Datamining Final Project Report

Bank Loan Data Prediction Using Machine Learning

Presented by

Team 5

Shang-Yi Cheng, Ting Yi Huang, Marcel Reitzle,

Tobias Michael Schäfer, Karan Sharma, Yen-Chun Tseng

Submitted to the Data Mining (I)

Prof. Dr. Christian Bizer University of Mannheim

**Abstract**

For a Bank in this century it is crucial to decide to which conditions loans can be given to which costumers or to whom to give a loan at all. To support this decision-making process, a bank might use information on the specific costumer in comparison to the characteristics of former costumers. The Bank in our case has already collected a dataset on costumer information. Unfortunately, the size of this dataset makes it impossible for a human analyst to come up with a significant classification without the use of modern data-mining techniques. Therefore, we used different classifiers to predict if a costumer pays back the loan he receives. Based on this and the specific customer profile, we could give a recommendation to the bank whether the bank should give the costumer a loan or not.

1 Application Area and Goals

Credit and default risks have been in the forefront of financial problems since the loan system came out. All the banks are trying to figure out effective business strategies to persuade customers to apply their loans. For this project, we use the analytical models we have learned in class to devise recommendations to potential lenders to help bank decide whether or not to fund new loan applications.

For us it would be of interest to figure out the specific dependencies between the bank loan characteristics. Meaning we will apply algorithms to find correlations between the loan status (“Fully Paid”, “Charged Off”) and the data on the consumer who received the loan. It would be interesting for us to figure out which consumer types guarantee to pay back the loan fully.The outcome we would like to produce will be characteristics or combinations of those that have a positive or that have a negative influence on the loan status. In other words, we want to predict in future who will be able to pay off their debt.

**Data mining process consist of three phases:**

* Data preparation

In this step, the data preparation process will take up to 90% of our time in the project. Once we identify the data sources, then we need to select, clean, construct, and format the data.

* The actual mining

We will aim to implement in-class taught algorithm such as K-Nearest-Neighbors, Naive Bayes, SVM, ANN, Decision Tree and RandomForest.

* Interpretation of the results

This step we will use the result of our evaluation methods and the scores as feedback. We will determine the best choice of preprocessing methods and features to use.

2 Data set

2.1 Data collection

For both training and testing datasets we will use kaggle as a source. The dataset called “Dataset for Bank Loan Prediction” will be the basis for our data mining approaches. It contains 19 variables:

* Loan ID
* Customer ID
* Loan Status
* Current Loan Amount
* Term
* Credit Score
* Annual Income
* Years in current job
* Home Ownership
* Purpose
* Monthly Debt
* Years of Credit History
* Months since last delinquent
* Number of Open Accounts
* Number of Credit Problems
* Current Credit Balance
* Maximum Open Credit
* Bankruptcies
* Tax Liens

Except for two of them, the ‘Credit Score’ and ‘Annual Income’, which are both at 80%, every column is populated to an extend of nearly 100%. The content of the columns can give various information on the loan of bank customers. Exemplary, it gives information on the customer’s years in current job and his/her annual income. Moreover, there are numbers on the “current loan amount”, which is the money that the customer currently owes the bank or the “monthly dept” of the costumer. We also find strings on the “home ownership”, the “purpose” of the loan and the “bankruptcies”. Moreover, there is a highly interesting attribute, which is the column named ‘credit score’. It is calculated out of different other scores like the payment history, amount owed, length of history and the types of credit used. All parts are accumulated with different weights to form the credit score.

2.2 Data cleaning

We want to either exclude or rewrite some of the columns or features that show too many blank and unreasonable data. For example, if the data is supposed to be a positive number, but it is recorded as negative number.

2.3 Data Selection

Once we identify the data sources, then we need to select, clean, construct, and format the data from the available data. For example, we can drop the data that isn’t necessary like ‘Loan ID’ and ‘Customer ID’.

2.4 Data Transformation

In this process, we have to transform and consolidate the data into different types that's suitable for mining. Normally this process includes normalization, aggregation, generalization, etc. After data integration, the available data is ready for data mining.

3 Preprocessing and Mining

* Describe approaches and parameter settings that you tried
* including evaluation setup and evaluation results
* including discussion of the results

4 Conclusion