



Yang Liu (Autor)

A theoretical and empirical study on the data mining process for credit scoring



Göttinger Wirtschaftsinformatik
Herausgeber: J. Biethahn · M. Schumann

Yang Liu

**A theoretical and empirical study
on the data mining process for credit scoring**

Band 42



Cuvillier Verlag Göttingen

<https://cuvillier.de/de/shop/publications/3168>

Copyright:

Cuvillier Verlag, Inhaberin Annette Jentsch-Cuvillier, Nonnenstieg 8, 37075 Göttingen,
Germany

Telefon: +49 (0)551 54724-0, E-Mail: info@cuvillier.de, Website: <https://cuvillier.de>

Contents

Figures -----	XI
Tables -----	XV
Abbreviations -----	XVII
1 Introduction -----	1
1.1 Motivations and aims -----	1
1.2 Organization of the dissertation-----	5
2 Background knowledge of credit scoring -----	7
2.1 Definition of credit scoring -----	7
2.1.1 Deductive and empirical credit scoring -----	7
2.1.2 Credit scoring and credit rating -----	10
2.2 The expanding application range of scoring techniques -----	16
2.2.1 Application areas -----	16
2.2.2 Business objectives -----	19
2.3 Data issues in credit scoring -----	21
2.3.1 Information for credit scoring-----	21
2.3.2 The time horizon of the sample data -----	25
2.3.3 Reject inference-----	28
2.3.4 The definition of the risk classes -----	30
2.4 Application issues in credit scoring-----	33
2.4.1 Practical uses of credit scoring models-----	33
2.4.2 Pros and cons in the practical application-----	35
3 Background knowledge of data mining and classification techniques -----	37
3.1 Introduction of data mining approach -----	37
3.1.1 Definitions of data mining -----	37
3.1.2 The general process of data mining-----	38
3.1.3 The application architecture of data mining-----	41
3.2 Classification decisions -----	42
3.2.1 Two ways of computer-based classification decisions -----	42
3.2.2 Classification techniques -----	45
3.2.2.1 Description of the classification procedure-----	45
3.2.2.2 Reasons for inaccurate classifications-----	46
3.2.3 Various groups of classification techniques -----	49

3.2.3.1	The professional backgrounds	49
3.2.3.2	The approaches of model generations	50
3.3	Classification algorithms for credit scoring	52
3.3.1	Discriminant Analysis: Bayesian Linear Discriminant Analysis	53
3.3.1.1	Basic Bayes' classification rule:	54
3.3.1.2	Linear Discriminant Analysis	55
3.3.2	Logistic Regression	57
3.3.3	Instance-Based Learning: k-Nearest-Neighbors algorithm	58
3.3.4	Model Trees: M5	60
3.3.5	Neural Networks: Multi-layer perceptron	64
3.3.6	Comparisons of the introduced algorithms	66
4	A framework of the data mining application process for credit scoring	71
4.1	Reasons for a process framework	71
4.2	Presentation of the general framework	72
4.2.1	The stage of the problem definition and data preparation	73
4.2.2	The stage of the data analysis and model building	75
4.2.3	The stage of the model application and validation	78
4.3	The process of empirical studies in later chapters	81
5	The evaluation subprocess	83
5.1	Reasons of the model evaluation	83
5.2	Criteria of the model evaluation	86
5.2.1	The classification accuracy	86
5.2.1.1	The estimation of true error rates	86
5.2.1.2	The validity of the sample data for error rate estimation	87
5.2.1.2.1	The independence between the train and test data	87
5.2.1.2.2	Sampling for population drift	89
5.2.1.2.3	The size of the test data	91
5.2.2	Practicability criteria	96
5.3	Methods of the model evaluation	98
5.3.1	Confusion matrix and two types of errors	98
5.3.2	The tradeoff of two types of error rates	100
5.3.2.1	ROC Curve	100
5.3.2.2	Cost function	103
5.3.2.3	The determination of the threshold value	105

5.3.3	Learning curve and incremental analysis	108
5.3.3.1	Learning curve	108
5.3.3.2	Incremental case analysis	109
5.3.3.3	Incremental complexity analysis	110
5.3.3.4	The basic phenomenology of incremental analysis	111
5.4	The empirical study of the credit scoring model evaluation	112
5.4.1	Problem definition and data preparation	113
5.4.1.1	Defining the problem	113
5.4.1.2	Defining the risk classes	115
5.4.1.3	Preparation of the data	118
5.4.2	The process of the evaluation	120
5.4.2.1	Incremental analysis	120
5.4.2.2	Comparison analysis	122
5.4.3	Results of the evaluation	124
5.4.3.1	Results of the incremental analysis	124
5.4.3.1.1	Linear discriminant analysis (LDA) and logistic regression (LR)	124
5.4.3.1.2	Neural networks (MLP)	125
5.4.3.1.3	K-nearest-neighbors (k-NN)	126
5.4.3.1.4	Model trees (M5)	127
5.4.3.2	Results of the comparison analysis	128
5.4.3.2.1	Classification accuracy	129
5.4.3.2.2	Practicability criteria	131
5.4.4	The practical usage of the scoring models	136
6	The Input-relevant subprocess	140
6.1	A general review of some tasks in the preprocessing of input data	140
6.1.1	Descriptive analysis	140
6.1.2	Transformation of measurement scales	141
6.1.2.1	Categorical to numeric	142
6.1.2.2	Numeric to categorical	145
6.1.3	Missing data treatment	146
6.2	Feature selection	148
6.2.1	Reasons of feature selection	148
6.2.2	Overview of feature selection methods	149
6.2.3	Machine learning feature selection methods	152

6.2.3.1	Feature ranking algorithms and best feature subset algorithms	152
6.2.3.2	Filter algorithms and wrapper algorithms	156
6.3	The empirical study of feature selection	157
6.3.1	Description of the problem and the data	157
6.3.2	The performance of models without feature selection	161
6.3.3	Data analysis with feature selection methods	163
6.3.3.1	Data analysis with “ReliefF” algorithm	163
6.3.3.1.1	Introduction of the algorithm	163
6.3.3.1.2	The selection of model parameters	164
6.3.3.1.3	The feature selection results	167
6.3.3.2	Data analysis with the correlation-based algorithm	170
6.3.3.2.1	Introduction of the algorithm	170
6.3.3.2.2	The feature selection results	171
6.3.3.3	Data analysis with the consistency-based algorithm	173
6.3.3.3.1	Introduction of the algorithm	173
6.3.3.3.2	The feature selection results	174
6.3.3.4	Data analysis with a wrapper algorithm	177
6.3.3.5	Comparison of the results	178
7	The model-relevant subprocess	183
7.1	Overview of various model combination techniques	183
7.1.1	The process of model combination	183
7.1.2	Theoretical basis of model combination techniques	185
7.2	Description of the idea of stacking technique	186
7.3	An empirical study of combining credit scoring models	188
7.3.1	The description of the data	188
7.3.2	Stacking models with different algorithms	190
7.3.3	Stacking models with different features and different algorithms	193
8	Summaries and prospects	196
Appendices		199
Appendix A.	Explanations of some variables	199
Appendix B.	Brief summary of software and toolsets used in this dissertation	200
Literature		201