Expert Systems with Applications 39 (2012) 2433–2442

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Does segmentation always improve model performance in credit scoring?

Katarzyna Bijak *, Lyn C. Thomas

School of Management, University of Southampton, Southampton SO17 1BJ, UK

ARTICLE INFO

Keywords:
Credit scoring
Segmentation
Logistic regression
CART
CHAID
LOTUS

ABSTRACT

Credit scoring allows for the credit risk assessment of bank customers. A single scoring model (scorecard) can be developed for the entire customer population, e.g. using logistic regression. However, it is often expected that segmentation, i.e. dividing the population into several groups and building separate scorecards for them, will improve the model performance. The most common statistical methods for segmentation are the two-step approaches, where logistic regression follows Classification and Regression Trees (CART) or Chi-squared Automatic Interaction Detection (CHAID) trees etc. In this research, the two-step approaches are applied as well as a new, simultaneous method, in which both segmentation and scorecards are optimised at the same time: Logistic Trees with Unbiased Selection (LOTUS). For reference purposes, a single-scorecard model is used. The above-mentioned methods are applied to the data provided by two of the major UK banks and one of the European credit bureaus. The model performance measures are then compared to examine whether there is improvement due to the segmentation methods used. It is found that segmentation does not always improve model performance in credit scoring: for none of the analysed real-world datasets, the multi-scorecard models perform considerably better than the single-scorecard ones. Moreover, in this application, there is no difference in performance between the two-step and simultaneous approaches.

1. Introduction

Thomas, Edelman, and Crook (2002) define credit scoring as "the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit" (p. 1). These models and techniques are used to assess the credit risk of bank customers (individuals as well as small and medium enterprises).

Depending on the data used to build models, there are different types of scoring. Application scoring is based on data from loan application forms while behavioural scoring is based on data on customers' behaviour stored in bank databases. A special type of the latter is credit bureau scoring. Credit bureaus are institutions that collect and analyse data on loans granted by banks operating in a given country (Anderson, 2007; Van Gestel & Baesens, 2009). Such data enable tracking the credit history of a customer in the banking sector. Credit bureau scoring is based on data on customers' credit histories. Application scoring can also be enriched with data from a credit bureau. As a rule, using such data increases performance of a scoring model (Van Gestel & Baesens, 2009).

A scoring model describes the relationship between customer's characteristics (independent variables) and his or her creditworthiness status (a dependent variable). A customer's status can be either "good" or "bad" (and sometimes also "indeterminate" or "other"). The most common form of scoring models is referred to as a scorecard. According to Mays (2004), the scorecard is "a formula for assigning points to applicant characteristics in order to derive a numeric value that reflects how likely a borrower is, relative to other individuals, to experience a given event or perform a given action" (p. 63). Scorecards are used to calculate scores and/or probabilities of default (PD). They are sometimes scaled to obtain a required relationship between scores and PD. A scoring model can consist of one or more scorecards. In the latter case, it can be referred to as a suite of scorecards. In order to develop such a multi-scorecard model, segmentation has to be applied.

It is commonly expected that segmentation will improve the model performance. Segmentation is often carried out using the two-step approaches, where logistic regression follows Classification and Regression Trees (CART) or Chi-squared Automatic Interaction Detection (CHAID) trees. In this research, these approaches were applied as well as Logistic Trees with Unbiased Selection (LOTUS). The latter is a new, simultaneous method, in which both segmentation and scorecards are optimised at the same time. A single-scorecard logistic regression model was used as a reference. All these methods were applied to the data provided by two of the major UK banks and one of the European credit bureaus. Once the models were developed, the obtained results were analysed to examine whether there is improvement in the model performance due to the segmentation methods used. Moreover, the segmentation contribution was assessed.

* Corresponding author. Tel.: +44 23 80598964.
E-mail address: K.Bijak@soton.ac.uk (K. Bijak).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.093

The paper is structured as follows. In the next section, the theoretical background of segmentation is presented as well as segmentation methods and other researchers' findings on its impact on the model performance. In the third section, the basics of logistic regression, CART, CHAID and LOTUS are introduced. In the fourth section, the datasets are described. The fifth section is on the research results. The sixth section is a discussion and the last section includes the research findings and conclusions.

2. Segmentation

2.1. Theoretical background

In credit scoring, segmentation can be defined as "the process of identifying homogeneous populations with respect to their predictive relationships" (Makuch, 2001, p. 140). The identified populations are treated separately in the process of a scoring model development, because of possible unique relationships between customer's characteristics and a dependent variable.

Nowadays segmentation is widely used in banking. There are various segmentation drivers, i.e. factors that can drive the division of a scoring model into two or more scorecards. Anderson (2007) classifies them into: marketing, customer, data, process and model fit factors. The first four factors reflect, respectively, the special treatment of some market segments, or customer groups, data issues (such as data availability) and business process requirements (e.g. different definitions of a dependent variable). The model fit relates to interactions within the data and using segmentation to improve the model performance. In this research, the focus is on segmentation which is driven by the model fit factors.

As far as segmentation is concerned, there are two key concepts: a segmentation basis and a segmentation method. A segmentation basis is a set of variables that allow for the assignment of potential customers to homogeneous groups. Segmentation bases can be classified as either general or product-specific, and either observable or unobservable (Wedel & Kamakura, 2000). As far as scorecard segmentation is concerned in this research, there is an unobservable product-specific basis. Once the segmentation is implemented, customers are grouped on the basis of their unobservable behavioural intentions to repay their loans or the relationship between their intentions and characteristics. On the date of grouping, it is not known whether the customers are going to repay or not.

According to Wedel and Kamakura (2000), there are six criteria for effective segmentation. It seems that three of them are especially important in credit scoring: identifiability (customers can be easily assigned to segments), stability and responsiveness (segments differ from each other in their response/behaviour). Unobservable product-specific bases, which contain behavioural intentions, are characterised by good identifiability, moderate stability and very good responsiveness (Wedel & Kamakura, 2000). The above-mentioned advantages make these bases promising as far as scorecard segmentation is concerned.

Segmentation methods can be classified as either associative (descriptive) or regressive (predictive) approaches (Aurifeille, 2000; Wedel & Kamakura, 2000). Since the ultimate goal is to assess the credit risk, the latter are applied in this research. There are two types of regressive approaches: two-step (a priori) and simultaneous (post hoc) methods (Aurifeille, 2000; Wedel & Kamakura, 2000). In the two-step approaches, segmentation is followed by the development of a regression model in each segment. In the simultaneous methods, both segmentation and regression models are optimised at the same time.

The two-step approaches are not designed to yield optimal results in terms of the prediction accuracy but rather to aid the understanding of overall strategy. On the other hand, the simultaneous methods give priority to a low, tactical level rather than to a high, strategic level of decision: the optimisation objective is to obtain the most accurate prediction, and not necessarily a meaningful and easily understandable segmentation (Desmet, 2001).

2.2. Segmentation methods

There is not much literature on segmentation methods in credit scoring. According to Siddiqi (2005), segmentation methods can be classified as either experience-based (heuristic) or statistical. As far as the experience-based methods are concerned, one approach is to define segments that are homogeneous with respect to some customers' characteristics. This allows for the development of segment-specific variables. For example, creating a segment of customers, who have a credit card, enables construction of such characteristics as credit limit used. Another approach is to define segments that are homogeneous with respect to the length of customers' credit history (cohorts) or data availability (thin/thick credit files). For instance, creating a segment of established customers allows building behavioural variables based on the data from the last 12 months, the last 24 months etc.

Furthermore, if there is a group (e.g. mortgage loan owners or consumer finance borrowers) that is expected to behave differently from other customers, or for whom the previous scoring model turned out to be inefficient, it is worth creating a separate segment for such a group. Moreover, customers can be grouped into segments in order to make it easier for a bank to treat them in different ways, e.g. by setting different cut-offs, i.e. score thresholds used in the decision making (Thomas, 2009).

Finally, segmentation can be based on variables (e.g. age) that are believed to have strong interactions with other characteristics (Thomas, 2009). This is a heuristic approach but it has been developed into statistical methods based on interactions. An alternative to segmentation based on a selected variable is to include all its interactions with the other variables in a single-scorecard model (Banasik, Crook, & Thomas, 1996). However, such a model has a large number of parameters and is less understandable than a multi-scorecard one.

The experience-based segmentation methods can help achieve various goals such as improving the model performance for a certain group of customers or supporting the decision making process. The experience-based segmentation may also allow for better risk assessment for the entire population of customers. However, there is no guarantee that segmentation, which intuitively seems reasonable, will increase the model performance (Makuch, 2001).

As far as statistical methods are concerned, segmentation is obtained using statistical tools as well as data mining and machine learning techniques. One approach is to do the cluster analysis (Siddiqi, 2005). The cluster analysis can be conducted using hierarchical clustering, the *k*-means algorithm or Self-Organising Maps (SOMs). Regardless of the algorithm applied, clustering is based on customers' characteristics. Therefore, customers with different demographic or behavioural profiles are classified into different segments. The resulting groups are homogeneous with respect to the characteristics but, since the customers' status is not used in segmentation, they do not need to differ in risk profiles.

Another approach is to use tree-structured classification methods such as CART or CHAID (VantageScore, 2006). In this approach, grouping is based on the customers' status, and thus segments differ in risk profiles. Both the cluster analysis and classification trees can constitute the first step in the two-step regressive approaches.

However, the classification trees often yield sub-optimal results (VantageScore, 2006). In 2006 VantageScore introduced a new, multi-level segmentation approach: combining experience-based segmentation (at higher levels) and segmentation based on a dedicated score (at lower levels). This score must be calculated using