iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52875

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Fraud Detection and Analysis for Insurance Claim Using Machine Learning

Apurva[1], Vaishnavi Patil[2], Pooja More[3], Kshtija Sakhare[4]
Dept. of Computing Science Eng., Zeal college of Eng. Pune, India

Abstract: ***Insurance fraud is an illegal conduct that is done on purpose in order to profit financially. This iscurrently the most serious issue that numerous insurance companies throughout the world are facing. The majority of the time, one or more gaps in the investigation of false claims has been identified as the primary factor. As a result, the requirement to use computer tools to stop fraud activities increased. Providing customers with a dependable and stable environment while significantly lowering fraud claims. We demonstrated the results of our research by automating the evaluation of insurance claims using a variety of data methodologies, where the detection of erroneous claims would be done automatically using Data Analytics and Machine Learning techniques.***

Keywords: Machine Learning, Data Analytics, Fraud Detection, Insurance Company's Reputation, CustomerSatisfaction.

I. INTRODUCTION

In essence, insurance fraud is deliberate deceptionthat can done by, against, or with the intent to defraud an insurance company or agent. It is a severe and urgent problem that is a threat becausefraudulent insurance applications put a greaterfinancial strain on the society through high premium prices. Recent research suggest that there is universal agreement that traditional methods of fraud identification are highly unreliable and imprecise. These worries prompt the machine learning and data analyticscommunity to focus on this issue and seek asolution. Similar to this, our proposed work accurately distinguishes between fraudulent and non-fraudulent claims so that only fraudulent cases need to be investigated and legitimate claims can be made quickly without wasting time or resources. This project aims to suggest the most accurate andsimplest way that can be used to fight fraudulent claims. The main problem with detecting fraudulent activities is the massive number of claims that run through the companies systems. This problem can also be used as an advantage if the officials were to take into account that they hold a big enough database if they combined the database of the claims. Which can be used in order to develop better models to flag thesuspicious claims.

As we live in a very materialistic world everyone is looking out to protect something they have or own inone way or another. People are willing to pay moneyas a contingent against the unknown loss that they might face. In the U.S alone the insurance industry is valued at 1.28 trillion dollars and the U.S consumer market losses at least 80 billion to insurance fraud every year. That causes the insurance companies to increase the cost of their policies which puts them ina less competitive position against the competition. This in turn also increased the threshold of the minimal payment for a policy since they can afford todo so while everyone is raising prices. This project will look into the different methods that have been used in solving similar problems to test outthe best methods that have been used previously. Searching if examining these methods and trying to enhance and build a predictive model that could flag out the suspicious claims based on the researching andtesting out the different models and comparing these models to come up with a simple enough time- efficient and accurate model that can flag out the suspicious claims without stressing the system it runs on.

A. Problem Statement

The main purpose of this project is to come up with a model to be used to find out if a certain insurance claim made is a fraud or not. The model will be designed after testing multiple algorithms to come up with the best model thatcan detect if a claim is fraud or not. This isaimed at the insurance companies as a pitch to come up with a more tailored model for their liking to their own systems. The model should be simple enough to calculate big datasets, yet complex enough to have a decent successful percentile.

The traditional method for the detecting frauds depends on the event of heuristics around fraud indicators. Supported these, the selection onfraud created is said to occur in either of situations like, uncertain things the principles are shown if the case should be interrogated for extra examination. In numerous cases, an inventory would be prepared with scores for various indicators of the occurred fraud. The factors for deciding measures and additionally the thresholds are tested statistically and periodically recalibrated. Associate aggregationand then price of the claim would verify necessity of case to be sent for extra examination .The challenge with abovestrategies is that they deliberately believe on manual mediation which might end in the next restrictions

B. Motivation

Basically businesses ought to obtain the responses to prevent fraud from happening or if that is out of the question, to watch it before important damage is finished at intervals 407 the strategy. In most of the companies, fraud is understood entirely once it happens. Measures arethen enforced to forestall it from happening over again. At intervals the given time that they can't resist at different time intervals, but Fraud detection is that the most effective suited issue forremoving it from the atmosphere and preventing from continuance. Previously frauds relatedinsurance detected or analyzed by manually usingthis method is not correct way to detect frauds related insurance, therefore increasing accuracy, precision, recall we proposed this project using machine learning algorithm. Peopleare making fool insurance companies by claimingwrong insurance claim so detecting this we did this project.

II. METHODOLOGIES

A. Methodologies of Problem solving

1) *Obtaining Dataset:* We collected dataset from different sources like haggle , Google. Also we created dummy dataset for analysis and detection of frauds. We used three typesof dataset raw dataset which is dummy dataset, processed insurance claim dataset and Integrated dataset.
2) *Loading Dataset:* For reading dataset we used panda's python library. We loaded dataset in csv File format. For visualizationthe features of dataset used python libraries which arematplotlib, seaborn plotly etc. We loaded dataset in jupyter notebook
3) *Preprocessing Dataset:* Preprocessing the dataset means data wrangling. In preprocessing dataset we reduced reduced redundancy of data. Under preprocessing step we cleaned the dataset, treated the null values, we did normalization of datta removed the outliers. Encoded variables categorical variables into continuous variables
4) *Comparative Analysis:* In comparative analysis we divided dataset intotraining and testing sets .we Visualized heat map of training and testing dataset. We compared model with Respect to accuracy parameter using different classifier .Finally we generated the ROC curve using the different classifier.
5) *Training and Validation:* For Training and Testing we splited dataset usingsklearn library into 80-20. For training we feuded or trained model with 100% accuracy for testing we got 65% accuracy. For visualizing training and testing dataset we used heat map.
6) *Calculating Accuracy:* Accuracy is defined as percentage of correct prediction for test data .Total accuracy is calculated using confusion matrix. Precision, recall and f1-Score is used for calculating the total accuracy. We got 65% accuracy of our model Total accuracy calculated random forest classifier.

B. Algorithm Used

Now, suitable model needs to Be constructed. A model is created by studying, practicing, and then using it.The model will be put to use and produce fraud detection. We went through these 6 processes

1) To create and train the model
2) Contextualize machine learning in your organization
3) Explore the data and choose the type of algorithm
4) Prepare and clean the dataset
5) Split the prepared dataset and perform crossvalidation
6) Perform machine learning optimization

The generic flow of machine learning data modelis presented below: An integrated researchapproach will be applied for this project. To explain the findings and conclusions of the paper and the various results, the project will employ some explanatory research methods in addition toexperimental research methods, some qualitativeresearch methods, and some quantitative research methods. We will start by determining what is crucial for managing the data in accordance withthe business that may use the model or the solutions. In this situation, an insurance provider will probably prioritize the financial aspects of each claim while giving no consideration to personal information when developing the model.43 The report will detail the available data, the many qualities, and how each attribute relates to determining whether or not this claim is fraudulent. What are the different forms of data that are available, and can the existing data be improved or changed without affecting the outcome of the end goal, which is identifying dubious claims? To do this, data cleaning is necessary. Some values or characteristics mayneed to be removed, or new values may need to be created by merging existing ones and using data integration techniques.

The coaching information consists of a group of coaching samples. Just in the case of supervised learning, every instance is often a base that incorporates the Associate in nursinginput object that's considered the vector, and also the output features a worth that acts as an indicator to run the model. A supervised learning rule initially accomplish

C. Random Forest

In essence, Random Forest is used for both classification and regression issues. It is an ensemble classification method and a supervised machine learning classifier. The more trees there are, the more precise the outcome would be. The RF machine learning Over-fitting is one of the decision tree algorithm's main problems. The decision tree appears to have remembered the data.Random Forest is utilized to prevent this and is an illustration of ensemble learning in action. The useof several repetitions of one or more algorithms is referred to as "ensemble learning." A "random forest" is a collection of decision trees.

D. XBoost

XGBoost is an open-source software library that implements optimized distributed gradient boosting machine learning algorithms under the Gradient Boosting framework. XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributedgradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting andis the leading machine learning library for regression,classification, and ranking problems.

It's vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and gradient boosting .Supervised machine learning uses algorithms to train a model to find patterns in adataset with labels and features and then uses the trained model to predict the labels on a new dataset'sfeatures.

III. SYSTEM ARCHITECTURE

Machine learning model is built with different algorithms that is trained by information and data set provided which predict new classification as "fraud" or "not" These algorithms implemented for building model that is trained using historical data and that predict unseen data with most matching features and then model is tested and validated to evaluate its performance. At first, we have taken the insurance claim dataset (raw) then, preprocessing activities are carried out to improve the quality of dataset .In preprocessing dataset we reduced reduce redundancy of data. Under preprocessing step we cleaned the dataset, the null values, we did normalization of data removed the outliers. Encoded variables dividing the data we used sclera library. In training phase development of modelusing machine learning is done. We calculated performance of model with respect to accuracy parameters .In testing phasemodel tested for unknown datasets. Result is calculated on basis of confusion matrix, precision recall and f1 score. And selected better model. The final model will detect the frauds caused.

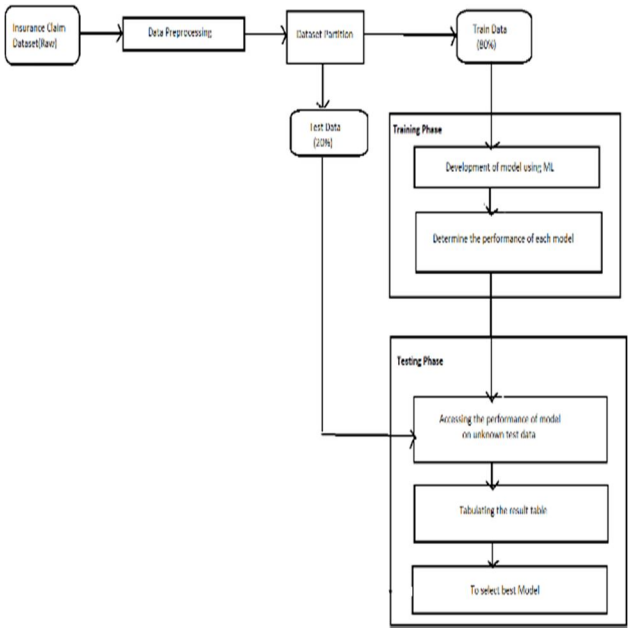
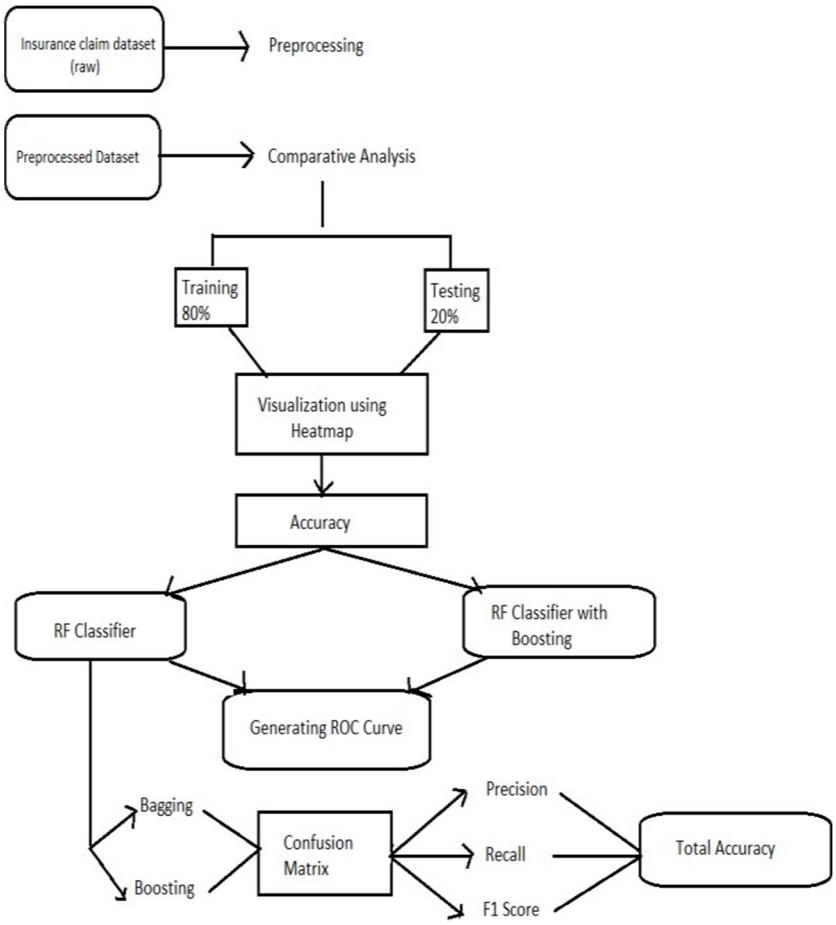


Fig (1).System Architecture

A. Flow Diagram



The categorization report includes a number of metrics that are crucial for assessing any model. Accuracy, precision, recall, and F1 are the included measures.
Accuracy: it is the ratio of correct predictions against total observations.
Precision: is the proportion of correctly made positive predictions to all positively observed data.
Recall: The ratio of correctly predicted positive observations to all of the observations in a class.
F1: is the average of the recall and precision scores.

Where,
TP=True Positive FP=False Positive TN=True NegativeFN=False Negative
The sickie-learn packages come within default parameters. The default parameters have resulted In undesired results, and so the tuning for the models have been done through tuningthe hyper parameter.

IV. RESULT

Heat map is data visualization technique that shows magnitude of phenomenon as color in two dimensions. The variation in color maybe by hue or intensity giving visual effects. Using a heat map visualize a confusion matrix, time-series moments, correlation matrix and SHAP interaction values. It highlight important relationships in data. • Training heat map the following heat map shows correlations between training dataset features. We haveplotted heat map using seaborne library.Seaborn is a datavisualization library based on matplotlib. It provides ahigh-level interface for drawing attractive graphs.Seaborn package allows the creation of a noted heat map which can be tweaked usingmatplotlib tools as per the requirements

V. FUTURE WORK

The machine learning models applied on thesedatasets were able to determine most of the fallacious cases with low false positive rate which suggests with cheap exactness. Certain knowledge sets had severe challenges around data quality, resulting in comparatively poor levels ofprediction. Given inherent characteristics of varied datasets, it would not be sensible to outlineoptimum algorithmic techniques or use feature engineering process for a lot of higher performance. The models would then be used for specific business context and user priorities. This helps loss management units to specialize in areplacement of fraud situations and thenguaranteeing that models square measure adaptingto spot them. However, it might be cheap to counselthat supported the model performancc on back- tcsting and talcnt to spot ncw frauds, thc sct of modcls work thc chcap suitc to usc within thc spacc of thc insurance claims fraud detection.

In order to compare the effectiveness of machine learning and deep learning methodologies, future research should focus on attempting to use an advanced or recently obtained dataset. Additionally, it is advised to utilize a different dataset in light of the fact that the one being used is unbalanced. Additionalevaluation should be done to determine feature relevance across various datasets that may ormay not have similar characteristics in order to develop a much more universal method to featureselection and focus. Because this research hasbeen done by using all features in the future, we will do the feature selection to measure the variance between the total and selected features.

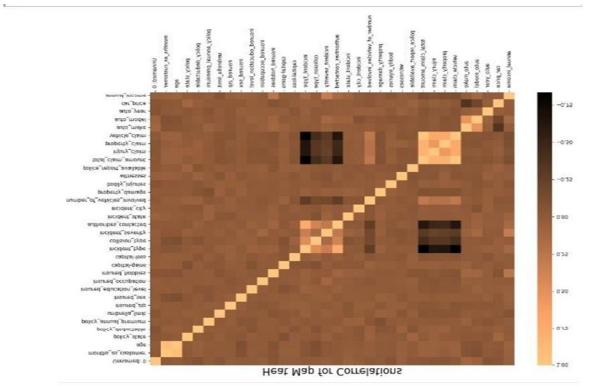
A. Applications

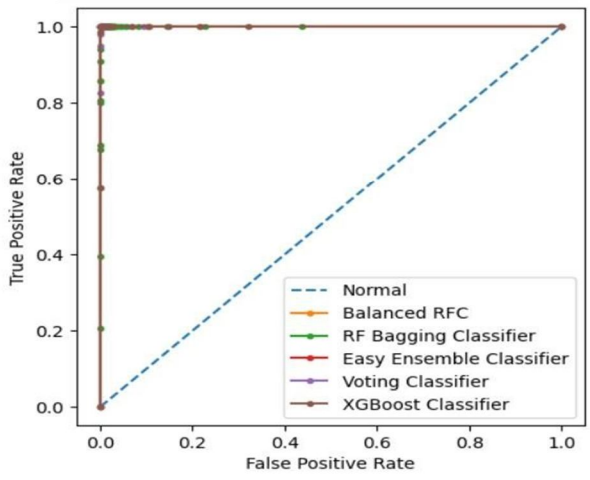
1) This project will help to determine frauds in insurance claim in faster way.

2) Since this system is loosely coupled and developed cohesively, it can be leveraged for a wide range of applications. The system can be Used in publically. Furthermore, if effectively scaled up, this system might be used in larger insurance company all over the world.

3) Early detection of frauds becomes possible. Early detection will help to reduce the loss of money.

4) The system can be used in publically. Furthermore, if effectively scaled up, this system Might be used in larger insurance company allover the world.





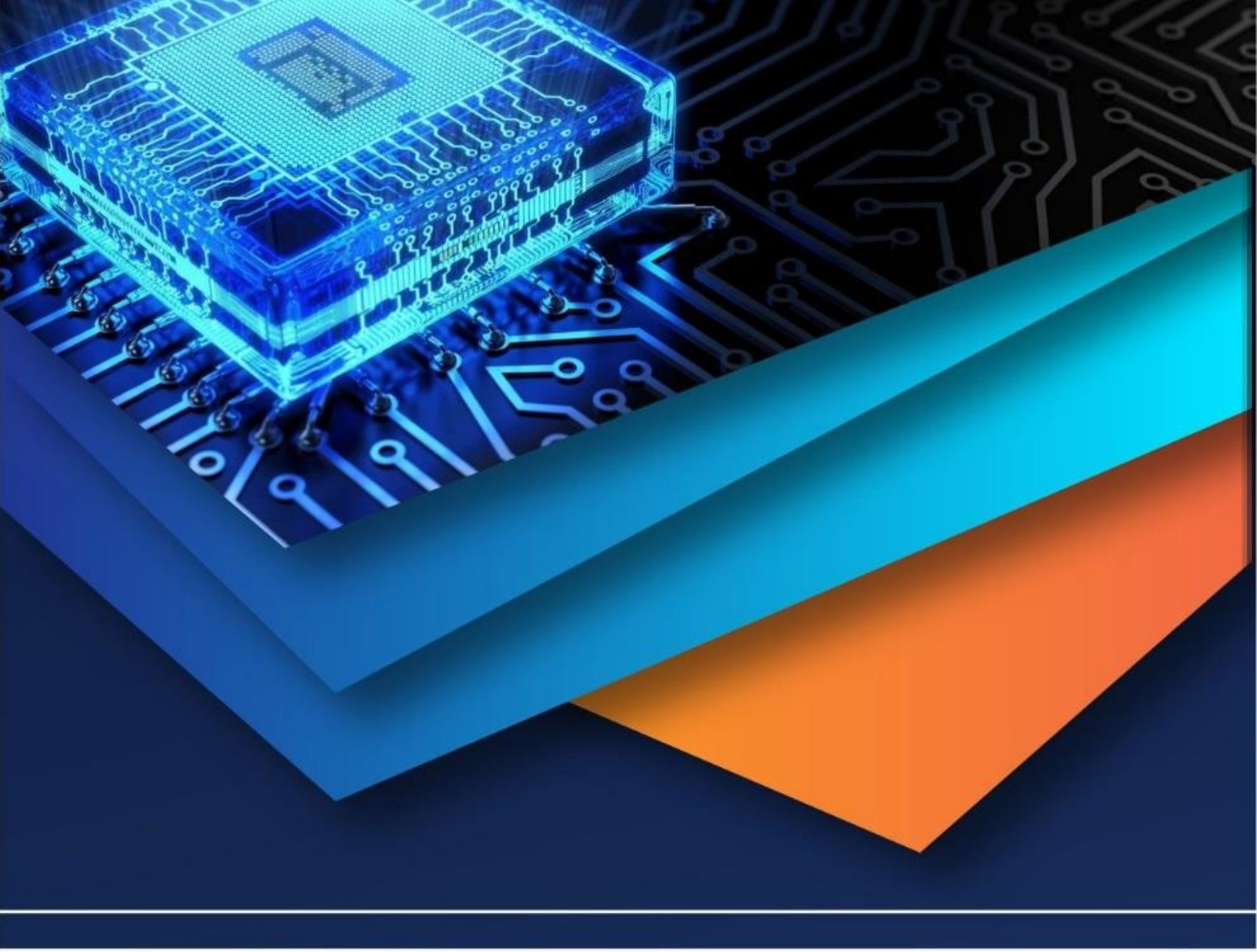
As the different countries around the world evolve into a more economical-based one, stimulating their economy is the goal. To fight these fraudsters and money launderers was quite a complex task before the era of machine learning but thanks to machine learning and AI we are able to fight these kinds of attacks. The proposed solution can be used in insurance companies to find out if a certain insurance claim made is a fraud or not. The model was designed after testing multiple algorithmsto come up with the best model that will detect if a claim is fraudulent or not. This is aimed at the insurance companies as a pitch to come up with a more tailored model for their liking to their own systems. The model should be simple enough to calculate big datasets, yet complex enough to have a decent successful percentile.

Under this project, we choose the sample of more than 1000 data and the data divided into training and testing data. We can see that, compared with the algorithms XGboost and random forest algorithms; have better performance than KNN. We just brought out the feature of machine learning algorithms. We worked with more algorithms and finally calculate which provide more accuracy, precision, and recall.

REFERENCES

[1] K. Usage Priya and S. Pushpa, "A Survey on Fraud Analytics Using Predictive Model in InsuranceClaims," Int. J. Pure Appl. Math., vol. 114, no. 7, pp.755–767, 2017.
[2] E. B. Belhadji, G. Dionne, and F. Tarkhani, "AModel for the Detection of Insurance Fraud," GenevaPap. Risk Insur. Issues Pract., vol. 25, no. 4, pp. 517–538, 2000, doi: 10.1111/1468-0440.00080.
[3] "Predictive Analysis for Fraud Detection." https://www.wipro.com/analytics/comparativeanalys is-of-machine-learning-techniques-for-
[4] %0Adetectin/.
[5] F. C. Li, P. K. Wang, and G. E. Wang, "Comparison of the primitive classifiers with extremelearning machine in credit scoring," IEEM 2009 - IEEE Int. Conf. Ind. Eng. Eng. Manag., vol.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY