

Jurij Weinblat

Prediction of highly lucrative companies using annual statements

A Data Mining based approach



Anchor Academic Publishing

disseminate knowledge

Weinblat, Jurij: Prediction of highly lucrative companies using annual statements: A Data Mining based approach, Hamburg, Anchor Academic Publishing 2015

Buch-ISBN: 978-3-95489-304-1

PDF-eBook-ISBN: 978-3-95489-804-6

Druck/Herstellung: Anchor Academic Publishing, Hamburg, 2015

Bibliografische Information der Deutschen Nationalbibliothek:

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Bibliographical Information of the German National Library:

The German National Library lists this publication in the German National Bibliography. Detailed bibliographic data can be found at: <http://dnb.d-nb.de>

All rights reserved. This publication may not be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Bearbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Die Informationen in diesem Werk wurden mit Sorgfalt erarbeitet. Dennoch können Fehler nicht vollständig ausgeschlossen werden und die Diplomica Verlag GmbH, die Autoren oder Übersetzer übernehmen keine juristische Verantwortung oder irgendeine Haftung für evtl. verbliebene fehlerhafte Angaben und deren Folgen.

Alle Rechte vorbehalten

© Anchor Academic Publishing, Imprint der Diplomica Verlag GmbH
Hermannstal 119k, 22119 Hamburg
<http://www.diplomica-verlag.de>, Hamburg 2015
Printed in Germany

Table of contents

LIST OF ILLUSTRATIONS.....	VII
LIST OF TABLES.....	VIII
LIST OF ABBREVIATIONS.....	IX
ACKNOWLEDGMENT.....	X
1. INTRODUCTION AND PROBLEM DESCRIPTION.....	11
1.1 INTENTION OF THIS STUDY.....	12
1.2 PROCEEDING.....	13
2. INTRODUCTION TO KEY FIGURE ANALYSIS.....	15
2.1 THE PRINCIPLE OF KEY FIGURES.....	15
2.2 THE CLASSICAL KEY FIGURE ANALYSIS APPROACH.....	15
2.3 MODERN KEY FIGURE ANALYSIS APPROACHES.....	16
2.4 LIMITATIONS OF ANNUAL REPORT ANALYSIS.....	19
3. THE AVAILABLE DATASET.....	21
3.1 DESCRIPTION OF THE DATASET.....	21
3.2 DATA CLEAN-UP.....	22
4. KEY FIGURE SELECTION.....	24
4.1 SIGNIFICANT KEY FIGURE REQUIREMENTS.....	24
4.2 THE SELECTED KEY FIGURES OF THIS ANALYSIS.....	25
4.2.1 Selected class variable.....	26
4.2.2 Selected qualitative key figures.....	28
4.2.3 Selected absolute key figures.....	30
4.2.4 Selected relative key figures.....	33
4.3 CLASS ANALYSIS.....	38
5. CLASSIFICATION TREES AND FORESTS.....	41
5.1 PRECONSIDERATIONS.....	41
5.2 CLASSIFICATION TREES.....	42
5.2.1 A simple example.....	43
5.2.2 Generation of classification trees.....	44
5.2.3 Pruning an existing tree.....	48
5.2.4 Relevant properties of CART trees.....	52
5.3 RANDOM FOREST.....	54
5.3.1 Classification process of a random forest.....	54
5.3.2 Generation of random forest.....	55
5.3.3 Relevant properties of random forests.....	57

6.	CLASSIFICATION RESULTS.....	59
6.1	CLASSIFICATION TREE RESULTS.....	59
6.1.1	Examination of the most precise tree	63
6.1.2	Key indicator importance ranking.....	63
6.1.3	Transfer to data from 2011	65
6.2	CLASSIFICATION FOREST RESULTS.....	68
6.2.1	Transfer to data from 2011	69
6.2.2	Key indicator importance ranking.....	74
7.	CONCLUSION	76
7.1	CRITICAL ASSESSMENT	76
7.2	OUTLOOK	77
	BIBLIOGRAPHY	79
	APPENDIX.....	86
1.	THIS STUDY’S PROCEDURE MODEL	86
1.1	DEFINITIONS OF “REFERENCE MODEL” AND “PROCESS MODEL”	87
1.2	THE CRISP-DM REFERENCE MODEL	88
1.2.1	The six phases of CRISP-DM	90
1.2.2	Assessment of CRISP-DM.....	93
2.	DATA EXTRACTION.....	95
3.	CLASS COMPARISON DIAGRAMS.....	97

List of illustrations

ILLUSTRATION 1: STRUCTURE OF THE AVAILABLE CSV-FILE.....	21
ILLUSTRATION 2: BOXPLOTS OF THE ABSOLUTE KEY FIGURES.....	31
ILLUSTRATION 3: PROPORTION OF CORPORATE FAILURE IN 1996-1997.....	32
ILLUSTRATION 4: BOXPLOTS OF THE RELATIVE KEY FIGURES.....	36
ILLUSTRATION 5: OCCURRENCE OF NAs AMONG THE KEY FIGURES.....	38
ILLUSTRATION 6: TRANSITION OF THE CLASS VARIABLE BETWEEN 2010 AND 2011.....	40
ILLUSTRATION 7: LUCRATIVENESS IN 2010 AND 2011 CONSIDERING NAs.....	40
ILLUSTRATION 8: CLASSIFICATION TREE OF THE EXAMPLE DATASET.....	43
ILLUSTRATION 9: TWO POSSIBLE SPLITS OF THE EXAMPLE DATASETS.....	46
ILLUSTRATION 10: OVERFITTING EXAMPLE.....	48
ILLUSTRATION 11: EXAMPLES'S VARIABLE IMPORTANCE VALUES OF THE TREE.....	54
ILLUSTRATION 12: RANDOM FOREST FOR THE EXAMPLE CONSISTING OF 100 TREES.....	55
ILLUSTRATION 13: EXAMPLE'S VARIABLE IMPORTANCE VALUES OF THE RANDOM FOREST.....	57
ILLUSTRATION 14: FALSE POSITIVE FALLACY.....	62
ILLUSTRATION 15: UNDERSAMPLED CLASSIFICATION TREE.....	63
ILLUSTRATION 16: CLASSIFICATION TREE'S IMPORTANCE RANKING.....	64
ILLUSTRATION 17: PRECISION VALUES FOR DIFFERENT MINIMUM VOTE COUNTS.....	71
ILLUSTRATION 18: FACTOR OF SHARE IMPROVEMENT FOR DIFFERENT MINIMUM VOTE COUNTS.....	72
ILLUSTRATION 19: INSTANCE NUMBERS FOR DIFFERENT VOTE LIMITS.....	73
ILLUSTRATION 20: IMPORTANCE RANKINGS OF THE CLASSIFICATION TREES AND FORESTS.....	74
ILLUSTRATION 21: THE SIX CRISP-DM PHASES.....	89
ILLUSTRATION 22: CLASS COMPARISON BOXPLOTS 1-8.....	97
ILLUSTRATION 23: CLASS COMPARISON BOXPLOTS 8-16.....	98
ILLUSTRATION 24: CLASS COMPARISON BAR DIAGRAM OF NATIONAL_LEGAL_FORM.....	98
ILLUSTRATION 25: CLASS COMPARISON BAR DIAGRAM OF LEGAL_FORM.....	99
ILLUSTRATION 26: CLASS COMPARISON BAR DIAGRAM OF COMPANY_INDEPENDENCE.....	99

List of tables

TABLE 1: IMPORTANT PARAMETERS OF THE CLASS VARIABLE	27
TABLE 2: SELECTED QUALITATIVE KEY FIGURES	29
TABLE 3: SELECTED ABSOLUTE KEY FIGURES	31
TABLE 4: SELECTED RELATIVE KEY FIGURES	36
TABLE 5: FOUR ENTRIES OF THE EXAMPLE DATASET.....	43
TABLE 6: CROSS VALIDATED CLASSIFICATION TREE RESULTS FOR 2010.....	62
TABLE 7: RPART'S PREDICTION OF LUCRATIVENESS FOR 2011.....	65
TABLE 8: RESULTS FOR THE REDUCED DATASET	67
TABLE 9: CROSS VALIDATED CLASSIFICATION FOREST RESULTS FOR 2010.....	68
TABLE 10: RANDOM FOREST'S PREDICTION OF LUCRATIVENESS FOR 2011	69
TABLE 11: RESULTS OF THE MEASURES TO IMPROVE THE PRECISION OF THE FORESTS.....	70

List of abbreviations

BvD	Bureau van Dijk Electronic Publishing GmbH
CART	Classification and Regression Trees
csv-file	Comma-separated-values-file
CV	Cross validation
DM	Data Mining
FN	False negative
FP	False positive
IQR.....	Interquartile range
NA.....	Not available
RF.....	Random forest
RM	Reference model
ROE	Return on equity
SQL	Structured Query Language
TN	True negative
TP	True positive

Acknowledgment

I would also like to thank my supervisor Prof. Dr. Andreas Behr for assisting me with this book. He provided me with valuable suggestions and gave me the opportunity to write about my favourite topic. Thank you very much!

I would also like to say thank you to my parents and my girlfriend Sarah for their support during my entire work.

1. Introduction and problem description

In literature, a lot of scientists describe how to use annual report data to predict whether a certain company is going to become bankrupt (Dimitras, Zanakis und Zopounidis 1996, 487–513). The reasons why this topic attracts such a high degree of scientific attention is rather obvious: The stability of the financial system depends on the ability of banks and other financial service providers to assess whether a certain firm will be able repay a loan or not. Furthermore, banks need this information to be able to calculate an adequate probability of default to identify a minimum interest rate for a concrete loan (Moro und Schäfer 2004).

Nevertheless, it is not only relevant to anticipate this worst case of bankruptcy, but also whether a regarded small firm will grow extraordinary in the next year and maybe even become a big company in the medium term. This is crucial information for private investors and fund managers who need to decide whether they should invest in a certain firm. Companies like Apple and Amazon have shown in the past that people who recognized the potential of such companies and bought their shares have earned a lot of money.

The prediction models, which are described in this paper, can also be used by politicians to identify companies which are eligible for funding. Because growing companies oftentimes hire many employees, it might be meaningful to facilitate their development process by selective subsidies to reduce unemployment. Furthermore, it is possible to question the prediction results of a financial analyst if he came to a different conclusion than a model.

Since annual reports are often publically available for free, it is reasonable to take advantage of them for such a prediction (Gräfer 1988, 52). Additionally, various information providers maintain huge databases with annual reports. A big data approach promises to further improve accuracy of predictions (Rauscher und Rockel 2001, 5). This paper introduces methods, which enable to generate knowledge out of these huge data sources to identify extraordinary lucrative firms.

To generate these prediction models, a data mining approach is used which is based on the approved CRISP-DM proceeding model for data mining processes. CRISP-DM ensures comparability and the consideration of best practices (Chapman, et al. 2000, 1-2). The prediction models are based on classification trees and forests because they have some very substantial advantages

over other methods like neural networks, which are frequently used in literature. For instance, the underlying algorithms of the used model do not require a certain distributional assumption, accept both quantitative and qualitative inputs, and are not sensitive with respect to outliers. But the two most important advantages are that a tree can be easily interpreted by users which is important for the previously described stakeholders because it is not easy to trust the results of a model which one does not understand (Löbbe 2001, 199). This is why a lack of understanding might impede the practical implementation of such a model. Besides that, the used algorithms can handle missing data which occur very often in the available dataset. In other analysis, these data entries would have been removed even if only one value is missing. This reduces the often already relatively small amount of available data and can reduce the model's accuracy (Neeb 2011, 67, Franken 2007, 5). This is not the case for the applied methods.

1.1 Intention of this study

The intention of this paper is to determine whether a stakeholder can use a classification tree or classification forest at the beginning of one year to identify German firms which will grow exceptionally in this year using annual reports' key figures from previous years. As a first step, key figures from the years 2007, 2008 and 2009 are used to generate different trees and forests which can predict whether a company grows outstandingly in 2010 or not. These models require the lucrativeness information from 2010 to be generated. To evaluate how well these unchanged models would work for the mentioned stakeholder at the beginning of the year 2011, they are also applied to data from 2008, 2009 and 2010 as a second step. This means that this time, the models are applied to more recent data to anticipate whether the regarded firms will grow intensively in 2011. Data from 2011 is only used to check the predictions' correctness and not to generate models. The best identified models are also compared and analysed.

These four particular years have been chosen because the available dataset "Amadeus" only contains a relatively small amount of more recent data. It is probably not necessary to regard more than three years for the generation of these models because it is shown in literature that this data is not able to noticeably improve prediction (Pytlik 1994, 94).