Unsupervised Learning

Yuval Lavie

May 7, 2020

Abstract

We apply unsupervised methods to detect anomalies,cluster and visualize the high dimensional data in the Credit Card Fraud dataset.

Contents

Part I
Introduction

0.1 Context

Anomaly detection is the process of labeling an event as one that differs from the norm. we usually have a dataset of unlabeled or semi-labeled data and are required to figure which instances are anomalies. we hope that the anomalies occur rarely and that their features are significantly different than the normal instances.

0.2 Problem

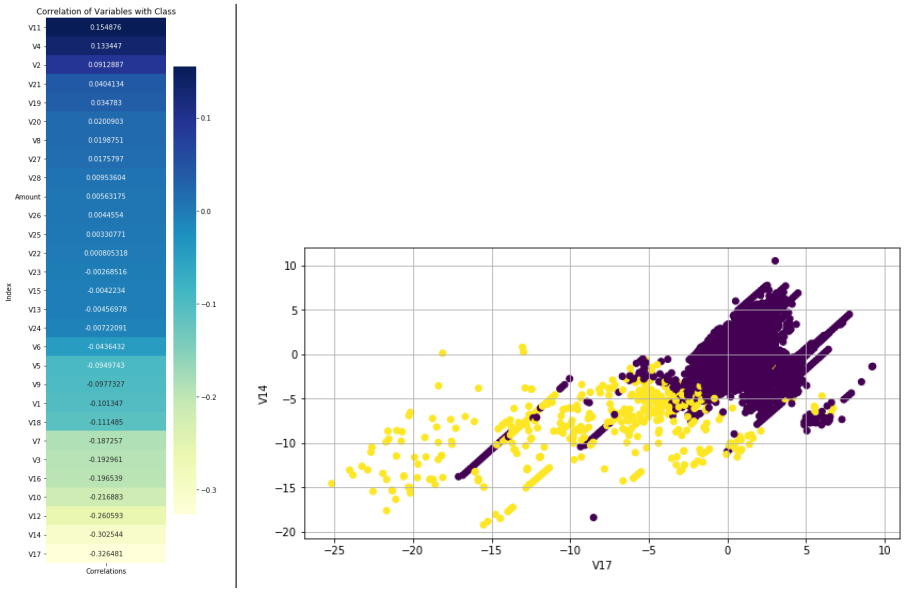
We are presented with the Credit Card Fraud dataset which is a highly imbalanced labeled dataset and are required to identity the patterns or anomalies within that set to figure out which transaction is fraud and which is legit.

0.3 Objective

We start by describing the statistics of the data set and move to visualization. we then apply unsupervised methods such as clustering and dimensionality reduction to visualize and analyze the data set and use the labels only to evaluate our models. we will not apply any supervised learning algorithm.

Part II
Statistics

Our data set consists of 31 columns where the first 30 are generated as a PCA decomposition of the original confidential data of credit card transactions. the 31th column is an imbalanced label variable where the fraudulent transactions are only 0.17% of the total transactions. We found that some features were correlated with the labels and visualize the data set and build our anomaly detection system based on these vectors. some features had moderate correlation with others but that had no affect on our system. the following graphics show the relation of the vectors, especially V14 and V17 to the labels.

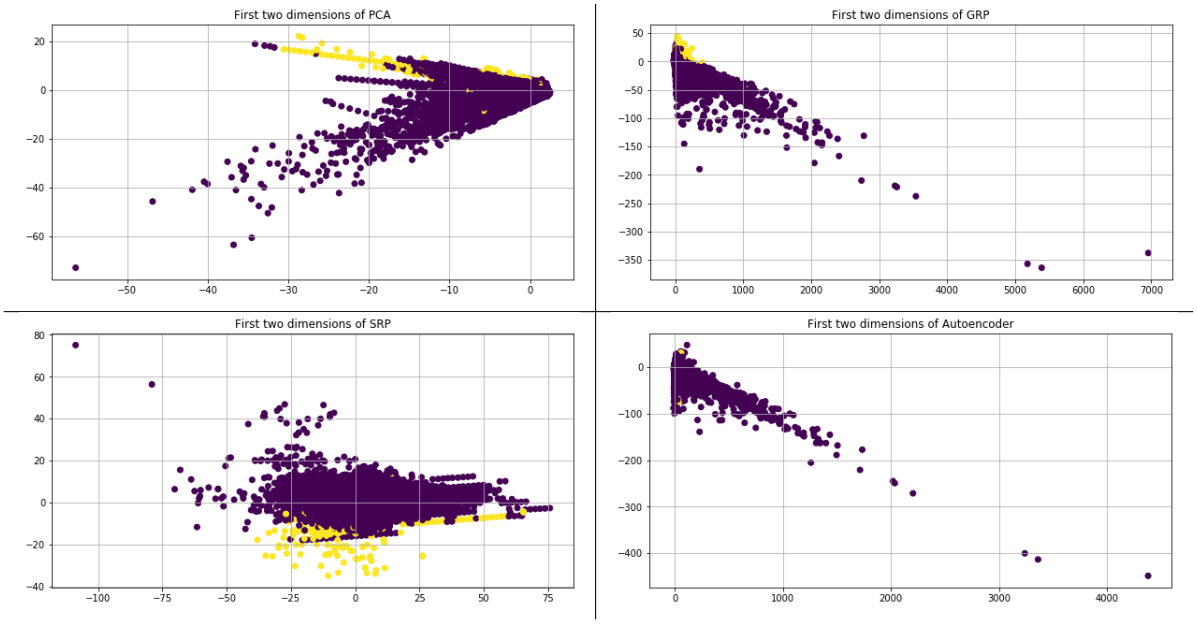


Part III

Visualization

The process of visualizing high dimensional data can be done via Dimensionality Reduction techniqued such as feature extraction, feature elimination and feature selection. by choosing V14 and V17 in the previous section we already demonstrated feature selection and we now show the results of the following algorithms:

1. Principal Component Analysis
2. Sparse Random Projection
3. Gaussian Random Projection
4. Autoencoder (10,15,22,29)



0.3.1 Insights

It seems as though the first two vectors of each dimensionality reduction technique indeed shows some relation between the labels and the features. some transactions may overlap but it certainly seems possible to find hidden patterns which reflect whether a transaction is legitimate or fraud.a soft clustering technique or a density based algorithm may be appropriate in these cases.

Part IV
Anomaly Detection

Principal Component Analysis is a technique widely used for lossy data compression. we define our compression/decompression scheme as an anomaly detection system and view the results. we define a score function which reflects the difference between an original vector and the reconstructed vector and call that our anomaly score. Let $S : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ defined as following:

$$S(v, v') = ||v - v'||_2^2$$

We build on the assumption that anomalies are different than the normal data and usually display different features and hope that compressing and decompressing the data would keep the general structure of the data but miss some refinements which we will classify as anomalies. since the data is highly imbalanced we will measure our success with precision, recall, f1 and the respective curve.

1 Results

We've split our dataset into training (80%) and test (20%) sets and follow the precision-recall curve to find the correct amount of dimensions to compress the data to. we use our dimensionality reduced representation fron part 1 where 6 vectors with the highest correlation to the labels were chosen and we follow the Presion-Recall curve and the EER point to compress the data with PCA into 5 dimensions and then decompress it.

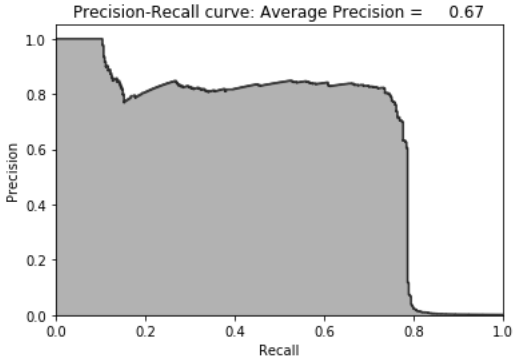
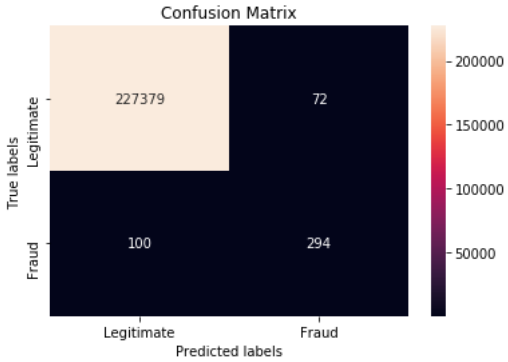


Figure 1:

1.0.1 Training set

We train our anomaly detection system to receive the following results

- Precision: 0.8
- Recall: 0.75
- f1-score: 0.77
- Mutual Information: 0.736 (p-value < 0.001)
- Jaccard Score: 0.718 (p-value < 0.001)



1.0.2 Test set

We show the following results on the test to prove that our system can detect new frauds not seen yet with the following results:

- Precision: 0.87
- Recall: 0.81
- f1-score: 0.84

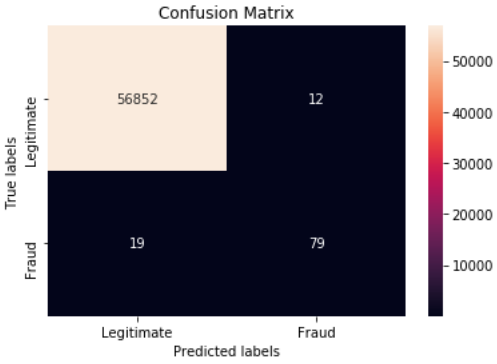


Figure 2:

1.0.3 Visualization

We visualize the the relation between the transactions, their respective anomaly scores assigned by the system and the threshold to show the skills of the system.

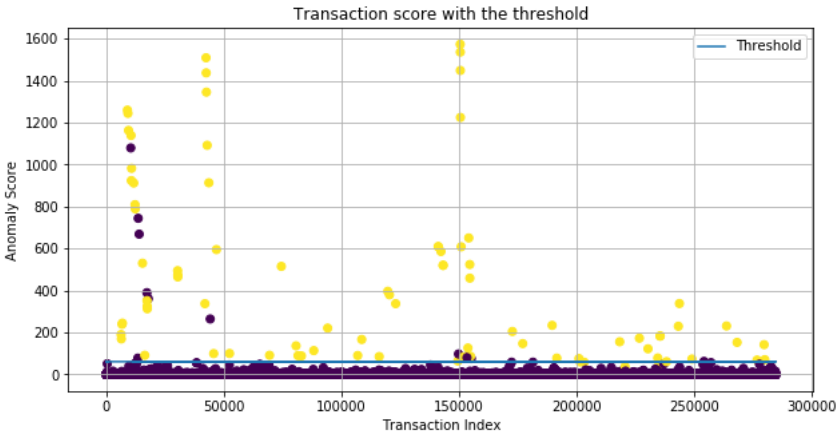


Figure 3:

Part V
Clustering

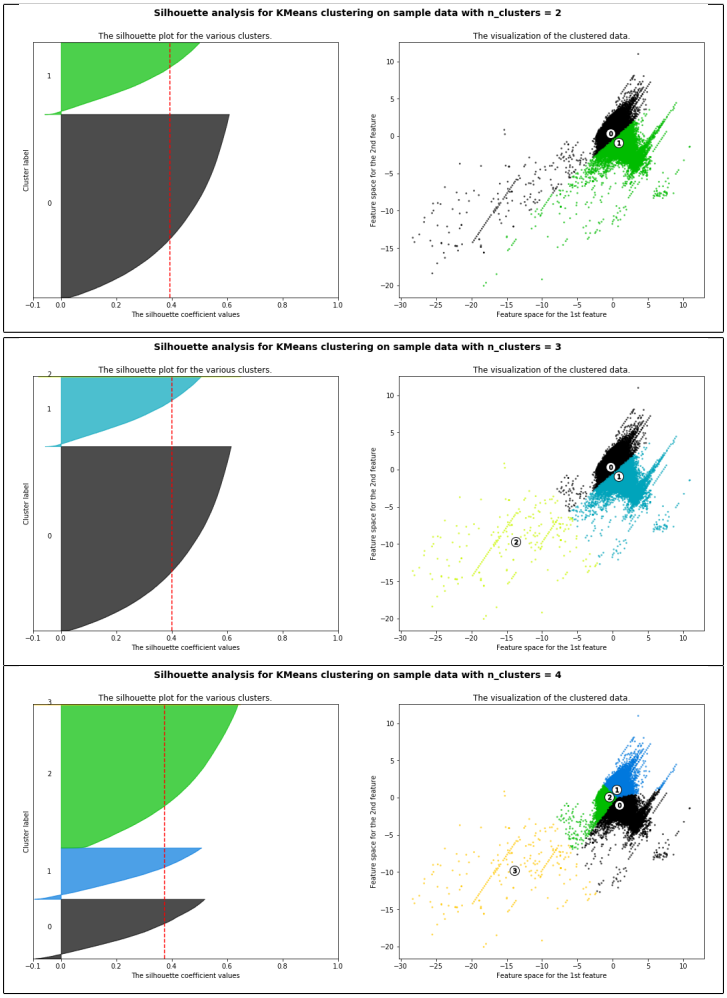
Clustering is the task of finding sub-populations in the data or in simplier words grouping data points in a manner which reflects their similiarty to other data points in the same group and their dissimiliarty to data points in different groups. We used several algorithms such as K-Means, Gaussian Mixture Models, DBScan, OC-SVM, Isolation Forest and Local Outlier Factor and found that K-Means generates the best results with respect to our data set.

1.1 Experiment

We divided our data set to training (80%) and test (20%) sets and use the training set without the labels to find how many groups exists in the data in hope to find a group of anomalies which would represent the fraudulent transactions. we evaluate our results on the test set.

1.1.1 Number of clusters

Silhouette analysis is a graphical way to understand the separation distance between the resulting clusters. it is a measure of how close each point is to points in the neighboring clusters. it has a range of [-1,1] where +1 means that a point is far away from other clusters, 0 indicates the point is on the decision boundary and -1 may indicate that a point is assigned to the wrong cluster.we iterated over a large number of clusters to find that surprisingly three clusters makes the most sense reflecting the fact that there may be many kinds of legitimate transactions that differ only by a little from each other but a small number of fraudulent transactions differ by alot as shown in the following graphs.



1.1.2 Evaluations

Evaluator	Value
Precision	0.91
Recall	0.59
f1-score	0.72
Silhouette	0.917
Homogeneity	0.515
Completeness	0.745
V-measure	0.609
ARI	0.715
AMI	0.609
Jaccard	0.558

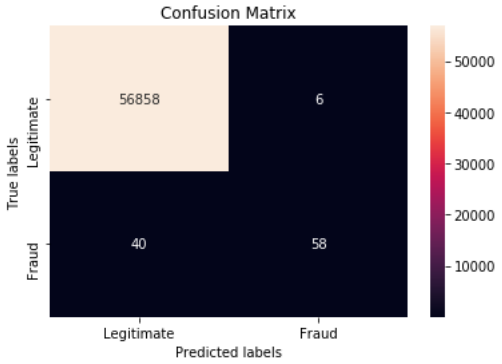


Figure 4:

1.1.3 Insights

We find that our assumptions based on our visualization and prior informations about anomalies indeed are representitive of our data set. our clustering found two major groups of legitimate transactions which only differ from each other by a little and a third small group which differs by a lot as seen by the silhouette graph of three clusters.we labeled our two similar groups as legitimate and third group as fraud and found that for all evaluation metrics with p-value less than 0.001 our clustering is efficient on the test set and has a high TPR and and low FPR. we note that our clustering is limited to finding transactions that were already labeled as fraudulent but we think that re-inspection of the data after our clustering may in fact reveal more fraudulent transactions that went under the radar. some points based on their respective silhouette score may be re-inspected to verify if their label is truely correct.