Christy McCanless

Customer default identification report

Credit One Data Team

**Table of Contents:**

Summary

The Data

The Process

The Results

**Summary:**

Businesses rely on Credit One to assess the credit worthiness of their customers seeking loans. A high rate of loan defaults costs our customers thousands of dollars, weakens our business relationships with them and ultimately damages our reputation in the industry.

Credit One’s Data Science team has been tasked with evaluating a sample of our customer’s data set containing known account defaults. Better detection and earlier identification of conditions that lead to default would enable us to make better recommendations to our customers.

The data set contained the information for 30,000 accounts with 6,636 or 22% of those accounts in default. We tested several predictive models with three variations of the data set and yielded an accuracy of 83% with the best model / data set combination. This means 5,508 of the defaults would have been identified and not recommended for loans. In other words, we would have predicted 1,128 of the given accounts would default which is a default rate of 4%.

**The Data:**

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. All values are in NT dollars (New Taiwan dollars).

This study reviewed the literature and used the following 23 variables as explanatory variables:

X1 = Amount of given credit ($): includes individual credit of consumer and their family (supplementary).

X2 = Gender (1 = male; 2 = female).

X3 = Education (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others).

X4 = Marital status (1 = married; 2 = single; 3 = divorce; 0=others).

X5 = Age (year).

X6 - X11: History of past payment. The measurement scale for the repayment status is: -2: No consumption; -1: Paid in full; 0: The use of revolving credit; 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. We tracked the past monthly payment records (from April to September, 2005) as follows:

X6 = repayment status in September, 2005;

X7 = repayment status in August, 2005;

X8 = repayment status in July, 2005;

X9 = repayment status in June, 2005;

X10 = repayment status in May, 2005;

X11 = repayment status in April, 2005;

X12-X17: Amount of bill statement:

X12 = amount of bill statement in September, 2005;

X13 = amount of bill statement in August, 2005;

X14 = amount of bill statement in July, 2005;

X15 = amount of bill statement in June, 2005;

X16 = amount of bill statement in May, 2005;

X17 = amount of bill statement in April, 2005;

X18-X23: Amount of previous payment:

X18 = amount paid in September, 2005;

X19 = amount paid in August, 2005;

X20 = amount paid in July, 2005;

X21 = amount paid in June, 2005;

X22 = amount paid in May, 2005;

X23 = amount paid in April, 2005;

Y = client's behavior; Y=0 then no default, Y=1 then default

**The Process:**

Original Data Set

1. Import necessary libraries and data set
2. Pre-Process data: Keep all features and observations as given
3. Set feature matrix and dependent variable
4. Create testing and training sets for feature matrix and dependent variable
5. Build and fit models: KNN - K Nearest Neighbors Classifier; RF – Random Forest Classifier; GB – Gradient Boosting Classifier; LOG – Logistic Regression (as Classifier); SVC – Support Vector Classifier
6. Make predictions

The Process: Modified Data Set

1. Import necessary libraries and data set
2. Pre-Process data:
   1. Remove observations with credit\_limit over $400,000
   2. Re-code the four possible responses for ‘other’ education to one response
3. Set feature matrix and dependent variable
4. Create testing and training sets for feature matrix and dependent variable
5. Build and fit models: KNN - K Nearest Neighbors Classifier; RF – Random Forest Classifier; GB – Gradient Boosting Classifier; LOG – Logistic Regression (as Classifier); SVC – Support Vector Classifier
6. Make predictions

The Process: Original Data Set with Backward Elimination

1. Import necessary libraries and data set
2. Pre-Process data: Keep all features and observations as given
3. Set feature matrix and dependent variable
4. Backward Elimination:
   1. Keep all features and observations as given
   2. Create regressor
   3. Fit model with all possible predictors
   4. Remove features with p-value over 0.05
   5. Total of 13 features removed
5. Create testing and training sets for feature matrix and dependent variable
6. Build and fit models: KNN - K Nearest Neighbors Classifier; RF – Random Forest Classifier; GB – Gradient Boosting Classifier; LOG – Logistic Regression (as Classifier); SVC – Support Vector Classifier
7. Make predictions

**The Results:**

Features remaining after Backward Elimination:

X1 = Amount of given credit ($): includes individual credit of consumer and their family (supplementary)

X2 = Gender (1 = male; 2 = female)

X3 = Education (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others)

X4 = Marital status (1 = married; 2 = single; 3 = divorce; 0=others)

X5 = Age (year)

X6 = repayment status in September, 2005

X7 = repayment status in August, 2005

X8 = repayment status in July, 2005

X12 = amount of bill statement in September, 2005

X18 = amount paid in September, 2005

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy of Predicted Models** | | | |
| Model | Original | Modifications | Backward Elimination |
| RF | 79.4% | 79.6% | 79.8% |
| SVD | 77.5% | 77.2% | 78.4% |
| GB | 80.6% | 81.0% | 83.0% |
| LOG | 77.6% | 77.3% | 78.4% |

**Conclusion and Recommendations:**

During our analysis, we found the following insights:

1. Strong predictors include: credit\_limit, education, marital status and gender
2. In the month prior to default, the payment status, amount billed and amount paid strongly predict account default
3. The customer credit score was not included but is a data point of interest

Using predictive models seem to improve the ability to detect potential defaults for our customers. We recommend that we use our model on a larger data set and, after confirming similar results to those found here, implement this model for future client evaluation.