Customs Data Science Use Case

Harmonized System Codes predictions using Machine Learning techniques

Buensod Julia
Campion Sebastien
Cluchague Maxime
Guillemin Loic

January 18, 2021

Contents

List of Figures

List of Tables

1 Introduction

1.1 Harmonized System Codes

The "HS" code stands for "Harmonized System". This six-digit code was developed by the World Customs Organization (WCO) to classify internationally traded goods. The HS is important since it provides a common framework to classify international goods, and thus identify their tariff rate category. Providing this code is (in most cases) compulsory for people/organisations desiring to import/export a good internationally.

The HS code is organized in big sections by type of material and processing level. For example, one section is dedicated to vegetable products, while another to "prepared foodstuff [...]". Another couple of examples would be the section dedicated to "textile and textile articles" and the other on "footwear [...]". In these sections, we find different chapters - 99 in total. These chapters describe the broad product category (e.g. live animals, pharmaceutical products, glass and glassware, ...), and correspond to the first two digits of the HS code. These chapters - or categories - are then split into subcategories (1,244 classes), which correspond to the four first digits of the HS code. Finally, these subcategories are further divided into sub-subcategories (5,224 classes) represented by the entire HS code - the six digits.

In this study, the HS code is the label we aim to predict with the various classification methods deployed in this set of notebooks. For simplicity purposes, we refer to the first category level (the chapter) as "C1" (2 digits), the second category level (the subcategory) as "C2" (4 digits), and finally the most granular level as "C3" (6 digits).

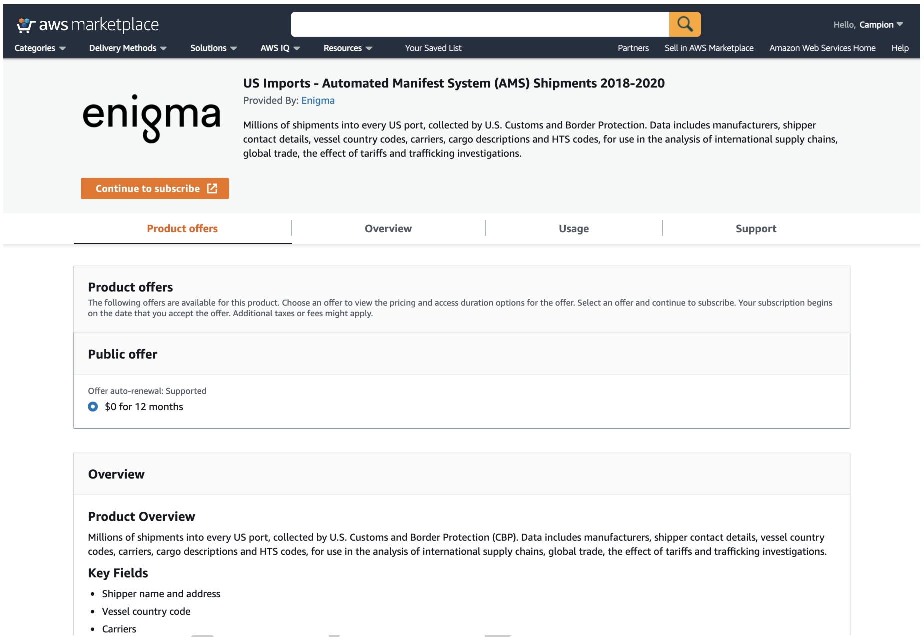
1.2 Problem statement

Millions of packages enter the European Union on a daily basis. These packages are classified based on their content into different categories represented by the HS codes. As mentioned in section 1.1, these codes determine the tariff applicable to the parcel. If the wrong HS code is assigned to a parcel, be it intentionally or not, this would result in a tariff issue at the customs. Custom teams have thus a limited amount of time to assess the risk of tariff non-compliance based on an enormous amount of items.

Our aim in this exercise is to develop models to classify the content of the parcels based on the available information (description of the content, information about the manufacturer and shipper, weight, etc.). We will evaluate several models arising from both Machine Learning and Deep Learning.

1.3 Source data

The data used for this exercise is the US shipment data. It is available in the AWS Data Marketplace: https://aws.amazon.com/marketplace/pp/prodview-stk4wn3mbhx24.



US Imports - Automated Manifest System (AMS) Shipments 2018-2020

Provided By: Enigma

Millions of shipments into every US port, collected by U.S. Customs and Border Protection. Data includes manufacturers, shipper contact details, vessel country codes, carriers, cargo descriptions and HTS codes, for use in the analysis of international supply chains, global trade, the effect of tariffs and trafficking investigations.

Product offers

The following offers are available for this product. Choose an offer to view the pricing and access duration options for the offer. Select an offer and continue to subscribe. Your subscription begins on the date that you accept the offer. Additional taxes or fees might apply.

Public offer

Offer auto-renewal: Supported

● $0 for 12 months

Overview

Product Overview

Millions of shipments into every US port, collected by U.S. Customs and Border Protection (CBP). Data includes manufacturers, shipper contact details, vessel country codes, carriers, cargo descriptions and HTS codes, for use in the analysis of international supply chains, global trade, the effect of tariffs and trafficking investigations.

Key Fields

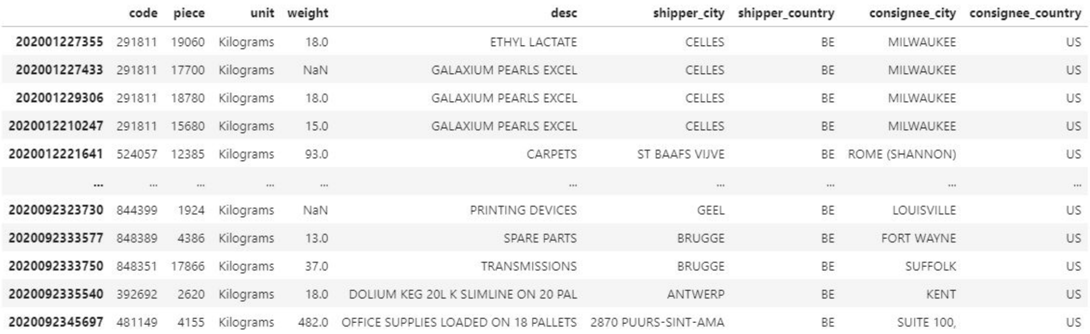
- Shipper name and address
- Vessel country code
- Carriers

Figure 1: Source data on the AWS marketplace

The data contains information such as:

- shipment ID
- cargo description
- shipment weight and weight unit
- number of pieces
- shipper contact details (incl. name, address, city and country)
- consignee contact details (incl. name, address, city and country)
- HS code (6-digit standard classifying globally traded products)

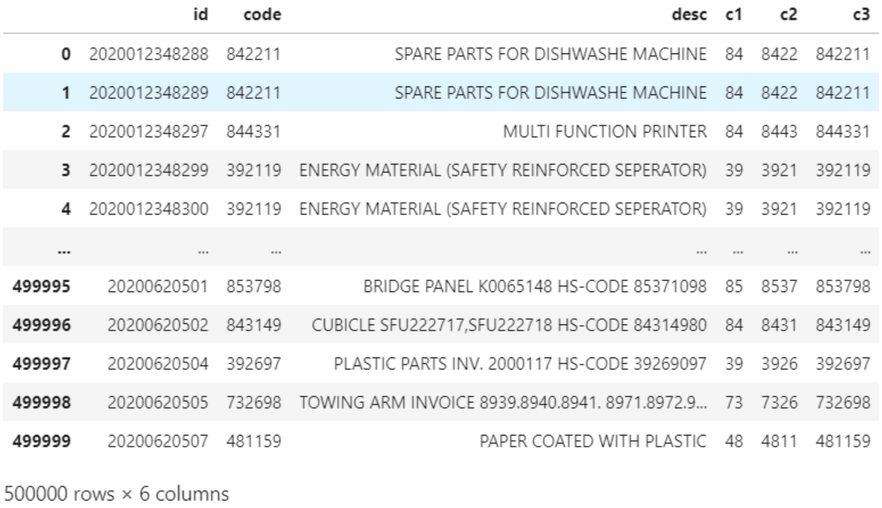
For our study, we work on a random sample of 166,286 rows due to time and data quality constraints. An extract of our dataset is illustrated in figure 2.



	code	piece	unit	weight	desc	shipper_city	shipper_country	consignee_city	consignee_country
202001227355	291811	19060	Kilograms	18.0	ETHYL LACTATE	CELLES	BE	MILWAUKEE	US
202001227433	291811	17700	Kilograms	NaN	GALAXIUM PEARLS EXCEL	CELLES	BE	MILWAUKEE	US
202001229306	291811	18780	Kilograms	18.0	GALAXIUM PEARLS EXCEL	CELLES	BE	MILWAUKEE	US
2020012210247	291811	15680	Kilograms	15.0	GALAXIUM PEARLS EXCEL	CELLES	BE	MILWAUKEE	US
2020012221641	524057	12385	Kilograms	93.0	CARPETS	ST BAAFS VIJVE	BE	ROME (SHANNON)	US
...	...	...	...	...	...	...	...	...	...
2020092323730	844399	1924	Kilograms	NaN	PRINTING DEVICES	GEEL	BE	LOUISVILLE	US
2020092333577	848389	4386	Kilograms	13.0	SPARE PARTS	BRUGGE	BE	FORT WAYNE	US
2020092333750	848351	17866	Kilograms	37.0	TRANSMISSIONS	BRUGGE	BE	SUFFOLK	US
2020092335540	392692	2620	Kilograms	18.0	DOLIUM KEG 20L K SLIMLINE ON 20 PAL	ANTWERP	BE	KENT	US
2020092345697	481149	4155	Kilograms	482.0	OFFICE SUPPLIES LOADED ON 18 PALLETS	2870 PUURS-SINT-AMA	BE	SUITE 100,	US

Figure 2: Extract of study data (complete dataset)

In a first step, we further reduce the scope of our data to 50,000 rows to assess different algorithms. In addition, we solely consider the description of the parcel, excluding the other fields. This new dataset is illustrated in figure 3. Based on this reduced dataset, we first aim to predict the first level of classification (C1), which only contains 82 classes. We focus on C1 because C2 and C3 are sub-classes of the former, and each contains more distinct classes than the upper classification level.



	id	code	desc	c1	c2	c3
0	2020012348288	842211	SPARE PARTS FOR DISHWASHE MACHINE	84	8422	842211
1	2020012348289	842211	SPARE PARTS FOR DISHWASHE MACHINE	84	8422	842211
2	2020012348297	844331	MULTI FUNCTION PRINTER	84	8443	844331
3	2020012348299	392119	ENERGY MATERIAL (SAFETY REINFORCED SEPERATOR)	39	3921	392119
4	2020012348300	392119	ENERGY MATERIAL (SAFETY REINFORCED SEPERATOR)	39	3921	392119
...	...	...	...	...	...	...
499995	20200620501	853798	BRIDGE PANEL K0065148 HS-CODE 85371098	85	8537	853798
499996	20200620502	843149	CUBICLE SFU222717,SFU222718 HS-CODE 84314980	84	8431	843149
499997	20200620504	392697	PLASTIC PARTS INV. 2000117 HS-CODE 39269097	39	3926	392697
499998	20200620505	732698	TOWING ARM INVOICE 8939.8940.8941. 8971.8972.9...	73	7326	732698
499999	20200620507	481159	PAPER COATED WITH PLASTIC	48	4811	481159

500000 rows × 6 columns

Figure 3: Extract of study data (sample data for exploratory phase)

In a second step, we base ourselves on this first model evaluation (based on C1) and select the best model. Next, we deploy the model on the entire data (our 166,286 rows) to obtain the final performance scores.

1.4 Data preparation

The HS code is available in a dedicated field. However, this field is empty for many items. Indeed, it is usual for HS codes to be provided in the description field instead of the dedicated HS code field. For these cases, we retrieved the code from the description using regular expressions (regex).

As mentioned in section 1.1, we divided the HS code in three parts, corresponding to the level of granularity of the classification: C1, C2, and C3. To facilitate our analysis, we have thus extracted these classes and added them in our dataset as three additional columns.

In addition, be it in our entire dataset or in the sample used for the evaluation models, we have only cleaned the description field upstream. Indeed, the parcel descriptions sometimes contained the HS code. Therefore, to prevent the training of the models from being biased, we removed any mention of C1, C2 and C3 from the description.

1.5 Data Exploration

To better understand our data, we start by exploring its distribution. Note that at this stage, we only look at C1. As can be seen in the illustrations below, the distribution of packages among the different categories is unbalanced. Indeed, the three most populated categories (84, 85, 73) account for 63% of the dataset (32%, 16% and 15% respectively) (see figures 4 and 5).

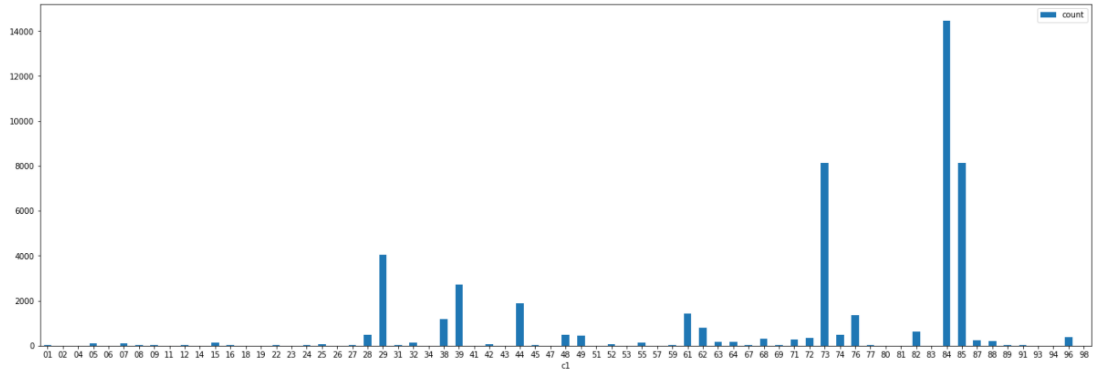


Figure 4: Number of parcels per category (C1)

These three top categories correspond to the HS chapters on ”Nuclear reactors, boilers, machinery and mechanical appliances, parts thereof” (84), ”Electrical machinery and equipment [...]; sound recorders and reproducers, television image and sound [...]” (84), and ”Articles of iron or steel” (73).

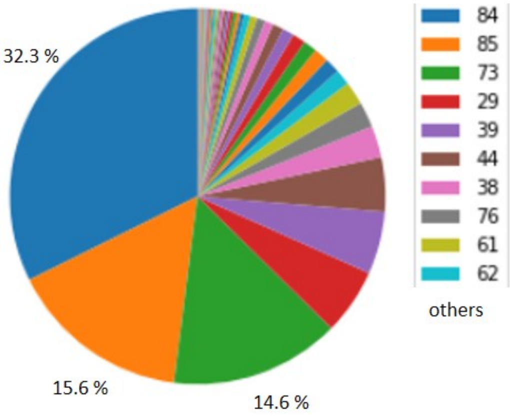
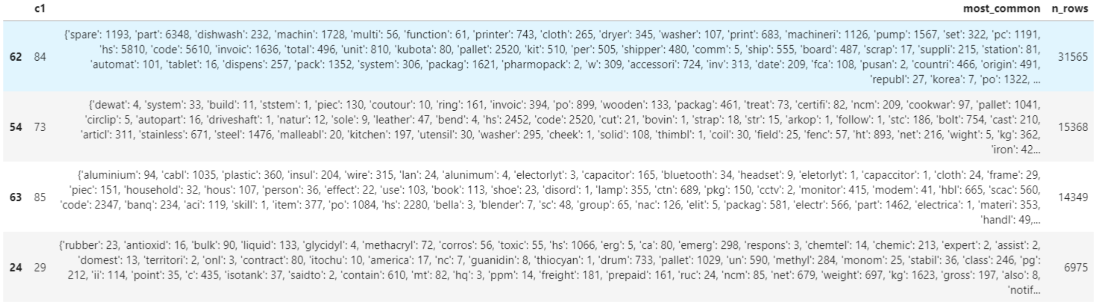


Figure 5: Distribution of parcels per category (C1)

We then wanted to see what were the most common words used in the parcel descriptions for each category (C1). To do so, we used the *nltk* library to clean the descriptions (i.e. make them lower-case, remove the punctuation and stop words, stem words) and tokenize them. As a result, we obtained a list of cleaned words for each item.

Next, we aggregated these lists by category (C1) and counted the number of occurrence of each word in the lists. The C1 categories have been sorted so that the most represented categories would appear first (see figure 6).



	c1	most_common	n_rows
62	84	{'spare': 1193, 'part': 6348, 'dishwash': 232, 'machin': 1728, 'multi': 56, 'function': 61, 'printer': 743, 'cloth': 265, 'dryer': 345, 'washer': 107, 'print': 683, 'machineri': 1126, 'pump': 1567, 'set': 322, 'pc': 1191, 'hs': 5810, 'code': 5610, 'invoic': 1636, 'total': 496, 'unit': 810, 'kubota': 80, 'pallet': 2520, 'kit': 510, 'per': 505, 'shipper': 480, 'comm': 5, 'ship': 555, 'board': 487, 'scrap': 17, 'suppli': 215, 'station': 81, 'automat': 101, 'tablet': 16, 'dispens': 257, 'pack': 1352, 'system': 306, 'packag': 1621, 'pharmopack': 2, 'w': 309, 'accessori': 724, 'inv': 313, 'date': 209, 'fca': 108, 'pusan': 2, 'countri': 466, 'origin': 491, 'republ': 27, 'korea': 7, 'po': 1322, ...	31565
54	73	{'dewat': 4, 'system': 33, 'build': 11, 'ststem': 1, 'piec': 130, 'coutour': 10, 'ring': 161, 'invoic': 394, 'po': 899, 'wooden': 133, 'packag': 461, 'treat': 73, 'certifi': 82, 'ncm': 209, 'cookwar': 97, 'pallet': 1041, 'circlip': 5, 'autopart': 16, 'driveshaft': 1, 'natur': 12, 'sole': 9, 'leather': 47, 'bend': 4, 'hs': 2452, 'code': 2520, 'cut': 21, 'bovin': 1, 'strap': 18, 'str': 15, 'arkop': 1, 'follow': 1, 'stc': 186, 'bolt': 754, 'cast': 210, 'articl': 311, 'stainless': 671, 'steel': 1476, 'malleabl': 20, 'kitchen': 197, 'utensil': 30, 'washer': 295, 'cheek': 1, 'solid': 108, 'thimbl': 1, 'coil': 30, 'field': 25, 'fenc': 57, 'ht': 893, 'net': 216, 'wight': 5, 'kg': 362, 'iron': 42...	15368
63	85	{'aluminium': 94, 'cabl': 1035, 'plastic': 360, 'insul': 204, 'wire': 315, 'lan': 24, 'alunimum': 4, 'electorlyt': 3, 'capacitor': 165, 'bluetooth': 34, 'headset': 9, 'eletorlyt': 1, 'capaccitor': 1, 'cloth': 24, 'frame': 29, 'piec': 151, 'household': 32, 'hous': 107, 'person': 36, 'effect': 22, 'use': 103, 'book': 113, 'shoe': 23, 'disord': 1, 'lamp': 355, 'ctn': 689, 'pkg': 150, 'cctv': 2, 'monitor': 415, 'modem': 41, 'hbl': 665, 'scac': 560, 'code': 2347, 'banq': 234, 'aci': 119, 'skill': 1, 'item': 377, 'po': 1084, 'hs': 2280, 'bella': 3, 'blender': 7, 'sc': 48, 'group': 65, 'nac': 126, 'elit': 5, 'packag': 581, 'electr': 566, 'part': 1462, 'electrica': 1, 'materi': 353, 'handl': 49,...	14349
24	29	{'rubber': 23, 'antioxid': 16, 'bulk': 90, 'liquid': 133, 'glycidyl': 4, 'methacryl': 72, 'corros': 56, 'toxic': 55, 'hs': 1066, 'erg': 5, 'ca': 80, 'emerg': 298, 'respons': 3, 'chemtel': 14, 'chemic': 213, 'expert': 2, 'assist': 2, 'domest': 13, 'territori': 2, 'onl': 3, 'contract': 80, 'itochu': 10, 'america': 17, 'nc': 7, 'guanidin': 8, 'thiocyan': 1, 'drum': 733, 'pallet': 1029, 'un': 590, 'methyl': 284, 'monom': 25, 'stabil': 36, 'class': 246, 'pg': 212, 'ii': 114, 'point': 35, 'c': 435, 'isotank': 37, 'saidto': 2, 'contain': 610, 'mt': 82, 'hq': 3, 'ppm': 14, 'freight': 181, 'prepaid': 161, 'ruc': 24, 'ncm': 85, 'net': 679, 'weight': 697, 'kg': 1623, 'gross': 197, 'also': 8, 'notif...	6975

Figure 6: Most common words per category (C1)

1.6 Technical challenges

As mentioned in the previous sections, our dataset is relatively small (166,286 rows) and unbalanced, which means that many classes - especially at the C3 level - are under-represented in our whole dataset. This is an issue for generalization during the algorithms training phase.

In addition, some descriptions are not explicit or quite vague, such as:

Unclear descriptions
”DISP C-TSTHS 491199”
”W-118472”
”01 EX 06X0BZ KIT”
”DECKING PO3028689”

Table 1: Example of unclear descriptions

Parcel descriptions can also be written in other languages, for instance:

Description in another language
”HOJAS DE NARANJA ORANGE LEAF TEA”
”HOLZFASERMATTEN”

Table 2: Example description in another language

These two problems will affect the performance of the model in terms of classification.

Finally, since our descriptions are free text, it adds another layer of complexity. Indeed, models are not able to manage free text so we need to encode the free text fields (vectorization) so that we can later develop models. The approach we use for text encoding is described in section 2.1.1.1. In the Deep Learning approach (see section 2.2), these techniques are embedded in the models (Word 2 Vect and CNN).

2 Initial assessment of models

As mentioned in our introduction in section 1.3, we begin our study with a sample of 50,000, only considering the description field. In this first exploratory step, we test several algorithms and evaluate their ability to predict C1 using the parcel description only.

The models we evaluate follow two approaches:

1. Machine Learning (ML)
2. Deep Learning (DL)

For the ML approach, we use a random forest, a logistic regression, and a singular value decomposition (SVD). For the DL approach, we apply a Convolutional Neural Network (CNN), once applying it at word level, a second time at character level. In addition, for this first assessment of different models, we solely focus on the parcel description field.

With all our models, we predict C1, C2 and C3 separately. In our experimental phase, we have also tested an approach whereby we extracted our predicted C1 and C2 based on our prediction for C3. However, our models' predictions where still better by predicting the different levels of classification separately.

> **Note:** The Jupyter notebooks containing the models can be downloaded in .ipynb or .pdf format from the SSA Platform Pilot's online documentation. Note that the access to the platform is restricted and credentials are needed.

2.1 Machine Learning approach

2.1.1 Data Preparation

Before we can start training our model, we must perform several data preparation steps on our data:

- make all text lower case
- tokenize: splitting the shipment description into words
- remove all punctuations
- remove stop words (e.g. 'of', 'and', 'the', ...)
- stem words: reducing words to their root form. For example, 'shipment', 'shipping' and 'shipped' all become 'ship'.

2.1.1.1 Vectorization with TF-IDF

After having done our preparation steps, we obtain a list of words. We still need to vectorize the text so that it can be used by our models. One of the techniques available for the vectorization is TF-IDF (Term Frequency — Inverse Document Frequency).

The TF-IDF technique allows to quantify the importance of a word in a document and corpus. It is often used in the context of information retrieval or text mining, and one of its popular applications is for text classification. The basic principle behind TF-IDF is that its value will increase proportionally to the number of times a word appears in a document, but will decrease in function of the number of documents in the corpus containing that word. This second part (IDF) allows to take into account that some words are more frequent in general.

2.1.1.2 Dimension reduction

At the end of the previous step (i.e. vectorization with TF-IDF) we obtain about 13,000 columns - keeping in mind that this was applied to a sample of data. Indeed, applying it on 100,000 rows, we obtain 20,000 columns which corresponds to a matrix of 15GB. Therefore, we use our sample of 50,000 rows to avoid such heavy matrices.

In addition, we use some methods to reduce the number of dimensions while keeping all useful information. One of the benefits for dimension reduction is to decrease the number of relationships between variables, and thus decrease the likelihood to overfit our models. Reducing dimensionality also allows to make processes less computationally intensive. We tried three different methods of dimension reduction: PCA, t-SNE, and TSVD.

Principal Components Analysis (PCA) is an unsupervised (linear) dimension reduction technique for high dimensional data. It works by reducing the dimensionality of highly correlated data by transforming the origin set of vectors into a new set: the principal component.
PCA aims to preserve the data's global structure by mapping clusters as a whole, which means that local structures might not be ignored. Therefore, PCA is more likely to be affected by outliers than the other two techniques we present here.

t-distributed stochastic neighbourhood embedding (t-SNE) is also an unsupervised (albeit non-linear) dimension reduction technique. This technique works by embedding the data points from a higher dimension to a lower one while keeping the neighborhood of that point. This way, local structures can be preserved.
By focusing on local structures, t-SNE is a flexible approach and often succeeds to find structures when other dimension reduction techniques cannot.

Truncated Singular Value Decomposition (TSVD) is the third unsupervised (linear) dimension reduction technique we use in this section. This technique does not center data before computing the SVD, which means it works efficiently with sparse matrices.

Our experiments indicated that PCA was the most appropriate dimension reduction technique in our context. Therefore, we applied it to our sample data.

2.1.2 Model training and evaluation

Following the dimension reduction step in section 2.1.1.2, we obtain a new dataset with 100 columns. The following models are applied on this new reduced dataset.

We split our data into a train (2/3 of the data) and a test (1/3 of the data) set. The train set is used to train the models, which are then evaluated on the test data. We build our models using the scikit-learn library.

Model	Package	Accuracy on C1
Random Forest	ScikitLearn - RandomForestClassifier	0.55
Logistic regression	ScikitLearn - LogisticRegression	0.45
Singular Value Decomposition (SVD)	ScikitLearn - OneVsRestClassifier	0.51

Table 3: Performance results of the models (ML approach)

2.2 Deep Learning approach

2.2.1 Model training and evaluation

We split our data into a train (2/3 of the data) and a test (1/3 of the data) set. The train set is used to train the models, which are then evaluated on the test data.

For our DL approach, we focus on a one dimensional Convolutional Neural Network (CNN). We build our models using keras.

There are two different ways for a CNN to consume ”raw” data:

- Encode the description field at **word level** - each word is associated with a unique integer. For this, we use the keras tokenizer and set a limit to a corpus of 20,000 words. Note that the word encoding doesn't support syntax errors.
- Encode the description field at the **character level** (ordinal encoding[1]) - each character is associated with a unique integer

An advantage of both models is that they do not require data preprocessing, except for text encoding.

Model	Package	accuracy on C1
Convolutional Neural Network (CNN) on word	Keras/Tensorflow	0.65
Convolutional Neural Network (CNN) on char	Keras/Tensorflow	0.61

Table 4: Performance results of the models (DL approach)

[1] With ordinal encoding, each value is assigned an integer. Note that with this method, the algorithm could introduce a bias, since the method introduces an arbitrary ordinal relation between the data. However, an advantage of ordinal encoding is that only one additional column is created. In contrast, one-hot encoding creates a large number of new features (one feature per possible value), which can result in the slowdown of the experiment. Considering all of the above, we opted for ordinal encoding given the available time and resources.

2.3 Preliminary results

At the end of this section 2, we are able to compare the performance of our different models. Table 5 summarizes the results obtained, their advantages and disadvantages, as well as the data preparation steps to be applied prior the training of the models.

Based on the results in Table 5, we found that the best performing model in our case is the **CNN (at word level) with an accuracy of 65%** for the first level of classification (C1) using a sample of our data (50,000 rows) and the parcel descriptions only. Therefore, we selected this model for our second step of the study: the application of the model on the complete data[2].

3 Improvements on CNN model

Based on the comparison of our models (see section 2.3), we only keep the most performing one: the CNN on words. Focusing on this model, we add additional data fields to improve the performance of our model.

3.1 Model improvement

For this part we increased the size of the samples used for the training phase. This allows a better generalization of the part of the classification algorithm. We therefore took a sample of size 166,286 rows. This data set will then be split into a set, a training dataset (2/3 the size of the dataset, i.e. 110,857 rows) as well as a test dataset (1/3 the size of the dataset, i.e. 55,429 rows).

3.2 Adding more features

In addition to the parcel description, we now consider:

- Number of pieces (Int)
- Total shipment weight value (Float)
- Total shipment weight unit (Category)
- Country (for consignee & consignor) (Category)
- City (for consignee & consignor) (Category)

We used one-hot encoding for the consignor and consignee countries, as well as the weight units. For the consignor and consignee cities, we used cardinal encoding.

[2] We acknowledge that additional algorithms might have been considered for this exercise, and that the CNN on words might not be the best model when testing on our whole dataset and with all the features (not only the description). However, due to time constraints we had to do a preselection of algorithms, which seemed appropriate for our classification task, as well as test them on a reduced sample.

Model	Preparation steps	Advantages	Disadvantages	Accuracy (predicting C1)
Machine Learning:				
Random Forest	tokenization, punctuation and stop words removal, stemming, TF-IDF, PCA (dimension reduction)	good for classification problems, no data normalization needed works well with both categorical and continuous values	storing issues due to TF-IDF (matrices of more than 50GB), difficulty to interpret results and to determine variables importance due to ensemble of decision trees	0.55
Logistic Regression	tokenization, punctuation and stop words removal, and stemming, TF-IDF, PCA (dimension reduction)	easy to implement, train and coefficients can be interpreted as feature importance indicators	assumes linearity between the attributes and the variables to be predicted, not appropriate for non-linear problems	0.45
SVD	tokenization, punctuation and stop words removing, and stemming, TF-IDF, PCA (dimension reduction)		not appropriate for non-linear problems	0.51
Deep Learning:				
CNN at word level	word encoding	no data preparation needed - except word encoding, text vectorization	does not support syntax errors, difficult to interpret because it is a black box	0.65
CNN at character level	character encoding	no data preparation needed - except character encoding, text vectorization	difficult to interpret because it is a black box	0.62

Table 5: Comparison of models performance on sample data

4 Final results

Following the two main improvements described in section 3, we thus applied our model on the whole dataset, including additional attributes, which have been encoded beforehand. Keeping a train / test ratio of 2/3 and 1/3 respectively, we obtained a train set of 110,857 rows and a test set of 55,429 rows.

4.1 Model performance

This section is dedicated to describing the final model performance. You will find in the appendix (figure 7) is the architecture of the CNN model used for C1.

Applying our evaluation on our test data, we obtained the following performances for the different class levels (i.e. C1, C2 and C3).

> **Note:** You can download the performance details for each class level (C1, C2 and C3) from the SSA Platform Pilot online documentation.

4.1.1 Final evaluation on C1

As a reminder, C1 corresponds to the two first digits of the HS-code. It is composed of 82 distinct classes.

	precision	recall	f1-score	support
weighted avg	83.34%	83.56%	83.11%	55429
macro avg	40.54%	34.40%	35.91%	55429

Table 6: Final performance results of CNN model for C1

Accuracy score: 83.56%

Figure 8 (see appendix) is the confusion matrix associated with the C1 label.

4.1.2 Final evaluation on C2

As a reminder, C2 corresponds to the four first digits of the HS-code. It is composed of 575 distinct classes.

Accuracy score: 70.18%

	precision	recall	f1-score	support
weighted avg	71.48%	70.18%	69.59%	55429
macro avg	29.48%	24.48%	25.70%	55429

Table 7: Final performance results of CNN model for C2

4.1.3 Final evaluation on C3

As a reminder, C3 corresponds to the full HS-code, i.e. six digits. It is composed of 4,799 distinct classes.

	precision	recall	f1-score	support
weighted avg	54.13%	54.46%	51.81%	55429
macro avg	18.53%	15.58%	15.47%	55429

Table 8: Final performance results of CNN model for C3

Accuracy score: 54.46%

4.2 Manual result investigation

After having evaluated our final model, we dive into the data to investigate, the cases where the category predicted by our model is not in line with the one indicated in our dataset. For simplicity, we refer to the former one as the "predicted" label, and the latter as the "declared" label.

We found that in many cases, our model disagreement with the declared label was not an error from the model but an indication that the parcel had been mislabelled in the original dataset, or that the description provided was very vague. Figure 9 (see appendix) illustrates cases where the parcel was incorrectly labelled in our dataset, but correctly (or partially correctly) classified by our model. This is interesting from a business application view, since these model errors could signal parcels, which are worth being investigated by the custom teams since it is probable that the parcel is actually mislabelled. This would be particularly relevant for categories which have significantly different associated tariffs.

Of course, the category predicted by our model can sometimes be wrong when the declared label is correct. We found this to be the case for categories with relatively little data. Therefore, with a bigger dataset, we expect our model to learn better and its ability to correctly predict classes to improve.

5 Conclusion

In this study, we have assessed several models and their ability to classify parcels using only descriptions in a first step. Based on this first evaluation, we found that the best model for our classification problem is a Convolutional Neural Network on words. Improving this model (by adding more features, and applying to the entire dataset) we obtained an accuracy of 83.42% on C1, 70.18% on C2 and 54.46% on C3.

This parcel classification study is a first draft. It is important to note that the dataset we were able to obtain where rows would at least contain a HS-code and parcel description is relatively small. More data would allow us to better train our models and perhaps obtain better performances on the class prediction.

6 Perspectives

6.1 Possible model improvements

The first track of possible developments concerns the improvement of the model. To get better predictions, several steps could be carried out. For example:

- Training our model on a bigger dataset
- Adding more fields to our dataset (e.g. name of consignee / consignor, consignee / consignor (geocoded) addresses, ...)
- Translate non-English descriptions to English
- Improving the model architecture
- Trying predicting the classification levels successively, meaning that C2 will be predicted based on the results for C1. Similarly, C3 should be predicted taking into account the predictions for C1 and C2.

6.2 Development of the model's business application

The second major axis of developments is about enhancing the business application of our classification model. As mentioned in section 4.2, this model can help custom teams identify packages likely to be mislabelled, and provide them with a correct model prediction when possible.

Based on this first study, some developments could further facilitate the task of custom teams. For instance:

- Indicating a confidence level when providing a category prediction could allow custom teams to better focus their attention on predictions with a low confidence level

- Extracting and adding the tariffs as an additional field could allow to identify parcels for which the declared and predicted class (HS-code) have big tariff differences and raise an alert for these cases in priority

A Appendix

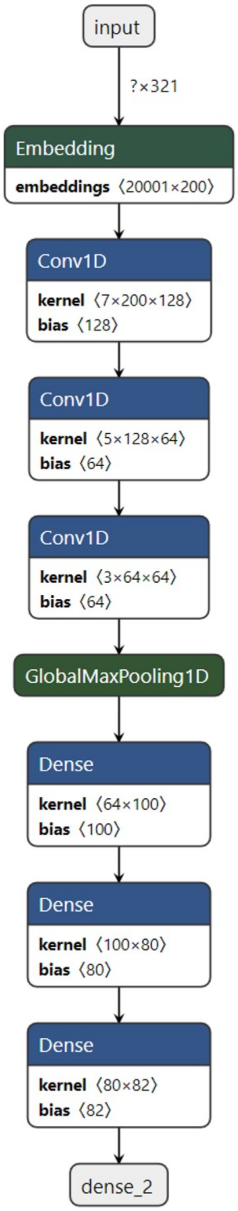


Figure 7: CNN architecture

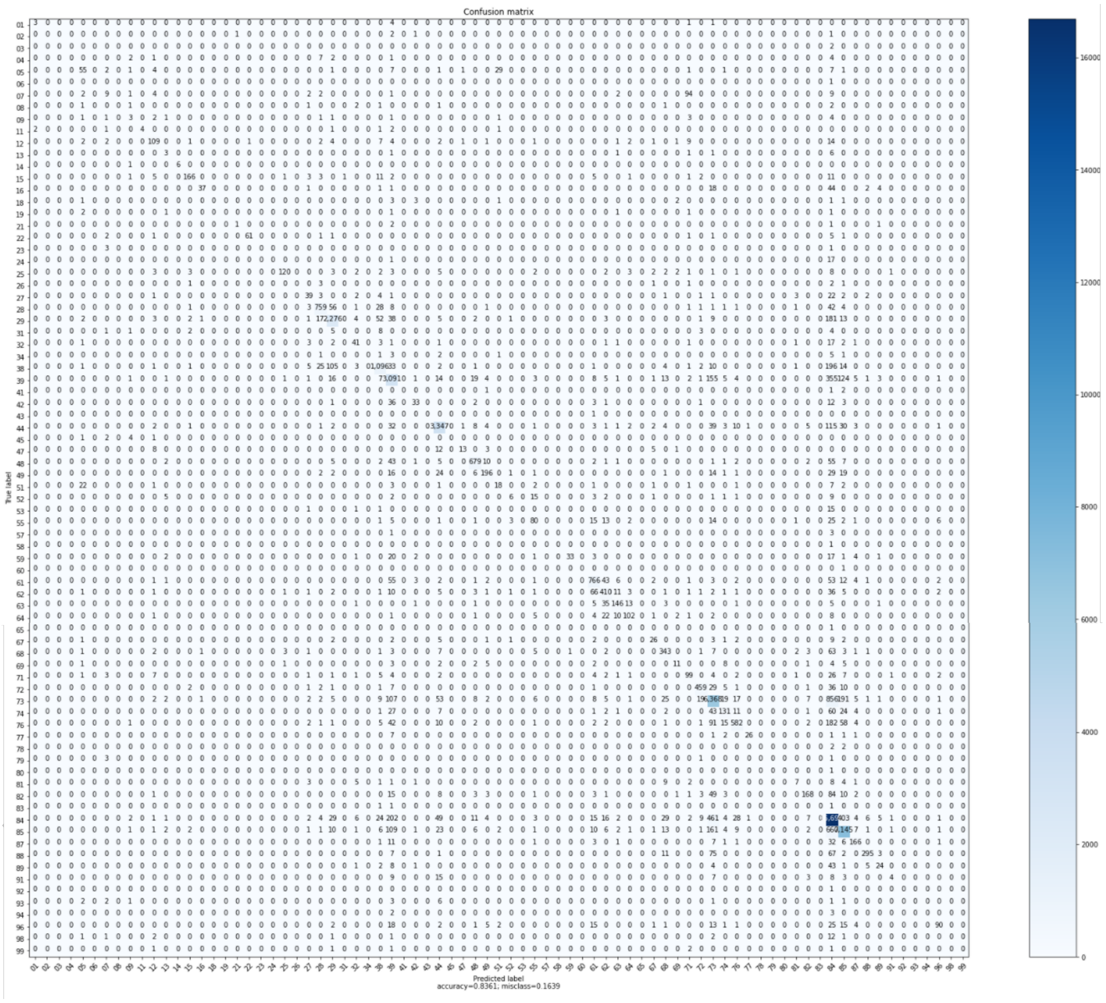
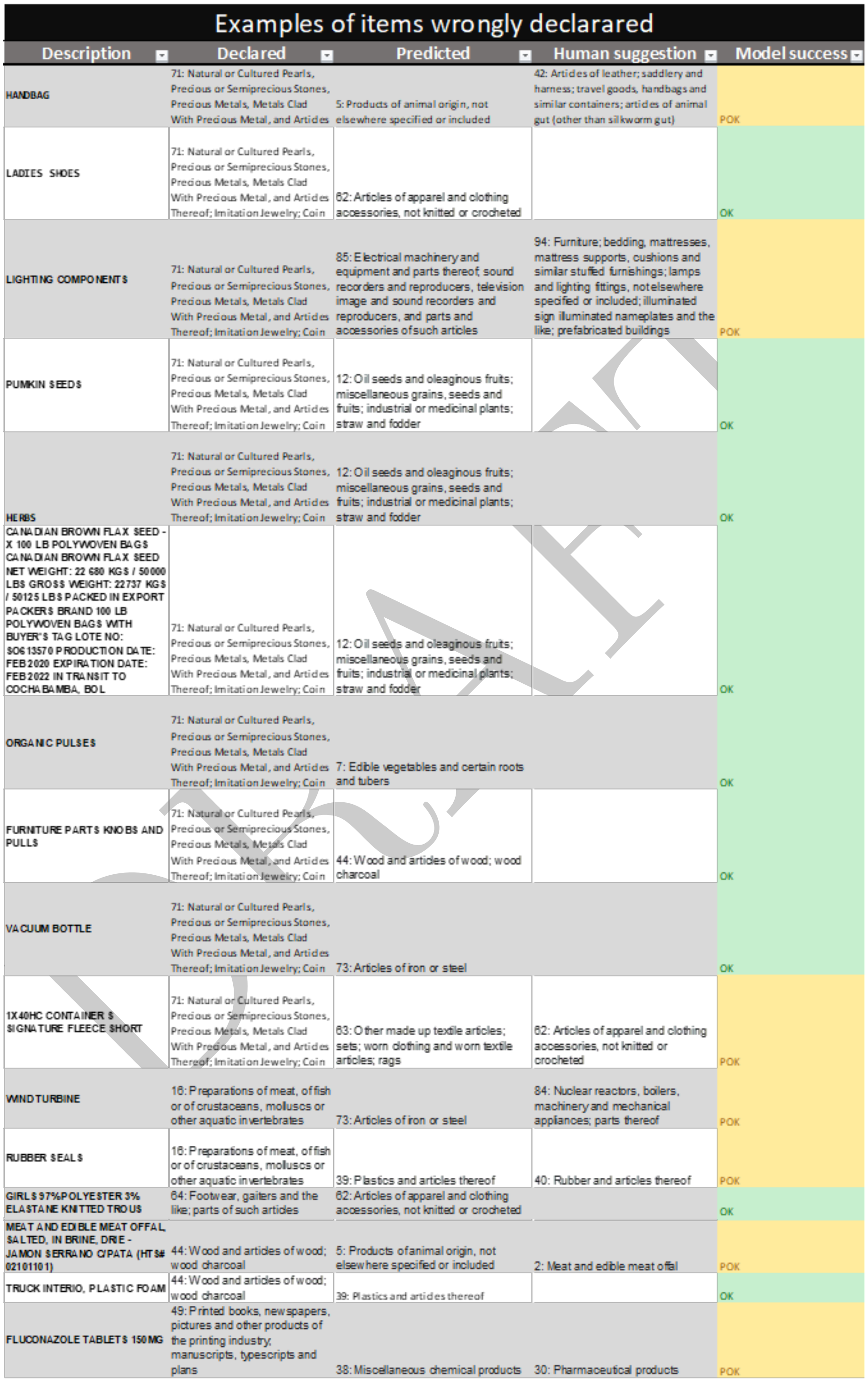


Figure 8: Confusion matrix for C1



Examples of items wrongly declarared				
Description	Declared	Predicted	Human suggestion	Model success
HANDBAG	71: Natural or Cultured Pearls, Precious or Semiprecious Stones, Precious Metals, Metals Clad With Precious Metal, and Articles	5: Products of animal origin, not elsewhere specified or included	42: Articles of leather; saddlery and harness; travel goods, handbags and similar containers; articles of animal gut (other than silkworm gut)	POK
LADIES SHOES	71: Natural or Cultured Pearls, Precious or Semiprecious Stones, Precious Metals, Metals Clad With Precious Metal, and Articles Thereof; Imitation Jewelry; Coin	62: Articles of apparel and clothing accessories, not knitted or crocheted		OK
LIGHTING COMPONENTS	71: Natural or Cultured Pearls, Precious or Semiprecious Stones, Precious Metals, Metals Clad With Precious Metal, and Articles Thereof; Imitation Jewelry; Coin	85: Electrical machinery and equipment and parts thereof; sound recorders and reproducers, television image and sound recorders and reproducers, and parts and accessories of such articles	94: Furniture; bedding, mattresses, mattress supports, cushions and similar stuffed furnishings; lamps and lighting fittings, not elsewhere specified or included; illuminated sign illuminated nameplates and the like; prefabricated buildings	POK
PUMKIN SEEDS	71: Natural or Cultured Pearls, Precious or Semiprecious Stones, Precious Metals, Metals Clad With Precious Metal, and Articles Thereof; Imitation Jewelry; Coin	12: Oil seeds and oleaginous fruits; miscellaneous grains, seeds and fruits; industrial or medicinal plants; straw and fodder		OK
HERBS	71: Natural or Cultured Pearls, Precious or Semiprecious Stones, Precious Metals, Metals Clad With Precious Metal, and Articles Thereof; Imitation Jewelry; Coin	12: Oil seeds and oleaginous fruits; miscellaneous grains, seeds and fruits; industrial or medicinal plants; straw and fodder		OK
CANADIAN BROWN FLAX SEED - X 100 LB POLYWOVEN BAGS CANADIAN BROWN FLAX SEED NET WEIGHT: 22 680 KGS / 50000 LBS GROSS WEIGHT: 22737 KGS / 50125 LBS PACKED IN EXPORT PACKERS BRAND 100 LB POLYWOVEN BAGS WITH BUYER'S TAG LOTE NO: S0613570 PRODUCTION DATE: FEB2020 EXPIRATION DATE: FEB2022 IN TRANSIT TO COCHABAMBA, BOL	71: Natural or Cultured Pearls, Precious or Semiprecious Stones, Precious Metals, Metals Clad With Precious Metal, and Articles Thereof; Imitation Jewelry; Coin	12: Oil seeds and oleaginous fruits; miscellaneous grains, seeds and fruits; industrial or medicinal plants; straw and fodder		OK
ORGANIC PULSES	71: Natural or Cultured Pearls, Precious or Semiprecious Stones, Precious Metals, Metals Clad With Precious Metal, and Articles Thereof; Imitation Jewelry; Coin	7: Edible vegetables and certain roots and tubers		OK
FURNITURE PARTS KNOBS AND PULLS	71: Natural or Cultured Pearls, Precious or Semiprecious Stones, Precious Metals, Metals Clad With Precious Metal, and Articles Thereof; Imitation Jewelry; Coin	44: Wood and articles of wood; wood charcoal		OK
VACUUM BOTTLE	71: Natural or Cultured Pearls, Precious or Semiprecious Stones, Precious Metals, Metals Clad With Precious Metal, and Articles Thereof; Imitation Jewelry; Coin	73: Articles of iron or steel		OK
1X40HC CONTAINERS SIGNATURE FLEECE SHORT	71: Natural or Cultured Pearls, Precious or Semiprecious Stones, Precious Metals, Metals Clad With Precious Metal, and Articles Thereof; Imitation Jewelry; Coin	63: Other made up textile articles; sets; worn clothing and worn textile articles; rags	62: Articles of apparel and clothing accessories, not knitted or crocheted	POK
WIND TURBINE	16: Preparations of meat, of fish or of crustaceans, molluscs or other aquatic invertebrates	73: Articles of iron or steel	84: Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof	POK
RUBBER SEALS	16: Preparations of meat, of fish or of crustaceans, molluscs or other aquatic invertebrates	39: Plastics and articles thereof	40: Rubber and articles thereof	POK
GIRLS 97%POLYESTER 3% ELASTANE KNITTED TROUS	64: Footwear, gaiters and the like; parts of such articles	62: Articles of apparel and clothing accessories, not knitted or crocheted		OK
MEAT AND EDIBLE MEAT OFFAL, SALTED, IN BRINE, DRIE - JAMON SERRANO C/PATA (HTS# 02101101)	44: Wood and articles of wood; wood charcoal	5: Products of animal origin, not elsewhere specified or included	2: Meat and edible meat offal	POK
TRUCK INTERIO, PLASTIC FOAM	44: Wood and articles of wood; wood charcoal	39: Plastics and articles thereof		OK
FLUCONAZOLE TABLETS 150MG	49: Printed books, newspapers, pictures and other products of the printing industry; manuscripts, typescripts and plans	38: Miscellaneous chemical products	30: Pharmaceutical products	POK

Figure 9: Manual verification of divergent classification cases