Fraud Detection in Auto Insurance Claims Using Machine Learning Algorithms

Pradip Ravindra Jagdale*, Manasi Manoj Sukale

Department of Statistics, Dr. D. Y. Patil, Art's Commerce and Science College, Pimpri, Pune, Maharashtra, India

Abstract: Insurance fraud is a major problem that threatens both the stability and fairness of insurance systems. This study explores how machine learning techniques—such as Logistic Regression, Decision Trees, Random Forest, and XGBoost—can be applied to identify fraudulent auto insurance claims. The models obtain great accuracy, precision, recall, and F1-score, demonstrating their capacity to distinguish between false and legitimate claims. The performance of the models is further enhanced and improved prediction accuracy is ensured by the use of advanced approaches like feature selection and hyperparameter tuning. Overall, by offering a thorough review of machine learning algorithms and their use in identifying fraudulent claims, this project makes a contribution to the field of auto insurance fraud detection. Insurance businesses can use the created models and procedures to improve their fraud detection processes, reduce financial risks, and safeguard their operations from fraudulent activity Using a real-world dataset from Kaggle, we applied preprocessing techniques, feature selection via Recursive Feature Elimination, and data balancing through SMOTE. Out of all the models tested, XGBoost showed the highest performance, achieving an accuracy of 89% and an F1-score of 87%. The paper highlights the effectiveness of AI-driven detection systems in minimizing financial loss, improving risk management, and ensuring fairness in insurance systems.

Keywords: Insurance fraud, Machine Learning, XGBoost, Auto claims, SMOTE

I. Introduction

When an act is committed with the purpose of securing a favorable but false result during an insurance claim, it is referred to as insurance fraud. The act of making a false insurance claim, as well as exaggerating losses, injuries, or damages in order to obtain benefits, is examples of insurance fraud.

Types of Insurance Fraud: Insurance fraud can take many forms, but it's generally classified into two main types: hard fraud and soft fraud. Hard fraud occurs when a person deliberately stages or fakes an accident, theft, or injury in order to make a false claim and get money from the insurance company.

1) Life insurance fraud: It entails purposeful deception or manipulation with the goal of obtaining monetary advantages from a life insurance policy through fraudulent methods.

2) Property Insurance Fraud: When someone commits property insurance fraud, they either steal or intentionally destroy items of personal property, buildings, or even cars in order to receive compensation from the insurance provider. This is typically done because the person is cash-strapped and the insurance payout is frequently higher than what the property would be worth if it were to be sold outright. When someone alleges property loss or damage, they may also exaggerate the claim by alleging ownership of objects they never actually owned or by representing them as newer and better-quality than they actually were.

3) Health insurance fraud: The term "health insurance fraud" describes deceptive practices primarily involving health insurance policies and claims. In order to get financial benefits to which they are not entitled, it happens when people or healthcare professionals purposefully deceive insurance companies or manipulate the healthcare system. There are many different types of health insurance fraud, which can involve patients, doctors, hospitals.

4) Auto Insurance Fraud: Auto insurance fraud involves dishonest actions related to car insurance policies. It happens when someone knowingly provides false information or tricks the insurance company to get money or benefits they shouldn't receive. Common examples include staging car accidents, exaggerating damage or injuries in claims, faking repair bills, using false documents, not paying full premiums, or illegally disposing of vehicles

Current detection methods to detect fraud:1) Advanced Analytics and Machine Learning: Modern fraud detection in the insurance sector is led by advanced analytics and machine learning. These tools provide sophisticated methods for evaluating huge quantities of data and spotting intricate patterns that can point to fraudulent activity. 2) Text Mining and Natural Language Processing: extual Data in Insurance Fraud Detection:

In the insurance sector, documents like claim forms, medical records, legal files, and other written materials often contain valuable information in unstructured text format. These tools help uncover hidden patterns and suspicious behaviour in claims, playing a key role in detecting fraudulent activities: i) Text Mining: In the context of insurance fraud, it can help identify unusual patterns, detect duplicate or suspicious claims, and flag inconsistencies in the information provided. ii) Natural Language Processing (NLP):

With NLP, insurers can analyse written statements, compare past claims, detect contradictions, and even understand the intent behind certain phrases — helping them spot potentially fraudulent claims. iii) Real-Time Monitoring and Alert Systems**:** These systems continuously track claim activities and instantly flag any behaviour that seems out of the ordinary. By using automated alerts and dashboards, insurance companies can quickly investigate and act on suspicious claims, reducing losses due to fraud.

II. Literature Review

1) Author: Natalia Markovskaia (Published on Jul. 09, 2020). The use of machine learning techniques for detecting insurance fraud is examined in this literature review. The field of artificial intelligence's machine learning presents some promising solutions for improving the efficacy and accuracy of fraud detection because identifying insurance fraud is difficult. This review discusses the benefits of machine learning, such as discovering previously undetected connections between variables that may be undetectable to individuals and forecasting new fraud schemes. The commonly utilized deep anomaly detection technique, in which models are built based on typical claims and applied to big datasets, is also highlighted in this work. For the purpose of detecting fraud, algorithms like Random Forest, Artificial Neural Networks, Support Vector Machines, K-Nearest Neighbors, and Logistic Regression are utilized. The review emphasizes the usage of automated fraud detection technology by corporate partners as well as the necessity of relationships between established businesses and cutting-edge startups in order to define the future of the insurance industry.

2) Daniel Aksman, Zackary Fitzgibbon , Alec Kneedler, Arseniy Pozdniakov (May 2020): This paper aims to enhance fraud detection in Russian banks by developing an emotion classification algorithm using voice data to reduce phone scam. The team conducted research on voice identification technologies and identified vocal features correlated with emotional states. In collaboration with The Financial University, four artificial intelligences were created, including one for client identification and three for detecting emotions to assist in identifying socially engineered individuals. The study focused on capturing the soundscapes of naturally anxious people. Authentic anxiety is crucial in fraud detection AI. because Fraudsters are likely to exhibit genuine anxious emotions. After collecting and extracting the data they tested the data on random forest, linear regression, and multi-layered perceptron (mesh neural network) models. And uses teqniques such as feature selection and scaling to increase accuracy.

3) Sharmila Subudhi Suvasini Panigrahi (July 2018):The article describes a cutting-edge method for employing data mining tools to find fraud in vehicle insurance claims. The suggested system entails the following steps: Feature Selection: To choose the most pertinent qualities from the initial dataset, an evolutionary algorithm-based feature selection method is used. After feature selection, A balanced dataset is produced as a result of the PFCM clustering's assistance in removing outliers from the majority class samples. Model Training: A collection of Weighted Extreme Learning Machine (WELM) classifiers is trained using the balanced dataset. These classifiers are produced using various WELM parameter combinations. Model Validation: On the balanced dataset, the WELM classifiers' effectiveness is verified using the 10-fold cross-validation approach. The dataset is split into 10 subsets using this manner, with 9 subsets being used for training and the ultimate subset being used for validation. The outcomes are averaged after the process is conducted numerous times. Fraud Detection: The best-performing model created from the preceding stage for classification is applied to the test set, taken from the chosen attribute set, for fraud detection. A dataset containing actual instances of vehicle insurance fraud is used to assess the proposed method. A comparative study is done to show how the suggested system is better than another approach. In addition, the article emphasizes the significance of fraud detection in vehicle insurance claims, the commonality of insurance fraud, and the difficulties in identifying fraudulent activity. In order to improve classification performance, this study addresses the problem of unbalanced class distribution in fraud detection.

4) Carol Hargreaves and Vidyut Singhania (January 2016): The paper examines the common issue of auto insurance fraud and highlights the significant expenses associated with manual fraud detection methods. It offers a methodology for detecting insurance fraud and suggests using data analytics to pinpoint important characteristics for fraud detection. To ascertain the significance of 31 variables in identifying fraud, the authors conducted statistical hypothesis testing to help in detecting fraud and to identify which of them are significant, they used the independent samples t-test for comparing mean of continuous variable and the Chi-Square test to check the association of categorical variable by using the software SPSS 22 summarized results based on significant variables.

III. Methodology

Data Source: https://www.kaggle.com/ Data Structure: This data is available in excel format; there are 39 columns 1000 rows in the data.

Data information: variables/features that comprise the dataset represent:

	Customer_duration	Years_old	Policy_unique_id	Policy_activation_date	Policy_region_state	Policy_combined_single_limit	Policy_deductable_amount
0	328	48	521585	10/17/2014	OH	250/500	1000
1	228	42	342868	6/27/2006	IN	250/500	2000
2	134	29	687698	9/6/2000	OH	100/300	2000
3	256	41	227811	5/25/1990	IL	250/500	2000
4	228	44	367455	6/6/2014	IL	500/1000	1000

Fig.1 data set

There are 21 categorical variables and 18 numerical variables in this data.

It is a classification problem. Predicting discrete or categorical outcomes based on input variables or attributes is a part of classification problems. The objective here is to categorize whether a claim for auto insurance is fraudulent or not. The "yes" and "no" values of the output variable correspond to the two unique classes or categories that the model is intended to predict. The dataset contains no missing values. Now, we proceed to remove the variables that are not needed for analysis.

Policy_unique_id, Policy_activation_date, Date_of incident, Incident-Location, Zip_code, Years_old, Policy_region_state, Insured_Gender, Incident-hour, Insured_kinship, Incident_location_state, Incident_location_city_ city, Capital_Gain, Capital_Loss. We have dropped above 14 columns, which are the least important for further analysis and could create noise in the analysis.

K-Modes Clustering Imputation Method:

The data imputation approach known as the k-modes clustering imputation method is typically applied to categorical or nominal data. The k-means clustering algorithm, which is frequently employed for numerical data, is extended by this technique.

DATA VISUALIZATION

A] Univariate analysis: 1) Plot of Type of Incident and Severity Level:
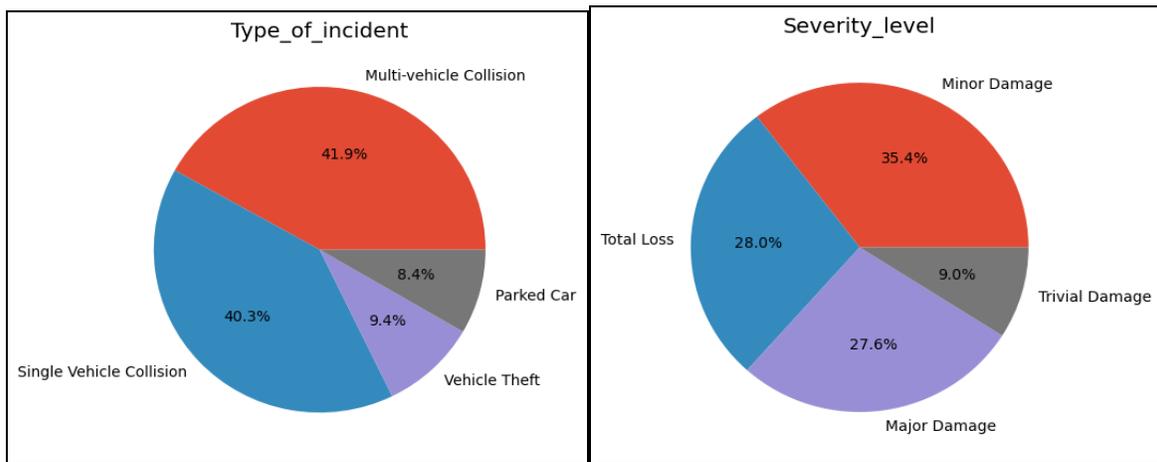

Fig 2 Type of Incident   Severity Level

Interpretation: The claims registered after the Multi-vehicle Collision is greater than the other 3 collisions. The occurrence of a multi-vehicle collision suggests that an accident involving multiple vehicles took place on a large scale, which could indicate a higher likelihood of damage or injuries compared to other types of incidents.
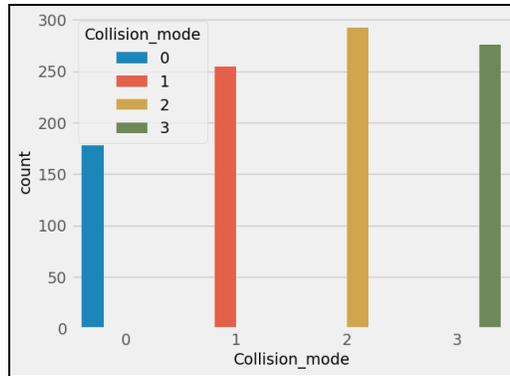
2)Plot of Collision Mode:

Fig.3   Collision Mode

0=other   1=front collision   2=rear collision   3=side collision

Interpretation: Suggests that rear-end accidents occur more frequently compared to other types of collisions. This may be caused by several factors, including tailgating, distracted driving, or sudden braking. Rear collisions can result in a wide range of damages, including bumper damage, trunk damage, or even structural issues. These repairs can be costly, prompting policyholders to file claims with their insurance company to cover the expenses.
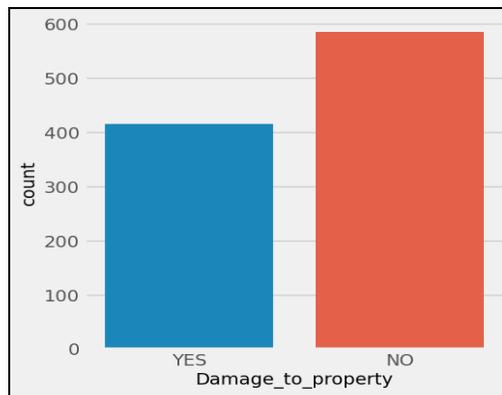
3) Plot of Damage to property:

Fig.4 **Damage to property**

Interpretation: The highest number of fraud cases is reported when major damage occurs. In simpler terms, fraud is most commonly associated with instances where significant damage has been done to the insured property
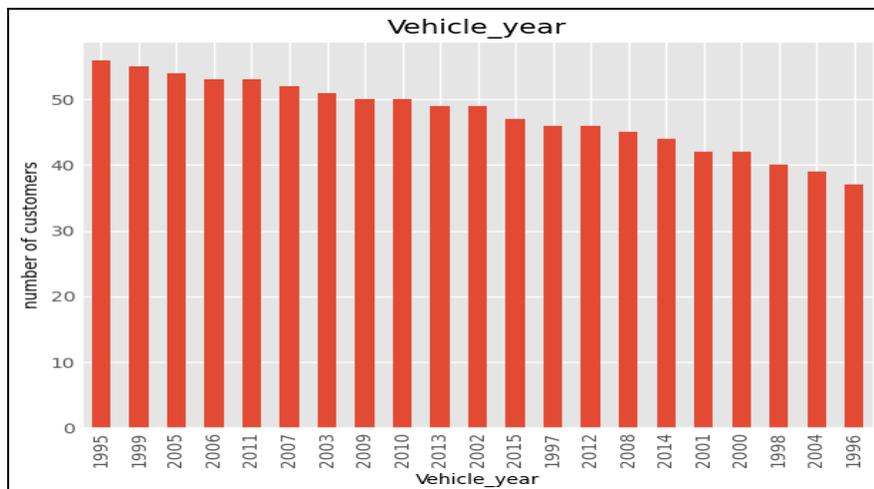
4) Plot of Vehicle year:

Fig.5 Vehicle year

Interpretation: The highest claims are made on vehicle manufactured in years 1995, 1999. The lowest claims are made on vehicle manufactured in years 1996, 2004.

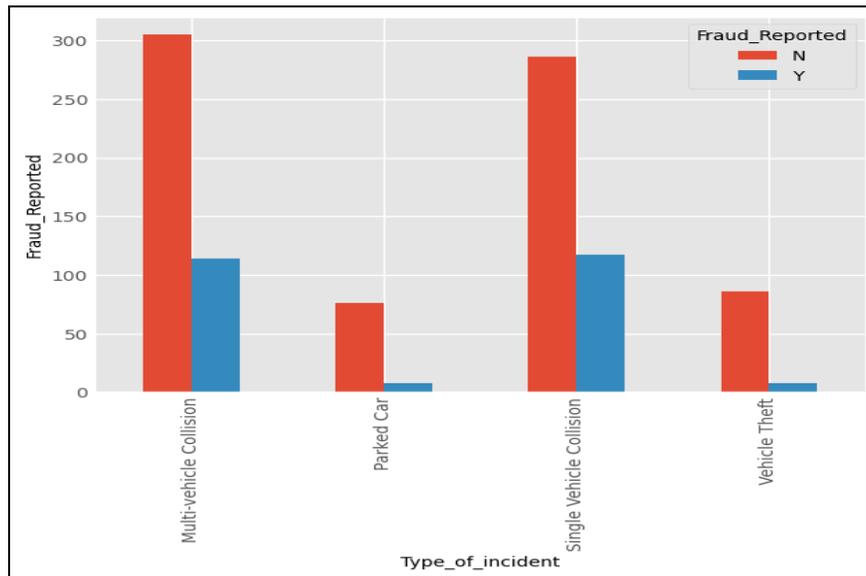**B] Bivariate analysis: 1) Plot of Fraud Reported and Type of incident:**

Fig.6 Fraud Reported and Type of incident

Interpretation: The analysis found that the highest number of fraud cases were reported when either a single vehicle collision or a multiple vehicle collision occurred. In simpler terms, the instances of fraud were most commonly observed in situations where there was either a collision involving only one vehicle or a collision involving multiple vehicles.
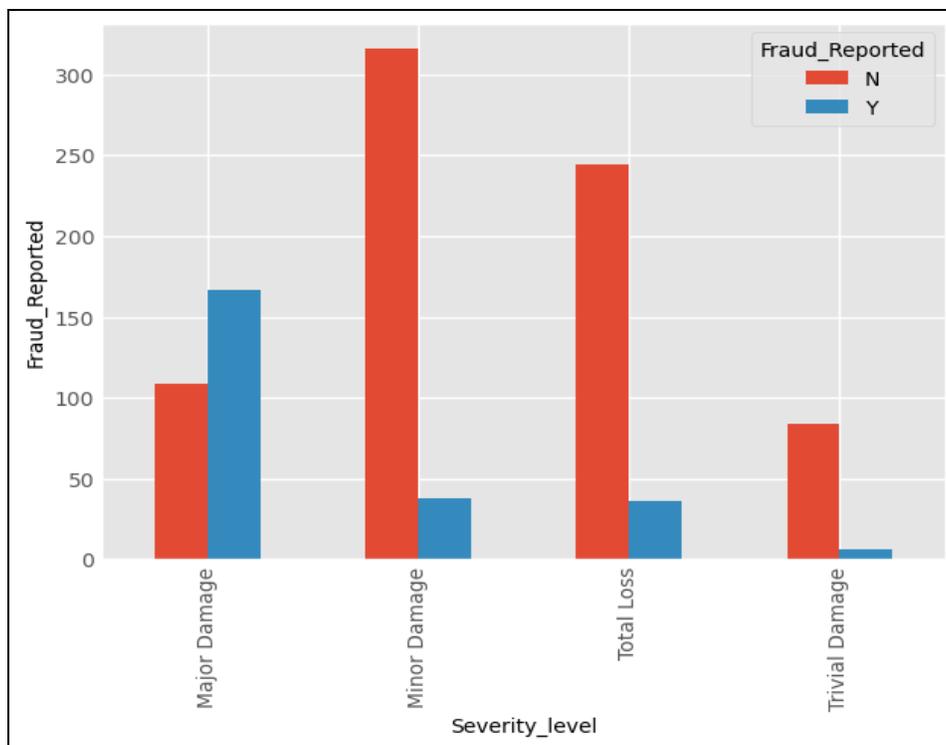
2) Plot of Severity level vs Fraud Reported:

Fig.7 Severity level vs Fraud Reported

Interpretation: The claims made on minor damage is higher than other 3 types of damages. The maximum number of fraud are reported when major damage is happened.
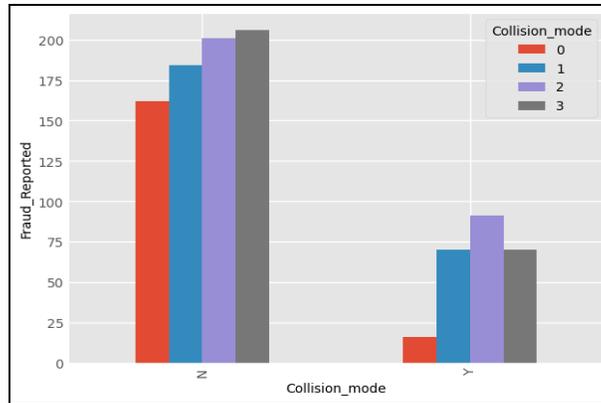
3)**Plot of Collision mode vs Fraud Reported:**

Fig 8 Collision mode vs Fraud Reported

Interpretation: The highest number of fraud cases, specifically 91 instances, are reported when rear collisions occur. In simpler terms, fraud is most frequently observed in cases where the collision involves the rear of the vehicle.

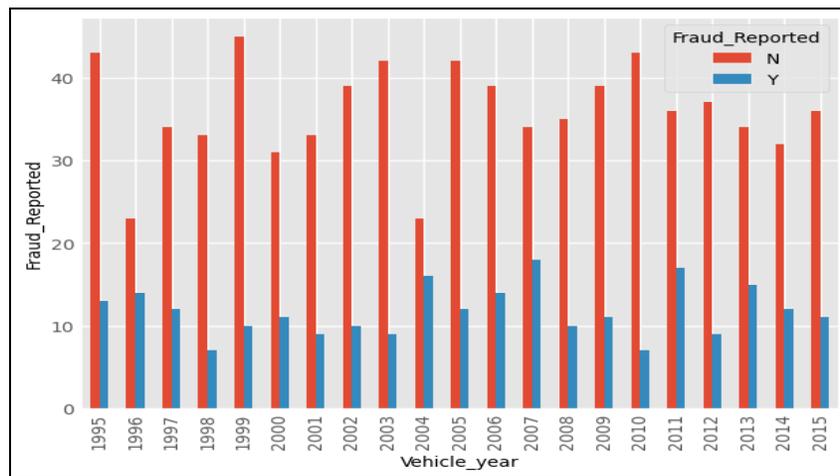4) **Plot of Vehicle year vs Fraud Reported:**

Fig.9 Vehicle year vs Fraud Reported

Interpretation: The maximum number of frauds are reported on claims which are made on vehicle manufactured in year 2007 and 2004.
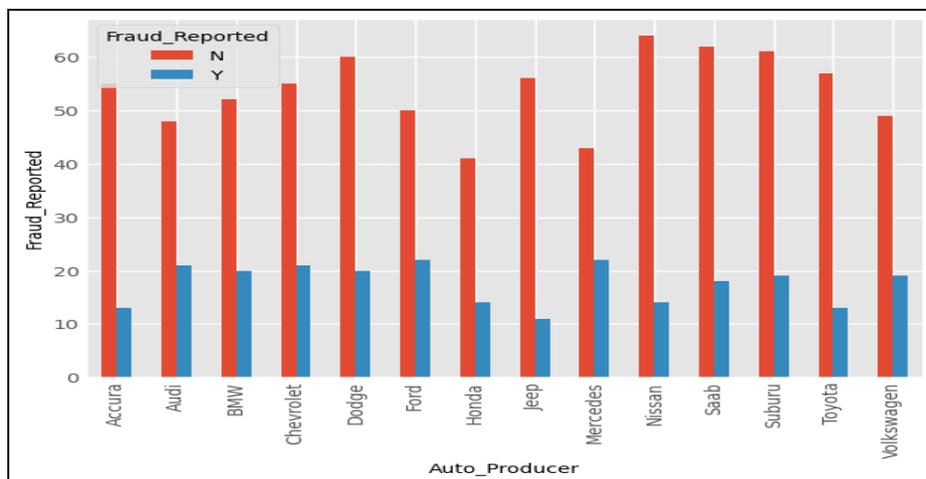
5) **Plot of Auto producer vs Fraud Reported:**

Fig.10 Auto producer vs Fraud Reported

Interpretation: The maximum numbers of claims registered on auto producers Dodge, Chevrolet, Saab and Suburu.
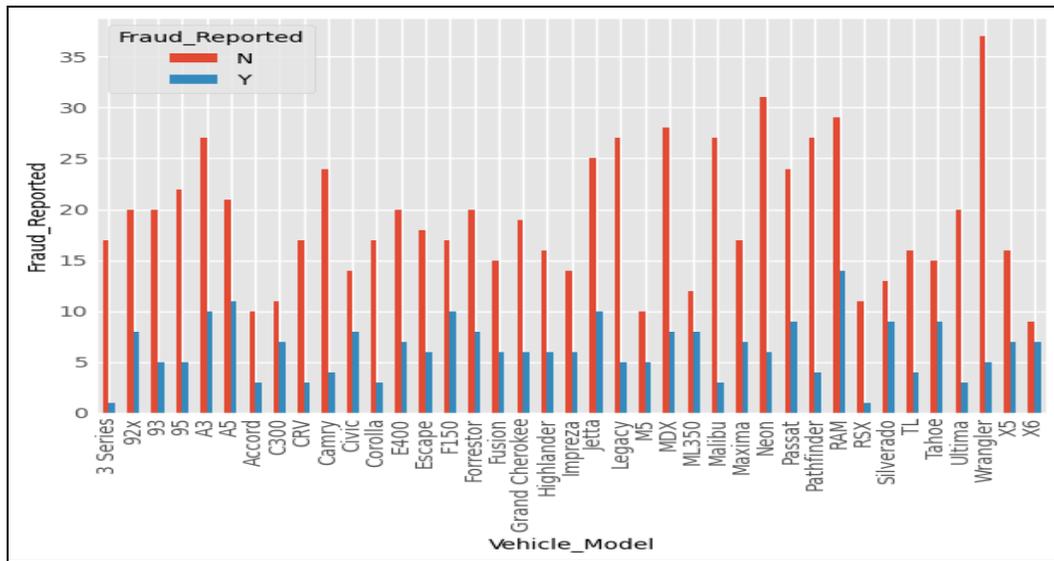
6) Plot of Vehicle Model vs fraud reported:

Fig.11 Vehicle Model vs fraud reported

Interpretation: The maximum fraudulent claims are made on auto model RAM and least claims are made on auto model RSX and 3 Series.

IV. Result and Discussion

Analysis using software: Feature Selection for Fraud Detection in Auto Insurance Claims

Choosing a subset of relevant characteristics from a broader pool of accessible features is a critical step in machine learning and data analysis. By concentrating on the most illuminating and discriminative elements, feature selection aims to optimize model performance, decrease over fitting, and improve interpretability.

Here we have used recursive feature elimination method for selecting most appropriate and useful features from dataset.

Label encoding:

According to this encoding technique, each distinct category or label within a categorical variable is given a distinct integer value. When categorical variables have an ordinal structure and the order of the categories matters, label encoding is very helpful.

Recursive Feature Elimination (RFE):

The most crucial characteristics from a given dataset are chosen using the feature selection technique known as recursive feature elimination (RFE). it is an iterative process that starts with all features and eliminates those that aren't important enough, based on the feature importance ranking of a chosen machine learning algorithm.

The selected features after using recursive feature elimination technique are:

```
['Customer_duration',
 'Policy_combined_single_limit',
 'Policy_deductable_amount',
 'Policy_premium_per_year',
 'Umbrella_insurance_limit',
 'Insured_profession',
 'Insured-Hobbies',
 'Insured_kinship',
 'Collision_mode',
 'Severity_level',
 'Contacted_Authorities',
 'Physical_injury',
 'Witnesses_present',
 'Total_Claim',
 'Bodily_Injury_Claim',
 'Property_Damage_Claim',
 'Vehicle_Damage_Claim',
 'Auto_Producer',
 'Vehicle_Model',
 'Vehicle_year']
```

Fig 12 features after using recursive feature elimination technique

Balancing the imbalanced data:
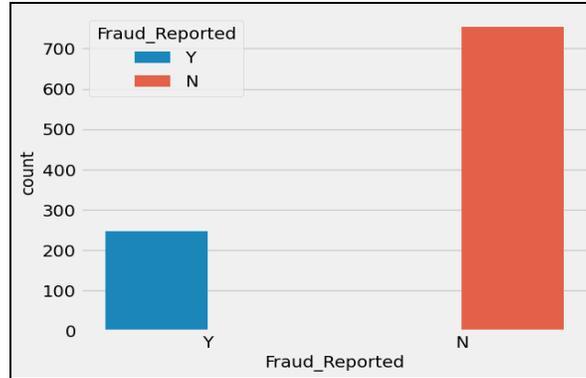

Fig 13 Balancing the imbalanced data

percentage of NO and YES

N   75.3

Y   24.7

This % indicates that the data is imbalanced

So, for balancing the dataset we have used SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE (Synthetic Minority Over-sampling Technique):

SMOTE is a technique used in machine learning to handle imbalanced data, especially when there are very few examples of the minority class. Class imbalance arises when one or more classes have disproportionately fewer occurrences than the other(s), which can result in biased models that have trouble correctly predicting the minority class.

Model creation and model evaluation: Hyper parameter tuning:

The process of choosing the best values for a machine learning model's hyperparameters is known as hyperparameter tuning Before training a model, the user needs to manually define the hyperparameters, as they are not learned from the data. These settings control how the model operates and influence its overall performance.

Model creation and evaluation using software:

1) Logistic Regression:

Here the dataset target variable that is Fraud_ Reported variable contains binary values that are 'yes' or 'no'

Confusion matrix:

$$[181 \quad 2]$$
$$[\ 67 \quad 0]$$

Accuracy = (True Positives + True Negatives) ÷ (All Predictions)

$$= [(181+0)/(181+2+67+0)]$$

$$= 0.724$$

Accuracy of model in percentage = 0.724*100 = 72%

Precision =0.989.

Recall =0.730.

F1 Score =2 * (0.989 * 0.730) / (0.989 + 0.730)

$$= 0.839$$

Confusion matrix has a true negative count of 0, which makes the calculation of specificity and F1 score undefined or not meaningful in this context.

Classification report:

```
              precision    recall  f1-score   support

           0       0.73      0.99      0.84       183
           1       0.00      0.00      0.00        67

    accuracy                           0.72       250
   macro avg       0.36      0.49      0.42       250
weighted avg       0.53      0.72      0.61       250
```

From above results we can see that if we fir the logistic model without using the SMOTE technique it will give non reliable results.

So, to overcome this situation now we have used logistic regression using SMOTE technique and the result we got as follows:

Confusion matrix =

$$[70 \ 72]$$
$$[71 \ 89]$$

Accuracy = (True Positives + True Negatives) ÷ (All Predictions)

$$159 / 302$$
$$\approx 0.526$$

Accuracy of model in percentage = 0.52*100 = 52%

Precision = TP / (TP + FP) = 70 / (70 + 72) ≈ 0.493

Recall = 70 / (70 + 71) ≈ 0.496

F1-Score) =

$$= 2 * (0.493 * 0.496) / (0.493 + 0.496)$$
$$= 0.494$$

We got accuracy of logistic model after using SMOTE technique as 52%

2) Decision Tree: It is a graphical visual representation of a collection of laws that aids in decision or prediction. Decision trees are useful in the area of fraud detection because they can capture intricate linkages and patterns in the data and offer transparent and comprehensible rules.

The results we got after using the decision tree are as follows:

The confusion matrix:

Model evaluation metrics	Before using hyperparameter	After using hyperparamter
Accuarcy	0.84	0.85
precision	0.86	0.845
recall	0.813	0.833
F1 score	0.835	0.84
specificity	0.87	0.86

3)Random Forest Classifier:

A machine learning ensemble approach called Random Forest Classifier combines several decision trees to produce predictions. It is an effective classification method that may be used for a variety of applications, including fraud detection.

Model evaluation metrics	Before using hyperparameter	After using hyper paramter
Accuracy	0.83	0.83
precision	0.89	0.84

recall	0.78	0.80
F1 score	0.84	0.83
specificity	0.89	0,86

4) XGB Classifier: The gradient boosting family of machine learning algorithms includes XGB Classifier, commonly referred to as XGBoost Classifier. Extreme Gradient Boosting, or XGBoost, is a method that combines the predictions of various weak models, like decision trees, to produce a highly accurate predictive model.

Model evaluation metrics	Before using hyperparameter	After using hyper paramter
Accuracy	0.88	0.89
precision	0.87	0.86
recall	0.86	0.88
F1 score	0.87	0.87
specificity	0.88	0.87

Comparing the accuracy of best models:

Model	Before using hyper paramter tuning	After hyper parameter tunning
Decision tree	0.84	0.85
Random forest classifier	0.83	0.83
XGB Classifier	0.88	0.89

XGB is chosen as the final model. Is not over fitting. Has highest accuracy as 89%. Has highest f1 score as 87%.
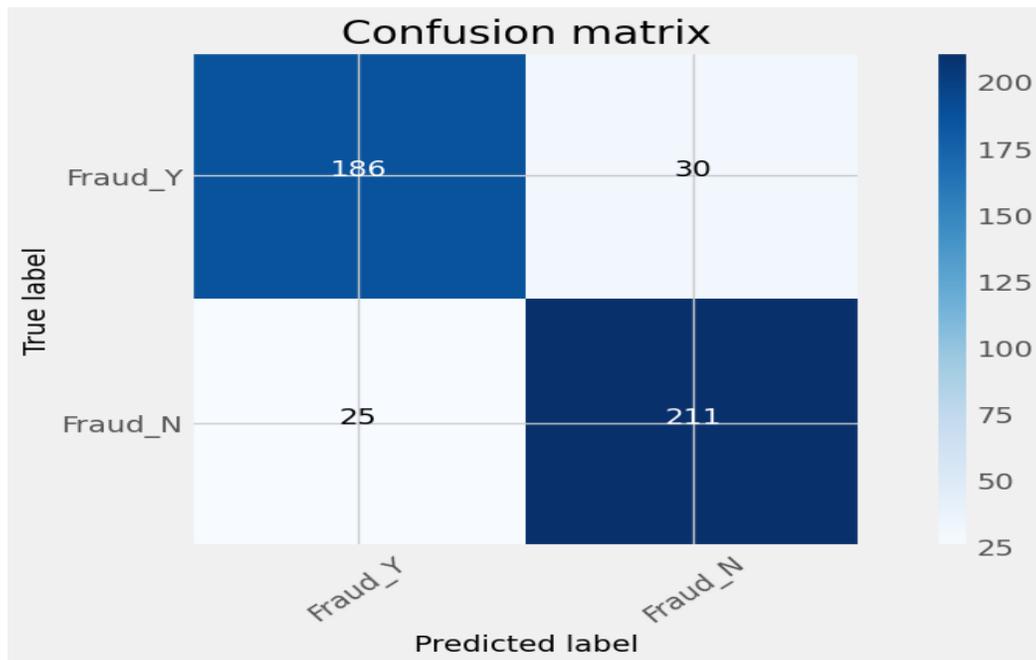
From the confusion matrix

Fig 14 confusion matrix

There are 211 instances where the model correctly predicted a fraud claim (actual fraud claim correctly identified as fraud). There are 30 instances where the model incorrectly predicted a fraud claim (non-fraud claim incorrectly identified as fraud). These are cases where the model flagged a claim as fraudulent, but it was actually not fraudulent. There are 186 instances where the model correctly predicted a non-fraud claim (actual non-fraud claim correctly identified as non-fraud). There are 25 instances where the model incorrectly predicted a non-fraud claim these are cases where the model failed to identify a fraudulent claim. With accuracy 89% and precision 0.87, this means that out of all the cases predicted as fraud, 87% were actually fraud.

	feature	imp
55	Severity_level_Total Loss	0.086019
53	Severity_level_Major Damage	0.085591
54	Severity_level_Minor Damage	0.085051
56	Severity_level_Trivial Damage	0.072787
11	Policy_combined_single_limit_250/500	0.039673
12	Policy_combined_single_limit_500/1000	0.031791
16	Insured_kinship_own-child	0.026057
57	Contacted_Authorities_Ambulance	0.021431
38	Insured-Hobbies_chess	0.021029
58	Contacted_Authorities_Fire	0.019403
5	Total_Claim	0.019229
10	Policy_combined_single_limit_100/300	0.018796
39	Insured-Hobbies_cross-fit	0.018775
42	Insured-Hobbies_golf	0.015259
1	Policy_deductable_amount	0.015221

Fig.15 Feature importance based on final model

The above table shows top 15 important features. according to the feature importance drivers who enjoyed hobbies like cross-fit or chess appeared to be more susceptible to fraud and severity level having highest feature importance among all features. The project's objectives were to minimize risk, protect the auto insurance company against false claims, increase accuracy, and make it possible for loss control units to provide greater coverage.

Five different classifiers were used in this project: logistic regression, Decision Tree, Random Forest, XGBoost. We also have used way of handling imbalance classes that is SMOTE. Hyperparameter tunning for optimizing the performance of a machine learning models who having higher accuracy. The best fitted model we got is XGB Classifier. The chosen model, XGBoost, showed promising performance in identifying insurance fraud in auto claims. It had an F1 score of 87% and an accuracy of 89%, demonstrating an excellent balance between recall and precision. These findings indicate that the model can distinguish between fraudulent and non-fraudulent claims with accuracy, reducing the financial losses brought on by fraudulent activity.

When a model excels on the training data but fails to generalize to new data, this is known as overfitting. To make sure the selected model is not overfitting the data, it was carefully assessed. by using appropriate hyperparameter tuning, the model was able to maintain good generalization performance. XGBoost is known for its proficiency in handling complicated datasets and effectively capturing non-linear relationships between variables. It uses a group of interconnected decision trees to generate reliable predictions. Additionally, XGBoost offers feature importance scores that let us pinpoint the elements that have the greatest influence on spotting fraud in auto insurance claims. Understanding the underlying patterns and behaviors connected to fraudulent operations may be possible with the help of this information. The created fraud detection system has the potential to greatly enhance the loss prevention strategies used by the auto insurance firm. A company can prevent financial losses and ensure fair and honest insurance procedures by accurately identifying fraudulent claims and taking the required measures. The technology can also help with coverage optimization by concentrating resources on valid claims and speeding up the processing of valid case.

Limitations and Future Scope:

Data Restrictions: The quality and quantity of the data are critical to the models' efficacy. The accuracy and generalizability of the results could be affected if the dataset having missing some crucial components. Additionally, the inclusion of incomplete or missing data might induce biases and impact the models' overall effectiveness. Limited sample size can also be the limitation of this study; larger data sets increase the stability of statistical models.

Evolving Fraud Patterns: Fraudulent procedures within the insurance sector keep evolving as fraudsters modify their strategies to stay undetected. Due to their training on historical data, the models created for this study might not have captured all of the new fraud trends. To stay up with the changing nature of fraud, constant monitoring and model update are required.

External elements: It's possible that the models created for this study don't take into consideration external elements or contextual data that could affect the likelihood of fraud. The prediction power of the models could be improved by incorporating other data sources or external variables relating to monetary conditions, legislative changes, or industry-specific trends. The study on fraud detection in vehicle insurance claims using different models and methodologies opens the way for additional investigation and advancements in a number of areas. Future potential areas of focus include:

Real-time Monitoring: Creating systems for detecting fraud in real-time that can examine insurance claims as they happen and quickly spot probable fraud. Proactive fraud identification and prevention can be made easier with the integration of data streams and the use of streaming analytics tools.

Continuous Model Monitoring and Updating: Implementing a framework for continual model monitoring and updating will help the system react to changing fraud trends and maintain its efficacy. Up-to-date data can be used for model validation and retraining on a regular basis to maintain the model's correctness and robustness.

Data Sharing in Collaboration: Encourage data exchange and collaboration among insurance companies to create datasets that are more extensive and diversified. By capturing a variety of fraud behaviors from the industry as a whole, this can aid in the development of more reliable fraud detection algorithms.

Reference

1. Natalia Markovskaia (Published on Jul. 09, 2020) Detecting Insurance Fraud with Machine Learning Detecting Insurance Fraud with Machine Learning - Plug and Play Tech Center"
2. Daniel Aksman, Zackary Fitzgibbon , Alec Kneedler , Arseniy Pozdniakov (M ay 2020)) Fraud Detection: Using Artificial Intelligence to Identify Suspicious Persons   Over the Phone (PDF) Fraud Detection Using Artificial Intelligence to Identify Suspicious Persons Over the Phone (researchgate.net)
3. Sharmila Subudhi,Suvasini Panigrahi (july 2018) Detection of Automobile Insurance Fraud Using Feature Selection and Data Mining Techniques (PDF) Detection of Automobile Insurance Fraud Using Feature Selection and Data Mining Techniques (researchgate.net)\
4. Carol Hargreaves and Vidyut Singhania (January 2016) Analytics for Insurance Fraud Detection: An Empirical Study (PDF) Analytics for Insurance Fraud Detection: An Empirical Study (researchgate.net)
5. 5)arkovskaia, N. (2020). *Detecting Insurance Fraud with Machine Learning*. Plug and Play Tech Center. https://www.plugandplaytechcenter.com/resources/detecting-insurance-fraud-machine-learning/
6. Aksman, D., Fitzgibbon, Z., Kneedler, A., & Pozdniakov, A. (2020). *Fraud Detection: Using Artificial Intelligence to Identify Suspicious Persons Over the Phone*. ResearchGate. https://www.researchgate.net/publication/341960155_Fraud_Detection_Using_Artificial_Intelligence_to_Identify_Suspicious_Persons_Over_the_Phone
7. Subudhi, S., & Panigrahi, S. (2018). *Detection of Automobile Insurance Fraud Using Feature Selection and Data Mining Techniques*. ResearchGate. https://www.researchgate.net/publication/326402184_Detection_of_Automobile_Insurance_Fraud_Using_Feature_Selection_and_Data_Mining_Techniques
8. Hargreaves, C., & Singhania, V. (2016). *Analytics for Insurance Fraud Detection: An Empirical Study*. ResearchGate. https://www.researchgate.net/publication/291972925_Analytics_for_Insurance_Fraud_Detection_An_Empirical_Study
9. FRISS. (2022). *Global Insurance Fraud Trends Report: Estimated 20% of Claims are Fraudulent*. https://www.friss.com/resources/global-insurance-fraud-trends-2022