Data misrepresentation detection for insurance underwriting fraud prevention

Félix Vandervorst[a,b,c], Wouter Verbeke[b], Tim Verdonck[c,d,*]

[a] *Allianz Benelux, Data Office, Koning Albert II Laan 32, Brussels 1000, Belgium*
[b] *KU Leuven, Faculty of Economics and Business, Naamsestraat 69, Leuven 3000, Belgium*
[c] *University of Antwerp, Department of Mathematics, Middelheimlaan 1, Antwerp 2020, Belgium*
[d] *KU Leuven, Department of Mathematics, Celestijnenlaan 200B, Leuven 3001, Belgium*

Abstract

Premium fraud concerns data misrepresentation committed by an insurance customer with the intent to benefit from an unduly low premium at the underwriting of a policy. In this paper, we propose a novel approach for evaluating the risk of underwriting premium fraud at the time of application in the presence of potentially misrepresented self-reported information. The aim of the approach is to support insurance companies in identifying fraudulent applications and their decisions to underwrite insurance contract propositions. Likewise, it can be use to make straight-through processing (i.e. automated) underwriting systems more fraudproof, by e.g., triggering a validation on applications prone to misrepresentations. Our approach is based on conditional density estimates for a set of validated contracts. The proposed approach does not require historical fraud labels and can adapt to changes in pricing policy. Moreover, the approach can be used to detect outliers in addition to predicting underwriting fraud and is extended to multivariate self-reported data. We further demonstrate a link between Shapley values in common conditional expectation problems and conditional density estimations to make our approach explainable. We report a case study involving motor insurance underwriting, in which a driver's identity and driving record can be misrepresented to benefit from an unduly low premium; the results indicate the effectiveness of the proposed approach for detecting and preventing underwriting fraud.

Keywords: Insurance Underwriting Fraud, Premium Fraud, Data Misrepresentation, Machine Learning, Nonlife Insurance.

*Corresponding author: Félix Vandervorst

Email addresses: felix.vandervorst@allianz.be (Félix Vandervorst), wouter.verbeke@kuleuven.be (Wouter Verbeke), Tim.Verdonck@uantwerpen.be (Tim Verdonck)

1. Introduction

Premium fraud concerns the intentional misrepresentation of information that is provided at the time of underwriting of an insurance contract in order to benefit from an unduly low premium. Premium fraud encompasses a variety of patterns and scenarios. In life insurance, for example, self-reported smoking status often has a material impact on expected mortality and, thus, on the premium price [1]. In the absence of an effective and credible strategy for confirming the true smoking status of the policyholder, the latter has a peculiar incentive to conceal this information. In motor insurance contracts, premium prices are often prohibitive for young and inexperienced drivers. Those drivers have a material incentive to overstate their driving experience, for instance, by having an older relative listed on the contract as the main driver on their behalf. [2] estimates that global motor insurance premium fraud amounts to 29 billion dollars annually. In workers' compensation, the company seeking insurance may misrepresent its payroll information to benefit from insurance coverage at a lower premium price [3]. The Economic Policy Institute reports that 10 to 20 percent of workers are misreported in the U.S.A., resulting in lower premiums paid by employers [4].

Premium fraud causes a loss of premiums, which will ultimately result in an increase in insurance premiums for all policyholders, including those who have not committed premium fraud. Another impact of premium fraud lies in the untruthful data collected by the insurance company and the knock-on impact on all decisions or actuarial models that are built on those data. Therefore, the ability to detect such fraud is essential.

Since the nineteen-nineties, researchers and practitioners have been investigating data-driven approaches for improving the power of fraud detection systems [5; 6; 7]. However, unlike many other types of fraud, insurance fraud is not self-revealing [8], i.e., the company will not gain knowledge on the fraudulent status of a claim or contract after an event without proactive and costly investigations. The non-self-revealing nature of insurance fraud (including premium fraud) poses a serious challenge for both its *identification* (the company cannot label past frauds or estimate the extent of the problem) and *prediction* (without labels, supervised learning is not possible, and unsupervised learning is difficult to validate). The literature on underwriting fraud is focused on the restricted setting in which one binary self-reported variable is potentially misrepresented and does not consider adaptation to changes in pricing policy.

In this paper, we contribute to the literature by presenting a novel approach for detecting premium fraud by evaluating data misrepresentation risk. The proposed approach is based on conditional density estimation (CDE) of a set of self-reported variables and is adaptive to exogenous changes in

pricing policy. Moreover, we present a case study based on real motor insurance data.

The remainder of this paper is structured as follows. In Section 2, we present a brief literature review on data misrepresentation and its applications in the insurance underwriting literature. Section 3 describes the methodology, in which we detail how premium fraud detection can be approached as a conditional density estimation problem with mixed data types, using the orthogonal basis projection trick of [9] and an adaptation of the random forest algorithm [10]. We also present how the pricing policy can be used in combination with CDE to represent financial motivation. In Section 4, we present an application of the presented methodology to real motor insurance data. Finally, Section 5 presents conclusions and directions for further research.

2. Literature Review

In [8], underwriting fraud is defined as covering, for example, "the dissimulation of information during application (application fraud) to obtain coverage or a lower premium (premium fraud), the deliberate concealment of existing insurance contracts covering the same property and casualty (P&C) risk, and underwriting coverage for fictitious risks" [8, p. 3]. The motivation behind premium fraud is the financial gain resulting from a misstatement in the underwriting information, such that the premium according to the misrepresented information is lower than the *true* premium.

The problem of data misrepresentation has a long history in the economic literature, in which researchers aim at estimating statistical models based on contaminated data [11]. The data misrepresentation problem is often formulated in terms of conditional densities [11]. For instance, in [12] the conditional density of the dependent variable Y, knowing the observed variable Z^* of its unobserved variable Z is formulated as

$$f(y|z^*) = \int f(y|z)g(z|z^*)dz$$

where $g(z|z^*)$ is the conditional distribution of the true value of $Z = z$, knowing the observed value $Z^* = z^*$, $f(y|z)$ the error-free conditional distribution and $f(y|z^*)$ the error-contaminated conditional distribution. The literature on measurement error is focused on contaminated samples of (Y, Z^*), while making distributional assumptions on the form of the misrepresentation [11]. To the best of our knowledge, [13] and [14] are the first studies to take this type of statistical approach to data misrepresentation in the context of insurance underwriting. They study the quantification of the misrepresentation of a true value Z (e.g., smoking status in a health insurance contract) by one corresponding binary self-reported random variable $Z^* \in \{0, 1\}$, compared to a parametric distribution of the loss $f(y|z)$ (e.g., gamma or Poisson) assumed in the insurance contract.

$$f(y|z^* = 1) = f(y|z = 1),$$
$$f(y|z^* = 0) = (1 - P(z = 1|z^* = 0))f(y|z = 0) + P(z = 1|z^* = 0)f(y|z = 1),$$

where $P(z = 1|z^* = 0)$ is the misrepresentation probability parameter. This equation is identifiable in the case of unidirectional misclassification, i.e., where $f(y|z^* = 1) = f(y|z = 1)$. Whereas this approach is informative on the prevalence of the problem in a population through the estimation of $P(z = 1|z^* = 0)$, it is of little use in improving the underwriting decision process because it is not a contract-dependent representation. To address this shortcoming, [15] proposes an extension of the framework established in [13] and [14] that includes other correctly reported variables, thereby providing a misrepresentation probability that is dependent on the contract details in the context of generalized linear models. However, the proposed models consider only discrete binary factors with unidirectional misclassification. The assumptions made in the above literature can be unrealistic and limiting in practical applications. Indeed, in many insurance applications, multiple variables may be self-reported, continuous or discrete. Moreover, insurance fraud is dynamic. Fraudsters swiftly capitalize on opportunities [8], such as changes of the pricing policy. The misclassification direction is not necessarily constant in this context. Also, the loss information is available only after a certain observation period.

An alternative approach to the data misrepresentation problem uses validation data (containing at least values of Z), although often not available [11]. When a validated dataset for (sets of) two paired variables (Z, X) is available, the problem of estimating the data misrepresentation of variable Z can be formulated as a CDE problem in which Z is the target variable and $f(z|x)$ is the conditional density one wishes to estimate. The data used for scoring insurance contracts are of mixed data types, including continuous variables (e.g., age) and discrete factors (e.g., car model, zip code). In the case that the target variable is binary, the posterior probability $P(Z = 1|X)$ coincides with the conditional mean $\mathbb{E}[Z|X]$ [16]. In that case, many well-documented methods are available that can handle a large number of covariates, in particular, Lasso regressions for mixed data types [17] or decision tree ensembles such as random forests or boosted trees, which are well suited for high-dimensional problems involving mixed data types. However, how to obtain an estimate $\hat{f}(z|x)$ for a continuous, potentially multivariate, variable Z is a less studied statistical problem. One class of approaches in the CDE literature relies on the use of orthogonal transformations of $f(z|x)$. The recent work [9] proposes a flexible conditional density estimator (FlexCode) that projects the target variable Z onto an orthogonal basis [18]. Each projection onto the basis is estimated via a well-known

conditional mean regression algorithm (e.g. linear models, decision trees) and the goodness-of-fit is evaluated with a variation of the integrated squared error (CDE loss). Another recent approach to conditional density estimation is to take the CDE loss of [9] as a variable splitting function in the random forest algorithm of [19], using a variant of a tree structure to estimate $f(z|x)$. The most important departure from the original algorithm is the use of the CDE loss of Equation (5), which is used as a splitting criterion in the construction of each tree t. Random forest CDE (RFCDE) can be interpreted as a weighted kernel density method in which the weights are obtained from a random forest. In this paper, we will focus on those two techniques, FlexCode and RFCDE. Alternative approaches exist for conditional density estimation, albeit sub-optimal in this context. The estimation of $\hat{f}(z|x)$ can be approached directly, via its conditional distribution $\hat{F}(z|x)$, or via its conditional quantile function $\hat{Q}_{z|x}(\alpha)$. For instance, quantile regression forests [20] estimate $\hat{Q}_{z|x}(\alpha)$ and $\hat{F}(z|x)$; however, the estimated distribution $\hat{F}(z|x)$ is nondifferentiable, and this method is not suited for multivariate responses. In this paper, we focus on methods in which the conditional density function $\hat{f}(z|x)$ is directly evaluated to avoid differentiability and invertibility conditions on $\hat{F}$ and $\hat{Q}_{z|x}(\alpha)$. Moreover, we have no prior information on the underlying distribution $f(z|x)$, which can potentially exhibit multimodality; therefore, we consider only nonparametric methods to avoid having to make assumptions on the shape of the distribution $f(z|x)$. In the nonparametric literature, a common approach for estimating the conditional density is to combine the estimation of the joint density $\hat{f}(z,x)$ with the estimation of the marginal density $\hat{f}(x)$ according to $\hat{f}(z|x) = \frac{\hat{f}(z,x)}{\hat{f}(x)}$. In one class of nonparametric methods, this estimation problem is addressed by means of kernel functions. [21] proposes a kernel density estimator whose bandwidth is estimated via cross-validation and that handles irrelevant covariates. However, this method does not scale well with the dimensionality and size of the dataset, making it impractical for larger datasets.

3. Methodology

3.1. Problem formulation

Let us define $\mathcal{U}$ as the true set of underwriting information and $\mathcal{U}^*$ as the reported set of underwriting information. An insurance premium $\mu(\mathcal{U}^*)$ is the coverage price offered by the insurance company based on the reported information. The job of an insurance underwriter is to assess whether an application with information $\mathcal{U}^*$ is truthful, that is, whether the reported information is equal to its true value $\mathcal{U}$. In particular, the underwriter must ensure that no misrepresentation has been committed in order to benefit from an unduly low premium, i.e., $\mu(\mathcal{U}^*) < \mu(\mathcal{U})$.

Furthermore, we define two subsets of underwriting information, namely, the self-reported information Z and the correctly measured information X, such that $\mathcal{U} = Z \cup X$. By construction, only the self-reported information is subject to misrepresentation; accordingly, $\mathcal{U}^* = Z^* \cup X$.

In Table 1, we present a sketch of a set of insurance contract applications where $dim(\mathcal{U}^*) = p+q$, $dim(Z^*) = q$, and $dim(X) = p$.

Table 1: Underwriting problem: Applications submitted to underwriting agents sequentially at times $t_1 < t_2 < ... < t_m$. If an application is validated by an agent, it becomes a contract.

	Premium	Self-Reported			...	Correctly Measured		
	$\mu(Z^*, X)$	Z_1^*	...	Z_q^*	...	X_1	...	X_p
Application 1, t_1:	1200	23	...	A	...	1100	...	18
Application 2, t_2:	800	20	...	B	...	1200	...	21
Application 3, t_3:	700	53	...	A	...	450	...	19
...								
Application m, t_m:	980	28	...	A	...	3400	...	32

The premium price $\mu(Z^*, X)$ is the price that is offered to the customer given the underwriting information, which characterizes the risk profile of the customer. This is the price that provides the customer with an incentive to misrepresent information in order to benefit from a lower premium $\mu(Z^*, X)$.

The distinction between self-reported and other information enables us to formulate the definition of premium fraud risk as follows:

$$\text{Premium Fraud Risk} := P(\mu(Z^*, X) < \mu(Z, X)) \tag{1}$$

3.2. Self-reported and correctly measured sets of underwriting information

Naturally, an insurance company would prefer that all required underwriting information include only non-self-reported data. In this case, $dim(X) = dim(U)$, all information is correctly measured, and the premium fraud risk in Equation (1) is nonexistent. However, the inclusion of self-reported data may add discriminative power and improve the pricing accuracy compared to the use of correctly measured data alone. Insurers need to find the right balance between exposure to data misrepresentation and pricing accuracy.

As explained earlier, data deemed to belong to the correctly measured set X are data for which the policyholder does not have the opportunity to misrepresent their values. The definition and scope of the correctly measured variables are specific to the context of each underwriting system. In motor

insurance, the detailed characteristics of vehicles may be sourced from third-party data or national registers. The details on the drivers, however, may be entirely self-reported. In property insurance, geographic and demographic variables can be sourced from external data, whereas data on the risk behaviors of the tenants of buildings are likely to be entirely self-reported. In health insurance, smoking status may be self-reported, whereas the age of an insured person could be verified against a national registry.

Depending on the problem at hand, some data that are technically reported by the policyholder could still be considered correctly measured. For instance, data that would certainly result in denial of insurance coverage at the occurrence of a claim are most unlikely to be misrepresented [22] since there is no financial incentive and, hence, no motivation.

The definitions of sets X and Z may vary over time, depending on the controls, opportunities and pricing policies. The relevant processes and data are also different from one insurance company to another.

For instance, the development of national registers of nonsensitive individual demographic data that are accessible to insurance companies may expand their sets of correctly measured data. On the other hand, data privacy regulations may require insurance companies to exclude some external data and thereby increase their reliance on self-reported data.

3.3. Conditional premium fraud risk

The primary objective of our premium fraud model is to provide a risk score based on self-reported information for a given application. Let us expand the previous definitions:

$$\begin{aligned}\text{Conditional Premium Fraud Risk} &:= P(\mu(Z^*, X) < \mu(Z, X) | X = x, Z^* = z^*) \\ &= \int_{\mathcal{Z}} \mathbb{1}_{\mu(z^*,x)<\mu(z,x)} f(z|x) d^q\mathbf{z} \end{aligned} \tag{2}$$

The above equation can be interpreted as the risk that the self-reported variables are misstated with respect to a certain pricing policy μ, given knowledge of the correctly measured information.

The indicator function $\mathbb{1}_{\mu(z^*,x)<\mu(z,x)}$ is straightforward to calculate given a commercial pricing policy μ. Note that a broader function $f(\mu(z^*, x), \mu(z, x))$ can also be considered, for instance, to give more weight to higher differences in premium. The conditional density function $f(z|x)$ is unknown and needs to be estimated.

3.4. Conditional density estimation

[9] proposes a flexible conditional density estimator (FlexCode) based on an orthogonal projection of Z such that the CDE problem is transformed into a series of I conditional expectation estimation

problems. The estimator is written as follows:

$$\hat{f}(z|x) = \sum_{i}^{I} \hat{\mathbb{E}}[\phi_i(z)|X]\phi_i(z), \tag{3}$$

where $\phi_i(z)$ is the projection of variable z onto the ith dimension of an orthogonal basis $\{\phi_i\}_I$ of size I. The conditional expectation $\hat{\mathbb{E}}[\phi_i(z)|X]$ can be estimated using any type of regression algorithm (e.g., a decision tree, linear model, or neural network). The choice of the optimal basis is evaluated based on the CDE loss.

The CDE loss [9] is the integrated square error between the true distribution $f(z|x)$ and the approximated distribution $\hat{f}(z|x)$, weighted by the marginal density of x.

$$\begin{aligned} L^{CDE}(f(z|x), \hat{f}(z|x)) &= \int\int (f(z|x) - \hat{f}(z|x))^2 dz dP(x) \\ &= C_f + \int\int \hat{f}(z|x)^2 dz dP(x) - 2\int\int \hat{f}(z|x) f(z|x) dz dP(x) \\ &= C_f + \int\int \hat{f}(z|x)^2 dz dP(x) - 2\int\int \hat{f}(z|x) f(z,x) dz dx \\ &= C_f + E_X\left[\int \hat{f}(z|x)^2 dz\right] - 2E_{X,Y}[\hat{f}(z|x)]. \end{aligned} \tag{4}$$

In practical real-world applications, unlike in simulation studies, the distribution $f(z|x)$ is unknown, which makes the estimation of the conditional density difficult to validate. The above formulation of the CDE loss is convenient because it isolates the $f(z|x)$ term in a constant that does not depend on the estimator $\hat{f}(z|x)$. The estimator of the above CDE loss in [9] is:

$$\hat{L}^{CDE}(f(z|x), \hat{f}(z|x)) = \frac{1}{m}\sum_{i=1}^{m}\int \hat{f}(z|x_i)^2 dz - \frac{2}{m}\sum_{i=1}^{m}\hat{f}(z_i|x_i), \tag{5}$$

which is evaluated on m observations held out for the purpose of estimating $\hat{f}(z|x)$, i.e., the validation set.

This approach conveniently relies on regression models, which are well known by practitioners and are suited for massive parallelization (each of the I regression models can be calculated independently). However, the basis size is assumed to be unique for all $X = x$, and using the CDE loss alone as the criterion for selecting the optimal basis size can be misleading for some problems [18]. Note also that the approach can generate spurious bumps or negative values that require postprocessing, as explained in [9].

One popular type of regression model that is suitable for high-dimensional sparse datasets of mixed data types (continuous and discrete variables) is a random forest estimator. A random forest estimator [19] of a conditional mean can be expressed as follows:

$$\hat{\mathbb{E}}[Z|X] = T^{-1} \sum_{t=1}^{T} \frac{\sum_{i=1}^{n} Z_i \mathbb{1}_{X_i \in R(x,\theta_t)}}{\sum_{i=1}^{n} \mathbb{1}_{X_i \in R(x,\theta_t)}},$$

where T is the number of decision trees, θ_t denotes the structure of tree t, and $R(x, \theta_t)$ is the region (leaf) delimited by tree t for observation x. This estimator can be used in Equation (3) to estimate each of the I projections of z.

Although the orthogonal projection framework proposed in [9] is designed to study univariate responses z, extensions to multivariate responses are theoretically possible, as pointed out in the paper, via tensor products. In the bivariate case, where $z = \{z_1, z_2\}$ and for two bases $\{\phi_i\}$ and $\{\phi_j\}$:

$$\hat{f}(z|x) = \sum_{i,j} \hat{\mathbb{E}}[\phi_i(z_1)\phi_j(z_2)|X]\phi_i(z_1)\phi_j(z_2).$$

However, neither the paper nor the related code details a cross-validation strategy for finding multiple basis sizes in the case of multiple variables. The number of models to evaluate is large, becoming $I * J$ with only two variables, and the numerical estimation of the CDE loss becomes a challenge in itself as the number of dimensions increases, necessitating the computation of a multivariate integral.

The random forest for conditional density estimation in the RFCDE method [10] is defined as follows:

$$\hat{f}(z|x) = \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \frac{\mathbb{1}_{X_i \in R(x,\theta_t)}}{\sum_{i=1}^{n} \mathbb{1}_{X_i \in R(x,\theta_t)}}\right)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{n} K_H(Z_i - z) \frac{\mathbb{1}_{X_i \in R(x,\theta_t)}}{\sum_{i=1}^{n} \mathbb{1}_{X_i \in R(x,\theta_t)}}, \tag{6}$$

where K_H is a kernel function with bandwidth matrix H.

An approximation is also adopted in [10] to enable the use orthogonal series in order to avoid the calculation of kernel densities at each split, as in [9], the size and type of the basis are parameters to be defined.

However, the CDE loss used in [9; 10] has no meaningful interpretation. It alone is not sufficient to evaluate the estimation of $f(z|x)$. We therefore introduce a complimentary evaluation metric in the next section to provide further confidence in and insight into the estimation of $f(z|x)$.

3.5. High-density region for conditional density diagnosis and outlier detection

High-density regions (HDR) are sometimes used in the conditional density estimation literature to diagnose the calibration of $\hat{f}(z|x)$, e.g., in [9].

The high-density region [23] of a conditional distribution for a given confidence level α is $R(f_\alpha) = \{z : \hat{f}(z|x) > f_\alpha\}$, where f_α is the largest constant such that $P(Z \in R(f_\alpha)|X = x) = 1 - \alpha$.

The empirical coverage $\hat{\alpha}(\alpha)$ is the average proportion of samples of Z that belong to the high-density region for their respective $\hat{f}(z|x)$ and a theoretical coverage level α, $\hat{\alpha}(\alpha) = \frac{1}{n}\sum_i^n \mathbb{1}_{z_i \in R_i(f_\alpha)}$. If $\hat{f}(z|x)$ is a good estimator of $f(z|x)$, then the empirical coverage should converge to the theoretical coverage for all α levels.

Based on the above definitions, we can define the HDR coverage loss as follows:

$$L^{HDR}(f(z|x), \hat{f}(z|x)) = \int (\alpha - \hat{\alpha}(\alpha))^2 d\alpha,$$

which is ≈ 0 if $\hat{f}(z|x) \approx f(z|x)$. Note that the relation between the empirical and theoretical coverages is a necessary but insufficient condition to determine the calibration of $\hat{f}(z|x)$. For instance, when used as an estimate of $\hat{f}(z|x)$, the prior distribution $\hat{f}(z)$ can satisfy this coverage test and yet be a poor and uninformative estimator of the posterior distribution. It is, however, an informative loss that can be used in combination with the CDE loss during the training phase of the algorithm or as a criterion for selecting an optimal basis size in Equation (3). It is also suited to the case of multivariate variables Z, similar to the CDE loss.

One peculiar byproduct of the use of conditional density estimation in Equation (2) and high-density regions is the ability to identify outliers in the self-reported values of Z, conditional on X. Using the same calibrated conditional density estimator $\hat{f}(z|x)$, an outlier is defined as $Z \notin R(f_\alpha)$ for a given α level. Moreover, high-density regions can be mapped to a "traffic light" approach to facilitate communication with nonstatisticians by means of intuitive code values [7, p. 282]. A traffic light approach can also be related to a governance policy; for instance, all data points in the red area, i.e., points that do not belong to the 99 percent high-density region, may require a second level of validation.

3.6. Pricing policy form μ and stagewise validation of information

The use of generalized linear models (GLMs) is a common standard in the insurance industry for calculating a pricing policy μ. We refer to [24] for a comprehensive overview. In this case, the pricing model can be expressed in a multiplicative form as $\mu(z, x) = \mu_0\mu_z(z)\mu_x(x)$, where μ_0 is a "baseline premium" and $\mu_z(z)$ and $\mu_x(x)$ are relative price increases/decreases in the values of the variables Z and X.

In this case, the premium fraud risk simplifies to

$$\int_{\mathcal{Z}} \mathbb{1}_{\mu_z(z^*)<\mu_z(z)} f(z|x) d^q\mathbf{z}. \tag{7}$$

This is convenient for the evaluation of the premium fraud risk in practice, as one requires only the multiplicative factor μ_z instead of the complete pricing model μ to evaluate the fraud risk.

In practice, we do not have a priori validated values for the self-reported values on an incoming contract. Instead, we estimate the likelihood $f(z|x)$ of each possible true value of the self-reported information in the region $\{z : \mu(z) > \mu(z^*)\}$. This region represents all values of z for which the premium price is greater than the price $\mu(z^*)$ and, therefore, the values for which there is a motivation to commit data misrepresentation.

The validation problem may be approached in a stagewise manner such that the premium fraud risk can be refined as the self-reported variables are validated. A conditional density model can provide guidance to an underwriter as he/she iteratively validates the self-reported values $Z_1^*, Z_2^*, ..., Z_q^*$.

For instance, in the case of motor insurance, one could first investigate the validity of a driver's identity as variable Z_1 and, as a second step, validate the claims history relating to that driver as a separate variable Z_2.

One advantage of this approach is that one can use the already-validated self-reported values to refine the subsequent prediction and analysis of other self-reported values, $\hat{f}(z_2|z_1, x)$.

$$P(\mu(Z_1^*, Z_2^*, X) < \mu(Z_1, Z_2, X)|X = x, Z_1^* = z_1^*, Z_2^* = z_2^*).$$

If the pricing policy is multiplicative, i.e., $\mu(Z_1, Z_2, X) = \mu_0\mu_{z_1}(Z_1)\mu_{z_2}(Z_2)\mu_x(X)$, then the conditional premium fraud risk is:

$$P(\mu_{z_1}(Z_1^*)\mu_{z_2}(Z_2^*) < \mu_{z_1}(Z_1)\mu_{z_2}(Z_2)|X = x, Z_1^* = z_1^*, Z_2^* = z_2^*)..$$

Once the value of Z_1^* has been validated, the premium fraud risk is updated to:

$$P(\mu_{z_2}(Z_2^*) < \mu_{z_2}(Z_2)|X = x, Z_1 = z_1, Z_2^* = z_2^*).$$

Note that this stagewise approach provides insight into the vulnerabilities of an underwriting system and, therefore, its exposure to premium fraud at different stages in the validation process. For instance, the variable Z_1 may be "more difficult" to predict (and Z_1^* may be more difficult to validate) than Z_2 (Z_2^*), and thus, the company may exert additional effort in the assessment of Z_1.

4. Empirical results

4.1. Data

Although there can be, in practice, a handful of self-reported variables in motor insurance contracts, two self-reported dimensions in particular are critically important in the determination of the insurance price: the identity of the driver (particularly his/her driving experience and age) and the driving record (represented on an ordinal scale in a bonus-malus insurance system). A bonus-malus insurance system is a merit-based system that reflects the past driving history on a single ordinal scale; for more details, we refer to [25].

The dataset that is used to evaluate the proposed approach contains a sample of 57 000 validated contract quotes (X_i, Z_i) spanning 4 years of history, where X_i represents the detailed information on the vehicle as certified by a third party and Z_i contains the self-reported information. As is usually assumed, the i pairs (X_i, Z_i) are assumed to be i.i.d. in nature. Note that validated contract quotes relate here only to new customers, contract renewals are therefore not included in the dataset.

A history of four years is selected here, as this corresponds to one generation of vehicles in Belgium. The correctly measured data X are composed of 38 variables describing vehicles' features (e.g., power, age of the car, height, brand, model, segment of car, number of doors, catalog value).

In the following, we present an application of the proposed methodology to motor insurance contracts where the self-reported variable Z is:

1. univariate and continuous: the driver's license age;
2. bivariate and continuous: the driver's license age and the driver age;
3. multivariate of mixed type: the driving record score and the driver's license age; and
4. univariate and ordinal: the driving record score.

The vehicle information is provided by a third party and deemed correctly measured.

4.2. Application

4.2.1. Univariate and continuous self-reported data: the driver's license age

The marginal distribution $f(z)$ of driver's license age provides baseline information on the expected driver's license age distribution. The estimator $\hat{f}(z)$ is depicted in Figure 1, where a traffic light approach is used to color four different high-density regions [23] according to three arbitrary levels α.

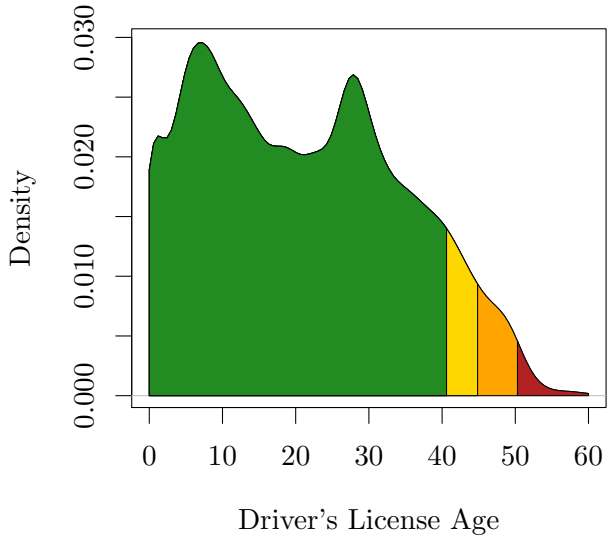

Figure 1: Estimator $\hat{f}(z)$ of the marginal distribution of driver's license age, estimated with a Gaussian kernel of bandwidth 1.3. The colored areas represent the high-density regions at levels of 90, 95 and 99 percent.

However, the marginal distribution $f(z)$ is not particularly helpful in validating incoming contracts because the estimated high-density regions are relatively spread out. One way of narrowing the high-density regions is to add the vehicle information X to estimate the conditional density $f(z|x)$ for a specific vehicle. If the vehicle is a "good predictor" of the driver's identity (here, the driver's license age), we expect $f(z|x)$ to define narrower high-density regions. The insight gained in this way is measured by the difference between the posterior distribution and the prior distribution. A common measure for comparing two distributions is the Kullback-Leibler divergence measure:

$$KL(\hat{f}(z|x), \hat{f}(z)) = \int \hat{f}(z|x) log(\frac{\hat{f}(z|x)}{\hat{f}(z)}) dz. \tag{8}$$

Measuring the added value obtained by adding the information X for predicting Z is strategically important for considering the acquisition of new correctly measured data to add to the system.

The orthogonal-based estimator of Equation (3) requires three main types of parameters to be set: the basis type ϕ, the basis size I and the parameters of the underlying I regression estimators.

Basis type: We consider the Fourier basis as used in [9], as we expect a relatively smooth function $\hat{f}(z|x)$. The Haar wavelet basis is a suitable candidate for discrete responses. For an overview of alternative suitable basis type choices, we refer to [26].

Basis size: The size of the basis is subject to a typical bias-variance trade-off, with a smaller basis size resulting in a smoother $f(z|x)$, while higher basis values produce less smooth functions.

The size of the basis is evaluated via cross-validation on the basis of the CDE loss as described in [9]. Note that a large basis size is undesirable because as many regression models must be trained as the basis size is large, leading to prohibitive computational costs if no parallelization is implemented. Moreover, we use the HDR coverage loss as a complementary loss, next to the CDE loss to determine the basis size I, as presented in the methodology section.

Regression model: Any common regression model for estimating a conditional mean can be considered. Decision tree-based methods are known to have high off-the-shelf performance with mixed data types of high dimensionality with a large number of observations and also are suited to parallelization in most cases. The authors of [9] have proposed FlexCode Random Forest (FlexCode-RF) for mixed data types, which implements Breiman's random forest algorithm as a regression function [19]. In practice, it is preferable to opt for an off-the-shelf algorithm with good expected performance with minimal tuning, since as many models must be trained as the basis size. We use the fast random forest implementation in [27].

A 5-fold cross-validation strategy is used here to validate the parameters used in the base random forest estimator. We refer to [28; 29] for an overview of parameter choice, heuristics, and selection in random forests. Note that there is no theoretical guarantee that a set of parameters proven optimal for conditional mean estimation is also optimal for conditional density estimation. This step is taken as the starting point for further parameter optimization and is also useful for comparing the different methods.

In underwriting decision problems, a good understanding of the estimated conditional density function is critical to assess (i) the soundness of the estimators, (ii) the insight gain achieved by adding X when predicting Z, and (iii) its explainability. Indeed, in this case, a single metric such as the CDE loss is insufficient to assess any of these criteria.

In Figure 2, we present the results of three models trained on a test set of $n/3$, i.e., 19k observations. We present the vanilla random forest algorithm used as the basis for the parameter choice as a point of comparison for a density estimator centered on its conditional mean estimate, albeit not an estimator of $\hat{f}(z|x)$ in itself. FlexCode-RF with a Fourier basis of size $I = 5$ (minimizing the CDE loss) is shown in the middle, and RFCDE is presented on the right.

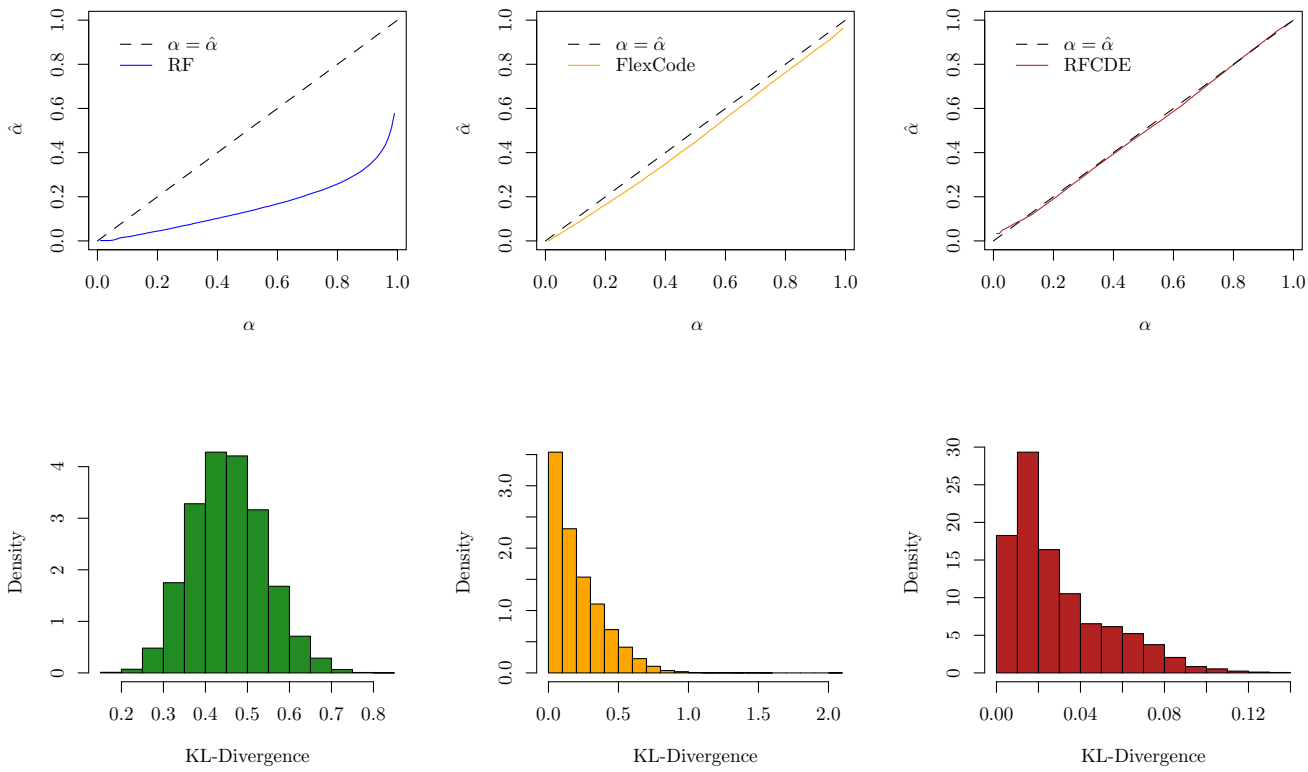

Figure 2: Theoretical versus empirical coverage plots and Kullback-Leibler divergence measures for the vanilla random forest algorithm (left), FlexCode-RF (middle) and RFCDE (right).

In Figure 2, one can observe that FlexCode and RFCDE are both close to the diagonal on the $\hat{\alpha} - \alpha$ coverage plot. However, FlexCode tends to diverge more from the prior distribution than RFCDE, as measured by the Kullback-Leibler divergence.

This can be explained by the use of basis transformation in FlexCode versus kernel density estimation in estimating the prior distribution. By contrast, in RFCDE, both the prior and posterior estimators use kernels.

Sample predictions from RFCDE compared to the prior distribution estimates for six contracts are presented in Figure 3.

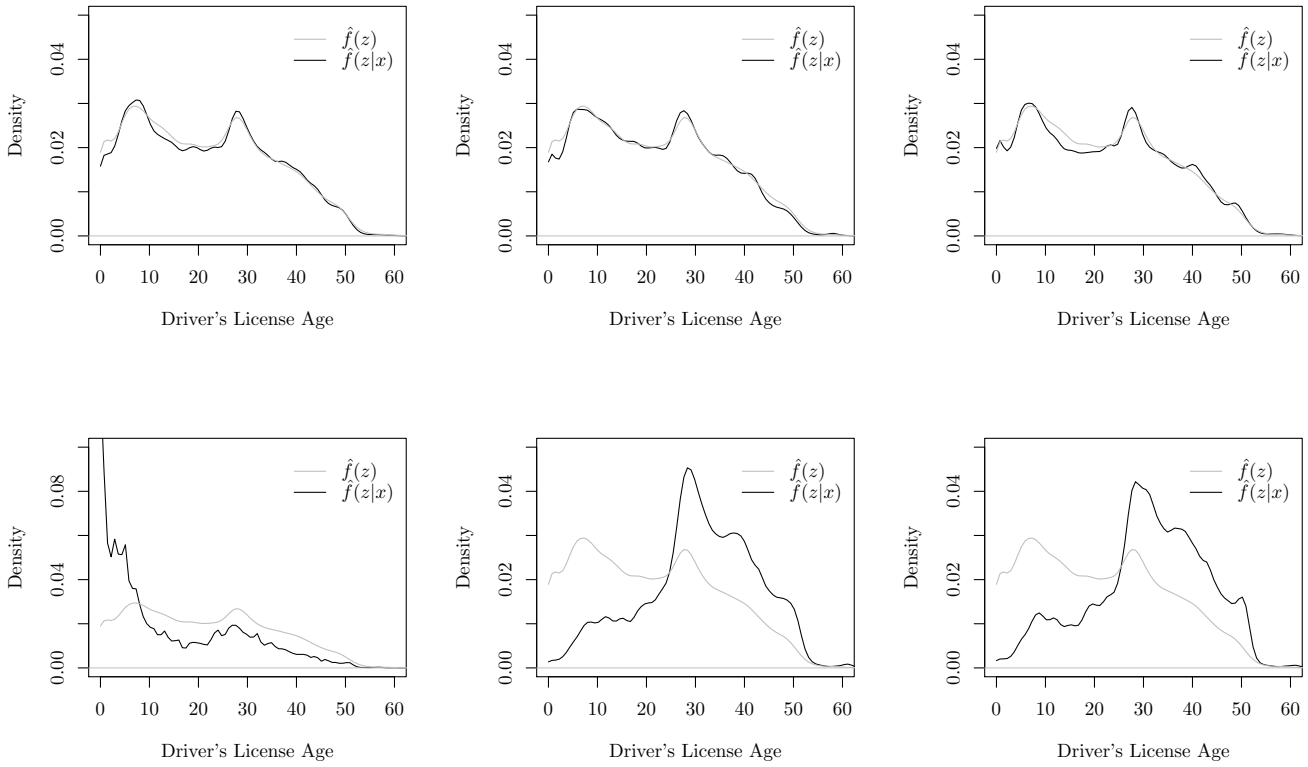

Figure 3: Sample predictions with the RFCDE estimate $\hat{f}(z|x)$ and the prior density $\hat{f}(z)$ in gray.

In Figure 3, the top and bottom rows of plots represent lower and higher Kullback-Leibler divergence measures, respectively, compared to the prior distribution $\hat{f}(z)$. Lower values of the Kullback-Leibler divergence indicate that the information X does not add much evidence on the true value of the self-reported variable. The distributions presented in the bottom row, however, indicate that the driver of the car is likely less experienced.

4.2.2. Bivariate and continuous self-reported data: the driver's license age and driver age

In the case that the self-reported variable Z is continuous and bivariate, the methods discussed above are extensible to multivariate prediction via the use of tensor products for orthogonal basis transformation, similarly to FlexCode [9], and with multivariate kernels for RFCDE [10].

Here, we briefly present a simple example with RFCDE, estimated via a kernel density method.

The CDE loss formulations of Equation (5) and Equation (6) are both applicable in multivariate cases. However, the bandwidth H is a 2-by-2 matrix instead of a single scalar as in the univariate

case. Duong [30] explains that the bandwidth matrix is usually parameterized as a diagonal matrix (constrained) or as a semi-positive definite and symmetric matrix (unconstrained).

For instance, the conditional distribution estimator of the prior distribution $\hat{f}(z)$ (where z is bivariate) is $\begin{pmatrix} 0.19 & 0 \\ 0 & 0.19 \end{pmatrix}$ in the constrained case and $\begin{pmatrix} 0.34 & 0.21 \\ 0.21 & 0.31 \end{pmatrix}$ in the unconstrained case, and the resulting densities $\hat{f}(z)$ with a Gaussian kernel are displayed in Figure 4 (with prescaling of the data and the Sum of Asymptotic Mean Squared Error pilot bandwidth selector; see [30] for a detailed explanation).

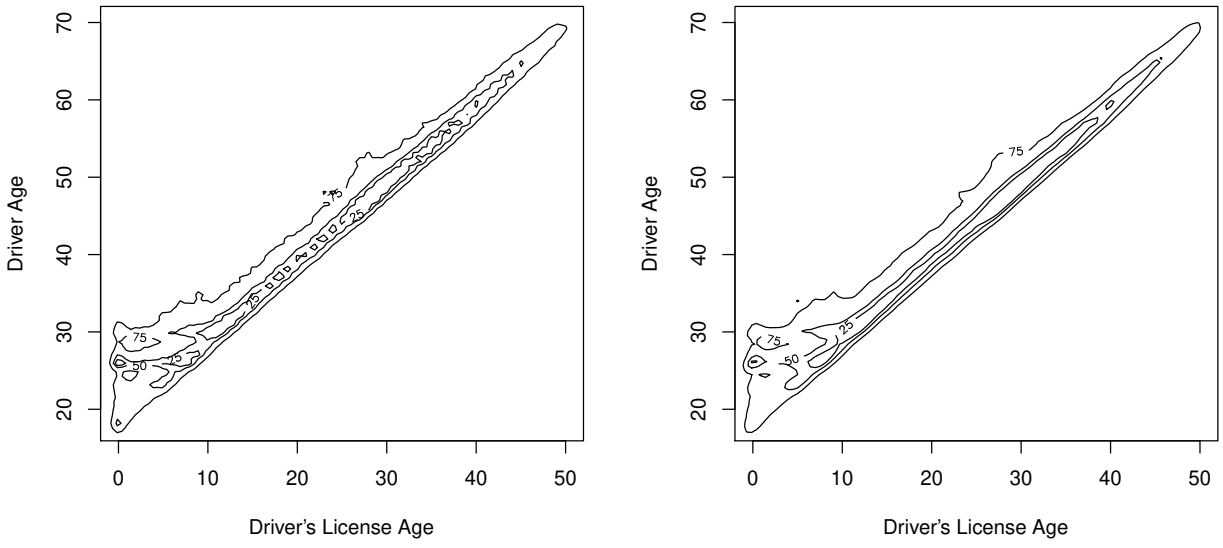

Figure 4: Kernel density estimation of the prior distribution $\hat{f}(z)$ with a constrained bandwidth matrix (left) and an unconstrained (right) density estimation of the prior.

There are many possible strategies for the parameterization of H, most of which are based on a numerical procedure (a "plug-in" estimator) or a cross-validation methodology; we refer to [30] for an overview. However, those methods estimate the matrix H for each density function $\hat{f}(z|x)$ and can exhibit numerical instabilities, which is undesirable, particularly when the model is intended to provide guidance to nonstatisticians.

A simple approach to guarantee the stability of the bandwidth estimation is to estimate the parameter of the prior density function $\hat{f}(z)$ and fix it for any conditional density function in Equation (6).

Figure 5 presents one sample prediction.

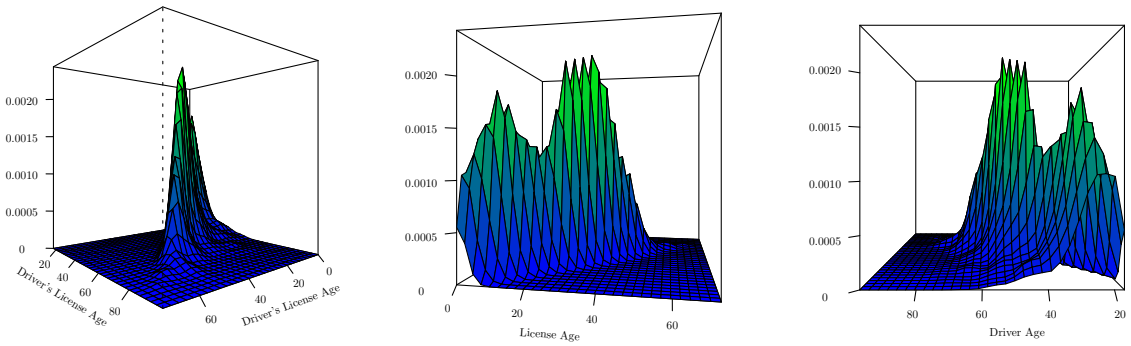

Figure 5: Conditional density function $\hat{f}(z|x)$ for a bivariate Z (driver age and license age). The bandwidth matrix H is the constrained diagonal matrix of Figure 4.

4.2.3. Bivariate mixed self-reported data: driver's license age and driving score

In the case that the self-reported variable Z is multivariate and of mixed type (discrete and continuous), computing the conditional premium fraud risk requires the evaluation of a joint conditional density $\hat{f}(z_1, z_2|x)$.

RFCDE is designed for one type of variable, and FlexCode is univariate. Here, we adapt the method of FlexCode to mixed-type multivariate responses using heterogeneous basis types, similar to the vector-valued extension sketched in [9], via tensor products. $\{\phi_i\}_I$ is a Fourier basis of size I, and $\{\phi_j\}_J$ is a Haar basis of size J. In this case, we set the size of the basis $\{\phi_j\}_J$ to the cardinality of the discrete variable. Figure 6 and Figure 7 present the CDE loss and HDR coverage loss for different values of the basis size I. In this case, we observe that both losses are minimized by a basis of size 3. Note that in this application, the postprocessing of the estimated densities is limited to removing negative values and uniform scaling of the density surface to ensure that it integrates to 1.

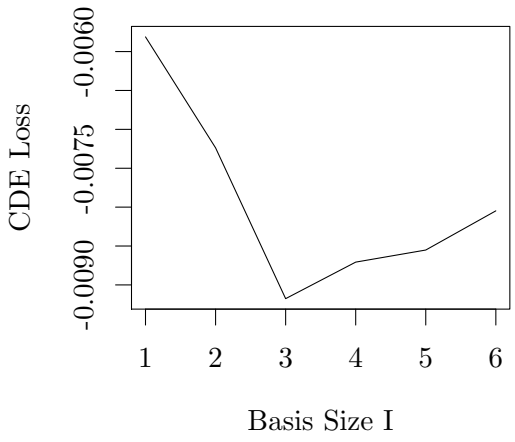

Figure 6: CDE loss on the test set.

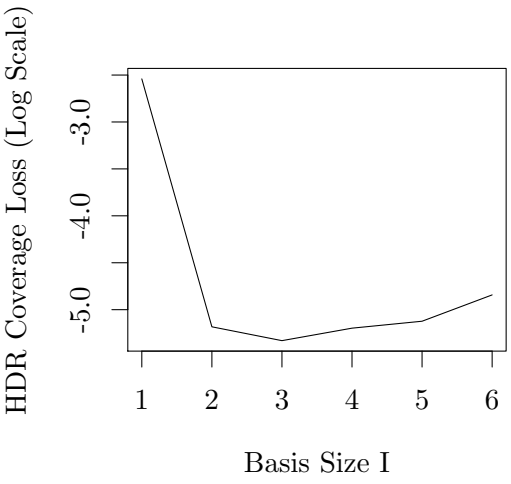

Figure 7: HDR coverage loss on the test set.

One sample prediction is depicted in Figure 8. We observe a high-probability mass at a low driving score, which corresponds to drivers with a good driving score. A second high-probability region is observed at a mid-range driving score, corresponding to the driving score of a new driver.

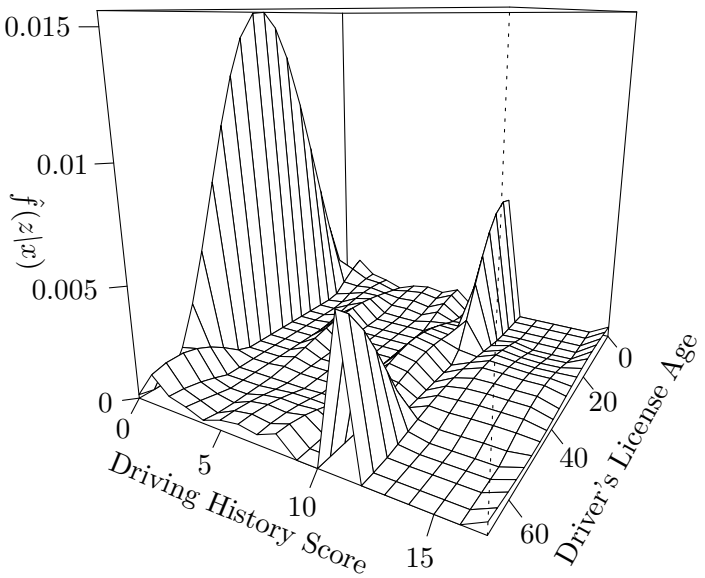

Figure 8: Sample prediction for bivariate self-reported data of mixed type.

4.3. Fraud score application

In this section, we briefly present an application of the fraud score in a simplified, univariate setting with the driver's license age as the self-reported variable and a pricing function which is

strictly decreasing with driver's license age (see Figure 13 in appendix). We apply the premium fraud risk model to a sample of new contract propositions, with a threshold of 5% increase in premium price to exclude small premium price differences.

The dataset that is used to evaluate the proposed approach contains a sample of 73 000 applications spanning 4 years of history. The applications contain data which resulted in a validation and were accepted by the insurance company, as well as applications which were rejected. Note that the sample used here contains applications which were validated and therefore used in the training of the conditional density estimation $\hat{f}(z|x)$.

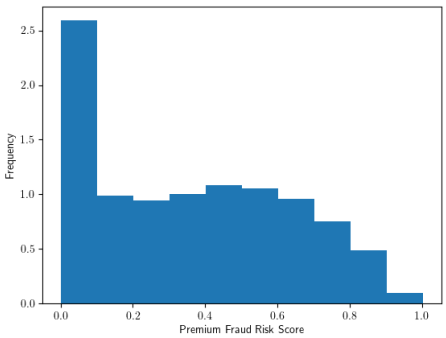

Figure 9: Histogram of premium fraud risk estimated on a sample new contract propositions.

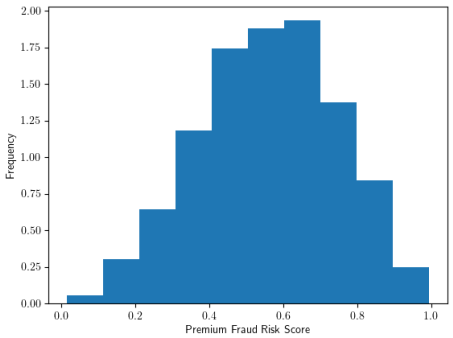

Figure 10: Histogram of premium fraud risk estimated on a "risky" sub-sample of new contract propositions.

Figure 9 presents the distribution of the fraud risk score on a sample of contracts to validate. A sub-segment of 2500 applications from this sample has been labeled by experts as presenting a larger premium fraud risk, although these applications have not confirmed to effectively involve fraud. This sub-segment contains old cars with a self-reported middle-aged main driver, who has already a contract for another vehicle. Therefore, the segment is classified as potentially containing *hidden young main drivers*. Figure 10 presents the fraud risk scores for this sub-segment. The difference in distributions suggests that our model does well in classifying those suspicious applications.

Additionally, in Figure 11, we present the CDEs of three randomly sampled applications from sub-segment of Figure 10. The self-reported value z^* for those applications is presented in red. The area filled corresponds to the region on the CDE where premium price would be 5% higher than with the self-reported value, and the surface represents the premium fraud risk. We provide in appendix an infographic summarizing the decision support systems' steps and discuss computation aspects.

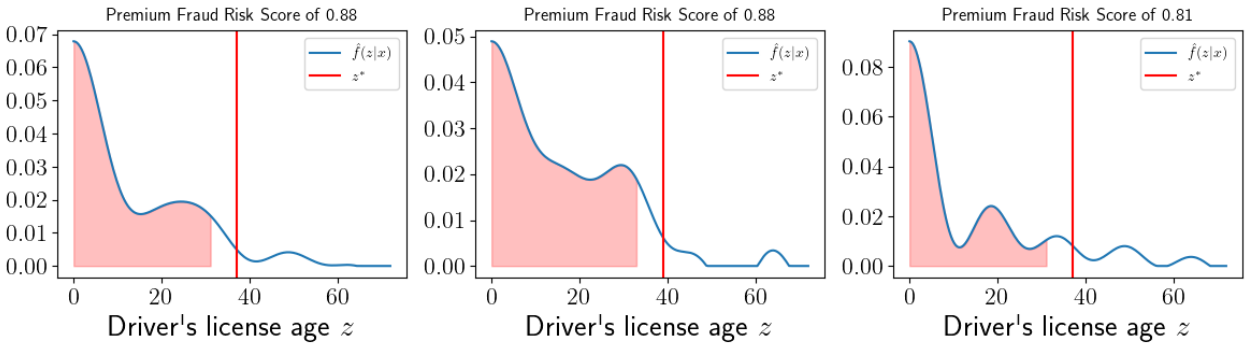

Figure 11: Sample premium fraud risk scores prediction and self-reported values of driver's license age.

4.4. Note on the explainability of $\hat{f}(z|x)$

The explainability of a model is of critical concern towards its adoption and is a challenging task, even for common regression models estimating the conditional mean $\mathbb{E}[Z|X]$.

In that regard, Shapley values from game theory is a popular approach for explaining black-box models [31]. Shapley values are attributed at the prediction level; typically, each feature in vector x is attributed a contribution Φ (a Shapley value) to the difference between $\hat{\mathbb{E}}[Z]$ and $\hat{\mathbb{E}}[Z|X]$ for a specific prediction.

In this section, we investigate the explainability of FlexCode and its input feature X. In particular we present an additive explanation to a CDE model of the form: $f(z|x) = \Phi_0^{f(z|x)}(z) + \sum_j \Phi_j^{f(z|x)}(z, x)$. We present in appendix the details and proofs on the methods to conveniently reuse the Shapley values from the regression models underlying FlexCode in the context of a CDE problem.

Figure 12 presents a sample prediction of $\hat{f}(z|x)$, where we are interested in explaining the difference between the prior distribution $\Phi_0(z)$ and the posterior distribution $\hat{f}(z|x)$ (CDE raw). This difference is explained by the contribution of each variable j, $\Phi_j^{f(z|x)}(x)$.

In this figure we observe the model predicts a higher density on less experienced driver compared to the prior distribution. One can learn for instance that for the lowest value of driver's license age (at $z = 0$), the power, height, length and car age have a positive contribution on the $\hat{f}(z|x)$ for observation x.

Likewise, the bump on the right side of Figure 12 may appear curious, we observe that the age of the car feature explains in part this prediction for this observation.

The displayed sum of the values $\sum_j \Phi_j^{f(z|x)}(z, x)$ with $j = 0, ..., p$ is equal to $\hat{f}(z|x)$ for all values of z, as expected by the additivity property of Shapley values.

The outcomes of model explanation also serves as justification for a model user to further inves-

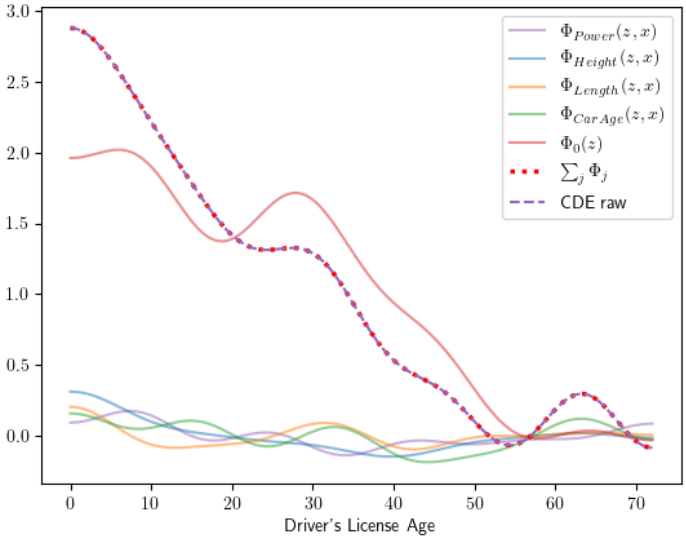

Figure 12: Variable contribution for a sample prediction, explaining the difference between the prior distribution $\Phi_0(z)$ and the estimator of a sample posterior distribution $\hat{f}(z|x)$ (CDE raw). Four sampled variables' Shapley values $\Phi_j^{f(z|x)}(z,x)$ are presented.

tigate a given element of a contract or justify the refusal of a contract proposition.

Moreover, model explanation is helpful from a modeling perspective in applying the Occam's razor principle to the input feature space X, for instance, to discard "dummy" variables that do not play any role in prediction. It can also be used in the assessment of adding new dimension(s) to the input variables X, e.g., sourcing additional variables on the vehicle.

4.5. Discussion

We have seen that conditional density estimation can be used as a means to estimate the plausibility of self-reported information in multiple formats (multivariate, mixed data types). The validation of a CDE system is inherently difficult because we do not observe the true conditional densities, only sample points. Criteria such as the CDE loss and high-density regions may be used to validate vastly different approaches, for instance, orthonormal projections (FlexCode) and weighted kernel density estimators (RFCDE), as two valid candidates, as we have seen in Figure 2.

The choice of the CDE technique depends on the nature of the self-reported data. In this paper, we apply the method to multivariate self-reported data of mixed type (discrete and continuous). Other applications may benefit from simpler approaches to CDE when the self-reported data are univariate or of nonmixed type or if a parametric form (e.g., a Gaussian distribution) can be reasonably assumed.

The efficiency of our method in fighting fraud depends on the predictability of the self-reported values based on the correctly measured values. In the application to motor insurance data presented here, the task can be summarized as guessing drivers' characteristics based on their vehicles' characteristics.

In this application, we have limited the correctly measured variables to vehicle information. Other correctly measured information could also be added to the system to enable better prediction of the self-reported values to further improve the effectiveness of the system.

5. Conclusions

In this paper, we have formulated *premium fraud detection* as a conditional density estimation problem. The estimation of conditional functions relies on the availability of a dataset of true values (X_i, Z_i) that has been previously validated by experts.

This formulation enables us to explicitly estimate the distribution of the self-reported variables and their relation to the pricing policy, thereby yielding a risk evaluation in the opportunity-motivation framework of [8] in the context of premium fraud.

Conditional density estimation can also be used for performing (conditional) anomaly detection to detect outliers across specific dimensions Z using high-density regions and a traffic light approach [7].

We have discussed state-of-the-art techniques for estimating the conditional densities of continuous variables in large datasets of mixed data types [9; 10], which has remained a challenging statistical problem to date. Notably, the use of a single criterion such as the CDE loss [9] appears insufficient to evaluate the overall calibration of the conditional densities. We have proposed an alternative loss function based on the high-density regions of [23] as a complementary loss function in CDE problems. We have presented how Shapley additive explanation [31; 32] can be adapted to CDE problems to make the estimated conditional densities explainable.

Moreover, we have discussed how the validation of a contract proposition can be approached in a stagewise manner. The understanding of the predictability of a given stage in the validation of a contract can provide insight into the weaknesses of the underwriting process and allow actions to be taken accordingly to improve the robustness of contract validation. For instance, a higher uncertainty around a driver's identity based on the car compared to a lower uncertainty around the prediction of a driving record based on the driver's identity and the car can motivate the acquisition of additional data predictive of the driver's identity or a change in the variable used for pricing.

Further developments in nonparametric conditional estimation methods could also be bench-

marked against the techniques and evaluation metrics used herein. Future research could investigate the use of the realized loss per contract as an additional source of information supplementary to our premium fraud risk score, thereby creating connections between the approaches of [13] and [14] and our own.

Our approach can be used by insurance companies to improve their decision making process in the validation of insurance contracts with potentially misrepresented self-reported information. For instance, in automated insurance underwriting ("*straight-through processing*"), our premium fraud risk score could be used to trigger a validation of the self-reported variable.

Our approach can also be applied for fraud detection in other, noninsurance domains in the application stage (*application fraud*), such as overstated income on credit applications to benefit from unduly low rates [33; 34; 35] or accounting misconduct [36].

Acknowledgment

We are grateful for the contribution of the Allianz Chair on Prescriptive Business Analytics in Insurance at KULeuven. Furthermore, we appreciate the opportunity that Allianz Benelux provides us to dedicate time to research activities.

6. Appendix

6.1. Pricing calculation and premium fraud examples in multiplicative pricing models

We present here two examples based on dummy data to illustrate (a) a premium calculation and (b) premium fraud when the pricing model has a multiplicative form. The pricing policy has the form $\mu(z,x) = \mu_0\mu_z(z)\mu_x(x)$, where μ_0 is a "baseline premium" and $\mu_z(z)$ and $\mu_x(x)$ are relative increases/decreases in the values of variables Z and X.

(a) *Premium calculation example*: Suppose that the baseline premium μ_0 is set to 100 monthly, a car considered safe translates into the pricing policy as $\mu_x(x) = 0.9$, and an inexperienced driver translates into the pricing policy as $\mu_z(z) = 1.2$. Then, the proposed price of the corresponding insurance policy is $100 * 0.9 * 1.2 = 108$ monthly.

(b) *Premium fraud example*: In Figures 13 and 14, one can read out, for instance, that the premium price $\mu(z^*, x)$ for a driver reported as having z_1^*= 20 years of driving experience and being z_2^*= 45 years old pays a premium of $\mu = \mu_0\mu_{z_1}(z_1^*)\mu_{z_2}(z_2^*)\mu_x(x)$, where $\mu_{z_1}(20) = 1.05$ and $\mu_{z_2}(45) = 0.99$. Then, if the "true" driver, who is the child of the reported driver and in fact has one year of driving experience ($\mu_{z_1}(1) = 1.30$) and is 25 years old ($\mu_{z_2}(25) = 0.95$), the premium that the true

driver should pay would increase by $\frac{\mu_{z_1}(z_1)\mu_{z_2}(z_2)}{\mu_{z_1}(z_1^*)\mu_{z_2}(z_2^*)}$= 18.8 percent over the premium currently being paid.

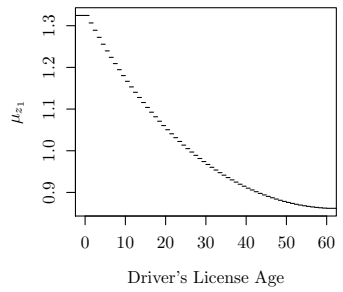

Figure 13: μ_{z_1}: Relative increase in premium price according to the driver's license age variable.

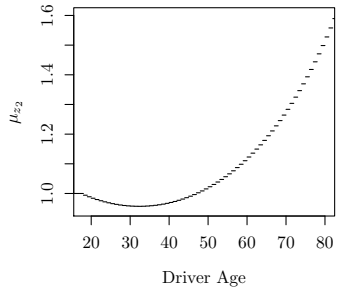

Figure 14: μ_{z_2}: Relative increase in premium price according to the driver age variable.

6.2. Detailed note on the explainability of $\hat{f}(z|x)$

Borrowing the notation from [32], the Shapley $\Phi_j^\delta(x)$ of a model δ for feature j and observation x equals:

$$\Phi_j^\delta(x) = \sum_{Q \subseteq S \setminus \{j\}} \mathcal{C}(\delta_{Q\cup\{j\}}(x) - \delta_Q(x)),$$

where Q is a subset (coalition) feature of $S = \{1, ..., p\}$, p is the dimension of the input feature X, j is the jth feature of X, $\mathcal{C} := \frac{|Q|!(|S|-|Q|-1)!}{|S|!}$, $(\delta_{Q\cup\{j\}}(x) - \delta_Q(x))$ measures the contribution of adding variable j to the coalition Q for a model δ and $\Phi_0^\delta = \delta_\emptyset(\emptyset)$. The model δ can be expressed as the sum of its Shapley values, $\delta(x) = \Phi_0^\delta + \sum_j \Phi_j^\delta(x)$ as per the additivity property of Shapley values.

Proposition 1: If $f(z|x)$ admits an orthonormal decomposition of the form $f(z|x) = \sum_i \phi_i(z)\mathbb{E}[\phi_i(Z)|X]$, $\Phi_j^{\mathbb{E}[\phi_i(Z)|X]}(x)$ is the Shapley value of the jth feature of the model $\mathbb{E}[\phi_i(Z)|X]$, and $\Phi_j^{f(z|x)}(x)$ the Shapley value of the jth feature of model $f(z|x)$ for a value of z, then $\Phi_j^{f(z|x)}(z, x) = \sum_i \phi_i(z)\Phi_j^{\phi_i(Z)}(x)$

proof:

$$\begin{aligned}\Phi_j^{f(z|x)}(z,x) &= \sum_{Q\subseteq S\setminus\{j\}} \mathcal{C}(f(z|x_{Q\cup\{j\}}) - f(z|x_Q)) \\ &= \sum_{Q\subseteq S\setminus\{j\}} \mathcal{C}(\sum_i \phi_i(z)\mathbb{E}[\phi_i(Z)|X_{Q\cup\{j\}}] - \sum_i \phi_i(z)\mathbb{E}[\phi_i(Z)|X_Q]) \\ &= \sum_i \phi_i(z) \sum_{Q\subseteq S\setminus\{j\}} \mathcal{C}(\mathbb{E}[\phi_i(Z)|X_{Q\cup\{j\}}] - \mathbb{E}[\phi_i(Z)|X_Q]) \\ &= \sum_i \phi_i(z)\Phi_j^{\mathbb{E}[\phi_i(Z)|X]}(x).\end{aligned}$$

This result is convenient as it enables to use the explanations of the underlying regression coefficients of models $\mathbb{E}[\phi_i(Z)|X]$.

Similarly to the above proposition, we can also relate the Shapley values $\Phi_j^{\mathbb{E}[Z|X]}(x)$ of the conditional expectation $\mathbb{E}[Z|X]$ to the Shapley values of the regression coefficients $\Phi_j^{\mathbb{E}[\phi_i(Z)|X]}(x)$, thanks to the relation $\mathbb{E}[Z|X] = \int zf(z|x)dz$. The result and proof are provided in appendix.

The conditional expectation $\mathbb{E}[Z|X]$ is related to the conditional density by $\mathbb{E}[Z|X] = \int zf(z|x)dz$.

Proposition 2: If $f(z|x)$ admits an orthonormal decomposition of the form $f(z|x) = \sum_i \phi_i(z)\mathbb{E}[\phi_i(Z)|X]$, $\Phi_j^{\mathbb{E}[Z|X]}(x)$ is the Shapley value of the jth feature of the model $\mathbb{E}[Z|X]$, and $\Phi_j^{\mathbb{E}[\phi_i(Z)|X]}(x)$ is the Shapley value of the jth feature of the model $\mathbb{E}[\phi_i(Z)|X]$, then $\Phi_j^{\mathbb{E}[Z|X]}(x) = \int z \sum_i \phi_i(z)\Phi_j^{\mathbb{E}[\phi_i(Z)|X]}(x)dz$

proof:

$$\begin{aligned}\Phi_j^{\mathbb{E}[Z|X]}(x) &= \sum_{Q\subseteq S\setminus\{j\}} \mathcal{C}(\mathbb{E}[Z|X_{Q\cup\{j\}}] - \mathbb{E}[Z|X_Q]) \\ &= \sum_{Q\subseteq S\setminus\{j\}} \mathcal{C}(\int zf(z|x_{Q\cup\{j\}})dz - \int zf(z|x_Q)dz) \\ &= \sum_{Q\subseteq S\setminus\{j\}} \mathcal{C}(\int z\sum_i \phi_i(z)\mathbb{E}[\phi_i(Z)|X_{Q\cup\{j\}}]dz - \int z\sum_i \phi_i(z)\mathbb{E}[\phi_i(Z)|X_Q]dz) \\ &= \int z\sum_i \phi_i(z) \sum_{Q\subseteq S\setminus\{j\}} \mathcal{C}(\mathbb{E}[\phi_i(Z)|X_{Q\cup\{j\}}] - \mathbb{E}[\phi_i(Z)|X_Q])dz \\ &= \int z\sum_i \phi_i(z)\Phi_j^{\mathbb{E}[\phi_i(Z)|X]}(x)dz.\end{aligned}$$

6.3. Note on the computational aspects of the decision support system

We add a few comments on the computational aspects of the decision support system using orthonormal bases with an infographic in the case of an univariate self-reported variable (illustrating section 4.3).

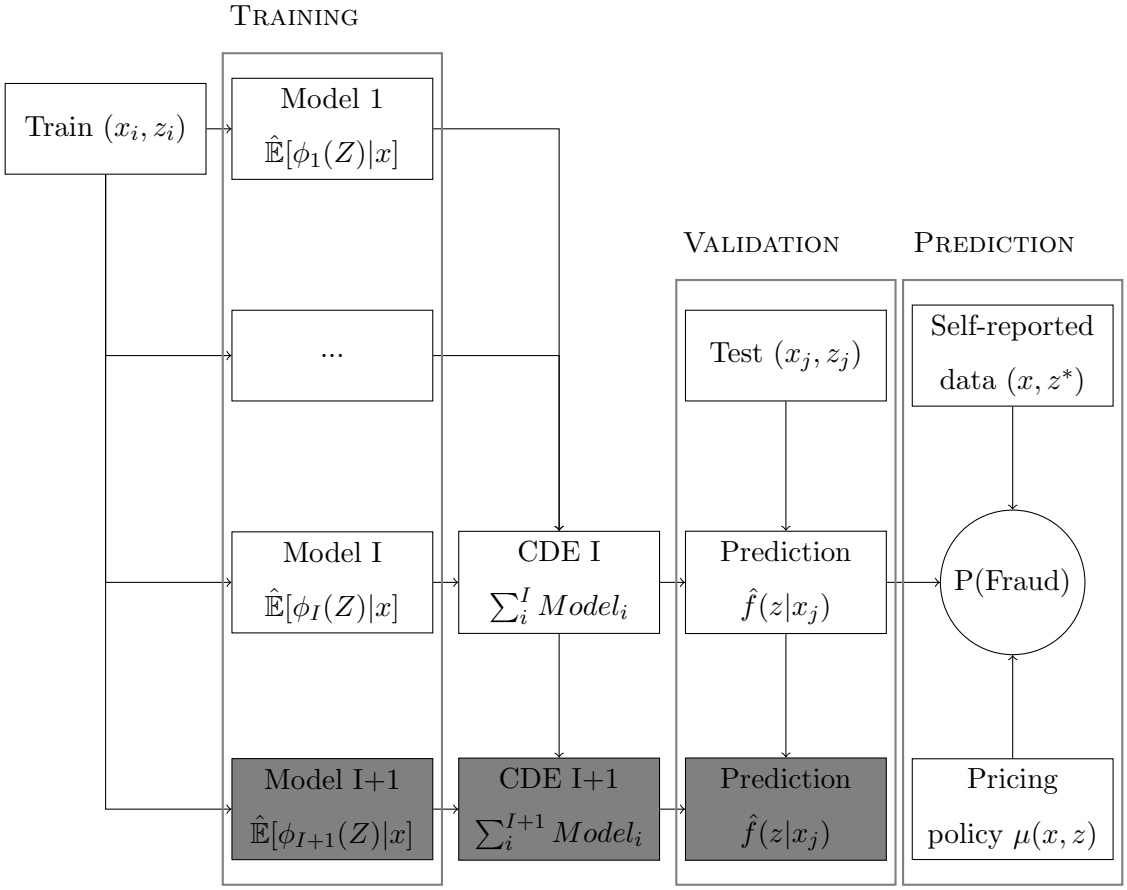

Figure 15: Overview of the decision support system for an univariate self-reported variable Z.

We distinguish three phases of the decision support system: training, validation and prediction. We assume the decision support system is intended to score new incoming contract propositions, hence the computational aspects of the prediction phase is more important than those attached to the training and validation phases.

The training and validation phases are proportional to the underlying regression model training and prediction times. In the above graph, the training phase and validation are implemented sequentially and stopped at model $I+1$ (where a decrease in the loss function is observed), hence the global computation time is roughly $I+1$ times the underlying model's training and prediction. This computation time can be reduced from $I+1$ to 1 times the underlying model's if the computational architecture allows to evaluate in parallel the different models. Similarly to the validation phase, the prediction phase requires a prediction of each underlying regression model, hence will be proportional to I times the prediction of the regression models (which can also be reduced with parallelization). The prediction step also requires the prediction from the pricing policy $\mu(x, z)$, for a grid of values of z. The pricing policy μ is typically a (generalized) linear model, whose prediction times are negligible.

In this application we used the fast implementation of random forest as underlying regression model. The algorithm is described in [27], where the performance are detailed and documented for different problems and dataset sizes.

Note that the choice of the basis type depends on the support and smoothness of variable Z, the self-reported of interest and can motivate variations around this theme.

References

[1] G. Schuman, Misrepresentation of smoking history in life insurance applications, Tort & Ins. LJ 30 (1994) 103.

[2] I. S. Office, The Challenge of Auto Insurance Premium Leakage, Tech. rep., Inc. Verisk Analytics (2017).

[3] R. A. Derrig, V. Zicko, Prosecuting insurance fraud—a case study of the massachusetts experience in the 1990s, Risk Management and Insurance Review 5 (2) (2002) 77–104.

[4] F. Carré, (in)dependent contractor misclassification, ECONOMIC POLICY INSTITUTE (2015).

[5] E. W. Ngai, Y. Hu, Y. H. Wong, Y. Chen, X. Sun, The application of data mining techniques

in financial fraud detection: A classification framework and an academic review of literature, Decision support systems 50 (3) (2011) 559–569.

[6] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM computing surveys (CSUR) 41 (3) (2009) 1–58.

[7] B. Baesens, V. Van Vlasselaer, W. Verbeke, Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection, John Wiley & Sons, 2015.

[8] S. Viaene, G. Dedene, Insurance fraud: issues and challenges, The Geneva Papers on Risk and Insurance-Issues and Practice 29 (2) (2004) 313–333.

[9] R. Izbicki, A. B. Lee, et al., Converting high-dimensional regression to high-dimensional conditional density estimation, Electronic Journal of Statistics 11 (2) (2017) 2800–2831.

[10] N. Dalmasso, T. Pospisil, A. B. Lee, R. Izbicki, P. E. Freeman, A. I. Malz, Conditional density estimation tools in python and r with applications to photometric redshifts and likelihood-free cosmological inference, Astronomy and Computing 30 (2020) 100362.

[11] S. M. Schennach, Recent advances in the measurement error literature, Annual Review of Economics 8 (2016) 341–377.

[12] Y. Hu, G. Ridder, Estimation of nonlinear models with mismeasured regressors using marginal information, Journal of Applied Econometrics 27 (3) (2012) 347–385.

[13] M. Xia, P. Gustafson, Bayesian regression models adjusting for unidirectional covariate misclassification, Canadian Journal of Statistics 44 (2) (2016) 198–218.

[14] R. M. Akakpo, M. Xia, A. M. Polansky, Frequentist inference in insurance ratemaking models adjusting for misrepresentation, ASTIN Bulletin: The Journal of the IAA 49 (1) (2019) 117–146.

[15] M. Xia, L. Hua, G. Vadnais, Embedded predictive analysis of misrepresentation risk in glm ratemaking models, Variance: Advancing the Science of Risk. in press.[Google Scholar] (2018).

[16] L. Devroye, L. Györfi, G. Lugosi, A probabilistic theory of pattern recognition, Vol. 31, Springer Science & Business Media, 2013.

[17] S. Devriendt, K. Antonio, T. Reynkens, R. Verbelen, Sparse regression with multi-type regularized feature modeling, Insurance: Mathematics and Economics 96 (2020) 248–261.

[18] M. Rosenblatt, Curve estimates, The Annals of Mathematical Statistics 42 (6) (1971) 1815–1842.

[19] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[20] N. Meinshausen, Quantile regression forests, Journal of Machine Learning Research 7 (Jun) (2006) 983–999.

[21] P. Hall, J. Racine, Q. Li, Cross-validation and the estimation of conditional probability densities, Journal of the American Statistical Association 99 (468) (2004) 1015–1026.

[22] R. S. Winsor, Misrepresentation and non disclosure on applications for insurance, Blaney McMurtry LLP (1995).

[23] R. J. Hyndman, Computing and graphing highest density regions, The American Statistician 50 (2) (1996) 120–126.

[24] P. De Jong, G. Z. Heller, et al., Generalized linear models for insurance data, Cambridge Books (2008).

[25] J. Lemaire, Bonus-malus systems in automobile insurance, Vol. 19, Springer science & business media, 2012.

[26] S. Mallat, A wavelet tour of signal processing, Elsevier, 1999.

[27] M. N. Wright, A. Ziegler, Eanger: A fast implementation of random forests for high dimensional data in c++ and r, Journal of Statistical Software 77 (1) (2017) 1–17.

[28] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2) (2016) 197–227.

[29] P. Probst, M. N. Wright, A.-L. Boulesteix, Hyperparameters and tuning strategies for random forest, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (3) (2019) e1301.

[30] T. Duong, et al., ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r, Journal of Statistical Software 21 (7) (2007) 1–16.

[31] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in neural information processing systems, 2017, pp. 4765–4774.

[32] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowledge and information systems 41 (3) (2014) 647–665.

[33] M. D. Pendley, G. Costello, M. Kelsch, The impact of poor underwriting practices and fraud in subprime rmbs performance, Fitch Ratings US Residential Mortgage Special Report (2007).

[34] A. Mian, A. Sufi, Fraudulent income overstatement on mortgage applications during the credit expansion of 2002 to 2005, The Review of Financial Studies 30 (6) (2017) 1832–1864.

[35] B. W. Ambrose, J. Conklin, J. Yoshida, Credit rationing, income exaggeration, and adverse selection in the mortgage market, The Journal of Finance 71 (6) (2016) 2637–2686.

[36] P. R. Hahn, J. S. Murray, I. Manolopoulou, A bayesian partial identification approach to inferring the prevalence of accounting misconduct, Journal of the American Statistical Association 111 (513) (2016) 14–26.