Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.755230

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Singireddy Gowthami
PG scholar, Department of Computer Science & Engineering, Holy Mary Institute of Technology & Science (Autonomous), Hyderabad, India.

Dr D prasad
Associate Professor, Department of Computer Science & Engineering, Holy Mary Institute of Technology & Science (Autonomous), Hyderabad, India.

ABSTRACT
A bank or financial services provider issues credit cards, which enable its holders to borrow money to pay for products and services from businesses that accept credit cards. Credit card firms must be able to recognize fraudulent credit card transactions in order to prevent customers from being charged for products they did not purchase. With everything being done online these days, there is a risk of card abuse and potential financial loss. By using machine learning approaches, data science may tackle challenges of this kind. It deals with credit card fraud detection and machine learning modeling of the dataset. Data is the primary component of machine learning, thus modeling previous credit card transactions with the data of those that turned out to be fraudulent is important. Subsequently, the constructed model is employed to identify the fraudulentness of a novel transaction. Sorting out whether or not there was fraud is the goal. Prior to applying a machine learning algorithm to the credit card dataset and determining the parameters and performance measures, the initial phase entails data analysis and pre-processing.
Keyword: Machine learning, Natural Language Processing, Artificial Intelligence.

1. INTRODUCTION
Risk of credit determining whether a customer would fail or have her credit collapse is the main focus of the board in banks. For example, the advantages of Taiwanese banks declined due to the overabundance of their important clientele, the land business. At that moment, the banks thought it would be beneficial to broaden their client base in order to gain more benefits. Rather of doing so, however, they began issuing Mastercards and enabling an ever-growing number of individuals to apply for them. Over time, their target clientele turned out to be Taiwan's youth. Due to the poor pay for young people, clients began to miss payments, and by February 2006, the total amount owed on Mastercards and other money cards was over 260 billion USD. This gave rise to several problems in Taiwan. Rates of self-destruction and other illegal activities started to rise in an effort to pay back the bank advances. It would have been considerably more effective and beneficial to predict the clients' needs before issuing them Mastercards based on certain characteristics. This would have helped avoid huge obligation problems. Banks currently look at hazard forecasting using a variety of order tactics, such as nave bayes and NLP. In the financial industry, FICO rating cards are a common risk management tactic. They use the personal information and data provided by customers to

predict future defaults and Mastercard borrowings. The bank would then have the option of deciding whether to grant the nominee a visa. Financial evaluations are a fair way to gauge the extent of risk.
The acceptance of credit, including credit cards, is essential to the modern economy. In today' s globalized world, using a credit card is commonplace, especially in developing countries lik e India. For moneylenders, approving credit continues to be a problem as it's hard to predict w hich customers will be a good credit risk and who won't. This is particularly true in developin g countries when the norms and guidelines from developed countries might not be applicable. Thus, it is necessary to look into effective ways for automatic credit approval that may help b ankers analyze customer credit. Every month, tens of thousands of credit card applications are submitted to each bank. In order to decide whether to offer a credit card to an applicant or not , banks must personally scan through each of these applications and carefully consider these v ariables. Owing to the laborious nature of this operation and the increasing probability of inac curacy with the volume of applications, banks are searching for prediction-based algorithms c apable of performing this function efficiently and precisely.In this work, we use a few machin e learning methods to predict whether an application will be granted a credit card or denied. I n order to better understand the elements that are essential for training the model, we first pre-processed the data and carried out extensive EDA. Furthermore, we have applied 10 machine learning algorithms to these pre-processed data sets in order to determine which model, taking into account the precision-recall trade-off, yields the best accurate findings. The following is t he arrangement of the essay. We have discussed the results of our literature review in Section II. We go into great length explaining the complete mechanism in Section III. We also present ed, examined, and contrasted the outcomes from several angles. We finished the results and o bservations at the end.

1.1OBJECTIVES

The purpose is to construct a machine learning model for Credit Card Fraud Prediction, to pot entially replace the updatable supervised machine learning classification models by predicting outcomes in the form of highest accuracy by comparing supervised algorithm.

2. LITERATURE SURVEY

Paruchuri Harish [1] Businesses aim to provide their clients with an increasing number of am enities. The ability to purchase products online is one of these conveniences. Customers may now purchase the necessary supplies online, but there is also a chance for fraud by dishonest p eople. Until the cardholder notifies the bank to restrict the card, thieves can steal any cardhold er's information and use it for online transactions. This article presents the various machine le arning techniques used to identify this type of transaction. According to the report, the primar y problem facing the financial sector that is becoming worse with time is CCF. An increasing number of businesses are transitioning to an online platform that enables consumers to do tran sactions online. Criminals can use this as a chance to steal credit card numbers or other person al information in order to conduct online transactions. Phishing and Trojan are the most often utilized methods for stealing credit card information. Therefore, to identify such activities, a f raud detection system is required.

Chandini S. B., Rajkumar N., Gagana P Rao, Nisarga K. S., Devika S. P. [2] These days, both public and private sector personnel are increasingly likely to use credit cards. Users make onli ne purchases of consumable durable goods with credit cards, moving money across accounts i n the process. Through phishing, Trojan viruses, and other means, the fraudster is tracking do wn the specifics of the user's transactions and using the card for illicit purposes. Users' sensiti ve information may be threatened by fraud. We have covered a number of techniques for iden tifying and stopping fraudulent activity in this essay. This will assist to enhance card transacti on security going forward. One of the main problems is credit card fraud, which affects many individuals. Numerous credit card customers are losing their money and private information a s a result of these illegal operations. In this essay, we've covered many methods for detecting and stopping credit card fraud. We've also spoken about how to strengthen security against fra udsters going forward to prevent illicit activity.

Hitendra Garg and Akanksha Bansal [3] Credit card transactions are one of the most often use d payment methods nowadays. Growing patterns of credit card-based financial transactions al so encourage fraud, which results in losses of billions of dollars on a global scale. Additionall y, a 35% rise in fraudulent transactions has been noted from 2018. To study fraud detection o perations, which involve analyzing behavior or irregularities in the transaction dataset to ident ify and disregard the suspected person's undesired conduct, a vast amount of transaction data i s accessible. To warn the user of fraudulent transactions, the suggested study provides a comp ressed summary of several methods for classifying fraudulent transactions from different data sets. Online transactions have become the most popular means of conducting financial transac tions over the past few decades, rising at a rapid pace. Threats rise in tandem with the expansi on of online commerce. Therefore, the proposed work summarizes the many methodologies u sed to detect the abnormal transaction in the dataset of credit card transaction datasets, bearin g in mind the security problem, nature, and anomaly in the credit card transaction. The greates t outcomes are being obtained when the data is balanced in every location.

H. Sangeetha, K. Saran Sriram, R. A. Karthikeyan, S. Abinayaa, and D. Piyush [4] The prima ry goal of this study was to identify actual credit card theft. For the qualifying data set, we mu st first gather the credit card data sets. Next, provide the customer credit card questions to test the set of data. Using the currently available data set and the previously assessed data set, the r andom forest algorithm classification technique is applied [1]. Ultimately, the precision of the outcome information is maximized. Subsequently, several variables will be processed in order to identify factors that influence fraud detection while examining the graphical model's depict ion. Based on precision, specificity, adaptability, and accuracy, the efficiency 17 method is ev aluated. The Random Forest Algorithm has shown to produce far more effective outcomes.

3. EXISTING SYSTEM

They put up a strategy they called the Information-Utilization-Strategy. INUM the correctness and convergence of an information vector produced by INUM are examined, as well as the ini tial design. By contrasting INUM with alternative approaches, its originality is demonstrated. In decision space, two D-vectors, or feature subsets, a and b, where Ai is the ith feature in a d ata set, are different, yet in objective space, they correspond to the same O-vector, y. Assume that decision-makers are only given a, and that a becomes useless because of an accident or ot her circumstances (such as a difficult extraction from the data set). Decision-makers then face difficulties. However, if they are given access to both feature subsets, they will have more opt ions to choose from that will work best for them. Put differently, decision makers may have g reater opportunities to make sure that their interests are well-served if there are more similar D -vectors available in the decision space. Consequently, it is crucial to solve MMOPs with both the greatest number of D-vectors for each O-vector and a decent Pareto front estimate.

4. PROPOSED SYSTEM

The suggested model aims to develop a classification model to classify if its fraud or not. The dataset of prior credit card instances is gathered and utilized to train the computer to recogniz e the issue. The first phase for comprises the examination of data where each and every colum n is evaluated and the appropriate measurements are done for missing values and other types o f data. Outliers and other values with little bearing are handled. Then pre-processed data is uti lized to develop the classification model where the data will be split into two parts one is for t raining and remaining data for testing purpose. Machine learning algorithms are performed on the training data where the model learns the pattern from the data and the model will deal wit h test data or new data and categorize whether its fraud or not .The algorithms are compared a nd the performance measure of the methods are calculated.

5. System Architecture

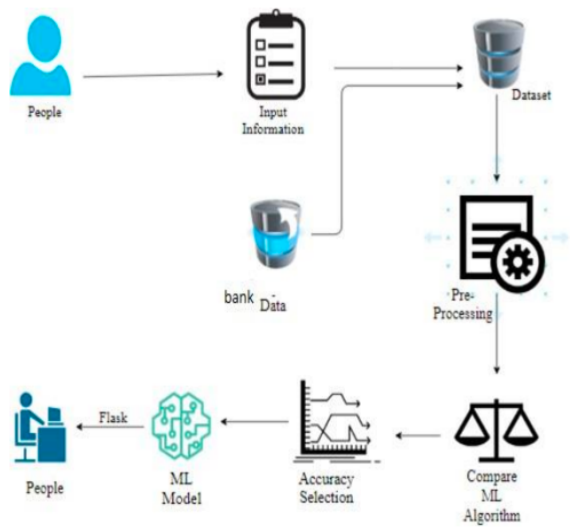

Fig.5.1. System Architecture

6. LIST OF MODULES

- Data Pre-processing
- Data Analysis of Visualization
- Comparing Algorithm with prediction in the form of best accuracy result
- Deployment Using Flask

6.1 MODULE DESCRIPTION

6.1.1. DATA PRE-PROCESSING

Validation approaches in machine learning are used to acquire the error rate of the Machine L earning (ML) model, which may be deemed as near to the genuine error rate of the dataset. If the data volume is large enough to be representative of the population, you may not require th e validation approaches. However, in real-world circumstances, to deal with samples of data t hat may not be a genuine representative of the population of supplied dataset. To discover the missing value, duplicate value and explanation of data type whether it is float variable or integ er. The sample of data used to offer an impartial evaluation of a model fit on the training datas et while tweaking model hyper parameters. The evaluation becomes increasingly biased when competence on the validation dataset is included into the model setup. The validation set is us ed to test a particular model, although this is for frequent evaluation. It as machine learning en gineers utilize this data to fine-tune the model hyper parameters. Data collection, data analysis , and the process of addressing data content, quality, and organization can add up to a time- co nsuming to-do list. During the process of data identification, it helps to understand your data a nd its qualities; this information will help you pick which method to employ to develop your model.

6.1.2. Data Validation/ Cleaning/Preparing Process

loading the specified dataset while importing the library packages. Analyzing the variable ide ntification by data type and form, assessing duplicate and missing values, etc. A validation da taset is a portion of data that was withheld from your model's training process. It is intended t o provide an estimate of model skill during model tuning. You may utilize processes to optim ize the utilization of test and validation datasets during model evaluation. Renaming the provi ded dataset, removing a column, and other actions are examples of data cleaning and preparat ion for uni-, bi-, and multi-variate analysis. Data cleaning procedures and methods differ depe nding on the kind of dataset. Finding and eliminating mistakes and abnormalities is the main o bjective of data cleaning, which aims to improve the value of data for analytics and decision-making.

6.1.3. Exploration Data Analysis Of Visualization

In applied statistics and machine learning, data visualization is a critical competency. In fact, t he main focus of statistics is on quantitative data descriptions and estimations. An essential se t of tools for developing a qualitative understanding is provided by data visualization. This ca n be useful for discovering trends, faulty data, outliers, and much more while examining and g etting to know a dataset. Plots and charts that are more visceral and stakeholders than measure ments of association or significance can convey and illustrate important relationships with the use of data visualizations, provided the user has some topic expertise. Data visualization and e xploratory data analysis are whole areas , and it will recommend a deeper dive into of the boo ks listed at the conclusion. When data is presented visually, such with charts and graphs, it ma y sometimes make sense. In applied statistics as well as applied machine learning, the ability t o display data samples and other types of information rapidly is crucial. It will explore the ma ny plot types that you should be aware of when using Python to visualize data and show you h ow to use them to enhance your understanding of your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.

7. OUTPUT RESULTS

```
import pandas as pd
import numpy as np
import tensorflow as tf
from tensorflow import keras
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
import matplotlib.pyplot as plt
%matplotlib inline

data = pd.read_csv('/content/test(HAND WRITTEN).csv')
test_data = pd.read_csv('/content/train.csv')



features = data.drop(["label"], axis=1)
labels = data ['label']
x_train, x_test, y_train, y_test = train_test_split(features, labels, test_size=0.2,random_state=42)
features
labels
print(features.describe())
labels.describe()
x_train
x_test
y_train
y_test

len(x_train)
x_train.loc[0].values
x_train.loc[20].values

y_train[0]

```

```
y_train[20]

labels.loc[20]

img = x_train.loc[0].values.reshape(28,28)
plt.title('Digit: '+ str(y_train[0]))
plt.imshow(img)


print(len(x_train))
print(len(y_train))
print(len(x_test))
print(len(y_test))

print(x_train.shape)
print(y_train.shape)
print(x_test.shape)
print(y_test.shape)

features.shape

x_train = x_train /255
x_test = x_test /255
x_train.loc[0].values


model = keras.Sequential([
    keras.layers.Dense(256, input_shape=(784,), activation= 'relu'),
    keras.layers.Dense(128, activation='relu'),
    keras.layers.Dense(100, activation='relu'),
    keras.layers.Dense(50, activation='relu'),
    keras.layers.Dense(30, activation='relu'),
```

```
    keras.layers.Dense(10, activation='softmax'),
])
model.compile(
    optimizer = 'adam',
    loss = 'sparse_categorical_crossentropy',
    metrics = ['accuracy']
)
callbacks = [
    keras.callbacks.ModelCheckpoint(filepath='model_weights.h5', save_weights_only=True),
    keras.callbacks.History()
]
history = model.fit(x_train, y_train, epochs=20, callbacks=callbacks)

epoch_accuracy = history.history['accuracy']
np.save('epoch_accuracy.npy', epoch_accuracy)


test_loss, test_accuracy = model.evaluate(x_test, y_test)
print('Test loss: ', test_loss)
print('Test accuracy: ', test_accuracy)


test_predictions = model.predict(x_test)
test_predictions[0]

np.argmax(test_predictions[0])

y_test[:5]

predicted_labels = [np.argmax(i) for i in test_predictions]
predicted_labels[:5]

```

```
test_predictions = model.predict(x_test)
predicted_labels = np.argmax(test_predictions, axis=1)

correct_predictions = np.sum(predicted_labels == y_test)
total_samples = len(y_test)
#Calculate Accuracy
accuracy = accuracy_score(y_test, predicted_labels)
print('Test Accuracy: ', accuracy)

#Calculate Precision
precision = precision_score(y_test, predicted_labels, average= 'weighted')
print('Precision: ',precision)

#Calculate recall
recall = recall_score(y_test, predicted_labels, average= 'weighted')
print('recall: ',recall)

#Calculate F1-score
f1 = f1_score(y_test, predicted_labels, average= 'weighted')
print('F1_Score: ',f1)

#Calculate Confusion Matrix
conf_matrix = confusion_matrix(y_test, predicted_labels)
print('Confusion Matrix: ', conf_matrix)

model.load_weights('model_weights.h5')
np.load('epoch_accuracy.npy')
```

8. CONCLUSION

The analytical procedure starts from data cleaning and processing, missing value, exploratory analysis and lastly model construction and assessment. The best accuracy on public test set is greater accuracy score will be discovered out. This program can assist to detect the Prediction of credit card fraud or not.

8.1 FUTURE WORK

Credit card fraud prediction to integrate with cloud model. To optimize the job to impleme nt in Artificial Intelligence environment.

REFERENCES

[1] Q. Wu, M. Zhou, Q. Zhu, Y. Xia, and J. Wen, ―MOELS: Multiobjective evolutionary list scheduling for cloud workflows,‖ IEEE Trans. Autom. Sci. Eng., vol. 17, no. 1, pp. 166–176, J an. 2020.

[2] L. Huang, M. Zhou, and K. Hao, ―Non-dominated immune-endocrine short feedback alg orithm for multi-robot maritime patrolling,‖ IEEE Trans. Intell. Transp. Syst., vol. 21, no. 1, p p. 362–373, Jan. 2020.

[3] X. Wang, K. Xing, C.-B. Yan, and M. Zhou, ―A novel MOEA/D for mul tiobjective sche duling of flexible manufacturing systems,‖ Complexity, vol. 2019, pp. 1–14, Jun. 2019.

[4] J. J. Liang, S. T. Ma, B. Y. Qu, and B. Niu, ―Strategy adaptative memetic crowding diffe rential evolution for multimodal optimization,‖ in Proc. IEEE Congr. Evol. Comput., Jun. 201 2, pp. 1–7.

[5] M. M. H. Ellabaan and Y. S. Ong, ―Valley-adaptive clearing scheme for multimodal opti mization evolutionary search,‖ in Proc. 9th Int. Conf. Intell. Syst. Design Appl., Nov. 2009, p p. 1–6.

[6] X. Li, ―Efficient differential evolution using speciation for multimodal function optimiza tion,‖ in Proc. Conf. Genetic Evol. Comput. - GECCO, Jun. 2005, pp. 873– 880.

[7] Y. Feng et al., ―Target disassembly sequencing and scheme evaluation for CNC machine tools using improved multiobjective ant colony algorithm and fuzzy integral,‖ IEEE Trans. Sy st., Man, Cybern. Syst., vol. 49, no. 12, pp. 2438–2451, Dec. 2019.

[8] L. Ma, X. Wang, M. Huang, Z. Lin, L. Tian, and H. Chen, ―Two-level master– slave RF ID networks planning via hybrid multiobjective artificial bee colony optimizer,‖ IEEE Trans. Syst., Man, Cybern. Syst., vol. 49, no. 5, pp. 861– 880, May 2019.

[9] X. Zhang, K. Zhou, H. Pan, L. Zhang, X. Zeng, and Y. Jin, ―A network reductionbased m ultiobjective evolutionary algorithm for community detection in large-scale complex network s,‖ IEEE Trans. Cybern., vol. 50, no. 2, pp. 703–716, Feb. 2020.

[10] Q. Fan and X. Yan, ―Solving multimodal multiobjective problems through zoning searc h,‖ IEEE Trans. Syst., Man, Cybern. Syst., early access, Oct. 9, 2019, doi: 10.1109/TSMC.201 9.2944338.

[11] K. Deb and S. Tiwari, ―Omni-optimizer: A procedure for single and multi- objective op timization,‖ in Proc. 3rd Int. Conf. Evol. Criterion Optim. Optim. (EMO), Mar. 2005, pp. 47–61

[12] R. Tanabe and H. Ishibuchi, ―A framework to handle multi-modal multi objective optim ization in decomposition-based evolutionary algorithms,‖ IEEE Trans. Evol. Comput., vol. 24, no. 4, pp. 720–734, Aug. 2020.