Loan Default Risk Prediction

Xiaoxue Chen

October 2024

1 Introduction

The possibility of default is threatening to lenders and borrowers alike, with significant consequences for both parties. For lenders, the costs of default are expensive, and may likely have secondary effects on its ability to raise capital in the future; for borrowers, default leads to punitive measures which have long-lasting effects. Knowing that default occurs more frequently in some types of debt than others, we seek to estimate the likelihood of default in peer-to-peer personal loans.

We demonstrate that Ridge and Lasso regression is an appropriate and preferred tool for our problem and its particular constraints. We show that Lasso regression can estimate the likelihood of default at least as well as alternative method which we consider: Ridge regression. Lasso allows us to overcome the limitations in the dataset and produce a more reliable predictive model.

The structure of the paper follows. Section 2 elaborates on the problem. Section 3 explains our methodology: our motivation for choosing it in particular and its mathematical underpinnings. Section 4 presents the data and section 5 the results. In section 6 we briefly discuss the results.

2 Problem: Predicting Default on Personal Loans

Lenders take many precautions to ensure that they only admit borrowers who will fully repay their debt. If lenders did not believe debt would be repaid, expected returns would be negative and investors would withdraw their funds. Credit would become unavailable or prohibitively costly. Despite admitting only borrowers they believe would repay, default still occurs — sometimes, very frequently. For example, in 2010 the average delinquency rate on single-family mortgages hit an all-time high, at 10.9%. Since then, it has steadily fallen to 2.5% in 2019, approximately the average during the 1990s. Mortgages are only one example of default in spite of extensive pre-conditions for borrowing.

Our problem is an exercise in supervised learning: using an extensive cross-section of credit-worthy borrowers — some of whom default but most of whom do not — we estimate the likelihood that a new borrower eventually fails to repay. Knowing this parameter estimate has practical finance applications for a lender: for example, whether to adjust its interest rate and fees structure, or calculating Value-at-Risk and compliance with regulatory requirements on stress-testing.

3 Methodology

Regression applications usually prefer ordinary least squares (OLS) regression; when the assumptions of the Gauss-Markov theorem are satisfied, OLS is the best linear unbiased estimator. However, there are three features of our problem which preclude using OLS. Ours is a prediction problem, so we frame each in the context of generalization error.

3.1 Model Specification: Logistic Regression

Logistic regression was chosen as the primary modeling approach because the target variable, loan status, is a binary outcome (e.g., "default" vs. "non-default"), making it well-suited for this type of classification task and ensuring that predictions remain within the range of 0 to 1. This makes it ideal for binary classification problems where the goal is to predict the likelihood of an outcome.

Additionally, logistic regression provides interpretable coefficients, allowing for insights into the relationship between predictor variables and the likelihood of the target outcome. This interpretability is crucial for understanding key drivers of loan status and making informed business decisions.

To predict the likelihood of loan default, we employ Logistic Regression. The link function is $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_{k+1} x_{1i}^2 + \beta_{k+2} x_{2i}^2 + ... + \beta_{2k+1} x_{1i} x_{2i} + \beta_{2k+2} x_{1i} x_{3i} + ... + \epsilon$. And the log-likelihood is given by: $L(\beta_0, \beta_1) = \prod_{i=1}^n p_i^{y_i}(1-p_i)^{1-y_i}$. Taking the log, we get $\ell(\beta_0, \beta_1) = \sum_{i=1}^n [y_i \log p_i + (1-y_i)\log(1-p_i)]$

Substituting $p_i = \frac{e^{\text{logit}(p_i)}}{1+e^{\text{logit}(p_i)}}$, we have log-likelihood:

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{e^{\text{logit}(p_i)}}{1+e^{\text{logit}(p_i)}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\text{logit}(p_i)}}\right)\right]$$

Normally, we do not have a closed form of the parameters $\hat{\beta}_0, \hat{\beta}_1, ..., \beta_{2k+n}$, they are often achieved using gradient descent or the Newton-Raphson method: $\beta^{(t+1)} = \beta^{(t)} - S'(\beta^{(t)})S(\beta^{(t)})^{-1}$, where S represents the score function and S' represents the Hessian matrix.

3.2 Regularization: Ridge/Lasso Regression

Ridge and Lasso regressions were chosen to mitigate multicollinearity and reduce overfitting because they retain the original features in the model, making them better suited for tasks where interpretability and inference are important. This allows us to identify the most important predictors of loan status while maintaining their direct relationships, which is critical for inference and understanding the key drivers of the outcome. In contrast, while PCA can also reduce overfitting by decreasing dimensionality, the principal components lose direct interpretability.

Ridge Regression

The mechanics of Ridge regression differ from OLS because Ridge regression added a penalty called the L_2 penalty in the minimization criterion: $\hat{\beta}_{\text{Ridge}} = \arg\min_\beta \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^T X_i)^2 + \lambda\|\beta\|_2^2$, where $\|\beta\|_2^2 = \sum_{j=1}^d \beta_j^2$. It turns out that the ridge regression has a closed-form solution similar to the least squares estimator and the spline: $\hat{\beta}_{\text{Ridge}} = \left(X^T X + n\lambda I_d\right)^{-1} X^T Y$.

The coefficients are moved toward 0 because in the matrix inverse, there is an additional $n\lambda I_d$ term. We will say that the ridge regression shrinks the estimator $\hat{\beta}_{\text{Ridge}}$ toward 0 (but does not reach 0).

The ridge regression can also be viewed as a Bayesian estimator (posterior mean). To see this, we assume the model $Y = \beta^T X + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ and place a prior over the parameter $\beta \sim N(0, \tau^2)$. Then you can show that the posterior mean is the ridge regression estimator with $\lambda = \frac{\sigma^2}{\tau^2}$.

Lasso Regression

Similarly to Ridge regression, Lasso regression also added a penalty term called ℓ_1 penalty: $\hat{\beta}_{\text{LASSO}} = \arg\min_\beta \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^T X_i)^2 + \lambda\|\beta\|_1 = \arg\min_\beta \hat{R}_n(\beta) + \lambda\|\beta\|_1$, where $\|\beta\|_1 = \sum_{j=1}^d |\beta_j|$ is the 1-norm of the vector β. If we normalize the covariates so that $X^T X = I_d$, the LASSO estimates can be written as $\hat{\beta}_{\text{LASSO},j} = \hat{\beta}_{\text{LS},j} \times \max\left\{0, 1 - \frac{n\lambda}{|\hat{\beta}_{\text{LS},j}|}\right\}$ they will be shrink to 0.

4 Data

We use publicly available loan and borrower data from LendingClub, an early online peer-to-peer lending platform. LendingClub facilitated peer-to-peer lending from its founding in 2006 until 2020, when it concluded the program to restructure itself as a digital bank. Our data exist because LendingClub took advantage of its unique status in a yet-unregulated industry. It crowd-sourced data analysis by publicly sharing its anonymized borrower data. These data begin with loans (and corresponding borrowers) issued in 2007 and conclude in the 4th quarter of 2018. Shortly thereafter, LendingClub removed the public data as it prepared to end the peer-to-peer platform and and restructure as a digital bank. The dataset has 2,839,891 complete cross-sectional observations which reflect the status of the loan as of 2018 Q4.

4.1 Unbalanced Classes

The dataset is highly imbalanced, with a significantly smaller proportion of defaulters compared to non-defaulters. In our original sample of $n = 2,839,891$, only about 1.7% of the loans are in default, which poses a challenge for predictive modeling. Machine learning algorithms tend to minimize overall error, which can lead to a bias towards the majority class and result in the minority class being under-predicted or ignored entirely.

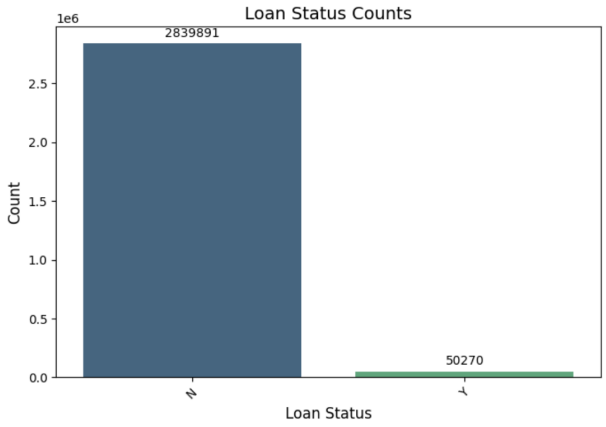


Figure 1: Distribution of Loan Settlement Flags

To address this issue, we do not use the complete dataset. The first is that it is neither necessary nor relevant, since we emphasize the usefulness of Ridge when there are many predictors relative to the sample size. It would also increase the intensiveness of computations. For this reason, we draw a random subsample of 50,270 observations. This subsample, which is the basis of our analysis, is presented below.

	Non-defaulters	**Defaulters**
Unbalanced	2,839,891	50,270
Balanced	50,270	50,270

Table 1: Class Imbalance

4.2 Feature Selection

Feature selection is a crucial step in the data preprocessing phase of building a predictive model. It involves identifying and retaining only the most relevant features while removing redundant, irrelevant, or noisy variables. In our dataset, there are certain variables that appear to be unreasonable and should be excluded from our model, for example,

1. id: This is a unique identifier for each record and does not carry any predictive information for the target variable.
2. url: Typically a URL string, which is irrelevant for modeling purposes.
3. hardship_flag: This variable is heavily imbalanced (e.g., most values are "no"), it will not contribute meaningful information. And it also might cause data leakage.

5 Results

5.1 Lasso Regularization Outperforms Ridge

We compare Ridge and Lasso regularization applied to logistic regression using metrics such as MSPE (Mean Squared Prediction Error), the number of nonzero coefficients, and the classification performance measured by accuracy and AUC-ROC. Both regularization techniques aim to improve model generalization and reduce overfitting, but Lasso emerges as the better option under the tested conditions.

Model	In-sample MSPE	Out-of-sample MSPE	Accuracy	Log-Loss
Ridge	0.049827	0.050612	0.942242	0.197033
Lasso	0.049797	0.050558	0.942272	0.197028
OLS	—	—	0.922431	0.197101

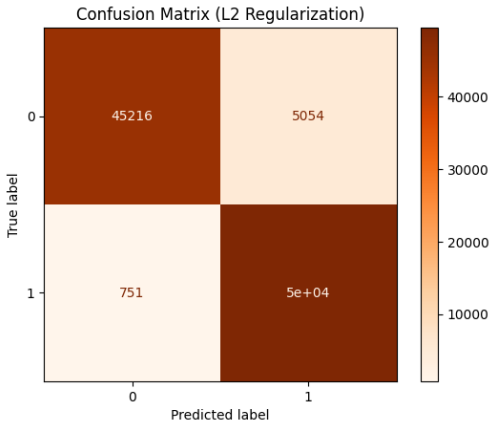
Table 2: Comparison of Ridge and Lasso Regularization

One key observation is that the Out-of-sample MSPE for Lasso (0.050558) is slightly lower than that for Ridge (0.050612). This reflects Lasso's ability to generalize slightly better when tested on unseen data. Similarly, the In-sample MSPE for Lasso (0.049797) is marginally lower than Ridge's (0.049827), showing consistent superiority in both training and test performance.

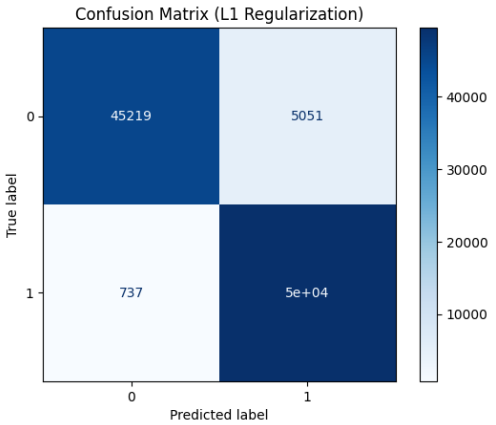
The classification performance of both methods is very close, with Lasso achieving a slightly lower Log-Loss (0.197033 vs. 0.197028) and similar accuracy (94.23% vs. 94.22%). While the differences are small, Lasso consistently performs better, making it the preferred choice.

When comparing the number of nonzero coefficients, both Ridge and Lasso retain the same number of features (81), suggesting that under the current regularization settings, Lasso does not lead to additional feature sparsity. However, it is worth noting that Lasso's strength lies in its ability to enforce sparsity, which could be more apparent under higher regularization strengths (larger λ values).

In conclusion, while both Ridge and Lasso perform well, Lasso offers slightly better generalization, probability validity, and classification performance, making it the preferable choice for this analysis



(a) Confusion Matrix for Ridge Regularization



(b) Confusion Matrix for Lasso Regularization

Figure 2: Confusion Matrix

5.2 Top Features

From Table 3 and Table 4, we can see that both models identify key financial variables, such as *collection_recovery_fee*, recoveries, and *loan_status*, as highly influential, indicating their robustness in influencing the target variable. Ridge regression retains all features by shrinking coefficients toward zero, making it ideal for handling multicollinearity and preserving all feature contributions. In contrast, Lasso regression introduces sparsity by driving some coefficients to zero, emphasizing a simpler and more interpretable model. Lasso also assigns larger magnitudes to influential features compared to Ridge, reflecting its feature selection process. Overall, Ridge is suited for scenarios where all features contribute meaningfully, while Lasso is advantageous for simplifying models and focusing on the most important predictors. These insights should be validated further through cross-validation or domain-specific evaluation.

Feature	**Ridge Coefficient**
collection_recovery_fee	5.7870
recoveries	-3.6377
loan_status	-2.1523
out_prncp	-1.4338
total_pymnt_inv	-1.3330
out_prncp_inv	-1.2080
total_rec_prncp	1.0908
last_fico_range_high	-0.8903
total_rec_int	0.8223
loan_amnt	-0.5624

Table 3: Top Features by Ridge Regression

Feature	Lasso Coefficient
collection_recovery_fee	6.4046
recoveries	-4.5063
loan_status	-2.1519
out_prncp	-1.7461
out_prncp_inv	-0.8964
last_fico_range_high	-0.8903
total_rec_int	0.5428
total_pymnt_inv	-0.5000
total_rec_prncp	0.4216

Table 4: Top Features by Lasso Regression

6 Conclusion

Logistic regression remains one of the most widely used methods for binary classification problems, making it particularly effective in predicting 0/1 outcomes such as loan default versus non-default. Its probabilistic framework allows for the interpretation of predicted probabilities, offering actionable insights into the likelihood of default. Additionally, logistic regression provides interpretable coefficients, making it an invaluable tool for understanding the relationship between predictor variables and target outcomes. This interpretability is critical for financial applications, where transparency and explainability are key considerations.

This study also demonstrated that both Ridge and Lasso regression are robust techniques for predicting the likelihood of loan default in a dataset of borrowers and their respective financial attributes. By comparing Ridge and Lasso alongside traditional logistic regression models, we emphasized their strengths in addressing multicollinearity and improving model generalization. We employed a cross-validation approach to fine-tune the regularization parameters and analyzed the trade-offs between interpretability and predictive accuracy. Additionally, we examined the sensitivity of the models to regularization and highlighted the differences in feature selection between Ridge and Lasso regression.

The findings of this paper underscore the importance of incorporating regularization techniques into risk modeling frameworks for financial institutions. By reducing overfitting and improving out-of-sample predictions, these methods enhance decision-making processes related to credit risk management. The implementation of such models provides a valuable tool for assessing default probabilities and optimizing asset management strategies, particularly in environments with high-dimensional or noisy datasets.