Freddie Mac Machine Learning

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. Logistic Regression

We begin our further analysis by predicting whether the NPV of an observation will be positive or negative. The covariates considered include Credit Score, Mortgage Insurance, Debt to Income ratio, among others. This model predicted that every mortgage would have a positive NPV, so it's accuracy measures were as follows:

Sensitivity	1
Specificity	0
Accuracy	0.98125

Additionally the ROC Curve can be found in Figure 1. The area under the curve was .6417, which is not very significant.

3. Linear Discriminant Analysis

We continue by applying Linear Discriminant Analysis. These results were the same as logistic regression, by predicting all rows as positive NPV.

Sensitivity	1
Specificity	0
Accuracy	0.98125

4. Quadratic Discriminant Analysis

Quadratic Discriminant Analysis begins predicting NPVs that are not all positive. The table of results is as follows:

	True -	True +
Predicted -	246	1587
Predicted +	3676	203619

This yields the following sensitivity, specificity, and accuracy:

Sensitivity	0.9922663
Specificity	0.0627231
Accuracy	0.9748336

It is worth noting that although a specificity of .06 seems poor, it is nearly 3 times more accurate than random guessing.

5. KNN Analysis

We continue through to KNN analysis, which produced even better specificity than QDA. This was the resulting confusion matrix:

	True -	True +
Predicted -	601	209
Predicted +	3321	204997

This yields the following sensitivity, specificity, and accuracy:

Sensitivity	0.1532381
Specificity	0.9989815
Accuracy	0.9831204

Of the four methods tested this had the best accuracy and impressive specificity for this dataset.

6. Cross Validated Results

We repeat the above methods using 5-fold cross validation.

6.1. Logistic Regression

Overall average accuracy was 0.981719 and took 15 seconds to run. All sensitivities were 1 and all specificities were 0.

6.2. LDA Analysis

Overall average accuracy was 0.9812459 and took 5 seconds to run. All sensitivities were 1 and all specificities were 0.

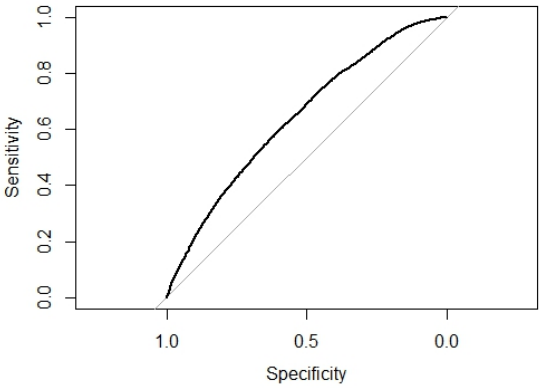

Figure 1. ROC Curve for Logistic Regression

6.3. QDA Analysis

Overall average accuracy was 0.9747953 and took 4 seconds to run. All sensitivities were .992 and specificities were between .049 and .075.

6.4. KNN Analysis

Overall average accuracy was 0.9806004 and took 9 minutes to run. All sensitivities were .998 and specificities were between .077 and .087.

7. Conclusion

These models show that QDA and KNN were better classifiers for our data than logistic or LDA. Additionally, cross validated results show lower accuracies than testing on your training data. However, this is to be expected.