Flip01 Project

Jingbao Luo

Nanjing University of Science and Technology

2023-07-30

Overview

Introduction

- Introduction

Method

- *iForest*

Data Exploration

- Data Visualization
- Data Visualization

Experiment

- Parameter adjustment
- Parameter adjustment
- Parameter adjustment

Conclusion

Introduction

Introduction

The main implementation in this projecct is utilizing the Isolation Forest algorithm to achieve credit card fraud detection. By continuously fine-tuning the algorithm through parameter adjustments, the article aims to improve the algorithm's performance.

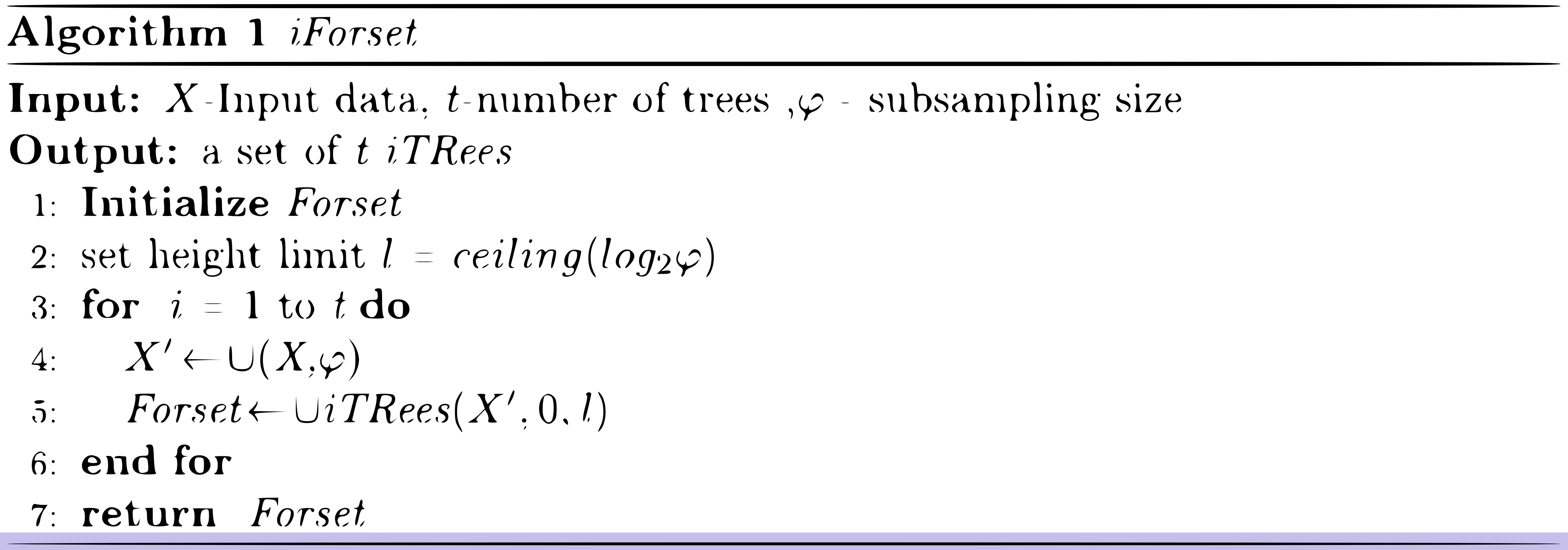
This project utilizes a dataset sourced from Kaggle to implement credit card fraud detection. The dataset contains transaction information of European cardholders during September 2013, conducted through credit cards. The dataset covers transactions that occurred within a span of two days, with a total of 284,807 transactions. Among these transactions, there were 492 cases of fraud, making the dataset highly imbalanced, where the positive class (fraudulent transactions) accounts for only 0.172

Method

iForest

The Isolation Forest algorithm, proposed in 2008 by Liu Fei, Zhou Zhihua, and others, does not rely on distance or density metrics to describe the differences between samples and other samples. Instead, it directly characterizes the so-called isolation level. Therefore, this algorithm is simple, efficient, and widely used in the industry.

The logic of the Isolation Forest algorithm is intuitive. It uses binary trees to split the data, and both sample selection and feature selection are performed using randomization. If a certain sample is an outlier, it may require very few iterations to be isolated.



```
Algorithm 1 $iForset$
Input: $X$-Input data, $t$-number of trees ,$\varphi$ - subsampling size
Output: a set of $t$ $iTRees$
Initialize $Forset$
set height limit $l = ceiling(log_2\varphi)$
for $i = 1$ to $t$ do
    $X' \leftarrow \cup(X,\varphi)$
    $Forset \leftarrow \cup iTRees(X',0,l)$
end for
return $Forset$
```

Data Exploration

Data Description

Table provides a basic overview of the data for each attribute, including measures such as the mean, maximum, minimum, and other indicators.

Table 1: Data description

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
count	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000
mean	94813.8596	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
std	47488.1460	1.9587	1.6513	1.5163	1.4159	1.3802	1.3323	1.2371	1.1944	1.0986	1.0888	1.0207	0.9992	0.9953	0.9586
min	0.0000	-56.4075	-72.7157	-48.3256	-5.6832	-113.7433	-26.1605	-43.5572	-73.2167	-13.4341	-24.5883	-4.7975	-18.6837	-5.7919	-19.2143
0.25	54201.5000	-0.9204	-0.5985	-0.8904	-0.8486	-0.6916	-0.7683	-0.5541	-0.2086	-0.6431	-0.5354	-0.7625	-0.4056	-0.6485	-0.4256
0.5	84692.0000	0.0181	0.0655	0.1798	-0.0198	-0.0543	-0.2742	0.0401	0.0224	-0.0514	-0.0929	-0.0328	0.1400	-0.0136	0.0506
0.75	139320.5000	1.3156	0.8037	1.0272	0.7433	0.6119	0.3986	0.5704	0.3273	0.5971	0.4539	0.7396	0.6182	0.6625	0.4931
max	172792.0000	2.4549	22.0577	9.3826	16.8753	34.8017	73.3016	120.5895	20.0072	15.5950	23.7451	12.0189	7.8484	7.1269	10.5268

	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
count	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000	284807.0000
mean	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	88.3496	0.0017
std	0.9153	0.8763	0.8493	0.8382	0.8140	0.7709	0.7345	0.7257	0.6245	0.6056	0.5213	0.4822	0.4036	0.3301	250.1201	0.0415
min	-4.4989	-14.1299	-25.1628	-9.4987	-7.2135	-54.4977	-34.8304	-10.9331	-44.8077	-2.8366	-10.2954	-2.6046	-22.5657	-15.4301	0.0000	0.0000
0.25	-0.5829	-0.4680	-0.4837	-0.4988	-0.4563	-0.2117	-0.2284	-0.5424	-0.1618	-0.3546	-0.3171	-0.3270	-0.0708	-0.0530	5.6000	0.0000
0.5	0.0481	0.0664	-0.0657	-0.0036	0.0037	-0.0625	-0.0295	0.0068	-0.0112	0.0410	0.0166	-0.0521	0.0013	0.0112	22.0000	0.0000
0.75	0.6488	0.5233	0.3997	0.5008	0.4589	0.1330	0.1864	0.5286	0.1476	0.4395	0.3507	0.2410	0.0910	0.0783	77.1650	0.0000
max	8.8777	17.3151	9.2535	5.0411	5.5920	39.4209	27.2028	10.5031	22.5284	4.5845	7.5196	3.5173	31.6122	33.8478	25691.1600	1.0000

Data Visualization

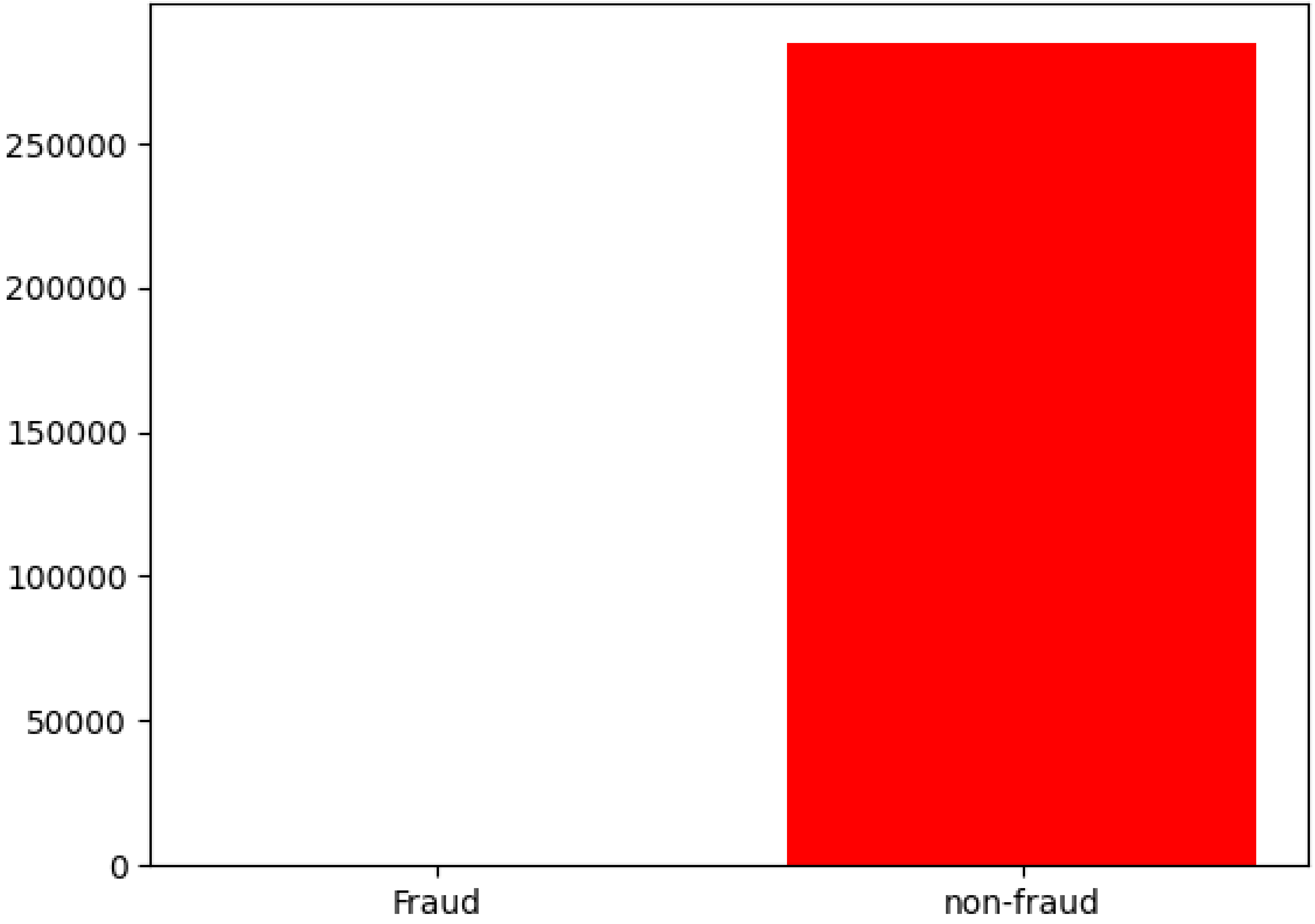


Figure 1: fraud or no fraud

Data Visualization

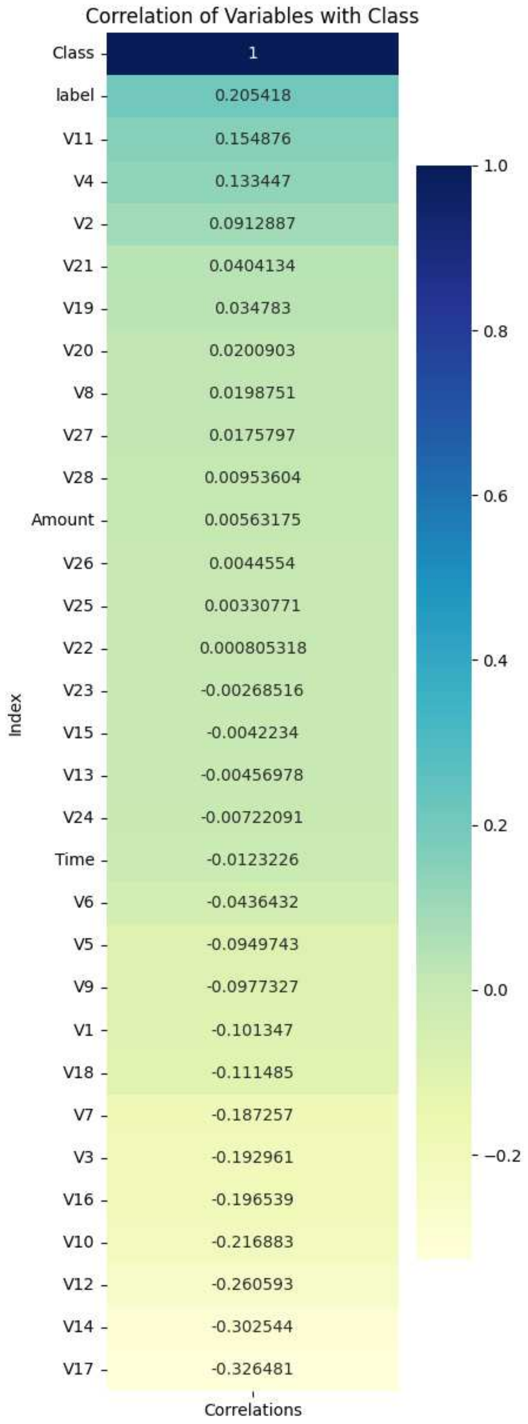


Figure 2: correlation heat plot

Experiment

Experiment

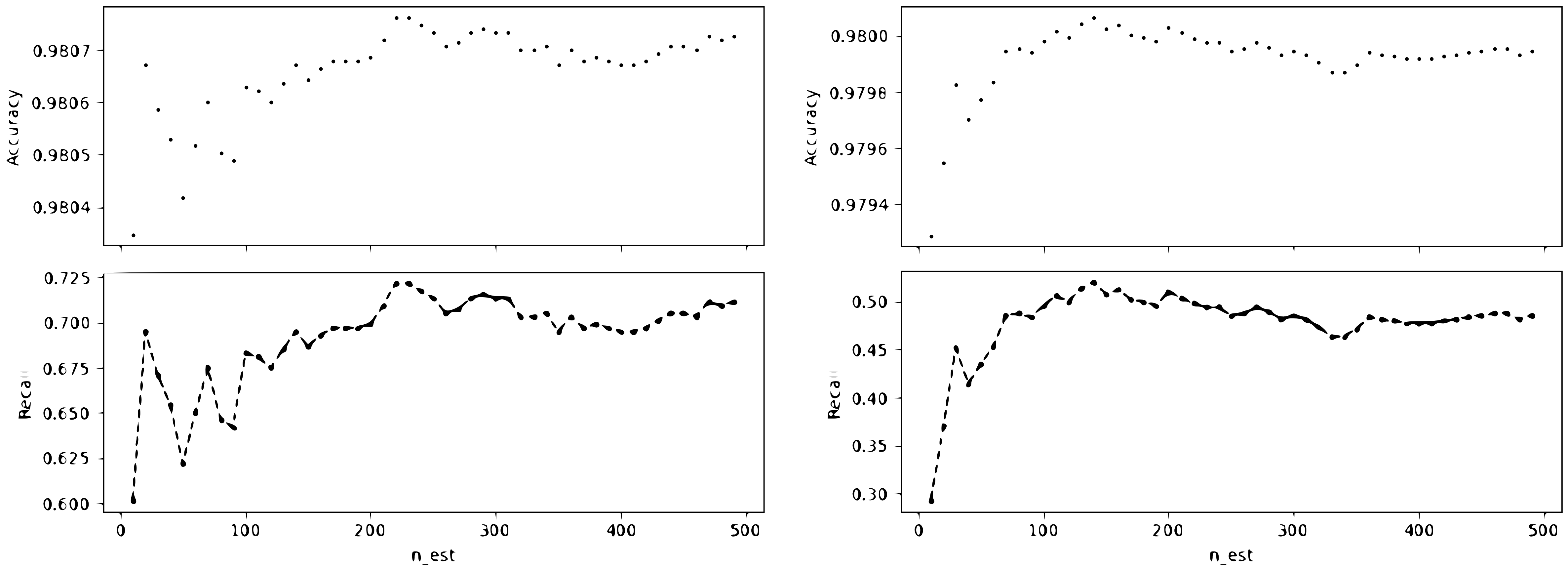
In the experimental design, we achieved the optimal performance of the model by continuously adjusting parameters. The evaluation of the training set's accuracy and recall mainly involved removing attributes and continuously changing the parameters of the isolation forest.

Table 2: parameter adjustment

Parameter	Accuarcy	Recall	Parameter	Accuarcy	Recall
none-parameter-adjust	0.963	0.821	none-parameter-adjust(feature _delete)	0.953	0.663
n_estimators-adjust(220)	0.981	0.722	n_estimators-adjust(feature _delete)(140)	0.980	0.520
n_features-adjust(11)	0.981	0.750	n_features-adjust(feature _delete)(11)	0.981	0.522
max_samples-adjust	0.981	0.829	nmax_samples-adjust(feature _delete)	0.980	0.634

Parameter adjustment

Introduction
Method
Data Exploration
Experiment
Parameter adjustment
Parameter adjustment
Parameter adjustment
Conclusion

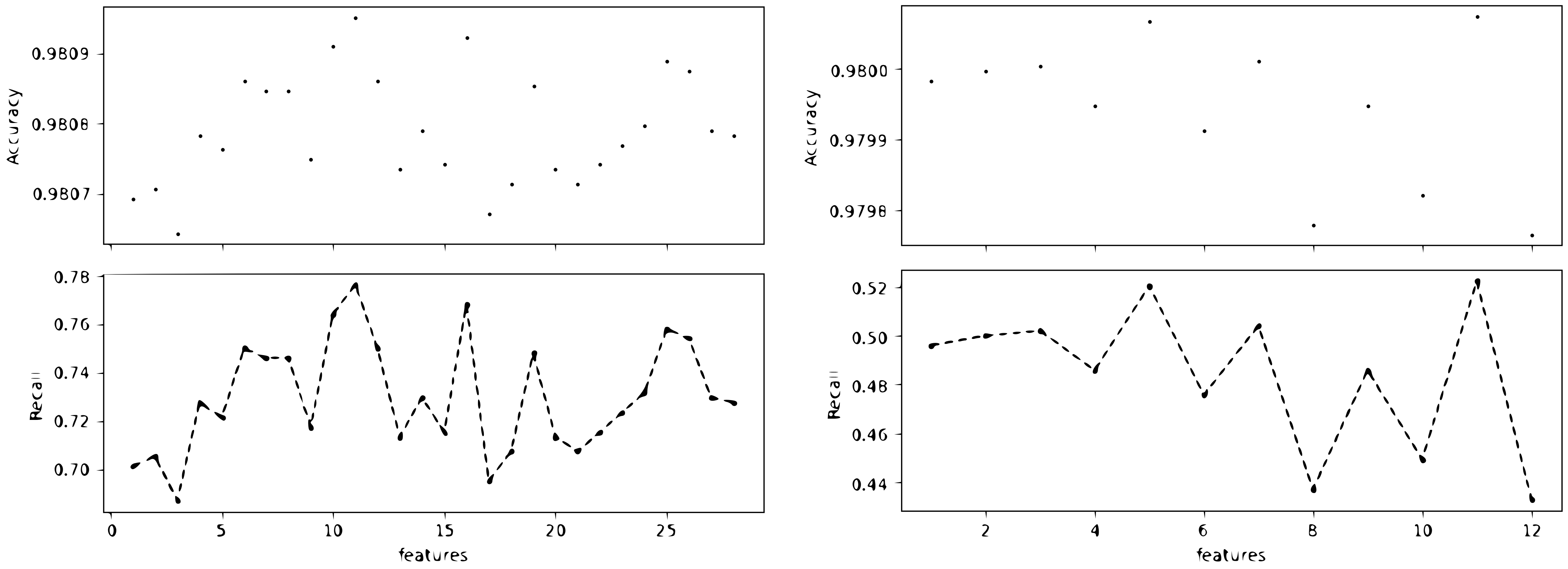


(a) n_estimators-adjust

(b) n_estimators-adjust (feature_delete)

Figure 3: n_estimators-adjust

Parameter adjustment

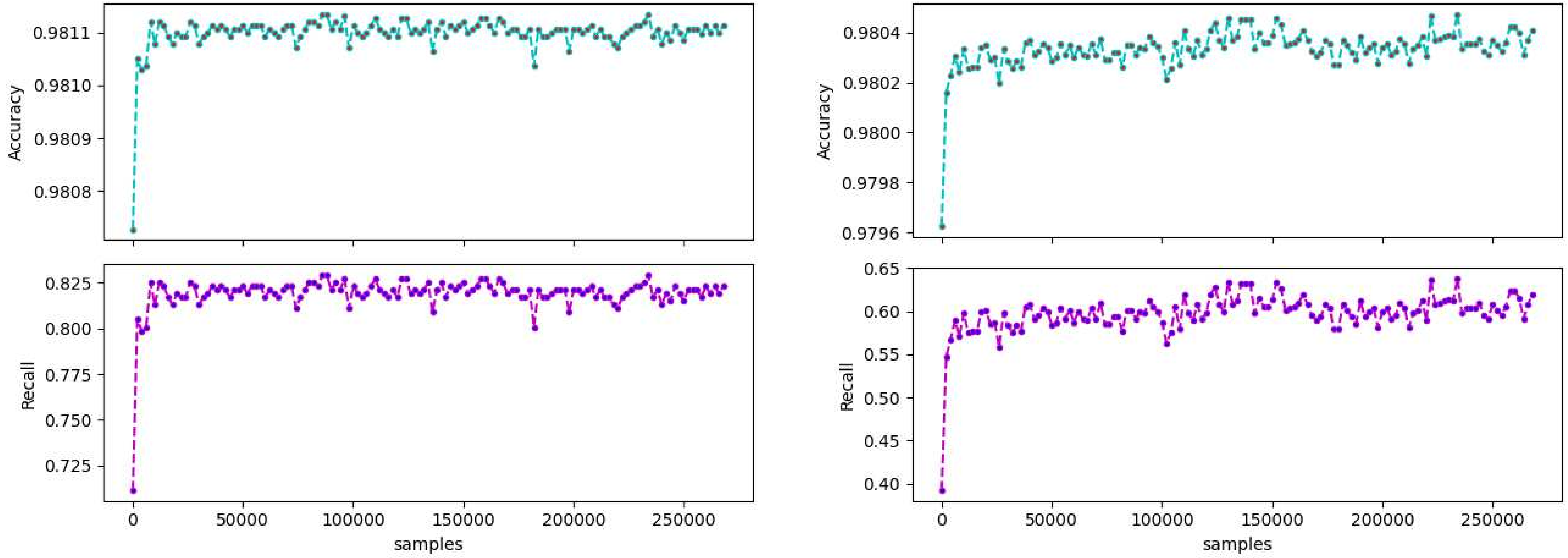


(a) n_features-adjust

(b) n_features-adjust (feature_delete)

Figure 4: n_efeatures-adjust

Parameter adjustment



(a) max_samples-adjust

(b) max_samples-adjust (feature_delete)

Figure 5: max_samples-adjust

Conclusion

Conclusion

Introduction
Method
Data Exploration
Experiment
Conclusion

The detection of credit card fraud data was achieved through Isolation Forest, and by continuously adjusting the parameters of the algorithm, a high level of accuracy and recall was ultimately reached, at 0.98 and 0.8, respectively.

However, a notable drawback of this method is the lack of a more in-depth analysis of attributes. This could mean that in the data preprocessing stage or feature selection process, there was insufficient exploration, understanding, and utilization of attribute information relevant to fraud. While optimizing the accuracy and recall of the model is undoubtedly important, equal attention should also be given to the importance and correlation of attributes, as well as their role in fraud detection.