



Universiteit  
Leiden  
The Netherlands

## Data science for tax administration

Pijnenburg, M.G.F.

### Citation

Pijnenburg, M. G. F. (2020, June 24). *Data science for tax administration*. Retrieved from <https://hdl.handle.net/1887/123049>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/123049>

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/123049> holds various files of this Leiden University dissertation.

**Author:** Pijnenburg, M.G.F.

**Title:** Data science for tax administration

**Issue Date:** 2020-06-24

# **Data Science for Tax Administration**

Mark Gerardus Franciscus Pijnenburg



Universiteit  
Leiden

Copyright 2020 by Mark Pijnenburg

Cover background image: Black framed eyeglasses by Jesus Kiteque on Unsplash

Typeset using  $\text{\LaTeX}$

Printed by Ridderprint — [www.ridderprint.nl](http://www.ridderprint.nl)

ISBN 978-94-6375-942-7

# Data Science for Tax Administration

## Proefschrift

ter verkrijging van  
de graad van Doctor aan de Universiteit Leiden,  
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,  
volgens besluit van het College voor Promoties  
te verdedigen op woensdag 24 juni 2020  
klokke 15:00 uur

door

MARK GERARDUS FRANCISCUS PIJNENBURG

geboren te Heerlen  
in 1979

**Promotores:** Prof. dr. ir. W. Kraaij  
Prof. dr. J.N. Kok

**Co-promotor:** Dr. W.J. Kowalczyk

**Promotiecommissie:**

Prof. dr. A. Plaat (voorzitter)

Prof. dr. T.H.W. Bäck (secretaris)

Prof. mr. dr. E.C.J.M. van der Hel - van Dijk RA (Nyenrode Business Universiteit)

Prof. dr. H. Blockeel (KU Leuven)

Prof. dr. A.P.J.M. Siebes (Universiteit Utrecht)

Prof. dr. ir. R. Arendsen

Dr. C.J. Veenman

Dr. A.J. Knobbe

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The Historical Link between Taxes and Data . . . . .	1
1.2	Tax Administrations in the Computer Age . . . . .	3
1.3	Analytics at the Netherlands Tax and Customs Administration . . . . .	5
1.4	Analytics and Related Concepts . . . . .	7
1.5	Research Questions . . . . .	9
1.6	Outline of Thesis . . . . .	12
1.7	Publications Related to Thesis . . . . .	14
<b>2</b>	<b>Analytics in Taxpayer Supervision</b>	<b>15</b>
2.1	Introduction . . . . .	16
2.1.1	Research objective . . . . .	16
2.1.2	Need for more effective taxpayer supervision . . . . .	16
2.1.3	Scope of the research . . . . .	17
2.1.4	Organization of this chapter . . . . .	17
2.2	Related Research . . . . .	18
2.2.1	Theories about modern supervision . . . . .	18
2.2.2	Analytical applications within tax administrations . . . . .	19
2.2.3	Managerial literature on analytics . . . . .	19
2.3	Practical Experiences with Analytics . . . . .	20
2.3.1	General experiences of tax administrations . . . . .	20
2.3.2	Practical experiences in banking and insurance . . . . .	21
2.4	Analytics for Taxpayer Supervision . . . . .	22

---

2.4.1	Activities in taxpayer supervision . . . . .	22
2.4.2	Classes of analytical techniques . . . . .	25
2.4.3	Matching supervision activities and analytical techniques . . . . .	27
2.4.4	A roadmap for analytics in compliance risk management . . . . .	29
2.5	Case Study: VAT refund risk model . . . . .	30
2.6	Conclusions . . . . .	32
2.7	Discussion . . . . .	33
<b>3</b>	<b>Extending Logistic Regression Models with Factorization Machines</b>	<b>35</b>
3.1	Introduction . . . . .	37
3.2	Related Research . . . . .	38
3.2.1	Existing Methods for Many Levels . . . . .	38
3.2.2	Factorization Machines . . . . .	39
3.3	Combining Logistic Regression with Factorization Machines . . . . .	39
3.3.1	Plain Logistic Regression (PLM) . . . . .	40
3.3.2	Logistic Regression with Grouping (LRG) . . . . .	40
3.3.3	LRFM1 . . . . .	41
3.3.4	LRFM2 . . . . .	41
3.3.5	LRFM3 . . . . .	41
3.3.6	LRFM4 . . . . .	42
3.4	Experiments . . . . .	42
3.4.1	Data Sets . . . . .	42
3.4.2	Model Quality Measures . . . . .	44
3.4.3	Settings Factorization Machines . . . . .	44
3.4.4	Results . . . . .	45
3.5	Conclusion and Discussion . . . . .	45
<b>4</b>	<b>Singular Outliers: Finding Common Observations with an Uncommon Feature</b>	<b>47</b>
4.1	Singular Outliers . . . . .	48
4.2	Related Work . . . . .	50
4.3	Singular Outlier Detection Algorithm . . . . .	51
4.3.1	The Algorithm . . . . .	51
4.3.2	Measurement of Characteristics of Singular Outliers . . . . .	53
4.4	Comparison . . . . .	54
4.4.1	Marks Data . . . . .	55
4.4.2	Istanbul Stock Exchange Data . . . . .	56
4.4.3	Wholesale Customers Data . . . . .	56
4.4.4	Polish Bankruptcy Data . . . . .	57
4.4.5	Algae Data . . . . .	58



---

4.5	Conclusion and Discussion . . . . .	58
<b>5</b>	<b>Are Similar Cases Treated Similarly? A comparison between process workers</b>	<b>61</b>
5.1	Introduction . . . . .	62
5.2	Related Work . . . . .	64
5.2.1	Process Drift . . . . .	64
5.2.2	Sequence Analysis . . . . .	65
5.3	Testing for similar treatment of similar cases . . . . .	65
5.3.1	Approach 1: Independence of Decisions . . . . .	66
5.3.2	Approach 2: Frequent Paths . . . . .	70
5.4	Experiments . . . . .	72
5.4.1	Synthetic data . . . . .	72
5.4.2	Tax Debt Collection data . . . . .	74
5.5	Conclusion and Future Research . . . . .	76
<b>6</b>	<b>Extending an Anomaly Detection Benchmark with Auto-encoders, Isolation Forests, and RBMs</b>	<b>79</b>
6.1	Introduction . . . . .	80
6.2	Theoretical Background . . . . .	80
6.2.1	Sparse Auto-encoder . . . . .	80
6.2.2	Isolation Forest . . . . .	83
6.2.3	Restricted Boltzmann Machine . . . . .	83
6.2.4	Type of outliers . . . . .	87
6.3	Experimental Setup . . . . .	88
6.3.1	General . . . . .	88
6.3.2	Setting hyper-parameters . . . . .	88
6.3.3	Data Sets . . . . .	90
6.4	Results . . . . .	92
6.5	Conclusions and Discussion . . . . .	97
<b>7</b>	<b>Various Topics</b>	<b>99</b>
7.1	Perceived Human Resources Practices and Time Allocation . . . . .	101
7.2	Reinforcement learning applied to tax debt collection . . . . .	102
7.3	Explaining risk models to non-experts . . . . .	104
7.4	Recommendations based on fuzzy sets . . . . .	107
<b>8</b>	<b>Conclusion</b>	<b>113</b>
8.1	Analytics in Taxpayer Supervision . . . . .	113
8.2	Logistic Regression and Factorization Machines . . . . .	114

8.3 Singular Outliers . . . . .	116
8.4 Are Similar Cases Treated Similarly? . . . . .	117
8.5 Extending an Anomaly Detection Benchmark . . . . .	117
8.6 Outlook . . . . .	118
<b>Summary</b>	<b>121</b>
<b>Samenvatting</b>	<b>123</b>
<b>Acknowledgements</b>	<b>125</b>
<b>Curriculum Vitae English</b>	<b>127</b>
<b>Curriculum Vitae Nederlands</b>	<b>129</b>
<b>Bibliography</b>	<b>140</b>
<b>Publication List Author</b>	<b>141</b>