Alternatives Managériales et Economiques
E-ISSN : 2665-7511
https://revues.imist.ma/?journal=AME

KOUACH et al. / Revue AME Vol 4, No 4 (Octobre, 2022) 96-113

Auto insurance fraud detection using unsupervised learning,

KOUACH, Y.[1], EL ATTAR, A.[2], EL HACHLOUFI, M.[3]

1. PhD student in Actuarial Science, Department of Statistics and Mathematics Applied to Economics and Management, Faculty of Juridical Sciences, Economic and Social-Ain Sebaa, Hassan II university, Casablanca, Morocco, yassinekouach-etu@etu.univh2c.ma
2. Professor- Researcher, Department of Statistics and Mathematics Applied to Economics and Management, Faculty of Juridical Sciences, Economic and Social-Ain Sebaa, Hassan II university, Casablanca, Morocco, abderrahim.elattar@univh2c.ma
3. Professor- Researcher, Department of Statistics and Mathematics Applied to Economics and Management, Faculty of Juridical Sciences, Economic and Social-Ain Sebaa, Hassan II university, Casablanca, Morocco, mostafa.elhachloufi@univh2c.ma

Submission date: 31/08/2022 **Acceptance date:** 13/10/2022

Abstract:

Insurance fraud represents a source of significant financial loss for insurance companies, particularly auto insurance fraud. Indeed, the insured seeks to increase his indemnity by fraudulent acts or declares a false claim. Consequently, the detection of automobile insurance fraud holds an important place in the strategy of the insurance company, since it allows to reduce the costs of claims and maintain a satisfactory profit. With this in mind, insurers are always looking to have more efficient detection systems in order to overcome the limits of traditional methods.

In this paper, we propose an automobile insurance fraud detection system based on Machine Learning. Two unsupervised learning algorithms will be used, namely Isolation Forest and Local Outlier Factor.

According to previous studies, these two algorithms are not yet applied in auto insurance fraud detection. Both algorithms belong to the anomaly detection algorithms family which is more suitable for this type of problem.

The results of this study show the possibility of developing a detection system, based on these two Machine Learning techniques, which serves to automatically detect fraud with considerable accuracy.

Key words: Automobile insurance, fraud, machine learning algorithms, unsupervised learning.

Apprentissage non supervisé pour la détection de fraude à l'assurance automobile

Résumé :

La fraude dans l'assurance représente une source de perte financière importante pour les compagnies d'assurance, particulièrement la fraude dans l'assurance automobile. En effet, l'assuré cherche à augmenter son indemnité par des actes frauduleux ou bien, il déclare une fausse réclamation. Dès lors, la détection de fraudes à l'assurance automobile tient une place importante dans la stratégie de la compagnie d'assurance étant donné qu'elle permet de réduire les coûts des sinistres et de préserver une rentabilité satisfaisante. Dans cette optique, les assureurs cherchent toujours à disposer de systèmes de détection plus performants afin de dépasser les limites des méthodes traditionnelles.

Dans cet article, nous proposons un système de détection de fraude à l'assurance automobile basé sur le Machine Learning. Deux algorithmes d'apprentissage non supervisés seront utilisés, en l'occurrence Isolation Forest et Local Outlier Factor.

Selon des études antérieures, ces deux algorithmes ne sont pas encore appliqués dans la détection de fraude à l'assurance automobile. Les deux algorithmes appartiennent à la famille des algorithmes de détection d'anomalies qui s'adapte mieux à ce type de problème.

Les résultats de cette étude montrent la possibilité de développer un système de détection, basé sur ces deux techniques de Machine Learning, qui serve à détecter automatiquement les fraudes avec une précision considérable.

Mots-clés : Assurance automobile, fraude, algorithmes d'apprentissage automatique, apprentissage non supervisé.

Introduction:

Insurance fraud, a scourge that is complex to control, concerns all types of insurance branches, in particular automobile insurance. Indeed, the automobile insurance branch is fertile ground for fraud because insurers adopt a rapid compensation process, which encourages the insureds to claim the indemnity for simulated accidents, false reports, or voluntary claims.

This significant phenomenon of fraud weighs heavily on the financial situation of insurance companies. In fact, indemnity for fraudulent claims requires the establishment of higher reserve allocations than the reserves without fraud. It should therefore be emphasized that fraudulent claims increase the claims burden of insurance companies which leads to the realization of significant losses in terms of profitability and therefore hinders the growth of insurance companies.

In order to avoid the negative impacts of fraud on the financial situation of insurance companies, insurers always seek to implement detection systems able to detect fraudulent attempts. However, traditional methods, such as the personal assessment of claims files by claims handlers, have limitations in terms of time and efficiency and sometimes require additional examinations. With this in mind, insurers are always looking to have more efficient detection systems in order to overcome the limits of traditional methods.

From this perspective, this article aims to develop a more relevant auto insurance fraud detection system using new technologies, in this case, Machine Learning. Our proposed model is an a posteriori fraud detection model which consists of detecting fraud after the declaration of the claim. We seek through this work to develop a fraud detection system in car insurance that allows insurers to protect themselves against financial losses by exploiting the unavoidable potential of Machine Learning techniques to detect anomalies.

To this end, we conducted a recent literature review of the use of machine learning in the detection of fraud in the financial field. According to the recent literature review, there are some techniques that are not yet used in fraud detection in the insurance sector and specifically in automobile insurance (Al-Hashedi & Magalingam, 2021; Hilal et al., 2022).

Subsequently, we choose to use two algorithms of unsupervised learning namely Isolation Forest and Local Outlier Factor. In order to choose the model that can identify whether a loss compensation claim is fraudulent or not fraudulent, we proceed to analyze their performance such as the accuracy rate.

This paper is declined into three parts: the first part stands for a literature review by presenting insurance fraud, the Machine Learning concept and the related works on the use of Machine

Learning in insurance and finance fraud detection. The second part provides the methodology adopted in this study by highlighting the machine learning models used, the performance measures in order to choose the best model, the process of developing indicators for detecting fraud in automobile insurance, as well as the data and material. The third part will be devoted to the exposure of the results of our empirical application in order to choose the best model as a detection system for automobile insurance fraud.

1. Literature review

1.1. Insurance fraud

Insurance fraud is any operation by an insured in bad faith seeking to take advantage of the guarantees of the insurance contract with illegal instruments. Fraud is a phenomenon that can appear at any stage in the life of the insurance contract. Indeed, at the underwriting phase, the insured may provide incorrect or incomplete information when applying for insurance in order to reduce the insurance premium. This type of fraud is named also premium fraud that concerns the intentional misrepresentation of information that is provided at the time of underwriting of an insurance contract in order to benefit from an unduly low premium (Vandervorst et al., 2022).

The fraud thus exists at the time of claims declaration by the exaggeration of the amount of the claim, declaration of simulated accidents that does not exist in reality, voluntary claims or declaration of an accident occurring before effective date of the insurance contract.

Fraud has a negative impact on the growth of insurance companies. In terms of tariffication, pricing based on historical fraudulent claims leads to higher insurance premiums. Moreover, bona fide lucky policyholders pay higher insurance premiums due to fraudulent policyholders which motivates policyholders to change the insurance company.

In terms of technical profitability, indemnity for claims with higher amounts or indemnity for claims that do not exist in reality requires the constitution of allocations of higher provisioning than provisioning without fraud. As a result, the claims burden increases and the technical profitability decreases.

This is why, the mission of detecting insurance fraud plays a crucial role since it makes it possible to offer fairer insurance premiums and according to the risk actually incurred without penalizing the insured in good faith. Thus, it helps reduce claims costs in terms of lower reserves constitutions.

1.2. Machine learning concept

Machine Learning is at the intersection of several disciplines namely statistics, probability, and computer science. It is a branch of Artificial Intelligence where the algorithms perform the tasks assigned by learning from their experience in executing these tasks (Ray, 2019) without being explicitly programmed (Joshi, 2020; Mahesh, 2019) with or without a teacher (Mitchell, 1997). Machine learning is a set of computational algorithms designed to emulate human intelligence in order to achieve a desired task to produce a particular outcome using input data (El Naqa & Murphy, 2015). Machine Learning techniques are used in many industries notably medical applications, agriculture, market forecast, finance, and banking.

There are several definitions of Machine Learning in the literature. Arthur Samuel, one of the fathers of machine learning (Ramasubramanian & Singh, 2019), defined machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed" (Samuel, 1959). For Tom Mitchell, Machine Learning is presented as (Mitchell, 1997) "A computer program is said to learn from experience (E) with respect to some class of tasks (T) and performance measure (P), if its performance at tasks in T , as measured by P , improves with experience E ".

Machine learning is carried out in three stages namely, representation, evaluation and optimization. The representation phase consists in finding the most suitable mathematical model. The evaluation phase measures the gap between the model and the reality of the test data. And finally, the optimization phase aims to reduce this gap (Kabak & Rajouani, 2021).

The Machine Learning algorithms are suitable for solving classification, regression, and clustering problems. The Machine Learning techniques are divided into four categories depending on the training data types. These categories are namely Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning.

- **Supervised learning:** The supervised algorithms are applied to predict the value of the output variables from labeled input data which contains both the independent and the target variable (Vaibhav, 2020).

- **Unsupervised learning:** In Unsupervised learning, the training data has only input data (El Naqa & Murphy, 2015) without any right answers (Gopinath et al., 2019) and any teacher (GERON, 2019; Mahesh, 2019). Indeed, the models are applied to unlabeled data (Joshi, 2020) where they try to learn relationships and structure from such data (Gareth et al., 2017) for the purpose of discovering patterns, e.g. grouping similar features (J. Wang & Biljecki, 2022). It is

often employed in applications where labeling is expensive or where it is not relevant (J. Wang & Biljecki, 2022).

- **Semi-supervised learning:** Semi-supervised learning is a combination of supervised and unsupervised techniques where the data are not fully labeled. In fact, the labeled data are used to infer the unlabeled portion (El Naqa & Murphy, 2015). Semi-supervised learning is important in the research on pattern recognition (Gao et al., 2022).

- **Reinforcement learning:** Reinforcement learning algorithms learn from their environments in order to get the best rewards in response to each action taken. Indeed, the agent interacts with the environment at a given instant, it executes an action (Gomes et al., 2022). Depending to this action, the next state will be determined. The cycle perception–action-reward progresses with time(Gomes et al., 2022).

1.3. Related works

In this section, we review some related works that introduced the use of machine learning in fraud detection in insurance and finance field.

Several studies have more recently performed fraud detection on financial area. (Raghavan & El Gayar, 2019) present in their study an empirical investigation comparing various machine learning and deep learning models for detection of fraudulent transaction by credit cards or online payments. They compare the performance of multiple machine learning methods such as k-nearest neighbor (KNN), Random Forest, and support vector machines (SVM), while the deep learning methods such as autoencoders, convolutional neural networks (CNN), restricted boltzmann machine (RBM) and deep belief networks (DBN) (Raghavan & El Gayar, 2019).

Moreover, they propose an ensemble model by combining the 3 best performing models using majority voting. Their work brings out that the best methods with larger datasets would be using SVMs, potentially combined with CNNs to get more reliable performance. For the smaller datasets, ensemble methods of SVM, Random Forest and KNNs can provide good enhancements. (Sadgali et al., 2019) propose a state of art in various fraud detection techniques, like machine-learning, proposed in the literature on different financial frauds in order to identify the methods that give the best performance. They based their comparison on some criteria such as the possibility that the technique can be run in real time, the accuracy, and the sensitivity. They observed that almost, all implemented algorithms, do not work in real time. Furthermore, the detection of credit card fraud uses several Machine Learning techniques for the detection of frauds of financial statements, it is based mainly on text processing techniques. They found that

hybrid fraud detection techniques are the most used, as they combine the strengths of several traditional detection methods.

The detection of insurance fraud is an area of research that arouses great interest because of the negative impact of fraud on the economic stability of insurance companies. (Viaene et al., 2005) report in their article the findings from applying Bayesian learning neural networks for personal injury protection automobile insurance claim fraud detection. In addition, they explored the explicative capabilities of neural network classifiers with automatic relevance determination weight regularization providing a way to determine the relative importance of each input to the trained neural network model.

(Karsenty, 2016) develops in his thesis an unsupervised method of RIDIT and PRIDIT to detect fraud in insurance. He used the RIDIT method to calculate a fraud suspicion score for each variable. After obtaining a score matrix with the individuals in rows and the variables in columns, he used PRIDIT method to calculate an overall fraud score per individual. The results given by the methods of RIDIT and PRIDIT show the relevance to use these methods in insurance fraud detection.

(Roy & George, 2017) present in their conference paper the theoretical framework of the use of machine learning in insurance claims fraud detection by revealing the feature of three classification algorithms namely Naive Bayes, random forest, and decision tree.

(Y. Wang & Xu, 2018) propose a deep learning model for automobile insurance fraud detection that uses Latent Dirichlet Allocation (LDA)-based text analytics. The authors used LDA method to extract the text features hiding in the text descriptions of the accidents appearing in the claims, and deep neural networks to detect the fraud claims in the data, which include the text features and traditional numeric features. They conclude that deep neural networks outperform widely used machine learning models, such as random forests and support vector machine.

In the study of (C. da Rosa, 2018) an empirical evaluation of several unsupervised machine learning approaches is performed in order to detect fraud in Medicare data. The author recommends using LOF in the case of unsupervised models to detect fraud on Medicare data. In Medicare field also (Deborah, 2019) applied Local Outlier Factor, Isolation Forest, in addition to Hierarchical Ascending Classification in order to detect fraudulent behavior during reimbursements in the context of the hospital sector.

(Severino & Peng, 2021) evaluated fraud prediction in property insurance claims using nine machine learning models based on real-world data from a major Brazilian insurance company. In addition, they compiled a general profile of confirmed fraudsters from the dataset and analyzed

the relative importance of the input variables according to each model using eXplainable Artificial Intelligence (XAI) methods. The results of their study showed that the random forest model achieved significantly better performance than the standard logistic regression and other machine learning methods. However, the deep neural network model outperformed the other models for the recall metric.

(Urunkar et al., 2022) implement a model based on machine learning algorithms that label and classify insurance claim. They used logistic regression, XGB, decision tree, random forest, and K nearest neighbor, to construct a model that detects if an insurance claim is fraudulent or not.

(Hajraoui & Zahi, 2022) highlight in their article the most relevant ex-post fraud detection indicators that would be integrated into an auto insurance fraud detection automation model. For that, the authors used two models of the literature. The first model is of (Belhadji & Dionne, 1997) model that used a Probit regression model, and the second model is of (Benedek & László, 2019) model that used decision tree, Naive Bayes, and neural network algorithms. The proposed model that used the relevant indicators from the two previous models reveals two indicators: a minor collision resulted in excessive repair costs and the type of incident.

According to the literature review, there are some algorithms of machine learning not yet been tested in automobile insurance fraud detection. In that light, we are looking to apply two unsupervised learning algorithms in occurrence Local Outlier Factor and Isolation Forest.

2. Methodology

2.1. Proposed models

In the framework of automobile insurance fraud detection, insurance companies seldom have the labeled data of the fraudulent claims that why supervised learning algorithms are not appropriate to detect future likely fraudulent claims. In this perspective, we propose to use the unsupervised learning techniques that are more suitable for this issue, especially the methods addressed to the anomaly detection problem such as Isolation Forest, Local Outlier Factor. The anomaly detection consists in finding the outliers and data points that are abnormal.

- **Isolation Forest:** Isolation Forest algorithm implemented using decision trees like random forests but they are built without teacher (without labels). This algorithm relies on the assumption that anomalies are few and different from normal points, and that recursive space partitioning should isolate anomalous data points (outliers) in an easier way with respect to normal data points (inliers) (Barbariol & Susto, 2022).

  The Isolation Forest algorithm starts dividing the data by randomly selecting an attribute q and a split value p, until there is no possibility to split. The Isolation tree constructed is a binary

tree where each node in the tree has exactly zero or two daughter nodes (Liu et al., 2008). The task of anomaly detection is to provide a ranking that reflects the degree of anomaly (Nofal et al., 2021). Thus, one way to detect anomalies is to sort data points according to their path lengths or anomaly scores, and anomalies are points that are ranked at the top of the list (Liu et al., 2008).

- **Local Outlier Factor:** The Local Outlier Factor (LOF) algorithm is an unsupervised density-based anomaly detection method. The algorithm finds outliers by calculating the local density deviation of a given data point. The determination of the outlier is judged based on the density between each data point and its neighbor points (Cheng et al., 2019). It is considered as outliers the points having a density appreciably lower than that of their neighbors. In other words, the algorithm detects local outliers by calculating the outlines as the degree of how isolated an object is from its surrounding neighbors (Choi et al., 2022).

2.2. Performance metrics

In our case to evaluate the performance of our models, we will use test data with some fraudulent claims in order to judge which model detects the fraudulent claims. For this reason, an evaluation of the performance of the proposed detection systems is considered.

Table 1. Confusion Matrix

		Actual	
		Negative	Positive
Prediction	Negative	TN True Negative	FP False Positive
	Positive	FN False Negative	TP True Positive

Source: Developed by the authors.

The performance metrics used in this article thanks to the confusion matrix (table 1) which provides the information on actual and predicted values, are:

- **Recall:**

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall measures the portion of claims that are correctly identified as fraudulent claims.

- **Precision:**

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Precision measures the portion of claims that are correctly identified as fraudulent among all the claims that are identified as positives.

- **Specificity:**

$$\boldsymbol{s}\text{pecificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Specificity measures the portion of claims that are correctly identified as not Fraudulent.

- **Accuracy:**

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Accuracy is a metric that combines the Precision and Recall metrics. Accuracy measures the portion of observations that are correctly classified among all predicted observations.

We will use also the Area Under Curve AUC score which is the area under the Receiver Operating Characteristic ROC Curve. The best fraudulent detection model according to AUC score is the model that has the larger area under the ROC curve.

2.3. Model conception

We are interested in the development of a detection system for automobile insurance fraud given the importance of this branch of insurance in the portfolio of insurance companies. Thus, we seek to develop this system using two Machine Learning algorithms in order to overcome the limits of traditional methods. Our proposed model is an a posteriori fraud detection model which consists of detecting fraud after the declaration of the claim.

Our work approach is part of a relational research that makes it possible to determine the relationship between the explanatory variables and the variable to be explained in this case, fraud reported.

The first step in the process as presented in figure 1 below, consists of data collection. Then we proceed to clean data by identifying the missing value.

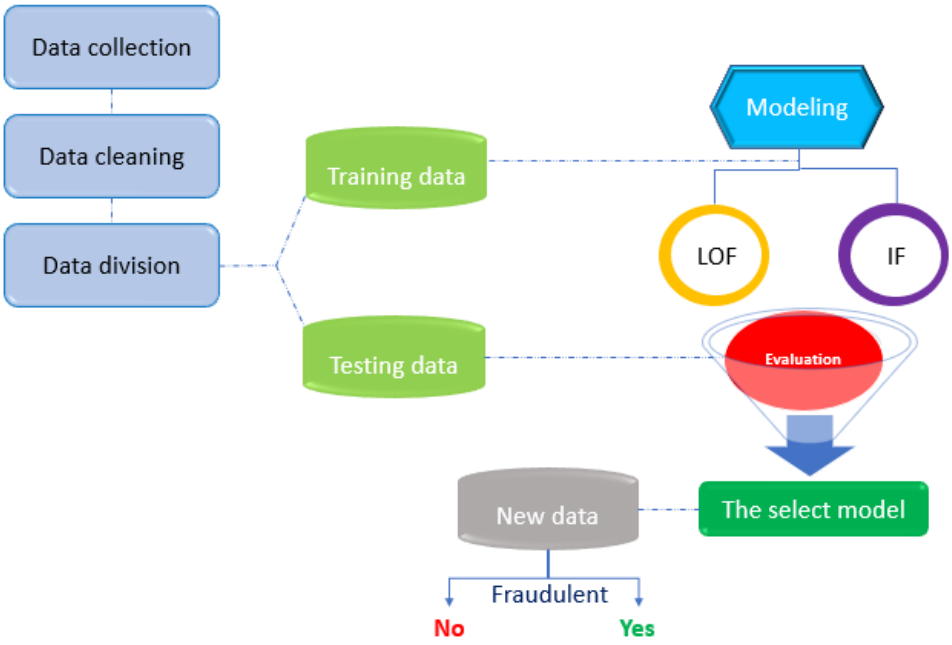
The application of machine learning algorithms needs to divide the database into two datasets, the training dataset intended for the development of the models and the test dataset intended for the evaluation of the models. 80% of the data is dedicated for model training and 20% of the data is dedicated for data testing. As far as the optimization of the parameters is concerned, the technique of cross-validation is used.

We proceed to the modeling of the fraud variable reported on the training base by the unsupervised machine learning algorithm, Isolation Forest. Then by the Local Outlier Factor algorithm.

We then calculate the performance measures namely recall, precision, accuracy and specificity.

Finally, we move to the evaluation step, in order to select the best model to be used in the detection of fraudulent claims in the new data in the future according to the performance metrics calculated.

Fig 1: The process of implementing the proposed automobile insurance fraud detection model



Source: Developed by the authors.

The data used in this article is open-source data for automobile insurance claims. It is constituted of 1000 observations and 39 columns including 38 explanatory variables and a response variable indicating the reported fraud, fraud or non-fraud. In our study, we will use just the explanatory variables without the response variable in the training phase because we use unsupervised learning methods. However, in the testing phase, we will use the response variable to evaluate the performance of the models.

After data processing, we retained only 18 explanatory variables (table 2), 17 variables from the initial database such as “incident type”, “collision type”, “incident severity”, “number of vehicles involved”, “witnesses”, “police report available”, and new explanatory variable indicates the period between the incident date and the date of subscription.

Table 2. Variables used for model conception

Variables	Description
Months_as_customer	number of months as insured by the insurance company {1,2…,k}
Number_of_vehicles_involved	number of vehicles involved in the accident {1,2…,k}
Bodily_injuries	number of injured {1,2…, k}
Witnesses	Number of witnesses {1,2…, k}
Injury_claim	Injury claim cost
Insured_sex	Sex of the insured {Male or Female}
Periode_sou_sin	The period of the claim
Policy_annual_premium	Annual premium
Vehicle_claim	Vehicle claim cost
Property_claim	Property claim cost
Insured_education	The insured education level
Insured_occupation	The insured job
Incident_type	Type of incident {parked car, single vehicle collision, vehicle thief…}
Collision_type	Type of collision {front collision, rear collision, side collision}
Incident_severity	The severity of the incident {minor damage, big loss…}
Property_damage	If the property damage exists {0,1}
Police_report_available	If the police report available {0,1}
Total_claim_amount	The total amount of the claim

Source: Developed by the authors.

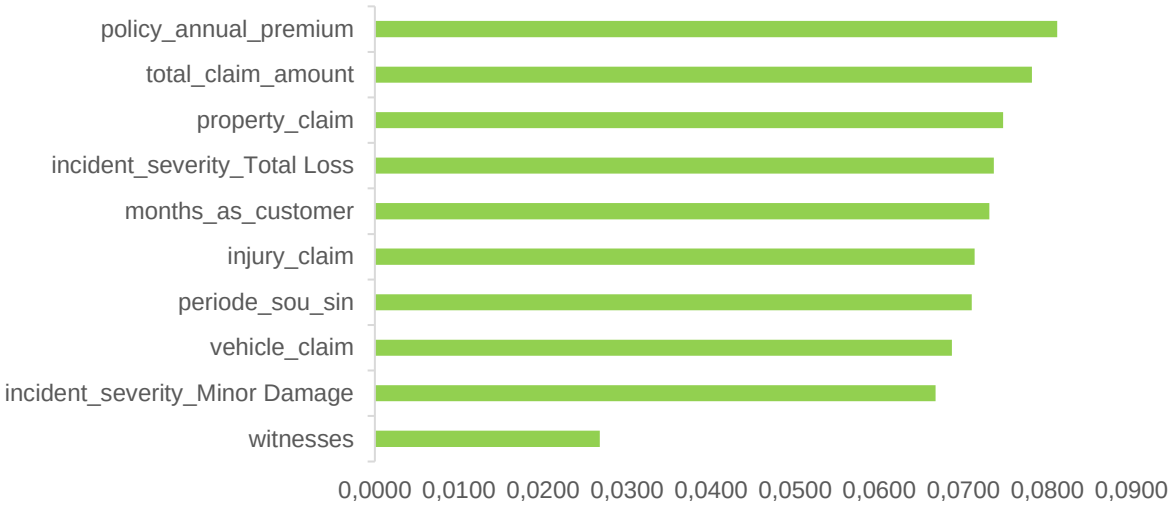
The auto insurance claim fraud detection model is implemented using Python programming software with the help of Scikit-learn library which is suitable for machine learning algorithms in Python.

3. Results and discussion

Before starting the implementation of the unsupervised learning algorithms adopted in this study to develop a fraud detection system in automobile insurance, we dwell on the importance of the explanatory variables influencing the response variable “fraud detected” (figure 2).

By using the supervised Machine Learning algorithm "Random Forest", we mention the first ten most important variables, namely the existence of witnesses, the period that the insured has spent as a customer in the insurance company, the period between insurance contract underwriting and declaration loss date, and the degree of severity of the loss. These variables qualified as being the most relevant and the most likely to lead to a suspicion of fraud.

Fig 2: The importance of explanatory variables according to random forest algorithm



Source: Based on Our results.

Firstly, we present the confusion matrix of the Local Outlier Factor algorithm (table 3) and the Isolation Forest algorithm (table 4) that allow us to calculate the performance measures namely Recall, Precision, Specificity, and accuracy.

Table 3. Confusion matrix of local Outlier Factor algorithm

		Actual	
		Negative	Positive
Prediction	Negative	171	49
	Positive	64	16

Source: Based on Our results.

According to the confusion matrix of Local Outlier Factor model (table 3), 16 claims are identified as fraudulent and they are fraudulent in the actual time. As well, 171 claims are identified as not fraudulent and they are not fraudulent in the actual time.

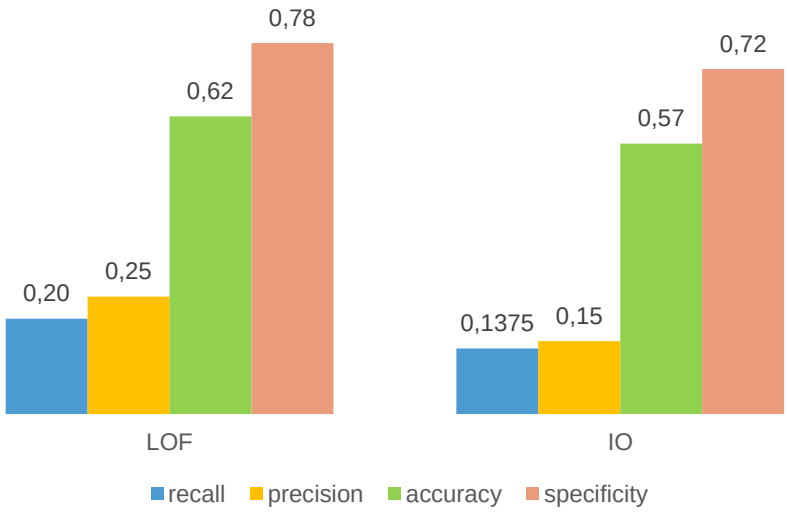
Table 4. Confusion matrix of Isolation Forest algorithm

		Actual	
		Negative	Positive
Prediction	Negative	159	61
	Positive	69	11

Source: Based on Our results.

As described in the confusion matrix of Isolation Forest model (table 4), this model identified 61 claims as not fraudulent but they are fraudulent in the actual time. Besides, the model identified 69 claims as fraudulent but they are not actually fraudulent.

Fig. 3. The performance metrics of Local Outlier and Isolation Forest



Source: Based on Our results.

The performance metrics adopted in this study with a view to choosing the best model are presented in figure 3 above.

Isolation Forest model reached a specificity of 0.72, in other words, 72% of non-fraudulent claims were identified as non-fraudulent. In addition, 13% of fraudulent claims were correctly detected according to recall, and 15% of non-fraudulent claims were effectively classified as non-fraudulent based on precision. In general, Isolation Forest attains to predict correctly 57% of the predicted observations with an accuracy rate equal to 0.57.

Regarding the Local Outlier Factor model, this model succeeds in identifying non-fraudulent claims with a specificity of 78%. With a recall equal to 0.20, this algorithm correctly detects 20% of fraudulent claims and 25% of non-fraudulent claims with an accuracy equal to 0.25. Overall, according to the accuracy rate which is 0.62, the LOF model classifies 62% of correctly predicted observations either as fraudulent or non-fraudulent claims.

Our work shows that the Local Outlier Factor algorithm outperformed the Isolation Forest algorithm in terms of fraud detection according to all performance measures likely Recall, Precision, Specificity, and Accuracy. Indeed, the first model manages to predict 20% of real fraudulent cases, this means that out of 100 really fraudulent cases it manages to detect 20 cases correctly. This algorithm has an error rate which is limited to 28%, among the 20 cases detected as fraudulent, 5 cases are incorrectly identified as fraudulent.

According to this study, we invite insurers to opt for unsupervised machine learning algorithms in order to detect fraud in automobile insurance. In detail, we favor the Local Outlier Factor model over the Isolation Forest model, despite the fact that these methods have not reached their best

performance wanted through this work due to the limitations of the database devoted to their training. Indeed, these algorithms require gigantic databases so that the models can train well and learn well in order to better predict fraudulent cases. Thus, it is necessary to pay attention to the imbalances of the data because the class of non-fraudulent cases is the majority, which sometimes leads to a convergence of the models towards the prediction of non-fraudulent cases instead of fraudulent ones.

Conclusion and prospects:

Fraud detection is a major challenge for insurance companies. Indeed, this phenomenon contributes to the increase in the cost of claims, hence a deterioration in the profitability of insurance companies. As a result, insurers are forced to develop more robust fraud detection systems in order to monitor new dishonest maneuvers by policyholders seeking to earn money by exercising the guarantees of their insurance contracts.

From this perspective, we have tried to propose an automobile insurance fraud detection system based on unsupervised Machine Learning algorithms. We applied Local Outlier Factor and Isolation Forest, two unsupervised techniques suitable for detecting anomalies and outliers.

The results of this article showed the performance of the Local Outlier Factor model in identifying fraudulent claims as opposed to the Isolation Forest model which did not give satisfactory prediction rates.

From a practical point of view, this article helps actuaries to test these two models as a fraud detection system within the insurance company. And from a theoretical point of view, this work allowed us to apply two unsupervised learning models that are not yet treated in the context of auto insurance fraud detection.

Among the limitations encountered during this work is the small size of the database which does not allow the algorithms to learn better from past data. In addition, there is the problem of imbalances of the data which gives rise to the classification of fraudulent claims as non-fraudulent claims wrongly because of the convergence of the model towards the majority response variable in this case non-fraudulent claims.

The first limitation remains to be dealt with the actuaries within the insurance companies which have colossal databases with a view to applying our proposed model in the future.

Regarding imbalanced data, it could be the subject of future research work by looking for new methods to be applied in order to remedy this imbalance issue and to provide databases suitable for the application of unsupervised Machine Learning algorithms for fraud detection.

Bibliography:

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 100402.

Barbariol, T., & Susto, G. A. (2022). TiWS-iForest: Isolation forest in weakly supervised and tiny ML scenarios. *Information Sciences*, *610*, 126–143. https://doi.org/10.1016/j.ins.2022.07.129

Belhadji, E. B., & Dionne, G. (1997). Développement d'un système expert de détection automatique de la fraude à l'assurance automobile. *Cahier de Recherche*, *97*, 04.

Benedek, B., & László, E. (2019). Identifying Key Fraud Indicators in the Automobile Insurance Industry Using SQL Server Analysis Services. *Studia Universitatis Babes-Bolyai*, *64*(2), 53–71.

C. da Rosa, R. (2018). *An evaluation of Unsupervised Machine Learning Algorithms for Detecting Fraud and Abuse in the U.S. Medicare Insurance Program | fau.digital.flvc.org* [Thesis master degree, Florida Atlantic University]. http://fau.digital.flvc.org/islandora/object/fau:40847

Cheng, Z., Zou, C., & Dong, J. (2019). Outlier detection using isolation forest and local outlier factor. *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, 161–168. https://doi.org/10.1145/3338840.3355641

Choi, J., Jeong, B., & Yoon, J. (2022). Identification of emerging business areas for business opportunity analysis: An approach based on language model and local outlier factor. *Computers in Industry*, *140*, 103677. https://doi.org/10.1016/j.compind.2022.103677

Deborah, H. (2019). *Implémentation d'un modèle de détection de fraude à l'assurance dans le cadre de soins hospitaliers*. ENSAE ParisTech.

El Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? In I. El Naqa, R. Li, & M. J. Murphy (Eds.), *Machine Learning in Radiation Oncology: Theory and Applications* (pp. 3–11). Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1

Gao, F., Gao, W., Huang, L., Xie, J., & Gong, M. (2022). An effective knowledge transfer method based on semi-supervised learning for evolutionary optimization. *Information Sciences*, *612*, 1127–1144. https://doi.org/10.1016/j.ins.2022.09.020

Gareth, J., Witten, D., Trevor, H., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*. springer.

GERON, A. (2019). *Le machine learning avec Scikit learn* (2nd ed.). dunod.

Gomes, G., Vidal, C. A., Cavalcante-Neto, J. B., & Nogueira, Y. L. B. (2022). A modeling environment for reinforcement learning in games. *Entertainment Computing*, *43*, 100516. https://doi.org/10.1016/j.entcom.2022.100516

Gopinath, R., Ajay, R., & Sanjay, C. (2019). *An introduction to machine learning*. springer.

Hajraoui, G., & Zahi, J. (2022). Indicateurs pertinents de détection automatique de fraude: Cas des compagnies d'assurance automobile. *Alternatives Managériales Economiques*, *4*(2), 420–436. https://doi.org/10.48374/IMIST.PRSM/ame-v4i2.32203

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). *Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances*.

Joshi, A. (2020). *Machine Learning and Artificial Intelligence*. Springer International Publishing. https://doi.org/10.1007/978-3-030-26622-6

Kabak, S., & Rajouani, B. (2021). Emergence de l'intelligence artificielle et incidences sur les gérants de portefeuille. *Alternatives Managériales Economiques*, *3*(4), 122–141. https://doi.org/10.48374/IMIST.PRSM/ame-v3i4.32569

Karsenty, J. (2016). *La détection de fraudes à l'assurance*. ENSAE ParisTech.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Mahesh, B. (2019). *Machine Learning Algorithms -A Review*. https://doi.org/10.21275/ART20203995

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Nofal, S., Alfarrarjeh, A., & Abu Jabal, A. (2021). A use case of anomaly detection for identifying unusual water consumption in Jordan. *Water Supply*, *22*(1), 1131–1140. https://doi.org/10.2166/ws.2021.210

Raghavan, P., & El Gayar, N. (2019). Fraud detection using machine learning and deep learning. *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 334–339.

Ramasubramanian, K., & Singh, A. (2019). *Machine Learning Using R: With Time Series and Industry-Based Use Cases in R* (2éme). Apress.

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 35–39. https://doi.org/10.1109/COMITCon.2019.8862451

Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, 1–6. https://doi.org/10.1109/ICCPCT.2017.8074258

Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. *Procedia Computer Science*, *148*, 45–54.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications*, *5*, 100074.

Urunkar, A., Khot, A., Bhat, R., & Mudegol, N. (2022). Fraud Detection and Analysis for Insurance Claim using Machine Learning. *2022 IEEE International Conference on Signal Processing,*

Informatics, Communication and Energy Systems (SPICES), *1*, 406–411. https://doi.org/10.1109/SPICES52834.2022.9774071

Vaibhav, V. (2020). *Supervised Learning with Python: Concepts and Practical Implementation Using Python*. Apress.

Vandervorst, F., Verbeke, W., & Verdonck, T. (2022). Data misrepresentation detection for insurance underwriting fraud prevention. *Decision Support Systems*, *159*, 113798. https://doi.org/10.1016/j.dss.2022.113798

Viaene, S., Dedene, G., & Derrig, R. A. (2005). Auto claim fraud detection using Bayesian learning neural networks. *Expert Systems with Applications*, *29*(3), 653–666.

Wang, J., & Biljecki, F. (2022). Unsupervised machine learning in urban studies: A systematic review of applications. *Cities*, *129*, 103925. https://doi.org/10.1016/j.cities.2022.103925

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87–95.