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

Mail. Model. Money: Two-Stage Targeting in Credit Mail Campaigns

Adele Barski and Jayme Perry
Applied Binary Classification (Dr. Gene Ray)

INTRODUCTION

Background

- Direct mail remains a valuable marketing channel in financial services, offering a tangible connection that digital ads often lack
- Despite its benefits, direct mail campaigns have low average response rates (~1%) and high costs, making targeting critical
- Financial institutions use predictive models to identify consumers likely to respond and manage credit reasonably
- Traditional credit scores offer limited insight; custom models using bureau data and behavioral indicators allow for more precise risk assessments
- Poor targeting increases the chance of defaults and wasted marketing spend, directly impacting profitability

Objective

- Use logistic regression to predict both response likelihood and credit risk in a unified framework
- Increase campaign ROI by targeting low-risk, high-response consumers

METHODS

Data Source

- Data provided by Atlanticus, containing 988,267 observations and 541 variables related to consumer credit behavior, including both non-responders and trade performance outcomes
- The final model was applied to an unseen out of time dataset also provided by Atlanticus

Data Cleaning & Pre-Processing

- **Missing Data Treatment:** Coded missing values (e.g., 9999s) were replaced with system missing values
- **Variable Filtering:** Variables missing more than 30% of values were removed
- **Imputation:** Median imputation was performed for all remaining missing numeric variables

Feature Engineering & Selection

- **Dimensionality Reduction:**
  - *Clustering*: Variable clustering reduced the predictor set to 78 cluster representatives
  - *Multicollinearity*: Variables with a VIF > 5 were excluded, resulting in 74 predictors explaining ~ 80% of variation in the data
- **Discretization:** Quantile-based binning was applied to improve model interpretability and stability, and to ensure monotonicity

Target Variable Construction

- Two binary outcomes were engineered:
  - GoodBad Response: Indicates response to the credit mailer
    - (1 = responded, 0 = did not respond)
  - GoodBad Credit: Indicates credit performance
    - (1 = defaulted, 0 did not default)

Data Partitioning

- The dataset was randomly split into:
  - 70% training, 15% testing, and 15% validation

Model Development

- Stepwise logistic regression was performed separately for both response and credit models using the training set
- Optimal thresholds were selected using F1 score analysis to balance precision and recall

Profitability Analysis

A rule-based profit function was designed by assigning dollar values to outcomes:

- Mailed, approved, and non-defaulting responder: **+$250**
- Mailed, approved, and defaulting responder: **-$600**
- Mailed, no response: **-$10**

Table 1: Frequency Distribution of Customer Responses to Mail Campaign

Customer Response	Frequency	Percent	% Ex. Non-Responses
Did Not Default	153,729	15.56%	62.73%
Defaulted	91,340	9.24%	37.27%
Did Not Respond	743,198	75.20%	
Total	988,267	100%	100%

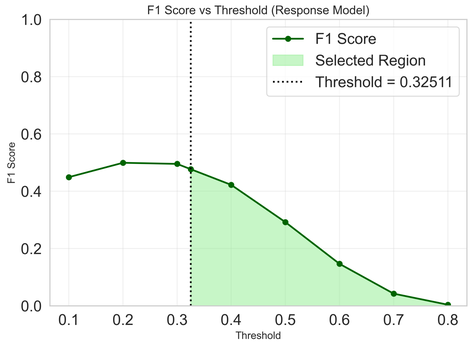

Figure 1: F1 Score vs Threshold for Classification Mailer Response Model

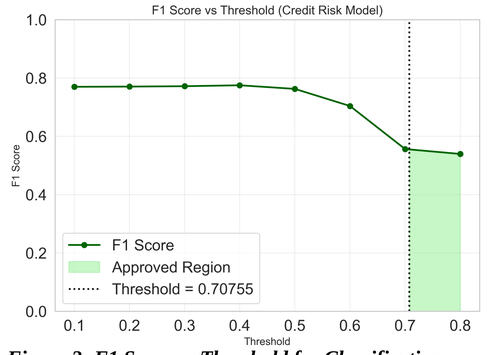

Figure 2: F1 Score vs Threshold for Classification Credit Risk Model

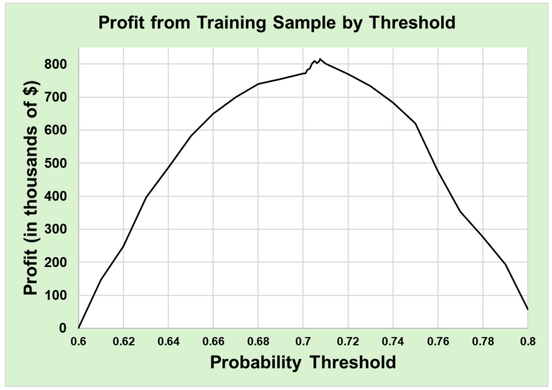

Figure 3: Profit-Threshold Curve Credit Risk Model

Figure 4: Percentage of Non-Defaulters Approved for Credit on Out of Time Sample

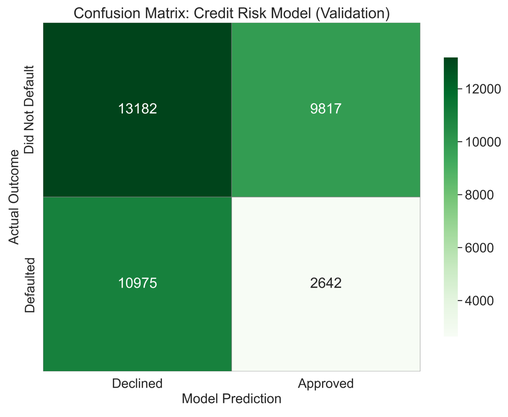

Figure 5: Confusion Matrix for Credit Model Outcome Validation Set

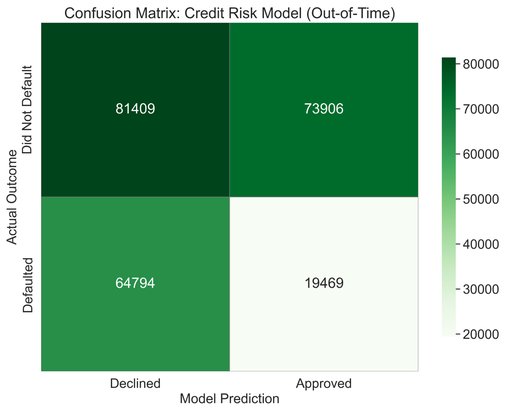

Figure 6: Confusion Matrix for Credit Model Outcome Out of Time Sample

RESULTS

Descriptive Analysis

- Table 1 shows the frequency distribution of the Binary target variable Customer Response. Of the ~25% of individuals who responded to the mailer, 62.73% did not default, while 37.27% did

Predictive Analysis

- **Variable Selection**
  - A total of 45 unique predictors were used across both models:
  - Response Model - 40 Predictors
  - Risk Model – 31 Predictors

  (Some variables used in both models)
- **Classification Thresholds**
  - Response Model: If probability of response >= 0.32511, the individual is selected to receive a mailer. (Figure 1)
  - Risk Model: If probability of not defaulting >=0.70755, the individual is approved for credit. (Figure 2)
- **Money Making**
- Validation profit: $184,570
  - Response AUC: .739 Risk AUC: .684
  - Precision: .789
- Out of time sample profit : $1,370,530
  - Response AUC : .747 Risk AUC: .682
  - Precision: .791 (Figure 4)

CONCLUSIONS

- Logistic Regression provides high interpretability and transparency, the primary reason it is used frequently in modeling credit data.
- More complex models may provide more accuracy in prediction but lack interpretability that logistic regression offers.
- Cost of simplicity- when building a parsimonious model, there is always a trade off with predictive power and high explainability.
- The model takes a deliberately conservative stance, prioritizing loss prevention over approval volume. This approach is aligned with our business objective: minimize downside risk in a lending environment where defaults are costly.
- Even though the model only caught about 21% of all defaulters, its real strength lies in precision, correctly approving ~79% of non-defaulters, avoiding costly errors.
- Looking ahead, this model can serve as the foundation for a more layered strategy, potentially combining it with alternative data, non-linear modeling, or adaptive thresholds to increase approval rates while maintaining responsible credit practices.

Connect with us on LinkedIn!

Adele Barski **Jayme Perry**