ISSN - 1303-099X

EGE AKADEMİK BAKIŞ

Ekonomi, İşletme, Uluslararası İlişkiler
ve Siyaset Bilimi Dergisi

EGE ACADEMIC REVIEW

Journal of Economics, Business Administration,
International Relations and Political Science

Cilt 24 • Sayı 1 • Ocak 2024
Volume 24 • Number 1 • January 2024

Yayınlayan / *Publisher*
Ege Üniversitesi, İktisadi ve İdari Bilimler Fakültesi
Bornova 35100 İZMİR / TÜRKİYE

E-mail: eab@mail.ege.edu.tr

Ege Akademik Bakış

Ege Akademik Bakış Dergisi, iktisat, işletme, uluslararası ilişkiler ve siyaset bilimi alanlarında çalışan akademisyenler, araştırmacılar ve profesyonellerin görüşlerini paylaştıkları bir forum oluşturmak amacıyla, bu alanlarda yapılmış olan uluslararası çalışmaları kapsamaktadır. Ege Üniversitesi İktisadi ve İdari Bilimler Fakültesi tarafından Ocak, Nisan, Temmuz ve Ekim aylarında olmak üzere yılda dört defa yayınlanan hakemli bir dergi olup, Türkçe veya İngilizce olarak kaleme alınmış tüm çalışmalar dergide yayınlanmak üzere gönderilebilir. Ege Akademik Bakış Dergisi aşağıdaki veri tabanlarınca taranmaktadır:

- EconLit (http://www.aeaweb.org/)
- ULAKBİM, Sosyal ve Beşeri Bilimler Veri Tabanı (http://www.ulakbim.gov.tr/)
- Emerging Sources Citation Index (ESCI)
- Director of Open Access Journals(http://www.doaj.org/)
- EBSCO Publishing (http://www.ebscohost.com/)
- PERO(http://knjiznica.irb.hr/pero)
- Scientific Commons(http://en.scientificcommons.org)
- WorldWideScience(http://worldwidescience.org)
- ProQuest(http://www.proquest.com)
- ASOS Index(http://www.asosindex.com)
- RePEc (http://www.repec.org)

Makaledeki görüşler yazarlarına aittir. Dergide yayınlanan makaleler kaynak göstermeden kullanılamaz.

Ege Academic Review includes international papers about economics, business administration, international relations and political science with the aim of providing a forum for academicians, researchers and professionals interested in these fields. This journal is subject to a peer-review process. Ege Academâic Review is published by Ege University Faculty of Economics and Administrative Sciences for four times in a year. Papers written in Turkish and English can all be sent in order to be published in the journal. The articles in Ege Academic Review are indexed/abstracted in:

- EconLit (http://www.aeaweb.org/)
- ULAKBİM, Social Sciences and Humanities Database (http://www.ulakbim.gov.tr/)
- Director of Open Access Journals(http://www.doaj.org/)
- EBSCO Publishing (http://www.ebscohost.com/)
- PERO(http://knjiznica.irb.hr/pero)
- Scientific Commons(http://en.scientificcommons.org)
- WorldWideScience(http://worldwidescience.org)
- ProQuest(http://www.proquest.com)
- ASOS Index(http://www.asosindex.com)
- RePEc (http://www.repec.org)

Authors are responsible for the content of their articles. Papers published in the journal can not be quoted without reference.

Volume 24 • Number 1 • January 2024

Cilt 24 • Sayı 1 • Ocak 2024

Contents

EGE AKADEMİK BAKIŞ / *EGE ACADEMIC REVIEW*

Cilt 24 • Sayı 1 • Ocak 2024
SS. 21/42
Doi: 10.21121/eab.1289964
Başvuru Tarihi: 29.04.2023 • Kabul Tarihi: 25.09.2023

Article Type: *Research Article*

Prepayment and Default Risks of Mortgage-Backed Security Collateral Pools

Tuğba GÜNEŞ[1], Ayşen APAYDIN[2]

ABSTRACT

Mortgage-backed securities (MBS) are structured financial products that are produced via securitization of mortgage loans. Due to the nature of securitization, all risks of mortgage loans are transferred from originators to MBS investors. Prepayment and default risks of mortgages lead to uncertainty in MBS cash flows and create a complex problem for valuation of these instruments. Therefore, estimating these mortgage termination risks has become the focus of valuation of MBS collateral pools. This study explores two questions by using a publicly open dataset provided by Fannie Mae. First, two machine learning algorithms (Random Forest and Multinomial Logit Regression) are used for classification to predict whether a mortgage loan is likely to be prepaid, defaulted or current. Afterwards, Competing Risks Cox Regression Analysis is performed to see determinants of when mortgage termination risks are likely to happen. It is found that not all mortgage borrowers behave optimally in their prepayment and default decisions. Therefore, in addition to refinancing incentive and negative equity which depend on variations in prevailing mortgage interest rates and housing prices, heterogeneity in mortgage borrowers' behaviors and loan characteristics, and also local economic factors are significantly important in estimating mortgage termination risks. It is worth noting that prominence role of mortgage payment delinquencies in particularly predicting defaults emphasizes the essential need of monitoring payments by servicers to keep safety of MBS investors and financial markets.

Keywords: Mortgage Risks, Mortgage-Backed Securities, Valuation Of Mbs Collateral Pools, Real Estate Finance, Machine Learning.

JEL Classification Codes: C14, C53, G17, G21, R30

Referencing Style: APA 7

INTRODUCTION

Value and yield of a financial instrument are the key factors considered for investment decisions in the fixed income securities market. Value of a fixed income security technically equals to present value of its expected cash flows. However, in the case of mortgage-backed securities (MBS), financial analysts and investors encounter with one of the most complex instruments in financial markets due to notorious risks in their collateral pool of loans: prepayment and default risks; i.e. mortgage termination risks (Hayre & Young, 2004).

MBS are structured financial products that are produced via securitization of mortgage loans. Mortgage loans are sold to special purpose vehicles under the true sale doctrine of securitization, which means that all risks of mortgage loans are transferred from originators (housing finance institutions, banks) to MBS investors. Mortgage borrowers are expected to make their payments in line with their loans' amortization schedule periodically, and investors can estimate value of the MBS they hold or plan to make investment by calculating present value of expected cash flows. Yet borrowers' payment behavior may vary significantly. Unscheduled early payments and/or mortgage default decisions result in uncertainty in cash flows of MBS pools and create a complex problem for valuation of these instruments.

Variations in housing prices and mortgage interest rates are the two major systematic factors influencing mortgage termination risks. Depressing housing values triggers mortgage defaults because continuation of repayments will be nonsense if negative equity occurs meaning that the value of house falls below the outstanding balance on the mortgage used for purchasing that property. In other words, despite their ability to make mortgage payments they strategically

[1] Ankara University, Graduate School of Natural and Applied Sciences, Department of Real Estate Development and Management, Ankara, gunest@ankara.edu.tr

[2] Ankara University, Graduate School of Natural and Applied Sciences, Department of Real Estate Development and Management, Ankara, aapaydin@ankara.edu.tr

This paper is extracted from Tugba Gunes's PhD dissertation entitled "Valuation of the Collateral Pool of Mortgage-Backed Securities and Automated Valuation Models" supervised by Prof. Dr. Ayşen Apaydın.

choose to default when they face with negative equity (Foote & Willen, 2018). Within a similar point of view, falling mortgage rates create an incentive for borrowers to prepay their loans by replacing current loan with a new one with lower interest rates (refinancing) (Lowell & Corsi, 2006; Spahr & Sunderman, 1992). Early studies in the literature focus on these two indicators, negative equity and refinancing incentive. They suggest that borrowers behave optimally based on these two indicators and decide whether to prepay or to go default or to continue repaying their loans in scheduled terms (Downing, Stanton, & Wallace, 2005; Dunn & McConnell, 1981). Roughly speaking, that borrowers take the most strategic decision providing the best economic advantageous lays behind the optimal behavior theory, and MBS are referred as financial instruments with embedded options that allow for prepayment and default. These studies adopt option-based theory developed by Black and Scholes (1973) and improved by Merton (1974) to estimate value of an MBS pool.

On the other hand, it is observed that many borrowers take suboptimal decisions (Weiner, 2016). Furthermore, optimal decision differs for every mortgager and should be distinguished from rational decision because not every optimal decision equals to rational decision. For instance, household debt and expenditures, moving from current house due to various life events like marriages, divorces, health conditions, and job or school changes of children are different for all borrowers (Spahr & Sunderman, 1992). In addition to borrower characteristics, mortgage loan features should be counted among the drivers of mortgage repayment behaviors. For example, unpaid principal balance amount, existence of prepayment penalty, amortization type, fixed or adjusted mortgage rates, and legislative regulations undoubtedly influence prepayment and default decisions.

Mortgage terminations are accompanied by transaction costs depending on borrower and loan features. In the case of prepayment via refinancing, borrowers need to consider prepayment penalties and new loan expenses. Having to move from home, finding a new mortgage loan or a rental house, adversely affected credit scores and therefore encountering of higher interest rates and/or not being given a new loan are among the adverse effects of default decisions (Foote & Willen, 2018). In short, any decision is accompanied with transaction costs which differ for mortgage loans with different features but these costs are not limited to economic burden. Psycho-social consequences of mortgage defaults are hardly ignorable because leaving home and being a person who is incapable of paying loan regardless of whether it is a strategic decision will hurt social status and/or mental health of borrowers (Agarwal, Ambrose, & Yildirim, 2015).

Timing is another important criterion to reach the optimal decision. Choosing the most optimal time either to prepay or to go default is crucial for obtaining the most optimal economic advantageous (Kalotay, Yang, & Fabozzi, 2004). Following variations in mortgage rates and housing prices, being aware of trends in financial markets, and understanding sophisticated financial engineering models and tools or getting consultancy from professionals may help borrowers.This reemphasizes the importance of borrower characteristics in terms of their financial decision skills, education and intelligence (Keys, Pope, & Pope, 2016).

Option-theoretical models cannot provide satisfactorily accurate predictions despite attempts of financial institutions by transferring professors studying in this specific field from universities and trying to develop a closed-form formula to determine the value of MBS pools. On the other hand, econometric models that are able to pay attention to borrower and loan characteristics, and local economic indicators are suggested in the literature (Sirignano, Sadhwani, & Giesecke, 2016). Econometric approach tries to model drivers of mortgage risks rather than directly targeting valuation of MBS pools because understanding the mortgage termination risks and their drivers are the core components of valuing MBS. Furthermore, modelling prepayment and default risks are key determinants from approval of loan applications and securitization to creating and rating MBS pools (McConnell & Buser, 2011).

Econometric models are criticized for their data-driven nature, requiring for dealing with huge datasets, and necessity of frequent updates (Weiner, 2016). Technological advances enable working with big data. Improved data accessibility and availability has allowed using machine learning algorithms in financial market analysis-including real estate and mortgage markets, and has been mitigating many drawbacks of the econometric models (Sirignano et al., 2016).

This study employs econometric modelling approach, and explores two questions by using a publicly open dataset provided by Fannie Mae, one of the two leading government sponsored entities in the United States. First, two machine learning algorithms (Random Forest and Multinomial Logit Regression) are used for classification to predict whether a mortgage loan is

likely to be prepaid, defaulted or current. Afterwards, Competing Risks Cox Regression Analysis is performed to see determinants of when mortgage termination risks are likely to happen. It is found that refinancing incentive and negative equity are the two of major determinants of prepayment and default risks respectively. However not all borrowers always take optimal decisions that provides economic advantages. Therefore, these two variables are not able to explain mortgage termination risks sufficiently without considering heterogeneities in borrowers' behaviors. Loan-to-value ratio, debt-to-income ratio, creditworthiness of borrowers, loan age and amount, variations in economic conditions and local default and prepayment rates are found among the major determinants of mortgage risks. It is worth mentioning that mortgage payment delinquencies are significantly important indicators particularly in predicting mortgage defaults. This obviously emphasizes the crucial importance of monitoring payments by servicers to keep safety of MBS investors and financial markets.

This paper is organized as follows. Section 2 provides a summary of existing literature. Section 3 explains mortgage termination risks, and Section 4 provides the details of methodology employed in this study. The data and empirical works are presented in Section 5, and finally Section 6 concludes this paper.

LITERATURE REVIEW

A closed-form formula determining value of complex MBS carrying prepayment and mortgage risks in their collateral pools has not yet constructed despite many attempts of both academicians and sector professionals (Rajashri, Davis, & McCoy, 2016). Majority of the literature focuses on the United States (US) since the country has the largest secondary mortgage market in the world and many types of securitized products. Furthermore, a good part of the literature is only interested in Agency-MBS, which carry guarantees[1] to the investors against losses arising from default risk on the underlying mortgages. Initial studies were only interested in prepayment risk, as the first MBS issuances were made by these institutions and therefore the reflection of default decisions on MBS collateral pools were considered similar to prepayments. Default risk was either ignored or accepted as prepayment (Huh & Kim, 2019). However, default risk has also been started to be included into the analyses along with the increase in private label MBS issuances and awareness of the seriousness of the default risk consequences (McConnell & Buser, 2011).

Another interesting aspect of the literature is that many of initial studies employ option-theoretic models. They assume that refinancing incentive for a mortgage borrower when market interest rate falls below the contract rate triggers refinancing (prepayment) decision, and negative equity that occurs when the value of collateral real estate falls below the outstanding balance on the mortgage loan pushes the borrower to go default (Hayre & Young, 2004; Kau, Keenan, & Li, 2011). Dunn and McConnell (1981) is considered as the first study attempting to value MBS by using option-pricing model of Black and Scholes (1973) and (Merton, 1974). They focus only on prepayments as they study on MBS with Ginnie Mae guarantee, and assume that all borrowers simultaneously prepay their mortgagers at the first moment when refinancing incentive occurs. Following studies take the attention to transaction costs of mortgage risks (Timmis (1985) and Johnston and Van Drunen (1988)). Professionals as well as the academics contribute to the literature, among which Davidson, Herskovitz, and Van Drunen (1988)'s model built for Merrill Lynch is one of the most famous ones. They accept the existence of suboptimal refinances and suggest that variations in transaction costs of borrowers are the major reason of these suboptimal decisions.

Relatively recent studies take heterogeneity in borrowers into account in addition to prepayment transaction costs. Kalotay et al. (2004) categorized borrowers into different groups based on their certain characteristics by arguing that borrowers with similar characteristics have also similar prepayment behaviors. Deng, Pavlov, and Yang (2005) assume that borrowers with similarities in their prepayment and default decisions live in close neighborhoods.

Predictive power of option-based models lags behind econometric models because of nonrealistic assumption of complete optimal behavior. That in addition to variations in mortgage interest rates and housing prices, borrower and loan characteristics, housing market conditions and financial and economic environment have

[1] Ginnie Mae, Fannie Mae and Freddie Mac are named as Agencies in the USA (the latter two are also known as government sponsored enterprises - GSEs). Ginnie Mae is a governmental institution while GSEs could be referred as quasi-governmental institutions. They were established by the Federal Government to enhance the housing sector conditions and economy in the US. Ginnie Mae provides full faith and credit guarantee of the US government for the MBS backed by mortgage loans issued under government agency programs. GSEs provide similar guarantee against default risk for the MBS they issue but this guarantee is not from the government but from themselves. However, their guarantee is known as "implicit guarantee" of the government as they have always been provided various privileges, and supported by the federal government. This implicit guarantee is proven with the rescue of GSEs by placing them into conservatorship of the Federal Housing Finance Agency (FHFA) in 2008.

significant impacts on prepayment and default rates has led to use of econometric models (Kalotay et al., 2004; Weiner, 2016). Loan level data availability has catalyzed adopting econometric models for estimating mortgage termination risks. Despite various econometric modelling-based studies performed in the late 1980s, models developed by Schwartz and Torous (1989) and Richard and Roll (1989) are accepted as the reference studies of the related literature. Schwartz and Torous (1989) employ proportional hazards model to estimate parameters of variables influencing prepayments of mortgage loans in MBS pools. They include refinancing incentive, burnout level of pool and seasonality in their model. Richard and Roll (1989) estimate annual prepayment rates with a multiplicative model built with four variables- refinancing incentive, loan age, seasonality and burnout level of pool. In the following studies features of geographical regions and outstanding mortgage balance are included in the models (Lowell & Corsi, 2006).

Some studies are interested in only prepayments (e.g. Schwartz and Torous (1989)), others focus on only default decisions (e.g. Quigley and Van Order (1991)). However, studying mortgage risks separately is found an unsound approach because occurrence of one risk makes the other impossible. Therefore, it is suggested that prepayment and default risks are competing risks and both should be modelled simultaneously (Bennett, Peach, & Peristiani, 2001). Competing Risk Analysis becomes like a standard procedure in modelling mortgage termination risks (Kau, Keenan, and Smurov (2006), Pennington-Cross (2010)).

Technological improvements allow making advanced model building studies using machine learning algorithms and big data for financial market analyses (Sirignano et al., 2016). For example, Groot (2016) and Mamonov and Benbunan-Fich (2017) analyze mortgage termination risks by using various machine learning algorithms. Sirignano et al. (2016) employ deep learning for classification of mortgages based on their risk exposures. By employing various machine learning algorithms for classification, Cowden, Fabozzi, and Nazemi (2019) focus on default rates of commercial real estate loans. Barbaglia, Manzan, and Tosetti (2023) compare the prediction accuracy of several machine learning algorithms modelling mortgage defaults in European mortgage markets. Zhu, Chu, Song, Hu, and Peng (2023) apply explanatory machine learning models to predict mortgage defaults. Some studies investigate specific subjects related to mortgage risks. For instance, Cooper (2018) compares default rates of modified and non-modified mortgage loans while Fout, Li, Palim, and Pan (2020) perform a similar study to compare default risks of high- and lower or mid- income borrowers. An, Deng, and Gabriel (2021) explore the impact of negative equity on default rates during the 2007 financial crises. A recent study by Blumenstock, Lessmann, and Seow (2022) applies machine learning techniques for survival analyses to predict mortgage terminations.

MORTGAGE TERMINATION RISKS

Prepayment and default risks, i.e. mortgage termination risks or mortgage risks, are the core of valuation of MBS because they crate uncertainties in cash flows on MBS collateral pools (LaCour-Little, 2008). Therefore, estimation of mortgage risks has become the focal point of any studies on valuation of MBS collateral pools. Besides, these risks are core components of mortgage markets from evaluation of mortgage loan applications in primary mortgage market to selection of mortgages to be securitized, credit enhancement, and rating MBS in secondary mortgage market. MBS investors need to spend a good deal of their resources on evaluation and estimation of these risks (Berliner, Quinones, & Bhattacharya, 2016).

Prepayment Risk

Prepayment risk is the probability of a mortgage loan will be fully paid off before its due date. Mortgage borrowers are given the chance of paying off their mortgage debt any time in return for bearing transaction costs although it depends to the legal regulations of the countries (Fabozzi, Bhattacharya, & Berliner, 2007; Rajashri et al., 2016). There are various systematic and idiosyncratic drivers behind prepayment behavior of mortgagors.

Prepayments occur in two ways: (1) refinancing, and (2) housing turnover. Additionally, there are also "payoffs" which means that mortgagors may pay off the mortgage loan debt with their own savings or non-credit resources before the due date. "Curtailments" or "partial prepayments" may occur when borrowers make extra payments in order to shorten the maturity period or reduce the outstanding balance. These two types of prepayments rarely happen and their share is quite minor compared to refinances and housing turnovers.

Refinancing is the replacement of an existing mortgage loan with a new loan without any change in the conditions of the collateral property. There are various types of refinances. One of the most common and well-known types is 'cash-in refinancing' which is preferred by a borrower when current mortgage rates substantially fall below the existing mortgage contract rate. In other words,

refinancing incentive arising from the declines in mortgage rates is the main motivation of cash-in refinances. Amount of the new loan equals to sum of current outstanding balance of the previous mortgage and transaction costs of prepayment. Cash-in refinancing provides mortgage rate and/or term advantageous for borrowers, however, creates a serious risk for MBS investors because principal payments are made unexpectedly early. Investors cannot gain return as much as they expected and also can reinvest at substantially lower prevailing interest rates (reinvestment risk) (Fabozzi et al., 2007).

of the existing loan and use remaining money for their other needs after paying off the current mortgage. A 'cash-out refinance' loan lets a borrower convert home equity into cash. Cash-out refinances tend to get higher when housing prices rises (Rajashri et al., 2016).

Figure 1 and Figure 2 illustrates share of cash-in refinances with mortgage interest rates, and share of cash-out refinances with housing price index, respectively (Data are provided from Freddie Mac (2020), FHFA (2021), and Freddie Mac (2021)). Figure 1 obviously shows the

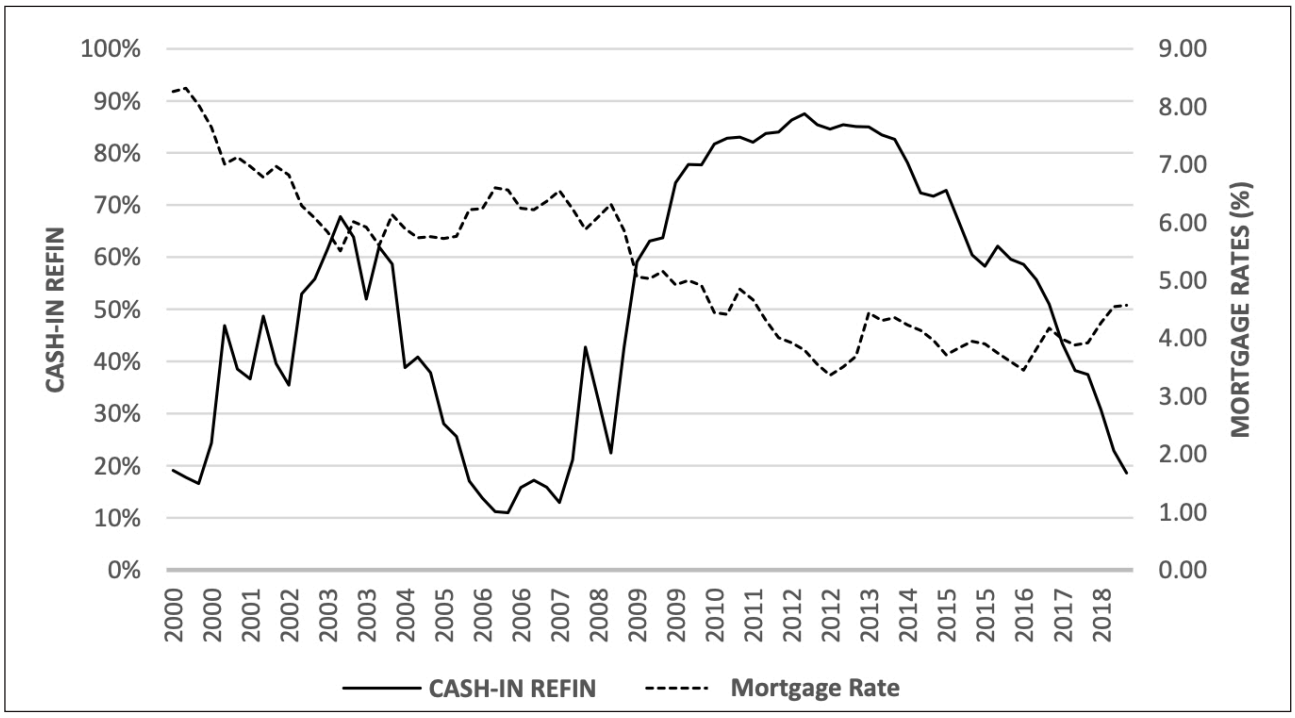

Figure 1. Mortgage Rates and Cash-in Refinancing Loans

Data Source: Freddie Mac (2020) and Freddie Mac (2021))

Refinancing incentive takes place at the center of option-based approach in MBS valuations. This theory assumes that borrowers refinance their current mortgages at the most optimal time when mortgage rates decline sufficiently. However, not all mortgagors can take rational decisions. Optimal decision requires borrowers to be fully informed and educated to follow and understand trends in financial markets. Also, they must guess the most optimal time for themselves because being late/early to refinance may result in suboptimal decisions (Kalotay et al., 2004; Keys et al., 2016).

It is observed in the markets that refinancing incentive is not the only driver of refinances. Borrowers apply for refinancing loans to meet their cash needs as well. Instead of a second-lien mortgage, they apply for a refinancing loan with a higher amount of than outstanding balance

negative correlation between mortgage interest rates and cash-in refinances while Figure 2 indicates the positive correlation between cash-out refinances and house prices.

Another type of prepayment occurs when home owners move from their residents. Housing turnover rate, second home sales as a percentage of total housing stock, provides a measure for this prepayment type in a country or a region (Rajashri et al., 2016). Seasonality influences prepayment speeds because in summers moving to another house becomes more available for people in terms of weather conditions and school term.

Default Risk

Default risk is the probability that mortgage borrowers will not make their payments on their loans

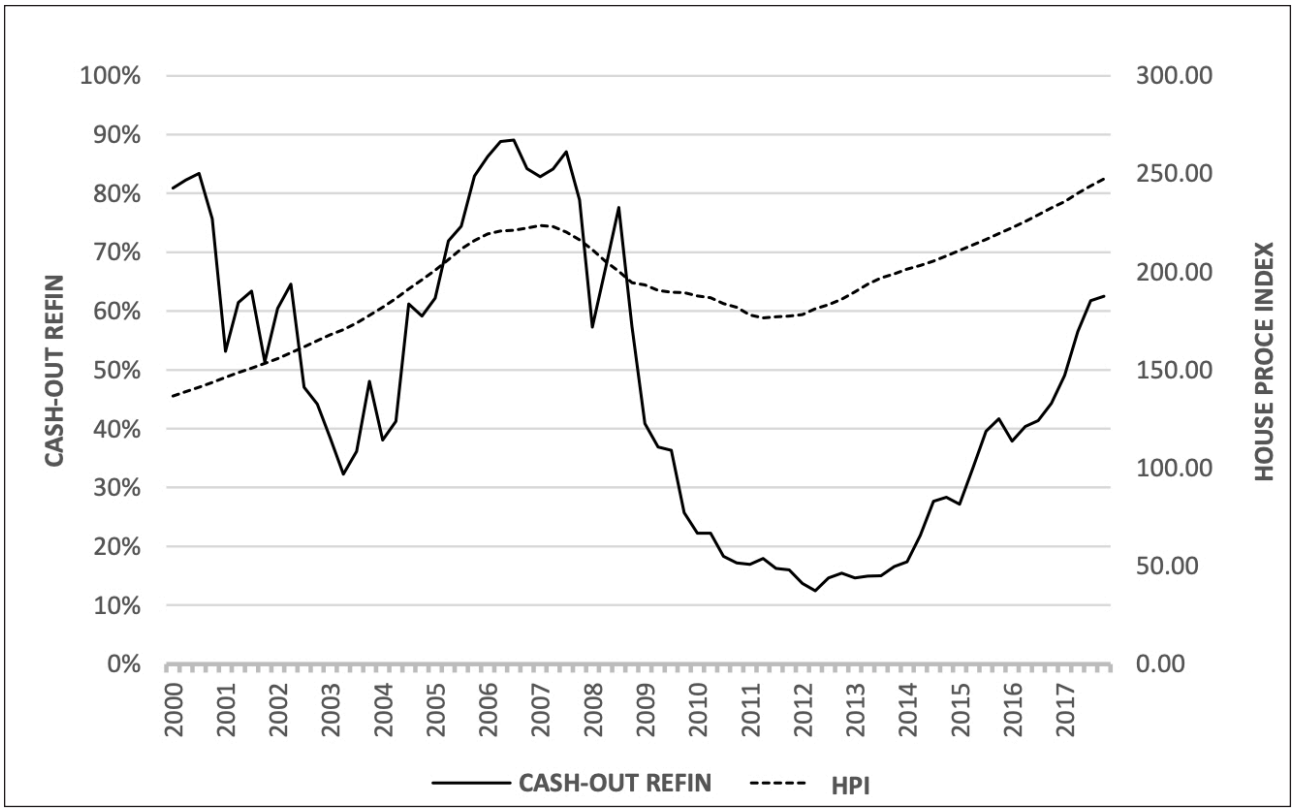

Figure 2. Housing Price Index and Cash-out Refinancing Loans

Data Source: FHFA (2021) and Freddie Mac (2021)

in exchange for giving the collateral property to the financial institution (Berliner et al., 2016). First studies explain mortgage defaults within the context of Black and Scholes (1973) and Merton (1974)'s option pricing theory, and assume that negative equity is the reason of any defaults. Borrowers strategically go to default when negative equity occurs when housing prices fall in the market. Another assumption of this frictionless/ ruthless option model is that there is no any cost or loss for borrowers other than losing their house, and they are able to obtain a new loan as much as they need with prevailing interest rates any time (Foote & Willen, 2018).

However, these assumptions are never always true. Negative equity alone cannot trigger default unless it deepens significantly (e.g., according to Foote and Willen (2018) mortgagors do not prefer unless negative equity reach to at least 35% or 40%) because people are unwilling to lose their home easily. Many studies show that defaults happen when an adverse life event accompanies with negative equity. This is called Double Trigger Model, which suggests that defaults and delinquencies occur if and only if negative equity and also an idiosyncratic shock adversely affecting households' capability of making payments happen together in the same household (Foote & Willen, 2018). Unemployment, income cuts, excessive financial stress, and also a serious disease or a death of a family member and divorces are the major shocks leading to delinquencies and defaults (Schelkle, 2018).

Furthermore, defaults are recorded in the financial history of borrowers for years and seriously harm their credibility. Credit institutions hesitate to lend a new loan such borrowers, or even if these borrowers are granted with a new loan, most probably the amount of the loan will be lower than they need and with higher interest rates (Demyanyk, 2017). Psycho-social consequences are worth to remember as well (Agarwal et al., 2015). So, consequences of default decisions are not limited to only losing homes. On the other hand, strategic defaults occur at a level of that to be underestimated (Gerardi, Herkenhoff, Ohanian, & Willen, 2018).

METHODOLOGY

Machine learning algorithms are quite popular in finance literature recent years (Sirignano et al., 2016). This study employs two supervised machine learning algorithms, Random Forest (RF) and Multinomial Logistic Regression (MNL), for classification of mortgages based on their repayment statuses. Random Forest is an ensemble machine learning algorithm using bagging

technique. There are many other ensemble machine learning algorithms with their own advantages and disadvantages. One fundamental advantage of the Random Forest over some of the other advanced learning models is that Random Forest is more interpretable and has better transparency (Tchuente & Nyawa, 2021). Multinomial Logistic Regression is commonly used in particularly predicting mortgage defaults in the literature (e.g. Mamonov and Benbunan-Fich (2017), Chen, Xiang, and Yang (2018), Barbaglia et al. (2023)), which allows comparing the results of this study with those of previous works. Afterwards, Competing Risks Cox Regression is performed to estimate marginal probability of prepayment and default risks of mortgage loans.

Classification with Machine Learning Algorithms

Logistic Regression is used to predict a categorical variable with two categories (binary variable). It estimates the probability of an event occurrence based on given a set of independent variables, and assumes a linear relationship between the binary dependent variable and the covariates which may include continuous variables. A transformation from probability to log-odds is applied to satisfy the linearity assumption, and model becomes as follows:

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (1)$$

When the dependent variable has more than two categories, it is called Multinomial Logistic Regression and used frequently for classification problems in the literature.

Random Forest (Breiman, 2001) is a tree-based ensemble machine learning algorithm that is used for solving classification and regression problems. Decision trees are the basis of any tree-based models. A decision tree is composed of leaves and nodes. Each independent variable is represented with the nodes. Independent variables are sorted based on their importance, and the most important one becomes the first node of the decision tree, and split the given dataset into subgroups. Covariates are reordered and second level nodes continue partitioning the dataset. This recursively splitting process based on certain criteria until the optimal classification is reached. Decision trees are able to be implemented and interpreted easily but prone to overfitting. Ensemble modelling methods are proposed to mitigate the overfitting problem. Random Forest, one of most powerful ensemble algorithms, averages predictions of many individual trees built on bootstrap samples. It is an extension of bagging (bootstrap aggregating) that fits many models on different subsets of a dataset by using randomly selected variables in each subsample. Random Forest is able to reduce overfitting problem of decision trees and handle nonlinearity and feature interactions (Kok, Koponen, & Martínez-Barbosa, 2017).

Classification works are started with splitting the dataset into two subsamples, 70% of the dataset (training data) is used to train the machine learning models, and prediction performances of the models are tested on the rest (30%) (testing data) (Hertzmann & Fleet, 2012). k-fold cross validation is a technique that is proposed in order to avoid overfitting risk. This technique splits the data into k sub-groups, and trains the model by using k-1 groups of data. The trained model is tested on the other sub-group (validation data), and this procedure is repeated k times (Berrar, 2018).

The training process of a model involves choosing the optimal hyperparameters. These hyperparameters are essential components of the training to improve the learning capability of the model. For instance, finding out the optimal number of trees to be included in a Random Forest is required for the training. Hyperparameters can be set heuristically or tuned via various techniques proposed in the literature. This study employs Random Search technique offered by Bergstra and Bengio (2012). The main idea behind this technique briefly is that after setting up a grid of hyperparameter values, model training is performed on randomly selected combinations of those values.

Confusion matrix, showing the predicted and actual classifications, helps in evaluating classification performances of machine learning models. For a binary classification, there are two classes as 'positive' and 'negative'. Confusion matrix can be written as follows:

$$Confusion\ Matrix: \begin{bmatrix} \varphi_{PP} & \varphi_{PN} \\ \varphi_{NP} & \varphi_{NN} \end{bmatrix} \quad (2)$$

where φ_{PP} is True Positive (TP: number of elements belonging to the class Positive and are classified in class Positive), φ_{NP} is False Positive (FP: number of elements are wrongly classified in class Positive), φ_{PN} is False Negative (FN: number of elements are wrongly classified in class Negative), and φ_{NN} is True Negative (TN: number of elements belonging to the class Negative and are classified in class Negative). Classification performance of a model is evaluated with various metrics. This study uses accuracy, sensitivity and specificity that are defined as follows (Alpaydin, 2020):

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

Competing Risks Cox Regression

Survival Analysis is used to analyze the expected duration of time until a certain event (e.g. time from surgery to death) occurs. In this study, that incidence is mortgage termination before the scheduled maturity date due to exposure of either prepayment or default risks. Suppose is a non-negative variable representing the duration time until the termination of a mortgage (survival time of a mortgage). Probability that a mortgage will continue to be paid after time , i.e. survival probability (also known as survival function) is as follows:

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx = 1 - F(t) \quad (6)$$

and, survival probability of a mortgage loan is 1 at t_0 ($\lim_{t\to 0} S(t) = 1$), and 0 (zero) as time approaches to infinity ($\lim_{t\to\infty} S(t) = 0$).

Hazard function is the probability of termination of a mortgage at time t, when this mortgage has not experienced the event (termination) until time t is as follows:

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \le T < t < \Delta t \backslash T \ge t}{\Delta t} = \frac{f(t)}{S(t)} \quad (7)$$

and Cumulative Hazard Function is defined as:

$$F(t) = P(T \le t) = \int_0^t f(x)dx = 1 - S(t), \qquad t > 0 \quad (8)$$

When the event of interest (exposing to mortgage termination risks) is not observed for some individuals (mortgage loans) before the study is terminated, survival times would be remain unknown for a subset of mortgages. This is called censoring, and if there had been no censored observations, time to event analysis would have been estimated by using regression analysis. As survival times of censored observations (mortgage loans) are exactly unknown, they should be taken into account while estimating survival function (Kaplan & Meier, 1958; Link, 1989). Both prepayments and defaults cause right censoring in the data. Some mortgages continue to be paid in the dataset, therefore they have not experienced prepayment or default yet. Also, if a mortgage is prepaid, then it cannot be defaulted vice versa. Therefore, defaulted (/prepaid) mortgages are accepted as censored for prepayment (/default) function.

The most common non-parametric method used to estimate the survival function is the Kaplan-Meier estimator but it is unable to consider the variables influencing the survival time. A flexible and semi-parametric method called Cox Regression is proposed to incorporate the covariates into the analyses (Cox, 1972). However, traditional survival analysis might be misleading if occurrence of a certain event depends on more than one reason. Competing Risks Cox Regression is proposed in such cases, and rather than using Kaplan-Meier estimator, Cumulative Incidence Function is proposed in estimating the marginal probability of the specific event of interest (Kalbfleisch & Prentice, 2011).

An event may occur due to one of reasons in competing risks analyses, and the time elapsed is observed only until the first (earliest) of these reasons occurs. Therefore, let denotes survival time ad is the reason of an event occurrence (for instance, mortgage termination is an event while the reasons are prepayment or default). Cumulative incidence function for the reason is as follows:

$$CIF_k(t) = P(T \le t, \delta = k) \quad (9)$$

Two methods are proposed for competing risks analyses in the literature, Cause Specific Hazard Function by Prentice et al. (1978) and Sub-distribution Hazard Function by Fine and Gray (1999). This study employs the Cause Specific Hazard Function that is defined as follows:

$$\lambda_k(t) = \lim_{\Delta t\to 0} \frac{P(t \le T < t + \Delta t,\ \ \delta = k \ \backslash \ T \ge t)}{\Delta t} \quad (10)$$

because Fine and Gray (1999) is interested in the occurrence of an event due to reason for the observations who have not experienced the reason while Prentice et al. (1978) provides the occurrence rate of an event due to reason for the observations that have not experienced any of the reasons.

DATA AND EMPIRICAL WORKS

The primary data for this study are provided by Fannie Mae for the period between 2000 and 2019 (Fannie Mae, 2019)[2]. Dataset consists of loans with fully amortizing,

[2] Data for the following years are seriously affected by the Covid19 because various changes and support schemes were launched to support borrowers during the pandemic, which are not fully reflected in the data.

single-family, 30-year, fixed rate mortgages. Data provide mortgage and borrower characteristics at the time of loan origination, and monthly payment performances of each loan. Mortgages that were modified or refinanced under certain programs are excluded from the dataset. Mortgage loans are defined as prepaid or defaulted according to Fannie Mae's instructions in the glossary file enclosed with the data. A loan is defined as prepaid if it is indicated as prepaid, repurchased or re-performing loan sale; and is defined as defaulted if it is indicated as third party sale, short sale, deed-in-lieu, note sale, or delinquent for 120 days or more. A stratified random sample of data based on loan origination month is used in the study. Under sampling method is employed to ensure a balance among mortgage payment status groups (prepaid, defaulted, and current) because number of defaulted loans is substantially smaller than prepaid and current loans (Drummond & Holte, 2003). The final dataset includes more than 455,000 loans, and Table 1 provides the number of observations for dependent variable per category of mortgage states.

Table 1. Mortgage States in the Dataset

Mortgage States	Sample
Current	152.766
Prepaid	163.315
Default	139.192
Total	**455.273**

Variables that are available for mortgage loan attributes at the time of origination, performance metrics for borrowers' payment behavior, and key economic indicators are listed in Table 2. Mortgage loans in the United States are originated through three channels. In other words, mortgage applications are made via either directly banks, or correspondents, or brokers. These loans can be used for purchasing a house or refinancing the current mortgages. Fannie Mae dataset provides the information on whether the collateral property is a second home or an investment for the borrower. Majority of the residential properties are single family homes in the USA. Since the number of other property types (condominiums, cooperative shares, planned urban developments and manufactured homes) are quite low in the dataset, they are collected under one category, "others". First time home-buyer flag represents whether it is the very first mortgage loan of the borrower. LTV and DTI represent the ratio of loan amount to collateral value, and the ratio of mortgage debt to borrower's income, respectively. Credit score is also known as FICO score that is a measure representing the credibility of the borrowers based on their financial behavior history. Loan age is the number of months since the mortgage origination date. Monthly refinancing incentive for each loan is calculated by distracting prevailing average mortgage interest rate from the contract rate. By following Demiroglu, Dudley, and James (2014), monthly negative equity for each mortgage is estimated as follows:

$$Negative\ Equity_{it} = EP_i - EP_i * HPA_t \quad (11)$$

where EP_i is estimated price of collateral property of loan i, which is calculated by dividing loan amount to loan-to-value ratio of loan i, and HPA_{it} is the house price appreciation rate for month t, calculated with house price index. Analyses are performed based on BEA Regions[3] because there are limited observations for several federal states in the dataset. Historical delinquency status of each loan is used in empirical analyses to represent borrowers' payment behaviors. Calculated prepayment and default rates at zip code level by using the original dataset (23.3 million mortgage loans) are used as the estimators of these two covariates. The dataset has both numerical and categorical variables. Continuous variables are used in their normalized versions. One-code encoding is used to transform categorical variables into numerical values

Results: Classification Analyses

Mortgage loans are classified based on their repayment status: prepaid, defaulted, and current. These three-state classification studies are performed with two machine learning algorithms, Random Forest (RF) method and Multinomial Logistic Regression (MNL). Random Search technique (Bergstra & Bengio, 2012) is employed to optimize the hyperparameters. The optimal hyperparameters for number of variables (mtry) and number of tress (ntree) for the final model turn out to be 24 and 400, respectively. Models are trained using the randomly selected 70% of the dataset (training data), and then tested on the rest of the data (testing data). By following the literature (López, López, and Ponce (2022) and Davis et al. (2022)), 5-fold cross validation is used for the evaluation of algorithm performances. The caret package of the R software is used to build the models. All mortgage loans are followed until either they exposed to a mortgage termination risk or the latest available month in the dataset.

[3] BEA regions are created by the US Bureau of Economic Analysis of the US Department of Commerce.

Table 2. Independent Variables

Variable	Short description
Loan origination channel	banks; correspondents; brokers
Loan purpose	purchase; cash-in refin; no cash-in refin
Occupancy status	owner occupied; second home or investment
Property type	single family homes; others
First time home-buyer	yes; no
Number of borrowers	one borrower; more than one borrower
Loan amount (USD)	US Dollar
Loan-to-Value ratio (LTV)	ratio of mortgage loan amount to property value
Debt-to-Income ratio (DTI)	ratio of mortgage debt to borrower income
Credit score of borrower	borrower's credibility measure
Mortgage interest rate	mortgage interest rate on the loan contract
Loan Age	number of months since origination date
Refinancing incentive	refinancing incentive monthly basis
Negative equity	negative equity monthly basis
Number of delinquencies	mortgage payment delinquencies
Unemployment rate	monthly unemployment rate at state level
House price index	monthly house price index at state level
Prepayment rate at zip code level	local prepayment rate
Default rate at zip code level	local default rate
Seasonality	winter, summer, spring, fall
BEA Region	8 regions based on state-level economic activity
Loan origination year	loan origination year

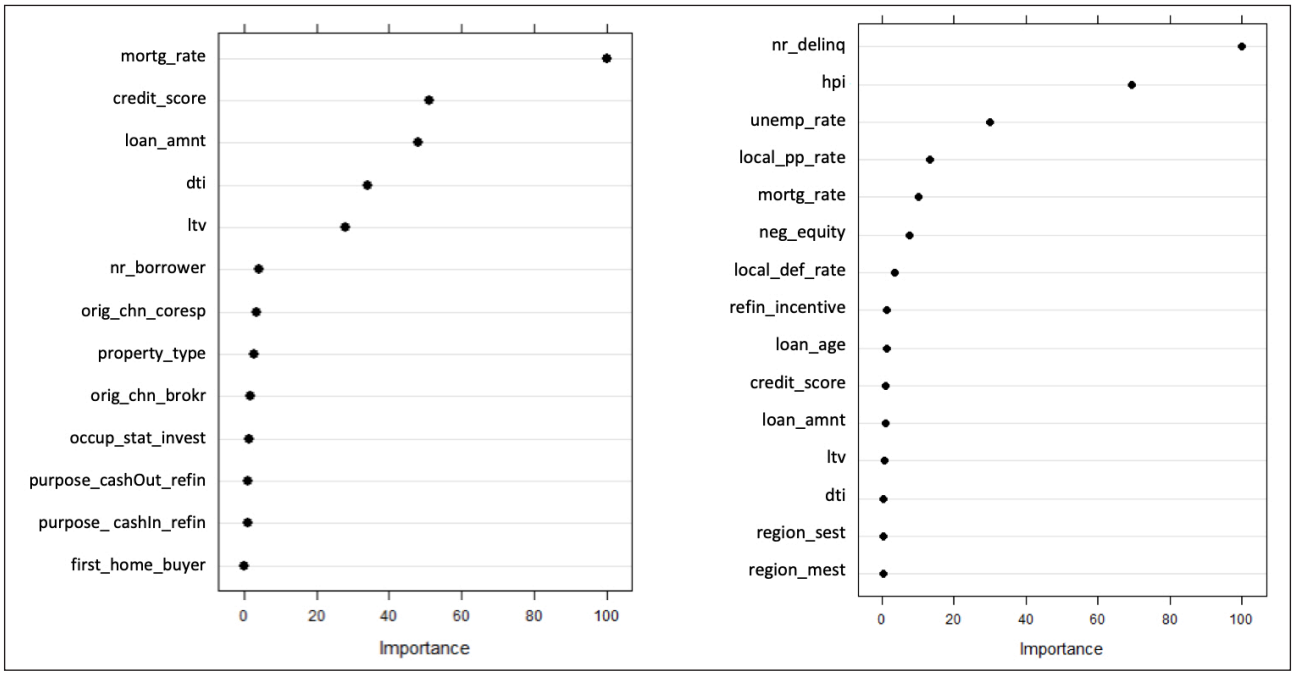

Figure 3. Variance Importance Plots for the Random Forest Model 1 (left) and Model 4 (right)

Table 3. Classification Results with Machine Learning Algorithms

		Random Forest			Multinomial Logit		
MODEL 1	Accuracy		0.699			0.688	
	p-value		p<0.001			p<0.001	
	Kappa		0.548			0.531	
	F1 score		0.799			0.803	
	Reference:	Current	Prepaid	Default	Current	Prepaid	Default
	Current	0.282	0.069	0.017	0.287	0.074	0.017
	Prepaid	0.035	0.198	0.069	0.031	0.200	0.086
	Default	0.019	0.092	0.219	0.018	0.085	0.201
	Sensitivity	0.839	0.551	0.718	0.854	0.556	0.660
	Specificity	0.869	0.838	0.841	0.862	0.817	0.852
MODEL 2	Accuracy		0.860			0.842	
	p-value		p<0.001			p<0.001	
	Kappa		0.790			0.762	
	F1 score		0.958			0.943	
	Reference:	Current	Prepaid	Default	Current	Prepaid	Default
	Current	0.332	0.008	0.017	0.330	0.018	0.017
	Prepaid	0.001	0.285	0.045	0.000	0.280	0.057
	Default	0.003	0.066	0.243	0.005	0.061	0.231
	Sensitivity	0.990	0.794	0.795	0.984	0.781	0.757
	Specificity	0.961	0.928	0.902	0.947	0.911	0.905
MODEL 3	Accuracy		0.871			0.851	
	p-value		p<0.001			p<0.001	
	Kappa		0.807			0.776	
	F1 score		0.969			0.949	
	Reference:	Current	Prepaid	Default	Current	Prepaid	Default
	Current	0.333	0.003	0.016	0.330	0.013	0.016
	Prepaid	0.000	0.293	0.045	0.000	0.284	0.053
	Default	0.002	0.063	0.245	0.005	0.061	0.236
	Sensitivity	0.993	0.817	0.802	0.987	0.939	0.937
	Specificity	0.971	0.930	0.907	0.959	0.990	0.982
MODEL 4	Accuracy		0.987			0.962	
	p-value		p<0.001			p<0.001	
	Kappa		0.981			0.943	
	F1 score		0.993			0.969	
	Reference:	Current	Prepaid	Default	Current	Prepaid	Default
	Current	0.332	0.001	0.000	0.333	0.005	0.014
	Prepaid	0.000	0.353	0.003	0.000	0.344	0.007
	Default	0.003	0.005	0.302	0.002	0.010	0.285
	Sensitivity	0.990	0.983	0.989	0.992	0.958	0.933
	Specificity	0.998	0.995	0.988	0.972	0.989	0.982

Model 1 is constructed using only mortgage attributes at the time of origination to anticipate what extent default and prepayments can be estimated during evaluation of mortgage applications, and Model 2 is performed by adding BEA regions and loan ages in the analyses. Model 3 is built by adding negative equity and refinancing incentive variables. Finally, local economic indicators and delinquency behaviors of borrowers are included in Model 4. Table 3 provides performance metrics of all these models including the confusion matrixes.

Accuracy scores of the Model 1 using the characteristics of mortgages and loans are found at around 69% for both RF and MNL algorithms. However, confusion matrixes reveal that classification performances are quite low. The Model 2 built with additional variables including the BEA regions where collateral property of mortgages located, mortgage origination years (vintages) and loan ages has a significantly higher accuracy score, 86% for the RF and 84% for MNL model; and provides better classification performances as well. Particularly the loan age variable[4] has a significant impact on predicting prepayment rates as the sensitivity score rises to almost 80% for 'prepaid' class in each machine learning algorithm, which supports that loan age is a major driver of mortgage prepayments.

Model 3 is constructed by adding negative equity and refinancing incentive. There is only a slight increase in the overall accuracy scores but the major increase occurs in sensitivity scores. Obviously, these two are significantly important variables to predict prepayments as prepayments, and defaults as defaults accurately. Delinquency experiences of the mortgages and economic factors are included in the final model. Both scores for overall accuracy and sensitivity are found quite high in the Model 4.

Normalized variable importance for the first and last Random Forest models are plotted with the caret library (Figure 3). Among the features of mortgages at the time of origination, mortgage interest rate is the variable that has the greatest effect on the mortgage termination risks in Random Forest Model 1. Borrower credibility, loan amount, and DTI and LTV scores are the other variables having the most importance in the model after the mortgage interest rate. When it comes to the latest model, delinquency behavior, local economic factors, refinancing incentive and negative equity are found among the most important variables in the Model 4. Mortgage interest rate, loan amount, credit score of borrowers, and LTV and DTI scores are still among the top 15 important variables. Loan age is definitely an important feature in the Model as it is a significant determinant of mortgage defaults and prepayments. That the seasoning is accompanied with delinquency behavior, refinancing opportunities and negative equity also implies the interaction among the variables. Therefore, mortgage features at the time of origination as well as the mortgage payment behavior and changing circumstances by time should be evaluated carefully to mitigate the impacts of mortgage risks exposures.

[4] Another model without the Loan Age variable is built during the empirical works, and sensitivity levels are found lower than 65% for both machine learning algorithms. These results may be obtained from the authors upon request.

Results: Competing Risks Cox Regression Analyses

Econometric models are required to be updated on a frequent basis because loan and borrower attributes, and economic conditions vary in time. For instance, the Basel Accords suggest twelve months for credit risk analyses. For prepayment and default modelling, first 12 or 24 months of observation period are recommended in the related literature (Fout et al., 2020) because prepayment and default rates show an increasing trend in the first years of the mortgage loans, and continue at a relatively constant level in the following years (Hayre & Young, 2004). As seen Figure 4, prepayment and default rates in the Fannie Mae dataset show a rising trend in the first years of mortgage loans. Therefore, Competing Cox Regression Analyses are performed for two years (24 months) of observation period (i.e. first two years of loans since their origination date) of the mortgage loans in the dataset by employing the Cause Specific Hazard Function offered by Prentice et al. (1978).

Model 1 is constructed with the variables of refinancing incentive and negative equity. Both are time varying covariates. In the Model 2, borrower and loan features at the time of origination (time invariant) are incorporated in the analyses. Model 3 is built with additional two variables, BEA regions and mortgage origination years (vintage). Finally, a set of time varying variables representing delinquency behaviors and one-month lagged regional economic indicators, are also taken into account in Model 4. All models are shown in Table 4. A variable with a positive/negative sign indicates that occurrence of risk (prepayment or default) will happen sooner/later; i.e. this variable has an effect of prolonging/shortening the survival time of a mortgage.

In line with theoretical expectations, both refinancing incentive and negative equity have positive a relationship with the occurrence of prepayment and default risks respectively, and parameters are found significantly important. The higher the refinancing incentive, the higher the risk of prepayment; and similarly the higher the negative equity, the higher the risk of default, as are stated in the literature (e.g. Sirignano et al. (2016) and Gerardi et al. (2018)).

Mortgage borrowers can apply for mortgages through the channel of banks or correspondents or mortgage brokers. Default rates are higher among the mortgages that are granted via the latter two channels in Table 4 because brokers have no responsibility about

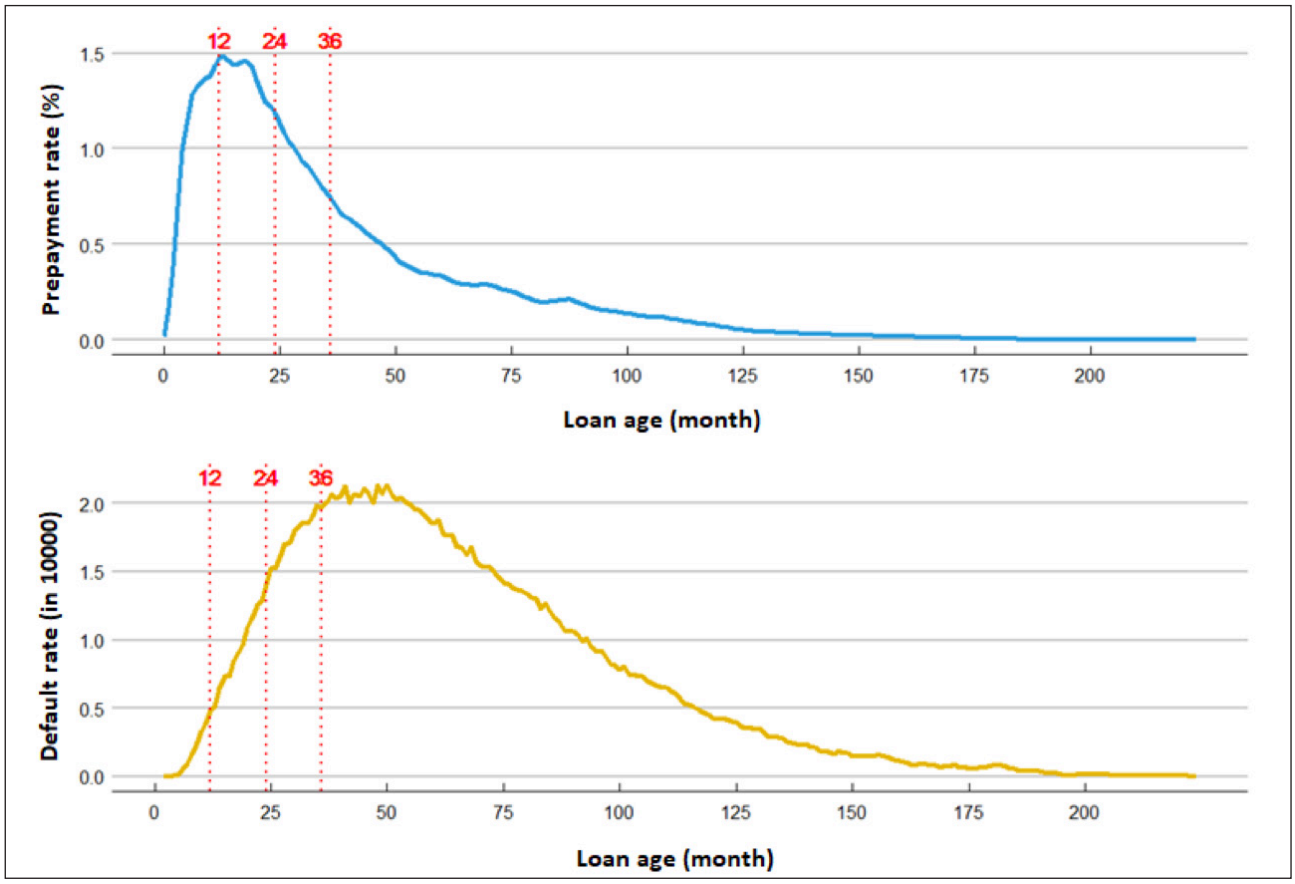

Figure 4. Prepayment and default rates based on Loan Age

loan repayments. The primary purpose of brokers is to generate high commission income by providing as much credit as possible, and they might take the advantage of information they have more about financial institutions and brokers for their own benefit. On the other hand, the results are found against the theoretical expectations for prepayment risk; exposure to prepayment risk is higher for mortgages originated via third party channels than banks. Cash-out refinancing loans might be lying behind this result.

In order to reduce their debt, mortgagors tend to add their current savings while making cash-in refinancing. Cash-out refinancing loans naturally increase borrowers' indebtedness which prevents making savings and therefore prepayments. Refinancing loans in prime mortgage market have higher default risk, which is supported in Table 4. Compared to mortgage loans granted for housing purchases, refinancing loans have lower prepayment and higher default risks.

Results for the occupancy status are consistent with theory. Mortgage loans used for purchasing home for investment purpose have higher default rates, and lower prepayment speeds. As the number of borrowers liable to pay the loan increases, prepayment risk increases and default risk decreases.

Similar to Patrabansh (2015)'s findings, probability of exposure to prepayment risk is found lower among the first-time home buyers, which might be arising from that first-time home buyers may be too young to have sufficient savings and income. On the other hand, there is no significant relationship found between first-time home buyers and default risk. Socio-economic and demographic factors undoubtedly influence housing acquisitions. Relationship between borrower attributes and mortgage payment behaviors could not be measured due to unavailability of data on these factors in the Fannie Mae dataset.

There is a positive correlation between loan amount and mortgage termination risks. Refinancing incentive occurs if the loan amount is large enough to cover refinancing costs. Besides, Sirignano et al. (2016) states that borrowers with high creditworthiness are able to obtain higher amount of loans, which implies getting refinancing loans might be easier for such borrowers. Also, when housing prices show a declining trend, negative equity occurs more rapidly for the loans with high volume. Therefore, findings about the relationship of loan amount with mortgage termination risks in Table 4 are consistent with theoretical expectations.

Credit scores of borrowers, and LTV and DTI ratios of the loans are counted as the leading determinants of mortgage

Table 4. Competing Risks Cox Regression Analyses Results (1/3)

	Model 1		Model 2		Model 3		Model 4	
	Prepay	**Default**	**Prepay**	**Default**	**Prepay**	**Default**	**Prepay**	**Default**
Refinancing incentive	0.557***	0.680***	0.460***	0.377***	0.690***	0.116***	0.588***	0.236***
	(0.004)	(0.006)	(0.006)	(0.008)	(0.01)	(0.014)	(0.007)	(0.01)
Negative equity	-0.185***	0.199***	-0.130***	0.273***	-0.016***	0.230***	-0.004	0.228***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.005)	(0.005)	(0.008)
Channel: correspondent			0.037***	0.161***	0.073***	0.136***	0.062***	0.151***
			(0.011)	(0.017)	(0.011)	(0.017)	(0.011)	(0.018)
Channel: broker			0.005	0.336***	-0.009	0.310***	0.008	0.277***
			(0.013)	(0.019)	(0.013)	(0.019)	(0.013)	(0.019)
Purpose: no-cash-in			-0.304***	0.437***	-0.250***	0.330***	-0.253***	0.483***
			(0.013)	(0.021)	(0.014)	(0.022)	(0.014)	(0.022)
Purpose: cash-in			-0.041***	0.433***	-0.139***	0.365***	-0.140***	0.400***
			(0.013)	(0.021)	(0.013)	(0.021)	(0.013)	(0.021)
Occupancy status: investment			-0.877***	0.166***	-0.608***	0.066***	-0.585***	0.338***
			(0.017)	(0.021)	(0.017)	(0.022)	(0.017)	(0.022)
Property type: others			-0.027**	-0.066***	0.107***	-0.114***	0.067***	0.011
			(0.011)	(0.018)	(0.011)	(0.018)	(0.012)	(0.018)
First home-buyer: No			0.315***	0.002	0.252***	0.017	0.252***	0.005
			(0.017)	(0.026)	(0.017)	(0.026)	(0.017)	(0.026)
More than 1 borrower			0.382***	-0.715***	0.271***	-0.676***	0.242***	-0.422***
			(0.01)	(0.016)	(0.01)	(0.016)	(0.01)	(0.016)

Table 4. Competing Risks Cox Regression Analyses Results (2/3)

	Model 1		Model 2		Model 3		Model 4	
	Prepay	**Default**	**Prepay**	**Default**	**Prepay**	**Default**	**Prepay**	**Default**
Loan amount			0.306***	0.224***	0.418***	0.185***	0.389***	0.170***
			(0.005)	(0.008)	(0.005)	(0.009)	(0.005)	(0.01)
LTV			-0.181***	0.575***	-0.183***	0.524***	-0.172***	0.641***
			(0.005)	(0.012)	(0.005)	(0.012)	(0.005)	(0.012)
DTI			-0.093***	0.189***	-0.027***	0.141***	-0.022***	0.128***
			(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.008)
Credit score			0.279***	-0.493***	0.242***	-0.493***	0.205***	-0.201***
			(0.005)	(0.008)	(0.006)	(0.008)	(0.006)	(0.008)
Interest rate (contract)			0.583***	0.270***	-0.192***	0.722***		
			(0.006)	(0.012)	(0.022)	(0.032)		
30-59 days delinquency							-0.161***	0.184***
							(0.008)	(0.002)
60-89 days delinquency							-0.142***	0.087***
							(0.016)	(0.001)
90-119 days delinquency							-0.043***	0.161***
							(0.016)	(0.001)
Unemployment rate							-0.028***	-0.035***
							(0.01)	(0.013)
House price index							-0.026**	-0.025
							(0.012)	(0.017)

Table 4 – Competing Risks Cox Regression Analyses Results (3/3)

	Model 1		Model 2		Model 3		Model 4	
	Prepay	**Default**	**Prepay**	**Default**	**Prepay**	**Default**	**Prepay**	**Default**
Regional prepayment rate							0.177***	-0.071***
							(0.002)	(0.01)
Regional default rate							-0.217***	0.109***
							(0.01)	(0.006)
Season: Winter							-0.040***	0.228***
							(0.014)	(0.021)
Season: Fall							-0.008	0.218***
							(0.013)	(0.021)
Season: Summer							0.090***	0.175***
							(0.013)	(0.021)
BEA region	No	No	No	No	Yes	Yes	Yes	Yes
Vintage	No	No	No	No	Yes	Yes	Yes	Yes
Observations	6,786,767	6,786,767	6,786,767	6,786,767	6,786,767	6,786,767	6,786,767	6,786,767
Likelihood ratio test	17339	16432	37383	30891	50846	33407	57159	73723
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AIC	1,104,605	465,523	1,084,586	451,090	1,071,171	448,622	1,064,876	408,324
BIC	1,104,622	465,539	1,084,717	451,208	1,071,511	448,929	1,065,294	408,702

* p < 0.1 ; ** p < 0.05 ; *** p < 0.01

Table 5. Financial crisis and mortgage termination risks

	Model 1	
	Prepayment	**Default**
Refinancing incentive	0.510***	0.589***
	(0.004)	(0.006)
Negative equity	-0.190***	0.187***
	(0.002)	(0.002)
Dummy variable for after crisis	-0.547***	-1.114***
	(0.011)	(0.020)
Observations	6,786,767	6,786,767
Likelihood ratio test	20173	20108
p-value	0.000	0.000
AIC	1,101,773	461850
BIC	1,101,799	461873

termination risks and mortgage approval decisions. LTV and DTI have a positive correlation with default risk while they are negatively correlated with prepayment risk. On the other hand, the higher the credit score, a measure of financial creditworthiness and payment ability of borrowers, the higher the probability of making prepayments and the lower probability of default. These findings are consistent with the literature (Mamonov & Benbunan-Fich, 2017).

Mortgage prepayment habits of the borrowers provide valuable insight into probability of mortgage termination risks (Agarwal et al., 2015; Sirignano et al., 2016). Mortgage defaults are particularly concentrated among borrowers with mortgage delinquencies in their credit history (Ahlawat, 2019; Pennington-Cross, 2010). As the number of delinquencies increases, default risk is also getting higher while prepayment risk falls (Table 4). Crucial importance of monitoring mortgage payment behaviors to offer loss mitigation exercises aiming to prevent default risk as much as possible is once again proven with these results.

Negative correlation between house prices and mortgage risks are consistent with theoretical anticipations. Declines in house prices might increase negative equity and therefore mortgage defaults, and also trigger willingness to move more luxurious or larger houses which results in an increase in prepayment rates. However, unemployment rate is found negatively correlated with default risk, which is against the theory. Similar results are also found in the literature (e.g. Danis and Pennington-Cross (2008)). Unemployment rates at federal states level cannot provide an exact intersection between negative equity and unemployment in a household. Therefore, without current employment status of borrowers, analyses would be misleading about the relationship between payment capability and default risk.

ROC curves and AUC values for prepayment and default models applied on testing data are provided in Figure 5. AUC values for prepayments increase from 0.61 to 0.79, and defaults from 0.64 to 0.75. Refinancing incentive and negative equity variables alone are able to explain mortgage risks under the level of 65% (61% for prepayments and 64% for defaults). Adding loan and borrower characteristics into the modelling increased AUC values significantly, this implies the heterogeneity in borrower behaviors and importance of loan age in explaining mortgage termination risks. Mortgage delinquencies and economic factors are indeed among the determinants of prepayment and default decisions as is seen with rising AUC values.

Many new regulations and amendments in legislative framework on mortgage markets from mortgage underwriting standards to securitization works were made after the financial crisis started in 2007. Therefore, Competing Risks Cox Regression Analyses are performed by adding a dummy variable representing the period after the crisis, and Model 1 is shown in Table 5. Both prepayment and default rates are significantly getting lower compared to the previous years but this effect is more obvious for the default rates. These findings might be interpreted as the fact that adoption of more stringent standards in mortgage underwriting process for particularly GSE loans, and transition to a more stable period reduce the uncertainties in the mortgage markets which has led to lower mortgage risks.

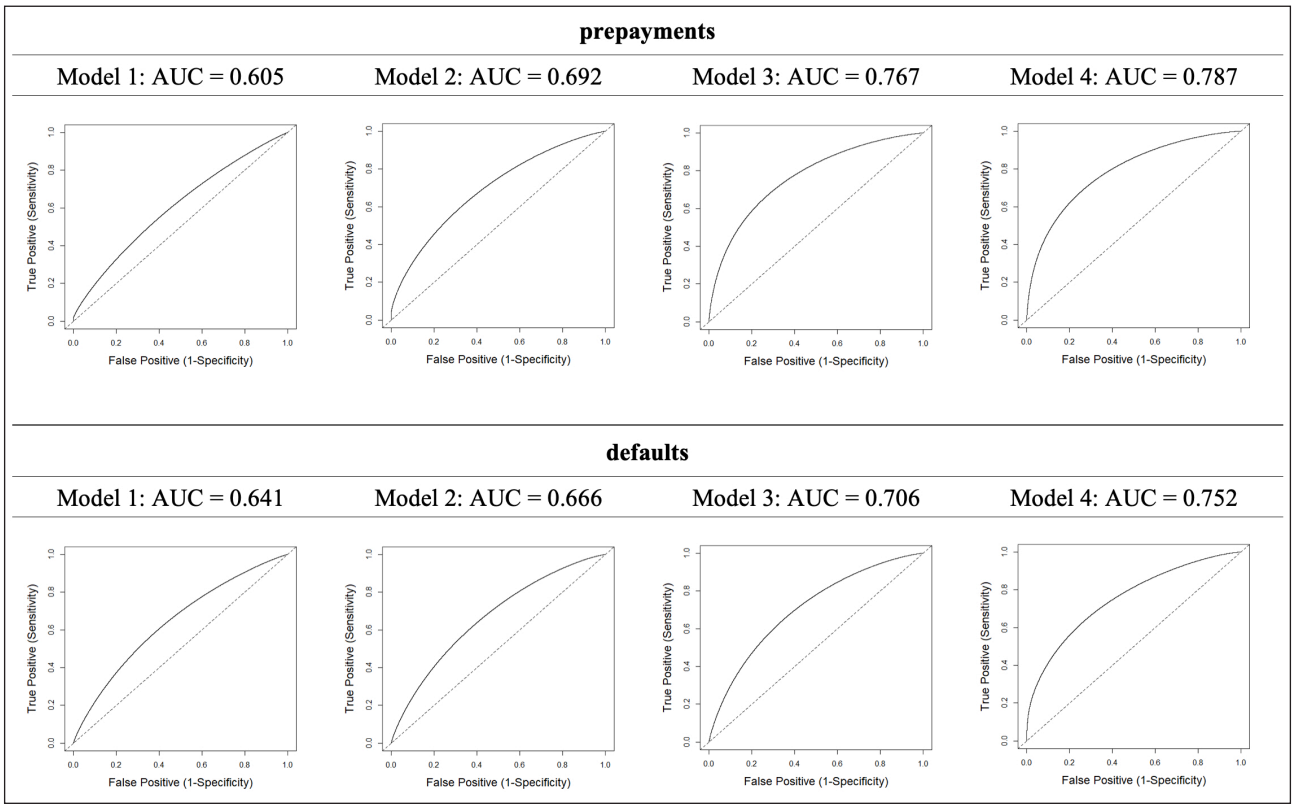

Figure 5. ROC curves and AUC values for prepayment and default models

CONCLUSION

Mortgage-backed securities (MBS) are produced through securitization, and mortgage termination risks are transferred from loan originators to MBS investors. Value of a fixed income security equals to the present value of its expected cash flows, however, valuation becomes a complex problem in the case of MBS due mortgage risks in their collateral pools because prepayment and default risks lead to uncertainty in the cash flows of MBS. Therefore, the mortgage risks become the core of valuation of MBS collateral pools.

There are both systematic and idiosyncratic factors behind the mortgage risks. Option-based models adopt the systematic risks view and focus on prevailing mortgage interest rates and housing prices. These models assume that borrowers behave in an optimal way while taking prepayment and default decisions, and try to explain prepayment speeds with refinancing incentive and default rates with negative equity. On the other hand, econometric models emphasize that not all borrowers take optimal decisions. In addition to refinancing incentive and negative equity, heterogeneity in borrower behaviors, loan attributes and local economic factors are suggested to be considered while predicting prepayment and default rates. This study employs econometric modelling view. After performing classification studies with machine learning algorithms (Random Forest and Multinomial Logistic Regression) and Competing Risks Cox Regression analyses to explain the prepayment and default rates, the study finds that refinancing incentive and negative equity variables alone are not sufficiently explain prepayment and default risks. Mortgage and borrower features (e.g. LTV and DTI ratios, loan amount, credit score of borrowers) and economic factors (e.g. house prices, unemployment rates, local prepayment and default rates) are significantly important indicators. Furthermore, borrowers' payment history of their current loans provides an import insight into whether they will make a default decision in the future because mortgage delinquencies are found important in predicting mortgage defaults. Therefore, role of servicers in monitoring payments closely to offer loss mitigation tools to potential defaulters is crucial to keep safety of both borrowers and financial markets.

REFERENCES

Agarwal , S., Ambrose, B. W., & Yildirim, Y. (2015). The subprime virus. Real Estate Economics, 43(4), 891-915.

Ahlawat, S. (2019). Evaluation of Mortgage Default Characteristics Using Fannie Mae's Loan Performance Data. The Journal of Real Estate Finance and Economics, 59(4), 589-616.

Alpaydin, E. (2020). Introduction to machine learning: MIT press.

An, X., Deng, Y., & Gabriel, S. A. (2021). Default option exercise over the financial crisis and beyond. Review of Finance, 25(1), 153-187.

Barbaglia, L., Manzan, S., & Tosetti, E. (2023). Forecasting loan default in Europe with machine learning. Journal of Financial Econometrics, 21(2), 569-596.

Bennett, P., Peach, R., & Peristiani, S. (2001). How much mortgage pool information do investors need? The Journal of Fixed Income, 11(1), 8-15.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

Berliner, B., Quinones, A., & Bhattacharya, A. (2016). Mortgage Loans to Mortgage-Backed Securities. In F. J. Fabozzi (Ed.), The Handbook of Mortgage-Backed Securities (Seventh Edition ed., pp. 3-29). Oxford, United Kingdom: Oxford Univeristy Press.

Berrar, D. (2018). Cross Validation. In S. Ranganathan, K. Nakai, & C. Schonbach (Eds.), Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics (Vol. 1, pp. 542-545): Elsevier.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. Journal of political economy, 81(3), 637-654.

Blumenstock, G., Lessmann, S., & Seow, H.-V. (2022). Deep learning for survival and competing risk modelling. Journal of the Operational Research Society, 73(1), 26-38.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Chen, J., Xiang, J., & Yang, T. T. (2018). Re-Default Risk of Modified Mortgages. International Real Estate Review, 21(1).

Cooper, M. J. (2018). A Deep Learning PredictionModel for Mortgage Default. Master of Engineering Thesis, University of Bristol, England.

Cowden, C., Fabozzi, F. J., & Nazemi, A. (2019). Default prediction of commercial real estate properties using machine learning techniques. The Journal of Portfolio Management, 45(7), 55-67.

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-202.

Danis, M. A., & Pennington-Cross, A. (2008). The delinquency of subprime mortgages. Journal of Economics and Business, 60(1-2), 67-90.

Davidson, A. S., Herskovitz, M. D., & Van Drunen, L. D. (1988). The refinancing threshold pricing model: An economic approach to valuing MBS. The Journal of Real Estate Finance and Economics, 1(2), 117-130.

Davis, R., Lo, A. W., Mishra, S., Nourian, A., Singh, M., Wu, N., & Zhang, R. (2022). Explainable machine learning models of consumer credit risk. Available at SSRN 4006840.

Demiroglu, C., Dudley, E., & James, C. M. (2014). State foreclosure laws and the incidence of mortgage default. The Journal of Law and Economics, 57(1), 225-280.

Demyanyk, Y. (2017). The impact of missed payments and foreclosures on credit scores. The Quarterly Review of Economics and Finance, 64, 108-119.

Deng, Y., Pavlov, A. D., & Yang, L. (2005). Spatial heterogeneity in mortgage terminations by refinance, sale and default. Real Estate Economics, 33(4), 739-764.

Downing, C., Stanton, R., & Wallace, N. (2005). An empirical test of a two-factor mortgage valuation model: how much do house prices matter? Real Estate Economics, 33(4), 681-710.

Drummond, C., & Holte, R. C. (2003). Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. Paper presented at the Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC.

Dunn, K. B., & McConnell, J. J. (1981). Valuation of GNMA mortgage-backed securities. The Journal of Finance, 36(3), 599-616.

Fabozzi, F. J., Bhattacharya, A. K., & Berliner, W. S. (2007). Mortgage-Backed Securities: Products, Structuring, and Analytical Techniques. Hoboken (Vol. 200): John Wiley & Sons.

Fannie Mae. (2019). Fannie Mae Single-Family Loan Performance Data, USA. Retrieved from https://capitalmarkets.fanniemae.com/credit-risk-transfer/single-family-credit-risk-transfer/fannie-mae-single-family-loan-performance-data

FHFA. (2021). House Price Index Datasets, USA. Retrieved from https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx#qpo

Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. Journal of the American statistical association, 94(446), 496-509.

Foote, C. L., & Willen, P. S. (2018). Mortgage-default research and the recent foreclosure crisis. Annual Review of Financial Economics, 10, 59-100.

Fout, H., Li, G., Palim, M., & Pan, Y. (2020). Credit risk of low income mortgages. Regional Science and Urban Economics, 80, 103390.

Freddie Mac. (2020). Mortgage Rates - Historical Data, . Retrieved from http://www.freddiemac.com/pmms/pmms_archives.html

Freddie Mac. (2021). Quarterly Refinance Statistics Archive, USA. Retrieved from http://www.freddiemac.com/research/datasets/refinance-stats/archive.page#archive

Gerardi, K., Herkenhoff, K. F., Ohanian, L. E., & Willen, P. S. (2018). Can't pay or won't pay? unemployment, negative equity, and strategic default. The Review of Financial Studies, 31(3), 1098-1131.

Groot, J. d. (2016). Credit risk modeling using a weighted support vector machine, Master Thesis, Universiteit Utrecht.

Hayre, L., & Young, R. (2004). Guide to mortgage-backed securities. Citigroup White Paper.

Hertzmann, A., & Fleet, D. (2012). Machine Learning And Data Mining Lecture Notes. Computer Science Department, University of Toronto.

Huh, Y., & Kim, Y. S. (2019). The Real Effects of Secondary Market Trading Structure: Evidence from the Mortgage Market. Available at SSRN 3373949.

Johnston, E., & Van Drunen, L. (1988). Pricing mortgage pools with heterogeneous mortgagors: Empirical evidence. Unpublished manuscript, University of Utah.

Kalbfleisch, J. D., & Prentice, R. L. (2011). The statistical analysis of failure time data (Vol. 360): John Wiley & Sons.

Kalotay, A., Yang, D., & Fabozzi, F. J. (2004). An option-theoretic prepayment model for mortgages and mortgage-backed securities. International Journal of Theoretical and Applied Finance, 7(08), 949-978.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.

Kau, J. B., Keenan, D. C., & Li, X. (2011). An analysis of mortgage termination risks: a shared frailty approach with MSA-level random effects. The Journal of Real Estate Finance and Economics, 42(1), 51-67.

Kau, J. B., Keenan, D. C., & Smurov, A. A. (2006). Reduced form mortgage pricing as an alternative to option-pricing models. The Journal of Real Estate Finance and Economics, 33(3), 183-196.

Keys, B. J., Pope, D. G., & Pope, J. C. (2016). Failure to refinance. Journal of Financial Economics, 122(3), 482-499.

Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. The Journal of Portfolio Management, 43(6), 202-211.

LaCour-Little, M. (2008). Mortgage termination risk: a review of the recent literature. Journal of Real Estate Literature, 16(3), 295-326.

Link, W. A. (1989). A model for informative censoring. Journal of the American statistical association, 84(407), 749-752.

López, A. L., López, E., & Ponce, H. (2022). Credit Risk Models in the Mexican Context Using Machine Learning. Paper presented at the Mexican International Conference on Artificial Intelligence.

Lowell, L., & Corsi, M. (2006). Mortgage Pass-Through Securities. In F. J. Fabozzi (Ed.), The Handbook of

Mortgage-Backed Securities (Sixth Edition ed., pp. 45-79). US: McGraw-Hill.

Mamonov, S., & Benbunan-Fich, R. (2017). What can we learn from past mistakes? Lessons from data mining the Fannie Mae mortgage portfolio. Journal of Real Estate Research, 39(2), 235-262.

McConnell, J. J., & Buser, S. A. (2011). The origins and evolution of the market for mortgage-backed securities. Annu. Rev. Financ. Econ., 3(1), 173-192.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. The Journal of Finance, 29(2), 449-470.

Patrabansh, S. (2015). The Marginal Effect of First-Time Homebuyer Status on Mortgage Default and Prepayment, FHFA Working Paper 15-2, USA.

Pennington-Cross, A. (2010). The duration of foreclosures in the subprime mortgage market: a competing risks model with mixing. The Journal of Real Estate Finance and Economics, 40(2), 109-129.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., & Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. Biometrics, 541-554.

Quigley, J. M., & Van Order, R. (1991). Defaults on mortgage obligations and capital requirements for US savings institutions: A policy perspective. Journal of Public Economics, 44(3), 353-369.

Rajashri, P. J., Davis, T., & McCoy, B. (2016). Valuation of Mortgage-Backed Securities. In F. J. Fabozzi (Ed.), The Handbook of Mortgage-Backed Securities: 7th Edition: Oxford University Press.

Richard, S. F., & Roll, R. (1989). Prepayments of fixed-rate mortgage-backed securities. Journal of Portfolio Management, 15(3), 73.

Schelkle, T. (2018). Mortgage default during the US mortgage crisis. Journal of money, credit and banking, 50(6), 1101-1137.

Schwartz, E. S., & Torous, W. N. (1989). Prepayment and the valuation of mortgage-backed securities. The Journal of Finance, 44(2), 375-392.

Sirignano, J., Sadhwani, A., & Giesecke, K. (2016). Deep learning for mortgage risk. arXiv preprint arXiv:1607.02470.

Spahr, R. W., & Sunderman, M. A. (1992). The effect of prepayment modeling in pricing mortgage-backed securities. Journal of housing research, 381-400.

Timmis, G. (1985). Valuation of GNMA mortgage-backed securities with transaction costs, heterogeneous households and endogenously generated prepayment rates. Carnegie-Mellon University.

Weiner, J. (2016). Modeling Prepayments and Defaults for MBS Valuation. In F. J. Fabozzi (Ed.), The Handbook of Mortgage-Backed Securities (Seventh Edition ed., pp. 531-559). Oxford, United Kingdom: Oxford Univeristy Press.

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. Data Science and Management.