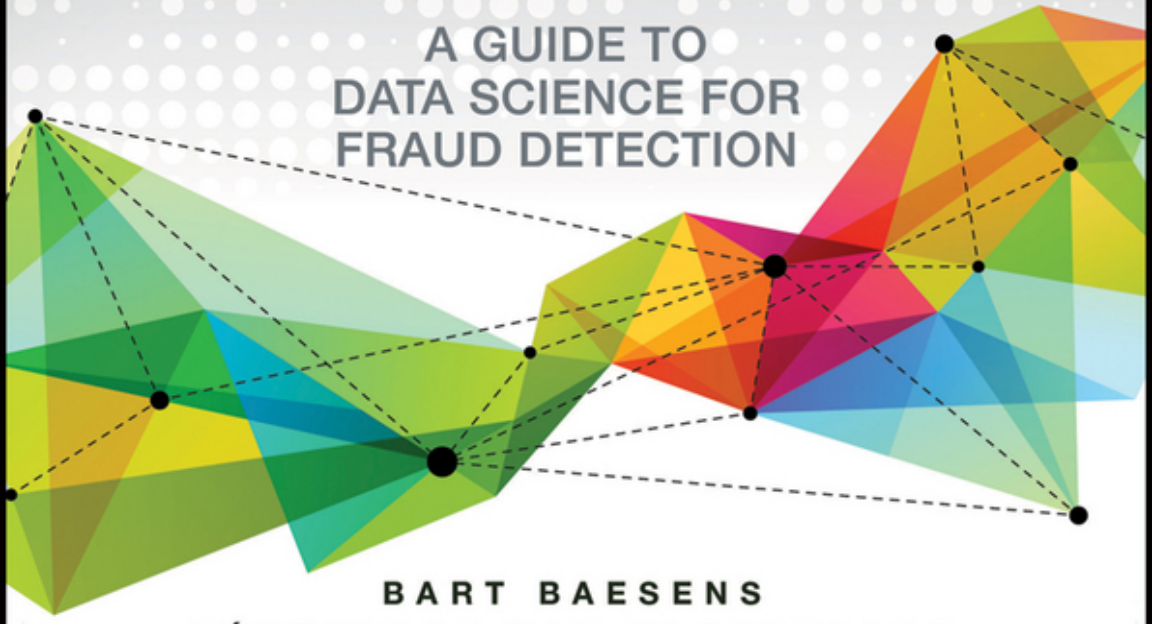
# FRAUD ANALYTICS

## USING DESCRIPTIVE, PREDICTIVE, AND SOCIAL NETWORK TECHNIQUES

### A GUIDE TO DATA SCIENCE FOR FRAUD DETECTION



BART BAESENS
VÉRONIQUE VAN VLASSELAER
WOUTER VERBEKE

WILEY

# Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques

# Wiley & SAS Business Series

The Wiley & SAS Business Series presents books that help senior-level managers with their critical management decisions.

Titles in the Wiley & SAS Business Series include:

*Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications* by Bart Baesens

*Bank Fraud: Using Technology to Combat Losses* by Revathi Subramanian

*Big Data Analytics: Turning Big Data into Big Money* by Frank Ohlhorst

*Big Data, Big Innovation: Enabling Competitive Differentiation through Business Analytics* by Evan Stubbs

*Business Analytics for Customer Intelligence* by Gert Laursen

*Business Intelligence Applied: Implementing an Effective Information and Communications Technology Infrastructure* by Michael Gendron

*Business Intelligence and the Cloud: Strategic Implementation Guide* by Michael S. Gendron

*Business Transformation: A Roadmap for Maximizing Organizational Insights* by Aiman Zeid

*Connecting Organizational Silos: Taking Knowledge Flow Management to the Next Level with Social Media* by Frank Leistner

*Data-Driven Healthcare: How Analytics and BI Are Transforming the Industry* by Laura Madsen

*Delivering Business Analytics: Practical Guidelines for Best Practice* by Evan Stubbs

*Demand-Driven Forecasting: A Structured Approach to Forecasting,* second edition by Charles Chase

*Demand-Driven Inventory Optimization and Replenishment: Creating a More Efficient Supply Chain* by Robert A. Davis

*Developing Human Capital: Using Analytics to Plan and Optimize Your Learning and Development Investments* by Gene Pease, Barbara Beresford, and Lew Walker

*The Executive's Guide to Enterprise Social Media Strategy: How Social Networks Are Radically Transforming Your Business* by David Thomas and Mike Barlow

*Economic and Business Forecasting: Analyzing and Interpreting Econometric Results* by John Silvia, Azhar Iqbal, Kaylyn Swankoski, Sarah Watt, and Sam Bullard

*Financial Institution Advantage and The Optimization of Information Processing* by Sean C. Keenan

*Foreign Currency Financial Reporting from Euros to Yen to Yuan: A Guide to Fundamental Concepts and Practical Applications* by Robert Rowan

*Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data Driven Models* by Keith Holdaway

*Health Analytics: Gaining the Insights to Transform Health Care* by Jason Burke

*Heuristics in Analytics: A Practical Perspective of What Influences Our Analytical World* by Carlos Andre Reis Pinheiro and Fiona McNeill

*Human Capital Analytics: How to Harness the Potential of Your Organization's Greatest Asset* by Gene Pease, Boyce Byerly, and Jac Fitz-enz

*Implement, Improve and Expand Your Statewide Longitudinal Data System: Creating a Culture of Data in Education* by Jamie McQuiggan and Armistead Sapp

*Killer Analytics: Top 20 Metrics Missing from Your Balance Sheet* by Mark Brown

*Predictive Analytics for Human Resources* by Jac Fitz-enz and John Mattox II

*Predictive Business Analytics: Forward-Looking Capabilities to Improve Business Performance* by Lawrence Maisel and Gary Cokins

*Retail Analytics: The Secret Weapon* by Emmett Cox

*Social Network Analysis in Telecommunications* by Carlos Andre Reis Pinheiro

*Statistical Thinking: Improving Business Performance*, second edition by Roger W. Hoerl and Ronald D. Snee

*Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics* by Bill Franks

*Too Big to Ignore: The Business Case for Big Data* by Phil Simon

*The Value of Business Analytics: Identifying the Path to Profitability* by Evan Stubbs

*The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions* by Phil Simon

*Understanding the Predictive Analytics Lifecycle* by Al Cordoba

*Unleashing Your Inner Leader: An Executive Coach Tells All* by Vickie Bevenour

*Using Big Data Analytics: Turning Big Data into Big Money* by Jared Dean

*Win with Advanced Business Analytics: Creating Business Value from Your Data* by Jean Paul Isson and Jesse Harriott

For more information on these and other titles in the series, please visit www.wiley.com.

# Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques

*A Guide to Data Science for Fraud Detection*

**Bart Baesens**
**Véronique Van Vlasselaer**
**Wouter Verbeke**

WILEY

Copyright © 2015 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the Web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at http://booksupport.wiley.com. For more information about Wiley products, visit www.wiley.com.

***Library of Congress Cataloging-in-Publication Data:***

Baesens, Bart.
Fraud analytics using descriptive, predictive, and social network techniques : a guide to data science for fraud detection / Bart Baesens, Veronique Van Vlasselaer, Wouter Verbeke.
pages cm. — (Wiley & SAS business series)
Includes bibliographical references and index.
ISBN 978-1-119-13312-4 (cloth) — ISBN 978-1-119-14682-7 (epdf) — ISBN 978-1-119-14683-4 (epub)
1. Fraud—Statistical methods. 2. Fraud—Prevention. 3. Commercial crimes—Prevention. I. Title.
HV6691.B34 2015
364.16' 3015195—dc23

2015017861

Cover Design: Wiley
Cover Image: ©iStock.com/aleksandarvelasevic

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

*To my wonderful wife, Katrien, and kids, Ann-Sophie, Victor, and Hannelore.*

*To my parents and parents-in-law.*

*To my husband and soul mate, Niels, for his never-ending support.*

*To my parents, parents-in-law, and siblings-in-law.*

*To Luit and Titus.*

# Contents

# List of Figures

# Foreword

Fraud will always be with us. It is linked both to organized crime and to terrorism, and it inflicts substantial economic damage. The perpetrators of fraud play a dynamic cat and mouse game with those trying to stop them. Preventing a particular kind of fraud does not mean the fraudsters give up, but merely that they change their tactics: they are constantly on the lookout for new avenues for fraud, for new weaknesses in the system. And given that our social and financial systems are forever developing, there are always new opportunities to be exploited.

This book is a clear and comprehensive outline of the current state-of-the-art in fraud-detection and prevention methodology. It describes the data necessary to detect fraud, and then takes the reader from the basics of fraud-detection data analytics, through advanced pattern recognition methodology, to cutting-edge social network analysis and fraud ring detection.

If we cannot stop fraud altogether, an awareness of the contents of this book will at least enable readers to reduce the extent of fraud, and make it harder for criminals to take advantage of the honest. The readers' organizations, be they public or private, will be better protected if they implement the strategies described in this book. In short, this book is a valuable contribution to the well-being of society and of the people within it.

Professor David J. Hand
Imperial College, London

# Preface

It is estimated that a typical organization loses about 5 percent of its revenues due to fraud each year. In this book, we will discuss how state-of-the-art descriptive, predictive and social network analytics can be used to fight fraud by learning fraud patterns from historical data.

The focus of this book is not on the mathematics or theory, but on the practical applications. Formulas and equations will only be included when absolutely needed from a practitioner's perspective. It is also not our aim to provide exhaustive coverage of all analytical techniques previously developed but, rather, give coverage of the ones that really provide added value in a practical fraud detection setting.

Being targeted at the business professional in the first place, the book is written in a condensed, focused way. Prerequisite knowledge consists of some basic exposure to descriptive statistics (e.g., mean, standard deviation, correlation, confidence intervals, hypothesis testing), data handling (using for example, Microsoft Excel, SQL, etc.), and data visualization (e.g., bar plots, pie charts, histograms, scatter plots, etc.). Throughout the discussion, many examples of real-life fraud applications will be included in, for example, insurance fraud, tax evasion fraud, and credit card fraud. The authors will also integrate both their research and consulting experience throughout the various chapters. The book is aimed at (senior) data analysts, (aspiring) data scientists, consultants, analytics practitioners, and researchers (e.g., PhD candidates) starting to explore the field.

Chapter 1 sets the stage on fraud detection, prevention, and analytics. It starts by defining fraud and then zooms into fraud detection and prevention. The impact of big data for fraud detection and the fraud analytics process model are reviewed next. The chapter concludes by summarizing the key skills of a fraud data scientist.

Chapter 2 provides extensive discussion on the basic ingredient of any fraud analytical model: data! It introduces various types of

data sources and discusses how to merge and sample them. The next sections discuss the different types of data elements, visual exploration, Benford's law, and descriptive statistics. These are all essential tools to start understanding the characteristics and limitations of the data available. Data preprocessing activities are also extensively covered: handling missing values, detecting and treating outliers, defining red flags, standardizing data, categorizing variables, weights of evidence coding, and variable selection. Principal component analysis is outlined as a technique to reduce the dimensionality of the input data. This is then further illustrated with RIDIT and PRIDIT analysis. The chapter ends by reviewing segmentation and the risks thereof.

Chapter 3 continues by exploring the use of descriptive analytics for fraud detection. The idea here is to look for unusual patterns or outliers in a fraud data set. Both graphical and statistical outlier detection procedures are reviewed first. This is followed by an overview of break-point analysis, peer group analysis, association rules, clustering, and one-class SVMs.

Chapter 4 zooms into predictive analytics for fraud detection. We start from a labeled data set of transactions whereby each transaction has a target of interest that can either be binary (e.g., fraudulent or not) or continuous (e.g., amount of fraud). We then discuss various analytical techniques to build predictive models: linear regression, logistic regression, decision trees, neural networks, support vector machines, ensemble methods, and multiclass classification techniques. A next section reviews how to measure the performance of a predictive analytical model by first deciding on the data set split-up and then on the performance metric. The class imbalance problem is also extensively elaborated. The chapter concludes by giving some performance benchmarks.

Chapter 5 introduces the reader to social network analysis and its use for fraud detection. Stating that the propensity to fraud is often influenced by the social neighborhood, we describe the main components of a network and illustrate how transactional data sources can be transformed in networks. In the next section, we elaborate on featurization, the process on how to extract a set of meaningful features from the network. We distinguish between three main types of features: neighborhood metrics, centrality metrics, and collective

inference algorithms. We then zoom into community mining, where we aim at finding groups of fraudsters closely connected in the network. By introducing multipartite graphs, we address the fact that fraud often depends on a multitude of different factors and that the inclusion of all these factors in a network representation contribute to a better understanding and analysis of the detection problem at hand. The chapter is concluded with a real-life example of social security fraud.

Chapter 6 deals with the postprocessing of fraud analytical models. It starts by giving an overview of the analytical fraud model lifecycle. It then discusses the traffic light indicator approach and decision tables as two popular model representations. This is followed by a set of guidelines to appropriately select the fraud sample to investigate. Fraud alert and case management are covered next. We also illustrate how visual analytics can contribute to the postprocessing activities. We describe how to backtest analytical fraud models by considering data stability, model stability, and model calibration. The chapter concludes by giving some guidelines about model design and documentation.

Chapter 7 provides a broader perspective on fraud analytics. We provide some guidelines for setting up and managing data quality programs. We zoom into privacy and discuss various ways to ensure appropriate access to both internal and external data. We discuss how analytical fraud estimates can be used to calculate both expected and unexpected losses, which can then help to determine provisioning and capital buffers. A discussion of total cost of ownership and return on investment provides an economic perspective on fraud analytics. This is followed by a discussion of in- versus outsourcing of analytical model development. We briefly zoom into some interesting modeling extensions, such as forecasting and text analytics. The potential and danger of the Internet of Things for fraud analytics is also covered. The chapter concludes by giving some recommendations for corporate fraud governance.