Mortgage Default Prediction with MILR

Yi-Chun Lo

June 1, 2024

1 Introduction

The mortgage market plays a crucial role in the global economy, particularly in the United States, where residential mortgage-backed securities (RMBS) are a significant component of the financial system. Understanding the factors that influence mortgage default is essential for financial institutions, policymakers, and investors. Accurate predictions of mortgage defaults can help in mitigating risks, formulating better lending policies, and enhancing the stability of the financial system.

1.1 Literature Review

Numerous studies have investigated the determinants of mortgage default, employing various methodologies and datasets. Early research by Quercia and Stegman (1992) highlighted the importance of borrower characteristics, loan terms, and economic conditions in predicting mortgage default. Their findings emphasized that higher loan-to-value (LTV) ratios and lower credit scores significantly increase the probability of default. Building on this foundation, Deng et al. (2000) incorporated macroeconomic factors such as interest rates and unemployment rates into their models, demonstrating that these external conditions play a crucial role in mortgage performance. They utilized hazard models to account for the time-varying nature of default risk, which provided a more dynamic understanding of the default process.

Recent advancements in machine learning have further enhanced the predictive accuracy of mortgage default models. Studies by Bajari et al. (2008) and Sirignano et al. (2016)) applied machine learning algorithms such as random forests and neural networks to mortgage datasets, achieving higher predictive power compared to traditional logistic regression models. These studies highlight the potential of machine learning techniques to capture complex, non-linear relationships within the data. However, the challenge of interpreting machine learning models remains a critical issue. To address this, researchers have combined machine learning with more interpretable models. For instance, Lessmann et al. (2015) demonstrated the efficacy of combining logistic regression with tree-based models to balance predictive accuracy and interpretability.

1.1.1 Multiple Instance Learning (MIL)

Multiple Instance Learning (MIL) is a type of supervised machine learning that deals with incomplete and uncertain label information (Bilal et al., 2022). MIL was first proposed by Dietterich et al. (1997). For demonstration, we introduce the MIL using the famous keychain problem. Suppose that, in one company, there are n staff, each with one key chain. For each keychain, there are m_i keys. The company manager knows which staff can open the room but does not know which key can open the room. In this case, there are n keychains (bags) each with m_i keys (instances), and $\boldsymbol{x}_{ij} \in \mathbb{R}^d$ is a d-dimensional predictor vector (width, length, and so on) used to describe the jth key in the ith keychain. Let Y_{ij} denote the label of the jth key of the ith keychain. Thus, if the jth key of the ith keychain can open the room, then $Y_{ij} = 1$, and otherwise, $Y_{ij} = 0$. This label can be predicted using $\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}, \ldots, \boldsymbol{x}_{1m_i}$ with a function f, say $Y_{ij} \sim Bernoulli(f(\mathbf{x}_{ij}))$. Moreover, considering that a keychain can open the room if at least one key can, the key label can be formulated as follows:

$$Z_i = \max_{j=1,\ldots,m_i} Y_{ij} = \mathrm{I}\left(\sum_{j=1}^{m_i} Y_{ij} > 0\right)$$

Dietterich et al. (1997) proposed the Axis-Parallel Rectangles (APR) method to solve the MIL problem, which can be seen as a decision tree where all leaves are labeled zero except for one leaf labeled one. Following this, subsequent approaches to MIL have developed more sophisticated methods. For example, Maron and Lozano-Pérez (1997) introduced the Diverse Density (DD) approach, which derives a basic profile of an individual from a sequence of images (bags) featuring that person. The DD approach assumes a different d-dimensional normal distribution for the instances' predictor vectors based on the bag's label. To handle large numbers of instances, Zhang and Goldman (2001) proposed the EM-DD approach, which combines Expectation-Maximization with Diverse Density. Additionally, Chen et al. (2016) introduced Multiple Instance Logistic Regression (MILR), applying logistic regression to multiple instance data in manufacturing. They utilized the Expectation Maximization (EM) algorithm for stable parameter estimation, demonstrating superior performance compared to naive approaches.

1.2 Research Objective

This project seeks to identify the key determinants of mortgage default by analyzing a comprehensive panel dataset using advanced statistical and machine learning techniques. Given the inherent complexities of mortgage data, including issues of loan origination before the observation period and potential censoring due to loan maturation or refinancing, we employ Multiple Instance Logistic Regression (MILR) to build predictive models.

The MILR model is particularly well-suited for this scenario due to the structure of our data. In the context of MIL, each borrower is considered a "bag" and each loan observation within a borrower's history is considered an "instance". Our goal is to classify whether a borrower will default based on their loan instances. This approach effectively manages grouped data, allowing us to capture the relationship between individual loan observations and overall borrower default risk. By leveraging the MILR model, we aim to improve the accuracy and interpretability of our predictions, providing valuable insights into the factors influencing mortgage defaults.

2 Data

The data set used in this study, provided by Baesens et al. (2016), is a panel dataset reporting origination and performance observations for 50,000 residential U.S. mortgage borrowers over 60 periods. The periods have been de-identified, reflecting real-world conditions where loans may originate before the start of the observation period and may be censored as they mature or are refinanced.

Table 1: Description of Features in the Mortgage Dataset

Feature	**Description**
`id`	Borrower ID
`time`	Time stamp of observation
`orig_time`	Time stamp for origination

first_time	Time stamp for first observation
mat_time	Time stamp for maturity
balance_time	Outstanding balance at observation time
LTV_time	Loan-to-value ratio at observation time, in %
interest_rate_time	Interest rate at observation time, in %
hpi_time	House price index at observation time, base year = 100
gdp_time	Gross domestic product (GDP) growth at observation time, in %
uer_time	Unemployment rate at observation time, in %
REtype_CO_orig_time	Real estate type condominium = 1, otherwise = 0
REtype_PU_orig_time	Real estate type planned urban development = 1, otherwise = 0
REtype_SF_orig_time	Single-family home = 1, otherwise = 0
investor_orig_time	Investor borrower = 1, otherwise = 0
balance_orig_time	Outstanding balance at origination time
FICO_orig_time	FICO score at origination time, in %
LTV_orig_time	Loan-to-value ratio at origination time, in %
Interest_Rate_orig_time	Interest rate at origination time, in %
hpi_orig_time	House price index at origination time, base year = 100
default_time	Default observation at observation time
payoff_time	Payoff observation at observation time
status_time	Default (1), payoff (2), and nondefault/nonpayoff (0) observation at observation time

This dataset represents a randomized selection of mortgage-loan-level data collected from the portfolios underlying U.S. residential mortgage-backed securities (RMBS) securitization portfolios (see Table 1).

2.1 Data Preprocessing

To prepare the data for analysis, several preprocessing steps were performed:

1. **Handling Missing Values**: Missing numeric values were replaced with the median of the respective column to ensure completeness of the dataset without introducing bias.

2. **Feature Engineering**: New features were created to capture the age of the loan, the time remaining to maturity, and the time since the first observation. These features help to understand the temporal aspects of the loans better (see Table 2).

3. **Label Creation**: A `default label` feature was created to indicate if a borrower had defaulted at any point in time, providing a clear target variable for prediction. The value 1 indicates a default, and 0 indicates no default.

4. **Data Cleaning**: Unnecessary columns, such as the original timestamps (`time`, `orig_time`, `first_time`, `mat_time`) and status indicators (`default_time`, `payoff_time`, `status_time`) were removed to streamline the dataset.

5. **Standardization**: Numeric features were standardized to have zero mean and unit variance. This step was crucial to ensure that all features contribute equally to the model training process.

Table 2: Newly Engineered Features

Feature	Description
`loan_age`	The age of the loan, calculated as the difference between the observation time and the origination time
`time_to_maturity`	The remaining time to loan maturity, calculated as the difference between the maturity time and the observation time
`time_since_first_obs`	The time since the first observation, calculated as the difference between the observation time and the first observation time

2.2 Data Exploration and Visualization

To gain a comprehensive understanding of the dataset, various data exploration and visualization techniques were employed.

2.2.1 Count Plot

The distribution of the default label was examined using a count plot (Figure 1). This visualization highlights a significant imbalance between defaulted and non-defaulted loans, with the majority of loans being non-defaulted.

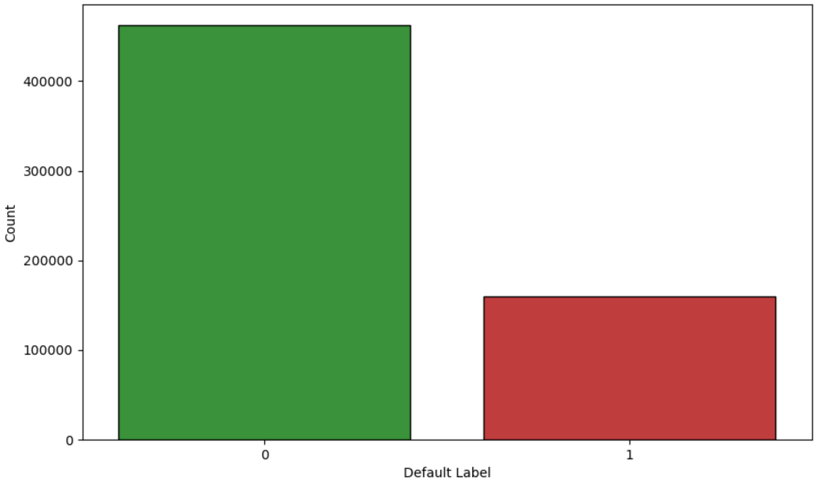

Figure 1: Count of Default Label

2.2.2 Histograms

Histograms were utilized to explore the distribution of key features. For instance, the histogram of loan age (Figure 2) reveals that most loans are relatively new.

Another important feature to consider is the time remaining to loan maturity. The histogram of time to maturity (Figure 3) shows that most loans have a time to maturity concentrated around 100 to 120 periods. This concentration suggests that the majority of loans are long-term loans, which might impact the risk of default.

Furthermore, the histogram of time since the first observation (Figure 4) shows a declining trend, with the highest frequency at the beginning and gradually decreasing over time. This suggests that a large

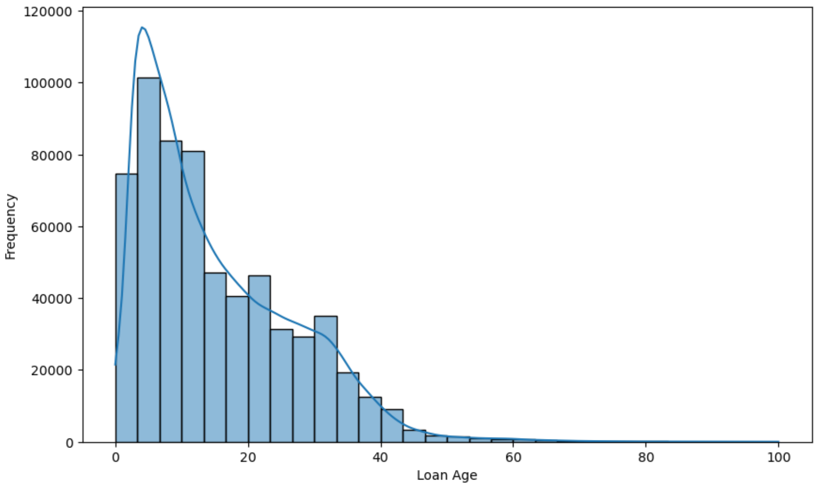

Figure 2: Distribution of Loan Age

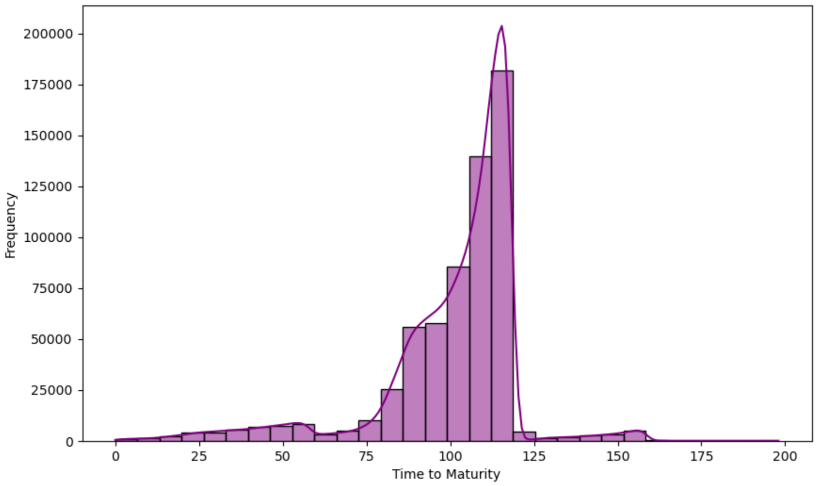

Figure 3: Distribution of Time to Maturity

number of observations are made shortly after the origination of the loan, and fewer observations are made as time progresses.

Finally, the histogram of LTV time (Figure 5) shows that most loans have a loan-to-value ratio concentrated around 75 to 100 percent. This suggests that many borrowers have relatively high loan-to-value ratios, which could impact their risk of default. Higher LTV ratios imply that borrowers have less equity in their homes, which might increase the likelihood of default, especially in times of economic stress.

2.2.3 Correlation Analysis

Table 3 reveals the relationships between various features and the default label. Key observations include:

- `time_to_maturity` has a positive correlation with the default label (0.2203), suggesting that loans with longer times to maturity are more likely to default.
- `LTV_time` (0.1916) and `interest_rate_time` (0.1811) also show positive correlations, indicating

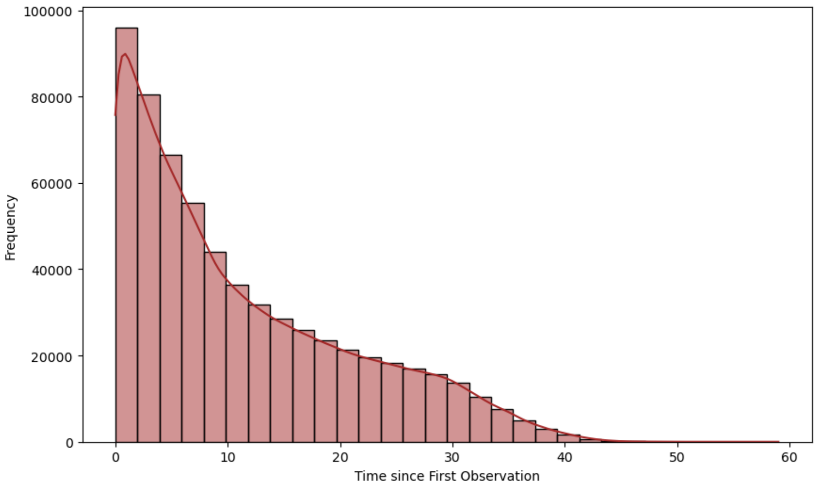

Figure 4: Distribution of Time since First Observation

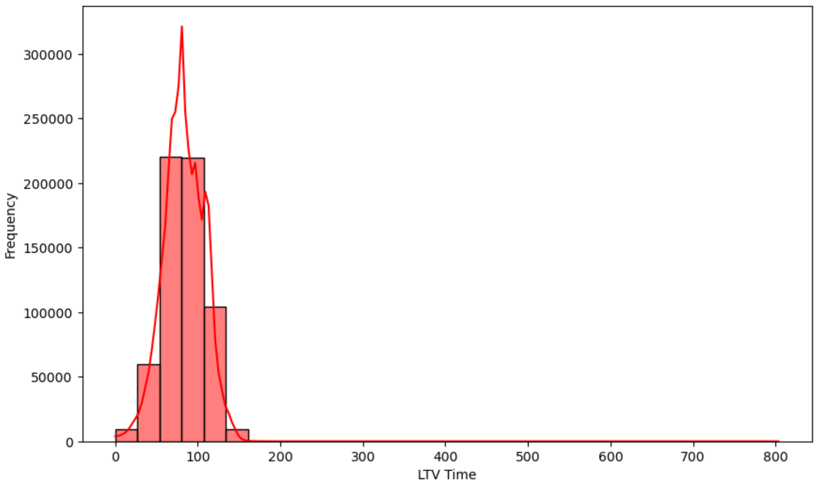

Figure 5: Distribution of LTV Time

that higher loan-to-value ratios and higher interest rates at the time of observation are associated with higher default risks.

- Conversely, `loan_age` (-0.2415) shows a negative correlation with the default label, suggesting that older loans are less likely to default, likely due to the accumulation of borrower equity over time.
- `FICO_orig_time` (-0.1303) and `gdp_time` (-0.1377) also show negative correlations, indicating that higher credit scores at origination and better economic conditions (GDP growth) are associated with lower default risks.

3 Method

3.1 Multiple Instance Logistic Regression (MILR)

Suppose the dataset consists of n bag binary responses, denoted as $Z_1, \ldots, Z_n$, where the ith bag contains m_i instances. For the jth instance of the ith bag, the instance label Y_{ij} follows a Bernoulli

Table 3: Correlation with Default Label

Feature	Correlation
default_label	1.0000
time_to_maturity	0.2203
LTV_time	0.1916
hpi_orig_time	0.1863
interest_rate_time	0.1811
LTV_orig_time	0.0855
hpi_time	0.0501
Interest_Rate_orig_time	0.0214
balance_time	0.0190
REtype_CO_orig_time	0.0108
REtype_PU_orig_time	0.0067
REtype_SF_orig_time	0.0002
balance_orig_time	-0.0062
investor_orig_time	-0.0110
uer_time	-0.0759
FICO_orig_time	-0.1303
gdp_time	-0.1377
time_since_first_obs	-0.2202
loan_age	-0.2415

distribution with the logistic regression form

$$Y_{ij} \sim Bernoulli(p_{ij}) \quad \text{where} \quad p_{ij} = \frac{e^{\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}}}$$

In the above equation, the parameter vector $\boldsymbol{\beta}$ corresponds to the model coefficients. Nevertheless, the labels of instances, Y_{ij}'s, are unobservable. Instead, we only observe the bag labels, which follow Bernoulli distributions denoted as

$$Z_i \sim Bernoulli(\pi_i) \quad \text{where} \quad \pi_i = 1 - \prod_{j=1}^{m_i}(1-p_{ij}). \tag{1}$$

The model parameter $\hat{\boldsymbol{\beta}}$ can be recalculated by the Expectation-Maximization (EM) algorithm. The EM algorithm iteratively estimates the parameters by maximizing the expected log-likelihood of the data. The steps are as follows:

E-Step: Expectation

The E-step of the EM algorithm involves calculating the expected value of the complete-data log-likelihood, given the observed data and the current estimates of the parameters. The complete data likelihood function for our MILR model is

$$L_c(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{j=1}^{m_i} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}.$$

The corresponding complete data log-likelihood is

$$\log L_c(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij}\log(p_{ij}) + (1-y_{ij})\log(1-p_{ij}).$$

To proceed with the E-step, we need to calculate the conditional expectation of the instance labels Y_{ij} given the observed bag labels Z_i and the current parameter estimates $\hat{\boldsymbol{\beta}}^{(t)}$. Since the instance labels Y_{ij} are unobserved, we need to compute their conditional expectation given the observed bag labels Z_i and the current parameter estimates $\hat{\boldsymbol{\beta}}^{(t)}$. This requires us to consider two cases based on the bag label Z_i:

1. **Case 1** ($Z_i = 0$)

   When the bag label $Z_i = 0$, it implies that all instances label are 0. Therefore, the conditional expectation of each instance label Y_{ij} is:

   $$\mathrm{E}(Y_{ij} \mid Z_i = 0, \hat{\boldsymbol{\beta}}^{(t)}) = 0.$$

2. **Case 2** ($Z_i = 1$)

   When the bag label $Z_i = 1$, it implies that at least one instance in the bag is positive. The conditional distribution of Y_{ij} given $Z_i = 1$ is used to compute the expectation:

   $$\mathrm{E}(Y_{ij} \mid Z_i = 1, \hat{\boldsymbol{\beta}}^{(t)}) = \frac{\hat{p}_{ij}^{(t)}}{1 - \prod_{k=1}^{m_i}(1 - \hat{p}_{ik}^{(t)})}$$

   We denote this conditional expectation as $\hat{\theta}_{ij}^{(t)}$:

   $$\hat{\theta}_{ij}^{(t)} = \frac{\hat{p}_{ij}^{(t)}}{1 - \prod_{k=1}^{m_i}(1 - \hat{p}_{ik}^{(t)})}.$$

Using these conditional expectations, we can now express the Q-function, which is the expected complete-data log-likelihood, as:

$$\mathrm{Q}\left(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(t)}\right) = \mathrm{E}\left(\log L_c(\boldsymbol{\beta}) \mid Z, \hat{\boldsymbol{\beta}}^{(t)}\right)$$

For each bag i, this becomes:

$$\mathrm{Q}_i(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(t)}) = \sum_{j=1}^{m_i} \hat{\theta}_{ij}^{(t)} \log(\hat{p}_{ij}) + \left(1 - \hat{\theta}_{ij}^{(t)}\right) \log(1 - \hat{p}_{ij})$$

M-Step: Maximization

In the M-step, we maximize the Q-function with respect to $\boldsymbol{\beta}$ to obtain the new parameter estimates $\hat{\boldsymbol{\beta}}^{(t+1)}$. The maximization step involves solving the following equation:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \mathrm{Q}\left(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(t)}\right)$$

Since the Q-function is nonlinear in $\boldsymbol{\beta}$, to simplify the maximization, we can apply a quadratic approximation to the Q-function. Friedman et al. (2010) suggest to take a Taylor expansion around the current estimate $\hat{\boldsymbol{\beta}}^{(t)}$:

$$\begin{aligned}\mathrm{Q}\left(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(t)}\right) &\approx -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \tau_{ij}^{(t)} \left(\phi_{ij}^{(t)} - \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}\right)^2 + \mathrm{R}\left(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(t)}\right) + \mathrm{C} \\ &\equiv \mathrm{Q}_2(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(t)}) + \mathrm{R}\left(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(t)}\right) + \mathrm{C}\end{aligned}$$

where C is a constant, $\mathrm{R}\,(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(t)})$ is the remainder term and

$$\phi_{ij}^{(t)} = p_{ij}^{(t)}(1 - p_{ij}^{(t)})$$

$$\tau_{ij}^{(t)} = \boldsymbol{x}_{ij}^{\top}\hat{\boldsymbol{\beta}}^{(t)} + \frac{\hat{\theta}_{ij}^{(t)} - p_{ij}^{(t)}}{\phi_{ij}^{(t)}}$$

Using this quadratic approximation, it becomes a weighted least squares problem. The coordinate descent algorithm (Friedman et al., 2010) updates each parameter $\boldsymbol{\beta}_d$ (where d indexes the components of the parameter vector $\boldsymbol{\beta}$) by holding the other parameters fixed and optimizing Q_2 with respect to $\boldsymbol{\beta}_d$. The update for $\hat{\boldsymbol{\beta}}_d$ at iteration $t+1$ is given by:

$$\hat{\boldsymbol{\beta}}_d^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\phi_{ij}^{(t)}\boldsymbol{x}_{ijd}\left(\tau_{ij}^{(t)} - \sum_{k\neq d}\boldsymbol{x}_{ijd}\hat{\boldsymbol{\beta}}_k^{(t)}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\phi_{ij}^{(t)}\boldsymbol{x}_{ijd}^2}$$

where $\boldsymbol{x}_{ijd}$ is the d-th element of $\boldsymbol{x}_{ij}$. In the context of predicting mortgage defaults, the MILR model is particularly well-suited because each borrower can be considered a bag, and each loan observation within a borrower's history can be considered an instance. Table 4 shows how each mathematical symbol relates to our scenario. By applying the MILR model, we can leverage the EM algorithm to iteratively estimate the model coefficients, thus providing a powerful tool to predict mortgage defaults while accounting for the inherent multiple instance structure of the data.

Table 4: Mathematical Symbols and Their Descriptions in Mortgage Default Prediction

Symbol	**Description**
n	The number of borrowers (bags) in the dataset.
m_i	The number of loan observations (instances) for the ith borrower.
$\boldsymbol{x}_{ij}$	The feature vector for the jth loan observation of the ith borrower. This includes attributes such as loan-to-value ratio, interest rate, balance, etc.
$\boldsymbol{\beta}$	The model coefficients that need to be estimated.
Y_{ij}	The unobserved instance label indicating whether the jth loan observation of the ith borrower is defaulted or not.
p_{ij}	The probability that the jth loan observation of the ith borrower is defaulted, given by the logistic function.
Z_i	The observed bag label indicating whether the ith borrower has defaulted on any loan at any point in time.
π_i	The probability that the ith borrower defaults on at least one loan, calculated from the instance probabilities.

3.2 Model Training and Evaluation

To train and evaluate the Multiple Instance Logistic Regression (MILR) model, the dataset was split into training and testing sets. Initially, the unique borrower IDs were partitioned into training and testing sets with a 70-30 split. This partitioning ensured that 70% of the borrowers were included in the training set, while the remaining 30% were allocated to the testing set. Once the model was trained, it was used to predict the default labels for the testing set. The predictions were made at the borrower (bag) level to align with the multiple instance learning framework.

To evaluate the performance of the model, a confusion matrix was generated from the predictions. Based on this confusion matrix, several metrics were calculated to assess the model effectiveness:

- **Accuracy**: The proportion of correctly predicted cases (both true positives and true negatives) out of the total cases.
- **Precision**: The proportion of true positive predictions out of all positive predictions, reflecting the model ability to avoid false positives.
- **Sensitivity (Recall)**: The proportion of true positive predictions out of all actual positives, indicating the model ability to identify all relevant cases.
- **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two metrics and offering a single measure of the model performance.

These metrics provided a comprehensive assessment of the model ability to predict mortgage defaults, highlighting its accuracy and effectiveness in identifying borrowers at risk of default.

4 Empirical Results

4.1 Model Performance

The performance of the MILR model was assessed using a confusion matrix, from which several key metrics were derived and are presented in Table 5.

Table 5: MILR Model Performance Metrics

Metric	**Value**
Accuracy	0.7253
Precision	0.5701
Sensitivity (Recall)	0.3613
F1 Score	0.4423

The results indicate that the MILR model achieved moderate accuracy in predicting mortgage defaults. The precision of 0.5701 suggests that over half of the positive predictions were correct, reflecting the ability to avoid false positives of the model. However, the sensitivity (recall) of 0.3613 indicates that the model missed a substantial portion of actual defaults, pointing to a need for improvement in identifying all relevant cases. The F1 Score of 0.4423, a balance between precision and recall, further highlights moderate performance of the model, suggesting that while it has some predictive power, there is significant room for enhancement in capturing mortgage defaults accurately.

4.2 Model Coefficients

The estimated coefficients of the MILR model provide insights into the significance of various features in predicting mortgage defaults. The Log-Likelihood of the model was -18749.918. The estimated coefficients along with their standard errors, Z values, and p-values are presented in Table 6. The significance codes are as follows: *** indicates p-value < 0.001, ** indicates p-value < 0.01, * indicates p-value < 0.05, and . indicates p-value < 0.1.

The coefficients indicate the importance of various features in predicting mortgage defaults:

- `LTV_time`: A positive coefficient (0.3780) and significant p-value (0.0000) suggest that higher loan-to-value ratios at the time of observation are associated with an increased likelihood of default. This aligns with the understanding that borrowers with higher LTV ratios have less equity in their homes, making them more vulnerable to defaulting.

Table 6: MILR Model Coefficients

Feature	**Estimate**	**Std. Error**	**Z Value**	**p-value**
`intercept`	-5.0295	0.0508	-98.9276	< 0.0001 ***
`balance_time`	0.2067	0.1947	1.0618	0.2883
`LTV_time`	0.3780	0.0827	4.5726	0.0000 ***
`interest_rate_time`	0.2635	0.0157	16.7742	< 0.0001 ***
`hpi_time`	-0.0168	0.0409	-0.4096	0.6821
`gdp_time`	-0.4483	0.0316	-14.1987	< 0.0001 ***
`uer_time`	-1.0468	0.0481	-21.7444	< 0.0001 ***
`REtype_CO_orig_time`	-0.1682	0.0475	-3.5407	0.0004 ***
`REtype_PU_orig_time`	-0.1650	0.0382	-4.3221	0.0000 ***
`REtype_SF_orig_time`	-0.1837	0.0263	-6.9754	< 0.0001 ***
`investor_orig_time`	0.1582	0.0327	4.8317	< 0.0001 ***
`balance_orig_time`	-0.1300	0.1984	-0.6555	0.5121
`FICO_orig_time`	-0.2870	0.0127	-22.5643	< 0.0001 ***
`LTV_orig_time`	-0.0029	0.0360	-0.0803	0.9360
`Interest_Rate_orig_time`	-0.0103	0.0098	-1.0495	0.2940
`hpi_orig_time`	0.2562	0.0482	5.3190	0.0000 ***
`loan_age`	0.0582	0.0331	1.7598	0.0784 .
`time_to_maturity`	0.1861	0.0198	9.3766	< 0.0001 ***
`time_since_first_obs`	-1.7005	0.0587	-28.9548	< 0.0001 ***

- `interest_rate_time`: The positive coefficient (0.2635) and highly significant p-value (< 0.0001) indicate that higher interest rates at the time of observation increase the probability of default. Higher interest rates can lead to higher monthly payments, potentially causing financial strain on borrowers.

- `gdp_time`: The negative coefficient (-0.4483) and significant p-value (< 0.0001) suggest that better economic conditions, as indicated by GDP growth, are associated with a lower risk of default. This is consistent with the idea that a stronger economy improves borrowers' financial stability and ability to meet mortgage payments.

- `REtype_CO_orig_time`, `REtype_PU_orig_time`, `REtype_SF_orig_time`: These real estate types have negative coefficients and significant p-values, indicating that borrowers with these property types are less likely to default compared to others. This may be due to the inherent stability and desirability of these property types.

- `investor_orig_time`: The positive coefficient (0.1582) and significant p-value (< 0.0001) indicate that loans originated by investors have a higher likelihood of default, which may reflect the higher risk associated with investment properties compared to owner-occupied properties.

- `FICO_orig_time`: The negative coefficient (-0.2870) and highly significant p-value (< 0.0001) suggest that higher credit scores at origination are associated with a lower likelihood of default. This is consistent with the understanding that borrowers with higher credit scores are generally more creditworthy and less likely to default.

- `time_to_maturity`: The positive coefficient (0.1861) and significant p-value (< 0.0001) imply that loans with longer remaining terms are more likely to default, possibly due to the increased uncertainty and risk over a longer time horizon.

- `time_since_first_obs`: The negative coefficient (-1.7005) and significant p-value (< 0.0001) suggest that the longer the time since the first observation, the lower the risk of default, which could indicate that borrowers who have been monitored for longer periods are generally more stable.

- Features like `balance_time`, `hpi_time`, `balance_orig_time`, `LTV_orig_time`, and `Interest_Rate_orig_time` were not significant predictors in this model, suggesting that they do not have a strong individual impact on the probability of default within this dataset.

4.3 Comparison with Logistic Regression

To further validate the performance of the MILR model, a comparison was made with a logistic regression model using LASSO regularization. The logistic regression model was trained with a LASSO penalty parameter of 0.00006, which was selected through 10-fold cross-validation. The comparison is presented in Table 7.

Table 7: Comparison of MILR and Logistic Lasso Models

Metric	**MILR**	**Logistic Lasso**
Accuracy	0.7253	0.4083
Precision	0.5701	0.3326
Sensitivity (Recall)	0.3613	0.9562
F1 Score	0.4423	0.4936

The results indicate that the MILR model outperforms the Logistic Lasso model in terms of accuracy (0.7253 vs. 0.4083) and precision (0.5701 vs. 0.3326), suggesting that the MILR model is better at correctly predicting non-default cases and avoiding false positives. Although the Logistic Lasso model has a higher sensitivity (0.9562 vs. 0.3613), this comes at the cost of a significant decrease in precision and overall accuracy. The high sensitivity of the Logistic Lasso model indicates that it is prone to predicting defaults even when the risk is low, leading to a high rate of false positives. In practical applications, such as mortgage default prediction, a model that can accurately and precisely identify non-default cases while maintaining a reasonable sensitivity is more valuable. This is because financial institutions are more interested in minimizing false positives to avoid unnecessary interventions and costs. Moreover, the higher accuracy of the MILR model demonstrates its robustness and reliability In conclusion, while the Logistic Lasso model shows high sensitivity, the MILR model's superior accuracy and precision make it a more effective and practical choice for predicting mortgage defaults, as it minimizes false positives and ensures more reliable overall predictions.

5 Conclusion

This study explored the application of MILR for predicting mortgage defaults using a comprehensive panel dataset. The MILR model demonstrated its effectiveness in handling the grouped structure of the data by treating each borrower as a bag and each loan observation as an instance. The performance of the model, evaluated through key metrics such as accuracy, precision, sensitivity (recall), and F1 score, indicated that while there is room for improvement, the MILR model provides a balanced approach to predicting mortgage defaults.

The significant coefficients identified in the MILR model provided valuable insights into the factors influencing mortgage defaults, such as loan-to-value ratios, interest rates, GDP growth, and property types. These findings align with existing research in several ways. Firstly, the study found that higher loan-to-value (LTV) ratios significantly increase the probability of default, consistent with the findings of Quercia and Stegman (1992), which underscores the importance of LTV as a critical factor in mortgage risk assessment. Secondly, the positive relationship between interest rates and default probability aligns with the work of Deng et al. (2000), who highlighted the role of macroeconomic factors in mortgage performance. Higher interest rates can lead to increased monthly payments, potentially causing financial strain on borrowers.

Furthermore, the negative coefficient for GDP growth found in this study is in line with Deng et al. (2000), who demonstrated that better economic conditions are associated with lower default risks. This suggests that macroeconomic stability plays a crucial role in the ability of borrowers to meet mortgage obligations. Lastly, higher credit scores at origination were found to be associated with lower default probabilities, supporting the conclusions of Lessmann et al. (2015) that borrower creditworthiness is a strong predictor of mortgage performance.

Overall, the superior performance in key metrics of the MILR model and its ability to handle the multiple instance structure of the data make it a promising tool for financial institutions and policymakers.

5.1 Future Research

Future research could focus on several key areas to enhance the predictive accuracy and robustness of the MILR model. One potential direction is to incorporate additional features that capture more nuanced aspects of borrower behavior and economic conditions. Integrating detailed borrower income data, employment history, and regional economic indicators could provide a more comprehensive understanding of the factors influencing mortgage defaults. For example, Ghent and Kudlyak (2011) demonstrated the significance of local economic conditions in predicting mortgage defaults, suggesting that regional economic indicators could improve model performance.

Another avenue for future research is to explore alternative modeling approaches that can enhance the sensitivity of the MILR model. Advanced machine learning techniques such as ensemble methods, deep learning models, or hybrid approaches that combine MILR with other predictive algorithms could be investigated. Guo and Berkhahn (2016) showed that ensemble methods and deep learning models can significantly improve predictive accuracy in financial applications, indicating their potential for mortgage default prediction.

Longitudinal studies that track changes in borrower behavior and economic conditions over time could also provide valuable insights into the dynamics of mortgage defaults. By examining the temporal evolution of default risk, researchers can develop more accurate predictive models that account for the time-varying nature of mortgage performance. Deng et al. (2000) emphasized the importance of longitudinal data in capturing the evolving risk factors associated with mortgage defaults, highlighting the value of such studies for future research.

By addressing these areas, future research can continue to improve the accuracy and applicability of models used for mortgage default prediction, ultimately contributing to more effective risk management strategies in the financial sector.

5.2 Practical Implications and Applications

The findings from this study have several practical applications for financial institutions, policymakers, and other stakeholders in the mortgage market. Financial institutions can use the MILR model to enhance their risk assessment and loan underwriting processes. By accurately identifying borrowers at higher risk of default, lenders can implement targeted interventions such as offering loan modifications, providing financial counseling, or adjusting loan terms to mitigate default risk.

Policymakers can leverage the insights gained from the MILR model to inform the development of regulations and policies aimed at promoting financial stability and protecting consumers. For example, understanding the impact of macroeconomic factors on mortgage defaults can help policymakers design effective measures to stabilize the housing market during economic downturns.

Additionally, investors in residential mortgage-backed securities (RMBS) can use the predictive capabilities of the MILR model to better assess the risk profile of their investment portfolios. By accurately predicting default probabilities, investors can make more informed decisions about portfolio allocation, risk management, and investment strategies.

References

Baesens, B., Roesch, D., and Scheule, H. (2016). *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS.* John Wiley & Sons.

Bajari, P., Chu, C. S., and Park, M. (2008). An empirical model of subprime mortgage default from 2000 to 2007. Technical report, National Bureau of Economic Research.

Bilal, M., Jewsbury, R., Wang, R., AlGhamdi, H. M., Asif, A., Eastwood, M., and Rajpoot, N. (2022). An aggregation of aggregation methods in computational pathology.

Chen, R.-B., Cheng, K.-H., Chang, S.-M., Jeng, S.-L., Chen, P.-Y., Yang, C.-H., and Hsia, C.-C. (2016). Multiple-instance logistic regression with lasso penalty. *arXiv preprint arXiv:1607.03615.*

Deng, Y., Quigley, J. M., and Van Order, R. (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica*, 68(2):275–307.

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Ghent, A. C. and Kudlyak, M. (2011). The role of local economic conditions in predicting mortgage defaults. *Regional Science and Urban Economics*, 41(3):216–227.

Guo, Y. and Berkhahn, F. (2016). Modeling financial time series with deep learning: A survey. *arXiv preprint arXiv:1612.05191.*

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

Maron, O. and Lozano-Pérez, T. (1997). A framework for multiple-instance learning. In *Proceedings of the 10th International Conference on Neural Information Processing Systems*, NIPS'97, page 570–576, Cambridge, MA, USA. MIT Press.

Quercia, R. G. and Stegman, M. A. (1992). Residential mortgage default: a review of the literature. *Journal of Housing Research*, pages 341–379.

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

Zhang, Q. and Goldman, S. (2001). Em-dd: An improved multiple-instance learning technique. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.