



2022 International Conference for Advancement in Technology (ICONAT 2022)



Fraud detection on bank payments

Paper ID:488

By

KAMMARI SANTHOSH,

PINKU RANJAN,

ARUN KUMAR,

SOMESH KUMAR

Abv-iiitm gwalior

Abstract

This project is based on fraud discovery and steps to automate it fully. Fraud discovery has become an essential priority for every bank. Fraud is increasing significantly, which results in substantial damages for the banks. Transactions cause new challenges for fraud detection due to the requirement of short processing time. The main work is a feasibility study of selected fraud detection approaches. By using models, this transaction is tested individually, and whatever suits them best has further proceeded. We first characterize a detection task: the dataset and its attributes, the metric choice, and some methods to handle such unbalanced datasets. Then we focus on the dataset shift. This refers to the fact that the underlying distribution generating the dataset evolves: For example, cardholders may change their buying habits over seasons, and fraudsters may adapt their strategies. Afterward, we highlighted different approaches used to capture the sequential properties of credit card transactions.

Keyword: Enterprise Modeling, Digital Transformation, Instant Payment., Fallacious Transaction, Online Shopping

Table of Contents

- INTRODUCTION
- LITERATURE SURVEY
- OBJECTIVES
- SYSTEM ARCHITECTURE
- DATA ANALYSIS AND PREPROCESSING
- EXPERIMENTAL ANALYSIS AND RESULTS
- ADVANTAGES
- LIMITATION
- FUTURE WORK
- REFERENCES

Introduction

Fraud detection in financial transactions has become an essential priority for banks. For this purpose, publications about existing approaches were analyzed to explore their utilization in instant payments.

Two methods used are:

- 1) analysis based on literature perspective
- 2) feasibility study of fraud detection.

Most of the fraud detection systems are based on artificial intelligence, pattern matching.

There needs to be a handle of incorrect data. Another problem is overlapping data. Many transactions may resemble fraudulent transactions when they are genuine transactions.

Even with very high accuracy, almost all fraudulent transactions can be misclassified. The system should take care of the amount of money lost due to fraud and the amount of money required to detect that fraud.

Literature Survey

Author	Title	Year	Publisher	Work
Simon Delecourt	Building a Fraud Detection System	2019	IEEE	Considered reactions of fraudsters to build a robust mobile fraud detection system
S.P.Maniiraj	fraud detection using ML	2019	IEEE	objective is to detect fraudulent transactions while minimizing the incorrect fraudulent classifications
Thamer Alquthami	Smart Meters Data Processing	2019	IEEE	The analysis is performed through a program that was developed in Python-Pandas

Literature Survey

Author	Title	Year	Publisher	Work
Treepatchara Tasnav- ijitvong, Phayung Meesad	Study on Machine Learning Techniques for Pay- ment Fraud Detection	2019	IEEE	the study conducted is with multiple machine learning techniques with the use of the synthesized dataset
N. Malini; M. Pushpa	Analysis on fraud iden- tification techniques based on KNN	2017	IEEE	Along with other tech- niques, the KNN algo- rithm and outlier detec- tion methods are im- plemented to optimize the best solution for the fraud detection problem.

OBJECTIVES

A fraud detection system's critical objective is to recognize suspicious events and reach them to an analyst while letting everyday transactions be automatically processed.

For years, financial institutions have been committing this task to rule-based methods that employ rule sets written by experts. But now, they frequently turn to a machine learning approach, as it brings significant enhancements to the process.

- **Higher accuracy of fraud detection.** As opposed to rule-based solutions, machine learning mechanisms have higher precision and return relevant results as they study multiple additional factors.
- **low manual work is needed for additional verification.** Improved accuracy leads to a decrease in the burden on analysts. As people are unable to check every transaction, even for a small bank

OBJECTIVES Contd..

- **Fewer false declines.** False declines happen when a system recognizes a legal transaction as suspicious and wrongfully cancels it.
- **Ability to identify new patterns and adapt to changes.** Unlike rule-based systems, ML aligns with the constant evolution of the environment; They enable analysts to remember new suspicious ways(pattern) and create new rules to prevent new types of scams.

System Architecture

Method used:

- a) We detect the fraudulent transactions from the Banksim dataset. This synthetically generated dataset consists of payments from various customers made in different periods and with different amounts.
- b) The primary dataset containing over 594643 records is collected, which has fraud and non-fraud transactions. data preprocessing:

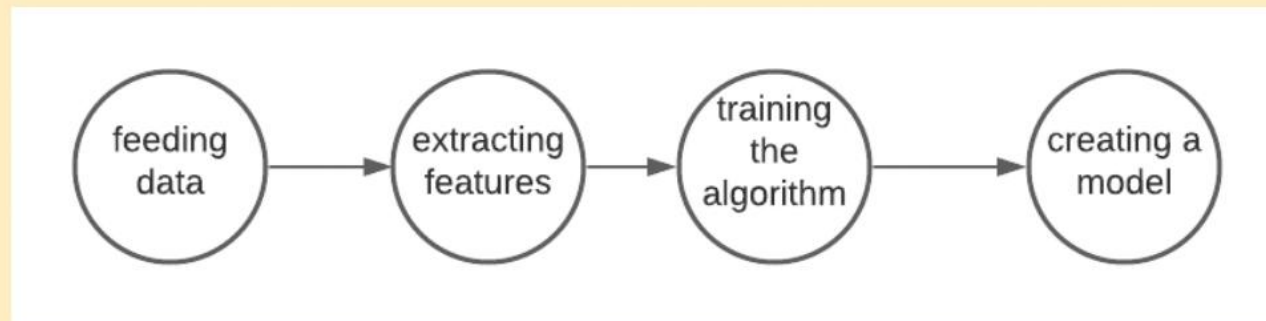


Figure: Basic Architecture

System Architecture Cont..

Data Preprocessing:

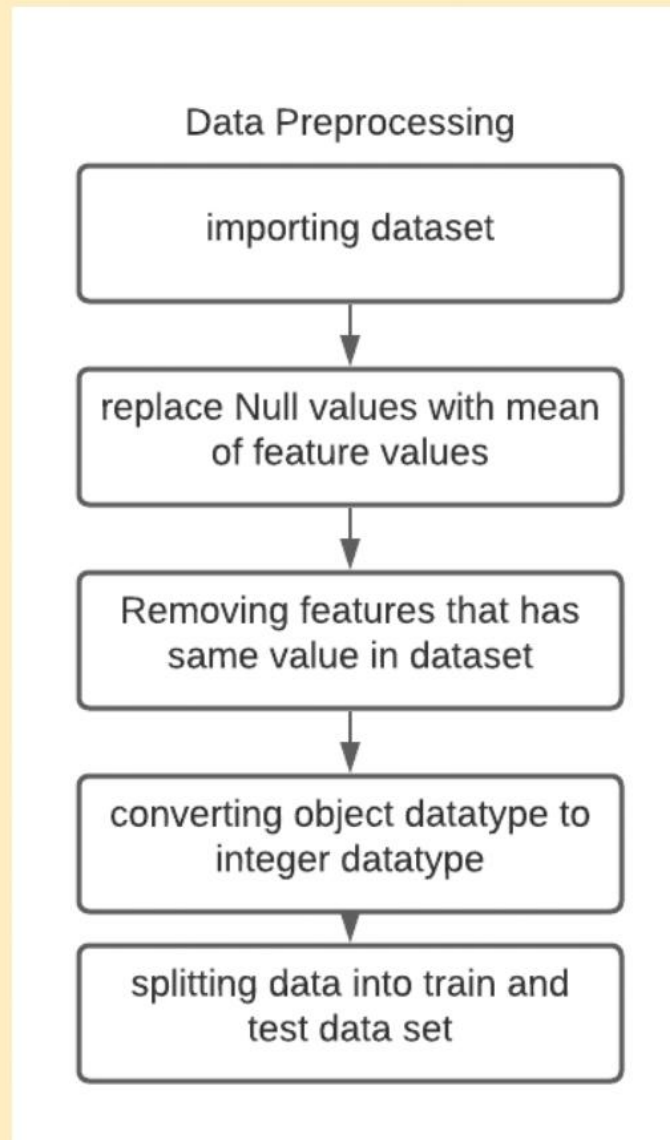


Figure: preprocessing

System Architecture Cont..

- After the preprocessing, a model with logistic regression is imposed on the dataset. The models are trained and tested and we determine the model efficiency on the basis of accuracy.
- At last, a baseline binary classification model using Logistic Regression is developed. The accuracies are evaluated and compared to determine the model with the best possible accuracy

The architecture of these model is stated in the flowchart that follows

System Architecture Cont..

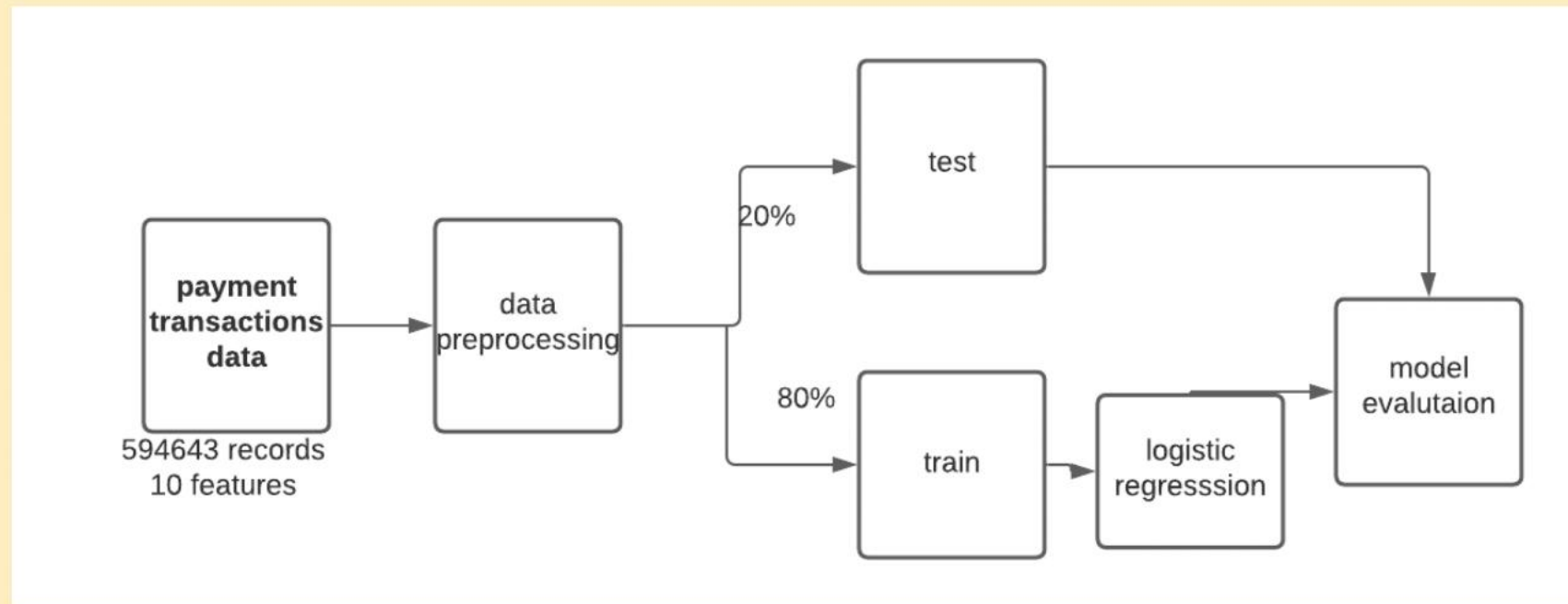


Figure: system architecture

DATA ANALYSIS

- using pandas data is fetched which consists of over 59000 records The data is analyzed and preprocessed by implementing a series of steps:

	step	customer	age	gender	zipcodeOri	merchant	zipMerchant	category	amount	fraud
0	0	'C1093826151'	'4'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	4.55	0
1	0	'C352968107'	'2'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	39.68	0
2	0	'C2054744914'	'4'	'F'	'28007'	'M1823072687'	'28007'	'es_transportation'	26.89	0
3	0	'C1760612790'	'3'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	17.25	0
4	0	'C757503768'	'5'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	35.72	0
5	0	'C1315400589'	'3'	'F'	'28007'	'M348934600'	'28007'	'es_transportation'	25.81	0
6	0	'C765155274'	'1'	'F'	'28007'	'M348934600'	'28007'	'es_transportation'	9.10	0
7	0	'C202531238'	'4'	'F'	'28007'	'M348934600'	'28007'	'es_transportation'	21.17	0

Figure: dataset

- **Search for Null values:**

There may a chance of empty values in the given data set, so we need to fetch the indexes of the null value and replace them with the mean of feature values;

```
No of Null values in given dataset = 0
```

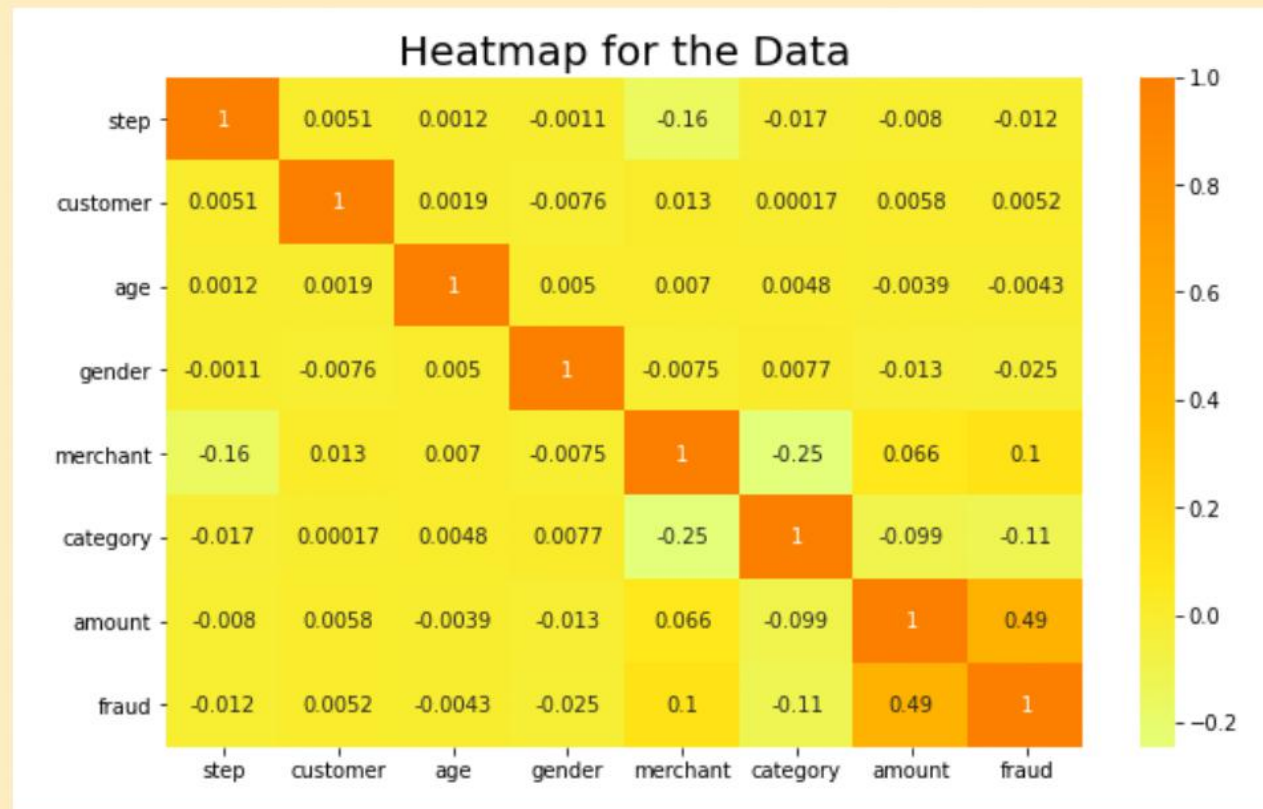
Figure: count of NULL values

From the above output, the dataset has zero Null values

DATA ANALYSIS Cont..

strength between each of feature


In the given dataset, we need find the strength between each and every feature,i.e proportionality rate between them.



the map value with second max value results in the most dependent feature

DATA ANALYSIS Cont..

Grouped by category:



	fraud	non-fraud	percentage_fraud
'es_transportation'	NaN	26.958187	0.000000
'es_food'	NaN	37.070405	0.000000
'es_hyper'	169.255429	40.037145	4.591669
'es_barsandrestaurants'	164.092667	41.145997	1.882944
'es_contents'	NaN	44.547571	0.000000
'es_wellnessandbeauty'	229.422535	57.320219	4.759380
'es_fashion'	247.008190	62.347674	1.797335
'es_leisure'	300.286878	73.230400	94.989980
'es_otherservices'	316.469605	75.685497	25.000000
'es_sportsandtoys'	345.366811	88.502738	49.525237
'es_tech'	415.274114	99.924638	6.666667
'es_health'	407.031338	103.737228	10.512614
'es_hotelservices'	421.823339	106.548545	31.422018
'es_home'	457.484834	113.338409	15.206445
'es_travel'	2660.802872	669.025533	79.395604

Figure: fraud vs non-fraud

DATA ANALYSIS Cont..

Fraudulent vs non-fraudulent dataset:

Again we can see in the histogram below that the fraudulent transactions are less in the count but more in amount.

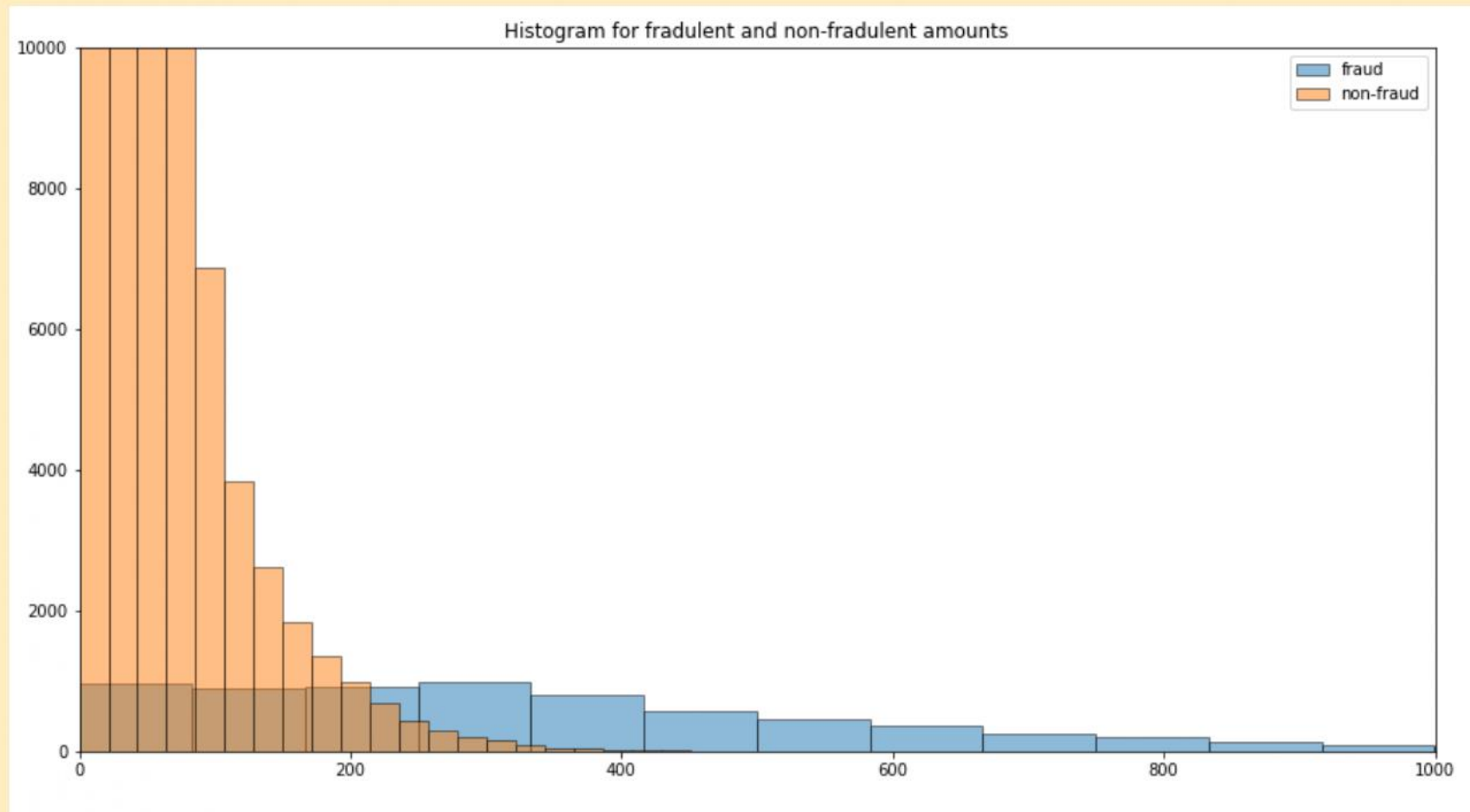
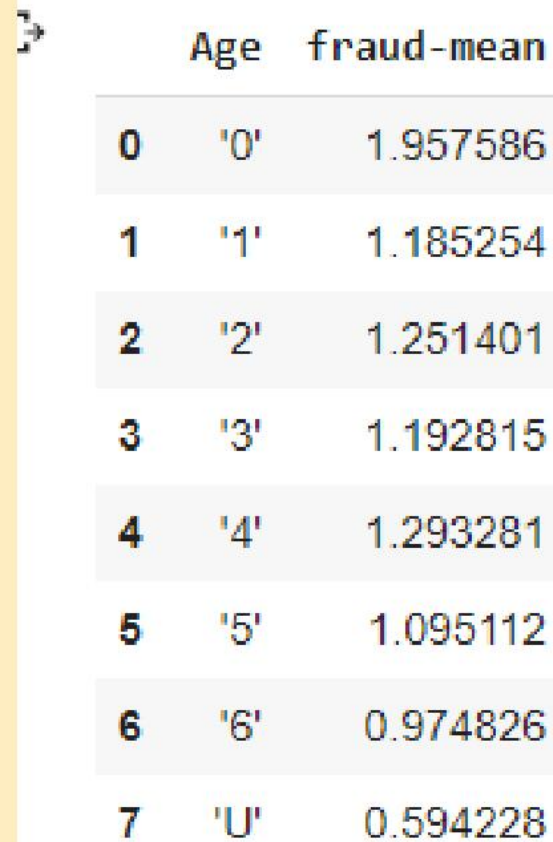


Figure: fraud vs non-fraud

DATA ANALYSIS Cont..

The dependency of age feature:



	Age	fraud-mean
0	'0'	1.957586
1	'1'	1.185254
2	'2'	1.251401
3	'3'	1.192815
4	'4'	1.293281
5	'5'	1.095112
6	'6'	0.974826
7	'U'	0.594228

Figure: Age vs fraud feature

Looks like fraud occurs more in ages equal and below 18

DATA PREPROCESSING

In this part we will preprocess the data and prepare for the training. There are only one unique zipCode values so we will drop them

```
dzip=dataset.zipcodeOri.nunique()
dmer=dataset.zipMerchant.nunique()

# erasing the zipCodeOri and zipMerchant values from the dataset"
dataset = dataset.drop(['zipcodeOri','zipMerchant'],axis=1)
print("Unique zipCodeOri values: ",dzip)
print("Unique zipMerchant values: ",dmer)

Unique zipCodeOri values: 1
Unique zipMerchant values: 1
```

Figure: dropping features

Undersampling Technique:

Using Undersampling Technique for balancing the dataset. Resulted counts show that now we have exact number of class instances (1 and 0). Finally , I will do a train test split for measuring the performance.

```
the shape of train_X: (14400, 7)
the shape of train_y: (14400,)

counts of label '1': 7200
counts of label '0': 7200
```

Figure: Undersampling

EXPERIMENTAL ANALYSIS AND RESULT:

The performance metrics used to evaluate the models are the accuracy and the loss. Our main aim was to increase the accuracy at the same time decrease the loss which means that the model is learning.

Logistic Regression Algorithm:

```
classifier = LogisticRegression(random_state = 0)
classifier.fit(X_train, y_train)
y_pred=classifier.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print("the confusion matrix is =")
print(cm)
print("accuracy per=",accuracy_score(y_test,y_pred)*100)

the confusion matrix is =
[[2071  89]
 [ 283 1877]]
accuracy per= 91.38888888888889
```

Figure: Logistic Regression

K-Nearest Neighbor Algorithm:

```
knn_cls=KNeighborsClassifier(n_neighbors=5,p=2)
knn_cls.fit(X_train,y_train)
y_pred=knn_cls.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print("the confusion matrix is =")
print(cm)
print("accuracy per=",accuracy_score(y_test,y_pred)*100)
```

the confusion matrix is =
[[2071 89]
 [228 1932]]
accuracy per= 92.66203703703704

Figure: K-Nearest Neighbor

Decision Tree Classification Algorithm:

```
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print(cm)
print("accuracy per=", accuracy_score(y_test, y_pred)*100)

[[2067  93]
 [ 75 2085]]
accuracy per= 96.11111111111111
```

Figure: Decision Tree Classification

Random Forest Algorithm:

```
rf=RandomForestClassifier(n_estimators=100,
                          max_depth=8,random_state=42,verbose=1,
                          class_weight="balanced")
rf.fit(X_train,y_train)
y_pred=rf.predict(X_test)
print("the confusion matrix is =")
print(cm)
print("accuracy per=",accuracy_score(y_test,y_pred)*100)
```



```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 c
the confusion matrix is =
[[2067  93]
 [ 75 2085]]
accuracy per= 96.64351851851852
```

Figure: Random Forest Algorithm

ADVANTAGES

- finding hidden and implicit correlations in the data
- automatic detection of possible fraud scenarios
- reduced number of verification measures
- real-time processing
- The amount of time required for the result and the cost is less when compared to manual testing.
- can be used for different transaction models for classification.

LIMITATIONS

The amount of time required for the result and the cost for manual testing and the limited availability of test kits. So we need a technique that provides the work faster and reduces the cost to check. Still, this model can return good accuracy and can be further improved.

FUTURE SCOPE

The findings obtained here are not in a generalized form that can be directly used in the global fraud detection problem. Here, we have considered a sample data set. As future work, some effective algorithms could be developed for the classification problem with variable misclassification costs. And the collection of more payment data will improve the models accuracy.

References

The references (1),(2),(3),(4),(5),(6) are cited

- [1] S. Delecourt, “Building a robust mobile payment fraud detection system with adversarial examples,” *IEEE*, 2019.
- [2] V. Jain, M. Agrawal, and A. Kumar, “Performance analysis of machine learning algorithms in credit cards fraud detection,” pp. 86–88, 2020.
- [3] S.P.Maniiraj, “credit card fraud detection using machine learning and data science,” *IEEE*, 2019.
- [4] hamer Alquthami; Abdullah M. Alsubaie; Murad Anwer, “Importance of smart meters data processing,” *IEEE*, 2019.
- [5] O. A. J. W. S. lawte; Hemaint Jaiman, “Comparative evaluation of credit card fraud detection using machine learning techniques,” *IEEE*, 2019.
- [6] S. Khatri, A. Arora, and A. P. Agrawal, “Supervised machine learning algorithms for credit card fraud detection: A comparison,” pp. 680–683, 2020.

Thank You