

CREDIT CARD FRAUD DETECTION USING DATA ANALYTIC TECHNIQUES

K. VENGATESAN¹, A. KUMAR, S. YUVRAJ, V. D. AMBETH KUMAR, AND S. S. SABNIS

ABSTRACT. The Banking sector offers many features to their customers like ATM card, Internet banking, Gold Loan, Education Loan, Debit card and Credit card for attracting many customers to open account in the bank. In this work, we are going to propose the system for credit card fraud detection using machine learning algorithm. Generally the credit cards are used by customer 24x7, so bank server is able to keep track of all the transactions using machine learning algorithms. It needs to find or predict the fraud detections. The Data set consists of all the features of every transaction and we need to classify whether each transaction happens legally or not. In this work also we compared accuracy of logistical regression and KNN algorithms.

1. INTRODUCTION

Today utilization of Credit Cards even in developing nations has become a typical situation. Individuals use it to shop, pay bills, and for online exchanges. However, with increase in the number of Credit Card clients, the instances of misrepresentation in Credit Card have additionally been on the rise. Master-card related scams has caused universally billions of dollars. Misrepresentation can be named any movement with the plan of double-dealing to get monetary benefit in any way without the information of the cardholder and the guarantor bank. Charge card extortion can be done from multiple points of view. By lost or taken cards, by delivering phony or fake cards, by cloning the first site, by eradicating or adjusting the attractive strip present at the card which contains the

¹*corresponding author*

Key words and phrases. Prediction, Classification, Logistical Regression, Credit Card Fraud.

client's data, by phishing, by skimming or by taking information from a dealer's side. Extortion location manages finding a misrepresentation movement among a great many certifiable ones, which in certainty advances a challenge. With proceeded with headway in fake systems it is critical to create powerful models to battle these fakes in their underlying stage just, before they can take to fruition. In any case, the significant test in growing such a model is, that the quantity of deceitful exchanges among the absolute number of exchange is a very modest number and consequently crafted by finding a deceitful exchange in a successful and productive manner is very irksome.

- (1) Application Frauds: When the fraudster gains control of the application system by getting to delicate customer nuances like mystery state and username and open a fake record. It generally happens as indicated by the extortion. When the fraudster applies for credit or another MasterCard all around in the name of the card holder. The fraudster takes the supporting reports to help or approve their phony application.
- (2) Electronic or Manual Credit Card Imprints: When the fraudster skims information that is put on the alluring bit of the card. This information is very mystery and by getting to it the fraudster may use it for beguiling trades in future.
- (3) CNP (Card Not Present): When the fraudster knows the expiry date and record number of the card, the card can be used without its certifiable physical having a place.
- (4) Counterfeit Card Fraud: It is usually tried through the route toward skimming. A fake appealing swipe card is made and it holds all the nuances of the principal card. The fake card is totally valuable and can be used to submit trades in future.
- (5) Lost and Stolen Card Fraud: In circumstances when the first card holder loses their card, it can get to the hands of fraudsters and they would then have the option to use it to make portions. It is hard to do this through machine as a pin number is required regardless; online trades are straightforward enough for the fraudster

The following Figure 1 shows the different types of classification algorithms like logistical regression, Naive Bayes, Decision Tree, Support vector machine, Random forest and K-Nearest Neighbours Algorithm.

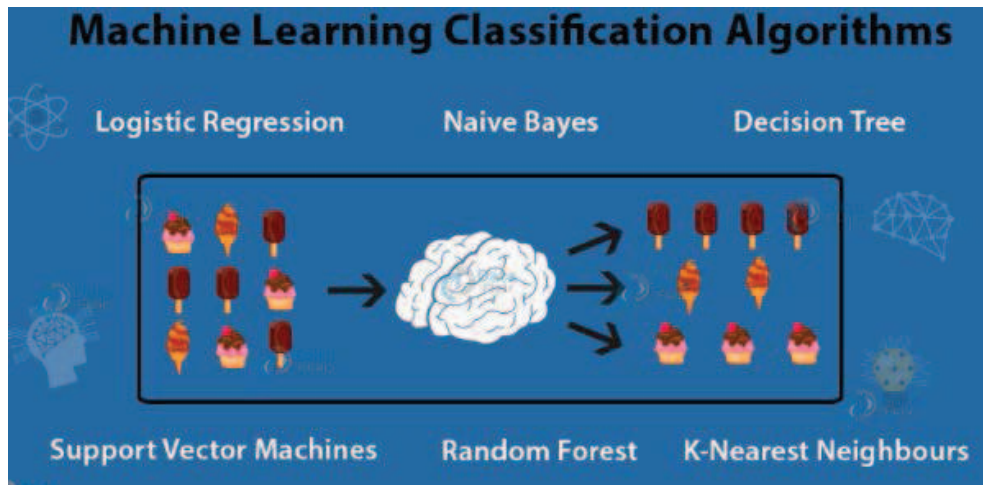


FIGURE 1. Classification algorithm types

2. PRELIMINARIES

Credit card fraud detection can be done using machine learning. Different Machine Learning approaches can be applied to this problem. In "Credit Card Fraud Detection: A case study" by Ayushi Agrawal, Shiv Kumar and Amit Kumar Mishra, combination of techniques is used like Genetic Algorithm, Behavior Based Technique and Hidden Markov Model. By this transaction is tested individually and whatever suits the best is further proceeded. And the foremost goal is to detect fraud by filtering the above techniques to get better result. For more details see [1–5].

In "Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier" by Masoumeh Zareapoor and Pourya Shamsolmoali]they trained various data mining techniques used in credit card fraud detection and evaluate each methodology based on certain design criteria. After several trial and comparisons; they introduced the bagging classifier based on decision tree, as the best classifier to construct the fraud detection model.

There are also many other techniques used to detect the fraudulent transactions made from a credit card. Like Credit card fraud detection using anti-k nearest neighbor algorithm and Behavior based credit card fraud detection using support vector machines.

3. PROPOSED SYSTEM

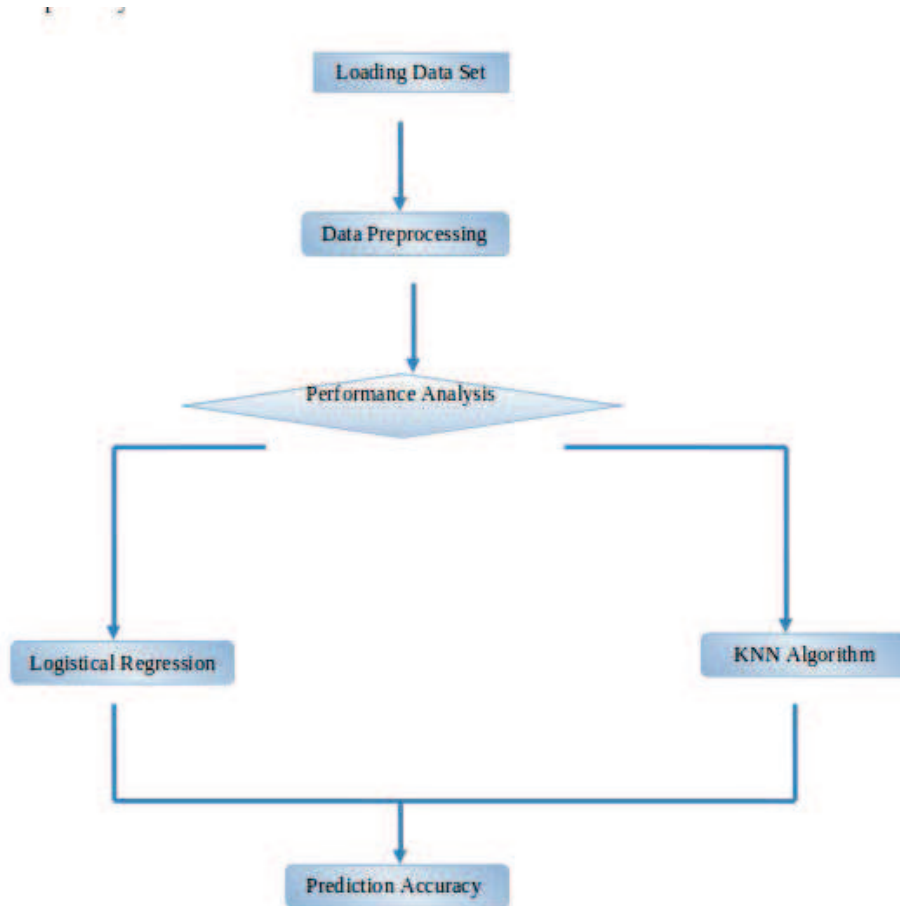
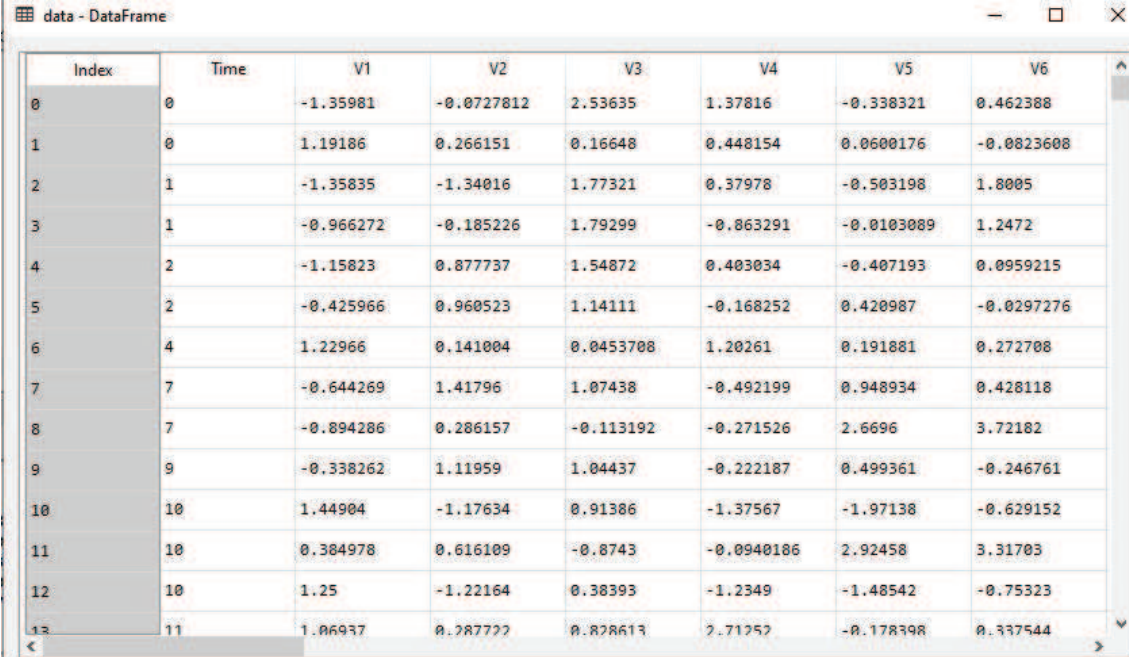


FIGURE 2. The Proposed System

The Figure 2 which represents the working model of the credit card fraud detection system, which takes dataset as input, then we need to apply for pre-processing techniques, which will find the error, noise or inconsistent values from data set and eliminated. This model works based on the classification algorithm, from which performance analysis should be measured using logistical regression and KNN algorithm, based on the training and test data set, also prediction accuracy is compared, which one will produce more accuracy.



Index	Time	V1	V2	V3	V4	V5	V6
0	0	-1.35981	-0.0727812	2.53635	1.37816	-0.338321	0.462388
1	0	1.19186	0.266151	0.16648	0.448154	0.0600176	-0.0823608
2	1	-1.35835	-1.34016	1.77321	0.37978	-0.503198	1.8005
3	1	-0.966272	-0.185226	1.79299	-0.863291	-0.0103089	1.2472
4	2	-1.15823	0.877737	1.54872	0.403034	-0.407193	0.0959215
5	2	-0.425966	0.960523	1.14111	-0.168252	0.420987	-0.0297276
6	4	1.22966	0.141004	0.0453708	1.20261	0.191881	0.272708
7	7	-0.644269	1.41796	1.07438	-0.492199	0.948934	0.428118
8	7	-0.894286	0.286157	-0.113192	-0.271526	2.6696	3.72182
9	9	-0.338262	1.11959	1.04437	-0.222187	0.499361	-0.246761
10	10	1.44904	-1.17634	0.91386	-1.37567	-1.97138	-0.629152
11	10	0.384978	0.616109	-0.8743	-0.0940186	2.92458	3.31703
12	10	1.25	-1.22164	0.38393	-1.2349	-1.48542	-0.75323
13	11	1.06937	0.287722	0.828613	2.71252	-0.178398	0.337544

FIGURE 3. Credit Card Data Set

4. RESULT AND DISCUSSION

The Figure 3 which represents the same credit card data set with 31 attributes and 284807 row last two columns like amount class are most important attribute for this fraud detection. Using python coding we are going to convert class attribute into two category like fraud and normal after conversion the count of fraud from the data set is 492 and count of normal from data set is 284315. By default the data set consists of all the attributes numerical type like float and integer, so it will be easy to apply regression and classification algorithm for numerical value.

5. PREPROCESSING

The Figure 4 represents the data set after preprocessing, initial data set contains total 31 columns after applying preprocessing, there is no any missing values is available in the source .

The Figure 5 which represents list top 14 transactions with maximum amount is fraud transaction from the Credit card dataset along with time column.

Index	0
Time	0
V1	0
V2	0
V3	0
V4	0
V5	0
V6	0
V7	0
V8	0
V9	0
V10	0
V11	0
V12	0
V13	0

FIGURE 4. Credit card dataset after processing

Index	Time	Amount	Class
176049	122608	2125.87	1
6971	9064	1009.68	1
249167	154278	1504.93	1
89190	62467	1402.16	1
81609	59811	1389.56	1
95597	65385	1354.25	1
199896	133184	1335	1
10690	18888	1218.89	1
249239	154309	1096.99	1
233258	147501	996.27	1
203528	134769	925.31	1
146790	87883	829.41	1
107637	70536	824.83	1
44001	41743	802.52	1

FIGURE 5. List of maximum amount fraud transaction

The Figure 6 shows list of fraud transaction along with transaction amount are greater than 100 from the credit card data set, result generated total 130 records. The maximum fraud transaction from the credit card is 2125.87.

The Figure 7 demonstrates the statistical information like count, mean, standard deviation, minimum and maximum values for each and every 31 attributes for the data set. Given data set final prediction attribute need to change into two categories like either transaction is normal or fraud, if the values class type

Index	Time	Amount	Class
623	472	529	1
4920	4462	239.93	1
6971	9064	1809.68	1
8972	12393	179.66	1
10630	17838	766.36	1
10690	18888	1218.89	1
10891	18675	188.78	1
16863	28242	730.86	1
20198	30852	104.81	1
23422	32745	717.15	1
27738	34684	125.3	1
30496	35953	111.7	1
31002	36170	111.7	1
39183	39729	776.83	1

FIGURE 6. Fraud transaction those amount more that 100

Index	Time	V1	V2	V3
count	284807	284807	284807	284807
mean	94813.9	3.91956e-15	5.68817e-16	-8.76907e-15
std	47488.1	1.9587	1.65131	1.51626
min	0	-56.4075	-72.7157	-48.3256
25%	54201.5	-0.920373	-0.59855	-0.890365
50%	84692	0.0181088	0.0654856	0.179846
75%	139320	1.31564	0.883724	1.0272
max	172792	2.45493	22.0577	9.38256

FIGURE 7. Statistical Information about Credit card Dataset

is 1 mean consider as fraud and 0 means normal, and we need to count list of values in each category.

6. LOGISTICAL REGRESSION

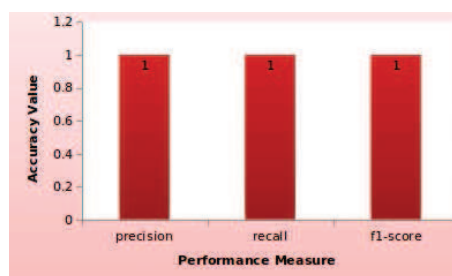


FIGURE 8. Classification report for normal Transaction

Classification is one of the most basic bits of composed learning. In this work, we will investigate the various course of action checks like essential lose the

confidence, perfect bayes, choice trees, optional timberlands and some more. We will experience the entirety of the check's association properties and how they work, and this can be observed in Figure 8.

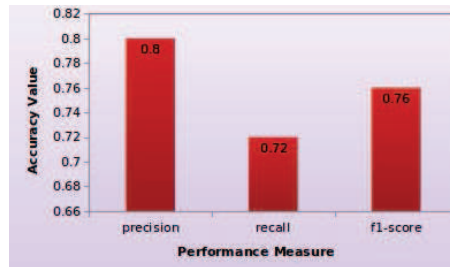


FIGURE 9. Classification report for fraud Transaction

The Figure 9 which represents the classification accuracy of credit card transaction for normal like class category will be 0. As per following three statistical measure like precision, recall and F1-Score value are calculated confusion matrix of training and test data set, the accuracy value of above three parameters are same value is 1.

```
In [140]: print(accuracy_score(y_test,y_pred))
0.9991673605328892

In [141]: print(classification_report(y_test,y_pred))
precision    recall  f1-score   support

0           1.00      1.00      1.00     99502
1           0.80      0.72      0.76       181
```

FIGURE 10. Logistical Regression accuracy

The Figure 10 which represents the fraud transaction details, X-axis represents the performance measure like precision, recall and f1-score, Y-axis represents the accuracy values corresponding attributes. The figure represents the accuracy value of logistical regression against training and test data set of credit card data set. The accuracy value is 0.999 calculated based on the confusion matrix and logistical regression.

The code for calculating accuracy is as follows:

```
from sklearn import linear_model
from sklearn.model_selection import train_test_split
x=data.iloc[:, : 1]
y=data[ 'Class ']
```

```

x_train , x_test , y_train , y_test = ...
train_test_split(x,y, test_size=0.35)
clf=linear_model.LogisticRegression(C=1e5)
clf.fit(x_train , y_train)
y_pred=np.array(clf.predict(x_test))
y=np.array(y_test)
from sklearn.metrics import confusion_matrix , ...
classification_report , accuracy_score
print(confusion_matrix(y_test , y_pred))
print(accuracy_score(y_test , y_pred))
print(classification_report(y_test , y_pred))

```

7. K-NEAREST NEIGHBOUR ALGORITHM

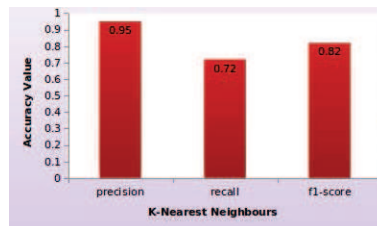


FIGURE 11. K Nearest Neighbours Algorithm

The Figure 11 which represents the KNN algorithm implementation of credit card data set X-axis represents various measurements parameter precisions, recall, and f1-score and Y-axis represents accuracy score of each and every parameter. As per the statistical measure the precision value is 0.95, recall value is 0.72 and f1-score value is 0.82. Figure 12 which represents the accuracy mea-

```

K-Nearest Neighbours
Confusion Matrix
tn = 28420 fp = 2
fn = 16 tp = 43
Scores
Accuracy --> 0.9993679997191109
Precision --> 0.9555555555555556
Recall --> 0.7288135593228338
F1 --> 0.8269230769230769

```

FIGURE 12. K-Nearest Neighbours Accuracy Measure

sure of KNN algorithm, which is calculate based on the training and test data set, as per result the accuracy value of KNN is 0.99.

```
knn=KNeighborsClassifier(n_neighbors = 5,...
algorithm="kd_tree", n_jobs = 1)
knn.fit(train_features , train_labels.ravel())
knn_predicted_test_labels=knn.predict(test_features)
#calculating confusion matrix for knn
tn , fp , fn , tp=confusion_matrix(test_labels ,...
knn_predicted_test_labels).ravel()
#scoring knn
knn_accuracy_score=accuracy_score(test_labels ,...
knn_predicted_test_labels)
knn_precison_score=precision_score(test_labels ,...
knn_predicted_test_labels)
knn_recall_score=recall_score(test_labels ,...
knn_predicted_test_labels)
knn_f1_score=f1_score(test_labels ,...
knn_predicted_test_labels)
#printing
print("")
print("K Nearest Neighbours")
print("Confusion Matrix")
print("tn =",tn , " fp =",fp)
print("fn =",fn , " tp =",tp)
print("Scores")
print("Accuracy    > ",knn_accuracy_score)
print("Precison    > ",knn_precison_score)
print("Recall      > ",knn_recall_score)
print("F1         > ",knn_f1_score)
```

8. CONCLUSION

The banking sector providing many services to customers, the credit card is one of the important service offering major banks. In spite of the way that there are a couple of fraud detection strategies open today anyway none can distinguish all fakes absolutely when they are truly happening, they, generally, remember it after the fraud has been submitted. This occurs considering the way that an infinitesimal number of trades from the total trades are truly phony in nature. So we need a developer that can perceive the bogus trade when it is happening with the objective that it might be ended immediately and that too in a base cost. So the huge task of today is to manufacture a precise, definite and speedy distinctive distortion revelation structure for Visa swindles that can recognize not simply fakes happening over the web like phishing and site cloning yet furthermore meddling with the charge card itself, for instance, it signals an alert when the modified charge card is being used. This is proposed with discussed about how machine learning algorithm will be used for credit card fraud detection, based on the customer's transaction. In this work accuracy of the system is predicted using machine classification algorithms like logistical regression and KNN algorithm based on the training and test data set, the KNN algorithm is produced best result such as statistical measure the precision value is 0.95, recall value is 0.72 and f1-score value is 0.82.

REFERENCES

- [1] A. A. PANSY KHURANA: *Credit Card Fraud Detection using Fuzzy Logic and Neural Network*, SpringSim, 2016.
- [2] R. M. JAMAIL ESMAILY: *Intrusion Detection System based on Multilayer Perceptron Neural Networks and Decision Tree*, in International Conference on Information and Knowledge Technology, 2015.
- [3] S. P. TANMAY KUMAR BEHERA: *Credit Card Fraud Detection: A Hybrid Approach using Fuzzy Clustering and Neural Network*, in International Conference on Advances in Computing and Communication Engineering, .
- [4] N. W. WEN -FANG YU: *Research on Credit Card Fraud Detection Model based on Distance Sum*, in International Joint Conference on Artificial Intelligence, Hainan Island, China, 2009.
- [5] E. D. Y. SAHIN: *Detecting Credit Card Fraud by Decision Trees*, in Proceedings of the International Multiconference of Engineers and Computer Science, Hong Kong, 2011.

SANJIVANI COLLEGE OF ENGINEERING
SAVITRIBAI PHULE PUNE UNIVERSITY
PUNE MAHARAHTRA, INDIA

DEPARTMENT OF COMPUTER SCIENCE
BANARAS HINDU UNIVERSITY
VARANASI, INDIA

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPT. OF ECE
CHENNAI, INDIA

DEPARTMENT OF CSE
PANIMALAR ENGINEERING COLLEGE
ANNA UNIVERSITY, CHENNAI, INDIA

SANJIVANI COLLEGE OF ENGINEERING
SAVITRIBAI PHULE PUNE UNIVERSITY
PUNE MAHARAHTRA, INDIA