Expert Systems with Applications 40 (2013) 5125–5131

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Consumer credit risk: Individual probability estimates using machine learning

Jochen Kruppa [a,1], Alexandra Schwarz [b,1], Gerhard Arminger [b], Andreas Ziegler [a,*]

[a] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Maria-Goeppert-Str. 1, 23562 Lübeck, Germany
[b] Schumpeter School of Business and Economics, University of Wuppertal, Gaußstraße 20, 42097 Wuppertal, Germany

ARTICLE INFO

Keywords:
Probability estimation
Random forest
Credit scoring
Probability machines
Logistic regression
Machine learning

ABSTRACT

Consumer credit scoring is often considered a classification task where clients receive either a good or a bad credit status. Default probabilities provide more detailed information about the creditworthiness of consumers, and they are usually estimated by logistic regression. Here, we present a general framework for estimating individual consumer credit risks by use of machine learning methods. Since a probability is an expected value, all nonparametric regression approaches which are consistent for the mean are consistent for the probability estimation problem. Among others, random forests (RF), k-nearest neighbors (kNN), and bagged k-nearest neighbors (bNN) belong to this class of consistent nonparametric regression approaches. We apply the machine learning methods and an optimized logistic regression to a large data set of complete payment histories of short-termed installment credits. We demonstrate probability estimation in Random Jungle, an RF package written in C++ with a generalized framework for fast tree growing, probability estimation, and classification. We also describe an algorithm for tuning the terminal node size for probability estimation. We demonstrate that regression RF outperforms the optimized logistic regression model, kNN, and bNN on the test data of the short-term installment credits.

1. Introduction

Credit scoring systems are an integrative part of companies risk management as they aim for preventing bad debt loss by identifying, analyzing, and monitoring customer credit risk (Brigham, 1992; Johnson & Kallberg, 1986). To measure the default risk involved by sales on credit, customers are assigned to certain risk classes based on their individual propensities to default on payment. The default probability can either be obtained externally or on basis of an internal scoring model. The main internal source of information on creditworthiness is a company's own accounting department which can provide data on a customer's previous payment behavior and individual characteristics, such as age, education, profession, and residence. Companies can also turn to commercial credit agencies which collect consumer data on criteria, such as unpaid bills, requests to pay issued by court order, enforcement procedures, and uncovered checks. These criteria normally serve as knock-out criteria as they deliver outright facts on a consumer's propensity to default on payment.

All or part of such data is used to construct a credit scoring model for predicting the default probability of new credits. The standard approach for estimating these probabilities is the logistic regression (logReg) model (Crook, Edelman, & Thomas, 2007; Thomas, Edelman, & Crook, 2002). However, several assumptions underlie these methods, and these assumptions are rather strict (Malley, Kruppa, Dasgupta, Malley, & Ziegler, 2012). First, all the important variables and supposed interactions must be entered correctly in the model. Otherwise, problems of model misspecification can arise. Second, the standard logReg model cannot deal with multicollinearity, i.e., high correlation between independent variables.

An alternative to the parametric logReg model are machine learning approaches, including neural networks, classification trees, random forests, or support vector machines (Arminger, Enache, & Bonne, 1997; Baesens et al., 2003; Brown & Mues, 2012), but they are used for classification in credit scoring only. However, default probabilities provide more detailed information about the creditworthiness of consumers than a binary or multicategory classification on creditworthiness.

In this paper, we apply machine learning methods to estimate the required default probability to a large data set of short-termed installment credits. We use the following trick to obtain consistent probability estimates. We embed the problem of estimating default probabilities into nonparametric regression. Any nonparametric regression approach yielding consistent estimates for the regression problem will also yield consistent estimates for the probability estimation problem, and these learning machines are called probability machines (Malley et al., 2012). The number of

* Corresponding author. Tel.: +49 451 500 2780.
E-mail address: ziegler@imbs.uni-luebeck.de (A. Ziegler).
[1] Equal contribution.

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.03.019

probability machines is large, and we concentrate on Random Forests (RF), k-nearest neighbors (kNN), and bagged k-nearest neighbors (bNN) because they are computationally fast, simple to implement, and they have already demonstrated their good performance in other problems (Kruppa, Ziegler, & König, 2012; Malley et al., 2012).

In the next section, we describe the data used in this article and the tuned logReg model (Section 2.1). The concept of probability estimation using machine learning relies on nonparametric regression and consistency of parameter estimates, which are both considered in Section 2.2. RF for probability estimation are based on probability estimation trees (PET). They are termed RF-PET and considered in Section 2.3. An important aspect of the RF-PET is the terminal node size, and we describe a tuning approach for the terminal node size. kNN and bNN are introduced in Section 2.4. We sketch methods for comparing the different approaches for estimating default probabilities (Section 2.5). Results of our analyses are shown in Section 3. Specifically, we demonstrate that RF-PET outperforms the optimized logReg model, kNN, and bNN on the test data of the short-termed installment credits.

Table 1
Description of variables of the consumer credit data set.

Variable	Values/levels
Date of delivery	Date
Date of first installment	Date
Applicant has ordered before, i.e., applicant is already customer	Yes; no
Residential region (larger regions, such as province or federal state)	Unordered with values 1–5
Type of telephone connection provided	Landline; cell/mobile
Gender	Female; male
Age (years)	Numeric
Delivery to different address (other than residential address)	Yes; no
Requested daytime of delivery	Afternoon; morning; none (N)
Delivery upon telephone appointment	Yes; no
Payment instrument for installments	Bank direct debit; postal slips
Credit card used for downpayment	Yes; no
Financed amount (euros)	Numeric
Downpayment made at order date (euros)	Numeric
Downpayment made at delivery (euros)	Numeric
Amount of each installment (euros)	Numeric
Number of installments (months)	13, 14, or 15

2. Methods

2.1. The data

The data set originates from a company which produces household appliances and offers its customers payment by installments. The company's management aims at improving the internal control of the cash flows from its accounts receivable portfolio because installment credits involve an increasing amount of bad debt loss. Approximately 13% of the total financed amounts remain uncollectible. It has to be noted that this unique data set is not distorted by a credit scoring system. This means that no systematic screening of the customers' credit standing had been implemented until the date of data retrieval, and all applications for installment purchase were accepted. Thus, we not concerned with problems of reject inference (Hand, 1998; Verstraeten & Van den Poel, 2005). The company operates in a Southern European country, where companies do not have access to individual information on customers' credit standing, employment status, or income, which usually serve as the most powerful predictors of consumer credit default (Bonne, 2000; Schwarz, 2008).

The data set consists of 64,524 installment purchases of household appliances, each of them paid off by 13, 14, or 15 consecutive monthly installments. The dependent variable y_i is the default event, being 1 if a customer defaults on payment, and 0, otherwise. In total, there are 9,155 (14.19%) default events. For the analysis, we randomly split the data into 2/3 training and 1/3 test data. In addition, we removed 3,365 installments from the data due to missings.

The potential covariates are displayed in Table 1, and they describe customer characteristics and their credit applications. The inference to be drawn from qualitative information may not be straightforward. Some of these links refer to a customer's need for anonymity, wanting himself to stay as non-transparent as possible. In this respect, a customer who is providing only a cell phone number, who does not want to be called for making a delivery appointment, who does not allow for collecting installments by bank direct debit etc. wants to act as self-determined as possible. Although this is not considered negative, in general, there is strong empirical evidence for these people being much more likely to default on payment (Bonne, 2000; Schwarz, 2008). Other characteristics give hints on income and employment, with younger customers having lower income on average. In contrast, customers who order delivery in the afternoon and/or to a different address are more likely to be employed, and thus should show a lower default risk. In addition, we expect an increasing default risk for higher financed amounts and low downpayments, which have to be made either when ordering or at delivery. The financed amount is an exact linear combination of the number of installments and the amount of each installment, i.e., one of these three variables has to be omitted in parametric models, such as the logReg model.

2.2. The reference model: optimized logistic regression

In logReg, we regress the customers' characteristics on the logistic transformation of the default probability

$$\ln\frac{p_i}{1-p_i} = \ln\frac{P(y_i=1|\boldsymbol{x}_i)}{1-P(y_i=1|\boldsymbol{x}_i)} = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

We optimize the regression model by a systematic stepwise selection of main and interaction effects of the given set of explanatory variables. In addition to the qualitative information given in Table 1, we use categorizations of financed amounts, downpayments, and age of the customers. This is motivated by the fact that we neither can suppose monotone, linear relations of these attributes and the risk of defaulting, nor do we observe continuous distributions of these variables. Table 2 shows the optimized logReg model developed on basis of the training data.

2.3. The fundamental idea of probability estimation using machine learning

Consider a sample of n individuals with a dichotomous dependent variable $y_i = 1$ if a credit of subject i is a default, and $y_i = 0$, otherwise. The covariables of subject i are denoted by $\boldsymbol{x}_i$. The aim is to estimate the default probability $P(y_i = 1|\boldsymbol{x}_i)$ of a credit given the variables x. Because $p(\boldsymbol{x}_i) = P(y_i = 1|\boldsymbol{x}_i) = E(y_i = 1|\boldsymbol{x}_i)$, the probability estimation problem is identical to the nonparametric regression estimation problem $f(\boldsymbol{x}) = E(y|\boldsymbol{x})$. Thus, as noted by Malley et al. (2012), any learning machine performing "well" on the nonparametric regression problem $f(\boldsymbol{x})$ will also perform "well" on the probability estimation problem $p(\boldsymbol{x}_i)$. More formally, a nonparametric regression function estimate $\widehat{f}(\boldsymbol{x}_i)$ is consistent if its mean square error converges to 0, i.e., $\lim_{n\to\infty} E(\widehat{f}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i))^2 = 0$. Consistency has been shown for many different machine learning