USING APRIORI ASSOCIATION RULES AND DECISION TREE ANALYSIS FOR DETECTING CUSTOMS FRAUD

Hamzah Qabbaah, George Sammour, Koen Vanhoof

***Abstract**:*

All over the world some people and companies try to avoid taxes whenever possible. Customs are not an exception to this. In this paper we investigate how customs fraud can be detected using data mining on logistics transaction data. We used both the Apriori algorithm for association rules and decision tree analysis to do so. We first transformed the continuous variables using both k-means clustering and CHAID decision tree analysis for the continuous variables in the data set. Analysis of the rules detected by our analysis indicates that it is possible via this methodology to find indicators of customs fraud cases. Moreover, it was possible to describe in detail the situation in which the fraud occurred (product type, country of origin, e-shipper, the weight and the price of the products shipped). The results of the decision tree analysis proved to verify the connection between both Apriori models used and showed similar results, and the evaluation using the classification measures (Accuracy, Precession, Recall and F1) based on the confusion matrix shows high percentages. This confirms the value of our conclusions.

***Keywords**: Customs fraud, CHAID decision tree, Apriori association rules, Data mining, Logistics.*

1. Introduction

Any act by which a person deceives, or attempts to deceive, the customs and thus evades, or attempts to evade, wholly or partly, the payment of import or export duties and taxes constitutes customs fraud (ACFE 2018) . Clearly both companies and governments would benefit from a system that could detect this as soon as possible in order to avoid hassle.

The literature on fraud detection focuses mainly on internet fraud by owners of websites to increase their income, the so-called ‘Click-fraud’. This fraud can have several origins. Some web-owners are simply dishonest and use automation to generate traffic to defraud advertisers (Metwally, Agrawal et al. 2007). Several fraud detection algorithms have been proposed for this problem. A new architecture for web fraud detection using an Apriori algorithm for association rule mining in a web advertising network was introduced by Tripathi et al. (Tripathi, Nigam et al. 2017). Zhang et al. presented a technique to detect duplicate clicks in jumping windows and sliding windows. They did so using two innovative algorithms that make only one pass over click streams (Zhang and Guan 2008). Haddadi defined bluff ads, which are a group of ads that join forces with the intention of increasing the effort level for click-fraud spammers (Haddadi 2010). Mittal et al. investigated a possible way to find fraud by falsely manipulating forbidden information on customers (Mittal, Gupta et al. 2006).

Finally, Triepels et al. investigated whether intelligent fraud detection systems can improve the detection of document fraud by miscoding and smuggling by analysing large sets of historical shipment data. They developed a Bayesian network that predicts the presence of goods on the cargo list of shipments. The predictions of the Bayesian network were compared with the accompanying documentation of the shipment to determine whether document fraud was perpetrated (Triepels, Daniels et al. 2018).

None of the above literature talks about customs fraud. To the best of our knowledge, our field of application is therefore novel. We used two algorithms to find indicators of customs fraud (detection being a novel dimension for the phenomenon as well). They are based on the Apriori association rules and decision tree analyses.

2. Problem statement and Research question

All over the world some people and companies try to avoid customs whenever possible. Therefore, detecting customs fraud is important part to logistics companies, their customers and governments in a world of growing e-commerce.

Consequently, the following research question has to be answered: How can customs fraud be detected using data mining on logistics transaction data? Developing a methodology capable to find some indicators of customs fraud in such logistics transactions on the basis of data mining technology is the objective of this research.

3. Methodology

To answer our research question we start from a real life dataset, we will create a proxy for customs fraud as we will explain in the data section later. Three models will be applied on the data. The Apriori algorithm for association analysis and decision tree analysis will be used for this fraud detection research. We will use Apriori analysis as it is one of the most frequently used methods for generating association rules and searching patterns in large databases (John and Shaiba 2019, Silva, Varela et al. 2019). It is very fast and memory friendly when generating rules comparing to other association rule algorithms. Moreover, it can be used as a supervised method to discover interesting patterns and rules for selected variable (Prithiviraj and Dr.R 2015). Decision tree analysis will be used to verify whether its results could establish a connection between the association rules resulting from the apriori analysis. The confusion matrix will be used in this study for the evaluation by determining the classification measures: Accuracy, precision, recall and F1. In this paper, we will first explain the details of the association rules and the Apriori algorithm and the CHAID decision tree analysis. Second, we will follow the research process from data collection via preprocessing of the data to explanation of the analyses and results.

3.1 Apriori Association Rules

Association Rules are a data mining technique. Basic association analysis deals with the simultaneous occurrence of several items with one another whereas deeper association analysis can take into consideration the quantity of the joint occurrence and sequence of occurrence, etc. The method for finding association rules through data mining involves several steps (Linoff and Berry 2011, Kotu and Deshpande 2014). The sequence is the following:

- Prepare the data: an association algorithm needs input data to be formatted in a particular format.
- Short-list frequently occurring item sets. Item sets are combination of items. An association algorithm limits the analysis to the most frequently occurring items. The final and meaningful rules are extracted in the next step.
- Finally, the algorithm generates and filters the rules based on the interest measure (Kotu and Deshpande 2014).

All association rules algorithms try to find frequently occurring item sets within a base of possible item sets. **The Apriori algorithm** is the most frequently used. It leverages some simple logical principles on the item sets to reduce the number to be tested for a certain support measure (John and Shaiba 2019). The algorithm is based on the principle that ‘If an item set is frequent, then all its subset items will be frequent’ (Tan, Steinbach et al. 2005). The name of the algorithm is based on the fact that the algorithm uses prior knowledge of properties of frequent items (Kotu and Deshpande 2014).

Apriori association rules employ an iterative approach known as a level-wise search, where k- item sets are used to explore (k +1) item sets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy a minimum support. The resulting set is denoted by L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-item sets can be found. The finding of each Lk requires one full scan of the database (Kotu and Deshpande 2014).

3.2 The Chi-Squared Automatic Interaction Detection (CHAID) algorithm

Decision tree analysis is a data mining technique that encompasses several algorithms to predict a dependent variable (Qabbaah, Sharawi et al. 2017). These predictions are determined by the influence of independent predictor variables (Qabbaah., Sammour. et al. 2019). It is “a structure that can be used to divide a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. Decision trees “use a set of rules for dividing a large heterogeneous dataset into smaller, more homogeneous groups with respect to a particular target variable” (Linoff and Berry 2011). With each successive division, the members of the resulting sets become more and more similar to one another”(Linoff and Berry 2011).

Each procedure goes as follows. “After the Dataset is split into n nodes, the process is repeated on each node until the data are completely partitioned or until a stopping rule is met. This means that there are no longer enough cases to partition, or only one case remains for the last node. The entire classification of data is then visualized through a graphical tree, which explains the interaction of x (x1, x2, …, xm) and y” (L. and Jr 2012). The classification procedures will attempt to provide as much separation as possible with regards to classifying data into correct cases (Long, Griffith et al. 1993). Graphically, decision trees will produce a tree T which is composed of a root node and child nodes (De'ath 2000) (Ripley and Hjort 1995). The tree is essentially a connected graphic that is inverted. However, the tree display is user specific (Loh and Shih 1997, L. and Jr 2012).

There are several decision tree classification procedures. The algorithm used in this research is **Chi-Squared Automatic Interaction Detector (CHAID)**. CHAID algorithm is one of the most popular classification procedures (Loh and Shih 1997) because it can manage both continuous and categorical variables and handle a very large dataset with high speed performance (Díaz-Pérez and Bethencourt-Cejas 2016). CHAID uses Chi-square test for splitting

With regards to splitting, CHAID is not bound to binary splits and allows for multiple level splits at each node. The algorithm searches for the best possible split for the different values of the independent variable then determines if the data should be merged to form a node or split in the data. Essentially, the best predictor is selected to begin splitting the data. During this process, CHAID selects the best predictor based on comparisons of the adjusted p-value for each predictor (L. and Jr 2012). This is a three step process for determining splitting. The splitting and merging procedures for CHAID are independent of one another.

The algorithm uses as sequential process where splitting and merging occur at the same time. The basis behind this aspect of the algorithm is rooted in the computation time of T. CHAID has difficulty in determining when and where to stop (Loh and Shih 1997). Stopping is also a three steps process determined by the chi-square statistic. As such, stopping occurs when the critical level of the chi-square test fails to meet the test.

4 Data and Data preprocessing

The research framework of this paper is: data collection, preprocessing of the data, analysing and obtaining the results.

The data used in this research were obtained from an international logistics services company. Cleaning and merging tables of the data has been applied in order to obtain the final data set. The total number of transactions in the final dataset is equal to the size of the sample (n=8,243). Some important elements have to be mentioned that guided us in selecting the variables used in the analysis models.

The product type, weight, quantity, destination and country of origin are logically used. This is amongst others based on what the customs regulation in Gulf Cooperation Council states (GCCStates 2018). We will use COD Value USD and Customs Values USD to define the fraud proxy. Table 1 shows the description of the variables used in this research.

Table 1. The description of the data variables.

Variable	Data type	Data description
Variables in the original dataset		
ID	Integer	The ID of the order
Weight In KG	Double	The weight of shipment in KG
CODValueUSD	Double	The amount of cash on delivery in USD
Product Group name	String	Product category group name of the shipment.
Product group ID	Integer	Product category group ID
Customs Value USD	Double	The declared value of the goods for customs department in USD
Destination	String	Destination city of shipment
Origin Country	String	Country of origin of the shipment
DestCountry	String	Country of destination of the shipment
ShipperID	Integer	The ID of the E-commerce companies
Created variables		
COD Customs amount USD	Double	The customs value of the goods measured using COD Value USD
Customs amount USD	Double	The customs value of the goods measured using the declared customs value USD
Weight K-mean class	Integer	Transformation classes of the weight using k-means
Weight CHAID class	Integer	Transformation classes of the weight using CHAID decision tree
Value USD K-mean class	Integer	Transformation classes of the COD value of using k-means
Value USD CHAID class	Integer	Transformation classes of the COD value using CHAID decision tree
Customs Fraud class	Boolean	Indicators of the fraud. Class:Yes(1),No(0)

The variables ‘COD Customs amount USD’ and ‘Customs amount USD’ have been calculated using the import duties of the products multiplied by the goods value (COD or the declared value to customs department) taking into consideration whether the shipment was sent within the same country or not. The variable ‘Customs Fraud class’ has been created by dividing the value of customs paid using the declared value of the goods to customs department (‘Customs amount USD’) over the new calculated customs value using the COD value which is the real amount of the goods that the customer should pay (‘COD Customs amount USD’). If the answer is less than 0.8 (indication of fraud) or greater than 1.25 (overpayment, not matched in our data) then we identified the Fraud class as 1 (there is fraud), otherwise the Fraud class was identified as 0 (no fraud).

All the variables should be nominal or categorical as an input for the Apriori association rules. Therefore, converting the continuous variables such as weight and price to nominal is essential. This transformation process will be performed using clustering and decision tree analysis as follows:

1. **Transformation using k-means clustering:** The main objective of K-means is to partition the dataset into k clusters in which each instance belongs to the cluster with the nearest mean (Qabbaah, Sammour et al. 2019).

   To transform the continuous variables taking into consideration ‘Fraud Class’ which is our goal in this research, we created a k-means cluster model with three attributes, ‘Fraud class’, ‘weight in KG’ and ‘COD Value USD’. Table 2 shows the results. The variables ‘Weight K-mean class’ and ‘Value USD K-mean class’ have been created for each transaction.

Table 2. Transformation results using K-means clustering.

Clusters	Avg.Weight In Kg	Avg.COD Value USD	Fraud class	Number of Items
Cluster 1	0.79383	72.166	1	551
Cluster 2	2.0341	131.96	1	778
Cluster 3	6.891	158.02	1	258

2.Transformation using decision tree: To transform the continuous variables taking into consideration ‘Fraud Class’ which is our goal in this research, we created two CHAID decision tree models. The dependent variable of the first model is ‘Fraud class’ and the independent variable is ‘Weight in KG’, whereas the dependent variable of the second model is ‘Fraud class’ and the independent variable is ‘COD Value USD’. Table 3 shows the categories of ‘Weight in kg’ variable and Table 4 shows the categories of ‘COD Value USD’ variable.

Table 3. ‘Weight in kg’ categories using CHAID decision tree analysis.

Class	Value	Class	Value
1	<= 0.230	**6**	(1.640 , 2.450]
2	(0.230 , 0.490]	**7**	(2.450 , 5.380]
3	(0.490 , 0.750]	**8**	(5.380 , 8.130]
4	(0.750 , 1.120]	**9**	(8.130 , 19.780]
5	(1.120 , 1.640]	**10**	>19.780

Table 4. ‘COD Value USD’ categories using CHAID decision tree analysis.

Class	Value	Class	Value
1	<= 39.5790	**5**	(134.702 , 168.061]
2	(39.579 , 98.397]	**6**	(168.061 , 263.188]
3	(98.397 , 117.657]	**7**	>263.188
4	(117.657 , 134.702]		

5 Analyses and Results

After applying the pre-processing phase, we are ready to use the data for further analyses. In this section, we will first present the results of the Apriori association rules analysis. Next we will show the results of the CHAID Decision tree analysis.

5.1 The analyses and the results of customs fraud detection apriori models (models 1 and 2)

Two models will be created for this fraud detection. Model-1 uses K-means transformation and model-2 uses CHAID transformation.

1. The variables used in **model-1** are (weight class k-mean, COD value class k-mean, product group name, destination, origin country, shipper-id, fraud class). Table 5 shows the Apriori run information for the first model using K-means transformation.

Table 5. Apriori run information for the first model using K-means transformation.

Instances	Minimum support	Minimum metric <confidence>
8234	0.01	0.1

Table 6 shows the summary of the important rules found as a result of the Apriori association rules algorithm using K-means transformation for the continuous variables.

Table 6. Summary of the important rules found when detecting customs fraud with model-1 using k-means transformation for the continuous variables.

N	The Rules
1	Weight class k-mean=2, COD class k-means=3, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 224 ==> _Fraud class=1. 224 cases, confidence:(1)
2	Weight class k-mean=2, COD class k-means=2, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 178 ==> _Fraud class=1. 178 cases, confidence:(1)
3	Weight class k-mean=2, COD class k-means=3, Destination=JED, Origin Country=HK, Shipper ID=599818. 168 ==> _Fraud class=1. 168 cases, confidence:(1)

N	The Rules
4	COD class k-means=3, Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 137 ==> _Fraud class=1. 137 cases, confidence:(1)
5	Weight class k-mean=2, COD class k-means=2, Product Group Name=Apparel, Destination=JED, Origin Country=HK Shipper ID=599818. 113 ==> _Fraud class=1. 113 cases, confidence:(1)
6	Weight class k-mean=2, Product Group Name=Shoes, Origin Country=HK, Shipper ID=599818. 115 ==> _Fraud class=1. 104 cases, confidence:(0.9)
7	Weight class k-mean=2, Destination=DHA, Origin Country=HK, Shipper ID=599818. 113 ==> _Fraud class=1. 101 cases, confidence:(0.89)
8	Product Group Name=Bag/Case, Origin Country=HK, Shipper ID=599818. 342 ==> _Fraud class=1. 126 cases, confidence:(0.37)
9	Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 1274 ==> _Fraud class=1. 607 cases, confidence:(0.48)
10	Product Group Name=Shoes, Origin Country=HK, Shipper ID=599818. 312 ==> _Fraud class=1. 133 cases, confidence:(0.43)
11	Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 1094 ==> _Fraud class=1. 453 cases, confidence:(0.41)
12	Weight class k-mean=2, Product Group Name=Bag/Case, Origin Country=HK, Shipper ID=599818. 129 ==> _Fraud class=1. 108 cases, confidence:(0.84)
13	Product Group Name=Apparel, Destination=DHA, Origin Country=HK, Shipper ID=599818. 272 ==> _Fraud class=1. 118 cases, confidence:(0.43)
14	Weight class k-mean=1, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 779 ==> _Fraud class=1. 169 cases, confidence:(0.22)

N	The Rules
15	Weight class k-mean=1, COD class k-means=1, Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 751 ==> _Fraud class=1. 151 cases, confidence:(0.2)
16	Weight class k-mean=1, COD class k-means=1, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 741 ==> _Fraud class=1. 131 cases, confidence:(0.18)
17	Weight class k-mean=1, Product Group Name=Beauty Supplies, Destination=JED. 106 ==> _Fraud class=1. 15 confidence:(0.14)

As an example, rule number one reads as follows: If the shipment sent from Hong Kong to Riyadh and bought via e-commerce-ID=599818 and the product category was ‘Apparel’ with price class= 3, and weight class=2 as shown in Table 6.15, then 224 transactions were indicated as a customs fraud case. All rules have to be interpreted in this way. So, rule number 8 reads as: If the origin country of the shipment is ‘Hong Kong’ and bought via e-commerce-ID=599818 and the product category was ‘Bag/case’, then 126 transactions are indicated as customs fraud cases.

2. The variables used in **model-2** are (weight class CHAID, COD value class CHAID, product group name, destination, origin country, shipper-id, fraud class). Table 7 shows the Apriori run information for the second model using CHAID transformation.

Table 7. Apriori run information for the second model using CHAID transformation.

Instances	Minimum support	Minimum metric <confidence>
8234	0.01	0.1

Table 8 shows the summary of the important rules found in model-2 as a result of the Apriori association rules algorithm using CHAID transformation for the continuous variables.

Table 8. Summary of the important rules found when detecting customs fraud with model-2 using CHAID transformation for the continuous variables.

N	The Rules
1	COD value class Chaid=5, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 135 ==> _Fraud class=1. 135 cases, confidence:(1)
2	COD value class Chaid=6, Product Group Name=Apparel, Origin Country=HK, Shipper ID=599818. 94 ==> _Fraud class=1. 94 cases, confidence:(1)
3	Weight class Chaid=6 COD, value class Chaid=5, Origin Country=HK, Shipper ID=599818. 93 ==> _Fraud class=1. 91 cases, confidence:(0.98)
4	Weight class Chaid=7, COD value class Chaid=5, Origin Country=HK, Shipper ID=599818, 99 ==> _Fraud class=1. 91 cases, confidence:(0.92)
5	COD value class Chaid=5, Destination=JED, Origin Country=HK, Shipper ID=599818. 112 ==> _Fraud class=1. 102 cases, confidence:(0.91)
6	Weight class Chaid=8, Product Group Name=Apparel, Origin Country=HK, Shipper ID=599818. 105 ==> _Fraud class=1. 88 cases, confidence:(0.84)
7	Weight class Chaid=7, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 166 ==> _Fraud class=1. 120 cases, confidence:(0.72)

N	The Rules
8	COD value class Chaid=4, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 229 ==> _Fraud class=1. 151 cases, confidence:(0.66)
9	COD value class Chaid=4, Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 159 ==> _Fraud class=1. 103 cases, confidence:(0.65)
10	Weight class Chaid=6 COD, value class Chaid=4, Product Group Name=Apparel, Origin Country=HK, Shipper ID=599818. 193 ==> _Fraud class=1. 121 cases, confidence:(0.63)
11	Weight class Chaid=6, Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 201 ==> _Fraud class=1. 114 cases, confidence:(0.57)
12	Weight class Chaid=6, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 264 ==> _Fraud class=1. 146 cases, confidence:(0.55)
13	Weight class Chaid=5, COD value class Chaid=3, Origin Country=HK, Shipper ID=599818. 171 ==> _Fraud class=1. 84 cases, confidence:(0.49)
14	Weight class Chaid=5, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 244 ==> _Fraud class=1. 119 cases, confidence:(0.49)

N	The Rules
15	Weight class chaid=6, COD value class chaid=4, Origin Country=HK, Shipper ID=599818. 245 ==> _Fraud class=1. 157 cases, confidence:(0.64)
16	COD value class Chaid=3, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 208 ==> _Fraud class=1. 95 cases, confidence:(0.46)
17	Product Group Name=Apparel, Destination=DHA, Origin Country=HK, Shipper ID=599818. 272 ==> _Fraud class=1. 118 cases, confidence:(0.43)
18	Product Group Name=Shoes, Origin Country=HK, Shipper ID=599818. 312 ==> _Fraud class=1. 133 cases, confidence:(0.43)
19	COD value class Chaid=3, Destination=JED, Origin Country=HK ,Shipper ID=599818. 214 ==> _Fraud class=1. 91 cases, confidence:(0.43)
20	Weight class Chaid=6, COD value class Chaid=3, Origin Country=HK, Shipper ID=599818. 208 ==> _Fraud class=1. 85 cases, confidence:(0.41)
21	Product Group Name=Bag/Case, Origin Country=HK, Shipper ID=599818. 342 ==> _Fraud class=1. 126 cases, confidence:(0.37)
22	Weight class Chaid=5, Destination=JED, Origin Country=HK, Shipper ID=599818. 290 ==> _Fraud class=1. 101 cases, confidence:(0.35)

N	The Rules
23	Weight class Chaid=2 COD, value class Chaid=2, Product Group Name=Apparel, Origin Country=HK, Shipper ID=599818. 314 ==> _Fraud class=1. 104 cases, confidence:(0.33)
24	Weight class Chaid=4, Product Group Name=Apparel, Origin Country=HK, Shipper ID=599818. 442 ==> _Fraud class=1. 127 cases, confidence:(0.29)
25	COD value class Chaid=2, Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 643 ==> _Fraud class=1. 169 cases, confidence:(0.26)
26	COD value class Chaid=2, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 629 ==> _Fraud class=1. 161 cases, confidence:(0.26)
27	Weight class Chaid=5, COD value class Chaid=2, Origin Country=HK, Shipper ID=599818. 427 ==> _Fraud class=1. 93 cases, confidence:(0.22)
28	Weight class Chaid=4, COD value class Chaid=2, Origin Country=HK, Shipper ID=599818. 514 ==> _Fraud class=1. 109 cases, confidence:(0.21)
29	Weight class Chaid=3, COD value class Chaid=2, Origin Country=HK, Shipper ID=599818. 407 ==> _Fraud class=1. 85 cases, confidence:(0.21)

Using this transformation, rule number 15 looks for instance as follows: If the origin country of the shipment is Hong Kong and bought via e-commerce-ID=599818 with a price value is between (117.657, 134.702] and the weight value between (1.640 , 2.450] then 157 transactions are indicated as customs fraud cases

Rule number 27 can be explained as follows: If the origin country of the shipment is Hong Kong and bought via e-commerce-ID=599818 and the weight value being between (1.120 , 1.640] and the price value being between (39.579 , 98.397] then 93 transactions are indicated as a customs fraud case

When we look at all the decision rules mentioned in both models, we can notice that most of the cases of customs fraud have some common characteristics. The origin of the shipments is Hong Kong, the customers were mostly customers of the same e-commerce-ID company (ID=599818) and the cases involved different product categories, but mostly ‘Apparel’ of different values. A practical conclusion for the forwarding logistics company might consequently be to look more carefully at who is operating the e-commerce site and maybe train the people there to understand the customs rules better and to apply them more carefully. Moreover improvement can be monitored.

5.2 The results of customs fraud decision tree model (model-3)

The goal of this model (**model-3**) is to try to detect customs fraud using CHAID decision tree analysis and verify whether the results of the two previous models using association rules actually are similar to this model and in how far the results support each other.

The dependent variable in this model is Fraud class. The independent variables are (Weight In KG, Product Group Name, COD Value USD, Origin Country, Shipper ID). Table 9 presents the summary of Model-3.

The total number of transactions present in each node is based on Fraud class (category ‘Yes-indication of Fraud’ or category ‘No-Not Fraud). The total number of nodes in the model is 24, while the number of terminal nodes is 17. Figure 1 shows the tree map of the Fraud model created.

We can conclude for instance that if the shipper ID is '599148' and weight in Kg <=0.230, there are 14 cases indicated as a fraud (node 5 explained). If the shipper ID is '599818' and the values of 'COD Value USD' are between (117.6-134.7) and the product group name are 'Beauty supplies' or 'Apparel', then there are 288 cases indicated as a fraud (node 19 explained).

We can draw some conclusions about the similarity between the rules conducted from association rules models and the decision tree model.

The rules of model-3 are in general somewhat similar to the rules conducted by the Apriori analysis. Some examples of this similarity are the following. Most of the cases were for instance attributed to Shipper ID= 599818 and occurred with the product groups 'Apparel', 'Shoes' and 'Bag/Case'. Moreover, 'weight in kg' and 'COD' variables show some similar values in the nodes of the CHAID model and the association rules. For instance, we can see the occurrence of the values 0.231, 0.49, 1.12 for the 'weight in kg' as variables in model-2 and model-3, and the values 98.39 and 134.7 for the 'COD' variable in both models as well. The transformation in model-1 was performed using k-means as shown in table One, the splitting of 'weigh in kg' and 'COD value USD' variables are different (in total 3 classes), but at the same time we can also see values of the 'COD' variable close to one another: the average of the first class in model-1 was 72.1 and the average of the second class was 131.9, values relatively close to the values found by model-3 (72.2 and 134.7). The rules performed by Apriori models (one and two) are more comprehensive than the results of the CHAID decision tree model. The decision tree model (model-3) splits the independent variables using the variable that has the most significant effect on the dependent variable one by one and sequentially.

Therefore, the results of Model-3 confirm our results in model-1 and model-2, but the rules extracted of the Apriori models are more comprehensive and are therefore capable of detecting fraud in deeper detail.

Table 9. Fraud class decision tree model summary.

Model Summary		
Specifications	**Growing Method**	CHAID
	Dependent Variable	Fraud class
	Independent Variables	Weight In KG, COD Value USD, Product Group Name, Origin Country, Destination, Shipper ID
	Validation	Cross Validation, n=25
Results	**Independent Variables Included**	Shipper ID, COD Value USD, Weight In KG, Product Group Name, Destination
	Number of Nodes	24
	Number of Terminal Nodes	17

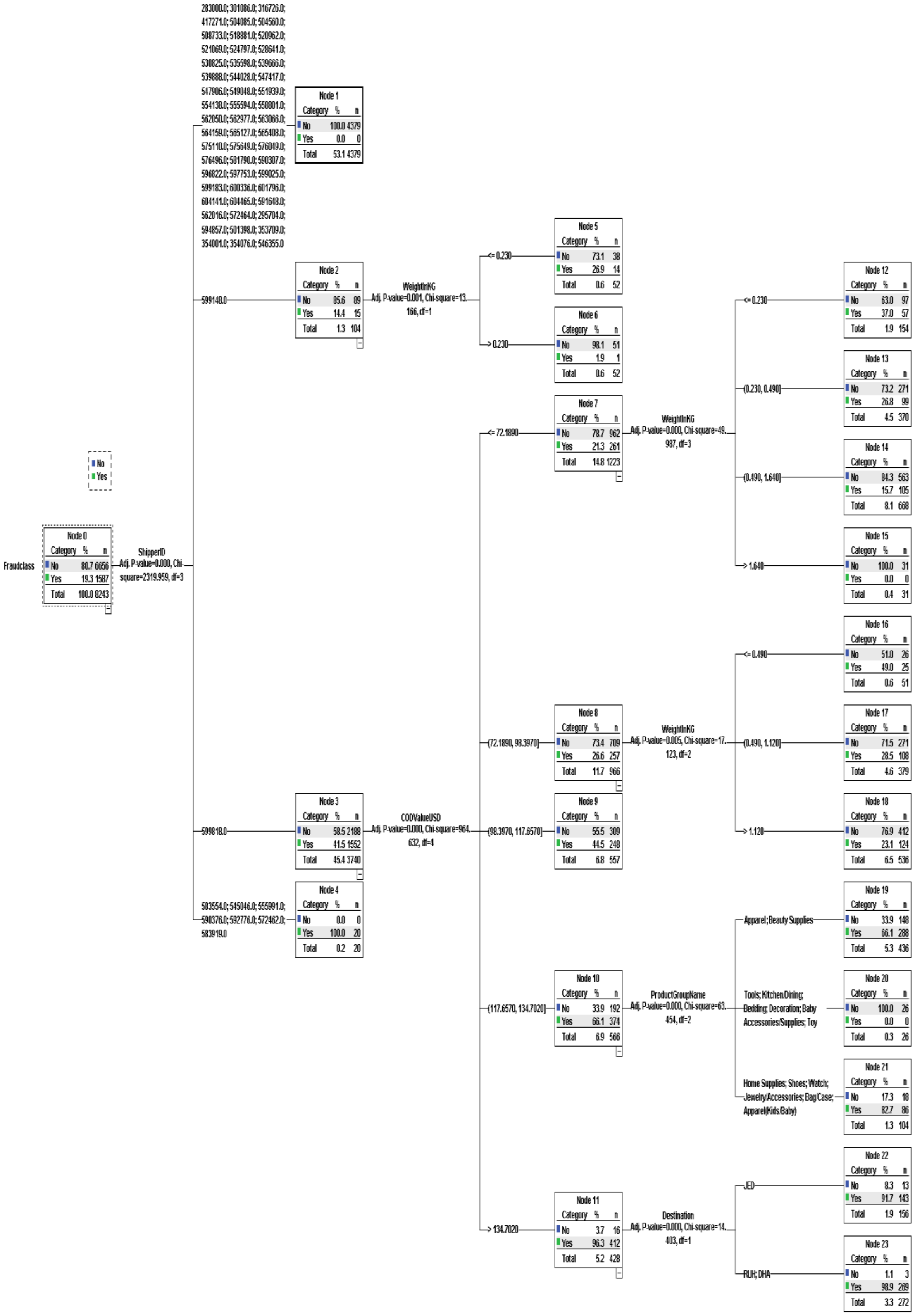

Figure 1. CHAID decision tree for Model 3

- **Evaluation the Customs fraud CHAID model using classification measures tests**

To evaluate the quality of the CHAID customs Fraud model, the classification measures (accuracy, recall, precision and f-score) have been used on the test data based on the confusion matrix.

Figure 2 shows the classification results on the test data of the CHAID customs Fraud model.

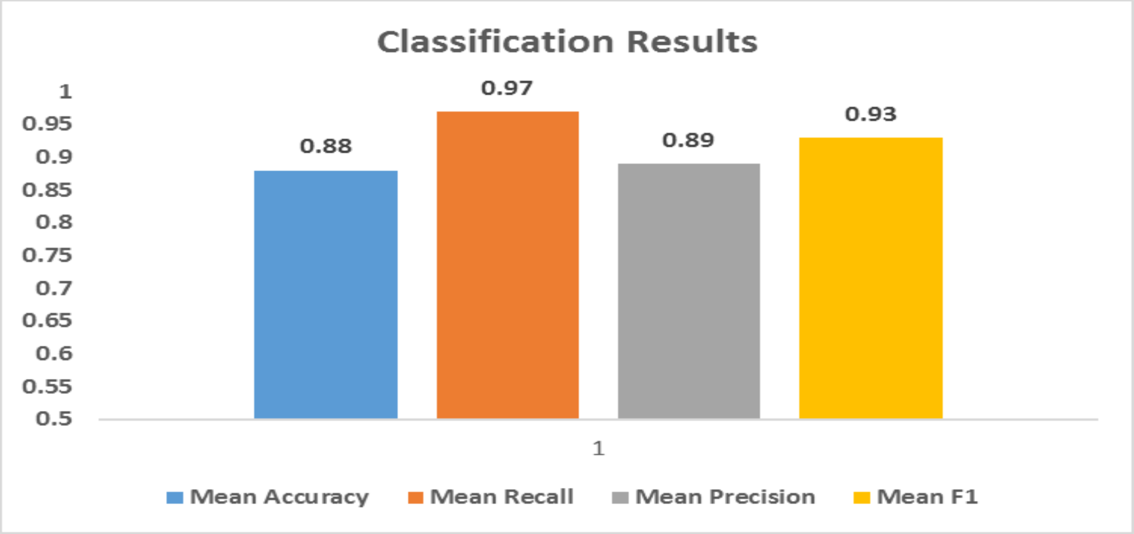

Figure 2: Classification results of the CHAID customs fraud model.

We can conclude that our CHAID model has a high average percentage of accuracy with 88%. The average of the recall, precision and f-score also have a high percentage with 97%, 89% and 93% respectively.

5.3 Discussion of the results of the custom fraud detection models

We have used the Apriori association rules and decision tree analysis on a partial section of our data base to answer the research question whether and how we can find indications of customs fraud using the data base only. After transformation of some of the continuous data using both k-means clustering and CHAID decision tree analysis, we have indeed succeeded in finding rules that allow to check whether there are cases of customs fraud in the data base. Some of these rules have been explained in detail.

This is highly significant for the users of the data base. Detecting customs fraud is indeed an important element in getting online business shipment systems within the realm of completely honest and trustworthy business on a global scale. This is important to all parties involved: forwarders, customers and governments. Forwarders like the company providing our database can more reliably offer fraud-proof systems to their customers, customers can faithfully rely on the forwarders that they will not be confronted with government claims afterwards and given the actual rules of customs application, and governments can trust forwarders more in not avoiding customs as fraud cases are detected before.

Our detection method is capable of doing so. To the best of our knowledge, it proves to be the first research method that detects fraud and proves to perform well. Moreover, the decision tree analysis proved to verify the connection between both Apriori models used and showed similar results, thus corroborating our research conclusions. Consequently this research constitutes an important contribution to both the scientific literature on data mining as a technique to find indication of customs fraud and to the practice of improving e-commerce behaviour.

6. Conclusion

In this section the major results of our different analyses are related to the research question will be presented.

Our research question read as follows: **‘How can customs fraud be detected using data mining on logistics transaction data?’**

The objective was to find evidence of potential customs fraud on different logistics trajectories based on the variables in the data base such as product categories and logistics routes.

This objective was achieved using both the Apriori algorithm for association rules and decision tree analysis. First we had to transform the continuous variables using both k-means clustering and CHAID decision tree analysis. This was done for the weight and price variables. Analysis of the rules detected in our analysis indicates that it is possible via this methodology to find indicators of customs fraud cases. Moreover, it was possible to describe in detail the situation in which the fraud occurs, indicating e-commerce customers (ID), types of goods (mostly apparel in our data base) and country of origin (mostly Hong Kong in our database).

Moreover, the results of the decision tree analysis proved to verify the connection between both Apriori models used and showed similar results, which confirm the value of our research conclusions.

This should allow logistics companies to offer a customs fraud service to their customers-partners, causing these customers and also government instances less hassle. As it is to our knowledge the first data mining technique capable of finding indications of customs fraud accurately, our research has a significant contribution to both data mining research and the logistics world.

Bibliography

ACFE (2018). REPORT TO THE NATIONS, 2018. GLOBAL STUDY ON OCCUPATIONAL FRAUD AND ABUSE.

De'ath, G. (2000). "Classification and regression trees: a powerful yet simple technique for ecological data analysis." Ecology v. 81(no. 11): pp. 3178-3192-2000 v.3181 no.3111.

Díaz-Pérez, F. M. and M. Bethencourt-Cejas (2016). "CHAID algorithm as an appropriate analytical method for tourism market segmentation." Journal of Destination Marketing & Management 5(3): 275-282.

GCCStates. (2018). "Common Customs Law for GCC States ", from https://www.fca.gov.ae/en/homerightmenu/ pages/gccstates.aspx?SelectedTab=14.

Haddadi, H. (2010). "Fighting online click-fraud using bluff ads." SIGCOMM Comput. Commun. Rev. 40(2): 21-25.

John, M. and H. Shaiba (2019). "Apriori-Based Algorithm for Dubai Road Accident Analysis." Procedia Computer Science 163: 218-227.

Kotu, V. and B. Deshpande (2014). Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner, Morgan Kaufmann Publishers Inc.

L., S. and B. Jr (2012). AN EMPIRICAL COMPARISON OF LOGISTIC REGRESSION TO DECISION TREE INDUCTION IN THE PREDICTION OF INTIMATE PARTNER VIOLENCE REASSAULT. Doctor of Philosophy, Indiana University of Pennsylvania.

Linoff, G. S. and M. J. A. Berry (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Wiley.

Loh, W.-Y. and Y.-S. Shih (1997). "SPLIT SELECTION METHODS FOR CLASSIFICATION TREES." Statistica Sinica 7(4): 815-840.

Long, W. J., J. L. Griffith, H. P. Selker and R. B. D'Agostino (1993). "A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain." Computers and Biomedical Research 26(1): 74-97.

Metwally, A., D. Agrawal, A. E. Abbad and Q. Zheng (2007). On Hit Inflation Techniques and Detection in Streams of Web Advertising Networks. 27th International Conference on Distributed Computing Systems (ICDCS '07).

Mittal, S., R. Gupta, M. Mohania, S. K. Gupta, M. Iwaihara and T. Dillon (2006). Detecting Frauds in Online Advertising Systems, Berlin, Heidelberg, Springer Berlin Heidelberg.

Prithiviraj, P. and P. Dr.R (2015). "A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study."

Qabbaah, H., G. Sammour and K. Vanhoof. (2019). Using K-Means Clustering and Data Visualization for Monetizing logistics Data. 2nd International Conference on new Trends in Computing Sciences (ICTCS), IEEE.

Qabbaah, H., L. Sharawi, G. Sammour and K. Vanhoof (2017). Development of Customs Prediction Model for Online Ordering. 2017 International Conference on New Trends in Computing Sciences (ICTCS), IEEE.

Qabbaah., H., G. Sammour. and K. Vanhoof. (2019). "DECISION TREE ANALYSIS TO IMPROVE E-MAIL MARKETING CAMPAIGNS." International Journal “Information Theories and Applications” 26(1): 303-330.

Ripley, B. D. and N. L. Hjort (1995). Pattern Recognition and Neural Networks, Cambridge University Press.

Silva, J., N. Varela, L. A. Borrero López and R. H. Rojas Millán (2019). "Association Rules Extraction for Customer Segmentation in the SMEs Sector Using the Apriori Algorithm." Procedia Computer Science 151: 1207-1212.

Tan, P. N., M. Steinbach and V. Kumar (2005). Association analysis: Basic concepts and algorithms.

Triepels, R., H. Daniels and A. Feelders (2018). "Data-driven fraud detection in international shipping." Expert Systems with Applications 99: 193-202.

Tripathi, D., B. Nigam and D. R. Edla (2017). "A Novel Web Fraud Detection Technique using Association Rule Mining." Procedia Computer Science 115: 274-281.

Zhang, L. and Y. Guan (2008). Detecting Click Fraud in Pay-Per-Click Streams of Online Advertising Networks. 2008 The 28th International Conference on Distributed Computing Systems.

Authors' Information

Hamzah Qabbaah *- PhD student at Research group of business informatics, Faculty of Business Economics, Hasselt University, B6b, Campus Diepenbeek, B-3590 Diepenbeek, Limburg,Belgium*

e-mail: hamzah.qabbaah@uhasselt.be

Major Fields of Scientific Research: Data mining, Digital marketing, Business Intelligence, Machine Learning, Knowledge Management

George Sammour *- Director of Quality Assurance and Accreditation Centre, Business Information Technology Department, Princess Sumaya University for Technology, Amman, Jordan.*

e-mail: George.Sammour@psut.edu.jo

Major Fields of Scientific Research: Data mining, Digital learning, lifelong learning, professional development

Koen Vanhoof *- Professor Dr., Head of the discipline group of Quantitative Methods, Faculty of Business Ecnomics, Universiteit Hasselt; Campus Diepenbeek; B-3590 Diepenbeek, Limburg,Belgium*

e-mail:koen.vanhoof@uhasselt.be

Major Fields of Scientific Research: data mining, knowledge retrieval