

Executive Summary

Predicting Missed Payments from Credit Card Clients

Team Members

Juergen Kritschgau and Song Gao

Overview

Credit card clients miss payments on their credit card debt for a variety of reasons. Being able to predict missed payments would allow banks, credit raters, and debt collectors to forecast their own operations, target interventions or financial products, and accurately appraise the value of credit card debt. In this project, we attempt to use a client's payment history over a 6 month and demographic information to predict whether the client will miss a credit card payment next month. We used data obtained from a Kaggle competition page to train different classification models, including logistic regressions, Bayesian models, Support Vector Classifiers, K-nearest neighbor models, and decision trees. We used cross-validation and classification accuracy to compare different classification models. Our primary finding is that no model is able to accurately predict whether or not a client will miss a payment.

Data set

We used the "Default of Credit Card Clients" data obtained from <https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients>. This data is publically available for use under the CC BY 4.0 license. This data has 30,000 instances where each row corresponds to the payment history and demographic information of a credit card client. A notable feature of this data set is that about 79% of credit card clients do not miss a payment, which is important for benchmarking our classification algorithms.

The features in this data set include:

- Age
- Sex
- Education
- Marital Status
- Credit Limit
- Payment Status over 6 months
- Bill Amounts over 6 months
- Payment Amounts over 6 months

We engineered features for:

- Credit Usage over 6 months defined as $(\text{Bill Amounts})/(\text{Credit Limit})$
- Proportions of Bill Paid over 6 months defined as $(\text{Payment Amount})/(\text{Bill Amount})$

Methods

Exploratory data analysis suggests that credit card clients who miss a payment have a declining payment status in the months prior to missing a payment. Payment status in the data essentially tracks how many payments behind a client is, suggesting that clients who have missed payments are more likely to miss payments in the future.

We trained a number of different classifiers, including logistic regressions, Bayesian models, Support Vector Classifiers, K-nearest neighbor models, and decision trees. In order to compare these models, we used 5-fold cross validation and accuracy (percentage of correctly predicted missed payments) as our primary performance measure.

Results

Our primary result is that none of these models performed particularly well in predicting missed payments. Support Vector Classifiers performed best, with an accuracy of 82%. However, given the class imbalance between clients who will miss a payment (21%) and clients who will not miss a payment (79%), scoring 82% is not impressive. In particular, classifying all credit card clients as not missing a payment would result in 79% accuracy and provides no insight into predicting missed payments.

Future Iterations

In the future, it may be fruitful to consider the probability with which a client will miss a payment rather than binary classification. Progress on this problem might also be made by attempting to detect clusters or similar groups of clients with different repayment behaviors.

There is also a potential shortfall in the data, where having more information about a client's credit card usage could improve predictions. For example, it could be helpful to know what kinds of products and services a credit card client is paying for with their credit card.

In a similar vein, more feature engineering or other machine learning models may also improve the classification results obtained with this data set.