Credit Card Fraud Detection

Lin Jiahong

Nanjing University of Science and Technology

2023-07-29

Overview

Problem Definition

Credit Card Fraud Detection

Data Analysis and Preprocessing

Unsupervised and Supervised Anomaly Detection Methods

Model Train and Result

Conclusion

Problem Definition

Credit Card Fraud Detection

Defn

Credit Card Fraud Detection aims to identify the presence of fraudulent credit card use through the characteristics of credit card use.

- Data covers Preprocessing features , Time and Amount.
- The data comes with a label of behavior category.

Class	row_num
All	284807
Normal	284315
Fraud	492

Data Analysis and Preprocessing

Overall data

- Time- Transaction time of day in seconds.
- 28 Features- ['V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10', 'V11','V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28']
- Amount - Real transaction amount
- Class - 0 for normal behavior, 1 for abnormal behavior
- This anomaly detection task dataset is imbalanced datasets.

Amount and Time Distribution

- Fraud is smaller in amount and more evenly distributed over time.

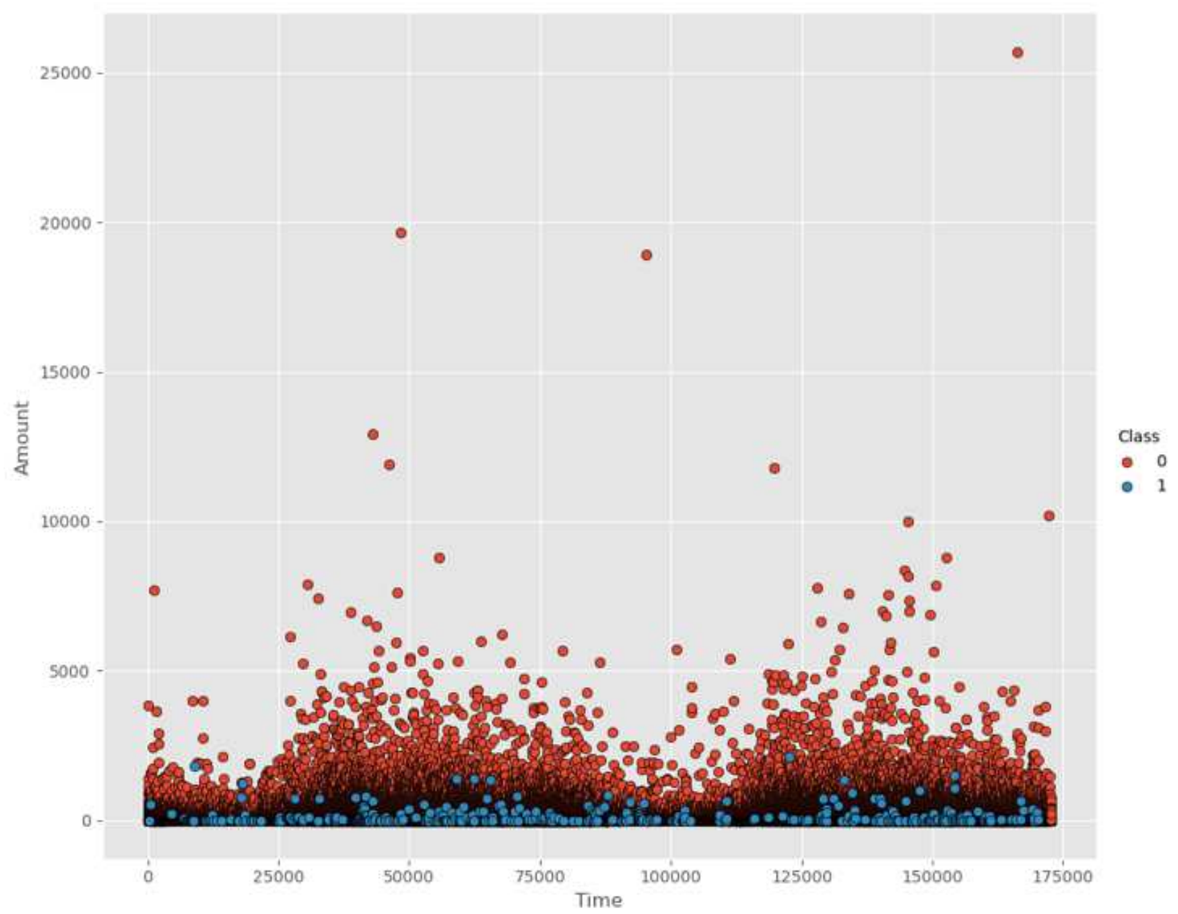


Figure 1: Amount and Time Distribution Scatterplot

Correlation Matrices

- Negative Correlations: V17, V14, V12 and V10 are negatively correlated.
  Positive Correlations: V2, V4, V11, and V19 are positively correlated.

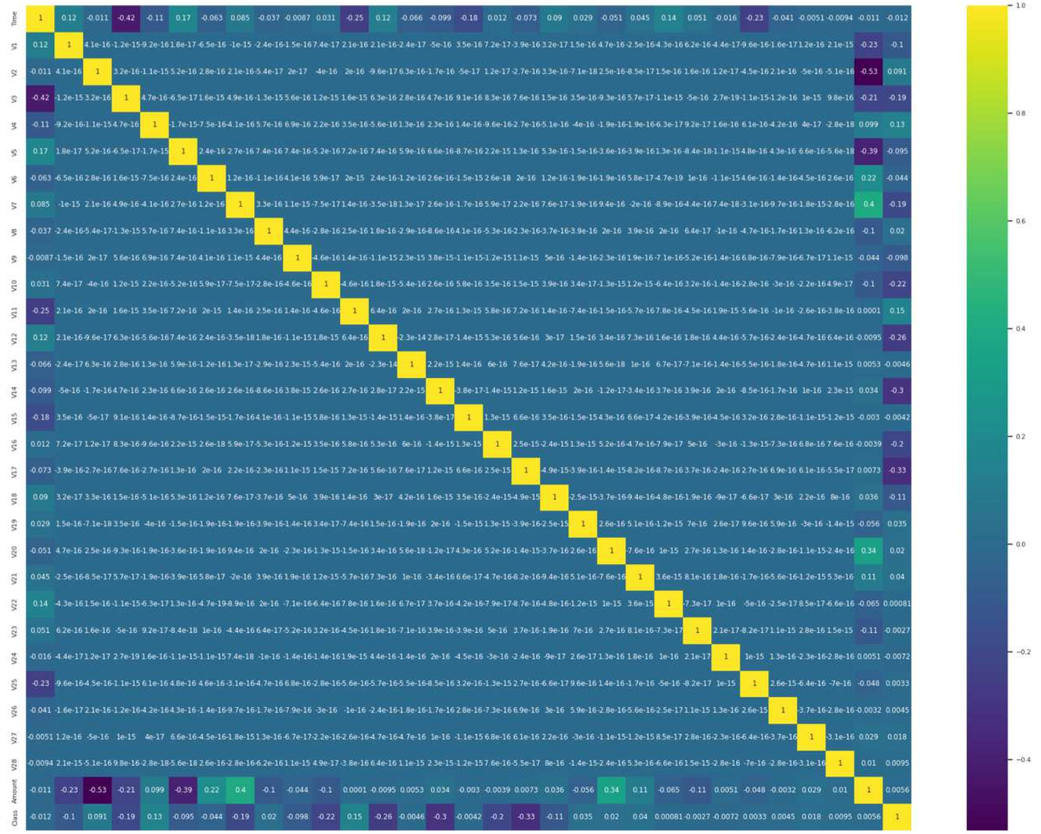


Figure 2: Correlation Matrices

PCA and t-SNE visualization

- There is an overlap between outliers and normal in PCA, while outliers are more independent in t-SNE.

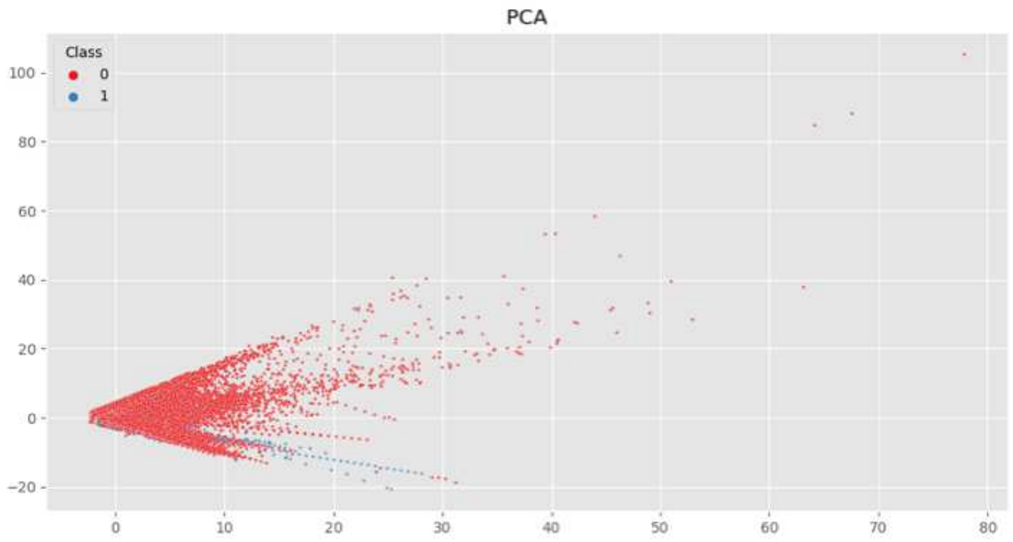


Figure 3: PCA

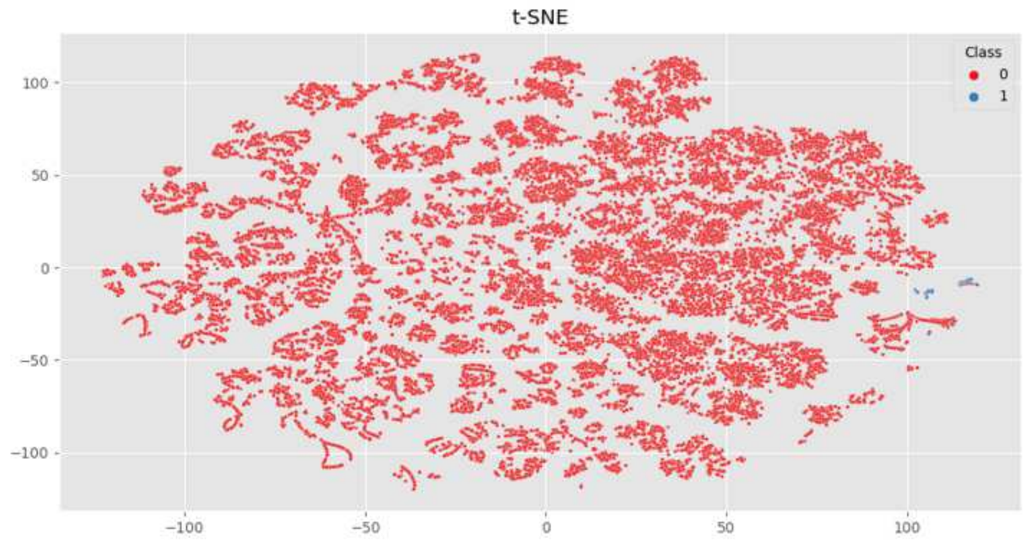


Figure 4: t-SNE

Data Preprocessing

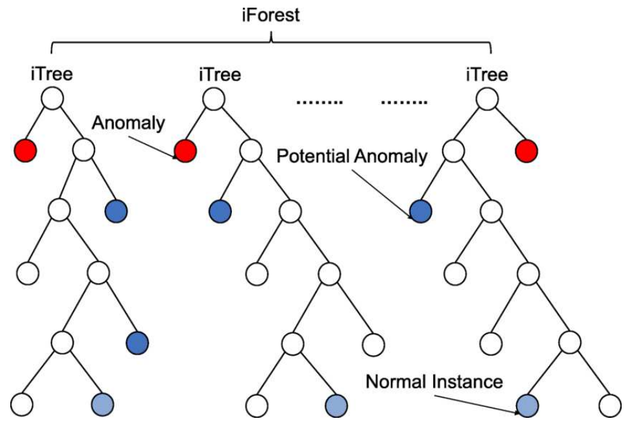
- Normalize the Amount and Time columns in the data.
- Divide the train set and test set.
- 30 Features- ['Time','Amount','V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10', 'V11','V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28'].

Data	row_num	Normal	Fraud
train	227845	227468	377
test	57339	56847	115

Unsupervised and Supervised Anomaly Detection Methods

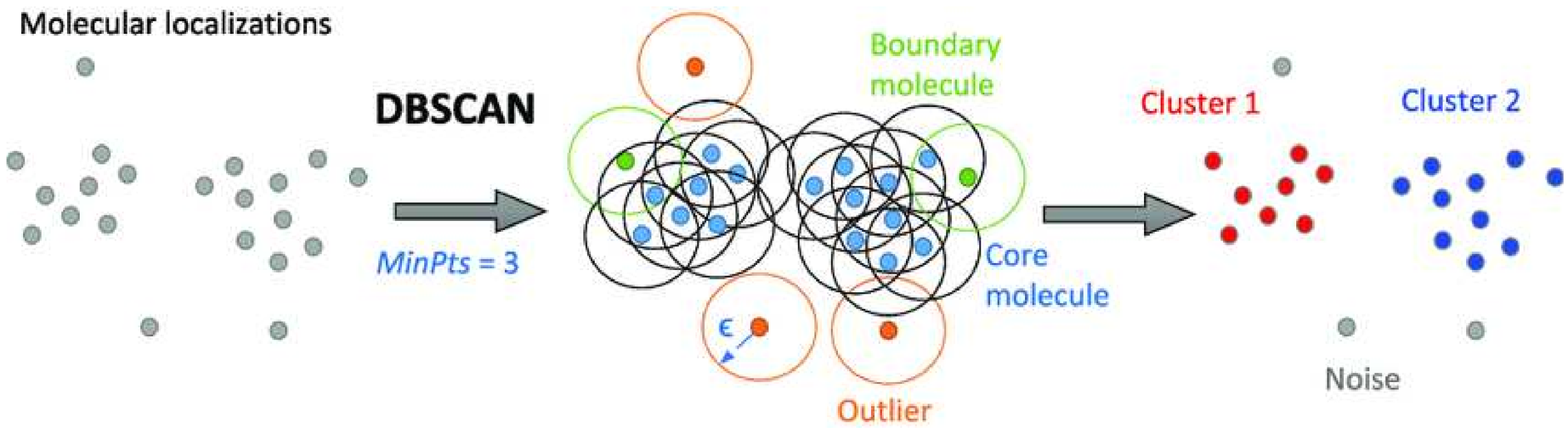
Unsupervised - Isolation Forest

- Isolation Forest(IF) is build based on decision trees. No pre-defined labels here. An unsupervised learning algorithm.
  1. Two quantitative properties of anomalous data points:Outliers are few and their features are very different from normal points.
  2. Not assume normal distribution and Detect outliers at a multi-dimensional level.
  3. Isolation Forest is computationally efficient: a low constant and a low memory requirement.
  4. Parameters - **Number of estimators**, **Max samples**, **Contamination**, **Max features**



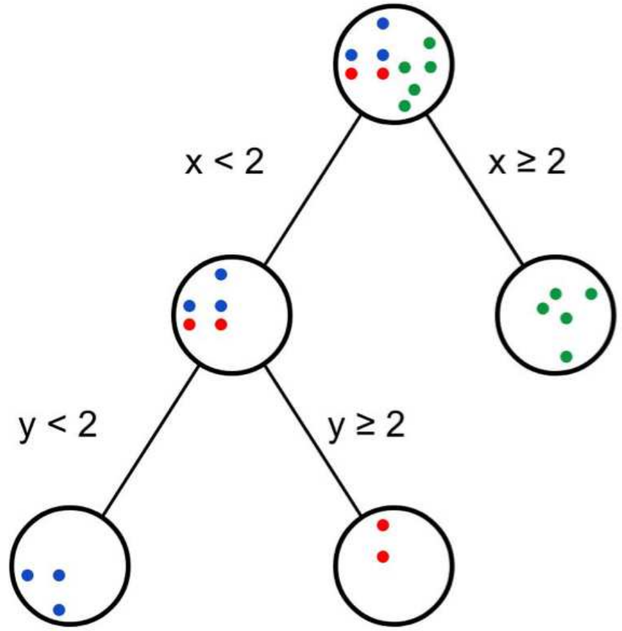
Unsupervised - DBSCAN

- DBSCAN is a powerful density-based data clustering algorithm.
  1. DBSCAN algorithm separates the high-density regions of the data from the low-density areas.
  2. Detect outliers by identifying noise.
  3. Parameters - **Epsilon**, **minPoints**



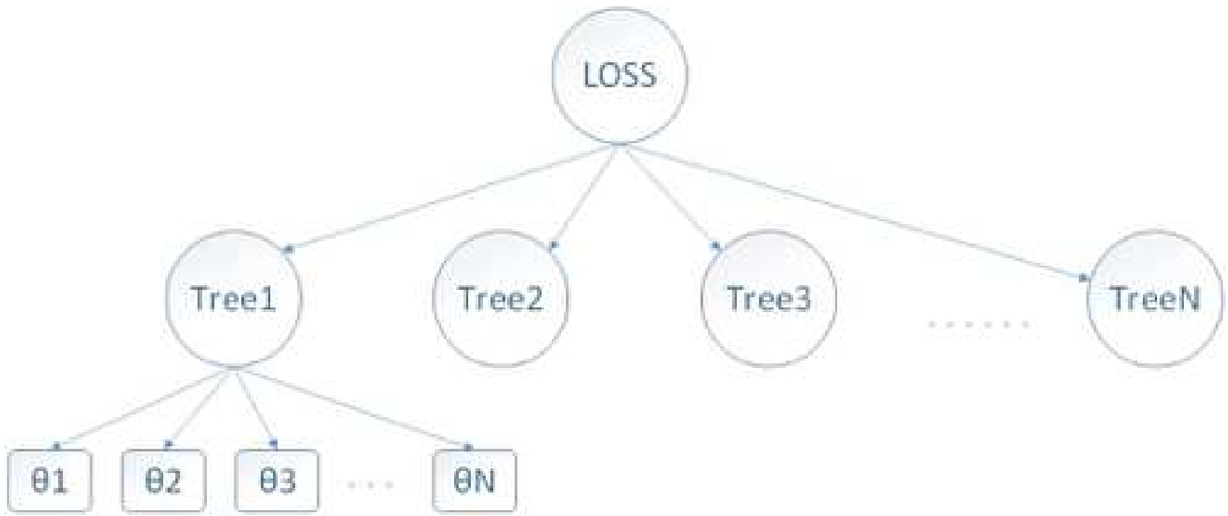
Supervised - Random Forest

- Random Forests perform classification by constructing multiple decision trees and combining their predictions.
  1. Constructing a decision tree from the train set.
  2. Prediction of test sets and feature importance exploration
  3. Parameters - **n_estimators**, **max_depth**, **min_samples_leaf**, **min_samples_split**



Supervised - XGBoost

- XGBoost (eXtreme Gradient Boosting) is a gradient boosting tree algorithm.
  1. The core principle is to combine multiple weak learners (decision trees) into one strong learner
  2. Decision trees are trained in an iterative manner to train new trees based on the residuals between the predictions and the actual labels of all the previous trees in order to gradually reduce the error.
  3. Parameters - **n_estimators**, **max_depth**, **learning_rate**, **subsample**, **colsample_bytree**



Model Train and Result

Model Train

- Each method has been subjected to a parameter network search and parameter tuning.Parameters not mentioned use default values.

Method	Type	Train	Test
Isolation Forest	*Unsupervised*	*all data*	*all data*
DBSCAN	*Unsupervised*	*all data*	*all data*
Random Forest	*Supervised*	*train*	*test*
XGBoost	*Supervised*	*train*	*test*

Method	Parameters
Isolation Forest	$n\_estimators = 1000, contamination = 0.00172, max_features = 1.0$
DBSCAN	$eps = 3.0, min\_samples = 10$
Random Forest	$n\_estimators = 100$
XGBoost	$n\_estimators = 100, learning\_rate = 0.3, max\_depth = 5$

Evaluation Method

Use **Accuracy**, **Precision**, **Recall**, and **F1** value to evaluate the model.

- TP - True Fraud TN - True Normal
  FP - False Normal FN - False Fraud

$$(1)\quad \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

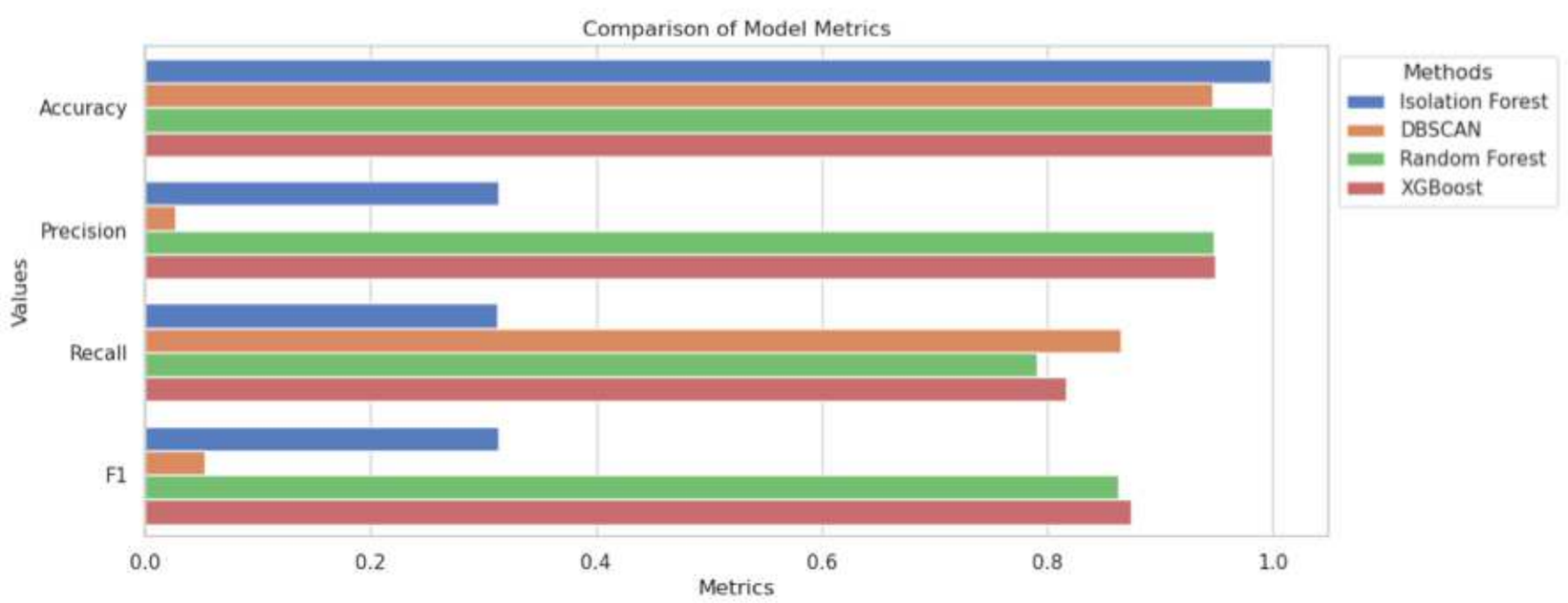
$$(2)\quad \text{Precision} = \frac{TP}{TP + FP}$$

$$(3)\quad \text{Recall} = \frac{TP}{TP + FN}$$

$$(4)\quad \text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Result

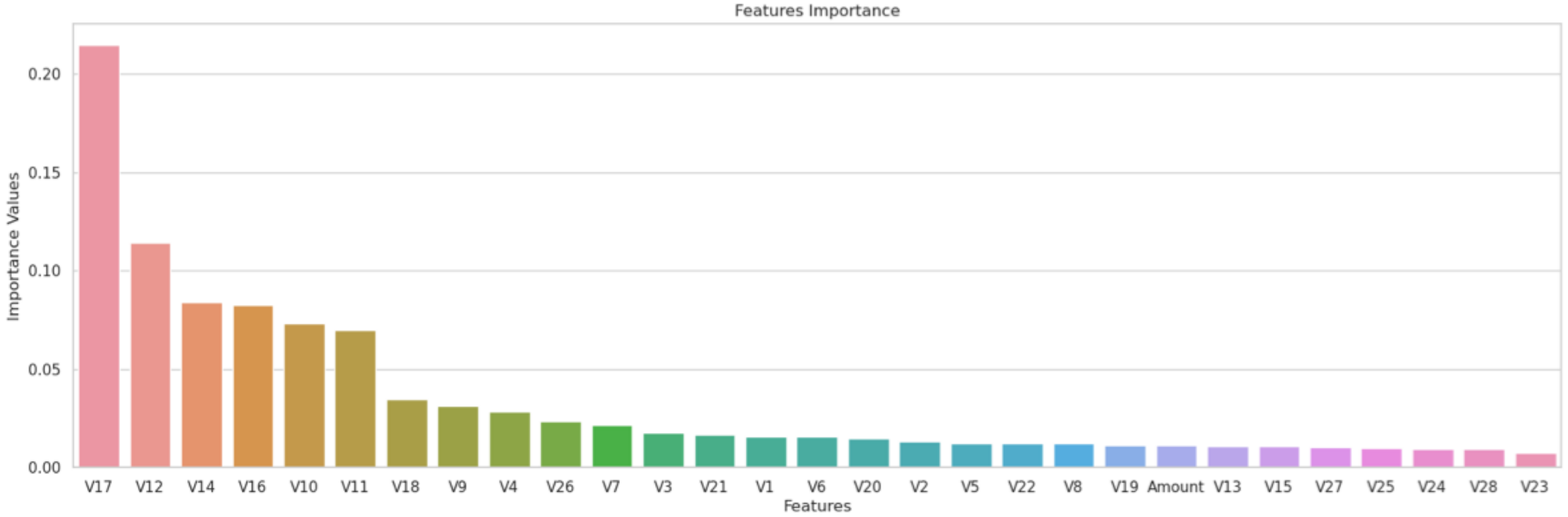
Method	Accuracy	Precision	Recall	F1	Time(s)
Isolation Forest	0.998	0.314	0.313	0.314	344
DBSCAN	0.946	0.027	0.865	0.053	182
Random Forest	0.999	0.948	0.791	0.863	314
XGBoost	0.999	0.949	0.817	0.874	56



Conclusion

Conclusion

- Supervised methods are superior to unsupervised methods.
- The performance of decision tree related methods is related to the number of decision trees and max depth.
- Based on correlation analysis and feature importance analysis, identifying credit card fraud is mainly related to features V4, V10, V11, V12, V14, and V17.



Questions?

Contact Information

Jiahong Lin

School of Economics and Management

Nanjing University of Science and Technology, China

KKSMI18@163.COM