

Deep Credit Risk

Machine Learning in Python

Daniel Rosch*
Harald Scheule^

*DANIEL ROSCH is a professor of business and holds the chair of statistics and risk management at the University of Regensburg, Germany.

+HARALD SCHEULE is a professor of finance at the University of Technology Sydney, Australia.

Contents

I	Principles of Data Learning	11
1	Deep Dive	12
1.1	Boost Your Practical Machine Learning Skills.....	12
1.2	Credit Risk Information.....	13
1.3	Hands-on Analysis.....	14
1.4	Basel, CECL, IFRS 9, DFAST, CCAR and Stress Tests.....	16
1.5	First Lessons from the COVID-19 Crisis.....	17
1.5.1	Credit Risk Analytics.....	17
1.5.2	Machine Learning.....	19
2	Python Literacy	20
2.1	Synopsis.....	20
2.2	Installation.....	20
2.2.1	Anaconda and IDEs.....	20
2.2.2	Packages.....	21
2.2.3	Coding Guidelines.....	22
2.3	First Look.....	23
2.4	Describing Data.....	24
2.5	Plotting.....	25
2.6	Generating New Variables.....	26
2.7	Transforming Variables.....	27
2.8	Subsetting Data.....	28
2.9	Resetting Indexes.....	30
2.10	Combining Data.....	30
2.10.1	Concatenating.....	30
2.10.2	Appending.....	31
2.10.3	Match Merging.....	31
2.10.4	Joining.....	32
2.11	Regression Models.....	32
2.12	numpy vs pandas.....	34
2.13	Packages and Basic Settings.....	36
2.14	Functions.....	36
2.14.1	versions.....	36
2.14.2	dataprep.....	37
2.14.3	woe.....	38
2.14.4	validation.....	38
2.14.5	resolutionbias.....	39
2.15	Sandbox Problems.....	39
3	Risk-Based Learning	40
3.1	Synopsis.....	40
3.2	Maximum-Likelihood Estimation.....	40
3.2.1	Example for Default Modeling.....	42
3.2.2	Practical Implementation.....	48
3.3	Bayesian Approaches.....	49
3.3.1	Distributions.....	50
3.3.2	Parameter Estimation.....	51

3.3.3	Example for Default Modeling	52
3.3.4	Analytic Computation	56
3.3.5	Markov-Chain-Monte-Carlo Simulation.....	58
3.4	Sandbox Problems.....	60
4	Machine Learning	61
4.1	Synopsis.....	61
4.2	Terminology.....	62
4.3	Cost Functions.....	62
4.4	Information and Entropy.....	63
4.5	Optimization: Gradient Descent.....	67
4.6	Learning and Validation.....	75
4.6.1	Train vs Test Split.....	75
4.6.2	Bias-Variance Tradeoff.....	75
4.6.3	Crossvalidation and Tuning.....	80
4.7	Practical Implementation.....	80
4.8	P-Value and ML Hacking.....	83
4.9	Sandbox Problems.....	83
II	Data Processing and Validation	84
5	Outcome Engineering	85
5.1	Synopsis.....	85
5.2	Outcomes.....	85
5.2.1	Default Events.....	86
5.2.2	Payoff Events.....	87
5.2.3	Loss Rates Given Default	87
5.2.4	Exposure Conversion Measures.....	88
5.3	Default Engineering.....	88
5.3.1	Time-Vintage-Age (TVA) Analysis.....	88
5.3.2	Multi-Lead Analysis	91
5.3.3	Multi-Period Analysis.....	92
5.4	LGD Engineering	94
5.4.1	Data Preliminaries.....	95
5.4.2	Resolution Period.....	95
5.4.3	Risk-Neutral LGD from Observed Workout Cash Flows.....	96
5.4.4	LGD Winsorizing.....	98
5.4.5	Indirect Workout Costs.....	99
5.4.6	LGD Discount Rates	99
5.4.7	Observed LGD.....	101
5.4.8	Resolution Bias	103
5.5	Sandbox Problems.....	108
6	Feature Engineering	109
6.1	Synopsis.....	109
6.2	Missing Feature Analysis.....	109
6.2.1	Option 1: Keeping Missing Values.....	110
6.2.2	Option 2: Deleting Missing Values.....	111
6.2.3	Option 3: Imputation of Missing Values.....	111

6.2.4	Which Option Should You Consider?.....	114
6.3	Feature Outlier Analysis.....	114
6.4	Scaling.....	116
6.4.1	Feature Ratios.....	116
6.4.2	Feature Scaling.....	116
6.5	Non-linear Feature Transformations.....	118
6.5.1	Option 1 : Polynomials	119
6.5.2	Option 2: Splines	120
6.5.3	Option 3: Categorization.....	121
6.5.4	Option 4: Weight-of-Evidence.....	122
6.5.5	Impact of Transformations on Fitted Default Rates.....	123
6.6	Feature Reduction by Aggregation.....	125
6.7	Feature Reduction by Clustering.....	128
6.7.1	Distance Measures.....	128
6.7.2	K-Means Clustering.....	129
6.7.3	Hierarchical Clustering	132
6.8	Feature Reduction by Principal Component Analysis.....	139
6.9	Function dataprep	142
6.10	Sandbox Problems.....	143
7	Feature Selection	144
7.1	Synopsis.....	144
7.2	Economic Feature Selection.....	144
7.2.1	Liquidity and Equity.....	145
7.2.2	Time, Vintage and Age.....	150
7.2.3	Environment.....	151
7.2.4	Function dataprep	151
7.3	Univariate Feature Selection	152
7.3.1	Means Test.....	152
7.3.2	F-Statistic.....	153
7.3.3	Association.....	154
7.3.4	WOE Correlations and Information Value	155
7.4	Model-based Feature Selection	157
7.4.1	Manual Selection.....	157
7.4.2	In (1) and Out (0) Selection.....	157
7.4.3	Regularization.....	159
7.5	Synthesis.....	162
7.6	Sandbox Questions.....	162
8	Validation	164
8.1	Synopsis.....	164
8.2	Qualitative Validation.....	164
8.3	Quantitative Validation.....	165
8.3.1	Basics.....	165
8.3.2	Backtesting as Part of Validation.....	167
8.3.3	Traffic Light Approach.....	168
8.4	Metrics for Discriminatory Power.....	168
8.4.1	Confusion Matrix.....	169
8.4.2	Accuracy Metrics	170
8.4.3	Classification Report.....	171

8.4.4	ROC Curve.....	172
8.4.5	Portfolio Dependence.....	177
8.5	Metrics for Calibration.....	180
8.5.1	Brier Score.....	180
8.5.2	OLS R-Squared	181
8.5.3	scikit-learn R-Squared.....	181
8.5.4	Binomial Test	182
8.5.5	Jeffrey’s Prior Test	184
8.5.6	Calibration Curve.....	186
8.5.7	Hosmer-Lemeshow.....	187
8.6	Metrics for Stability.....	188
8.7	Function validation	189
8.8	Other Outcomes.....	191
8.9	Validation Study.....	191
8.9.1	Data Preparation and Feature Engineering.....	191
8.9.2	Fitting of CandidateModels and Validation	192
8.9.3	Comparing ROC Curves Out-of-Time.....	196
8.9.4	Model Stability.....	197
8.9.5	Practical Recommendations.....	199
8.10	Sandbox Problems.....	200

III Default, Payoff, LGD and EAD Modeling 201

9	Default Modeling	202
9.1	Synopsis.....	202
9.2	Data Preparation.....	202
9.3	Random Credit Defaults and Expectations.....	202
9.4	Default Models for Probability of Defaults.....	204
9.4.1	Link Function.....	204
9.4.2	Logit/Probit Models and GLMs.....	206
9.4.3	GLM/Logit Model.....	207
9.4.4	GLM/Probit Model.....	208
9.4.5	Comparison GLM/Logit vs. GLM/Probit Model.....	209
9.4.6	Multivariate Interactions	211
9.4.7	Comprehensive Model	213
9.4.8	statsmodels vs. scikit-learn.....	215
9.5	Forecasting PDs.....	218
9.5.1	Training and Test Sample.....	218
9.5.2	Fitting for Training Sample.....	218
9.5.3	Crisis Prediction for Test Sample.....	219
9.6	Crisis PDs.....	221
9.6.1	Asymptotic Single Risk Factor	221
9.6.2	Point-in-Time PDs.....	224
9.6.3	Through-the-Cycle PDs.....	226
9.6.4	Basel Capital.....	228
9.7	Stress-Testing	229
9.7.1	EBA and FRB Stress-Testing.....	229
9.7.2	Scenario-Based Stress-Testing.....	230
9.7.3	Parameter-Based Stress-Testing.....	232

9.8	Low Default Portfolios.....	236
9.8.1	Independent Default Events.....	236
9.8.2	Margin of Conservatism.....	239
9.8.3	Dependent Defaults.....	240
9.8.4	Scaling to the Mean.....	242
9.9	Sandbox Problems.....	243
10	Payoff Modeling	244
10.1	Synopsis.....	244
10.2	Data Preparation.....	244
10.3	Payoff Models.....	244
10.3.1	IFRS 9.....	246
10.3.2	Selection Control	247
10.3.3	Other Models.....	251
10.4	Sandbox Problems.....	251
11	LGD Modeling	253
11.1	Synopsis.....	253
11.2	Data Preparation.....	253
11.3	Linear Regression.....	254
11.4	Transformed Linear Regression.....	258
11.5	Fractional Response Regression.....	260
11.6	Beta Regression.....	263
11.7	Forecasting LGDs.....	266
11.7.1	Training and Test Sample.....	266
11.7.2	Fitting for Training Sample.....	267
11.7.3	Crisis Prediction for Test Sample.....	268
11.8	Sandbox Problems.....	269
12	Exposure Modeling	270
12.1	Synopsis.....	270
12.2	Data Preparation.....	271
12.3	Non-distressed Exposures.....	271
12.3.1	Computation of Credit Conversion Measures.....	271
12.3.2	Fitting of Credit Conversion Measures	272
12.3.3	Fitting Exposures.....	275
12.4	Exposures at Default.....	276
12.4.1	Computation of Credit Conversion Measures.....	276
12.4.2	Fitting of Credit Conversion Measures	278
12.4.3	Fitting Exposures.....	280
12.5	Sandbox Problems.....	281
IV	Machine Learning for PD and LGD Forecasting	282
13	Standalone Techniques	283
13.1	Synopsis.....	283
13.2	Data Preparation.....	283
13.3	Logistic Regression	284
13.3.1	Classical Logistic Regression.....	284

13.3.2	Logistic Regression with Regularization	286
13.3.3	Logistic Regression with Regularization and Hyperparameter Tuning.....	288
13.4	K-Nearest Neighbors.....	296
13.4.1	Idea.....	296
13.4.2	Practical Implementation.....	302
13.4.3	Hyperparameter Tuning.....	304
13.5	Naive Bayes.....	306
13.5.1	Idea.....	306
13.5.2	Practical Implementation.....	307
13.6	Decision Trees.....	308
13.6.1	Idea.....	308
13.6.2	Practical Implementation.....	309
13.6.3	Hyperparameter Tuning.....	311
13.6.4	Visualization of Trees.....	313
13.7	Support Vector Machines.....	314
13.7.1	Idea.....	314
13.7.2	Practical Implementation.....	314
13.8	Synthesis.....	315
13.9	Sandbox Problems.....	316
14	Neural Networks and Deep Learning	317
14.1	Synopsis.....	317
14.2	Neural Networks and Deep Learning	317
14.2.1	Idea.....	317
14.2.2	Simple Network without Hidden Layer.....	318
14.2.3	Neural Network with Hidden Layers and Non-Linearity.....	323
14.2.4	Practical Implementation.....	326
14.3	Synthesis.....	333
14.4	Sandbox Problems.....	334
15	Ensemble Techniques	335
15.1	Synopsis.....	335
15.2	Data Preparation.....	335
15.3	Bagging.....	336
15.3.1	Idea.....	336
15.3.2	Practical Implementation.....	337
15.4	Boosting.....	338
15.4.1	Idea.....	338
15.4.2	Practical Implementation.....	339
15.5	Random Forests	340
15.5.1	Idea.....	340
15.5.2	Practical Implementation.....	340
15.5.3	Hyperparameter Tuning.....	341
15.6	Boosted Trees.....	343
15.6.1	Summary of Approaches.....	343
15.6.2	Adaptive Boosted Trees	344
15.6.3	Stochastic Gradient Boosting.....	345
15.6.4	Light GBM.....	346
15.7	Voting Classifier.....	347
15.8	Synthesis.....	348

15.9	Sandbox Problems.....	349
16	Machine Learning for LGD	350
16.1	Synopsis.....	350
16.2	Data Preparation.....	350
16.3	Regression.....	351
	16.3.1 Linear Regression.....	351
	16.3.2 Regression with Regularization.....	355
16.4	K-Nearest Neighbors.....	361
16.5	Decision Trees	364
16.6	Random Forests	368
16.7	Boosted Trees.....	371
	16.7.1 Adaptive Boosted Trees	371
	16.7.2 Light GBM.....	372
16.8	Support Vector Machine	373
16.9	Neural Networks.....	374
16.10	Voting Regressor.....	378
16.11	Synthesis.....	380
16.12	Sandbox Problems.....	381
V	Synthesis: Lifetime Modeling, IFRS 9/CECL, Loan Pricing and Credit Portfolio Risk	382
17	Multi-period Modeling	383
17.1	Synopsis.....	383
17.2	Outcomes by Age.....	383
17.3	Roll Rate Analysis.....	385
	17.3.1 Rating Classification Criteria.....	385
	17.3.2 Rating Class Formation	386
	17.3.3 Single-Period Rating Migrations.....	388
	17.3.4 Multi-Period Rating Migrations.....	390
	17.3.5 Multi-Period Cumulative Default Rates.....	390
	17.3.6 Multi-Period Marginal Default Rates.....	391
17.4	Age Feature Models.....	393
	17.4.1 PDsbyAge.....	393
	17.4.2 Multi-Period Forecasting of PDs.....	395
17.5	Survival Time Models.....	398
	17.5.1 Probability Density Function, Survival Probability and Hazard Rate.....	398
	17.5.2 Cross-Sectional Dataset.....	399
	17.5.3 Cox Proportional Hazard Model	400
17.6	Other Risk Measures.....	407
17.7	Sandbox Problems.....	407
18	Expected Credit Losses	408
18.1	Synopsis.....	408
18.2	Expected Loss Concepts.....	408
	18.2.1 Basel vs. CECL/IFRS 9.....	408
	18.2.2 One-Period ExpectedLosses.....	409

18.2.3	Lifetime Expected Losses	409
18.3	Credit Risk Modeling of Age and Time.....	410
18.3.1	Default Rates by Age and Time.....	411
18.3.2	PDs by Age and Time.....	411
18.4	Multi-period Forecasting of Time-Varying Features.....	414
18.4.1	Time-Varying Features.....	414
18.4.2	Time Series Tests.....	414
18.4.3	Vector Autoregressions.....	415
18.4.4	Model Fitting.....	417
18.4.5	Multi-period Forecasting.....	417
18.5	Multi-period Forecasting of PDs.....	420
18.6	Computing Expected Lifetime Loss.....	423
18.6.1	Expected Losses.....	425
18.6.2	Expected Lifetime Losses	425
18.7	IFRS 9 Significant Increase in Credit Risk	425
18.8	Loan Pricing and Other Economic Models.....	428
18.9	Sandbox Problems.....	432
19	Unexpected Credit Losses	434
19.1	Synopsis.....	434
19.2	Unexpected Loss or Credit-Value-at-Risk	434
19.3	Basel Calibrations.....	435
19.4	Asset Correlation.....	439
19.4.1	Probit-Linear Model without Features.....	439
19.4.2	Probit-Linear Regression with Features.....	443
19.5	Credit Portfolio Loss Distributions.....	445
19.5.1	Infinitely Granular Portfolio.....	445
19.5.2	Limited Granularity: Numerical Integration.....	448
19.5.3	Limited Granularity: Monte-Carlo Simulation.....	451
19.5.4	Comparison of Approaches.....	454
19.6	Applications.....	455
19.6.1	Expected Loss.....	456
19.6.2	Value-at-Risk.....	456
19.6.3	Expected Shortfall.....	457
19.7	Sandbox Problems.....	459
20	Outlook	460
20.1	Where Do We Stand Today?.....	460
20.2	Roles of Machines.....	460
20.3	Where Next?	461
20.4	Keep in Touch.....	461
	About the Authors	463
	Bibliography	464