Comparing Various Models to Improve Default Detection on Small-Loan Business

Roland DePratti
CCSU Data Science

1. Executive Summary.

Objective:

Our small loan business (loans equal or below $40,000) loses $499.81 per loan, due to the large amount of defaults. This project examines a set of 87,805 loans from our database, to improve the way the bank identifies loans that may default. We are hoping that better identification of potential defaults will turn this into a profitable business category. Five models were built and compared to meet that purpose.

Key Findings:

1. For every $1 we earn from a non-defaulted loan, we lose $3.30 from a defaulted loan. 26% of loans in the sample defaulted, representing 22,719 individual loans, with an average loss of $14,514 per loan. This represents a total loss of $329.7M, which offsets the $286.9M revenue we earned from the good loans and resulted in a net loss of $43.9M. This translates to a cost of $499.81 per loan. It is important that we fine-tune our selection process to better eliminate our loan defaults.

2. To develop a model to best identify potential defaulters, we used five algorithms: CART, C5.0, Treebag, Adaboost, and Random Forest. We used the CART algorithm in our earlier study. This work examines 4 additional algorithms that use an ensemble approach. Ensemble algorithms develop multiple models for the same problem and then compare results to generate a final prediction. This approach can provide more accurate results for some problems. C5.0 is an ensemble algorithm that uses CART as its starting point. The other 3 algorithms use a different starting point.

3. The loan data was adjusted to account for the fact that, as mentioned in item 1, a customer incorrectly classified as not a defaulter presents a larger financial risk than a good customer classified as a defaulter.

4. All models had an accuracy rate, which measures the accuracy of identifying both Defaulters and Non-defaulters, of approx. 50% or greater (CART(48.54), C5.0(53.7), Treebag(60.19), Random Forest(61.43), and Adaboost(62.59)). However, the CART-based models did a better job of identifying True Defaulters, CART(86.27) and C5.0(80.14). As mentioned earlier, missing defaulters provide the greatest financial risk. In our sample alone, applying one of the CART-based models would result in a decrease in default costs of 284.5M and 264.3M, respectively.

5. Incorporating the different costs of misclassifications, all tested models converted our calculated baseline cost per customer of $499.81 to a revenue per customer (CART(635.79), C5.0(678.64), Treebag(432.84), Random Forest(384.88), and Adaboost(397.10)). The CART and C5.0 models delivered better revenue per customer numbers (635.79 and 678.64). The C5.0 model improvement of $53 per customer over the CART model was due to identifying close to the same number of defaulters (18,207 vs 19,600), but doing a better job of identifying non-defaulters (28,478 vs 23,018). Based on the higher cost of a defaulted loan, applying the C5.0 model will result in a $103.5M gain to our revenue (59.6M revenue vs -43.9M loss). This improves the cost per customer from $499.81 cost per customer to $678 revenue per customer. This is a 236% improvement, making small-segment loans profitable.

Next Steps:

Incorporate the C5.0 developed model into the loan process.

2. **Comparison of results from** CART, treebag, adaboost, random forests, and C5.0 models**.**

	CART	**Treebag**	**Adaboost**	**RF**	**C5.0**
***TN**	23018	39446	42783	41425	28478
FP	42068	25640	22303	23661	36608
***FN**	3119	9318	10544	10207	4512
TP	19600	13401	12175	12512	18207
Accuracy	0.4854	0.6019	0.6259	0.6143	0.5317
Error Rate	0.5146	0.3981	0.3741	0.3857	0.4683
Sensitivity	0.8627	0.5899	0.5359	0.5507	0.8014
Specificity	0.3537	0.6061	0.6573	0.6365	0.4375
Precision	0.3178	0.3433	0.3531	0.3459	0.3322
Recall	0.8627	0.5899	0.5359	0.5507	0.8014
F_1	0.4645	0.4340	0.4257	0.4249	0.4697
F_2	0.6424	0.5158	0.4856	0.4924	0.6248
$F_{0.5}$	0.3638	0.3746	0.3790	0.3737	0.3762
Total Revenue	-55,825,890	-38,005,380	-34,867,320	-33,794,202	-59,588,208
Revenue per Customer	-635.79	-432.8384	-397.0995	-384.8779	-678.6425

3. Predicting the Holdout Validation Set.

Of the five algorithms, CART, C5.0, Treebag, Adaboost, and Random Forest, all developed models had an accuracy of approx. 50% or greater. However, the CART based models did a better job of identifying True Defaulters, CART(86.27) and C5.0(80.14).

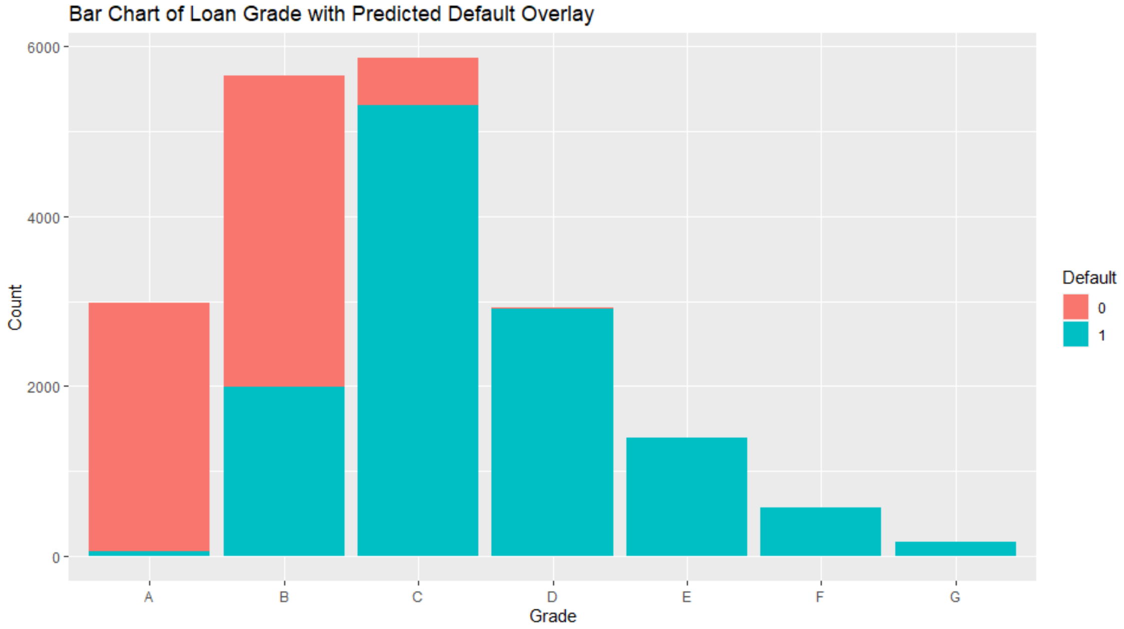
When taking misclassification costs into account, the CART and C5.0 models delivered better revenue per customer amounts ($635.79 and $678.64). Applying the C5.0 model will result in a $103.5M gain to our revenue (59.6M revenue vs -43.9M loss). This improves the results per customer from a $499.81 cost per customer to a $678 revenue per customer. This is a 236% improvement, making small-segment loans profitable. For those reasons, I applied the model developed using C5.0 to the holdout dataset.

The following steps were applied to process the holdout data set of 19,561 new loan records to predict their outcomes:

a) All continuous predictors were standardized to accommodate predictors that may have greatly varying data ranges.

b) The C5.0 algorithm was applied to the new dataset to predict defaulting. This algorithm is based on the CART algorithm but adds the boosting ensemble method.

The plot below compares the predictions generated by our new model to the grade that our current process assigned to each loan. It demonstrates that it matches closely to the assigned loan grade from the current process, while improving performance by being more selective on higher grades.


Bar Chart of Loan Grade with Predicted Default Overlay
Count
A
B
C
D
E
F
G
Grade
Default