ISSN: 2456-1134 www.isjcresm.com
Vol-08 Issue-01 March 2023

# FRAUD DETECTION AND ANALYSIS FOR INSURANCE CLAIM USING MACHINE LEARNING

**[1] B.Raghupathi, [2] Dr.P.Satish Reddy,[3] Asif Ahmed Algur, [4] KARNE ANUSHKA**

[1,2,3] Assistant Professors,Department of Computer Science and Engineering, Kasireddy Narayanreddy College Of Engineering And Research, Abdullapur (V), Abdullapurmet(M), Rangareddy (D), Hyderabad - 501 505

[4]student,Department of Computer Science and Engineering, Kasireddy Narayanreddy College Of Engineering And Research, Abdullapur (V), Abdullapurmet(M), Rangareddy (D), Hyderabad - 501 505

**ABSTRACT**

Insurance Company working as commercial enterprise from last few years have been experiencing fraud cases for all type of claims. Amount claimed by fraudulent is significantly huge that may causes serious problems, hence along with government, different organization also working to detect and reduce such activities. Such frauds occurred in all areas of insurance claim with high severity such as insurance claimed towards auto sector is fraud that widely claimed and prominent type, which can be done by fake accident claim. So, we aim to develop a project that work on insurance claim data set to detect fraud and fake claims amount. The project Implement machine learning algorithms to build model to label and classify claim. Also, to study comparative study of all machine learning algorithms used for classification using confusion matrix in term soft accuracy, precision, recall etc. For fraudulent transaction validation, machine learning model is built using PySpark Python Library.

## I. INTRODUCTION

Insurance fraud is a significant and growing problem worldwide, leading to substantial financial losses for both insurance companies and policyholders. It is estimated that fraudulent claims account for billions of dollars in losses annually, placing a heavy burden on the insurance industry and increasing premiums for honest customers【1】. Traditional methods of fraud detection, which often rely on manual

ISSN: 2456-1134 www.isjcresm.com
Vol-08 Issue-01 March 2023

investigations and rule-based systems, are increasingly inadequate in dealing with the sophisticated techniques employed by fraudsters today【2】. Machine learning has emerged as a powerful tool in the fight against insurance fraud, offering the ability to analyze large volumes of data and identify patterns that may indicate fraudulent activity【3】. Unlike traditional methods, machine learning models can learn from historical data, detecting complex and subtle patterns that would be difficult for human investigators to identify【4】. By continuously learning from new data, these models can adapt to evolving fraud tactics, making them more effective in identifying both known and emerging fraud schemes【5】.

The application of machine learning in fraud detection typically involves the use of supervised learning techniques, where models are trained on labeled datasets containing both fraudulent and legitimate claims【6】. These models can then classify new claims as either suspicious or legitimate, based on the features and patterns they have learned【7】. In addition to supervised learning, unsupervised learning methods, such as clustering and anomaly detection, are also used to identify outliers or unusual patterns in data that may indicate fraud【8】.

This project focuses on the development of a machine learning-based system for fraud detection and analysis in insurance claims. The system will leverage various machine learning algorithms, including decision trees, random forests, and neural networks, to build a robust model capable of accurately identifying fraudulent claims【9】. The proposed system will not only detect fraud but also provide insights into the patterns and characteristics of fraudulent claims, enabling insurance companies to refine their fraud prevention strategies and reduce financial losses【10】.

## II.EXISTING SYSTEM

Machine learning is usually abbreviated as metric capacity unit. The study of machine learning includes computers with the implicit capability to be trained whereas not being expressly programmed. This capacity unit focuses on the expansion of pc programs that has enough capability to alter, that square measure once unprotected to the new

ISSN: 2456-1134 www.isjcresm.com
Vol-08 Issue-01 March 2023

information. Metric capacity unit algorithms square measure generally classified into 3 main divisions that square measure supervised learning, unattended learning and reinforcement learning. Data processing a neighborhood of machine learning has advanced considerably within the current years.

Data mining focuses at analysing the whole data obtained. Furthermore data processing makes an attempt to seek out the realistic patterns in it. On the contrary, within the different of getting the knowledge for world understanding is within the processing applications like machine learning, it uses the knowledge to locate patterns in information and improvise the program actions thereby. Mainly within the supervised machine learning is that the objective of deducing which means from label on the information used for the coaching.

The coaching information consists of a group of coaching samples. Just in case of supervised learning, every instance are often a base which incorporates Associate in Nursing input 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 accomplishes a groundwork task from the sample information then tries to construct a short lived perform, therefore it will plot new input vectors.

The supervised learning algorithms square measure conspicuously employed in large choice of application areas. Associate in Nursing best setting altogether the chance assist the rule to accurately mark the class labels for close instances and therefore a similar aspires supervised learning rule to chop back from the knowledge to the enclosed objects in terribly good manner.

**Disadvantages**

- The system is not implemented Convex-NMF based Supervised Spammer Detection with Social Interaction (CNMFSD).

- The system is not implemented any ml classifier for test and train the datasets.

## III.PROPOSED SYSTEM

The influence of the feature engineering, feature choice parameter modification area unit explored with an aim of achieving superior prophetic

ISSN: 2456-1134 www.isjcresm.com
Vol-08 Issue-01 March 2023

performance with superior accuracy. The assorted machine learning techniques area unit utilized in the development of accuracy of detection in unbalanced samples. As a system, the info are divided into 3 completely different segments. These area unit loosely coaching, testing and validation.

The algorithmic program is trained on partial set of knowledge and parameters. These area unit later changed on a validation set. This may be studied for evaluation and performance on the particular testing dataset. The high acting models area unit formerly tested with numerous random splits of knowledge. This helps to confirm the consistency in results the approach discussed above comprises of three layers.

• Data Pre-processing step: In this step, the data is ready in order that are often employed in code with efficiency. Extraction of the dependent and freelance variables from the given dataset. Then the dataset is split as coaching and checking victimisation train test split module from sklearn library. Feature scaling is completed therefore on get correct results of predictions

• Fitting Logistic Regression to the Training set: LogisticRegression category of the sklearn library is employed. Classifier object is made and accustomed work the model to the supply regression Predicting the test result: The model is well trained on the training set, the result is predicted by using test set data.

• Test accuracy of the result: Confusion matrix is employed to judge the check accuracy. In this model of fraud detection, the prediction is completed therefore on check if deceitful dealings is claimed as deceitful and the other way around.

• Visualizing the test set result: Adjust the model fitting parameters, and repeat tests. Adjust the model fitting parameters, and repeat tests. Adjust the options or machine learning algorithmic program and repeat tests.

**Advantages**

• Different models are tested on the dataset once it is obtained and cleaned.

• On the basis of the initial model performance, different features of the model are engineered and tested again.

• Once all the options area unit designed, the model is made and run victimisation completely different completely different values and victimisation different iteration procedures.

ISSN: 2456-1134 www.isjcresm.com
Vol-08 Issue-01 March 2023

• A predictive model is created that predicts if an insurance claim is fraudulent or not.

• Binary Classification task takes place which gives answer between YES or NO. This report deals with classification algorithm to detect fraudulent transaction.

## IV. MODULES

### Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Train & Test Data Sets, View Trained Accuracy in Bar Chart, View Trained Accuracy Results, View Type, Find Type Ratio, Download Predicted Datasets, View Type Ratio Results, View All Remote Users.

### View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

### Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like register and login, predict type, view your profile.

## V.CONCLUSION

The implementation of a machine learning-based system for fraud detection and analysis in insurance claims represents a significant advancement in the efforts to combat insurance fraud. By leveraging the power of machine learning, the proposed system can analyze vast amounts of data and identify complex patterns that indicate fraudulent activity, providing a more efficient and accurate alternative to traditional fraud detection methods.

The system's ability to continuously learn and adapt to new fraud tactics ensures that it remains effective in an ever-evolving threat landscape. Additionally, the insights generated by the system can help insurance companies better understand the nature of fraudulent claims, enabling them to develop more targeted and proactive fraud prevention strategies. As the insurance industry

continues to face the challenges posed by fraud, the adoption of machine learning technologies will be crucial in mitigating financial losses and protecting the integrity of the insurance system.

## VI.REFERENCES

1. Viaene, S., & Dedene, G. (2004). Insurance fraud: Issues and challenges. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 29(2), 313-333.
2. Derrig, R. A. (2002). Insurance fraud. *Journal of Risk and Insurance*, 69(3), 271-287.
3. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.
4. Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559-569.
5. Nian, R., Zhang, Y., Tayal, A., Coleman, T., & Li, J. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *Journal of Financial Crime*, 23(4), 950-961.
6. Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22, 85-126.
7. Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004). Survey of fraud detection techniques. *IEEE International Conference on Networking, Sensing and Control*, 2, 749-754.
8. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60, 19-31.
9. Guo, W., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach. *ACM SIGKDD Explorations Newsletter*, 6(1), 30-39.
10. Smith, G., & Wheeler, P. (2020). Machine learning and AI in the insurance industry. *Journal of Insurance Regulation*, 39(6), 1-19.