# Insurance Claim Fraud Detection Using Machine Learning

# V Sai Nirmal Vignu, Vasanth Chelapaka, P V S R Naidu, G Pavan Kumar Varma

# Department of Information Technology ,Sagi Ramakrishnam Raju Engineering College Bhimavaram, Andhra Pradesh

## **INTRODUCTION**

Many insurance firms are currently dealing with the problem of fraudulent insurance claims, which results in significant financial losses each year. These frauds have a negative impact on society because the losses are covered by raising policyholder premiums. The typical claim investigation process, which is time consuming and tiresome and frequently results in erroneous outcomes, has also been cited as a major contributor. As a result, we present a framework for automated fraud detection based on machine learning and the weighted XGBoost algorithm in this study. The goal is to accurately identify fraud claims in a shorter amount of time. Data analysis is used to validate, sanitise, and retrieve useful data throughout the process. As a result, employing this structure, an insurance firm can preserve its credibility in the eyes of the public.

Fraudulent claims can be highly expensive for insurers. Therefore, it is important to know which claims are correct and which are not. It is not doable for insurance companies to check all claims personally since this will cost simply too much time and money. Insurance fraud detection is a challenging problem, given the variety of fraud patterns and relatively small ratio of known frauds in typical samples. While building detection models, the savings from loss prevention needs to be balanced with the cost of false alerts. Machine learning techniques allow for improving predictive accuracy, enabling loss control units to achieve higher coverage with low false positive rates.A comparison study has been performed to understand which ML algorithm suits best to the dataset. We are able to cut losses for the insurance company. Less losses equates to more earning.

Insurance fraud is described as unethical behaviour carried out by individuals in order to achieve a favourable outcome from an insurance company. There are several types of insurance available, such as health insurance, agricultural insurance, business insurance, life insurance, auto insurance, and so on. Only Auto Insurance Fraud Detection is included in this research.

This form of insurance fraud is divided into two categories.

1. Hard Insurance Fraud: This is a sort of fraud in which an accident does not occur, but the individuals intend to file an insurance claim anyhow.
2. Opportunistic Insurance Fraud: This sort of fraud is also known as soft insurance fraud. If an accident or any other type of car collision occurs, the insured individual may file a claim for insurance by raising the severity of the vehicle's damage. When compared to hard insurance fraud, soft insurance fraud is more common.

Types of insurance fraud claims

* 1. Inflating insurance claims.
  2. Staging accidents.
  3. Submitting claim forms for injuries or damage that never occurred.
  4. False reports of stolen vehicles

The following are some of the possible satiations in which fraud can occur:

1. Situations such as drinking and driving, engaging in high-risk activities, and driving while using a cell phone, to name a few. If an accident occurs during these circumstances, insurance companies are not obligated to provide coverage; but, if an accident occurs during these circumstances, the individual will file a claim for coverage.
2. No accident has occurred; yet, an individual will fabricate a lovely narrative about the event and tell it to the insurance company.
3. An accident has happened, but the amount of damage to the car is little; yet, an insured person will file an insurance claim claiming a large amount of loss.

And the major problem in this is imbalanced data as fraud claims are less when compare to non-fraudulent claims, Imbalance class problems are common in many industries. Many a times, we are interested in a minority class against another much bigger class or classes. For instance, classification of other types of frauds, classification of defective goods, classification of at-risk teenagers, identifying high potential employees, identifying people of interest such as terrorist, just to name a few.

## **About the Dataset**

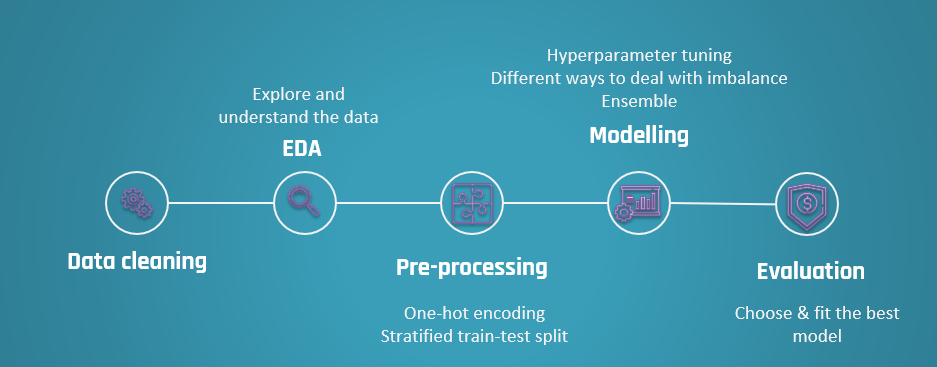
Data source: [https://www.kaggle.com/roshansharma/insurance-claim](https://www.kaggle.com/roshansharma/insurance-claim" \t "_blank)

The inspiration for this project was to perform classification on imbalance class data sets, in particular fraud. Fraud datasets are very hard to come by and often unlabelled due to its sensitive nature.

The current data set was labelled with n=1000 samples. Unlike many other data sets, this one was less popular with only the author and one other having a notebook of it on Kaggle, making this data set one that was rather novel in nature. The data set consist of 1000 auto incidents and auto insurance claims from Ohio, Illinois and Indiana from 01 January 2015 to 01 March 2015. Before any cleaning or feature engineering, the data set has a total of 39 variables. It is not stated if this data is from multiple insurance companies or just one company. However, throughout the report, "the insurance company" will be used to refer to the origin of this data.

The obvious con of this data set is the small sample size. However, there are still many companies who do not have big data sets. The ability to work with what is available is crucial for any company looking to transition into leveraging data science. In the 2017 MIT tech review, EmTech presentation, Professor Andrew Ng penned a cyclical diagram on the white board and explained that many companies start off with some small data and develop a product which have users, which in turn leads to generation of more products. In similar vein, companies may start off with a small data set and build towards a bigger data set as time goes by. Compared to a company that waits for the day when it has a huge data set, the company that started with a small data set and worked on it will more likely succeed earlier in its data science journey and reap its rewards.

## **METHODOLGY**



This project's purpose is to develop a model that can detect motor insurance fraud. The difficulty with machine learning fraud detection is that scams are significantly less prevalent than legitimate insurance claims. Unbalanced class categorization is the term for this sort of issue.

The data was cleaned and exploratory data analysis was performed before modelling. The data was then pre-processed in preparation for modelling. The models were tested after modelling, and the best-fitting model was chosen using the F1 score and the ROC AUC score. The final fitted model's performance was reviewed in further depth, and its best aspects were highlighted. The experiment ended with a reminder of the value of the research and what had been accomplished, as well as some limits.

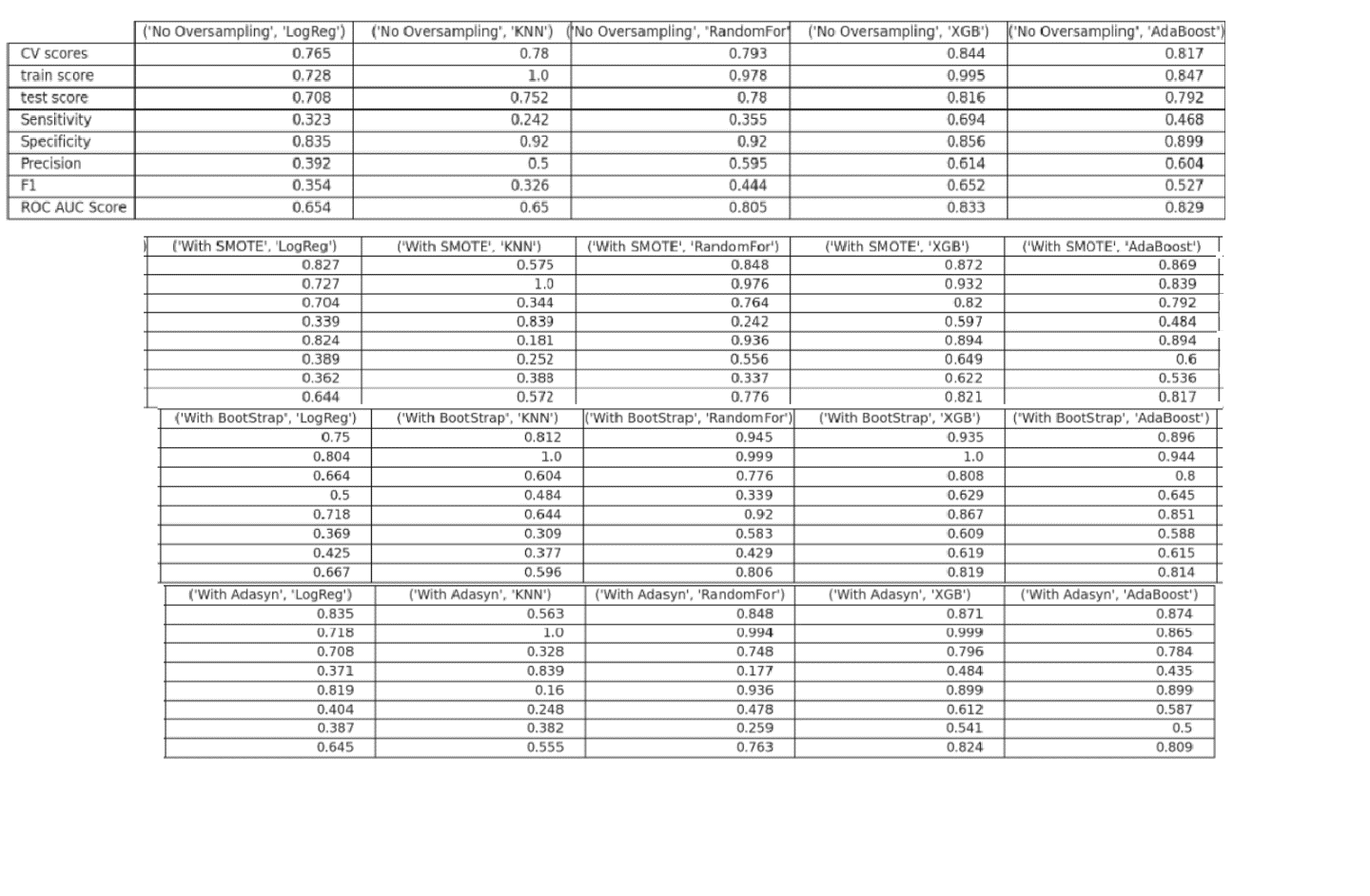
Several models like Logistic Regression, KNN, Random Forest, XgBoost, AdaBoost were put to the test using various strategies like Weighting using Hyper parameter tuning, Oversampling using SMOTE, ADASYN, Bootstrapping for dealing with imbalance datasets

As our dataset is imbalance, accuracy is not a good measure of success. A high accuracy can be achieved by a poor model that only selects the majority class, hence, not detecting and measuring the accuracy of classifying the class of interest. In fact, predicting only the majority class will give an accuracy of 75%, specificity of 100% but a sensitivity of 0%.

If we make a naive prediction that all claims are frauds, so that no frauds escape our watch, we will have a score as shown below:

* Sensitivity: 1.0
* Specificity: 0.0
* Precision: 0.248
* F1 score: 0.397
* ROC AUC Score: 0.50

As identifying as many frauds as possible is the goal, the F1 score of 0.397 was used as a baseline. However, investigations into frauds can be time consuming and expensive and may even affect customer experience. Thus, ROC AUC score will also be used to measure how well we distinguish between Fraud and legit claims. The baseline ROC AUC score is 0.50. I am to have a ROC AUC of at least 0.70.



* Comparing across models that were fitted on, no oversampling dataset, on SMOTE dataset, on ADASYN dataset and on bootstrapped dataset, I picked the logistic regression, KNN and CART model.
* The best logistic regression model was the one trained on the ADASYN dataset. It had a F1 score of 0.403 and an AUC of 0.646
* The best KNN model was the one trained on the ADASYN dataset. It had a F1 score of 0.42 and an AUC of 0.60.
* CARTs performed very well on this data set. However, since CARTs' predictions are all highly correlated, the best of them were selected. The best CART model was the weighted XGBoost on the dataset with no oversampling. It yields a F1 score of 0.65 and an AUC of 0.83.
* Overall, the ADASYN method seem to have the best F1 scores when compared to the other oversampling methods.

The final fitted model is the weighted XGBoost on the dataset with no oversampling. The best estimators of the model are as follows:

* Scale\_pos\_weight: 3.054054054054054,
* Reg\_lambda (L2 regularization weight): 0.1,
* Reg\_alpha (L1 regularization weight): 0.05,
* N\_estimators: 650,
* Max\_depth: 4,
* Gamma: 1,
* Eta: 0.3

The model had a training accuracy score of 0.992 and a test accuracy of 0.816. The high accuracy score hint of a low bias (it is only a hint as accuracy is not a good measure of bias in imbalance class problems). An accuracy score difference of 0.176 between train and test is relatively small. Thus, this model can be said to have low variance and is generalizable on unseen data.

**Conclusion and Limitations**

Fraud accounted for between 15 percent and 17 percent of total claims payments for auto insurance bodily injury in 2012, according to an Insurance Research Council (IRC) study. The study estimated that between \$5.6 billion and \$7.7 billion was fraudulently added to paid claims for auto insurance bodily injury payments in 2012, compared with a range of \$4.3 billion to \$5.8 billion in 2002.

This project has built a model that can detect auto insurance fraud. In doing so, the model can reduces loses for insurance companies. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims.

Five different classifiers were used in this project: logistic regression, K-nearest neighbours, Random forest, XGBoost, AdaBoost. Four different ways of handling imbalance classes were tested out with these five classifiers: model with class weighting and hyperparameter tuning, oversampling with SMOTE, oversampling with ADASYN and oversampling with bootstrapping.

The best and final fitted model was a weighted XGBoost what yelled a F1 score of 0.65 and a ROC AUC of 0.83. The model performed far better than the baseline F1 score of 0.397 and ROC AUC target of 0.5. The model's F1 score and ROC AUC scores were the highest amongst the other models. In conclusion, the model was able to correctly distinguish between fraud claims and legit claims with high accuracy.

The study is not without limitations. Firstly, this study is restricted by its small sample size. Statistical models are more stable when data sets are larger. It also generalizes better as it takes a bigger proportion of the actual population. Furthermore, the data only capture incident claims of 3 states from 01 January 2015 to 01 March 2015. This means that we do not know the proportion of auto insurance policy holder who had no incidents compared to those who had incidents. We are also restricted to incidents between 2 months which may not be an accurate picture of the year. This is important as certain time of the year may correlate to higher incident rates such as St. Patrick’s Day or other holidays. Future studies may investigate acquiring a larger data set with multiple years. However, due to the sensitive nature of fraud and confidential information tagged to such data, this may remain a challenge.

**REFERENCES**

1. ” Insurance Claim Analysis Using Machine Learning Algorithms” – Rama Devi Burri et all, IJITEE 2019
2. ” A Survey Paper on Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques” – Pinak Patel et all, IRJET 2019
3. ”Auto Insurance Fraud Detection”- Kavya Priya , Anusha Y , Amrutha T , Harsha R , Harshitha, IJARCCE 2020.
4. “For Real? Auto Insurance Fraud Claim Detection with Machine Learning.”- [IceAsher Chew](https://iceasherchew.medium.com/?source=post_page-----efcf957b38f3--------------------------------), Towards DataScience 2020.