Using Natural Language Processing to Classify Administrative Data of Purchased Products

Gergely Attila Kiss

02.28.2023

Abstract

The motivation of the paper is to apply a novel data source to support the estimation of consumption expenditure and consumer price data in Hungary. Compared to similar approaches in the international literature the transaction data does not include bar code identifiers that makes the identification of consumption categories more difficult. My approach is based on Natural Language Processing combined with Machine Learning (ML) models. I experimented with unsupervised and supervised ML models to find the appropriate approach for classifying the data. My results suggest a general pattern where unsupervised models find high level patterns and supervised models are better for the detailed classification. From the tested supervised models three dominate in terms of accuracy: logistic regressions, random forest ensembles and multilayered perceptrons.

1 Introduction

This paper is connecting to the recent trend of innovations on how to support household consumption expenditure estimation and Consumer Price Index. In the international environment there are similar innovative projects that use scanner data and bar code based classification of purchased products ([Harms and Spinder 2019], [Martindale et al. 2020] and [Sands 2021]). However, the data at the disposal of HCSO are more alike with [Myklatun 2019] and [Roberson 2021] therefore I took them as role models for my analysis. The main reason behind this general pattern of innovation is that surveys have decreasing accuracy because of the observed rise in the non-response rate, sometimes even with more and better targeted incentives, and the opportunity of using private/administrative big data sources at the disposal of National Statistics Institutes (NSIs). The Hungarian Central Statistical Office (HCSO) experienced the same and now is developing a unique solution for supporting the consumption expenditure estimates in Hungary. In the case of the Hungarian data sources they do not record the bar codes thus the problem is somewhat different fundamentally leading me to the Natural Language Processing (NLP) based methods.

The HCSO's data comes from a cooperation with the National Tax and Customs Administration (NTCA) and the cooperation's goal is to be able to create more timely experimental statistics on the CPI and the gross purchased goods, that should be a highly relevant for consumption expenditures too. In this innovative project multiple data sources are used, most importantly data from the Online Cash Registers(OCR) and the Online Invoice System(OIS). Based on these data sources, we aim at building a Machine Learning based methodology to identify the appropriate COICOP and CN categories.

During my research I experimented with several ML techniques so far to see what could bring an optimal solution in creating the appropriate depth and precision for the classifications. Experiments with Unsupervised models, especially Latent Dirichlet Allocation, show that it cannot provide the necessary depth nor precision for a general solution although it preforms reasonably well for some categories when the text data is detailed enough. However, supervised learning provides a set of tools that, for my surprise too, performed exceptional on our test samples even though the train sets were small and unbalanced across categories. The three best performing model types are logistic regressions, random forests and multi layer perceptrons.

The structure of the paper continues as follows: in Section 2, I will introduce in a little bit more detail the data source and structure with emphasis on the unique aspects of the data. Section 3 will describe the modelling techniques and their corresponding results, first for the unsupervised approach and then for the supervised. Finally, in Section 4, I conclude and emphasis the lessons learnt during the experiments as well as describe the approach I would like to take in the future.

2 Data

The original data is the NTCA's property and is a highly sensitive data, as it contains the time and location of the purchase and in case of the Online Invoice System all the necessary data required to create that invoice. While also is extremely large as the Cash Register and Invoice System combined should contains all the legal transactions conducted in the country in real time and historically back until 2018. Therefore, even the HCSO's data access is restricted and we are currently working to create an appropriate virtual environment to be able to analyze the data more efficiently, as the current solutions, I will present later on, were only preliminary for the project and helped us to create a sound base for the cooperation.

It must be noted that most of the results are created on a non-representative sample of the data. The NTCA shared with us 3 different data samples, two from the Cash Register and one from the Invoice System. The two Cash Register samples are: 1) all the purchases at gas stations for July of 2021, and 2) a more restricted one on all the different item names from 2 of the largest retail merchants in Hungary. While the Invoice System data contains the purchases at the single largest e-tailer's of Hungary for the same month of July 2021. Also,

as notable from the 2^{nd} row of Table 1 the purchased goods are sometimes aggregated in a manner that the items with the exact matching names and unit prices are summed up.

The data structure in the two database the Online Cash Register and the Online Invoice System are actually very similar. The basic structure can be seen in Table 1, of course the table excludes the sensitive variables and shows only the ones the models rely on. The subtle difference between the Cash Register and the Invoice system is that the latter has more detailed item names and thus I thought it would be a more reliable source to experiment with the unsupervised approaches.

Table 1: Data Structure

Item name	Price(HUF)	Measurement Unit	Amount
Juicy Apples	235	kg	0.2
OMV Diesel	496	l	208,76
Parkside MPT	33599	pc	1
chca.bar.snk	524	amt	2
pc012	-150		10
...	...	...	...

The most important consequences of the structure above are 1) our sources do not include bar code identifiers and thus I used NLP techniques on the recorded item names. And 2) these item names are given by the retailer itself and many times, especially in the Cash Register, are just abbreviations of the product's name. That is what I incorporated in the above Table 1 with the variations on the item names, which makes manual coding and thus learning sample creations difficult too. All-in-all the data is very detailed and contains a lot of useful information that inspired the HCSO to invest in finding a sound method that could help later on the estimation procedure of household consumption expenditure. Even if the end product is just an auxiliary measurement to have a sanity check on our surveyed numbers.

3 Modelling

In this section I will cover the preprocessing of the data, as well as the unsupervised and supervised analysis ant their results. Starting off with the general ideas of preprocessing the text data, that deserves a side note just because it is so different than the usual small texts that come to mind. Then continue with the algorithm used for the unsupervised method and finishing with the results of the supervised modelling techniques.

In many cases when NLP related papers write about small texts they are usually related to twitter data and their processing is done by using word vectors and TF-IDF representations most commonly. The previously described data, similarly to Roberson 2021, however makes an exception from the above as the recorded texts are tend to be even shorter than tweets (our texts are about 5-50 character long vs. previous upper limit of 150 characters) and usually do not have the coherence of a sentence. Thereof, we choose the simplest approaches to preprocess the texts, e.g. One-Hot-Encoding, and frequency based Bag-of-Words. Furthermore, I experimented with some usual cleaning steps (e.g. delete numbers and words less than 3 character, etc.) just to came to the recognition that they actually worsen the accuracy of the models as they delete a part of important information that helps the learning. The final set of learning sample included the result of the preprocessing steps merged with the non text variables of the original data, resulting in a combined matrix.

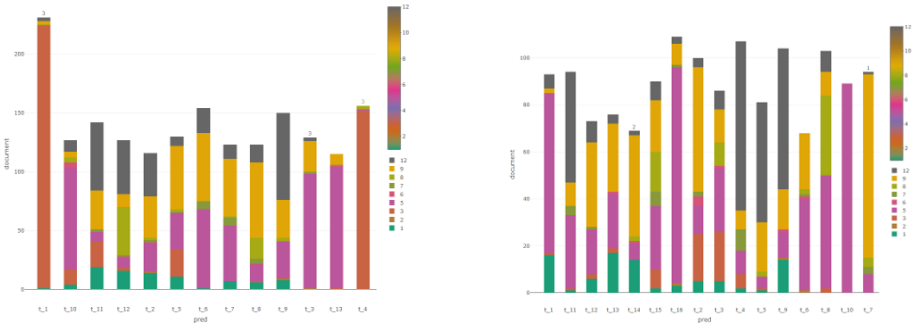
3.1 Unsupervised modelling

Let me turn to the first approach I experimented with unsupervised algorithms. There were three methods that I wanted to try on the Invoice System data Latent Dirichlet Allocation (LDA), Randomforest clustering and KMeans. In my case LDA was the dominant one as the other two were significantly less effective in finding proper clusters that could translate into COICOP categories. This goes with the literature where LDA is a usual benchmark in case of text based clustering, also known as topic modelling. Therefore, I turned to LDA and developed an iterative algorithm on Invoice System data that could be used to predict some of the categories.

The main idea behind the algorithm is that if after teaching the LDA to cluster the data into a given number of topics then I can relate a topic to a COICOP category if that consists mainly or unanimously of that category. After finding such topics I filter them out of the learning sample and re-teach the LDA to estimate another set of topics. I do this iteration as long as I can find eligible topics. After all the iterations I will have a set of estimators that can be used to predict if a text belongs to a given topic and therefore to a COICOP category. It must be noted that this approach is not an elegant or efficient way as to be able to decide which topic is eligible a labelled sample is necessary for measuring or eye-balling the structure of a topic.

In Figure 1 a two iteration process' decision graphs are shown. The columns are the topics and the colors are the COICOP divisions of the observation. In Panel 1a there are two columns eligible for separation the first and the last both showing mostly observations from clothing (orange). While in the second iteration there is only one topic to be filtered out, the one before the last which mainly consists of household appliances (purple). After the two steps in this case I have not find any new eligible topics.

To measure the two iterations by the standards of accuracy, recall and precision I created another manually labeled small random sample where these measures could be calculated, while also did a prediction for the whole Invoice



(a) First iteration (b) Second Iteration

Figure 1: LDA results

System data. The first round's two topic worked reasonably well in predicting clothes in the small sample resulting in a .97 precision and a .85 recall score. While the second round's topic was noticeably less accurate, especially in the false negative domain, with 1.0 precision, but .17 recall. This small recall for the second iteration should mainly come from the large dispersion of the COICOP category across the created topics. In the whole data case only the precision numbers could be calculated and for the first iteration it was similar while for the second iteration it decreased significantly to .6.

All the above implies that this approach cannot be a solution in itself for HCSO's problem of creating a classification to use for detailed expenditure estimation or CPI calculation. The main reason for that is even if the algorithm works with proper classification accuracy the implied COICOP divisions are much broader than what is needed as the division are on a 2-digit level and our publications require at least 4, but rather a 6-digit level classifications. However, the experiment is not in vain using LDA to imply meaningful topics could be used as a first step in more sophisticated approaches. It could be used as a first step to mine keywords that could help a string distance related more traditional type of classification or could be used to predict observations that belong to some division to filter the data and decrease the workload of creating a learning sample for the specific sub-divisional categories under that division.

3.2 Supervised modelling

Based on the previous lessons I also approached the classification problem with supervised modelling for products in the Cash Register data. Based on the previous lessons after creating the labelled sample I filtered it to include only the top 10 most frequent categories to get ahead of class imbalance and large dispersion type issue like in case of the second iteration of LDA. However, class imbalance somewhat persisted on as the first most frequent category(fuels) take more than 60% of the sample still, although the rest was quite evenly spread. Then I had to chose the types of models to be used for that I created model "families" to

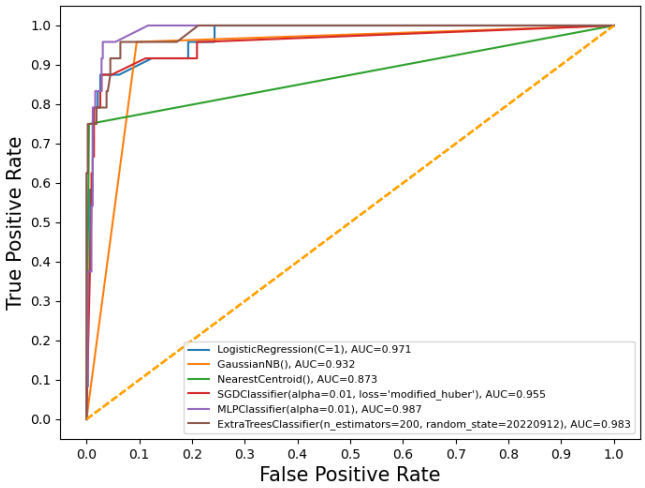


Figure 2: Soft Drink Classifier Example

be estimated where a "family" consisted of similar type models. Altogether, there was 6 of these model families: linear type, Bayesian type, neighbourhood type, SVM type, decision tree type and multi layered perceptron in itself. I used F2 score for hyperparameter tuning to increase the weight of recall in the parameter selection. It's purpose was to guard the results against the problem observed with the second round iteration of the LDA based algorithm.

The analysis itself is based on a horse race type view because I was not sure which model or model types would perform the best for this specific data. I first conducted the race in the model family to find the best model for a given category from a family. Where I used the ROC-AUC score as the measure of goodness of fit. And second, I compared the best model of each family to the best model of other families and decided on which model is the best for the given category. This second comparison can be seen on Figure 2 for classifiers of purchased soft drinks at gas stations. For the rest of the most frequent categories the comparisons resulted in almost identical plots with some variations of what is the family race winner model.

The result of the across family comparison was not straight forward for several reasons. First, as it can be seen on Figure 2 there is no clear winner in the sense of being significantly better and dominating the field. Second, there can be a decision made along just the ROC-AUC but comparing more than one statistics would provide a more sound basis for model selection. However, it only made the comparison more even for the 3 best type of models. These are the logistic regression, random forest ensembles and multilayered perceptrons. Out of these three none could show the best measurements for each of the 10 most frequent categories and decide in a later step of the project where I also have to take computation times into account.

4 Summary

To sum up, this paper is a starting point of a project that is to classify a massive data set in order to be a useful input for the household consumption expenditure and Consumer Price Index calculations by classifying the purchased products. In the focus is machine learning approaches as they can provide some scalable solutions. The most important results for this project is the fact that unsupervised modelling can not be a way to create the required classification although it still can be useful for implying higher hierarchical structure or providing some keywords for further analysis, and that from the supervised models there is a set of handful models that are providing the best results.

The further plans for the project is to extend the data in two dimensions: first, in time and retest the best 3 supervised algorithms, although this would require a large enough learning sample from where the models could learn more global patterns for categories. Second, the cooperation should expand the data gradually to other easily classifiable vendors or vendor types one at a type (eg. tobacco shops, clothing merchants, IKEA, etc.).

References

Harms, Alexander and Siemen Spinder (2019). "A comprehensive view of machine learning techniques for CPI production". In.

Martindale, Hazel et al. (2020). "Semi-supervised machine learning with word embedding for classification in price statistics". In: *Data amp; Policy* 2, e12. DOI: 10.1017/dap.2020.13.

Myklatun, Kristian Harald (2019). "Utilizing Machine Learning in the Consumer Price Index". In: Helsinki.

Roberson, Andrea (2021). "Applying Machine Learning for Automatic Product Categorization". In: *Journal of Official Statistics* 37.2, pp. 395–410. DOI: http://dx.doi.org/10.2478/JOS-2021-0017.

Sands, Helen (June 2021). *Research into the use of scanner data for constructing UK consumer price statistics*. URL: https://www.ons.gov.uk/economy/inflationandpriceindices/articles/.