Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 2, August 2023
32

SEQUENCE EMBEDDING HELP DETECT INSURANCE FRAUD

AyazKhan
Dept. of computer Science andEngineering
SRM Institute of Science and Technology
Chennai
Indiaak6031@srmist.edu.in

Abhinash Bhengra
Dept.of computer Science and Engineering
SRM Institute of Science and Technology
Chennai
Indiaab2983@srmist.edu.in

Dr.S.Raja Ratna
AssistantProfessor
Dept.ofcomputerScienceandEngineering
SRMInstituteofScienceandTechnology
Chennai
Indiagracelinrr@gmail.com

Abstract—Fraud is costly and costly for businesses and customers in the banking and insurance industry. Examples include fraudulent credit card purchases and false insurance claims. Fraudulent claims are estimated to account for approximately 10% of the claims and claims settlement costs faced by the insurance industry each year. The expansion and prevalence of digitization in the financial and insurance industries is generating huge data sets, especially text data, that can be exploited for fraud detection. Unlike existing machine learning techniques, the text embedding architecture proposed in this article can help detect fraudulent claims.We demonstrate the method using a data set from a large international health insurance company.

Keywords—*Deep learning architectures and misleading claims.*

I. INTRODUCTION

The detection and prevention of fraud have become critical issues for businesses in recent years, as defrauding activities can cause significant financial losses and affect a company to experience a fall in reputation. Machine learning techniques have emerged as powerful tools to analyze fraud data and identify potential fraudulent activities. However, the use of these techniques presents several challenges that need to be addressed to achieve accurate and reliable results. In particular, billing data is often in an unstructured format, fraud data is highly imbalanced, and billing is not timed, making it difficult to select appropriate classification methods and performance indicators. This research paper aims to explore these challenges in detail and propose advanced techniques to overcome them,such as deep learning and anomaly detection algorithms, to improve the efficiency and accuracy of fraud detection systems.There search will investigate the current state-of-the-art techniques used to detect and prevent fraud, including statistical and machine learning methods, and explore their limitations in detecting complex fraudulent activities. The paper will also examine the potential of using advanced techniques such as natural language processing and one-clas classification to address the challenges posed by unstructured billing data and highly imbalanced fraud data. Additionally, there search will analyze the impact of the number of products billed on the bill in the fraud detection process and propose suitable solutions to handle this variability.

The objective of this research is to contribute to the development of a fraud detection systems that can better protect businesses from the financial and reputational harm caused by fraudulent activities.

II. LITERATURESURVEY

Fraud detection is a critical issue in the insurance industry, and statistical and machine learning techniques have emerged as powerful tools to analyze bigdata and identify potential fraudulent activities.

In 2017, Weweilin et al proposed an Insurance Big Data Analysis using Ensemble Random Forest Algorithm. The study found that while random forest algorithms can handle large volumes of data, they are less interpretable than single decision trees.

In2019,Wen Shiand WeiNeng Chen proposed an Adaptive Distribution Algorithm for Multipolicy Insurance Investment Planning. The research focused on anticipating unforeseen events and adapting to changing market conditions to improve investment planning.

In 2020, Chun Yan et al. proposed a Determination Model based on the CNN-HVSM Algorithm for car Insurance Investment Planning. The study found that small changes in the dataset can make the tree structure unstable,which can cause variance in the model.

In 2022, Khyati Kapadiyal et al. proposed a Block chain and Ai-SVM Empowered Health care Insurance Fraud Detection system. The study highlights the importance of choosing a good kernel function for SVMmodels.

Finally,in2021,Tallal Omarand Mohamed Zohdy proposed Health Insurance Premiums for People of Different Ages using Data-Driven Prediction with Clustering Application. The study found that clustering algorithms require high memory to store all of the training data.

In summary, the literature survey reveals that statistical and machine learning techniques have been extensively studied for insurance fraud detection.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 2, August 2023
32

The studies indicate the importance of considering model interpretability, adapting to changing market conditions, handling variance,selecting suitable kernel functions,and managing memory usage when dealing with large datasets.

III. EXISTING SYSTEMS

The existing system proposes a novel approach to fraud detection by utilizing auto encoders to identify fraud cases based on the total reconstruction error A-RE)of the underlying data. A neural network that is trained to regenerate input data with minimum error is know as Auto encoder, which makes them well-suited for detecting anomalies incomplex and unstructured datasets. By calculating the A-REof each data point, the system is able to identify instances with high reconstruction errors, which are indicative of potentially fraudulent activity.

Usage of such kind of system provides major advantages and one of them is that the processing time of datasets get reduced, which is a critical factor in fraud detection where timely action is often necessary. However, it also has the disadvantage of relying on existing methods with competing accuracies, which may limit its effectiveness in certain scenarios. Additionally, only part of the classifier model has been implemented, which may further limit its accuracy and generalizability.

Despite these limitations, the use of auto encoders for fraud detection has shown promising results in previous studies. Furthermore, the proposed approach can be easily extended and improved upon by incorporating additional machine learning techniques, such as feature engineering or ensemble learning. Overall, the existing system provides a useful starting point for further research in the field of fraud detection and prevention

IV. PROPOSEDSYSTEM

The proposed model aims for improvement in accuracy and performance which can detect fraud using machine learning algorithms. The system will employ the use of Adaboost and Catboost algorithms to enhance the accuracy of fraud detection.

The prediction method of the proposed system wills use these algorithms to analyze the data and detect fraudulent activities. Adaboost algorithm is used as boosting algorithm that is used for combining multiple weak classifiers and this creates a strong classifier, while Catboost algorithm is a gradient boosting algorithm that can handle categorical features better than other algorithms.

The advantages of this proposed system are numerous. Firstly,the prediction accuracy is expected to be significantly higher than existing methods, thereby reducing the rate of false positives and false negatives.

Secondly, the system will require less manpower and time to analyze large datasets,leading to a more efficient and effective fraud detection system.Lastly,the proposed system is less error-prone, providing a more reliable and trustworthy result.

In conclusion, the proposed system utilizing Adaboost and Catboost algorithms both bring significant enhancement in the accuracy and efficiency of fraud detecting systems in the insurance industry.

V. METHODOLOGY

The proposed approach for classification analysis involves several steps that aim to transform a large and complex dataset into a simplified, pre-processed dataset that can be effectively used for classification. The first step involves utilizing a dataset of Sequence Embedding Help,which includes a sequence of items, each of which has a corresponding embedding.

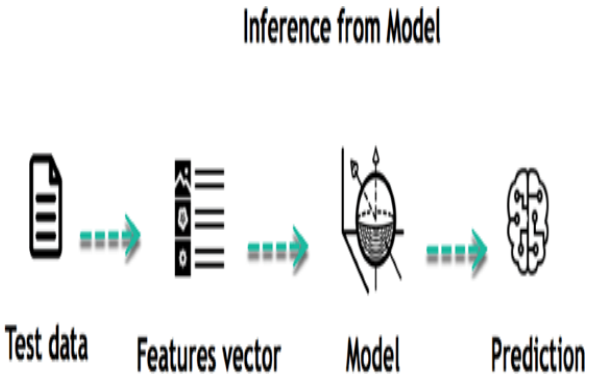

Fig1:Basic flowofthe project

The next step is to filter out any irrelevant data and reduce noise in the dataset by including only the relevant

Attributes needed for analysis .Once the new dataset has been constructed, it needs to be pre-processed, which involves several steps such as normalization, scaling, and feature extraction.

The dataset must be divided into training data and testing data, post pre-processing. Now our training data will be used to train the model for classification and testing data is used for evaluating the performance of the model.

The classification algorithm is applied to the testing data after training the model with the training data. Various machine learning algorithms can be used for classification depending on the specific needs of the analysis. Some example can be SVMs(support vector machines), decision trees, logistics regression, and random forests.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 2, August 2023
32

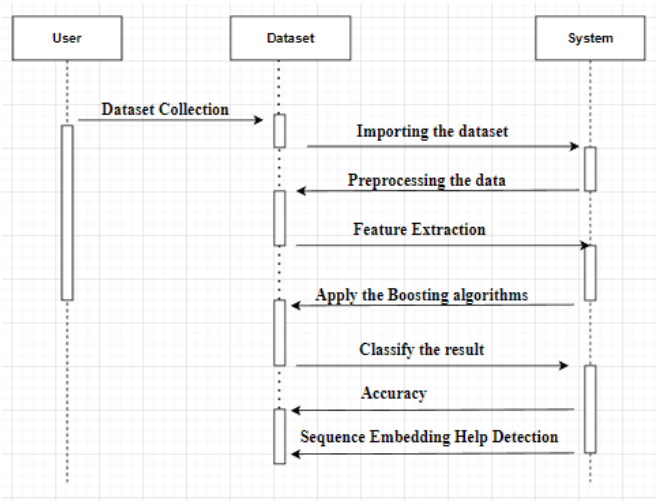

Fig2: SequentialDiagram

The final stage is for evaluating the performance of the model and this is done using accuracy metrics for example F1 score, precision, and AUC(area under the curve). Now following theses steps and using such powerful classification algorithm, it is possible to achieve high levels of accuracy in predicting the outcomes of the analysis.

The approach of this proposed model is useful in different fields for example fraud detection, language processing, and image recognition. One of the most important remark can be stated that the model accuracy is majorly dependent on the dataset quality, the algorithm used, and the techniques applied for pre-processing. Hence, it is crucial to ensure that the dataset is relevant, of good quality, and well-processed to obtain accurate results.

VI. MODULEDESCRIPTION

A. Collection of data

Data collection is a process that gathers information on Sequence Embedding Help from a variety of sources, which is then utilised to create machine learning models. The type of data that is being used in this work is Sequence Embedding Help data with features. The purpose of this stage is to make the selection of the subset available and accessible in which we will be working. The challenging part in machine learning is when working with large amounts of data and in this, the desired solution is already known. The type of information for which the desired outcome is already is called Labelled Data.

B. Pre-processing of data

Data undergoes a few steps in data pre-processing to format, clean, and sample the data. There are three stages in data pre-processing:

In general, the format of that data does not allow us to deal according to our needs. The file format of data that we want is likely to be a relational database or text file, but the data may be different file format. This stage helps to format the data according to the requirements.

1) Cleaning

The process of replacing the missing data is known as Data Cleaning. There might be data with insufficient instances and missing information which does not require for addressing the issue. This stage is needed to eliminate such unwanted information.

2) Sampling

The precaution need to be taken care of when we get more access to the data which is not required. Algorithms may require more compute and memory to run as well as take significantly longer to process larger volumes of data. Instead of taking the complete dataset, the better approach would be to take a smaller representative of selected data which may be much faster and this helps with exploring and testing.

C. Extraction of features

The next step is to A process of attribute reduction is feature extraction .Feature extraction actually alters the attributes as opposed to feature selection, which ranks the current attributes according to their predictive relevance.The original attributes are linearly combined to generate the changed attributes,or features.Finally,the Classifieralgorithm is used to train our models. We make use of the acquired labelled dataset. The models will be assessed using the remaining labelled data we have. Pre-processed data was categorised using a few machine learning methods. Randomforest classifiers were selected.

D. Evaluating the model

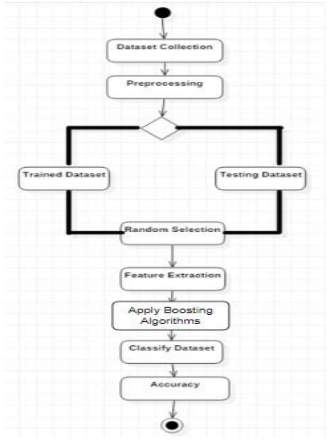

Fig3:ActivityDiagram

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 2, August 2023
32

The techniques that are being used to access the model in data science are Hold Out and Cross-Validation. To prevent overfitting, both strategies use a test dataset.

The performance of each categorization is estimated based on its average performance. The desired outcome has to be take a form that was predicted earlier. And the graph representation of data is shown according to category.

AlgorithmsUsed:

- ADABoost
- CatBoost

In the process of model development, a step is involved and the step is called model evaluation. Knowing the model which provides the data and predicts how our model will perform is very much useful for the future. In the field of data science, performing an evaluation using a training dataset is not acceptable because this can lead to a too optimistic and over fitted models.

CAT Boost (Categorical Boosting) is a machine learning algorithm designed for handling datasets with categorical features. It is an extension of the gradient boosting method that uses an optimized gradient descent algorithm to learn a model. CATboost can handle complex datasets with categorical features and also handle large datasets. Because of this reason, it has been used in real-world applications and even in various machine learning competitions. It has a ability to handle missing values and its auto match and ling of categorical features, which reduces the need for manual feature engineering.

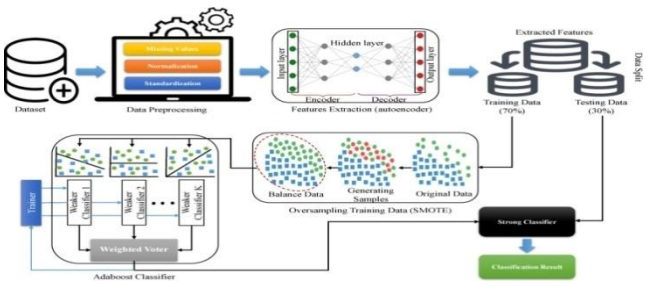

Fig4:ADABOOST

1) ADA Boost

The machine learning algorithm used for classification and regression problems is known as Adaptive Boosting(AdaBoost). AdaBoosting is an ensemble learning which tends to combine the outputs of a weak learner to produce a stronger learner.AdaBoost is the algorithm used to improvement of the performance of weak learner and produces strong learner that is more accurate than any of the individual weak learners. It is known for its ability to handle imbalanced datasets and its resistance to overfitting.

CONCLUSION

In conclusion,the proposed approach of utilizing sequence embedding for fraud detection involves a series of steps that are designed to transform a large and complex dataset into a simplified, pre-processed dataset that can be effectively used for classification. The approach involves filtering irrelevant data, pre-processing, dividing the dataset into training and testing data, applying a powerful classification algorithm ,and evaluating the model's performance using accuracy metrics.

REFERENCE

[1] Riya Roy and Thomas George K, "Detecting Insurance claims fraud using machine Learning Techniques, 2017 International Conference on circuits Power and Computing Technologies[ICCPCT]

[2] LutaoZheng,GuanjunLiu,ChungangYan,andChangjun Jiang, "Transaction Fraud DetectionBased on Total Order Relation and BehaviorDiversity

[3] SangeetaMittalandShivaniTyagi,"PerformanceEvaluation of Machine Learning Algorithms forCreditCard Fraud Detection",978-1-5386-5933-5/19/$31.00©2019IEEE

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 2, August 2023
32