Kay Quiballo | MSDS 411 | 05.31.2023

MSDS 411 Assignment 4: Anomaly Detection

Abstract. An executive summary of the research.

In this assignment, we explore different unsupervised methods of anomaly detection in the context of sales transactions. Using variables such as quantity and value of sales from the 2017, 2021 Torgo case study, we are tasked with identifying the best methods identifying anomalies in a dataset. Using KNN distances, DBSCAN/LOF, Isolation Forests, and metrics such as precision/recall/F1, we make our informed recommendation to the management problem.

Introduction. Why are you conducting this research?

This assignment involves the study of anomaly detection in the context of sales data. By looking at financial transactions reported by sales representatives, we are tasked with using different unsupervised methods such as DBSCAN/LOF and isolation forest to identify anomalies. Fraud can be deliberately misreported by sales representatives in order to receive larger commissions.

Literature review. Who else has conducted research like this?

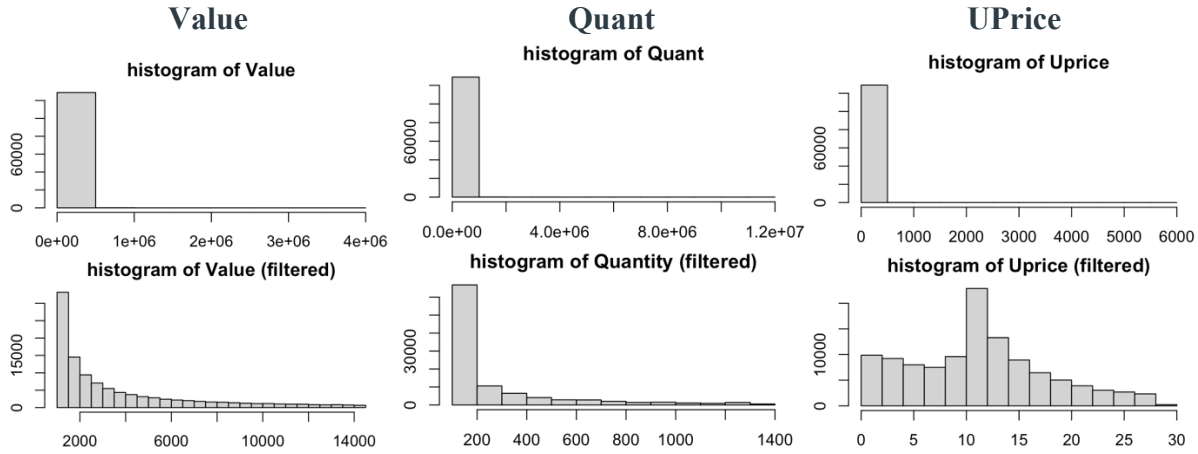
Training and test data comprise a systematic sample of sales observations from a case study reported by Torgo (2017, 2021).

Methods. How are you conducting the research?

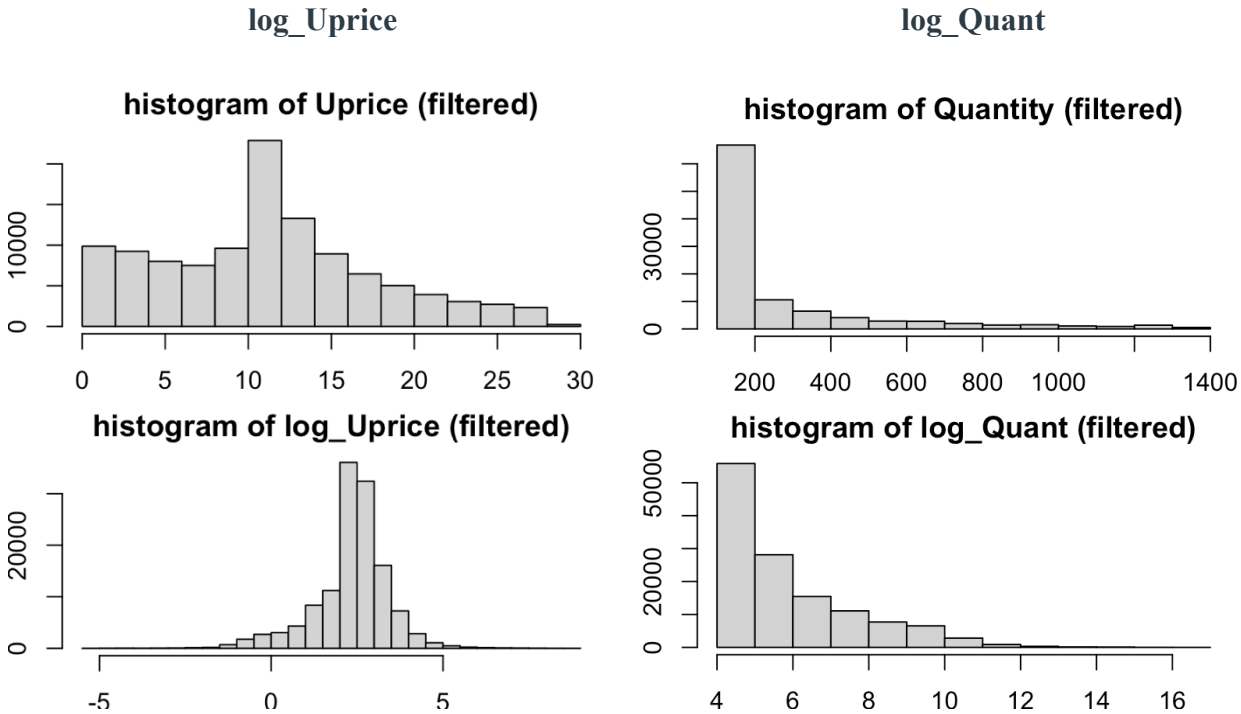
From the Assignment instructions: “The training data consist of 133,731 unlabeled sales transactions across 798 products. Test data include 15,732 inspected sales transactions across these same products, where manual inspection found 14,462 transactions to be normal and 1,270 fraudulent.” Each transaction is labeled either normal (ok) or anomaly (fraud).

The data consist of the following variables:
ID = sales representative identification (not to be used in the anomaly detection model)
Prod = product number
Quant = quantity of the product that the sales representative claims to have sold
Val = the value of the total sale that the sales representative claims to have sold (currency units are unknown but most likely €)
Insp = result of inspection with known values in the test set ("ok" or "fraud") and missing/unknown ("unkn") in the training set

Let us look at the distribution of data:



The unsupervised methods we will be using in the assignment can only consider numeric values, so we will only look at the numeric variables from the training dataset. Both Value and Quant show a highly skewed right distribution with large outliers, we used the 1.5*IQR method to determine the maximum cutoff and determine a more reliable distribution. We also combined Value divided by Quant to determine the Unit Price and used this in place of Value. After filtering using the IQR method, we see the distributions of Value and Quant are still skewed right.

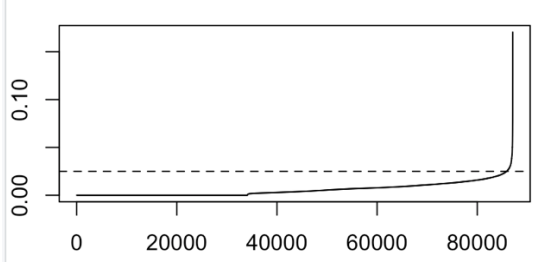


After considering the uneven distributions of data, we will scale the data we want to use in analysis using a log transformation. This makes a distribution of log_Uprice more normally distributed as well as brings in the tail end of log_quant. Given the imbalanced data, we will be looking at the F1 score which combines both precision and recall.

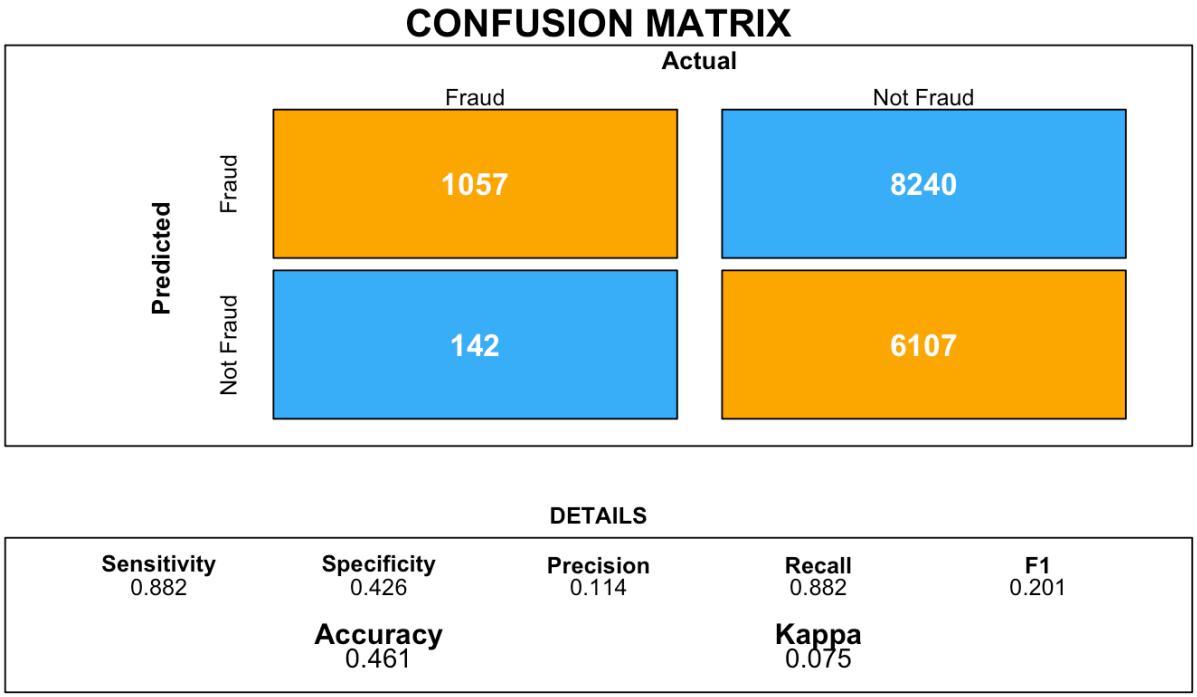
Results. What did you learn from the research?

1. DBSCAN Results

KNN- distance plot with eps=0.025

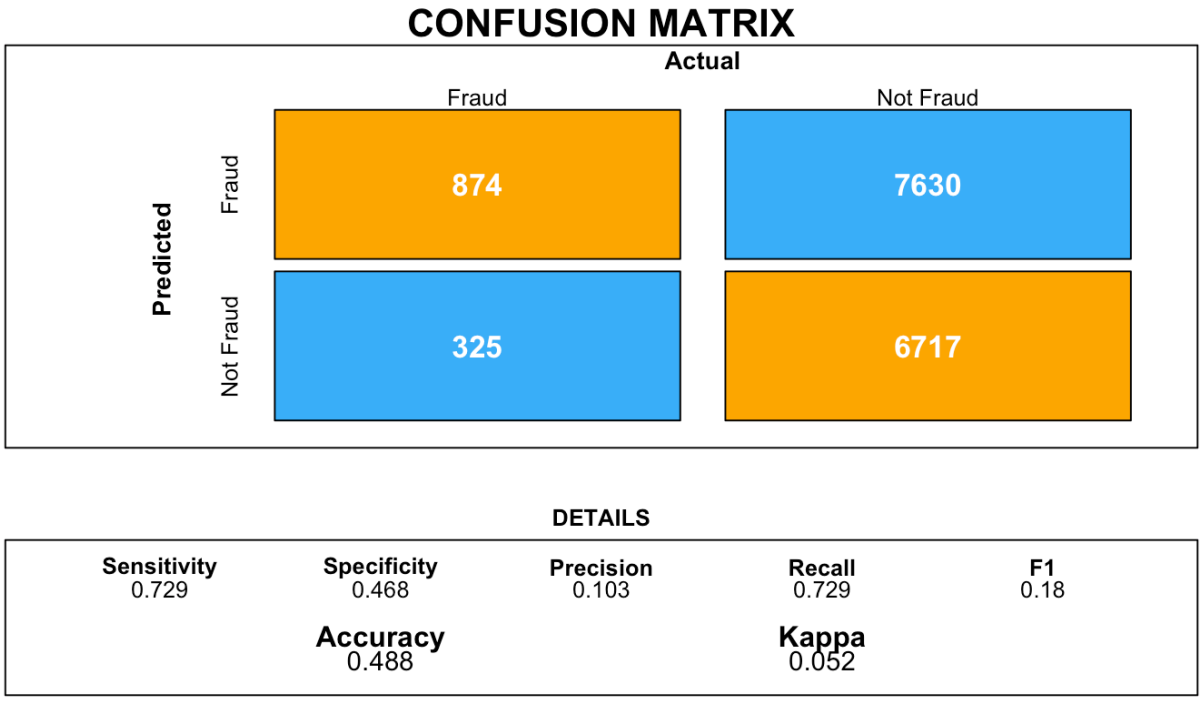


By examining the distance plot, we can determine the optimal value for epsilon is 0.025. Arbitrarily, we choose minPts for our DBSCAN model to be 5. The results are the following:



We note that the model has an accuracy of 0.461. However, with the imbalance of data, we care more about precision (0.114), recall (0.882), and F1 (0.201). We will compare these metrics across models.

LOF Results



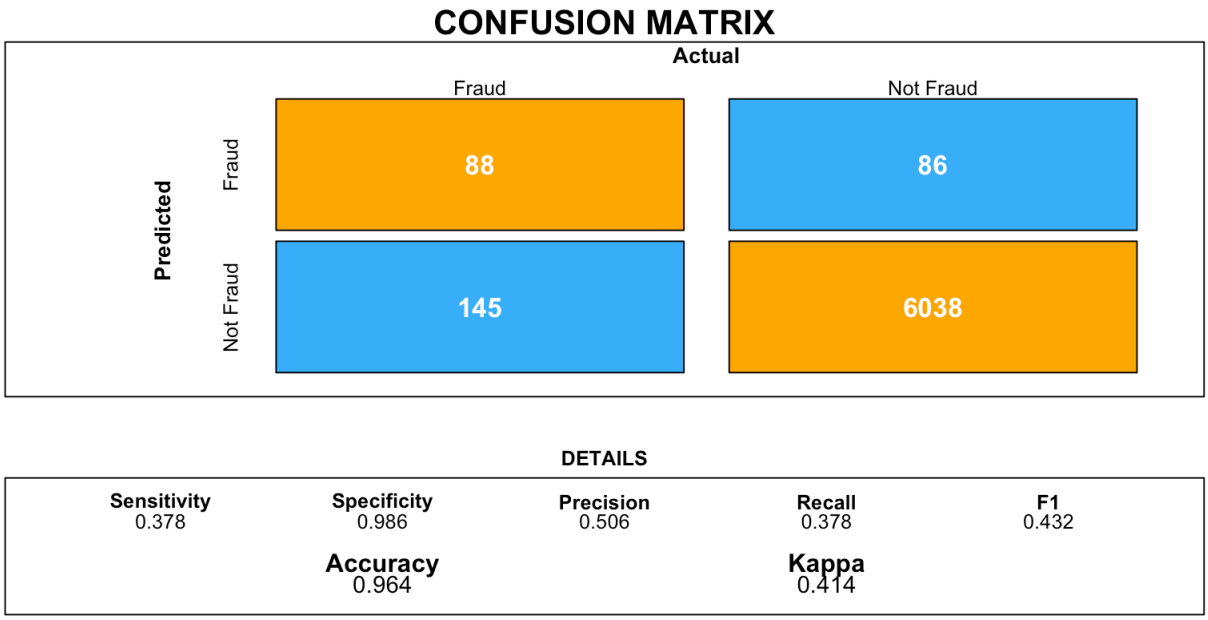
From the Local Outlier Factor Score analysis, we yield similar results to the DBSCAN: We note that the model has an accuracy of 0.488. The metrics of interest are precision (0.103), recall (0.729), and F1 (0.180). We will compare these metrics across models.

Isolation Forest Results:

There are 3 types of forests that were run in this analysis. The plain isolation forest has the highest AUROC area under the curve which indicates it is the best fitting analysis.

Model	AUROC
Isolation Forest	0.8682328
Density Isolation Forest	0.8484808
Fair-Cut Forest	0.8548061

Here are the results from the isolation forest:



From the isolation forest we see a high accuracy (0.964) and specificity (0.986). The metrics of interest are precision (0.506), recall (0.378), and F1 (0. 432).

Conclusions. So, what does it all mean?

	Precision	Recall	F1
DBSCAN	0.114	0.882	0.201
LOF	0.103	0.729	0.180
Isolation Forest	0.506	0.378	0.432

In our results, precision is lower if we have too many false positives. Recall is lower if we have too many false negatives. We see that DBSCAN/LOF have lower precisions and are therefore more willing to label transactions as fraud even if they are not. The isolation forest however has a higher precision but a lower recall, meaning that it is more willing to label transactions as "ok" even if it is fraud. However, relatively, the isolation forest does this much less often as indicated by the F1 score which is significantly higher than that of the DBSCAN/LOF.

In the context of the management problem, ideally we would like an unsupervised anomaly detector that has a high precision and a high recall. With the give and take of these metrics, our best solution is the isolation forest which has the highest F1 value. Although accuracy and specificity are not the most important metrics of choice due to the imbalance of data, this method does really well across both in identifying overall fraud vs non-fraud (accuracy=0.964, specificity=0.986) which is also better than DBSCAN/LOF. That being said, we must take caution as this method may label not detect as many fraudulent transactions as DBSCAN/LOF (but this is at the cost of overlabeling transactions fraudulent that are actually "ok").

My advice to the management of this problem is to implement some form of an isolation forest to help identify some of the fraudulent anomalies, but then have a human set of eyes or another model to sift through the results to determine which of the following identified fraudulent activity are indeed fraudulent.