KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

It's All About the Money: Maximizing a Profit Function Using Applied Logistic Regression

Victoria M. Chaney and Juliet M. Womack

Faculty Advisor: Dr. Jennifer L. Priestley

INTRODUCTION

Logistic regression is the modeling algorithm of choice in the finance and credit industry. Compared to more complex algorithms, like neural networks, logistic regression enables analysts, lenders, and consumers alike to understand and communicate about the model's outputs and implications. More specifically, logistic regression can be used to predict a 2-class outcome, such as not default or default. Using the probability values produced from a logistic regression model, the model can then be optimized to meet the goal(s) of the company. The present analysis focuses on using logistic regression to predict the credit risk of 1.2 million sub-prime consumers. These predictions are later used to find the optimal cut-off value to maximize a profit function resulting in an average profit of $113.12 per consumer.

METHODS

Datasets:

- **CPR (credit performance)** with 1.4 million observations and 339 predictors of credit performance.
- **PERF (post hoc performance)** with 17.2 million observations and 18 variables characterizing consumer performance after a credit product was given to them. The delinquency identifier variable (DELQID) and credit limit (CRELIM) variables were the only variables used from this dataset.
- Consumers were uniquely identified with the MATCHKEY variable.

Creating the Binary Dependent Variable (GOODBAD):

- The maximum DELQID of each consumer was used to categorize consumers into two classes, good or bad credit risk consumers, based on their "worst" behavior (see Table 1).
  - **Good credit risk consumers:** maximum DELQID ≤ 2
  - **Bad credit risk consumers:** maximum DELQID > 2

Variable Imputation and Reduction:

- Predictor variables with 40% or more missing/coded values were removed from the dataset to retain variables with "true" values (values that do not need to be imputed).
- For the remaining predictor variables, any value that was more than 5 standard deviations away from the mean was imputed with the median.
- **Variable clustering** grouped variables that represented the same concept (e.g., co-finance accounts, department store accounts, auto-finance accounts, etc.). The variable with the lowest 1-R^2 value in each cluster was retained.
- **Variation inflation factor analysis** was used to ensure the predictor variables retained from clustering were not redundant of one another.

Variable Transformations (see Table 2):

- Binning was used to establish monotonic relationships between each predictor variable and the binary dependent variable.
- **Monotonic relationships** are characterized by a consistent positive or negative relationship with the binary dependent variable (see Figure 1).
- Continuous and discrete variables (with 20 or more values) were binned into groups of equal frequency and equal width.
- Discrete variables with less than 20 values were collapsed to achieve monotonicity.
- An odds and a log odds transformation were applied to the variables once a monotonic relationship was established with the binary dependent variable.

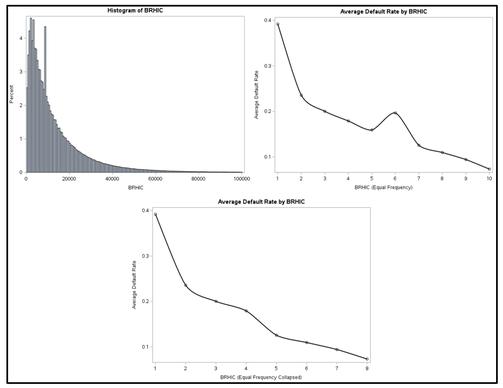
Logistic Regression:

- The data were split into 80% training, to build the model, and 20% validation, to quantify the model's performance on unseen data.
- Logistic regression models were run to identify the strongest form of each predictor variable using the Wald chi-square value to quantify strength. The form with the highest chi-square value was retained.
- The strongest predictor variables were then used to train a single logistic regression model and the variables were reduced to produce a parsimonious model (i.e., 5 to 7 variables).

Table 1: Distribution of GOODBAD by DELQID

	DELQID								
GOODBAD	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
0	751,874	226,001	56,954	0	0	0	0	0	1,034,829
1	0	0	0	31,390	24,419	19,425	20,257	125,109	220,600
Total	751,874	226,001	56,954	31,390	24,419	19,425	20,257	125,109	1,255,429

Figure 1: Monotonic Binning for BRHIC*



*BRHIC = Total high credit bank revolving accounts

Figure 2: ROC Curve for Final Model

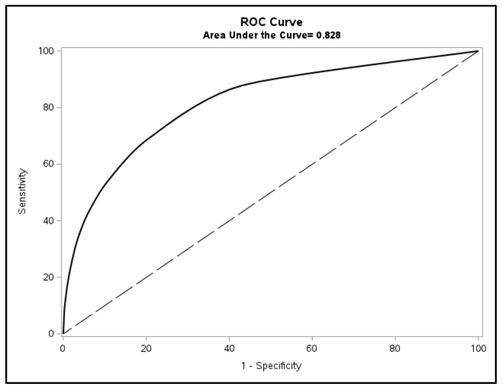


Figure 3: Profit Function Cut-off Value Optimization

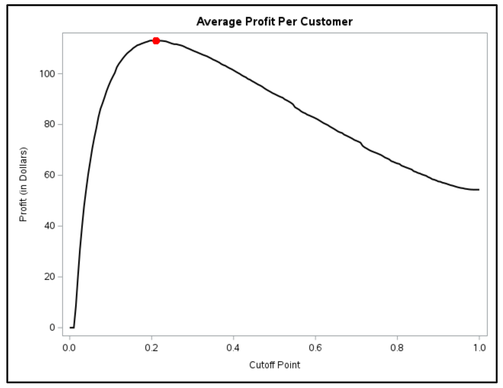


Figure 4: KS Curve for Final Model

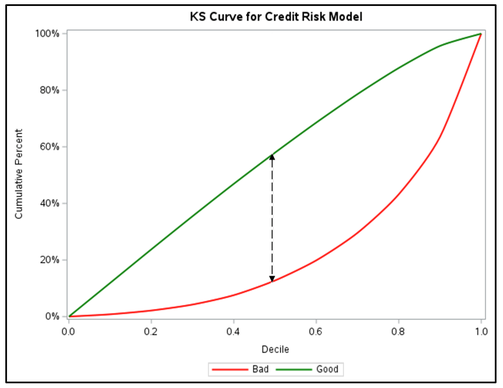


Table 2: Variable Transformations

Variable Type	Variable Forms
Continuous or Discrete (≥ 20 values)	1. Equal frequency 2. Equal frequency odds 3. Equal frequency log odds 4. Equal width 5. Equal width odds 6. Equal width log odds
Discrete (< 20 values)	1. Collapsed 2. Collapsed odds 3. Collapsed log odds
Binary / Effectively Binary	1. Odds 2. Log odds

Table 3: Predictor Variables in the Model

Parameter	Estimate	Chi-Square
Intercept	-5.3894	63732.7479
Number of Bank Revolving Accounts with > 75% High Balance (Equal Width; Odds)	3.8879	11579.5773
Number of Bank Revolving Accounts Currently Satisfactory (Equal Frequency; Odds)	2.0225	10843.8944
% of Satisfactory Accounts to Bank Revolving Accounts (Equal Width)	-0.1557	15722.7202
% of Open Trades in 24 to Total Open Trades (Equal Width)	0.2559	10529.5782
Total Balance for Revolving Trades Reported Within 6 Months (Equal Width)	0.3665	17715.0134
Number of Accounts in 90+ Days in the Past 24 Months (Binary)	0.6628	7421.0677
Number of Trades in 30 or 60 Days (Binary)	0.7168	11035.4937

Equation 1: Profit Function Used to Maximize Profit (Cut-off = 0.21)

$$profitability = (\#of\ True\ Negatives)(\$250) - (\#of\ False\ Negatives)(\$1{,}123)$$

Table 4: Confusion Matrix for Final Model (Cut-off = 0.21)

Actual	Predicted		
	Good Credit Risk	**Bad Credit Risk**	**Total**
Good Credit Risk	169,106	37,879	206,985
Bad Credit Risk	14,799	29,441	44,240
Total	183,905	67,320	251,225

CONCLUSIONS

- The present analysis demonstrates how binary logistic regression can be applied to a real-world business problem to help lenders maximize their profit.
- The final model had 7 predictor variables (see Table 3).
- Percent concordance, which was equal to the area under the curve (AUC; see Figure 2), was used to quantify the model's performance.
  - The **percent concordance** was 82.8%. This indicates that 82.8% of the good-bad pairs had the good credit risk consumer with the lower probability of default and the bad credit risk consumer with the higher probability default.
- **AUC** indicates the model's ability to separate the predictions of the two classes. The model is capable of correctly classifying a random consumer as a bad credit risk approximately 83% of the time.
- The profitability function was provided by the stakeholder (see Equation 1):
  - **True negatives**, predicted good credit risk consumers who are truly good credit risk consumers, yields $250 in profit.
  - **False negatives**, predicted good credit risk consumers who were actually bad credit risk consumers, yields -$1,123 in profit.
  - True positives and false positives do not yield any profit, and thus, are not included in the profitability function.
    - **True positives** are predicted bad credit risk consumers who are actually bad credit risk consumers. These consumers are not extended credit, and thus, no profit is lost or gained.
    - **False positives** are predicted bad credit risk consumers who are actually good credit risk consumers. These consumers are likely to go to a competitor lender.
- Evaluating the profitability function at cut-off values from 0 to 1 in increments of 0.005, resulted in the profitability function being maximized when the cut-off value = 0.21 (see Table 4 and Figure 3).
  - In other words, consumers who have a probability of default at 0.21 or below should be extended credit to maximize profit.
  - The profitability function, when maximized, results in an average profit of $113.12 per consumer.
- The **KS Curve** (see Figure 4) indicates the model can distinguish between 49.3% of good and bad credit risk consumers.
- Depending on the financial goals of the lender, this model is capable of yielding profit for the lender.
- The model can correctly classify 66% of bad credit risk consumers and 81% of good credit risk consumers. Different cut-off points, in the future, could be tested to lower the number of misclassified good credit risk consumers, but may result in a loss of profit overall.