NUS National University of Singapore | NUS BUSINESS SCHOOL

DBA3713 Group 3 Assignment 1

Name	Matriculation No.
Ammar Bin Hussein Bagharib	A0218111X
Lo Hei Ting	A0188435U
Loh Yee Shing, Bryan	A0188158N
Aaron Yuen Sze Tian	A0188252Y

Regression Results

Portfolio Overview

The LendingClub data set consists of 24,999 unique loan listings, of which 12,628 are used in our model. The average interest rates and loan amounts of these loans are 13.5% and $14,529.10. Out of these loans, 77.1% of the loans are fully paid off while the others are not.

Result Summary

	Basic Regression	**L1 Regression**
Number of Features	116	49
AUC	0.686	0.698
MLE Loss	0.497	0.488
Expected Payoff	127.61	152.79

Table 1

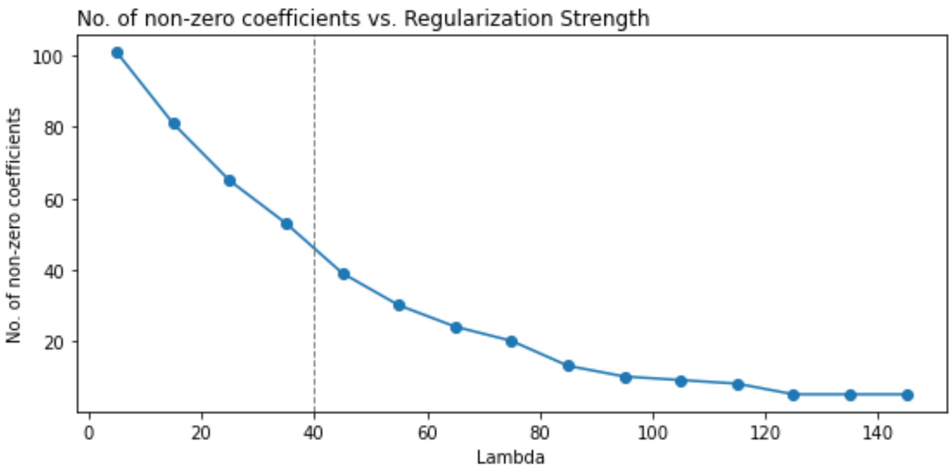
Basic Logistic Regression

As seen above in Table 1 above, we compare the results obtained from the logistic regression model and the L1 regularised model based on the ‘test’ data. The AUC, MLE Loss and Expected Payoff in the Basic Logistic Regression model are 0.686, 0.497 and 127.61 respectively. The logistic regression model we ran comprises of 116 features. By utilizing every feature in the dataset, such complex models is less intuitive and deters commercial use cases.

Tuning L1 Regularization Strength

To minimize the risk of overfitting in our model training, the use of L1 regularization helps to reduce the flexibility of the coefficients, and modify certaiize the risk of overfitting in our model training, the use of L1 regularization helps to reduce the flexibility of the coefficients, and modify certain coefficients to be 0. In other words, we omit some features in the model which would help to reduce overfitting as we focus only on the “more important” features that affect the loan status outcome.

By inputting different values of lambda, we tune the regularization strength, and we observe that the greater the value of lambda, the lesser the number of non-zero coefficients we are left with in the final model. A lambda value of 0 is equivalent to not regularizing.



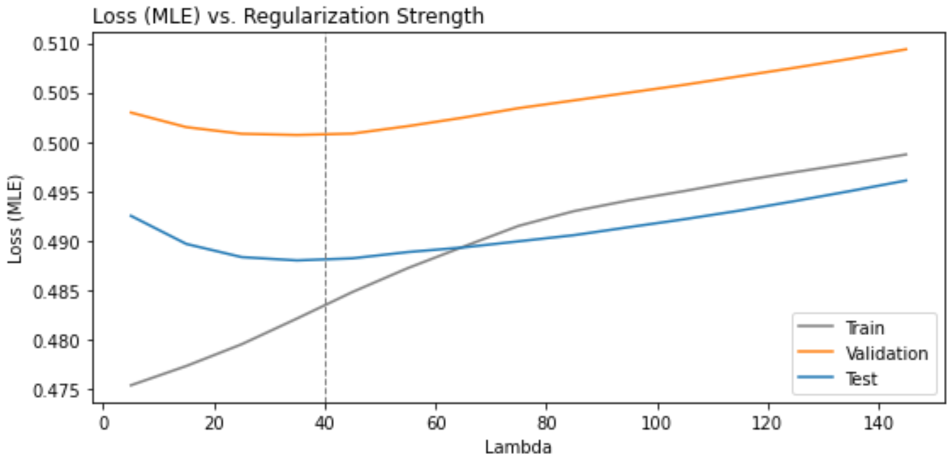
This can also be explained through the L1 cost function essentially comprising of:

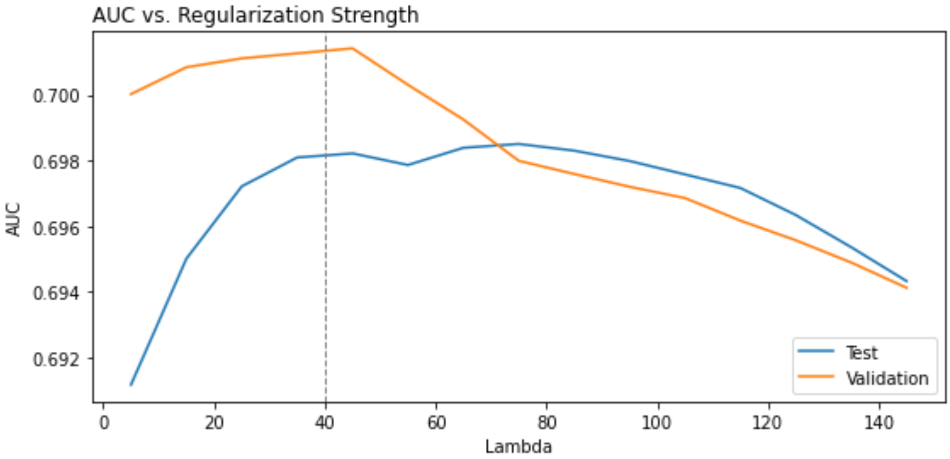
$$Training\ Loss + \lambda \sum_{i=0}^{m} |b_i|$$

In the above equation, lambda represents a hyperparameter that determines the strength of the penalizing term of the L1 regression model i.e., $\lambda \sum_{i=0}^{m} |b_i|$. When lambda is 0, the cost function is completely explained by the MLE loss, with our regression model being overly complex. On the other hand, as we increase lambda, we make our model simpler by selecting the more important features (whose beta coefficients do not shrink to zero).

Selecting The Optimal Lambda:

To find the optimal lambda that minimizes the risk of overfitting and underfitting, we compare the loss functions and AUC scores across different lambda values. The lambda value, which gives us the lowest MLE loss and/or highest AUC score, corresponds to the optimal lambda value for our training data. In this case, the optimal lambda corresponds to 40.

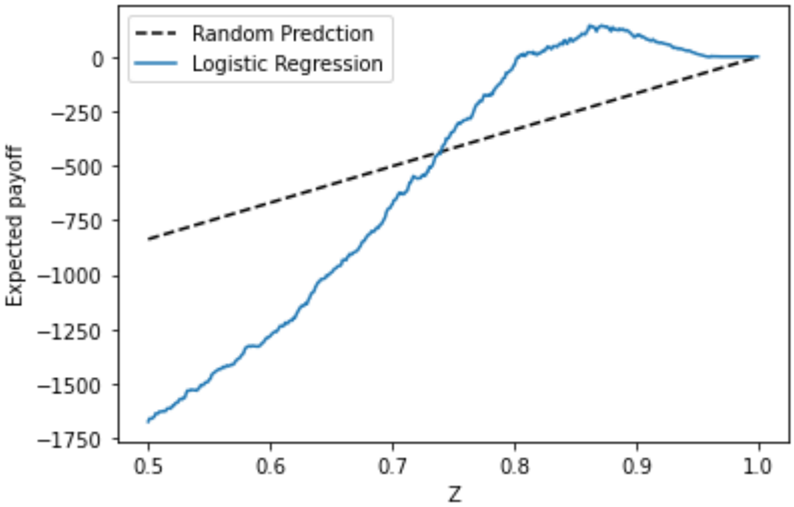




Furthermore, referring to Table 1, we can clearly observe how the through regularising the model, we have now managed to achieve a lower MLE loss whilst enhancing the expected payoffs. Furthermore, based solely on the test data, considering how the AUC has increased through L1 regularisation, we can say that the L1 regularisation has improved our model's predictability.

Expected Payoff

Selecting the optimal lambda addresses the issue of tuning the L1 model to best distinguish between the good and bad borrowers optimally. However, we still need to identify the probability threshold that assigns between good and bad borrowers (Z), conditional on maximizing expected payoff. Depending on the cost parameters, the payoff distribution will implicate the expected payoff on varying Z values. The result is visualized in the chart below, and we can observe that the highest expected payoff can be obtained by an optimal Z of 87.1%.



How should we use the model to determine which loans to accept?

From the model, we should accept people whose predicted probability of paying back fully (Q) is higher than the optimal Z (i.e., 87.1%) as the expected payoff will be greater than zero due to increased credit worthiness.

How should we use the model results to set interest rates?

The model that has been trained in this exercise is equivalent to a credit risk model, and we can use the model to evaluate the credit worthiness of each borrower. In the lending industry, it is common practice to set interest rate based on the credit worthiness of a borrower as a means of managing credit risk. The riskier the borrower, the higher the interest rate is set to offset the higher credit risk.

The following is a brief methodology on how LendingClub can evaluate the credit worthiness of each borrower based on the model trained during the earlier phase of this exercise.

Given a set of loan characteristics (i.e., final features used in the Lasso model), we can compute the probability of a borrower being a good borrower, Q. From this, the probability of default (PD) can be derived by taking complement (1-Q).

PD is a form of credit risk measure of the borrower – the higher the PD, the riskier the borrower is. The PD is a useful measure as it summarizes every piece of information of the borrower into a single number, since it is a function of the features used in the Lasso model, such as borrower's past loan activity, financial metrics, FICO scores etc.

With this, LendingClub will be able to evaluate the credit worthiness of each borrower based on the PD and set interest rates according to the borrower's risk profile.