Delinquency flagging and probability of default of loan

by Raghav Pandit

Submission date: 10-May-2021 08:38AM (UTC+0530)
Submission ID: 1582293412
File name: Schwalt-literature_report_MiniProject.docx (694.73K)
Word count: 2734
Character count: 14655

Delinquency flagging and probability of default of loan

Raghav V Pandit
Computer Science & Engineering
PES University
Bangalore, India
raghavvpandit81@gmail.com

Mahin Mohan
Computer Science & Engineering
PES University
Bangalore, India
mahinmohan27@gmail.com

VR Badri Prasad
Computer Science &Engineering
PES University
Bangalore, India
badriprasad@pes.edu

***Abstract*— The key scheme of this project is to flag fraudulent transactions and to calculate the probability of default of an individual. Calculation, flagging and gaining insights on the numerous factors using machine learning techniques like neural networks and statistical models like logistic regression that are a cause, from quite obvious factors like geo-location, transaction amount, purchase specifications, along with others, in case of fraud flagging and transaction history, credit score, bank balance and others in case of calculation of probability of default.**

***Index Terms*—Delinquency, Fraud Detection, Insolvency, Neural Network, Logistic regression**

I. Introduction

In the modern world, we notice a lot of credit card fraudulent activities all around the world and banks spend a tremendous amount of time and money in identifying such crimes. "Surveys conducted recently all over the world point out that cyber security risk is a major concern among market participants and it ranks first in the DTCC Systemic Risk Barometer and second in the 2017 H2 systemic risk." (Bank of England, 2017). In the Nilson report, it shines light on the technology disruption in the plethora of payment gateways specially in credit card fraud which amounts to a whopping $30 Billion worldwide in 2020. Due to advancement in tech the number of transactions that take place has increased exponentially in recent years. As the defenses have been updated with the modernization of tech so has the tech of fraudsters, who are dynamically trying to be innovative.

On the other hand banks also need to deal with providing loans by checking the client's/customer's credibility to avoid the insolvency problem which is based on multiple criteria, the most important amongst them being the probability of default. Probability of default (PD) is essentially the credit risks in the finance world. It gives an estimate of the likelihood that a banks client who has taken a loan will not be able to meet its debt.

Default as a term basically implies that an account holder who has paid an obligation.

Banks could benefit from a machine learning-based or statistical-based model for fraud detection and reducing the insolvency problem. This would mean the model could be trained to detect fraudulent transactions and calculate probability of default of a client within more than one type of transaction and application, simultaneously.

In this day and age of rising cybercrimes it is of paramount importance that the banks and other financial companies invest and spend resources in protecting the customer's money from being misplaced and misused, and protect their business by evaluating the customer's credibility. That's where fraud detection and PD calculation models play a key role in every banking company to eradicate the above stated problems.

II. Review of Literature

In the paper by Apapan Pumsirirat(et al., 2018) it highlights the fact that investigators of fraud, banks and electronic gateway systems such as Google Pay and other online transaction gateways have a systematic and sophisticated fraud detection system to prevent fraud activities that change dynamically and at a very fast pace. Using this information, the lesson is to bring about stand out patterns of fraudulent behavior that have undergone significant modifications with respect to its previous models. It goes on to highlight the types of Fraud i.e. anomaly detection and misuse detection. Anomaly detection techniques identifies the factors that stand out in a typical fraudulent transaction and is used to train the models to detect these novel frauds. Misuse fraud detection system uses data history that is stored it the database if the financial institutions and trains the models to detect whether it's a fraudulent one or not which is what we have implemented in our project.

The experiment studied in this paper brings together three datasets to study and understand the performance of the binary classifiers.

The Authors of this paper give a substantial explanation of deep learning as, “Deep learning is the state-of-the-art technology that recently attracted the IT circle’s considerable attention. The deep learning principle is an Artificial neural network that has many hidden layers. Conversely, non-deep learning feed forward neural networks have only a single hidden layer” (Apapan Pumsirirat et al., 2018)

Before the case taken up here, on the study of AE and RBM, this paper also provides the various advantages and disadvantages of using different models for fraud detection. Refer TABLE 1.

TABLE I. COMPARISON FRAUD DETECTION TECHNIQUE

Fraud Detection Techniques	Advantage	Disadvantage
K-nearest Neighbor Algorithm	KNN method can be used to determine anomalies in the target instance and is easy to implement.	KNN method is suitable for detecting frauds with the limitations of memory.
Hidden Markov Model (HMM)	HMM can detect the fraudulent activity at the time of the transaction.	HMM cannot detect fraud with a few transactions.
Neural Network	Neural networks have learned the previous behavior and can detect real-time credit card frauds.	Neural networks have many sub-techniques. So, if they pick-up this which is not suitable for credit card fraud detection, the performance of the method will decline.
Decision Tree	Decision Tree can handle non-linear credit card transaction as well.	Decision Tree have many type of input feature, DT can be constructed using different induction algorithm like ID3, C4.5 and CART. So, the cons are how to bring up induction algorithm to detect fraud as well. DT cannot detect fraud at the real time of transaction.

In another study by Charles Kwofie(et al., 2019) probability of default (PD) in industries is the focusand how it is computed utilizing history of transaction market based data which outlines ease for pecuniary inspection. The paper goes on to suggest many ML and statistical models that provide aid to thoroughly analyze credit risks like the probability of default, migration risk and many more. Each of these models is trivial for calculating and estimating credit risk of not only the industry but also the individual, however, the key model to be focused on is PD, i.e., employed in this paper. Even though this is PD of a firm it is interesting that the idea in these papers can be correlated to calculating PD of an individual too. Like how the evaluation of probability of default of a firm is the initial step while surveying the credit exposure and potential misfortunes faced by a firm so is evaluation of a person’s transaction history, credit score, and other key factors to judge the credibility of that individual

In another study by Antoine Bouveret(2019)The top grossing cyber issues faced by the financial firms is discussed. This paper showcases an in-depth documentation of cybercrime around the world in financial institutions by analyzing the various cyber incidents like data breaches, fraud and business disruption, etc.

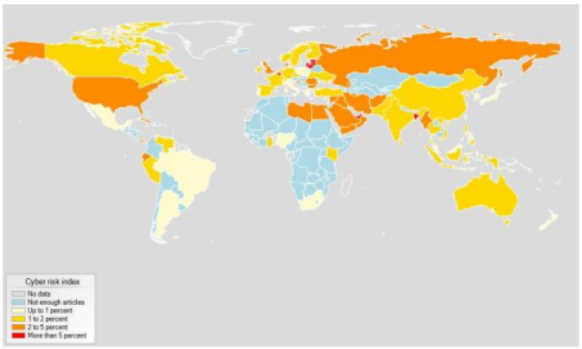
Interesting points regarding data on cyber incidents being scarce and only few quantitative analyses of cyber risk being made is emphasized. Further the Antoine Bouveret goes on to talk about how the data on cyber risk is notoriously sparse and scarce therefore firms are having very little incentive and resources to report them as they have very little incentive to keep a record.

This paper provides a framework to assess cyber risk for financial institutions in addition to a qualitative and quantitative overview giving us a glimpse of how vast the problem is and a scalable measure to evaluate the company’s security with regard to these cyber risks.

On the other hand, it goes on to show that losses faced by these various firms due to cyber-attacks are frequently independent and also has a low frequency and a high impact, which is tagged as a blackout scenario. “The firms are impacted by cyber-attacks through the three main criteria of information security, that is confidentiality, integrity and availability” (Antoine Bouveret, 2019). It is proven that after the cyber-attack, the loss of risk of confidence could be high for financial sectors

The International Telecommunication Unit (ITU), an agency of the United Nations, provides a global cyber security index for the world.

Figure 1: Measure of cyber risk for banks



"Among financial institutions, it's shown that banks account for the bulk of the attacks (91 percent of the attacks), followed by insurance companies (7 percent), among banks, retail banking activities (39 percent of the total) and credit cards services (25 percent) were the main business lines targeted." (ITU, 2019)

The paper highlights various techniques and patterns for the cyber-crimes, for instance text-mining techniques which provides for each case and background information in text form.

"Risk, as a term is defined as a combination of consequences and likelihood in which, likelihood is a function of threat levels a company faces and the ease of exploitation of existing vulnerabilities present" (ISO, 2011).

The repercussions dealt by the firms after cyber-attacks are staggering because these days financial activities, i.e. transactions are highly dependent on online and non-cash payment methods, using different payment gateways.

The finance industry has become such a huge target as it has a large interconnected network and a target sight for fraudsters.

There has been an entire section dedicated to fraud in this paper. Fraud has been explained as Cyber-attacks that can be used for delinquent purposes, as proven recently by theft using SWIFT. [7, Figure 2].

Access to confidential information, including the clients' transaction credentials are being used by cyber-criminals. In the dataset presented in the paper [7], cyber-crime related delinquencies, accounts for about 90 percent of reported losses of a firm. Average losses are around $66 million due to loss of data, with a median at $4.7 million. [7, Figure 2]

Figure 2: Recent cyber attacks using SWIFT

Over the last three years, at least ten attacks were based on the SWIFT system— a messaging system used by financial institutions for financial transactions. Hackers accessed the victims' SWIFT credentials and sent fraudulent payment orders on behalf of the target (EM banks) to the hackers' bank accounts—in some cases transiting through AE banks and central banks. Initial losses amounted to USD 336 Million, while actual losses were around USD 87 Million, as some orders were frozen and some money was recouped.

Table 3: Cyber-attacks using SWIFT

Institutions	Date	Initial losses (USD million)	Current estimated losses* (USD million)
Banco del Austro (Ecuador)	Jan. 2015	12.2	9.4
Bangladesh Central Bank	Feb. 2016	81	66
Union Bank of India	Jul. 2016	171	0
TP Bank (Vietnam)	May 2016	1	0
Akbank (Turkey)	Dec. 2016	4	4
Far Eastern International Bank (Taiwan, Province of China)	Oct. 2017	60	0.5
NIC Asia Bank (Nepal)	Oct. 2017	4.4	0.6
Globex (Russia)	Dec. 2017	1	0.1
Unidentified bank (Russia)	Dec. 2017	Unknown	6
City Union Bank (India)	Jan. 2018	2	Unknown

Sources: ORX News, Financial Times. * Current estimated losses are based on publicly available information. Targeted institutions are in the process of recovering the losses through legal proceedings.

The appalling scenario in which the occurrence of such events was twice the peak observed in the year 2013, It is seen that average losses of such companies would amount to $268 billion (26 percent of net income) and risk indicators would vary between $352 and 539 billion (34 to 52 percent of the net income). [7, Figure 2]

III. Dataset

There are 2 datasets each for one functionality. Both consists of enumerable indicators.

The fraud detection dataset encompasses factors credit card number, transaction geo location, latitude and longitude of both the client and merchant and city to list a few. The data provides a view on the various transactions that has fared over the span of almost 2 years, it aims to provide clarity on the transaction details of the different customers of a bank. This is simulated credit card transaction dataset containing legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants.[2]

Data mining is a crucial course of action in discovery of knowledge that involves theories, tools and procedures for unveiling the schemes in the given dataset. It is of the essence

to understand the principle behind the methods so that there is a pertinent fit of the tools and techniques used with the data and the objective of pattern recognition.

On receiving a loan application from a client, based on the profiling report of the applicant, the bank has to gauge the credibility risk the client adds to the firm.

"Two types of risks are associated with the bank's decision If the applicant is a good credit risk, i.e., is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank and If the applicant is a bad credit risk, i.e., is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank." (Antoine Bouveret, 2019)

Both the datasets used were cleared of not having Null values, cleaning in terms of replacing the ' ' with '_' was done and few of the columns that were considered were dropped using the results obtained by doing PCA.

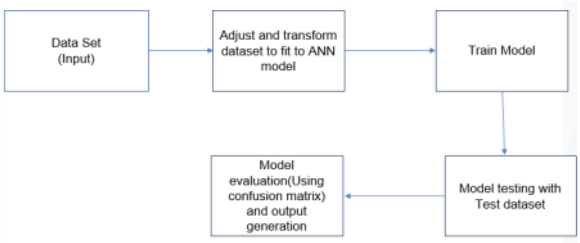
IV. Problem Statement and Approach

On initial exploration into the financial sector and the major problems faced by them, a glaring domain of delinquency and insolvency was brought to notice, specifically the amount of money that goes into financial security of a financial firm was shocking, in which Fraudulent transaction flagging and calculation of probability of default of loan of an individual was intriguing.

As there are 2 functionalities 2 approaches, suitable to that particular problem statement was chosen.

Artificial Neural Network was used in flagging of delinquent transactions in reference to Apapan Pumsirirat and Liu Yan's paper where they compare other machine learning models and found that neural networks provided a better accuracy. Columns like gender, merchant and other columns were dropped as they are not as significant as factors like geolocation and amount transacted and so on. Keras, a python library was used to build the ANN framework. Encoding the Categorical data to numeric data using LabelEncoder from sklearn, Feature scaling using Standard scaler using sklearn Training the ANN model using adam as an optimizer and binary crossentropy as the last layer (output layer) loss function, with batch size 30 and 50 epochs, Testing and calculation of accuracy of model using accuracy_score() then predict the fraudulency of a transaction and evaluate the performance using a confusion matrix.

Figure 3



Logistic regression was used as an approach to solve the insolvency problem with respect to calculation of probability of default. The required parameters are then applied to the logistic regression model estimator implemented from the class sklearn.linear_model. The obtained values of the intercept and coefficients are fed to the function which computes the value of Z in the formula Pr(default = 1/X) = 1 / 1 + exp(-Z) where Z = w0 + w1. LimitBalance + w2.Age (where w0,w1 and w2 are constants obtained from the logistic regression model). This value of Z is then fed to the function that computes the probability default for any input value of the parameters of previous credit amount and the age. This also provides necessary statistical visual plots along with the probability of default which are convenient for banks to utilize in order to determine the customer's credibility.

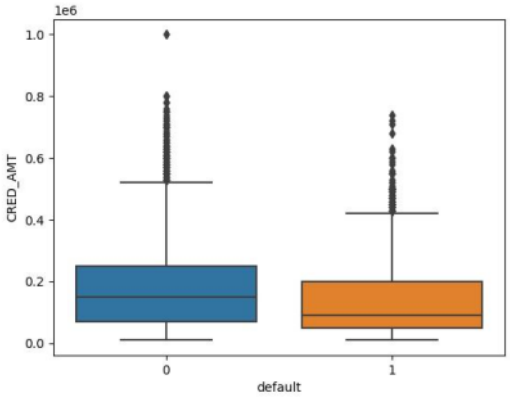
Figure 4

.

V.RESULT & FUTURE SCOPE

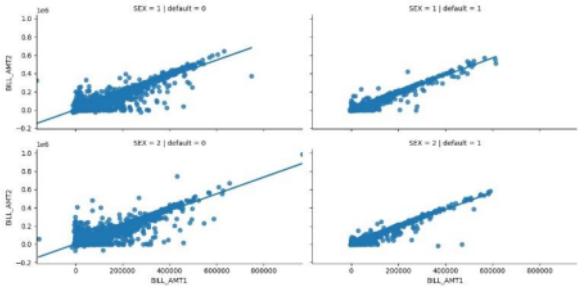
We conclusively evaluated the probability of default of an individual and appended the prediction of the neural network model for the possible fraudulent transactions. Major factors that effected the probability of default were identified and visually represented using the graphs represented below

Figure5



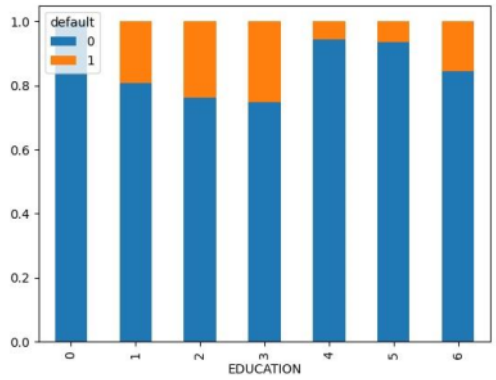
The box plot (in Figure 5) depicts the fraction of customers with a similar credit amount who have a default close 1 and the rest who have a default close to 0.

Figure 6



The conditional plot (in Figure 6) indicates the variation of bill_amt1 which is the the bill amount of the first month and the bill-amt2 corresponding to the second month for males/females according to their respective defaults.

Figure 7



The percentage breakdown of the different types of students (in Figure 7) according to their default values is shown in the above figure

With respect to fraudulent flagging its conclusively proven in "*Credit Card Fraud Detection Using Deep Learning"* by Apapan Pumsirirat, et al that factors of fraud transaction is dynamic in nature so the Artificial Neural Network model handles the allotment of weights to the different criteria based on the the past transactions fed into the model.

The models prepared has a scope to perform live fraud tracking and send an instant notification to the client directly if a transaction is flagged as delinquent and to also store all this data into the banks database to intern train and improve the model for increased accuracy. We hope to extend the probability of default model to be able to handle large data and extend it to be integrated into the banking apps so that not only can the bank, but the bank's clients can themselves can check how financially fit they are to avail any loan from that particular bank.

ACKNOWLEDGMENT

We would like to express out profound gratitude to Dr. Prof.V R Badri Prasad for encouraging us with this opportunity and guiding us along the way. We would also like to thank the Computer Science and Engineering department at PES University, for always inspiring us to venture into various domains and get hands on experience.

REFERENCES

[1] UCI Machine Learning Repository. (2017, Nov. 29). *Stalog (German credit data) dataset* [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

[2] Kaggle. (2020, September). *Credit Card Transactions Fraud Detection Dataset* [Online]. Available: https://www.kaggle.com/kartik2112/fraud-detection?select=fraudTest.csv

[3] Single Hidden Layer Neural Network [Online]. Available: http://nicolamanzini.com/single-hidden-layer-neural-network/

[4] Charles Kwofie, 1Caleb Owusu-Ansah and 2Caleb Boadi Predicting the Probability of Loan-Default: An Application of Binary Logistic Regression 1Department of Statistics, University of Ghana, P.O. Box LG 25, 2 School of Business, University of Ghana, P.O. Box LG 25, Legon-Accra, Ghana- 2019

[5] Amir Ahmad Dar1* , N. Anuradha2 and Shahid Qadir3 Estimation of pd of diffferent firms by Journal of Global Enterpreneurship Research- 2019

[6] Apapan Pumsirirat, Liu Yan,School of Software Engineering, Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine, Tongji University, Shanghai, China (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 1, 2018

[7] Cyber Risk for the Financial Sector: A Framework for Quantitative Assessment, Antoine Bouveret-ResearchGate-2019

[8] Financial Stability report, Bank of England, 2017

Delinquency flagging and probability of default of loan

ORIGINALITY REPORT

11%	8%	6%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	3%
2	www.imf.org Internet Source	3%
3	www.kaggle.com Internet Source	1%
4	Shivangi Gupta, Greeshma Karanth, Niharika Pentapati, V R Badri Prasad. "A Web Based Framework for Liver Disease Diagnosis using Combined Machine Learning Models", 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020 Publication	1%
5	pdfs.semanticscholar.org Internet Source	1%
6	Amir Ahmad Dar, N. Anuradha, Shahid Qadir. "Estimating probabilities of default of different firms and the statistical tests", Journal of Global Entrepreneurship Research, 2019 Publication	<1%

7	Apapan Pumsirirat, Liu Yan. "Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine", International Journal of Advanced Computer Science and Applications, 2018 Publication	<1 %
8	"Intelligence in Big Data Technologies—Beyond the Hype", Springer Science and Business Media LLC, 2021 Publication	<1 %
9	online.stat.psu.edu Internet Source	<1 %
10	Wisdom Akpalu, Mintewab Bezabih. "Tenure Insecurity, Climate Variability and Renting out Decisions among Female Small-Holder Farmers in Ethiopia", Sustainability, 2015 Publication	<1 %

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		