Credit Risk Modelling: Shrinkage Methods and Lasso Selection in PD Modelling (Part 1)

written by Hikmat Abdulazizov Hikmət Əbdüləzizov

In the dynamic environment of the financial industry, the accurate assessment of credit risk is as a pivotal factor in maintaining the stability and sustainability of lending institutions. As the global economy continually evolves, so too do the complexities and challenges associated with credit risk management. In this context, the utilization of advanced econometric techniques and diverse datasets has become important for financial institutions seeking to maintain the precision and reliability of their credit risk models. Credit risk refers to the potential that a borrower may fail to meet their financial obligations, which would lead to a financial loss for the lender. It is essentially the risk that arises from the uncertainty of whether borrowers will repay their loans and/or meet their contractual obligations. Effective credit risk management is crucial for maintaining the stability and solvency of financial institutions and for sustaining a healthy financial system. The field of credit risk modeling involves developing sophisticated statistical and mathematical models to predict and quantify credit risk, helping lenders to make informed decisions about lending and managing their overall risk exposure. The integration into such models of advanced statistical and machine learning techniques allows for a more detailed evaluation of creditworthiness, which enables financial institutions to make informed lending decisions and optimize capital allocation.

Default

In the context of credit risk in banking, default, with a 90-day overdue criterion, occurs when a borrower fails to make a payment on their loan or credit obligation for a period exceeding 90 days (3 consecutive unpaid months in the case of the Azerbaijani market or one similar to it). This extended delinquency is considered a substantial breach of the loan agreement, indicating a heightened level of credit risk. Financial institutions commonly use the 90-day threshold as a significant milestone for classifying borrowers as being in default.

The accurate econometric estimation of Probability of Default (PD) is a critical component in assessing credit risk. In recent years, there has been a growing interest in admitting the power of shrinkage methods and Lasso selection techniques to gain the precision and stability of PD models. Conventional PD modeling typically involves traditional statistical techniques, such as logistic regression. In logistic regression, the relationship between the predictor variables (features related to a borrower's creditworthiness) and the binary target outcome (default or non-default) is modeled. The coefficients obtained from logistic regression represent the influence of each predictor on the likelihood of default. Advantages of this model are the easy interpretability of its coefficients and its simplicity. Logistic regression is a well-established and widely used method. However, it might be the oldest classifier model.

Another challenge is that logistic regression assumes a linear relationship between predictors and the log-odds of default, and it may not capture complex, non-linear relationships in the data. In Azerbaijan's outdated banking sector, even the largest banks still use only this model. Moreover, the selection of important variables becomes incredibly burdensome with this model. Imagine you have 10 binary variables. Then you need to have 2^(10)=1.024 possible models. If you have 20 features, it becomes 1.048.576 possible models, which is impossible to compute in a constrained time interval. We will

deal with this problem in Part 2 of the paper.

Thus, this study focuses on the application of shrinkage methods, particularly Ridge and Lasso regression, in PD modelling. Shrinkage methods are employed to mitigate the problem of overfitting, which often spoils traditional statistical models in credit risk assessment. By incorporating penalty terms, these methods can effectively regularize the model, reducing the impact of noisy or multicollinear predictors while preserving the essential information.

Furthermore, Lasso selection, a variant of the shrinkage technique, aids in feature selection and model interpretability. By inducing sparsity in the model coefficients, Lasso automatically identifies the most relevant variables, leading to more parsimonious and interpretable models. This attribute is of utmost importance in credit risk modelling, as it enables financial institutions to pinpoint the key factors driving default probabilities.

This study explores the benefits and challenges associated with shrinkage methods and Lasso selection in PD modelling. We compare the predictive performance and stability of these models against traditional approaches, such as logistic regression and decision trees, using real-world credit datasets. Additionally, we delve into the interpretability of Lasso-selected features and the implications for risk management.

The results of this research demonstrate that the integration of shrinkage methods and Lasso selection in PD modelling leads to more robust and accurate credit risk assessments. These techniques not only improve the model's predictive power but also simplify the model's complexity, making it more accessible for stakeholders. As the financial industry continues to evolve, embracing these advanced methodologies is crucial for managing credit risk effectively and making well-informed decisions in lending and investment.

This is a very important topic since these calculations in the end directly affect the bank profits. Moreover, a LASSO approach can be applied not only in banks but in various industries that need reasonable feature selection and forecasting. We took the data from the open source and it consists of 307.512 individuals (or loans) and 122 features. Thus, the dataset is very rich, and not only can or should the methods supported by its use in this study be applied to any bank in Azerbaijan, but also the dataset gives an idea of what should be recorded as data since the Azerbaijani financial market is still in a state of primitive incubatory in terms of data and modelling culture.

The series of articles, of which this is the first, will provide a substantially useful methodology and approach to the market and researchers. This series will contain two articles: In the first paper we will show the somewhat conventional modelling approach and its difficulties together with Ridge of shrinkage modelling. In the second paper, we will dig deep into the LASSOism of Robert Tibshirani (1996). In that way we will understand how effectively we can decrease the time and complexity using modern techniques.

Data and Methodology

As we note above, the dataset used here is open source and consists of around 300.000 individuals and their rich features. Around 10% of the data is ones (1=a defaulted individual). Therefore, this is what we use for predicting the targeted default outcome (yes/no, 1s and 0s). The answer is rather simple: Whatever we have in the dataset and "more." "More" here means that we can generate more features than we have using what we already have. For example, if we have age data, we can square or cube it to create more variables which account for non-linearity. We do not recommend using the weight of evidence (WOE) transformation to make monotonic relations for such kind of variables. Below is a very small portion of data used for your attention. In general, banks

have similar data for econometric analysis. The most important and valuable data, of course, are delinquency data as well.

Table 1. Features

ID	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	AGE	FLAG_OWN_CAR
100002	1	Cash loans	M	43	N
100003	0	Cash loans	F	26	N
100004	0	Revolving loans	M	34	Y
100006	0	Cash loans	F	63	N

Here, we only use cash loans as revolving loans need to be addressed separately. We separate data into train set, validation set and test set. We need the validation set for tuning the hyperparameter, namely the cutoff. We need cutoff since our logistic regression predicts probabilities of default, and we must transform them to zeros and ones by using the cutoff parameter (between zero and one) to test it on the test set. We use backward elimination for conventional logistic regression to be time-effective. We end up with 20 variables out of more than 100 features. To have fewer variables, we used p-values less than 1% for selection of the features. Of course, we could end up with more if we use p-values of less than 5%.

The p-value is a probability associated with a statistical test. It quantifies the evidence against a null hypothesis. In hypothesis testing, the null hypothesis represents a default assumption that there is no effect or no difference. The p-value indicates the probability of obtaining the observed results or more extreme results when the null hypothesis is true. If p is less than or equal to the predecided significance level alpha: Reject the null hypothesis and find that the variable is significant. If $p > alpha$: Fail to reject the null hypothesis, thus it is an unimportant variable. We want to emphasize the age variable a bit more. It is in the form of $(a*x-b*x^2)$. It is downward parabola. This means that our default response variable has nonlinear relationship to

age. At some point, the default probability increases with age and after some age it decreases. This observation might be related to the well-known fact that earnings are maximized at the median age. That is why we do not recommend linear transformations such as weight of evidence for these kinds of variables.

Next, we need to validate the hyperparameter cutoff using the validation set. This is a significant issue since we do not know what to call default or non-default after predicting probabilities of default between 0 and 1 with the training data set. A function predicting classes could be used but would make the program choose the cutoff value for probabilities as 0.5, which is not justified due to oversimplification. A cutoff value of 0.5 means above 0.5 class is predicted as 1, otherwise as 0.

Although it seems a priori normal, we need to check each dataset for justification of such a cutoff. Here we need to introduce criteria for which we obtain cutoff values and then compare models. We can use the AUC (area under the curve), but we need more universal criteria to make the models comparable with machine learning counterparts such as random forest or decision tree.

Three concepts need to be defined: accuracy, sensitivity, and specificity. Accuracy is the portion that has been captured by the model correctly. If the accuracy is 90%, then 90% of the zeros and ones were predicted correctly. Sensitivity measures the rate at which a model correctly predicts true positive (TP) rates(1s). On the contrary, specificity measures the rate at which a model correctly predicts negatives (0s).

Graph 1. Confusion Matrix

| | **Predicted Value** | |
	Positive	Negative
Actual Value Positive	True Positive	False Negative
Actual Value Negative	False Positive	True Negative

Three criteria are used to choose the cutoff values for each of the three models. The first is the cutoff that maximizes accuracy. The second maximizes both sensitivity and specificity, minimizing the distance between the upper left corner of the ROC curve graph and the curve itself. An area under the curve equal to 0.5 indicates a random classifier, which has no predictive power. Using Gini instead of AUC is also not recommended.

Graph 2. ROC curve

The last criterion is the cost-minimizing cutoff, which minimizes the self-defined cost function. This cost function sums up false negatives and false positives to achieve a situation where false negatives are three times as costly as false positives. In this case, it means that the cost of mislabeling a default as non-default is much more costly than mislabeling a non-default as default since the former is riskier. The cutoff hyperparameter has a huge importance since it directly affects the profit of the financial institution, the bank.

Results

Here we enter the world of shrinkage models. In Part 2, we will explore them better but here we need them to compare the models with conventional logistic regression. Shrinkage models, also known as regularization or penalized regression models, are statistical techniques used to address issues such as overfitting and multicollinearity by adding a penalty term to the regression coefficients. These models aim to improve the stability and generalization performance of the model by shrinking or regularizing the estimated coefficients towards zero. Two common types of shrinkage models are Ridge Regression and Lasso Regression. They are also known to improve overall mean squared error by introducing extra bias in the model. There is an alpha parameter (between zero and one) in tuning shrinkage models in cases of penalized logistic regression. Since there are an infinite quantity of real

numbers between 0 and 1, we employ only 11 of them by increasing the alpha parameter by 0.1 each time and compare models with their respective AUCs (area under ROC curve). To be concise, we do not employ so called pure machine learning models here.

Graph 3. Models and Predictive Powers

Prediction Power
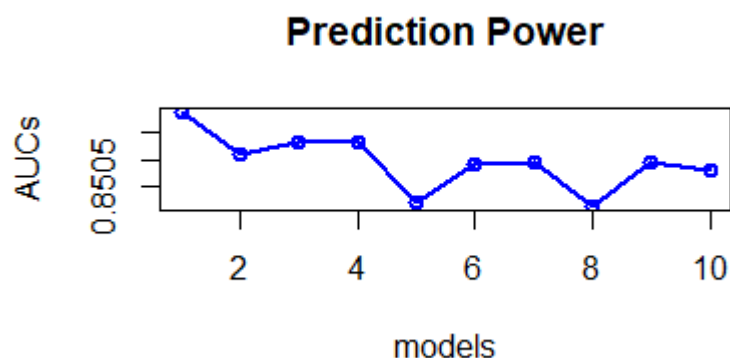
Let's look at the numbers of the selected models to pin them down to one model.

Table 2. Models' Results

MODEL 2 alpha = 0.1					
Accuracy Cutoff	0.32	Cost Minimum Cutoff	0.24	Minimum Distance Cutoff	0.09
Accuracy	0.83	Accuracy	0.81	Accuracy	0.76
Sensitivity	0.21	Sensitivity	0.29	Sensitivity	0.73
Specificity	0.93	Specificity	0.88	Specificity	0.67

MODEL 6 alpha = 0.6					
Accuracy Cutoff	0.35	Cost Minimum Cutoff	0.19	Minimum Distance Cutoff	0.11
Accuracy	0.88	Accuracy	0.83	Accuracy	0.73
Sensitivity	0.25	Sensitivity	0.27	Sensitivity	0.76
Specificity	0.96	Specificity	0.83	Specificity	0.62

MODEL 8 alpha = 0.8					
Accuracy Cutoff	0.43	Cost Minimum Cutoff	0.22	Minimum Distance Cutoff	0.07
Accuracy	0.87	Accuracy	0.79	Accuracy	0.79
Sensitivity	0.32	Sensitivity	0.31	Sensitivity	0.71
Specificity	0.96	Specificity	0.81	Specificity	0.72

All of the above depict elastic-net shrinkage models that lie between Lasso and Ridge (alpha does not take values 0 and 1 exactly but between them). I would present these to the managers of a bank and let them choose between cutoffs since it gives flexibility to them as well. However, academically speaking, I would choose the minimum-distance cutoff with the highest sensitivity since that is the most powerful. After selecting our model, we can test it on the test dataset of course. Then, we can apply our final model to the outstanding portfolio to find the future defaults and calculate our expected credit loss.

To conclude the first paper, we have shown that using shrinkage methods improves model accuracy and give us smaller errors. In general, the stubborn usage of the oldest models is not recommended. For further research in Part 2, we will dig deeper into feature selection via LASSOism and machine learning.

References:

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp.267–288.