DS 5220 Project: Predicting Credit Default

EJ Wong and Ryan Houseman

April 22, 2025

1 Milestone 1

1.1 Step 1A

Initially we look to complete an exploratory data analysis of the data sampled from to understand basic trends and balance of the target outcomes. Since this is a classification problem with a binary outcome we look to initially preform a multinomial logistic regression with our co-variates. We will include a train, test, and validation split using k-fold cross validation. To improve our model and minimize the effect on unimportant predictors we will implement Least Absolute Shrinkage and Selection Operator (LASSO).

1.2 Step 1B

Based on information that we have not worked on previously in class and that we have not learned, we look to implement two other methods for this binary classification task. The first is stepwise regression for logistic regression which will assist in feature selection of the predictors. The second will be a decision tree for classifiers which will be evaluated with standard classification metrics.

1.3 Step 1C

1.3.1 Data Sources:

Data Source: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients
Literature Review Used to discuss data: https://doi.org/10.1016/j.eswa.2018.01.012

1.3.2 Target Variable:

Default Payment Next Month (Binary Categorical Variable): Outcome of a default payment for the next month (1 if default, 0 if not)

1.3.3 Co-variates:

Data and descriptions taken from Yeh (2009) and Asuncion Newman (2007).

LIMIT BAL: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit

SEX: Gender (1 = male; 2 = female).

EDUCATION: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

MARRIAGE: Marital status (1 = married; 2 = single; 3 = others).

AGE: Age (year).

PAY 0-6: History of past payment. Authors tracked the past monthly payment records (from April to September, 2005) as follows: PAY 0 = the repayment status in September, 2005; PAY 2 = the repayment status in August, 2005;...; PAY 6 = the repayment status in April, 2005. The measurement scale for the repayment status is: 1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; ...; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

BILL AMT 1-6: Amount of bill statement (NT dollar). BILL AMT 1 = amount of bill statement in September, 2005; BILL AMT 2 = amount of bill statement in August, 2005;...; BILL AMT 6 = amount of bill statement in April, 2005.

PAY AMT 1-6: Amount of previous payment (NT dollar). PAY AMT 1 = amount paid in September, 2005; PAY AMT 2 = amount paid in August, 2005;...; PAY AMT 6 = amount paid in April, 2005.

2 Milestone 2

Please see r code files in repo for this along with canvas submission. In this case we implemented a LASSO logistic regression.

3 Milestone 3

3.1 Model Used

Decision Tree with Cross Entropy Loss

3.2 Math Derivations

Cross Entropy Loss Function: $-\sum_{0,1} p(x)log(q(x))$ where $p(x)$ is the true probability distribution and $q(x)$ is the predicted probability distribution.

Minimization (Maximization in this case as we want to maximize the "loss"):

$\frac{\partial}{\partial q(x)} - \sum_{0,1} p(x)log(q(x)) = -\sum_{0,1} \frac{p(x)}{q(x)}$

3.3 Code

Please see r code files in repo for this.

4 Repository

Link to the repository: `https://github.com/edwardjwong08/ds5220_project`

5 Bibliography

Asuncion, A., Newman, D. J. (2007). UCI machine learning repository. Irvine, CA: School of Information and Computer Science, University of California. https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

Hongliang He, Wenyu Zhang, Shuai Zhang (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. Expert Systems with Applications, Volume 98, Pages 105-117, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.01.012.

Yeh, I-Cheng. "Default of Credit Card Clients." UCI Machine Learning Repository, 2009, https://doi.org/10.24432/C55S3H.