FRAUD DETECTION AND ANALYSIS FOR INSURANCE CLAIM USING MACHINE LEARNING

G. HARI PRIYA [1], K. SUPRAJA [2]

[1]Assistant Professor, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS), NH-5 Bypass Road, Gudur, Tirupati – 524101.

[2]PG Scholar, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS), NH-5 Bypass Road, Gudur, Tirupati – 524101.

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.

Index Terms—Machine Learning, K-Nearest Neighbor, Internet of Things.

I. INTRODUCTION

Insurance fraud is a claim made for getting improper money and not actual amount of money from insurance company or any other underwriter. Motor and insurance area unit two outstanding segments that have seen spurt in fraud. Frauds is classified from a supply or nature purpose of read. Sources is client, negotiator or internal with the latter two being a lot of essential from control framework purpose of reads. Frauds cowl vary of improper activities that a private might commit so as to attain the favorable outcome from an underwriter. Frauds is classified into nature wise, for example, application, inflation, identity, fabrication, contrived, evoked accidents etc. This could vary from

staging incident, misrepresenting matters as well as pertinent members and therefore reason behind finally the extent of injury occurred. Probable things might embrace packing up for a state of affairs that wasn't lined beneath the insurance. Misrepresenting the context of an event. This might embrace transferring blames to the incidents wherever the insured set is accountable, failure to require approved the security measures. Increased impact of the incident .Inflated measure of the loss occurred through the addition of not much related losses or/and attributing inflated price to the increased losses[1][2][3].

II. LITERATURE SURVEY

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 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

A Model for the Detection of Insurance Fraud

The aim of this article is to develop a model to aid insurance companies in their decision-making and to ensure that they are better equipped to fight fraud. This tool is based on the systematic use of fraud indicators. We first propose a procedure to isolate the indicators which are most significant in predicting the probability that a claim may be fraudulent. We applied the procedure to data collected in the Dionne Belhadji study (1996).

Comparison of the primitive classifiers with extreme learning machine in credit scoring

With the rapid growth in the credit industry, credit scoring classifiers are being widely used for credit admission evaluation. Effective classifiers have been regarded as a critical topic, with the related departments striving to collect huge amounts of data to avoid making the wrong decision. Finding effective classifier is important because it will help people make an objective decision instead of them having to rely merely on intuitive

experience. This study proposes two well-known classifiers, namely, K-Nearest Neighbor (KNN),

An Empirical Comparison of Supervised Machine Learning Algorithms for Internet of Things Data

Internet of Things(IoT) is one of the rapidly growing fields and has a wide range of applications such as smart cities, smart homes, connected wearable, connected health-care, and connected automobiles, etc. These IoT applications generate tremendous amounts of data which needs to be analyzed to draw useful inferences required to optimize the performance of IoT applications. The artificial intelligence(AI) and machine learning (ML) play the significant role in building the smart IoT

Quick Review of Machine Learning Algorithms

A Machine learning is predominantly an area of Artificial Intelligence which has been a key component of digitalization solutions that has caught major attention in the digital arena. In this paper author intends to do a brief review of various machine learning algorithms which are most frequently used and therefore are the most popular ones. The author intends to highlight the merits and demerits of the machine learning algorithms from their application perspective to aid in an informed decision making towards selecting the appropriate learning algorithm to meet the specific requirement of the application.

III. PROPOSED SYSTEM

The overview of our proposed system is shown in the below figure.

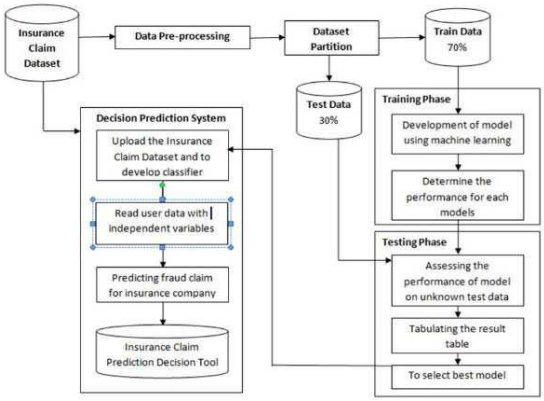

Fig. 1: System Overview

Implementation Modules

Service Provider Module

In this module, service provider login to the system using valid username and password. After login successful, he can perform the following operations like train and test dataset, view trained and tested accuracy, view trained and tested accuracies results using charts, view prediction insurance claim type, view prediction type ratio, and view remote users.

Remote User

In this module, the remote user register to the system, and login to the system valid username, and password. After login successful, he can perform view profile, predict the insurance claim type.

Implementation Algorithms

Support Vector Machine

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. An SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier

Navie Bayes

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object

Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

IV. RESULTS

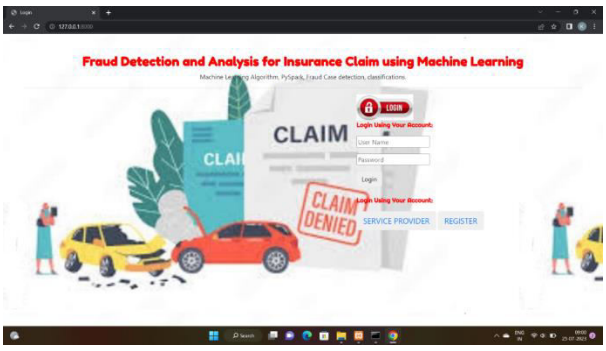

Fig. 2: Home Page

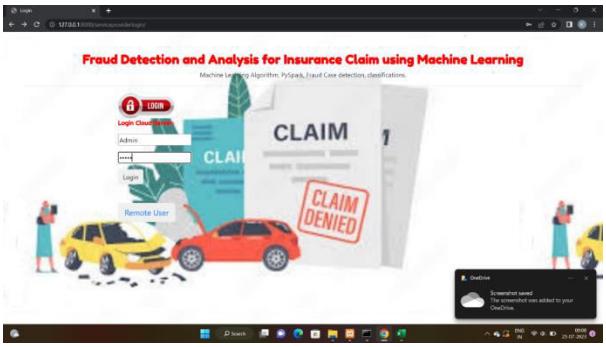

Fig. 3: Service Provider Login

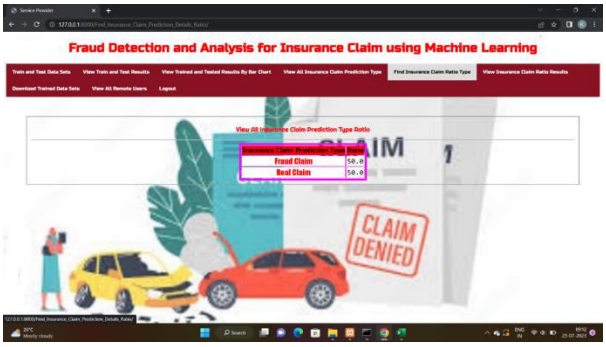

Fig. 4: Ratio of Fraud and

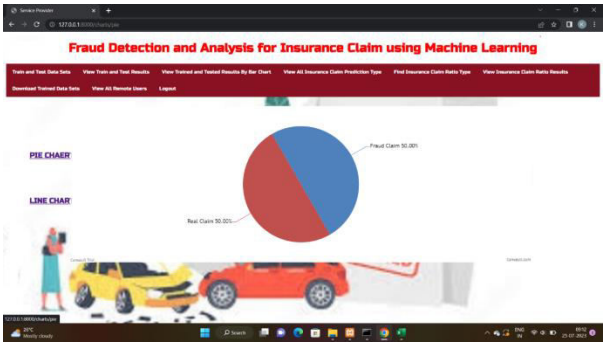

Fig. 5: Pie chart of Ratio

V. CONCLUSION

The machine learning models that square measure mentioned which square measure applied on these datasets 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 of prediction. Given inherent characteristics of varied datasets, it would not be sensible to outline optimum 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 a replacement fraud situations and then guaranteeing that models square measure adapting to spot them. However, it might be cheap to counsel that supported the model performance on back-testing and talent to spot new frauds, the set of models work the cheap suite to use within the space of the insurance claims fraud detection.

REFERENCES

[1] K. UlagaPriya and S. Pushpa, “A Survey on Fraud Analytics Using Predictive Model in Insurance Claims,” Int. J. Pure Appl. Math., vol. 114, no. 7, pp. 755–767, 2017.

[2] 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.

[3] “Predictive Analysis for Fraud Detection.” https://www.wipro.com/analytics/compara

tiveanalysis-of-machine-learning-techniques-for-%0Adetectin/.

[4] F. C. Li, P. K. Wang, and G. E. Wang, "Comparison of 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.

[5] V. Khadse, P. N. Mahalle, and S. V. Biraris, "An Empirical Comparison of Supervised MachineLearning 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.

[6] 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.

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

AUTHORS

G. Hari Priya has received B.sc in computer science (2014) and M.Sc in Computer science in Madras university in (2016),M.phil (computer science) st. peters university in 2017., M.Tech (CSE) Swetha institute of Technology and science in JNTU University (2019). She is dedicated to teaching field from the last 3 years. She has guided P.G students. At present in working as Assistant professor in Audisankara College of Engineering and Technology, Gudur, Tirupati(Dt), Andhra pradesh, India.

K. Supraja has Pursuing her MCA from Audisankara College of Engineering and Technology (Autonomous) Gudur, affiliated to JNTUA in 2023, Andhra pradesh, India.