Data science for tax administration

Pijnenburg, M.G.F.

Summary

This dissertation is about applications of data science techniques to the administration of taxes. These applications can provide effective and efficient processes within a tax administration or improve the compliance of taxpayers. On the one hand, existing techniques are assessed, and on the other, new techniques have been developed.

After an introductory chapter, the question is addressed which processes in the tax domain are suitable for improvement with data science applications (Chapter 2). Because the entire tax domain is extensive, we limit ourselves to an important subdomain: *taxpayer supervision*. A subtle picture of the (im-)possibilities of data science is created by making an inventory of the supervisory activities and investigating for each activity which data science techniques can add value. The chapter also provides an overview of techniques that can be deployed. As such it is applicable for tax authorities that want to improve their supervisory activities with data science.

In Chapter 3 an existing data science technique, *logistic regression*, is extended. Logistic regression has difficulty in getting information efficiently from certain types of data, the so-called 'categorical variables with many levels'. Examples are 'place of residence' or 'occupation'. Because these variables occur regularly with tax authorities and because logistic regression is often used as well, this extension is a useful addition. Categorical variables with many levels occur also outside the domain of taxation, so this extension may add value there as well. Probably the same techniques of incorporating categorical variables with many levels can also be applied to techniques other than logistic regression, but this has not been investigated further.

In Chapter 4 we look at anomaly detection techniques (i.e. finding outliers in data). Tax data often includes a typical type of anomalies (also called *outliers*) that are not well detected by existing techniques. In the chapter this type of outliers is described and a new technique has been developed that finds these anomalies. Within tax administrations, the technique makes sense to detect tax fraud or to improve the

data quality by detecting errors in data entry.

Chapter 5 is about similar treatment of similar cases. For large organizations, such as tax authorities, it is not easy to ensure that equal cases always receive equal treatment. In particular for those processes that require a human assessment ('Knowledge Intensive Processes'). A first step to arrive at similar treatment is to *test* whether similar cases are treated similarly and, if not, where in the process the dissimilar treatment occurs. For this purpose, two new procedures are introduced in Chapter 5. These procedures belong to the domain of *process mining*.

Science consists of discovering new knowledge, but also organizing that knowledge in logical frameworks. In Chapter 6 the latter topic is looked at: three little-used / new anomaly detection techniques are being tested on a benchmark, recently published by Goldstein and Uchida [43]. This comparison helps in understanding the strengths and weaknesses of various anomaly detection techniques. Surprisingly, the three new techniques are able to match or improve the benchmark for 30% of the data sets. In addition, attention is given to the relation between the new methods and the more traditional anomaly detection methods. Moreover we look at what type of anomalies are found.

In Chapter 7 four smaller topics are discussed. The chapter gives a good impression of the variety of data science topics within the tax domain. The first subsection deals with a topic from HR Analytics, a sub-area within data science that focuses on applications within the Human Resources (HR) domain. The next subsection looks at an application of Reinforcement Learning techniques within the collection process of a tax authority. Reinforcement Learning is the branch of Artificial Intelligence that focuses on learning strategies for board games and computer games. The third topic is about explaining data science models to end-users. The last topic concerns applying ideas from the Fuzzy Set community in tax domain.

The dissertation ends with conclusions and a short outlook.