A New Cost-Sensitive Expert System for Auto-Insurance Fraud.

Raphaël Zerbato, Georges Dionne, François Bellavance

HEC Montréal

February 16, 2025

Abstract

Ex-post moral hazard leads to less than optimal risk allocations between economic agents. In the insurance context, fraud is defined as an agent misrepresenting his true loss. Economists have demonstrated that the role of an audit strategy is twofold: to detect fraudulent claims as early as possible and to deter potential future fraud. In recent years, insurance companies have invested heavily in integrating machine-learning (ML) algorithms into their existing fraud detection processes as advanced expert systems. While powerful, such supervised expert systems cannot serve as the sole base of an audit policy as they fail to estimate the overall fraud rate faced by an insurance company. The Optimal Audit Model (OAM) developed by Dionne et al. (2009) is a cost-sensitive expert system that accounts for the elasticity of the fraud rate in response to a change in audit policy. Our contributions are fourfold: (1) we adapt the OAM model to function with any potential machine-learning algorithms, significantly reducing the total expected cost of fraud; (2) we modify the signal treatment of OAM to create σOAM, a new cost-sensitive framework that fully leverages

the enhanced predictive capabilities of modern ML classifiers; (3) we demonstrate that σOAM outperforms other widely tested cost-sensitive expert systems by being more aggressive toward groups of claimants with higher expected fraud rates; (4) we evidence the increased importance of law firms in patterns identifying fraud.

Keywords— Auto-insurance fraud, Machine-Learning, Cost-sensitive model, Optimal Audit, Classification, Red-Flags

JELnumbers— D81, D82, G22

1 Introduction

Over the years, insurance companies have spent a considerable amount of resources establishing in-house business processes to audit suspicious claims. Current fraud detection processes are comprised of sequential layers, each feeding information to the next one. The critical *triage* layer is composed of human teams who decide if a claim should undergo a full audit. Given the volume of claims supplied, a triage team has very little time to evaluate each claim. Insurance claims are complex, high-dimensional entities, making short assessments inefficient. As automated expert systems offer an instantaneous overview of a high number of dimensions and produce consistent recommendations, they are prime candidates to assist triage experts in their workflow. However, an audit decision must primarly be a business decision. We propose a new expert system that integrates machine learning (ML) classifiers to produce an estimated fraud probability. Once a fraud probability is estimated, an audit decision is recommended based on three monetary factors: the cost of investigation, the expected cost of the claim and the estimated deterrence effect on potential fraud entailed by the chosen audit policy.

Our study sits at the crossroad between economic and data-science literature. The use of expert systems in insurance fraud detection have been theorized in economic literature following the assessment that an optimal contract induces an audit-zone [1]. Classifying algorithms such as logistic regression, tree, Bayesian classifiers, boosted tree, etc,[2] have

[1] See Picard (2013), Mookherjee and Png (1989), Townsend (1979)

[2] Viaene et al. (2002, 2004), Majhi et al. (2019), Sun et al. (2007)

been tested as expert systems in this setting. A comprehensive review of the use of data-science in automobile insurance fraud is offered in Banulescu-Radu and Kougblenou (2024).

We modify and extend the Optimal Audit Model (OAM) developed by Dionne et al. (2009) to the Extreme Gradient Boosting (XGB) framework [3]. Our main contribution is to introduce σOAM, an extension of OAM that better utilize the signal emitted by the XGB model. We compare this new cost-sensitive model to existing cost-sensitive algorithms and analyze the observed increase in cost-savings. This study also contributes to the red-flags indicators literature which focuses on selecting key features that best predict fraud [4].

The rest of this paper develops as follows: the next section presents the literature review. Section 3 introduces the database upon which the model was trained. We then introduce the OAM and our modified version σOAM, its parametrization methodology, and its adaptation to the XGB framework. Section 5 is composed of five different subsections. They present: the list of red-flag indicators ubiquitous in our models, the performance of the base classifiers, the performances of the four tested cost-sensitive frameworks, and a deeper analysis of the best two performing frameworks. Finally, we present a sensitivity analysis to assess the influence of the parametrization on our global results. Section 6 concludes our study.

We show that the OAM model benefits more from the introduction of machine-learning models compared to the logistic regression model. We note an increase in performance measured both in dollars saved and in classification metrics. The extension σOAM increases these performances by directly relying on the output of the classifiers to create the discrete vector signal defined in Dionne et al. (2009). σOAM outperforms the two previously tested cost-sensitive frameworks by targeting audits specifically at groups of claimants with a higher likelihood of fraud. Additionally, we highlight the increasing involvement of law firms in the claims process.

[3] See Chen and Guestrin (2016)

[4] See Weisberg and Derrig (1998), Belhadji et al. (2000), Dionne and Belhadji (1996)

2 Literature Review and Main Contribution

2.1 Literature Review

According to the definition provided by the Canadian government, fraud to insurance consists in *"...any act committed with the intent to fraudulently obtain payment from an insurer. Insurance crimes range in severity, from slightly exaggerating claims to deliberately causing accidents or damage"*[5]. As this definition points out, if insurers could have perfect information about the outcome of a claim, fraud would not be a major concern. An insurance claim is a complex object, involving many actors such as claimants, medical teams, body-shops, legal advisors, third-parties, etc. The signal produced by each claim is inherently multidimensional and can be distorted at various levels. Insurance companies have invested substantial amounts of human and financial resources to reduce their information gap with insureds.

The two main parties of each insurance contract are the insured and the insurance company. Insureds find value in insurance products because they allow for the smoothing of wealth in case of unpredicted events. Effectively, the insured, or agent, enters into a contract with an insurance company, or principal, to share an exogenous risk between the two parties. An insurance contract is characterized by a lost amount x, a reimbursement function $I(x)$, and a premium P paid by the agent to the principal. Costs associated with the processing of a claim can be modeled either as a constant C or as a function of the loss $C(x)$.

Designing an optimal audit strategy has been a major focus in the economic literature since economists started to model interactions between principals and agents through contract theory. Concerning the insurance industry, the main focus has been to characterize an optimal contract producing a Pareto-efficient allocation of risk between each party (Arrow (1964, 1974)). In particular, Raviv (1979) showed that contracts with deductibles strictly dominate coinsurance contracts in the presence of proportional transaction costs and the

[5]https://www.canada.ca/en/financial-consumer-agency/services/financial-toolkit/insurance/insurance-3/11.html

absence of asymmetry of information.

The loss of a claim is not strictly exogenous as the agent can select a level of effort that minimizes his expected loss. When effort is not observable by the insurer, this situation leads to a moral hazard problem. As the principal can only partially observe the agent's activity, the optimal contract must provide an incentive to increase the agent's effort. Mirrlees (1976) shows that a deductible is a valid incentive to encourage such a high level of effort. Holmstrom (1979) proves that moral hazard warranties the use of a monitoring set-up to further strengthen the agent's commitment.

Incorporating monitoring costs, Townsend (1979) produced the now canonical Costly State Verification model (CSV). In this framework, only the agent is fully aware of the loss and can be tempted to dissimulate its true value to the principal. In the context of the insurance industry, the insured could be tempted to exaggerate the magnitude of the loss. The model assumes that auditing is always reliable in unveiling the true nature of the loss.

The desired outcome is an optimal consistent contract where lying about the realization of the loss is always a dominated strategy from the agent's perspective. A CSV optimal contract is characterized by an audit region which triggers an automatic investigation for any loss falling into it. In an insurance setup, the CSV model recreates a deductible type of contract. This creates a non-audit region that includes small losses that have negative actuarial value once the audit cost is included.

Many authors further refined the CSV model. Gollier (1987) includes a fixed cost structure of audit creating nuisance claims with negative actuarial values which cannot be audited and hence are not insured. Expanding on the principal-agent model developed by Grossman and Hart (1983), Mookherjee and Png (1989) derived the conditions under which random auditing dominates deterministic auditing. The authors showed that substantive penalties to audited fraudulent claims act as a sufficient deterrent to potential fraudsters and reduce the global cost of audit.

Furthermore, when claimants have the possibility to manipulate signals[6] through third-

[6]See Subelj et al. (2011), Bodaghiand and Teimourpour (2018)

party activities[7], Crocker and Morgan (1998) show that the resulting optimal contract entails a level of misrepresentation for all losses but the extreme ones. The authors demonstrate that if the misrepresentation monotonically increases in the loss level, then the optimal contract is also monotonic and provides over-insurance for small losses and under-insurance for large losses. Dionne and Gagné (2001) empirically highlighted the relation between a high deductible and a propensity to inflate losses in the auto-insurance industry.

Since the volume of claims prohibits any credible promise of full audit of any significant loss region, random auditing seems to be a dominating strategy (Mookherjee and Png (1989)). So far, the auditing decision was solely based on the amount claimed. Any methods improving the random odds to audit a fraudulent claim would simply further decrease the cost endured by insurance companies. Insurance claims are complex objects involving a large number of different parties. Levying extra information through expert systems is a cost-efficient and scalable practice widely used in the industry.

A large part of the empirical literature focuses on independent variables, or red-flags selection to detect fraud. Weisberg and Derrig (1998) proposed a consolidated list of red-flag indicators for bodily injury claims after conducting interviews with adjusters and special investigators. Belhadji et al. (2000), following the study done in Dionne and Belhadji (1996), used a Probit model to identify a list of significant indicators to predict fraud in the property damage auto-insurance claims.

Since its development by Chen and Guestrin (2016), the XGBoost classifier has been well tested on insurance datasets (see Pesantez-Narvaez et al. (2019)). XGBoost is an ensemble learning framework where base learners are used sequentially to navigate a selected loss function. Each base learner is used to further refine the classification of the sum of the predictions made by all the previous base learners, focusing on rectifying classification errors of the previous iterations. Base learners here are either decision trees or linear regression models. To interpret the results obtained by XGB, Shapley Value is often used to rank the independent variables in terms of their contribution to the prediction.

[7]Garage, medical examiner, towing etc...

Shapley value analysis, developed by Lloyd Shapley within game theory literature, is now widely used to interpret machine learning models. As described in Lundberg et al. (2018) and Lundberg and Lee (2017), SHapley Additive exPlanation, or SHAP, value analysis is a consistent and model agnostic method to evaluate the importance of individual features in the prediction. The fundamental idea behind the SHAP method is to sequentially add the contribution ϕ_i of each feature to the general expected value $E[f(x)] = 1$ given a vector of characteristics x to recreate the model's prediction $f(x)$. For a non-linear model $f(x)$, the order in which features are introduced matters. Hence, all combinations of order of features need to be evaluated and averaged to obtain a global estimate of the contribution of each individual feature to the model's output. Individual feature contributions are then ranked by their absolute averaged score. We use the R package *"SHAPforxgboost"*[8] to isolate a set of size n top contributing variables based on their averaged absolute contribution.

2.2 Main Contribution

The selection of a classifier used as the expert system becomes now a crucial aspect. A large list of classifiers has been studied, such as regressions, trees and boosted trees, Bayesian predictors, shallow and deep neural networks, SVM, KNN, etc.[9]. Supervised and unsupervised versions of these learning methods have both been applied[10]. A classifier's optimality varies regarding the various parameters of the problem, such as interpretability, data availability, noise in data, prediction robustness, and cost reduction. Cost-sensitive classifiers are a type of classifier which aims to minimize the cost associated with the predictions, rather than maximizing a classification metric[11].

Dionne et al. (2009) offer a unified approach to fraud detection by incorporating the score signal into an asymmetric information setting with monitoring costs similar to Townsend's CSV model. The Optimal Audit Model (OAM) allows for ex-post signal manipulation by the claimant but also takes into account the deterrence effect of an audit policy, the idea

[8] https://liuyanguu.github.io/post/2019/07/18/visualization-of-shap-for-xgboost/references

[9] See Viaene et al. (2002, 2004), Majhi et al. (2019)

[10] See Nian et al. (2016) and Brockett et al. (2002)

[11] See Sun et al. (2007), Höppner et al. (2022)

being that if an insurance company decides to fully audit every claim, no one is likely to ever fraud again. The OAM is akin to other cost-sensitive frameworks because it selects a threshold based on a cost metric rather than prediction accuracy and, to our knowledge, is the only cost-sensitive model accounting for the entire cost of fraud through the deterrence effect.

In Dionne et al. (2009), the optimal threshold is derived from the set of binary predictors with significance levels above 99% in a logistic regression to predict fraud. OAM was tested on a sample consisting of 2,000 original claims from a European company. The estimated total cost of fraud for the company over a one-year period amounted to 622 million euros. The model contributes to a 60% reduction in the expected total cost for the company. Thanks to the inclusion of residual fraud, the model demonstrates that investigation units should not be evaluated merely on profit, as it is optimal to audit claims with negative expected profit for deterrence effect. Our paper prolongs the work of Dionne et al. (2009) by incorporating new machine learning classifiers into the OAM framework, offering an extension to the original theoretical OAM and testing it against modern cost-sensitive frameworks.

Two cost-sensitive frameworks are tested: CSboost and Elkan threshold. CSboost is a cost-sensitive version of XGBoost developed by Höppner et al. (2022). Using the results obtained by Elkan (2001), the authors modified the XGBoost algorithm to a cost-sensitive boosting algorithm where each newly added weak learner is trained to minimize the Average Expected Cost. The algorithm is implemented via the R package *"CSboost"*, developed by Höppner et al. (2022). The algorithm does not allow for missing entries in the data, resulting in an extra loss of observations.

Elkan-Set threshold is a cost-sensitive method developed by Elkan (2001). The method consists of selecting a prediction threshold p^* that minimizes the expected cost of audit, accounting for the various costs of classification, including the cost of a false-positive classification. The optimal prediction threshold is calculated on a calibrated output. We use the isotonic regression developed by Jiang et al. (2011) as the calibrating function.

Our modified OAM, σOAM, saves 34.6% of the cost of fraud, which is evaluated as the

distance between perfect prediction and the corner solution of not auditing. The second-best performing cost-sensitive model using an ML classifier manages a gain of 30.2%. The difference in performance is due to the deterrence effect of the OAM framework, which results in targeting groups of claimants where more cases of false negatives are present.

3 Data

The database, provided by a Canadian insurance company, is comprised of 70,813 claims with *Accident Benefit* (AB) elements opened from 2012 to 2020. An AB claim consists of anyone *"... who has been injured in a motor vehicle collision can claim for accident benefits. Pedestrians and bicyclists who have been in contact with a vehicle qualify for accident benefits."*[12]

Particular care has been taken in selecting the independent variables among the 800 features originally available. In our context, only variables available early in a claim's life were selected. This requirement limits our final performance but better reflects the actual performances of the models in practice. A claim's life cycle spans from its opening to the day it is closed. Our limitations arise from two reasons. The first reason has to do with the difficulty of recovering any amount already disbursed to a fraudulent claim, reducing any potential gain from an audit. The second reason pertains to information availability. If a claim is suspicious to an adjuster, the said adjuster can ask to obtain various extra-specific variables. These variables may artificially inflate the classifier's performance and act as biased predictors.

3.1 Dependent variable

Claims are labeled by three classes: *"Suspicious"* with 2% of the population, *"Fraud"* with 3%, *"No_label"* with 95% of the population. *"Suspicious"* and *"Fraud"* claims are observations that have been transferred to the National Fraud Tracking System (NSTF) and not been discarded after 5 days. *"Suspicious"* claims are claims where clear suspicious

[12]https://www.aboutkidshealth.ca/article?contentid=1266&language=english

activities were detected but no definite amount of money retrieved or not disbursed as opposed to *"Fraud"* claims.

Claims are sent to the triage phase following the appearance of one or several leads. The main sources of leads in terms of volume are from human adjusters who handle the regular process of claims, in-house and external detection models, and the watch-list suppliers list. These various sources produce different quality of leads, resulting in diverse shades of suspiciousness. It was unfortunately not possible to group observations according to their source origin with our database.

3.2 Literature indicators

A part of the literature on fraud detection for bodily injuries has been dedicated to the discovery of meaningful early indicators. Weisberg and Derrig (1998, 1991) asked adjusters and investigators to grade claims respectively on a suspicious/fraud scale. Each possible subset from the indicators list of up to ten variables was then fitted through a regression analysis. The best five models to predict suspicious/fraudulent claims were kept and further analyzed. The authors noted a core set of 29 indicators among the 65 tested variables.

A few remarks should be made concerning the indicators from Weisberg and Derrig (1998). The authors aimed to recreate the classification made by human teams to understand the factors that were consistently looked upon. These indicators were created through interviews where adjusters were answering *Yes* or *No* questions related to their claims. Some indicators are therefore subjective and only appear late in the claim, such as: *"Claimant was uncooperative"* or *Long disability for a minor injury.* These types of indicators cannot be used in the framework of this study, as our goal is to develop an early detection system. Furthermore, observations are here aggregated at the contract level, making the tracking of a single injury difficult. Finally, the data extraction process itself needs to follow strict regulations concerning confidentiality and security, making specific data extractions difficult. presents all 65 variables used in Weisberg and Derrig (1998). It indicates the 22 indicators that can be recreated using our partner database.

3.3 Final raw variables dimensions

Once as many of the known indicators are replicated from the insurer's database, the first step in filtering the remaining features is to discriminate the variables known to be late in the detection process, too sparsely populated, incoherent, or redundant. There is an asymmetry of knowledge between suspicious/fraudulent and normally processed claims, as extra information is often required for the former. It is hence not surprising to observe an uneven distribution of information, with post-fact variables being more and more sparsely populated. Approximately 65% of the feature variables have fewer than 50% of observations missing, resulting in around 370 sparse variables that needed to be discarded.

The final 126 extracted features can be grouped into 10 distinct sets, controlling for diverse aspects of a claim. presents them before preprocessing and feature engineering. Some variables are not directly used by the classifiers but are necessary for the preprocessing and feature engineering steps. and give a detailed focus on analyzing the features composing the final database. Features are analyzed following two categories: continuous or categorical. The final database before one-hot encoding is composed of 70,412 observations with 126 feature variables. After one-hot encoding the factor variables, the number of feature variables rises to 207. One-hot encoding consists of breaking up a categorical variable comprised of many categories into a set of dummy variables, one for each previous category.

4 Models

4.1 Logistic Regression

Equation (1) presents the logistic model. $\hat{y}$ is the latent output of the linear model. By default, when $\hat{y} > 0.5$ then the claim should be audited. Variables are grouped into families that represent different dimensions of the claim.

$$\begin{aligned}\hat{y} = \alpha_1 Veh_Age + \alpha_2 PolReport + \alpha_3 CoseOfLoss + \alpha_4 HowReported \\ + \alpha_5 PointOfImpact + \alpha_6 Occupation + \alpha_7 LawVar + \alpha_8 Towing \\ + \alpha_9 AccidentCharact + \alpha_{10} VehCat + \alpha_{11} VehMake + \alpha_{12} BodyShop \\ + \alpha_{13} Educ + \alpha_{14} Reserve + \alpha_{15} Injuries\end{aligned} \tag{1}$$

Table 10 in Appendix C presents the details of the independent variables chosen for the regression.

4.2 XGB Classifiers

Developed by Chen and Guestrin (2016), ExtremGradientBoosting (XGB) is a supervised learning algorithm which can be applied to classify instances in classes. XGB is an ensemble learning framework that relies on the iterative use of weak classifiers [13] to correct its loss function.

Given a convex and differentiable loss function $l(y_i, \hat{y}_i)$, where $\hat{y}_i$ represents the model's prediction, XGBoost introduces a base learner $f_t(\mathbf{x})$ at iteration t in order to minimize the approximate general loss function $\hat{\imath}^t$. Here, $\mathbf{x}$ denotes the vector of independent variables. The goal at each iteration is to refine the model by adding $f_t(\mathbf{x})$, which further reduces the loss function, improving the model's overall predictive accuracy.

$$\hat{\imath}^t = \Sigma_{i=1}^{n} \left[l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{2}$$

Eq(2) is the second order Taylor expansion of the exact loss function at point $f_t(\mathbf{x_i})$ where y is the true label, $\hat{y}$ is the assigned label by the model, $\mathbf{x_i}$ is the vector of independent variables for observation i, and f_t is the t^{th} learner added to the prediction of $\hat{y}$. The $\Omega(f_t)$ is a regularization term described in (3). The regularization $\Omega(f_t)$ penalizes the model by the number of leaves T and the weight given to each observation at a given leaf w. The parameters γ and λ are left to the discretion of the user and respectively control the size of the base learner and the smoothing of individual predictions.

[13] Commonly classification tree or linear regression

$$\Omega(f_t) = \gamma * T + \frac{1}{2}\lambda||w||^2 \tag{3}$$

Eq(2) adds to the previous loss function $l(y_i, \hat{y}_i^{t-1})$, the added weight of the t^{th} base learner $f_t(\mathbf{x_i})$ for observation $i, i \in [1, 2, ..., n]$. Note that the term $l(y_i, \hat{y}_i^{t-1})$ in (2) is a constant and hence does not contribute to the optimization. Therefore (2) simplifies to (4):

$$\hat{\imath}^t = \Sigma_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{4}$$

where g_i and h_i are respectively the first and second order derivatives of $l(y_i, \hat{y}_i^t)$ at point $\mathbf{x_i}$. The extra weight added by the t^{th} base learner for observation i is simply the weight of the leaf containing the observation i. (4) can be rewritten in (5) and (6). Let $I_j = \{i/q(\mathbf{x_i}) = j\}$ the leaf j containing the observation $\mathbf{x_i}$:

$$\hat{\imath}^t = \Sigma_{i=1}^n \left[g_i w_i + \frac{1}{2} h_i w_i^2 \right] + \gamma T + \frac{1}{2}\lambda \Sigma_{j=1}^T w_j^2 \tag{5}$$

$$= \Sigma_j^T \left[\Sigma_{i \in I_j}(g_i) w_j + \frac{1}{2}(\Sigma_{i \in I_j} h_i + \gamma) w_j^2 \right] + \gamma T \tag{6}$$

where (6) is obtained from (5) by summing the added weight by leaf instead of by observation. For a fixed structure of tree $q(\mathbf{x})$ for each base learner, one can extract from (6) the optimal form for w_j^*, the weight of leaf j:

$$w_j^* = -\frac{\Sigma_{i \in I_j} g_i}{\Sigma_{i \in I_j} h_i + \lambda} \tag{7}$$

Replacing the optimal weight w_j^* in (6) results in the scoring function $\hat{\imath}^t$. $\hat{\imath}^t$ is used to select which base learner $f^t(\mathbf{x})$ is optimal to add at iteration t as described in (8).

$$\hat{\imath}^t(q) = -\frac{1}{2}\Sigma_{j=1}^T \frac{\Sigma_{i \in I_j} g_i}{\Sigma_{i \in I_j} h_i + \lambda} + \gamma T \tag{8}$$

Appendix F further details the splitting procedure to select the variables in each base classification tree.The hyper-parameters are selected after a grid search cross-validation

procedure using a statistical approach of the F1 score performance measure developed by Goutte and Gaussier (2005) and detailed in Appendix G with results in Appendix H [14]

4.3 Optimal Audit Model (OAM)

This section presents the OAM of Dionne et al. (2009) and its extension σOAM. An audit strategy is defined as a function mapping each claim to a recommendation of audit. This function uses as input a signal emitted by a base classifier such as a logistic regression or any other machine learning classifier. An audit function is said *"Optimal"* when it minimizes the total expected cost of fraud.

OAM differs from typical cost-sensitive models as it encompasses an economic approach at its roots. The fraudulent behavior is modeled at the customer level by a bi-variable utility function $\mathcal{U}(W, \omega)$ where W refers to the wealth of the customer and ω refers to a fixed moral cost associated with the action of fraud. We assume $\mathcal{U}^{'}(W) \geq 0$ and $\mathcal{U}^{''}(W) \leq 0$ and $\mathcal{U}^{'}(\omega) \leq 0$. ω follows a distribution on $\mathbb{R}+$ over the entire population of claimants.

Figure 1 displays the decision facing a claimant. The claimant will defraud if and only if his expected utility is higher than the status quo, that is (9) holds. Let A be the event that a claim is audited. If a claim is audited, its true nature is revealed. The expected utility is a function of: $P(A/F)$, the probability of audit conditional on the claim being fraudulent, B a financial penalty imposed on a claimant caught being fraudulent, t the average compensation the claimant can expect from a successful fraudulent claim and W_0 the claimant's wealth after the insurance premium is paid.

[14] $F_1 = \dfrac{2\text{tp}}{2\text{tp} + \text{fp} + \text{fn}}$, where tp = true-positive, fp = false-positive and fn = false-negative

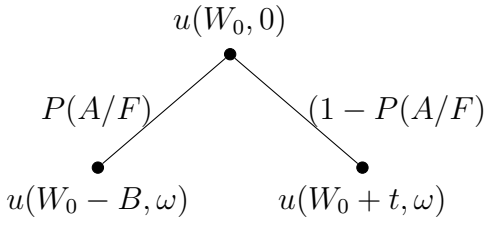


Figure 1: Claimant's choice

$$(1 - P(A/F))U(W_0 + t, \omega)) + P(A/F)U(W_0 - B, \omega)) \geq U(W_0, 0) \tag{9}$$

Each claimant is segmented into θ groups. In the original study, θ segmentation uses claimants' characteristics known by the insurer independent of the distribution of moral cost ω. In this study, θ are formed following the amount of AB Premium charged to the claimants, reflecting the risk segmentation done by the insurance company on its pool of claimants.

When (9) becomes an equality, the claimant is indifferent between the certain and uncertain state. The function $\phi(P) : [0, 1] \rightarrow \mathbb{R}+$, links the probability $P(A/F)$ to the moral cost ω corresponding to such indifference. Appendix I gives additional details on the characteristics of the $\phi(P)$ function.

An audit strategy is determined by two key factors: the θ group to which the claimant belongs and the signal emitted by the claim itself. Each claim's signal is represented by a vector σ of binary variables with dimension k. Since every $\sigma_i \in \Sigma$ is a binary variable, the total number of possible signal configurations is finite, with a cardinality of $2^k = l$. These signals, derived from the base classifier, reflect its top predictors. Let P_i^f and P_i^n represent the probabilities that the signal σ takes configuration i, conditional on the claim being fraudulent (F) or non-fraudulent (N), respectively.

$$P_i^f = P(\sigma = \sigma_i/F) \tag{10}$$

$$P_i^n = P(\sigma = \sigma_i/N) \tag{11}$$

Signals emitted are then ordered by their historical ratio $\frac{P_i^f}{P_i^n}$ as displayed in (12).

$$\frac{P_1^f}{P_1^n} < \frac{P_2^f}{P_2^n} < ... < \frac{P_i^f}{P_i^n} < ... < \frac{P_l^f}{P_l^n} \tag{12}$$

Dionne et al. (2009) demonstrate that the optimization problem of selecting the optimal audit policy for each θ involves determining a rank i in (12). Claims with a signal ranked higher than i should be audited, while those with a signal ranked lower than i

should be approved without further investigation. The chosen rank i reflects the level of aggressiveness of the audit policy for a given θ group.

Let $q : \Theta \times \Sigma \to [0, 1]$ denote the audit strategy implemented by the insurance company. The audit strategy $q(\theta, \sigma)$ specifies whether a claim with signal σ should be audited based on the claimant's group θ_i. Define $Q^f(\theta)$ and $Q^n(\theta)$ as the respective probabilities of auditing a fraudulent or non-fraudulent claim for a claimant of type θ. These probabilities, $Q^f(\theta)$ and $Q^n(\theta)$, are determined by the chosen audit strategy $q(\theta, \sigma)$, as shown in (13) and (14).

$$Q^f(\theta) = \Sigma_{i=1}^{l}(P_i^f q(\theta, \sigma_i)) \tag{13}$$

$$Q^n(\theta) = \Sigma_{i=1}^{l}(P_i^n q(\theta, \sigma_i)) \tag{14}$$

The recommended audit policy aims to minimize the total expected cost of fraud, which consists of two main components: the expected cost of investigation (15) and the expected cost of residual fraud (16). The model excludes opportunistic fraud, assuming that a claimant either fully fabricates a fraudulent injury claim or the injury is genuine. Therefore, π represents the probability of a claim being a legitimate Accident Benefit (AB) claim. Let $H(\omega/\theta)$ denote the conditional cumulative distribution of the moral cost for claimants in group θ, and $h(\omega/\theta)$ its corresponding density function.

$$c\pi E_\theta[Q^n(\theta)] + c(1-\pi)E_\theta[Q^f(\theta)H(\phi(Q^f(\theta)/\theta))] \tag{15}$$

$$t(1-\pi)E_\theta[(1-Q^f(\theta))H(\phi(Q^f(\theta))/\theta)) \tag{16}$$

Adding (15) and (16) and minimizing the sum over $q(\theta, \sigma)$ gives the global optimization problem in (17).

$$\begin{aligned} Min_{q(\theta,\sigma)} E_\theta \Bigg[& c\pi[Q^n(\theta)] + c(1-\pi)Q^f(\theta)H(\phi(Q^f(\theta)/\theta)) \\ & + t(1-\pi)\Big((1-Q^f(\theta))H(\phi(Q^f(\theta))/\theta))\Big)\Bigg] \end{aligned} \tag{17}$$

Proposition 1 of Dionne et al. (2009) shows that, under limited conditions, the problem of solving (17) consists in finding an optimal severity index i in (12) for each θ. Then the optimization problem becomes (18):

$$Min_{i_\theta} E_\theta \Bigg[c\pi\mu(i^*_\theta) + (1-\pi)H(\phi(\lambda(i^*_\theta))/\theta) \Big(c\lambda(i^*_\theta) + t(1-\lambda(i^*_\theta))\Big)\Bigg] \tag{18}$$

where $\gamma(i) = \Sigma_{i=1}^{l}(P_i^f)$ and $\mu(i) = \Sigma_{i=1}^{l}(P_i^f)$. Due to the discretization of the top five predictors to form the σ signal vectors, it is possible to have a claim in the test-set whose signal σ is unknown in the training set. For such observations, no predictions are possible as the optimal threshold *"i"* in (12) cannot be trained.

4.4 Parametrization of OAM

Dionne et al. (2009) describes an implementation methodology that is followed here. The authors start by approximating the average fraud rate Γ by extracting $\hat{\pi}$, the observed probability of filling a claim, fraudulent or not. Assuming a global fraud rate of $P(F/\theta)$, the authors deduce the probability of observing a fraudulent claim $\hat{\Gamma}(0,\theta) = \hat{\pi}P(F/\theta)$ and the probability of a genuine accident $\pi = \hat{\pi}(1 - P(F/\theta))$.

$$\Gamma(Q,\theta) = (1-\pi)H(\phi(Q^f_\theta)/\theta) \tag{19}$$

$$\eta(Q,\theta) = \frac{-Q^f_\theta \phi'(Q^f_\theta) h(\phi(Q^f_\theta)/\theta)}{H(\phi(Q^f_\theta)/\theta)} \tag{20}$$

$\hat{\Gamma}(0,\theta)$ is the probability of observing a fraudulent claim observed with the actual audit policy. To estimate the effect of a deviation from the current audit policy, we must calculate the elasticity of the fraud rate with respect to the investigation strategy: $\eta(Q,\theta)$ in (20). Dionne et al. (2009) highlights the difficulty of estimating such a function as it depends on the estimation of the moral cost distribution and on a parametrization of the function ϕ: $H(\phi(q)/\theta)$. The authors opted for the following approximation in (21):

$$\Gamma(Q,\theta) = \hat{\Gamma}(\theta)(1-Q)^{\gamma(\theta)} \tag{21}$$

(21) is used to approximate (18) into (22). The parameter $\gamma(\theta)$ controls the deterrence effect per θ or the sensitivity of a group's fraud rate to a change in the audit policy.

$$Min_{i_\theta} E_\theta \Bigg[c\pi\mu(i^*_\theta) + \hat{\Gamma}(\theta)(1-\lambda(i^*_\theta))^{\gamma(\theta)} \Big(c\lambda(i^*_\theta) + t(1-\lambda(i^*_\theta)) \Big) \Bigg] \tag{22}$$

In our study, θ groups are formed by increasing deciles of Accident Benefit[15] (AB) premiums charged to the insured. Each observation is given an index from 1 to 10, 10 being the group with the highest AB premium paid by the insured. Table 1 gives the population of claims per θ with our data as well as the thresholds forming each group.

The probability of filling an AB claim, $\hat{\pi}(\theta)$, is parameterized using all the active policies in 2019. Active policies are classified and summed per θ groups. All AB claims with a date of loss in 2019 are extracted from the main train database, grouped and counted per θ. The ratio of the two extractions gives an estimate of the probability that an active policy would fill an AB claim per θ. The global probability of filling an AB claim is about 1% for the year 2019. This is lower than the average probability of being involved in an accident of about 1.9%, given by the local government in 2018. Table 1 shows an expected increase in filling rate $\hat{\pi}$ in relation to the premium group of the claimant. A higher AB premium does reflect a higher risk of injury overall as is depicted in Table 1.

In Dionne et al. (2009), $P(F)$ was given by the industrial partner for a general representative customer at 8%. In consultation with our industrial partner, we assume a global fraud rate $P(F)$ of 15% and distribute it at θ levels following the methodology presented in Appendix J. A sensitivity analysis is later conducted to test this base assumption. All results presented below in this article are calculated with this 15% hypothesis. We also consider that all investigated claims were truly fraudulent although no amount of money could be recovered. This hypothesis is also relaxed later when the target population is limited to claims from which an amount of dollars was recovered.

Table 1 shows a steady increase in the probability of fraud by index of θ group. Claims with higher AB premiums are more likely to be fraudulent. The observed fraud rate in the

[15] AB claims are claim with bodily injuries.

Table 1: Population and parameters per θ

θ	Pop	$\hat{\pi(\theta)}$	Thres$	P(F/θ)	$\pi(\theta)$	$\Gamma(0,\theta)$	$t(\theta)$	$c(\theta)$
1	15153	0.68%	0	11.6%	0.60%	0.07%	16997	2219
2	5303	0.66%	221	11.0%	0.59%	0.07%	15538	1639
3	5036	0.70%	283	11.7%	0.62%	0.08%	15032	1698
4	4970	0.78%	341	12.4%	0.68%	0.09%	14604	1662
5	5011	0.77%	402	13.4%	0.66%	0.10%	14813	1793
6	5349	0.95%	472	14.1%	0.81%	0.13%	14821	1868
7	5813	1.0%	558	15.4%	0.85%	0.15%	14653	1941
8	6478	1.2%	670	16.8%	1.0%	0.21%	15227	2112
9	7836	1.4%	834	18.6%	1.1%	0.26%	14947	2144
10	9463	1.5%	1137	22.3%	1.2%	0.35%	15061	2365

Note: Column *Pop* is the population per θ in the Train-set. $\hat{\pi(\theta)}$ is the probability of filling an AB claim. *Thres$* is the AB premium lower limit that form the θ group. *P(F/θ)* is the probability of fraud. $\pi(\theta)$ is the probability of accident. $\Gamma(0,\theta)$ is the probability of observing a fraudulent claim. $t(\theta)$ is the average compensation per claim and $c(\theta)$ is the average cost of investigation.

database is about 5% and of which 37% is concentrated in the last AB premium decile. The estimated residual fraud rate is stable at around 9.5% per θ

The average compensation per claim, $t(\theta)$, is directly observed at the claim level. The database is comprised of 153 sub-files disbursement variables which track any amount disbursed by the insurer on a claim at the sub-policy level. The total amount is summed by θ and then averaged. Due to outliers, we impose a cap of 50000$ in total compensation per claim[16]. This is not a strong assumption as any claim reaching a higher amount is automatically audited. We observe higher compensations for claims belonging to smaller AB premium. This is explained by a density of compensation with heavier tail for smaller AB premiums.

Based on estimates from our industrial partner, the average cost of investigation, $c(\theta)$, includes three components: the fixed salary costs for an adjuster and online investigation ($(700\$ + 300\$)$ per claim), additional medical examinations directly observed in the data, and extra surveillance costs ranging from $4,000 to $7,000.

We are left with estimating the cost and the expected cost per claim of extra-surveillance which corresponds to expenditures such as hiring private detectives. Appendix K presents a detailed breakdown of the methodology upon which the expected extra-surveillance costs are estimated. The primary driver of global investigation costs is the additional medical expenditure, which is directly observed in the data and increases for θ groups with higher AB premiums.

4.5 Accident signal σ

After having parameterized the OAM, we need to input the signal emitted by the classifier. The original model from Dionne et al. (2009) allows for σ accident signals to only be a set of binary variables. The optimization process is hence simplified to selecting the optimal threshold *"i"* in (12)[17]. Dionne et al. (2009) used a logistic regression and selected the top significant variables at the 1% level to form the signal vector σ. Since variables were

[16]The 3rd quantile is at 20000$

[17]Proposition 1 in Dionne et al. (2009)

only binaries, they formed a finite set of possible signals. In our case, main predictors can be continuous or multilevel. In such cases, variables are discretized into quartiles or one-hot encoded to generate a finite set of possible signal vectors σ. Shapley value analysis is used to identify the top predictors from the ML models tested. Unlike in Dionne et al. (2009), such features in our data are most likely not binary. One-hot encoding is carried out during the preprocessing for all factorial features extracted. Continuous variables are segmented into quantiles and then one-hot encoded themselves to form the signal vector.

4.6 σOAM

We propose an extension to OAM called σOAM. In σOAM the signal σ is created directly from the output of the classifier. An ML classifier produces a continuous output that is used to rank observations. Traditionally, a cut-off function is optimized to select a threshold that maximizes a chosen classification metric. We use the optimization function (22) as the cutoff selector. To obtain a finite set of potential signals, we discretize the output vector of the ML classifiers into centiles which form the signal sub-groups. The optimization problem is similar to (18) but *"i"* corresponds to the index of the signal sub-group in (12) that minimizes the expected cost of fraud. This extension provides several advantages.

The first advantage relates to a steep improvement in performances. An upgraded ML classifier will directly impact the accuracy of the predictions of the OAM layer and increase the performances in terms of dollars saved. The original OAM's F1 score relies solely on the information contained in the top variables contributing the most to the ML classifier's prediction. However, two ML classifiers may have a similar list of ranked predictors but different classification performances. Table 9 in Appendix B shows the top predictors of each model in every configuration. In the Default configuration, which corresponds to the raw database presented in Section 3, the XGB framework with the Linear and Tree booster have almost an identical list of top predictors. Much of the $\{Model, Configuration\}$ pair have a similar list of top predictors. As shown in Table 4 of Section 5, the decrease in F1 scores after applying the optimization function (22) is more pronounced for OAM than for σOAM.

Table 2 details the different predictions made by the linear booster in the default configuration using σOAM or OAM. There are a total of respectively 120 and 100 i subgroups for each model. Column Rank shows the global aggressiveness of both OAM and σOAM. A higher rank means that a signal with a lower ratio $\frac{P_i^f}{P_i^n}$ is audited. We see that σOAM audits from ranks 14 ($i = 86$) to 16 ($i = 82$) onward, sending to be audited about 67.6% of the fraudulent claims and 12% of the non-fraudulent claims. OAM model is more aggressive, auditing claims with signal ranking from 34, ($i = 22$) to rank 38 ($i = 100$). This audit policy leads to a similar audit rate of fraudulent claims of 66.8% while a higher 19.4% of non-fraudulent claims are audited. This difference of 7.4% in the percentage of non-fraudulent audited claims makes σOAM outperform OAM in terms of amount saved.

There are additional advantages to using σOAM to OAM. Results show that σOAM is more robust than OAM to different configurations but is also less difficult to parameterize as a different choice of discretization of selected continuous variables would lead to a different set of results. Another advantage is that predictions are always possible on a test set since all signals are included in the training set. In OAM if a continuous variable is selected among the top predictors, it is tempting to increase the precision of the model by creating smaller subgroups of index i instead of the 4 quartiles used in this study. However, this will increase the probability of receiving unknown groups of signals from the test set making predictions impossible. σOAM is not hindered by such dilemma and will directly benefit from any upgrade made to the classifying technology.

5 Results

Two types of results are obtained in the optimal audit model. The first consists of the F1 score generated by the base classifiers [18] as a measure of the efficiency of the classification of claims on the test set. The second is the cost-saving estimate resulting from the use of the OAM. The test set is comprised of 14.080 observations in the default configuration and is created through a stratified-sampling procedure implemented by the R-package

[18] XGBoost tree,dart and linear and the Logistic regression

Table 2: σOAM and *OAM* Linear booster in default configuration

	σOAM					OAM				
θ	Rank	i groups	Thres	% F Audit	% NF Audit	Rank	i groups	Thres	% F Audit	% NF Audit
1	11	90	86	0.58	0.08	31	19	22	0.60	0.16
2	12	89	86	0.61	0.09	32	101	104	0.60	0.17
3	13	88	86	0.64	0.10	33	130	104	0.60	0.17
4	14	86	86	0.66	0.11	34	22	31	0.64	0.18
5	15	87	86	0.68	0.12	35	104	31	0.66	0.19
6	16	82	86	0.69	0.13	36	31	31	0.66	0.19
7	17	85	86	0.71	0.14	37	26	31	0.69	0.19
8	18	80	86	0.72	0.15	38	100	26	0.69	0.20
9	19	84	87	0.74	0.16	39	21	26	0.70	0.21
10	20	83	82	0.75	0.17	40	18	100	0.70	0.22

Note: Rank is the increasing indexed σ subgroups created by each claim. As the rank increases, the True Positive Rate decreases. Thres is the optimal signal threshold below which no claim should be audited. % F Audit and % NF Audit are respectively the percentage of fraudulent and non-fraudulent claims audited.

Table 3: Base classifier Default Configuration

Config Name	Classifiers	# features	AB_label	Law
Default	Tree/Dart/Linear & logis	190	Investigated	Present
No-Law (NL)	Tree/Dart/Linear & logis	186	Investigated	Absent
Fraud-Only (FO)	Tree/Dart/Linear & logis	190	Fraud	Present

Note: Column *"Config Name"* displays the name of the configuration. *"Classifiers"* is the type of classifiers used. *"# features"* is the number of variables used in the train-set. In column AB_label, "Investigated" indicates that all suspicious claims are considered fraudulent. In column AB_label "Fraud" indicates that only observations with an amount of dollars saved are used as the target population. *"Law"* is the presence or absence of law variables in the training-set.

"SamplingStrata" developed by Barcaroli (2014)[19]. Appendix L gives the list of target variables used to obtain an accurate representative test-set.

Table 3 presents the 3 configurations tested here. The *"Default"* configuration includes all the 208 variables after preprocessing. The *"Fraud-Only"* configuration removes all observations that are labeled *"Suspicious"*[20]. Finally, the *"No-Law"* configuration removes all variables related to the presence of a law firm in the claim process. Law variables[21] represent a potential bias as a claimant might solicit a lawyer because he was signified by the insurer of a potential investigation. Because of this bias, the default configuration already censors the law variables of observations that had lawyers arriving 36 days or after the date of the loss. Despite this censorship, law variables are constantly among the top predictors of fraud and a *"No_law"* configuration is tested to gauge the robustness of the models.

[19]For a comprehensive introduction to the package please refer to: https://www.istat.it/it/files/2014/06/SamplingStrata-An-R-Package-for-the-Optimization-of-Stratified-Sampling.pdf

[20]All observations where no amount of $ where recovered, around 50% of the target population

[21]See Table 10 in Appendix C

5.1 Red-flags indicators

Out of 126 variables[22] in the dataset, there is a set of 25 core variables which are ubiquitous in predicting fraudulent claims. Out of these 25 variables, 8 of them are present in the list of top 8 main predictors between 5 and up to 11 of the 12 tested models. Despite the censorship of 36 days, we notice that 4 of these 8 features control for different aspects of lawyers on a claim. Selected law variables control for law-firms with: a high specialty in insurance claims, their presence on a claim, the number of lawyers present and the number of days before their first arrival on a claim[23]. In the *"Default"* configuration, these variables are set to 0 if the first lawyer arrived on the case after 36 days of the loss as a way to control for lawyers hired because a claim was already declared suspicious.

The other 3 main predicting variables are: the number of years the claimant has held their driving license, the number of days since the claimant has been insured with the industrial partner, [24] and abnormal channels of reporting the accident to the insurer [25]. The *"No-Law"* configuration adds red flags related to the types of injuries such as the presence of a back injury, a high psychological trauma and the delay between the date of loss and the first physical injury declared [26].

Between the 4 different classifiers tested in both *"Default"* and *"Fraud-Only"* configurations, there are 40 different potential predictors to be extracted. Appendix A shows that a succinct list of 8 variables already fills 36 slots of the 40 possible slots. Only one red flag is added to the list [27]. σ signals in OAM between the two configurations are hence very similar as they are formed by the top 5 predictors of each model.

Weisberg and Derrig (1998) listed 29 core predictors to detect fraud in bodily-injury auto-insurance claims. Due to differences in granularity between the data in our study and the indicators controlled by Weisberg and Derrig (1998), a one-to-one mapping is

[22] See Appendix C

[23] Law_rep_agrr, Law_rep_none, DTd_los_law, Nbr_lawyers

[24] The two variables are correlated because of regular client turnover

[25] Drv_licence_years, DTd_dol_Oicd, HR_unknown

[26] Inj_Back, MostSeverePsyMax, DTd_FAB_Dol

[27] Law_rep_moderate

difficult. However, out of the 25 main predictors of this study, 6 belong to dimensions cited by adjusters as indicative of fraud in Weisberg and Derrig (1998). Another 3 core indicators were tested but not selected in Weisberg and Derrig's core indicators. All but 2 of the 8 variables that form the most important subset of core predictors belong to the red-flags list selected by Weisberg and Derrig (1998). Notably, Law variables were tested in Weisberg and Derrig (1998) but were not cited by adjusters at the time as key indicators. This underlines a clear evolution in the industry as the importance of law firms in claims settlement has increased significantly, sometime leading to fraudulent practices as evidenced in this study. Appendix A lists all the 25 key predictors.

5.2 Classifiers Results

The stand-alone $F1$ column in Table 4 presents the raw classification score that reflects how well each model performs at the task of distinguishing between a suspicious and a regular claim. Table 4 shows an overall better performance of the XGB family classifiers (Dart, Linear and Tree) compared with the Logistic regression with higher classification performances for each classifier in every configuration except the pair $\{$*"Fraud-Only", Tree*$\}$. Out of the 3 base classifiers, Dart shows the overall best performance with an average F1-score of 36.1. In comparison, the average F1-score obtained by the Logistic regression is 27.0.

The average best result is obtained in the *Default* configuration with an average F1-score of 0.37 over the four classifiers. A significant decrease in F1 scores is observed when transitioning from the *Default* to the *Fraud-Only* or *No-Law* configurations. ML classifiers outperform the Logistic regression in pure classification results in 11 out of the 12 tested pairs of $\{Config, Classifiers\}$. Overall, the Dart booster is more stable and has a higher classification performance than the 3 other tested base classifiers.

Table 4 shows that σOAM outperforms OAM. The average $F1_{OAM}$ for the two models is respectively 29.4 and 21.6. We clearly see the stability of using σOAM compared to OAM. Comparing the Linear and Tree classifier in the Default configuration in Table 4, we see a slight advantage in classification performances for the Linear booster with an F1 score of 39.8 compared to 38.4. After the cost-sensitive OAM is applied, the two classifiers have

Table 4: Optimal Audit Model results

			OAM		σOAM		Elkan		CSboost	
Config	Classifier	F1	$F1_{OAM}$	Sav$	$F1_{OAM}$	Sav$	$F1_{OAM}$	Sav$	$F1_{OAM}$	Sav$
Default	Logistic	0.336	0.287	540,509	0.31	2,073,495	0.350	2,659,740	×	×
	Dart	0.401	0.302	2,220,976	0.385	3,530,975	0.402	2,973,675	0.297	1,482,911
	Linear	0.398	0.251	849,518	0.361	3,270,789	0.389	3,034,209	0.297	1,397,911
	Tree	0.384	0.27	1,521,527	0.358	2,289,313	0.339	2,580,260	0.299	1,402,410
Fraud-Only	Logistic	0.323	0.183	-1,250,041	0.229	-122,484	0.31	1,154,900	×	×
	Dart	0.383	0.212	-326,884	0.306	1,257,100	0.349	1,566,847	0.28	751,628
	Linear	0.356	0.172	-1,735,863	0.272	780,840	0.358	1,504,290	0.28	762,392
	Tree	0.318	0.197	-563,346	0.318	1,143,741	0.293	1,077,348	0.28	751,628
No-Law	Logistic	0.153	0.195	-1,490,538	0.222	461,022	0.257	1,058,786	×	×
	Dart	0.301	0.185	-1,048,756	0.287	1,738,962	0.308	1,991,014	0.095	313,487
	Linear	0.281	0.167	-2,668,173	0.263	1,106,568	0.265	1,397,390	0.091	298,356
	Tree	0.254	0.181	-1,890,283	0.221	1,073,631	0.242	1,076,887	0.091	12,653

Note: Columns $\{Config, Classifiers\}$ introduce the configuration and the classifier. *Default* configuration has all variables related to the presence of lawyers and type of law firms on the claims in the training set and also includes all suspicious claims as fraudulent. *Fraud-Only* limits the target population to claims with amount of money recovered. *No-Law* removes all variables related to law from the training set. The first row introduces the Cost-sensitive framework OAM, σOAM, Elkan and CSBoost,.... *F1* is the F1-score of the algorithm used on the test-set. Column $F1_{OAM}$ is the F1 score after the minimization of cost is applied and $Sav\$$ is the amount saved to the corner solution of not-auditing introduced in Table 5. A negative amount reflects that the corner solution of *No-audit* is more profitable.

inverted $F1_{OAM}$ performances, respectively 25.1 vs 27. As a result, the Linear Booster achieves the poorest performances in terms of the amount saved in OAM with 849,518$ to 1,521,527$ to the corner solution of not-audit. As σOAM selects its optimal threshold directly on the output of the classifiers, the higher prediction power of the Linear booster is less distorted with an $F1_{OAM}$ of respectively 36.1 and 35.8. Similarly higher amounts are saved as the classification becomes less degraded.

5.3 Performances of the cost-sensitive frameworks

Following Bahnsen et al. (2016), we assess the performance of the 4 cost-sensitive frameworks (OAM, σOAM, Elkan and CSBoost) through the cost entailed by their usage compared with the default situation of not-auditing. This allows for a fair comparison of performances across configurations as test-sets may differ. Table 5 presents the expected costs associated with the 3 corner solutions[28] of perfectly predicting fraud, auditing every claim, and not auditing any claim for each configurations/cost-sensitive framework. (23) through (25) show how each cost in Table 5 is calculated. Although the same test-set is used across all models, some corner costs differ. By construction, *"Fraud-Only"* configuration has a different test-set reference cost. Moreover, Logistic models do not accept missing values, resulting in further discarding some observations and variables. Columns Sav$ in Table 4 present the result of the difference between the minimized cost of fraud obtained by the corresponding classifiers in (22), and the corner solution of *No-Audit*.

$$All - Audit = (TP + FN) \cdot \bar{c} + (FP + TN) \cdot (\bar{c} + \bar{t}) \tag{23}$$

$$No - Audit = (FP + FN + TP + TN) \cdot \vec{t}\,) \tag{24}$$

$$Perfect - Prediction = (TP + FN) \cdot \bar{t} + (FP + TN) \cdot \bar{c} \tag{25}$$

where $\bar{t}$ and $\bar{c}$ are respectively the average compensation and cost of investigation per θ. FP, TP, FN and TN are *False-Positive*, *True-Positive*, *False-Negative* and *True-Negative* instances.

Analyzing column Sav$ in Table 4, σOAM model, paired with the Dart classifier, saves

[28]Not accounting for residual fraud

Table 5: Test-DataBase key costs of corner solutions in $

Model	Population	Perfect-Prediction	All-Audit	No-Audit
Def/NL	14,073	207,545,209	234,465,168	217,389,852
FO	13,741	209,807,439	237,073,839	215,634,936
FOLogi	11,872	179,397,549	202,898,204	184,496,828
NLLogi	12,411	180,991,409	204,654,570	190,037,764
DefLogi	12,335	180,102,741	203,655,137	188,608,623
CSboost Def/NL	11,087	160,086,400	180,790,992	168,436,333
CSboost FO	10,531	157,856,590	178,308,085	162,433,178

Note: Abbreviations Def, NL and FO refer respectively to the configuration, *Default*, *No-Law* and *Fraud-Only* presented in Table 4. Abbreviations Logi and CSboost refer to the classifiers used: Logistic regression or Cost-Sensitive boosting. *Population* shows the number of observations for each test-set. *Perfect-Prediction* is the cost of perfectly predicting all known fraudulent claims. *All-Audit* corresponds to auditing every claims and *no-Audit* is the cost of not having a fraud detection process.

$3,530,975, which is approximately 35% of the excess cost associated with fraud and auditing valid claims. Regardless of the configuration and the classifier, σOAM and Elkan cost-sensitive frameworks always produce the first and second top amounts saved. As mentioned in Table 4, *Default* configuration has all variables related to the presence of lawyers and type of law firms on the claims in the training set and also includes all suspicious claims as fraudulent. *Fraud-Only* limits the target population to claims with amounts of money recovered. *No-Law* removes all variables related to law from the training set.

Table 4 shows that over the 12 pairs {*Config,Classifiers*}, σOAM model averages savings of $1,550,329$\$ of the extra cost, and $1,799,102$\$ excluding Logistic model, respectively 17.4% and 21.4% of the extra cost. Similarly, Elkan model averages savings of $1,839,612$\$ and $1,911,324$\$ excluding Logistic model, respectively 18.1% and 22.6% of the extra cost. Although Elkan model is more stable than σOAM when the configuration changes from *Default* to *Fraud-Only* or *No-Law*, σOAM produces the top two results with pairs {*Default,Dart*} and {*Default, Linear*} and outperforms Elkan in the *"Default"* configuration. The following section expands on the comparisons between these two cost-sensitive frameworks.

The vector of the top 5 predictors selected in the *"Default"* and *"Fraud-Only"* configurations are almost the same for each classifier(see Table 8 in Appendix A). Consequently, in the *OAM* framework, the main difference between the two configurations is their target population. The average F1-scores obtained by the 4 base classifiers [29] in Table 4 decreases only slightly from 0.379 in the *"Default"* configuration to 0.345 in the *"Fraud-Only"* configuration.

However, the decrease in the average F1-score of the 4 base classifiers between the *"Default"* and *"Fraud-Only"* configurations in the OAM framework is steeper from 0.27 to 0.19 as observed in Table 4. Despite the similarities between these configurations, the steeper decrease in performance after applying the OAM framework suggests that the top predictors are less effective at classifying *"Fraud"* observations than *"Suspicious"* ones.

[29]Logis, Dart, Tree, Linear

In Table 4, except for 5 instances out of 27, *Save*$ results obtained by machine-learning classifiers are higher than their Logistic counterparts. Out of all classifiers, Dart has the highest averaged saved amount of $1,370,995$, 16.4% of the extra cost, and the top two highest performances across all 3 configurations and 4 cost-sensitive frameworks. The Logistic regression has the lowest averaged saved amount of $565,043$, 5% of the expected cost of fraud, across its 3 cost-sensitive frameworks. Overall, machine-learning classifiers produce higher and more stable savings than the Logistic regression.

In σOAM, the signal is created directly from the output of the algorithm instead of derived from the top 5 predictors. The resulting predictions produce better performances both in terms of F1-score and Sav$ in Table 4 with an average increase of respectively 37% and 199% across all configurations. The predictions are significantly more robust to configuration changes, with average F1-score increases of 37% and 47% for the *Fraud-Only* and *No-Law* configurations, respectively. The increase in performances occurs for all classifiers but further highlights the gap in performances between the logistic regression and ML predictors. Overall, the increase in performances obtained from OAM to σOAM far outweighs the increase in performances obtained by an ML classifier to the logistic regression, which emphasizes the role of an optimal cost-sensitive model.

5.4 Elkan vs σOAM

Until now, comparisons between cost-sensitive frameworks were carried out as if there were no residual fraud. Given the partner's estimated global fraud rate of 15%, σOAM is more aggressive than optimal when the observed fraud rate of 5% is assumed to be accurate. Classical cost-sensitive models cannot be parametrized to any assumed fraud rate. To fully compare σOAM and Elkan model, we need to assume the observed fraud rate of 5% labeled in the data is accurate.

At a 5% global fraud rate, we observe in Table 6 that σOAM saves $3,363,068$ compared to not-auditing, representing about 34.6% of the extra cost of fraud. Elkan model saves $2,973,675$, 30.2% of the extra cost of fraud, a reduction of $389,393$ when compared to σOAM. Table 6 details the difference between the audit policies of the two frameworks.

Table 6: σOAM vs Elkan with an assumed 5% fraud rate

			Elkan					σOAM				
			Sav\$:2,973,675\$, F1Score: 0.40					Sav\$:3,363,068\$; F1Score: 0.40				
θ	$\frac{t}{c}$	F%	TP	FP	TN	FN	Thrs	TP	FP	TN	FN	Thrs
1	7.66	1.76	15	28	3171	47	0.130	15	24	3175	47	0.190
2	9.43	1.15	3	17	1013	6	0.106	1	5	1025	8	0.374
3	8.84	1.90	8	21	943	7	0.113	5	9	955	10	0.190
4	8.78	2.62	13	16	916	17	0.114	12	12	920	18	0.130
5	8.29	3.78	15	28	898	27	0.121	15	28	898	27	0.141
6	7.95	4.54	19	31	934	27	0.126	21	35	930	25	0.141
7	7.56	5.90	20	45	1014	42	0.132	27	81	978	35	0.101
8	7.19	7.50	42	54	1146	51	0.139	53	123	1077	40	0.109
9	6.96	9.56	50	75	1376	81	0.144	78	200	1251	53	0.109
10	6.35	13.6	112	105	1485	155	0.158	183	363	1227	84	0.092

Note: columns $\frac{t}{c}$ is the ratio of the compensation to the investigation cost per θ. $F\%$ is the fraud-rate and TP, FP, TN and FN are respectively the true-positive, false-positive, true-negative and false-negative number of claims. $Thrs$ is the calibrated probability threshold above which a claim is audited. A higher threshold means a less aggressive audit policy.

The F1-scores of Elkan σOAM are similar at around 0.40. σOAM audit policy targets more aggressively the θ groups that have a higher fraud rate. On the other hand, Elkan model sets its threshold solely based on the cost-ratio. Given that investigation costs are at an order of magnitude less than compensation, σOAM focuses on diminishing the anticipated number of false negative instances thanks to its capacity to anticipate subsequent fraud rates.

Tables 15 and 16 in Appendix M present similar results for *"Fraud-Only"* and *"No-Law"*. Assuming the observed fraud rate to be correct also increases the performances in these other 2 configurations. In the *"Fraud-Only"* configuration, the fraud rate is set to 3% as explained in Section 3. Under these conditions, σOAM performs similarly to Elkan model with a saving of $1,536,980$\$ compared with $1,566,847$\$. Under the *"No-Law"* configuration, σOAM framework achieves a $2,023,254$\$ saving compared to the $1,991,014$\$ performance of the Elkan framework. Tables 15 and 16 in Appendix M show σOAM displays increased aggressiveness vis-à-vis θ with higher fraud rate, limiting the number of false-negative and behaving similarly than in the *"Default"* configuration.

5.5 Sensitivity analysis

Table 7 presents the sensitivity of the results obtained in Table 4 to assumptions made on the global fraud rate and γ, the sensitivity of the fraud rate to the audit policy. The sensitivity analysis is carried out with Dart base learner and σOAM as the cost-sensitive framework.

We observe an inverse relationship between γ and the aggressiveness of the OAM as shown by column $\#Adt$: number of audited claims. Because of the higher deterrence effect, fewer observations need to be audited to obtain an optimal cost. Inversely, a higher global fraud rate leads to a more aggressive audit policy as the cost of residual fraud increases. Cost performances are globally stable across the different parameters.

Table 7: σOAM with Dart base learner, Sensitivity Analysis at 15% fraud rate

P(F):		5% Fraud		10% Fraud		15% Fraud		20% Fraud	
Conf	γ	Sav$	#Adt	Sav$	#Adt	Sav$	#Adt	Sav$	#Adt
FO	1	1,246,400	1931	1,110,314	2547	717,753	3085	271,410	3529
FO	2	1,481,432	1549	1,307,614	1984	1,257,100	2327	1,143,011	2550
FO	3	1,633,840	1293	1,453,016	1972	1,423,141	1820	1,312,336	1972
NL	1	1,902,179	1163	1,529,541	1394	1,083,588	1741	708,874	2075
NL	2	2,023,254	746	2,042,559	1116	1,738,962	1215	1,611,014	1362
NL	3	2,008,734	683	2,104,657	832	2,085,044	984	2,028,554	1089
Def	1	3,523,823	1623	3,523,823	2023	3,180,470	2386	2,987,844	2682
Def	2	3,363,068	1290	3,473,061	1513	3,530,975	1726	3,442,445	1915
Def	3	3,160,843	997	3,431,421	1245	3,495,284	1386	3,506,196	1527

Note: column *Conf* shows the configuration's name. γ is the elasticity of the fraud rate to the optimal audit policy. Sav$ is the amount of dollars saved to the corner solution of non-auditing for a global fraud rate $P(F)$ of X%, #Adt is the number of claims audited

6 Conclusion

We have extended OAM to recent machine-learning classifiers and compared its performances to other cost-sensitive methods. ML classifiers produce better and more stable performances than the Logistic regression when incorporated into OAM. Our second contribution was to extend the OAM model to a more general fraud detection model σOAM that produces better classification and saving performances across the three different configurations and four classifiers tested.

In the default configuration, the total cost of fraud is estimated at $9,844,643$\$. When the Logistic model is applied $504,509$\$ and $2,073,495$\$ are saved respectively in OAM and in the σOAM . The best ML classifier, Dart, without a cost-sensitive threshold selection saves $3,176,912$\$. σOAM with the Dart produces the best performance with $3,530,975$\$. σOAM significantly outperforms the original OAM and other tested cost-sensitive algorithms. Furthermore, σOAM shows better performances than the Elkan-Threshold frame-

work when for residual fraud and a deterrence effect that is, to our knowledge, foreign to any existing cost-sensitive algorithm.

We also bring forward evidence of the new importance taken by law firms in claim settlements. A more accurate parametrization would lead to further improvements. OAM offers a framework to calculate an optimal audit policy from any type of continuous output. However, detecting fraudulent claims could be enhanced by combining different classes of data and algorithms. A key enabling factor to fraud is the influence of claim suppliers such as medical examiners, lawyers or mechanics. Combining complex network predictions on graph data with XGB predictions and tabular data could enhance the prediction power of the output and further refine OAM. The sensitivity of the fraud rate to the audit policy is approximated by the parameter γ that would benefit from empirical insights to its behavior. Further refining the claimant's choice to fraud by introducing a measure of risk aversion for the insured and an endogenous deterrence effect in the setting of the optimal economic model could also be interesting developments.

References

Arrow, K. J. (1964). "The Role of Securities in the Optimal Allocation of Risk-bearing", pp. 91–96. ISSN: 00346527, 1467937X.

– (1974). "Optimal insurance and generalized deductibles". *Scandinavian Actuarial Journal* 1, pp. 1–42. DOI: 10.1080/03461238.1974.10408659.

Bahnsen, A. Correa, D. Aouada, A. Stojanovic, et al. (2016). "Feature engineering strategies for credit card fraud detection". *Expert Systems with Applications* 51, pp. 134–142. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2015.12.030.

Banulescu-Radu, D. and Y. Kougblenou (2024). "Data science for insurance fraud detection: a review". In: *Handbook of Insurance*. Springer. Chap. Forthcoming.

Barcaroli, G. (2014). ""SamplingStrata: An R Package for the Optimization of Stratified Sampling"". In: vol. "61". "4", "1–24". DOI: "10.18637/jss.v061.i04".

Belhadji, E., G. Dionne, and F. Tarkhani (2000). "A Model for the Detection of Insurance Fraud". *The Geneva Papers on Risk and Insurance - Issues and Practice* 25, pp. 517–538.

Bodaghiand, A. and B. Teimourpour (2018). "Automobile Insurance Fraud Detection Using Social Network Analysis: Case Studies in Social Networks and Beyond". *Applications of Data Management and Analysis*, pp. 11–16. DOI: 10.1007/978-3-319-95810-1_2.

Brockett, P., R. Derrig, L. Golden, et al. (2002). "Fraud Classification Using Principal Component Analysis of RIDITs". *Journal of Risk and Insurance* 69, pp. 341–371. DOI: 10.1111/1539-6975.00027.

Chen, T. and C. Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". *CoRR* abs/1603.02754. DOI: 10.1145/2939672.2939785. arXiv: 1603.02754.

Crocker, K. J. and J. Morgan (1998). "Is Honesty the Best Policy? Curtailing Insurance Fraud through Optimal Incentive Contracts". *Journal of Political Economy* 106.2, pp. 355–375. ISSN: 00223808, 1537534X.

Dionne, G. and E. Belhadji (1996). "Évaluation de la fraude à l'assurance automobile au Québec". *Assurance* 64, pp. 365–394.

Dionne, G. and R. Gagné (2001). "Deductible Contracts against Fraudulent Claims: Evidence from Automobile Insurance". *The Review of Economics and Statistics* 83.2, pp. 290–301. ISSN: 00346535, 15309142.

Dionne, G., F. Giuliano, and P. Picard (2009). "Optimal Auditing With Scoring: Theory and Application to Insurance Fraud". *Management Science* 55, pp. 58–70. DOI: 10.1287/mnsc.1080.0905.

Elkan, C. (2001). "The Foundations of Cost-Sensitive Learning". *Proceedings of the Seventeenth International Conference on Artificial Intelligence: 4-10 August 2001; Seattle* 1.

Gollier, C. (1987). "Pareto-optimal risk sharing with fixed costs per claim". *Scandinavian Actuarial Journal* 1987.1-2, pp. 62–73. DOI: 10.1080/03461238.1987.10413818.

Goutte, C. and E. Gaussier (2005). "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation". *Lecture Notes in Computer Science* 3408, pp. 345–359. DOI: 10.1007/978-3-540-31865-1_25.

Grossman, S. J. and O. D. Hart (1983). "An Analysis of the Principal-Agent Problem". *Econometrica* 51.1, pp. 7–45. ISSN: 00129682, 14680262.

Holmstrom, B. (1979). "Moral Hazard and Observability". *The Bell Journal of Economics* 10, pp. 74–91.

Höppner, S., B. Baesens, W. Verbeke, et al. (2022). "Instance-dependent cost-sensitive learning for detecting transfer fraud". *European Journal of Operational Research* 297.1, pp. 291–300. ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2021.05.028.

Jiang, X., M. Osl, Jihoon J. Kim, et al. (2011). "Smooth Isotonic Regression: A New Method to Calibrate Predictive Models". *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science* 2011, pp. 16–20.

Lundberg, S., G.G. Erion, and S.I Lee (2018). "Consistent Individualized Feature Attribution for Tree Ensembles". *CoRR* abs/1802.03888. arXiv: 1802.03888. URL: http://arxiv.org/abs/1802.03888.

Lundberg, S. and S.I. Lee (2017). "A unified approach to interpreting model predictions". *CoRR* abs/1705.07874. arXiv: 1705.07874. URL: http://arxiv.org/abs/1705.07874.

Majhi, S., S. Bhatachharya, R. Pradhan, et al. (2019). "Fuzzy clustering using salp swarm algorithm for automobile insurance fraud detection". *Journal of Intelligent & Fuzzy Systems* 36, pp. 2333–2344. DOI: 10.3233/JIFS-169944.

Mirrlees, J. (1976). "The Optimal Structure of Authority and Incentives Within an Organization". *Bell Journal of Economics* 7, pp. 105–131. DOI: 10.2307/3003192.

Mookherjee, D. and I. Png (1989). "Optimal Auditing, Insurance, and Redistribution". *The Quarterly Journal of Economics* 104.2, pp. 399–415.

Nian, K., H. Zhang, A. Tayal, et al. (2016). "Auto insurance fraud detection using unsupervised spectral ranking for anomaly". *The Journal of Finance and Data Science* 2.1, pp. 58–75. DOI: 10.1016/j.jfds.2016.03.001.

Pesantez-Narvaez, J., M. Guillen, and M. Alcañiz (2019). "Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression". *Risks* 7.2. ISSN: 2227-9091. DOI: 10.3390/risks7020070.

Picard, P. (2013). "Economic analysis of insurance fraud". *Handbook of Insurance: Second Edition*, pp. 349–395. DOI: 10.1007/978-1-4614-0155-1_13.

Raviv, A. (1979). "The Design of an Optimal Insurance Policy". *The American Economic Review* 69.1, pp. 84–96. ISSN: 00028282.

Subelj, L., S. Furlan, and M. Bajec (2011). "An expert system for detecting automobile insurance fraud using social network analysis". *Expert Systems with Applications* 38.1, pp. 1039–1052. DOI: 10.1016/j.eswa.2010.07.143.

Sun, Y., M. S. Kamel, Andrew K.C. Wong, et al. (2007). "Cost-sensitive boosting for classification of imbalanced data". *Pattern Recognition* 40.12, pp. 3358–3378. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2007.04.009.

Townsend, R. M. (1979). "Optimal contracts and competitive markets with costly state verification". *Journal of Economic Theory* 21.2, pp. 265–293. ISSN: 0022-0531. DOI: 10.1016/0022-0531(79)90031-0.

Viaene, S., R. Derrig, B. Baesens, et al. (2002). "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection". *Journal of Risk and Insurance* 69, pp. 373–421. DOI: 10.1111/1539-6975.00023.

Viaene, S., R. Derrig, and G. Dedene (2004). "A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis". *Knowledge and Data Engineering, IEEE Transactions on* 16(5), pp. 612–620. DOI: 10.1109/TKDE.2004.1277822.

Weisberg, H. and R. Derrig (1998). "Quantitative methods for detecting fraudulent automobile bodily injury claims". *Risques* 35, pp. 75–101.

Weisberg, H. and R. Derrig. (1991). "Fraud and Automobile Insurance: A Report on Bodily Injury Liability Claims in Massachusetts". *Journal of Insurance Regulation* 9.4, p. 497.

Appendices

Appendix A

Table 8 presents the main predictors extracted from the base classifiers. The first column introduces the name of the feature. Features marked with (✓) are among the top five predictors for the corresponding base classifiers. Features marked with ✓are among the top eight predictors while features marked with × are not in the top eight predicting variables. The column *Derrig* indicates that the main predictor belongs to the traditional red-flags for body injury claims listed by Weisberg and Derrig (1998). A boxed reference (*ACC*8) indicates the red-flags belong to the core list indicators selected by Weisberg and Derrig (1998). A regular reference (CLT5) indicates that the corresponding indicator was tested in Weisberg and Derrig (1998)[30].

[30]Annex B, table 9

Table 8: Optimal Audit Model results

VarName	Tree	Dart	Line	Logis	FOLogis	FODart	FOTree	FOLine	NLLogis	NLDart	NLLine	NLTree	Derrig
Law_rep_none	✓	✓	✓	✓	✓	✓	✓	✓	NA	NA	NA	NA	CLT5
DTd_dol_Oicd	✓	✓	✓	×	×	✓	✓	✓	×	✓	✓	✓	INS7
DRV_licence_yrs	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	NA
ToL	✓	×	×	×	×	×	✓	×	×	×	×	✓	$ACC8$
HR_Unknown	✓	×	✓	×	×	×	✓	✓	×	✓	✓	✓	$INJ1$
Nbr_Lawyers	✓	✓	✓	✓	✓	✓	✓	✓	NA	NA	NA	NA	NA
Law_rep_agrr	✓	✓	×	✓	✓	✓	✓	×	NA	NA	NA	NA	CLT5
DTd_los_law	✓	✓	×	✓	✓	✓	✓	×	NA	NA	NA	NA	CLT1
Shop_Cat_Others	×	✓	×	×	×	×	×	×	×	✓	×	×	$CLT2$
Law_Rep_Mod	×	×	×	×	✓	✓	×	×	NA	NA	NA	NA	CLT5
Nbr_Claims_Pol	×	×	×	×	×	✓	×	×	✓	×	×	×	$ACC8$
Unk_ToL	×	×	✓	×	×	×	×	✓	×	×	✓	×	NA
Min_injsum_rat	×	×	✓	×	×	×	×	×	×	×	✓	×	$INJ1$
Sum_injmax	×	×	✓	×	×	×	×	✓	×	×	✓	×	$INJ1$
Min_injmin	×	×	×	×	×	×	×	✓	×	×	✓	×	$INJ1$
TotPartCad_rat	×	×	×	×	×	×	×	×	×	×	×	✓	$ACC10$
Inj_Back	×	✓	×	✓	×	×	×	×	✓	✓	✓	✓	$INJ1$
HR_Other	×	×	×	×	×	×	×	×	✓	×	×	×	NA
HR_247	×	×	×	×	×	×	×	×	✓	×	×	×	NA
HR_fax	×	×	×	×	×	×	×	×	✓	×	×	×	NA
HR_phone	×	×	×	✓	×	×	×	×	✓	×	×	×	NA
MostSeverePsyMax	×	×	×	×	×	×	×	×	×	✓	×	×	$INJ2$
DTd_FAB_DoL	×	×	×	×	×	×	×	×	×	✓	×	×	$INJ6$
Nbr_Days_Law_Miss	×	×	×	✓	✓	×	×	×	×	×	×	×	CLT1
male_clmnt	×	×	×	×	×	×	×	×	×	✓	×	×	NA

Appendix B

Table 9: List of the indicators used by Weisberg and Derrig (1998). The column *Suspicious* and *Fraud* indicate the indicator stood-out respectively for the adjusters or the investigators. The *In study* column indicates the name of the corresponding variable created to match the indicator in this study.

Code	Indicators	Suspicious	Fraud	In study
ACC1	No report by police officer at scene	Yes	Yes	Yes
ACC2	No witnesses to accident	Yes	No	No
ACC3	Rear-end collision	No	No	Yes
ACC4	Single vehicle accident	No	No	Yes
ACC5	Controlled intersection collision	No	No	Yes
ACC6	Claimant was in parked vehicle	No	No	Yes
ACC7	Two drivers were related or friends	No	No	No
ACC8	Late-night accident	No	Yes	Yes
ACC9	No plausible explanation for accident	Yes	Yes	Yes[41]
ACC10	Claimant in and old, low-value vehicle	Yes	No	Yes
ACC11	A rental vehicle involved in accident	No	No	No [44]
ACC12	No tow from scene despite severely damaged car	No	No	Yes
ACC13	Site investigation raised questions	No	No	No [47]
ACC14	Property damage was inconsistent with accident	No	No	Yes
ACC15	Very minor impact collision	Yes	No	Yes
ACC16	Claimant vehicle stopped short	No	Yes	Yes
ACC17	Claimant vehicle made unexpected maneuver	No	No	Yes
ACC18	Insured/claimant versions differ	No	No	No

Continued on next page

Table 9: List of the indicators used by Weisberg and Derrig (1998). The column *Suspicious* and *Fraud* indicate the indicator stood-out respectively for the adjusters or the investigators. The *In study* column indicates the name of the corresponding variable created to match the indicator in this study. (Continued)

Code	Indicators	Suspicious	Fraud	In study
ACC19	Insured felt set up, denied fault	Yes	No	No
CLT1	Retained an attorney very quickly	No	No	Yes
CLT2	Had a history of previous claims	No	No	Yes
CLT3	Gave address as hotel or PO Box	No	No	No
CLT4	Was an out-of-state resident	No	No	No
CLT5	Retained a "high-volume" attorney	No	No	Yes
CLT6	Was difficult to contact/uncooperative	No	No	No
CLT7	Was one of three or more claimants in vehicle	No	Yes	Yes
CLT8	Was resident of high claim town	No	No	No
CLT9	Avoided use of telephone or mail	No	No	No
CLT10	Was unemployed	No	No	Yes
CLT11	Appeared to be "claims-wise"	Yes	Yes	No
INS1	Had a history of previous claims	No	No	Yes
INS2	Gave address as hotel or PO Box	No	No	Yes
INS3	Readily accepted fault for accident	No	No	No
INS4	Was acquainted with other-vehicle occupants	No	No	No
INS5	Was not willing to provide a sworn statement	No	No	No
INS6	Was difficult to contact/uncooperative	Yes	No	P-F
INS7	Accident occurred soon after policy effective date	No	No	Yes
INS8	Appeared to be "claims-wise	No	No	No

Continued on next page

Table 9: List of the indicators used by Weisberg and Derrig (1998). The column *Suspicious* and *Fraud* indicate the indicator stood-out respectively for the adjusters or the investigators. The *In study* column indicates the name of the corresponding variable created to match the indicator in this study. (Continued)

Code	Indicators	Suspicious	Fraud	In study
INJ1	Injury consisted of strain/sprain only	No	Yes	No
INJ2	No objective evidence of injury	Yes	Yes	No
INJ3	Police report showed no injury or pain	No	Yes	No
INJ4	Claimant refused to appear for IME	No	Yes	P-F
INJ5	No emergency treatment was given for the injury	No	No	No
INJ6	Non-emergency treatment was delayed	No	Yes	No
INJ7	First non-emergency treatment was by a DC	No	No	No
INJ8	Activity check cast doubt on injury	No	No	P-F
INJ9	Injuries were inconsistent with police report	Yes	No	No
INJ10	IME suggests injury was unrelated to accident	No	No	No
INJ11	Unusual injury for this auto accident	No	No	No
INJ12	Evidence of an alternative cause of injury	No	Yes	No
TRT1	Large number of visits to a chiropractor	Yes	No	P-F
TRT2	DC provided 3 or more modalities on most visits	Yes	No	P-F
TRT3	Large number of visits to a physical therapist	No	No	P-F
TRT4	MRI or CT scan but no inpatient hospital charges	No	No	P-F
TRT5	Use of "high volume" medical provider	No	No	Yes
TRT6	Significant gaps in course of treatment	No	No	P-F
TRT8	Treatment was unusually prolonged	No	No	P-F
TRT7	IME questioned extent of treatment	No	No	P-F

Continued on next page

Table 9: List of the indicators used by Weisberg and Derrig (1998). The column *Suspicious* and *Fraud* indicate the indicator stood-out respectively for the adjusters or the investigators. The *In study* column indicates the name of the corresponding variable created to match the indicator in this study. (Continued)

Code	Indicators	Suspicious	Fraud	In study
TRT9	Medical audit raised questions about charges	No	No	P-F
LW1	Claimant worked for self or family member	No	Yes	Yes
LW2	Employer wage differs from claimed wage loss	No	No	P-F
LW3	Claimant recently started employment	No	No	No
LW4	Employer unknown/hard to reach	No	No	No
LW5	Lost wages statement looked unofficial	No	No	P-F
LW6	Long disability for a minor injury	Yes	No	P-F

Appendix C

Table 10: List of variables extracted before preprocessing and feature engineering.

Variable Names			
Date Variables			
Date_of_Loss	First_AB_dt	First_PD_dt	First_Lawyer_dt
Last_Move_dt	PD_Accepted_Date	original_inception	
Claimant and Policy Characteristics			
female_clmnt	male_clmnt	Nbr_Claims_on_pol	age_claimant_max
Nbr_Drivers_on_prem_qty	Drv_Licence_Years	Endt_AB_Chg_Dt	Chg_AB_end_ct
Claimants Education			
Edu_Less_High_School	Edu_College	Edu_High_School	Edu_Missing
Edu_University			
Claimants Employment			
Occ_Retired	Occ_Middle_Mgmt	Occ_Clerical	Occ_Sales
Occ_Senior_Mgmt	Occ_Truck_Driv	Occ_Machine_Op	Occ_Mecanic

Continued on next page

Table 10: List of variables extracted before preprocessing and feature engineering. (Continued)

Variable Names			
Occ_Const_Trades	Occ_Labour	Occ_Dentist	Occ_Physician
Occ_Teacher	Occ_Engineer	Occ_Farmer	Occ_Fire_Fighter
Occ_Homeperson	Occ_Ins_Adjuster	Occ_Lawyer	Occ_Nurse
Occ_Secretary	Occ_Psycho	Occ_Police	Occ_Ins_Agent
Occ_Retail	Occ_Self_Employed	Occ_ft_Student	Occ_Other
Occ_Unknown	Occ_Proff	Occ_Child	
Law variables			
Nbr_Lawyer	law_rep_agrr	law_rep_mod	law_rep_none_pass
Accident characteristics			
Police_Report_Exists	Cause_of_Loss	How_Reported	POIPrimary[53]
Annual_Km_Qty	Nbr_Claimants	Nbr_Claimants_AB	Nbr_Towing
Highway	Time_of_Loss	Loss_Province_cd	LossType
Insd_Fault_%	Shop_Category	RepairerStateProvince	Nbr_Third_Party
Nbr_Driver			

Continued on next page

Table 10: List of variables extracted before preprocessing and feature engineering. (Continued)

Variable Names			
Vehicle and damages			
TotalReplaceLaborCAD	TotalLaborCAD	TotalRepairLaborCAD	TotalPartCAD
Veh_Age	Vehicle_Category	Vehicle_Type	Drivable_Audatex
totalloss	Vehicle_Price		
Injury variables			
Nbr_Injury_Max	sum_injsum	psy	phys
Driver_Injured	Insured_Injured	Pre_Physical	Pre_Psycho
Days_Before_Opened_Psy_Max	Days_Before_Opened_Phys_Max	MostSeverePhysMIN	MostSeverePsyMIN
MostSeverePsyMAX	MostSeverePhysMAX	MinPsy	MinPhys
MaxPhys	MaxPsy	max_injmax	min_injmax
sum_injmax	nb_injmax	max_injmin	min_injmin
sum_injmin	nb_injmin	max_injsum	min_injsum
nb_injsum	Nbr_Injury	Inj_Back	Inj_Chest
Inj_Neck	Inj_Shoulder	Inj_Head	Inj_Other

Continued on next page

Table 10: List of variables extracted before preprocessing and feature engineering. (Continued)

Variable Names			
Reserve and disbursement variables			
First_Dart_Sum	First_Dart_Max	Res_at_AB_open	Res_TenDays_AB

The 10 groups are: *Date* comprised of 7 variables that control the difference in days between the date of loss and the inception of the policy, the first AB exposure, the first Property-Damage (PD) exposure and the date of the first lawyer to appear on the claims.

The second type of variables relates to some characteristics of the policy or the claimants. It is comprised of 8 variables, that control for the number of male and female claimants, the number of claims on the policy, the oldest claimant in the claim, the number of drivers allowed on the policy, the number of years of driving experience the main beneficiary of the contract posses and the number of recent changes in the AB coverage choice.

The third and fourth types are related respectively to the level of education of the main beneficiary and his type of employment. The fifth type is comprised of variables pertaining to the type of law firm involved on the claim. Such variables track the level of *"aggressiveness"* with which law firms specialised in insurance claims disputes settle the case.

The sixth family of variables controls for some specific accident characteristics such as the type of accident ("rear-end","improper-left-turn", etc...), the manner with which the accident was reported to the insurance company, the point of impact on the insured car and other characteristics such as the body shop chosen by the claimant. The seventh type of variables records the severity of the damages and some indications on the vehicle involved in the crash. It has 10 distinct variables. The first four accounts for the

estimated cost of labor and the cost of replacing the various part of the vehicle. The remaining 6 variables relate to the type, age, price and drivability of the vehicle. The two next families of variables control respectively for the severity of injury sustained in the accident and the estimated financial cost placed in reserve by the insurance company.

The variable *Cause_of_Loss*, from the **Accident characteristics** family, controls the reason behind the accident: *"rear-end accident"*, 36% of observations, *"improper-left-turn"*, 9.5%, *"Bicyclist/pedestrian"*, etc. From the same family, *How_reported* controls for the means by which our Industrial Partner was made aware of the incident. *POIPrimary* describes the primary point of impact on the claimant's vehicle while *Shop_Category* identifies the level of previous collaboration with the garage associated with the accident.

The family **Injury variables** depicts the severity of the injuries. There are two types of injuries: psychological and physical. 62000 claims have at a least one listed physical injuries and only 14000 claims have at least one psychological injuries. Injuries are listed by person, body location and trauma severity as reported after a medical examination. As claims are aggregated at the contract level, there might be multiple injured per claim. Therefore, variables such as *min_injmax* or *max_injsum* refer to the least sever injury respectively (*min_*) sustained by the maximum injured, or the most injured claimant (inj*max*), and to the sum of the all most sever injuries sustained on the claim. The same logic applies to all similar variables. *Pre_Psycho* and *Pre_Physical* indicates if there are any physical or psychological injuries listed for the policy holder of the contract while *psy* and *phys* control for the number of psychological and physical injuries on the claim. The family **Reserve and disbursement variables** control for the dollars put in reserve following our industrial partner actuarial predicted cost of treatment: *Res_at_AB_open* and such estimate after ten days. Disbursed amounts refer to amount spent on two specified policies: *amp* and *air*. Policy *amp* refers to medical payments and policy *air* refers to income replacement. As many people may be injured on the same claim, the prefix *max/min* refers again to the average amount paid for the maximum/minimum injured person on the claim. The prefix *freq* refers to the frequency of payments

so far and the suffix number corresponds to the number of days at which the snapshot has been taken. We only limit ourselves up to fifteen days after the date of loss to avoid capturing post-fact information.

A number of variables can be categorized as global. Variables, such as *sum_injsum*, *Res_at_AB_open*, *phys* and *psy* or the variables capturing the severity of the accident such as *TotalReplaceLaborCAD* give an overview of the claim by controlling for the general severity and number of injures, the estimated global cost of the AB side and the severity of the damages sustained by the vehicle.

Appendix D

Appendix D focuses on analysing the features composing the final database. Features are analysed following two categories: continuous or categorical. The final database before one-hot encoding is composed of 70412 observations with 126 feature variables. After one-hot encoding the factor variables, the number of feature variables rises to 207.

This section analyses the categorical variables at value level present in the database. There are in total a 62 categorical variables, and after one-hot encoding, a total of 426 level-variables of which 345 have investigated/fraud observations. Each level-variable forms a subset of the general population. In this section, each such subset's investigated/fraud rate is compared to the remaining population's investigated/fraud rate through a two-proportion z-test The tested null hypotheses is that the investigated/fraud rate is equal in the sub-sample formed by observations bearing such level and in the remaining population. The z-test formula is as follow:

$$z = \frac{i_{sample} - i_{pop}}{\left(\frac{i_{tot} * n_{tot}}{n_{sample}} + \frac{i_{tot} * n_{tot}}{n_{pop}}\right)^{\frac{1}{2}}} \tag{26}$$

Table 11 below presents the results of the top twenty five levels ranked by decreasing *P.value*[55]. The first column present the variable and the tested level separated by a "_". Columns *"InvesSamp"* and *"InvesPop"* are respectively the percentage of investigated claims belonging to the tested level and the remaining population respectively. The column *conf.low* is the 95% lowest bound of the investigation rate for the level population. The *SampPop* is the number of claims in the level. The P-value column is not presented because all P-values are below one per thousands. The same analysis is conducted with the different preprocessing of *AB_fraud_label* where all *"Investigated"* claims are considered as *"No_label"*. The columns *FraudSamp*, *FraudPop*, *FraudConf.low* are similar to the one

[55]There are respectively 114 and 141 levels with investigation or fraud rate higher than the remaining population at the 1% level

described before but for the subset of *"Fraud"* claims.

As an example the first line of Table 11 shows: *"law_rep_agrr_1"* which refers to the variable *"aggressive law firm is on the claim"* with "_1", indicating that one such law firm is on the claim. There are 2546 claims with such level and this subset has an investigation rate of 33%. The remaining population of 67866 has an investigated rate of above 4%. With a 95% confidence interval, the lowest estimate of the investigated rate for the subset of claims with strictly one aggressive law firm is 27%.

Table 11 shows the similarities between the sets of significant levels in predicting *"Investigated"* or *"Fraud"*. Out of the set of 25 variables, 19 are common between the two sets. Table 11 also illustrates the importance taken by law variables with eleven out of the main 25 levels for both preprocessing. Among these law variables, one or more aggressive law firm being present on a claim seems to indicate a higher investigation or fraud. Observations with multiple back or neck injuries are also associated with higher investigation and/or fraud rate. Similarly observations with multiple male claimants and no female claimants have a significant higher investigated and fraud rate.

The variable *ShopCat* controls for the type of body shop used in the claim. **ShopCat_other** refers to body shops that do not belong to the rely network of our industrial partner. This level is ranked forty-one with a 1% significance when it concerns fraud and is therefore more significant for predicting investigation. The variable *OCC_HomePerson* controls how many claimants are not working. Observations with such characteristics have a significantly higher investigating rate than the remaining population. The fraud rate difference is also significant at the 1% level but is ranked forty-nine and is out of this table scope. The variable *OCC_SelfEmployed* has an opposite pattern: both difference in fraud and investigated rate are significant at the 1% level but the ranking P.value is lower for fraud than for investigated claims. Claims with multiple accident benefit claimants, presence of male claimants and absence of female claimants all have significant difference in term of investigated rate compare to the general population.

Table 11: Analysis of categorical variables

Index	VarName	InvesSamp	InvesPop	InvesConf.low	FraudSamp	FraudPop	FraudConf.low	SampPop
1	law_rep_agrr_1	33.39	4.42	27	22.55	2.38	19	2546
2	law_rep_agrr_2	60.22	5.25	50	40.88	2.96	33	274
3	law_rep_none_pass_0	12.00	1.99	10	7.02	1.02	6	24441
4	NBR_DAYS_LAW_Miss_0	11.71	2.18	9	6.67	1.23	5	24278
5	NBR_LAWYER_1	13.68	3.50	1	7.03	2.17	4	13565
6	law_rep_mod_1	13.06	4.45	8	7.32	2.54	4	8323
7	INJ_BACK_2	14.03	4.99	8	8.43	2.81	5	3678
8	law_rep_mod_2	22.30	5.27	15	14.55	2.97	10	825
9	ShopCat_Other	8.37	4.34	4	-	-	-	19628
10	INJ_BACK_3	26.26	5.32	18	16.36	3.01	10	495
11	law_rep_agrr_3	68.00	5.42	51	44.00	3.08	28	50
12	INJ_OTHER_2	12.84	5.12	7	7.80	2.89	4	3115
13	male_clmnt_2	12.50	5.18	6	6.86	2.95	3	2768
14	law_rep_agrr_NA	23.56	5.35	0.15	-	-	-	433
15	INJ_NECK_2	10.86	5.12	0.05	6.30	2.90	0.03	4209
16	law_rep_none_pass_NA	28.34	5.38	0.18	17.00	3.06	0.10	247
17	OCC_HOMEPERSON_1	10.32	5.09	0.04	-	-	-	5030
18	INJ_OTHER_1	7.13	4.43	0.02	4.28	2.38	0.02	26893

Continued on next page

Table 11: Analysis of categorical variables (Continued)

Index	VarName	InvesSamp	InvesPop	InvesConf.low	FraudSamp	FraudPop	FraudConf.low	SampPop
19	NBR_CLAIMANTS_AB_3	12.62	5.23	0.06	8.00	2.94	0.04	2250
20	NBR_LAWYER_2	9.32	5.05	0.04	6.56	2.73	0.03	6886
21	law_rep_mod_NA	24.22	5.38	0.15	-	-	-	322
22	female_clmnt_0	7.06	4.51	0.02	-	-	-	26404
23	PREXS_Yes	6.84	4.36	0.02	-	-	-	31356
24	INJ_OTHER_3	18.92	5.37	0.11	12.95	3.03	0.07	502
25	PRE_PSYCHO_1	8.69	5.08	0.03	5.08	2.87	0.02	7461
26	OCC_RETIRED_0	5.85	2.39	0.00	0.03	62537		
27	male_clmnt_3	19.07	5.38	0.00	0.10	451		
28	INJ_BACK_1	6.60	4.46	0.00	0.02	33012		
29	male_clmnt_1	6.58	4.48	0.00	0.02	32927		
30	EDU_MISSING_3	16.81	5.37	0.00	0.09	595		
31	INJ_NECK_3	16.58	5.37	0.00	0.09	597		
32	NBR_CLAIMANTS_AB_NA	15.51	5.36	0.00	0.08	722		
33	INJ_BACK_4	31.25	5.42	0.00	0.18	112		
34	INJ_SHOULDER_2	14.50	5.35	0.00	0.07	862		
35	HR_Call_Centre	7.26	4.91	0.00	0.02	16609		
36	law_rep_agrr_4	70.59	5.45	0.00	0.44	17		

Continued on next page

Table 11: Analysis of categorical variables (Continued)

Index	VarName	InvesSamp	InvesPop	InvesConf.low	FraudSamp	FraudPop	FraudConf.low	SampPop
37	EDU_HIGH_SCHOOL_2	11.20	5.31	0.00	0.05	1901		
38	OCC_RETAIL_1	9.59	5.27	0.00	0.03	3161		
39	OCC_SELF_EMPLOYED_1	9.41	5.27	0.00	0.03	3367		
40	OCC_HOMEPERSON_2	16.26	5.40	0.00	0.08	449		
41	INJ_SHOULDER_1	7.13	5.03	0.00	0.02	14682		
42	law_rep_mod_3	26.05	5.43	0.00	0.14	119		
43	PRE_PHYSICAL_2	10.81	5.35	0.00	0.04	1536		
44	INJ_CHEST_0	5.78	3.49	0.00	0.02	60657		
45	EDU_LESS_HIGH_SCHOOL_1	9.03	5.30	0.00	0.03	3199		
46	EDU_HIGH_SCHOOL_1	6.82	5.03	0.00	0.01	17249		
47	NBR_THIRD_PARTY_0	6.00	4.41	0.00	0.01	46597		
48	VehMake_ACURA	11.66	5.37	0.00	0.05	1012		
49	VehCat_Mid__Foreign	7.38	5.18	0.00	0.02	9178		
50	NBR_CLAIMANTS_4	9.66	5.34	0.00	0.03	2030		
51	INJ_SHOULDER_3	26.51	5.44	0.00	0.12	83		
52	NBR_CLAIMANTS_AB_4	12.39	5.40	0.00	0.05	694		
53	HR_Fax	9.60	5.35	0.00	0.03	1886		
54	INJ_HEAD_0	5.74	3.78	0.00	0.02	60554		

Continued on next page

Table 11: Analysis of categorical variables (Continued)

Index	VarName	InvesSamp	InvesPop	InvesConf.low	FraudSamp	FraudPop	FraudConf.low	SampPop
55	VehMake_HONDA	7.72	5.26	0.00	0.02	5801		
56	NBR_CLAIMANTS_AB_2	6.95	5.16	0.00	0.01	11885		
57	VehCat_Unknown	6.52	5.06	0.00	0.01	19645		
58	OCC_OTHER_2	11.54	5.40	0.00	0.04	780		
59	PRE_PSYCHO_2	14.44	5.42	0.00	0.06	360		
60	PhysInjuryMissEv_Evaluation	5.55	0.96	0.00	0.04	69053		
61	NBR_CLAIMANTS_3	7.36	5.26	0.00	0.02	6806		
62	VehMake_LowPoP	6.59	5.12	0.00	0.01	16427		
63	EDU_MISSING_2	7.84	5.32	0.00	0.02	4185		
64	OCC_MACHINE_OP_1	10.10	5.39	0.00	0.03	1139		
65	male_clmnt_4	24.00	5.45	0.00	0.10	75		
66	RepStateProv_O	9.81	5.39	0.00	0.03	1244		
67	EDU_UNIVERSITY_0	5.60	3.22	0.00	0.02	66310		
68	HR_After_Hours_Service	7.41	5.31	0.00	0.01	5315		
69	OCC_UNKNOWN_3	13.62	5.43	0.00	0.05	323		
70	LossProv_CD_ON	5.54	1.55	0.00	0.03	69059		
71	female_clmnt_2	7.58	5.33	0.00	0.02	4290		
72	EDU_MISSING_4	17.73	5.44	0.00	0.07	141		

Continued on next page

Table 11: Analysis of categorical variables (Continued)

Index	VarName	InvesSamp	InvesPop	InvesConf.low	FraudSamp	FraudPop	FraudConf.low	SampPop
73	NBR_CLAIMANTS_5	11.11	5.42	0.00	0.03	558		
74	VehMake_NISSAN	8.26	5.38	0.00	0.02	2178		
75	Insured_INJURED_NA	9.93	5.41	0.00	0.03	826		
76	PRE_PHYSICAL_3	18.45	5.45	0.00	0.06	103		
77	AccidentSeverityEval_Missing	6.47	5.24	0.00	0.01	12808		
78	Drivable_Audatex_Missing	6.47	5.24	0.00	0.01	12808		
79	LossType_Missing	6.47	5.24	0.00	0.01	12808		
80	POI_Missing	6.47	5.24	0.00	0.01	12808		
81	RepStateProv_Missing	6.47	5.24	0.00	0.01	12808		
82	TotLoss_Audatex_Missing	6.47	5.24	0.00	0.01	12808		
83	INJ_NECK_4	16.67	5.44	0.00	0.06	132		
84	Driver_INJURED_NA	9.72	5.41	0.00	0.03	854		
85	INJ_OTHER_4	15.97	5.44	0.00	0.05	144		
86	CLDE_Hit_&_run	9.18	5.41	0.00	0.02	1078		
87	NBR_CLAIMANTS_AB_5	14.86	5.44	0.00	0.05	175		
88	OCC_OTHER_1	6.34	5.25	0.00	0.01	13919		
89	OCC_UNKNOWN_2	7.88	5.39	0.00	0.01	2233		
90	OCC_LABOUR_1	9.53	5.42	0.00	0.02	808		

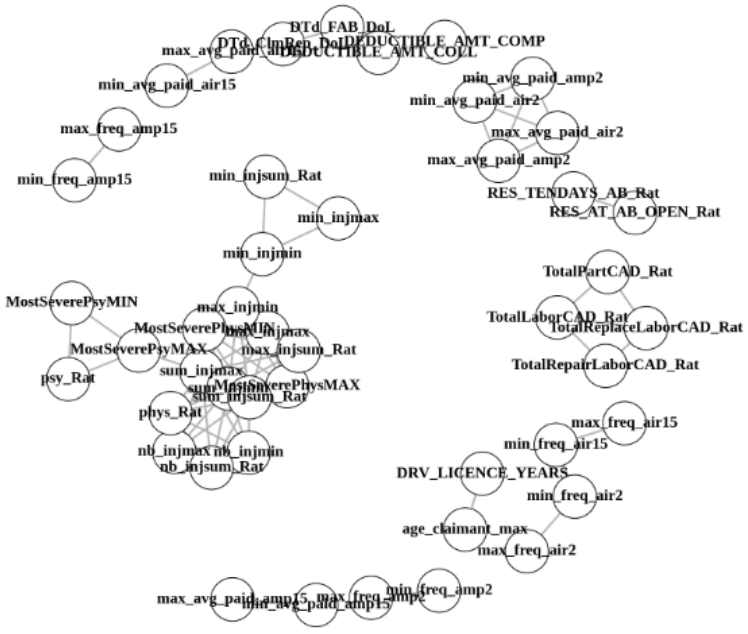


Figure 2: Correlation Network.

Appendix E

Appendix E analyses the 50 continuous variables present in the database. The first part analyses the correlation structure within the feature variables. The second part tests the difference in mean of each continuous variables between subsets of Investigated/Fraud observations and No-label observations.

Figure 2 presents all correlation above a 60% threshold extracted from the correlation matrix in Figure 3. We can clearly see clusters of variables correlated together. The biggest cluster is comprised of variables treating the severity of injuries present on the claim. This cluster cam be further disaggregated into three smaller cluster. The most central one is linked to physical injuries with variables such as *phys_Rat*[56] and *MostSeverePhysMAX*[57] which are strongly linked with global injury indicators such as *max_injsum_Rat*[58] or *sum_injsum_Rat*[59]. This indicates that the global severity of the injuries is driven more by physical injuries rather than psychological ones. To the north of the central cluster, the first smaller cluster of three variables *min_injmin*, *min_injsum_Rat* and *min_injmax* controls for the injuries with smallest gravity. This sub-cluster is linked to the central cluster by the edge between *min_injmin* and *max_injmin* which are respectively the less and most severe injury sustained by the less injured claimant. The third sub-cluster is composed of psychological injuries and is linked to the central cluster by an edge with *sum_injmax*. The variable *sum_injmax* is the sum of all the injuries sustained by the most critically

[56]Number of physical injuries sustained in the claim

[57]The most severe physical injury sustained on the claim

[58]The sum of the maximum injuries sustained by all claimants normalised by the number of AB claimants

[59]The sum of the severity of all the injuries of all the claimants normalised by the number of AB claimants

injured claimant. Psychological variables being most related to *sum_injmax* can be interpreted as that the most severely hurt claimant is most often the one suffering from psychological injuries.

The second main cluster is composed of four variables all related to damages on the car: *TotalPartCAD_Rat*, *TotalLaborCAD_Rat*, *TotalRepairLaborCAD_Rat*, *TotalPartCAD_Rat*, normalised by the price of the vehicle. The final cluster of interest shows the relation between the variables *Dtd_FAB_DoL* and *Dtd_ClmRep* which control for the difference in days between the first AB elements and the date of loss and for the difference in days between the first report of the claim and the date of loss. The high correlation between the two suggests that few claims have property damage elements opened long before the accident benefits elements.

Table 12 shows the results of student-test between the mean of investigated and no label claims. The null hypothesis being that the sample mean, the mean of investigated claims, is equal to the population mean, the mean of the no label claims. The results are presented in Table 12. Variables are ordered by increasing P-values. There are 45 significant variables at the 1% levels and 11 variables not significant at the 5% levels.

Table 12: Analysis of continuous variables

Index	VarName	Pop_Mean	Pop_Sd	Sample_Mean	Sample_Sd	p.value
1	Veh_Age	7.59	5.14	8.24	5.03	0.00
4	TIME_TO_EXPIRY	189.56	105.54	195.83	105.74	0.00
5	nb_injmin	2.05	1.28	2.53	1.42	0.00
6	MostSeverePhysMAX	0.119	0.0638	0.13	0.0553	0.00
7	max_injmax	0.125	0.0681	0.147	0.067	0.00
8	min_injmax	0.0852	0.0439	0.0906	0.0356	0.00
9	sum_injmax	0.25	0.19	0.35	0.21	0.00
10	nb_injmax	2.23	1.29	2.94	1.38	0.00
13	NBR_CLAIMS_ON_POL	2.12	2.29	2.31	2.88	0.00
14	age_claimant_max	45.76	19.13	42.07	17.65	0.00
15	DAYS_BEFORE_OPENED_PHYS_MAX	65.64	144.48	101.39	178.52	0.00
19	ToL	505.97	468.64	603.13	480.66	0.00
22	phys_Rat	1.86	1.06	2.12	1.05	0.00
23	psy_Rat	0.23	0.55	0.58	0.77	0.00
24	RES_AT_AB_OPEN_Rat	9823.52	15596.60	8238.07	9724.99	0.00
25	RES_TENDAYS_AB_Rat	6935.84	15386.09	6013.56	8603.30	0.00
31	TotalLaborCAD_Rat	0.09	0.10	0.07	0.09	0.00
34	DTd_Los_LAW	1.76	6.34	6.85	10.84	0.00

Continued on next page

Table 12: Analysis of continuous variables (Continued)

Index	VarName	Pop_Mean	Pop_Sd	Sample_Mean	Sample_Sd	p.value
2	DEDUCTIBLE_AMT_COLL	372.99	360.60	325.20	409.46	0.00
3	FTP_AB	690.70	2857.56	1084.92	864.15	0.00
11	max_injmin	0.12	0.07	0.13	0.07	0.00
12	sum_injmin	0.23	0.19	0.29	0.21	0.00
16	MostSeverePhysMIN	0.11	0.06	0.12	0.05	0.00
17	min_avg_paid_amp15	31.72	1102.04	8.16	176.57	0.00
18	max_avg_paid_amp15	32.37	1109.45	8.32	177.25	0.00
20	DRV_LICENCE_YEARS	23.40	15.73	16.19	12.61	0.00
21	DAYS_BEFORE_PD	16.36	152.25	5.34	30.77	0.00
26	nb_injsum_Rat	2.09	1.24	2.69	1.33	0.00
27	sum_injsum_Rat	0.23	0.18	0.32	0.20	0.00
28	min_injsum_Rat	0.08	0.04	0.08	0.04	0.00
29	max_injsum_Rat	0.12	0.07	0.14	0.07	0.00
30	TotalPartCAD_Rat	0.13	0.18	0.10	0.15	0.00
32	TotalReplaceLaborCAD_Rat	0.07	0.08	0.05	0.07	0.00
33	TotalRepairLaborCAD_Rat	0.02	0.02	0.02	0.02	0.00
37	DTd_DoL_Oicd	2744.76	2682.88	1407.94	1844.30	0.00
38	DTd_FAB_DoL	40.63	234.52	25.23	75.41	0.00

Continued on next page

Table 12: Analysis of continuous variables (Continued)

Index	VarName	Pop_Mean	Pop_Sd	Sample_Mean	Sample_Sd	p.value
39	MostSeverePsyMIN	0.03	0.06	0.07	0.09	0.00
40	MostSeverePsyMAX	0.03	0.07	0.07	0.09	0.00
41	min_avg_paid_amp2	9.34	446.36	2.84	103.41	0.01
42	min_injmin	0.08	0.04	0.08	0.04	0.01
43	max_avg_paid_amp2	9.35	446.37	2.99	104.56	0.01
44	DEDUCTIBLE_AMT_COMP	317.28	317.66	300.15	383.10	0.01
45	max_avg_paid_air15	20.31	1497.55	5.36	97.86	0.01
46	min_avg_paid_air15	20.17	1497.04	5.36	97.86	0.01

From Table 12, all variables in controlling for the severity of injuries are significant at the 1% level significance. Among the most significant of those injury variables are these that relates to the most severely injured person present in position seven through ten. The most injured persons on investigated claims have on average more serious and more numerous injuries. It is however interesting to put the level of severity of injuries with the amount put in reserve. At position 24 and 25, variables *Res_ab_Open_Rat* show that investigated claims have on average a reserve 6000 lower than the reserve of not investigated claims. The difference at ten days is 9000. Reserve amount are calculated to predict the total amount a claim will cost to our industrial partner using severity of injuries and other confidential inputs. Table 12 shows a discrepancy between the level of injury severity and the amount in reserve for investigated claims with on average stronger injuries but lower estimate of final cost.

Topping the significant list is the age of the vehicle on the policy. The older the vehicle is the most investigated the claim is which is consistent with discussion with our industrial partner's analysis. The variable *ToL* for time of loss shows investigated claims occurring on average a hundred minutes later in the day compare to none investigated claims. Similarly investigated claims are on average earlier on their contract's life by six days as shown by variable *Time_to_Expiry*, and have a history of previous claims on the contract as displayed by the variable *NBR_CLAIMS_ON_POL*. Table 12 also shows that the oldest claimant is on average three years younger for investigated claims.

Apart from the three main clusters presented above, most other clusters link variables of similar type together such as frequencies and amount disbursed by policies: *"max_avg_paid_"* or *"max_freq_amp_"* amounts in reserve normalised by the number of AB claimants when the AB side of the claim is opened and ten days later, *Res_at_AB_open_Rat* and *Ten-Days_Rat* or the maximum age of the claimant and the number of years the policy holders has had his licence. A cluster of interest consists in *max/min_avg_paid_amp/air_2* controlling respectively disbursed amounts in special medical expenditures and income replacement two days after the date the observation was reported. All four variables are correlated

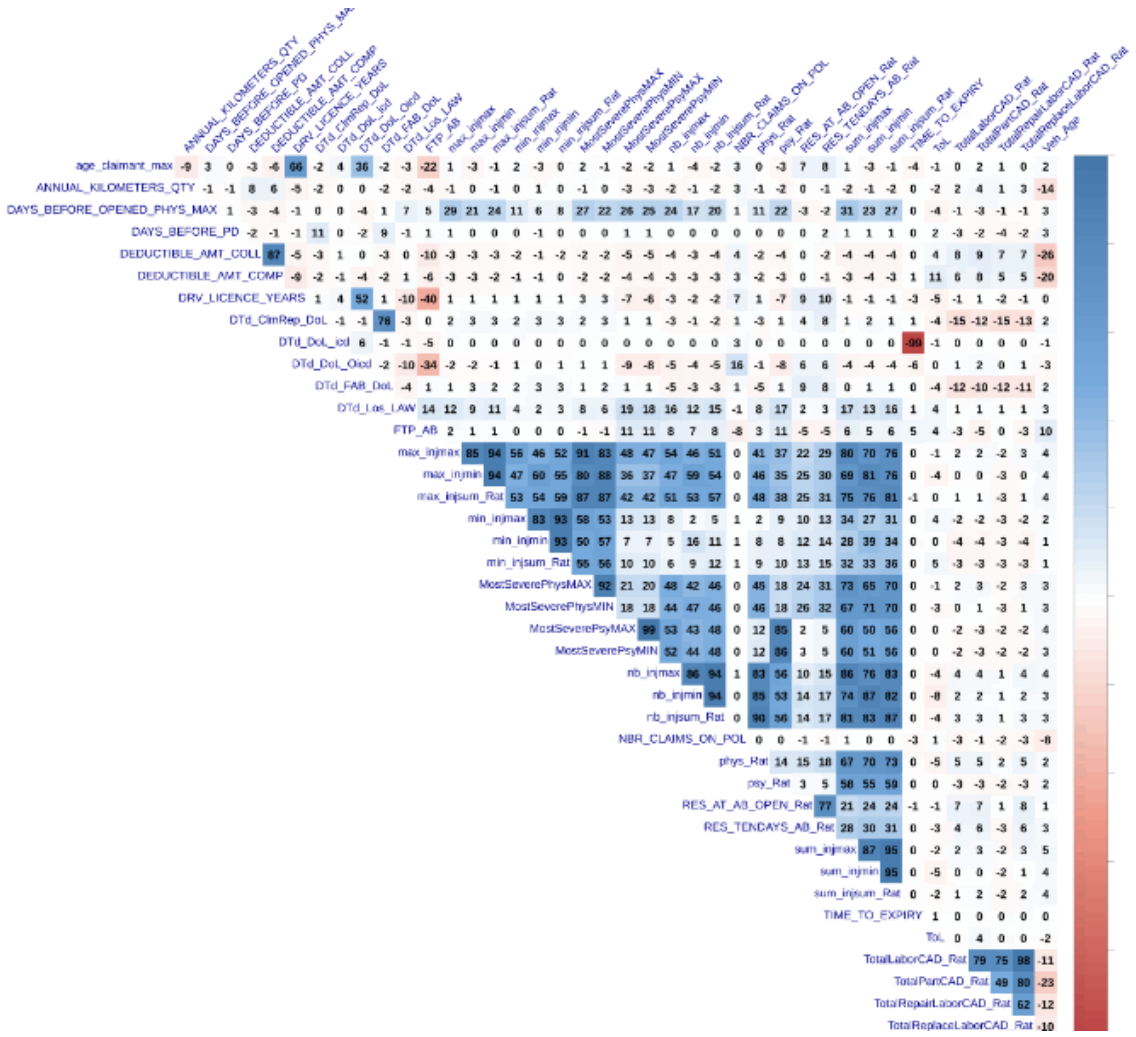


Figure 3: Full Correlation Matrix

above the 60% level, indicating than early income replacement is often accompanied by extra investments in medical expenditures.

Variable *Veh_Age* is negatively correlated at about 20% to deductible amounts paid for collision and comprehensive coverage as well as to all variables related to car damages and annual mileage. Accident Benefit premium is negatively correlated at around 35% to the number of years with driving licence as well as the number of years the claimant has been insured with our industrial partner and by extension at about 20% to the claimant's age. Variables *Res_at_AB_open_Rat* and *Res_TenDays_AB_Rat*, controlling for the amount of dollars in reserve respectively when the injury is declared and after ten days, are positively correlated at around 20 % and 30% to the variables controlling for the global health state of the most and the least severely hurt claimant as well as the sum of all injuries on the claim normalised by the number of claimants. These correlations were to be expected and represent a sign of the quality of the data.

Interestingly, the variable *Days_Before_Open_Phys_Max* is positively correlated at around the 25% level to most injury severity indicators. This indicates that the latest the most severe physical injury is declared, the higher the severity of such injury is. Similarly the variable *Dtd_DoL_Law* controlling for the number of days between the presence of lawyer and the date of loss is positively correlated at the 15% level with most of the severity of injury indicators. The variable *Dtd_DoL_Law* is censored after thirty-six days. This correlation can not distinguish if the level of severity of injuries increases after a lawyer appears or serious injury lead to lawyers arriving later in the claims but still before thirty-six days.

Appendix F

Chen and Guestrin (2016) also use equation 8 to evaluate a potential split for a tree by calculating the difference in $\hat{\imath}^t(q)$ between the original tree and the original tree after the split. The authors obtain equation 27 where I_L and I_R are instances in respectively the left and right nodes after a split.

$$\iota_{split} = \frac{1}{2}\left[\Sigma_{j=1}^{T}\frac{\Sigma_{i\in I_L} g_i}{\Sigma_{i\in I_L} h_i + \lambda} + \Sigma_{j=1}^{T}\frac{\Sigma_{i\in I_R} g_i}{\Sigma_{i\in I_R} h_i + \lambda} - \Sigma_{j=1}^{T}\frac{\Sigma_{i\in I} g_i}{\Sigma_{i\in I} h_i + \lambda}\right] + \gamma T \qquad (27)$$

As it is impossible to list all tree structures $q(x)$, XGBoost starts from a single leaf and construct a tree base-learner by selecting a variable k from a random sub-sample V of all the feature variables available using (27). The authors offer two different methods to select the best value for a potential split among all variables k in the sub-sample V. The first method is called the *Exact Greedy Algorithm for Split Finding* which is impractical for many real life applications. The second method, used here, is called the *Approximate Algorithm for Split Finding.*

Approximate Algorithm for Split Finding consists in calculating weighted quantiles for each variables of the sub-sample. Using the notation from Chen and Guestrin (2016), we define $D_k = \{(x_{1k}, h_1), (x_{2k}, h_2), ..., (x_{nk}, h_n)\}$ the ordered tuple over feature variable k of each observation value paired with the corresponding second derivative. (28) is the rank function that weights each quantile by the value of the second derivative h.

$$r_k(z) = \frac{1}{\Sigma_{(x,h)\in D_k}(h)}\Sigma_{(x,h)\in D_k, x<z}(h) \qquad (28)$$

Potential split points $\{s_{k1}, s_{k2}, ..., s_{kl}\}$, for each feature k are selected following a criteria dependent on ϵ, an hyper parameter regulating the XGBoost algorithm as described in (29).

$$r_k(s_{k,j})r_k(s_{k,j+1}) \leq \epsilon, s_{k1} = min_i(x_{ik}), s_{kl} = max_i(x_{ik}) \tag{29}$$

Once a potential split point s_k1 for each feature v is selected, (27) is used to evaluate the $\{k, s_{ki}\}$ combination that most decreases Equation (2), where k refers to the feature variable chosen and s_{ki} the value upon which the split is done. Chen and Guestrin (2016) finally address issues related to multiple potential optimal split and implementation as well as sparsity robust trees that we leave to the curious reader.

In this study, we implement the XGBoost algorithm using the R-package: *"xgboost: Extreme Gradient Boosting"*[60]. Hyper-parameters η, γ, max_depth, λ, subsample, α, $scale_pos_weight$ are fixed through a cross-validation procedure described in the following section. η, α and $scale_pos_weight$ parameters are additional hyper-parameters not mentioned during the introduction of the boosting model. η is the equivalent of a learning step for a gradient descent optimization algorithm. Each new booster is equivalent to direction selection in the optimization space, while η limits the step taken in the selected direction. α is similar to the γ parameter and puts a weight on newly added leaves. The $scale_pos_weight$ parameter controls for the penalty weight of mis-classifying a positive sample when the data is imbalanced. The sampling_method used is fixed to the *"uniform"* method where each columns has a similar chance of being sub-sampled.

Appendix G

In this section, we describe the methodology of Goutte and Gaussier (2005) followed here to select the most stable set of hyper-parameters. The authors offer a statistical interpretation to the *Precision* (31) and *Recall* (32). We follow their methodology to select the combination of hyper-parameters that would produce the most stable results on

[60]https://xgboost.readthedocs.io/en/latest/parameter.htmlglobal-configuration

new data.

$$F1score = 2 * ((precision * recall)/(precision + recall)) \tag{30}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{31}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{32}$$

Table 13: Confusion Matrix

Assignment	z	+	-
Label l	+	TP	FP
	-	FN	TN

(31) and (32) can be reformulated in statistical terms as $p = P(l = +/z = +)$ and $r = P(z = +/l = +)$. The authors underline the important shift in interpretation arising from the reformulation of (31) and (32): *"...in the original formulation, p and r are just formulas calculated from the observed data, ... , p and r (...) are now parameters of a (primitive) generative model. Thus, the usual expressions arise only as estimates of these unknown parameters."*[61]

Table 13 introduces the confusion matrix. The vertical dimension represents the prediction made by the model while the horizontal dimension represent the true label of the observation. The top right corner cell is the portion of observations that have been correctly classified as positive, hence True-Positive. The other cells are from right to left and top to bottom: False-Positive, False-Negative and True-Negative classification. Assuming that TP conditional on FP, TP and FN follow Binomial Distribution and a Beta prior for Precision and Recall, the authors obtained (33) through (37).

[61] page 4, paragraph 3 of Goutte and Gaussier (2005)

$$p = \frac{X}{X+Y} \tag{33}$$

$$r = \frac{X}{X+Z} \tag{34}$$

$$\text{with } X \sim \Gamma(TP + \lambda, h) \tag{35}$$

$$Y \sim \Gamma(FP + \lambda, h) \tag{36}$$

$$Z \sim \Gamma(FN + \lambda, h) \tag{37}$$

$$\tag{38}$$

From (33) the authors can deduce that F1 score is a combination variable as described in (39).

$$F_1 = \frac{U}{V+U} \tag{39}$$

$$\text{with } U \sim \Gamma(TP + \lambda, 2h) \tag{40}$$

$$V \sim \Gamma(FP + FN + 2\lambda, h) \tag{41}$$

Thanks to (39), F1 has now a density with a set of parameters $\{TP, FN, FP, \lambda, h\}$ from which we can simulate an empirical density of the distribution of each $F1score$ obtained during the cross-validation procedure. An $F1$ score is preferred to another one following the criteria described in (42).

$$\hat{P}(F_1^1 > F_1^2) = \frac{1}{L}\Sigma_{i=1}^{L} I(f_i^1 > f_i^2) \tag{42}$$

We now present the steps that are implemented in this study to select the set of parameters that produces the best distribution of the $F1$ score.

- Retrieve TP, FP and FN for the entire Training database through the CV process

- Simulate a sample of the Gamma laws (L=10000) based on the TP, FP and FN from each set of parameters. Γ and h are arbitrarily chosen to be 2.
- Simulated Distribution of F1 score are then compared by empirically calculating the probability in (42).
- All pairwise score are calculated and then the dominating set of parameters is selected.
- The model is then retrained on the train database using the set of selected parameters.
- Predictions are made on the testing set.

Appendix H

The cross-validation procedure applied to XGB model searches the most performative combination of parameters. The search grid is comprised of 1458 combinations and controls for hyper-parameters described in table 14. The first line introduces the grid itself while subsequent rows give the best parametrization selected after the cross-validation for each configuration in table 3. Specific parameters for *Linear* booster are: *Updater* = *coord_descent* and *Sample_Type* = *uniform*. For *Dart* booster, specific parameters are: *FeatSelecter* = *random* and *Normalize_type* = *tree*.

Table 14: Cross-validation grid

Model	λ	η	γ	MaxDepth	Subsample	Weight	DropRate
Grid	$\{0,1,2\}$	$\{.3,.4,.5\}$	$\{.3,.4,.5\}$	$\{4,5,6\}$	$\{.7,.8\}$	$\{5,10,20\}$	$\{.08,.1,.12\}$
Tree	2	.4	.3	4	.8	5	NA
Dart	0	.3	.5	4	.8	5	.1
Linear	0	.5	.4	5	.7	5	NA

Continued on next page

Table 14: Cross-validation grid (Continued)

Model	λ	η	γ	MaxDepth	Subsample	Weight	DropRate
FOTree	1	.3	.4	4	.8	5	NA
FODart	1	.3	.5	4	.7	5	.1
FOLinear	0	.5	.5	5	.8	5	NA
NLTree	0	.3	.5	4	.8	10	NA
NLLinear	0	.5	.4	5	.7	10	NA
NLDart	2	.4	.4	4	.8	5	.08

Appendix I

The function $\phi(P)$ is truncated at P_0 such that $\phi(P) = 0$ if $P_0 < P < 1$ where P_0 is defined by (43). P_0 is the audit probability at which someone with no moral cost will be deterred to fraud. Any more aggressive audit strategy that P_0 is inefficient as at any audit level such that $P > P_0$, there is simply no fraud being committed.

If the penalty to defraud B is zero, then $P_0 = 1$. The observation signifies that when there are no penalties B, individuals with no moral cost are deterred from defrauding only if the audit strategy consists in investigating all claims: $P_0 = 1$. If the penalty to defraud $B > 0$ then $0 < P_0 < 1$.

$$(1 - P_0)U(W_0 + t, 0) + P_0U(W_0 - B, 0) = U(W_0, 0) \tag{43}$$

The function ϕ has characteristics as described in (44) through (46):

$$\phi(0) > 0 \tag{44}$$

$$\phi(P_0) = 0 \tag{45}$$

$$\phi^{'}(P) < 0 \text{ if } 0 < P < P_0 \tag{46}$$

(44) shows that if the audit rate is non-existent, claimants with a positive moral cost ω will be indifferent between the status-quo and fraud and (43) will be true. (45) reflects the absence of fraud when the audit level is at its maximum P_0, ω belonging to $\mathbb{R}+$. (46) is the rate at which an aggressive audit will deter fraud.

Appendix J

The following methodology describes the spreading and smoothing of the residual undetected fraud over all θ:

- A 15% fraud rate gives 10561 fraudulent claims out of 70142 observations.
- $10561 - 3848 = 6713$ undetected frauds.
- Undetected_fraud_rate $P(\hat{F}/\theta)$ is $\frac{6713}{(70142-3848)} = 10.00\%$ and is distributed equally on all θ
- let N be the non-fraud population of a given θ
- let I be the fraud population of a given θ
- $P(F/\theta) = \frac{P(\hat{F}/\theta)*N+I}{I+N}$

Appendix K

We present here the methodology with which the expected cost of special investigation is estimated. After discussing with our industrial partner, we base our estimate on the hypothesis that around 10% of investigated claims necessitate a deeper special investigation.

Over a period of eight years[62], 1200 claims have been specially investigated resulting in a global probability of investigation of 1.7%. This global probability is then disaggregated over all θ groups following the procedure below:

- From the data, we can observe the probability $P(\theta/AEX > 0)$ where $AEX > 0$ indicates that the observation has special examination costs.
- Hypothesis: $P(\theta/AEX > 0) = P(\theta/Surveillance)$.
- By Bayes law, we get $P(Surveillance/\theta) = \frac{P(\theta/Surveillance)*P(\theta)}{P(S)}$.
- If some θ do not have observations with $AEX > 0$ because of low population then a minimum 0.1% is enforced.

The cost of special investigation is modelled after the total cost of compensation $t(\theta)$. The assumption is that a cost of special investigation would on average not exceed a certain fraction of the total compensation that the claim will cost. Moreover our industrial partner indicated that most special investigations have a cost in $[4000\$, 7000\$]$.

Based on these observations, the range of compensation at claim level is calculated and segmented in fifty parts. Similarly the range of special investigation cost is segmented in fifty distinct parts. A mapping between the two range allows to obtain the final special investigation costs. The cost is obtained as follow: an observation belonging to the 1^{st} segment in t range will be clustered at midpoint of the 1^{st} segment in special surveillance and so on for each clusters. We now have a cost and a probability at the claim level. This allows to calculate the expected cost of special investigation and aggregate it at the θ level.

Figure 4 shows the break down of investigation cost $c(\theta)$ per component at θ level.

[62]The database lasts from 2012-2020

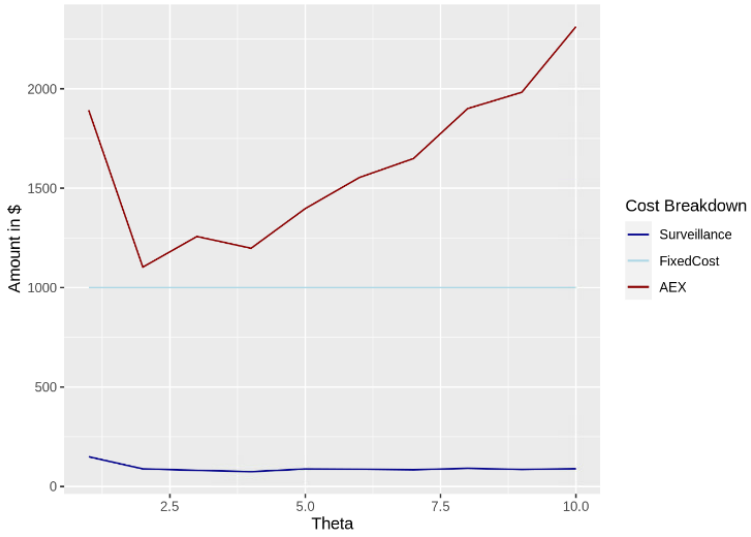


Figure 4: Break down of global $c(\theta)$

Appendix L

The algorithm stratifies the data into categories according to a list of discretized variables. Stratas are created according to dimensions: *"How_reported"*, *Cause_of_Loss* and *Occupation*. For each created strata, the algorithm samples the minimum number of observations necessary so that the first and second moments of some target variables are within a 95% interval from the full population distributions. These target dimensions are: 15 dimensions related to disbursement of policies :"PAID_SF_ADE","PAID_SF_AOX","PAID_SF_ADP","PAID_SF_AEX and 6 general purposes variables *age_claimant_max*, *RES_AT_AB_OPEN_Rat*, *AB_fraud_label2*, *Veh_Age*, *TIME_TO_EXPIRY*, *NBR_CLAIMS_ON_POL*.

Appendix M

Table 15: σOAM vs Elkan in Fraud-Only

			Elkan					σOAM				
θ	$\frac{t}{c}$	F%	TP	FP	TN	FN	Thrs	TP	FP	TN	FN	Thrs
1	7.65	3.34	11	60	3256	33	0.130	7	33	3283	37	0.230
2	9.52	2.42	1	17	1039	3	0.105	1	5	1051	3	0.294
3	8.99	2.96	3	21	975	5	0.111	2	11	985	6	0.230
4	8.77	3.41	11	24	883	3	0.114	10	16	891	4	0.193
5	8.27	3.97	11	17	926	8	0.121	10	11	932	9	0.193
6	8.02	4.34	8	30	920	16	0.125	8	19	931	16	0.193
7	7.61	5.28	17	7	51	1020	0.131	15	43	1028	21	0.155
8	7.24	6.27	30	73	1049	32	0.138	29	69	1053	33	0.155
9	6.91	7.13	37	95	1277	47	0.145	39	114	1258	45	0.115
10	6.27	9.35	75	137	1433	68	0.159	87	217	1353	56	0.079

Columns $\frac{t}{c}$ is the ratio of the compensation to the investigation cost per θ. $F\%$ is the fraud-rate and TP, FP, TN and FN are respectively the true-positive, false-positive, true-negative and false-negative number of claims. $Thrs$ is the calibrated probability threshold above which a claim is audited. A higher threshold means a less agressive audit policy.

Table 16: σOAM vs Elkan in No-Law

			Elkan					σOAM				
θ	$\frac{t}{c}$	F%	TP	FP	TN	FN	Thrs	FP	FN	TP	TN	Thrs
1	7.66	1.76	5	29	3170	57	0.130	7	60	2	3192	0.211
2	9.43	1.15	3	22	1008	6	0.106	6	7	2	1024	0.211
3	8.84	1.90	5	26	938	10	0.113	14	12	3	950	0.203
4	8.78	2.62	13	40	892	17	0.114	25	20	10	907	0.206
5	8.29	3.78	8	46	880	34	0.121	39	34	8	887	0.200
6	7.95	4.54	22	61	904	24	0.126	58	26	20	907	0.200
7	7.56	5.90	23	83	976	39	0.132	88	40	22	971	0.113
8	7.19	7.50	40	130	1070	53	0.139	179	48	45	1021	0.089
9	6.96	9.56	55	196	1255	76	0.144	275	65	66	1176	0.075
10	6.35	13.63	139	324	1266	128	0.158	498	85	182	1092	0.064

Columns $\frac{t}{c}$ is the ratio of the compensation to the investigation cost per θ. $F\%$ is the fraud-rate and TP, FP, TN and FN are respectively the true-positive, false-positive, true-negative and false-negative number of claims. $Thrs$ is the calibrated probability threshold above which a claim is audited. A higher Threshold means a less agressive audit policy.