Big Data, Network Structure, and the Nonbank Mortgage Sector

Paulo Issler, Richard Stanton, Johan Walden and Nancy Wallace

Abstract

Mortgages issued by bank and non-bank lenders have recently shown very different prepayment, default and delinquency behavior, in line with the warnings of Kim, Laufer, Pence, Stanton, and Wallace (2018). In this paper we plan to use both classical regression and machine-learning techniques to understand in detail the performance difference between bank- and non-bank-originated mortgages. We also plan to build a simple theoretical model of mortgage origination consistent with these empirical findings.

1. Research Strategy

Kim, Laufer, Pence, Stanton, and Wallace (2018) discuss the potential risks posed to the U.S. economy by the rise of non-bank mortgage lenders, which are currently responsible for approximately half of all mortgage lending, but are largely exempt from the regulation and oversight to which their bank counterparts are subject. Their critical dependence on credit to finance mortgage origination and the costs of mortgages in default makes them extremely vulnerable to liquidity shocks in the event of a crisis. Indeed, the much larger share of the market that they represent today means that they arguably represent a greater risk to the financial system now than pre-crisis. Consistent with these concerns, mortgages issued by bank and non-bank lenders have recently shown very different prepayment behavior. In this paper, we plan to use both classical regression and machine-learning techniques, including random forests and deep-learning algorithms, to understand in detail what leads to these performance differences between bank- and non-bank-originated mortgages, including the structure of the financiual network linking the firms (see Stanton, Walden, and Wallace, 2014, 2018, who show that network structure may induce firms to endogenously separate into high- and low-quality regions.)

Since there are no unique mortgage loan identifiers in the U.S., we develop an algorithm using machine-learning techniques to match loans found in two independent datasets: McDash which contains loan-level data compiled by Black Night Financial Services, and ATTOM which provides detailed property transaction and ownership information in addition to a time-series history of all recorded mortgage-lien events such as new mortgage originations, prepayments, REO, foreclosure, short sales, and arms-length sales and loan payoffs.

Our merging process applies a modified k-nearest neighbor classifier (see Hastie, Tibshirani, and Friedman, 2009; James, Witten, Hastie, and Tibshirani, 2015) using `sklearn.neighbors` in Python to fit radial kernels using `BallTree`. The k-nearest-neighbors classifier implements learning based on the 25 nearest neighbors in the corresponding zip code in the McDash data. We represent each loan in each data set with a thirteen-element data vector. Each element is assigned a category subscore between 0 and 1. Each subscore is then squared to achieve a greater penalty for matches on key elements such as the loan amount. The category subscore is then scaled by a factor which

represents the category's importance to the match quality relative to other elements used in the match. Each category factor is an integer between 0 and 100, and the sum of the category factors is 100. Our scoring algorithm (get.score in Python's sklearn) takes into account the 13 different elements of each matched pair of loans to calculate a score. The score roughly corresponds to the estimated error for each match, measured in hundredths of a percent. For fixed-rate GSE loans originated between 2009 and 2015, we obtain a 90% merge rate. We then further merge the ATTOM-McDash data to HMDA data using standardized lender names. Thus, our final merged data set for this analysis is composed of about 23 million GSE and FHA/VA mortgages. Our dataset includes all the origination characteristics of each loan, accurate information on the true lender of record for each loan, information on whether the lender is a bank or a nonbank and the exact address of the property.

In addition to the static origination information for each loan, we also have month to month information on the ex post performance (prepayment, delinquency, and default) of the loans. Thus, our analysis data set is composed of about 1.8 billion loan month performance records. These data will be used to estimate the differential prepayment and default performance of bank and nonbank loans using various versions of competing risk models.

Our second data set focuses on the bank and nonbank labor markets. We have scraped data from S & P Capital IQ for the managers and directors of banks and nonbanks. We have these data for 2,500 managers and directors giving us information on their education (degrees, educational institution), the length of their employment, their prior employment history, and the bankruptcy incidence of prior employment. Following Stanton et al. (2014, 2018), these data will be applied to a network analysis of the quality of the managerial labor market that produced the bank and nonbank loans and their relative performance quality by firm. Our ex post analysis of the underwriting inputs and outputs of the post crisis mortgage market has never before been undertaken at this level of granularity.

2. Computing Resources

The datasets used in this analysis are stored on a 7-node Hadoop cluster, each server configured with 24 cores, 128 GB of RAM, and 20 TB of storage. Data analysis will additionally be performed on a 4-server High Performance Computing Cluster, each server configured with 32 cores and 512 GB of RAM, and on 2 GPU servers, each with 56 cores, 5000 CUDA cores, 24 GB of high-speed memory, 256 GB of RAM, and 10 TB of SSD storage. Initial data analysis on the Hadoop cluster will be performed using Spark and the Impala database on top of the Hadoop Distributed File System (HDFS), and subsequent analysis will be done in Python.

References

Hastie, T., R. Tibshirani, and J. Friedman, 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition (Springer, New York).

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2015, *An Introduction to Statistical Learning with Applications in R* (Springer, New York).

Kim, Y. S., S. M. Laufer, K. Pence, R. Stanton, and N. Wallace, 2018, Liquidity crises in the mortgage market, *Brookings Papers on Economic Activity* (Spring), 347–406.

Stanton, R., J. Walden, and N. Wallace, 2014, The industrial organization of the U.S. residential mortgage market, *Annual Review of Financial Economics* 6, 259–288.

Stanton, R., J. Walden, and N. Wallace, 2018, Mortgage loan-flow networks and financial norms, *Review of Financial Studies* 31, 3595–3642.