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

Table of contents

List of illustrations

List of tables

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).