Realtime Credit Card Fraud Detectin Using Machine Learning

Abstract Credit card fraud events take place frequently and then result in huge financial lsses [1]. The number f nline transactins has grwn in large quantities and nline credit card transactins hlds a huge share f these transactins. Therefre, banks and financial institutins ffer credit card fraud detectin applicatins much value and demand. Fraudulent transactins can ccur in varius ways and can be put int different categries. This paper fcuses n fur main fraud ccasins in realwrld transactins. Each fraud is addressed using a series f machine learning mdels and the best methd is selected via an evaluatin. This evaluatin prvides a cmprehensive guide t selecting an ptimal algrithm with respect t the type f the frauds and we illustrate the evaluatin with an apprpriate perfrmance measure. Anther majr key area that we address in ur prject is realtime credit card fraud detectin. Fr this, we take the use f predictive analytics dne by the implemented machine learning mdels and an API mdule t decide if a particular transactin is genuine r fraudulent. We als assess a nvel strategy that effectively addresses the skewed distributin f data. The data used in ur experiments cme frm a financial institutin accrding t a cnfidential disclsure agreement.

Keywrds credit card frauds, fraud detectin system, fraud detectin, cnfidential disclsure agreement, realtime credit card fraud detectin, skewed distributin.

I. INTRDUCTIN

Fraud has been increasing drastically with the prgressin f statefart technlgy and wrldwide cmmunicatin. [5] Fraud can be avided in tw main ways: preventin and detectin. Preventin avids any attacks frm fraudsters by acting as a layer f prtectin. Detectin happens nce the preventin has already failed. Therefre, detectin helps in identifying and alerting as sn as a fraudulent transactin is being triggered. Recently, cardntpresent transactins [6] in credit card peratins have becme ppular amng web payment gateways. Accrding t the Nilsn Reprt in ctber 2016, mre than $31 trillin were generated wrldwide by nline payment systems in 2015, increasing 7.3% than 2014. Wrldwide lsses frm credit card fraud have been rising t $21 billin in 2015, and will pssibly reach $31 billin by 2020. [3] Hwever, there has been an extreme increase in fraudulent transactins that affect the ecnmy dramatically. Credit card fraud can be classified int several categries. The tw types f frauds that can be mainly identified in a set f transactins are Cardntpresent (CNP) frauds and Cardpresent (CP) frauds. Thse tw types can be described further by bankruptcy fraud, theft/cunterfeit fraud, applicatin fraud, and behaviural fraud. ur study aims at addressing fur

fraud natures that belng t the CNP fraud categry described abve and we prpse a methd t detect thse frauds real time.

Machine learning is this generatins slutin which replaces such methdlgies and can wrk n large datasets which is nt easily pssible fr human beings. Machine learning techniques fall int tw main categries; supervised learning and unsupervised learning. Fraud detectin can be dne in either way and nly can be decided when t use accrding t the dataset. Supervised learning requires prir classificatin t anmalies. During the last few years, several supervised algrithms have been used in detecting credit card fraud.

The data which is being used in this study is analyzed in tw main ways: as categrical data and as numerical data. The dataset riginally cmes with categrical data. The raw data can be prepared by data cleaning and ther basic preprcessing techniques. First, categrical data can be transfrmed int numerical data and then apprpriate techniques are applied t d the evaluatin. Secndly, categrical data is used in the machine learning techniques t find the ptimal algrithm.

This paper cnsists f selecting ptimal algrithms fr the fur fraud patterns thrugh an extensive cmparisn f machine learning techniques via an effective perfrmance measure fr the detectin f fraudulent credit card transactins.

The rest f this paper is presented as fllws. Sectin 2 presents the literature review. Sectin 3 prvides the experimental methdlgy including results. Finally, cnclusins and discussins f the paper are presented in Sectin 4.

II. LITERATURE REVIEW

In earlier studies, many appraches have been prpsed t bring slutins t detect fraud frm supervised appraches, unsupervised appraches t hybrid nes; which makes it a must t learn the technlgies assciated in credit card frauds detectin and t have a clear understanding f the types f credit card fraud. As time prgressed fraud patterns evlved intrducing new frms f fraud making it a keen area f interest fr researchers. The remainder f this sectin describes single machine learning algrithms, machine learning mdels and fraud detectin systems that were used in fraud detectin. The prblems that came acrss the review have analyzed fr the

later use f implementing an efficient machine learning mdel.

With the analysis f varius detectin mdels, past researchers have fund many prblems regarding fraud detectin. In [14] and [3] they have mentined Lack f reallife data as a huge issue. Real life data are lacking because f the data sensitivity and privacy issues. Papers [3] and [7] have studied Imbalance data r skewed distributin f data. The reasn behind this is having quite a less amunt f frauds when cmpared t nnfrauds in the transactin datasets. Paper [3] states that data mining techniques take time t execute when dealing with big data. verlapping f data is anther majr drawback in preparatin f credit card transactin data. Accrding t paper [2] and [7] the issue ccurs due t sme scenaris when the legitimate transactins lk exactly like fraudulent transactins. In anther way, fraudulent transactins may appear as legitimate transactins. Als, they have cme acrss the difficulty in dealing with categrical data. When cnsidering the credit card transactin data, mst f the features have categrical values. In this case, almst all the machine learning algrithms d nt supprt the categrical values. In [3][4] they have mentined chice f detectin algrithms and feature selectin as a challenge in detecting frauds since mst f the machine learning algrithms take much time fr training purpses than predicting. Anther key issue that affects financial fraud detectin is the feature selectin. It aims t filter ut the attributes that mst describes the aspects f fraud detectin and its characters. In paper [7] they have highlighted fraud detectin cst and lack f adaptability as challenges in the fraud detectin prcess. When cnsidering a system, the cst f fraudulent behaviur and the preventin cst shuld be taken int cnsideratin. Lack f adaptability ccurs when the algrithm is expsed t new types f fraud patterns and nrmal transactins. Effectiveness can change accrding t the prblem definitin and its specificatins, s having a gd understanding f the perfrmance measure is necessary [4].

There are different kinds f mdels implemented fr credit card fraud detectins. In thse mdels, different algrithms have been used.

Adapting the fraud detectin system t newly intrduced frauds can be prblematic whether t retrain the machine learning mdel due t drastic changes in the fraud patterns, als may be cstly and risky. Fr instance, Tyler et al. extended a framewrk prpsed in [12], implemented the mdel and the mdel was applied t a realwrld transactin lg. T address the classificatin prblem Lgistic Regressin (LR) has been used. The instances f fraudulent transactins have been discretized int strategies by using Gaussian Mixture Mdels (GMMs). Here synthetic minrity versampling technique was used t address the class imbalance. T stand ut the significance f estimates in ecnmic value sensitivity analysis has been used. The results have prven that a practical methd which uses minimal steps t retrain a mdel culd functin as same as a classifier that typically retrains every rund [13].

There is anther mdel called RiskBased Ensemble (RBE) that can handle the data cnsisting f issues and give utstanding results. Fr handling imbalanced data, a highly efficient bagging mdel has been used. T handle the implicit nise in the transactin dataset they have used Naive Bayes algrithm [9]. Peter et al. evaluated several deep learning algrithms with respect t their efficacy. The fur tplgies are Recurrent Neural Netwrks (RNNs), Gated Recurrent Units (GRUs), Lng Shrtterm Memry (LSTMs), and Artificial Neural Netwrks (ANNs). In their prject in additin t data cleaning and ther data preparatin steps, they have vercme class imbalance and scalability prblems by using undersampling. T discver which hyperparameters had the highest influence n the perfrmance f the mdel, the sensitivity analysis was carried ut. They have discvered that the perfrmance f the mdel was affected by the size f the netwrk. They cncluded that larger the netwrk it shwed better perfrmance. [11]

Credit card data have the issue f skewed distributin which is als knwn as the class imbalance. Accrding t Andrea et al., their prject addresses class imbalance including ther issues such as cncept drift and verificatin latency. They have als illustrated the mst relevant perfrmance matrix that can be used in credit card fraud detectin. The achievement f the research als includes a frmal mdel and a pwerful learning strategy fr addressing the verificatin latency and an alert and feedback mechanism. Accrding t experiments they have declared the precisin f the alerts as the mst imprtant measure [15].

Chee et al. used twelve standard mdels and hybrid methds which use AdaBst and majrity vting methds t achieve better accuracy rates in credit card fraud detectin [16]. They were evaluated using bth benchmark and realwrld data. A summary f the strengths and limitatins f the methds were evaluated. The Matthews Crrelatin Cefficient metric (MCC) has been taken as the perfrmance measure. T evaluate the rbustness f the algrithms nise was added t the data. Als, they have prved that the majrity vting methd was nt affected by the added nise.

The analysis carried ut n highly imbalanced data in paper [17] shw that KNN shws utstanding perfrmance fr sensitivity, specificity and MCC, except fr accuracy. The paper [18] discussed cmmnly used supervised techniques and they have prvided a thrugh evaluatin f supervised learning techniques. Als, they have shwn that all algrithms change accrding t the prblem area.

Fraud detectin system presented in paper [19] is built t handle class imbalance, the frmatin f labelled and unlabeled, and prcessing f large datasets. The prpsed system was able t vercme all the challenges.

III. EXPERIMENTAL METHDLGY

A. Data descriptin

The dataset was created cmbining tw data surces; the fraud transactins lg file and all transactins lg file. The fraud transactins lg file hlds all the nline credit card fraud ccurrences while all transactins lg file hlds

all transactins stred by the crrespnding bank within a specified time perid. Due t the cnfidential disclsure agreement made between the bank and the authrs f the paper, sme f the sensitive attributes such as card number were hashed. When evaluating the cmbined dataset, the shape f the data was much skewed due t the imbalanced numbers f legitimate transactins and fraudulent ccurrences. The file with the fraud cases had 200 recrds while the transactin lg file had 917781 recrds. Attributes f the tw data surces are as fllws.

Cllected raw data were first divided int 4 data sets accrding t its fraud pattern. This prcess was dne with the infrmatin gained by the bank.