Mortgage Default Prediction with MILR

Yi-Chun Lo

Department of Statistics, National Taipei University

June 1, 2024

Outline

Introduction

- The mortgage market plays a crucial role in the global economy, particularly in the United States.
- 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 mitigate risks, formulate better lending policies, and enhance financial system stability.

Literature Review

Numerous studies have investigated the determinants of mortgage default using various methodologies and datasets.

- Quercia and Stegman (1992) highlighted the importance of borrower characteristics, loan terms, and economic conditions.
- Deng et al. (2000) incorporated macroeconomic factors such as interest rates, using hazard models for a dynamic understanding of default risk.
- Recent advancements in machine learning have further enhanced the predictive accuracy of mortgage default models (Bajari et al. (2008), Sirignano et al. (2016)).
- Combining machine learning with interpretable models has shown promising results (Lessmann et al. (2015)).

Multiple Instance Learning (MIL)

Multiple Instance Learning (MIL) is a type of supervised learning that deals with cases where label information is incomplete or uncertain (Bilal et al., 2022). MIL was first proposed by Dietterich et al. (1997) and is often illustrated with the keychain problem: In this problem, n staff each have one keychain with m_i keys. The manager knows which staff can open the room but not which specific key can. Each keychain is a "bag" and each key within a keychain is an "instance". The goal is to predict whether a keychain (bag) can open the room based on the properties of the keys (instances).

Research Objective

- Identify the key determinants of mortgage default by analyzing a comprehensive panel dataset.
- Employ Multiple Instance Logistic Regression (MILR) to build predictive models.
- Leverage the MILR model to improve the accuracy and interpretability of predictions.
- Provide valuable insights into the factors influencing mortgage defaults.

Data

The dataset includes origination and performance observations for 50,000 U.S. mortgage borrowers over 60 periods. Provided by Baesens et al. (2016), the data reflects real-world conditions with loans originating before the observation period and potential censoring due to loan maturation or refinancing.

- `id`: Borrower ID
- `orig_time`, `first_time`, `mat_time`: Various loan timestamps
- `balance_time`, `LTV_time`, `interest_rate_time`: Financial metrics at observation
- `hpi_time`, `gdp_time`, `uer_time`: Economic indicators at observation

Data

- `REtype_CO_orig_time`, `REtype_PU_orig_time`, `REtype_SF_orig_time`: Real estate types at origination
- `investor_orig_time`, `balance_orig_time`, `FICO_orig_time`: Borrower and loan metrics at origination
- `LTV_orig_time`, `Interest_Rate_orig_time`, `hpi_orig_time`: Financial metrics at origination
- `default_time`, `payoff_time`, `status_time`: Loan status observations

Data Preprocessing

- **Handling Missing Values**: Replaced missing numeric values with the median of the respective column.
- **Feature Engineering**: Created new features such as loan age, time to maturity, and time since first observation.
- **Label Creation**: Created a default label indicating if a borrower had defaulted at any point.
- **Data Cleaning**: Removed unnecessary columns to streamline the dataset.
- **Standardization**: Standardized numeric features to have zero mean and unit variance.

Data Exploration and Visualization

A significant imbalance between defaulted and non-defaulted loans, with the majority of loans being non-defaulted.

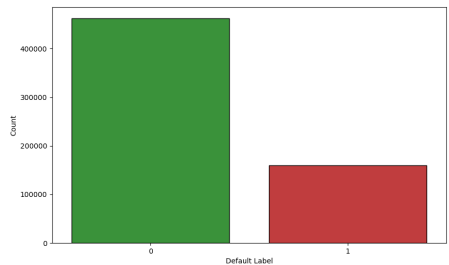

Figure: Count of Default Label

Data Exploration and Visualization

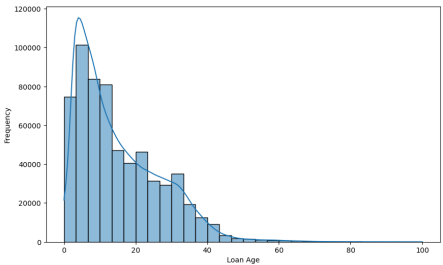

Figure: Distribution of Loan Age

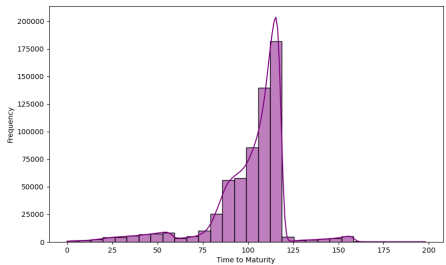

Figure: Distribution of Time to Maturity

Data Exploration and Visualization

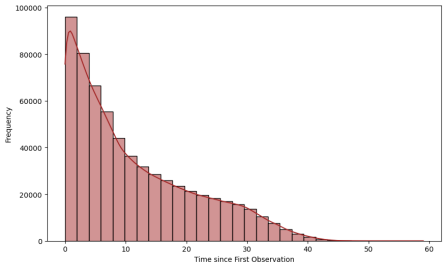

Figure: Distribution of Time Since First Observation

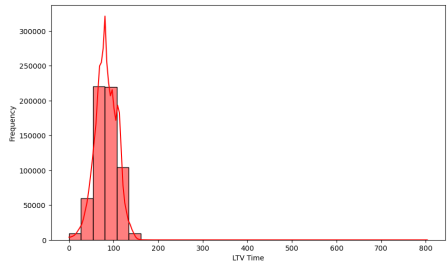

Figure: Distribution of LTV Time

Correlation Analysis

Key observations from correlation analysis:

- Positive correlation between `time_to_maturity` and default label (0.2203).
- Positive correlations for `LTV_time` (0.1916) and `interest_rate_time` (0.1811) with default label.
- Negative correlations for `loan_age` (-0.2415), `FICO_orig_time` (-0.1303), and `gdp_time` (-0.1377) with default label.

Multiple Instance Logistic Regression (MILR)

- MILR is used to predict bag labels (borrower default) based on instance labels (loan observations).
- The model uses logistic regression within the MIL framework.
- Each borrower is considered a "bag" and each loan observation within a borrower's history is an "instance."
- The Expectation-Maximization (EM) algorithm iteratively estimates model parameters, maximizing the expected log-likelihood of the data.

MILR Model Formulation

- For the jth instance of the ith bag, the instance label Y_{ij} follows a Bernoulli distribution with the logistic regression form:

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

- The bag label Z_i follows a Bernoulli distribution:

$$Z_i \sim Bernoulli(\pi_i) \quad \text{where} \quad \pi_i = 1 - \prod_{j=1}^{m_i}(1 - p_{ij}) \tag{2}$$

Expectation-Maximization (EM) Algorithm

- **E-Step:** Calculate the conditional expectation of the instance labels given the observed bag labels and the current parameter estimates.
- **M-Step:** Maximize the Q-function with respect to β to obtain new parameter estimates.

$$\hat{\beta}^{(t+1)} = \arg\max_{\beta} \mathrm{Q}\left(\beta \mid \hat{\beta}^{(t)}\right) \tag{3}$$

The Q-function is the expected complete-data log-likelihood:

$$\mathrm{Q}\left(\beta \mid \hat{\beta}^{(t)}\right) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \hat{\theta}_{ij}^{(t)} \log(\hat{p}_{ij}) + \left(1 - \hat{\theta}_{ij}^{(t)}\right) \log(1 - \hat{p}_{ij}) \tag{4}$$

Model Training and Evaluation

- **Data Split:** The dataset was partitioned into training (70%) and testing (30%) sets.
- **Performance Metrics:**
  - **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.
  - **Sensitivity (Recall):** The proportion of true positive predictions out of all actual positives.
  - **F1 Score:** The harmonic mean of precision and recall.

Model Performance

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

The MILR model achieved a moderate accuracy of 72.53%. Precision was 57.01%, indicating over half of the positive predictions were correct. Sensitivity was 36.13%, showing the model missed some actual defaults. The F1 score of 44.23% suggests room for improvement.

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 ***

Model Coefficients

Feature	Estimate	Std. Error	Z Value	p-value
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 ***

Model Coefficients

Feature	Estimate	Std. Error	Z Value	p-value
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 ***

Comparison with Logistic Regression

The logistic regression model was trained with a LASSO penalty parameter of 0.00006, which was selected through 10-fold cross-validation.

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

Conclusion

- The MILR model effectively handles the grouped structure of mortgage data.
- Significant coefficients provide insights into key factors influencing defaults.
- Findings align with existing research:
  - Quercia and Stegman (1992) highlighted the importance of LTV ratios.
  - Deng et al. (2000) demonstrated the impact of interest rates and GDP growth.
  - Lessmann et al. (2015) emphasized the relevance of credit scores.
- MILR's superior accuracy and precision make it a practical choice for predicting mortgage defaults.

Future Research

- Incorporate additional features such as detailed borrower income data, employment history, and regional economic indicators to provide a more comprehensive understanding of mortgage defaults (Ghent and Kudlyak, 2011).
- Explore advanced machine learning techniques, including ensemble methods, deep learning models, and hybrid approaches that combine MILR with other algorithms to enhance predictive accuracy and model robustness (Guo and Berkhahn, 2016).
- Conduct longitudinal studies to capture evolving risk factors in mortgage defaults and improve the temporal accuracy of predictive models, providing deeper insights into the dynamics of default risk over time (Deng et al., 2000).

Practical Implications and Applications

- Financial institutions can improve risk assessment and loan underwriting by identifying high-risk borrowers, allowing for targeted interventions like loan modifications or financial counseling.
- Policymakers can develop regulations to promote financial stability and protect consumers. Insights into macroeconomic factors affecting mortgage defaults can guide measures to stabilize the housing market.
- Investors in RMBS can better assess the risk profile of their portfolios. Accurate default predictions enable informed decisions about portfolio allocation and risk management, enhancing investment strategies.

References I

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.

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.

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*.

References II

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.

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*.

Mortgage Default Prediction with MILR

Yi-Chun Lo

Department of Statistics, National Taipei University

June 1, 2024