Journal of Mathematics and Modeling in Finance (JMMF)
Vol. 2, No. 1, Winter & Spring 2022
Research paper

Using local outlier factor to detect fraudulent claims in auto insurance

Maryam Esna-Ashari[1], **Farzan Khamesian**[2], **Farbod Khanizadeh**[3]

[1] Insurance Research Center, Tehran, Iran
esnaashari@irc.ac.ir

[2] Insurance Research Center, Tehran, Iran
khamesian@irc.ac.ir

[3] Insurance Research Center, Tehran, Iran
khanizadeh@irc.ac.ir

Abstract:
Given the significant increase in fraudulent claims and the resulting financial losses, it is important to adopt a scientific approach to detect and prevent such cases. In fact, not equipping companies with an intelligent system to detect suspicious cases has led to the payment of such losses, which may in the short term lead to customer happiness but eventually will have negative financial consequences for both insurers and insured. Since data labeled fraud is really limited, this paper, provides insurance companies with an algorithm for identifying suspicious cases. This is obtained with the help of an unsupervised algorithm to detect anomalies in the data set. The use of this algorithm enables insurance companies to detect fraudulent patterns that are difficult to detect even for experienced experts. According to the outcomes, the frequency of financial losses, the time of and the type of incident are the most important factors to in detecting suspicious cases.

Keywords: Unsupervised algorithm, Fraud detection, Auto insurance, Classification.
MSC2010 Classifications: 91B30, 68T09.

1 Introduction

Insurance frauds refer to those groups of behaviors committed by a fraudster with the goal of illegal financial profiteering. In doing so, different and various tactics are devised and utilized, including providing incorrect information, forgery of documents and evidence, collusion, and creating cases of exaggerated or intentional losses and damages. It seems obvious that the consequences of such actions do not just affect insurance companies, and its financial pressure will subsequently influence the insureds negatively by enhancing the premiums. Car insurance scams are known to be highly common, and according to a survey conducted in 2017 by Nerdwallet Co., one in ten Americans provides incorrect information to the insurance

[2]Corresponding author
Received: 22/06/2022 Accepted: 25/07/2022
http://dx.doi.org/10.22054/jmmf.2022.68662.1058
An American company in the field of finance and a provider of reports and materials related to various financial products, including credit cards, banking affairs, investment, and insurance.

company when applying for a car insurance policy. Lying about the annual traveled kilometers (mileage) is recognized as one of the most common forms of frauds regarding that 40% of those surveyed have admitted to reporting less than the actual distance traveled at the time of issuance of the policy insurance; also, 27% of the individuals have excluded one of their licensed drivers from their insurance policy, and in 20% of cases, incorrect information has been provided to the insurance company about the car that the person is using. Insurers have decided to pay particular attention to the scientific research done in the field of insurance frauds due to the heavy financial consequences of such frauds for insurance companies. Detecting and identifying frauds is indeed one of the most challenging issues in the world. In this regard, the use of artificial intelligence and machine learning algorithms seem to be dramatically needed more than ever with the ultimate goal of enhancing the accuracy of identifying fraudulent cases. Machine learning algorithms are often classified into supervised and unsupervised categories. The training data are labeled in the supervised algorithms, or in other words, the dataset has a target variable with a label tagged. In contrast, existing labels are not shown to the training algorithm in the unsupervised algorithms and there is no dependent variable, and all features play the role of independent variables ([8]).

2 Literature Review

Machine learning is a new approach to extracting knowledge from data. Machine learning is an interdisciplinary skill equipped with specializations in statistics, artificial intelligence, and computer science, which is also referred to by terms such as statistical learning and predictive analyses ([12, 15, 26]). The first applications of the machine learning topic may be traced back to the 1990s with the advent of the email filtering system. At that time, computers learned how to distinguish and separate spam emails from non-spam emails without human intervention ([10, 21]). In general, some of the machine learning applications are as follows:

(i) In problems where achieving the existing solutions require a great deal of time and writing a long list of rules.

(ii) In complex problems with no specific algorithm and solution available for solving them.

(iii) In problems that face many changes and fluctuations; in such conditions, machine learning algorithms can adapt themselves to changes quickly.

(iv) Extracting templates and information from a large amount of data.

Machine learning methods can be also classified into four main categories based on the amount and type of supervision as follows:

(i) Supervised learning

(ii) Unsupervised learning

(iii) Semi-supervised learning

(iv) Reinforcement learning

In the supervised learning, the system assumes that the desired answer to our problem is not currently available but exists in the historical data we have, and the task of the supervised learning algorithm is to identify and find the solution and achieve it through historical data ([9]). Unsupervised algorithms are those algorithms in which the input data is known while no output data is given to the algorithm. In these circumstances, the unsupervised algorithm searches specific and certain patterns in the dataset ([14]). Some algorithms are also capable of working with training data that is not thoroughly labeled. In fact, labeled data is combined with more unlabeled data in the algorithm. The semi-supervised learning is used partly due to the high cost and time required for the labeling process. The algorithms for this method of machine learning are usually a combination of supervised learning and unsupervised learning ([29]). The reinforcement machine learning algorithm is trained with the approach of learning from the environment, and if it's done well, it will receive rewards, and its ultimate goal is to maximize those rewards ([27]).

Numerous scientific studies have been conducted on the discovery of car insurance frauds since the 1990s, including the research by [6,11,17]. Remarkable efforts have been made in recent years using data mining methods to detect and identify damages and car insurance frauds. For example, classification algorithms such as the decision tree, neural networks, logistic regression, support vector machine, and genetic algorithm have been utilized to identify suspected fraud cases in the studies performed by [2, 3, 7, 23–25, 28]. Some of the research done by using supervised algorithms to detect car insurance frauds are also given in Table 1. However, unsupervised methods and algorithms are known as the most extensively used strategies in the insurance industry and especially in the field of fraud detection ([13, 16]). This arises from the predominant structure associated with the dataset. In fact, the amount of tagged (labeled) data is not high given the cost and time constraints, and many fraud detection projects inevitably do their research on unlabeled data, and through unsupervised algorithms. For example, the clustering algorithm has been utilized to identify suspicious cases in the research by [19, 20, 22]. Also, social networks have been analyzed to detect frauds in a paper by [4]. [18] has focused on identifying anomalous cases in his paper. The Local Outlier Factor (LOF) algorithm is one of the unsupervised algorithms, which is capable of providing very good results.

There are two differences in this research compared to other articles: the first is the use of an unsupervised method to detect outlier points. Most of the researches in this field have used supervised algorithms, however those that have employed unsupervised learning have limited themselves to classical clustering models. The

Table 1: The results of using supervised algorithms to detect car insurance frauds

Model used	Year	Author
(Decision Tree) 78%	2011	Bhowmik
(Logistics) 67% (Neural network) 77%	2018	Kowshalya and Nandhini
(Support vector) 59% (Neural network) 79% (Decision tree) 60%	2018	Subudhi and Panigrahi
(Decision tree) 85% (K-nearest neighbor) 72% (Support vector) 72% (AdaBoost) 80% (Random forest) 83% (Naïve Bayes) 81%	2022	Rukhsar et al.

second difference of this paper is the use of the data set that includes a wide range of independent variables. In fact, the variables describe the features related to the culprit, victim, car, accident and the insurance policy.

3 Methodology

Detecting anomalies is recognized as one of the major applications of unsupervised algorithms. The concept of anomaly refers to the patterns that differ from known patterns and records ([8]). Detecting anomalies is dramatically important, which often provides applied and critical information in different areas. The absence of a response or dependent variable is the prominent feature of this type of machine learning method. In fact, these algorithms seek to discover the patterns and relationships within the datasets. In other words, only a set of features is received as input (x) and the data output is not already available to the analyst. Then, these models look for hidden patterns in the unlabeled datasets ([8, 15]). Clustering-related algorithms are known as the most important and common types of these methods. In this paper, due to the available data structure and the absence of response variable, we have employed one of the unsupervised algorithms which is briefly introduced below.

3.1 Local Outlier Factor

The Local Outlier Factor (LOF) algorithm, which is considered a clustering model, is an approach that considers neighbors of a particular point and examines its density, and then compares it with the density of other points ([11]). This algorithm

actually detects the local outlier points; i.e., the outlier and abnormal points are compared to their neighbors, and not to the distribution of the total data. An instance of this method is illustrated in Figure 1:

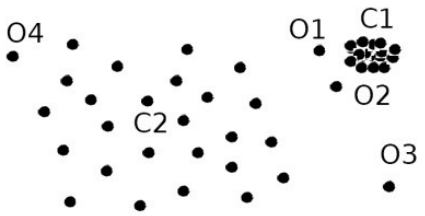

Figure 1: Anomalies in the Local Outlier Factor Algorithm

In Figure 1, O1 and O2 are local outlier data compared to C3, and O3 is a global outlier data point. The k-nearest neighbor method is used in this algorithm. For a point such as x in the dataset, its k-nearest neighbor is defined as follows:

$$N_k(x) = \{y | y \in D, dist(x, y) \leq dist_k(x)\}.$$

This set encompasses all points whose distance from x is less than their k- distance from that point. Pay attention to the following definition to better understand the above relationship.

Definition 3.1. The distance between two data points, p and o can be calculated using the n-dimensional Euclidean space as follows:

$$dist(p, o) = \sqrt{\sum_{i=1}^{n} (p_i - o_i)^2} \ .$$

Now, consider the dataset D and a positive integer k. Then, for a point like p, the value of k- distance p is equal to the distance, $dist(p, o)$, between p and the farthest point of the neighboring data (sample) such as o ($o \in D$) that holds in the following conditions:

i. Minimum k data point (sample) such as $o' \in D - \{p\}$ retains the inequality, $dist(p, o') \leq dist(p, o)$.

ii. Maximum k-1 data point (sample) such as $o' \in D - \{p\}$ retains the inequality, $dist(p, o') < dist(p, o)$.

Using this algorithm requires two other concepts known as Reachability Distance and Local Reachability Density, which are defined as follows:

Definition 3.2. Suppose k is a positive integer. The reachability distance of a data point x from the data point y is defined as the following equation:

$$reachdist_k(x \leftarrow y) = \max\{dist_k(x), dist(x, y)\}.$$

Figure 2 shows an example of a reachability distance when the value of k is equal to 5. If the actual distance between the data point p4 and the data point o is shorter

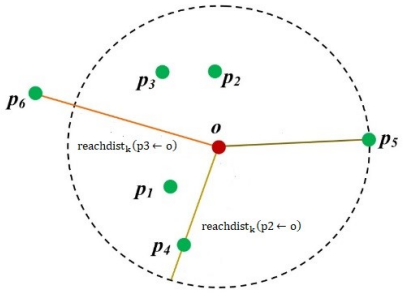

Figure 2: Reachability Distance for k=5

than the distance $dist_k(o)$, the reachability distance of the data point p4 will be $dist_k(o)$. On the other hand, if the actual distance between the data point p6 and data point o is longer than the distance $dist_k(o)$, then, the reachability distance of the data point p6 is the same as the actual distance. Such a smoothing operation is utilized to decrease the statistical fluctuations of $dist(p, o)$ for points close to o. The smoothing rate is controlled by the value of k.

Two parameters are used to define the concept of density in the density-based clustering algorithms namely; 1. the minimum number of data points and 2. the volume. Here, we define the local reachability density as the following equation using the concept of the minimum number of points:

$$lrd_k(x) = \frac{||N_k(x)||}{\sum_{x \in N_k(x)} reachdist_k(x \leftarrow y)}.$$

Finally, we come to the definition of the Local Outlier Factor (LOF), which determines anomalies:

$$LOF_k(x) = \frac{\sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)}}{||N_k(x)||} = \sum_{y \in N_k(x)} lrd_k(y) \cdot \sum_{y \in N_k(x)} reachdist_k(y \leftarrow x).$$

In fact, the LOF for an observation is the average ratio of the reachability distance of the mentioned observation and its k-nearest neighbors. The larger the number than 1, the more likely the data would belong to an anomalous dataset. This algorithm is a technique that seeks to utilize the idea of the k-nearest neighbors to detect outliers. A score is assigned to each sample in this method. This score is based on the degree of isolation or the likelihood of being an outlier of that data due to the size of the sample local neighborhood. The samples with the highest scores are probably the outlier data. The content presented in this section has been examined more accurately and inclusively in the research done by [1, 5].

3.2 Dataset

The datasets containing 50,000 samples, 2008-2009, were used in this paper regarding insurance claims. The number of variables used are 15, whose descriptions are provided in Table 2.

Table 2: list of variables

Code	Variable Name
CtyNam	City name
InCty	Area in which an accident happened
MapCarTypCod	Type of vehicle
Age	Age of culprit
AgeLoser	Age of victim
IsLcnsFit	Type of license
UsgCod	Car usage
IsMale	Gender of victim
CusMaleCod	Gender of culprit
CarGrpCod	Car group
ThrLosTyp	Type of damage to victim
AcdTypCod	Type of accident
CtyNam	County

The features can be divided into four main groups namely, car, culprit, victim, accident and the policy features. Each variable has different number of classes as follows: CtyNam (18), MapCarTypCod (5), InMale (2), CusMaleCod (2), AgeLoser (8), Age (8), Days (4), Hour (7), PrdDte (5), CarGrpCod (22), ThrLosType (6), IsLcnsFit (2), AcdTypCod (6) , InCty (2) and CtyName (22). It is worth noting that some of the variables have been coded from the beginning by central insurance of Iran and provided to the research team. The rest were categorized and coded by the authors as needed.

4 Findings

Doing preprocessing on the data is one of the most important and main steps of data analysis and pattern extraction from the dataset. In fact, preprocessing helps to enhance the accuracy and efficiency of the model. The preprocessing stages depend on the available datasets and may vary from one dataset to another. The following measures were done in this regard according to the data set.

- Combining the features
  The inputs related to two or more variables were combined and integrated aimed at achieving applied and analyzable information.

- Eliminating noise from the dataset
  A feature with conflicting information had been filled for some points in the dataset. For example, in some records, the pickup truck has been put in both the passenger and cargo groups. These cases were either corrected or deleted.
- Assigning numerical codes to variables
  Appropriate coding was made for nominal variables (such as vehicle type, type of use, etc.).
- Grouping of numerical independent variables
  Appropriate grouping was done for some continuous variables to enhance the efficiency of the model.

After entering the datasets as the input of the algorithm, 762 samples of anomalies were identified and extracted as the output of the algorithm. The principal component analysis method was employed to display anomalies, which results are depicted in Figure 3.

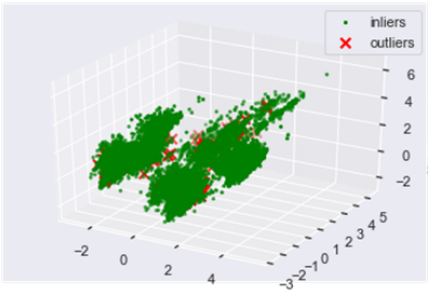

Figure 3: Anomalies and normal points

As seen in Figure 3, the points with unusual behavior that are identified as anomalies are marked in red and normal samples are marked in green. Note that these outliers are related to those accidents that have very similar characteristics to normal claims. This makes it difficult to identify suspicious cases, even for experts. However, the density based mechanism of the LOF provides a strategy to distinguish the hidden effective features. This can be shown by comparison between different claims.

Bellow, we compare the trend in anomalous points with the normal cases based on the algorithm outcomes:

In Figure 4, the difference between normal claims (right column) and fraudulent claims (left column) can be seen for all features used in the analysis. As one can see, in some cases, there is no significant differences between normal and fraudulent claims with respect to a single variable. For example, concerning type of vehicle, the majority of cars belong to the same class regardless of whether the claim is suspicious or not. This is presented in more detail in the next section.

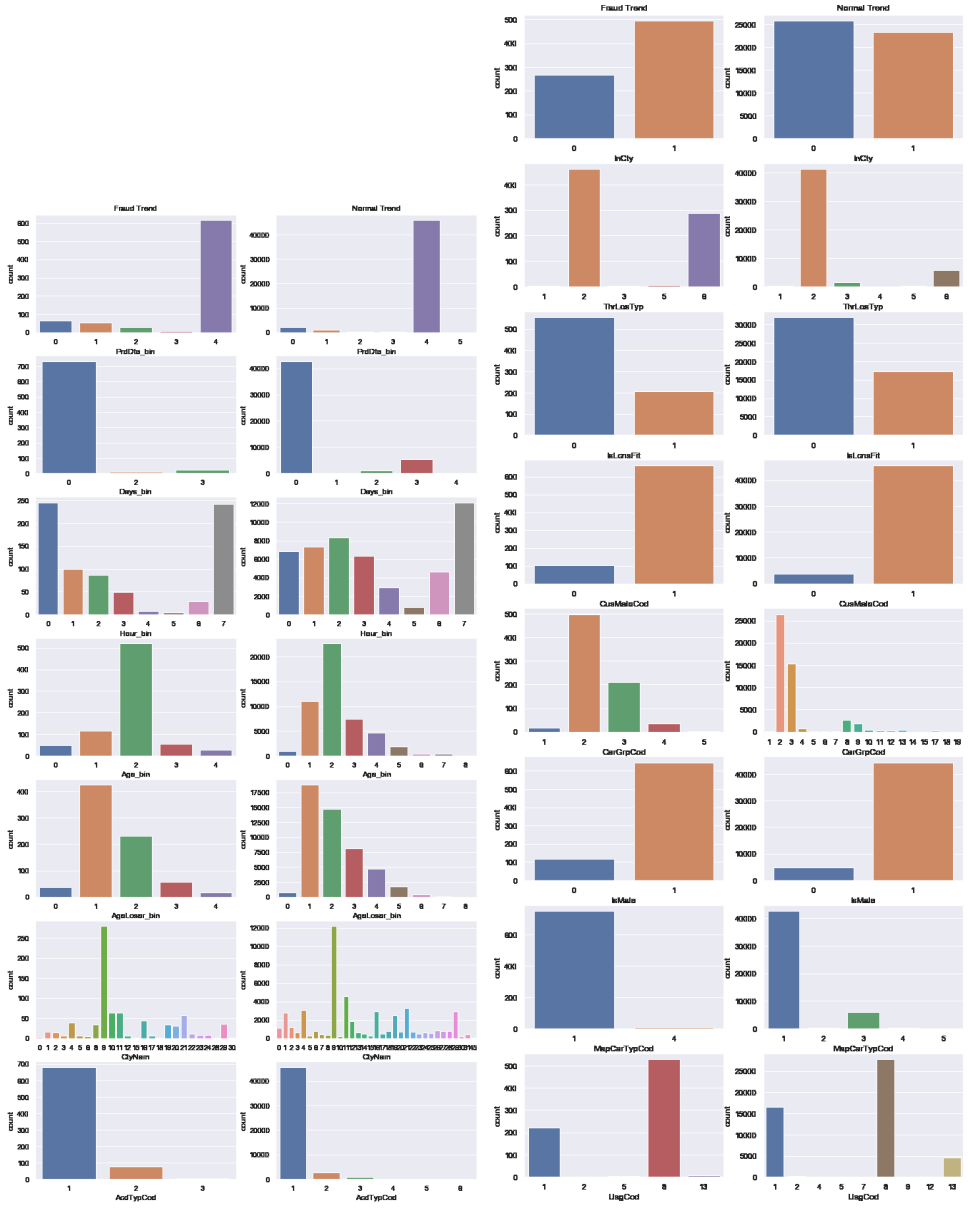

Figure 4: Structure of features for suspected and normal claims

5 Discussions

The vast majority of studies regarding fraud detection are conducted based on supervised algorithms. This is due to the easier evaluation of the model performance and avoiding the complexities associated with unsupervised algorithms. However,

the accuracy and reliability of the supervised models are doubtful due to the nature of the datasets available in the field of insurance frauds. In this regard, this paper is one of the few studies that has focused on detecting fraud patterns through unsupervised algorithms.

The local outlier factor algorithm was employed to detect suspicious frauds in this paper. This algorithm indeed helps professionals to discover unusual patterns in those claims which apparently seem very normal. This algorithm is capable of discovering highly suspicious files that are kept so hidden as to not be identified even by highly experienced claim adjusters. Based on the model results, the influential variables in identifying suspicious cases were extracted as follows:

Out of normal claim cases, 10% include financial losses, while in suspicious cases this number reaches 62%. In other words, the frequency of financial losses in fraudulent cases is higher than 50% compared to the normal ones. This difference is less (20%) concerning the incident occurrence variable; however, it is still seen as a sound indicator for the model to predict and detect suspicious cases. In fact, among the normal and fraud suspected cases, 12% and 32% of incidents have occurred between 12 pm and 3 pm, respectively. No significant differences were found between suspicious and normal cases regarding individuals gender. Another factor contributing to identifying the hidden patterns of fraud is the type of incident variable. In this regard, only 2% of the damages are related to vehicle collisions with humans in normal cases, while this type of incident in suspicious cases accounts for 13%.

The main models, used in most research papers concerning insurance frauds in car insurance, are decision tree, neural network, Naïve Bayes, logistic model, support vector, and the k-nearest neighbor. On average, the lowest and highest accuracy rates of these models are obtained as 40% and 86%, respectively. The fluctuation found between the accuracy rates of different models has made the use of the supervised algorithms face a great challenge. Besides the small number of the labeled data, the low number of the total samples in the dataset (in some papers, less than 20,000 samples) also exacerbates the randomness of the model results. This has led the researchers to preferably use unsupervised models such as the LOF algorithms. The variable of the time of the incident was also examined in this report as an influential factor in identifying suspicious cases, which may be seen as an innovation compared to common variables found in most studies (such as age, gender, history of insurer claims).

It is worth noting that these results can be provided to the relevant experts and professionals as a guide on the importance of access to appropriate variables for predicting the model. The dynamic nature of machine learning models is obviously one of their strengths; thus, over time, the proposed model has the ability to provide new variables for identifying suspicious cases. In this regard, insurance companies are recommended to use this algorithm as an effective artificial intelligence system in identifying suspected cases of damage to gradually discover the definitive fraud

files and add them to the labeled datasets. This allows us to use the integration of supervised and unsupervised algorithms in the future for identifying fraud cases more accurately and, as a result, prevent insurance companies from losing money and increasing premiums.

Bibliography

[1] O. Alghushairy, R. Alsini, T. Soule, X. Ma, *A review of local outlier factor algorithms for outlier detection in big data streams*, Big Data and Cognitive Computing, 5(1), (2020) 1.

[2] M. Asghari Oskoei, F. Khanizadeh, A. Bahador, *Application of Data Mining through Machine Learning Algorithms to Study Effect of Car Features in Predicting Financial Claim of Motor Third Party Liability Insurance*, Iranian Journal of Insurance Research, (2020) 35(1) (in Persian).

[3] T. Badriyah, L. Rahmaniah, I. Syarif, *Nearest neighbour and statistics method based for detecting fraud in auto insurance*, In 2018 International Conference on Applied Engineering (ICAE) , (2018) 1-5.

[4] A. Bodaghi, B. Teimourpour, *Automobile insurance fraud detection using social network analysis*, In: Moshirpour M., Far B., Alhajj R. (eds) Applications of Data Management and Analysis. Lecture Notes in Social Networks. Springer, Cham. , (2018) 11-16.

[5] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander, *LOF: identifying density-based local outliers*, In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, (2000) 93-104.

[6] L. Caron, G. Dionne, *Insurance fraud estimation: more evidence from the Quebec automobile insurance industry*, In Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation, (1999) 175-182.

[7] S.B. Caudill, M. Ayuso, M. Guillén, *Fraud detection using a multinomial logit model with missing information*, Journal of Risk and Insurance, 72(4), (2005) 539-550.

[8] V. Chandola, A. Banerjee, V. Kumar, *Anomaly Detection: A Survey*, ACM Computing Surveys. vol, 41, (2009) 15.

[9] P. Cunningham, M. Cord, S.J. Delany, *Supervised learning. In Machine learning techniques for multimedia* , Springer, Berlin, Heidelberg (2008) 21-49.

[10] E.G. Dada, J.S. Bassi, H. Chiroma, A.O. Adetunmbi, O.E. Ajibuwa, *Machine learning for email spam filtering: review, approaches and open research problems*, Heliyon, 5(6), (2019) e01802.

[11] M.I. Dixon, *Recent initiatives in the prevention and detection of insurance fraud*, Journal of Financial Crime, (1997).

[12] I. El Naqa, M.J. Murphy, *What is machine learning?*, In machine learning in radiation oncology (pp. 3-11). Springer, Cham, (2015).

[13] M. Goldstein, S. Uchida, *A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data*, PloS one, 11(4), (2016) e0152173.

[14] T. Hastie, R. Tibshirani, J. Friedman, *Unsupervised learning*, In The elements of statistical learning, (2009) 485-585.

[15] G. James, D. Witten, T. Hastie, R. Tibshirani, *An introduction to statistical learning (Vol. 112, p. 18)*, New York: springer, (2013).

[16] F. Khanizadeh, F. Khamesian, A. Bahiraie, *Customer Segmentation for Life Insurance in Iran Using K-means Clustering*, International Journal of Nonlinear Analysis and Applications, 12(Special Issue), (2021) 633-642.

[17] O.M. Kurland, *Combating insurance fraud*, Risk Management, 39(7), (1992) 52-54.

[18] K. Nian, *Unsupervised Spectral Ranking for Anomaly Detection*, (Master's thesis, University of Waterloo), (2014).

[19] K. Nian, H. Zhang, A. Tayal, T. Coleman, Y. Li, *Auto insurance fraud detection using unsupervised spectral ranking for anomaly*, The Journal of Finance and Data Science, 2(1), (2016) 58-75.

[20] S.M. PALACIO, *Abnormal pattern prediction: Detecting fraudulent insurance property claims with semi-supervised machine-learning*, Data Science Journal, 18(1), (2019).

[21] J. PROVOST, *Nave-bayes vs. rule-learning in classification of email*, University of Texas at Austin, (1999).

[22] S. SUBUDHI, S. PANIGRAHI, *Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection*, Journal of King Saud University-Computer and Information Sciences, 32(5), (2020) 568-575.

[23] M. VASU, V. RAVI, *A hybrid under-sampling approach for mining unbalanced datasets: applications to banking and insurance*, International Journal of Data Mining, Modelling and Management, 3(1), (2011) 75-105.

[24] S. VIAENE, R.A. DERRIG, B. BAESENS, G. DEDENE, *A comparison of state of the art classification techniques for expert automobile insurance claim fraud detection*, Journal of Risk and Insurance, 69(3), (2002) 373-421.

[25] S. VIAENE, R.A. DERRIG, G. DEDENE, *A case study of applying boosting Naive Bayes to claim fraud diagnosis*, IEEE Transactions on Knowledge and Data Engineering, 16(5), (2004) 612-620.

[26] H. WANG, Z. LEI, X. ZHANG, B. ZHOU, J. PENG, *Machine learning basics*. Deep learning, (2016) 98-164.

[27] M.A. WIERING, M. VAN OTTERLO, *Reinforcement learning*, Adaptation, learning, and optimization, 12(3), (2012) 729.

[28] W. XU, S. WANG, D. ZHANG, B. YANG, *Random rough subspace based neural network ensemble for insurance fraud detection*, In 2011 Fourth International Joint Conference on Computational Sciences and Optimization, (2011) 1276-1280.

[29] X. ZHU, A.B. GOLDBERG, *Introduction to semi-supervised learning*, Synthesis lectures on artificial intelligence and machine learning, 3(1), (2009) 1-130.

How to Cite: Maryam Esna-Ashari[1], Farzan Khamesian[2], Farbod Khanizadeh[3], *Using local outlier factor to detect fraudulent claims in auto insurance*, Journal of Mathematics and Modeling in Finance (JMMF), Vol. 2, No. 1, Pages:137–148, (2022).