Deep Credit Risk

Machine Learning in Python

Daniel Rösch*
Harald Scheule†

*DANIEL RÖSCH 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.

Deep Credit Risk

Machine Learning in Python

1st edition, 2020

ISBN: 9798617590199

Daniel Rösch, Harald Scheule
c/o AutorenServices.de
König-Konrad-Str. 22
D-36039 Fulda

Contents

Part I

Principles of Data Learning

1 Deep Dive

1.1 Boost Your Practical Machine Learning Skills

"Deep Credit Risk — Machine Learning in Python" aims at starters and pros alike to enable you to:

- Understand the role of liquidity, equity and many other key banking features;
- Engineer and select features;
- Predict defaults, payoffs, loss rates and exposures;
- Predict downturn and crisis outcomes using pre-crisis features;
- Understand the implications of COVID-19;
- Apply innovative sampling techniques for model training and validation;
- Deep learn from Logit Classifiers to Random Forests and Neural Networks
- Do unsupervised Clustering, Principal Components and Bayesian Techniques;
- Build multi-period models for CECL, IFRS 9 and CCAR;
- Build credit portfolio correlation models for value-at-risk and expected shortfall; and
- Run over 1,500 lines of pandas, statsmodels and scikit-learn Python code
- Access real credit data and much more ...

This book has five parts. In the first part, "Principles of Data Learning", we study the single mortgage dataset underlying this book, refresh our Python skills and compare Risk-based and Machine Learning. In the second part, "Data Processing and Validation", we develop the basic components and cover Output Engineering, Feature Engineering, Feature Selection and Validation. In the third part, "Default, Payoff, LGD and EAD Modeling", we develop a complete suite of statistical techniques, including models for probabilities of default and payoff (PD and PP), loss rates given default (LGD) and loan exposures at default (EAD). In the fourth part, "Machine Learning for PD and LGD Forecasting", we develop a suite of machine learning (ML) concepts, which have the advantage of greater machine processing and hence, greater scalability and flexibility. We cover K Nearest Neighbor, Naive Bayes, Bagging and Boosting, Trees including Random Forests, Neural Networks and Deep Learning including Support Vector Machines. In the fifth part, "Synthesis: Lifetime Modeling, IFRS 9/CECL, Loan Pricing and Credit Portfolio Risk", we bring everything together and calculate expected losses for loan loss provisioning and loan pricing as well as unexpected credit losses for economic and Basel capital.

We apply rigorous standards as we train methodologies using pre-crisis information and test/validate models with crisis and post-crisis information.

This may be the first comprehensive book in the area of credit risk analytics using Python. Python is an up-and-coming programming language that has become popular in many areas of data analytics in a very short time. Python is popular as it is free to use and offers some advantages for Machine Learning. We start with a first introduction to Python in the next chapter.

This book is aimed at credit analysts in financial institutions, fin-techs and prudential regulators, as well as credit risk researchers. Data, codes and more are available on the accompanying website: **www.deepcreditrisk.com**.

1.2 Credit Risk Information

Commercial banks keep data in a number of repositories that correspond to steps in the lending process: origination data, performance data, modification data payoff/retention data, default data/workout data and maturity data. These sources may include:

- Origination/underwriting data: relates to the origination or underwriting of loans;
- Performance data: relates to multiple reviews, which are generally conducted monthly, quarterly or annually and cover the period from the origination time to the latest review;
- Modification data: relates to the modification time;
- Payoff/retention data: relates to the payoff time;
- Maturity data: relates to the maturity time and covers administrative matters such as release of collateral and various accounting activities;
- Default/workout data: relates to the period from default to resolution and collects all cash flows that are paid or received by the bank during this time. Times are continuous and resolution periods can cover multiple periods (e.g., up to ten years).

External data may include:

- Macroeconomic information: time varying information that is identical at a given period for a number of borrowers. Macroeconomic information may be stratified by country, state, statistical areas and other units;
- Population statistics: these are generally time-varying;
- There are many other data sources which may include business filings, data from social networks, data from external experts such as ratings agencies, property appraisers, activity profiles of payment systems or transport systems.

The general aim is to construct a panel data. For the panel data, we observe information features and risk outcomes like default, payoff, loss rates and exposures for loans ($i = 1, ..., I$) and time periods ($t = 1, ..., T$). Generally speaking, there are a number of steps needed before we can process data.

First, raw data needs to be measured. Credit risk models are developed for scaling: estimated models are used to measure the risk of new borrowers and loans as well as borrowers and loans for which the risk factors have changed since the last review. Any information that is used in a model should be observable and hence measurable for the borrowers and loans. For the data sources, this means that the sensor that has been used in the past to collect the raw information should be available to collect the same information going forward. The sensor information is then transferred to numbers or strings, collected in an array and processed. Should the sensing or the transformation to numbers or strings fail, values will be missing and have to be treated in a second stage.

Second, identifiers that are common across number arrays and other credit relevant information need to be created. Different identifiers may exist. Two very important ones are the ultimate credit

subject (examples are borrowers or loans) and time. Examples for time in lending are application time, origination time, observation time, payoff time, default time and maturity time.

In identifying the ultimate credit subject, one may consider the loan or the borrowers and the question is whether the risk should be measured at the loan-level or the borrower-level. The answers to these questions are often bank-specific and an economic analysis is required for most consumer and corporate loans. Things get more complicated for a few high-net-worth individuals and large firms as these maintain many accounts and credit-risk relevant relationships with other subjects. Even the definition of borrower is not trivial as the borrower may be part of a larger holding structure, family or benefit from guarantees and an analysis at the parent level may be sensible.

Third, the resulting array of credit information is converted into a dataframe in a format that allocates rows as combinations of ultimate credit subject and time. Time is generally observation time but in special applications, such as building score cards for loan applications, it may also be loan origination time. Information is analyzed by computing relations. These relations can be reciprocal or one-way. In most of this book we analyze the impact of independent variables (also known as covariates, risk factors, explanatory variables, features and right-hand side variables) on dependent variables (also known as responses, outputs, outcomes and left-hand side variables).

1.3 Hands-on Analysis

We are working with one comprehensive dataset on US mortgages. The data is collected and provided by International Financial Research (see www.internationalfinancialresearch.org). The dataset is a randomized selection of mortgage loan-level data collected from the portfolios underlying US RMBS securitization portfolios.

We are further working with module `dcr.py`, which contains a number of credit risk functions that are repetitively used throughout this book. These functions make a great addition to any Python library.

To get started:

- Download data `dcr.csv` from www.deepcreditrisk.com;
- Download module `dcr.py` from www.deepcreditrisk.com;
- Place data and file in directory `'C:/TMP'` You can save both files in a different directory but need to then change all paths in dcr.py from `'C:/TMP'` to your preferred path.;
- Save all codes (e.g., Jupyter Notebooks) in the same directory.

The dataset is in panel form, reporting origination and performance observations for 5,000 residential US mortgage borrowers over 60 periods. The central variables are `id` and `time`. The data size was chosen to allow for efficient computing of a number of machine learning techniques. A larger dataset is available at www.deepcreditrisk.com.

The quarterly periods have been de-identified and observation periods run from 1 to 60. The observation period starts at the beginning of the millennium and includes the Global Financial Crisis (GFC) in period 27 approximately. Origination times prior to the start of the observation period have negative numbers. For example, `orig_time=-2` implies an origination time of two period prior to the observation window. Within this dataset we observe business cycles with economic upturns and downturns.

Outcome observations are observed over a time period after features are recorded. For example, default, payoff and status events are observed one period after features in the same row. LGD and related recoveries are observed between the default and resolution time.

As in the real world, loans may originate before the start of the observation period. This is possible if loans are transferred between banks and investors as in securitization. The loan observations may thus be censored as the loans mature or borrowers refinance.

The information is organized in the following order:

- Borrower IDs;
- Time stamps;
- Information features at observation time;
- Information features at loan origination;
- Outcome observations.

More specifically, key variables are:

- `id`: borrower id;
- `time`: time stamp of observation;
- `orig_time`: time stamp for origination;
- `first_time`: time stamp for first observation;
- `mat_time`: time stamp for maturity;
- `res_time`: time stamp for resolution;
- `balance_time`: outstanding balance at observation time;
- `LTV_time`: loan to value ratio at observation time, in %;
- `interest_rate_time`: interest rate at observation time, in %;
- `rate_time`: risk-free rate, in %;
- `hpi_time`: house price index at observation time, base year=100;
- `gdp_time`: GDP growth at observation time, in %;
- `uer_time`: unemployment rate at observation time, in %;
- `REtype_CO_orig_time`: real estate type — condominium: 1, otherwise: 0;
- `REtype_PU_orig_time`: real estate type — planned urban developments: 1, otherwise: 0;
- `REtype_SF_orig_time`: real estate type — single family home: 1, otherwise: 0;
- `investor_orig_time`: investor borrower: 1, otherwise: 0;
- `balance_orig_time`: outstanding balance at origination time;
- `FICO_orig_time`: FICO score at origination time, in %;
- `LTV_orig_time`: loan to value ratio at origination time, in %;
- `Interest_Rate_orig_time`: interest rate at origination time, in %;
- `state_orig_time`: US state in which the property is located;
- `hpi_orig_time`: house price index at observation time, base year=100;
- `default_time`: default observation at observation time;
- `payoff_time`: payoff observation at observation time;
- `status_time`: default (1), payoff (2) and non-default/non-payoff (0) observation at observation time;
- `lgd_time`: LGD, at default time, assuming no discounting of cash flows;
- `recovery_res`: sum of all cash flows received during resolution period.

Python language is case-sensitive. Please ensure that you use the identical upper- and lowercase character as indicated. We make use of all variables, the most important of which used in this book are `default_time` (outcome variable) as well as `LTV_time`, `FICO_orig_time` and `gdp_time` (feature variables).

A central variable is the loan to value ratio (LTV). The data includes LTV at loan origination time (`LTV_orig_time`) and at observation time (`LTV_time`). The two ratios are related. `LTV_time` includes `LTV_orig_time`, the change in house prices (measured by the house price index, HPI) and amortization measured by the loan balance at loan origination (`balance_orig_time`) and at observation time (`balance_time`). The formula for the calculation is:

$$\text{LTV\_time} = \frac{\text{balance\_time}}{\text{house price at time}}$$

The house price at time can be approximated as follows:

$$\text{house price at time} = \text{house price at origination} * \frac{\text{hpi\_time}}{\text{hpi\_orig\_time}}$$

With:

$$\text{house price at origination} = \frac{\text{balance\_orig\_time}}{\text{LTV\_orig\_time}}$$

We later convert the LTV ratio to home equity for economic discussions. Most variables are generally observed for all observations. Exceptions are `lgd_time`, `recovery_res` and `res_time` which are only observed for `default_time=1` and if the resolution process is complete.

1.4 Basel, CECL, IFRS 9, DFAST, CCAR and Stress Tests

Deep credit risk analytics is the basis of a number of regulations for financial institutions that all have an impact on the amount of capital an institution holds. Important regulations are:

- Basel: minimum amount of required Tier I and Tier II capital. Basel may include the various reforms (Basel I to Basel III), and a number of nationally issued guidance notes;
- Current Expected Credit Loss (CECL), IFRS 9: loan loss provisioning and eligible amount of available Tier I capital;
- National stress tests (e.g., Dodd-Frank Act Stress Test (DFAST) or Federal Reserve Bank (FRB) stress tests in the US): requirement of additional capital buffers;
- Comprehensive Capital Analysis and Review (CCAR): requirement of additional capital buffers.

All frameworks are related and large banks require the estimation of PDs, LGDs and EADs. However, the estimation details may differ. Basel requires through-the-cycle PDs, Downturn EADs and Downturn LGDs. CCAR, CECL and IFRS 9 require Lifetime PDs and EADs/LGDs that are based on current economic circumstances and are forward-looking, i.e., take future expectations into account. DFAST and FRB stress tests require stressed PDs, LGDs and EADs.

It is possible to meet all requirements within the frameworks presented in this book. Further, financial institutions may compute the minimum amount of economic capital. Economic capital is important

for setting loan prices and provide incentives that are aligned with stakeholder values. Banks that work off credit prices that are too low attract risks that are too high through adverse selection. Bank that use prices that are too high lose market share and are unable to offset fixed costs.

1.5 First Lessons from the COVID-19 Crisis

1.5.1 Credit Risk Analytics

At the time of writing this book the COVID-19 Crisis has started and it was evident that there will be a long-lasting impact on many areas of human life including bank lending and credit risk. There are a number of aspects that are important for credit risk analytics and we believe that we cover some in this book.

The crisis has led to a sudden shock to employment and incomes that is unprecedented in most banks' data including the Global Financial Crisis (GFC). Many employees have been made redundant, stood down (i.e., kept employed at zero income), or lost a part of their entitlements. The follow-on effects were the inability of private and corporate borrowers to make rent payments and private and corporate investors to service loans. Governments have tried to offset liquidity constraints by distributing money (sometimes called helicopter money), lowering taxes and offering liquidity facilities. We discuss liquidity to some degree in this book by measuring feature `cep_time` (see the chapter on feature selection). There are many aspects and proxies for liquidity and banks often consider working capital ratios, debt to income ratios and more sophisticated measures from transaction data.

Equity next to liquidity is a central aspect. The COVID-19 crisis had at the beginning limited impact on asset and hence wealth values. Many believe that there will be also a shock to asset markets. We discuss equity in this book using feature `equity_time`.

Time effects are very relevant, this includes the origination time, age and the state of the economy.

There is most "meat on the bone" in credit risk models to carefully collect and interpret information. For example, the pandemic is global as many systems that we have created. The impact on the travel industry, global supply chains for many labor, finance, commodities, manufactured goods and service will become key features in explaining credit risk outcomes yet most models built are poor on this end. Many identified features are missing in today's models and can be added to the models.

That said the correct use of methodologies is key from a model design perspective. For example, models built during economic upturns struggle predicting credit risk outcomes during downturns.

In this book, we try to provide and showcase solutions for many questions. The following table shows potential solutions to some challenges that may come from the COVID-19 crisis impact:

Challenge	Solution	Chapter
Calculating Crisis PDs without downturn data	Model-based measurement of crisis PDs, Parameter-based stress-testing (Margin of conservatism, Bayesian approach)	Chapter 9 Default Modeling — Crisis PDs/Stress-testing, Chapter 3 Risk-based Modeling — Bayesian Approaches

Challenge	Solution	Chapter
Calculating Crisis PDs with downturn data	Scenario-based stress-testing, Parameter-based stress-testing (Regime-switching models)	Chapter 9 Default Modeling — Stress-testing
Liquidity as a driver of default	Estimation of models with liquidity as feature; Inclusion of additional liquidity feature (e.g., income over non-discretionary expenses)	Chapter 5 Outcome Engineering, Chapter 9 Default Modeling
Equity as a driver of default	Estimation of models with equity as feature; Inclusion of additional equity feature (e.g., home improvement/deterioration)	Chapter 5 Outcome Engineering, Chapter 9 Default Modeling
Impact of time effects	TVA Analysis: control for vintage and age effects through dummy variables or other features that describe the origination process and for time effects through macroeconomic features	Chapter 5 Outcome Engineering, Chapter 9 Default Modeling
Low default portfolios	Most prudent estimators/Margin of conservatism	Chapter 9 Default Modeling — Low default portfolios
Validation of pre-crisis models	Backtesting: split training and validation sample along time dimension	Chapter 8 Validation
Ability of machine learning models to predict defaults for severe downturns	Backtesting of machine learning approaches	Chapter 13 Standalone Techniques, Chapter 14 Neural Networks and Deep Learning, Chapter 15 Ensemble Techniques, Chapter 16 Machine Learning for LGD
Adequacy of model estimates for application	Comparison of Basel capital with expected loss	Chapter 9 Default Modeling – Crisis PDs

Our approach includes four main elements. First, model calibrations should include the latest credit data. Second, features that are identified as drivers of credit risk outcomes should be included in the models. For example liquidity, equity and network effects should be added to the feature space. Third, validation should focus on backtesting and include a train-test split along the time line. This is known as backtesting. Fourth, the adequacy of model estimates for applications like capital adequacy, loan loss

provisioning and loan pricing needs to be vetted.

1.5.2 Machine Learning

Fintech is short for “financial technology” and is a worldwide innovation boom that has created multi-billion dollar industries in many countries. Fintech is seen as the future of financial services and challenged existing financial service providers like banks, insurers and pension funds by lower variable costs. The services are generally offered digitally and online via mobile apps. Back-end operations, including credit risk analytics are often to a higher degree automatized. This industry has been at the forefront of applying advanced technique in credit risk analytics

The outbreak of the pandemic COVID-19 let to an unprecedented decline in many financial markets globally. The pandemic was to a large degree unexpected but two issues have become already apparent. First, fintech firms have in many cases lost more market value than existing banks as they have issued loans to higher risk borrowers and often at lower lending standards. Second, advanced econometric techniques had a limited ability to predict the crisis.

In defense of the fintech industry, we would like to stress that the growth — like with many other innovations — was driven by investor over-excitement, which perhaps led to lending standards that were below model standards. Further, credit risk generally realizes in time-lags and we have not seen the full impact of the pandemic on credit risk. Traditionally, economic downturns result in unemployment, a lower wealth and liquidity of borrowers declining housing and other collateral values. The outcomes are greater future likelihoods and magnitudes of credit losses. Whilst advanced models may have failed in an immediate crisis prediction, they may adjust quicker to the new risk levels than traditional models.

It is possible that an upcoming credit crisis will be interpreted as a failure of fintechs and advanced credit risk approaches. We take a neutral position in this book and judge all techniques based on pandemic levels. We rely on the Global Financial Crisis, as a severe economic shock, and analyze the performance of these advanced approaches as if they had been built prior to the crisis and use-tested in the aftermath.

The fintech and banking industry are under great stress and the industry is likely to transform as a consequence. However, we believe that new business models will continue to challenge existing lenders by using new technology. This book will help to provide a common understanding and enable credit risk analysts to benefit from opportunities that will come with it.