Decision Support Systems xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

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

Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach

Johannes Paefgen [a], Thorsten Staake [b,c], Frédéric Thiesse [d,*]

[a] ITEM-HSG, University of St. Gallen, Dufourstrasse 40a, 9000 St. Gallen, Switzerland
[b] MIS and Energy-Efficient Systems, University of Bamberg, An der Weberei 5, 96047 Bamberg, Germany
[c] Information Management, WEV G 217, Weinbergstrasse 58, 8092 Zurich, Switzerland
[d] IS Engineering, University of Wuerzburg, Josef-Stangl-Platz 2, 97070 Wuerzburg, Germany

A R T I C L E I N F O

Article history:
Received 15 October 2012
Received in revised form 26 April 2013
Accepted 3 June 2013
Available online xxxx

Keywords:
Location data
Actuarial decision-making
Pay-as-you-drive insurance
Classification analysis
Logistic regression
Neural networks
Decision trees

A B S T R A C T

Vehicle sensor data enable novel, usage-based insurance premium models known as 'Pay-As-You-Drive' (PAYD) insurance, but pose substantial challenges for actuarial decision-making because of their inherent complexity and volume. Based on a large real-world sample of location data from 1572 vehicles, the present study proposes a classification analysis approach that addresses (i) the selection of predictor variables, (ii) the presence of class skew and time-variant prior distributions, and (iii) the suitability of classifier scores as an aggregated actuarial rate factor. Using raw location data, we derive a set of 15 predictor variables that we use to train and compare logistic regression, neural network, and decision tree classifiers. We find that while neural networks exhibit superior classification performance, logistic regression is better suited from an actuarial viewpoint in several ways. In sum, our results clearly demonstrate the potential of high-resolution exposure data for reducing the complexity of PAYD insurance pricing in practice.

1. Introduction

'Ubiquitous Computing' [54], 'Pervasive Computing' [47], 'Things that think' [27], 'Ambient Intelligence' [1], 'Silent Commerce' [21] — a plethora of novel terms presage the arrival of a paradigm shift in information processing. Common to all these concepts is the shared vision of a future world of everyday physical objects equipped with sensors and networking capabilities, whose quantity is expected to surpass the number of current Internet infrastructure entities by orders of magnitude [13]. From a management perspective, the appeal of this vision is grounded in the hope of closing the gap between real-world objects and their digital counterparts in the network, which may ultimately lead to a state of 'real world awareness' [29] of enterprise information systems. Ubiquitous wireless sensor technologies hold the unique potential of providing firms with a continuous stream of fine-granular and timely information on physical events within the organization and beyond [2].

A current example of this development is the increasing availability and commercial usage of vehicle sensor data. Today's road vehicles are equipped with an abundance of sensors that provide information on their location, situation, and state. Location information in particular, and the vehicle trajectories derived from it, have become an important cornerstone of applications such as road infrastructure optimization [43] and resource planning strategies in a vehicle fleet [46,50]. In this paper, we consider the use of vehicle location information in actuarial decision-making, a concept frequently referred to as 'Pay-As-You-Drive' (PAYD) insurance [17]. The idea underlying PAYD insurance contracts is to utilize location data in order to re-calculate car insurance premiums at periodic intervals – for example, on a monthly basis – based on the policyholders' individual driving patterns. The hope among insurers is that PAYD insurance will allow for tariffs that (i) more closely reflect actual risk exposure of road vehicles, and (ii) are adaptive over time, thus providing policyholders incentives to minimize risk. With PAYD insurance, information asymmetry between insurers and policyholders is reduced, which mitigates adverse selection and moral hazard [10]. For instance, prior studies have shown that respective tariffs have a significant impact on safe driving behavior among young drivers [7]. Furthermore, substituting conventional lump-sum premium payments with flexible rates brings about a better allocation of mobility-related costs. PAYD insurance has therefore also been associated with macroeconomic benefits such as insurance affordability, higher consumer surplus, improved traffic safety, and reduced externalities [36]. However, in order

* Corresponding author at: Information Systems Engineering, Julius-Maximilian University of Wuerzburg, Josef-Stangl-Platz 2, 97070 Wuerzburg, Germany. Tel.: +49 931 3180789; fax: +49 931 3181268.
E-mail addresses: johannes.paefgen@unisg.ch (J. Paefgen), thorsten.staake@uni-bamberg.de (T. Staake), frederic.thiesse@uni-wuerzburg.de (F. Thiesse).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.06.001

to reap the alleged benefits of PAYD insurance, providers need to overcome the substantial challenge of adapting actuarial decision-making to incorporate vehicle sensor data, which differs substantially in complexity and volume from established variables.

Multivariate tariff models in motor insurance commonly follow a two-step procedure in the determination of premiums. First, actuaries define a set of discrete tariff classes. These classes are inferred either directly from categorical variables or by applying bounded-interval rules on continuous variables [3,25,30]. Pricing variables (*rate factors*) include gender or age of policyholders, among others. Secondly, for each tariff class, actuaries estimate the expected loss per policy over a certain period, referred to as the *pure premium*, based on distribution characteristics of historical claims data [16]. Tariff classes should accordingly exhibit significant variations in the corresponding pure premiums. Depending on the amount of claims data available to an insurer, there is a trade-off between the number of differentiating tariff classes and the accuracy of premium estimation within these classes.

Collecting vehicle sensor data allows insurers to generate various risk-related variables such as mileage, time of day, and the type of road a vehicle traverses. Since each additional rate factor increases the dimensionality of an actuarial table and multiplies the previous number of tariff classes by its number of categories or intervals, insurers must carefully consider which of these variables constitute suitable rate factors in PAYD insurance. An important precondition for PAYD insurance is hence determining suitable levels of data aggregation that combine several variables. Ideally, one might derive a one-dimensional aggregated variable that adds only one further dimension to actuarial tables. Such an aggregated variable may also be a suitable substitute for rate factors that have been ruled discriminatory in the wake of recent changes in insurance regulation in several regions. However, the existing body of literature on the topic of sensor data-derived rate factors is still very limited. Despite the rapidly growing interest among actuarial practitioners, there appears to be no consensus on how to approach the problem of PAYD insurance variable assessment and aggregation. Against this backdrop, we address this research gap by a classification analysis approach to rate factor aggregation. For the purpose of model building and evaluation, we build upon an extensive data sample gathered under real-world conditions from 1572 vehicles. Our results confirm the value of sensor data to the insurance business and show that the performance of different classification approaches varies considerably in the PAYD insurance scenarios.

The remainder of the paper is organized as follows. In the next section, we canvass key aspects of state-of-the-art actuarial decision-making in motor insurance and review related work on rate factors obtained from vehicle sensor data. Section 3 details the development of classifiers using logistic regression, neural networks, and decision trees. Sections 4 and 5 present the process of data collection and discuss the results from the empirical evaluation. The paper closes with conclusions and limitations to our work, and outlines suggestions for further research.

2. Theoretical background

2.1. Conventional motor insurance rate factors

In order to differentiate the risk of insurance policies, actuaries use a set of rate factors to separate policies into groups (i.e., tariff classes). The construction of tariff classes is ultimately a clustering task [3]. Each tariff class corresponds to a certain combination of rate factor categories or intervals in the case of continuous rate factors. For each tariff class, actuaries analyze historical claims data to arrive at a reliable estimate of the corresponding pure premium, that is, the minimum required payment per policy to cover the expected losses from its class [16,36]. An important distinction is usually made between claim frequency, claim type, and claim amount as different dependent variables to be estimated [24]. For each of these variables, actuaries estimate a non-linear regression model from the historical claims in a tariff class. For a detailed review of specific regression techniques used to model the different dependent variables, we refer the reader to the review by Lord & Mannering [37].

Once the estimates for frequency, type, and amount distributions within a tariff class become available, actuaries calculate pure premiums as the expected claims value for a given period and tariff class. In automotive insurance, different coverage types exist such as liability, personal injury protection, protection from underinsured accident counterparties, and comprehensive insurance [26]. As our research is concerned with rate factors that define tariff classes rather than the calculation of pure premiums within these classes, we disregard coverage type in our analysis without a loss of generality. For each tariff class, pure premiums for a specific coverage type are independently calculable from historical claims of that type.

Actuarial decision-making in conventional motor insurance incorporates a broad range of established rate factors. These are roughly divisible into driver-related and vehicle-related predictor variables, all of which are typically obtained from a potential policyholder before the conclusion of an insurance contract through administering a questionnaire. Which specific rate factors actuaries choose for this questionnaire depends on an insurer's specific business policies as well as insurance regulations. Possible driver-related criteria include age, gender, nationality, and family status, while vehicle-related factors typically comprise the model and make of the car, its registration date, and cubic capacity [25,45]. While driver-related factors are typically more indicative of differences in claim *frequencies* within a tariff class, vehicle-related factors are more relevant to the estimation of claim *amounts* as they represent the residual value of an insured vehicle. Annual mileage constitutes a separate type of rate factor that is also referred to as the *exposure* of a vehicle [57]. However, while empirical evidence suggests that mileage is highly relevant for predicting accident risk [4,11], it is typically difficult to obtain correct mileage values directly from policyholders. Researchers have found that vehicle owners at times underreport their annual mileage, which may be attributed at least to some extent to elevated insurance rates associated with higher annual mileage [55].

2.2. PAYD insurance rate factors and vehicle sensor data

The rationale behind the PAYD insurance concept is that objective data on the actual usage of a vehicle – usually location information – allow for a more refined differentiation of premiums, which is also adaptive to changes in insured risk over time. Today, several tariffs require policyholders to inform the insurer if their annual mileage is significantly different from that of the previous year. However, PAYD insurance considers a much broader variety of variables and is consequently regarded as more objective than self-reported data. From another perspective, usage-based variables are also important as a substitute for established rate factors in insurance. For example, the European Court of Justice has ruled that from December 2012 at the latest, unisex premiums are mandatory for newly concluded insurance contracts (i.e., gender is to be disregarded as a rate factor) [52]. Similar provisions are likely for other factors such as nationality. The omission of such rate factors considerably impedes insurers' ability to differentiate between risks. PAYD insurance rate factors are potentially a nondiscriminatory alternative, and insurers who use them can more adequately price policies and gain competitive advantage.

Despite the impact of PAYD insurance on insurers' business practices and the substantial amount of PAYD insurance policies contracted by pioneering insurers, there are only few published studies that specifically address the relationship between actual vehicle sensor data and accident risk. Jun et al. [31,32] report an analysis of location data collected from 167 vehicles over a 14 month duration, 26 of which were involved in accidents during the study period. The authors derive mileage and velocity variables as derivatives of vehicle position after Kalman filter

Please cite this article as: J. Paefgen, et al., Evaluation and aggregation of pay-as-you-drive insurance rate factors: a classification analysis approach, Decision Support Systems (2013), http://dx.doi.org/10.1016/j.dss.2013.06.001