On a strategy to develop robust and simple tariffs from motor vehicle insurance data

Andreas Christmann
University of Dortmund, Department of Statistics, D-44221 Dortmund, Germany

Abstract

The goals of this paper are twofold: we describe common features in data sets from motor vehicle insurance companies and we investigate a general strategy which exploits the knowledge of such features. The results of the strategy are a basis to develop insurance tariffs. The strategy is applied to a data set from motor vehicle insurance companies. We use a nonparametric approach based on a combination of kernel logistic regression and $\varepsilon-$support vector regression.

Key words: Classification; Data Mining; Insurance tariffs; Kernel logistic regression; Machine learning; Regression; Robustness; Simplicity; Support Vector Machine; Support Vector Regression.

1 Introduction

Insurance companies need estimates for the probability of a claim and for the expected sum of claim sizes to construct insurance tariffs. In this paper we consider statistical aspects for analyzing such data sets from motor vehicle insurance companies. Some of the results may also be useful for other areas, e.g. in credit risk scoring, customer relationship management (CRM) or for CHURN analyses.

In Section 2 the statistical objectives are given. Section 3 describes characteristics of data sets from motor vehicle insurance companies. Section 4 contains the proposed strategy. Section 5 briefly describes kernel logistic regression and $\varepsilon-$support vector regression, which both belong to statistical machine learning methods based on convex risk minimization, *cf.* Vapnik (1998). In Section 6 the results of applying the strategy to a data set from 15 motor vehicle insurance companies are described. Here, we use a non-parametric approach based on a combination of kernel logistic regression and $\varepsilon-$support vector regression. Section 7 contains a discussion.

2 Statistical objectives

In this section we describe common statistical problems in analyzing data from motor vehicle insurance companies.

An insurance company collects a lot of information for each single policy holder for each year or period. Often dozens or even hundreds of variables are available from each customer. Most of this information belongs to one of the following categories:

- personal information: e.g. name, surname, type of policy, policy number, other insurance policies
- demographic information: e.g. gender, age, place of residence, postal zip code, population density of the region the customer is living in, occupation type
- driver information: e.g. main user, driving distance within a year, car kept in a garage
- family information: e.g. age and gender of other people using the same car
- history: e.g. count and size of previous claims, property damage, physical injury, occurrence of a loss
- motor vehicle: e.g. type, age, strength of engine
- response information: claim (yes/no), number of claims, claim size.

In practice, the claim size is not always known exactly. E.g. if a big accident occurs in November, the exact claim size will often not be known at the end of the year and perhaps not even at the end of the following year. Possible reasons are law-suits or the case of physical injuries. In this case, a statistician will have to use more or less appropriate estimations of the exact claim size to construct a new insurance tariff for the next year. Hence, the empirical distribution of the claim sizes is in general a mixture of really observed values and of estimated claim sizes.

Further, some explanatory variables may have imprecise values. *E.g.* there is a variable describing the driving distance of a customer within a year. The customer has to choose between some categories, e.g. below 9000 kilometers, between 9000 and 12000 kilometers, between 12000 and 15000 kilometers, etc. There are reasons making it plausible, that a percentage of these values are too small, because it is well-known that the premium of an insurance tariff increases for increasing values of this variable.

An insurance company is interested in determining the actual premium charged to the customer. In principle, the actual premium is the sum of the pure premium plus a safety loading plus administrative costs plus the desired profit. In this paper the focus will be on the **pure premium** denoted by Y. We assume that $Y \geq 0$ almost surely. The pure premium y_i for each customer is computed by

$$\text{pure premium} = \frac{\text{sum of individual claim sizes}}{\text{number of days under risk}/360} \ .$$

Useful information of explanatory variables is included in the vector $x \in \mathbb{R}^p$. Our **primary response variable** is the conditional expectation of the pure premium Y given the information of the explanatory variables, i.e. $\mathrm{E}(Y|X = x)$. The **secondary response variable** is the conditional probability $\mathrm{P}(Y > 0|X = x)$ that the policy holder will have at least one claim within one year given the information contained in the explanatory variables.

An estimate for the expected pure premium should have the following four properties.

- It is **fair**. The expectation of the estimated pure premium $\mathrm{E}(\hat{Y}|(X = x)$ should be approximately unbiased for the whole population and also in subpopulations of reasonable size.

- It has a **high precision**. One precision criterion is the mean squared error (MSE) defined by $\mathrm{E}((Y - \hat{Y})^2|X = x)$, which should of course be small. But other measures of precision may also be interesting, e.g. a trimmed version of the MSE or a chi-square statistic of Pearson-type for subpopulations of reasonable size.

- It is **robust** against moderate violations of the statistical model assumptions, and the impact of outliers on the estimation is bounded.

- It has a **simplicity** property, because too complex tariffs with many interactions terms may only have a reduced practical importance.

If the tariff is not fair, there will be two cases. A high positive bias is of course bad from the view point of the policy holder, because the premium is too high and he or she will have to pay too much money. At first sight this case looks great for the insurance company, but there is a danger that the customer will turn to another insurance company. If there is a large negative bias, this is good from the view point of the customer. In general, an insurance company avoids this case, but the company is interested in increasing its share of the market.

3 Characteristics of the data

In this section we describe some of the characteristics of a data set from the Verband öffentlicher Versicherer in Düsseldorf, Germany. The data set contains data from 15 insurance companies in non-aggregated structure. For reasons of data protection and confidentiality, there is no indicator variable describing which customer belongs to which insurance company. Overall, the data set has approximately a size of 3 GB as a compressed SAS data set and contains information from more than 4.6 million policy holders. A single policy holder may contribute more than one case to the data set, *e.g.* if he or she was involved in more than one claim. Using an identification number one can determine for each policy holder, whether he or she had a claim, sum up the total claim sizes, and compute then the pure premium. The data set contains more than 70 explanatory variables; most of them are discrete. Even nowadays, a reasonable statistical analysis strategy has to take into account this considerable size and not all software packages can deal with such big data sets.

There are approximately $275,000$ claims overall. Approximately 94.37 percent of the customers had no claim, 5.35 percent had one claim, and 0.26 percent had two claims. The others had three or even more claims. Most claims are from the year 2001. Note, that therefore some of the claim sizes are probably only estimates. Approximately 5 percent of the claims are older ones (up to 10 years).

An explanatory data analysis (EDA) shows that there are many complex dependencies between various variables. There are empty cells because not all combinations of the discrete explanatory variables are present, and of course, missing values occur as well. In the following we briefly mention some interesting results

Table 1: Description of the primary response variable pure premium.

Pure premium	% obs.	% of total sum	Mean	Median	Std Dev
total			364	0	21996
0	94.9	0	0	0	0
(0,2000]	2.2	6.7	1097	1110	520
(2000,10000]	2.4	27.1	4156	3496	1940
(10000,50000]	0.4	19.8	18443	15059	8911
>50000	0.07	46.4	234621	96417	784365

of the EDA for the observed pure premium values. The total mean of the pure premium values taking no explanatory variables into account is approximately 360 EUR. However, the empirical distribution of these values has an atom in zero and is extremely skewed. Table 1 shows that approximately 95% of all policy holders had no claim within the period under consideration. Approximately 2.2% of the policy holders had a positive pure premium up to 2000 EUR, 2.4% had a moderate pure premium of size between 2000 EUR and 10000 EUR. Only 0.4% of the customers had a high pure premium between 10000 EUR and 50000 EUR, but the sum of their pure premium values contribute almost 20% to the overall sum. It is interesting to note, that only 0.07% of all policy holders had a pure premium value higher than 50000 EUR, but they contribute 46.4% to the grand sum. The maximum observed pure premium was 28 million EUR for a customer who was only 36 days under risk and had a claim of 2.8 million EUR. The highest individual claim size was above 6 million EUR.

The distribution of the primary response variable is skewed even if one restricts to the interval $(0, 10000]$ EUR, see Figure 1. The density of the best fitting gamma distribution based on maximum likelihood estimation is shown. This figure draws attention to four peaks in intervals near by 1200, 1800, 2400, and 3000 EUR. Although this is only a univariate description of the data, it is interesting to note that the distance between sequential peaks is of comparable size. One possible explanation is that there are some standard procedures or flat rates to deal with minor or moderate accidents such that claim sizes within these intervals occur with an increased probability. It is clear from Table 1, that extreme pure premium values above 50000 EUR contain a lot of valuable information concerning the main response variable, although the number of extreme high values is small. Methods from extreme value theory can be very helpful in this case. Consider the Generalized Pareto distribution (GPD) with distribution function

$$\begin{aligned} G_{\xi,\beta} &= 1-(1+\beta^{-1}\xi x)^{-1/\xi} && \text{if} \quad \xi \neq 0 \\ &= 1-\exp(-x/\beta) && \text{if} \quad \xi = 0\,, \end{aligned}$$

where $x \geq 0$ if $\xi \geq 0$, and $x \in [0, -\xi^{-1}]$ if $\xi < 0$, *cf.* Embrechts, Klüppelberg and Mikosch (1997) and Celebrián, Denuit and Lambert (2003). The parameter ξ is called the Pareto index. The QQ plots given in Figure 2 show that the GPD can be useful to fit extreme pure premium values. The upper left plot shows a QQ plot on logarithmic scales for all 3343 pure premium values higher than 50000 EUR. The

Figure 1: Histogram of pure premium. Only positive values less than or equal to 10000 EUR are shown.

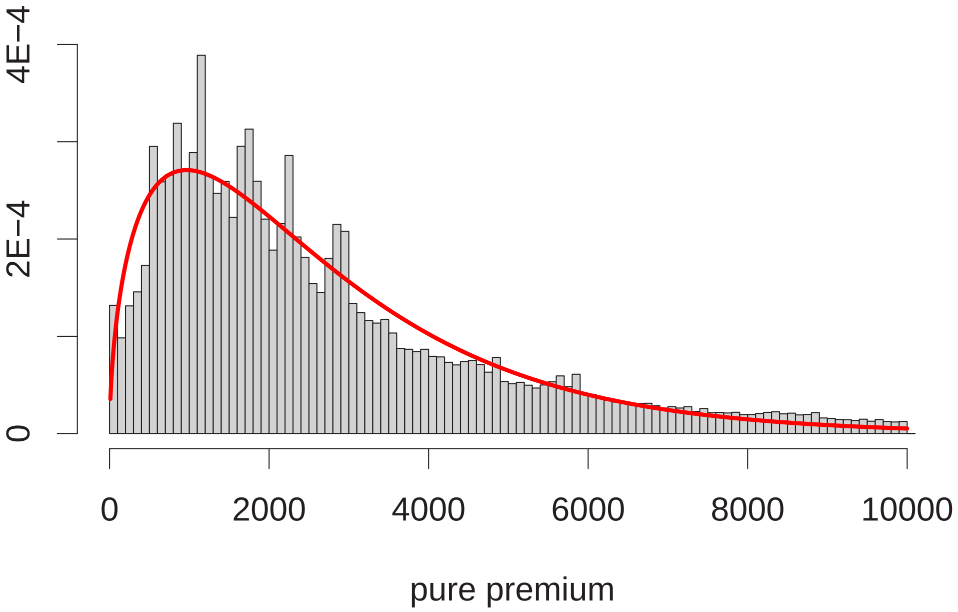


Table 2: Maximum likelihood estimates for the parameters ξ and β of the Generalized Pareto distributions used to model extreme pure premium values higher than 50000 EUR. Confidence intervals at the 95% level are given in parenthesis.

data	$\hat{\xi}$	95% CI	$\hat{\beta}$	95% CI
all	0.804	$(0.743, 0.865)$	50035.42	$(46813.48, 53257.37)$
knot 1	0.887	$(0.788, 0.985)$	53357.66	$(47959.25, 58756.08)$
knot 2	0.635	$(0.541, 0.728)$	51214.80	$(45919.53, 56510.07)$
knot 3	0.852	$(0.720, 0.984)$	44742.58	$(38588.19, 50896.97)$

other three plots given in Figure 2 give the corresponding plots for pure premium values higher than 50000 EUR, which belong to the main 3 knots of a regression tree constructed on the basis of a few important explanatory variables. The Pareto indix for the data set of knot 2 differ from the Pareto indices from the other two knots, see Table 2. Beirlant, de Wet and Goegebeur (2003) proposed an alternative method to fit extreme claim amounts.

Figure 2 shows the relative frequency that a claim occurred and the pure premium stratified with respect to the age of the main user of the car. The smooth curve is the fit provided by $\varepsilon-$support vector regression, *cf.* Vapnik (1998). There is an interesting non-monotone relationship between the age of the main user and both response variables. It is well-known that young drivers in Germany, say between 18 and 24 years old, have an increased claim frequency. As was expected, the plot also shows an increased claim frequency for elderly drivers. An interesting subgroup is the age group around 50 years. These main users also show an increased claim frequency and increased pure premium values. One plausible explanation is, that a

Figure 2: QQ plots with respect to the Generalized Pareto distribution (GPD) for pure premium values higher than 50000 EUR. Upper left: for all such data points ($n = 3343$). Upper right and lower: for data points belonging to the main 3 knots of a regression tree ($n = 1416$, $n = 1175$, and $n = 752$, respectively).

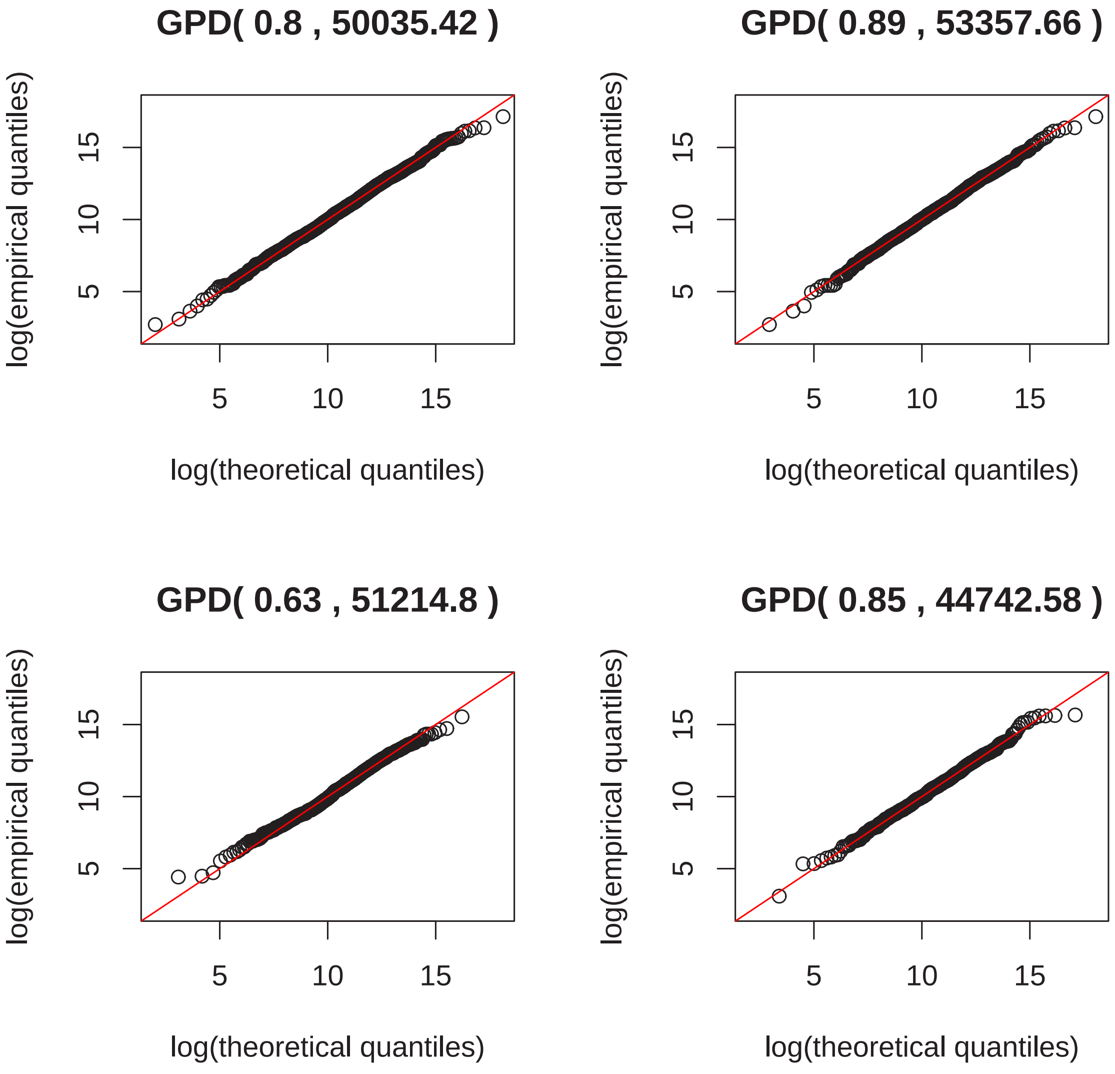


Figure 3: Relative frequency of claims and pure premium stratified by age of the main user. The solid curve joins the fitted values based on $\varepsilon-$support vector regression. The dashed curves show nonparametric bootstrap confidence bands at the 95% level.

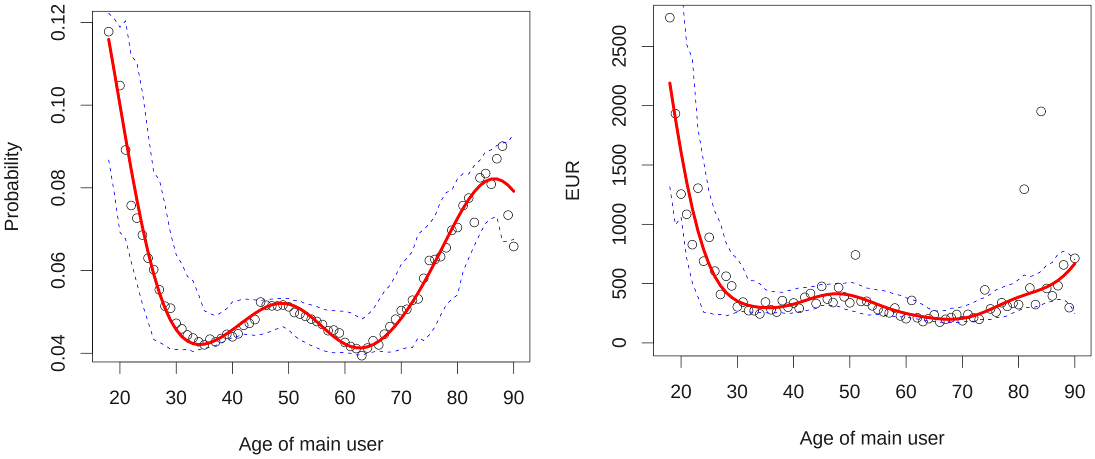
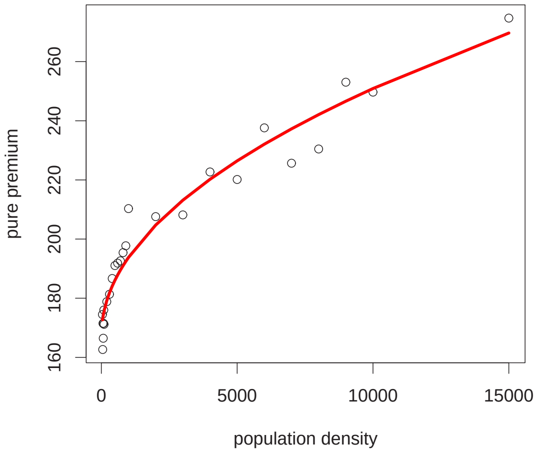


Figure 4: Scatterplot of pure premium vs. population density.



considerable percentage of main users from this age group have children between 18 and 24 years old, who have already an own driving license but are using the parents' car.

Figure 4 indicates a monotone but non-linear relationship between population density of the region the customer is living in and the pure premium. The regression curve was fitted with $\varepsilon-$support vector regression, *cf.* Vapnik (1998). However, there is additional geographical information in the data set not explained by the population density alone. A map of the mean pure premium values computed in regions of Germany defined by the first three digits of the postal zip codes shows that there is a geographical clustering effect, see Figure 5. Typically the pure premium is increased in big German cities, e.g. in Berlin, Hamburg, Munich or Frankfurt, as was to be expected from the previous figure. However, the so-called Ruhr-Area (see the 6. biggest German city Dortmund, which is in the eastern part of the Ruhr-Area), has a similar geographical size and a similar number of inhabitants than Berlin, but the average of the pure premium values is much lower than in Berlin. On the other hand, in the north-east of Germany the mean claim sizes are increased, too, although the population density is relatively low.

4 Strategy

The following fairly general strategy has the goal to detect hidden structure in data sets from motor vehicle insurance companies by exploiting certain characteristics of such data sets.

- Most of the policy holders have no claim within a year or a certain period.
- The claim sizes are extremely skewed to the right, but there is atom in zero.
- There is a complex, high dimensional dependency structure between variables.
- There are only imprecise values available for some explanatory variables.
- Some claim sizes are only estimates.
- The data sets to be analyzed are huge.
- Extreme high claim amounts are rare events, but contribute enormously to the total sum of all claims.

We prefer statistical procedures with good robustness properties due to the fourth and fifth property in the list.

As before, let Y denote the pure premium of a customer within a year and x the vector of explanatory variables. In the first step we construct an additional stratification variable C by defining a small number of classes for the values of Y with a high amount of interpretability. For example, define a discrete random variable C with five possible values:

$$\begin{aligned} C &= 0, && \text{if } Y = 0 && \text{'no claim'} \\ &= 1, && \text{if } Y \in (0, 2000] && \text{'low pure premium'} \\ &= 2, && \text{if } Y \in (2000, 10000] && \text{'medium pure premium'} \\ &= 3, && \text{if } Y \in (10000, 50000] && \text{'high pure premium'} \\ &= 4, && \text{if } Y > 50000 && \text{'extreme pure premium'}. \end{aligned}$$

Of course, it depends on the application how many classes should be used and how reasonable boundaries can be defined. We will not address this problem here.

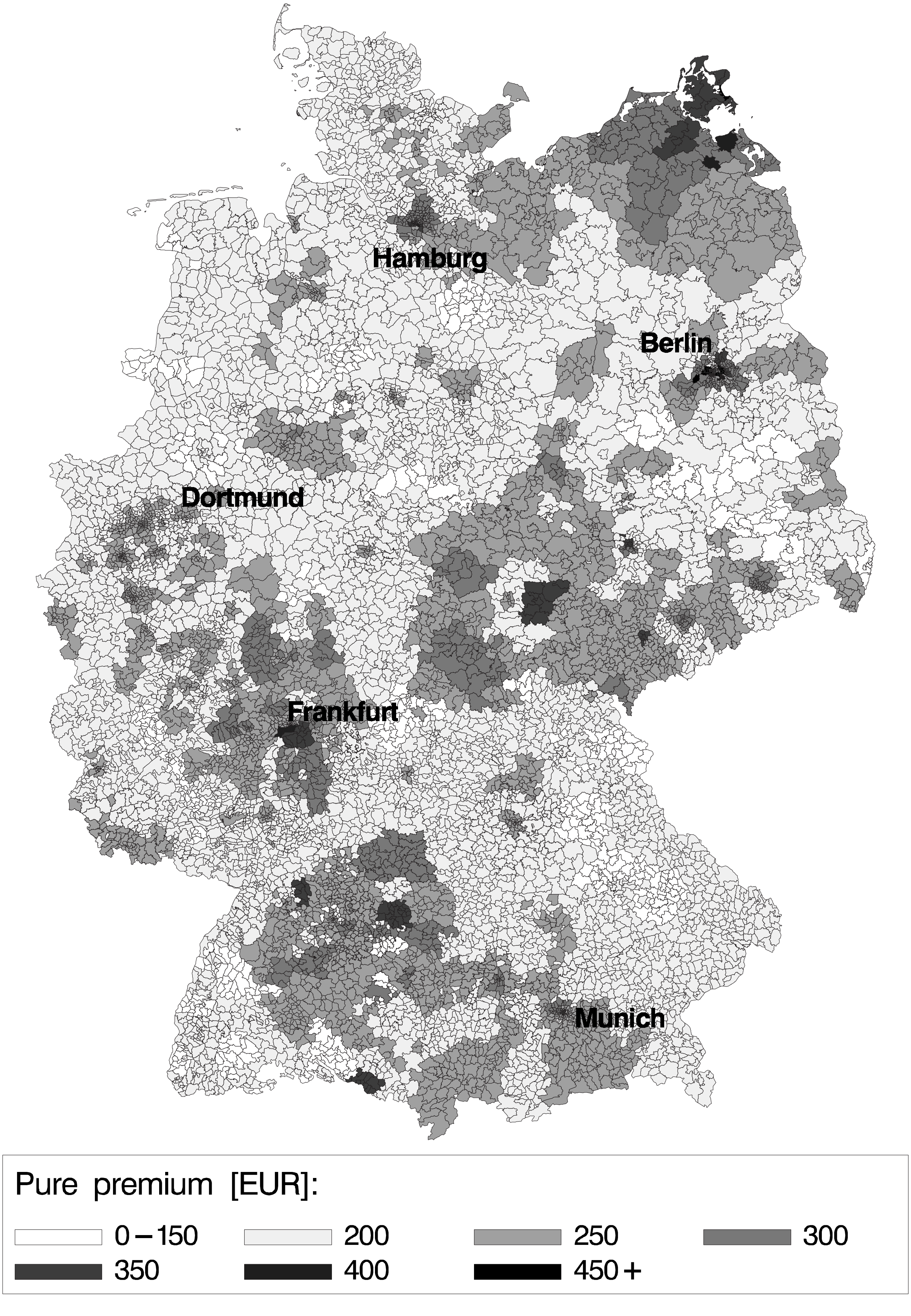
Note, that given the information that no claim occurred, it holds that

$$\mathrm{E}(Y|C = 0, X = x) \equiv 0\,. \tag{1}$$

Using (1) we can write the conditional expectation of Y given $X = x$ by

$$\begin{aligned} \mathrm{E}(Y|X = x) &= \mathrm{P}(C > 0|X = x) \times \\ & \textstyle\sum_{c=1}^{k} \mathrm{P}(C = c|C > 0, X = x) \cdot \mathrm{E}(Y|C = c, X = x)\,, \end{aligned} \tag{2}$$

Figure 5: Map of Germany for the average of the pure premium values up to 50000 EUR within geographical regions defined by the first three digits of the postal zip code.



and we denote this formula as strategy A. Note, that in (2) the summation starts with $c = 1$. Hence, it is only necessary to fit regression models to small subsets of the whole data set, see Table 1. However, one has to estimate the conditional probability $\mathrm{P}(C = 0|X = x)$ and the multi-class probabilities $\mathrm{P}(C = c|C > 0, X = x)$ for $c \in \{1, \ldots, k\}$, e.g. by a multinomial logistic regression model or by kernel logistic regression. If one splits the total data set into three subsets for training, validating, and testing, one only has to compute *predictions* for the conditional probabilities and the corresponding conditional expectations for *all* data points. Bias reduction techniques applied to the validation data set may be helpful to reduce a possible bias of the estimates.

From our point of view the indirect estimation of $\mathrm{E}(Y|X = x)$ via strategy A has practical and theoretical advantages over direct estimation of this quantity. Insurance companies are interested in estimating the terms in (2), because they contain additional information: the probability that a customer has at least one claim (which is our secondary response variable), the conditional probabilities $\mathrm{P}(C = c|C > 0, X = x)$, and the conditional expectations $\mathrm{E}(Y|C = c, X = x)$. The strategy circumvents the problem, that most observed pure premium values y_i are 0, but $\mathrm{P}(Y = 0|X = x) = 0$ for many classical approaches based on a gamma or log-normal distribution. A reduction of computation time is possible, because we only have to fit regression models to a small subset (say 5%) of the data set. The estimation of conditional class probabilities for the whole data set is often much faster than fitting a regression model for the whole data set. It is possible, that different explanatory variables show a significant impact on the response variable Y or on the conditional class probabilities for different classes defined by C. This can also result in a reduction of interaction terms. In principle, it is possible to use different variable selection methods for the $k + 1$ classes. This can be especially important for the class of extreme pure premium values: because there may be only some hundreds or a few thousands of these rare events in the data set, it is in general impossible to use all explanatory variables for these data points. Finally, the strategies have the advantage, that different techniques can be used for estimating the conditional class probabilities $\mathrm{P}(C = c|X = x)$ and for estimating the expectations $\mathrm{E}(Y|C = c, X = x)$ for different values of C. Examples for reasonable pairs are:

- Multinomial logistic regression + Gamma regression
- Robust logistic regression + semi-parametric regression
- Multinomial logistic regression + ε-Support Vector Regression (or $\nu-$SVR)
- Kernel logistic regression (KLR) + ε-Support Vector Regression (or $\nu-$SVR)
- Classification trees + regression trees
- A combination of the pairs given above, where some additional explanatory variables are constructed as a result of classification and regression trees.

Even for data sets with several million of customers it is not possible to fit simultaneously all high-dimensional interaction terms with classical statistical methods such as logistic regression or gamma regression, because the number of interaction terms increases too fast.

The combination of kernel logistic regression and $\varepsilon-$support vector regression (*cf.* section 5), both with an RBF kernel has the advantage, that important interaction terms are fitted automatically without the need to specify them manually.

We like to mention that some statistical software packages (e.g. R) may run into trouble in fitting multinomial logistic regression models for large and high dimensional data sets. Two reasons are that the dimension of the parameter vector can be quite high and that a data set with many discrete variables recoded into a large number of dummy variables will perhaps not fit into the memory of the computer. To avoid a multinomial logistic regression model one can consider all pairs and then use pairwise coupling, *cf.* Hastie and Tibshirani (1996).

Of course, the law of total probability offers alternatives to (2). The motivation for the following alternative, say strategy B, is that we first split the data into the groups 'no claim' versus 'claim' and then split the data with 'claim' into 'extreme pure premium' and the remaining $k-1$ classes:

$$\begin{aligned} \mathrm{E}(Y|X=x) &= \mathrm{P}(C>0|X=x) \times \qquad (3)\\ &\quad \{\mathrm{P}(C=k|C>0,X=x)\cdot \mathrm{E}(Y|C=k,X=x) + \\ &\quad\ [1-\mathrm{P}(C=k|C>0,X=x)] \times \\ &\quad\ \sum_{c=1}^{k-1} \mathrm{P}(C=c|0<C<k,X=x)\cdot \mathrm{E}(Y|C=c,X=x)\}\,. \end{aligned}$$

This formula shares with (2) the property, that it is only necessary to fit regression models to subsets of the whole data set. Of course, one can also interchange the steps in the above formula, which results in strategy C:

$$\begin{aligned} \mathrm{E}(Y|X=x) &= \mathrm{P}(C=k|X=x)\cdot \mathrm{E}(Y|C=k,X=x) \qquad (4)\\ &\quad + [1-\mathrm{P}(C=k|X=x)]\cdot \{\mathrm{P}(C>0|C\neq k,X=x) \times \\ &\quad\ \sum_{c=1}^{k-1} \mathrm{P}(C=c|0<C<k,X=x)\cdot \mathrm{E}(Y|C=c,X=x)\}\,. \end{aligned}$$

Note, that two big binary classification problems have to be solved in (4), whereas there is only one such problem in (3).

5 Kernel logistic regression and $\varepsilon-$support vector regression

In this section we briefly describe two modern methods based on convex risk minimization in the sense of Vapnik (1998), see also Schölkopf und Smola (2002).

In statistical machine learning the major goal is the estimation of a functional relationship $y_i \approx f(x_i) + b$ between an outcome y_i belonging to some set $\mathcal{Y}$ and a vector of explanatory variables $x_i = (x_{i,1}, \ldots, x_{i,k})' \in \mathcal{X} \subseteq \mathbb{R}^p$. The function f and the intercept parameter b are unknown. The estimate of (f, b) is used to get predictions of an unobserved outcome y_{new} based on an observed value x_{new}. The classical assumption in machine learning is, that the training data (x_i, y_i) are independent and identically generated from an underlying unknown distribution P for a pair of random variables (X_i, Y_i), $1 \leq i \leq n$. In applications the training data set is often quite large, high dimensional and complex. The quality of the predictor $f(x_i) + b$ is measured by some loss function $L(y_i, f(x_i) + b)$. The goal is to find a predictor $f_{\mathrm{P}}(x_i) + b_{\mathrm{P}}$ which minimizes the expected loss, *i.e.*

$$\mathrm{E}_{\mathrm{P}}\, L(Y, f_{\mathrm{P}}(X) + b_{\mathrm{P}}) = \min_f \mathrm{E}_{\mathrm{P}}\, L(Y, f(X) + b), \tag{5}$$

where $\mathrm{E}_{\mathrm{P}}\, L(Y, f(X) + b) = \int L(y, f(x) + b) d\mathrm{P}(x, y)$ denotes the expectation of L with respect to P. We have $y_i \in \mathcal{Y} := \{-1, +1\}$ in the case of binary classification problems, and $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ in regression problems.

As P is unknown, it is in general not possible to solve the problem (5). Vapnik (1998) proposed to estimate (f, b) as the solution of a empirical regularized risk. His approach relies on three important ideas: (1) restrict the class of all functions f to a broad subclass of functions belonging to a certain *Hilbert space*, (2) use a *convex* loss function L to avoid computational intractable problems which are NP-hard, and (3) use a *regularizing term* to avoid overfitting. Let $L : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ be an appropriate convex loss function. Estimate (f, b) by the solution of the following empirical regularized risk:

$$(\hat{f}_{n,\lambda}, \hat{b}_{n,\lambda}) = \arg \min_{f \in \mathcal{H},\, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i) + b) + \lambda \|f\|^2_{\mathcal{H}}, \tag{6}$$

where $\lambda > 0$ is a small regularization parameter, $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) of a kernel k, and b is an unknown real-valued offset. The problem (6) can be interpreted as a stochastic approximation of the minimization of the following theoretical regularized risk:

$$(f_{\mathrm{P},\lambda}, b_{\mathrm{P},\lambda}) = \arg \min_{f \in \mathcal{H},\, b \in \mathbb{R}} \mathrm{E}_{\mathrm{P}}\, L(Y, f(X) + b) + \lambda \|f\|^2_{\mathcal{H}}\,. \tag{7}$$

In practice, it is often numerically better to solve the dual problem of (6). In this problem the RKHS does not occur explicitly, instead the corresponding kernel is involved. The choice of the kernel k enables the above methods to efficiently estimate not only linear, but also non-linear functions. Of special importance is the Gaussian radial basis function (RBF) kernel

$$k(x, x') = \exp(-\gamma \|x - x'\|^2)\,, \quad \gamma > 0, \tag{8}$$

which is a universal kernel on every compact subset of $\mathbb{R}^d$, *cf.* Steinwart (2001).

For the case of binary classification, popular loss functions depend on y and (f, b) via $v = y(f(x) + b)$. Special cases are:

- Support Vector Machine (L1-SVM): $L(y, f(x) + b) = \max\{1 - y(f(x) + b), 0\}$
- Least Squares (L2-SVM): $L(y, f(x) + b) = [1 - y(f(x) + b)]^2$
- Kernel Logistic Regression (KLR): $L(y, f(x)+b) = \log(1+\exp[-y(f(x)+b)])$
- AdaBoost: $L(y, f(x)+b) = \exp[-y(f(x)+b)]$, *cf.* Freund and Schapire (1996) and Friedman, Hastie and Tibshirani (2000).

Kernel logistic regression has the advantage, that it estimates $\log(\frac{P(Y=+1|X=x)}{P(Y=-1|X=x)})$, i.e. $P(Y = +1|X = x) = (1 + e^{-[f(x)+b]})^{-1}$, such that scoring is possible. Note, that the support vector machine 'only' estimates whether $P(Y = +1|X = x)$ is above or below $\frac{1}{2}$.

For the case of regression, Vapnik (1998) proposed the $\varepsilon-$support vector regression ($\varepsilon-$SVR) which is based on the $\varepsilon-$insensitive loss function

$$L_\varepsilon(y, f(x) + b) = \max\{0, |y - [f(x) + b]| - \varepsilon\},$$

for some $\varepsilon > 0$. Note, that only residuals $y - [f(x) + b]$ lying outside of an $\varepsilon-$tube are penalized. Strongly related to $\varepsilon-$support vector regression is $\nu-$support vector regression, *cf.* Schölkopf und Smola (2002).

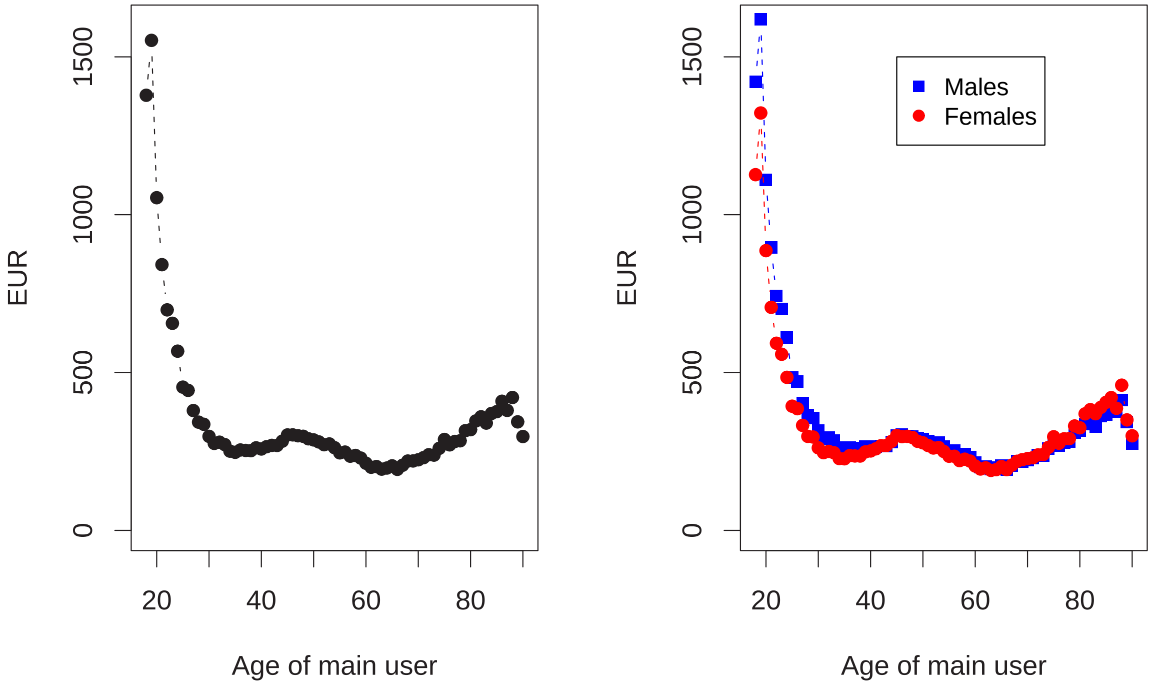
6 Application

This section describes some results of applying the strategy A, i.e. (2), to the data set from the Verband öffentlicher Versicherer in Düsseldorf, Germany, see Section 3. We use a nonparametric approach based on the pair (KLR, $\varepsilon-$SVR) to fit the observed pure premium values. The following 8 explanatory variables were selected: gender of the main user, age of the main user, driving distance within a year, geographic region, a variable describing whether the car is kept in a garage, a variable for the population density computed via the first 3 digits of the postal zip code, a variable related to the number of years the customer had no claim, and the strength of the engine.

We used a binary logistic regression model to fit the conditional probabilities that $P(C > 0|X = x)$ (SAS, PROC LOGISTIC). The classical estimator in binary logistic regression is the maximum likelihood estimator. If the data set has complete separation or quasi-complete separation, the maximum likelihood estimates do not exist. Unfortunately, most statistical software packages do not check whether the maximum likelihood estimates exist. Christmann and Rousseeuw (2001) developed software to check whether the data set has complete or quasi-complete separation. Christmann, Fischer and Joachims (2002) compared such methods with methods based on Vapnik's support vector machine with a linear kernel. Rousseeuw and Christmann (2003) proposed the hidden logistic regression model, which is strongly related to logistic regression, and investigated estimates which always exist.

The conditional probabilities $P(C = c|C > 0, X = x)$, $c \in \{1, 2, 3, 4\}$, were estimated via kernel logistic regression. First, we estimated the probabilities for all pairs $P(C = j|C \in \{i, j\}, X = x)$, where $1 \leq i < j \leq 4$. Keerthi et al. (2002) developed a fast dual algorithm for kernel logistic regression for the case of pattern recognition. Rüping (2003) implemented this algorithm in the program myKLR. We used myKLR to estimate the probabilities for these pairs. Then the multi-class probabilities $P(C = c|C > 0, X = x)$, $c \in \{1, 2, 3, 4\}$, were computed with pairwise coupling, *cf.* Hastie and Tibshirani (1996). The expected conditional pure premium $E(Y|C > 0, X = x)$, $c \in \{1, 2, 3, 4\}$, was estimated using $\varepsilon-$support vector regression (svm in the R-package e1071, Leisch et al., 2003). We used the exponential radial basis function kernel both for kernel logistic regression and for $\varepsilon-$support vector regression. Tuning constants were set to the values proposed by

Figure 6: Results of applying strategy A, part I. Left: expected pure premium stratified by age of the main user. Right: expected pure premium stratified by gender and age of the main user.



Cherkassky and Ma (2004). The data set was split up into subsets: training (50%, $n \approx 2.2E6$), validation (25%, $n \approx 1.1E6$), and test (25%, $n \approx 1.1E6$).

We restrict attention to the explanatory variable age of the main user, but of course the calculations were done for all 8 explanatory variables simultaneously. For all data points in the test data set, the estimates of the expected pure premium $\mathrm{E}(Y|X = x)$ and of the conditional probabilities and conditional expectations defined in (2) were computed. Figure 6 to Figure 8 show the averages of these estimates stratified by the age of the main user or by gender and age of the main user.

Figure 6 shows the estimates of $\mathrm{E}(Y|X = x)$ stratified by age, which have the same sharp peak for young people, a moderate peak around 50 years, and an increase for elderly people which were already visible by a univariate analysis, *cf.* Figure 3. There is an interaction term for young people, say for the age group 18 to 24 years, see the right plot in Figure 6. The expected pure premium for 18 to 20 years old males is approximately 300 EUR higher than for females of the same age. This is a rather big difference, because the base risk is approximately 360 EUR. Our method based on strategy A using a combination of the two nonparametric methods kernel logistic regression and $\varepsilon-$support vector regression was able to detect this interaction term although we did *not* model such an interaction term. It was confirmed by the Verband öffentlicher Versicherer, that this interaction term is not an artefact, but typical for their data sets.

However, from our point of view the main strength of strategy A becomes visible if one investigates the conditional probabilities and conditional expectations which were fit by (2). Figure 7 shows the conditional probabilities stratified by gender and age of the main user. The probability of a claim, i.e. $\mathrm{P}(Y > 0|X = x)$, shows again a similar shape than the corresponding curve in Figure 3. However, the conditional probability of a claim in the interval $(0, 2000]$ EUR given the event that a claim occurred, increases for people of at least 18 years, *cf.* the subplot for $\mathrm{P}(C = 1|C > 0, X = x)$ in Figure 7. This is in contrast to the corresponding subplots for moderate, high or extreme pure premium values, see the subplots for $\mathrm{P}(C = c|C > 0, X = x)$, $c \in \{2, 3, 4\}$ in Figure 7. Especially the last two subplots show, that young people have a 2 to 3 times higher probability in producing a claim than more elderly people.

The effect of gender and age of the main user on the expected pure premium given the knowledge of C, i.e. whether it is a small, moderate, high or extreme pure premium, is shown in Figure 8. Because $\mathrm{E}(Y|C = 0, X = x) \equiv 0$, we dropped this subplot in Figure 8. The impact of these two explanatory variables is much lower than for the conditional probabilities. Nevertheless, the subplots for $\mathrm{E}(Y|C = c, X = x)$, $c \in \{1, 2, 3\}$, show again that young people have a higher expected pure premium than more elderly people, even if one conditions with respect to the class variable C. Please note, that the $y-$axis of the subplots for $\mathrm{E}(Y|C = c, X = x)$ are different, otherwise it would be hard to see any differences between the curves for males and females. A comparison of the estimated quantities shown in Figures 7 and 8 with the corresponding empirical averages and relative frequencies (not given here) of a univariate description of the data set shows that the fit using the pair (KLR, $\varepsilon-$SVR) was quite successful.

Concluding, people belonging to the age group 18 to 24 years have a higher pure premium and a higher probability to have a claim than more elderly customers. Strategy A offers a lot of additional information. The application of the pair kernel logistic regression and $\varepsilon-$support vector regression was able to detect automatically an interaction term for young people with respect to gender and a moderate increased pure premium for main users with an age around 50 years.

7 Discussion

In this paper certain characteristics of data sets from motor vehicle insurance companies were described. It was proposed to estimate the pure premium in motor vehicle insurance data in an indirect manner. There are several advantages of this strategy in contrast to a straightforward estimation of the expected pure premium. The strategy exploits knowledge of certain characteristics of data sets from motor vehicle insurance companies and estimates conditional probabilities and conditional expectations given the knowledge of an auxiliary class variable C describing the magnitude of the pure premium. This proposal offers additional insight into the structure of the data set, which is not visible with a direct estimation of the expected pure premium alone. Such additional information is valuable for insurance companies and can also be useful for aspects not related to the construction of insurance tariffs, *e.g.* in the context of direct marketing. Further, different estimation techniques and different variable selection methods can be used for the classes of pure premium defined by the auxiliary variable C.

Figure 7: Results of applying strategy A, part II. Pure premium and conditional probabilities stratified by gender and age of the main user.

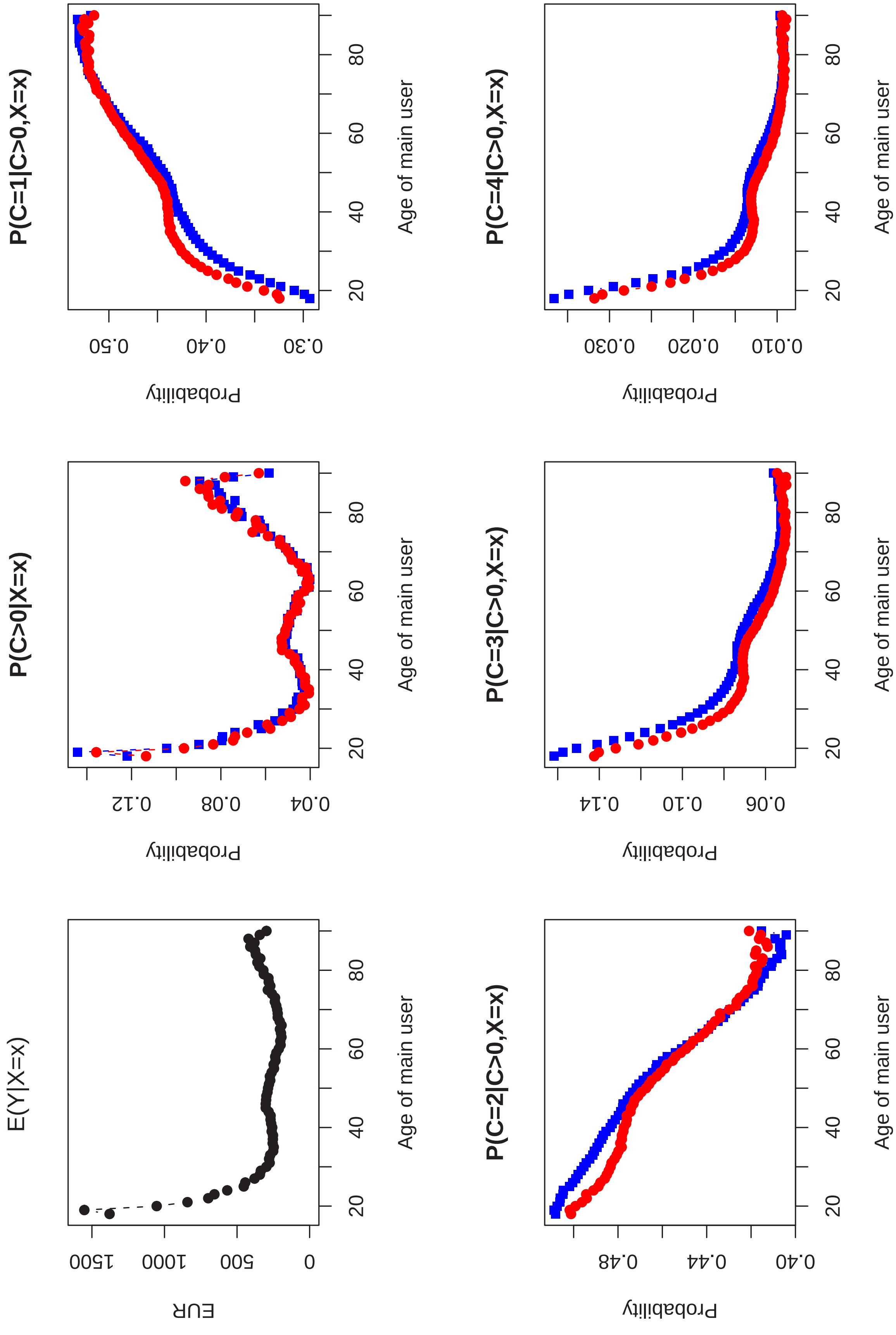
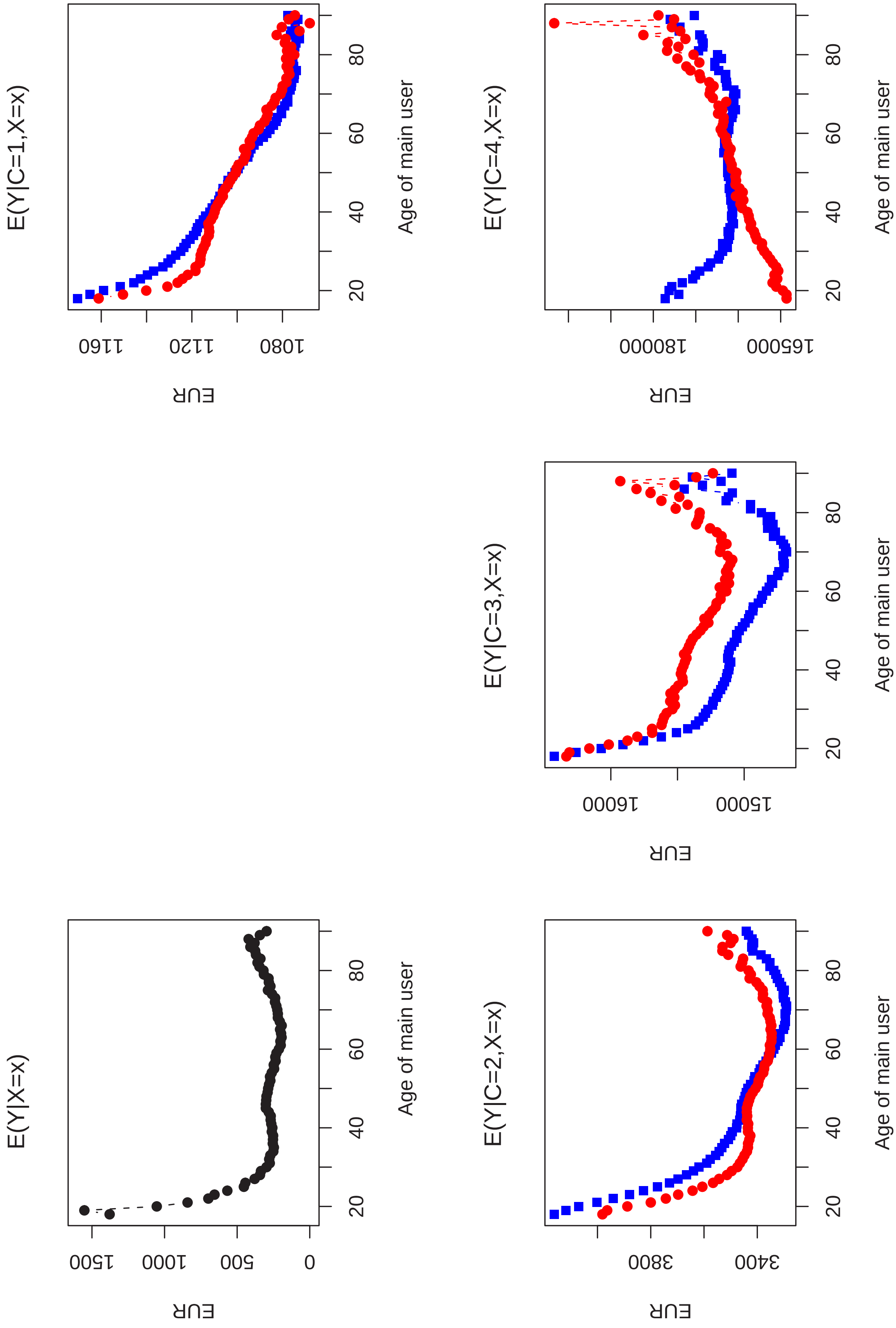


Figure 8: Results of applying strategy A, part III. Pure premium and conditional expectations stratified by gender and age of the main user.



The application of the proposed strategy was illustrated for a large data set containing data from 15 motor vehicle insurance companies from Germany. We used a nonparametric approach based on a combination of kernel logistic regression and $\varepsilon-$support vector regression. Both techniques belong to the class of statistical machine learning methods based on convex risk minimization, see Vapnik (1998), Schölkopf and Smola (2002), and Hastie, Tibshirani, Friedman (2001). Such methods can fit quite complex data sets and have good prediction properties. The combination of these methods was able to detect an interesting interaction term and violations of a monotonicity assumption without the necessity that the researcher has to model interaction terms or polynomial terms manually.

Christmann and Steinwart (2003) showed that a large class of such convex risk minimization methods have good robustness properties. Special cases are kernel logistic regression and support vector machine. Robustness is an important aspect in analyzing insurance data sets, because some explanatory variables may only be measured in an imprecise manner and some reported values of the claim size are only estimates and not the true values. On the other hand, in contrast to some other areas of applied statistics, extreme high response values can not be dropped from the data set because this would systematically underestimate the expected pure premium.

In the literature it is often recommended to determine the tuning constants for kernel logistic regression and for $\varepsilon-$support vector regression, say c_{cost}, λ, γ, and ε, via a grid search or by cross-validation. Such methods can be extremely time-consuming for such a big data set we were dealing with in section 6. Some preliminary experiments are indicating that these constants can be chosen in a reasonable way if they are determined as the solution of an appropriate optimization problem, e.g. by minimizing the mean squared error of the predictions for the validation data set or by using a chi-squared type statistic similar to the one used by the Hosmer-Lemeshow test for checking goodness-of-fit in logistic regression models, c.f. Hosmer and Lemeshow (1989) or Agresti (1996). The Nelder-Mead algorithm, c.f. Nelder and Mead (1965), worked well in this context for some test data sets, but a systematic investigation of this topic is beyond the scope of this paper.

Although there were approximately 3300 customers with an observed pure premium above 50000 EUR in the data set we used for illustration purposes, an attractive alternative to $\varepsilon-$support vector regression for this subgroup is extreme value theory, *e.g.* by fitting generalized Pareto distributions of the main knots of a regression tree or more sophisticated methods.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures"). I am grateful to A. Wolfstein and Dr. W. Terbeck from the Verband öffentlicher Versicherer in Düsseldorf, Germany, for making available the data set and for helpful discussions.

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* Wiley, New York.

Beirlant, J., de Wet, T., Goegebeur, Y. (2002). Nonparametric Estimation of Extreme Conditional Quantiles. Katholieke Universiteit Leuven, Universitair Centrum voor Statistiek. Technical Report 2002-07.

Celebrián, A.C., Denuit, M., Lambert, P. (2003). Generalized Pareto Fit to the Society of Actuaries' Large Claims Database. *North American Actuarial Journal*, **7**, 18-36.

Cherkassky, V., Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, **17**, 113-126.

Christmann, A., Fischer, P., Joachims, T. (2002). Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Computational Statistics*, **17**, 273-287.

Christmann, A., Rousseeuw, P.J. (2001). Measuring overlap in logistic regression. *Computational Statistics and Data Analysis*, **37**, 65-75.

Christmann, A., Steinwart, I. (2003). *On robust properties of convex risk minimization methods for pattern recognition.* University of Dortmund, SFB-475, Technical Report 15/2003.

Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modelling Extreme Events for Insurance and Finance.* Springer, Berlin.

Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. In: *Machine Learning: Proceedings of the 13th International Conference*, Morgan Kauffman, San Francisco, pp. 148-156.

Friedman, J., Hastie, T., Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.*, **28**, 337-407.

Hastie, T., Tibshirani, R. (1996). *Classification by Pairwise Coupling.* Stanford University. Preprint. `http://www-stat.stanford.edu/∼hastie`.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer, New York.

Hosmer, D.W., Lemeshow, W. (1989). *Applied Logistic Regression.* Wiley, New York.

Keerthi, S.S., Duan, K., Shevade, S.K., Poo, A.N. (2002). *A fast dual algorithm for kernel logistic regression.* National University of Singapore. Preprint. `http://guppy.mpe.nus.edu.sg/∼mpessk`

Leisch, F. et al. (2003). R package e1071. `http://cran.r-project.org`.

Nelder, J. A., Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, **7**, 308-313.

Rousseeuw, P.J., Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*, **43**, 315-332.

Rüping, S. (2003). *myKLR - kernel logistic regression.* University of Dortmund. Department of Computer Science.
http://www-ai.cs.uni-dortmund.de/SOFTWARE.

Schölkopf, B., Smola, A. (2002). *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press.

Scovel, J.C., Steinwart, I. (2003). Fast rates for support vector machines. Los Alamos Technical Report LA-UR-03-9117.

Steinwart, I. (2001). On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of Machine Learning Research*, **2**, 67-93.

Vapnik, V. (1998). *Statistical Learning Theory.* Wiley, New York.