Anomaly Detection in Payments

IE 406 : Machine Learning

Group no. 18

Assigned By

Prof. M.V. Joshi

Problem Statement

In dataset, there are transactions made by credit cards in September 2013 by european cardholders. As name of the project suggests, The goal is to separate fraudulent and normal transactions.

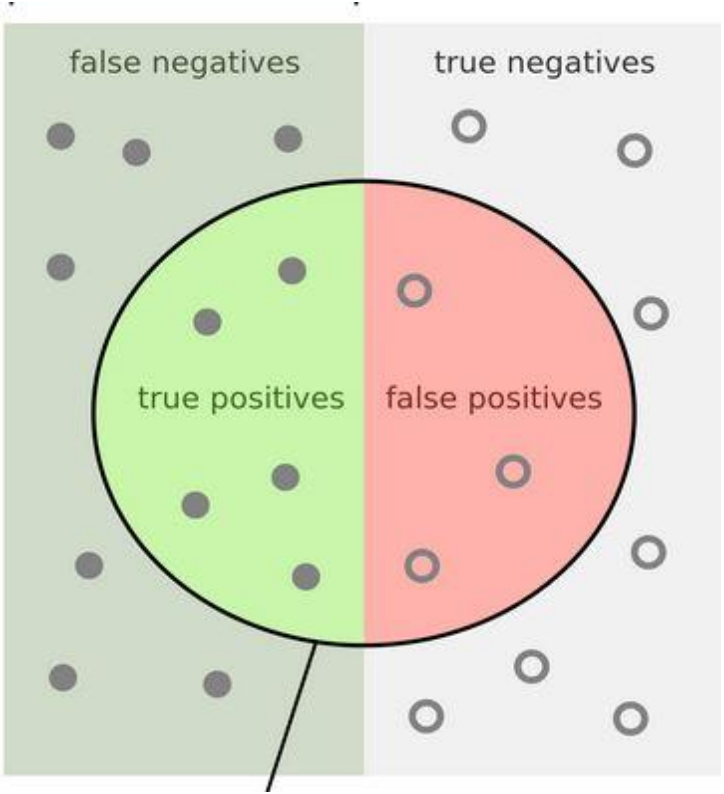
- Logistic Regression
- Support Vector Machine
- Isolation Forest
- Local Outlier Factor

Imbalanced Dataset

The dataset is highly unbalanced, the frauds account for 0.172% of all transactions. we have 492 frauds out of 284,807 transactions.

So, We can’t use accuracy to compare models.

Recall



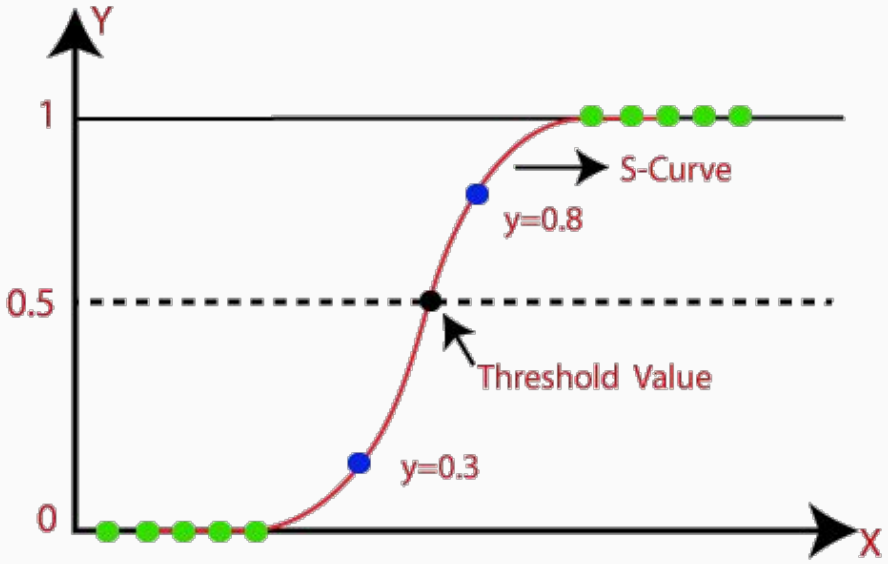
$$\text{Recall} = \frac{\text{}}{\text{}}$$

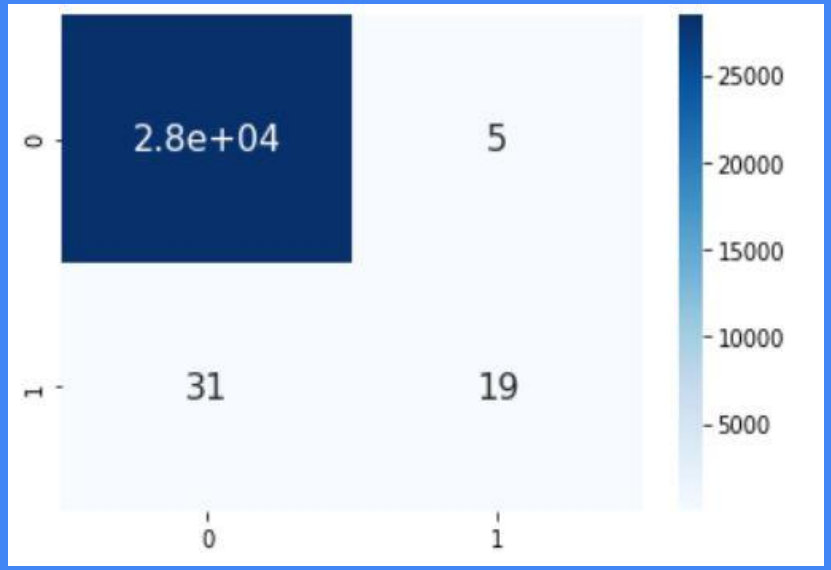
Recall : proportion of actual Fraud was identified correctly.

In other words

If model has 0.7 recall then it correctly identifies 70% of all Fraud.

Logistic Regression

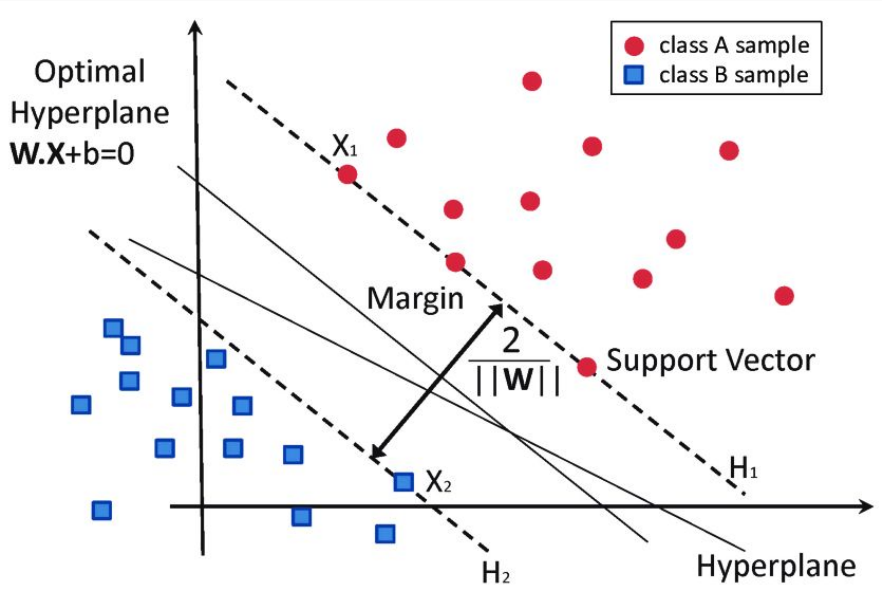


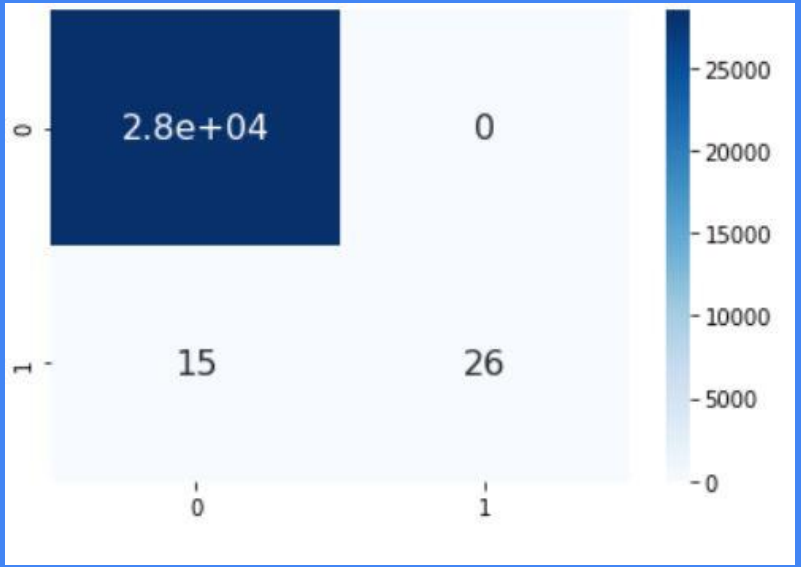


Recall :- 38.0%

Error :- 0.126%

Support Vector Machine (SVM)





Recall :- 63.41%

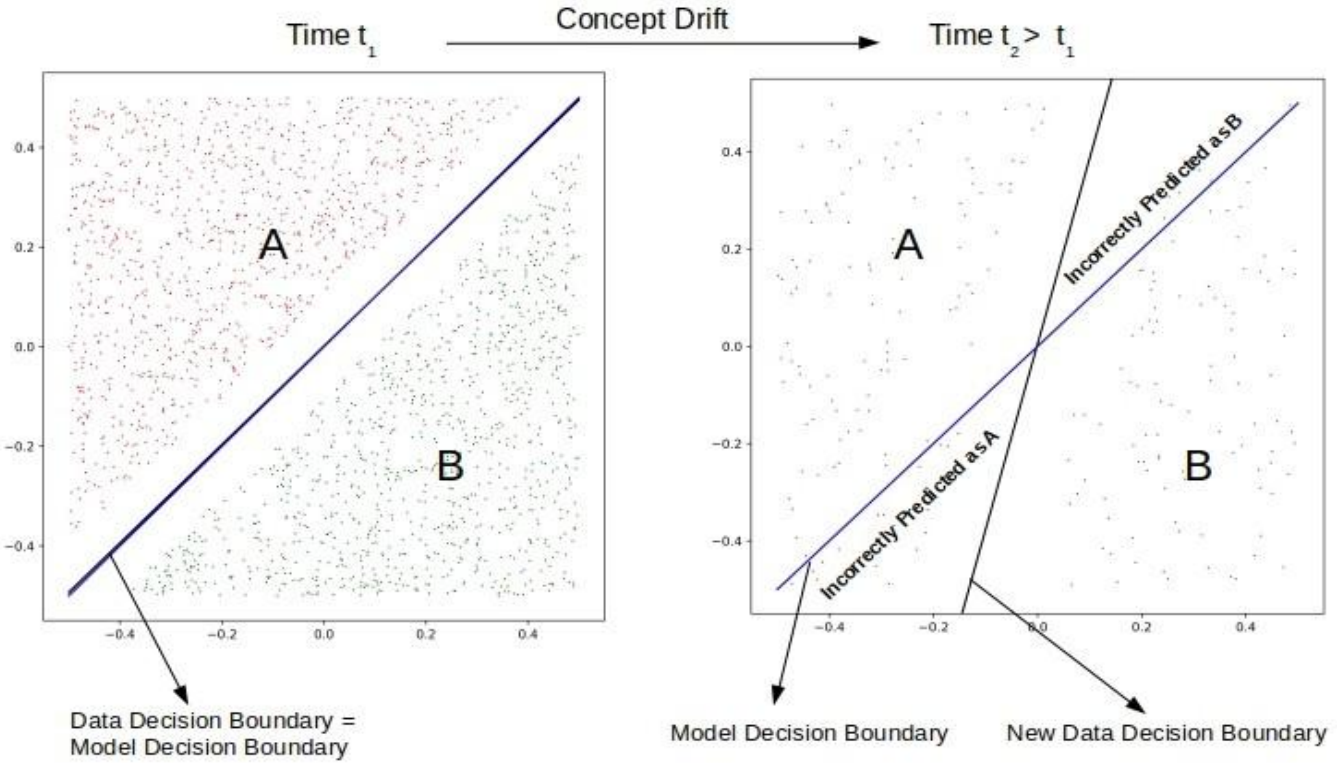
Error :- 0.052%

Imbalanced dataset

Classifier algorithms like SVM and Logistic Regression have a bias towards majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class.

Why supervised Algo won’t work : Concept drift

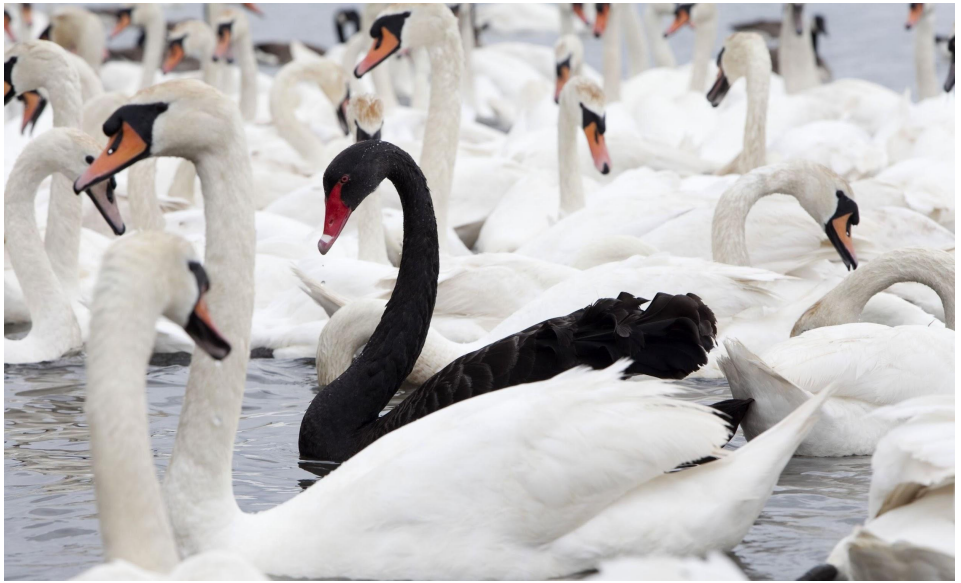
Concept Drift means that the properties of the target variable, which the model is trying to predict, change over time.



What is Anomaly ?

Anomalies are data patterns that have different data characteristics from normal instances.

Frauds are anomalies in payment transactions.



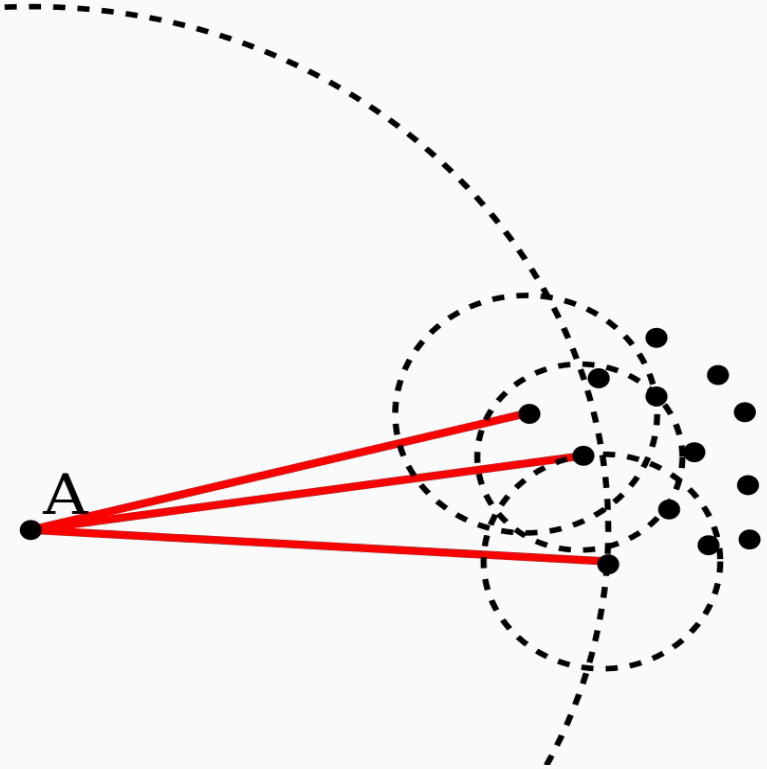
So, Frauds are few and different

Local Outlier Factor (LOF)

Local Outlier Factor (LOF)

Basic idea of LOF

comparing the local density of a point with the densities of its neighbors. A has a much lower density than its neighbors. So,it is outlier.



Local Outlier Factor (LOF)

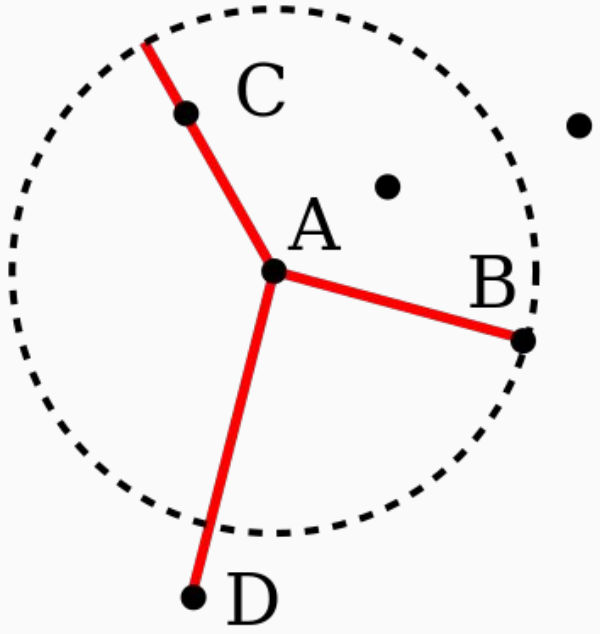
k_distance(A) = distance from point A to its kth nearest neighbor

N(A) = set of k nearest neighbors

Reachability distance

rd(A,B) = max{ k_distance(A), d(A,B) }

Objects *B* and *C* have the same reachability distance (k=3), while *D* is not a *k* nearest neighbor



Local Outlier Factor (LOF)

Local Reachability distance :

$$lrd(A) = 1 \,/\, \text{mean}(rd(A,i))$$

where i belongs to N(A)

Outlier_score :

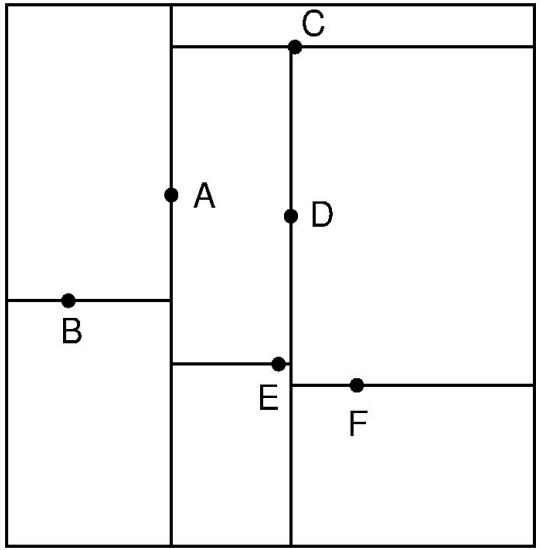
$$lof(A) = \text{mean}(lrd(i)) \,/\, (lrd(A))$$

where i belongs to N(A)

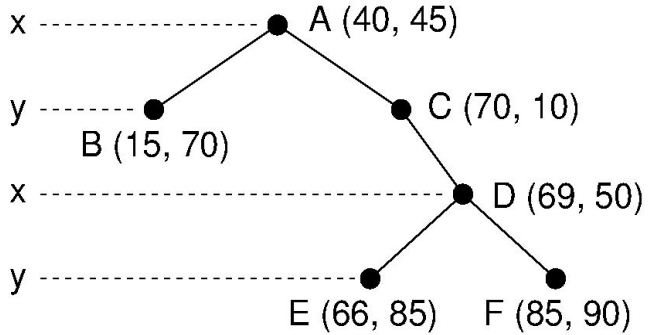
Local Outlier Factor (LOF) implementation : KD tree

KD tree is a space partitioning **data structure** for organizing points in k dimensional space.

Every non-leaf node can be thought of as implicitly generating a splitting hyperplane that divides the space into two parts, known as half-spaces.

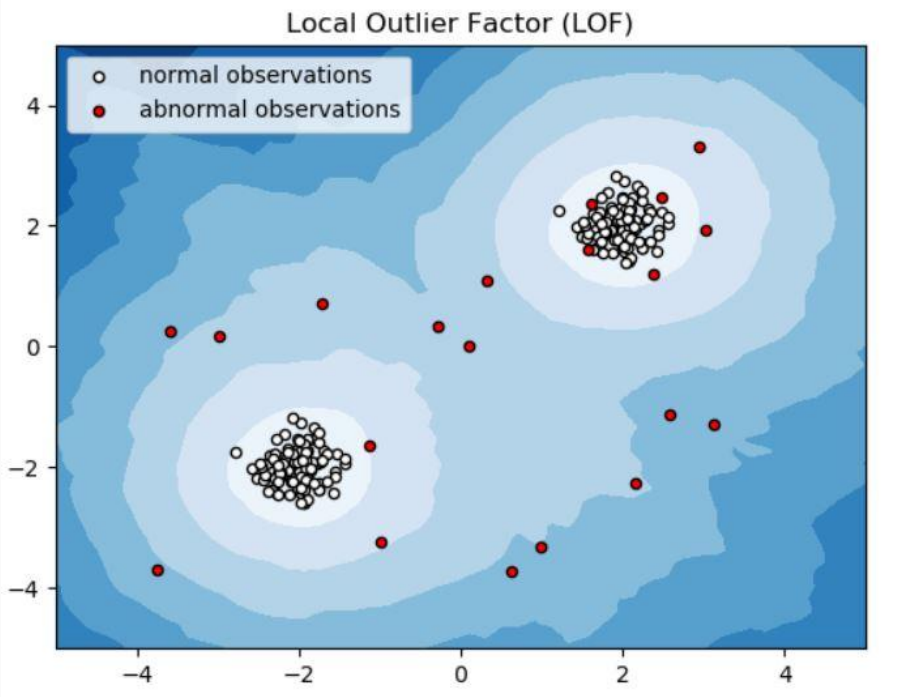


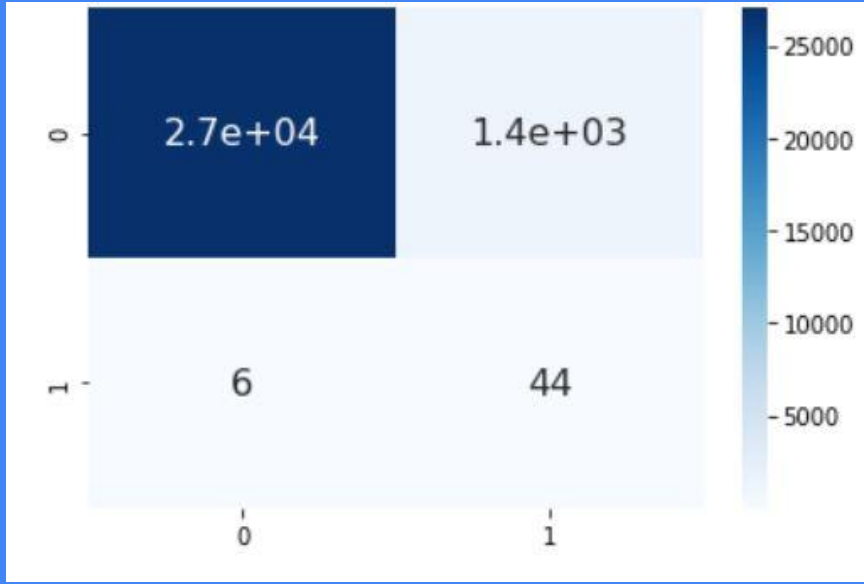
(a)



(b)

Local Outlier Factor(LOF)





Recall :- 88.0%

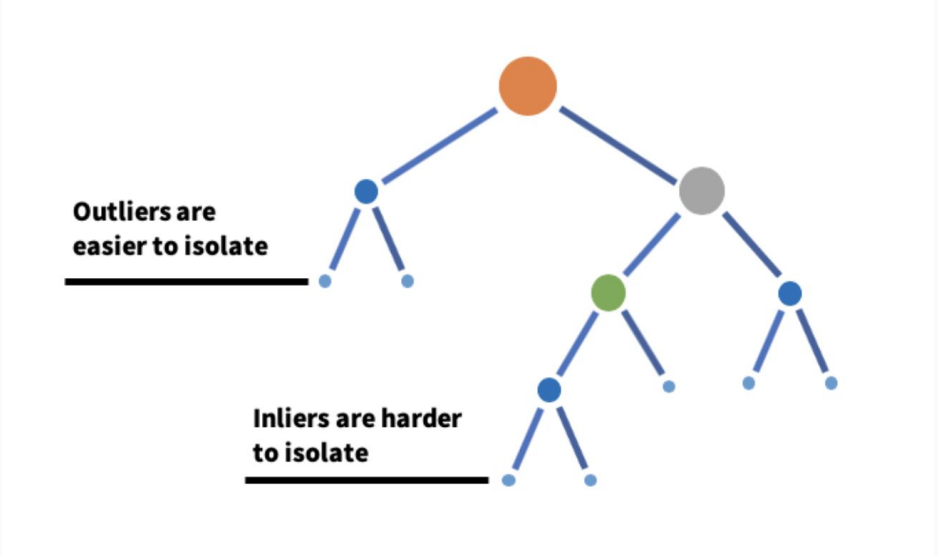
Error :- 11.42%

Isolation Forest

Isolation Forest

Let’s use random tree for partitioning of transactions.

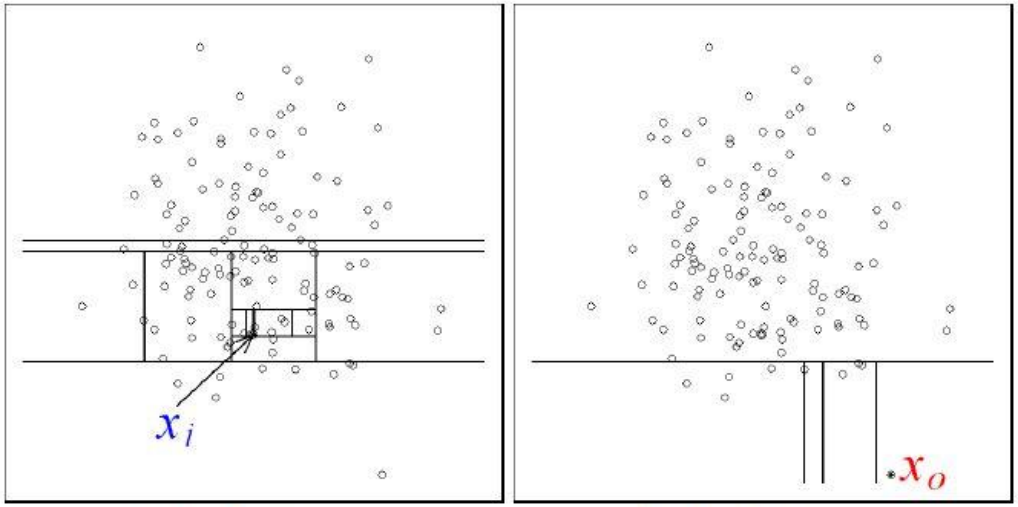
frauds are more likely to be separated in early partitioning because they are different.



Isolation Forest

Partitions are generated by randomly selecting an attribute and then randomly selecting a split value between the maximum and minimum values of the selected attribute.

Isolation Forest



(a) Isolating x_i (b) Isolating x_o

(a) a normal point xi requires twelve random partitions to be isolated

(b) an anomaly xo requires only four partitions to be isolated

Isolation Forest - Implementation

n = number of instances

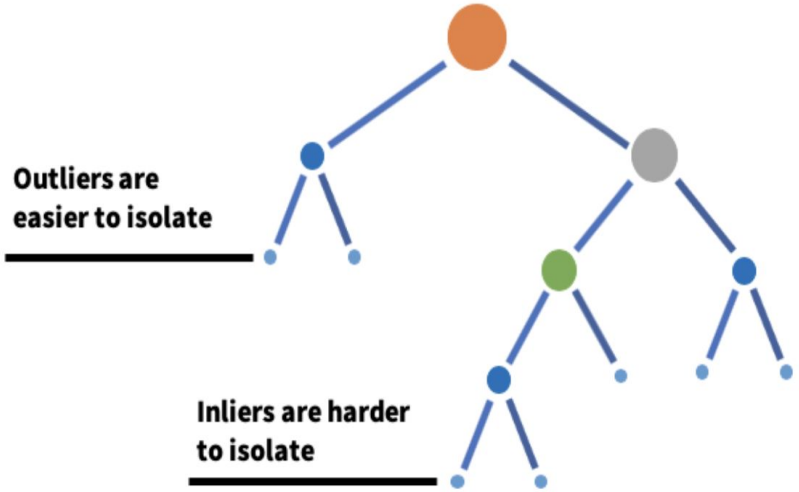
c(n) = Normalization constant

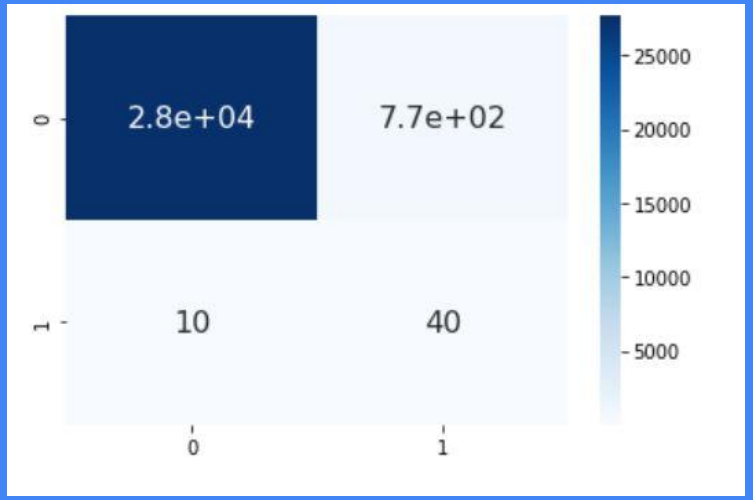
path length h(x) : edges between the root node and a terminating node

Anomaly Score:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Isolation Forest





Recall :- 80.0%

Error :- 2.724%

Previous Works / Dataset / References

- Dataset : https://www.kaggle.com/mlg-ulb/creditcardfraud
- Local Outlier Factor, Isolation Forest : https://ieeexplore.ieee.org/document/8741421
- Logistic regression, SVM, Random Forest : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8757212
- LOF : https://en.wikipedia.org/wiki/Local_outlier_factor
- Isolation forest: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest
- Imbalanced dataset effect on svm: http://www.cs.ox.ac.uk/people/vasile.palade/papers/Class-Imbalance-SVM.pdf
- Imbalanced dataset effect on LR: https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/


Thank You..!