Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/20307
DOI URL: http://dx.doi.org/10.21474/IJAR01/20307

RESEARCH ARTICLE

FINANCIAL DISTRESS PREDICTION USING MACHINE LEARNING

Padma Magar
Malini Kishor Sanghvi College of Commerce.

Manuscript Info

Manuscript History
Received: 25 November 2024
Final Accepted: 28 December 2024
Published: January 2025

Abstract

Financial distress occurs when a company or individual is unable to generate adequate revenue or money to fulfil or payback its financial commitments. This research looks at how Machine Learning can be used to identify personal financial distress. Financial frauds are a rising problem in the financial services industry with far-reaching implications, while numerous techniques have been developed, Machine Learning has been used to automate the processing of large volumes of complicated data in finance systems. In the identification of distress, Artificial Intelligence has played a significant role in the financial industry. Predicting different frauds or patterns is a big data challenge that is made more difficult by two factors: first, the profiles of normal and fraudulent behaviour vary regularly, and second, cybercrime data sets are highly skewed. This research explores and compares the performance of Different Machine Learning Models on publicly available dataset. Dataset of 15,000 individuals is sourced from public repository by Lending.com. The Algorithms are implemented on the raw and pre-processed data and the outcome of these Algorithms/Models is evaluated based on accuracy, sensitivity, specificity and precision.

Introduction:-

This research tries to predict if an individual can face financial distress over the period of next 2 years. This information can be very important to financial institutions which will cater its services to such individuals. If this research can predict the financial distress of an Individual, then that data can be used by financial institutions to limit or deny services to any individual who can face financial distress in near future. Financial fraud is a rising problem in the government, business organisations, and the financial industry, with far-reaching consequences. The heavy reliance on internet technologies in today's environment has accelerated financial transactions. As online transactions have become a more common way of payments, emerging computational approaches for dealing with financial services difficulties have gotten a lot of attention. Many credit scoring systems and tools are available to help organisations such as credit card industry, retail sector, e-commerce services, insurance, and other industries to avoid fraud. It is difficult to be absolutely confident of an application's or transaction's real intention and legality. The most effective method is to use mathematical algorithms to search for probable fraud evidence in the existing data. The procedure of identifying those individuals that are suspected is converted into two classes of real class and distress class, various algorithms and models are developed and deployed to solve such tasks as deep neural networks, frequent item set mining, machine learning models, migrating bird’s optimization algorithm, logistic regression, Support Vector Machines, decision tree and random forest.

Corresponding Author:- Padma Magar
Address:- Malini Kishor Sanghvi College of Commerce.

These problems are quite prevalent in the financial world, yet they are also hard to resolve. First, it's difficult to match a pattern for data set because of the fact that there is just a little amount of data. Second, several data collection items with separate truncations may likewise fit within acceptable conduct patterns. There are also several limitations to the problem. First of all, data sets are not easily accessible to the Public, and the outcomes of study are often obscured and monetized, making the results unavailable. Datasets with actual published studies are not mentioned in previous research. Furthermore, it is more difficult to develop techniques by limiting the interchange of ideas and methodologies in these studies as a result of the security issue. Finally, the data sets are always changing, which makes it possible to distinguish the profiles of ordinary and malicious behaviours that the legal transaction in the past has been or is still a fraud. This study examines the four approaches of machine learning, decision tree, vector support, logistic regression and random forests, followed by a joint comparison to assess which model was best performed.

Literature Survey:-

In [1] This paper represents a case study involving the prediction of fraud, which shows that before modelling, data standardisation is used and with the results obtained from the use of unattended learning networks and deep neural fraud detection networks that clustering characteristics can minimise neural input. The use of standardised data with already trained data can also provide intriguing outcomes. In this study, unsupervised learning was applied and new techniques for fraud prediction were designed and the findings accurately improved.

In [2] A new comparative measure was created in this study, which effectively aggregates evaluation metrics. A cost-sensitive method based on Bayes is presented with the proposed cost measurement. When comparing this approach with other state-of-the-art algorithms, up to 25% gains are achieved. The data collected for this research was based on transactional real-life data of a major Global firm, and personal data was kept confidential. The accuracy of an algorithm is about 60%. This effort was aimed at developing an algorithm and reducing costs. The result was a 26% increase, with Bayes' least risk approach.

In [3] To identify fraudulent transactions, several current approaches based on pattern recognition, deep neural networks, machine learning, artificial intelligence and others have been developed and are continuously being developed. All these techniques require a thorough and clear knowledge, which will undoubtedly lead to an effective system. This paper comprises an examination of several techniques and an evaluation of each methodology on the basis of certain performance standards. The survey in this paper aimed to assess each technique's efficiency and sensitivity. The relevance of this study is carrying out a study to assess multiple algorithms so that the best way to solve the problem is determined.

In [4] In this study, a comparison of artificial intelligence models is done, as well as a comprehensive explanation of the created fraud detection system, such as the Naive Bayes Classifier and the model on Bayesian Networks, the deep neural network model. Finally, judgments regarding the outcomes of the models' evaluative testing are reached. Using the Bayesian Network, it was found that the number of lawful truncations was higher or equal to 0.68, indicating that their accuracy was 68 percent. The purpose of this work is to compare artificial intelligence models, along with a general parametrization of the produced system, and to indicate the specificity of each model, as well as recommendations for improving the model.

In [5] Nutan and Suman supported the theory of what is fraud, types of fraud such as telecommunications, fake bankruptcy, and how to detect it in their review on fraud detection. They also explained numerous algorithms and methods for detecting fraud, including the Glass's Algorithm, Bayesian networks, Hidden Markov model, Decision Tree, and others. They offer detailed explanations of how the algorithms operate as well as mathematical explanations. The goal of this study is to identify fraud in a dataset collected from ULB website by utilising Logistic regression, Decision trees, and other models to evaluate their accuracy, sensitivity, specificity, and precision, and compare them to the best feasible model to address the fraud detection problem.

Background

The ability of a system to learn and improve without explicit programming is machine learning. It includes the development of computer systems that can use data for their own learning. That a classifier algorithm may be described as an algorithm for classification, especially when implemented, as well as a mathematical function that is implemented in categories by an algorithm and maps input data. It is a supervised learning example, which provides a training set of correctly accepted observations.

Logistic Regression:
Logistic regression is a supervised classification technique predicting the likelihood of a binary variable depending on the independent variable in the data set. The probabilities of a result with two values, zero or one, yes or no, false or true, are predicted using logistic regression. Linear regression is like a straight-line regression, but logistic regression generates a sigmoid curve. Based on a predictor or an outcome variable, the logistic regression produces sigmoid curves which represent zero to one value based on logarithmic functions. Regression is a model with a category dependent variable which analyses the link between several independent variables. The logistic regression models, including binary, multiple and binomial logistic models, are many variants. The Binary Logistic Regression Model calculates the probability of a binary response depending on one or more variables.

SVM (Support Vector Machine):
SVM is a method of machine learning regression and classification. It is a supervised form of learning that captures information. Modelling SVM involves two steps: training a data set to build a model, and then predicting information from a test data set by using that model. The SVM model depicts the training data points as points in the n-dynamically spatial range, then maps them in a way that separates the points of various classes from the broadest range possible. In the SVM technique, each data item is treated as a point for n-dimensional space, where n is the number of characteristics and the value of each feature is the value of a specific co-ordinate. The classification is then performed by locating the hyperplanes which separate the two groups clearly.

Decision Tree:
Decision tree is an algorithm that provides a tree-like graph or model of decisions and their likely consequences for probabilistic choosing. This method uses conditional assertions of control. It is an algorithm for an objective function, which represents an alpha function in the decision tree. These algorithms are famous for inductive learning and have been used to various applications efficiently. It assigns a label to a new block, indicating whether the class label is valid or false, then test the transaction value against the decision tree, and then trace the journey from the root node to that item's output/class label. Decision rules determine the results of the contents of the leaf node. In principle, rules are 'If condition 1 and condition 2 are true, but condition 3 is wrong, the result is false'. This decision tree makes it easy to understand and analysis and enables the insertion of additional scenarios, making it easier to establish the worst, best and expected values for diverse situations.

Random Forest:
Random Forest is a technique of regression and classification. It is a group of decision tree classifiers. A fraction of the training sample is sampled altered so that each node splits all the exercises in a single tree and then one decision tree by random subset. Also, it is remarkably quick even for large-scale sets of training and data in random forests with every tree being trained independently. This technique provides a good evaluation of the generalisation error and resists overfitting. The relevance of variables may be determined naturally in the Random Forest using a random forest in a regression or classification task.

Methodology:-

The initial data is derived from the data source and the validation is done on the data set, where the redundancy is removed, empty spaces are filled into columns and the required variable is converted in factors or classes. The K-fold crosses are now validated and randomly separated into k sub-samples of the same size. The validation of the model is retained as a subsample, while the rest of the k sub-samples are used as training data. Logistic regression, decision-making, SVM and random forest models will be developed, and precision will be tested, and a comparison will be made. Sensitivity will then be evaluated.

The data set comes from a public repository that is updated for peer-to-peer financial services by Lending.com. The dataset contains information of 15000 individuals with their financial history. The data set is severely imbalanced, and 0.18 percent of the data is skewed to the negative class. It comprises solely numerical (continuous) input variables that are transformed into 12 main components according to the Principal Component Analysis (PCA). And in this study a total of 8 input functions are used. A variable in each profile usage, indicating customer financial situation combined with days of month, hours of days of day, geographical sites or type of the merchant in whom the transaction takes place is the typical behavioural of the individual. Confidentiality problems cannot provide the specifics and context of characteristics. The time function saves the seconds between each transaction and the first transaction in the dataset. The transaction value is the 'amount' feature. Feature 'class' is the binary class target class and takes value 1 in positive (failure) and value 0 in negative (fail) cases (non-fraud).

Four classification models were trained in this study based on logistic regression, SVM, decision-trees and Random Forest. 80% of the data set is utilised for training to assess these models, whilst 20% are used for validation and testing. The performance of the four classifiers is assessed using accuracy, sensitivity, specificity, precision. In every set of a sample the true positive, true negative, false positive and false negative rates are represented in the table below and a confusion matrix format is also shown. The precision and specificity ratings of several true negatives are inaccurately high in the table.

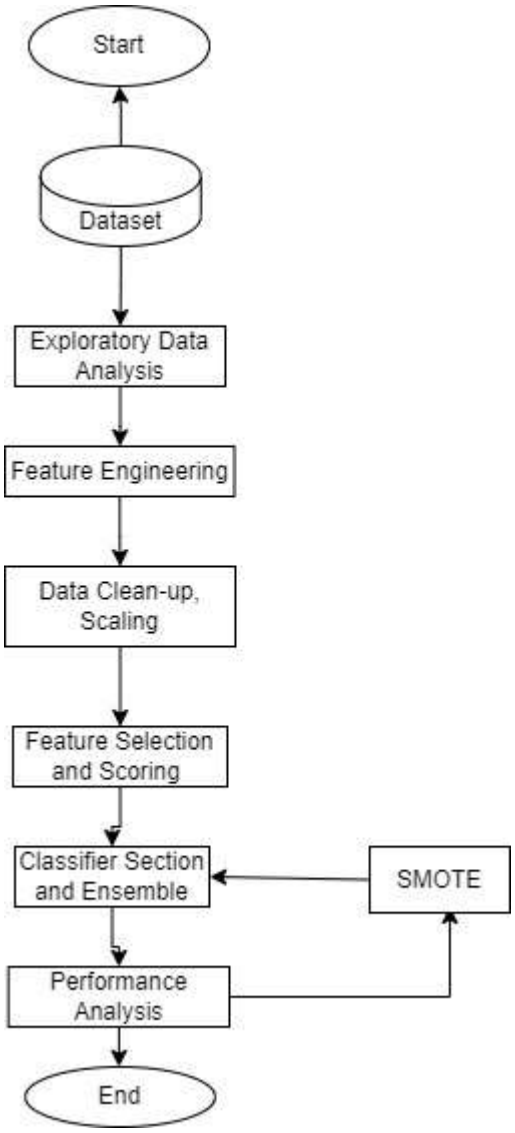

Figure 1:- Architecture.

Results:-

From the studies, it has come to the knowledge that the logistic model is 97.7 percent accurate, while the SVM is 97.5 percent accurate as well as the decision tree is 95.5 percent accurate, however, the Random Forest with highest outcomes have achieved. 98.6 percent precision. Interpreting form different model performance metrics, it comes to light that model was overfitting the training data because of bias inherited form the dataset. After SMOTE was applied, Model performance was seen to be improved. Random Forest was seen to be the best performer on the dataset.

Metrics	Logistic Regression	SVM	Decision Tree	Random Forest
Accuracy	0.977	0.975	0.955	0.986
Sensitivity	0.965	0.973	0.949	0.991
Specificity	0.923	0.912	0.893	0.982

Precision	0.996	0.995	0.963	0.994

Table 1:- Performance Metrics.

Conclusion:-

Though there are many identity verification methods available today none is able to identify all frauds entirely while they are actually occurring, they generally detect it until the fraud has been perpetrated. This happens because a very minuscule number of transactions from the total transactions are actually fraudulent in nature. With more learning information, the Random Forest Algorithm will do faster, but velocity will be impaired in experimentation and implementation. It would also assist to implement more pre-processing methods. The support vector machine software already comes from unbalanced data sets issue and needs a higher preliminary processing rate to achieve superior outcomes at the outcomes as seen by Support vector machine. The requisite to develop a successful hybrid system is to combine costly training techniques with incredibly precise and exact outcomes with an enhancement method to reduce system costs and rapidly train the machine. The selection of hybrid methods depends on how the fraud sensing device works and the workplace

References:-

[1] Raj S.B.E., Portia A.A., Analysis on credit card fraud detection methods, Computer, Communication and Electrical Technology International Conference on (ICCCET) (2011), 152-156.
[2] Jain R., Gour B., Dubey S., A hybrid approach for credit card fraud detection using rough set and decision tree technique, International Journal of Computer Applications 139(10) (2016).
[3] Dermala N., Agrawal A.N., Credit card fraud detection using SVM and Reduction of false alarms, International Journal of Innovations in Engineering and Technology (IJIET) 7(2) (2016).
[4] Phua C., Lee V., Smith, Gayler K.R., A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119 (2010).
[5] Bahnsen A.C., Stojanovic A., Aouada D., Ottersten B., Cost sensitive credit card fraud detection using Bayes minimum risk. 12th International Conference on Machine Learning and Applications (ICMLA) (2013), 333-338.
[6] Carneiro E.M., Dias L.A.V., Da Cunha A.M., Mialaret L.F.S., Cluster analysis and artificial neural networks: A case study in credit card fraud detection, 12th International Conference on Information Technology-New Generations (2015), 122-126.
[7] Hafiz K.T., Aghili S., Zavarsky P., The use of predictive analytics technology to detect credit card fraud in Canada, 11th Iberian Conference on Information Systems and Technologies (CISTI) (2016), 1-6.
[8] Sonepat H.C.E., Bansal M., Survey Paper on Credit Card Fraud Detection, International Journal of Advanced Research in Computer Engineering & Technology 3(3) (2014).
[9] Varre Perantalu K., Bhargav Kiran, Credit card Fraud Detection using Predictive Modeling (2014).
[10] Stolfo S., Fan D.W., Lee W., Prodromidis A., Chan P., Credit card fraud detection using meta-learning: Issues and initial results, AAAI-97 Workshop on Fraud Detection and Risk Management (1997).
[11] Maes S., Tuyls K., Vanschoenwinkel B., Manderick, B., Credit card fraud detection using Bayesian and neural networks, Proceedings of the 1st international naiso congress on neuro fuzzy technologies (2002), 261-270.
[12] Chan P.K., Stolfo S.J., Toward Scalable Learning with NonUniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection, In KDD (1998), 164-168.
[13] Rousseeuw P.J., Leroy A.M., Robust regression and outlier detection, John wiley & sons (2005).
[14] Wang C.W., Robust automated tumour segmentation on histological and immunohisto chemical tissue images, PloS one 6(2) (2011).
[15] Sait S.Y., Kumar M.S., Murthy H.A. User traffic classification for proxy-server based internet access control,IEEE 6th International Conference on Signal Processing and Communication Systems (ICSPCS) (2012), 1-9.