Personal Loan Fraud Detection Based on Hybrid Supervised and Unsupervised Learning

[1]Prof. Usha B. Nandwani, [2]Miss. Archana Deshmukh, [3]Miss. Deepti Gangurde, [4]Mr. Anuj Satavase

[1]Asst.Professor ,[1234] UG Student, [1234]Computer Engg. Dept. Shivajirao S.Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.

[1]ushanandwani@gmail.com,[2]archu.deskhmukh789@gmail.com,[3]gangurdedeepti@gmail.com, [4]anujsatavase1998@gmail.com.

Abstract- In recent years, have been witnessing a dramatic increase on the personal loan for consumption, due to the rapid development of e-services, including e-commerce, e-finance and mobile payments. Resulting from a lack of effective grid verification and supervision, will inevitably leads to largescale losses caused by credit loan fraud . Considering how the difficulty of manual inspection and the verification on the large amount of credit card exchanges, machine learning methods are to be commonly used to detect fraudulent exchanges automatically. In order to filter the useless information and to preserve the useful information without knowing the meaning of the data, this paper combines Kernel Principal Component Analysis (Kernel PCA) together with XGBoost(algorithm) and proposes a new hybrid supervised & unsupervised learning model, KPXGBoost. There use network search to abstain from over-fitting and look at the exhibition of both XGBoost and P-XGBoost and other traditional AI techniques . It turns out that P-XGBoost totally outperforms the XGBoost in case of fraud detection, which provides a totally new perspective privacy to detecting the fraud behaviour while protecting the clients.

Keywords : supervised learning, unsupervised learning, Extreme Gradient Boosting, principal component analysis .

I. INTRODUCTION

Fraud refers to the misuse of profitable organization’s system without necessarily leading to direct lawful concern. Fraud is an universal act in order to deceive another person or organization for financial benefits. The fraud committed by individuals exterior for the organization is called as customer fraud or external fraud where when a fraud is committed by top-level management is known as management fraud or internal fraud. Credit card fraud is an unauthorized account activity by a person for which the account is not proposed. It is additionally characterized as when an individual takes another individual Visa for individual reasons while the owner of the card and the card backer don't know about the way that the card being utilized. The people utilizing the card has not in the least having the sorting out with the card holder or the guarantor has no target of making the reimbursements for the buy they done. Information mining alludes to concentrate or mining information from huge measure of information. Information mining related with (a)supervised learning dependent on preparing information of known misrepresentation and real cases and (b)unsupervised learning with information that are not to be extortion or legitimate. The improvement of new blackmail revelation techniques are made more troublesome because of the limit of the exchanging of contemplations in deception distinguishing proof. The best financially savvy choice is coax out potential ideas of misrepresentation from the accessible information utilizing numerical logical calculations. In present outcome, carrying out viable misrepresentation anticipation strategies from the start spot and identification procedure if there should arise an occurrence of disappointment of preventive measures is not any more an upper hand yet .[4],[7],[11]

II. AIMS AND OBJECTIVE

a) Aim

The Paper aim is to plan and build up an extortion discovery strategy for Streaming Transaction Data, and , to examine the past exchange subtleties of the clients and concentrate the personal conduct standards. This for cardholders who are bunched into various gatherings dependent on their exchange sum. These is best savvy technique is to coax out potential ideas of misrepresentation from the past accessible information utilizing the numerical logical calculations. Later various classifiers are prepared absurd independently

b) Objective

The objective is to Detect misrepresentation consequently. Permitting the Streaming and the capacity for distinguishing on the web misrepresentation continuously .Hence Less time required for confirmation strategies likewise permitting Identifying covered up relationshipsin information.

III. LITERATURE SURVEY

The literature survey deals with the topics and the researches that would help to understand the existing systems that are similar to this paper. The objective of this literature survey is to analyze the related work to this paper and mechanisms used in previous studies.

Paper 1: Detecting Fraud, Corruption, and Collusion in International Development Contracts :

This paper depicts a proof-of-idea of a totally machine-controlled extortion, defilement, and agreement association for recognizing hazard in worldwide advancement contracts.We built up this technique related to the planet Bank bunch - the greatest global improvement bank - to support the time and value power of their examination strategy. , side by side of by input exploitation verifiable monetary honor information and past examination results, our classifier relegates a "hazard score" to UN organization contracts. during this paper, there's blessing a proof-of-idea for a totally machine-controlled extortion, defilement, and intrigue association for global advancement contracts , while past work has fixated absolutely on discovery charge card misrepresentation, this work is extra expansive, focused on separating not exclusively extortion anyway furthermore debasement and agreement, that square measure normally more strong to find. this technique, side by side of by contribution from rehearsed UN office examiners, joins along 20 years of UN office examination.

Paper 2: Semi-Supervised Anti-Fraud Models for Cash Pre-Loan in Internet Consumer Finance :

The research provides insights into the current status of the microfinance industry in the digital economy, describes the core challenges of online microlending in the area of assessing creditworthiness, reviews options and opportunities of building user type styles in anti-fraud scoring in the microloan system. The banking sector takes various measures to reduce credit risks: fewer loans per one borrower, use of collateral as loan security, involving guarantors of loan repayment, loans insurance, etc., using borrower credit rating tools. Moreover, in general terms, the business process of bank consumer lending embraces the following phases: receiving an application, interviewing a potential borrower, studying creditworthiness through various sources and risk assessment, drawing up a contract, making a resolution on loan issue, performance of the loan agreement.[12]

Paper 3: Social Media Analytics for Better Detection of Fraudulent Applications for Online Microfinance Loans:

This exploratory study aims to address the problem that cash loan fraud customers are difficult to detect manually. Cash loan is a new consumption model in the concept of Internet consumer finance(ICF). Manual detection of fraudulent customers requires a lot of manpower and time, and often causes great losses to financial institutions, so our group did the research mentioned above.In this paper, it is proposed a Semi-supervised Pre-loan Fraud Detection (SPFD) system via investigating various supervised and unsupervised learning algorithms on basis of 285,771 applicants' desensitized data from MUCFC (a Chinese ICF company. In SPFD, feature selection methods consist of KL Divergence.[10]

IV. EXISTING SYSTEM

The framework depicts the spellbinding models, that the unaided learning capacities. These capacities don't foresee target esteem, however they center more around the inborn design, relations, interconnectedness, and so forth Self-Organizing Maps a self-arranging map (SOM) a neural organization strategy however utilized unaided learning. It permits clients to imagine information from high dimensional to low dimensional. Gathering strategy for information dealing with (GMDH) an inductive learning calculation for displaying complex frameworks. GDMH is a self-sorting out approach that tests progressively muddled models and assesses them utilizing some outer rule on independent pieces of the information test . Anomaly discovery techniques (OD) a totally different from the conventional perception strategy. Anomaly strategy is utilized to recognize strange conduct in a framework utilizing an alternate system. . Affiliation rule investigation (AR) were characterized on exchange sets. Thickness based spatial bunching of uses with clamor (DBSCAN) a thickness based grouping calculation which can be utilized to sift through exceptions and find groups of discretionary shapes.

V. COMPARATIVE STUDY

Table No .1 Comparative Study

SR NO.	PAPER TITLE	AUTHOR NAME	METHOD	ADVANTAGE	DISADVANTAGE
1.	Detecting Fraud, Corruption, and Collusion in International Development Contracts	Emily Grace , Ankit Rai , Elissa Redmiles , Rayid Ghani	fully automated from data collection through pre-processing and modeling	70% achievement rate in anticipating charges that will be validated, a 84% increment from the current examination achievement rate	Time Consuming
2.	Social Media Analytics for Better Detection of Fraudulent Applications for Online Microfinance Loans	Valentina Kuznetsova, Iskandar Azhmuhamedov ,oleg Protalinskiy	social media to retrieve information for decision-making, including in the banking sector,	provides more reliable data than classic borrower data collection by credit organizations	Difficult to understand
3.	Semi-Supervised Anti-Fraud Models for Cash Pre-Loan in Internet Consumer Finance	Wanlin Sun , Ming Chen, Jie-xia Ye , Yuhang Zhang , Cheng-zhong Xu Yangqing Zhang , Yaonan Wang , Wen Wu , Peng Zhang , Feipeng Qu	supervised and unsupervised learning algorithms	Good Approach Explained	Testing and evaluation is time consuming

VI. PROBLEM STATEMENT

Different computational methods have been stated for detecting the fraud by computing various parameters for each kind of algorithm and the computing time representing with graphical view. In existing system fraud detection is done using ID3 and support vector machine algorithms and a survey stating the percent of fraud happened and defining different parameters and comparing different parameters for the algorithms. The system which i had proposed is fraud detection using supervised learning algorithms that is decision tree learning algorithm and logistic regression and XGBOOST for Machine learning.

VII. PROPOSED SYSTEM

The System proposed in these paper states about the Card exchanges are new when contrasted with past exchanges made by a client. There originality is a most inconvenient issue in authentic when are called thought drift issues . Idea floating can be said as a variable which changes over the long haul and unforeseenly. We can see the essential highlights that are caught when any exchange is made. Attribute Named 1)Transaction ID - Recognizable proof Number of an exchange ,2) Cardholder ID - Extraordinary Identification number given to the Cardholder,3)Amount - Sum moved or credited in a specific exchange by the client, 4) Time- Subtleties like time and date ,to indentify and when the exchange was made 5) Label - To indicate whether the exchange is real or false. These showed the Raw features of credit card transactions.

The Attributes of Data set of the systems include 1) Time -Time in seconds to determine the passes between the current exchange and first exchange, 2) Amount - Transaction Amount,3) Class for fraud or not Fraud .

VIII. ALGORITHM

```
Step 1: Start
Step 2: Data Preprocess ip dataset
Step 3: Use logistic regression to train model:
X = new_df.drop('Class', axis=1)
y = new_df[['Class']]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.33, random_state=44)
print('X_trainshape=>'+str(X_train.shape))
print('y_train shape =>'+ str(y_train.shape))
print('X_test shape => ' + str(X_test.shape))
print('y_test shape => ' + str(y_test.shape))
X_train shape => (659, 30)
y_train shape => (659, 1)
X_test shape => (325, 30)
y_test shape => (325, 1)
Step 4: Calculate Accuracy, Macro avg ,Weighted avg using XGBOOST
Step 5: FRAUDUALENT Report has been calculated.
Step 6: End
```

IX. MATHEMATICAL MODEL

1. XGBOOST

The algorithm XGBoost is the recently ruling the applied AI and the Kaggle rivalries for the organized and plain information. XGBoost is the execution of the slope supported choice trees intended for speed and furthermore execution.[

$$\textit{Real Value (label)Known from the training data} \\ \textit{– set}$$

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i , \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

$$\textbf{Can be seen as } \mathbf{f}(\mathbf{x} + \Delta\mathbf{x})\ \textit{where}\ x = \hat{y}_i^{(t-1)}$$

XGBoost is a choice tree-based outfit Machine Learning calculation that utilizes an angle boosting system shows the eqaution ... The calculation separates itself in the accompanying manners: A wide scope of uses: Can be utilized to settle relapse, characterization, positioning, and client characterized forecast issues.[9]

2. Logistic Regression

Logistic Regression is the supervised classification method Calculated Regression is the managed grouping strategy which returns the likelihood of the double reliant variable that is anticipated from the autonomous variable of dataset that is calculated relapse foresee the likelihood for a result which has two qualities either zero or one, yes or no and bogus or valid. The condition addresses the calculated relapse in numerical structure.Fig 1. Shows the plot of X Vs Y for regression .[1],[9]

$$p = \frac{e^{\alpha+\beta_n X}}{1 + e^{\alpha+\beta_n X}}$$

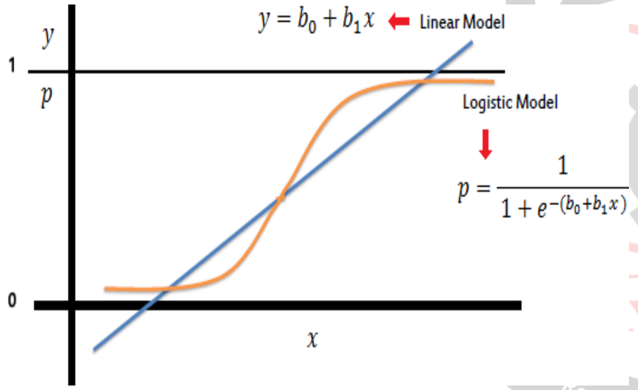


Fig 1. Plot X Vs Y

3. Decision Tree

Decision tree is the calculation that utilizes a tree like chart or model of choices and their potential results to foresee about ultimate conclusion, this calculation utilizes restrictive control articulation.

$$Entropy(S) = \sum_{i=1}^{n} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_v|}{|S|} Entropy(S_v)$$

The equations show gain is maximum or entropy is minimum after that to make a decision tree node shows fig 2. containing that attribute and lastly recursion is performed on subsets using remaining attributes .[9]

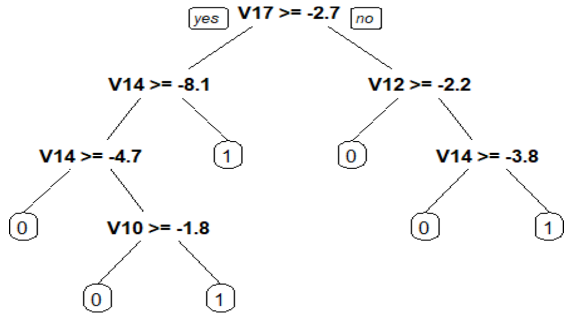


Fig.2 Decision Tree

X. SYSTEM ARCHITECTURE

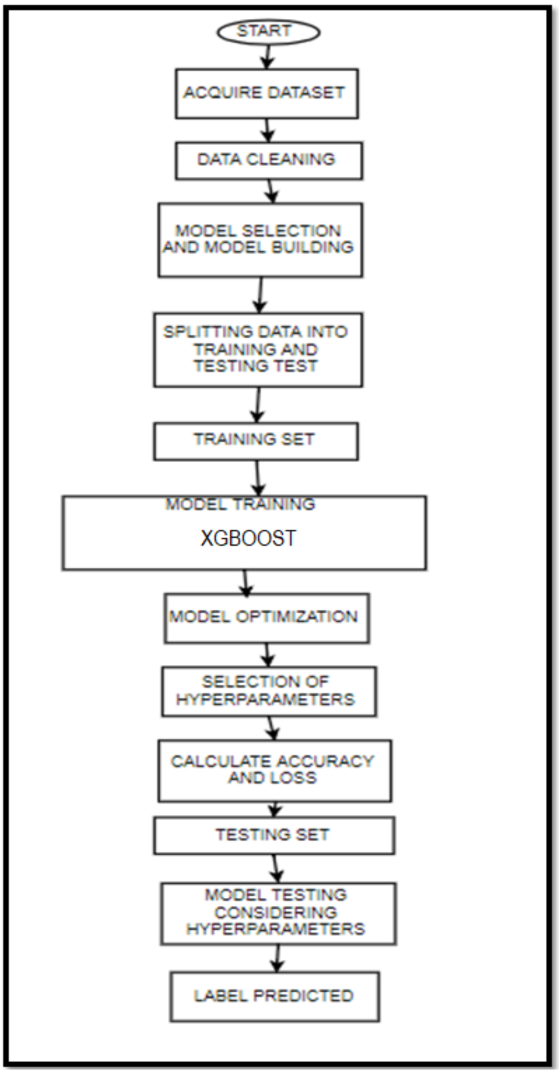

START
ACQUIRE DATASET
DATA CLEANING
MODEL SELECTION AND MODEL BUILDING
SPLITTING DATA INTO TRAINING AND TESTING TEST
TRAINING SET
MODEL TRAINING
XGBOOST
MODEL OPTIMIZATION
SELECTION OF HYPERPARAMETERS
CALCULATE ACCURACY AND LOSS
TESTING SET
MODEL TESTING CONSIDERING HYPERPARAMETERS
LABEL PREDICTED

Fig.3: System Architecture

Description: Above fig 3.shows 3 types of phases:

1. Acquiring Phase 2. Training phase 3. Testing Phase

1) Acquiring Phase : Aquire Datasets and cleaning previous data ,selecting model and building an efficient Model

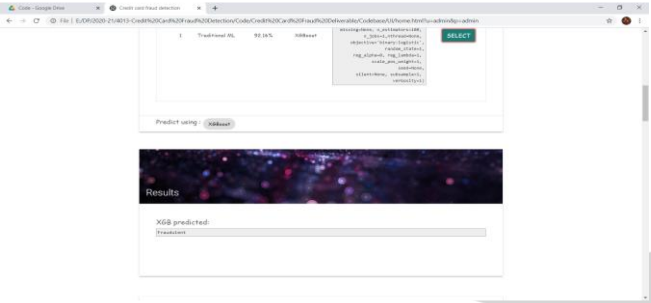
2) Training Phase: using XGBOOST model for Training

3) Testing Phase : Testing Model while considering various parameters

XI. ADVANTAGES

This model achieves 92.615 % accuracy . This Model allows Users to actively involved in the development. Since in this philosophy a functioning model of the framework is given, the clients improve comprehension of the framework being created. Fast client criticism is accessible that leads towards better arrangements. Missing usefulness can be distinguished effectively Confusing of troublesome capacities can be recognized Requirement Validation, Quick Implementation of deficient yet Functional Application.

XII. DESIGN DETAILS


Code - Google Drive
Credit card fraud detection
File | E:/DP/2020-21/4013-Credit%20Card%20Fraud%20Detection/Code/Credit%20Card%20Fraud%20Deliverable/Codebase/UI/home.html?u=admin&p=admin
SELECT
XGBoost
Results

1 Traditional ML 92.16% XGBoost

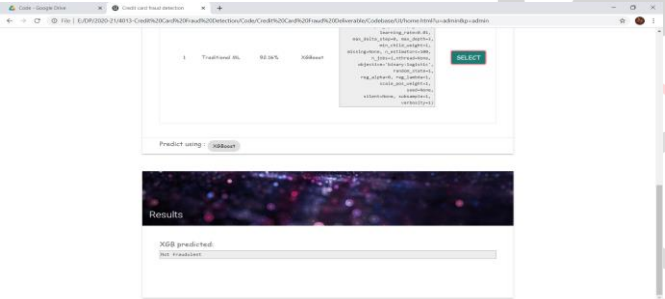
```
missing=None, n_estimators=100,
n_jobs=1, nthread=None,
objective='binary:logistic',
random_state=1,
reg_alpha=0, reg_lambda=1,
scale_pos_weight=1,
seed=None,
silent=None, subsample=1,
verbosity=1)
```

Predict using :

XGB predicted:

Fraudulent

Fig.4: Fraud data Prediction


Code - Google Drive
Credit card fraud detection
File | E:/DP/2020-21/4013-Credit%20Card%20Fraud%20Detection/Code/Credit%20Card%20Fraud%20Deliverable/Codebase/UI/home.html?u=admin&p=admin
SELECT
XGBoost
Results

1 Traditional ML 92.16% XGBoost

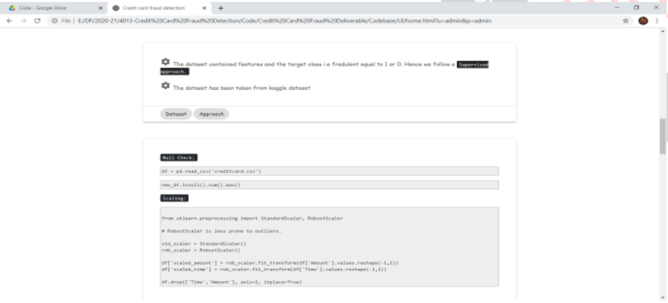
```
learning_rate=0.01,
max_delta_step=0, max_depth=3,
min_child_weight=1,
missing=None, n_estimators=100,
n_jobs=1, nthread=None,
objective='binary:logistic',
random_state=1,
reg_alpha=0, reg_lambda=1,
scale_pos_weight=1,
seed=None,
silent=None, subsample=1,
verbosity=1)
```

Predict using :

XGB predicted:

Not Fraudulent

Fig.5: No Fraud Prediction


Code - Google Drive
Credit card fraud detection
File | E:/DP/2020-21/4013-Credit%20Card%20Fraud%20Detection/Code/Credit%20Card%20Fraud%20Deliverable/Codebase/UI/home.html?u=admin&p=admin
Dataset
Approach

The dataset contained features and the target class i.e fradulent equal to 1 or 0. Hence we follow a Supervised approach.

The dataset has been taken from kaggle dataset

Null Check:

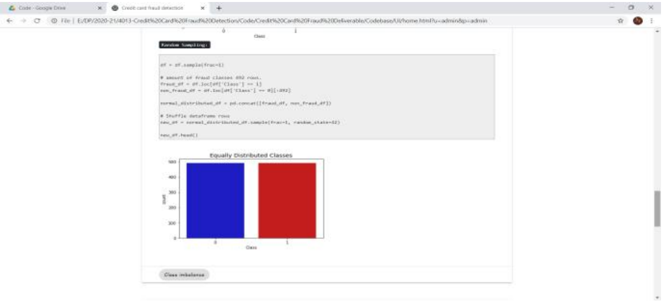
```
df = pd.read_csv('creditcard.csv')
```

```
new_df.isnull().sum().max()
```

Scaling:

```
from sklearn.preprocessing import StandardScaler, RobustScaler

# RobustScaler is less prone to outliers.

std_scaler = StandardScaler()
rob_scaler = RobustScaler()

df['scaled_amount'] = rob_scaler.fit_transform(df['Amount'].values.reshape(-1,1))
df['scaled_time'] = rob_scaler.fit_transform(df['Time'].values.reshape(-1,1))

df.drop(['Time','Amount'], axis=1, inplace=True)
```

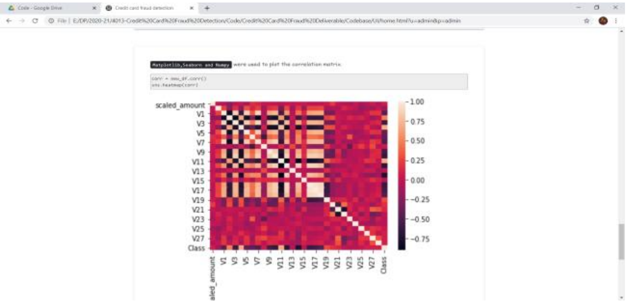
Fig.6 : Data Science


Code - Google Drive
Credit card fraud detection
File | E:/DP/2020-21/4013-Credit%20Card%20Fraud%20Detection/Code/Credit%20Card%20Fraud%20Deliverable/Codebase/UI/home.html?u=admin&p=admin
Class
Equally Distributed Classes
500
400
300
200
100
0
count
0
1
Class
Class imbalance

Random Sampling:

```
df = df.sample(frac=1)

# amount of fraud classes 492 rows.
fraud_df = df.loc[df['Class'] == 1]
non_fraud_df = df.loc[df['Class'] == 0][:492]

normal_distributed_df = pd.concat([fraud_df, non_fraud_df])

# Shuffle dataframe rows
new_df = normal_distributed_df.sample(frac=1, random_state=42)

new_df.head()
```

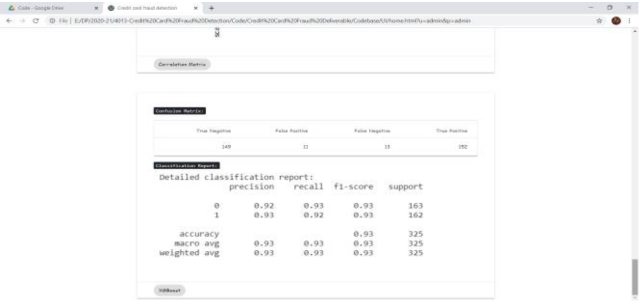
Fig.6 : Random Sampling


Code - Google Drive
Credit card fraud detection
File | E:/DP/2020-21/4013-Credit%20Card%20Fraud%20Detection/Code/Credit%20Card%20Fraud%20Deliverable/Codebase/UI/home.html?u=admin&p=admin
scaled_amount
V1
V3
V5
V7
V9
V11
V13
V15
V17
V19
V21
V23
V25
V27
Class
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Matplotlib,Seaborn and Numpy were used to plot the correlation matrix.

```
corr = new_df.corr()
sns.heatmap(corr)
```

Fig.7: Correlation matrix for for attributes


Code - Google Drive
Credit card fraud detection
File | E:/DP/2020-21/4013-Credit%20Card%20Fraud%20Detection/Code/Credit%20Card%20Fraud%20Deliverable/Codebase/UI/home.html?u=admin&p=admin
Correlation Matrix
XGBoost

Confusion Matrix:

True Negative	False Positive	False Negative	True Positive
149	11	13	152

Classification Report:

```
Detailed classification report:
              precision    recall  f1-score   support

           0       0.92      0.93      0.93       163
           1       0.93      0.92      0.93       162

    accuracy                           0.93       325
   macro avg       0.93      0.93      0.93       325
weighted avg       0.93      0.93      0.93       325
```

Fig.8 : Comparative Study

XIII. CONCLUSION

Thus we have tried to implement the paper “Hanlin wen, Fangming Huang”, Personal Loan Fraud Detection Based on Hybrid Supervised and Unsupervised Learning” published in IEEE 2020, which proposes a novel hybrid supervised and unsupervised learning method for developing a credit card fraud detection algorithm under the condition of protecting clients’ privacy. An unsupervised learning based on kernel principal component analysis is proposed to decompose the dimension of the dataset. At that point XGboost are joined into the extortion discovery calculation to play out an expectation result. The execution of a cross breed approach is to utilizes unaided anomaly scores to broaden the list of capabilities of a misrepresentation location classifier.

Thus our model can give detailed classification report with accuracy over 0.93 precision with macro avg 0.93 and weight avg 0.93 increasing the possible outcome efficiency.

In the event that one of these or blend of calculation is applied into bank Visa extortion identification framework, the likelihood of misrepresentation exchanges can be anticipated not long after Visa exchanges by the banks. Furthermore, a progression of hostile to extortion methodologies can be embraced to keep banks from incredible misfortunes previously also, diminish hazards. This paper gives commitment towards the charge card misrepresentation identification utilizing the directed learning calculations.

REFERENCES

[1] Andrew. Y. Ng, Michael. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression Advances in neural information processing systems, vol. 2, pp. 841-848, 2002.

[2] A. Shen, R. Tong, Y. Deng, "Application of classification models on credit card fraud detection", Service Systems and Service Management 2007 International Conference, pp. 1-4, 2007.

[3] A. C. Bahnsen, A. Stojanovic, D. Aouada, B. Ottersten, "Cost sensitive credit card fraud detection using Bayes minimum risk", Machine Learning and Applications (ICMLA). 2013 12th International Conference, vol. 1, pp. 333-338, 2013.

[4] I .M. Azhmukhamedov, O.N. Vybornova Introduction of metrics for risk assessment and management // Caspian Journal: Management and High Technologies. - 2015. No. 4(32). pp. 10-22.

[5] F. N. Ogwueleka, "Data Mining Application in Credit Card Fraud Detection System", Journal of Engineering Science and Technology, vol. 6, no. 3, pp. 311-322, 2011.

[6] G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel, A. Riyaz, "A Machine Learning Approach for Detection of Fraud based on SVM", International Journal of Scientific Engineering and Technology, vol. 1, no. 3, pp. 194-198, 2012, ISSN ISSN: 2277-1581.

[7] K. Chaudhary, B. Mallick, "Credit Card Fraud: The study of its impact and detection techniques", International Journal of Computer Science and Network (IJCSN), vol. 1, no. 4, pp. 31-35, 2012, ISSN ISSN: 2277-5420.

[8] M. J. Islam, Q. M. J. Wu, M. Ahmadi, M. A. SidAhmed, "Investigating the

Performance of Naive-Bayes Classifiers and KNearestNeighbor Classifiers", IEEE International Conference on Convergence Information Technology, pp. 1541-1546, 2007.

[9] R. Wheeler, S. Aitken, "Multiple algorithms for fraud detection" in Knowledge-Based Systems, Elsevier, vol. 13, no. 2, pp. 93-99, 2000. (IJCSMC), vol. 4, no. 4, pp. 92-95, 2015, ISSN ISSN: 2320-088X.

[10] “Social Media Analytics for Better Detection of Fraudulent Applications for Online Microfinance Loans” in Wanlin Sun , Ming Chen, Jie-xia Ye , Yuhang Zhang , Cheng-zhong Xu Yangqing Zhang , Yaonan Wang , Wen Wu , Peng Zhang , Feipeng Qu .

[11] “Personal Loan Fraud Detection Based on Hybrid Supervised and Unsupervised Learning” Hanlin Wen , Fangming Huang 978-1-7281-4111-4/20/$31.00 ©2020 IEEE

[12] “Semi-Supervised Anti-Fraud Models for Cash Pre-Loan in Internet Consumer Finance” Wanlin Sun1 , Ming Chen1∗ , Jie-xia Ye1 , Yuhang Zhang1 , Cheng-zhong Xu3 Yangqing Zhang2 , Yaonan Wang2 , Wen Wu2 , Peng Zhang2 , Feipeng Qu2 978-1-5386-8500-6/19/$31.00 ©2019 IEEE