Baruah, Pallav Kumar, Satya Sai Mudigonda, Rohan Yashraj Gupta, Sankar Krishna, Eswar Prem Sai Gupta Maturi, Srinand Hegde, Phani Krishna Kandala, and Sumanth Chebrolu. 2025. "Integrating Machine Learning Models with Business Rule Triggers to Boost Performance in Health Insurance Fraud Detection: A Case Study." *Variance* 18 (May).

Other

Integrating Machine Learning Models with Business Rule Triggers to Boost Performance in Health Insurance Fraud Detection: A Case Study

Pallav Kumar Baruah[1a], Satya Sai Mudigonda[1b], Rohan Yashraj Gupta[1c], Sankar Krishna[1d], Eswar Prem Sai Gupta Maturi[1e], Srinand Hegde[1f], Phani Krishna Kandala[1g], Sumanth Chebrolu[1h]

[1] Centre of Excellence in Actuarial Data Science, Sri Sathya Sai Institute of Higher Learning

Keywords: Claims, XG-Boost, GBM, Naïve Bayes, GLM, Health insurance, Business rules/triggers, Data imbalance, Decision tree, Random forest

Variance

Vol. 18, 2025

Health insurance fraud is a significant problem for the insurance industry, where it causes billions of dollars in annual losses. This article describes a novel approach to fraud detection in health insurance that integrates machine learning models with business rule triggers to identify unusual patterns in claims data and flag them for further investigation. Combining machine learning models with business rule triggers greatly enhanced performance across all models. Notably, the approach substantially improved the ability of a model to identify fraudulent cases, leading to a significant increase in effectiveness. This improvement promises to help the insurance industry mitigate the financial impact of fraud.

1. INTRODUCTION

Insurance fraud has been a persistent problem in the commercial world, and health insurance fraud is among the most concerning because healthcare fraud and abuse (Becker, Kessler, and McClellan 2005; Yang and Hwang 2006) are major concerns in many countries (Joudaki et al. 2015). In 2018, $3.6 trillion was spent on healthcare in

[a] Prof. Pallav Kumar Baruah completed his PhD in mathematics at SSSIHL in 1994. He has held various positions at the Department of Mathematics and Computer Science, including the Head of Department. He is currently the Dean of Sciences at SSSIHL. He was awarded the Innovation Award, 2014 by NVIDIA, HiPC Goa. He has won a number of awards in various national and international forums for best paper and posters.

[b] Satya Sai Mudigonda is a senior tech actuarial consultant, adjunct professor, and coordinator of Center of Excellence for Actuarial Data Science at SSSIHL. He has 32 years of senior management experience in actuarial science and technology. His international exposure covers nine countries, including the U.S., U.K., and India. He has managed 36 large projects worth $25 million USD and holds five professional qualifications: AIAI, PMP, CPCU, INS, and API. Satya has published 36 research papers, contributed seven articles, and presented at conferences. He has guided 27 postgraduate dissertations and mentored three PhD scholars.

[c] A highly skilled and experienced actuarial data science professional with a PhD, Rohan Yashraj Gupta is an Associate of the Institute of Actuaries and Associate of Society of Actuaries with over 6 years of experience in the areas of pricing, experience studies, R modeling, and predictive analytics. His expertise in actuarial data science is evident from his eight international journal papers, one book chapter, and presentations at conferences such as Insurance Data Science, CAS Spring Meeting, and ASTIN Actuarial Colloquia. He has developed over five web apps/dashboards for dynamic reporting and visualizations.

[d] Sankar is an accomplished actuary with a strong proficiency in data science. His extensive work experience encompasses reserving, Solvency II, risk management, and the application of machine learning and deep learning in reserving. He excels at integrating actuarial and data science techniques to solve complex actuarial problems. Sankar is also a distinguished recipient of the prestigious CAS Trust Scholarship and holds a master's degree in mathematics, with a specialization in actuarial science.

[e] Eswar Prem is an actuarial science postgraduate student at SSSIHL. He has 3 actuarial exams (Exam1, Exam 2, Exam MAS-1) to his credit. He has worked on two research projects and has completed the CAS Mentor-led Summer Program.

[f] Srinand Hegde is an actuarial science postgraduate student at SSSIHL. He has 3 actuarial exams (CS1-IAI, Exam FM-SOA, MAS1 - CAS) to his credit. He has worked on two research projects and has completed the CAS Mentor-led Summer Program.

[g] Phani Krishna Kandala, with over 12 years in insurance and reinsurance, excels in casualty underwriting. He is an Associate of the Institute of Actuaries U.K. (AIA) and Actuaries Institute India (AIAI), leveraging expertise in data science, analytics, and financial risk management. Mentoring PhD scholars and guiding actuarial projects, Phani thrives in dynamic environments, collaborating across divisions, and advising on strategy and risk. Passionate about creating organizational value through analytical rigor and meticulous attention to detail, he welcomes opportunities to connect and contribute.

[h] Sumanth Chebrolu is a postgraduate student specializing in actuarial science at SSSIHL. He has cleared three CAS papers and completed one VEE requirement. His recent projects have focused on interpretable machine learning, notably a case study in health insurance fraud detection.

the United States, representing billions in health insurance claims. The National Health Care Anti-Fraud Association estimates that financial losses due to healthcare fraud total tens of billions of dollars each year. A conservative estimate is that fraud losses represent 3% of total healthcare expenditures, with some government and law enforcement agencies placing the loss as high as 10% of the annual health outlay, which amounts to more than $300 billion.[1] In 2019, India had over $5.4 billion in insurance fraud loses. Fraudulent health insurance claims can range from 15% to 35% of overall claims.[2]

Losses due to insurance fraud present a major business challenge for many reasons:

- It is difficult to determine the exact value of fraud-related losses. Although precise numbers prove elusive, our study highlights the capacity to enhance our proficiency in estimating losses by strategically applying machine learning models.
- Detected case numbers are much lower than the actual number of fraudulent cases (Joudaki et al. 2015; Kang et al. 2016). Our research illustrates that incorporating triggers into machine learning models proves instrumental in augmenting the detection of fraudulent cases.
- An estimated 8.5% of industry-generated revenue is lost to fraud. Our models produced a higher level of accuracy in detecting fraud cases and concurrently identified a higher volume of fraudulent events. This could conceivably contribute to reducing the percentage of revenue lost; however, our study did not address this aspect.

Existing anti-fraud solutions include having claims managers investigate, with assistance from regulatory agencies (e.g., the Insurance Fraud Bureau and the Insurance Regulatory and Development Authority of India), or using specialized services and tools provided by private organizations. But these approaches have limitations:

- Fraud detection is subjective and depends on the claims investigator's expertise.
- Although a few algorithms and statistical methods (J. Li et al. 2008) have been developed, they are poorly adapted to the problem (Major and Riedinger 2002).

The health insurance sector is growing rapidly and generating increasingly massive amounts of data. Unfortunately, many companies have legacy computer systems that do not capture sufficient details to identify and combat fraud.

Traditional healthcare fraud detection methods rely heavily on auditing and expert inspection, which are costly, inefficient, time consuming, and require significant human intervention. These methods typically focus on specific claims characteristics and pay little attention to relationships between factors. The healthcare insurance industry needs more efficient and effective fraud detection methods that can process vast amounts of data (Mesa et al. 2009).

The basis of this study is the extensive fraud detection research conducted by the SOA Research Institute (Lieberthal et al. 2018). Our study focused on customer-level fraud in healthcare by analyzing claim and policyholder records from the Ayushman Bharat universal health coverage scheme. We applied standard performance metrics to evaluate the performance of several machine learning models developed to detect fraudulent claims (M. E. Johnson and Nagarur 2016). We also compared fraud detection effectiveness in machine learning models (Bauder, Khoshgoftaar, and Seliya 2017; Bauder, da Rosa, and Khoshgoftaar 2018) both with and without integrated business rules.

Our results showed that adding triggers to the data improved machine learning model effectiveness for detecting fraudulent healthcare claims (S.-H. Li et al. 2012). By identifying fraud using an efficient and accurate method, we can reduce financial losses and improve the quality of patient care. This research contributes to the ongoing effort to develop more effective fraud detection methods in healthcare, which will ultimately benefit patients, healthcare providers, and insurers alike.

1.1. ACTUARIAL METHODS AND SKILL SETS IN FRAUD DETECTION

Actuaries assess the financial risks of insurance policies. Insurance fraud not only results in significant financial losses for insurance companies but also increases policyholders' premiums. One important risk mitigation method is claims control, where actuaries use data analysis to identify and quantify patterns of fraudulent behavior in insurance claims data, then develop strategies to mitigate these risks.

Actuaries are uniquely qualified to identify and analyze factors that contribute to insurance fraud. They have a thorough understanding of data, product value chains, claims, and policyholder behavior. Using their understanding of the motivations behind fraudulent behavior, actuaries design insurance products and pricing structures that deter fraud and minimize losses (Vanhoeyveld, Martens, and Peeters 2020). Actuaries also work closely with claims adjusters and other experts to investigate suspicious claims and identify potential cases of fraud. However, traditionally, actuaries are consumers of fraud analytics output. Research in preventing and detecting fraud using actuarial data science techniques will open opportunities for actuaries in this area (Richman 2018).

[1] National Health Care Anti-Fraud Association. https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/

[2] https://paytminsurance.co.in/health-insurance/articles/health-care-fraud-in-india-causes-and-preventive-measures/

Table 1. ML and DL model comparisons

ML Methods	DL Methods
Typically use simpler models (e.g., decision trees, support vector machines), so they generally require less computational power and are more interpretable.	Employ complex neural networks with multiple layers, so they demand significant computational resources and are less intuitive and interpretable.
ML is an AI algorithm that allows systems to learn from data.	DL is an ML algorithm that uses deep (more than one layer) neural networks to analyze data and provide output accordingly.
Traditional ML models, such as logistic regression or decision trees, typically have a shallow architecture with a limited number of layers. These models are designed to operate on handcrafted features and assume that the features are informative and relevant to the task.	DL models, such as neural networks, have a deep architecture with multiple layers. Network depth allows them to automatically learn hierarchical representations of features. Each layer captures different levels of abstraction, enabling the model to extract complex patterns from raw data.
Example In traditional ML, like logistic regression, feature engineering (Khurana et al. 2017; Zheng and Casari 2018) involves manually selecting and creating relevant features, which are carefully chosen based on domain knowledge and an understanding of the problem.	Example In DL, particularly with neural networks, there is often no need for explicit feature engineering (Zheng and Casari 2018). The model is fed raw data, such as unstructured claim documents, images of medical records, or other raw representations of information. The deep neural network automatically learns hierarchical representations of features at different levels of abstraction.

Ultimately, actuaries' role in insurance fraud detection is critical to the success and sustainability of the insurance industry (Gupta, Mudigonda, et al. 2019). By effectively detecting and preventing fraud, actuaries help maintain insurance companies' financial stability and their ability to provide affordable and reliable coverage to policyholders.

1.2. MACHINE LEARNING TO DETECT HEALTH INSURANCE FRAUD

Health insurance industry growth has generated a massive amount of data. Unfortunately, many companies still rely on outdated systems that fail to capture sufficient details to identify and prevent fraudulent activity (Househ and Aldosari 2017). Most companies manage and control fraud; consequently, only a few fraud cases are ever identified, and those are often discovered years after they occurred.

To combat this problem, machine learning (ML) and deep learning (DL) fraud detection techniques are becoming increasingly popular in the health insurance sector (Gomes, Jin, and Yang 2021). These techniques analyze large amounts of data more accurately and faster (Zhou et al. 2020; Srinivasan and Arunasalam 2013) than methods that rely on humans manually analyzing data to detect patterns and anomalies (Sadiq et al. 2017; van Capelleveen et al. 2016) that may indicate fraudulent activities (Kose, Gokturk, and Kilic 2017). ML models recognize patterns in claims data that suggest fraud, such as abnormal billing patterns or excessive billing for a specific service, while DL models scrutinize medical records and claims data to identify potential fraud and flag the records for further investigation (Gao et al. 2018).

Moreover, ML and DL models can adapt to new fraud patterns and continue to learn from new data (Verma, Taneja, and Arora 2017), which is particularly critical in an industry where new types of fraud emerge quickly and require prompt detection and response (Thornton et al. 2013). Overall, the use of ML and DL techniques in health (J. M. Johnson and Khoshgoftaar 2019) insurance fraud detection (M. E. Johnson and Nagarur 2016) has the potential to lower fraud losses, improve the efficiency of identifying fraudulent activity, and reduce costs for insurance companies and consumers (Yoo et al. 2012). Table 1 describes distinctions between the ML and DL methods.

These distinctions make certain tasks uniquely suited for either ML or DL methods. Our study emphasized ML methods; we did not thoroughly explore DL techniques. However, it is important to acknowledge that DL methods have the potential to effectively identify fraud and may serve as a compelling area for future research.

2. DATA

2.1. DATASET

Ayushman Bharat, also known as Pradhan Mantri Jan Arogya Yojana, is the world's largest group health insurance scheme. It was launched by the government of India in September 2018, with the aim to provide health insurance coverage to economically vulnerable families in India. Eligible beneficiaries are entitled to free treatment for various medical conditions at empaneled hospitals across India. The program currently covers more than 100 million families (approximately 500 million individuals), making it the world's largest government-funded healthcare program.

In addition to providing healthcare coverage, Ayushman Bharat also aims to establish health and wellness centers across the country, with the goal of promoting preventive healthcare and early detection of illnesses. This scheme is a major step toward achieving universal health coverage in India, which is a United Nations key sustainable development goal.

The data used for this research are from August 2019 to August 2020. We initially had two datasets: claims data and policy data. We later combined these to form a single dataset on which we conducted data preprocessing steps, such as missing values management (Frane 1976; Lin and Tsai 2020; Patidar and Tiwari 2013), feature scaling, and dimensionality reduction. During the dimensionality reduction step, we eliminated features with zero standard deviation and selected the top 20 unique values from columns with a high number of distinct values.

Zero standard deviation indicates that the feature values were constant, thereby offering no meaningful impact on the target variable (i.e., fraud) predictions. Consequently, these features were removed from consideration.

Certain categorical features exhibited a multitude of unique values, prompting us to narrow our focus to the top 20 (i.e., most frequent) occurrences. Values with limited recurrence are likely to have a minimal impact on predicting fraud. To facilitate model training in subsequent stages, we planned to convert these categorical features into factors through encoding. However, encoding all the unique values in a column with numerous possibilities would result in an extensive dataset, which would require computationally demanding training. Therefore, we opted to retain only the top 20 unique values for certain columns, ensuring a more manageable dataset for training models in later stages.

2.2. REASONS FOR FRAUD

We identified policyholder fraud reasons from the data and classified them into seven categories (Dietz and Snyder 2007; Villegas-Ortega, Bellido-Boza, and Mauricio 2021; Ekin, Musal, and Fulton 2015).

2.2.1. USING THE WRONG DIAGNOSIS TO JUSTIFY PAYMENT

To obtain medical reimbursements, individuals must complete claim forms that detail the services they received, which are typically based on a medical diagnosis. However, it is possible for policyholders to collude with medical service providers (Pande and Maas 2013) to deliberately manipulate the information provided to the insurance company to obtain reimbursement.

Reasons identified under this category included the following:

- The policyholder, in collaboration with the medical service provider, filed a claim for a covered procedure, even though the treatment received was not a covered benefit under the policy.
- The policyholder underwent an outpatient department procedure (Liou, Tang, and Chen 2008) but applied for reimbursement under the surgical package.
- Medical documents that do not suggest any illness were submitted as proof of illness.
- Photos that suggest surgery has not been done were submitted as proof of surgery.
- The policyholder had surgery on a body part not covered by the insurance policy but submitted a claim for reimbursement that falsely stated the surgery was performed on a covered body part.

2.2.2. PRICE AND DOCUMENT MANIPULATION

Manipulating documents, such as clinical exams, certificates, medical prescriptions (Aral et al. 2012; Victorri-Vigneau et al. 2009), and other related documents, with the intention of obtaining an economic benefit, is a common practice (Fang and Gong 2017; Shin et al. 2012).

Reasons identified under this category included the following:

- Providing documents from a pre-procedure instead of post-procedure as proof of the procedure. For example, x-rays, percutaneous transluminal coronary angioplasty frozen images.
- Not submitting the relevant documents/photos for the package applied for. For example, hemodialysis chart, discharge summary, clinical notes, vital charts, treatment charts, mortality audit form, referral slips.
- Sending the same document/photos for different dates.
- Submitting reports that were not signed by any specialist.
- Mismatched billing and discharge summary dates.
- Date tampering on submitted documents.

2.2.3. BILLING FOR SERVICES NOT PROVIDED

Patients may file fraudulent medical claims on their own or in collaboration with acquaintances or healthcare providers to obtain reimbursement for medical expenses they did not actually incur (Fang and Gong 2017; Shin et al. 2012).

Reasons identified under this category included the following:

- Applied for reimbursement for a procedure that was not actually done. Examples include delivery, CT scan, surgery, stent, ventilator, ICU admission.
- Applying for reimbursement of specialist charges when there was no specialist available in that hospital.
- Applying for reimbursement for a procedure advised by the doctor, but the procedure was not done because the patient left against medical advice.
- Applying for reimbursement for a procedure done in a hospital where the patient was initially admitted, but the procedure was not done because the patient was referred to another hospital.
- Applying for reimbursement for a procedure, where the patient died before the procedure.
- Registering multiple treatment packages on the same date to inflate the claim amount.
- Applying for reimbursement for two procedures on the same day, when those two procedures cannot be done on the same day.
- Seeking a package cost increase after discharge to inflate the claim amount.

2.2.4. *OPPORTUNISTIC FRAUD*

Opportunistic fraud involves taking advantage of a real claim to introduce fictional preexisting or previous damages.

Reasons identified under this category include the following:

- Providing photos from a past treatment.
- Providing documents from treatments that were done before the insurance coverage started.
- Applying for reimbursement even though the patient was not eligible for the scheme based on location.
- Package enhancement was taken earlier than the minimum number of days before which enhancement is allowed.
- Applying for reimbursement of a claim that has already been covered under another package.
- Duplicate claims.
- Applying for packages that require a minimum number of hospital admission days when the actual number of admitted days was below the minimum.

2.2.5. *IDENTITY FRAUD*

Identity fraud involves acquiring and using another person's health insurance card to receive medical treatment or other services. This may occur either with or without the knowledge or consent of the card owner.

Reasons identified under this category include the following:

- A male patient applied for a delivery package.
- Names on the discharge summary and clinical notes did not match.
- Family members of the insured were admitted instead of the insured.

2.2.6. *MISREPRESENTING ELIGIBILITY*

Patients may provide false or misleading information about themselves or their dependents to fraudulently obtain medical coverage for which they are not entitled.

Reasons identified under this category include the following:

- Patients not revealing their alcohol habits when applying for insurance.
- Misrepresenting the patient's age to obtain a package for which they are ineligible based on age.

2.2.7. *DELAY*

Delay refers to intentionally delaying claims submission or withholding necessary information to obtain a greater benefit from the insurance provider.

Reasons identified under this category include the following:

- Preauthorization request raised after the patient was admitted to the hospital.
- Preauthorization was generated after the patient was discharged from the hospital.
- Admitting the patient several days after the approval.
- Submitting the claim several days after discharge.
- Claims submission delay ranging from 30 to 120 days.
- Procedure was performed before the procedure approval date.

Table 2. Data overview

Descriptions	Values
Sample size (nrow)	109,193
No. of variables (ncol)	20
No. of numeric/integer variables	5
No. of text variables	10
No. of date variables	5

Table 3. Target variable summary

Fraud	Frequency
0	105,930
1	3,263

2.3. EXPLORATORY DATA ANALYSIS

Table 2 provides an overview of the data and data types.

Table 3 shows the number of fraudulent claims, which comprise about 3% of the dataset.

2.4. DATA DICTIONARY

Table 4 contains the data dictionary.

3. BUSINESS RULE TRIGGERS

Triggers are specific sets of activities that indicate potential fraud and raise suspicion. Identifying triggers is a complex task that requires a comprehensive understanding of the business processes involved. Moreover, triggers can vary depending on the stage at which the fraud is committed. To identify triggers and generate trigger data for the analysis, we built R functions based on several business rules. We then added the trigger data to the original dataset. The 10 triggers used in the analysis are described below.

3.1. CLAIM AMOUNT FRAUD TRIGGER

A predetermined amount for each procedure is typically agreed upon by the insurer and medical service provider (Pande and Maas 2013), particularly for cashless claim processing. A claim amount that exceeds the established procedure-specific amount triggers a fraud alert.

3.2. HOSPITAL ADMISSION DAYS FRAUD TRIGGER

A specific number of hospital admission days is deemed reasonable for each procedure. If the actual number of hos-

Table 4. Data dictionary

Feature	Description	Category
insured_id	Insured's unique ID	Character
benefit_type	Benefit type (medical/surgical)	Character
treatment_start_date	Treatment start date	Date
treatment_end_date	Treatment end date	Date
claim_reported_date	Date the claims were reported	Date
approved_allowed_amount	Approved/allowed amount	Numeric
medical_service_provider_id	Hospital's unique ID	Character
no_of_days_stayed	Number of days hospitalized	Numeric
primary_diagnosis_code	Unique diagnostic code	Character
primary_procedure_code	Unique procedure code	Character
hospital_type	Public or private hospital	Character
category	Type of family eligible for scheme	Character
policy_commencement_date	Policy start date	Date
policy_termination_date	Policy end date	Date
birth_date	Insured's age	Numeric
gender_code	Insured's gender 1-Male 0-Female	Character
residence_location	District in which the insured stays	Character
claim_count_pa	Total number of claims in a year	Numeric
hospital_location	District where the hospital is located	Character
fraud	Claim fraud status: fraud (1) / not fraud (0)	Numeric

pital admission days exceeds the established duration for the given procedure, a fraud alert is triggered.

3.3. AGE FRAUD TRIGGER

Specific medical procedures have expected age ranges assigned. When a claimant's age falls outside the expected range for a particular procedure, a fraud alert is triggered. For this function to operate correctly, age must be correctly specified in both the claims file and the triggers file.

3.4. GENDER FRAUD TRIGGER

Certain medical procedures are specific to a particular gender, so a procedure performed on a claimant of the wrong gender generates a fraud alert. For instance, a gynecological procedure performed on a male claimant would trigger an alert. To operate correctly, gender must be specified correctly in both the claims file and the triggers file.

3.5. CLAIM COUNT FRAUD TRIGGER

The number of medical treatments for a specific condition are typically limited for policyholders, depending on the nature of the condition and treatment. If a policyholder has an unreasonably high frequency for a specific treatment, it may indicate fraud. This function detects all policyholders with an excessive number of treatments for the given condition.

3.6. CLOSE PROXIMITY FRAUD TRIGGER

A close proximity claim is when the treatment start date is very close to the policy start date. This condition will trigger a fraud alert.

3.7. TREATMENT DATE VALIDITY FRAUD TRIGGER

Each policy has a start date and an end date within which the claim event (treatment) must occur. If the treatment date is outside these dates, a fraud alert is triggered.

3.8. CLAIM REPORTING DELAY FRAUD TRIGGER

Each covered condition has a treatment start date and a treatment end date. The claim must be reported within the permissible number of days after the treatment end date (discharge date). A claim reported date that is outside the permissible limit will trigger a fraud alert.

3.9. EMPANELED HOSPITALS (MEDICAL SERVICE PROVIDERS) FRAUD TRIGGER

Insurers typically select hospitals to serve their policyholders through a process known as empanelment, whereby the insurer verifies that the hospital has the necessary facilities and has agreed to the specified tariff for each treatment. A fraud alert is triggered if the hospital specified in a claim is not included in the list of empaneled hospitals (Pande and Maas 2013).

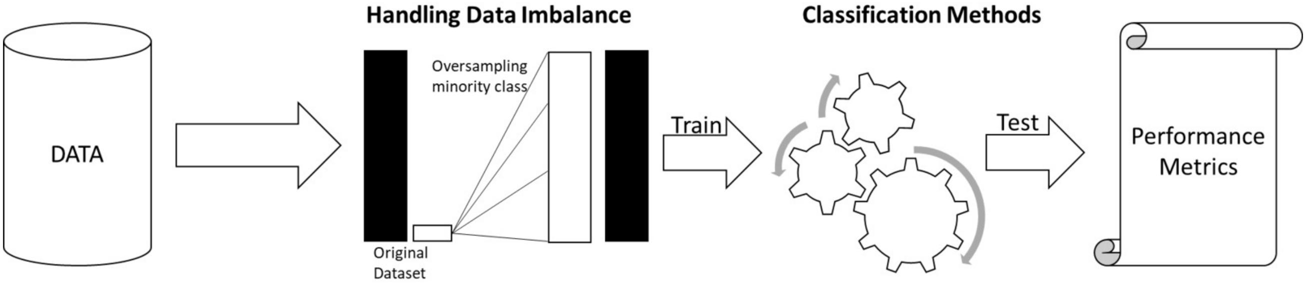

Figure 1. Schematic representation of M1.

3.10. HOSPITAL DISTANCE FRAUD TRIGGER

It is reasonable to anticipate that a policyholder will receive treatment at the nearest hospital. If the distance between the policyholder's residence and the hospital location exceeds a predetermined threshold, a fraud alert is triggered.

4. METHODOLOGY

The methods adopted for fraud detection are described below (M. E. Johnson and Nagarur 2016; Kirlidog and Asuk 2012).

- First, we addressed the problem of data imbalance through several oversampling techniques.
- Then we conducted analyses
  - Methodology 1: ML models applied to data without triggers (M1).
  - Methodology 2: ML models applied to data with triggers (M2).

4.1. HANDLING DATA IMBALANCE

Data imbalance is a common problem in health insurance claims data used for fraud detection. It occurs when the number of fraudulent claims in the dataset is significantly smaller than the number of legitimate claims (Mohammed, Rawashdeh, and Abdullah 2020; Thabtah et al. 2020).

With an imbalanced dataset, the fraud detection model trained on the data may be biased toward predicting that most claims are legitimate, since this will result in a high overall accuracy score. Consequently, the model may fail to detect fraudulent claims, leading to significant financial losses for insurance companies (Hassan and Abraham 2016; Gupta, Mudigonda, and Baruah 2021).

To address this problem, we used oversampling techniques where the minority class (fraudulent claims) would be oversampled until both classes had equal proportions.

It is important to note that while these techniques improve the performance of fraud detection models on imbalanced data, they cannot completely solve the problem. Therefore, it is important to carefully evaluate model performance (Seliya, Khoshgoftaar, and Van Hulse 2009) and continuously monitor the models to ensure that they are accurately identifying fraudulent claims (Frane 1976).

4.2. M1: ML MODELS APPLIED TO DATA WITHOUT TRIGGERS

First, we handled the data imbalance problem using four oversampling techniques (Aviñó, Ruffini, and Gavaldà 2018): random oversampling examples (ROSE), synthetic minority oversampling (SMOTE) (Chawla et al. 2002), majority weighted minority oversampling (MWMOTE) (Barua et al. 2014), and adaptive synthetic sampling (ADASYN) (He et al. 2008).

Then we used six classification models (Koh and Tan 2011) of supervised learning methods: decision tree, random forest (Ho 1995), XGBoost, naive Bayes, gradient boosting machine (GBM) (Gupta, Sai Mudigonda, et al. 2019), and generalized linear model (GLM) (H. Chen and Lindsey 1998; Hutcheson 2001) on the data without triggers.

Figure 1 shows the steps involved in M1.

4.3. M2: ML MODELS APPLIED ON DATA WITH TRIGGERS

First, we generated data for the 10 trigger functions based on the business rules described in Section 3 and added the data columns to the original dataset.

Then, as with M1, we addressed the data imbalance problem using the same four oversampling techniques (ROSE, SMOTE (Chawla et al. 2002), MWMOTE (Barua et al. 2014), and ADASYN (He et al. 2008)) and applied the same six classification (Ali, Shamsuddin, and Ralescu 2015) models of supervised learning methods (decision tree (Patidar and Tiwari 2013), random forest (Ho 1995), XGBOOST, naive Bayes, GBM, and GLM (H. Chen and Lindsey 1998; Hutcheson 2001).

Brief descriptions of the ML models used in this study are as follows:

1. Decision trees represent a type of algorithm used in decision-making processes. The algorithms pose a series of questions that partition the data into subsets, and the process persists until a final decision or prediction is reached. Decision trees are visualized as hierarchical structures that resemble an inverted tree, where each node signifies a question and each branch corresponds to a possible answer. The aim is to create

a model that effectively navigates through data, making informed decisions based on a systematic series of inquiries. Decision trees are widely used because they are easily interpreted and can handle both categorical and numerical data, making them a valuable tool for several domains, including finance, healthcare, and ML applications. *Classification and Regression Trees* (Breiman et al. 1984) presents a useful overview of decision trees.

2. Random forests (Ho 1995) comprise a combination of tree predictors such that each tree depends on random vector values sampled independently and with the same distribution for all trees in the forest (Breiman 2001). After many trees are generated, they vote for the most popular class. Instead of using the entire dataset, a sample of the dataset is used to train each of the trees, which is also the case for decision trees. Furthermore, not all variables are used for splitting the nodes. These steps ensure that the model does not overfit the data.
3. XGBoost is a highly effective and widely used ML method. Boosting grows trees sequentially by using information from previous trees, which differs from bagging methods, where bootstrapping is used. XGBoost is an optimized distributed gradient boosting (Guelman 2012) library designed for efficient and scalable training of ML models. It is an ensemble learning model that combines predictions from multiple weak models to produce a stronger prediction. XGBoost was introduced by T. Chen and Guestrin (2016) and has demonstrated success in various ML competitions in the Kaggle data science platform. To boost regression trees, we chose the number of splits to include in our tree, then created multiple trees of that size, where each new tree after the first was created by predicting the residuals of the previous tree.
4. GLM (H. Chen and Lindsey 1998; Hutcheson 2001) is a framework for extending linear regression models (Gupta et al. 2020) to handle different types of data distributions (Lu and Boritz 2005). It generalizes linear regression by incorporating a link function and a probability distribution to model a broader range of relationships.
5. Naive Bayes is a probabilistic classification algorithm rooted in Bayes' theorem. It leverages a foundational assumption of feature independence given the class label, thereby simplifying computational processes. Widely applied in text classification, spam filtering, and categorization tasks, naive Bayes is an effective framework for making informed decisions based on probabilistic reasoning. It handles diverse datasets with efficiency, making it a preferred choice in scenarios where feature independence assumptions align with the nature of the data.
6. GBM is a boosting algorithm that constructs a sequence of weak models, strategically amalgamating them to form a robust predictive model. The optimization process involves learning from the errors of preceding models, iteratively refining accuracy over successive iterations. GBM's iterative approach to model building enhances predictive performance by focusing on areas of misclassification, progressively refining the overall predictive capability. This algorithm is widely acknowledged for its ability to deliver high accuracy and resilience across diverse datasets, making it a valuable tool in predictive modeling and ML applications.

5. RESULTS

Performance of the six ML algorithms on data without triggers (N) and data with triggers (Y) is summarized in Table 5. Each algorithm was trained using the four synthetic oversampling methods discussed previously, resulting in 48 combinations.

Our primary objective was to minimize the financial damage resulting from fraudulent activities. Therefore, we aimed to increase the true positive rate and decrease the false negative rate in our classification model. While false positives may result in additional costs associated with conducting further investigations, these costs are generally less than the losses incurred due to false negatives. Since false negatives are more expensive than false positives, we used the F2 score to compare performance across models, because it prioritizes recall (true positive rate) over precision, which better captures model performance in identifying the minority class.

Table 6 shows the F2 score improvement for all model combinations. The rows in bold indicate the most improved combination for each ML model.

Based on the F2 score, the XGBoost model with ADASYN (He et al. 2008) balancing applied to the data with triggers (Y) demonstrated the highest performance, with a score of 0.9267. Overall, all models showed improved performance when applied to data with triggers, with the decision tree model with SMOTE (Chawla et al. 2002) showing the most improvement.

This finding may have resulted from including the trigger data, which provided the model with additional information, leading to better performance. Important features and information were extracted from the trigger data, which helped the model better predict fraud.

Figure 2 depicts the categories the model assigns to a claim and the related errors, where the positive class indicates fraud and the negative class indicates not fraud.

- TP (true positive): Fraud identified as fraud by the model
- TN (true negative): Not fraud identified as not fraud by the model
- FP (false positive): Not fraud identified as fraud by the model
- FN (false negative): Fraud identified as not fraud by the model

Figure 3 shows that the models run on data with triggers had improved F2 scores compared with the models run on data without triggers.

Table 5. Comparative results of ML models

ML Method	Oversampling Techniques	Trigger	Accuracy	Sensitivity	Specificity	Precision	Recall	F1	F2	ROC-AUC	PR-AUC
Decision Tree	ADASYN	N	0.6648	0.6153	0.7143	0.6829	0.6153	0.6473	0.6277	0.6695	0.3222
Decision Tree	ADASYN	Y	0.7205	0.6217	0.8191	0.7745	0.6217	0.6898	0.6473	0.7580	0.3195
Decision Tree	SMOTE	N	0.7132	0.3003	0.9485	0.7686	0.3003	0.4318	0.3420	0.6524	0.3097
Decision Tree	SMOTE	Y	0.8349	0.6607	0.9342	0.8512	0.6607	0.7439	0.6916	0.7974	0.3167
Decision Tree	MWMOTE	N	0.7727	0.6502	0.8952	0.8612	0.6502	0.7410	0.6837	0.7994	0.3182
Decision Tree	MWMOTE	Y	0.8610	0.8187	0.9032	0.8943	0.8187	0.8548	0.8328	0.8713	0.3226
Decision Tree	ROSE	N	0.6485	0.5409	0.7561	0.6892	0.5409	0.6061	0.5653	0.6913	0.3190
Decision Tree	ROSE	Y	0.7996	0.6960	0.9032	0.8779	0.6960	0.7764	0.7261	0.8078	0.3192
Random Forest	ADASYN	N	0.7597	0.7757	0.7455	0.7308	0.7757	0.7526	0.7663	0.8393	0.8482
Random Forest	ADASYN	Y	0.8909	0.9047	0.8780	0.8737	0.9047	0.8890	0.8984	0.9479	0.9506
Random Forest	SMOTE	N	0.7922	0.7069	0.8433	0.7302	0.7069	0.7184	0.7115	0.8412	0.7330
Random Forest	SMOTE	Y	0.9169	0.8729	0.9434	0.9026	0.8729	0.8875	0.8786	0.9676	0.9476
Random Forest	MWMOTE	N	0.8410	0.8419	0.8402	0.8397	0.8419	0.8408	0.8415	0.9103	0.9185
Random Forest	MWMOTE	Y	0.9203	0.9158	0.9249	0.9257	0.9158	0.9207	0.9177	0.9678	0.9715
Random Forest	ROSE	N	0.6836	0.6827	0.6845	0.6860	0.6827	0.6844	0.6834	0.7438	0.7343
Random Forest	ROSE	Y	0.8343	0.9015	0.7864	0.7507	0.9015	0.8192	0.8667	0.9054	0.9156
XG Boost	ADASYN	N	0.7414	0.8193	0.6941	0.6196	0.8193	0.7056	0.7697	0.7976	0.6917
XG Boost	ADASYN	Y	0.9162	0.9351	0.8989	0.8944	0.9351	0.9143	0.9267	0.9649	0.5788
XG Boost	SMOTE	N	0.8586	0.8550	0.8602	0.7351	0.8550	0.7905	0.8280	0.9044	0.6056
XG Boost	SMOTE	Y	0.9221	0.8988	0.9350	0.8850	0.8988	0.8919	0.8960	0.9631	0.5513
XG Boost	MWMOTE	N	0.8507	0.9484	0.7880	0.7418	0.9484	0.8325	0.8984	0.9267	0.6350
XG Boost	MWMOTE	Y	0.9177	0.9233	0.9124	0.9112	0.9233	0.9172	0.9208	0.9666	0.5729
XG Boost	ROSE	N	0.6804	0.7063	0.6602	0.6174	0.7063	0.6589	0.6866	0.7537	0.7122
XG Boost	ROSE	Y	0.8297	0.8811	0.7905	0.7622	0.8811	0.8174	0.8544	0.9010	0.6912
GLM	ADASYN	N	0.6971	0.7071	0.6880	0.6730	0.7071	0.6896	0.7000	0.7330	1.0000
GLM	ADASYN	Y	0.7034	0.7790	0.6600	0.5674	0.7790	0.6566	0.7250	0.7787	1.0000
GLM	SMOTE	N	0.6522	0.5149	0.7966	0.7269	0.5149	0.6028	0.5468	0.7431	1.0000
GLM	SMOTE	Y	0.8395	0.8277	0.8449	0.7047	0.8277	0.7613	0.7998	0.8724	1.0000

ML Method	Oversampling Techniques	Trigger	Accuracy	Sensitivity	Specificity	Precision	Recall	F1	F2	ROC-AUC	PR-AUC
GLM	MWMOTE	N	0.6666	0.6190	0.7776	0.8665	0.6190	0.7222	0.6565	0.7405	1.0000
GLM	MWMOTE	Y	0.8399	0.8652	0.8178	0.8052	0.8652	0.8341	0.8525	0.9094	1.0000
GLM	ROSE	N	0.6626	0.6790	0.6489	0.6167	0.6790	0.6464	0.6656	0.7338	1.0000
GLM	ROSE	Y	0.8070	0.8731	0.7608	0.7185	0.8731	0.7883	0.8371	0.8649	1.0000
Naive Bayes	ADASYN	N	0.6934	0.6714	0.7220	0.7578	0.6714	0.7120	0.6871	0.6928	0.9993
Naive Bayes	ADASYN	Y	0.6835	0.7181	0.6585	0.6038	0.7181	0.6560	0.6919	0.7424	0.9826
Naive Bayes	SMOTE	N	0.6023	0.4686	0.7666	0.7115	0.4686	0.5650	0.5029	0.6831	0.9933
Naive Bayes	SMOTE	Y	0.7834	0.6861	0.8464	0.7433	0.6861	0.7136	0.6968	0.7953	0.9801
Naive Bayes	MWMOTE	N	0.6745	0.6192	0.8254	0.9064	0.6192	0.7358	0.6611	0.7241	0.9874
Naive Bayes	MWMOTE	Y	0.8153	0.8099	0.8209	0.8241	0.8099	0.8169	0.8127	0.8324	0.9907
Naive Bayes	ROSE	N	0.6335	0.6236	0.6452	0.6736	0.6236	0.6477	0.6330	0.6967	0.9989
Naive Bayes	ROSE	Y	0.7830	0.8060	0.7631	0.7453	0.8060	0.7744	0.7931	0.7999	0.9880
GBM	ADASYN	N	0.7330	0.7760	0.7016	0.6551	0.7760	0.7105	0.7484	0.7572	0.7123
GBM	ADASYN	Y	0.7719	0.7805	0.7638	0.7563	0.7805	0.7682	0.7756	0.8362	0.7122
GBM	SMOTE	N	0.6346	0.4978	0.8012	0.7531	0.4978	0.5994	0.5340	0.7288	0.6166
GBM	SMOTE	Y	0.8220	0.7304	0.8834	0.8078	0.7304	0.7671	0.7447	0.8870	0.6166
GBM	MWMOTE	N	0.7236	0.7738	0.6890	0.6321	0.7738	0.6958	0.7406	0.7792	0.7123
GBM	MWMOTE	Y	0.8696	0.8830	0.8572	0.8523	0.8830	0.8673	0.8766	0.9169	0.7123
GBM	ROSE	N	0.6424	0.7081	0.6082	0.4846	0.7081	0.5754	0.6483	0.7131	0.7123
GBM	ROSE	Y	0.8075	0.8529	0.7725	0.7432	0.8529	0.7943	0.8285	0.8619	0.7123

Table 6. F2 score improvement for various models

		F2 Score		
		Without Triggers (N)	With Triggers (Y)	Improvement (Y-N)
Decision Tree	ADASYN	0.627706	0.64727	0.019563
	SMOTE	**0.341957**	**0.691621**	**0.349664**
	MWMOTE	0.683707	0.832806	0.149098
	ROSE	0.565263	0.726076	0.160813
Random Forest	ADASYN	0.766289	0.898366	0.132077
	SMOTE	0.711475	0.878644	0.167169
	MWMOTE	0.841494	0.917737	0.076243
	ROSE	**0.683367**	**0.866669**	**0.183302**
XG Boost	ADASYN	0.769666	0.926701	0.157035
	SMOTE	0.827989	0.896021	0.068031
	MWMOTE	0.898357	0.920819	0.022462
	ROSE	**0.686558**	**0.854447**	**0.167889**
GLM	ADASYN	0.700041	0.724952	0.024911
	SMOTE	**0.546762**	**0.799755**	**0.252992**
	MWMOTE	0.656543	0.852497	0.195954
	ROSE	0.665585	0.837094	0.171509
Naive Bayes	ADASYN	0.687059	0.691931	0.004872
	SMOTE	**0.502909**	**0.696826**	**0.193917**
	MWMOTE	0.661109	0.812692	0.151583
	ROSE	0.633022	0.793091	0.160069
GBM	ADASYN	0.748396	0.775561	0.027165
	SMOTE	**0.534029**	**0.744655**	**0.210626**
	MWMOTE	0.740569	0.87664	0.136071
	ROSE	0.648306	0.828461	0.180156

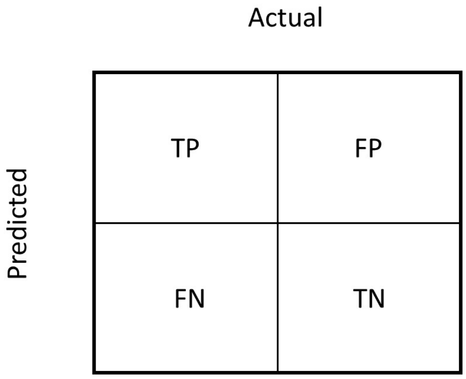

Figure 2. Confusion matrix.

For example, for XGBoost with ADASYN (He et al. 2008) and trigger data, the claim_reported_delay_flag, distance, gender_flag, and hospital_distance_flag, which were generated from the trigger data, were all important for predicting fraudulent claims. This shows that integrating trigger functions into ML models improves model performance. In this model the F2 score increased from 0.77 to 0.93.

Figure 4 shows the results of XGBoost with ADASYN (He et al. 2008) without trigger data.

Figure 5 shows the results of XGBoost with ADASYN (He et al. 2008) with trigger data.

6. CONCLUSION

Our results show that including trigger data generally leads to improved model performance. However, the extent of this improvement is influenced by the model and imbalance method used. Therefore, it is crucial to select the optimal combination of model and imbalance method based on the specific requirements of the task. Additionally, the effectiveness of triggers on model improvement may also depend on the characteristics of the trigger data being used. Therefore, it is important to carefully evaluate the data before incorporating triggers. The primary objective of our project's model was to optimize for the highest F2 score, making it the paramount criterion for selecting the XGBoost model integrated with the ADASYN (He et al. 2008) imbalance technique. Notably, the XGBoost algorithm was adept at enhancing model performance without overfitting the training data, which is another factor to consider when

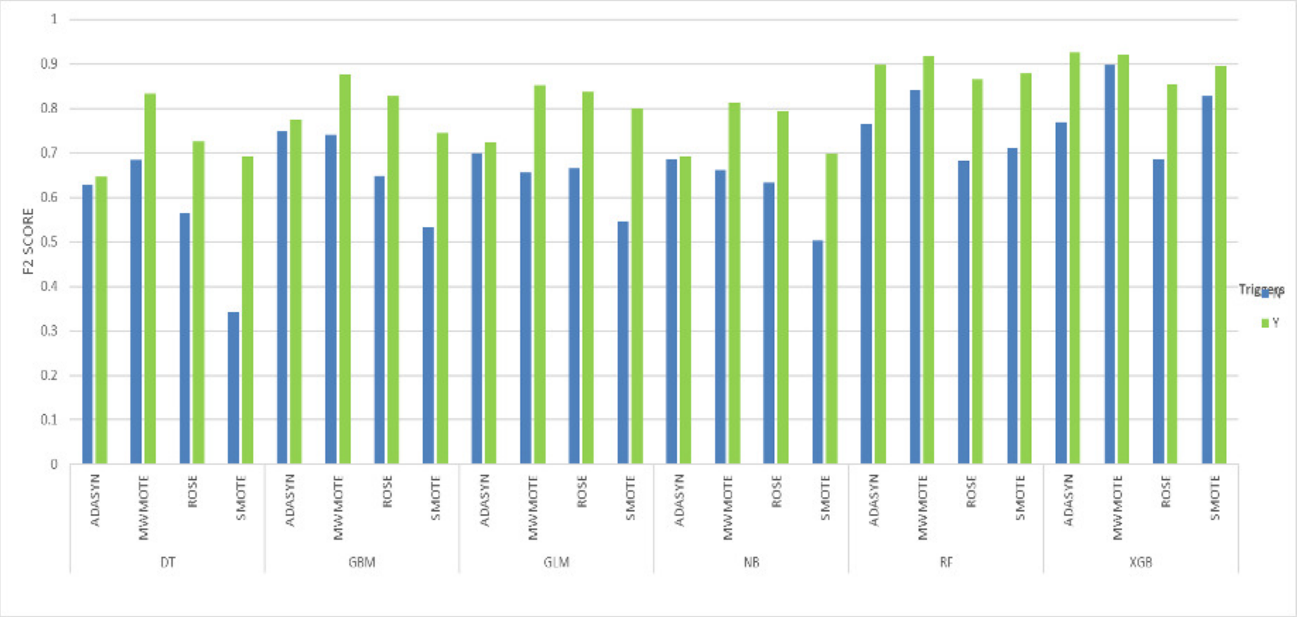

Figure 3. F2 Score plot for all models.

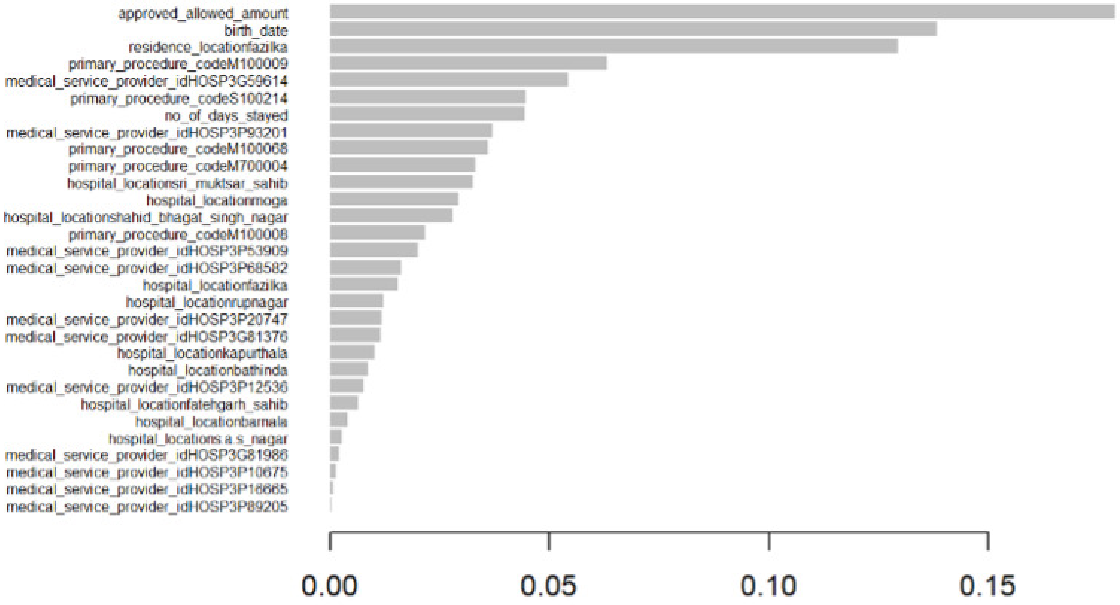

Figure 4. Feature importance: XGBoost ADASYN without triggers.

choosing XGBoost with ADASYN (He et al. 2008) as the optimal model.

Future studies could expand our work by

- applying our method of integrating triggers into ML models to enhance fraud detection in other business domains,
- investigating the interpretability of ML models to improve fraud detection,
- performing network analyses to leverage relationship-based information for detecting organized fraud, and
- performing fraud analysis from a healthcare provider prospective (subject to data availability).

Submitted: April 10, 2023 EDT. Accepted: January 03, 2024 EDT. Published: May 16, 2025 EDT.

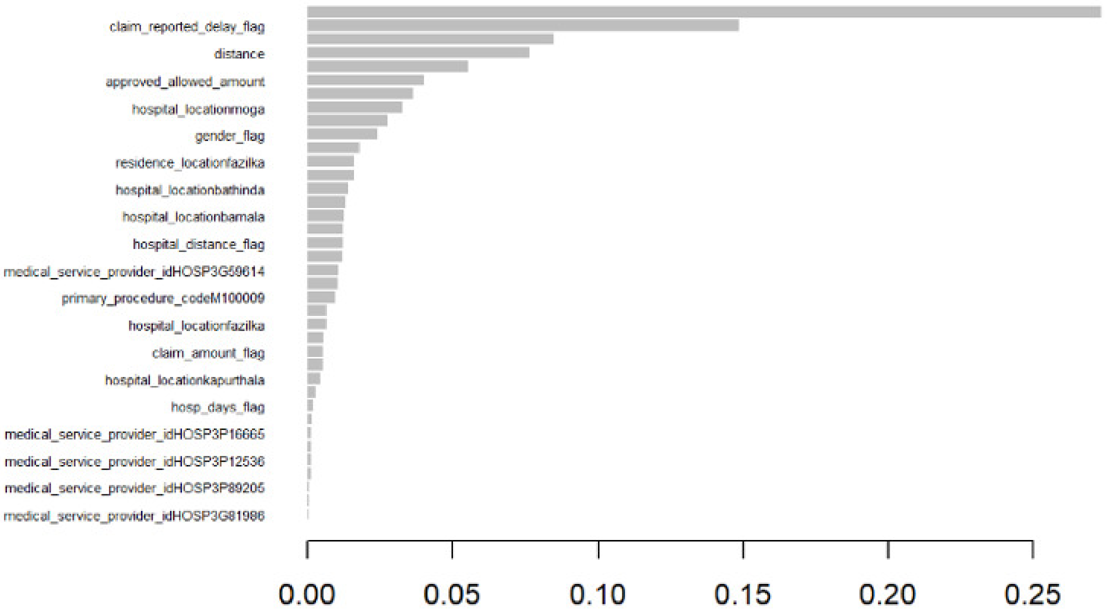

Figure 5. Feature importance: XGBoost ADASYN with triggers.

REFERENCES

Ali, A., S. M. Shamsuddin, and A. L. Ralescu. 2015. “Classification with Class Imbalance Problem: A Review.” *International Journal of Advances in Soft Computing and Its Applications* 5 (3): 176–204.

Aral, K. D., H. A. Güvenir, İ. Sabuncuoğlu, A. R. Akar, I. Sabuncuoĝlu, and A. R. Akar. 2012. “A Prescription Fraud Detection Model.” *Computer Methods and Programs in Biomedicine* 106 (1): 37–46. https://doi.org/10.1016/j.cmpb.2011.09.003.

Aviñó, L., M. Ruffini, and R. Gavaldà. 2018. “Generating Synthetic but Plausible Healthcare Record Datasets.” Preprint, arXiv. https://doi.org/10.48550/arXiv.1807.01514.

Barua, S., M. M. Islam, X. Yao, and K. Murase. 2014. “MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning.” *IEEE Transactions on Knowledge and Data Engineering* 26 (2): 405–25. https://doi.org/10.1109/TKDE.2012.232.

Bauder, R., T. M. Khoshgoftaar, and N. Seliya. 2017. “A Survey on the State of Healthcare Upcoding Fraud Analysis and Detection.” *Health Services and Outcomes Research Methodology* 17:31–55. https://doi.org/10.1007/s10742-016-0154-8.

Bauder, R., R. da Rosa, and T. Khoshgoftaar. 2018. “Identifying Medicare Provider Fraud with Unsupervised Machine Learning.” In *IEEE International Conference on Information Reuse and Integration*, 285–92. https://doi.org/10.1109/IRI.2018.00051.

Becker, D., D. D. Kessler, and M. McClellan. 2005. “Detecting Medicare Abuse.” *Journal of Health Economics* 24 (1): 189–210. https://doi.org/10.1016/j.jhealeco.2004.07.002.

Breiman, L. 2001. “Random Forests.” *Machine Learning* 45:5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. 1st ed. Routledge. https://doi.org/10.1201/9781315139470.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. “SMOTE: Synthetic Minority Over-Sampling Technique.” *Journal of Artificial Intelligence Research* 16:321–57. https://doi.org/10.1613/jair.953.

Chen, H., and J. K. Lindsey. 1998. “Applying Generalized Linear Models.” *Technometrics* 40 (2): 156. https://doi.org/10.2307/1270654.

Chen, T., and C. Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In *KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785.

Dietz, D. K., and H. Snyder. 2007. “Internal Control Differences Between Community Health Centers That Did or Did Not Experience Fraud.” *Research in Healthcare Financial Management* 11 (1): 91–103.

Ekin, T., M. Musal, and L. V. Fulton. 2015. “Overpayment Models for Medical Audits: Multiple Scenarios.” *Journal of Applied Statistics* 42 (11): 2391–2405. https://doi.org/10.1080/02664763.2015.1034659.

Fang, H., and Q. Gong. 2017. “Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked.” *American Economic Review* 107 (2): 562–91. https://doi.org/10.1257/aer.20160349.

Frane, J. W. 1976. “Some Simple Procedures for Handling Missing Data in Multivariate Analysis.” *Psychometrika* 41 (3): 409–15. https://doi.org/10.1007/BF02293565.

Gao, Y., C. Sun, R. Li, Q. Li, L. Cui, and B. Gong. 2018. “An Efficient Fraud Identification Method Combining Manifold Learning and Outliers Detection in Mobile Healthcare Services.” *IEEE Access* 6:60059–68. https://doi.org/10.1109/ACCESS.2018.2875516.

Gomes, C., Z. Jin, and H. Yang. 2021. “Insurance Fraud Detection with Unsupervised Deep Learning.” *Journal of Risk and Insurance* 88 (3): 591–624. https://doi.org/10.1111/JORI.12359.

Guelman, L. 2012. “Gradient Boosting Trees for Auto Insurance Loss Cost Modeling and Prediction.” *Expert Systems with Applications* 39 (3): 3659–67. https://doi.org/10.1016/j.eswa.2011.09.058.

Gupta, R. Y., S. S. Mudigonda, and P. K. Baruah. 2021. “TGANs with Machine Learning Models in Automobile Insurance Fraud Detection and Comparative Study with Other Data Imbalance Techniques.” *International Journal of Recent Technology and Engineering* 9 (5): 236–44. https://doi.org/10.35940/ijrte.E5277.019521.

Gupta, R. Y., S. S. Mudigonda, P. K. Baruah, and P. K. Kandala. 2020. “Implementation of Correlation and Regression Models for Health Insurance Fraud in Covid-19 Environment Using Actuarial and Data Science Techniques.” *International Journal of Recent Technology and Engineering* 9 (3): 699–706. https://doi.org/10.35940/ijrte.C4686.099320.

Gupta, R. Y., S. S. Mudigonda, P. K. Kandala, and P. K. Baruah. 2019. “A Framework for Comprehensive Fraud Management Using Actuarial Techniques.” *International Journal of Scientific & Engineering Research* 10 (3): 780–91.

Gupta, R. Y., S. Sai Mudigonda, P. K. Kandala, and P. K. Baruah. 2019. “Implementation of a Predictive Model for Fraud Detection in Motor Insurance Using Gradient Boosting Method and Validation with Actuarial Models.” In *2019 IEEE International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development*, 1–6. https://doi.org/10.1109/INCCES47820.2019.9167733.

Hassan, A. K. I., and A. Abraham. 2016. “Modeling Insurance Fraud Detection Using Imbalanced Data Classification.” *Advances in Intelligent Systems and Computing* 419:117–27. https://doi.org/10.1007/978-3-319-27400-3_11.

He, H., Y. Bai, E. A. Garcia, and S. Li. 2008. "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning." In *2008 IEEE International Joint Conference on Neural Networks*, 1322–28. https://doi.org/10.1109/IJCNN.2008.4633969.

Ho, T. K. 1995. "Random Decision Forests." In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–82. https://doi.org/10.1109/ICDAR.1995.598994.

Househ, M., and B. Aldosari. 2017. "The Hazards of Data Mining in Healthcare." In *Studies in Health Technology and Informatics*. https://doi.org/10.3233/978-1-61499-781-8-80.

Hutcheson, G. 2001. "Generalized Linear Models." In *The SAGE Dictionary of Quantitative Management Research*, 133–34. SAGE. https://doi.org/10.4135/9781446251119.n40.

Johnson, J. M., and T. M. Khoshgoftaar. 2019. "Medicare Fraud Detection Using Neural Networks." *Journal of Big Data* 61 (1): 63. https://doi.org/10.1186/S40537-019-0225-0.

Johnson, M. E., and N. Nagarur. 2016. "Multi-Stage Methodology to Detect Health Insurance Claim Fraud." *Health Care Management Science* 19 (3): 249–60. https://doi.org/10.1007/s10729-015-9317-3.

Joudaki, H., A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab. 2015. "Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study." *International Journal of Health Policy and Management* 5 (3): 165–72. https://doi.org/10.15171/ijhpm.2015.196.

Khurana, U., F. Nargesian, H. Samulowitz, E. B. Khalil, and D. Turaga. 2017. "Learning Feature Engineering for Classification." *Proceedings of the Twenty-Sixth Joint Conference on Artificial Intelligence*. https://doi.org/10.24963/ijcai.2017/352.

Kirlidog, M., and C. Asuk. 2012. "A Fraud Detection Approach with Data Mining in Health Insurance." *Procedia - Social and Behavioral Sciences* 62:989–94. https://doi.org/10.1016/j.sbspro.2012.09.168.

Koh, H. C., and G. Tan. 2011. "Data Mining Applications in Healthcare." *Journal of Healthcare Information Management* 19 (2): 65. https://doi.org/10.4314/ijonas.v5i1.49926.

Kose, I., M. Gokturk, and K. Kilic. 2017. "An Interactive Machine-Learning-Based Electronic Fraud and Abuse Detection System in Healthcare Insurance." *Applied Soft Computing* 36 (October): 283–99. https://doi.org/10.1016/j.asoc.2015.07.018.

Li, J., K. Huang, J. Jin, and J. Shi. 2008. "A Survey on Statistical Methods for Health Care Fraud Detection." *Health Care Management Science* 11 (3): 275–87. https://doi.org/10.1007/s10729-007-9045-4.

Li, S.-H., D. C. Yen, W.-H. Lu, and C. Wang. 2012. "Identifying the Signs of Fraudulent Accounts Using Data Mining Techniques." *Computers in Human Behavior* 28 (3): 1002–13. https://doi.org/10.1016/j.chb.2012.01.002.

Lieberthal, R. D., J. Ai, S. D. Smith, and R. L. Wojciechowski. 2018. "Examining Predictive Modeling Based Approaches to Characterizing Health Care Fraud." In *7th Annual Conference on Economics and Public Health*. https://ashecon.confex.com/ashecon/2018/meetingapp.cgi/Paper/5997.

Lin, W. C., and C. F. Tsai. 2020. "Missing Value Imputation: A Review and Analysis of the Literature (2006–2017)." *Artificial Intelligence Review* 53 (2): 1487–1509. https://doi.org/10.1007/s10462-019-09709-4.

Liou, F.-M., Y.-C. Tang, and J.-Y. Chen. 2008. "Detecting Hospital Fraud and Claim Abuse Through Diabetic Outpatient Services." *Health Care Management Science* 11 (4): 353–58. https://doi.org/10.1007/s10729-008-9054-y.

Lu, F., and J. E. Boritz. 2005. "Detecting Fraud In Health Insurance Data: Learning to Model Incomplete Benford's Law Distributions." *Machine Learning: ECML 2005* 3720:633–40.

Major, J. A., and D. R. Riedinger. 2002. "EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud." *Journal of Risk & Insurance* 69 (3): 309–24. https://doi.org/10.1111/1539-6975.00025.

Mesa, F. R., A. Raineri, S. Maturana, and A. M. Kaempffer. 2009. "Fraud in the Health Systems of Chile: A Detection Model." *Pan American Journal of Public Health* 25 (1): 56–61. https://doi.org/10.1590/s1020-49892009000100009.

Mohammed, R., J. Rawashdeh, and M. Abdullah. 2020. "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results." In *11th International Conference on Information and Communication Systems*, 243–48. https://doi.org/10.1109/ICICS49469.2020.239556.

Pande, V., and W. Maas. 2013. "Physician Medicare Fraud: Characteristics and Consequences." *International Journal of Pharmaceutical and Healthcare Marketing* 7 (1): 8–33. https://doi.org/10.1108/17506121311315391.

Patidar, P., and A. Tiwari. 2013. "Handling Missing Value in Decision Tree Algorithm." *International Journal of Computer Applications* 70 (13): 31–36. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a6fb36e2d0e2e5bfcc8024cf036e48ca5710ab66.

Richman, R. 2018. "AI in Actuarial Science." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3218082.

Sadiq, S., Y. Tao, Y. Yan, and M. L. Shyu. 2017. "Mining Anomalies in Medicare Big Data Using Patient Rule Induction Method." In *2017 IEEE Third International Conference on Multimedia Big Data*. https://doi.org/10.1109/BigMM.2017.56.

Seliya, N., T. M. Khoshgoftaar, and J. Van Hulse. 2009. "A Study on the Relationships of Classifier Performance Metrics." In *21st IEEE International Conference on Tools with Artificial Intelligence*, 59–66. https://doi.org/10.1109/ICTAI.2009.25.

Shin, H., H. Park, J. Lee, and W. C. Jhee. 2012. "A Scoring Model to Detect Abusive Billing Patterns in Health Insurance Claims." *Expert Systems with Applications* 39 (8): 7441–50. https://doi.org/10.1016/j.eswa.2012.01.105.

Srinivasan, U., and B. Arunasalam. 2013. "Leveraging Big Data Analytics to Reduce Healthcare Costs." *IT Professional* 15 (6): 21–28. https://doi.org/10.1109/MITP.2013.55.

Thabtah, F., S. Hammoud, F. Kamalov, and A. Gonsalves. 2020. "Data Imbalance in Classification: Experimental Evaluation." *Information Sciences* 513:429–41. https://doi.org/10.1016/j.ins.2019.11.004.

Thornton, D., R. M. Mueller, P. Schoutsen, and J. van Hillegersberg. 2013. "Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection." *Procedia Technology* 9:1252–64. https://doi.org/10.1016/j.protcy.2013.12.140.

van Capelleveen, G., M. Poel, R. M. Mueller, D. Thornton, and J. van Hillegersberg. 2016. "Outlier Detection in Healthcare Fraud: A Case Study in the Medicaid Dental Domain." *International Journal of Accounting Information Systems* 21:18–31. https://doi.org/10.1016/j.accinf.2016.04.001.

Vanhoeyveld, J., D. Martens, and B. Peeters. 2020. "Customs Fraud Detection: Assessing the Value of Behavioural and High-Cardinality Data Under the Imbalanced Learning Issue." *Pattern Analysis and Applications* 23 (3): 1457–77. https://doi.org/10.1007/s10044-019-00852-w.

Verma, A., A. Taneja, and A. Arora. 2017. "Fraud Detection and Frequent Pattern Matching in Insurance Claims Using Data Mining Techniques." In *2017 Tenth International Conference on Contemporary Computing (IC3)*, 1–7. https://doi.org/10.1109/IC3.2017.8284299.

Victorri-Vigneau, C., K. Larour, D. Simon, and P. Jolliet. 2009. "Creating and Validating a Tool Able to Detect Fraud by Prescription Falsification from Health Insurance Administration Databases." *Therapie* 64 (1): 27–31. https://doi.org/10.2515/therapie/2009004.

Villegas-Ortega, J., L. Bellido-Boza, and D. Mauricio. 2021. "Fourteen Years of Manifestations and Factors of Health Insurance Fraud 2006–2020: A Scoping Review." *Health Justice* 9 (26): 1–23. https://doi.org/10.1186/s40352-021-00149-3.

Yang, W.-S., and S.-Y. Hwang. 2006. "A Process-Mining Framework For The Detection Of Healthcare Fraud And Abuse." *Expert Systems with Applications* 31 (1): 56–68. https://doi.org/10.1016/j.eswa.2005.09.003.

Yoo, I., P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua. 2012. "Data Mining in Healthcare and Biomedicine: A Survey of the Literature." *Journal of Medical Systems* 36:2431–48. https://doi.org/10.1007/s10916-011-9710-5.

Zheng, A., and A. Casari. 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly.

Zhou, S., J. He, H. Yang, D. Chen, and R. Zhang. 2020. "Big Data-Driven Abnormal Behavior Detection in Healthcare Based on Association Rules." *IEEE Access* 8:129002–11. https://doi.org/10.1109/ACCESS.2020.3009006.

APPENDIX

All datasets, code, and trigger functions developed and used as part of this project can be accessed from the below GitHub Repo:

https://github.com/RohanYashraj/CAS-Project-Health-Insurance-Fraud-Detection

Access can be provided to the same upon request.