Boosting Credit Risk Models

Bart Baesens[1,2], Kristien Smedts[1]

[1]Faculty of Economics and Business

KU Leuven

Belgium

[2]Southampton Business School

University of Southampton

United Kingdom

{Bart.Baesens;Kristien.Smedts}@kuleuven.be

Abstract

In this article, we give various recommendations to boost the performance of credit risk models. It is based upon more than two decades of research and consulting on the topic. Building credit risk models typically entails four steps: gathering and preprocessing data, modelling of probability of default (PD), Loss Given Default (LGD) and Exposure at Default (EAD), evaluating the credit risk models built and then the deployment step to put them into production. We give recommendations to boost credit risk models during each of these steps. Furthermore, we also define and review model risk as an all-encompassing challenge one needs to be properly aware of during each step of the process. We conclude by presenting a research agenda of topics we believe are in high need for further investigation and study.

Keywords: credit risk; Probability Of Default (PD), Loss Given Default (LGD), Exposure At Default (EAD), Basel, IFRS 9

Introduction

Following the 2008 financial crisis, risk management models and processes have been improved in order to be better equipped to navigate new crises (BIS 2015). Undoubtedly, this will have had an impact on how financial institutions have absorbed the COVID-19 pandemic, and are currently dealing with difficult macroeconomic conditions and a fragile geopolitical situation. Notwithstanding enhanced risk management practices, the current uncertainties call for increased attention of credit risk managers. Rating agencies hint at growing defaults and worsening probabilities of default. S&P (2023) notes that year-to-date defaults have been at the highest level since 2009. In addition, they argue that higher credit risk is not only observed for the lower rated instruments, but also for investment grade debt. Similarly, Swiss RE (2022) forecasts that the rate of default will increase, even reaching levels of 15% for high yield instruments in a severe recession scenario. Finally, also the European Banking Authority points at a rising share of loans that show significant increases in credit risk (EBA, 2022). In addition, due to restrained access to bank credit, we have witnessed a boom in private credit which, coupled with less transparency and weaker regulation, could hide additional credit risks (Moody's, 2023)). As well-functioning credit markets are crucial to foster economic growth, efficient and effective credit risk management should therefore be high on the agenda of policymakers, regulators, and lending firms (BIS, 2018; Bessley, Roland and Van Reenen, 2020). That is also why regulators continue to urge financial institutions to proactively push them to strong credit risk management practices (ECB Supervision blog, 2020).

Much effort therefore goes into the estimation of credit risk and the determination of economic capital to cover for unexpected credit losses. In doing so, financial researchers are very much stimulated by real-world events. For example, Brooks and Schopohl (2018) show that research on credit risk has surged after the global credit crisis with an aim to provide answers to this unprecedented realisation of credit losses. As a result, tremendous progress has been made in the measurement of credit risk and in the tools available to measure it. While credit risk was initially limited to the estimation of simple notional amounts, it was soon clear that such view on credit risk was ignoring its complexity. Credit risk analysis has therefore evolved first to risk-weighted amounts, later to the inclusion of internal and external credit ratings, and ultimately to the development of complex internal portfolio credit models.

While different institutions all build and use their proprietary credit models, they all share the same key input variables: probability of default (PD), loss given default (LGD), and credit exposure (CE) or exposure at default (EAD) (Baesens, Rösch and Scheule, 2016). Probability of default is the likelihood that the counterparty will default on its obligations typically in the next 12 months. The higher this likelihood to default, the higher the credit risk. Loss given default is the fractional loss due to default. The lower the recovery rate, the higher the LGD and credit risk. Finally, the credit exposure is the market value of the claim. At default, this is also referred to as exposure at default. The larger the exposure, the more credit risk the bank faces.

The challenge of estimating credit risk goes beyond the precise estimation of the above risk drivers. Not only is there is need for an accurate estimation of these risk elements, in addition, there is also a need for an accurate aggregation of the different determinants, and for an aggregation at the level of the portfolio. Still today, this poses many challenges to both scholars and practitioners. Continued attention and effort to resolve these challenges is crucial: the economic capital that banks assign to cover for unexpected credit losses should be estimated as meticulously as possible, since the cost of overcapitalization as well as undercapitalization is very high. Overcapitalization is costly with, at individual bank level, the risk of a low return on equity and thus of being competed out of the market (Bouteille and Coogan-Pushner, 2021) and at market level, the risk of reduced capital supply (see, e.g.,

Gopalakrishnan, Jacob and Mohapatra, 2021, and Hyun and Rhee, 2011). Undercapitalization, on the other hand, increases the risk of a bank failure, and potentially jeopardizes the continuity of the bank, see e.g. Demirgüç-Kunt, Detragiache and Merrouche (2013), and Beltratti and Stulz (2012).

This trade-off and fragility is also recognized by policymakers. Because of the important role of banks, the systemic risk in the sector, and their business model with relatively low levels of capital, banks operate in a highly regulated environment (Hull, 2018). Apart from the economic capital calculation for internal purposes, banks therefore need to be in rule with the Basel regulations and with accounting prescriptions such as IFRS 9.

With the outbreak of the credit crisis in 2008, it became clear that risk, and in particular credit risk, had been largely underestimated. In response to this, new Basel requirements have been introduced that significantly strengthened bank solvency (BIS, 2017). Not only do banks need to hold more and more qualitative capital, they also need to dispose of more liquidity.[1] In addition, banks also need to report more elaborately about the risks they face, both on-balance as well as off-balance. With regard to credit risk, the standardized approach to measure credit risk has been revised and the use of internal models has been constrained. Where banks always had quite some flexibility in developing their own internal risk models, regulators nowadays acknowledge that some standardisation is desirable to reduce unwarranted variability in banks' Risk-Weighted Asset (RWA) calculations and to limit the potential benefits of using internal models. This has, among others, led to the introduction of a leverage ratio and a capital floor, independent of the outcome of the internal risk models (BIS, 2017).

In addition, reporting requirements for banks have drastically changed. The introduction of IFRS 9 has, to a large extent, aligned the accounting treatment of credit risk measures such as loan loss provisions and reserves and the valuation of financial instruments with the Basel practices. The insights from risk management can therefore be the basis for informative reporting practices. In addition, this transition to a more forward-looking approach as also used in risk management will allow for more integration between the traditional risk calculations and reporting requirements, and thus between the risk and finance functions. As noted by Lim, Woods, Humphrey and Seow (2017) such integration and interaction between both are required for effective risk management.

Building upon the above insights, it is hence of key importance to have good performing credit risk models, amongst others for economic capital calculation, loan provisioning and reporting. In this paper, we contribute to the literature by filling two gaps. First, we provide a concise yet focused literature overview of the state of the art in credit risk modelling. As research on credit risk spreads across a wide variety of disciplines ranging from accounting, finance, and banking, to analytics, machine learning, operations research (OR), and even legislation and ethics, we find this to be an important gap to close. Secondly, we give clear recommendations on how to boost the performance of contemporary credit risk models crystallising new research directions and applications in the field.

In what follows, we start by outlining a generic credit risk model architecture and discuss key model requirements. Next, we introduce four key elements where performance improvements can be made: data, model, evaluation, and deployment. In terms of data, we zoom in on new data sources and featuring engineering. At the model level, we review the added value of Bayesian and deep learning. In terms of evaluation, we elaborate on profit-driven modelling. For deployment, we discuss model

[11] Chiaramonte and Casu (2017) find that the new liquidity requirements seem effective in reducing bank fragility. The effectiveness of strengthened capital requirements, however, seems to be limited to large financial institutions. This supports the differential regulatory treatment of large and systemically banks within Basel III.

use and stress testing. We conclude by defining model risk and illustrating how it transversally relates to all four elements discussed: data, model, evaluation, and deployment.

Credit Risk Model Architecture

Contemporary credit risk models of lenders worldwide are typically built using a three-layer framework. This framework is mostly not only used to build credit risk models, but also to validate, backtest, benchmark and stress test them (Baesens, Rösch and Scheule, 2016).

Level 0 is the data that feeds into the models. This could be socio-demographic data (Lessmann, Baesens, Seow and Thomas, 2015), credit facility data (Loterman, Brown, Martens, Mues and Baesens, 2012), transactional data (Baesens, Rösch and Scheule, 2016), external data (De Cnudde et al., 2019; Bartov, Faurel and Mohanram, 2022; Stevenson, Mues and Bravo, 2022) or expert based (also called soft) data (Ozdemir and Miu, 2009; Doumpos and Figueira, 2019; Luong, Scheule, and Wanzare, 2023). All these data sources need to be adequately combined and preprocessed (e.g., missing values, outliers, categorization, transformations) before being fed to level 1. The impact of these activities is not to be underestimated as they have a direct impact on the performance of the credit risk models that will be built higher up in the architecture.

At level 1, discrimination models are built to discriminate obligors or credit facilities in terms of default (PD), loss (LGD), or exposure (EAD, CE) risk. Model discrimination essentially aims at sorting or ranking obligors or exposures. For PD modelling, logistic regression is the reference technique. It is used to build both application and behavioural scoring models which yield credit scores for both new and existing obligors. Typically, these models discriminate fairly well with on average between 10 to 15 predictors and area under the Receiver Operating Characteristic (AUC) values ranging between 70% to 85% (Baesens et al., 2003; Lessmann, Baesens, Seow and Thomas, 2015). For LGD/EAD modelling, commonly used techniques are linear regression, regression trees, and mixture models. These models are typically much harder to estimate with R-squared performance metrics seldomly above 20% which is quite unfortunate given the fact that both LGD/EAD have a linear impact on unexpected and expected losses (Loterman, Brown, Martens, Mues and Baesens, 2012).

Level 2 is the ratings and calibration level where the scores or output of level 1 are categorized into more robust default, loss, and exposure ratings, accompanied by calibrated risk measurements which go beyond discrimination and precisely quantify the probability (in case of PD) or percentage (in case of LGD and EAD) (Tasche, 2013; Bequé, Coussement, Gayler and Lessmann, 2017). These calibrations are typically done using both historical data as well as forward-looking expectations about, e.g., the global economy and the market(s) the lender operates in. Moreover, for LGD and EAD the calibration should be based on economic downturn assumptions since, contrary to PD, no conservative adjustment is made in the Merton capital requirements formula underlying the Basel regulation (Van Gestel and Baesens, 2009).

Credit Risk Model Requirements

Credit risk models typically have to satisfy various requirements. In what follows, we elaborate on statistical performance, interpretability, profitability, regulatory compliance, and ecological impact.

A first one concerns statistical performance. The discrimination power of PD models is usually evaluated using the AUC, Gini (which equals 2×AUC-1) or Kolmogorov-Smirnov (KS) statistic (which equals the maximum vertical distance between the ROC curve and the diagonal). For LGD/EAD models it is quantified using the mean squared error (MSE), mean absolute deviation (MAD), Pearson correlation, or (adjusted) R-squared. The calibration performance of PD models can be measured by

means of the Brier score or various test statistics such as the binomial, Hosmer-Lemeshow, normal or Vasicek test (Thomas, 2009). For LGD/EAD models, it can be evaluated by means of a Student's t-test or a Wilcoxon signed rank test, often combined with bootstrapping procedures (Loterman, Brown, Martens, Mues and Baesens, 2012).

Another key requirement of credit risk models is interpretability (Molnar, 2022). This is not only needed to provide a possible explanation to customers whose credit has been denied, but also to foster trust and transparency since complex models (e.g., ensemble methods, deep learning) are typically still very much frowned upon and hence distrusted by both model users and regulators. As already mentioned, for PD modelling, logistic regression is the international industry standard. Throughout our research (Baesens et al., 2003; Lessmann, Baesens, Seow and Thomas, 2015), we found it to work surprisingly well, if not best, even when compared to more powerful techniques such as ensemble or deep learning methods. Though logistic regression is already quite interpretable in itself, a very intuitive but less well-known way to further improve its readability is to use nomograms which we highly recommend (Zlotnik and Abraira, 2015). Also for LGD/EAD modelling, linear regression, regression trees, and mixture models are considered very interpretable (Baesens, Rösch and Scheule, 2016).

Numerous studies find that credit risk is negatively related to profitability (Miller and Noulas, 1997; Athanasoglou, Brissimis and Delis, 2008; Petria, Capraru, and Ihnatov, 2015). More specifically, good credit risk models should result into fewer non-performing loans, better financial solvency and hence lower default probability of the lender itself. While it is indeed true that credit risk losses weigh on the profitability of the lender, the relation between credit risk, and its risk drivers, on the one hand, and profitability on the other hand is not one-to-one, and is conditional on the credit product and market that is analysed. In the context of revolving credit facilities, Andreeva, Ansell and Crook (2007) show that probability of default is non-linearly related to profitability, with crucial information conveyed by the survival probability of default and of second purchase. Related, numerous papers find that it is not traditional credit scoring methods with a focus on PD that are most linked to profitability, but rather profitability-driven scoring models with a focus on internal rates of return (Verbraken, Bravo, Weber and Baesens, 2014).

Given their strategic impact, credit risk models should also be regulatory compliant. This not only relates to, e.g., the Basel and IFRS 9 guidelines that have been introduced, but also to data that feeds into these models which should be in accordance with privacy, ethical, and fairness guidelines as we discuss below (Martens, 2022). In fact, these guidelines constitute a vast, continuously evolving, and controversial area of research with multiple and often conflicting definitions and/or regulatory requirements. Verma and Rubin (2018) critically summarize and illustrate the most prominent definitions of fairness and illustrate these using a logistic regression model for application scoring. Andreeva and Matuszyk (2019) illustrate that the inclusion of gender as a predictor for credit scoring is statistically significant and though it does not affect model performance (due to other variables proxying gender effects), it does impact the proportion of accepted women/men. Using data on US mortgages, Fuster, Goldsmith-Pinkham, Ramadorai and Walther (2022) find that Black and Hispanic borrowers are disproportionately less likely to benefit from machine learning. Lee and Floridi (2021) consider fairness in a relational way and use US mortgage data to illustrate how their methodology better captures ethical trade-offs. Makhlouf, Zhioua, and Palamidessi (2020) introduce a list of criteria for the applicability of machine learning fairness notions and provide a decision diagram to navigate these. Kozodoi, Jacob, and Lessmann (2022) revisit statistical fairness criteria for credit scoring, survey algorithmic ways of embedding fairness in model development, and empirically compare different fairness processors for profit-oriented credit scoring. They give recommendations for fair credit

scoring and clarify the profit-fairness trade-off. Furthermore, not only when developing credit models, but also when using them, one needs to be careful of any behavioural biases, e.g., loan approval (Office of the Comptroller of the Currency, 2023). Among others, Beck, Behr and Guettler (2013) find gender differences in the management of credit risk, with female loan officers having a lower likelihood of granting non-performing loans. While the use of hard information, such as scoring models and expert systems can weaken such behavioural biases, Reynal-Querol and Garcia-Montalvo (2020) show that gender differences exist even when there is relatively little room for soft information. Related, Witzany (2017) argues that an effective risk organization is crucial for excellent credit models to pay off. The loan officers making the ultimate credit assessment need to do this in both an unbiased and independent manner.

One should also consider the ecological impact of credit risk models. Though this is not a concern for the rather simple predictive methods (e.g., linear/logistic regression, decision trees) we discussed above, it may become an issue to deep learning methods which may be considered to process the newer sources of data or do feature engineering as we discuss below. These models typically require tons of parameters to be estimated on energy-consuming hardware accelerators, leaving behind a heavy carbon footprint (Getzner, Charpentier and Günnemann, 2023) which can be quantified in terms of energy in kWh, CO2 emission in grams, and the equivalent in km driven by a car (Anthony, Kanding and Selva, 2020).

Finally, note that there exist mutual trade-offs between these various criteria as visualized in Figure 1. Let us elaborate on some of them. Statistically accurate credit risk models are not necessarily the most profitable as they have been trained with another objective function (Stripling, vanden Broucke, Antonio, Baesens and Snoeck, 2018; Höppner, Stripling, Baesens, vanden Broucke and Verdonck, 2020) or rely on complex black box (e.g., deep learning) techniques (Gunnarsson, vanden Broucke, Baesens, Óskarsdóttir and Lemahieu, 2021) which not only gives them a heavy ecological footprint, but also makes them not regulatory compliant.

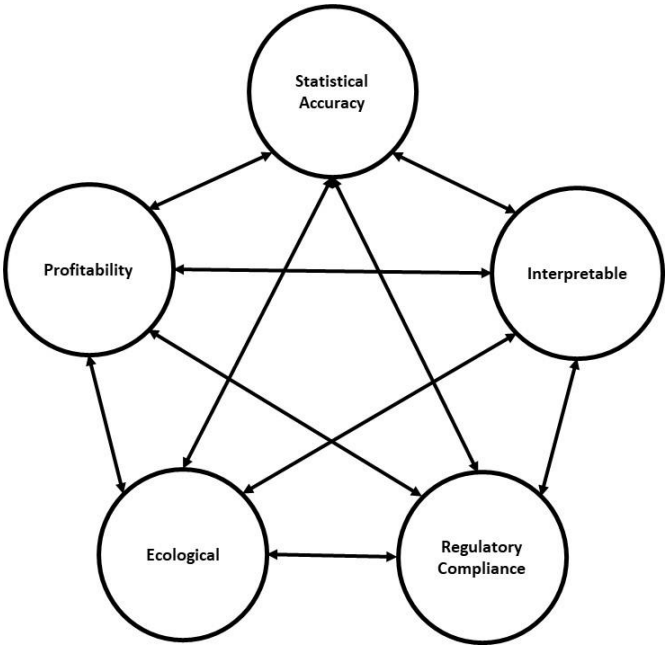

Figure 1 Trade-offs between credit risk model requirements.

Boosting Credit Risk Models

We conducted various benchmarking studies for both PD and LGD modelling. Baesens et al. (2003) analyse eight application credit scoring data sets with 17 classification techniques. Lessmann, Baesens, Seow and Thomas (2015) repeat the same study and analyse eight data sets with 41 classification techniques. In Loterman, Brown, Martens, Mues and Baesens, 2012, we study LGD modelling for retail and corporate loans using six data sets and 24 techniques. We are not aware of any in-depth benchmarking studies for EAD/CE modelling, but based on smaller-scale studies (Brown, Mues and Thomas, 2010), the initial results and trends seem quite similar to the ones found for LGD modelling.

Essentially, many of these studies illustrate that simple techniques such as linear/logistic regression and decision trees are usually very competitive with more complex techniques such as ensemble methods and/or deep learning on structured or tabular data. Hence, in order to boost the performance of credit risk models, either in terms of discrimination or calibration, one needs to carefully reconsider the various steps of model development and see where key improvements can be made. In what follows, we start from the data and illustrate various new sources that could be beneficial for credit risk modelling. We then continue and zoom into feature engineering to creatively enrich and/or transform data to better predict credit risk. Next, we focus on the modelling step itself. We review Bayesian learning as a promising set of methods to deal with both expert and data patterns which is especially relevant for small or low default credit portfolios. We also extensively discuss how deep learning can play a role in credit risk modelling. We further zoom in on the evaluation criteria used and advocate the role of profit in credit risk modelling. We conclude by elaborating on model risk, which essentially transversally encompasses all model development steps.

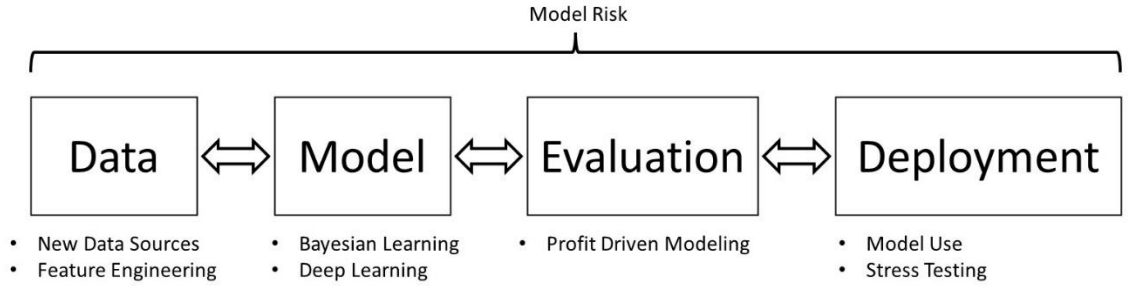

Figure 2 Boosting Credit Risk Models.

We elaborate on each of the above in what follows.

New Data Sources

In what follows, we elaborate on various data sources that could be beneficial for boosting credit risk models. We do wish to highlight that some of these may come with privacy and/or ethical concerns and should be thoroughly evaluated against (inter)national legislation before they can be used. In fact, many regulators worldwide are currently struggling and lagging behind to develop accompanying legislation and guidelines on what data can be used for what prediction task (e.g., credit, fraud, marketing, HR).

Text data is a first interesting source to consider to boost credit risk models. Stevenson, Mues and Bravo (2021) integrate loan officer text statements with standard credit scoring to predict small business loan default. They find text to be predictive, but less relevant when combined with traditional structured credit scoring data. Zhang, Wang, Zhang, and Wang (2020) develop a peer-to-peer (P2P)

lending model based on textual data derived from loan descriptions. Using deep learning, they illustrate how textual data further boosts the AUC performance when combined with traditional data on two real-life data sets. Similarly, Kriebel and Stitz (2022) use deep learning on user-generated descriptions from loan applications from Lending Club to improve credit default predictions in peer-to-peer lending. Nguyen et al. (2021) successfully leverage SEC filings containing long-form textual information about the quality of companies and their outlook, and combine it with financial statement and market data using stack ensembling and bagging to predict corporate credit ratings. Other interesting textual data to consider for credit risk modelling could be news reports or social media posts. Cathcart, Gotthelf, Uhl, and Shi (2020) construct a global news sentiment indicator from the Thomson Reuters News Analytics database and find it to be predictive of credit default swap (CDS) returns and thus default risk for sovereign exposures. They find a negative relation between CDS returns and news sentiment. In other words, positive sentiment is typically followed by a decrease in returns. Aguilar, Ghirelli, Pacce, and Urtasun (2021) construct a new newspaper-based daily sentiment indicator for Spain which is significantly correlated to GDP, which in turn can be helpful to calibrate and/or stress test credit risk models. Bartov, Faurel and Mohanram (2022) find that aggregate Twitter opinion can predict changes in CDS spreads and credit ratings.

Also network data can be used for credit risk modelling. Óskarsdóttir, Bravo, Sarraute, Vanthienen and Baesens (2019) use call detail record (CDR) data for PD modelling. The research was triggered by the insight that people tend to call those in their economic circle. Using a unique credit card data set with not only traditional sociodemographic data but also calling behaviour, they find that combining CDR data with traditional data for credit risk modelling significantly increases both the statistical performance as well as profitability. In fact, the benefit of using CDR data for credit scoring is twofold. First, it can boost the performance of credit risk models. However, it can also assist people living in developing countries, minorities, immigrants, and young people to get access to credit since these typically do not have any historical data available to make a credit decision. The higher availability of mobile phone data for this population provides an interesting alternative for credit scoring, hereby facilitating credit access. This clearly illustrates that what may initially look like a privacy concern, could also imply interesting opportunities or in this case the fostering of financial inclusion. Óskarsdóttir and Bravo (2021) introduce a framework for creating a multilayer bipartite network based on geographic location and economic activity of borrowers to model credit risk. Introducing a personalized multilayer PageRank centrality measure, they illustrate that including this network data improves the AUC by more than 10% over non-network credit risk models. Muñoz-Cancino, Bravo, Ríos and Graña (2023) build networks based on family ties formed by marriages or parent/child relationships and corporate ties (based on co-ownership, employment, and transactional services) for a Latin-American country. They introduce a new framework to combine graph representation learning (i.e., feature engineering, graph embeddings, and graph neural networks) for application and behavioural credit scoring and illustrate its superior performance in terms of AUC and KS when compared to traditional methods. Their results clearly demonstrate the potential of network data for credit scoring thin-file borrowers (i.e., borrowers with only limited historical credit data).

On-line data is another interesting data source to consider. Biatat, Crook, Calabrese and Hamid (2021) illustrate the predictive power of email usage and psychometric variables for PD modelling. Rozo, Crook, and Andreeva (2023) illustrate how web browsing variables can enhance the predictive accuracy of PD models. Social media data could also be a potentially interesting source for credit scoring. De Cnudde et al. (2019) study the use of Facebook data for credit scoring in microfinance. Their research finds clear clusters of befriended (non-) defaulters. Another key finding is that both defaulters and non-defaulters tend to like similar Facebook pages. Yang, Yuan and Lau (2022) use deep learning techniques to identify personality traits of borrowers through their social media posts and

illustrate the identified psychometric patterns to be predictive of credit risk. Similar to the CDR data we discussed above, all these variables can be especially interesting to score applicants with little to no credit history.

Gebru et al. (2017) portray the demographic and economic makeup of neighbourhoods using Google Street View. They illustrate how socioeconomic attributes such as income, race, education, and voting patterns can be inferred from cars detected in Google Street View images using deep learning. More specifically, more foreign cars in a neighbourhood implies a higher average income. German and Japanese cars (Lexus in particular) are found in areas with high median household income, whereas American cars (such as Buicks, Oldsmobiles, and Dodges) are associated more with lower median incomes. Besides Google Street View, also LiDAR (short for "Light Detection And Ranging" or "Laser Imaging, Detection, And Ranging") data can be considered for geo-demographic profiling. LiDAR data provides detailed three-dimensional elevation maps of landscapes and is also available as open source. Stevenson, Mues and Bravo (2022) propose the use of deep learning to create embeddings such that this data can easily be used for geo-demographic analysis. They illustrate the significant performance of their approach for predicting seven English indices of deprivation (income, employment, education, health, crime, barriers to housing & services, and living environment) in the Greater London area. The deep learning features or embeddings learned using either data sources (Google Street View or LiDAR) could be interesting add-ons in, e.g., LGD models for mortgages which often use these characteristics to accurately estimate the Loan-to-Value (LTV) ratio based on well-calibrated real-estate value estimates.

Google Trends is another interesting source as it provides search data and popular search terms across e.g., time and geography. More specifically, it shows how frequently a search term is entered into Google's search engine relative to the site's total search volume over a given period of time. As such, it can be used for nowcasting to forecast the present or near future. This allows to spot trends more quickly than traditional channels such as government organizations or statistical reporting agencies. A popular example is forecasting unemployment based on Google searches with key terms such as jobs, unemployment, unemployment benefits, social security, curriculum vitae, motivation letter. González-Fernández and González-Velasco (2020) use Google Trends data for constructing a sentiment index for bank credit risk which is significantly and positively related to bank sector CDS. Woloszko (2020) describes a methodology to develop an OECD Weekly Tracker of GDP providing a weekly nowcast of GDP growth rates by analysing Google Trends data using neural networks for the OECD and G20 countries. Google Trends data is also used by Bargaglia et al. (2022) for nowcasting GDP under the COVID pandemic. To summarize, Google Trends data can be very useful for PD, LGD, and EAD calibration, especially in an IFRS 9 setting where these measurements should be point in time or based on the reporting date. It can also be useful to anticipate macroeconomic downturns more quickly and do stress testing.

Also data related to sustainability could be worthwhile to consider, especially for sovereign credit portfolios. More specifically, Moody's (2020) report that sustainability risks have contributed to 36% of sovereign defaults since 1997. As an example, climate change can both directly and indirectly affect sovereign risk through public finances. It raises the cost of capital for climate-vulnerable countries and as such jeopardises debt sustainability. Anand, Vanpée, and Baesens (2023) study the significance of sustainability risks in predicting Moody's sovereign credit ratings. Using XGBoost combined with Shapley values and LIME for interpretation (Molnar, 2022), they find a significant yet complex, non-monotonic, and country-specific relationship between sustainability and credit risk. It can be expected that similar effects also play in corporate credit portfolios.

Asset correlations are another important data element that play a crucial role in calculating capital requirements. Within the IRB approach, these values are either fixed (e.g., 15% for mortgages and 4% for qualifying revolving exposures) or computed as a positive function of firm size, and a decreasing or constant function of the likelihood of default (BIS, 2005). As a consequence, the asset correlation estimates are rather static and do not reflect the empirically observed variability. Lee, Lin and Yang (2011) show that asset correlations are industry-specific, procyclical, and asymmetric. A similar result is found by Cho and Lee (2022), who stress that this leads to an underestimation of the asset correlations, and thus to capital cushions being too small during economic downturns.

All of the above data sources focus on creating better predictors for credit risk modelling. However, also the target variables themselves, PD, LGD, and EAD are worth reconsidering. PD is typically tackled as a binary classification problem assuming a preset definition of default which is typically 90 days in payment arrears as set in the Basel accord. Another approach is to predict the number of days overdue using a Tobit regression model as done by Brezigar-Masten, Masten and Volk (2021). They illustrate that modelling of days past due can be useful for default stage allocation in an IFRS 9 setting. Korangi, Mues, and Bravo (2023) successfully use deep learning for mid-cap company default prediction for multiple time horizons, integrating different sources of data measured with different frequencies. Another way of specifying the PD target is to predict the time of default using survival analysis techniques. Predicting when borrowers default can also be very valuable information for profit scoring and customer lifetime value (CLV) modelling (see, e.g., Dirick, Baesens and Claeskens, 2017). In terms of the LGD target definition, the most common challenges remain what discount factor to use, how to factor in incomplete workouts and indirect costs and how to qualify an economic downturn scenario for calibration (Scheule, Jortzik, 2019; Betz, Kellner, Rösch, 2018). Especially given current economic conditions, some lenders are getting worried about a significant portion of future LGDs exceeding 100%. This poses an interesting research challenge and some preliminary, mostly unpublished attempts have been made at developing structural models for LGD. As to the EAD or credit conversion factors (CCF) target definition, the most important issues relate to CCFs exceeding 100% and also the qualification of an economic downturn scenario.

Feature Engineering

As we discussed above both structured as well as unstructured data can be used for credit risk modelling. Structured data can easily be represented in a tabular format with rows typically depicting obligors, companies, countries etc., and columns representing variables (also called predictors or covariates) that represent socio-demographic data, credit facility data, transactional data, contextual data, etc. Popular examples of unstructured data are text, images, LiDAR, and graph data as discussed above.

Credit risk models can be boosted by carefully engineering the data taking into account the characteristics of the analytical technique that will be used to analyse it (Baesens, Höppner, and Verdonck, 2021). Structured data engineering can be decomposed into instance (also called observation) and feature engineering. Instance engineering is particularly relevant to PD modelling where we typically start from skewed data sets due to the low number of defaulters. This requires the adoption of special preprocessing routines, which are basically all variants of under- or oversampling. Throughout our empirical research and industry experience, we find that undersampling and SMOTE are usually the preferred methods to deal with skewed data sets (Zhu, Baesens, and van den Broucke, 2017). Hence, we do not expect much to be gained from yet another new instance engineering method.

Structured feature engineering concerns the careful crafting of features from variables in the data to improve either the statistical performance or interpretability of credit risk models (Baesens, Verdonck, Óskarsdóttir, and vanden Broucke, 2021). The most popular example of feature engineering in credit scoring is undoubtedly coarse classification or categorization which can be applied to both categorical or continuous variables. For categorical variables, it reduces the number of categories to deal with, hence requiring fewer parameters (or dummy variables) to be estimated and thus yielding more robust predictive models. For continuous variables, coarse classification can be beneficial to model non-monotonic effects of variables (e.g., age) on default risk. Coarse classified variables are then typically featurised using Weights-of-Evidence (WoE) coding which essentially encodes a monotonic relationship between the variable and the default target (Thomas, 2009; Thomas, Crook, Edelman, 2017). Another popular featurisation example is creating ratio features measuring, e.g., liquidity, solvency, and profitability in corporate credit risk which are typically highly predictive but also come with various caveats. First, special care should be taken such that the risk score can be obtained as a continuous function of the ratio. More specifically, it should be clear what happens if the denominator can become zero either during estimation or production time. Another problem relates to the sign of the numerator and denominator. If both numerator and denominator can have positive and negative signs (e.g. think about the net earnings/equity ratio) a positive ratio value may stem from either both a positive numerator and denominator, or both a negative numerator and denominator which obviously has not the same credit risk implications. Hence, this may require some tweaking of the ratio (e.g., restraining the denominator between a small value close to zero and a sufficiently high value or using the difference of both instead). Finally, some ratio variables tend to have fat-tailed distributions which may also need special preprocessing such as categorisation (Van Gestel, Baesens and Martens, 2024).

Temporal trend features are also an example of structured feature engineering measuring, e.g., evolutions in loan-to-value (LTV) ratios or behavioural scores which often provide early warnings of significant credit events (Baesens, Rösch and Scheule, 2016). Another key challenge in credit risk modelling is creating features for high-dimensional categorical variables such as SIC code and ZIP code which are both typically very relevant for PD, LGD, and EAD modelling. The traditional way of treating these is to group them using, e.g., a decision tree or Chi-squared analysis and then create features using the popular Weights-of-Evidence (WoE) method (Baesens, Rösch and Scheule, 2016). Although newer methods based on deep learning generated embeddings have been proposed, their empirical superiority has not been clearly demonstrated (Cheng and Berkahn, 2016). In some cases, the categorical feature can be re-encoded to an underlying continuous equivalent. For example, instead of using ZIP codes, Fernandes and Artes (2016) introduce kriging based on the latitude and longitude to define a feature measuring the spatial dependence among SMEs in their local neighbourhood in terms of default and demonstrate its significance in a traditional logistic regression model.

Also, variable transformations can be considered a type of feature engineering, such as the logarithmic, exponential, Box-Cox, and powerful Yeo-Johnson transformations (Yeo and Johnson, 2000). Essentially, these transformations create features that can model an exponential, saturation or other complex relationship with the target. We believe much more research is needed in structured feature engineering to understand which type of features work well for what type of portfolio, and credit measure (PD, LGD, and EAD), either in terms of statistical performance, interpretability, or profit (Van Gestel et al. 2005).

Unstructured feature engineering aims at generating numerical representations, also called embeddings, for text, images, geospatial, LiDAR, graph data, etc. Often this is done using powerful deep learning neural networks (e.g., convolutional neural networks, LSTMs, transformers, graph neural

networks, etc.) trained to generate condensed representations of the data by squeezing it into a lower dimensional space without significant loss of semantics. Unfortunately, many of these embeddings generated are merely black box numbers, which poses a serious research challenge in credit risk modelling due to the lost interpretability.

Bayesian Learning

Bayesian learning is gently starting to find its way in credit risk. One of the key strengths of Bayesian learning is that it allows to combine domain expertise or prior knowledge (often expressed as a probabilistic distribution) with patterns learned from data. This is particularly interesting in settings where there is a lack of data (such as exposures to sovereigns, banks, project finance, and new credit products) which often also coincides with a low number of defaulters, i.e. the well-known low default portfolio (LDP) problem. Bayesian learning also facilitates reasoning with missing data which is another of its key strengths.

Bijak and Thomas (2015) use Bayesian methods to build a single hierarchical model for LGD for retail credit and illustrate how this facilitates the generation of a predictive distribution for LGD as well as the quantification of downturn and stressed LGDs. Jobst, Kellner and Rösch (2020) use Bayesian learning to estimate the LGD of European sovereigns which is a typical example of a small data set with only two defaults: Greece and Cyprus. They extend their research to PD in Jobst and Rösch (2021) where they use Bayesian learning to analyse Eurozone sovereign real-world default probabilities and correlations, and compare regulatory and economic capital requirements. Tasche (2013) uses Bayesian learning for PD modelling of LDPs. He illustrates how upper confidence bounds for PDs can be obtained as quantiles of a Bayesian posterior distribution for a prior that is more conservative than the uninformed neutral prior. Bargaglia et al. (2022) use a Bayesian model averaging framework with an innovative selection prior to forecast GDP as we discussed above.

A Bayesian network is a graphical model or directed acyclic graph (DAG) with nodes depicting variables and arcs representing dependencies. It can be learned from data, expert knowledge or a combination of both. The network comes with conditional probability tables which can be used for inferencing. In other words, it can compute the probability of each node value for each configuration of other nodes even in case some are missing. The hardest part of building a Bayesian network is finding the network structure or DAG. The conditional probability tables are usually learned using either a maximum likelihood, simulation, or Expectation-Maximization (EM) approach. Bayesian networks are often referred to as probabilistic white-box models and are thus very promising for credit risk modelling, not only in terms of accuracy but also in terms of interpretability. Baesens, Egmont-Petersen, Castelo and Vanthienen (2002) constructed Bayesian networks using Markov Chain Monte Carlo (MCMC) search to estimate default probabilities for a retail credit data set. Ballester, López, and Pavia (2023) use Bayesian networks to study systemic credit risk transmissions among European sectoral CDS and find that the network relationships learned using conditional independence tests explain between 5% and 40% of systemic risks. Masmoudi, Abid, and Masmoudi (2019) learn Bayesian networks by combining expert knowledge and a Hill climbing procedure for a retail portfolio and find them to be highly interpretable for default risk.

Deep Learning

Deep learning is a research area that has been extensively researched and applied in numerous fields with great success (LeCun, Bengio, Hinton, 2015). Deep learning methods are based on a neural network backbone and tailor the architecture, objective function, and training methods to a variety of

tasks and data sources. They have been used for credit risk modelling with mixed success as we elaborate in what follows.

Gunnarsson, vanden Broucke, Baesens, Óskarsdóttir and Lemahieu (2021) compare state-of-the-art deep learning techniques to two ensemble methods (i.e. random forests and XGBoost) and two conventional methods for credit scoring (i.e. logistic regression and decision trees) on 10 data sets in terms of AUC, Brier score, partial gini and profit. All data sets are tabular or hence structured data. As a first conclusion, it is found that XGBoost is the best performing classifier on all performance measures considered, except for one where random forests was the best performing classifier. Secondly, deep networks with a number of hidden layers, i.e., deep learning, do not outperform shallower networks with one hidden layer. Therefore, one can conclude that deep learning algorithms do not seem to be appropriate methods for credit scoring and that two ensemble methods, XGBoost and random forests, should be preferred over the other credit scoring methods when statistical performance is the sole objective. Biatat, Crook, Calabrese and Hamid (2021) also find deep learning neural networks, i.e., multilayer perceptrons (MLPs) with four hidden layers, to not outperform more traditional machine learning methods such as XGBoost for credit scoring. However, as mentioned above, credit risk models should be interpretable. Hence, ensemble methods such as XGBoost or random forests can be used either as benchmarks or to find out if there are complex non-linearities or interactions in the data which one can then try to manually feature engineer using any of the methods discussed above (e.g., a Yeo-Johnson transformation) into a traditional logistic regression model. Though to the best of our knowledge, no similar benchmarking studies have been conducted for LGD/EAD modelling, we do not expect the findings to be any different with respect to the non-existing added value of deep learning techniques on structured data.

While the merits of deep learning on structured data are marginal to non-existing, this is definitely not the case for unstructured data. Earlier on we already discussed the predictive potential of text data, social media posts, Google Street View data, LiDAR data, and call graphs for boosting credit risk models. All of these can be analysed using deep learning techniques such as transformers, convolutional, or Long Short-Term Memory (LSTM) networks. These techniques typically learn condensed data representations, also called embeddings, which can then be used as additional features in a PD, LGD or EAD model to further boost its performance. Table 2 lists the deep learning architectures used for some of the data sources we described above.

Type of data used	**Type of deep learning method**	**Reference**
Loan officer text statements	BERT (Bidirectional Encoder Representations from Transformers	Stevenson, Mues and Bravo (2021)
Loan description text	Transformer	Zhang, Wang, Zhang, and Wang (2020)
Network data based on networks based on family ties and corporate ties	Graph convolutional networks; graph autoencoders	Muñoz-Cancino, Bravo, Ríos and Graña (2023)
Google Street View images	Convolutional neural networks	Gebru et al. (2017)
LiDAR data	Convolutional neural networks	Stevenson, Mues and Bravo (2022)
Social media posts	BERT; Multilayer Perceptron (MLP)	Yang, Yuan and Lau (2022)
User-generated descriptions from loan applications	Convolutional neural networks, recurrent neural networks, convolutional recurrent neural	Kriebel and Stitz (2022)

	networks, average embedding neural networks, BERT, robustly optimized BERT pretraining approach (RoBERTa)	
Mixed frequency data	LSTMs, Temporal Convolutional Networks, Transformer Encoder model for Panel data (TEP)	Korangi, Mues, and Bravo (2023)

Table 1 Deep learning architectures for unstructured data.

Based on these preliminary and limited studies combined with our own research experience, it can be conservatively stated that transformers are especially suited for textual data, convolutional networks for imagery, and graph neural networks for network data.

Recent developments in generative large language models (LLMs) (e.g., ChatGPT) which use a transformer-based deep learning architecture to automatically generate text trained on a massive corpus of text have catalysed public interest and awareness (Brown et al., 2020). Generative LLMs can be beneficial in various ways for credit risk modelling. First of all, they can be of assistance to help credit risk modellers build analytical PD, LGD, and EAD models by providing coding assistance, generation, and debugging facilities in SQL, R, Python, SAS, etc. Next, they can obviously also contribute in parsing and analysing unstructured data and verifying how this can help to build better credit risk models. Generative LLMs can also help in uncovering newer, previously unused sources of data, or any unexplored yet interesting combination thereof, to further boost the performance of credit risk models. They might also come in handy to help generate insights into the functioning of (black-box) credit risk models as alternative tools to, e.g., Shapley values and LIME (Molnar, 2022). They can also be leveraged to provide model documentation for already developed models and, as such, contribute to better model governance. Generative LLMs can also be helpful for stress testing as we discuss later. More specifically, they can be prompted to help re-enact historical stress scenarios. When we prompted ChatGPT 3.5 on April 13th, 2023 about "What exactly happened during the credit crisis of 2008 to GDP, inflation, unemployment, and house prices?", we got a quite extensive answer in terms of percentage changes of these measures for the US and Europe which was summarized as "Overall, the credit crisis of 2008 had a significant impact on the global economy, leading to a decrease in GDP, an increase in unemployment, a decline in housing prices, and low inflation rates.". Besides historical scenarios, these models can also contribute to devising hypothetical stress scenarios as we discuss below. It is also quite probable that in the foreseeable future generative LLMs will be used to generate draft credit scores, ratings, or even credit offers (e.g., in terms of interest rates, duration, collateral, guarantees) which can then be further finetuned with other (e.g., financial) information. Finally, also note that deep generative models (not language based) have already been successfully used to derive behavioural data based on application credit scoring data which essentially facilitates cross-selling or marketing campaigns at loan origination (Mancisidor, Kampffmeyer, Aas, and Jenssen, 2022).

Profit Driven Modelling

Many analytical techniques adopted for credit risk optimize a statistical objective function such as the Mean Squared Error (MSE) for a continuous target (e.g., LGD or EAD), or a maximum likelihood function such as the cross-entropy error for a categorical target (e.g., credit rating, PD). However, what really matters for modern-day lenders is the bottom-line impact measured as either profit gains or cost cuts. Obviously, this assumes a carefully crafted profit formula. Petrides, Moldovan, Coenen, Guns, and Verbeke (2022) introduce a method to estimate misclassification costs and evaluate a range of cost-

sensitive learning methods in terms of their ability to boost the profitability of application scorecards. Using data from a Romanian non-banking financial institution, they find that cost-sensitive models improve profitability across three business channels with a single-digit improvement for two of the channels and a double-digit boost for the third one. Verbraken, Bravo, Weber and Baesens (2014) introduce a profit-based classification performance measure for PD modelling which enables the selection of the most profitable PD model and provides the optimal cutoff point for the accept/reject decision. The profit formula considers both the expected profits and losses of credit granting. It builds further upon the Expected Maximum Profit (EMP) (Verbraken, Verbeke, Baesens (2012)) and finds a trade-off between the expected losses (based on EAD and LGD) and the operational income generated by the loan. They illustrate that the proposed profit measure outperforms classification accuracy and area under the ROC curve for selecting model parameters in terms of both accuracy and monetary value

A profit formula can be used in three possible ways during credit risk modelling (Vanderscheuren, Verdonck, Baesens, Verbeke, 2022). A first obvious one is for model evaluation where analytical models are being built in the traditional way but evaluated on a test set in terms of profit. A second option is to use profit for model selection such as determining the optimal values of (hyper-) parameters or features in a profit-driven way. For example, Kozodoi, Lessmann, Papakonstantinou, Gatsoulis, and Baesens (2019) use the Expected Maximum Profit (EMP) for feature selection using (regularized) logistic regression and extreme gradient boosting on ten credit scoring data sets. They illustrate that the resulting scorecards have a higher expected profit using fewer features than traditional feature selection strategies. Following the same idea, Maldonado, Bravo, López, and Pérez (2017) successfully use the EMP to propose a profit-driven approach for feature selection and classifier construction using linear Support Vector Machines on a Chilean credit scoring data set. The third option is the most ambitious one where the idea is to directly embed the profit measure into the estimation of the model parameters rather than optimizing a business irrelevant statistical objective function such as the MSE or cross-entropy. Stripling, vanden Broucke, Antonio, Baesens and Snoeck (2018) introduce ProfLogit which applies a real-coded genetic algorithm to maximize a profit measure when constructing a logistic regression model for churn prediction. Building upon this line of research, Höppner, Stripling, Baesens, vanden Broucke and Verdonck (2020) develop ProfTree, an extension of classical decision trees directly optimizing profit instead of traditional impurity measures (e.g., entropy, gini) and illustrate its monetary superiority again in a churn prediction setting. Both ProfLogit and ProfTree can be easily applied for credit scoring as well.

Model Use

To fully unleash the power of credit risk models, it is also important that they are used as much as possible by the firm. This very idea was actually early on already articulated in the Basel accord as the use test. Looking at contemporary firms, there is definitely room for boosting the impact of credit risk models by widening their scope of usage.

Besides provisioning and regulatory capital calculation, modern day credit risk models can also contribute to the credit approval decision by setting the decision threshold in a profit-driven way. In Verbraken, Bravo, Weber and Baesens (2014), we calculate the optimal cutoff value by taking into account the expected profits and losses of credit granting as quantified by the PD, LGD, and EAD. Taking this one step further, one can use the three credit risk parameters to do risk-based pricing or set the price or conditions (e.g., duration and collateral) of the loan based on the quantified risk.

Ideally, LGD models are detailed enough to fully understand the mechanics of the debt collection process. This allows to extend their usage towards pro-active debt collection and mitigating losses

whenever accounts tend to go into default. So, Mues, De Almeida and Thomas (2019) use a Bayesian Markov Decision Process (MDP) model to find an optimal policy of what action (e.g., telephone calls, formal letters, legal proceedings) to take when, for how long, and in what order given the current information on the individual debtor's repayment performance thus far in order to maximize the recovery rate or minimize the LGD.

Bringing the risk management approach, with its advanced risk models, more ingrained into the business is highly desirable. In particular, Landier, Sraer and Thesmar (2009) argue that risk management failures are not only driven by the inherent difficulty of measuring credit risk, but also by the organizational flaws that follow from the independence of the risk management function from the revenue-generating business lines. In particular, this separation of roles feeds a different appreciation of risk, certainly when different risk measures are being used. In addition, the effectiveness of risk management will crucially depend upon its power and status. Among others, Keys, Mukherjee, Seru and Vig (2009) show that default rates on mortgage loans are lower in banks where the risk manager's power and status are strong. Ellul and Yerramilli (2013) show that banks with a strong risk function experience lower tail risks during crisis years. Related, Kok, Müller, Ongena and Pancaro (2023) find that banking supervision is more effective in managing credit risk in institutions with a strong risk management culture.

Finally, a major challenge lies with the integration of credit risk with other risk models, i.e., the market risk model, the operational risk model, and the liquidity risk model, to consolidate all the risk exposures within the bank, and come to a comprehensive estimate of the amount of capital that is needed to cover the different risk losses (Hartmann, 2010). Clearly, just adding the different risk exposures ignores the complex non-linear dependencies, and thus diversification effects, that exist between the different risk types. One solution consists of using copulas to link the separate marginal loss distributions (see a.o. Rosenberg and Schuermann, 2006). An alternative approach consists of building an integrated risk model, which explicitly allows for the integration of dependence through different common risk drivers (Grundke, 2010, Bellini, 2013, Zhu, Wei and Li, 2021). While considered more accurate, the latter approach requires a good understanding of the common risk factors. With a lack of market valuations in credit risk, this remains an important and challenging unresolved task.

Stress Testing

Since the 2008 crisis, stress testing has become part of the standard toolkit of risk management practices and regulation. Initially, regulators had introduced stress testing as a crisis solution tool to identify capital shortfalls and increase market discipline. Nowadays, they use stress testing more as a crisis prevention tool to identify hidden vulnerabilities and to assess the resilience of individual banks and by extension, the banking system as a whole against adverse, mostly exogenous circumstances such as macroeconomic downturns or recessions (Baesens, Roesch and Scheule, 2016). If the regulatory stress test exercises unveil significant weaknesses, it can then lead to more stringent capital requirements. In addition, the compulsory reporting on the results of stress testing has an important information dissemination role. Numerous authors find that the disclosure of stress test results indeed affects market behaviour and as such increases market transparency and enhances market discipline (e.g., Ahnert, Vogt, Vonhoff and Weigert, 2020 and Fernandes, Igan and Pinheiro, 2020). While such market disciplining role of stress tests is not confirmed by Kok, Muller, Ongena and Pancaro (2023), these authors do show the effectiveness of stress testing in reducing risk-taking behaviour by banks. In particular, they show that banks that participate in a supervisory stress test reduce their credit risk exposures as compared to non-participating banks.

Apart from its usage by regulators, also financial institutions increasingly use stress tests for risk management purposes (Stein, 2012). It complements the traditional risk quantification methods that are mostly backward-looking. Stress testing then provides valuable forward-looking insights into the risk profile and draws attention to tail risk vulnerabilities in the business model and strategy. It is also key to understand the risks that come with financial innovation and for which no historical experience is available.

Stress testing is conducted using either sensitivity analysis or scenario analysis. Sensitivity analysis gauges the impact of an adverse change in one or a few risk factors on the overall credit risk of a portfolio, e.g., what is the effect of an x% increase in PDs, LGDs, CCFs, or correlations. Since the horizon in sensitivity analysis is typically short, it mimics an instantaneous shock. Technically, sensitivity analysis is thus quite straightforward and does not require too many resources. It is therefore appropriate when insights are needed either in a short period of time, or an a frequent basis. (BIS, 2000).

To capture the complexity and interconnectedness of risks, scenario analysis is more appropriate. This involves the simultaneous move in a range of risk factors, reflecting an extreme event that may occur in the future. In historical scenario analysis one relies on an actual event that occurred in the past, with the advantage of being able to use historical data and having to make fewer judgements or extrapolations. In hypothetical scenario analysis, on the other hand, one devises an event that has not yet happened (Breuer and Summer, 2020).

Hypothetical scenario analysis is the most complex, yet relevant way of stress testing in modern-day business environments. Only this approach can effectively account for the new types of risk that have recently emerged, such as pandemics, crypto-currencies, and strong geopolitical tensions. In this respect, the 2021 SSM-wide stress test unveiled that a prolonged COVID 19 impact would have a severe impact on capital buffers, mainly through loan losses (European Banking Authority, 2021). In addition, climate risk and the interconnectedness of the global financial world have explicitly been put forward as the new directions for macroprudential stress testing (de Guindos, 2021). As a consequence, the 2022 EBA stress test was focused on climate risk and it was found that many banks are still at an early stage in terms of factoring climate risk into their credit risk models and stress testing methodologies; the exercise also highlighted important data gaps and inconsistencies among institutions (European Central Bank, 2022). Key challenges here are how to predict losses over longer time periods (e.g., 30 years), how to quantitatively relate scenarios (e.g., drought, heat, and flood risk scenarios) to PD, LGD and EAD and how to anticipate changes in customer behaviour which may in turn affect PD, LGD and EAD. In addition, both Ferrari, Van Roy and Vespro (2021) and Grundke, Pliszka, and Tuchscherer (2020) demonstrate that model and estimation risk is considerable in credit stress test modelling. Extensive robustness checks are then crucial to be able to correctly interpret and use the results of stress test exercises.

Obviously, working out a comprehensive and relevant hypothetical stress scenario is a daunting and challenging task both in terms of identifying and quantifying all risks involved. Often times it boils down to extrapolating PD, LGD, and EAD models into unknown far away territory with little to no historical data nor business expert knowledge to guide us. As a simple example, most banks do not have data available on inflation shocks which are yet very actual and relevant. Also the effects of (hypothetical) stress scenarios stretch beyond any lender's boundaries, so supervisors should be well aware of the interconnectedness of the financial system so as to be able to properly identify knock-on failure effects and conservatively quantify the systemic risk. Finally, another important challenge relates to how much of the results of stress tests should be disclosed to financial markets (Caleb, Davis, Korenok and Lightle, 2022).

Model Risk

Fair Isaac Corporation (FICO) and the market intelligence firm Corinium released a report in 2021 which assesses how well companies are doing in adopting responsible artificial intelligence (AI). Some highlights from the report[2]:

- 65% of companies cannot explain how specific AI model decisions or predictions are made;
- 73% struggle to get executive support for prioritizing AI ethics;
- only 20% actively monitor their models in production;
- 30% of organizations report an increase in adversarial and other attacks against their models.

The numbers clearly highlight the fact that all AI or analytical models come with a certain degree of model risk. vanden Broucke and Baesens (2021) define model risk as: "*the risk of expected or unexpected loss resulting from the inadequate development or usage of analytical models across all business units and activities of the company.*" In fact, it was George Box (Box, 1976) who said that "*all models are wrong, but some are useful*" which nicely summarizes the essence of model risk. Obviously, this also pertains to PD, LGD, and EAD models. As shown in Figure 2, model risk essentially transversally encompasses the entire credit risk modelling process.

Knowing that model perfection is impossible is a good starting insight before using them; especially given the fact that credit risk models are being used for various purposes such as credit approval, provisioning, capital calculation, risk-based pricing, collections management, and credit limit setting. Model risk can occur during the various stages of building credit risk models. Table 2 provides an overview of the various components of model risk together with some examples in credit risk: data risk, specification risk, development risk, validation risk, operational risk, security risk, and managerial risk. Since the table only provides some mere examples, it is clear that model risk is material and should be properly accounted for.

Given its diverse and continuously changing components, it is a utopic ambition to unambiguously quantify model risk in a number and translate that to provisions and capital. A first step is being able to identify and qualify the various sources of model risk and trying to remedy it as much as possible. Next, whenever a source of model risk is identified, special caution and conservativeness (e.g., in terms of rating assignment, calibration, or additional buffers) should be adopted. For example, Baviera (2022) finds a significant dependency between PD and LGD in a corporate setting, and to cater for this type of model risk argues to apply a scaling factor of at least 1.4 to the risk-weighted assets (RWA).

[2] https://business-of-data.com/reports/state-of-responsible-ai-2021

Type of Model Risk	Examples in Credit Risk	References
Data risk	• Reject inference in PD modelling • Incomplete workouts in LGD modelling • Poor data quality • Lack of (default) data	• Banasik and Crook, 2007; Ehrhardt et al., 2021 • Rapisarda and Echeverry, 2013 • Moges, Dejaeger, Lemahieu, and Baesens, 2013 • Tasche, 2013 ; Van Gestel et al., 2005
Specification risk	• Imperfect definition of PD/LGD/EAD • Multicollinearity • Ratio discontinuities	• Van Gestel and Baesens, 2009; Brezigar-Masten, Masten, Volk, 2021 • Stine and Foster, 2017 • Van Gestel, Baesens Martens, 2024
Development risk	• Data leakage between train and test set • Correlation between PD, LGD and EAD • Parameter risk • Lack of model documentation	• Kapoor, S., Narayanan, 2022 • Da-Rocha Lopes and Nunes, 2010; Bellotti, 2017; Do, Rösch, Scheule, 2018 • Claußen, Rösch, Schmelzle, 2019 • Baesens, 2014; Baesens, Rösch and Scheule, 2016
Validation risk	• Unexpected signs of predictors • Wrong evaluation metrics • Overfitting	• Stine and Foster, 2017 • Verbraken, Bravo, Weber, Baesens, 2014 • Baesens, 2014
Operational risk	• Data drift • Output drift • Model overrides	• Rahmani et al., 2023 ; Baesens, Rösch and Scheule, 2016 • Nikolaidis, Doumpos, Zopounidis, 2017 • Angilella, S., Mazzù, S., 2019
Security risk	• Credit application fraud • Model exfiltration • Model backdooring	• Phua, Gayler, Lee, V., Smith-Miles, 2009 • Tidjon and Khomh, 2022 • Goldblum et al., 2022
Managerial risk	• Transition to Basel IRB approach • Regulation risk: Basel versus IFRS 9 • Model governance	• Merikas, Merika, Penikas, Surkhov, 2020 • Temin, 2016 • Doddi, 2021; OSFI, 2023; Baesens, Rösch and Scheule, 2016

Table 2 Model Risk in Credit Risk Modelling.

Suggestions for Future Research

Having conducted a literature review of more than 150 papers further augmented with our own recent research findings and industry experience, we can conclude by crystallising various lines for future research in credit risk modelling. To do so, we refer back to Figure 1 as our overall framework.

At the data level, we think many of the new data sources studied need further replication in other settings with other data from other portfolios (e.g., in terms of industry, geography) to improve the external validity of the findings. Also many of the data sources have been considered in isolation. Hence, more comprehensive studies where they are combined and their joint effect on credit risk investigated, would be highly welcomed. We also believe more research potential resides in feature engineering, especially in structured and unstructured feature engineering. More specifically, the development of newer methods that properly balance out the requirements of statistical accuracy, interpretability, profitability, ecological footprint, and regulatory compliance (see Figure 1) is an interesting topic for future research.

From a modelling perspective, we believe the potential of Bayesian learning in credit risk is still far bigger than explored up until now and consider this an important avenue for more research. Key research challenges are the better quantification and formalisation of expert knowledge through, e.g., prior distributions or draft DAG networks but also the usage thereof for small or low default portfolios. The first studies on the use of deep learning on unstructured data for credit risk are very promising. Also here, more research is needed to reinforce the generalisability of the findings. Given the ever-expanding set of newly developed deep learning techniques, it would be nice to have crystal clear guidelines on when to use what technique on what type of credit risk data, hence warranting the necessity of rock-solid benchmarking studies. Finally, the emergence of generative LLMs (e.g., ChatGPT) will offer great potential for not only the modelling of credit risk but also the management thereof and hence needs to be more thoroughly investigated.

Profit-driven evaluation and building of credit risk models provides an interesting and relevant new angle on the problem. Preliminary attempts have been made in the area of PD modelling. It would be interesting to further extend this research also to LGD and EAD modelling by first carefully thinking about how to define profit and then subsequently embed it in the development of the predictive models themselves. Given that a lender's actions typically impact this (e.g., think about actions considered during a workout LGD process), we are convinced reinforcement learning and/or Markov Decision Processes have a lot of value to add in studying this. Furthermore, extending profit driven model evaluation and building towards survival analysis models (for PD, LGD, or EAD) could also be an interesting yet challenging topic to study.

Widening the usage of credit risk models beyond regulatory purposes is another important avenue for further research. One popular example is risk-based pricing. Though this topic has raised interest from credit risk researchers for quite some time already, we still feel there is room for further study. One interesting approach concerns the application of causal machine learning in the pricing of loans. The idea here would be to study the effect of different treatments (in terms of, e.g., optimal interest rate, duration, guarantees, and collateral) on the customer successfully taking up the offer whilst simultaneously maximizing profit and thus minimizing default risk. Reinforcement learning and/or Markov Decision processes are definitely also worth further exploring in the context of LGD collection processes. A better understanding of the dynamics and potential cross-fertilisation effects between the risk management function and the revenue-generating business lines is highly desirable. Finally, a better understanding and quantification of the dependencies between credit, market, operational, and liquidity risks also prioritises high on the research agenda.

Despite its relevance and actuality, we find stress testing a much under-researched area. This is undoubtedly triggered by the complexity of the research questions such as how to integrate and stress various types of risk at aggregate/corporate level, how to deal with new types of emerging risks, how to adopt a much longer time horizon when developing stress scenarios and also how to validate these scenarios and adequately act upon the results thereof. We believe new research on stress testing should be characterized by a close collaboration between three types of partners: research institutions (e.g., universities), regulators, and the industry. Data pooling initiatives across firms, industries, and geographical regions are much needed to better qualify and quantify systemic risk and subsequently come up with new stress testing regulation.

As mentioned, model risk is an all encompassing term essentially spanning the entire model development process. The emergence of new technologies and exogeneous problems (e.g., deep learning, generative LLMs, climate risk, cybersecurity) continuously introduces new types of model risk that should first be identified and then properly quantified and managed. A key research challenge remains what to do with the outcomes of a model risk exercise, more specifically how to translate these into elevated capital floors and/or additional buffers.

Finally, most of the research on newer data sources, modelling techniques, evaluation metrics, model usage, and stress testing is still largely focussed on PD. Given the linear impact of LGD and EAD on capital requirements and provisions, we strongly advocate future research to focus more on these two key credit risk parameters, especially given the fact that the performance of contemporary LGD and EAD models is still very moderate with R-squared values seldom exceeding 20%. Also studying correlations between these risk parameters as well as asset correlations for both retail and corporate portfolios and how these may break down during periods of macroeconomic stress are relevant topics worth thoroughly investigating. We also plead for more reproducibility of existing credit risk research and ask for more publicly available data sets properly anonymised using the necessary procedures.

Conclusion

In this article, we survey 158 papers for credit risk modelling which we combine with our own research and industry experience to come up with a research agenda of topics that will help boost credit risk models. We start from a generic credit risk model architecture and corresponding model requirements. We review various new data sources and discuss feature engineering as a way to enrich the data for improved credit risk modelling. From a modelling perspective, we discuss how Bayesian learning and deep learning can contribute to better credit risk models depending on the source of data and modelling problem. In terms of evaluation, we advocate the use of profit-driven modelling instead of merely statistical modelling and evaluation. We also plead to use credit risk models more transversally for various activities and across all business units of the firm. We review some key challenges in stress testing and underscore its complexity. We introduce model risk as a concern spanning the entire credit risk modelling process and highlight various components thereof. The paper concludes by giving precise guidelines and topics for future research.

Acknowledgments

The authors wish to thank Harry Scheule (University of Technology, Sydney), Daniel Rösch (University of Regensburg), and Christophe Mues (University of Southampton) for their valuable feedback.

Appendix

DATA	**Text Data**
	• Loan officer text statements
	  ○ Stevenson, Mues and Bravo (2021)
	• Loan descriptions
	  ○ Zhang, Wang, Zhang, and Wang (2020)
	• SEC filings
	  ○ Nguyen et al. (2021)
	• News reports
	  ○ Cathcart, Gotthelf, Uhl, and Shi (2020)
	  ○ Aguilar, Ghirelli, Pacce, and Urtasun (2021)
	• Social media posts
	  ○ Bartov, Faurel and Mohanram (2022)
	Network data
	• CDR data
	  ○ Óskarsdóttir, Bravo, Sarraute, Vanthienen and Baesens (2019)
	• Geographic location/economic activity
	  ○ Óskarsdóttir and Bravo (2021)
	• Family ties
	  ○ Muñoz-Cancino, Bravo, Ríos and Graña (2023)
	• Corporate ties
	  ○ Muñoz-Cancino, Bravo, Ríos and Graña (2023)
	On-line data
	• E-mail
	  ○ Biatat, Crook, Calabrese and Hamid (2021)
	• Web browsing
	  ○ Rozo, Crook, and Andreeva (2023)
	• Social media data
	  ○ De Cnudde et al. (2019)
	  ○ Yang, Yuan and Lau (2022)
	Geographical data
	• Google Street View
	  ○ Gebru et al. (2017)
	• LiDAR
	  ○ Stevenson, Mues and Bravo (2022)
	Google Trends data
	• González-Fernández and González-Velasco (2020)
	• Woloszko (2020)
	• Bargaglia et al. (2022)
	Sustainability data
	• Moody's (2020)
	• Anand, Vanpée, and Baesens (2023)
	Asset correlation
	• Lee, Lin and Yang (2011)
	• Cho and Lee (2022)
	Target
	• PD
	  ○ Brezigar-Masten, Masten and Volk (2022)
	  ○ Dirick, Baesens and Claeskens (2017)

	Feature Engineering
	• General
	  ○ Baesens, Höppner, and Verdonck (2021)
	  ○ Baesens, Verdonck, Óskarsdóttir, and vanden Broucke (2021)
	• Instance engineering
	  ○ Zhu, Baesens, and van den Broucke (2017)
	• Structured feature engineering
	  ○ Coarse classification
	  ▪ Thomas (2009); Thomas, Crook, Edelman (2017)
	  ○ Ratio features/
	  ▪ Van Gestel, Baesens and Martens (2024)
	  ○ Trend features
	  ▪ Baesens, Rösch and Scheule (2016)
	  ○ High-dimensional categorical variables
	  ▪ Cheng, Berkahn, 2016
	  ▪ Fernandes and Artes (2016)
	  ○ Variable transformations
	  ▪ Yeo and Johnson (2000)
	  ▪ Van Gestel et al. (2015-
	• Unstructured feature engineering
	  ○ LeCun, Bengio, Hinton (2015)

MODEL	**Bayesian learning**
	• PD
	  ○ Jobst and Rösch (2021)
	  ○ Tasche (2013)
	  ○ Baesens, Egmont-Petersen, Castelo and Vanthienen (2002)
	  ○ Masmoudi, Abid, and Masmoudi (2019)
	• LGD
	  ○ Bijak and Thomas (2015)
	  ○ Jobst, Kellner and Rösch (2020)
	Deep learning
	• Transformers (e.g., BERT)
	  ○ Stevenson, Mues and Bravo (2021)
	  ○ Zhang, Wang, Zhang, and Wang (2020)
	  ○ Yang, Yuan and Lau (2022)
	  ○ Kriebel and Stitz (2022)
	  ○ Korangi, Mues, and Bravo (2023)
	• Convolutional networks
	  ○ Gebru et al. (2017)
	  ○ Stevenson, Mues and Bravo (2022)
	  ○ Kriebel and Stitz (2022)
	  ○ Korangi, Mues, and Bravo (2023)
	• Graph neural networks
	  ○ Muñoz-Cancino, Bravo, Ríos and Graña (2023)
	• Recurrent neural networks
	  ○ Kriebel and Stitz (2022)
	• LSTMs
	  ○ Korangi, Mues, and Bravo (2023)
	• Generative models
	  ○ Mancisidor, Kampffmeyer, Aas, and Jenssen (2022)

EVALUATION	**Profit driven modelling**

	• Profit formula for PD
	  ○ Petrides, Moldovan, Coenen, Guns, and Verbeke (2022)
	  ○ Verbraken, Bravo, Weber and Baesens (2014)
	• Profit driven model selection
	  ○ Kozodoi, Lessmann, Papakonstantinou, Gatsoulis, and Baesens (2019)
	  ○ Bravo, López, and Pérez (2017)
	• Profit driven model estimation
	  ○ Stripling, vanden Broucke, Antonio, Baesens and Snoeck (2018)
	  ○ Höppner, Stripling, Baesens, vanden Broucke and Verdonck (2020)

DEPLOYMENT	**<u>Model Use</u>**
	• Credit approval
	  ○ Verbraken, Bravo, Weber and Baesens (2014)
	• Debt collection
	  ○ Mues, De Almeida and Thomas (2019)
	• Risk management
	  ○ Landier, Sraer and Thesmar (2009)
	  ○ Keys, Mukherjee, Seru and Vig (2009)
	  ○ Ellul and Yerramilli (2013)
	• Risk integration
	  ○ Hartmann (2010)
	  ○ Rosenberg and Schuermann (2006)
	  ○ Grundke (2010)
	  ○ Bellini, 2013, Zhu, Wei and Li (2021)
	
	<u>Stress Testing</u>
	• Sensitivity Analysis
	  ○ Baesens, Rösch and Scheule, 2016
	• Scenario Analysis
	  ○ Baesens, Rösch and Scheule, 2016
	  ○ Breuer and Summer, 2020
	  ○ European Central Bank, 2022
	• Disclosure
	  ○ Caleb, Davis, Korenok and Lightle, 2022

References

[1] Aguilar, P., Ghirelli, C., Pacce, M., Urtasun, A. (2021). Can news help measure economic sentiment? An application in COVID-19 times. Economics Letters, 199, 109730. doi: doi.org/10.1016/j.econlet.2021.109730.

[2] Ahnert, L., Vogt, P., Vonhoff, V., Weigert, F. (2020). Regulatory stress testing and bank performance. European Financial Management, 26(5), 1449-1488. doi: doi.org/10.1111/eufm.12267.

[3] Anand, A., Vanpée, R., Baesens B. (2023). Sovereign credit risk modeling using machine learning: a novel approach to sovereign credit risk incorporating private sector and sustainability risks. Journal of Credit Risk, 19(1), 105-154. doi: doi.org/10.21314/JCR.2022.008.

[4] Andreeva, G., Ansell, J., Crook, J. (2007). Modelling profitability using survival combination scores. European Journal of Operational Research, 183, 1537-1549. doi: doi.org/10.1016/j.ejor.2006.10.064.

[5] Andreeva, G., Matuszyk, A. (2019). The law of equal opportunities or unintended consequences? The effect of unisex risk assessment in consumer credit. Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(4), 1287-1311. doi: doi.org/10.1111/RSSA.12494.

[6] Angilella, S., Mazzù, S. (2019). A credit risk model with an automatic override for innovative small and medium-sized enterprises. Journal of the Operational Research Society, 70(10), 1784-1800. doi: doi.org/10.1080/01605682.2017.1411313.

[7] Anthony, L. F. W., Kanding, B., Selva, R. (2020). Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. doi: doi.org/10.48550/arXiv.2007.03051.

[8] Arnould, G., Avignone, G., Pancaro, C., Zochowski, D. (2021). Bank funding costs and solvency. The European Journal of Finance, 28(10), 931-963, doi: doi.org/10.1080/1351847X.2021.1939753.

[9] Athanasoglou, P., Brissimis, S., Delis, M. (2008). Bank-specific, industry-specific and macroeconomic determinants of bank profitability. Journal of International Financial Markets, Institutions and Money, 18(2), 121-136. doi: doi.org/10.1016/j.intfin.2006.07.001.

[10] Baesens, B. (2014). Analytics in a Big Data World: The Essential Guide to Data Science and its Applications. Wiley, ISBN: 1118892704.

[11] Baesens, B., Egmont-Petersen, M., Castelo, R., Vanthienen, J. (2002). Learning Bayesian Network Classifiers for Credit Scoring using Markov Chain Monte Carlo Search. Proceedings of the Sixteenth International Conference on Pattern Recognition (ICPR'2002). IEEE Computer Society.

[12] Baesens, B., Höppner, S., Verdonck, T. (2021). Data Engineering for Fraud Detection. Decision Support Systems,150. doi: doi.org/10.1016/j.dss.2021.113492.

[13] Baesens, B., Rösch, D., Scheule, H. (2016). Credit Risk Analytics - Measurement Techniques, Applications and Examples in SAS. Wiley.

[14] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627-635. doi: doi.org/10.1057/palgrave.jors.2601545.

[15] Baesens, B., Verdonck, T., Óskarsdóttir, M., vanden Broucke, S. (2021). Special issue on feature engineering editorial. Machine Learning. doi: doi.org/10.1007/s10994-021-06042-2.

[16] Ballester, L., López, J., Pavia, J. M. (2023). European systemic credit risk transmission using Bayesian networks. Research in International Business and Finance, 65, 101914. doi: doi.org/10.1016/j.ribaf.2023.101914.

[17] Banasik, J., Crook, J. (2007). Reject inference, augmentation, and sample selection. European Journal of Operational Research, 183(3), 1582-1594. doi: doi.org/10.1016/j.ejor.2006.06.072.

[18] Bargaglia, L., Frattarolo, L., Onorante, L., Pericoli, F., M., Ratto, M., Pezzoli, L.,T. (2022). Testing big data in a big crisis: Nowcasting under COVID-19. International Journal of Forecasting. doi: doi.org/10.1016/j.ijforecast.2022.10.005.

[19] Bartov, E., Faurel, L., Mohanram, P. (2022). The Role of Social Media in the Corporate Bond Market: Evidence from Twitter. Management Science. doi: doi.org/10.1287/mnsc.2022.4589.

[20] Baviera, R. (2022). The measure of model risk in credit capital requirements. Finance Research Letters, 44. doi: doi.org/10.1016/j.frl.2021.102064.

[21] Beck, T., Behr, P., Guettler, A. (2013). Gender and banking: Are women better loan officers? Review of Finance, 17(4), 1279-1321. doi: doi.org/10.1093/rof/rfs028.

[22] Bellini, T. (2013). Integrated bank risk modeling: A bottom-up statistical framework, European Journal of Research, 230(2), 385-398. doi: doi.org/10.1016/j.ejor.2013.04.031.

[23] Bellotti, A.T. (2017). Estimating Unbiassed Expected Loss, with Application to Consumer Credit. SSRN Electronic Journal. doi: doi.org/10.2139/ssrn.2916145

[24] Beltratti, A., Stulz, R. (2012). Why did some banks perform better during the credit crisis? A cross-country study of the impact of governance and regulation. Journal of Financial Economics 105(1), 1-17, doi: doi.org/10.1016/j.jfineco.2011.12.005.

[25] Bequé, A., Coussement, K., Gayler, R., Lessmann, S. (2017). Approaches for credit scorecard calibration: An empirical analysis. Knowledge-Based Systems, 134, 213-227. doi: doi.org/10.1016/j.knosys.2017.07.034.

[26] Bessley, T., Roland, I., Van Reenen, J. (2020). The aggregate consequences of default risk: evidence from firm level data. ECB Working paper series no. 2425, 78pp.

[27] Betz, J., Kellner, R., Rösch, D. (2018). Systematic Effects among Loss Given Defaults and their Implications on Downturn Estimation. European Journal of Operational Research, 271, 3, 1113-1144. doi: doi.org/10.1016/j.ejor.2018.05.059.

[28] Biatat, V., A., D., Crook, J., Calabrese, R., Hamid, M. (2021). Enhancing credit scoring with alternative data. Expert Systems with Applications, 163, doi: doi.org/10.1016/j.eswa.2020.113766.

[29] Bijak, K., Thomas, L.C. (2015). Modelling LGD for Unsecured Retail Loans Using Bayesian Methods. The Journal of the Operational Research Society, 66, 2. doi: doi.org/10.1057/jors.2014.9.

[30] BIS (2000). Stress testing by large financial institutions: Current practice and aggregation issues, 44pp.

[31] BIS (2005). An explanatory note on the Basel II IRB risk weight functions, 19pp.

[32] BIS (2015). Developments in credit risk management across sectors: current practices and recommendations, 35pp.

[33] BIS (2016). International convergence of capital measurement and capital standards. A revised framework, 347pp.

[34] BIS (2017). Capital Floors: The design of a framework based on standardized approaches. Consultative Document, 12pp.

[35] BIS (2017). Finalising Basel III Reforms: In Brief, 9 pp

[36] BIS (2018). Structural changes in banking after the crisis. CGFS paper no. 60, 120pp.

[37] Bouteille, S., Coogan-Pushner (2021). The handbook of credit risk management: Originating, assessing, and managing credit exposures, 2nd ed, Wiley.

[38] Box, G. E. (1976). Science and statistics. Journal of the American Statistical Association, 71(356), 791-799.

[39] Breuer, T., Summer, M. (2020). Systematic stress tests on public data. Journal of Banking and Finance, 118. doi: doi.org/10.1016/j.jbankfin.2020.105886.

[40] Brezigar-Masten, A., Masten, I., Volk, M. (2021). Modeling credit risk with a Tobit model of days past due. Journal of Banking & Finance, 122. doi: doi.org/10.1016/j.jbankfin.2020.105984.

[41] Brooks, C., Schopohl, L. (2018). Topics and trends in finance research: What is published, who publishes it and and what gets cited? The British Accounting Review 50(6), 615-637, doi: doi.org/10.1016/j.bar.2018.02.001.

[42] Brown, I., Mues, C., Thomas, L.C.. (2010). Regression model development for exposure at default (EAD). Proceedings of the 24th European Conference on Operational Research.

[43] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D, Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877—1901, Curran Associates, Inc. doi: doi.org/10.18653/v1/2021.mrl-1.1.

[44] Caleb, C., Davis, D., Korenok, O., Lightle, J. (2022). Stress tests and information disclosure: An experimental analysis. Journal of Banking and Finance. doi: doi.org/10.1016/j.jbankfin.2022.106691.

[45] Cathcart, L., Gotthelf, N. M., Uhl, M., Shi, Y. (2020). News sentiment and sovereign credit risk. European Financial Management, 26(2), 261-287. doi: doi.org/10.1111/eufm.12219.

[46] Cheng, G., Berkahn, F. (2016). Entity Embeddings of Categorical Variables. doi: doi.org/10.48550/arXiv.1604.06737.

[47] Chiaramonte, L., Casu, B. (2017). Capital and Liquidity Ratios and Financial Distress. Evidence from the European Banking Industry. The British Accounting Review, 49(2), 138-161. doi: doi.org/10.1016/j.bar.2016.04.001.

[48] Cho, Y. and Lee, Y. (2022). Asymmetric asset correlation in credit portfolios. Finance Research Letters, 49, 103037. doi: doi.org/10.1016/j.frl.2022.103037.

[49] Claußen, A., Rösch, D., Schmelzle, M. (2019). Hedging parameter risk. Journal of Banking & Finance, 100, 111-121. doi : doi.org/10.1016/j.jbankfin.2019.01.003.

[50] Da-Rocha Lopes, S., Nunes, T. (2010). A simulation study on the impact of correlation between LGD and EAD on loss calculation when different LGD definitions are considered. Journal of Banking Regulation, 11, 156-167. doi: doi.org/10.1057/jbr.2010.7.

[51] De Cnudde, S., Moeyersoms, J., Stankova, M., Tobback, E., Javaly, V., Martens, D. (2019). Who cares about your Facebook friends? Credit scoring for microfinance. Journal of Operational Research Society, 70(3), 353-363. doi: doi.org/10.1080/01605682.2018.1434402.

[52] de Guindos, L. (2021). Macroprudential stress testing under great uncertainty. In: Is macroprudential policy resilient to the pandemic? Financial Stability Review, 24, 87pp.

[53] Demirgüç-Kunt, A., Detragiache, E., Merrouche, O. (2013). Bank Capital: Lessons from the Financial Crisis. Journal of Money Credit and Banking, 45, 1147-1164. doi: doi.org/10.1111/jmcb.12047.

[54] Dirick, L., Baesens, B., Claeskens, G. (2017). Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research Society, 68(6), 652–665. doi: doi.org/10.1057/s41274-016-0128-9.

[55] Do, H. X., Rösch, D., Scheule, H. (2018). Predicting loss severities for residential mortgage loans: A three-step selection approach. European Journal of Operational Research, 270(1), 246-259. doi: doi.org/10.1016/j.ejor.2018.02.057.

[56] Doddi, H. (2021). What Is AI Model Governance? Forbes.

[57] Doumpos, M., Figueira, J.R. (2019). A multicriteria outranking approach for modeling corporate credit ratings: An application of the Electre Tri-nC method. Omega, 82, 166-180. doi : doi.org/10.1016/j.omega.2018.01.003.

[58] EBA (2022). Risk dashboard – data as of Q3 2022, 55pp.

[59] Ehrhardt, A., Biernacki, C., Vandewalle, V., Heinrich, P., Beben, S. (2021). Reject Inference Methods in Credit Scoring. Journal of Applied Statistics, 48. doi: doi.org/10.1080/02664763.2021.1929090.

[60] Ellul, A., Yerramilli, V. (2013). Stronger risk controls, lower risk: Evidence from U.S. bank holding companies. Journal of Finance, 68(5), 1757-1802. doi: doi.org/10.1111/jofi.12057.

[61] European Banking Authority (2021). 2021 EU-wide stress test: Results, 64pp.

[62] European Central Bank (2022). 2022 Climate Risk Stress Test, 54pp.

[63] European Central Bank Supervision blog 2020. Who pays the piper calls the tune: The need for and benefit of strong credit risk management. December 4.

[64] European Systemic Risk Board. (2020). Macro-financial scenario for the 2020 EU-wide banking sector stress test.

[65] Fernandes, G.B., Artes, R. (2016). Spatial dependence in credit risk and its improvement in credit scoring. European Journal of Operational Research, 249(2), 517-524. doi: doi.org/10.1016/j.ejor.2015.07.013.

[66] Fernandes, M., Igan, D., Pinheiro, M. (2020). March madness in wall street: (What) does the market learn from stress tests? Journal of Banking & Finance, 112, 105250. doi: doi.org/10.1016/j.jbankfin.2017.11.005.

[67] Ferrari, S., Van Roy, P., Vespro, C. (2021). Sensitivity of credit risk stress test results: Modelling issues with an application to Belgium. Journal of Financial Stability, 52, 100805. doi: doi.org/10.1016/j.jfs.2020.100805.

[68] Fitzpatrick, T., Mues, C. (2021). How can lenders prosper? Comparing machine learning approaches to identify profitable peer-to-peer loan investments. European Journal of Operational Research, 294(2), 711-722. doi: doi.org/10.1016/j.ejor.2021.01.047.

[69] Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A. (2022). Predictably Unequal? The Effects of Machine Learning on Credit Markets. Journal of Finance, 77(1), 5-47. doi: doi.org/10.1111/JOFI.13090.

[70] Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E.,L., Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. Proceeding of the National Academy of Sciences, 114(50) 13108-13113. doi: doi.org/10.1073/pnas.1700035114.

[71] Getzner, J., Charpentier, B., Günnemann, S. (2023). Accuracy is not the only Metric that matters: Estimating the Energy Consumption of Deep Learning Models. doi: doi.org/10.48550/arXiv.2304.00897.

[72] Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, D., Song, D., Madry, A., Li, B., Goldstein, T. (2022). Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), 1563-1580. doi: 10.1109/TPAMI.2022.3162397.

[73] González-Fernández, M., González-Velasco, C. (2020). An alternative approach to predicting bank credit risk in Europe with Google data. Finance Research Letters, 35, 1-6. doi: doi.org/10.1016/j.frl.2019.08.029.

[74] Gopalakrishnan, B., Jacob, J., Mohapatra, S. (2021). Risk-sensitive Basel Regulations and Firms' access to credit: Direct and Indirect Effects. Journal of Banking & Finance, 126: 106101, doi: doi.org/10.1016/j.jbankfin.2021.106101.

[75] Grundke, P. (2010). Top-down approaches for integrated risk management: How accurate are they? European Journal of Operations Research, 203(3), 662-672. doi: doi.org/10.1016/j.ejor.2009.09.015.

[76] Grundke, P., Pliszka, K., Tuchscherer, M. (2020). Model and estimation risk in credit risk stress tests. Review of Quantitative Finance and Accounting, 55, 163-199. doi: 10.1007/s11156-019-00840-5.

[77] Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't. European Journal of Operational Research, 295(1),292-305. doi: doi.org/10.1016/j.ejor.2021.03.006.

[78] Hartmann, P. (2010). Interaction of market and credit risk, Journal of Banking & Finance, 34(4), 697-702. doi: doi.org/10.1016/j.jbankfin.2009.10.013.

[79] Höppner, S., Stripling, E., Baesens, B., vanden Broucke, S., Verdonck, T. (2020). Profit Driven Decision Trees for Churn Prediction. European Journal of Operational Research, 284(3), 920-933. doi:doi.org/10.1016/j.ejor.2018.11.072. doi: doi.org/10.1016/j.ejor.2018.11.072.

[80] Hull, J. (2018). Risk Management and Financial Institutions. 4th ed., John Wiley & Sons.

[81] Hyun, J.S., Rhee, B.K. (2011). Bank capital regulation and credit supply. Journal of Banking & Finance, 35: 323-330. doi:10.1016/j.jbankfin.2010.08.018.

[82] Jobst, R., Kellner, R., Rösch, D. (2020). Bayesian loss given default estimation for European sovereign bonds. International Journal of Forecasting, 36, 1073–1091. doi: doi.org/10.1016/j.ijforecast.2019.11.004.

[83] Jobst, R., Rösch, D. (2021). Euro zone sovereign default risk and capital-a Bayesian approach. Journal of Fixed Income, 31(3), 41–65. doi: doi.org/10.3905/jfi.2021.1.124.

[84] Kapoor, S., Narayanan, A. (2022). Leakage and the Reproducibility Crisis in ML-based Science. doi: doi.org/10.48550/arXiv.2207.07048.

[85] Keys, B., Mukherjee, T., Seru, A., Vig, V. (2009). Financial regulation and securitization: Evidence from subprime loans, Journal of Monetary Economics, 56(5), 700-720. doi: doi.org/10.1016/j.moneco.2009.04.005.

[86] Kok, C., Müller, C., Ongena, S., Pancaro, C. (2023). The disciplining effect of supervisory scrutiny in the EU-wide stress test. Journal of Financial Intermediation, 53, 101015. doi: doi.org/10.1016/j.jfi.2022.101015.

[87] Korangi, K., Mues, C., Bravo, C. (2022). A transformer-based model for default prediction in mid-cap corporate markets. European Journal of Operational Research, forthcoming. doi: doi.org/10.1016/j.ejor.2022.10.032.

[88] Korangi, K., Mues, C., Bravo, C. (2023). A transformer-based model for default prediction in mid-cap corporate markets. European Journal of Operational Research, 308(1), 306-320. doi: doi.org/10.1016/j.ejor.2022.10.032.

[89] Kozodoi, N., Jacob, J.,Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research, 297(3), 1083-1094. doi: doi.org/10.1016/J.EJOR.2021.06.023.

[90] Kozodoi, N., Lessmann, S., Papakonstantinou, K., Gatsoulis, Y., Baesens, B. (2019). A multi-objective approach for profit-driven feature selection in credit scoring. Decision Support Systems, 120, 106-117. doi: doi.org/10.1016/j.dss.2019.03.011.

[91] Kriebel, J., Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. European Journal of Operational Research, 302(1), 309-323. doi: doi.org/10.1016/j.ejor.2021.12.024.

[92] Landier, A., Sraer, D., Thesmar, D. (2009). Financial risk management: When does independence fail? American Economic Review, 99(2), 454-458. doi: doi.org/10.1257/aer.99.2.454.

[93] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep Learning. Nature, 521 (7553), 436-444. doi: doi.org/10.1038/nature14539.

[94] Lee, M. S. A., Floridi, L. (2021). Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs. Minds and Machines, 31(1), 165-191. doi: doi.org/10.1007/s11023-020-09529-4.

[95] Lee, S.-C., Lin, C.T., Yang, C.K. (2011). The asymmetric behavior and procyclical impact of asset correlations, Journal of Banking & Finance, 35, 2559-2568. doi: doi.org/10.1016/j.jbankfin.2011.02.014.

[96] Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136. doi: doi.org/10.1016/j.ejor.2015.05.030.

[97] Lim, C.Y., Woods, M. , Humphrey, C., Seow, J.L. (2017). The Paradoxes of Risk Management in the Banking Sector. The British Accounting Review, 49(1), 75-90. doi: doi.org/10.1016/j.bar.2016.09.002.

[98] Loterman, G., Brown, I., Martens, D., Mues, C., Baesens, B. (2012). Benchmarking Regression Algorithms for Loss Given Default Modeling. International Journal of Forecasting, 28(1), 161-170. doi: doi.org/10.1016/j.ijforecast.2011.01.006.

[99] Luong, T.M., Scheule, H., Wanzare, N. (2023). Impact of mortgage soft information in loan pricing on default prediction using machine learning. International Review of Finance, 23(1), 158-186. doi: doi.org/10.1111/irfi.12392.

[100] Makhlouf, K., Zhioua, S., Palamidessi, C. (2020). Machine learning fairness notions: Bridging the gap with real-world applications. Information Processing and Management, 58(5). doi: doi.org/10.1016/j.ipm.2021.102642.

[101] Maldonado, S., Bravo, C., López, J., Pérez, J. (2017). Integrated framework for profit-based feature selection and SVM classification in credit scoring. Decision Support Systems, 104, 113-121. doi: doi.org/10.1016/j.dss.2017.10.007.

[102] Mancisidor, R. A., Kampffmeyer, M., Aas, K., Jenssen, R. (2022). Generating customer's credit behavior with deep generative models. Knowledge-Based Systems, 245, 108568. doi: doi.org/ 10.1016/j.knosys.2022.108568.

[103] Martens, D. (2022). Data Science Ethics: Concepts, Techniques, and Cautionary Tales. Oxford University Press, ISBN: 0192847279.

[104] Masmoudi, K., Abid, L., Masmoudi, A. (2019). Credit risk modeling using Bayesian network with a latent variable. Expert Systems with Applications, 127, 157-166. doi : doi.org/10.1016/j.eswa.2019.03.014.

[105] Merikas, A., Merika, A., Penikas, H., I., Surkhov, M., A.. (2020). The Basel II internal ratings based (IRB) model and the transition impact on the listed Greek banks, 22. doi: doi.org/10.1016/j.jeca.2020.e00183.

[106] Miller, S., Noulas, A. (1997). Portfolio mix and large-bank profitability in the US. Applied Economics, 29(4), 505-512. doi: doi.org/10.1080/000368497326994.

[107] Moges, H.T., Dejaeger, K., Lemahieu, W., Baesens B. (2013). A multidimensional analysis of data quality for credit risk management: New insights and challenges. Information & Management, 50(1), 43-58. doi : doi.org/10.1016/j.im.2012.10.001.

[108] Molnar, C. (2022). Interpretable Machine Learning. ISBN: 979-8411463330.

[109] Moody's (2023). Private credit: A growing market with growing risks.

[110] Moody's. (2020). The Causes of Sovereign Defaults. Sovereign Default Series report.

[111] Muñoz-Cancino, R., Bravo, C., Ríos, S.A., Graña, M. (2023). On the combination of graph data for assessing thin-file borrowers' creditworthiness. Expert Systems with Applications, 123(A). doi: doi.org/10.1016/j.eswa.2022.118809.

[112] Nguyen, C. V., Das, S. R., He, J., Yue, S., Hanumaiah, V., Ragot, X., Zhang, L. (2021). Multimodal Machine Learning for Credit Modeling. In 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), 1754-1759), IEEE. doi: doi.org/10.1109/COMPSAC51774.2021.00262.

[113] Nikolaidis, D., Doumpos, M., Zopounidis, C. (2017). Exploring Population Drift on Consumer Credit Behavioral Scoring. In: Grigoroudis, E., Doumpos, M. (eds) Operational Research in Business and Economics. Springer Proceedings in Business and Economics. Springer, Cham. doi: doi.org/10.1007/978-3-319-33003-7_7.

[114] Office of the Comptroller of the Currency (2023). Fair Lending. https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/fair-lending/index-fair-lending.html.

[115] OSFI. (2023). A Canadian Perspective on Responsible AI.

[116] Óskarsdóttir, M., Bravo, C. (2021). Multilayer Network Analysis for Improved Credit Risk Prediction. Omega, 150. doi: doi.org/10.1016/j.omega.2021.102520.

[117] Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., Baesens B. (2019). The Value of Big Data for Credit Scoring: Enhancing Financial Inclusion using Mobile Phone Data and Social Network Analytics. Applied Soft Computing, 74, 26-39. doi: doi.org/10.1016/j.asoc.2018.10.004.

[118] Ozdemir, B., Miu, P. (2009). Basel II Implementation: A Guide to Developing and Validating a Compliant Internal Risk Rating System. McGraw Hill.

[119] Petria, N., Capraru, B., Ihnatov, I. (2015). Determinants of banks' profitability: Evidence from EU 27 banking Systems. Procedia Economics and Finance, 20, 518-524. doi: doi.org/10.1016/S2212-5671(15)00104-5.

[120] Petrides, G., Moldovan, D., Coenen, L., Guns, T., Verbeke, W. (2022). Cost-sensitive learning for profit-driven credit scoring. Journal of the Operational Research Society, 73(2), 338-350. doi: doi.org/10.1080/01605682.2020.1843975.

[121] Phua, C., Gayler, R., Lee, V., Smith-Miles, K. (2009). On the communal analysis suspicion scoring for identity crime in streaming credit applications. European Journal of Operational Research, 195(2), 595-612. doi: doi.org/10.1016/j.ejor.2008.02.015.

[122] Rahmani, K., Thapa, R., Tsou, P., Chetty, S. C., Barnes, G., Lam, C., Tso, C. F. (2023). Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. International Journal of Medical Informatics, 173. doi: doi.org/10.1016/j.ijmedinf.2022.104930.

[123] Rapisarda, G., Echeverry, D. (2013). A nonparametric approach to incorporating incomplete workouts into loss given default estimates. Journal of Credit Risk, 9(2), 47-61. doi: doi.org/10.21314/JCR.2013.159.

[124] Reynal-Querol, M., Garcia-Montalvo, J. (2020). Gender and credit risk: A view from the loan officer's desk, CEPR Press Discussion Paper (14500).

[125] Rösch, D., Scheule, H. (2020). Deep Credit Risk-Machine Learning in Python. Independently Published, United States.

[126] Rosenberg, J., Schuermann, T. (2006). A general approach to integrated risk management with skewed, fat-tailed risks. Journal of Financial Economics, 79(3), 569-614. doi: doi.org/10.1016/j.jfineco.2005.03.001.

[127] Rozo, B., J., G., Crook, J., Andreeva, G. (2023). The role of web browsing in credit risk prediction. Decision Support Systems, 164. doi: doi.org/10.1016/j.dss.2022.113879.

[128] S&P (2023). This month in credit. March vol. 3, 14pp.

[129] Scheule, H., Jortzik, S. (2019). Benchmarking LGD discount rates. Journal of Risk Model Validation, 14, 1. doi: doi.org/10.2139/ssrn.3673120.

[130] So, M., Mues, C., De Almeida, F., A., Thomas, L. (2019). Debtor level collection operations using Bayesian dynamic programming. Journal of the Operational Research Society, 70(8), 1332-1348. doi: doi.org/10.1080/01605682.2018.1506557.

[131] Stein, R.M. (2012). The role of stress testing in credit risk management. Journal of Investment Management, 10(4), 64-90.

[132] Stevenson, M., Mues, C., Bravo, C. (2022). Deep residential representations: Using unsupervised learning to unlock elevation data for geo-demographic prediction. ISPRS Journal of Photogrammetry and Remote Sensing, 187, 378-392. doi: doi.org/10.1016/j.isprsjprs.2022.03.015.

[133] Stevenson, M.P., Mues, C., Bravo, C. (2021). The value of text for small business default prediction: a deep learning approach. European Journal of Operational Research, 295(2), 758-771. doi: doi.org/10.1016/j.ejor.2021.03.008.

[134] Stine, R., Foster, D. (2017). Statistics for Business: Decision Making and Analysis. Pearson, third edition. ISBN-10: 0134497163.

[135] Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., Snoeck, M. (2018). Profit maximizing logistic model for customer churn prediction using genetic algorithms. Swarm and Evolutionary Computation, 40, 116-130. doi: doi.org/10.1016/j.swevo.2017.10.010

[136] Swiss Re (2022). Economic insights. Rising defaults: "zombie firms" will be the first to fall. No. 25, 2pp.

[137] Tasche D. (2013). Bayesian estimation of probabilities of default for low default portfolios. Journal of Risk Management in Financial Institutions, 6, 302-326. doi: dx.doi.org/10.2139/ssrn.2048818.

[138] Tasche, D. (2013). The art of probability-of-default curve calibration. Journal of Credit Risk, 9(4), 63-103. doi: doi.org/10.48550/arXiv.1212.3716.

[139] Temin, J. (2016). The IFRS 9 Impairment Model and its Interaction with the Basel Framework, Moody's Analytics Risk Perspectives.

[140] Thomas, L.C. (2009). Consumer credit models: Pricing, profit, and portfolios. Oxford University Press.

[141] Thomas, L.C., Crook, J., Edelman, D. (2017). Credit scoring and its applications. SIAM-Society for Industrial & Applied Mathematics; 2nd Revised edition. ISBN: 1611974550.

[142] Tidjon, L.N., Khomh, F. (2022). Threat assessment in machine learning based systems. doi: doi.org/10.48550/arXiv.2207.00091.

[143] Van Gestel, T., Baesens, B. (2009). Credit risk management: Basic concepts. Oxford University Press.

[144] Van Gestel, T., Baesens, B., Martens, D. (2024). Predictive analytics techniques and applications in credit risk modelling. Oxford University Press.

[145] Van Gestel, T., Baesens, B., Van Dijcke, P., Suykens, J.A.K., Garcia, J., Alderweireld, T. (2005). Linear and non-linear credit scoring by combining logistic regression and support vector machines. Journal of Credit Risk, 1(4), 31-60. doi: doi.org/10.21314/JCR.2005.025.

[146] vanden Broucke, S., Baesens, B. (2021). Managing model risk. ISBN: 979-8521686988.

[147] Vanderscheuren, T., Verdonck, T., Baesens, B., Verbeke, W. (2022). Predict-then-optimize or predict-and-optimize? An empirical evaluation of cost-sensitive learning strategies. Information Sciences, 594, 400-415. doi: doi.org/10.1016/j.ins.2022.02.021.

[148] Verbraken, T., Bravo, C., Weber, R., Baesens B. (2014). Development and application of consumer credit scoring models using profit-based V classification measures. European Journal of Operational Research, 238(2), 505-513. doi: doi.org/10.1016/j.ejor.2014.04.001.

[149] Verbraken, T., Verbeke, W., Baesens, B. (2012). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Transactions on Knowledge and Data Engineering, 25. doi: doi.org/ 10.1109/TKDE.2012.50.

[150] Verma, S., Rubin, J. (2018). Fairness definitions explained. Proceedings of the International Workshop on Software Fairness, 1-7. doi: doi.org/10.1145/3194770.3194776.

[151] Witzany, J. (2017). Credit risk management. In: Credit risk management. Springer, Cham, doi: doi.org/10.1007/978-3-319-49800-3_2.

[152] Woloszko, N. (2020). Tracking activity in real time with Google Trends. OECD Working Paper (1634).

[153] Yang, K., Yuan, H., Lau, R. Y. K. (2022). PsyCredit: An interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing. Expert Systems with Applications, 198. doi: doi.org/10.1016/j.eswa.2022.116847.

[154] Yeo, I. K., Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. Biometrika, 87(4), 954-959.

[155] Zhang, W., Wang, C., Zhang, Y., & Wang, J. (2020). Credit risk evaluation model with textual features from loan descriptions for P2P lending. Electronic Commerce Research and Applications, 42. doi: doi.org/10.1016/j.elerap.2020.100989.

[156] Zhu, B., Baesens, B., van den Broucke, S. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. Information Sciences, 408, 84-99. doi: doi.org/10.1016/j.ins.2017.04.015.

[157] Zhu, X., Wei, L., Li, J. (2021). A two-stage general approach to aggregate multiple bank risks. Finance Research Letters, 40, 101688. doi: doi.org/10.1016/j. frl.2020.101688.

[158] Zlotnik, A., Abraira, V. (2015). A general-purpose nomogram generator for predictive logistic regression models. The Stata Journal, 15(2), 537-546. doi: doi.org/10.1177/1536867X1501500212.