International Journal on Applications in Engineering and Technology
Volume 8 : Issue 2 : May 2022, pp 32 – 37 www.aetsjournal.com ISSN (Online) : 2455 - 0523

Anti-Fraud Model for Internet Loan prediction

MOHAN KUMAR.M[1], PAVAN KUMAR.K[2], KALAIYARASAN.M[3], BRINDA.B.M[4]

[1,2,3] Undergraduate student, Department of Computer Science and Engineering,
Paavai College of Engineering

[4] Assistant Professor, Department of Computer Science and Engineering,
Paavai College of Engineering

Abstract: - Internet finance is increasingly popular. However, bad debt has become a serious threat to Internet financial companies. The fraud detection models commonly used in conventional financial companies is logistic regression. A large public loan dataset, e.g. Lending club, for example, to explore the potential of applying deep neural network for fraud detection. An XGBoost algorithm is employed to select the most discriminate features. After that, we propose to use a synthetic minority oversampling technique to deal with the sample imbalance.

Keywords: Internet finance, loan fraud detection, deep learning, financial model.

1. INTRODUCTION

Machine learning (**ML**) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

Internet fraud methods are increasing dramatically in recent years, Internet lending companies face an unprecedented risk of online fraud. The rules of this model were manually constructed by the fraud experts from the bank. Edge and Sampaio proposed a set of a financial fraud modeling language (FFML) for better describing and combining fraud rule sets to assist fraud analysis.

The rules of this model were manually constructed by the fraud experts from the bank. Sanchez *et al.* proposed to use association rules to detect fraud and help risk analysts extract more fraud rules. Edge and Sampaio proposed a set of a financial fraud modeling language (FFML) for better describing and combining fraud rule sets to assist fraud analysis. However, the rule-based models require sufficient and accurate expertise knowledge and can not be updated timely to new frauds. To this end, machine learning models have been introduced for fraud detection. Gosh and Reilly uses neural net- works to detect credit card fraud. Kokkinaki proposed decision trees and Boolean logic functions to characterize normal transaction patterns to detect fraudulent transactions. Peng *et al.* compared nine machine learning models for fraud detection. The results demonstrate linear logistic and Bayesian networks are more effective. Lei and Ghorbani proposed a new clustering algorithm namely improved com- petitive learning network (ICLN) and supervised an improved competitive learning network (SICLN). Sahin *et al.* designed a decision tree based on cost sensitivity. Halvaiee and Akbari proposed to use an AIRS improved algo- rithm for fraud detection. However, these traditional machine learning methods heavily rely on manual subjective rules and easily lead to model risk. These methods also tend to overfit due to the imbalance training dataset with serious pol- lution by noises. Thus, ensemble learning methods have also been introduced to integrate different models for complicated fraud detection. Louzada and Ara proposed a bagging ensemble model that integrates k-dependence probabilistic networks. The results show that the proposed ensemble model has stronger modeling capabilities. Carminative *et al.* proposed a combination of semi-supervised and unsupervised fraud and anomaly detection methods, mainly using a histogram-based outlier score (HBOS) algorithm to modelthe user's past behavior.

A. XGBoost Algorithm

XGBoost stands for Extreme Gradient Boosting, which was proposed by the researchers at the University of Washington. It is a library written in C++ which optimizes the training for Gradient Boosting. Before understanding the XGBoost, we first need to understand the trees especially the decision tree**:**

B. Decision Tree:

A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be *"learned"* by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions.

C. Bagging:

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

Each base classifier is trained in parallel with a training set which is generated by randomly drawing, with replacement, N examples(or data) from the original training dataset, where N is the size of the original training set. The training set for each of the base classifiers is independent of each other. Many of the original data may be repeated in the resulting training set while others may be left out.Bagging reduces overfitting (variance) by averaging or voting, however, this leads to an increase in bias, which is compensated by the reduction in variance though.

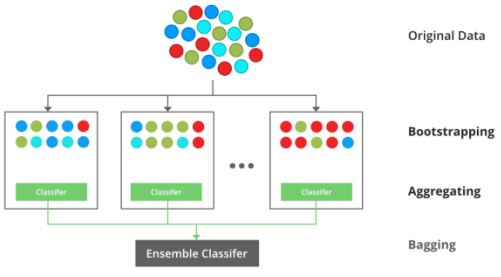

Fig -1: Bagging classifier

D. Boosting:

Boosting is an ensemble modeling, technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

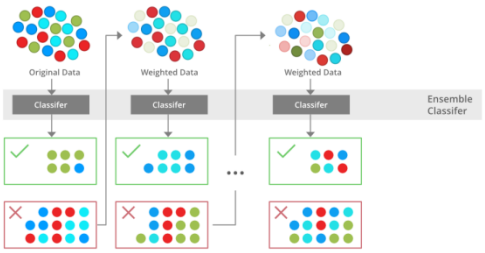

Fig -2: Boosting

E. Gradient Boosting

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels. There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees). Traditional machine learning methods, on the other hand, rely primarily on manual subjective criteria and are prone to model risk. These approaches also have a tendency to overfit due to an unbalanced training dataset with significant noise pollution. As a result, ensemble learning methods have been created to incorporate many fraud detection models. The bagging ensemble model suggested by Loused and Are integrates k-dependence probabilistic networks. The results suggest that the proposed ensemble model is more capable of modeling. Carminative et al. suggested a hybrid of semi-supervised and unsupervised fraud and anomaly detection approaches, primarily based on the usage of a histogram-based outlier score (HBOS) algorithm to represent the user's previous behavior. Deep learning techniques have recently gotten a lot of academic and corporate interest, and they offer a new perspective on financial data processing. Convolutional neural networks were utilized by Fu et al. to efficiently minimize feature redundancy. To et al. propose a fraud detection system based on deep feature representation. To combine prior knowledge with the deep network, Greiner and Wang pointed out that before getting the loan, the borrower is likely to conceal information that is not beneficial to him or even fictitious favorable information. The borrower is likely to default unilaterally after receiving the loan.

II. METHODOLOGY

A. Detection of financial fraud on the internet:

The financial industry's expansion has surely been boosted by the reform of Internet technology. The hazards of Internet finance are rising in tandem with the emergence of diverse Internet finance models. From the standpoint of fraud, the most typical types of Internet financial fraud are as follows:

- **Identity theft:** Criminals steal personal financial information from users in order to carry out fraudulent financial transactions or remove funds from their accounts.
- **Investment fraud:** False, misleading, or fraudulent information is used to sell investments or securities.
- **Mortgage and loan fraud:** The borrower uses false information to open a mortgage or loan, or the lender uses a high-pressure sales strategy to sell the mortgage or loan or predatory loan to users.

B. Cleaning of data and feature selection:

Cleaning of data and selecting features are critical prerequisites for developing a feature-rich anti-fraud model. The model identification efficiency will be substantially enhanced if this part of the task is adequately handled. We start by removing any incorrect variables or variables with the same value that aren't reasonable. Then we determine out what values are lacking and mark these holds as needing to be filled. After that, we perform a random experiment to extend and derive the missing variables. After that, we perform a random experiment to extend and derive the missing variables. Then, to pick key characteristics, we use an XGBoost model. As a consequence, there are two independent variables and one target variable.. The samples in each category are severely skewed after picking the characteristics. We propose an oversampling synthetic minority oversampling approach (SMOTE) for processing in this respect. The data is then normalized in preparation for modeling.

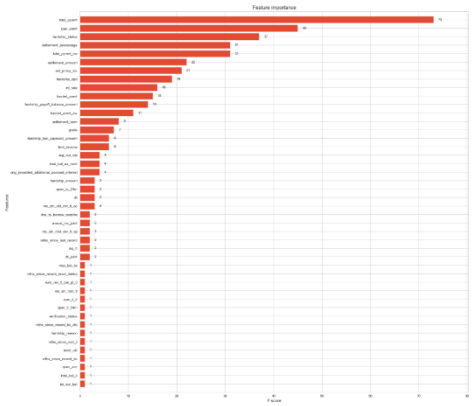

Fig -3: The XGBoost algorithm calculates the feature significance

C. Internet fraud detection with a deep neural network:

We simulate the Internet fraud detection job using a simple yet efficient deep neural network after correctly screening out the discriminated features. An input layer, a few hidden layers, and an output layer make up a deep neural network (DNN). The input layer is the first, the output layer is the final, and the layers in between are all concealed layers. There are numerous neurons in each layer. The input data determines the number of neurons in the input layer. Other layers' neuron counts will be modified based on the current condition. The number of concealed layers, which is usually greater than one, can be customized. Layers are frequently totally linked, meaning that each neuron in one layer may communicate with all neurons in the next. The neurons of the layer do not communicate with one another. Three hidden layers make up the deep neural network. Self-learning is achieved by deep neural networks via forwards propagation and back propagation. The process of feeding samples into the neural network, passing through the hidden layer, and eventually receiving results from the output layer is known as forward propagation. The loss function's outcome may be used to assess the model's fitting degree. The loss function is usually the mean square error between the output layer result and the sample label. In back propagation, the gradient descent approach is widely used to reduce iterative loss function optimization. The value of the loss function is continually changing, and parameters such as weights and offset values are regularly modified during this process.

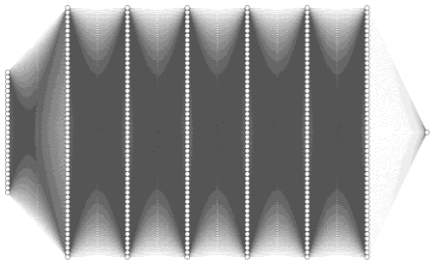

Fig -4: The architecture of the deep neural network

III. EXPERIMENTATION

A. Description of data:

The experimental data is drawn from a huge public lending dataset made available by Lending Club. A total of 200,000 data records were collected from the fourth quarter of 2016 to the second quarter of 2017. There are 145 characteristics in each record, including 107 floating-point values and 38 character values. The first dataset has a size of 228.1 MB.

The samples in the dataset are divided into two groups: 70% were used for training and the remaining 30% were used for testing. XGBoost is used to choose discriminating characteristics. The Internet fraud detection algorithm is trained using thirty carefully chosen characteristics. First, every feature data is normalized. These characteristics are then loaded into a deep neural network, which has an input layer, six hidden layers, and an output layer.

B. Empirical analysis:

We can see that the minimum loan amount is $1,000 and the maximum loan amount is $40,000. The loan amount is mostly concentrated around $10,000, with a median of $12,000. This business specialises in small loans. Intuitively, the larger the loan amount, the greater the risk.

There are two types of loan products in this dataset: 36-month and 60-month. The proportion of loans with a 60-month loan is 24.45 percent, and the proportion of loans with a 36-month loan is 75.55 percent. As a general rule, the longer the loan term, the higher the risk. The second is the payback of a credit card. One who uses a loan to consolidate debt and pay off credit cards has a limited cash flow. Before shifting to P2P platform loans, such customers are likewise unable to make loans. These consumers have a poor ability to repay their loans. Default is a greater chance. For other loans, the risk must be evaluated further based on the customers' other characteristics. Customers' credit ratings are divided into seven categories by Lending Club, ranging from A to G. Customers that have an A credit rating have the best credit ratings. Customers with a credit score of G have the lowest credit scores. Customers with a good credit score are less likely to default on their payments.

At the moment, Category B customers have the highest credit rating, followed by Category C and Category A. These three groups account for 81.12% of the total. Customers with credit ratings of E, F, and G account for 5.42 percent of the total. Lending Club's credit department has tighter controls over applicants' credit histories.

C. Design choices:

To implement the models discussed in this study, we use the Tensor flow framework. In the buried layers, the number of nodes is crucial. The nodes in the hidden layers are determined using the empirical formula below.

$$N \quad \frac{N_S}{(1)(\alpha \times (N_i + N_o))}$$

where Ni denotes the number of neurons in the input layers, No denotes the number of neurons in the output layers, and Ns is the number of samples in the training set, which is typically set between 2 and 10. When calibrating a deep learning model, batch size and epochs are critical. The accuracy of the model for various combinations of batch size and epochs. The model achieves the highest accuracy when batch size is 1000 and epoch is 200. To train our model for performance evaluation, we set the batch size to 1000 and the epochs to 200. To train the deep neural network, we compare several optimization strategies. SGD, RMSprop, AdaGrad, Adadelta, Adam, Adamax, and Nadam are examples of optimization algorithms often used in deep learning. The accuracy of various optimization strategies when applied to the DNN model. The performance of the optimizers is similar, with the exception of the AdaGrad optimizer, which has a lower accuracy rate. The Adam optimizer outperforms the other six optimizers by a little margin. As a result, the Adam optimizer is used to train the DNN model.

D. Experimental outcomes:

The activation function was set to the ReLU function and the sigmoid function. The batch size is 1000, the number of epoch is 200, and the Adam function is used as the optimizer. The model's loss function is set to binary cross-entropy. After 175 iterations, the accuracy changes little. After 175 epochs, the model's accuracy is pretty high. The loss is also quite minimal at this time. The training accuracy is 0.9763 and the training loss is 0.0835 as a final result. The testing accuracy is 0.9771 and the loss is 0.0789, indicating that the suggested model's generalization ability is rather excellent. AUC and KS values, such as AUC 0.97 and KS 0.94, are used to evaluate the model. We can observe that the model is capable of generalization and has a high level of model stability.

E. Discussion and evaluation of performance:

We compare the deep neural network's performance to four regularly used models, including logistic regression (LS), support vector machine (SVM), decision tree (DT), and random forest (RF). Table 3 lists the evaluation outcomes, which include AUC, KS, and ACC. the deep neural network outperforms the comparisons, despite the fact that the metrics are comparable. The decision tree's AUC is the smallest, the KS level is medium, and the ACC is the smallest. The results are slightly less impressive than those of other models. This suggests that the decision tree may be harmed by minor perturbations in the financial data. Logistic regression has the second highest AUC, with the lowest KS and the highest ACC.

IV. IMPLEMENTATION

A. Dataset acquisition:

The data set was gathered from the Kaggle website and is organized into three categories: training, validation, and testing. This will divide our dataset into training, validation, and testing sets according to the above-mentioned ratio: 80% for training, 10% for validation, and 20% for testing. The original dataset comprised of 162 40x scanned slide photos. There is a significant imbalance in the class data, with nearly twice as many negative data points as positive data points.

3.2.Preprocessing:

Preprocessing is the process of reducing the image's dimension. We provide our network the input picture volume structure, where depth is the number of color channels in each image.The picture is resized based on the size of the deep learning layer's rows and columns**.**

B. Feature extraction:

we'll use a CNN (Convolutional Neural Network) to design the network, which we'll call Cancer Net. The following procedures are carried out by this network.3 CONV filters should be used. These filters should be stacked on top of each other. Max-pooling should be done. Use separable convolution by depth (more efficient, takes up less memory).

C. Training of features:

The Adam's training programmer is being executed. The training approach uses adaptive momentum as an optimizer for gradients with epochs. You'll be training on benign data and testing on malignant data, which isn't typical. It's breast cancer, ordered by size, and the things at the beginning are more likely to be benign, while the ones towards the end are more likely to be malignant.
We use Kara's technique to generate the model based on feature vectors.

D. Testing and performance evaluation:

We can split the model using a test set that is 30% of the original data set because this function is implemented. The input simply specifies the size of the input and is referred to as D (see the X_ train shape code above).The dense layer, on the other hand, is where the main work is done: it takes the input and performs a linear transformation to produce a size 1 output. The sigmoid activation function is the linear transformation we wish to use such that the result is between 0 and 1
The module includes loss per iteration, training loss, and validating loss. The accuracy and sensitivity of the data being evaluated.

V. EXPERIMENTAL SETUP

- A bubble chart is another name for a DFD. It is a basic graphical formalism that may be used to depict a system in terms of the data it receives, the processing it does on that data, and the data it generates as output.
- One of the most essential modeling tools is the data flow diagram (DFD). It's used to represent the system's many components. The system process, the data used by the process, an external entity that interacts with the system, and the information flows in the system are all examples of these components.
- DFD depicts how information flows through the system and is transformed through a sequence of transformations. It's a graphical representation of data flow and the changes that occur when data goes from input to output.
- DFD is sometimes referred to as a bubble chart. At any level of abstraction, a DFD may be used to depict a system. DFD can be divided into levels, each representing a different level of information flow and functional detail.

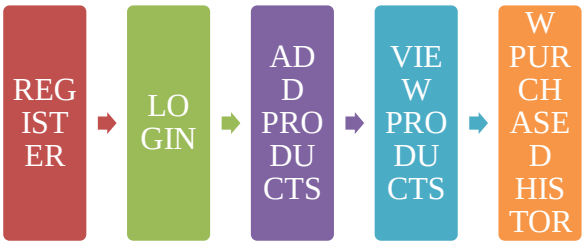

Fig -5: Flow of data

VI. SYSTEM ARCHITECTURE

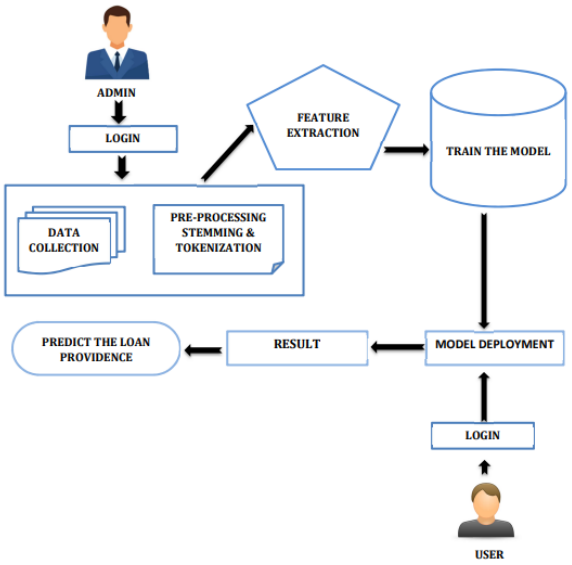

Fig -6: system Architecture

VII. CONCLUSIONS

We take the real customer information of the public loan data set of the lending club company as a sample. Then, we build a deep learning based Internet fraud detection model. We introduce the main parameters of the model and optimizes to find the optimal parameter combination of the model. Finally, the most popular logistic regression in the financial industry as well as other comparisons are used as a baseline to evaluate the performance of the proposed model. The results reveal the deep neural network achieves better performance, which is promising to be used in the financial industry for Internet fraud detection. The results show that the deep neural network performs better, indicating that it might be useful in the banking industry for detecting Internet fraud. For blacklists and whitelists, we want to work with established Internet financial technology companies and institutions in China in the future. The deep neural network, when paired with blacklists and whitelists, as well as expert anti-fraud rules, has the potential to improve fraud detection.

REFERENCES

[1] B. N. N. Li, X. Wang, R. Wang, T. Zhou, R. Gao, E. J. Ciaccio, and P. H. Green, ''Celiac disease detection from videocapsule endoscopy images using strip principal component analysis,'' *IEEE/ACM Trans. Com- put. Biol. Bioinf.*, early access, Nov. 15, 2020, doi: 10.1109/TCBB.2019. 2953701.

[2] X. Li, L. Bai, Z. Ge, Z. Lin, X. Yang, and T. Zhou, ''Early diagno- sis of neuropsychiatric systemic lupus erythematosus by deep learning enhanced magnetic resonance spectroscopy,'' *J. Med. Imag. Health Infor-mat.*, vol. 11, no. 5, May 2021.

[3] J. Wang, Z. Xie, Y. Li, Y. Song, J. Yan, W. Bai, T. Zhou, and J. Qin, ''Real- township between health status and physical fitness of college students from south China: An empirical study by data mining approach,'' *IEEE Access*,vol. 8, pp. 67466–67473, 2020.

[4] C. Li, S. Tang, H. K. Kwan, J. Yan, and T. Zhou, ''Color correction based on CFA and enhancement based on retina with dense pixels for underwater images,'' *IEEE Access*, vol. 8, pp. 155732–155741, 2020.

[5] C. Li, S. Tang, J. Yan, and T. Zhou, ''Low-light image enhancement via pair of complementary gamma functions by fusion,'' *IEEE Access*, vol. 8,pp. 169887–169896, 2020.

[6] C. Li, S. Tang, J. Yan, and T. Zhou, ''Low-light image enhancement based on quasi-symmetric correction functions by fusion,'' *Symmetry*, vol. 12, no. 9, p. 1561, Sep. 2020.

[7] G. Xiao, G. Tu, L. Zheng, T. Zhou, X. Li, S. H. Ahmed, and D. Jiang, ''Multi-modality sentiment analysis in social Internet of Things based on hierarchical attentions and CSATTCN with MBM network,'' *IEEE Internet Things J.*, early access, Aug. 10, 2021, doi: 10.1109/JIOT.2020.3015381.

[8] D. Jiang, G. Tu, D. Jin, K. Wu, C. Liu, L. Zheng, and T. Zhou, ''A hybrid intelligent model for acute hypotensive episode prediction with large-scaledata,'' *Inf. Sci.*, vol. 546, pp. 787–802, Feb. 2021.

[9] F. Louzada and A. Ara, ''Bagging k-dependence probabilistic networks: An alternative powerful fraud detection tool,'' *Expert Syst. Appl.*, vol. 39, no. 14, pp. 11583–11592, Oct. 2012.

[10] M. Carminati, R. Caron, F. Maggi, I. Epifani, and S. Zanero, ''BankSealer: A decision support system for online banking fraud analysis and investi- gation,'' *Comput. Secur.*, vol. 53, pp. 175–186, Sep. 2015.

[11] K. Fu, D. Cheng, Y. Tu, and L. Zhang, ''Credit card fraud detection using convolutional neural networks,'' in *Proc. Int. Conf. Neural Inf. Process.*, Oct. 2016, pp. 483–490.

[12] B. Tu, D. He, Y. Shang, C. Zhou, and W. Li, ''Deep feature representa- tion for anti-fraud system,'' *J. Vis. Commun. Image Represent.*, vol. 59, pp. 253–256, Feb. 2019.

[13] M. E. Greiner and H. Wang, ''Building consumer-to-consumer trust in E- Finance marketplaces: An empirical analysis,'' *Int. J. Electron. Commerce*,vol. 15, no. 2, pp. 105–136, Dec. 2010.

BIOGRAPHIES

MOHAN KUMAR M is an undergraduate student, department of computer science and engineering in paavai college of engineering.

PAVAN KUMAR K is an undergraduate student, department of computer science and engineering in paavai college of engineering.

KALAIYARASAN M is an undergraduate student, department of computer science and engineering in paavai college of engineering.

Mrs. BRINDA B.M is an Assistant Professor, department of computer science and engineering in paavai college of engineering.