12
(720,740]	18.89	1.86	0.20	0.63	0.39	1.87	−3.61	−2.07	−1.42	−0.32	1.55	0.29	−1.47	−0.59	−0.94	0.94
(740,760]	14.60	−0.15	−0.57	−0.70	−0.40	−0.15	−2.89	−1.68	−2.20	−1.39	−0.47	−0.62	−1.93	−1.60	−0.76	0.76
(760,780]	9.63	0.07	0.28	0.10	0.20	0.07	−0.30	−0.26	−0.13	−1.52	−0.11	0.84	−0.11	−0.18	−0.54	0.54
(780,800]	6.74	−0.45	0.05	0.01	−0.06	−0.45	1.54	0.90	1.12	−0.82	−0.51	1.30	0.90	0.34	−0.41	0.41
(800,900]	5.42	−0.73	−0.08	0.07	0.05	−0.74	2.81	2.13	1.86	0.53	−0.75	1.64	1.59	0.76	−0.47	0.47

Panel B: Out-of-sample (2009–2015)

FICO bucket	Actual avg delq rate (%)	Model-predicted delinquency rate - actual delinquency rate (%)													95% C.I.	
		GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK	Lower bound	Upper bound
(0,620]	87.63	−5.40	−2.27	−4.66	−5.09	−5.41	−7.04	−8.29	−1.91	−15.14	−3.97	−9.01	−4.54	−3.44	−3.73	3.73
(620,680]	56.56	−2.53	−0.90	−0.68	−0.46	−2.54	3.13	2.09	0.24	2.88	−1.61	−4.17	0.89	0.85	−1.41	1.41
(680,700]	35.68	1.53	0.82	0.17	0.28	1.53	2.49	−0.90	−0.04	2.82	1.67	−1.12	0.87	1.03	−1.69	1.69
(700,720]	25.53	2.68	0.64	0.98	0.55	2.68	−5.08	−1.66	−0.41	1.39	2.58	0.02	−1.51	−0.46	−1.42	1.42
(720,740]	21.31	−0.42	−2.11	−1.67	−1.92	−0.43	−6.05	−4.42	−3.69	−2.64	−0.80	−2.10	−3.81	−2.89	−1.27	1.27
(740,760]	14.26	0.38	0.02	−0.19	0.14	0.38	−2.45	−1.19	−1.79	−0.90	0.03	−0.03	−1.48	−1.12	−0.97	0.97
(760,780]	9.38	0.03	0.22	0.03	0.14	0.04	−0.24	−0.19	−0.08	−1.47	−0.13	0.85	−0.09	−0.22	−0.68	0.68
(780,800]	6.33	0.04	0.53	0.53	0.43	0.04	2.04	1.36	1.59	−0.37	−0.03	1.80	1.40	0.85	−0.50	0.50
(800,900]	5.94	−1.30	−0.67	−0.51	−0.52	−1.31	2.33	1.62	1.35	0.03	−1.30	1.05	1.07	0.20	−0.64	0.64

Panel C: Out-of-time (2016)

FICO bucket	Actual avg delq rate (%)	Model-predicted delinquency rate - actual delinquency rate (%)													95% C.I.	
		GLM1	GLM2	GLM3	GAM	RLGR	RF	XGB	NN	ADA	SVM	KNN	HC	STACK	Lower bound	Upper bound
(0,620]	62.07	3.15	6.12	6.79	5.87	3.53	1.48	4.67	8.02	−0.32	5.21	−0.56	5.43	7.03	−17.66	17.66
(620,680]	59.22	−2.60	−1.85	−1.99	−1.39	−2.36	1.00	1.93	−0.49	2.00	−1.39	−3.40	−0.49	−0.77	−1.78	1.78
(680,700]	36.91	2.63	0.67	−0.32	0.18	2.87	2.46	−0.80	−0.84	3.64	2.89	1.26	0.43	0.41	−2.43	2.43
(700,720]	26.25	3.02	−0.17	−0.04	−0.22	3.28	−7.17	−3.08	−2.35	1.48	2.88	1.35	−3.20	−1.94	−2.08	2.08
(720,740]	19.62	2.02	−0.58	−0.34	−0.38	2.24	−5.43	−3.91	−3.13	−0.58	1.67	0.66	−2.98	−1.85	−1.90	1.90
(740,760]	15.52	0.20	−0.96	−1.17	−0.74	0.36	−4.17	−3.01	−3.24	−1.59	−0.22	−0.19	−2.86	−2.26	−1.62	1.62
(760,780]	11.21	−0.71	−0.96	−1.23	−1.03	−0.58	−1.88	−1.89	−1.80	−2.62	−0.96	−0.18	−1.64	−1.67	−1.29	1.29
(780,800]	8.46	−1.38	−1.22	−1.31	−1.34	−1.27	0.17	−0.57	−0.40	−2.04	−1.53	0.23	−0.51	−1.07	−1.03	1.03
(800,900]	7.67	−2.47	−2.11	−2.02	−1.99	−2.40	0.89	0.08	−0.21	−1.35	−2.55	−0.15	−0.44	−1.35	−1.29	1.29

References

Adams, N. M., Anagnostopoulos, C., & Hand, D. (2012). *Measuring classification performance: The H-measure package*. London: Imperial College Technical report.

Addo, P. M., Guegan, D., & Hassan, B. (2018). *Credit risk analysis using machine and deep learning models*. Ca' Foscari University of Venice *Working paper*.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175–185.

&, Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*, 627–635.

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Breiman, L. (2000). Some infinity theory for predictors ensembles. *Some infinity theory for predictors ensembles*: 577. US: UC Berkeley Technical Report 577.

Breiman, L. (2004). Consistency for a sample model of random forests. *Consistency for a sample model of random forests*: 670. US: UC Berkeley Technical Report 670.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance, 72*, 218–239.

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. In *In Proceedings of the twenty-first international conference on Machine learning ACM: 18*.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD conference on knowledge discovery and data mining* (pp. 785–794).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273–297.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*, 1447–1465.

David, R. H., Edelman, D. B., & Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. *IMA Journal of Mathematics Applied in Business & Industry, 4*, 43–51.

Desai, V. S., Crook, J. N., & Overstreet, G. A., Jr. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research, 95*, 24–37.

Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*, 3–42.

Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. *European Journal of Operational Research, 249*, 427–439.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal–Japanese Society For Artificial Intelligence, 14*, 771–780.

Friedman, J., Hastie, T., & Tibshirani, R. (2001), *The elements of statistical learning* (1, pp. 337–387). Springer series in statistics.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annuals of Statistics, 29*(5), 1189–1232.

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

Genuer, R., Poggi, J.-. M., & Tuleau, C. (2008). *Random forests: Some methodological insights*. INRIA Research Report RR-6729.

Hand, D. J. (2009a). Mining the past to determine the future: Problems and possibilities. *International Journal of Forecastin,g, 25*, 441–451.

Hand, D. J. (2009b). Measuring classifier performance, a coherent alternative to area under the ROC curve. *Machine Learning, 77*, 103–123.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman & Hall/CRC.

Kennedy, K., Namee, B. , . M. ., & Delany, S. J. (2013). Using semi-supervised classifiers for credit scoring. *The Journal of the Operational Research Society, 64*(4), 513–529.

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

Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems With Applications, 102*, 207–217.

Lessmann, S., Baesens, B., Seow, H.-. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*, 124–136.

Li, Y., Bellotti, T., & Adams, N. (2017). Machine learning performance over long time frame. In *Paper presented at credit scoring and credit control conference*. University of Edinburgh.

Li, Y., Wang, X., Djehiche, B., & Hu, X. (2020). Credit scoring by incorporating dynamic networked information. *European Journal of Operational Research, 286*, 1103–1112.

Martens, D., Baesens, B., Gestel, T. V., & Venthienan, J. (2007). Comprehensive credit scoring models using rule extraction from support vector machines. *Europeans Journal of Operational Research, 183*, 1466–1476.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspective,s, 31*(2), 87–106.

Nikolaou, N., Edakunni, N., Kull, M., Flach, P., & Brown, G. (2016). Cost-sensitive boosting algorithms: Do we really need them. *Machine Learning, 104*, 359–384.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 11*, 169–198.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & Grisel, O. (2011). Scikit-learn: Machine Learning in Pytho. *Journal of Machine Learning Research, 12*, 2825–2830.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers, 10*(3), 61–74.

Qi, M., Zhang, X., & Zhao, X. (2014). Unobservable systematic risk factor and default prediction. *Journal of Banking & Finance, 49*, 216–227.

Sirignano, J. A., Sadhwani, A., & Giesecke, K. (2018). *Deep learning for mortgage risk*. Univ. of Illinois at Urbana-Champaign *Working paper*.

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). Credit scoring and its applications. *The Society for Industrial and Applied Mathematics*.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3–28.

Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research, 238*, 505–513.