Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 995–1003
Expert Systems with Applications
www.elsevier.com/locate/eswa

Analysis of healthcare coverage: A data mining approach

Dursun Delen *, Christie Fuller, Charles McCann, Deepa Ray

William S. Spears School of Business, Department of Management Science and Information Systems, Oklahoma State University, North Classroom Building #378, 700 North Greenwood Avenue, Tulsa, OK 74106, USA

Abstract

The existing disparity in the healthcare coverage is a pressing issue in the United States. Unfortunately, many in the US do not have healthcare coverage and much research is needed to identify the factors leading to this phenomenon. Hence, this study aims to examine the healthcare coverage of individuals by applying popular machine learning techniques on a wide-variety of predictive factors. Twenty-three variables and 193,373 records were utilized from the 2004 behavioral risk factor surveillance system survey data for this study. The artificial neural networks and the decision tree models were developed and compared to each other for predictive ability. The sensitivity analysis and variable importance measures are calculated to analyze the importance of the predictive factors. The experimental results indicated that the most accurate classifier for this phenomenon was the multi-layer perceptron type artificial neural network model that had an overall classification accuracy of 78.45% on the holdout sample. The most important predictive factors came out as income, employment status, education, and marital status. Using two popular machine learning techniques, this study identified the factors that can be used to accurately classify those with and without healthcare coverage. The ability to identify and explain the reasoning of those likely to be without healthcare coverage through the application of accurate classification models can potentially be used in reducing the disparity in healthcare coverage.

Keywords: Healthcare coverage; Data mining; Prediction; Classification; Neural networks; Decision trees

1. Introduction

Healthcare coverage in general and the existing disparity in this coverage in specific is a pressing issue in the United States. Many in the US do not have healthcare coverage, and much research has been conducted, and much more is needed to identify the factors leading to this disparity in coverage. Previous work has identified two key situations where understanding of these factors is beneficial (Cunningham & Ginsburg, 2001). First, given that these factors exist at both the state and local level, it is imperative that the individuals responsible for funding decisions correctly interpret the reasons that the uninsured rates may be elevated. Second, it is important that the individuals are able to determine if the uninsured rates may be elevated simply due to unchangeable characteristics within the population.

While existing research has identified factors which differ between those with and without coverage, it has not progressed to building discriminatory models to separate them from each other. Further, despite identification of differences on some factors, the disparity has not been reduced (Glover, Moore, Probst, & Samuels, 2004). This study advances the current research by building classification models to identify those belonging to each group – those that do and those that do not have healthcare coverage. Such a model may eventually be used to help reduce healthcare coverage disparity. Similar techniques have previously been introduced in a more general manner for use in targeting customers in the insurance industry (Wu, Kao, Su, & Wu, 2005).

Identifying those without healthcare coverage is important, as those without coverage may have reduced access to medical care (Monheit & Vistnes, 2000) and may have

* Corresponding author. Tel.: +1 918 594 8283; fax: +1 918 594 8281.
E-mail address: dursun.delen@okstate.edu (D. Delen).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.10.041

more preventable hospitalizations (Services USDoHaH, 2003). Lack of healthcare coverage has also been linked to poor health and early death. This issue has been exacerbated by the clear increasing trend of the number of uninsured in the last 20 years (Cunningham & Ginsburg, 2001; Herring, 2005). From 1989 to 1999, there was an 18.4% increase in the number of people without insurance. The level of uninsured has grown to approximately 16% of the population as of 2003 (Jonk et al., 2005). It has been estimated that 60 million people are uninsured at some point during a given year (Cunningham & Ginsburg, 2001). The factors leading to the increasing number of uninsured have been studied at both the state and local level (Cunningham & Ginsburg, 2001), and across many socio-demographic factors (Hendryx, Ahern, Lovrich, & McCurdy, 2002; Holtz-Eakin, 2002). Lifestyle factors may also contribute to this problem.

Gender is one socio-demographic variable that has been found to be related to healthcare coverage (Carrasquillo, Carrasquillo, & Shea, 2000; Monheit & Vistnes, 2000; Shi, 2000). Many studies have found that females are less likely to be insured than males (Hendryx et al., 2002; Holtz-Eakin, 2002; Monheit & Vistnes, 2000), though the opposite has also been found (Carrasquillo, Himmelstein, Woolhandler, & Bor, 1999; Nelson, Bolen, Wells, Smith, & Bland, 2004). Race or ethnicity may also be a factor contributing to healthcare coverage disparity (Carrasquillo et al., 2000; Lucas, Barr-Anderson, & Kington, 2003; Monheit & Vistnes, 2000), with minorities generally having less healthcare coverage (Carrasquillo et al., 1999; Glover et al., 2004; Monheit & Vistnes, 2000). According to the results of multiple studies, those with lower incomes are less likely to have healthcare coverage (Cardon & Hendel, 2001; Carrasquillo et al., 1999; Lucas et al., 2003), and having healthcare coverage also has a relationship with employment status (Schmidt & Deichert, 1996) and type of employment (Krieger, Barbeau, & Soobader, 2005). There is also a difference among states, region of the country, or even county in the rate of healthcare coverage (Cardon & Hendel, 2001; Carrasquillo et al., 1999; Nelson et al., 2004; Schmidt & Deichert, 1996). For example, in the Northeast and Midwest, one is more likely to be insured while in the South or the West, one is more likely to be uninsured (Cardon & Hendel, 2001). Some studies suggest that younger adults have less insurance (Cardon & Hendel, 2001; Carrasquillo et al., 1999). Also, studies (Jonk et al., 2005; Woolhandler et al., 2005) that looked at differences between veterans and non-veterans found that fewer veterans had less insurance relative to the remaining population. Education and marital status have also been found to have a relationship with this insurance disparity (Shi, 2000). Disabled Americans may lack coverage (Landerman et al., 1998). For some populations studied, the role of household size in this disparity has also been examined (Glover et al., 2004; Pol, Mueller, & Adidam, 2002).

In addition to the socio-demographic factors described above, lifestyle may also play a part in the existing disparity. Those that are already in poor health may have reduced coverage (Cunningham & Ginsburg, 2001; Glover et al., 2004). These studies did not specify whether poor health included mental and/or physical health. Smoking status has been studied in relation to type of insurance coverage held (King & Mossialos, 2005). Exercise and alcohol consumption are additional lifestyle variables that have previously been used in classification models for insurance policy purposes (Chae, Ho, Cho, Lee, & Ji, 2001). Further, the extent to which someone is a risk taker may affect whether they secure healthcare coverage (Cunningham & Ginsburg, 2001).

Past studies have often looked at a subgroup, such as the near-elderly (Monheit, Vistnes, & Eisenberg, 2001) or immigrants (Herring, 2005), of the overall US population and have succeeded in identifying variables that seem to contribute to the disparity in healthcare coverage for the subgroup. This study will look at healthcare coverage disparity across the population of the US, rather than within a smaller subgroup of the population. It will also address the numerous possible contributing factors, both socio-demographic and lifestyle, and their contribution to the growing disparity in healthcare coverage. Further, the study will utilize machine learning techniques in building classification models. Previously, the issue of healthcare coverage disparity has been studied using primarily statistical techniques such as logistic regression (Glover et al., 2004) and basic descriptive statistics (Shi, 2000; Woolhandler et al., 2005). For many years linear regression has been the primarily used technique in capturing and representing functional relationships between dependent and independent variables, largely because of its well-known statistically explainable optimization strategies. However, in many problem scenarios, the model accuracy suffers as the assumed linear approximation of a function is not valid. With current technology, machine learning techniques can easily model such scenarios as healthcare coverage, as is addressed in this paper. These techniques are not constrained by the Gauss–Markov assumption (such as multicollinearity and normality) which is a major concern for more traditional models (Uysal & Roubi, 1999). Previously, these techniques have been used to study other healthcare issues such as factors affecting inpatient mortality (Chae, Kim, Tark, Park, & Hoa, 2003) and influencing prenatal care (Prather et al., 1997). In this study, we have attempted to build an accurate classification model, using machine learning techniques. The model could then be used to predict whether or not an individual has healthcare coverage based on specific socio-demographic and lifestyle information as well as the importance of the various factors in the model.

2. Research method

2.1. Data

The data source that was used for this project was the Behavioral Risk Factor Surveillance System 2004 Survey