Freddie Mac Cross Validation

Thomas Billman

1. Introduction

Freddie Mac holds a large portion of the United States of America's home mortgages. As such looking for trends in the values and performance of these loans is crucial. Our dataset consists of data collected at the time loans were issued and a calculated Net Present Value (NPV) that acts as a proxy for the total value of each loan. This paper summarizes a subset of the different variables found in the dataset. The dataset can be found on my GitHub at the following address: https://tinyurl.com/ycpr8lrj

2. KNN Selection of K

We begin this project by testing different values of K for accuracy. For this I decided to use 5-fold cross validation in order to evaluate choices of K by testing accuracy rather than training accuracy. I tested values of 1,3,5,7,9, and 11.

K	Accuracy
1	0.9806052
3	0.9806052
5	0.9806195
7	0.98061
9	0.98061
11	0.98061

This code took 37 minutes to run. This shows that all choices of K amongst this set have essentially identical accuracies. We will continue this analysis using K = 3.

3. 5-Fold Cross Validation

In this section we apply 5-fold cross validation to the Logistic, LDA, QDA, and KNN models previously selected.

Model	Acc	Sens	Spec	Run Time
Logistic	0.9812	1	0	10 s
LDA	0.9812	1	0	4.7 s
QDA	0.9747	0.9922	0.06339	4.2 s
KNN	0.9806	0.9978	0.08076	6.5 m

This shows that Logistic and LDA predicted all observations as positive NPV to achieve the greatest overall accuracy. However, I would pick KNN as a model given that it can predict negative NPVs at a value over four times their relative occurrence in the population. Additionally, the accuracies for these models can be found in Figure 1

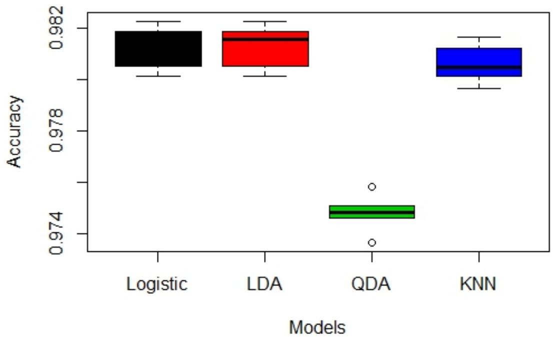

Figure 1. Accuracies for 5-Fold CV

4. Leave-One-Out Cross Validation (LOOCV)

We continue this cross validation by utilizing LOOCV. This is the least biased as it does not involve any randomization. However, it is also the most computationally intensive. Due to our large dataset it was necessary to utilize parallel computation techniques and High Performance Computing resources. The results of our LOOCV were as follows:

Model	Acc	Sens	Spec	Time to Run
Logistic	0.9812	1	0	1.1 days
LDA	0.9812	1	0	7.5 hours
QDA	0.9748	0.9922	0.07503	6.6 hours
KNN	0.9805	0.9972	0.09778	26 minutes

It is also worth noting that any boxplots for these accuracies would be degenerate as accuracies in LOOCV can only take values of 1 or 0. Due to this lack of interpretation, they are not included in this paper.

5. Conclusion

These models show that KNN was our best classifier for our data, particularly when utilizing LOOCV.