Marcus Baulig

Machine Learning in Statistics

Forecasting Applications in Corporate Finance

SHAKER
VERLAG

Machine Learning in Statistics – Forecasting Applications in Corporate Finance

Dissertation
zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft
(doctor rerum politicarum)
der Fakultät für Empirische Humanwissenschaften und
Wirtschaftswissenschaft der Universität des Saarlandes

vorgelegt von

Marcus Baulig

Tag der Disputation:	14. Juni 2019
Dekan:	Univ.-Prof. Dr. Stefan Strohmeier
Erstberichterstatter:	Univ.-Prof. Dr. Ashok Kaul
Zweitberichterstatter:	Univ.-Prof. Dr. Dieter Hess

Berichte aus der Statistik

Marcus Baulig

Machine Learning in Statistics

Forecasting Applications in Corporate Finance

Shaker Verlag
Düren 2019

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: Saarbrücken, Univ., Diss., 2019

Printed in Germany.

ISBN 978-3-8440-6820-7
ISSN 1619-0963

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren
Phone: 0049/2421/99011-0 • Telefax: 0049/2421/99011-9
Internet: www.shaker.de • e-mail: info@shaker.de

Abstract

Given the growing awareness of machine learning outside of computer science in both academia and business in recent years, I examine two corporate financing forecasting problems and show how machine learning models compare to established literature models. The forecasting of future corporate bankruptcies serves vicariously as a classification problem and the forecasting of future corporate earnings as a regression problem, which are both traditionally approached with econometric techniques such as logistic and linear regression.

I forecast bankruptcies and earnings using a diverse set of machine learning models—ranging from subset selection models to highly flexible Boosting models—and combine these models to stacked ensembles. Ensemble learning can potentially outstrip the performance of individual machine learning models and to date has not been investigated for bankruptcy and earnings forecasts. Besides the focus on high performing models, I create highly interpretable logistic and linear regression models with the most predictive variables assessed over all machine learning models.

For both forecast problems, stacked ensembles show in their optimal calibration the best forecast performance, exceeding established literature models by at least 5% in terms of standard evaluation criteria. Individual machine learning models achieve at least an improvement of 4% and my self-generated regression models at least 1%.

Despite the already capable performance of established literature models in the field of corporate finance already, the utilization of machine learning successfully enriches the way researchers can transform data into forecasts. The engineering of new data with machine learning methods has the potential to be an even more influential approach for future researchers.

Acknowledgments

I would like to express my special gratitude to Ashok Kaul, my first advisor, for the numerous academic and professional opportunities he has provided that have pointed the way ahead for me. Among those opportunities, I wish to highlight my time abroad in China, the US, and South Africa, all of which have found their way into this thesis. I am also very grateful to Dieter Hess, my second advisor, for his outstanding supervision from the very beginning of my PhD studies.

My colleagues Christian Agethen, Nathalie Neu-Yanders, and Manuel Schieler provided me not only with inspiring academic ideas, but also with a very pleasurable atmosphere at the chair. I truly appreciate you all and the time we have spent together. A special thanks also goes to Andrew Yanders for his proofreading of my thesis.

Finally, I would like to thank everyone who has accompanied me on a personal level these last years: most of all my parents, Felix Malzahn, Matthias Schindler, and Jani Coetzee.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AMEX	American Stock Exchange
ARIMA	Autoregressive Integrated Moving Average (Model)
AUC	Area under the (ROC) Curve
BOOST.GEN	Generalized Boosting
BOOST.XG	Extreme Gradient Boosting
BS	Breir-Score
BSS	Best Subset Selection
BSTEP	Backward Stepwise Selection
CV	Cross-validation
CRSP	Center for Research in Security Prices
D	Cross-deviance
Diff.	Difference
ENET	Elastic Net Regression
EP	Earnings Persistence Model (Li and Mohanram, 2014)
FPR	False Positive Rate
FSTEP	Forward Stepwise Selection
GAM	Generalized Additive Models
GDP	Gross Domestic Product
HVZ	Hou et al. (2012)
iid	Independent and identically distributed
Ind.-Spec.	Industry-Specific
KNN	K-Nearest Neighbors
Lasso	Least Absolute Shrinkage and Selection Operator (Lasso Regression)
LCI	Lower Confidence Interval
LDA	Linear Discriminant Analysis
LOG	Logistic Regression
LOOCV	Leave-One-Out Cross-Validation
LM	Li and Mohanram (2014)

MARS	Multiple Adaptive Regression Splines
MSE	Mean Squared Error
NASDAQ	National Association of Securities Dealers Automated Quotations
NN	Neural Networks
NYSE	New York Stock Exchange
OLS	Ordinary Least Squares Regression
PS	Per Share
RF	Random Forests
Ridge	Ridge Regression
RIM	Residual Income Model (Li and Mohanram, 2014)
ROC	Receiver Operating Characteristics (Curve)
RSS	Residual Sum of Squares
SVM	Support Vector Machine
TPR	True Positive Rate
TREE	Decision Tree
UCI	Upper Confidence Interval
VAR	Vector Autoregressive (Model)
VIM	Variable Importance Model
VIMnl	Variable Importance Model (non-linear)
WRDS	Wharton Research Data Service