

# FRAUD DETECTION ON BANK PAYMENTS

*A project report submitted in partial fulfillment of the requirements  
for Mini Project*

*by*

**Kammari Santhosh(2018IMT-043)**

*Under the Supervision of*

**Dr. Pinku Ranjan**



विश्वजीवनामृतं ज्ञानम्

**ABV INDIAN INSTITUTE OF INFORMATION  
TECHNOLOGY AND MANAGEMENT  
GWALIOR-474 015  
2020**

Signature of the Supervisor

---

## 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.

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

## ACKNOWLEDGEMENTS

I am highly indebted to Dr. Pinku Ranjan for giving me the autonomy of functioning and experimenting with ideas. I would like to take this opportunity to express my profound gratitude to his not only for his academic guidance but also for his personal interest in my project and constant support coupled with confidence boosting and motivating sessions which proved very fruitful and were instrumental in infusing self-assurance and trust within me. The nurturing and blossoming of the present work are mainly due to her valuable guidance, suggestions, astute judgment, constructive criticism and an eye for perfection. My mentor always answered myriad of my doubts with smiling graciousness and prodigious patience, never letting me feel that I am novice by always lending an ear to my views, appreciating and improving them and by giving me a free hand in my project. It's only because of her overwhelming interest and helpful attitude, the present work has attained the stage it has. Finally, I am grateful to our Institution and colleagues whose constant encouragement served to renew my spirit, refocus my attention and energy and helped me in carrying out this work.

(Kammari Santhosh)

---

# TABLE OF CONTENTS

<b>1</b>	<b>CHAPTER 1</b>	<b>4</b>
1.1	INTRODUCTION . . . . .	4
1.2	BACKGROUND INFORMATION . . . . .	4
1.3	LITERATURE SURVEY . . . . .	5
<b>2</b>	<b>METHODOLOGY</b>	<b>6</b>
2.1	OBJECTIVES . . . . .	6
2.2	MOTIVATION . . . . .	6
2.3	SALIENT FEATURES . . . . .	7
2.4	DATASETS . . . . .	7
2.5	SYSTEM ARCHITECTURE . . . . .	7
2.6	SETUP AND TOOLS . . . . .	9
<b>3</b>	<b>CHAPTER 3</b>	<b>10</b>
3.1	DATA ANALYSIS . . . . .	10
3.2	DATA PREPROCESSING . . . . .	14
3.3	EXPERIMENTAL ANALYSIS AND RESULT: . . . . .	15
<b>4</b>	<b>CONCLUSION</b>	<b>18</b>
4.1	ADVANTAGES . . . . .	18
4.2	LIMITATIONS . . . . .	18
4.3	FUTURE SCOPE . . . . .	18
<b>5</b>	<b>REFERENCES</b>	<b>19</b>

---

## List of Figures

1	simple modeling . . . . .	7
2	Data preprocessing . . . . .	8
3	system architecture . . . . .	9
4	heatmap . . . . .	10
5	count of fraud data . . . . .	11
6	fraud vs non-fraud . . . . .	12
7	fraud vs non-fraud . . . . .	13
8	dropping features . . . . .	14
9	transforming features . . . . .	14
10	Undersampling . . . . .	15
11	Logistic Regression . . . . .	15
12	K-Nearest Neighbor . . . . .	16
13	Decision Tree Classification . . . . .	16
14	Random Forest Algorithm . . . . .	17

## List of Tables

1	Summary of Related Work . . . . .	5
2	Accuracy matrix . . . . .	17

---

# 1 CHAPTER 1

## 1.1 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.

## 1.2 BACKGROUND INFORMATION

Fraud detection is a collection of actions taken to stop money or property from being obtained through dishonest acts.

Fraud can be engaged in different ways and many productions. The majority of detection methods connect various fraud detection datasets to build an attached sketch of both valid and non-valid payment data to secure a decision. This decision must consider IP address, geolocation, device identification, BIN data, global latitude/longitude, historic transaction patterns, and the actual transaction information. In practice, this means that traders and issuers deploy analytically based responses that use domestic and outside data to apply a set of business rules or logical algorithms to detect fraud.

### 1.3 LITERATURE SURVEY

Author	Title	Year	Published	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	credit card fraud detection using machine learning and data science	2019	IEEE	the objective here 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
Treepatchara Tasnav-ijitvong; Panit Suwimonseatein; Phayung Meesad	Study on Machine Learning Techniques for Payment 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 identification techniques based on KNN and outlier detection	2017	IEEE	Along with other techniques, the KNN algorithm ,outlier detection methods are implemented to optimize the best solution for the fraud detection problem.

Table 1: Summary of Related Work

---

## 2 METHODOLOGY

### 2.1 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
- **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.

### 2.2 MOTIVATION

- As humans, we can't differentiate the fraud transaction from any kind of dataset, but a model can do it easily, The increase in e-transaction turns the customers to face a number of risks of security breaching.
- The major problem in the online payment system is the management of fraudulent entries companies are investing in anti-fraud systems
- Building a framework for fraud risk management and a robust system of internal financial control can help organizations to reduce the risk of loss through fraud and financial crime.
- many organizations employ or engage internal audit and risk management experts to ensure a strong anti-fraud framework. For a number of years, retail businesses have been looking at a mixture of technological solutions and internal control reviews to detect and prevent such fraud.

---

## 2.3 SALIENT FEATURES

- The Scraped dataset is keenly reviewed and selected before reading them to ensure better training and accuracy.
- This system is developed to prevent people from fraud in bank transactions.
- The Transactions with features that don't deviate from the norm are allowed for processing.
- After learning a particular pattern ,the model can predict the fraud payment
- Many of the systems use both rules and machine learning techniques to achieve higher efficiency.
- Their accuracy is evaluated and compared to the binary baseline classifier using LogisticRegression.

## 2.4 DATASETS

To train the above model datasets must be retrieved

- A Dataset containing the information of bank transactions and the dataset consists of about 59000 records
- The Dataset is a collection of bank transactions of both fraud and non fraud payments.

## 2.5 SYSTEM ARCHITECTURE

The Method used:

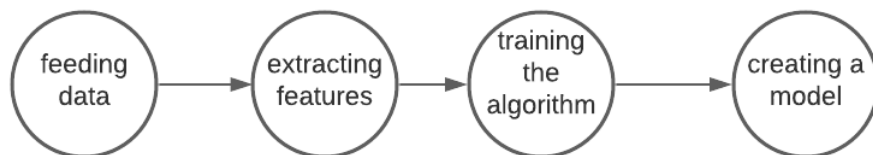


Figure 1: simple modeling

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.



---

## Data preprocessing:

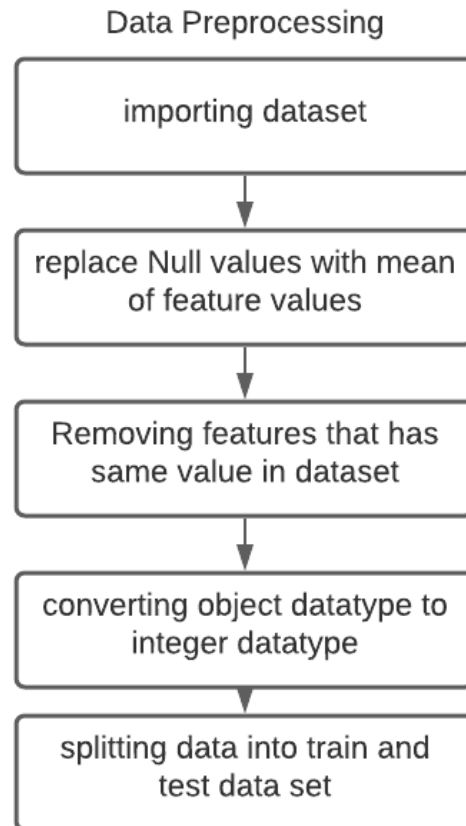


Figure 2: Data preprocessing

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.

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

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.

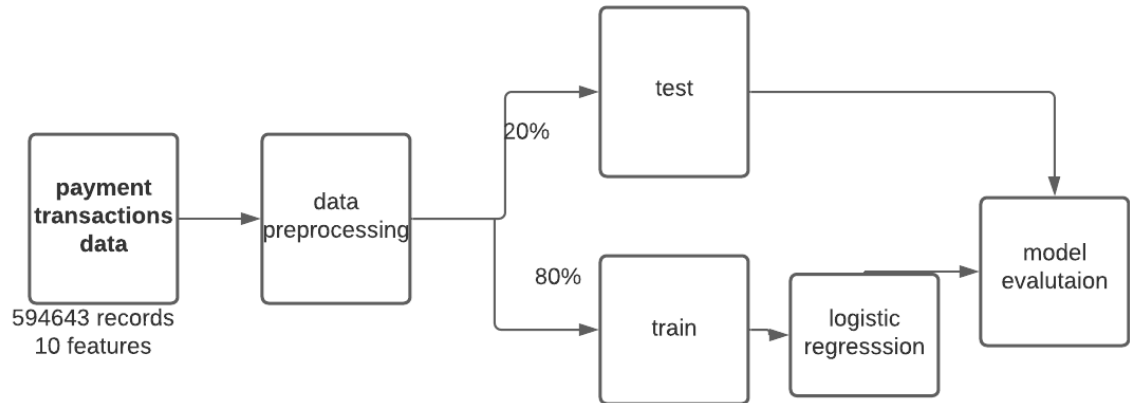


Figure 3: system architecture

## 2.6 SETUP AND TOOLS

- Operating System: Windows 10
- Required Software
  - Python 3.0
  - Google Collaboratory: An Online Python IDE for Machine Learning
  - Creately: A website providing tools to build the System Architecture and other related
  - Lucidchart: A website providing tools to build Gantt Chart designs
- REQUIRED LIBRARIES:
  - Python
  - Numpy
  - Pandas
  - Seaborn
  - Sklearn
  - CPU or GPU(recommended)

---

## 3 CHAPTER 3

### 3.1 DATA ANALYSIS

Importing the dataset, The data is fetched using a python library named pandas which consists of over 59000 records **Importing data set:**data is converted into the tabular form using a pandas data frame

**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; From the above output, the dataset has zero Null values **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.

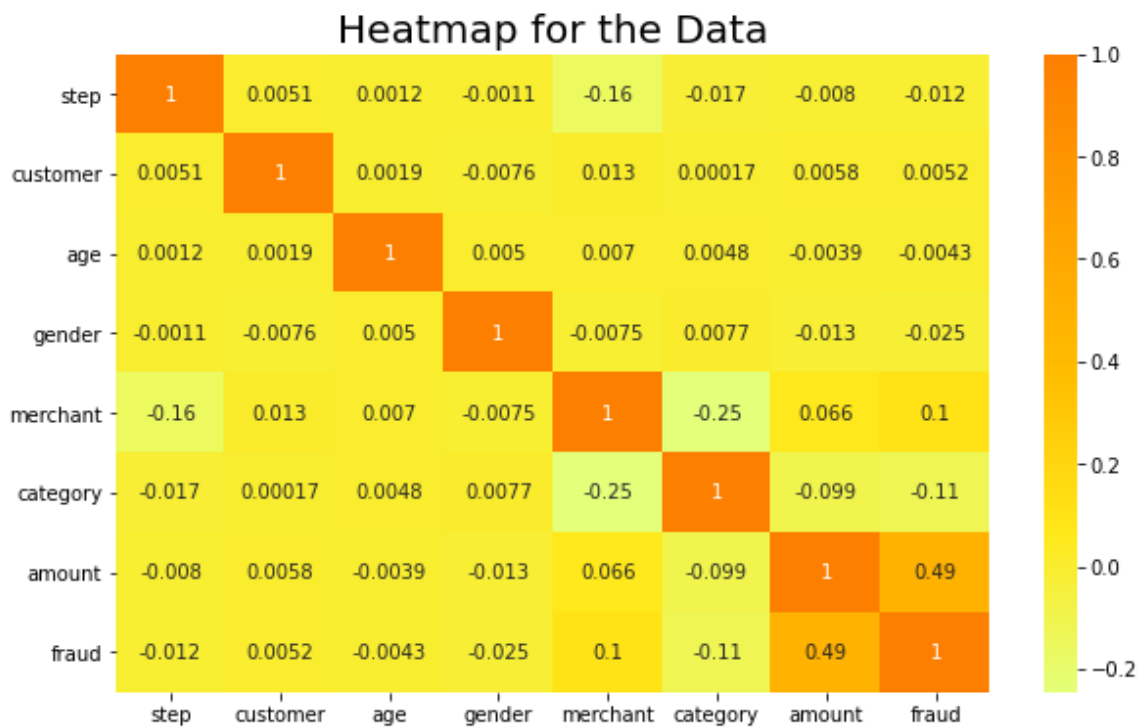


Figure 4: heatmap

With the above heatmap:

- 1) feature itself has proportionality constant=1,which is not useful for prediction
- 2) the map value with second max value results in the most dependent feature

---

counting no of fraud data in Dataset Grouped by category: Our hypotheses

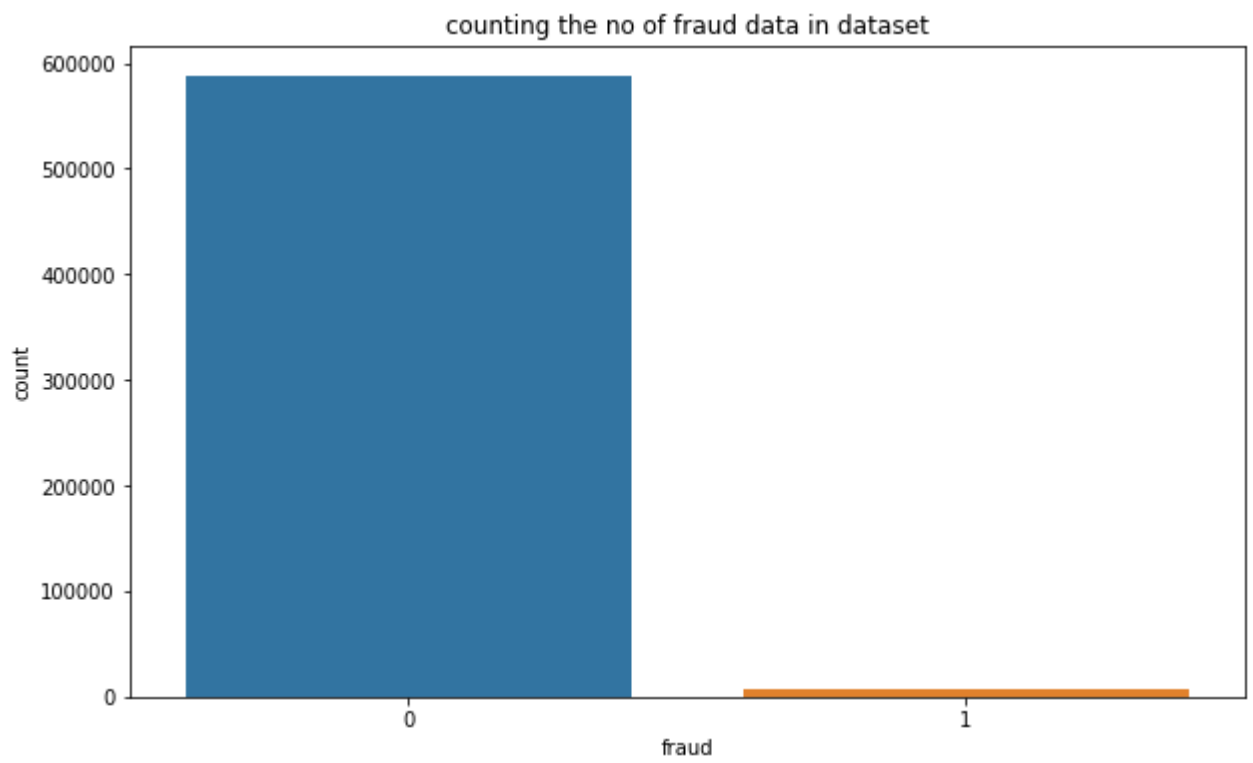



Figure 5: count of fraud data

is for fraudsters choosing the categories which people spend more.



	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 6: fraud vs non-fraud

but as we can see in the table above we can say confidently say that a fraudulent transaction will be much more (about four times or more) than average for that category.

Average amount spend it categories are similar; between 0-500 discarding the outliers, except for the travel category which goes very high

---

**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.

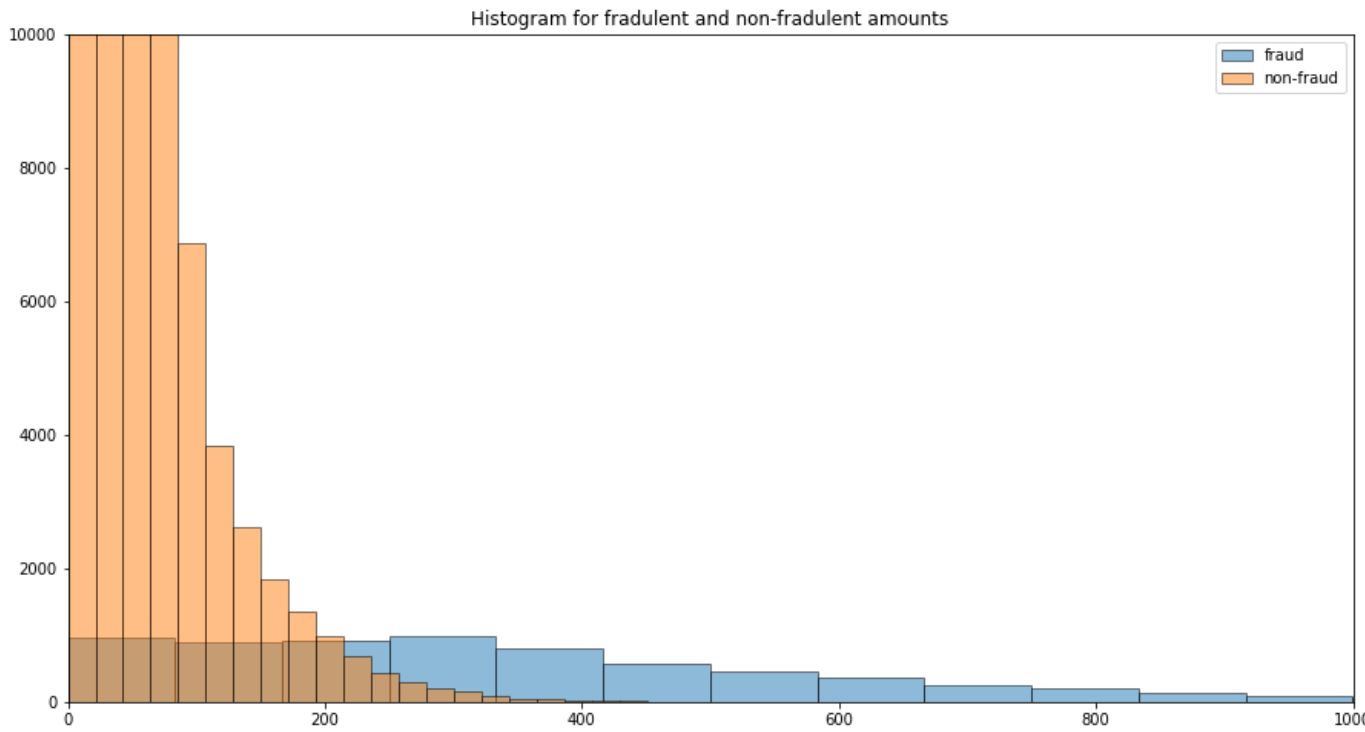


Figure 7: fraud vs non-fraud

---

## 3.2 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 Here we will trans-

```
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 8: dropping features

form categorical features into numerical values. It is usually better to turn these type of categorical values into dummies because they have no relation in

	step	customer	age	gender	merchant	category	amount	fraud
0	0	210	4	2	30	12	4.55	0
1	0	2753	2	2	30	12	39.68	0
2	0	2285	4	1	18	12	26.89	0
3	0	1650	3	2	30	12	17.25	0
4	0	3585	5	2	30	12	35.72	0

Figure 9: transforming 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 10: Undersampling

### 3.3 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 11: 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 12: 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 13: 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 co
the confusion matrix is =
[[2067  93]
 [ 75 2085]]
accuracy per= 96.64351851851852
```

Figure 14: Random Forest Algorithm

## Comparing Accuracies of the four Models:

MODEL	ACCURACY
Logistic Regression	91.38
K-Nearest Neighbor	92.66
Decision Tree Classification	96.11
Random Forest Algorithm	96.64

Table 2: Accuracy matrix

---

## 4 CONCLUSION

### 4.1 ADVANTAGES

:

- finding hidden and implicit correlations in the data
- automatic detection of possible fraud scenarios
- reduced number of verification measures
- real-time processing

### 4.2 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.

### 4.3 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 model's accuracy.

---

## 5 REFERENCES

[1],[3],[6],[2],[5],[4]

### References

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