Machine Learning Engineer Nanodegree

Capstone Project: Anomaly Detection

Luis Chapa
June 30th, 2018

I. Intro

1.1 Project Overview

The Bank is in the way to search for protection to their clients against fraud operations. The most common operations they provide to their clients are: loans, payment-orders and transactions.

The Bank managers has shared an anonymized dataset with the Machine Learning Team. Which includes information for those operations as well as their clients geography location and type of credit cards.

1.2 Problem Statement

So far, The Bank managers have not registered fraud operations. It means that there is no a **ground truth** yet. We were asked to help them in determine an **novelty threshold** in order to start identifying anomalies and/or novelties (outliers) in the dataset.

Because there is no ground truth already recorded, the solution to their problem is going to be an Unsupervised Feature Learning using two diferent approaches. Both approaches is going to be using reconstructing errors as decision parameter for classifying each data point as **novelty** or **normal** point.

By identifying novelties, The Bank managers are able to take actions on those outstanding operations, then validate the operation and label as fraud or not. Once the labeled dataset migth be populated the machine learning algorithms could be updated and retrained, but that is part of a different project, so far the threshold is needed.

1.3 Metrics

In this project, we will be determining the **novelty threshold** as follows:

1. Benchmark Model:

$$T = \lambda * max(distance(\forall_x, \forall_c))$$

$$max(distance(x, \forall_c)) \underset{False}{\overset{True}{\gtrless}} T$$

The threshold T is determined by the max distance of each data point against each center of the clusters, then this number is reduced by a factor λ in order to catch outstanding points in regards to the cluster centers. Once the threshold T is calculated, each data point is labeled as novelty or normal accordingly if it is beyond this threshold in regards of the max distance to each center.

Note: In the proposal, we selected a Silhouette Score, however this score is calculated for all points. Instead of that joining score we determined a metric that could be calculated for each point individually.

2. Solution Model:

$$T = \lambda * max(distance(f(x), h(r))$$

$$distance(f(x), h(r)) \underset{False}{\overset{True}{\gtrless}} T$$

In this model, the threshold T is calculated by taking the max distance (loss) in between the original point and the reconstructed by the autoencoder.

The $distance$ function, used in both models is the Euclidean Distance between two multidimensional points.

2. Data Collect

The Bank managers has provided us a normalized and anonymized database to train the models on.

Note: For the propose of this Nanodegree Capstone Project, this dataset was downloaded from:

lpetrocelli/1999 Czech Financial Dataset - Real Anonymized Transactions Updated Jun 1, 2017 Version: d10118e4 Public Domain Licensehttps://data.world/lpetrocelli/czech-financial-dataset-real-anonymized-transactions.

The data base has the following tables, each one of them is relevant to describe the behavior of the customers of The Bank, in order to detect those data points that out stand the mean and might be labeled as anomalies:

- Account (4500 records): static characteristics of an account.
- Client (5369 records): characteristics of a client.
- Disposition (5369 records): Relation on a client with an account.
- Order (6471 records): characteristics of a payment order.
- Transaction (1056320 records): Transaction on an account.
- Loan (682 records): Loan granted for a given account.
- Credit card (892 records): Credit card issued to an account.
- Demographic data (77 records): Demographic characteristics of a district.

Figure 1: Entity-relation diagram the bank database here is the entity -relation diagram of the tables of the database that was provided by The Bank managers. We can appreciate how the data is related to each other. From this diagram we can group by the **account_id**, because each operation relates to this field. The dataset is normalized, the dataset should be joined into a single dataset and dealing with missing values.

After the corresponding joins, we got a single-table dataset as follows:

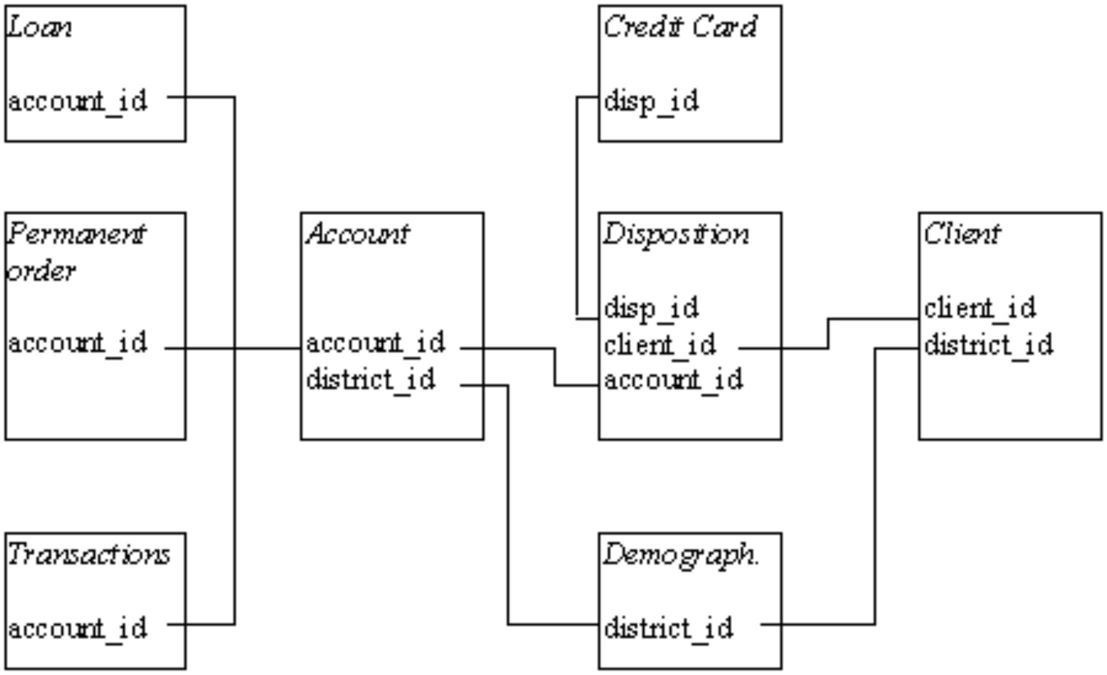

FIGURE 1: ENTITY-RELATION DIAGRAM THE BANK DATABASE

Samples: *2243458*
Features: *23*

3. Exploratory analysis and visualisations.

There are three principal operations The Bank managers emphasize on. These operations are the ones are more prone to risk of fraud.

1. Loan.

In **Figure 2: Loan per account** we can see the amount per loan grouped by account. At first sight it is difficult to figure out a kind of relation among the amounts and the account_id's which has been order consecutively.

However we can find a kind of group among the last thousand account_id's. But the same amount in loans.

Also, it is noticeable that with this simple scatter plot some discordant points start emerging among the others.

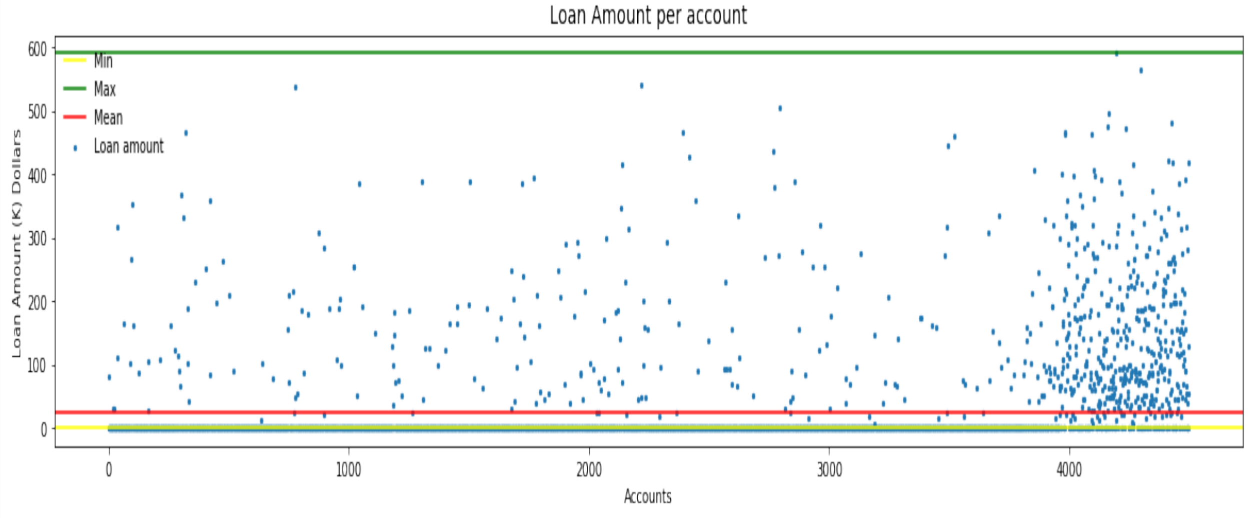

Figure 2: Loan per account (Minimum, Mean, Maximum)

2. Payment Orders.

In **Figure 3: Payment Orders per account** we can see the amount per order grouped by account. As well as in the Figure 2, it is difficult to figure out a kind of relation among the amounts and the account_id's which has been order consecutively.

Here, in the Order plot, it is not identified any kind of group, pattern or relations in the set.

But, as well as in the Figure 2, here it is noticeable some discordant points.

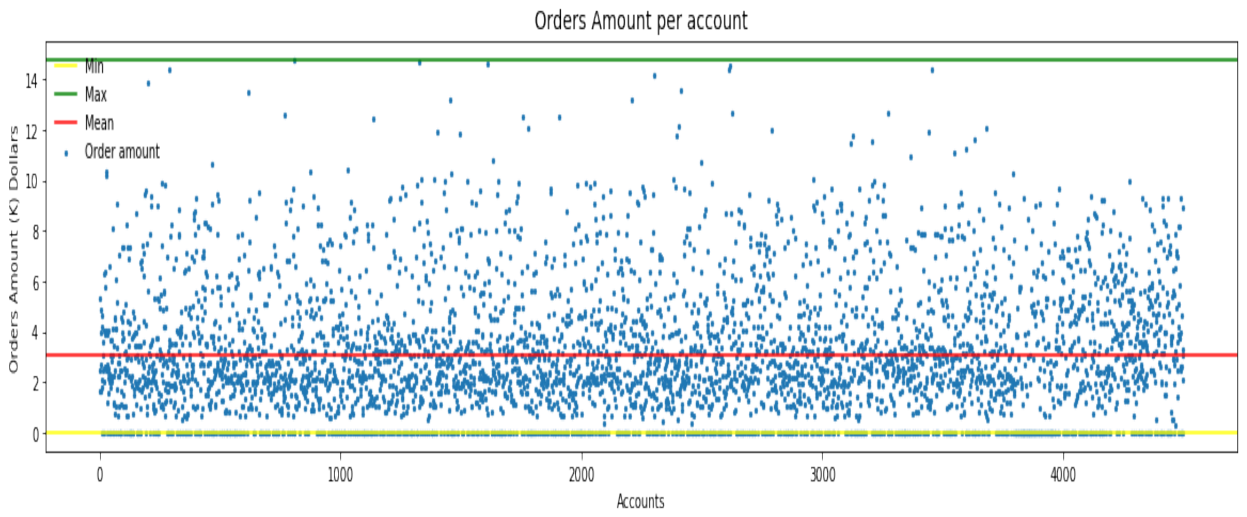

Figure 3: Payment Orders per account (Minimum, Mean, Maximum)

3. Transactions.

In **Figure 4: Transactions per account** we can see the amount per order grouped by account. It is not noticeable patters, relations or groups either. But some outstanding points away from the mean.

However in Figure 2, Figure 3 and Figure 4, it is set 2 limits the maximum, the minimum and the mean. So, despite the no-relation found with this method, we can start figuring out how the data points, in this case, amounts of money, are distributed a long the accounts.

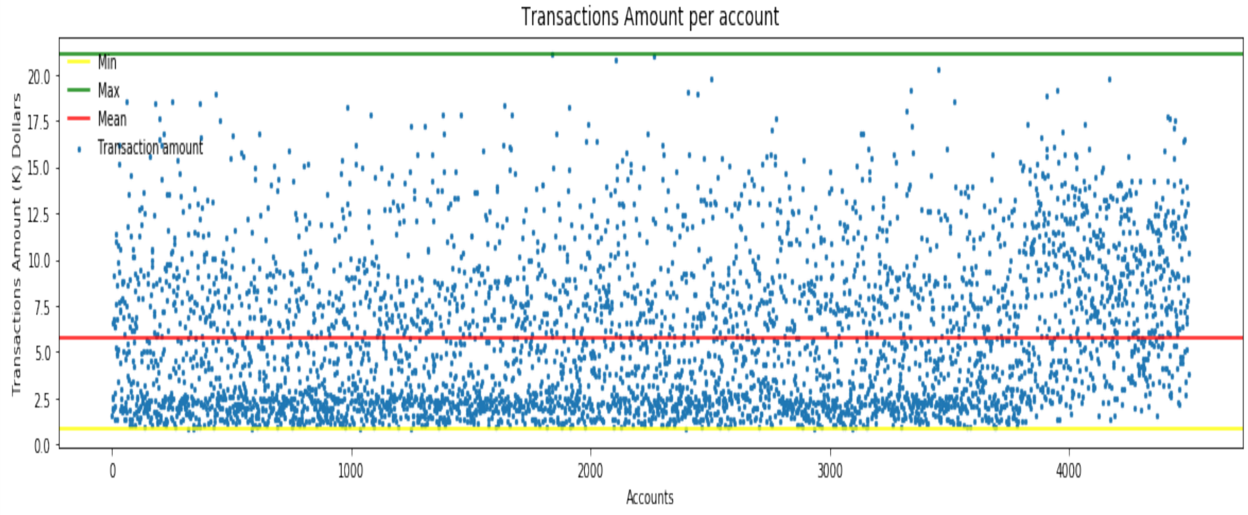

Figure 4: Transactions per account (Minimum, Mean, Maximum)

Let's see now, another point of view of the scatter plots: The Scatter Matrix.

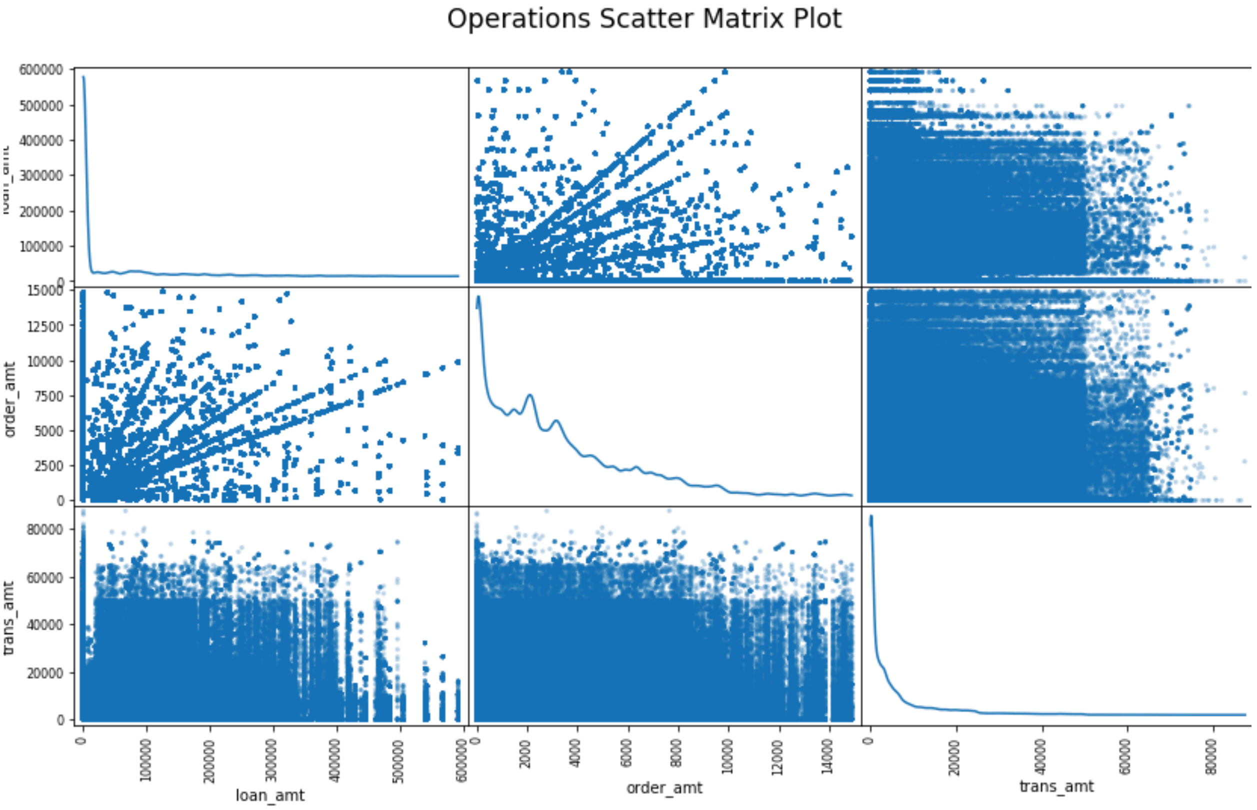

FIGURE 5 OPERATIONS SCATTER MATRIX PLOT

Figure 5 Operations Scatter Matrix Plot shows that there is a group-related pattern of two operations: order and loan, and we can notice that transactions are totally unrelated to the others.

So, it seems that the more common algorithms are going to find hard the search for patterns. That's why we has proposed two models for this investigation of patterns and novelties in this dataset:

1. Benchmark Model, **PCA** (Principal Component Analysis) for selecting features and **GMM** (Gaussian Mixture Model) for grouping the previuos selected features.

2. Solution Model, the bet is for this model, it is a **Deep Autoencoder**, which is using neural networks to work.

More on this in the coming sections.

4. Benchmark model

As it was discussed at the beginning of this report, The Bank has not labeled operations as fraud or normal so far. So, the we are helping them with this labeling process in order to detect likely fraudulent operation in advanced and act accordingly.

However The Bank managers have estimated a rule of thumb: one of a thousand operations must be investigate further. So if the dataset counts with:

$$6471(orders) + 1056320(transactions) + 682(loans) = 1{,}063{,}473(operations)$$

then: $1{,}063{,}473/1000 = 1063(novelties)$ might be identified for further investigation.

4.1 Algorithms and techniques.

The objective of this project is to find a way of detecting novelties on-time among the operations clients perform in The Bank. Such thing can be achieved setting a Threshold T in order to label operations. After that, human-being intervention is needed to validate and adjust that threshold.
The threshold T, in this approach, will be determined by the following technique:

1. Apply a PCA to the data set in order to select only the features that best describes the dataset.
2. Apply a GMM to the features previously selected in order to group features.
3. Measure the euclidean distance of each data point to each group center and set the maximun distance.

4. Reduce by a factor λ this maximun distance and then establish the Threshold T.
5. By compare the Threshold T to the max distance of each point to any of the centers, decide if it is beyond and should be marked as novelty.

4.2 Preprocessing.

In this section, a couple of preprocessing is applied to the data set.

1. The categorical columns are changed to dummies.
2. The columns that hold monetary values, are log-transformed in order to make the distributions less skewed.

4.3 Implementation.

PCA or Principal Component Analysis is a statistical algorithm for dimensionality reduction by orthogonal transformation to a data set of possibly correlated variables that explain as much of the variability of the dataset.

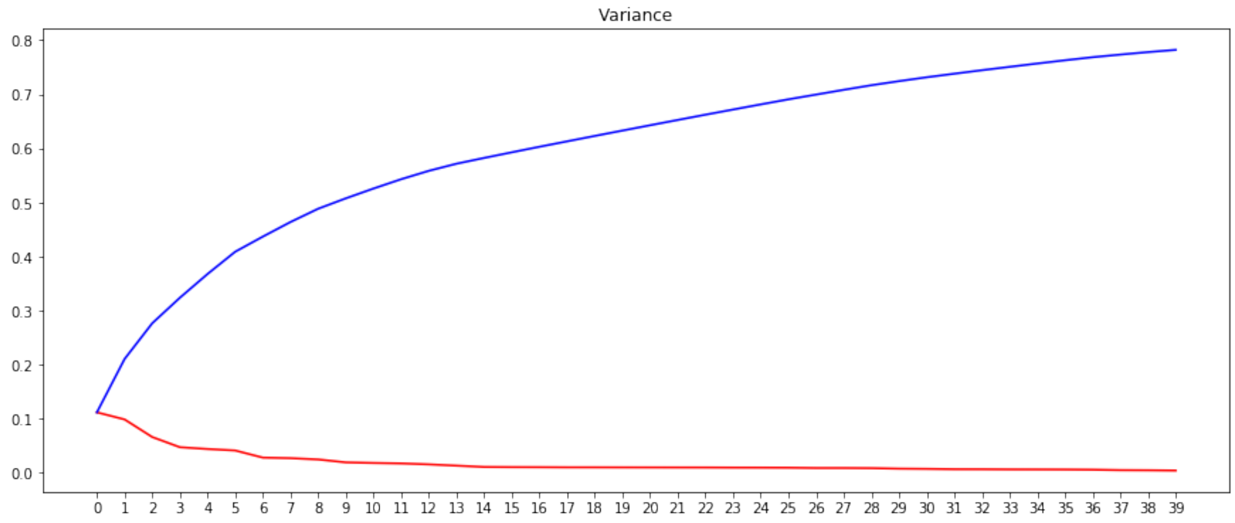

Figure 6: Scree Plot

In **Figure 6: Scree Plot** can be noticed an "elbow" by the component 3. So, by taking the first 3 components [0,1,2] that are before this one accordingly to this graph is a good choice. Which is also convenient to make a 3D plot and keep going in visualize identifying novelties.

The total amount of variability explained by these 3 components is 27.64%. We have determined this is a good trade-off in between the amount of variability explained and the benefit of 3D plot we can get by taking these 3 Principal Components.

In **Figure 7: 3D PCA plot** we can visualize clusters, and novelties are barely noticed apart in the space, in other words points away from any of the clusters. Let's apply the algorithms to calculate mathematically those points.

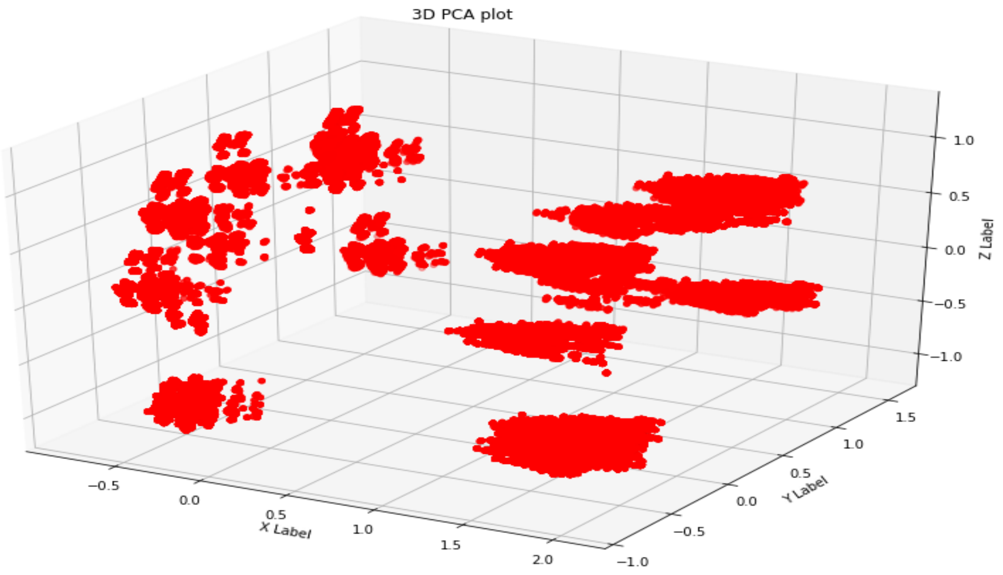

Figure 7: 3D PCA plot

GMM or Gaussian Mixture Model proposes that all data points belongs to one of the finite mixture of gaussians distributions with unknown parameters.

In **Figure 8: 3D Novelties detected by threshold** it is noticed a total of **244377** data points classified and depicted as novelties by this model.

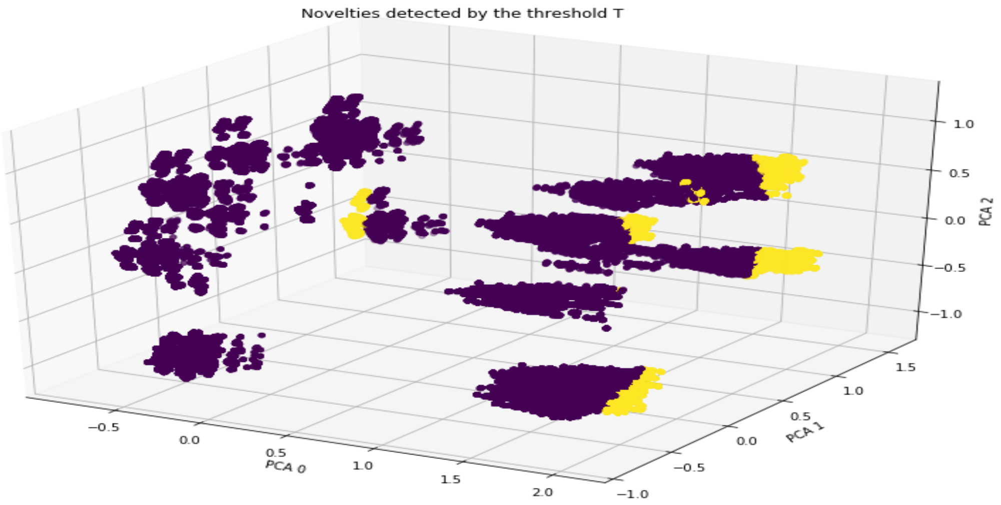

Figure 8: 3D Novelties detected by the threshold T

4.4 Refinement

The novelty detector was tested with 3 different values of the λ parameters in order to establish the threshold T.

As expected by increasing λ the threshold T gets narrower resulting in less number of novelties detected.

Lambdas	0.7	0.8	0.9
Novelties	2231549	1597725	244377

However, is was not enough for The Bank Managers as they told Machine Learning team, there are still too many novelties to be analyzed and labeled by humans (>≈1000).

5. Solution Model.

Once the Benchmark Model has shown up its results, let's propose a different approach, let's see this Solution Model which is using Deep Autoencoders to catch distant data points from the mean.

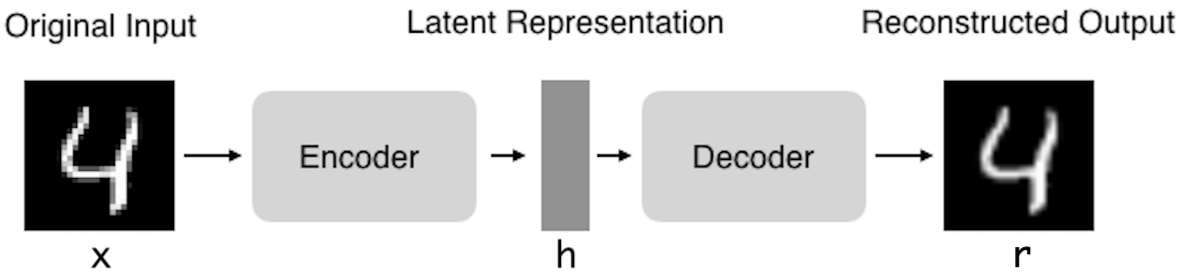

FIGURE 9: ARCHITECTURE OF AUTOENCODER

The Benchmark Model only takes into account the 27.64% of the variability by selecting the Principal Components. This might be enough, but in this approach, the model allows a more advanced technique by taking the whole set of features, this way a wider picture of the dataset is taken into account, thus making this approach smarter. So, the system is able to identify those novelties from non-linear related datasets (as shown in Figures 2,3 and 4) easier and more reliable.

Note: Zocca, Valentino, et.al. Python Deep Learning. Birmingham, UK. Pack Publishing LTD. April 2017. Chapter 9 'Anomaly Detection'

5.1 Algorithms and Techniques.

The Deep Autoencoders are layers of neural networks that copy the input data to the output. But in the middle, there is a layer that compress the most relevant features of the input (Encode) into a latent representation h and then reconstructs (Decode) from this latent representation to the data again fully fledged of all features r. Of course, the reconstructed data has some reconstruction error, and this parameters is going to be taken into account to label a datapoint as novelty. This is because the distant datapoints have a higher reconstruction error.

In **Figure 9: Architecture of Autoencoder** we can see the explanation above.

Note: I'd like to mention the source of my explanation and image about the Autoencoders: Nathan Hubens (February 25, 2018) Deep inside: Autoencoders [Blog post]. Retrieve from: https://towardsdatascience.com/deep-inside-autoencoders-7e41f319999f

So, this algorithms are apply as follow:

1. Train the Deep Autoencoder with the data_set.
2. Pass each data point through it and get the reconstruction error for each one.
3. Calculate the max reconstruction error.

4. Get the threshold T by reducing the max reconstruction error by a λ factor.
5. Compare each point reconstruction error to the Threshold T and decide if it should be marked as novelty.

5.2 Preprocessing

Recall from the Benchmark Model where a couple of preprocessing where applied, the same resulting transformed dataset is going to be used here because the same preprocessing is valid for this Solution Model.

5.3 Implementation.

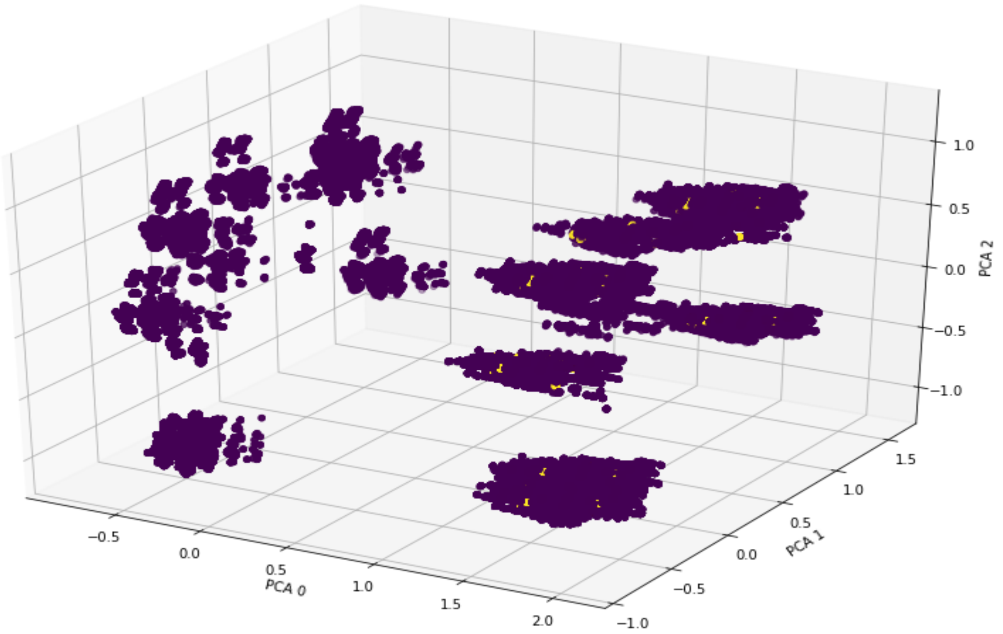

Figure 10: 3D Novelties detected by the threshold T in Autoencoder model

In **Figure 10 Novelties detected by threshold T in Autoencoder model**, we barely see the only 3466 novelties detected. This is a less than novelties detected by the GMM as this algorithm is much more precise.

In order to be able to graph a 3D plot like the one above, the PCA dataset with the first 3 principal components were used, but, depicting the points detected by the Autoencoder.

5.4 Refinement.

As expected, by increasing the λ parameters the threshold T is narrower resulting in less number of novelties detected.

In this Solution model, The Bank Manager accepted the number of novelties detected because it is a more manageable number of novelties to be analyzed by humans and start labeling each one of them and established a solid ground truth.

Lambdas	0.7	0.8	0.8
Novelties	21648	3466	300

6. Results.

6.1 Model Evaluation and Validation.

The topology of Deep Autoencoders is as big as the combinations among layers, filters and connections can be done in between them. Because of computationally limitations the validation of them was limited to understand the purpose of some of them, the one chosen is described below:

1. The one chosen is the "Vanilla" one, which means it is the simplest topology one.
2. It is only composed by the input layer with the same number of features in the dataset, the encoder with significant less size in order to let them learn the principal features, and the output layer with the same size as the number of features in the data set.
3. A dropout layer was added with a rate of 0.3, in order to avoid overfitting the dataset.

Also, three values of the λ parameter were validated in both models, the explanation of that in the following section where comparing the two models is done.

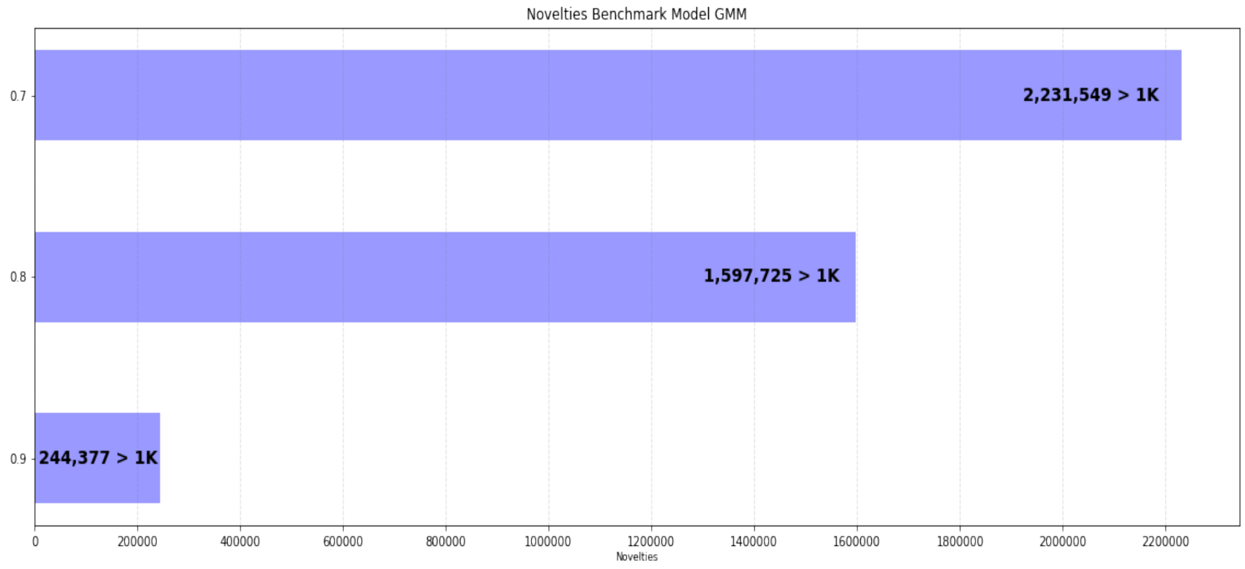

Figure 11: Novelties Benchmark Model GMM

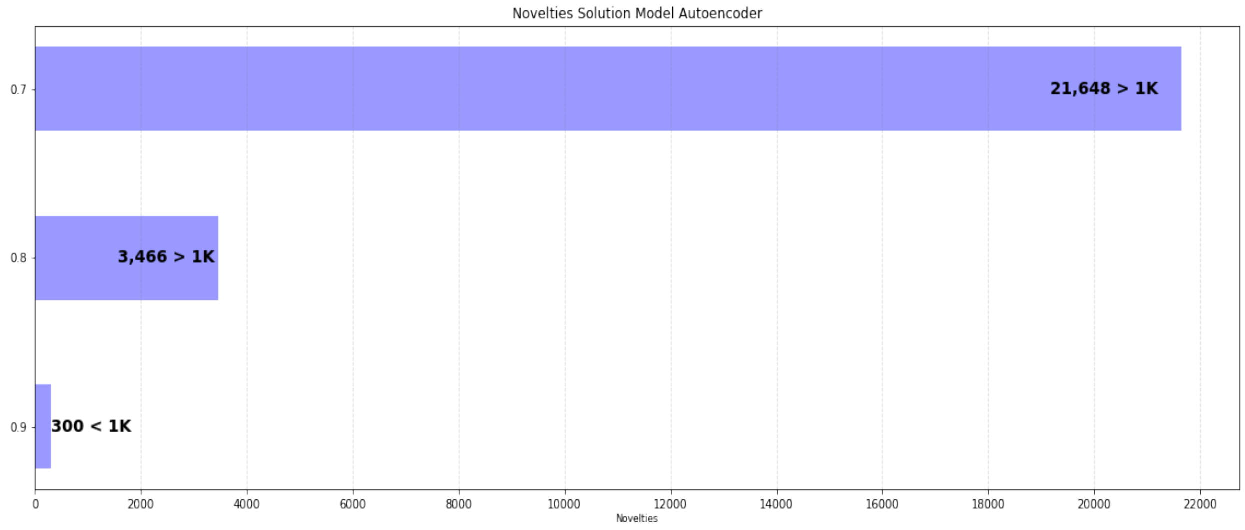

Figure 12: Novelties Solution Model Autoencoder

6.2 Justification.

The experimentation was done taking three values [0.7,0.8,0.9] for the λ parameter. With same value applied to both models we can notice the following:

In Figure 11: Novelties Benchmark Model GMM, the **Benchmark Model (GMM)** with λ=[0.7] we notice that almost the entire data set was identied as novelty, which is not acceptable at all. Nor with λ=[0.8], although it decreases the number of novelties identified, it is still not acceptable. The most significant approach was obtained with λ=[0.9], although the number of novelties identified stills exceeds the rule of thumb set by the managers (≈1000).

In Figure 12: Novelties Solution Model Autoencoder, the other hand, the **Solution Model (Autoencoder)** with λ=[0.7] got a result greater than the rule of thumb. But the model got a much better result with the following two values λ=[0.8,0.9] identifying less than the rule of thumb (≈1000) for the both values.

So, the **Solution Model (Autoencoder)** gives The Bank manager a good start for further research in identifying anomalies in their client's operations and act accordingly.

7. Conclusion.

7.1 Free-Form Visualization.

A long the implementation of the models, several 3D plots were depicted in order to see how the number of datapoints identified as novelties were changing along the algorithms were applied to the data set.

- Figure 8: 3D Novelties detected by threshold.
- Figure 10 Novelties detected by threshold T in Autoencoder model.

7.2 Reflection.

This project was about a kind of identifying outstanding data points in a dataset, in order to start a baseline for label anomalies in the same dataset. Those anomalies are going to be used as warning against fraud in the 3 main operations The Bank serves to its clients.

Above all, The Bank Managers has established a rule of thumb accordingly to the industry past experience of 1 fraud operation among a thousand.
Taking that rule of thumb into account, a couple of models were proposed, in order to compare one against the other a determine the better performance of the one proposed as Solution.

First model was a Gaussian Mixture Model who was the benchmark, the second was a Deep Autoencoder using neural networks. In both, in order to determine a novelty point, an Euclidean Distance was calculated and a Threshold proposed in order to reach the number of novelties accordingly to the rule of thumb.

The interesting thing in this project was about having learnt how powerful and useful the neural networks and the such big different topologies could be to solve practical problems in a dataset.

I found hard to solve the problem, the fact that The Bank does not have a ground truth. Such important component for Machine Learning projects came to my mind as a fundamental piece to have a successful project.
However, the lacking of ground truth does not stop this approach that helped a lot to The Bank in order to achieve their goal of protecting its clients.

7.3 Improvements.

Despite of the fact this is the first version of a fraud protection system, and the problem was partially solved. A more accurate Solution model should be trained once the ground truth is recorded.

Once that ground truth is recorded for a considerable period of time, a Deep Encoder Model with a similar topology can be trained to calculate new parameters and the parameters of this Solution Model can be used as benchmark for that one.

Also, in this project if we had understood how the **likelihood** can be calculated, we have applied that formula for identify novelties.