EXPLAINABLE MACHINE LEARNING FOR INSURANCE CLAIM FRAUD DETECTION: A COMPARATIVE PYTHON-BASED STUDY

Ermira MEMETI[1*], Eftim ZDRAVEVSKI[1], Shkurte LUMA-OSMANI[1], Florinda IMERI[1]

[1]*Department of Informatics, Faculty of Natural Sciences and Mathematics, University of Tetova*
[*]*Corresponding author e-mail: ermira.memeti@unite.edu.mk*

Abstract

Today insurance fraud is a big problem for insurance companies since it is challenging to detect potential fraud cases within a large amount of claims data. Traditional rule-based models can detect already known fraud cases, but they often fail to detect new ones and can flag too many false positives. In this paper, several machine learning algorithms will be compared to detect insurance claim fraud. The analysis will focus on comparing models' performance, handling class imbalances, interpretability and feature engineering. A publicly available dataset with 1,000 cases of insurance claims will be used in the analysis and at the same time, eight classification algorithms will be considered: majority-class baseline, Gaussian Naive Bayes, k-NN, weighted logistic regression, decision tree, random forest, gradient boosting, and XGBoost. To ensure responsible modeling, the study leaves out direct identifiers and sensitive variables like sex, ZIP code, education, occupation, hobbies, and relationship status. The analysis includes missing-value treatment, date-based feature engineering, one-hot encoding, stratified train-test splitting, 5-fold stratified cross-validation, and model evaluation using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. The results indicate that XGBoost achieved the highest cross-validated F1-score, while the decision tree achieved the strongest F1-score on the holdout test set. Random forest had the highest holdout ROC-AUC, and weighted logistic regression was a useful and interpretable benchmark. SHAP analysis found that factors like incident severity, vehicle claim amount, policy age, incident state, vehicle year, annual premium, and other claim-related variables affected the model's predictions.

Keywords: insurance fraud detection; machine learning; explainable AI; XGBoost; SHAP; Python.

Introduction

Insurance companies work with large amounts of data, in which case they make decisions about claims, risk, underwriting, pricing, and customer service based on careful analysis. When we talk about such activities, we must mention that fraud detection is a key part of them. For honest policyholders, it is known that fraudulent claims can raise costs, lower profit and can lead to higher premiums. Insurance fraud can occur across many different types of insurance such as auto, health, property, and even life insurance. It can even involve exaggerated damages, staged incidents, false documents, incorrect information, or coordinated attempts to manipulate the claims process.

As everything is digitalizing, even claim records are becoming digital, in such a case insurance companies have new opportunities to use AI and ML for fraud detection. Instead of using algorithms based on rules that were created by people, now they can build models that help them identify suspicious patterns in sensitive data that are used by insurance companies. There are a lot of patterns that are hard to spot by hand and here comes machine learning which helps us to identify them. However, the word perfect does not fit with machine learning because it still comes with its own issues. The most difficult part is explaining a decision that was made, why such a decision?

In this way, there is a balance between predictive performance and model interpretability as two important aspects of modeling. Is known that complicated models might give very good results, but it is not easy for people working in one company to understand them. An example of a simpler model that exists is logistic regression. This model is easier to explain but it fails when given complex data. In the world of insurance, fairness, accountability, and transparency

matter more than ever, being able to interpret models is a practical need (European Insurance and Occupational Pensions Authority, 2021).

This paper presents a Python-based comparative analysis of machine learning focused on the domain of insurance companies. Python is used in order to compare different machine learning methods, and the key point is detecting claim fraud.

The proposed methodology uses a public insurance claims dataset and trains some classification algorithms under identical conditions. The main aim is to check if these models will be able to detect fraudulent insurance claims effectively and give clear explanations regarding the most critical features for the problem at hand?

How does having an uneven number of fraud and non-fraud cases affect how we judge model performance?

Can a model that is easy to understand also achieve high recall and give helpful explanations for fraud detection?

Which claim features have the biggest impact on classifying fraud?

This paper does not introduce a new fraud detection algorithm. Rather, it presents an analysis framework for the evaluation of machine learning methods for identification of fraudulent claims in insurance companies' domain. It starts with a description of a Python-based pipeline that includes the main steps of analysis. It covers every step, from data preparation to training and evaluation and finally the interpretation of the results. The next step is a comparison of different models, including both simple and advanced ones. The models that are included are: logistic regression, decision tree, random forest, gradient boosting, and XGBoost. The third step is the focus on evaluating model performance using metrics that are suitable for imbalanced fraud detection, such as recall, F1-score, ROC-AUC, and confusion matrix analysis, rather than relying on traditional accuracy only. Lastly, the study adds explainability and responsible AI through anonymization. It removes potentially sensitive variables and uses SHAP to explain the selected XGBoost model (Lundberg & Lee, 2017).

Related Work

Machine Learning in Insurance Fraud Detection

Nowadays the number of actual incidents have increased a lot and compared to fraud, it is much bigger but, fraud can have a big financial impact and this is why insurance fraud detection is an important research area. Old, traditional fraud detection systems usually rely on expert-defined rules, such as unusually high claim amounts, suspicious timing, sometimes even missing police reports or even repeated claims by the same claimant. Those systems can be used to identify known fraud patterns, but the problem stands in the fact that they are limited by the rules that have been manually created by people based on their needs. Based on this, as fraud behavior changes, rule-based systems know to fail while detecting new or more complex patterns.

Machine learning offers a more flexible and better approach because it searches and learns from historical claim data to find patterns in past claim data. The supervised learning methods use datasets that include both insurances of fraud and no fraud. Some of the techniques that are used are: logistic regression, decision trees, random forests, support vector machines, gradient boosting, and neural networks etc. Since labeled fraud data can be hard or limited to get, unsupervised methods such as clustering and anomaly detection, are also relevant. Debener, Heinke, and Kriebel (2023) bring up the fact that supervised fraud detection has been investigated a lot, while unsupervised fraud detections are even less discovered in insurance situations. This emphasizes the need for comparative studies that inspect the strengths and weaknesses of different model types.

Class Imbalance in Fraud Detection

For fraud detection we can say that it usually is an imbalanced classification problem because of the fact that there are a smaller number of fraudulent claims than legitimate ones (Chawla et al., 2002). In this situation, accuracy can be misleading. For example, if we take into consideration a model that predicts every claim as non-fraudulent, it might achieve high accuracy because most claims are legitimate, but it would fail to detect any fraud. For this reason, fraud detection studies must also use metrics like precision, recall, F1-score, ROC-AUC, and confusion matrix values.
For showing how many real fraud cases the model finds, there is an important factor which is recall. But we again have a problem because of the fact that very high recalls can also lead us to more false positives and this surely creates extra work for investigators and can affect customers.
For showing how many flagged claims are fraudulent, the other important factor was used which is precision. Also, it is known that F1-score helps balance the precision and recall. In practice, the best balance varies on how many cases the insurer can examine, the cost of false positives, and the cost of missed fraud.

Explainability and Trust in Insurance AI

Explainability is an important point when we talk about claims because in insurance the results from one model can affect how claims are treated, which cases are explored, and how customers are treated. For insurers, it is so important to understand the reason why a certain model marks a claim as a suspicious one. In 2017 Lundberg and Lee developed a new method known as SHapley Additive exPlanations (SHAP). The idea behind this lies in explaining a model's predictions by showing how much each attribute influences.
In insurance domain, the data is highly sensitive that even good precision sometimes is not enough which means that high accuracy is not always sufficient. Besides good precision, models need to be clear, reviewable and allow humans to oversee. The current dataset that is used focuses on fairness, accountability, transparency, safety, and compliance. These are important values which matter because insurance choices impact customers' financial protection and rights. This is why we say that fraud detection models should support decisions and not just auto-generate them. Final choices should always be reviewed and confirmed by people.

Methodology

Dataset

The study uses an insurance claims dataset with 1,000 claim records which is publicly available. The variable used for showings if a claim was labeled as fraudulent or non-fraudulent is fraud_reported. The dataset includes both numerical and categorical information related to policies, insured people, incidents, claim amounts, vehicles, and report status. The dataset used in this study is insurance_claims.csv, published by Aqqad (2023) on Mendeley Data.
All descriptive statistics reported in this section were calculated directly from the dataset. Therefore, the reported fraud rate should be understood only as the fraud rate within this selected public dataset, not as a general fraud rate for the insurance industry.
The dataset is imbalanced. Out of 1,000 records, 247 claims are labeled as fraudulent and 753 are labeled as non-fraudulent. Based on this, now we can say that fraudulent claims represent 24.7% of the dataset, while non-fraudulent claims represent 75.3%.

Table 1. Dataset summary

Dataset characteristic	**Value**
Total records	1,000
Fraudulent claims	247
Non-fraudulent claims	753
Fraud rate	24.7%
Initial columns	39
Responsible feature selection	Applied
Sensitive/proxy variables removed	Yes
Train-test split	75% / 25%
Cross-validation	5-fold stratified cross-validation

Data and Feature

The workflow was Python. Missing with question treated as Numerical were replaced median, while missing values with the most category.

Preprocessing Responsible Selection

preprocessing implemented in values marked marks were missing data. missing values with the categorical were replaced frequent

The variables policy_bind_date and incident_date were converted into date format. Based on these variables, three new features were created:
policy_age_days: the number of days between the policy binding date and the incident date;
incident_month: the month in which the incident occurred;
incident_dayofweek: the day of the week on which the incident occurred.
To reduce the risk of memorization, direct identifiers and highly specific location or date values were removed. These variables included policy_number, policy_bind_date, incident_date, and incident_location.
A responsible feature selection step was also applied. Variables such as insured_zip, insured_sex, insured_education_level, insured_occupation, insured_hobbies, and insured_relationship were excluded from the final modeling pipeline because they contain sensitive personal information. These variables included
Besides the fact that it is so important for the fraud detection models to aim for predictive performance, it is also very important for them to be able to reduce the risk of unfair decisions. This is why this step was included. By removing these variables, the analysis becomes more fitting for an insurance context where fairness, accountability, and transparency are very important.
There is a significant problem with using different models because they also vary on the way they read the variables. For solving this problem, in this study there was used the one-hot encoding. The final feature matrix included variables such as claim-related, policy-related, incident-related, and vehicle-related. The dataset needed to be split into 75% training data and 25% testing data using stratified sampling, so that the fraud and non-fraud proportions were preserved in both subsets.

Models

The eight classification models used in this study are: a majority-class baseline, Gaussian Naive Bayes, k-Nearest Neighbors with k = 15, weighted logistic regression, decision tree, random forest, gradient boosting, and XGBoost. The reason why the majority-class baseline was included was to prove that accuracy alone can be deceptive in an imbalanced fraud detection problem. On the other hand, logistic regression and decision tree were used as more interpretable models, while random forest, gradient boosting, and XGBoost were included as stronger tree-based ensemble methods. Where appropriate, class weighting was applied to reduce the effect of class imbalance and improve the models' ability to detect fraudulent claims.

Evaluation Metrics and Cross-Validation

The metrics that were used to assess the models are: accuracy, precision, recall, F1-score, ROC-AUC, and the value of the confusion matrix. Each of them was used for a different and important reason.
Accuracy – calculates the total percentage of correctly predicted outcomes
Precision – represents the total number of fraudulent claims
Recall – describes the decision of fraudulent claims
The confusion matrix reports the number of true negatives, false positives, false negatives, and true positives (Fawcett, 2006).
In fraud detection, accuracy alone can be misleading, this is the reason why recall and F1-score are very important. A model may achieve high accuracy by predicting most claims as non-fraudulent but still fail to identify fraud. At the same time, a model with high recall but low precision may create too many false alarms. Therefore, model selection should consider both statistical performance and practical usefulness.
To improve the reliability of the results, 5-fold stratified cross-validation was used. Stratification preserved the proportion of fraudulent and non-fraudulent claims in each fold. In order to assess the model stability, cross-validation was used, while a separate holdout test set was used to obtain final confusion matrix results. All calculations that are shown in tables above as results were generated using the same Python preprocessing and modeling pipeline to confirm reproducibility.
Experimental Results
As mentioned above, for the fact that the dataset is imbalanced, the results needed to be interpreted mainly through recall, F1-score, and ROC-AUC and could not be interpreted only with accuracy. But, on the other hand, even though accuracy is not used as the main source for model selection, it is still included for completeness. There is a serious problem in fraud detection because a model can appear accurate if it predicts most claims as non-fraudulent, but the same model can still fail while trying to detect actual fraud cases. This is the reason why the comparison focuses on how well each model recognizes the smaller fraud class while keeping false positives and false negatives at an appropriate level.
The experimental results can be seen below on Tables 2 and 3. Table 2 presents the 5-fold stratified cross-validation results, while Table 3 presents the results from the single holdout test set.

Table 2. Cross-validated model performance

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC-AUC**
Majority baseline	0.753	0.000	0.000	0.000	0.500
Gaussian Naive Bayes	0.670	0.264	0.203	0.227	0.590
k-Nearest Neighbors (k=15)	0.756	0.624	0.057	0.101	0.661
Weighted Logistic Regression	0.749	0.496	0.656	0.564	0.757
Decision Tree	0.781	0.570	0.514	0.532	0.700
Random Forest	0.766	0.557	0.328	0.402	0.752

Gradient Boosting	0.765	0.534	0.425	0.467	0.741
XGBoost	0.780	0.553	0.595	0.570	0.755

Table 3. Holdout test-set model performance

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC-AUC**	**TN**	**FP**	**FN**	**TP**
Majority baseline	0.752	0.000	0.000	0.000	0.500	188	0	62	0
Gaussian Naive Bayes	0.656	0.306	0.306	0.306	0.580	145	43	43	19
k-Nearest Neighbors (k=15)	0.748	0.400	0.032	0.060	0.652	185	3	60	2
Weighted Logistic Regression	0.752	0.500	0.597	0.544	0.720	151	37	25	37
Decision Tree	0.780	0.552	0.597	0.574	0.675	158	30	25	37
Random Forest	0.760	0.526	0.323	0.400	0.747	170	18	42	20
Gradient Boosting	0.776	0.562	0.435	0.491	0.719	167	21	35	27
XGBoost	0.756	0.508	0.516	0.512	0.713	157	31	30	32

For the fact that most claims were non fraudulent, the majority-class baseline achieved an accuracy of 0.752 on the holdout test set. However, it did not detect any fraudulent claims. This is why we have values 0.000 of recall and F1-score. This is a confirmation of we already mentioned: accuracy alone is not a reliable measure for evaluating fraud detection models on imbalanced data. Gaussian Naive Bayes was the one that detected more fraud cases, but on the other hand, the problem was that it produced many false positives, which would create unnecessary investigation work in a real insurance setting. The k-Nearest Neighbors model achieved reasonable accuracy, but it detected only 2 out of 62 fraudulent claims, making it unsuitable for this task. The comparison shows that model performance depends on the evaluation metric. XGBoost achieved the strongest cross-validated F1-score, suggesting stable performance across validation folds. On the separate holdout test set, however, the decision tree achieved the highest F1-score, while random forest achieved the highest ROC-AUC. Weighted logistic regression also remained useful because it provided competitive results while being easier to interpret. These findings show that fraud detection models should not be selected based on a single metric. Instead, cross-validation results, holdout test performance, recall, F1-score, ROC-AUC, and interpretability should be considered together. Since XGBoost achieved the strongest cross-validated F1-score and is well suited for tree-based explanation methods, it was selected for the SHAP-based explainability analysis.

Explainability Analysis

Because XGBoost achieved the strongest F1-score during cross-validation and is also well suited for SHAP interpretation, it was selected for the explainability analysis. SHAP was used to identify which variables had the greatest influence on the model's fraud predictions. By showing how much each feature contributes to the final prediction, SHAP makes it easier to understand the behavior of a complex tree-based model.
The highest SHAP values were observed for incident_severity_Major Damage, vehicle_claim, policy_age_days, incident_state_WV, auto_year, policy_annual_premium, incident_state_NY, capital-gains, property_claim, and total_claim_amount.
Figure 1 presents the SHAP feature importance chart for the XGBoost model. The figure shows that major incident severity was the most influential feature, followed by vehicle claim amount, policy age, incident state, vehicle year, annual premium, and other claim-related variables. B y taking a deeper look at figure 1, it gets clearer which factors had the deepest impact on the model's fraud predictions.

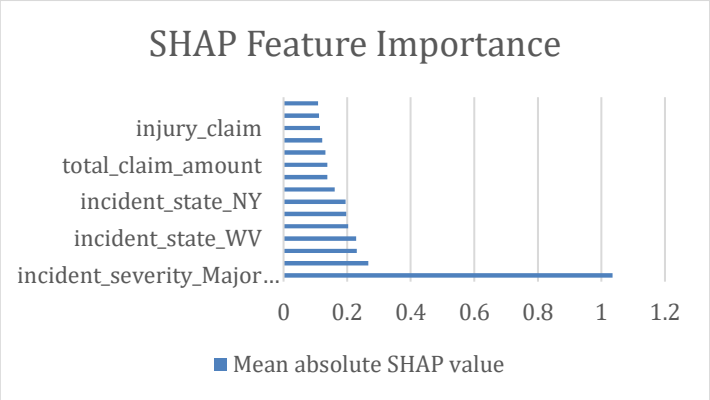

Figure 1. SHAP feature importance for the XGBoost fraud detection model

Looking at the figure 1, we can see that based on the SHAP analysis, it is clear that severity was the most significant feature in the XGBoost model.
This makes sense when we are talking about insurance fraud, because more serious incidents may be connected to higher claim amounts, more complex investigations, or unusual reporting patterns. Additional important variables that are associated with the claims, policies, vehicles or incidents include vehicle claim amount, policy age, incident state, vehicle year, annual premium, capital gains, property claim amount value, total claim amount value, months as customer, injury claim, incident hour, and witnesses.
An important step in the process of developing out final model included removing variables that could be sensitive or act as proxies for sensitive information before training the models. From a deeper analysis that we did, we saw that some of these variables had potential to affect the model's interpretation. By removing them, out model complies with responsible AI principles. This does not demonstrate that the model is entirely unbiased, but it addresses potential ethical issues.
The explainability results also show another reason why the interpretability of an algorithm is important. A model that is interpretable is able to provide valuable insights into the factors that contribute to a particular prediction being made. Indeed, SHAP explanations may help claim investigators understand in more details why a claim was selected as suspicious. However, these explanations should not be treated as final proof of fraud, instead they should be used only as supplementary information.

Discussion

However, it turns out that the performance of the model depends on the evaluation method and the metric selected. Thus, the 5-fold stratified cross-validation results show that XGBoost had the best, which means the highest F1-score, followed closely by weighted logistic regression. In other words, we can say that the average balance between precision and recall was higher for XGBoost when we compare it to others. However, the results that were gathered from the test set showed that the decision tree performed better. This proves one more time that fraud detection models must be analyzed using multiple (at least two) metrics.
Since the biggest number of claims used in the dataset was without any fraud, the accuracy of the model with majority class baseline was quite high. However, this does not mean that it was useful for detecting fraudulent claims, since it was unable to detect them. Thus, the results confirmed that accuracy can be an unreliable measure in fraud detection, especially when the data is imbalanced. The findings showed that simpler models, such as the decision tree and weighted logistic regression, can still perform well when class imbalance is taken into account. Although XGBoost algorithm was chosen for the SHAP analysis because it showed an impressive cross-validated F1-score and tree-based nature.
From all the results that we have, we can conclude that the prediction performance model may not always be the easiest one to explain or use. This was conducted based on the fact that weighted logistic regression was much easier to interpret compared to other models but it did not produce the best result.
Based on all the results, now we can say that in practice SHAP analysis can help us see what factors affected the predictions in XGBoost. The most important factors were related to the incidents, policies, vehicles, and claims.
This model will let investigators decide which claims need closer review. It should not be used for automatic rejection of the claims. Claims that have higher potential for fraud could be sent to special investigators, while the ones that have lower risk could continue through the regular process. Keeping people involved in this process can make fraud detection more efficient, and at the same time keep the whole procedure human and accountable.

Limitations

An insurance company can contain big amount of data and based on that, the dataset that this study uses is relatively small. Based on that, the results may not match when considered big amount of data but, the reason why it was taken is that it was a public one.
Besides the fact that insurance companies nowadays can have big amount of data, they can also have more complex data than the ones that the dataset we used has, they can even include images, documents etc.
Besides the fact that 5-fold stratified cross-validation was used to improve reliability, again the fact that the study relies on a single dataset represents another limitation.
For future work, it is important to test the approach to additional insurance datasets that include more than 1,000 records. By implementing this, we remove two limitations at once, the first one and this one.
SHAP makes the model easier to understand, but it does not prove that the model is fair. Some proxy effects may remain, even after removing sensitive variables.
Applying this type of model in practice would require careful threshold selection, continuous monitoring, feedback loops, human review, and proper regulatory governance.

Conclusions

This paper presented a Python-based comparative study of machine learning methods for insurance claim fraud detection using a public insurance claims dataset. The experiment combined responsible feature selection, stratified train-test splitting, 5-fold stratified cross-validation, holdout test-set evaluation, and SHAP-based explainability. The results show that XGBoost achieved the strongest cross-validated F1-score, while the decision tree achieved the highest holdout F1-score and random forest achieved the highest holdout ROC-AUC. Weighted logistic regression remains useful for being an understandable model benchmark.
After the SHAP analysis, now it is clearer that the severity of the accident is the most dominant attribute among all available attributes. This paper proves that comparing models, analyzing class imbalance, explainability, and responsible feature selection can help a lot to improve the insurance fraud detection. For future work, researchers must consider larger and more complex datasets and also different models.

Data and Code Availability

The complete dataset that was used for this study is easily accessible online and can be downloaded at any time (Aqqad, A. (2023)). All descriptive statistics, class distributions, performance metrics, confusion matrices, and SHAP values for feature importance were generated from this dataset using the Python code described in the methodology section. In order to enable reproducibility, all Python notebooks, preprocessing steps, random seeds, training and testing splits, and other needed parameters can be obtained from the corresponding author by making a reasonable request.

References

[1]. Aqqad, A. (2023). *insurance_claims* [Data set]. Mendeley Data. https://doi.org/10.17632/9 92mh 7dk9y.2
[2]. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324
[3]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
[4]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785– 794. https://doi.org/10.1145/2939672.2939785
[5]. Coalition Against Insurance Fraud. (2022). *The impact of insurance fraud on the U.S. economy*.
[6]. Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk and Insurance, 90*(3), 743–768. https://doi.org/10.1111/jori.12427
[7]. European Insurance and Occupational Pensions Authority. (2021). Artificial intelligence governance principles: Towards ethical and trustworthy artificial intelligence in the European insurance sector.
[8]. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
[9]. Insurance Information Institute. (n.d.). Facts + statistics: Insurance fraud. Retrieved May 1, 2026, from https://www.iii.org/fact-statistic/facts-and-statistics-insurance-fraud
[10]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.
[11]. National Association of Insurance Commissioners. (2023). *Model bulletin on the use of artificial intelligence systems by insurers*.
[12]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.
[13]. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.