Data science for tax administration

Pijnenburg, M.G.F.

8 Conclusion

In previous chapters, some selected topics in the field of data science and taxation have been treated in detail. These selected topics are limited in their scope, but can be considered as small bricks, that will allow science eventually to construct a solid building. In this chapter, we will look again at the conclusions of the various chapters, and try to place them in a wider context. To do so, we will look at each chapter individually and pay special attention to the research question underlying the chapter. After that, we will take a step back and reflect shortly on the future perspective of data science and taxation.

8.1 Analytics in Taxpayer Supervision

The research of this chapter started with the research question

> **Research Question** *What data science / analytical techniques can be used in* taxpayer supervision *and what contributions may be expected from these techniques?*

A list of the techniques that can be used can be found in Table 2.4. Concerning the expectations, the main contribution of the chapter is to bring realism in the accomplishments that can be expected from applying analytics to taxpayer supervision. Analytics will not replace the paradigm of Compliance Risk Management in taxpayer supervision, but has potential to support many current supervision activities within that framework. So, neither the claim that analytics is just a hype and will hardly contribute to taxpayer supervision, nor the claim that analytics will fundamentally change taxpayer supervision is supported.

The conclusion that analytics complements current activities, instead of replacing them, seems to have plausibility for other domains as well. Think for example about internal processes of tax administrations or other supervision tasks of governments.

At the same time, it is essential not to generalize this conclusion outside the realm of analytics. If we think of data science in general (recall Figure 1.7 for the difference between analytics and data science in this thesis), this also contains the field of Artificial Intelligence (AI) that is currently making fast progress towards reproducing certain aspects of human intelligence. Technologies developed in AI have potential for changing essential aspects of our society. And if society will change, taxation will be adjusted to it as well.

Figure 8.1 reflects our believes about the direct and indirect influence of developments in data science on taxation. We believe that the direct application of data science by tax administrations to improve its processes will most likely lead to more effectiveness (and some gain in efficiency), without changing fundamental paradigms (blue line in Figure 8.1). This is in line with the conclusions of chapter 2. In contrast, applications of data science in the wider society may change some fundamental aspects of society, and thereby indirectly changing fundamental aspects of taxation (brown dotted line in Figure 8.1).

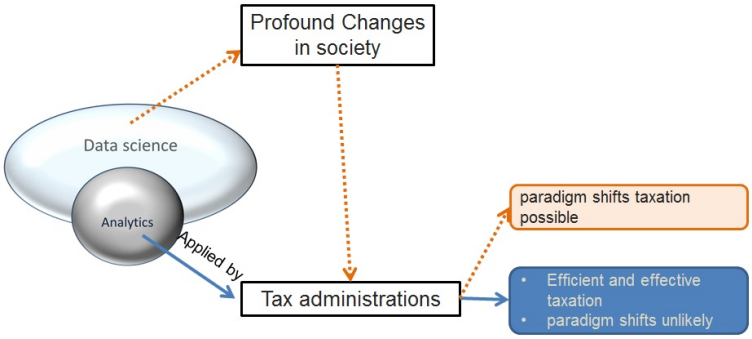

Figure 8.1: Two ways that data science influences a tax administration

8.2 Logistic Regression and Factorization Machines

The research question that was addressed in this chapter was:

> **Research Question** *Can we improve on current audit selection models by incorporating categorical variables with many values (like 'industry sector') that are valued by experts doing manual audit selection?*

The reason to address this question was the (at that time) surprising observation that in the manual selection of audits, categorical variables with many values were highly valued, while they were not incorporated in standard risk models at the NTCA.

After completing the research, it can be said that current audit selection models *can* be improved by including categorical variables with many values. The risk model that had started the research (a standard VAT audit selection model) could be improved by nearly 10%. We developed a technique that uses the recently developed Factorization Machines to achieve this goal.

The contribution is important beyond the actual problem at the NTCA. Many data scientists encounter categorical features with many levels and are looking for the right way to incorporate them in risk models or other applications. This is evident when looking for instance at the questions posed on this topic at Stack Overflow (www.stackoverflow.com), a main question and answer site in the domain of computer science. The solution described in the chapter, solves the problem in a satisfying way. The solution described transforms the categorical features in a data preprocessing step into numerical features. These numerical features can then be used in any classification or regression algorithm.

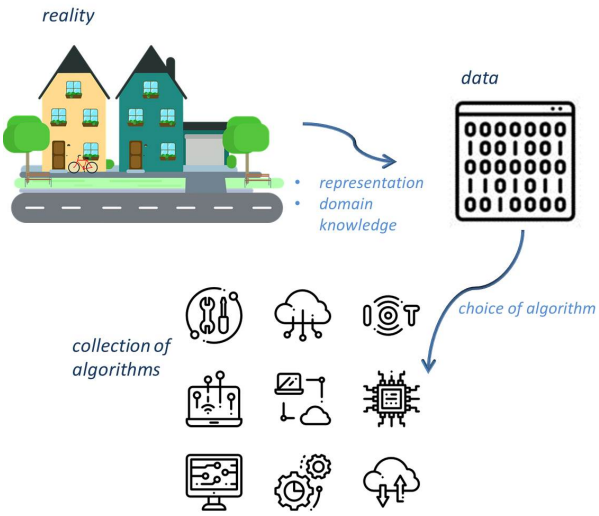

Figure 8.2: Visualization of the relation between the real world and data, stressing the importance of the way of representing the real world.

The idea of Factorization Machines to present categorical levels as vectors in a multidimensional space is attractive by itself: a multidimensional space contains a natural distance measure, allowing to quantify subtle differences between various levels. A spectacular example of this kind is presented by *word2vec* [101], a group

of models used in word embedding. These models are able to present each word as a vector, such that differences between vectors have an intrinsic meaning, like the famous example of *king - man + woman = queen*, where *king* is the vector belonging to the word 'king', *man* the vector belonging to the word 'man', et cetera. The plus and the minus sign in the equation are the usual addition and subtraction of vectors.

The representation of (interactions of) categorical features as numeric vectors is an example of an unconventional way of presenting reality in data. In this thesis, we met another example of the importance of a good data representation when discretizing numerical data to make it suitable as input for a Restricted Boltzmann Machine, see 'thermometer encoding' in Chapter 6 on page 83. Figure 8.2 shows that in general the representation of reality as input for an algorithm is an essential intermediate step. Data representation is often given little attention, but is important. We expect more good results in many application areas if more attention is paid to the representation of the data.

8.3 Singular Outliers

When applying standard anomaly detection algorithms to find tax fraud, we noticed that most of the anomalies found were not related to fraud. This sparked the following research question:

> **Research Question** *Can we develop a new anomaly detection algorithm tailored to find the outliers that are of interest to a tax administration?*

The answer is affirmative. In Chapter 4 we argue that interesting outliers do not necessarily have anomalous values for all features. Observations with anomalous values for one or a few features might be even of more interest in the domain of tax fraud detection or detecting data quality issues. The phenomenon of singular outliers is not only restricted to the tax domain. Also in other fraud types, like credit card fraud, clues are often derived from one or a few features.

The algorithm (SODA) that we developed is the first algorithm developed to detect this type of outliers. We successfully applied a previous version of the algorithm to select VAT field audits.

Since it is a first algorithm. we can imagine that other algorithms may be found in the future that can perform even better. For instance, existing anomaly detection algorithms may be adapted to find singular outliers. Also, the current algorithm may be refined by by taking other values for p instead of 2 and 1 in the Local Euclidean Manhattan Ratio, see Section 4.5.

At this place we want to point to a similarity between singular outlier detection and sub-group discovery. Sub-group discovery can be seen as clustering observations

on a subset of the features [8]. Before the arrival of sub-group discovery, many clustering techniques existed that clustered observations taking into account *all* features. Sub-group discovery became popular when it was realized that in some applications, interest lies in observations that form clusters when restricted to a few features. This is comparable with singular outlier detection where outliers are to be found for a few features.

8.4 Are Similar Cases Treated Similarly?

'Similar treatment of similar cases' is important for tax administrations. The phrase took a prominent place in the strategic five-year plan of the NTCA. Although it is easy to state this principle theoretically, it is not easy to impose it in practice in a large organization. Many factors may influence the treatment of taxpayers, especially for Knowledge Intensive processes, that are abundant at tax administrations. For this reason, we formulated our fifth research question as follows,

Research Question *Can we develop a data science technique that assists in testing similar treatment of similar (tax) cases?*

The test procedures introduced in Chapter 5, see pages 70 and 72, give a tool to test on similar treatment of similar cases. The procedures expect a process log as input, extract features and frequencies, and then apply a sequence of χ^2-tests. In case the Null hypothesis of similar treatment is rejected, the values of the individual χ^2-statistics allow to pinpoint the places in the process that most contributed to the rejection. The test procedures that are developed belong to *process mining*.

Process mining is still a relatively young field of research, but has a lot of potential for tax administrations and other organizations. Since processes are more and more controlled by state of the art case management systems that record valuable data, the tools of process mining can be used to optimize these processes. However, almost no tools exist yet for automatically checking *constraints* of processes, like the similar treatment of similar cases. We think that the use of statistical techniques, as is done in Chapter 5, can contribute further to process mining.

8.5 Extending an Anomaly Detection Benchmark

The topic of Chapter 6 is somewhat different compared to the other chapters. Its contribution lies in comparing the performance of existing algorithms. The comparison of algorithms contributes to the general body of knowledge, since science is not only about discovering new facts, but also about placing these facts in an agreed, coherent framework.

Comparing different algorithms also has practical value. A frequently asked question from data scientists is what algorithm should be used to solve a problem at hand. We can answer such questions, as soon as we know the advantages and disadvantages of various algorithms, and in what situations the strength of a particular algorithm can be utilized. The formal research question underlying chapter 6 is:

Research Questions *Unsupervised anomaly detection algorithms play an important role in (tax) fraud detection. What algorithms can be expected to work well under what conditions?*

The main finding of the chapter is that some exotic algorithms like Restricted Boltzmann Machines, or some new algorithms like Isolation Forests perform well on a range of data sets. For some data sets the investigated algorithms beat the current best algorithm. The results of the chapter are summarized in Table 6.3.

An interesting finding of the chapter is that simple algorithms for detecting anomalies, such as k-nearest neighbors and Histogram Based Outlier Score, perform well on a large number of data sets. This suggests a practical strategy to first try simple algorithms for a problem and then to build further from these results with more specific algorithms, like the ones tested in Chapter 6.

8.6 Outlook

Tax authorities around the world will continue to be large information processing organizations. If current trends of an increasingly dynamic society, increasing citizen expectations and internationalization continue, the tax administrations will face enough challenges. Fortunately, the constant developments in computer science can offer instruments to deal with an increasing workload. The application of data science in the tax authorities has started, but has not yet reached the highest maturity levels. In the near future, we expect progress in at least the following topics, that are of interest for other organizations as well,

- automatically categorizing incoming documents,
- developing new algorithms for detecting (tax) fraud in networks of agents,
- privacy-preserving data mining.

Certainly, not all challenges of data science and taxation are of a technical nature. When tax administrations are embracing data science, organizational and ethical questions arise, like,

- How to position data science in relation to classical ICT?

- How to organize teams of data scientist for optimal cooperation?
- How to organize a portfolio process?
- How to organize quality assurance and quality control?

The organizational questions may be of interest to Business Science.

Moreover ethical questions are playing an important role already, and will become even more important in the coming years. These are questions like,

- How to preserve the privacy of citizens and how to conform to privacy guidelines?
- In what cases do we allow (semi-) automatic decision making and in what cases do we want to avoid this?
- How can we develop effective controls that prevent the emergence of a police state?

The right for the protection of privacy is stated in various legal documents, like the Convention for the Protection of Human Rights and Fundamental Freedoms, article 8, see [28].

The question of automatic decision making is relevant since computer systems may miss important information about the context, and may therefore take a wrong decision. Moreover, it may be hard for a citizen to dispute such an automatic decision once human process workers have been replaced by cheaper, more efficient machines. Currently, the view is emerging that intelligent computer systems should support human decision making, and not take over complex decisions with a wide impact, see for instance the European Union Communication on Building Trust in Human-Centric Artificial Intelligence [35].

The power that data science and artificial intelligence may give to governments, is subject to debate currently [37, 84]. Although this power may do much good, it may be used by a future government to tightly control its citizens and to undermine democratic principles, like individual freedoms. Regulatory frameworks may be needed to prevent if one wants to avoid this.