

INNOVATION, ENTREPRENEURSHIP AND MANAGEMENT SERIES

BIG DATA, ARTIFICIAL INTELLIGENCE AND DATA ANALYSIS SET



Volume 1

Big Data for Insurance Companies

**Edited by
Marine Corlosquet-Habart
Jacques Janssen**

ISTE

WILEY

Big Data for Insurance Companies

**Big Data, Artificial Intelligence
and Data Analysis Set**

coordinated by
Jacques Janssen

Volume 1

**Big Data for Insurance
Companies**

Edited by

Marine Corlosquet-Habart
Jacques Janssen

ISTE

WILEY

First published 2018 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

© ISTE Ltd 2018

The rights of Marine Corlosquet-Habart and Jacques Janssen to be identified as the authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

Library of Congress Control Number: 2017959466

British Library Cataloguing-in-Publication Data

A CIP record for this book is available from the British Library

ISBN 978-1-78630-073-7

Contents

Foreword	xi
Jean-Charles POMEROL	
Introduction	xiii
Marine CORLOSQUET-HABART and Jacques JANSSEN	
Chapter 1. Introduction to Big Data and Its Applications in Insurance	1
Romain BILLOT, Cécile BOTHOREL and Philippe LENCA	
1.1. The explosion of data: a typical day in the 2010s	1
1.2. How is big data defined?	4
1.3. Characterizing big data with the five Vs.	5
1.3.1. Variety	6
1.3.2. Volume	7
1.3.3. Velocity	9
1.3.4. Towards the five Vs: veracity and value	9
1.3.5. Other possible Vs.	11
1.4. Architecture.	11
1.4.1. An increasingly complex technical ecosystem.	12
1.4.2. Migration towards a data-oriented strategy.	17
1.4.3. Is migration towards a big data architecture necessary?	18
1.5. Challenges and opportunities for the world of insurance	20
1.6. Conclusion	22
1.7. Bibliography	23

Chapter 2. From Conventional Data Analysis Methods to Big Data Analytics	27
Gilbert SAPORTA	
2.1. From data analysis to data mining: exploring and predicting	27
2.2. Obsolete approaches	28
2.3. Understanding or predicting?	30
2.4. Validation of predictive models	30
2.4.1. Elements of learning theory	31
2.4.2. Cross-validation	34
2.5. Combination of models	34
2.6. The high dimension case	36
2.6.1. Regularized regressions	36
2.6.2. Sparse methods	38
2.7. The end of science?	39
2.8. Bibliography	40
Chapter 3. Statistical Learning Methods	43
Franck VERMET	
3.1. Introduction	43
3.1.1. Supervised learning	44
3.1.2. Unsupervised learning	46
3.2. Decision trees	46
3.3. Neural networks	49
3.3.1. From real to formal neuron	50
3.3.2. Simple Perceptron as linear separator	52
3.3.3. Multilayer Perceptron as a function approximation tool	54
3.3.4. The gradient backpropagation algorithm	56
3.4. Support vector machines (SVM)	62
3.4.1. Linear separator	62
3.4.2. Nonlinear separator	66
3.5. Model aggregation methods	66
3.5.1. Bagging	67
3.5.2. Random forests	69
3.5.3. Boosting	70
3.5.4. Stacking	74
3.6. Kohonen unsupervised classification algorithm	74
3.6.1. Notations and definition of the model	76
3.6.2. Kohonen algorithm	77
3.6.3. Applications	79
3.7. Bibliography	79

Chapter 4. Current Vision and Market Prospective	83
Florence PICARD	
4.1. The insurance market: structured, regulated and long-term perspective	83
4.1.1. A highly regulated and controlled profession	84
4.1.2. A wide range of long-term activities	85
4.1.3. A market related to economic activity.	87
4.1.4. Products that are contracts: a business based on the law	87
4.1.5. An economic model based on data and actuarial expertise	88
4.2. Big data context: new uses, new behaviors and new economic models.	89
4.2.1. Impact of big data on insurance companies	90
4.2.2. Big data and digital: a profound societal change	91
4.2.3. Client confidence in algorithms and technology.	93
4.2.4. Some sort of negligence as regards the possible consequences of digital traces	94
4.2.5. New economic models.	95
4.3. Opportunities: new methods, new offers, new insurable risks, new management tools	95
4.3.1. New data processing methods	96
4.3.2. Personalized marketing and refined prices	98
4.3.3. New offers based on new criteria	100
4.3.4. New risks to be insured	101
4.3.5. New methods to better serve and manage clients	102
4.4. Risks weakening of the business: competition from new actors, “uberization”, contraction of market volume	103
4.4.1. The risk of demutualization	103
4.4.2. The risk of “uberization”	104
4.4.3. The risk of an omniscient “Google” in the dominant position due to data	105
4.4.4. The risk of competition with new companies created for a digital world.	105
4.4.5. The risk of reduction in the scope of property insurance	106
4.4.6. The risk of non-access to data or prohibition of use.	107
4.4.7. The risk of cyber attacks and the risk of non-compliance	108
4.4.8. Risks of internal rigidities and training efforts to implement	109

4.5. Ethical and trust issues	109
4.5.1. Ethical charter and labeling: proof of loyalty	110
4.5.2. Price, ethics and trust.	112
4.6. Mobilization of insurers in view of big data.	113
4.6.1. A first-phase “new converts”	113
4.6.2. A phase of appropriation and experimentation in different fields	115
4.6.3. Changes in organization and management and major training efforts to be carried out	118
4.6.4. A new form of insurance: “connected” insurance	118
4.6.5. Insurtech and collaborative economy press for innovation	121
4.7. Strategy avenues for the future	122
4.7.1. Paradoxes and anticipation difficulties	122
4.7.2. Several possible choices.	123
4.7.3. Unavoidable developments	127
4.8. Bibliography	128

Chapter 5. Using Big Data in Insurance 131

Emmanuel BERTHELÉ

5.1. Insurance, an industry particularly suited to the development of big data	131
5.1.1. An industry that has developed through the use of data.	131
5.1.2. Link between data and insurable assets	136
5.1.3. Multiplication of data sources of potential interest	138
5.2. Examples of application in different insurance activities	141
5.2.1. Use for pricing purposes and product offer orientation.	142
5.2.2. Automobile insurance and telematics	143
5.2.3. Index-based insurance of weather-sensitive events.	145
5.2.4. Orientation of savings in life insurance in a context of low interest rates	146
5.2.5. Fight against fraud	148
5.2.6. Asset management	150
5.2.7. Reinsurance	150

5.3. New professions and evolution of induced organizations for insurance companies	151
5.3.1. New professions related to data management, processing and valuation	151
5.3.2. Development of partnerships between insurers and third-party companies	153
5.4. Development constraints	153
5.4.1. Constraints specific to the insurance industry	153
5.4.2. Constraints non-specific to the insurance industry	155
5.4.3. Constraints, according to the purposes, with regard to the types of algorithms used	158
5.4.4. Scarcity of profiles and main differences with actuaries	159
5.5. Bibliography	161
List of Authors	163
Index	165

Foreword

Big data is not just a slogan, but a reality as shown by this book. Many companies and organizations in the fields of banking, insurance and marketing accumulate data but have not yet reaped the full benefits. Until then, statisticians could make these data more meaningful: through correlations and the search for major components. These methods provided interesting, sometimes important, but aggregated information.

The major innovation is that the power of computers now enables us to do two things that are completely different from what was done before:

- accumulate individual data on thousands or even millions of clients of a bank or insurance company, and even those who are not yet clients, and process them separately;
- deploy the massive use of unsupervised learning algorithms.

These algorithms, which, in principle, have been known for about 40 years, require computing power that was not available at that time and have since improved significantly. They are unsupervised, which means that from a broad set of behavioral data, they predict with amazing accuracy the subsequent decisions of an individual without knowing the determinants of his/her action.

In the first three chapters of this book, key experts in applied statistics and big data explain where the data come from and how they are used. The second and third chapters, in particular, provide details on the functioning of learning algorithms which are the basis of the spectacular results when using massive data. The fourth and fifth chapters are devoted to applications in the insurance

sector. They are absolutely fascinating because they are written by highly skilled professionals who show that tomorrow's world is already here.

It is unnecessary to emphasize the economic impact of this study; the results obtained in detecting fraudsters are a tremendous reward to investments in massive data.

To the best of my knowledge, this is the first book that illustrates so well, in a professional context, the impact and real stakes of what some call the “big data revolution”. Thus, I believe that this book will be a great success in companies.

Jean-Charles POMEROL
Chairman of the Scientific Board of ISTE Editions

Introduction

This book presents an overview of big data methods applied to insurance problems. Specifically, it is a multi-author book that gives a fairly complete view of five important aspects, each of which is presented by authors well known in the fields covered, who have complementary profiles and expertise (data scientists, actuaries, statisticians, engineers). These range from classical data analysis methods (including learning methods like *machine learning*) to the impact of *big data* on the present and future insurance market.

Big data, megadata or massive data apply to datasets that are so vast that not only the popular data management methods but also the classical methods of statistics (for example, inference) lose their meaning or cannot apply.

The exponential development of the power of computers linked to the crossroads of this data analysis with artificial intelligence helps us to initiate new analysis methods for gigantic databases that are mostly found in the insurance sector as presented in this book.

The first chapter, written by Romain Billot, Cécile Bothorel and Philippe Lenca (IMT Atlantique, Brest), presents a sound introduction to big data and its application to insurance. This chapter focuses on the impact of megadata, showing that hundreds of millions of people generate billions of bytes of data each day. The classical characterization of big data by 5Vs is well illustrated and enriched by other Vs such as variability and validity.

In order to remedy the insufficiency of classical data management techniques, the authors develop parallelization methods for data as well as possible tasks thanks to the development of computing via the parallelism of several computers.

The main IT tools, including Hadoop, are presented as well as their relationship with platforms specialized in decision-making solutions and the problem of migrating to a given oriented strategy. Application to insurance is tackled using three examples.

The second chapter, written by Gilbert Saporta (CNAM, Paris), reviews the transition from classical data analysis methods to big data, which shows how big data is indebted to data analysis and artificial intelligence, notably through the use of supervised or non-supervised learning methods. Moreover, the author emphasizes the methods for validating predictive models since it has been established that the ultimate goal for using big data is not only geared towards constituting gigantic and structured databases, but also and especially as a description and prediction tool from a set of given parameters.

The third chapter, written by Franck Vermet (EURIA, Brest), aims at presenting the most commonly used actuarial statistical learning methods applicable to many areas of life and non-life insurance. It also presents the distinction between supervised and non-supervised learning and the rigorous and clear use of neural networks for each of the methods, particularly the ones that are mostly used (decision trees, backpropagation of perceptron gradient, support vector machines, boosting, stacking, etc.).

The last two chapters are written by insurance professionals. In Chapter 4, Florence Picard (Institute of Actuaries, Paris) describes the present and future insurance market based on the development of big data. It illustrates its implementation in the insurance sector by particularly detailing the impact of big data on management methods, marketing and new insurable risks as well as data security. It pertinently highlights the emergence of new managerial techniques that reinforce the importance of continuous training.

Emmanuel Berthel  (Optimind Winter, Paris) is the author of the fifth and last chapter, who is also an actuary. He presents the main uses of big data in insurance, particularly pricing and product offerings, automobile and telematics insurance, index-based insurance, combating fraud and reinsurance. He also lays emphasis on the regulatory constraints specific to the sector

(Solvency II, ORSA, etc.) and the current restriction on the use of certain algorithms due to an audibility requirement, which will undoubtedly be uplifted in the future.

Finally, a fundamental observation emerges from these last two chapters cautioning insurers against preserving the mutualization principle which is the founding principle of insurance because as Emmanuel Berthel  puts it:

“Even if the volume of data available and the capacities induced in the refinement of prices increase considerably, the personalization of price is neither fully feasible nor desirable for insurers, insured persons and society at large.”

In conclusion, this book shows that big data is essential for the development of insurance as long as the necessary safeguards are put in place. Thus, this book is clearly addressed to insurance and bank managers as well as master’s students in actuarial science, computer science, finance and statistics, and, of course, new master’s students in big data who are currently increasing.

Introduction to Big Data and Its Applications in Insurance

1.1. The explosion of data: a typical day in the 2010s

At 7 am on a Monday like any other, a young employee of a large French company wakes up to start her week at work. As for many of us, technology has appeared everywhere in her daily life. As soon as she wakes up, her connected watch, which also works as a sports coach when she goes jogging or cycling, gives her a synopsis of her sleep quality and a score and assessment of the last few months. Data on her heartbeat measured by her watch are transmitted by WiFi to an app installed on her latest generation mobile, before her sleep cycles are analyzed to produce easy-to-handle quality indicators, like an overall score, and thus encourage fun and regular monitoring of her sleep. It is her best night's sleep for a while and she hurries to share her results by text with her best friend, and then on social media via Facebook and Twitter. In this world of connected health, congratulatory messages flood in hailing her “performance”! During her shower, online music streaming services such as Spotify or Deezer suggest a “wake-up” playlist, put together from the preferences and comments of thousands of users. She can give feedback on any of the songs for the software to adapt the

Chapter written by Romain BILLOT, Cécile BOTHOREL and Philippe LENCA.

upcoming songs in real time, with the help of a powerful recommendation system based on historical data. She enjoys her breakfast and is getting ready to go to work when the public transport Twitter account she subscribes to warns her of an incident causing serious disruption on the transport network. Hence, she decides to tackle the morning traffic by car, hoping to avoid arriving at work too late. To help her plan her route, she connects to a traffic information and community navigation app that obtains traffic information from GPS records generated by other drivers' devices throughout their journeys to update a real-time traffic information map. Users can flag up specific incidents on the transport network themselves, and our heroine marks slow traffic caused by an accident. She decides to take the alternative route suggested by the app. Having arrived at work, she vents her frustration at a difficult day's commute on social media. During her day at work, on top of her professional activity, she will be connected online to check her bank account balance and go shopping on a supermarket's "drive" app that lets her do her shop online and pick it up later in her car. Her consumer profile on the online shopping app gives her a historical overview of the last few months, as well as suggesting products that are likely to interest her. On her way home, the trunk full with food, some street art painted on a wall immediately attracts her attention. She stops to take a photo, edits it with a color filter and shares it on a social network similar to Instagram. The photo immediately receives about 10 "likes". That evening, a friend comments on the photo. Having recognized the artist, he gives her a link to an online video site like YouTube. The link is for a video of the street art being painted, put online by the artist to increase their visibility. She quickly watches it. Tired, she eats, plugs in her sleep app and goes to bed.

Between waking up and going to sleep, our heroine has generated a significant amount of data, a volume that it would have been difficult to imagine a few years earlier. With or without her knowledge, there have been hundreds of megabytes of data flow and digital records of her tastes, moods, desires, searches, location, etc. This *homo sapiens*, now *homo numericus*, is not alone – billions of us do the same. The figures are revealing and their growth astonishing: we have entered the era of big data. In 2016, one million links were shared, two million friend requests were made and three million