HY-484
Fraud Detection

Charalambos Varsamis, Apostolos Mavrogiannakis

January, 2022

Using LaTeX.

Introduction

What is Fraud Detection?

Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretenses. Fraud detection is applied to many industries such as banking or insurance. In banking, fraud may include forging checks or using stolen credit cards. Other forms of fraud may involve exaggerating losses or causing an accident with the sole intent for the payout. In this project we dive into more details on fraud transactions using Real-World Datasets.More specifically, we use transaction information to detect suspicious activity.

Why Machine Learning for Fraud Detection

Machine learning is a set of methods and techniques that let computers recognise patterns and trends and generate predictions based on those. When it comes for fraud decisions, results must come fast! Machine learning is like having several teams of analysts running hundreds of thousands of queries and comparing the outcomes to find the best result. In the same way, machine learning can often be more effective than humans at uncovering non-intuitive patterns or subtle trends which might only be obvious to a fraud analyst much later.

Supervised Machine Learning

Supervised learning is the types of machine learning in which machines are trained using well labelled training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

Unsupervised Machine Learning

Unsupervised learning is a type of algorithm that learns patterns from untagged data. The hope is that through mimicry, which is the primary way young children learn the machine is forced to build a compact internal representation of its world and can generate imaginative content.These algorithms discover hidden patterns or data groupings without the need for human intervention.Unsupervised learning models are utilized for three main tasks—clustering, association, and dimensionality reduction.

Graph Analysis Metrics

In order to assist our models, we are going to calculate some Graph Analysis Metrics that will help us detect even more fraudulent transactions.

- **PageRank** works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

- **Closeness Centrality** is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes.

- **EigenVector Centrality** is a measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores.

Demonstration of our approaches

Dataset Structure

Credit Card

Dataset: Synthetic Financial Datasets

Synthetic datasets generated by the PaySim mobile money simulator

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.00	160296.36	M1979787155	0.00	0.00	0	0

Figure 1: Table Row Example

- step - maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).
- type - CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
- amount - amount of the transaction in local currency.
- nameOrig - customer who started the transaction
- oldbalanceOrg - initial balance before the transaction
- newbalanceOrig - new balance after the transaction
- nameDest - customer who is the recipient of the transaction
- oldbalanceDest - initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).
- newbalanceDest - new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).
- isFraud - This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system.

- isFlaggedFraud - The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.

Synthetic Financial Datasets For Fraud Detection with PCA

Dataset: Credit Card

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation.Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Approaches with Supervised Machine Learning

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by the majority voting. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set.Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

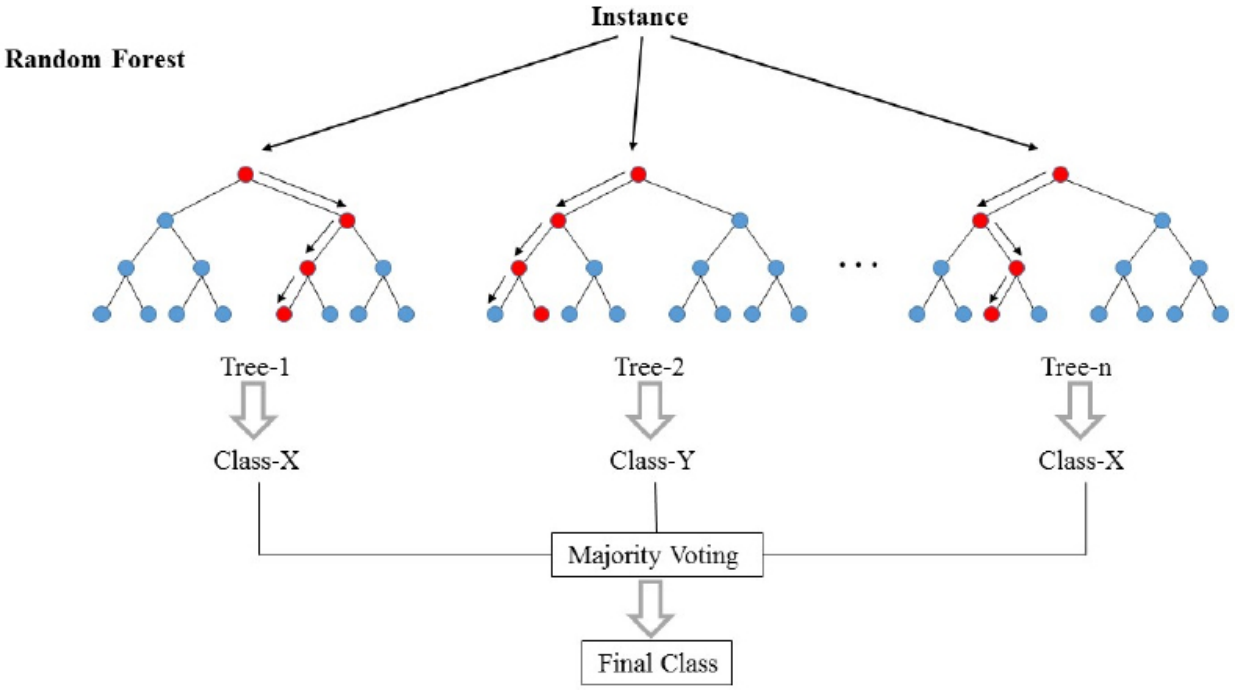


Figure 2: Random Forest Example

Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

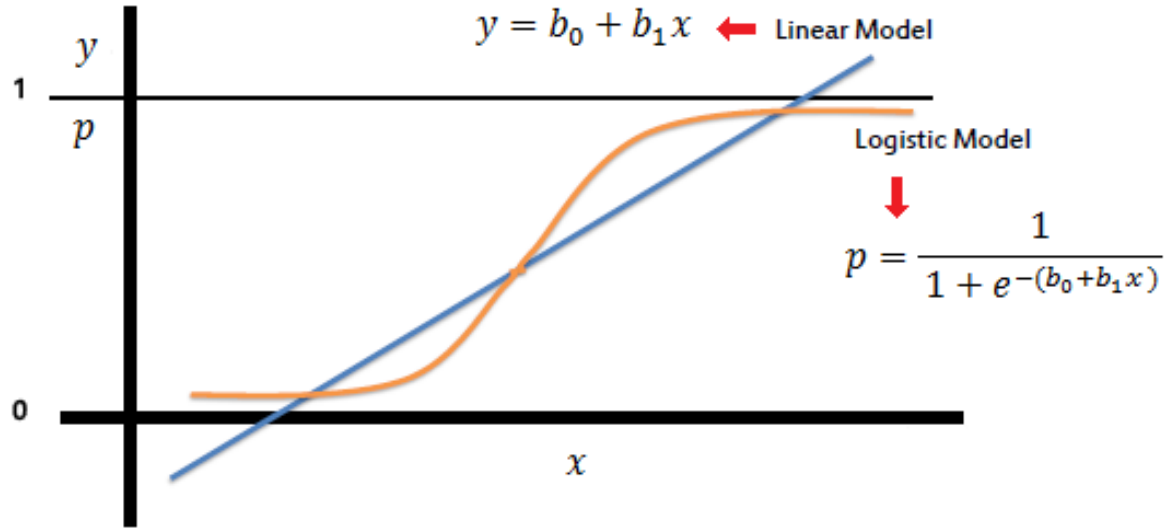


Figure 3: Logistic Regression example

Adaptive Boosting

AdaBoost, short for Adaptive Boosting, is a statistical classification meta-algorithm. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

Approaches with Unsupervised Machine Learning

Isolation Forest

Isolation forest is an anomaly detection algorithm. It detects anomalies using isolation (how far a data point is to the rest of the data), rather than modelling the normal points.Instead of trying to build a model of normal instances, it explicitly isolates anomalous points in the dataset. The main advantage of this approach is the possibility of exploiting sampling techniques to an extent that is not allowed to the profile-based methods, creating a

very fast algorithm with a low memory demand.At the basis of the Isolation Forest algorithm, there is the tendency of anomalous instances in a dataset to be easier to separate from the rest of the sample (isolate), compared to normal points. In order to isolate a data point, the algorithm recursively generates partitions on the sample by randomly selecting an attribute and then randomly selecting a split value for the attribute, between the minimum and maximum values allowed for that attribute.

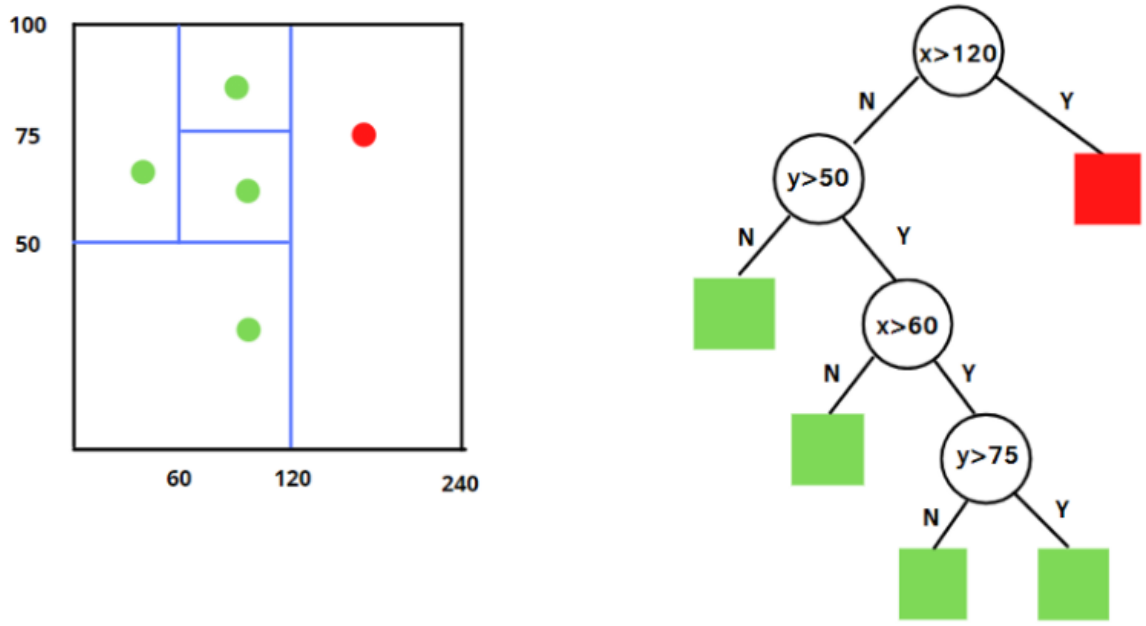


Figure 4: Isolation Forest example

K-Means Clustering

K-means clustering uses "centroids", K different randomly-initiated points in the data, and assigns every data point to the nearest centroid.To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

A cluster refers to a collection of data points aggregated together because of certain similarities.

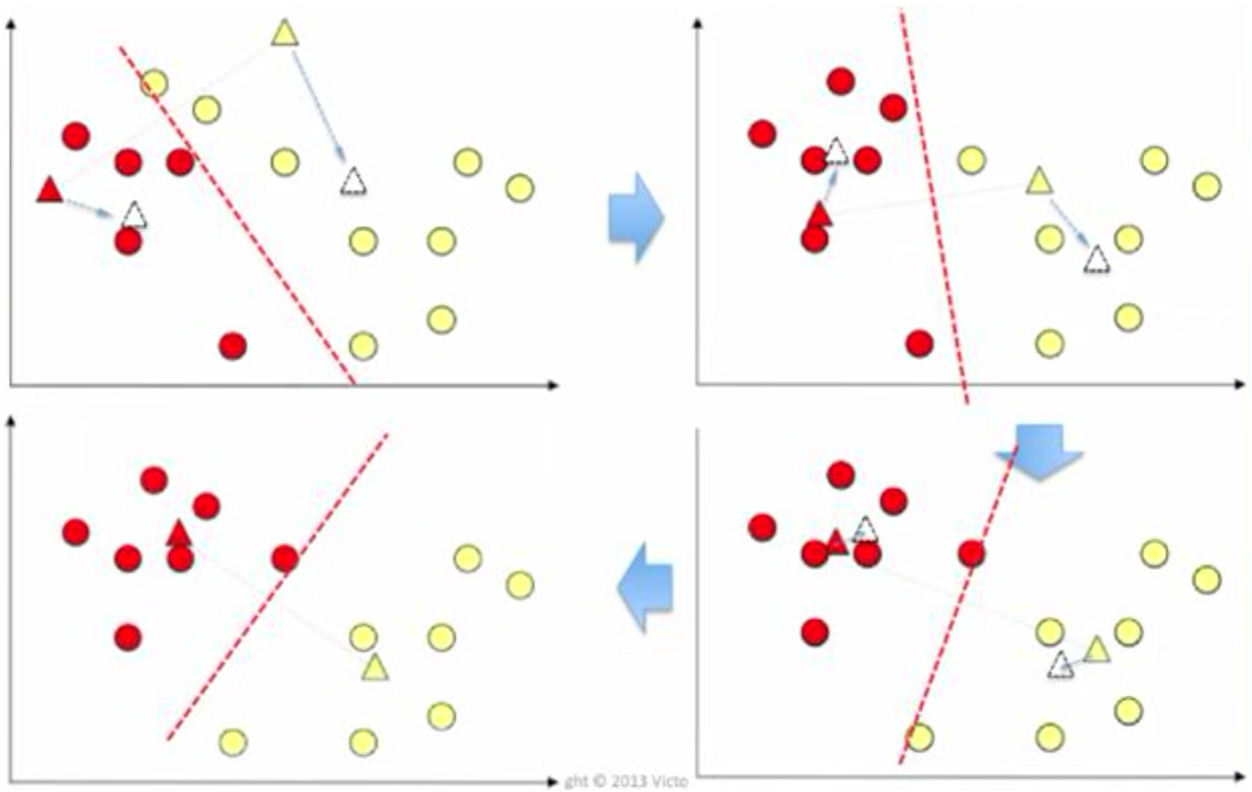


Figure 5: K-Means Clustering example

Compare Models

In order to be able to compare our approaches, we need to clarify what is a confusion matrix and a ROC curve.

What is Confusion Matrices

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

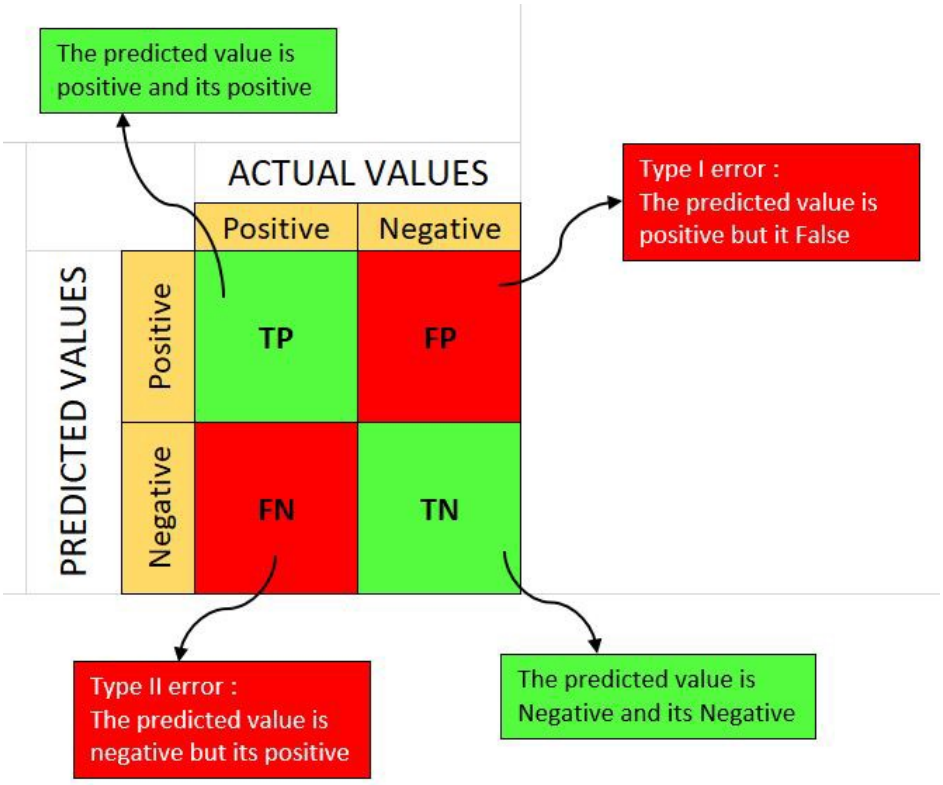


Figure 6: Confusion Matrix

This matrix is necessary in order to calculate the ROC Curve.

What is ROC Curve & AUC

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection. The false-positive rate is also known as probability of false alarm and can be calculated as $(1 - \text{specificity})$.

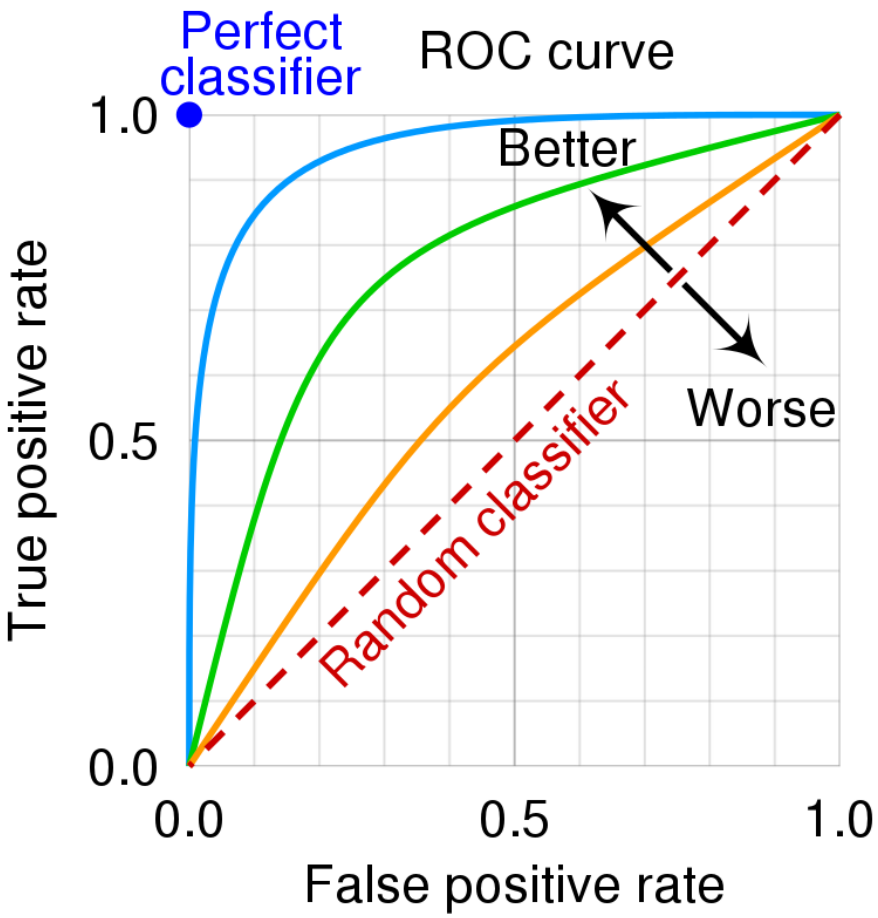


Figure 7: ROC Curve

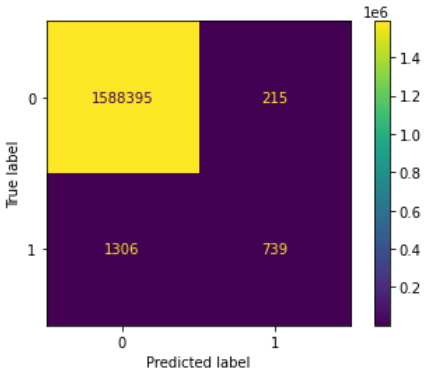
Another important definition is AUC. AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

Comparison

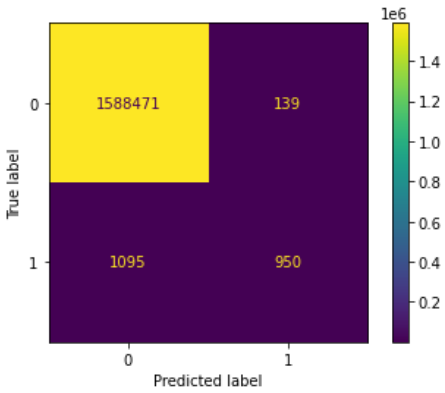
In this section we show the confusion matrices calculated in each approach and we compare them using a ROC Curve.

Supervised Learning

Firstly, we are going to prove that adding our Graph Analysis Metrics actually help the models improve their predictions.



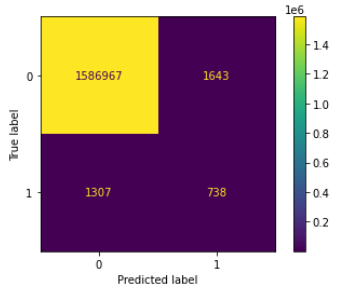
(a) Adaptive Boosting without metrics



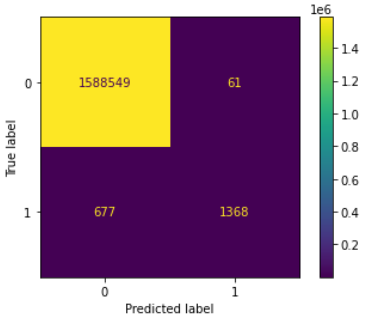
(b) Adaptive Boosting with metrics

As we can see in the figure, after inserting Graph Analysis Metrics in our model, the True Positives increase by 200+ fraudulent transactions, but at the same time, False Positive and False Negative transactions decrease.

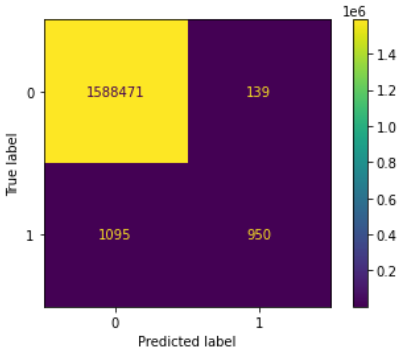
Now that we have proved that metrics are significantly helpful, let's compare all supervised models together.



(a) Percentage storage utilization



(b) standard deviation



(c) execution time

Figure 9

Clearly, the Random Forest Model is the most accurate model in our research. Whereas, Logistic Regression did a poor job detecting fraudulent transactions, it predicted half the number of true positives compared to Random Forest. Adaptive Boosting did a decent job, detecting 950 true positives, coming really close to Random Forest.

Looking at the ROC Curve, Logistic Regression is not the model we should pick in our Fraud Detection Systems. However, Adaptive Boosting & Random Forest came pretty close, with Random Forest taking the lead.

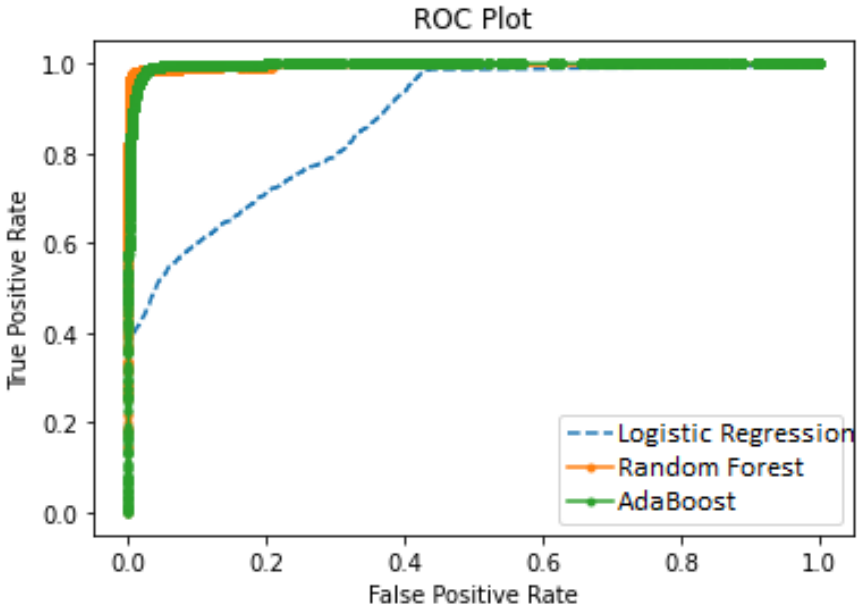


Figure 10: ROC Curve - Supervised Learning

Before someone picks a model for the fraudulent predictions, we should mention that Random Forest executed in 16 min, whereas Adaptive Boosting only executed in 6. As we mentioned in the introduction, time is a huge factor in fraud detection systems. As a result, someone would pick Adaptive Boosting, sacrificing accuracy in order to gain time efficiency.

Unsupervised Learning

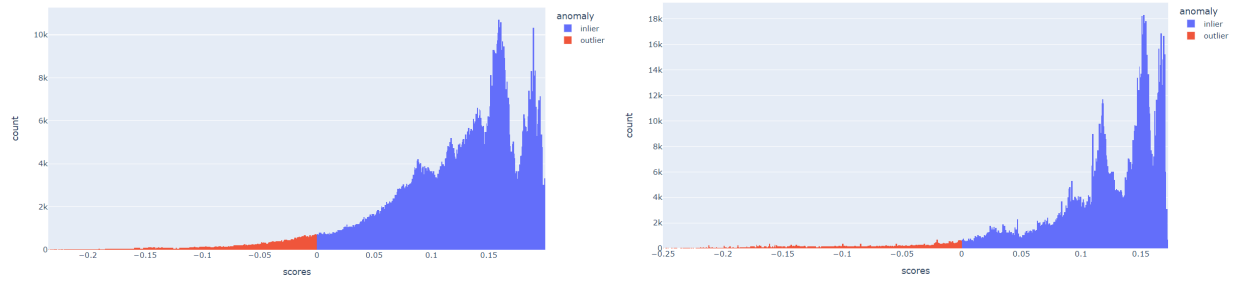
We have two datasets that were used to test the Unsupervised Learning models.These are:

- Credit Card Data
- Synthetic Financial Datasets For Fraud Detection(with PCA columns)

Credit Card Data

Below we can see two executions of Isolation Forest.In these two executions we have assumed a 5% abnormal values.The difference between these two executions is that (a) is without Principal Component analysis and (b) is with Principal Component Analysis. The aim was to observe if PCA could improve the final results. From the Histogram we cannot distinguish which of

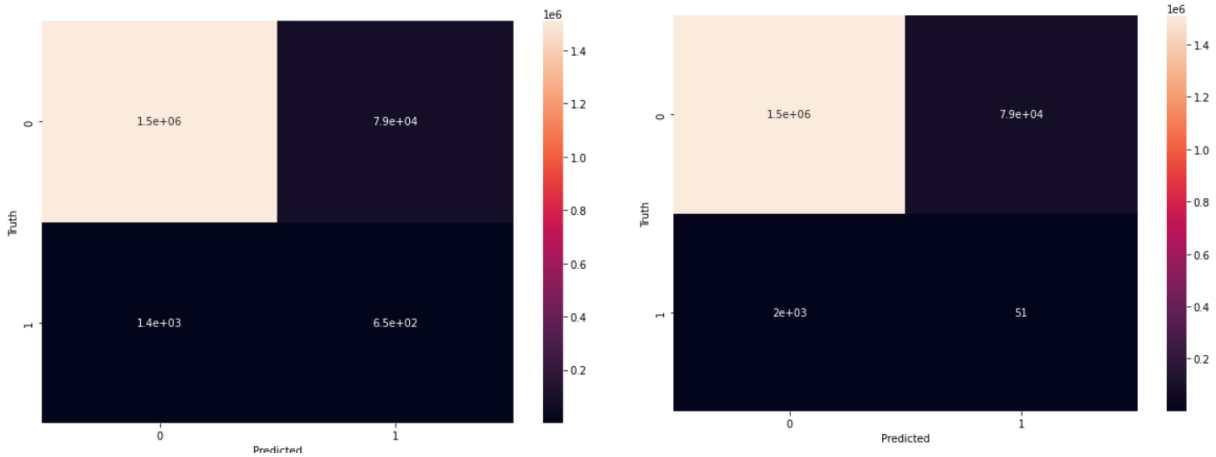
the two are better.In the execution with the PCA we can see that the counts per score value increases in a lower ratio than in the one with the PCA.In the execution with the PCA the counts per score have way higher increment ratios.



(a) Isolation Forest without PCA

(b) Isolation Forest with PCA

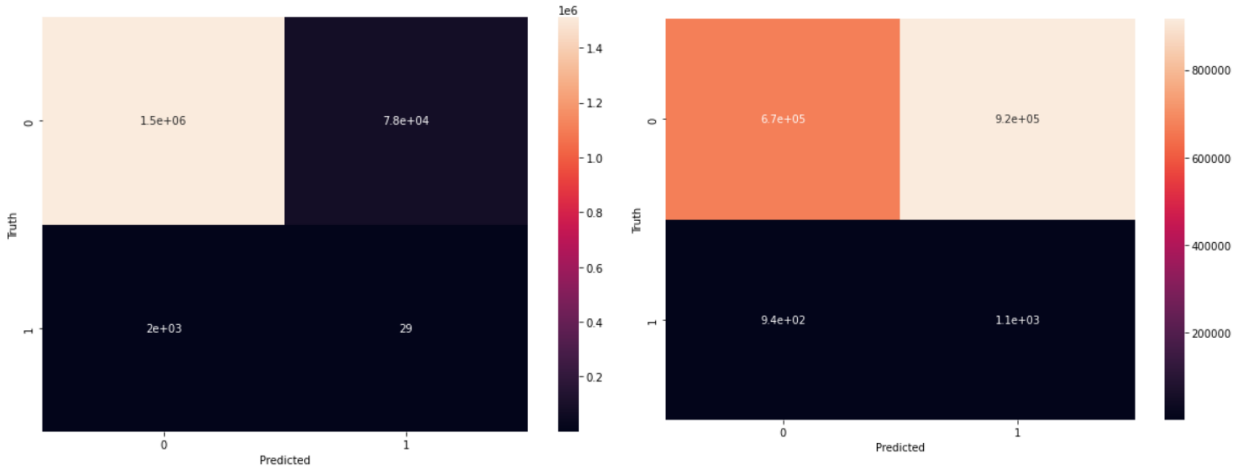
Below we can see the Confusion matrices for the two executions and we can clearly see that in (a),the True Positives increase by 600(False Negatives decrease by 600), which means that the execution without the PCA gives better results.



(a) Isolation Forest without PCA

(b) Isolation Forest with PCA

Now lets see how K-Means does. Below we can see that the model fails to identify which transactions are fraudulent and which are not in both executions(with and without PCA). Without PCA it succeeds at finding more non-fraudulent transactions but it fails to find the fraudulent ones. The execution with the PCA increases the number of fraudulent transactions recognized but it decreases the non-fraudulent predicted(True Positives).We can say that this model is producing very bad results.



(a) K-Means Clustering without PCA

(b) K-Means Clustering with PCA

Let's compare the ROC Curves of these models:

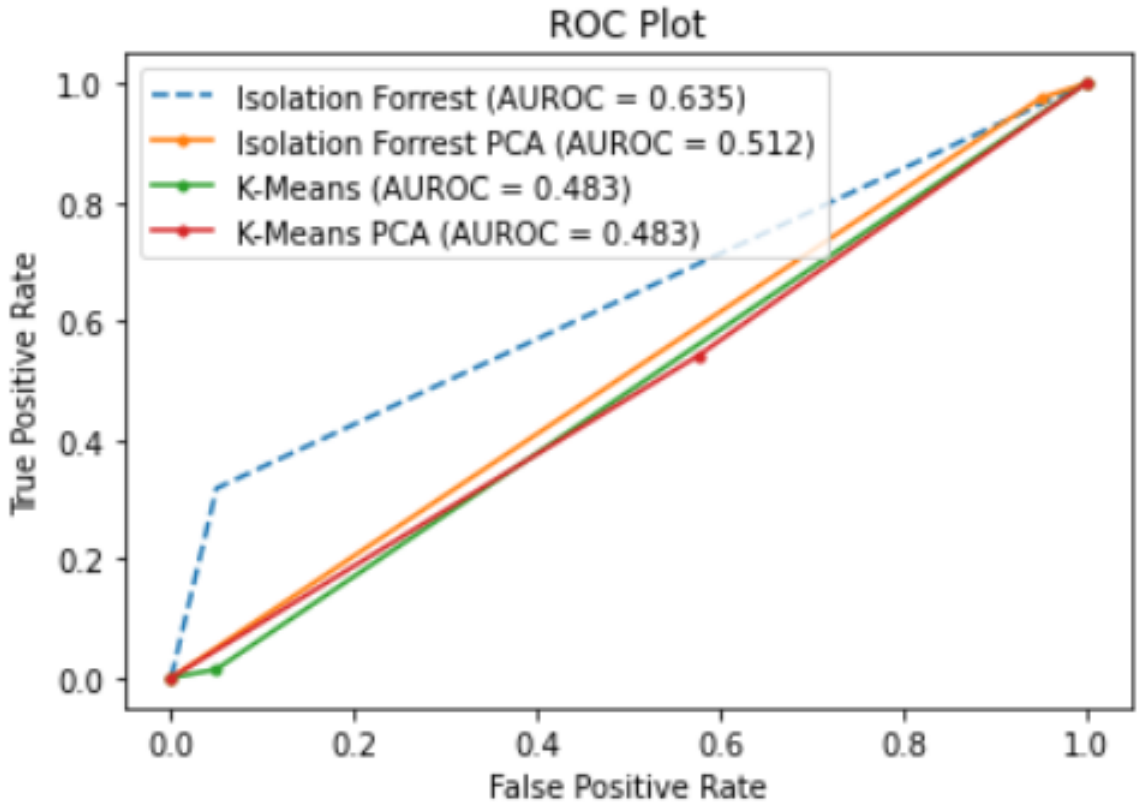
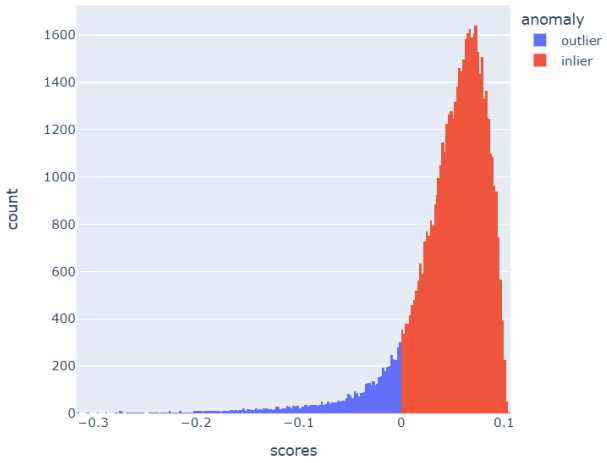


Figure 14: ROC Curve

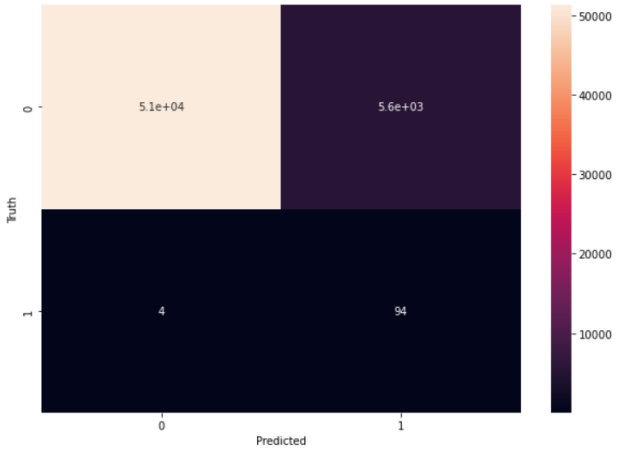
We can see that none of them gives good results, but the one that gave the best among these is the Isolation Forest model without the Principal Component Analysis.

Synthetic Financial Datasets For Fraud Detection(with PCA columns) Because the results were not good and it might have been due to the nature of the dataset we are going to check the results of the model with another dataset.This dataset contains only columns that have been transformed with PCA

Below we can see the histogram and the Confusion Matrix from the execution of Isolation Forest.We can see that it finds mostly all the True Negatives but it also categorizes many non-fraudulent transactions as fraudulent(5600 FP).



(a) Isolation Forest with PCA Confusion Matrix



(b) Isolation Forest with PCA

Now lets see how K-means does.We can see that K-means fails to identify fraudulent transactions.Unfortunately it finds all transactions non-fraudulent.

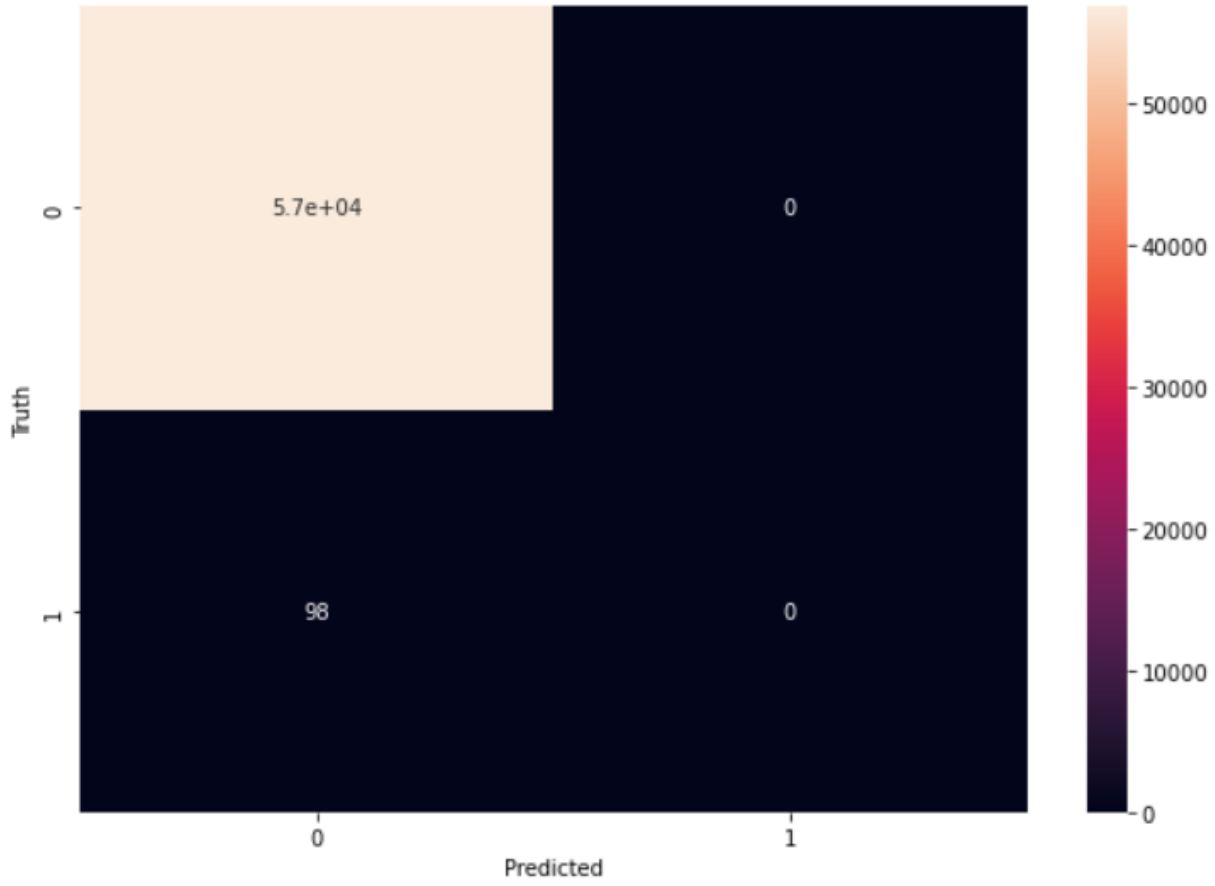


Figure 16: K-Means Clustering

Let's compare the ROC Curves of these models.Isolation Forest gives some fairly good results but that may be because the fraudulent transactions are only 98(and it finds the 94 of them).

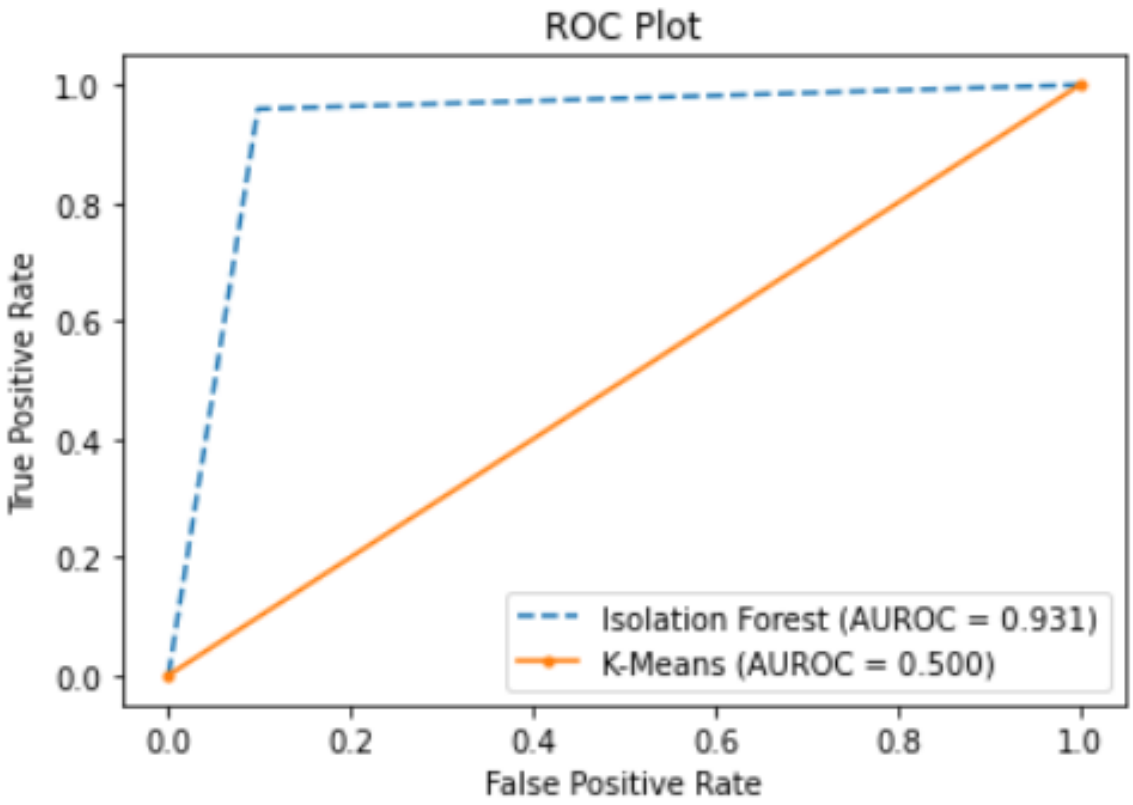


Figure 17: ROC Curve