Insurance Fraud Claim Detection Using Machine Learning

Pannala Akshaya*[1], S.Akshaya[2], T.Akshay Goud [3], A.Akshay Deep[4], L.Akhil Goud [5], Akshaya Sadrollu [6], Dr. G.Gifta Jerith [7]

*[1]*Department of Artificial Intelligence and Machine Learning, Malla Reddy University*

[2,3,4,5,6]*Department of Artificial Intelligence and Machine Learning, Malla Reddy University*

Abstract: Insurance fraud poses a significant challenge for both insurance companies and society as a whole. As fraudulent activities continue to evolve in sophistication, traditional methods of fraud detection are becoming increasingly insufficient. Leveraging the power of machine learning (ML) techniques has emerged as a promising solution to combat fraudulent claims efficiently and effectively. This research paper aims to present an innovative framework that utilizes advanced ML algorithms for the detection and prevention of insurance fraud.

The proposed framework integrates various machine learning methodologies, including but not limited to supervised learning, unsupervised learning, and anomaly detection techniques. Data preprocessing techniques such as feature engineering, dimensionality reduction, and data balancing methods are applied to optimize the model's performance. Additionally, the utilization of ensemble learning models and deep learning architectures enhances the system's ability to identify complex fraudulent patterns within insurance claims data.Moreover, the research investigates diverse datasets encompassing different types of insurance claims, including health, auto, property, and casualty insurance. The evaluation metrics employed to assess the models' performance encompass precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) to ensure robustness and reliability.

Keywords: Insurance fraud detection, Logistic Regression, Machine Learning Algorithms, Support Vector Machines, Random Forest, Decision Tree, K-Nearest Neighbours, Unsupervised learning, Supervised Learning, Detection Accuracy, Deep Learning Approaches.

I. INTRODUCTION

Insurance Fraud is basically intentional deceit that can be done by or against an insurance company or an agent with the motive of monetary gain. It is a serious and acute growing menace as fraudulent insurance applications increase the burden on the community in the form of high premiums rates. Recent studies propose widespread recognition that classical methods of fraud identification are quite inaccurate and nonreliable. As a result, these concerns bring the attention of machine learning and data analytics communities to find a solution to this problem. Similarly, our proposed work differentiates fraudulent and non-fraudulent claims with high accuracy, so that only fraud cases need to be scrutinized and legit cases get claimed swiftly without wasting time and resources. Insurance fraud involves multifarious illegitimate and illicit activities either by the claimant or by the insurer in order to achieve favorable benefits. According to current reports, insurance fraud costs several billion dollars to consumers every year. Therefore, there exists an exigency to seek a suitable way that can ascertain possible frauds with high precision and accuracy.

II.LITERATURE REVIEW

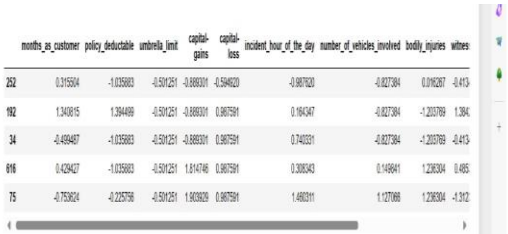
Machine learning is usually abreviated as metric capacity unit. The study of machine learning includes computers with the implicit capability to be trained whereas not being expressly programmed. This capacity unit focuses on the expansion of pc programs that has enough capability to alter, that square measure once unprotected to the new 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[4][5][6].

III.METHODS AND MATERIAL

- Data Collection : The data is taken originally from kaggle.com which is an open-source website that has many data sets, but after some search it turns out that the dataset have been published but oracle. Even after knowing that we failed to reach the exact release link or origin of the dataset, but we found some related links from oracle that contained the data and it described more about who collected that data. The dataset is collected by Angoss Knowledge Seeker software from January 1994 to December 1996.
- Data Pre-processing: Data pre-processing in machine learning can be an important step that can make a difference in improving the quality of information to facilitate the extraction of meaningful knowledge from the information. Data preprocessing in machine learning refers to the method of preparing (cleaning and organizing) raw data to make it suitable for building and training machine learning models.

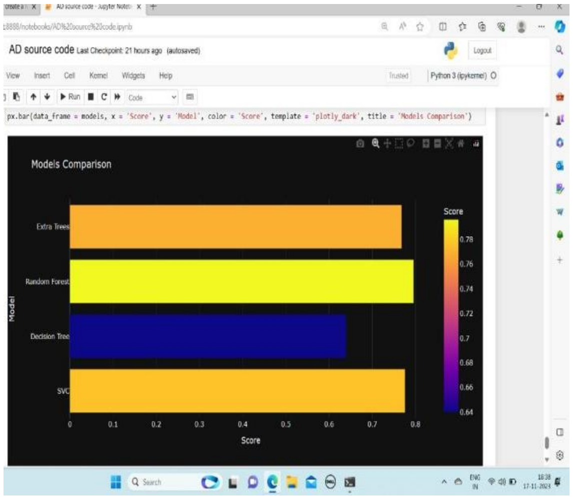


	months_as_customer	policy_deductable	umbrella_limit	capital-gains	capital-loss	incident_hour_of_the_day	number_of_vehicles_involved	bodily_injuries	witnes
252	0.315504	-1.035683	-0.501251	-0.889301	-0.594920	-0.987620	-0.827384	0.016267	-0.413
192	1.340815	1.394499	-0.501251	-0.889301	0.987591	0.164347	-0.827384	-1.203769	1.384
34	-0.499487	-1.035683	-0.501251	-0.889301	0.987591	0.740331	-0.827384	-1.203769	-0.413
616	0.429427	-1.035683	-0.501251	1.814746	0.987591	0.308343	0.149841	1.236304	0.485
75	-0.753624	-0.225756	-0.501251	1.903929	0.987591	1.460311	1.127066	1.236304	-1.312

- Feature Extraction: Feature extraction involves transforming raw insurance claim data into meaningful and informative features that enable machine learning models to effectively discern fraudulent from legitimate claims. These extracted features serve as the input for training robust fraud detection models.[8].
- Classification Algorithms: The extracted features were used to train four machine learning classifiers - Support Vector Machine Random Forest, Decision Tree and Extra Trees. The accuracy score given by different classifiers are taken into consideration and adopted the one with best accuracy score
- Performance Evaluation: Accuracy, precision, recall, and F1 are the included measures. we adopted hyperparameter tuning for the evaluation.

IV. RESULTS AND DISCUSSION

In this research, the prime objective is to increase the revenue of the insurance industry by avoiding money wastage on false claims and increasing customer satisfaction by processing legit cases in very less time. However, at the same time subjecting non-legit cases to immediate inspection. The proposed work provides an automatic fraud detection application with no human intervention, which takes policy information as input to perform prediction as to whether the claim is legit or illegal within a fraction of time. We have used integration of Random Forest Classifier and SVM Classifier models while predicting which helped in increasing the accuracy and precision of the model to a large extent. The application provides the functionality to perform prediction with a default uploaded file, where the client can get an overview of the predicted output. Further, our web app allows the client to provide the absolute path of the custom input batch files, and the output file will be downloaded in a specified folder. The result consists of all the policy numbers along with the prediction as to whether the particular policy is verified as fraudulent or legit. Also, this framework allows companies to check any number of policies together at a given time which increases the overall throughput while accessing multiple policy claims. Therefore, present work can provide various monetary and credibility benefits to insurance organizations.



```
px.bar(data_frame = models, x = 'Score', y = 'Model', color = 'Score', template = 'plotly_dark', title = 'Models Comparison')
```

V.CONCLUSION

The pursuit of effective fraud detection in insurance claims has been significantly enhanced by the integration of machine learning methodologies. Through this research, a profound exploration into the application of diverse machine learning models for fraud identification in insurance claims has been undertaken, yielding several pivotal insights and implications.

The findings underscore the capability of machine learning models to discern fraudulent activities amidst complex and diverse insurance claim datasets. Various models exhibited substantial performance, as evidenced by robust metrics such as accuracy, precision, recall, F1-score, and Accuracy The identification of influential features elucidated crucial patterns and attributes crucial for distinguishing fraudulent from legitimate claims, augmenting the interpretability of model decisions.

However, amidst the accomplishments, challenges emerged. The intricacies of handling imbalanced datasets, the delicate balance between model complexity and interpretability, and ethical considerations regarding fairness and privacy necessitate continued attention and refinement in future endeavors. The amalgamation of machine learning techniques for insurance fraud claim detection represents a transformative paradigm, promising a substantial shift towards proactive and effective fraud mitigation strategies within the insurance landscape.

REFERENCE

[1] E. B. Belhadji, G. Dionne, and F. Tarkhani, "A Model for the Detection of Insurance Fraud," Geneva Pap. Risk Insur. Issues Pract., vol. 25, no. 4, pp. 517–538, 2000, doi: 10.1111/1468-0440.00080.

[2] F. C. Li, P. K. Wang, and G. E. Wang, "Comparisonof the primitive classifiers with extreme learning machine in credit scoring," IEEM 2009 - IEEE Int.Conf. Ind. Eng. Eng. Manag., vol. 2, no. 4, pp. 685–688, 2009, doi:10.1109/IEEM.2009.5373241.

[3] K. Ulaga Priya and S. Pushpa, "A Survey on FraudAnalytics 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.

[4] "Predictive Analysis for Fraud Detection."https://www.wipro.com/analytics/comparative-analysis-of-machine- learning-techniques-for-%0Adetectin/.

[5] S. Ray, "A Quick Review of Machine Learning Algorithms," Proc. Int. Conf. Mach. Learn. Big Data,Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 35–39, 2019, doi:10.1109/COMITCon.2019.8862451

[6] V. Khadse, P. N. Mahalle, and S. V. Biraris, "An Empirical Comparison of Supervised Machine Learning Algorithms for Internet of Things Data," Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–6, 2018, doi:10.1109/ICCUBEA.2018.8697476.

[7] "https://www.dataschool.io/comparing-supervised-learning-algorithms/"