A Predictive Analytics Approach to Estimate Home Loan Risk and Allocate Loans

Jinyu Chen, Bharat Agarwal, Alejandro Hoyos Orozco, Esteban Molina, Isha Sharma, Matthew A. Lanham

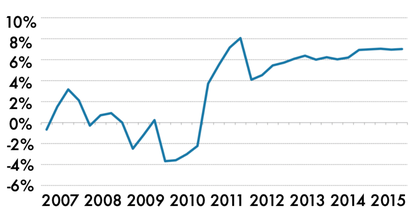
Purdue University Krannert School of Management

chen2409@purdue.edu; bagarwa@purdue.edu; ahoyosor@purdue.edu; molina25@purdue.edu; sharm258@purdue.edu; lanhamm@purdue.edu

Abstract

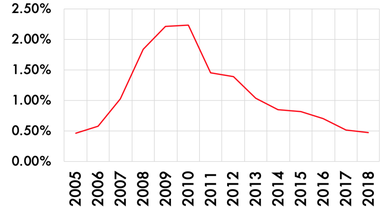
We develop a predictive model for home lending institutions to estimate and evaluate riskiness and creditworthiness of their customers. We use sophisticated ML algorithms to accurately predict loan defaulters. Our model also ranks customers based on actual historical default performance using a probability calibration plot. We aim to provide intelligent decision-support for financial decision-making using quantitative finance and predictive analytics methodologies. Our research allows estimating risk accurately and in a visually interpretable way that is fair to both parties.

Introduction



Source: https://articles2.marketrealist.com/2015/09/understanding-role-credit-economy/

Figure 1. YoY Credit Growth



Source: https://www.statista.com/statistics/798766/foreclosure-rate-usa/

Figure 2. Foreclosure Rate

The collapse of the housing market in 2007-08 was due to bad credit risk management.

So, how can our predictive model help?

For the ***banks***, our model helps in evaluating potential clients, managing credit risk, and provides the ability to issue more loans, essentially increasing their interest income.

For a ***consumer***, a smaller home loaner who might be unable to get a loan in the past, can now have a better opportunity to get the loan.

For the ***overall economy***, as credit leads to an increase in spending, it increases income levels, in turn leading to higher GDP which facilitates economic growth.

Research questions:

- Which machine learning models fairly, rank customers for home loans?
- What is the relationship among statistical performance and business performance (i.e. probability calibration plots)?

Literature Review

Study	All factors considered	Time Efficiency	Feature Selection	Accuracy
Dukic, 2011		√	√	√
Li, 2011	√		√	√
Zhou, 2013	√	√		√
Benyacoub, 2014	√	√	√	
Our Study	√	√	√	√

Table 1. Literature review summary by methods' features

We compare and contrast our study to previous papers, to demonstrate that other evaluations should be considered besides strictly statistical metrics.

Methodology

Figure 3 outlines our study design, starting from data collection, data cleaning, data pre-processing, feature creation and selection, model/approach selection, cross-validation design, and model assessment/performance measures.

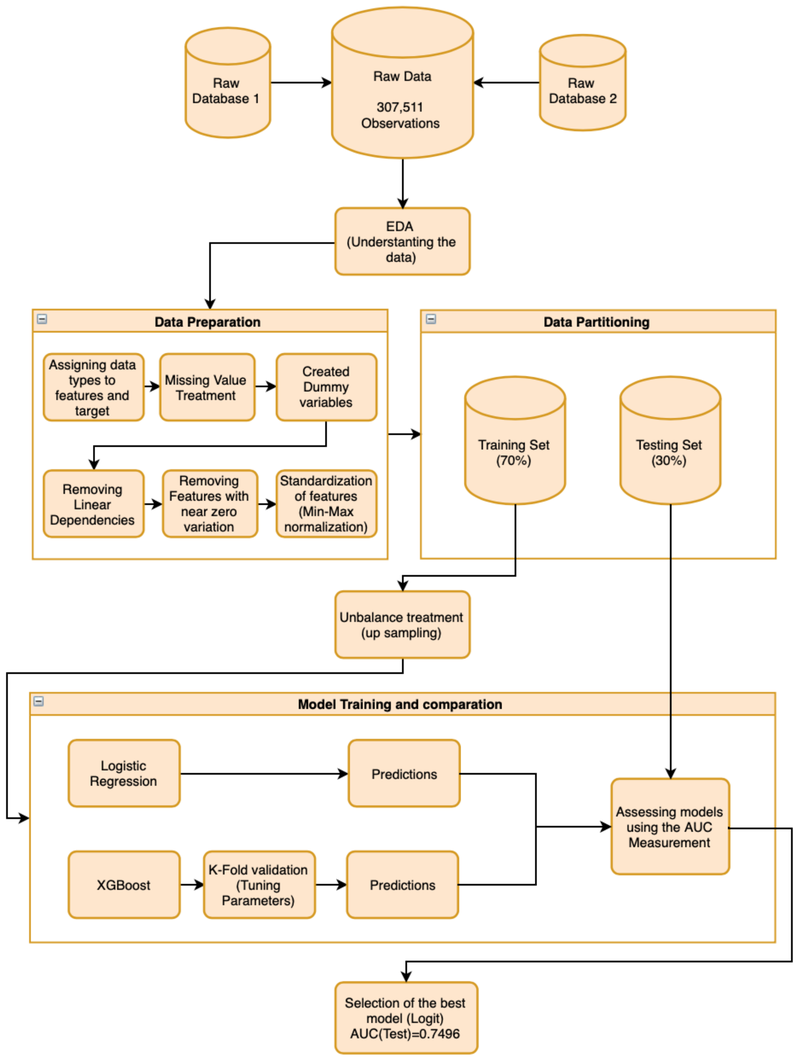


Figure 3. Study Design

Results

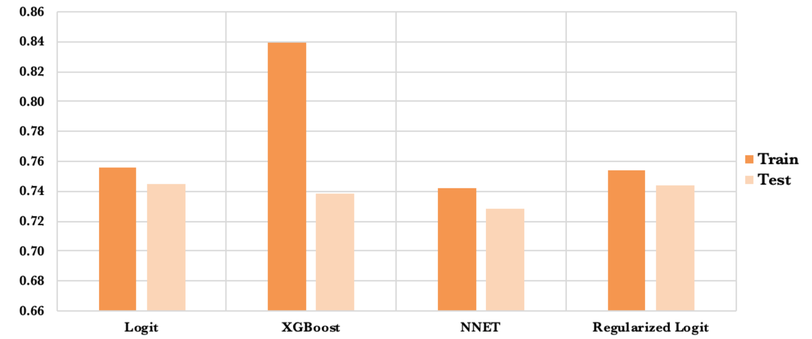


Figure 4. Model Evaluation (AUC)

		Actual	
		Default	No-Default
Predicted	Default	3,895	20,880
	No-default	1,786	46,324

Sensitivity	0.68562
Specificity	0.68930

Figure 5. Confusion Matrix

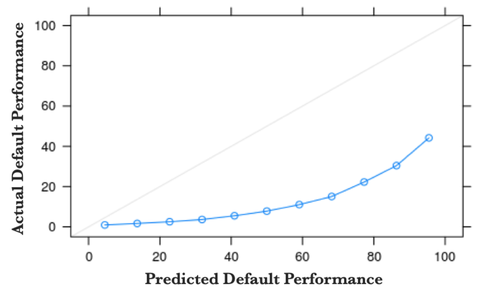


Figure 6. Probability Calibration Plot

Conclusions

As credit allocation affects the banks, the consumers and the overall economy, managing credit-risk is of critical importance. Applying the different machine learning models, and analyzing their performance in each case, we concluded that Logistic Regression is the best ML model to identify loan defaulters accurately. Using the AUC, we found our model correctly predicts the default performance for 68.5% of the loan applicants.

Acknowledgements

We thank Professor Matthew Lanham and his MS BAIM graduate students for constant guidance on this project.