Using Some Data Mining Approaches with Application on Insurance Data

Research extracted from a Master thesis of Insurance

By

Dr. Aya Shehata Mahmoud

Lecturer of Statistics, Mathematics, and Insurance

Faculty of Commerce, Benha University

Dalia Sherif Shaban

Demonstrator of Statistics, Mathematics, and Insurance

Faculty of Commerce, Benha University

dalia.mousa@fcom.bu.edu.eg

Dr. Zohdy Mohammed Nofal

Professor of Statistics, Mathematics, and Insurance

Faculty of Commerce, Benha University

Scientific Journal for Financial and Commercial Studies and Research (SJFCSR)

Faculty of Commerce – Damietta University

Vol.5, No.2, Part 1., July 2024

APA Citation:

Mahmoud, A. S.; **Shaban**, D. S. and **Nofal**, Z. M. (2024). Using Some Data Mining Approaches with Application on Insurance Data, ***Scientific Journal for Financial and Commercial Studies and Research***, Faculty of Commerce, Damietta University, 5(2)1, 913- 937.

Website: https://cfdj.journals.ekb.eg/

Using Some Data Mining Approaches with Application on Insurance Data

Dr. Aya Shehata Mahmoud؛ Dalia Sherif Shaban and Dr. Zohdy Mohammed Nofal

Fraud considered as the most common problem in insurance companies. Detecting frauds is a difficult problem for insurance companies. This study presents a statistical and data mining techniques. The statistical and data mining techniques helps in predicting fraud in this data. The data was cleaned and pre-processed by removing duplication, filling the missing data, managing the categorical data by label encoding and detecting the outliers. Then the data was split into train and test data. After that, using the standardization feature scaling for the data. Finally, the data was evaluated by some data mining models and the best two models are the Adaptive Boost and Gradient Boost. The Ada Boost model achieves the highest values of accuracy (95.556%), recall (92.308%), precision (87.805%), F1_score (90%) and MCC (Matthews Correlation Coefficient) (87.190%). the Gradient Boost model achieves the second highest values of accuracy (92.778%), recall (76.923%), precision (88.235%), F1_score (82.192%) and Matthews Correlation Coefficient MCC (77.976%). So, a new model was proposed in this research called GA which is a combination of Gradient Boost and Adaptive Boost by the hybrid classifier.

Keywords

Data mining, Fraud Detection, Car Insurance, Hybrid Classifier

1- Introduction

Insurance fraud is a major issue. It's challenging to recognize fraud allegations. We will use some data from auto insurance to show how we can build a prediction model that can determine whether or not an insurance claim is a fraud. Numerous research projects have been conducted on data mining techniques.

Numerous programmers and mathematicians use a variety of techniques to solve the problem. Data mining made this feasible by offering a methodology that teaches computers to identify patterns in data rather than pre-programming them with equations that represent these patterns. We were able to improve our ability to identify patterns in higher dimensions, offer statistical backing for our research, and make predictions in a variety of domains by teaching computers to recognize patterns in data. For many years, the insurance industry has employed a variety of data mining algorithms, including Decision Tree (DT), Random Forest (RF), Adaptive Boost (Ada Boost), Gradient Boost (GrBoost) and Extra Tree (ET), for classifying data. These algorithms have been successful in characterizing, analyzing, and properly predicting the outcomes.

This paper is organized into seven sections where Section 2 represents related work. Section 3 illustrates the methods and materials of this paper for predicting fraud and this section is divided into four subsections: dataset collecting, data pre-processing (filling the missing data, label encoding, and detecting the outliers), data visualization, and feature scaling technique. Section 4 represents the building of data mining models. Section 5 illustrates the model evaluation and selection. Section 6 represents the proposed new model of the insurance data. Finally, Section 7 represents conclusion of the entire work of this paper.

2- Related work

Various studies have been conducted on the prediction-making process and the assessment of data mining model performance in categorization across various domains, including insurance science. An overview of a few of these publications that review work related to insurance science and data mining models is provided below:

Bhowmik (2011) predicted and presented fraud using decision tree-based algorithms and naïve Bayesian classifiers. He examined the confusion matrix-derived model performance metrics. Confusion matrix-derived performance measurements include accuracy, recall, and precision. Because of its significantclass skew, it is a trustworthy performance metric in many crucial fraud detection application domains.

Tao et al. (2012) developed a dual membership fuzzy support vector machine model for the purpose of identifying insurance fraud. Each sample is given a dual membership during the SVM training process based on the distance between the sample mean vector and itself; the dual membership that is assigned can be used to describe the imprecision of insurance fraud data. The empirical findings demonstrate that the fuzzy support vector machine model with dual membership outperforms other conventional insurance fraud identification models.

Senousy et al. (2019) offered a compelling and original model that explains how the Egyptian social insurance dataset is pre-processed using supervised learning methods. For the purpose of determining which of the three algorithms is more accurate and efficient, they have selected the Decision Tree, Naïve Bayes, and CN2 Rule Inducer algorithms. Following algorithm application, the outcomes demonstrated that the Decision Tree and CN2 Rule Inducer algorithms outperform the Naïve Bayes algorithm in predicting which individuals are covered by the social insurance program and which are not.

Abdelhadi et al. (2020) focused on cutting-edge statistical approaches and data mining algorithms that are the best approach for handling missing information in order to create a precise model to anticipate auto insurance claims using machine learning techniques. They developed the prediction model utilizing Extreme Gradient Boosting (XGBoost), Decision Trees (DT), Naïve Bayes classifiers, Artificial Neural Networks (ANN), and Kaggle's public datasets, which comprise 30240 cases and 12 variables. The outcomes of the experiment demonstrated that the model produced appropriate results. Of the four models, the XGBoost model and Resolution Tree had the highest accuracy, with 92.53% and 92.22%, respectively.

3- Methods and material

This section describes the methodology of the proposed model.

Data Collection

Collecting data for applying data mining models is the first step in data mining pipeline. The "insurance_claims.csv" dataset is an extensive compilation of insurance claim documentation. A claim is represented by each row, and its numerous attributes are represented by the columns. The dataset highlights attributes such as policy number, age, months_as_customer, and so forth. The fraud reported variable is the primary focus.

Data on claims were obtained from multiple insurance companies and included a wide range of insurance categories, such as auto, home, and personal injury. The record of each claim offers a comprehensive look into the claimant's history, the details of the claim, any related paperwork, and the opinions of insurance experts.

The dataset provides a detailed insight into the complexity of every claim by including particular signs and factors that were taken into account during the claims assessment. Certain identifying information has been anonymized for privacy purposes and in compliance with the participating insurance providers. Each entry is linked to a distinct ID rather than names or direct identifiers, protecting data.

Data pre-processing

Data pre-processing is an important process of data mining. It refers to the cleaning, transforming, and integrating of data. In this process, raw data is converted into an understandable format and made ready for further analysis. The aim is to improve data quality and make it up to mark for specific tasks. Table 1 represents a statistical analysis of the fraudulent dataset for numerical features. Table 2 represents a statistical analysis of the fraudulent dataset for categorical features.

Table 1 Statistical analysis of the fraudulent dataset for numerical features.

Features	Mean	Std.	Min	Max
months_as_customer	203.954000	115.113174	0.000000	479.000000
age	38.948000	9.140287	19.000000	64.000000
policy_number	546238.648000	257063.005276	100804.000000	999435.0000
policy_deductable	1136.000000	611.864673	500.000000	2000.00000
policy_annual_premium	1256.406150	244.167395	433.330000	2047.59000
insured_zip	501214.48800	71701.610941	430104.00000	620962.000
capital-gains	25126.100000	27872.187708	0.000000	100500.000
capital-loss	-26793.70000	28104.096686	-111100.0000	0.000000
incident_hour_of_the_day	11.644000	6.951373	0.000000	23.000000
number_of_vehicles_involved	1.83900	1.01888	1.00000	4.00000
bodily_injuries	0.992000	0.820127	0.000000	2.000000
witnesses	1.487000	1.111335	0.000000	3.000000
total_claim_amount	52761.94000	26401.53319	100.00000	114920.000
injury_claim	7433.420000	4880.951853	0.000000	21450.0000
property_claim	7399.570000	4824.726179	0.000000	23670.0000
vehicle_claim	37928.950000	18886.252893	70.000000	79560.0000
auto_year	2005.103000	6.015861	1995.000000	2015.00000

Table 2 Statistical analysis of the fraudulent dataset for categorical features.

Features	Class	Count
police_report_available	NO YES	686 314
policy_state	OH IL IN	352 338 310
policy_csl	250/500 100/300 500/1000	351 349 300
insured_sex	FEMALE MALE	537 463
insured_education_level	JD High School Associate MD Masters PhD College	161 160 145 144 143 125 122

insured_occupation	machine-op-inspct	93
	prof-specialty	85
	tech-support	78
	sales	76
	exec-managerial	76
	craft-repair	74
	transport-moving	72
	other-service	71
	priv-house-serv	71
	armed-forces	69
	adm-clerical	65
	protective-serv	63
	handlers-cleaners	54
	farming-fishing	53
insured_hobbies	reading	64
	exercise	57
	paintball	57
	bungie-jumping	56
	movies	55
	golf	55
	camping	55
	kayaking	54
	yachting	53
	hiking	52
	video-games	50
	skydiving	49
	base-jumping	49
	board-games	48
	polo	47
	chess	46
	dancing	43
	sleeping	41
	cross-fit	35
	basketball	34
insured_relationship	own-child	183
	other-relative	177
	not-in-family	174
	husband	170
	wife	155
	unmarried	141
incident_severity	Minor Damage	354
	Total Loss	280
	Major Damage	276
	Trivial Damage	90

authorities_contacted	Police Fire Other Ambulance None	292 223 198 196 91
incident_state	NY SC WV VA NC PA OH	262 248 217 110 110 30 23
incident_city	Springfield Arlington Columbus Northbend Hillsdale Riverwood Northbrook	157 152 149 145 141 134 122
property_damage	? NO YES	360 338 302
police_report_available	? NO YES	343 343 314
auto_make	Saab Dodge Suburu Nissan Chevrolet Ford BMW Toyota Audi Accura Volkswagen Jeep Mercedes Honda	80 80 80 78 76 72 72 70 69 68 68 67 65 55
auto_model	RAM Wrangler A3 Neon MDX	43 42 37 37 36

	Jetta Passat A5 Legacy Pathfinder . . . 3 Series X6 M5 Accord RSX	35 33 32 32 31 . . . 18 16 15 13 12
incident_type	Multi-vehicle Collision Single Vehicle Collision Vehicle Theft Parked Car	419 403 94 84
collision_type	Rear Collision Side Collision Front Collision	470 276 254
fraud_reported	N Y	753 247

Data pre-processing includes several actions such as:

Data cleaning is the first step in data pre-process to clean the data from duplication.

Handling the missing values missing values in our data take another form as they take the form of a question mark “?” and are not completely empty (nan). Then, by knowing the variables that contain this tag “?”, we will convert them to empty values (nan) so that we can process them. We have filled in the blanks using the mode.

The encoding (managing the categorical data) label encoded was used in this data to convert the categorical variables into numerical one. Table 3 represents the label encoder for the fraudulent dataset.

Table 3 The label encoder for the fraudulent dataset.

Features	Class	Count
police_report_available	NO	0
	YES	1
policy_state	OH	0
	IL	1
	IN	2
policy_csl	250/500	0
	100/300	1
	500/1000	2
insured_sex	FEMALE	0
	MALE	1
insured_education_level	JD	0
	High School	1
	Associate	2
	MD	3
	Masters	4
	PhD	5
	College	6
insured_occupation	machine-op-inspct	0
	prof-specialty	1
	tech-support	2
	.	.
	.	.
	.	.
	protective-serv	11
	handlers-cleaners	12
	farming-fishing	13
insured_hobbies	reading	0
	exercise	1
	paintball	2
	.	.
	.	.
	.	.
	sleeping	18
	cross-fit	19
	basketball	20
insured_relationship	own-child	0
	other-relative	1

	not-in-family	2
	husband	3
	wife	4
	unmarried	5
incident_severity	Minor Damage	0
	Total Loss	1
	Major Damage	2
	Trivial Damage	3
authorities_contacted	Police	0
	Fire	1
	Other	2
	Ambulance	3
	None	4
incident_state	NY	0
	SC	1
	WV	2
	VA	3
	NC	4
	PA	5
	OH	6
incident_city	Springfield	0
	Arlington	1
	Columbus	2
	Northbend	3
	Hillsdale	4
	Riverwood	5
	Northbrook	6
property_damage	NO	0
	YES	1
police_report_available	NO	0
	YES	1
auto_make	Saab	0
	Dodge	1
	Suburu	2
	Nissan	3
	Chevrolet	4
	Ford	5
	BMW	6
	Toyota	7
	Audi	8

	Accura Volkswagen Jeep Mercedes Honda	9 10 11 12 13
auto_model	RAM Wrangler A3 Neon . . . M5 Accord RSX	0 1 2 3 . . . 36 37 38
incident_type	Multi-vehicle Collision Single Vehicle Collision Vehicle Theft Parked Car	0 1 2 3
collision_type	Rear Collision Side Collision Front Collision	0 1 2
fraud_reported	N Y	0 1

Detecting outliers the isolation forest clustering technique was used to detect the outlier in the dataset. It works based on decision tree and it isolates the outliers. If the result is -1, it means that this specific data point is an outlier. If the result is 1, then it means that the data point is not an outlier.

1. We added the scores and the anomaly columns.
2. We knew the total number of outliers.

```
Total number of outliers is: 100
Then dropping the outliers
```

3. Dropping the outliers.

Data visualization

Data visualization is the exercise of translating records into a visual context, including map or graph, to make data easier for the human brain to understand and pull insights from. The principle aim of data visualization is to make it easier to identify patterns, trends and outliers in large data sets. The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics. Figure 1 represents the correlation matrix for the fraudulent data. Figure 2 represents the features importance level of the fraudulent data.

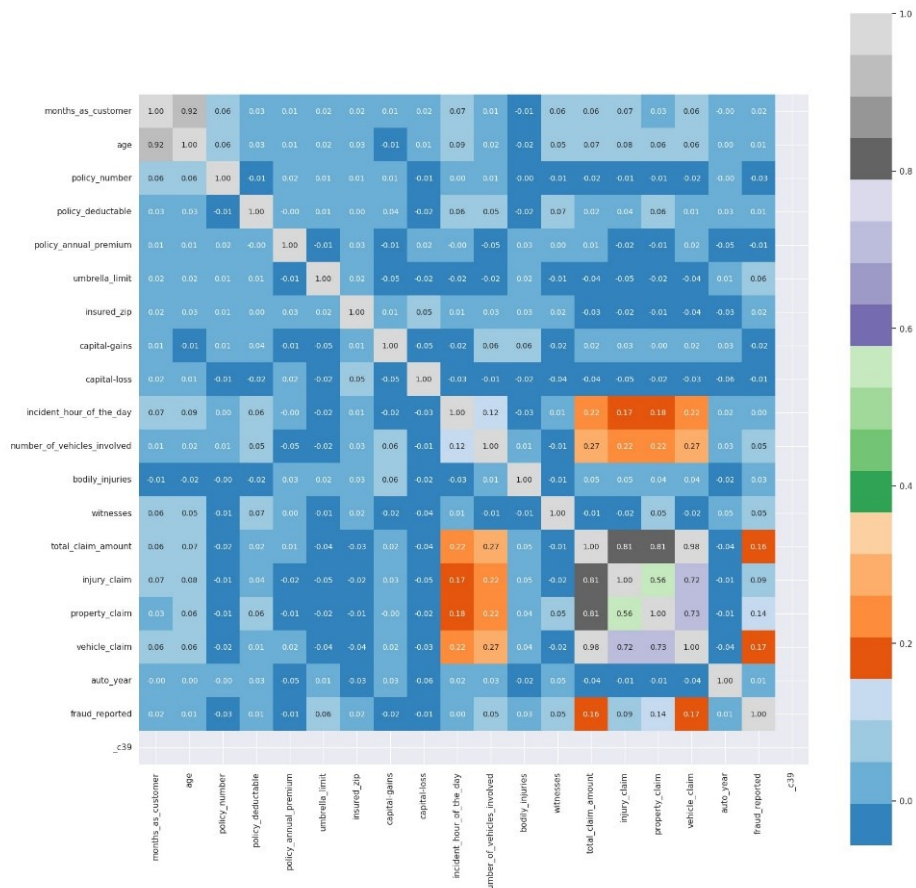


Figure 1 Correlation matrix for the fraudulent Data.

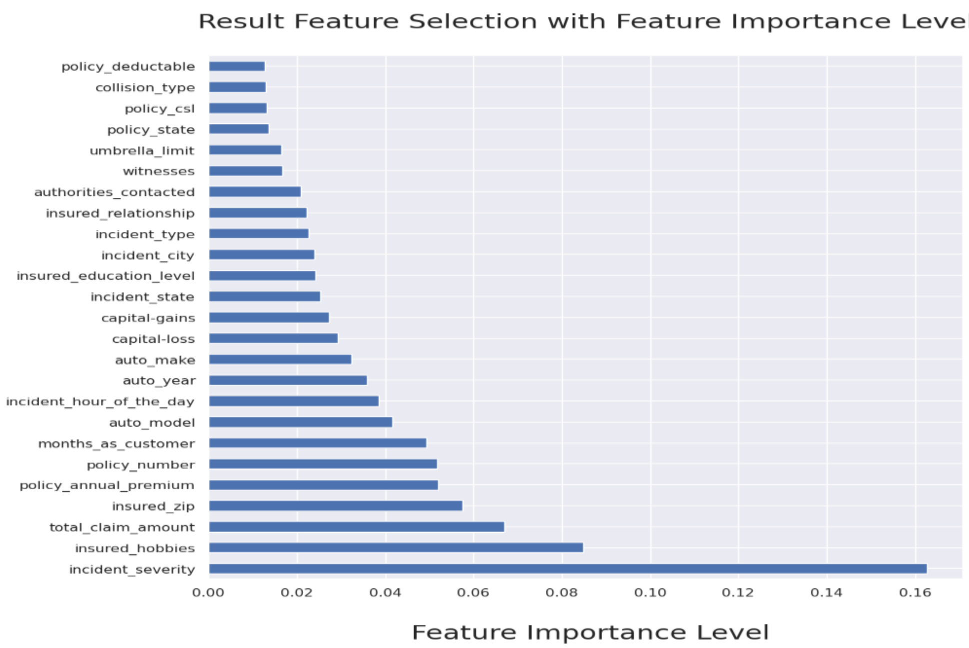


Figure 2 The importance level of features for the fraudulent data.

Feature scaling techniques

Feature scaling is the act of scaling all variables, or features, to ensure that they take values on the same scale. To achieve the scaling, we can use the feature scaling technique, which involves obtaining the feature's mean and standard deviation. We take this action to stop one characteristic from dominating the other and the data mining model from ignoring it.

Standardization: it is the method that we used in our data to be scaled and its form as follow:

$$\grave{x} = \frac{x-\bar{x}}{\sigma} \quad (1)$$

Where σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector. Standardization results in a distribution that has a standard deviation of 1 and mean of 0.

4- Model building

Data is split into two categories, referred to as training and testing data, before building of a data mining model. We never permit the exposure of testing data in order to train the model; only training data is exposed. We utilize the model to compute the predictions over the testing data after it has been trained using that data. To do this, we first define the independent variable, X, and the dependent variable, y. we will now build the data mining models by using some of data mining classification algorithms such as Decision Tree (DT), Random Forest (RF), Adaptive Boost (Ada Boost), Gradient Boost (GrBoost), and Extra Tree (ET).

Ada Boost operates in a step-by-step fashion without utilizing bootstrap sampling. Instead, each classifier is fitted on a customized version of the original dataset before being combined to generate a powerful classifier. The AdaBoost classifier is represented by the following equation:

$$d(\bar{x}_i) = sign(\sum_{j=1}^{N_c} \alpha^j C_j(\bar{x}_i)) \quad (2)$$

Where N_c is the number of base models used in this ensemble method, α^j the weight of each sub-classifier and $C_j(\bar{x}_i)$ is the predicted class of $\bar{x}_i$ by classifier C_j.

GrBoost fixes the errors of the Ada Boost models and this is the only way it differs from Ada Boost. Instead of assigning varying weights to the instances based on how accurately they were classified, the models that come after attempt to forecast the residuals of the preceding group of models. Thus, in gradient boosting, the models that come after are selected based on how little the prior ensemble of models' residual error is. The next models will concentrate on correctly classifying situations that were previously misclassified by minimizing the residual error.

Decision tree, although decision trees are a supervised learning technique, they are primarily employed to solve classification problems. A binary decision tree's construction begins at the root node, or initial decision node, as shown in figure 3.7 above. Comprises the complete dataset as well as two or more sub-trees/branches (the splitting is determined by the impurity measurements). The features of the dataset are represented by the decision nodes, the decision rules by the branches, and the classification result by each leaf node.

When the target is a classification outcome taking the values $0,1,...,i-1$, for a node j, representing a region D_j with observations N_j, P_{ji} is the proportion of class i observations in the node can be calculated as follows:

$$P_{ji} = \frac{1}{N_j} \sum_{y \in D_j} I\,(y = i) \qquad (3)$$

Random forest is a well-liked data mining algorithm that can be applied to both classification and regression problems in data mining. It is based on the idea of ensemble learning, which is the process of combining multiple classifiers to solve a complex problem and enhance the performance of the model. Random Forest is defined as a classifier that contains multiple decision trees on different subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

5- Model evaluation and selection

Model evaluation the process of assessing the models using different performance measurement criteria. The model's efficiency may be clearly shown through evaluation, which also helps in the selection of the most effective model for making predictions. In this research, as we utilizing five algorithms, “DT”, “RF”, “Ada Boost”, GrBoost”, and “ET”, These models have gone through a phase of model evaluation. For practical illustration, this study makes considerable use of the Scikit-learn Python packages. We employ multiple performance measurement metrics, such as the Confusion Matrix, which subsequently facilitates the computation of Accuracy, Precision, Recall, F1_Score, and Matthews Correlation Coefficient (MCC). Furthermore, we employ Area Under the Curve (AUC).

Table 4 Classification reports formula.

$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	$Precision = \frac{TP}{TP + FP}$
$Recall = \frac{TP}{TP + FP}$	$F1\_score = \frac{2 * Recall * Precision}{Recall + Precision}$
$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	$AUC = \frac{(x_{i+1} - x_i)(y_i + y_{i+1})}{2}$

Which TP refers to Total Positive, TN is a Total Negative, FP is a False Positive (type one error), and FN is a False Negative (type two error).

According to AUC formula, where x represents the values of FPR (False Positive Rate) and y represents the values of TPR (True Positive Rate) which the x_s must be ranked.

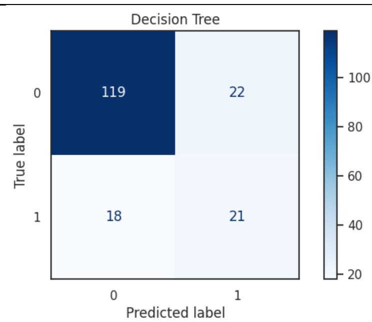
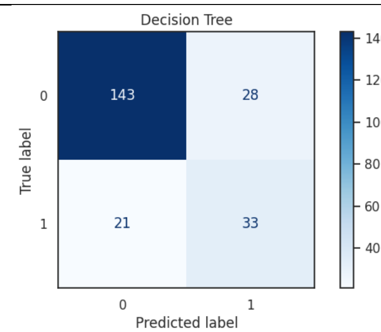
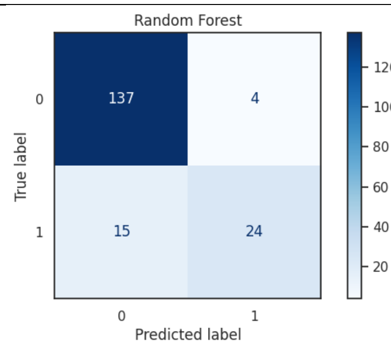
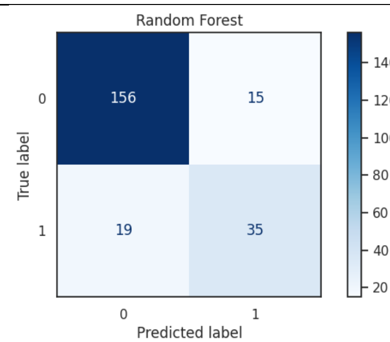
The performance of each classifier has been evaluated before and after balancing data. This data is an imbalance data. The imbalance data was converted into balance one by using SMOTE technique. The results of the evaluation are as shown below

Table 5 The imbalance data results evaluation report for fraudulent data.

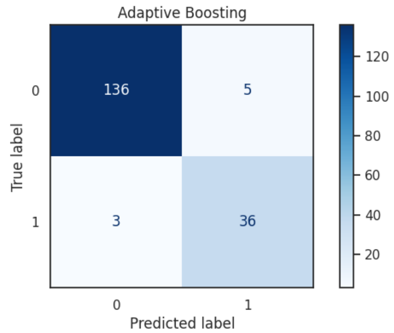
Model	Accuracy	Recall	Precision	F1_score	MCC	AUC
DT	0.77778	0.53846	0.48837	0.51219	0.36949	0.69
RF	0.89445	0.61538	0.85714	0.71642	0.66725	0.79
Ada Boost	0.95556	0.92308	0.87805	0.9	0.87190	0.94
GrBoost	0.92778	0.76923	0.88235	0.82192	0.77976	0.87
ET	0.88333	0.56410	0.84615	0.67692	0.62783	0.77

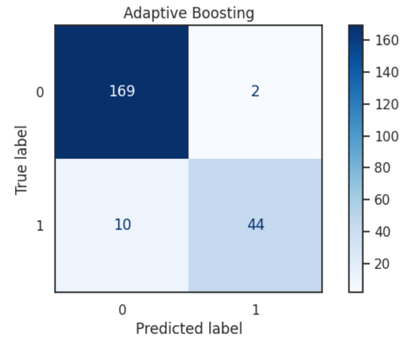
Table 6 The evaluation report for fraudulent data after applying SMOTE technique.

Model	Accuracy	Recall	Precision	F1_score	MCC	AUC
DT	0.78223	0.61112	0.54098	0.57391	0.42981	0.72
RF	0.84889	0.64815	0.7	0.67308	0.57572	0.78
Ada Boost	0.94667	0.81481	0.95652	0.88001	0.85049	0.90
GrBoost	0.90667	0.77778	0.82353	0.79999	0.73971	0.86
ET	0.84889	0.66667	0.69231	0.67925	0.58063	0.79

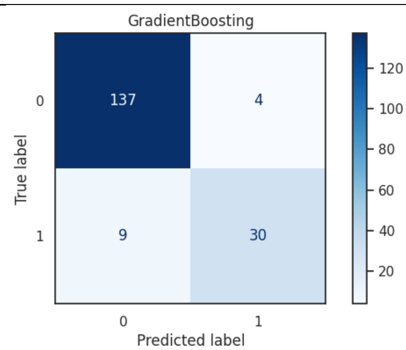
<table>
<tr><th>Model</th><th>Confusion matrix “imbalance”</th><th>Confusion matrix “SMOTE”</th></tr>
<tr><td>DT</td><td>
</td><td>
</td></tr>
<tr><td>RF</td><td>
</td><td>
</td></tr>
</table>

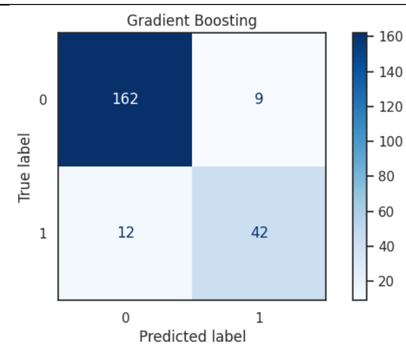
Ada Boost



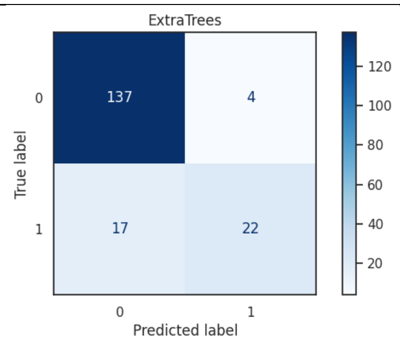


GrBoost





ET



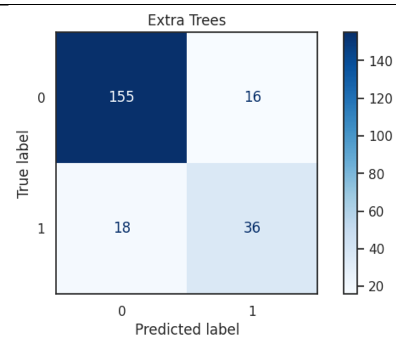


Figure 3 The confusion matrices for models of the fraudulent data before and after balancing.

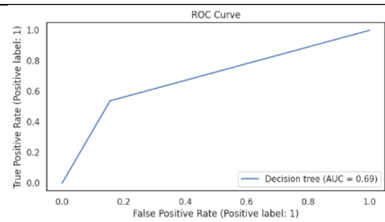
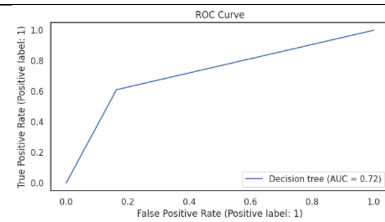
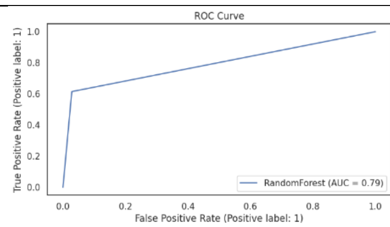
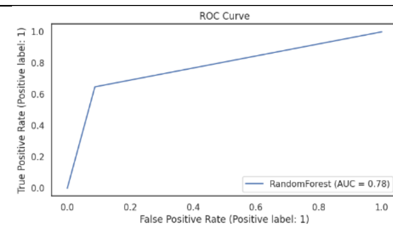
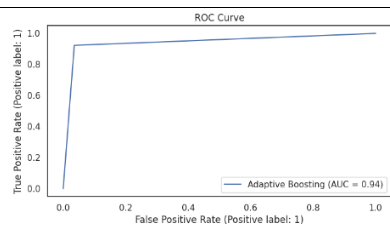
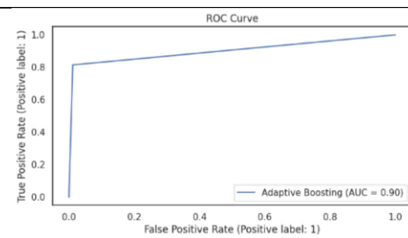
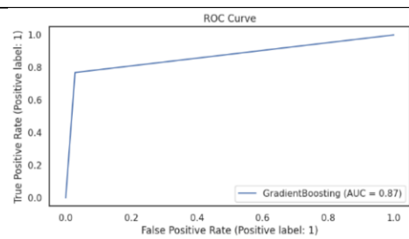
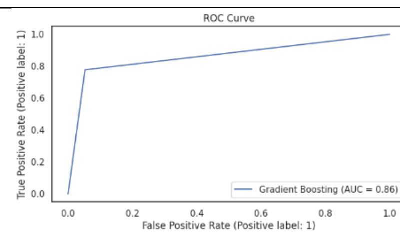
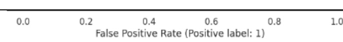
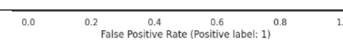
Model	ROC "imbalance"	ROC "SMOTE"
DT		
RF		
Ada Boost		
GrBoost		
ET		

Figure 4 The ROC for models of the fraudulent data before and after balancing.

6- The proposed new model

The proposed model is a hybrid model by the *voting classifier* from the ski-learn library. There are two types of voting classifier: soft and hard. in this research we used the hard voting classifier.

Hard voting, sometimes referred to as majority voting, is the straightforward process of adding up each base model's predictions and designating the class with the highest number of votes as the final forecast. It works well in classification jobs where the classes are mutually exclusive and discrete.

$$\hat{y} = argmax(N_C(y_t{}^1).N_C(y_t{}^2).\dots.N_C(y_t{}^n)) \quad (4)$$

The new model is called GA model which is a combination of two models (GrBoost+Ada Boost). Figure 5 represents the confusion matrix and ROC for hybrid classifier for the data.

- With 93.333% accuracy, we take a closer look at the confusion matrix:
- 140 transactions were classified as valid that were actually valid.
- 1 transaction were classified as fraud that were actually valid (type one error).
- 11 transactions were classified as valid that were fraud (type two error).
- 28 transactions were classified as fraud.

Table 7 represents The evaluation report for the proposed GA model of the fraudulent data.

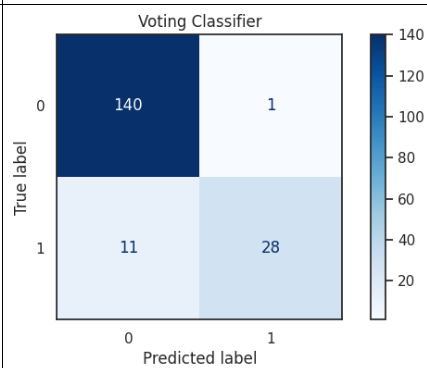
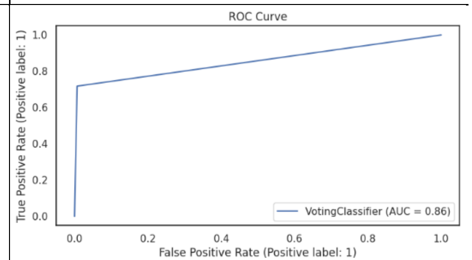
Model	Confusion matrix	ROC
Hybrid classifier (GA)		

Figure 5 The confusion matrix and ROC for hybrid classifier for the fraudulent data.

Table 7 The evaluation report for the proposed GA model of the fraudulent data.

Model	Accuracy	Recall	Precision	F1_score	MCC	AUC
The proposed GA model	0.93333	0.71795	0.96552	0.82353	0.79659	0.86

7- Conclusion

Insurance faces many problems and various objectives, and fraud is one of the insurance problems and providing insurance services is one of its objectives. Through the two types of data that we used and analyzed using data mining, as data mining is very important to predict whether fraud or to improve the quality of insurance services provided. This study uses a various of data mining classification techniques to identify and predict the target class (fraud_report). The proposed approach consists of several stages including data loading and initial exploration, data cleaning and pre-processing, Exploratory Data Analysis (EDA), feature scaling, modeling and model evaluation.

According to this data (insurance fraud detection data), the ensemble classifiers are the best models that give us the optimal results with imbalance data. The Ada Boost model achieves the highest values of accuracy (95.556%), recall (92.308%), precision (87.805%), F1_score (90%) and MCC (87.190%). the Gradient Boost model achieves the second highest values of accuracy (92.778%), recall (76.923%), precision (88.235%), F1_score (82.192%) and MCC (77.976%). By comparing the results that we obtained with the results of this research (Njeru, A. M. (2022)), we can say that we obtained higher results than it as the uppermost model result is Ada boost and XGBoost with accuracy value of 84.5%, recall value of 69%, precision value of 64% and F1_score value of 67% with imbalance dataset.

But the proposed model called GA (combination of GrBoost and Ada Boost) that applied after the SMOTE technique with hard voting classifier gives 93.333% accuracy, 71.795% recall, 96.552% precision, 82.353% F1_score and 79.659% MCC.

References

1. Abdelhadi, S., Elbahnasy, K., & Abdelsalam, M. (2020). A proposed model to predict auto insurance claims using machine learning techniques. *Journal of Theoretical and Applied Information Technology*, *98*(22).
2. Al Hammadi, S. A. (2022). Use of Data Mining Techniques to Detect Fraud in Procurement Sector (Doctoral dissertation, The British University in Dubai (BUiD)).
3. Bhowmik, R. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences*, *2*(4), 156-162.
4. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.
5. Gupta, M., & Aggarwal, N. (2010). Classification techniques analysis. In *National Conference on Computational Instrumentation* (pp. 128-131).
6. Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class adaboost. Statistics and its Interface, 2(3), 349-360.

7. Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering, 17(3), 299-310.
8. Kasture, P., & Gadge, J. (2012). Cluster based outlier detection. *International Journal of Computer Applications*, *58*(10).
9. Komarek, P. (2004). *Logistic regression for data mining and high-dimensional classification*. Carnegie Mellon University.
10. Larose, D. T., & Larose, C. D. (2014). Discovering knowledge in data: an introduction to data mining (Vol. 4). John Wiley & Sons.
11. Leopord, H., Cheruiyot, W. K., & Kimani, S. (2016). A survey and analysis on classification and regression data mining techniques for diseases outbreak prediction in datasets. *Int. J. Eng. Sci*, *5*(9), 1-11.
12. Leung, K. M. (2007). Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, *2007*, 123-156.
13. Lever, J. (2016). Classification evaluation: It is important to understand both what a classification metric expresses and what it hides. Nature methods, 13(8), 603-605.
14. Njeru, A. M. (2022). *Detection of Fraudulent Vehicle Insurance Claims Using Machine Learning* (Doctoral dissertation, University of Nairobi).
15. Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6).
16. Salmi, M., & Atif, D. (2022). Using a data mining approach to detect automobile insurance fraud. In *Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2021)* (pp. 55-66). Cham: Springer International Publishing.
17. Sarapardeh, A. H., Larestani, A., Menad, N. A., & Hajirezaie, S. (2020). *Applications of artificial intelligence techniques in the petroleum industry*. Gulf Professional Publishing.
18. Senousy, Y., Hanna, W. K., Shehab, A., Riad, A. M., El-Bakry, H. M., & Elkhamisy, N. (2019). *Egyptian Social Insurance Big Data Mining Using Supervised Learning Algorithms*. *Rev. d'Intelligence Artif.*, *33*(5), 349-357.

19. Tao, H., Zhixin, L., & Xiaodong, S. (2012). *Insurance fraud identification research based on fuzzy support vector machine with dual membership*. In *2012 international conference on information management, innovation management and industrial engineering* (Vol. 3, pp. 457-460). IEEE.
20. Tarmizi, N. D. A., Jamaluddin, F., Bakar, A. A., Othman, Z. A., Zainudin, S., & Hamdan, A. R. (2013). Malaysia dengue outbreak detection using data mining models. *Journal of Next Generation Information Technology (JNIT)*, *4*(6), 96-107.
21. Tongesai, M., Mbizo, G., & Zvarevashe, K. (2022). Insurance Fraud Detection using Machine Learning. In *2022 1st Zimbabwe Conference of Information and Communication Technologies (ZCICT)* (pp. 1-6). IEEE.
22. Zhang, C., & Ma, Y. (Eds.). (2012). Ensemble machine learning: methods and applications. Springer Science & Business Media.

استخدام بعض طرق التنقيب عن البيانات مع التطبيق على بيانات التامين

المستخلص

يعتبر الاحتيال المشكلة الأكثر شيوعا في شركات التأمين. حيث يعد اكتشاف عمليات الاحتيال مشكلة صعبة بالنسبة لشركات التأمين. فان هذه الدراسة تقدم التقنيات الإحصائية واستخراج البيانات. وتساعد التقنيات الإحصائية واستخراج البيانات في التنبؤ بالاحتيال في هذه البيانات. وتم تنظيف البيانات ومعالجتها مسبقًا عن طريق إزالة التكرار او الازدواجية وملء البيانات المفقودة وإدارة البيانات الفئوية عن طريق ترميز العلامات واكتشاف القيم المتطرفة. ثم تم تقسيم البيانات إلى بيانات التدريب والاختبار. بعد ذلك، يتم استخدام ميزة التوحيد القياسي للبيانات. وأخيرا تم تقييم البيانات من خلال بعض نماذج التنقيب عن البيانات وأفضل نموذجين هما المقياس المتطور للتكيُّف (Adaptive Boost) والنموذج المحفز للتدرج و(Gradient Boost) حيث يحقق نموذج المقياس المتطور للتكيُّف (Ada Boost) أعلى قيم الدقة (٩٥,٥٥٦٪)، والاستدعاء (٩٢,٣٠٨٪)، والدقة (٨٧,٨٠٥٪)، و(90%) F1_score، و MCC (87.190%). ويحقق نموذج المحفز للتدرج (Gradient Boost) ثاني أعلى قيم الدقة (٩٢,٧٧٨٪)، والاستدعاء (٧٦,٩٢٣٪)، والدقة (٨٨,٢٣٥٪)، و(82.192%) F1_score، و MCC (77.976%). لذلك، تم اقتراح نموذج جديد في هذا البحث يسمى GA وهو عبارة عن مزيج من Gradient Boost وAdaptive Boost بواسطة المصنف الهجين.

الكلمات المفتاحية

استخراج البيانات، كشف الاحتيال، التأمين على السيارات، المصنف الهجين.