Alternative Credit Scoring of Micro-, Small and Medium-sized Enterprises

HONG KONG MONETARY AUTHORITY
香港金融管理局

ASTRI
香港應用科技研究院

Table of Contents

Table of Contents

Acknowledgements

An Advisory Panel of experts, industry players and lending business stakeholders was formed to provide advice on the development of an alternative credit framework for banks in Hong Kong. The members of the Advisory Panel have contributed substantially to the applied research presented in this white paper, through working group discussions, information- and experience-sharing, and participation in experiments and proof-of-concept implementations.

No.	Member of the Advisory Panel (in alphabetical order)	Representatives
1	AsiaPay	Mr Edward Ng, Mr Issac Leung and Mr Jeff Tam
2	Bank of China (Hong Kong) Limited	Mr Hsu Chi (Matthew) Hung, and Mr Tony Leung
3	CRD Association	Mr Kuwahara Satoshi and Dr Lan H. Nguyen
4	CRIF	Mr Vincenzo Resta and Mr Giordano Giulianelli
5	Deloitte Advisory (Hong Kong) Limited	Dr Paul Sin and Ms Joanne Lam
6	Dun & Bradstreet (Hong Kong) Limited	Mr James Chen
7	Ernst & Young Advisory Services Limited	Mr Sky So
8	Euler Hermes Hong Kong Services Limited	Dr Pablo Crotti
9	HKT Limited	Ms Monita Leung and Mr Sam Yee
10	HKTV Mall	Ms Jessie Cheng
11	Joint Electronic Teller Services Limited	Mr Angus Choi and Mr Ricky Lau
12	Moody's Analytics Hong Kong Limited	Mr Lawrence Antioch and Ms Miriam Chan
13	Nova Credit Limited	Mr Samuel Ho and Mr Albert Lo
14	Ping An OneConnect Bank (Hong Kong) Limited	Mr Choi Ka-fai
15	PwC Hong Kong	Mr Henry Cheng
16	Standard Chartered Bank (Hong Kong) Limited	Ms Vicky Kong, Ms Winnie Tung, and Mr Andrew Sim
17	TransUnion Limited	Mr Benny Siu and Mr Eric Cheung
18	Welab Bank	Ms Michelle Yim and Mr Paul Yip
19	ZA Bank	Mr Kelvin Tam

Foreword
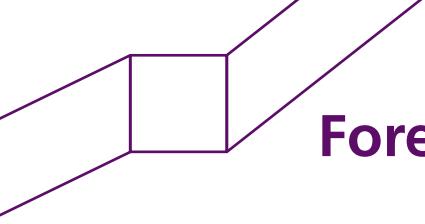
In Hong Kong, micro, small and medium-sized enterprises (MSMEs) play a vital role in supporting the economic development of the local business sector. There are currently over 340,000 MSMEs, accounting for over 98% of all enterprises and about 45% of the private sector employment in Hong Kong.

Despite their importance to Hong Kong's economy and employment, some MSMEs may encounter difficulties when accessing finance for their business growth and operations. Compared with large corporations, MSMEs may not have sufficient credit history and readily available financial records. Without such data and visibility of their business operations, it may be difficult for banks to assess MSMEs' credit worthiness. As a result, MSMEs may find it more difficult to obtain financing than larger corporations, which impedes their business expansion.

In an attempt to address the issue, this year, the Fintech Facilitation Office (FFO) of the Hong Kong Monetary Authority commissioned the Hong Kong Applied Science and Technology Research Institute (ASTRI) to conduct a study and explore the use of financial technologies to develop an alternative credit scoring framework for MSME lending businesses. A panel of industry experts were invited to participate in the study to contribute their insights, discuss the benefits and challenges of the proposed framework, and share potential industry use cases.

To facilitate and promote a wider use of alternative data, data availability and data sharing infrastructures are crucial. In 2018, the HKMA formulated the Open Application Programming Interface (API) Framework to facilitate data exchange between banks and third-party service providers (TSPs). Two years on, we are exploring a new data strategy and will consider building a new financial infrastructure, namely Commercial Data Interchange (or 'CDI'). CDI is a consent-based infrastructure that enables more direct, secure and efficient data flow between banks and sources of commercial data to enhance inclusive finance in Hong Kong. With CDI, we anticipate that enhanced financial products and services could be offered to MSMEs which are in full control of their own digital footprint.

Foreword
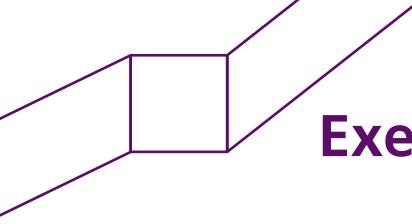
Lastly, we would like to express our gratitude to all the industry experts who contributed thematic articles and supported the experimental testing of the machine learning algorithms with ASTRI during the study. We hope that this paper will offer the industry some useful reference when considering the adoption of alternative credit scoring.

Edmond Lau
Senior Executive Director
Hong Kong Monetary Authority

Executive Summary

Background

Micro-, small and medium-sized enterprises (MSMEs) in Hong Kong are key players in the city's economy, and the primary source of employment. However, they often encounter difficulties borrowing from banks due to Hong Kong's lack of a credit information infrastructure and the significant burden faced by banks in conducting credit scoring and monitoring related processes. Against this background, the Hong Kong Monetary Authority (HKMA) engaged the Hong Kong Applied Science and Technology Research Institute (ASTRI) to explore how new financial technologies could be used to develop an alternative credit scoring framework for banks' MSME lending businesses.

The value of a non-traditional/alternative approach for evaluating the creditworthiness of MSMEs is gaining recognition in both developed and emerging economies throughout the world. Alternative credit scoring represents an emerging approach for both challenger and incumbent banks that enables innovative credit underwriting processes to be developed based on the analysis of alternative data through fintech.

The objectives of this white paper

This white paper first describes the meaning of alternative credit scoring and then how it works, and explains why it can help the banking industry and MSMEs in credit scoring. The paper also lays out the technological components needed to handle and process the alternative data used in alternative credit scoring. Further, it proposes building an effective alternative credit scoring ecosystem for banks and providers of alternative data in Hong Kong that can handle data management, credit scoring automation, and monitoring, and suggests steps that need to be taken by the players in the ecosystem to support this proposal.

Executive Summary

The intention of this paper is to promote the adoption of alternative credit scoring by banks in Hong Kong, with a view to improving access to finance for MSMEs and helping banks to improve the business scale of their existing MSME financing services. This paper could be used as a basic blueprint for banks looking to kickstart the adoption process. Various alternative credit scoring capabilities are required as part of the adoption process, including the following.

- **Data Management:** the ability to collect and manage alternative data for credit scoring.

- **Platform Automation:** the ability to develop a software platform to achieve the automation of credit underwriting and the monitoring of MSMEs' loan applications.

- **Model Innovation:** the ability to formulate alternative credit scoring models using AI and machine learning.

Structure of this paper

The contents presented in this paper are organised as follows:

- Part One: The creditworthiness of micro-, small and medium-sized enterprises

MSMEs are facing difficulties in getting loans from banks that use a conventional approach to assess their creditworthiness. These difficulties include the lack of financial and operating data for underwriting, governance weaknesses, and ineffective risk management capabilities. On the other hand, banks also find the conventional approach inefficient and costly for the processing of MSME loans that involve relatively small loan amounts. To address these issues, various types of data from third-party data providers, known as "alternative data", can be used to determine the creditworthiness of MSMEs. The new approach of utilising alternative data to evaluate a borrower's financial soundness and repayment capability is known as "alternative credit scoring". It aims to provide an all-round perspective on an MSME's creditworthiness and to augment the credit score generated by the conventional credit scoring approach.

As alternative credit scoring is directly driven by the nature and content of the alternative data, its implementation is determined by the type of alternative data used. A method for classifying alternative data is described in this white paper, based on the input provided by the members of the Advisory Panel. The white paper also describes and gives examples of two major classifications of alternative data, transactional data (e.g. cashflow data) and non-transactional data (e.g. company credit analysis reports).

As the benefits of using alternative data for credit scoring have become apparent to financial lenders, alternative credit scoring has been gradually adopted by banks, mission-driven lenders and fintech lenders in recent years. A wide variety of examples of alternative credit scoring being implemented in different countries are described in this paper.

- Part Two: Fintech infrastructural components for alternative credit scoring

To support the credit underwriting process for loan applications using alternative credit scoring, alternative data first needs to be collected from relevant data providers. Specific data fields are then extracted from the alternative data to enable default prediction to be performed with machine learning. A fintech infrastructure therefore needs to be developed that can source the data needed, structure the data into the format required for various machine learning models, manage data privacy concerns, and support final decision-making for loan applications.

Among the various types of alternative data, this paper focuses on the use of transactional data to assess the creditworthiness of MSMEs using machine learning models. The workflow for developing machine learning models for default prediction is a major fintech infrastructural component, and comprises the preparation of model and data, model building, and evaluation of the results.

Another key infrastructural component is an online lending platform. This is required to provide an execution environment for processing incoming alternative data, executing the machine learning model, and performing continuous reassessment of creditworthiness based on any up-to-date alternative data received.

Executive Summary

Addressing data privacy concerns in the fintech infrastructure is also critical for securing a stable and reliable supply of alternative data from third-party data providers. Part of this white paper outlines possible ways of addressing data privacy challenges arising from the need for data-sharing between banks and data providers by utilising privacy-enhancing technologies.

- Part Three: Technical evaluation of machine learning models

To demonstrate the technical feasibility of using machine learning for alternative credit scoring, two sets of experiments were conducted with different datasets of MSMEs. The first set aimed to evaluate the performance of the latest machine learning algorithms. Nine selected machine learning algorithms were tested on a rich dataset containing bank account data of MSMEs in Japan. The second set of experiments was intended to explore the technical feasibility of the industry-specific alternative credit scoring framework that is being proposed as a basic reference for the industry. To test the proposed framework, a proof-of-concept scenario was carried out based on the datasets of three participating organisations in Hong Kong, which include a bank and two third-party data providers (a point-of-sale payment data provider and an Internet payment data provider).

The key insights gained from the technical evaluation of the experiments are summarised below:

- The selected machine learning algorithms demonstrated different predictive power, but generally all were able to make effective default predictions based on different datasets of MSMEs' bank account information, including the dataset of MSME cashflow information.

- Banks could develop machine learning models that would achieve desirable short-term monthly prediction results, based on MSMEs' monthly bank statement data and MSMEs' monthly transactional data of cashflow collected from third-party data providers.

- Effective machine learning models for making short-term monthly predictions of problematic financial situations could be developed, based on MSMEs' transactional data from third-party data providers. The relevant data providers depending on the business sector of MSMEs could be used by banks for credit scoring.

- Part Four: Roadmap ahead

In summary, this paper first describes what is alternative credit scoring and how its usage can tackle the difficulties faced by MSMEs in getting loans from banks. It then outlines the fintech infrastructural components required to support the automation of credit underwriting by combining alternative and conventional credit scoring. To demonstrate the technical feasibility of using machine learning to develop alternative credit scoring, different machine learning models were tested and evaluated through various experiments on different MSME datasets.

The final part of this paper offers a roadmap for the adoption of alternative credit scoring in Hong Kong, and suggests three areas for future development. Firstly, continuous support by the government and infrastructure facilitation for data sharing are critical to maintain the availability of alternative data. Secondly, the continuous development of innovative machine learning models is required to enhance the handling of model validation, performance, data privacy, fairness, and interpretability. Finally, a centralised data-sharing platform could facilitate an ecosystem that would expedite the adoption of alternative credit scoring by banks in Hong Kong.

Part One:

The creditworthiness of micro-, small and medium-sized enterprises

The difficulties faced by small businesses wishing to get loans have limited their development and therefore affected the healthy growth of the overall economy. This section of the white paper first explains why difficulties arise for micro-, small, and medium-sized enterprises (MSMEs) seeking loans from banks that use a conventional approach to assess their creditworthiness. It describes how the conventional credit scoring approach is limited in its ability to deal with loan requests from MSMEs. It then introduces the concept of "alternative credit scoring" as an approach that is better suited to the loan needs of MSMEs. Alternative credit scoring offers banks the ability to expand the range of data that they use to assess an entity's creditworthiness. Whereas conventional credit scoring uses a limited range of financial data (mainly financial records), alternative credit scoring takes advantage of new technology to obtain and use new kinds of data (known as "alternative data") that can throw valuable light on an entity's creditworthiness. This alternative data may include information about, for example, an entity's trade payments, sales transaction records, credit analysis reports, and the behavioural traits of its business principals.

Alternative credit scoring is directly driven by the nature and content of the alternative data. The classification of the alternative data described in this paper therefore offers a generic way of categorising different alternative credit scoring approaches. Innovative approaches to credit scoring based on different alternative data are catching the attention of the financial industry. Notable industry examples are also described in this section to illustrate some of the initial efforts being made around the world to implement alternative credit scoring.

Part One

1. Challenges and opportunities relating to credit scoring for MSMEs

MSMEs generally face difficulties in getting loans from banks. This is because the conventional credit scoring process adopted by banks relies on analysing the financial statements of MSMEs to evaluate their creditworthiness. Many MSMEs fail to provide enough credible financial data to meet the requirements of conventional credit scoring. This section describes various challenges faced by banks in obtaining the financial data they need about MSMEs. It then explains why banks have begun to develop new, alternative approaches to credit scoring that leverage AI and machine learning.

1.1 Micro-, small, and medium-sized enterprises in Hong Kong

Different countries use different criteria to define micro-, small, and medium-sized enterprises (MSMEs). These criteria include an enterprise's assets, number of employees, sales turnover, and industry sector. The definition adopted by the Hong Kong SAR government is that an MSME[1] is a company that employs fewer than 50 persons (for non-manufacturing businesses) or 100 persons (for manufacturing businesses). MSMEs include micro-enterprises, which employ fewer than 10 persons[2]. Financial institutions also commonly classify MSMEs according to their annual turnover and loan size.

In Hong Kong, MSMEs in total account for more than 98% of business establishments and employ about 46% of the workforce in the private sector. Their continuing vitality and positive business performance are crucial for the continuing development of the local economy. However, financial institutions are cautious about lending to MSMEs because of the difficulty of obtaining credit information about them. Many MSMEs cannot borrow money without paying high interest rates on loans or offering tangible collateral.

1. A Report on Support Measures for Small and Medium Enterprises. (n.d.). TID. https://www.tid.gov.hk/english/aboutus/publications/smes/smes04_chapter2.html)
2. LCQ3: Measures to assist micro-enterprises and small and medium-sized enterprises. (2012). GovHK. https://www.info.gov.hk/gia/general/201205/30/P201205300299.htm

1.2 The conventional credit scoring approach

To measure the creditworthiness of a company, lenders use a credit scoring model that calculates the probability that a borrower will fail to repay loans in the future, known as the probability of default (PD). To facilitate better decision-making, lenders use mathematical models known as credit scorecards to quantitatively estimate whether a borrower is likely to display negative credit behaviour such as loan default, bankruptcy, or delinquency. Lenders decide whether to approve a loan by comparing the borrower's score with the cutoff score in the scorecard.
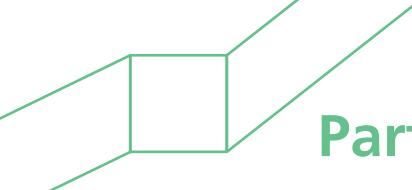
Since the pioneering works of Beaver (1966)[3], Altman (1968)[4], and Ohlson (1980)[5], statistical techniques have been applied to credit risk analysis. To statistically determine the significant predictors of default, the conventional approach mainly focuses on various financial ratios and financial structures based on data extracted from the financial statements of the borrower. These typically include the ratio of the loan to the borrower's total assets, the current ratio, the leverage ratio, the liquidity ratio, and the profitability ratio.

1.3 Key challenges in assessing MSME credit risk

Insufficient financial and operating data for underwriting

Financial institutions seeking to assess the creditworthiness of MSMEs face a serious constraint and source of inefficiency in the lack of transparent management information available. Many banks prefer to lend to large enterprises rather than MSMEs because these enterprises are able to provide clear audited financial statements. It can be difficult for banks to evaluate MSMEs because they often do not have solid accounting systems in place.

3.   Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71–111.

4.   Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609.

5.   Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 109–131.

In general, the main problem for lenders and MSME borrowers is asymmetry of information. MSMEs typically have a scanty digital and financial footprint compared with their larger counterparts. Large enterprises' financial information is generally more accessible to lenders seeking to estimate their creditworthiness. It is relatively costly for MSMEs to organise and present their financial documents in a form acceptable to banks for credit scoring. Making available a comprehensive, efficient, and lower-cost credit risk evaluation infrastructure is needed to address the current credit constraints on MSMEs.

Problems that banks typically encounter in MSME credit scoring include:

- Insufficient financial/audit invoicing information

- Lack of supply-chain information

- Lack of operational transparency

- Lack of know your customer (KYC) information concerning the business principal, which could enable lenders to evaluate potential adverse effects of his or her personal financial habits

Weaknesses in corporate governance

Governance weaknesses expose MSMEs to legal, regulatory, reputational, and investment risks. For example, an irresponsible or imprudent guarantee provided by an MSME to associated companies or external parties may become a liability that adversely affects its own operations. MSME owners may sometimes expand recklessly based solely on short-term goals. Such poor investment decisions are detrimental to the company's financial health.

Lack of effective risk management capabilities

Non-accounting data and risk control models may be missing or be inadequate within their risk control functions. Data relating to MSMEs is typically complex, diverse, and massively multidimensional, and frequently changes. This means that traditional data analytic methods and credit scoring techniques used by financial institutions are unable to generate warning signals about an MSME's operating status in a timely and accurate manner. For example, a sudden economic downturn may cause an MSME's upstream counterparts to impose unreasonable payment terms, while its downstream counterparts may default on payables or loans. Traditional credit scoring techniques are not agile or dynamic enough to assess the ability of MSMEs to withstand risk and weather adverse events like this.

Inefficient processes and infrastructure of financial institutions

Some financial institutions are still operating in a traditional manual mode. Conducting due diligence usually involves a great deal of manual work, including interviewing key personnel, going through paper files, and carrying out field visits. Given the relatively small loan amounts sought by MSMEs, the time and effort that lenders spend assessing their creditworthiness is usually disproportionate to the returns available. For this reason, financial institutions much prefer to process sizeable loans for large corporations.

Heavy reliance on collateral

Due to the difficulty in accessing relevant data, most loans to MSMEs are backed by collateral. This is usually of the brick-and-mortar variety, as its value can be easily assessed and liquidated. This approach is simple but always uncertain because of the volatility of collateral valuation (usually within the property market) and the control over such collateral. The collateral value cannot be precisely marked to market at any given time. Moreover, in volatile conditions such as the 2008 financial crisis in the US, when there was a sudden and huge contraction of collateral values, many loans backed by collateral are recalled. The volatility of collateral value is an extra factor that financial institutions have to consider in assessing loan applications.
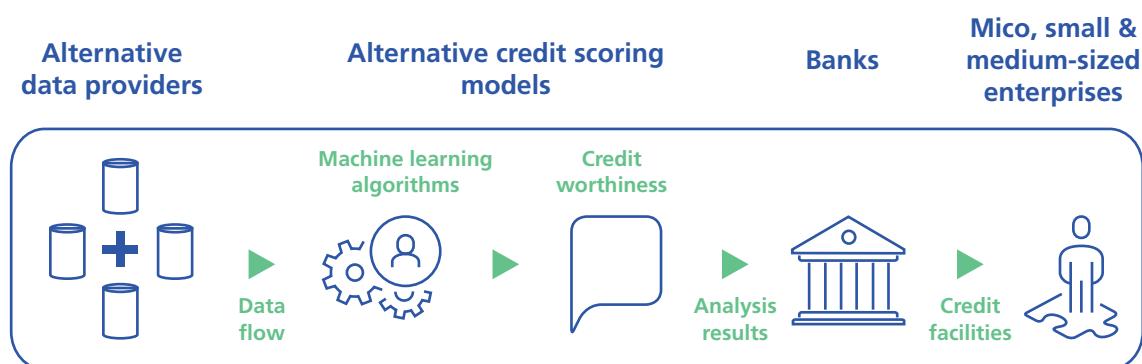
1.4 How does alternative credit scoring benefit banks and MSMEs?

Conventional credit scoring relies on specific financial ratios that are formulated based on asset value information and the financial statements of loan applicants. This conventional approach requires MSMEs to have a good credit history or asset-based collateral. Utilising alternative data as a substitute for traditional asset-based data to determine the creditworthiness of an MSME is an emerging approach for credit scoring. Alternative data is surrogate data from third-party data providers such as telecom companies, utility companies and social media platforms. It can also include analysed data from a wide variety of unconventional evaluation methods, for example data based on psychometric analysis that can be used to evaluate an individual's ability and willingness to pay, or data tracing digital activities on social media that can be used to evaluate the potential operational risk of a business. Utilising alternative data to evaluate a borrower's financial soundness and repayment capability is known as "alternative credit scoring". This new assessment approach aims to provide an all-round perspective on MSMEs' creditworthiness, and to augment the credit score generated by conventional credit scoring.

Alternative credit scoring not only lowers the barrier for MSMEs to acquire loans but also creates opportunities for banks to automate their credit underwriting processes. Driven by new trends in digital transformation, alternative credit scoring is significantly changing management practices in the credit industry. The conventional practice of credit scoring typically involves the manual input of personal, financial, and historical data into scorecards. Although this type of manual assessment has been successful in the past, it currently lacks two elements necessary to keep pace with regulatory, technological, and client-based changes.

The first element is the speed at which the assessment needs to be executed. With the technical advancements of computational hardware, current computers can support assessment models with higher complexity. Thus, financial institutions can perform credit scoring in a much shorter time, and can accelerate any upcoming processes if re-assessments are required. The second element is the larger amounts of data available for credit risk assessment than were fully available in the past, due to the current state of technology and access to cloud storage solutions. In the past (when storage solutions were unavailable or very expensive), only data considered to be relevant were kept. Hence, previous models that used only a few features (or variables) have now been replaced with more advanced models that use hundreds of features.

In the last decade, the use of artificial intelligence and machine learning has allowed diverse industries (and not simply financial services) to accelerate and improve their credit risk assessment processes. Access to high-end solutions on the cloud (e.g. the latest CPUs and GPUs) is one factor that has helped to accelerate the credit scoring process, while continuous improvements in the use of multiple algorithms for credit scoring has largely contributed to building more accurate models and generating more accurate credit risk assessments. By combining more powerful computational hardware and a larger amount of data for analysis, artificial intelligence and machine learning have provided many and diverse industries with better tools for assessing risk.

Figure 1.1 Workflow of alternative credit scoring by machine learning

Figure 1.1 shows how banks can leverage machine learning to automate the workflow of alternative credit scoring. With the benefit of operational cost-effectiveness, the manual processes of handling data and performing credit scoring can be replaced by a system that supports the collection of data from the providers of alternative data, the formulation of credit scoring results by machine learning algorithms, and the visualisation of analysis results to support decision-making for loan applications.

Acknowledgements for contributions to this section:
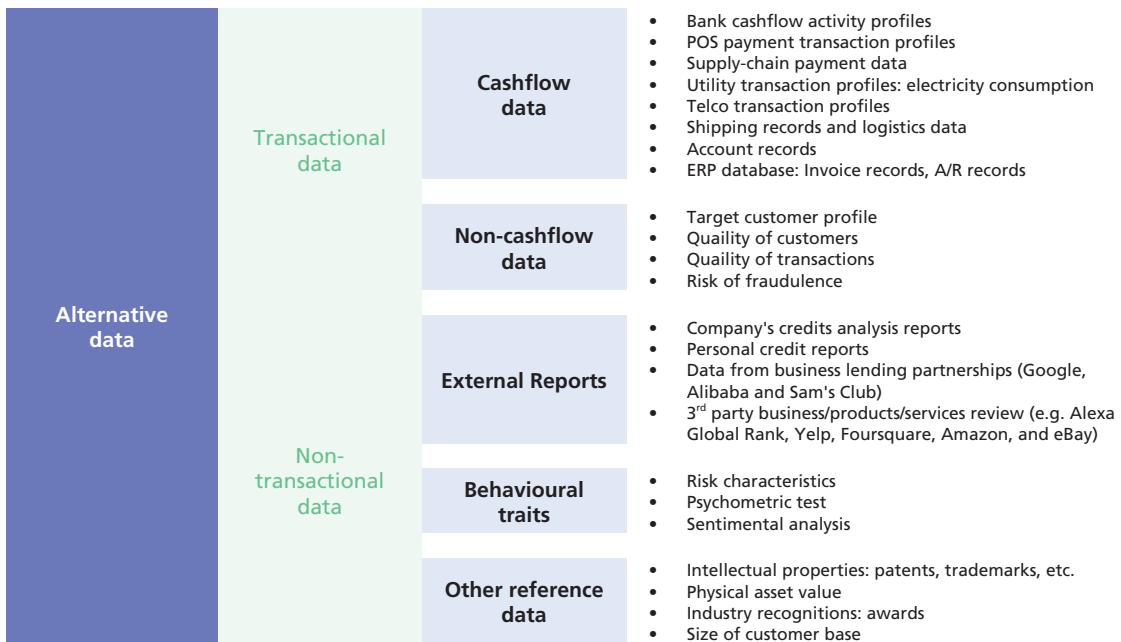
Company	Contributions
Nova Credit	Information on the key challenges involved in assessing MSME credit risk

2. Data for alternative credit scoring

Alternative data for credit scoring can take different forms. This section outlines a way for banks to identify and classify alternative data for the purpose of assessing an entity's creditworthiness, based on the input collected from the members of the Advisory Panel of this white paper. It describes and gives examples of two major classifications, transactional data (e.g. cashflow data) and non-transactional data (e.g. company credit analysis reports). It then explores the benefits and challenges that using alternative data can bring for lenders engaged in credit scoring.

2.1 Classification of alternative data for evaluation of creditworthiness

The conventional approach to credit scoring often uses financial ratios derived from financial statements and other third-party data to predict possible loan defaults within one to three years. Although this approach works for sizeable companies, it is not feasible for assessing MSMEs that lack sufficient reliable financial data to support a prediction. To tackle this challenge, other information cues can be used to assess the loan-repayment ability of MSMEs. These information cues are known as "alternative data". Alternative data are used to give lenders a better appreciation of the creditworthiness of MSMEs. Alongside conventional credit scoring data, alternative data can be used to generate supporting information that may give lenders a competitive edge in decisions about lending to MSMEs.

Alternative data	Transactional data	Cashflow data	• Bank cashflow activity profiles • POS payment transaction profiles • Supply-chain payment data • Utility transaction profiles: electricity consumption • Telco transaction profiles • Shipping records and logistics data • Account records • ERP database: Invoice records, A/R records
		Non-cashflow data	• Target customer profile • Quaility of customers • Quaility of transactions • Risk of fraudulence
	Non-transactional data	External Reports	• Company's credits analysis reports • Personal credit reports • Data from business lending partnerships (Google, Alibaba and Sam's Club) • 3rd party business/products/services review (e.g. Alexa Global Rank, Yelp, Foursquare, Amazon, and eBay)
		Behavioural traits	• Risk characteristics • Psychometric test • Sentimental analysis
		Other reference data	• Intellectual properties: patents, trademarks, etc. • Physical asset value • Industry recognitions: awards • Size of customer base

Figure 1.2 Classification of alternative data

Alternative data can take various forms, ranging from data based on observations of the borrower's operations to data relating to the business principal's personal risk characteristics and credibility. To come up with a method for classifying the different types of alternative data that are currently being used by industry players, input from the members of the Advisory Panel of this white paper (made up of experts, industry players and lending business stakeholders in Hong Kong) was collected. This is summarised in Figure 1.2.

2.2 Transactional data

"Transactional data" refers to the records of business activities between a company and its customers. They usually include revenue-related information (cashflow data) and nonmonetary-related information (non-cashflow data). The behaviour trends generated by an analysis of revenue-related information can be used to assess the latest financial status of a company. At the same time, nonmonetary-related information can produce insights useful in predicting a company's creditworthiness.

With machine learning techniques currently being used to perform trend analysis and default prediction, transactional data are becoming a promising type of alternative data for credit scoring. Opportunities to acquire transactional data are being created by open banking and OpenAPI initiatives, while financial institutions can also source transactional-based data from third-party data providers.

2.2.1 Cashflow data

Bank cashflow activity profiles
'Bank transactional data' refers to all of a bank account's cash inflows and outflows. Bank accounts are typically used by MSMEs for receiving revenue and settling payments. Generally, bank transactional data are held by banks in large quantities and are of credible quality. Only recently did banks realise that they could take advantage of this huge amount of data. Transactional data are also relatively easy to retrieve, as their recording is fully automated and they can be transferred between banks with the consent of the bank account's owner. With access to an MSME's bank accounts, banks can formulate a new type of credit scoring model using the MSME's bank transactional data.

Payment transactional data
In the retail industry, customer payment transactional data represent the sales activities of an MSME. Alternative credit scoring can be performed on transactional data to identify whether the revenue of an MSME for any given year/season/month is good or not. Although payment transactional data are not directly related to defaults on loans, they can be used to predict trends and patterns in an MSME's revenues. For example, with access to an MSME's payment transaction history, providers of online payment service platforms in the US (such as PayPal, Amazon, and Square) can offer it alternative lending services based on its sales data.

In the trading and logistics industries, supply-chain payment profiles capture the ability of MSMEs to pay their suppliers on time. Similarly, invoice information records can be used to analyse the status of and trends in an MSME's revenue streams.
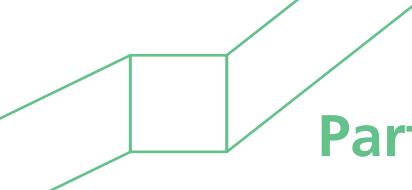
2.2.2 Non-cashflow data

Besides cashflow data, transactional data records carry information that is not related to cashflow but can be used to determine the quality of an MSME's business, i.e. non-cashflow data.

- Quality of customers — The payment identifications of the transaction records can be analysed to identify the profiles of customers and ascertain if they are recurring or one-time customers.

- Target customer profile — The transaction values and patterns in the transaction records can be used to categorise the spending profiles of customers.

- Quality of transactions — The percentages of cancelled, reversed, and voided transactions can reveal the quality of engagement with customers.

- Risk of fraudulence — Non-cashflow data provide credible sources of information relating to risk factors and problematic transactions, for fraud detection and credit-related analysis.

2.2.3 An example of transactional data: Telecommunication company transactional data

Telecommunication providers capture various data points relating to mobile phone usage. Telecommunication company (telco) transactional data contain information about call duration, call origin locations, call destinations, porting history, monthly bill payments, handsets used, number of missed calls, and more. Both cashflow and non-cashflow data are included in the transactional records. Cashflow data support the analysis of an MSME's cash outflow patterns, and non-cashflow data facilitate the analysis of an MSME's existing customer profiles and its patterns of business behaviour.

Cashflow data:

- Billing/payments data — Billing data can be used to study an MSME's payment history. Data such as whether all telco bill payments have been made on time or if there is a habitual delay can provide insights into an MSME's behavioural tendency to make late payments.

- Mobile wallet and e-commerce transactions — Telco data can provide insights into e-commerce transactions, company websites, E-money usage, use of financial applications, Internet advertising, and social reach, all of which are available from Internet usage history. Nowadays, many MSMEs use mobile wallets to send and receive money, especially if they are exposed to retail customers. The number of transactions, ticket size, and frequency of e-commerce of an MSME can provide good insights into its cash flow and the stability of this flow. These are very valuable inputs for evaluating an MSME.

Non-cashflow data:

- Call and SMS data — Frequently dialled numbers show who and where calls are being made to. If the numbers are those of stakeholders related to the business, this is evidence that the business is genuine.

- Subscription data — Subscription choices (e.g. the choice between a prepaid account and a post-paid account and the choice of value-added services (VAS)), and the kinds of mobile handsets used, can help determine an MSME's level of affluence. If the MSME owner maintains a prepaid account with erratic top-ups, this may indicate irregular income. On the other hand, a post-paid account shows a monthly commitment to pay and stability in cashflow if paid regularly and without delays.

- Customer data — Telco data can be used to verify the information provided in a loan application by an MSME or its owners, including the company office address and/or personal addresses.

2.3 Non-transactional data

2.3.1 Data from credit reference agencies
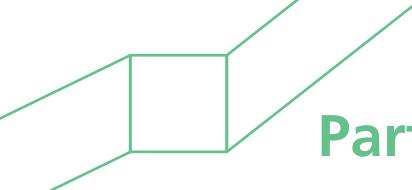
Third-party data providers can provide historical and current data that may add value to existing available in-house data. To illustrate this, consider a situation where a credit scoring model is required to assess the viability of a business that includes, for example, information on the number of years the business has been running and the number of times it has failed to pay its rent. Some of this information is readily available, such as the number of years it has been in business, but historical data regarding rent payments is lacking. In this case, a third-party data provider (such as a credit reference agency) may be able to supply the missing data points for the credit scoring model, if available. In summary, a third-party data provider can add value to current data available in-house. This value may come from supplying missing data or adding new information that is currently unavailable in-house.

Company credit analysis reports
Quantitative data from credit bureau reports include reference data that can be very useful for assessing the creditworthiness of MSMEs. The Commercial Credit Reference Agency (CCRA) was established in Hong Kong under an industry initiative supported by the Hong Kong Monetary Authority. The CCRA in Hong Kong is an organisation that collects information about the indebtedness and credit history of business enterprises and makes this information available to lending institutions. After receiving consent from an MSME, lending institutions can check with the CCRA about the MSME's credit record to help them assess the loan application. The CCRA increases lending institutions' knowledge of borrowers' credit records, expedites the loan approval process, and helps strengthen lending institutions' credit risk management. The fact that information about borrowers can be exchanged by lending institutions also incentivises borrowers to repay their loans and helps to reduce the overall default rate.

Personal credit history
The personal credit history of a borrower's key personnel can provide lenders with a view into the lending and repayment behaviour of these individuals. As these people are the decision-makers in the company, their credit behaviour is likely to have an influence on the company's credit behaviour. For example, if the proprietor is repaying debt regularly without default, then it is very likely that the company will do the same.

The relevant parameters for evaluating MSMEs that are available from personal credit histories include:

- Repayment history — Repayment history is very important because it says a great deal about owners' behavioural characteristics and their willingness to make timely payments. An owner's credit history can be used to ascertain whether that person makes consistent repayments and displays responsible credit behaviour. Seeing a positive attitude towards repayment and a good repayment history on the part of the owner will give financial institutions the confidence to extend credit with low credit risk in cases where the owner makes decisions on behalf of the MSME.

- Personal debt history — This considers whether an individual has credit available that they are not using and whether they can access such credit if the business needs it. This can be calculated by dividing an owner's outstanding debt balances by their total available revolving credit. If the personal debt usage is larger than the total revolving credit, then the credit risk of the MSME will be on the higher side; if smaller than the total revolving credit, then the credit risk of the MSME will be lower.

- Types of credit — Prompt repayments by an individual who has a basket of different types of credit (such as a credit card and a personal loan) show that individual's ability to manage different kinds of loans/credits/liabilities. A business may require a number of different kinds of loans, such as a working capital loan and a loan for capital investments. If the owner demonstrates the ability to manage different types of loans in his or her individual credit life, then the credit risk of the MSME will be considered lower; if they have a poor record in this area, then the credit risk of the MSME will be considered higher.

- Length of credit history — If the owner has a long credit history, it indicates that the business is in the hands of a person with a good deal of experience in handling credit. If this is coupled with consistent on-time repayments, it indicates that the owner will be likely to repay business loans on time if credit is extended. This will also lower the credit risk.

- New credit enquiries — Each time borrowers request a loan, lenders look to obtain a personal credit report on them. These hard inquiries create a record in the applicant's credit file which normally remains there for a considerable amount of time. A large number of hard inquiries in a personal credit report may indicate that the individual has had many credit applications rejected. This creates doubt, and is an indicator of high risk for MSME lending.

2.3.2 Behavioural traits of business principals

A psychometric test is a standardised tool used to objectively assess traits that are not visible on a physical level (such as personality, intelligence, motivators, and needs). A psychometric test is used to understand an individual by better understanding their personality, achievement orientation, intelligence, needs, or motivators, for instance. These constructs are used for the psychometric test as they are usually found to be consistent, they can be mapped in an individual, and they can be used for profiling.

Psychometric tests can provide valuable insights into an individual's ability and willingness to repay loans. They can also provide valuable insights into an MSME's future by assessing the personalities of its key individuals. Depending on the legal structure of the MSME, the key person whose personality influences the MSME's prospects will be different.

- Sole Proprietorship: The sole proprietor is responsible for all acts performed in the capacity of business owner. Thus, the relevance of personality assessment is very high.

- Partnership: Although certain forms of partnership (e.g. Limited Liability Partnerships) make the owner responsible for only certain types of debts, the majority of the decisions are taken by the partners. Hence a personality assessment of the partners can provide major insights into the prospects of the enterprise.

- Limited company: A limited company has its own corporate identity, and the company's liability is not the liability of its shareholders. Thus, personality assessment is not especially relevant, and is not needed for this class of MSMEs.

An example of a psychometry assessment from CRIF

A notable example of the deployment of psychometry assessment for MSMEs is a platform called MAP. The platform, offered by CRIF, can help map credit risk based on an individual's personality. The psychometric test is essentially a gamified quiz that engages users by putting them in everyday life situations and assessing their personality traits based on their responses to these situations.

MAP adopts a multi-layered decision framework that utilises a metadata score, an application score, and a psychometric score to assist in decision-making for loan applications.

- Metadata score: The metadata component monitors aspects of the response behaviour during the test, such as the speed of response and the number of times the answers are changed, to assess the reliability of the test. The metadata component also looks at the consistency of responses and at possible scenarios where the applicant may be trying to trick the system (e.g. by going too fast, changing responses, or "faking good"). These aspects are captured during the entire test, and a customer that does not appear to be genuine will be flagged as unreliable.

- Application score (A-score): This is a traditional risk assessment methodology that uses demographic information about the applicant such as age and education. It is a proven methodology, commonly used by financial institutions. MAP provides a plug-in placeholder to collect information related to the application scorecard. This placeholder reduces the need for multiple touchpoints, and all of the information required for making decisions can be captured in one go.

- Psychometric score (P-score): This is an innovative alternative type of risk scoring that evaluates the applicant's personality type. The psychometric component helps to uncover hidden traits. It studies approximately 10 personality traits, such as discipline and guilt-proneness. The questions in the psychometric test do not ask about the individual's money-spending behaviour directly, but are indirect and cover everyday life situations in the region where the applicant lives. Thus, the test is customised for certain cultural and regional contexts.

2.4 Benefits and challenges for lenders of using alternative data for credit scoring

There are both benefits and challenges for lenders in adopting alternative data for credit scoring.

Benefits:

- Information advantage

  Alternative credit data can present more insights into a MSME's creditworthiness. Banks can use it to make better-informed decisions on approving loan applications by MSMEs.

- Credibility
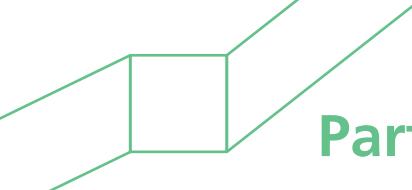
  Data from third-party sources are more reliable, whereas financial data are subject to accounting manipulation, making it much easier to reduce lending risk with alternative credit data.

- Fraud detection

  Alternative data are available through digital automation, and machine learning can help to detect abnormal patterns of business operations, with which lenders are able to detect fraud and implement risk mitigation measures.

- Continuous monitoring

  With alternative data, the lender can monitor the borrower's actual business situation, getting a more complete and comprehensive view of a consumer's creditworthiness.

Challenges

- Data quantity

    There must be enough alternative data available to build the machine learning models. By nature, alternative data is more difficult to process than financial data because the data format is often unstructured.

- Data quality

    It is extremely important to guarantee the quality of the alternative data when creating a reliable risk assessment model, as data points containing no value or having a large variance can compromise the output of the model.

- Data privacy

    Personal data or data that in aggregate can be used to piece together an individual's identity have become a lightning rod for regulators. Data protection and privacy laws will likely be applicable if alternative data sources contain personal data.

- Model fairness

    Using the correct data in the machine learning model is crucial for ensuring the appropriateness of that model. Indeed, the dataset is often the first place where bias is introduced into a model, and this situation also applies to alternative data.

- Special engineering efforts

    The adoption of alternative data requires engineering efforts in the areas of data science and machine learning. Lack of relevant human resources will hinder the development of alternative credit scoring, because banks need talents in these areas to adopt this new approach.

These challenges and benefits coexist. There are still many uncertainties regarding the future adoption of alternative data for use in credit scoring. A key question is how best to combine the use of alternative data with conventional financial data to improve credit scoring performance.

Acknowledgements for contributions to this section:

Company	Contributions
CRIF	Information on personal credit data, telco data, and psychometric tests for MSMEs

3. Industry examples of alternative credit scoring

As an emerging credit scoring approach for MSMEs, alternative credit scoring has been gradually adopted in recent years. The rate of adoption is however not evenly distributed worldwide. Alternative credit scoring is being used not only by banks but also by other loan providers, including mission-driven lenders and fintech lenders through their online lending platforms. Examples of the adoption of alternative credit scoring by industry players in different countries suggest that the implementation of this innovative credit scoring approach will continue to pick up.

3.1 The global landscape of MSME loan lenders

Banks

The adoption of alternative credit scoring varies between countries. Some banks in Japan use MSMEs' cashflow data for credit scoring, including Resona, Mitsubishi UFJ, Sumitomo SBI Net Bank, Mizuho Bank, and a few others. Resona Bank, one of the four Japanese megabanks, has recently introduced a credit line that only requires MSMEs to have bank accounts at the bank for a certain period to be eligible for loan screening. The credit line does not require collateral, guarantees, or the submission of financial statements. Screening is mainly done by machine learning algorithms that analyse the cash movements of the applicant's bank account. The financial costs range from 3% to 9%, only slightly higher than the rates for traditional loans. The lending amount is capped at 10 million yen, or approximately US$100,000 (1 US$ = 100 yen), apparently to cover the potential risk associated with this newly introduced assessment method.

Most mid-tier and cross-regional commercial banks in China, such as Xinwang, Bohai, and Fubon, use data such as income tax and business tax records for credit scoring. Online application approvals are usually based on credit data from the People's Bank of China as well as the tax bureau, which consist of the company's revenue data for the past few months. The use of automated underwriting algorithms means that such applications usually take only a few minutes to complete.

MSME lending is creating opportunities for challenger banks. Equipped with the latest fintech and brand-new credit underwriting approaches, virtual banks are attracting MSME financing business by offering smaller loans and lower default interest rates to borrowers. One of the virtual banks in Hong Kong, PAO Bank, started offering loan products to MSMEs in 2020 based on AI and innovative use of alternative data.

Mission-driven lenders

Sometimes called "community development financial institutions" (CDFIs), mission-driven lenders are located throughout the US. Originally, they were non-profit small business lenders that offered loans to businesses left behind by the traditional financing market. These lenders focus on "thin file" business owners in underinvested communities, who often lack collateral, financial documentation, or a reliable credit history. Loan sizes typically range from US$500 to US$5 million, with most falling between US$30,000 and US$100,000. The nature of this business means that a large part of the loan underwriting process is still based on specialised, judgemental input. CDFIs usually possess in-depth knowledge about the specific industry, operating methodology, and geography of the segment concerned, making them experts with the ability to assess loans for MSMEs in that particular environment even in the absence of financial data. Apart from credit scoring, CDFIs can at times give very specific advice to MSMEs to help enhance their business operations and thus increase their ability to pay back the loans. Having a CDFI in a specific industry or cluster in the community is the most valuable asset on which these CDFI lending businesses are built. These lenders also focus more on automation, for example by using open financial platforms such as Plaid to capture financial and transactional data. Some notable CDFIs include Connect2Capital, Community Reinvestment Fund, Accion, and Opportunity Fund.

Fintech lenders (online lending platforms)
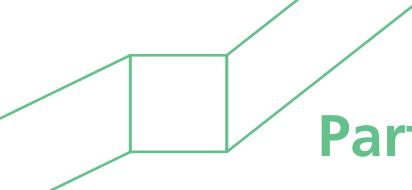
The rise of fintech lenders in the US, the UK, and China has rejuvenated the MSME lending ecosystem. These lenders operate their MSME lending business either using the online direct model or the peer-to-peer lending model (also called "marketplace lending"). These fintech lenders adopt a great deal of new technology and utilise big data. As a result, loan decisions are usually guided by automated underwriting systems that may be pulling data from credit reports but that also make use of alternative sources, such as real-time business accounting information, payment and sales history, logistics and supply-chain data, and online customer reviews. A series of advanced analytics models (i.e. machine learning) have also been designed and implemented to support credit decision-making, anti-fraud, market analytics, and risk segmentation. Some notable fintech lenders are:

- China: WeBank, Ant Financial, and Meituan;

- US: Kickfurther, Kabbage, and OnDeck;

- UK: OakNorth and Funding Circle.

3.2 Case studies in the implementation of alternative credit scoring

3.2.1 Hong Kong

The HKMA allows both conventional and virtual banks in Hong Kong to use new models and techniques enabled by big data techniques and consumer behavioural analytics to approve and manage related credit risks. For example, a number of virtual banks have started to adopt data-driven income estimation models to inform their credit underwriting and lending decisions, rather than collecting income proof from applicants. Similarly, a few other virtual banks are exploring the use of business transaction data combined with supervised machine learning to circumvent the time-consuming conventional loan approval process for MSMEs.

Case study of PAO Bank's deployment of alternative credit scoring in Hong Kong
Ping An OneConnect Bank (PAOB) is one of the eight licensed virtual banks in Hong Kong, which aim to promote financial inclusion. To offer small businesses in Hong Kong more efficient banking by reducing the time and effort needed to apply for financing, PAOB and Tradelink Electronic Commerce Limited (Tradelink) have partnered to co-create a simple and convenient banking service for these small businesses, utilising synergies between each other's strengths in data, analytics, and technology.

Tradelink has been a leading provider of Government Electronic Trading Services (GETS) since 1997. It provides an efficient channel and a robust platform for importers and exporters in Hong Kong to electronically lodge trade compliance documents such as Import and Export Declarations, Dutiable Commodities Permits, and Electronic Cargo Manifests. The platform captures business data that include a number of features of value to bank's risk assessment and account monitoring of borrowers from the trading sector, a major economic segment densely populated with small businesses. First, the data reflect the past and current business conditions of individual trading companies vis-à-vis their peers. Second, the data are current, as they are recorded when the company lodges trade declarations; changes and unusual activities can be identified automatically and in a more timely manner compared with the use of, say, data such as financial statements or even management accounts. Third, the data are reliable, as they are related to the actual shipment of goods that the company is required to declare under the relevant legislation.

Using big data techniques on these unique business data, PAOB developed a new credit underwriting approach in which customers that fit the risk appetite of the bank can be identified in advance. This is different from conventional practices in which difficulties can arise in the course of the lengthy information enquiry and credit approval process that takes place between bank and customer. Based on this approach, PAOB rolled out its first credit product, Trade-Connect Loan. This had a "5-Day Service Pledge", meaning that HK$1,000 cash compensation would be given if the credit approval and loan disbursement together took longer than five working days. Apart from lending, PAOB has launched an SME mobile banking app that supports 24/7 remote account opening, with an embedded in-house eKYC solution. Put together, these solutions effectively tackle the major pain points in customer onboarding.

Case study of the CCRA's use of an API for deployment of alternative credit scoring in Hong Kong

In 2004, Dun & Bradstreet was appointed as the CCRA service provider for the Hong Kong banking industry. Currently, Dun & Bradstreet sends CCRA reports to Authorised Institutions (AIs) in relatively flexible formats (paper, PDF, HTML). Based on the "Open API Framework for the Hong Kong Banking Sector", one of seven initiatives by the HKMA announced in September 2017 to prepare Hong Kong for a new era in Smart Banking, Dun & Bradstreet has proposed a "CCRA API" to the major banks that would allow them to obtain CCRA information programmatically, thus enabling digital transformation of their loan approval process. With this capability, Hong Kong banks will be able to develop faster, cheaper and more automated processes, using algorithms that incorporate advances in machine learning/artificial intelligence, to improve or augment their loan approval decision-making, with straight-through processing of the master data from front to back. This is a tool that could drive competition locally and globally for Hong Kong as a global financial centre and improve the allocation of capital to the MSME sector.

Dun & Bradstreet also suggested that it would be feasible to further expand the scope of the CCRA API to incorporate other relevant alternative information — specifically litigation data, payment data, registration data, corporate linkages, and predictive analytics — in order to improve the quality, accuracy, and precision of the credit scoring process available to Hong Kong banks.[6]

---

6.   The proposals for the CCRA API have been submitted and are currently under review by the HKAB.

3.2.2   China

Case study of WeLab's deployment of alternative credit scoring in mainland China
WeLab is a leading fintech company in Asia, with operations in Hong Kong, mainland China, and Indonesia. It offers purely online credit solutions to over 46 million users across groups such as salaried individuals, sole proprietors of MSMEs, and car owners.

When undertaking credit scoring of MSMEs in China, two types of alternative data are typically obtained:

1. Merchant business data

   Fintech lenders usually work with two major types of loan sourcing channels (or third-party data providers) to assess the creditworthiness of MSMEs, namely E-marketplaces and Point-of-Sale service providers. Depending on their partnership agreement, these channels can provide data about merchant profiles, merchant popularity, refund histories, and sales data on the transaction level or on a summarised level.

   To protect the interests of the channel and the privacy of its merchant customers, in most cases, the fintech lender will engage the channel to develop a pre-screening model that performs credit scoring based on the data of the channel's MSME clients. The credit scoring is carried out by the channel on-site so that no confidential data is leaked. MSMEs with favourable credit scoring outcomes are labelled as whitelisted merchants by the channel. This pre-screening model not only reduces credit risk but also increases the approval rate, thus enhancing the customer experience by offering loan promotions/options only to whitelisted merchants from the channel.

2. Personal credit history of the business owner

   Some data vendors can provide the credit history of business owners, gathered from non-bank FIs. Information provided may include delinquencies, number of loan enquiries and/or number of existing loans, fraud blacklists, and details of individuals who have defaulted. Some large vendors can also provide an alternative credit score for business owners, developed on the basis of demographic data and spending behaviour.
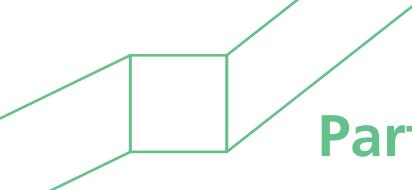
When an MSME applies for a loan, API calls are triggered to obtain the required response parameters from different data providers for credit scoring. The data are then transformed and fed into a credit decision engine for further modelling and decision-making.

This alternative credit scoring can deliver the following advantages:

- The alternative data can provide a source of additional inferences regarding the merchant's creditworthiness, especially for MSMEs.

- The processes for MSMEs to take out small-sized loans are simpler compared with the conventional credit scoring approach.

- FIs do not need to verify the business registration details of the business, as the channel has already done this.

However, there remain challenging areas for which fintech companies are still actively developing solutions:

- Currently available technology makes it difficult to accurately assess the fraud risk associated with customers colluding with e-merchants to artificially inflate their business volume. A proxy parameter involves checking the refund history of the e-merchants.

- Data relating to a merchant whose business runs on multiple E-marketplaces need to be aggregated before a full picture of the merchant's sales can be gained. This means longer processing times, which may impact the applicant's experience.

- The cost of the business and the net income of the merchant are not always available, which affects the accuracy of the merchant's debt-to-income ratio assessment.

- From a monitoring perspective, because the pre-screening model is usually developed at multiple sites, the role of and responsibility for ongoing monitoring of the consistency of the outcomes of the pre-screening model needs to be clearly defined.

3.2.3 **Europe**

Case study of CRIF's deployment of alternative credit scoring in Europe
CRIF is a global company specialising in credit information solutions, business information, outsourcing services, and credit management solutions. Founded in 1988, CRIF has achieved a solid international presence and now serves lenders and business clients on five continents. Following the implementation of the Second Payment Service Directive (PSD2), CRIF obtained registration as an Account Information Service Provider in 31 European countries, enabling it to provide services in support of a growing number of credit institutions and non-financial companies.

An example of an opportunity captured from the implementation of PSD2 is highlighted in the following "5 Ws of Open Banking" case study:

WHO: A medium-sized, multi-regional banking group made up of several companies operating in all financial sectors, and having a strong focus on investing in innovation and sustainability.

WHERE: Italy.

WHAT: The scope of the activity was the computation of a score and a set of KPIs, based on SME bank account holder transactional data, focusing on statistical performances and business benefits with respect to the likelihood of future repayment and seizing new business development opportunities through cross-selling, up-selling, and anti-churn actions.

WHEN: A trial that involved conducting an alternative credit scoring with the banking group was performed over a period of 10–12 weeks prior to the formal implementation of the Open Banking Directive.

WHY: The main benefits derived from the activity can be summarised as follows:

- Fine-tuning of the categorisation of bank account transactions — creating a granular taxonomy of over 240 categories offering additional elements supporting the generation of customer insights.

- Use of bank account transaction data to supplement existing information assets and refine the evaluation of SME creditworthiness: the result of the trial showed that the performance of the credit scoring improved significantly when using the categorisation of bank account transactions.

- The categorisation process, based on machine learning and AI detection methodologies, also brought tangible benefits relating to different stages of the lender journey. For instance, it identified different types of commercial opportunities that resulted in approximately 7,100 commercial actions. Some examples are:

- Cross-selling opportunities: 22.6% of the bank's customers became the target of specific financial products based on an analysis of companies' seasonal behaviour or international trading records.

- Consultancy opportunities for business development: about 19% of the target customers of the bank had amongst their clients companies that were not part of the bank's client base and had good credit references.

Overall, the use of alternative data such as transactional data combined with machine learning methodologies helped to speed up the creditworthiness assessment process, both in terms of evaluating the financial needs of companies and assessing the sustainability of new credit lines. This in turn enhanced the organisation's business development, especially by providing insights and client profiling derived from improved transaction classification.

Case Study of the deployment of alternative credit scoring by CCDS (Commercial Credit Data Sharing) in the UK
Small and medium enterprises (SMEs) comprise 99.9% of all private-sector businesses in the UK and play an essential role in the country's economy. The Commercial Credit Data Sharing ("CCDS") scheme was launched in 2016 as part of the UK Government's commitment to supporting SME growth. The scheme requires nine leading banks to share credit information on SMEs with four designated Credit Reference Agencies appointed by HM Treasury, of which one is Dun & Bradstreet.

These four designated Credit Reference Agencies receive credit data from the nine banks and manage, cleanse, and match the data against their existing records. Lenders can use this completely new package of information to build a picture of a UK business, to review a business's financial performance, and to make robust lending decisions more quickly.

Under the CCDS Scheme, competition can be introduced to encourage new entrants in SME lending to address the financing gap within the sector. The scheme covers not just the smallest UK businesses but any company with a turnover of up to £25 million, which includes 99.9% of the entire UK business population.

3.2.4   United States

Case Study of the deployment of alternative credit scoring by SBFE (the Small Business Financial Exchange) in the US
In the United States, the SBFE (Small Business Financial Exchange) provides local lenders (banks, credit unions, and credit card issuers) with credit reports to help them make better credit decisions. Formed in 2001, the SBFE is a highly trusted business data exchange governed by the small business lending industry and managed independently from the credit reporting agencies.

SBFE's single-feed, multi–Certified Vendor model is designed to support the safe and secure growth of small businesses by delivering a highly accurate and comprehensive picture of small businesses built from SBFE Data™ across a broad ecosystem. It does this by providing SBFE Data™ to four SBFE Certified Vendors™, data which enables lenders to develop a varied set of risk management solutions and contingency options. One of these certified vendors is Dun & Bradstreet.

On top of the existing data attributes that SBFE provides, it also provides payment data that can give lenders a more complete overview of the financial situations of the companies under consideration. Additional scores and ratings based on different data attributes can be used to further quantify and simplify the credit scoring. Table 1.1 shows examples of the benefits of putting alternative data and the existing SBFE key attributes together for credit scoring:

Table 1.1 Benefits of using alternative data and the existing SBFE key attributes

SBFE Key Attribute	Available Payment Data (Alternative Data)	Benefits of using the two fields together
# of Open Accounts	Total Payment Experiences in D&B's file	Provides a more complete view of the total number of obligations the loan applicant must assess in terms of lending and credit activity
Total Balance on Open Accounts	Now Owes	Gives a more complete current view of the company's debts arising from lending and trade obligations, as reported to SBFE and D&B
Max Account Balance	Highest Now Owing	Enables the lender to compare, contrast, and summarise the highest tradeline balances to assess aggressive credit requests
Max Credit Limit/ Original Amount	Largest High Credit	Provides an understanding of the highest credit line/loan amount and trade invoices owed at one time in the financial tradeline — this reveals what have other creditors have provided
D&B SBFE Score	Commercial Credit Score	Indicates whether the company will pay its lending obligations on time vs other trade obligations. Again, helpful for ascertaining the likelihood of being paid on time

Acknowledgements for contributions to this section:

Company	Contributions
Nova Credit	Information about alternative MSME loan Lenders
Dun & Bradstreet	Information on the case studies described
PAO Bank	Information on the case study described
WeLab Bank	Information on the case study described
CRIF	Information on the case study described

Part Two:

Fintech infrastructural components for alternative credit scoring

Banks wishing to use alternative credit scoring to make decisions on loan applications need to first perform a series of processes. First, alternative data needs to be collected from relevant third-party data providers. The alternative data then needs to be pre-processed so that relevant data fields are extracted that can be used to run the machine learning models for credit scoring. Based on the results of the machine learning models, the prediction results of alternative credit scoring are augmented by the conventional credit scoring results. Finally, the decisions about loan application approval can be taken. A fintech infrastructure therefore needs to be developed to support these processes.

This paper envisions a fintech infrastructure for alternative credit scoring containing three essential elements: a supply of alternative data, a machine learning model for predicting default, and an online lending platform that supports the automation of credit underwriting. The alternative data is what enables the machine learning model to deliver credit insights into MSMEs. Having identified both transactional and non-transactional types of alternative data, this paper focuses on the use of transactional data by machine learning models to assess the creditworthiness of MSMEs. The
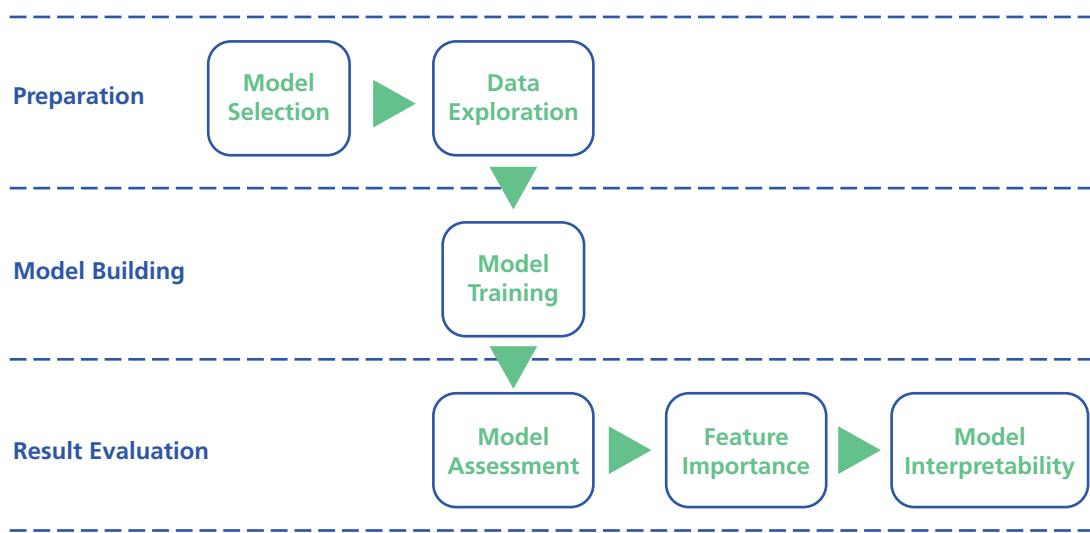
development of an accurate machine learning model is thus a key infrastructural component requiring skilful engineering design work in terms of preparation, model building, and the evaluation of results. An online lending platform is required to provide an execution environment that is able to process incoming alternative data, execute the machine learning model that has been developed, and continuously reassess the MSME's creditworthiness based on any up-to-date alternative data received.

This new fintech infrastructure will enable banks to assess and monitor the creditworthiness of MSMEs continuously. However, it brings both opportunities and challenges. The ability to conduct continuous monitoring gives banks a new instrument for mitigating risk relating to problematic loans. On the other hand, obtaining a stable and reliable supply of alternative data from third-party data providers is a challenge. Concerns around data privacy are also a potential hindrance to alternative credit scoring because of the need to share alternative data with non-banking entities. These data privacy concerns need to be addressed by utilising new advances in privacy-enhancing technologies.

1. Development of machine learning models for default prediction

Machine learning models enable synthetic, effective, and dynamic indexes to be calculated that can facilitate rapid, informed decision-making in a changing and increasingly competitive environment. Machine learning models are the best candidates for the efficient extraction of value from data. They are constantly evolving as part of the search to cut analysis times, reduce discretionary power in the process, improve model development and validation, and produce high-performance risk indicators. Alternative credit scoring using machine learning requires the right procedures and processes to be in place. This section introduces the common pipelines and tools required to empower AI/machine learning for alternative credit scoring.

Figure 2.1 The machine learning model development workflow

To support decision-making about loan applications and satisfy risk management requirements, there are three phases involved in developing a machine learning model for predicting default, as shown in Figure 2.1. First, the preparation phase selects suitable machine learning algorithms for model development and explores the alternative data available for credit scoring. Second, based on the selected algorithms and the pre-processed data, the model needs to be trained. Finally, once the model has been executed against the pre-processed data, a sequence of steps to evaluate the result must be carried out, consisting of model assessment, output consideration, and model interpretability.

1.1 Model Selection

There is no single answer to the question of which machine learning algorithm is best to use for alternative credit scoring, because default prediction is highly dependent on the type of alternative data available. Model selection therefore needs to be an exploratory process involving the continuous evaluation of multiple machine learning models.

This paper focuses on the nine machine learning algorithms that were used for the experiments described in Part Three of this paper. These algorithms are Logistic Regression, Random Forest, Extra-Trees, CatBoost, LightGBM, XGBoost, K-Nearest Neighbours, Convolutional Neural Networks, and Stacking. Please refer to Appendix B for detailed descriptions of these machine learning algorithms.

Technologies such as cloud computing and cloud storage mean that it is now possible to incorporate more variables and a wider range of data in (alternative) credit scoring models. Broadening the data universe can be useful, but it also adds to the model complexity. Once the model needs to compare variables comprising numbers or characters (alphanumeric), which may have discrete or continuous distributions, it becomes important to decide which model generates the most accurate predictions of the probability of default. Of the nine selected machine learning algorithms described in this paper, Logistic Regression is the model traditionally used for credit scoring. It performs very well when the data fields in the dataset are linearly related to one another. The other algorithms can be classified as either ensemble machine learning algorithms or neural network algorithms. There are generally three categories of ensemble machine learning algorithms, namely bagging, boosting, and stacking. Extra-trees and Random Forest are kinds of decision tree learning. Random Forest belongs to Bagging, while CatBoost, LightGBM, and XGBoost belong to Boosting. Stacking is an ensemble learning technique that uses multiple machine learning algorithms to obtain better prediction accuracy than could be obtained from any of the constituent machine learning algorithms alone. K-Nearest Neighbours (KNN) is a non-parametric method and its output depends on the average of K nearest neighbours. As for Convolutional Neural Networks (CNN), it is a kind of neural network algorithm.

Empirical observation indicates that Random Forest and XGBoost are more popular machine learning algorithms in recent years. XGBoost has won many machine learning competitions[7]. In practice, the best machine learning algorithm will depend on the problem that needs to be solved. All machine learning algorithms have their respective pros and cons as alternative credit scoring models. A more detailed description of the nine selected algorithms can be found in Appendix B.

1.2 Data Exploration

In alternative credit scoring, the data available will dictate what type of model to employ and what information can be gained from the model output. Not all data are made equal. Nowadays, in-house data (e.g. data collected internally by a bank) may comprise data ranging from public financial information (such as an MSME's business structure and the number of years it has been registered) to temporal data (such as its cashflow, claims, and overdrafts). The process of credit scoring modelling requires rigorous and detailed exploratory data analysis (EDA)[8] to distinguish the variables and optimise the model.

When collecting data for a new project, sometimes the data are database snapshots, with the snapshot comprising all fields present in the database. Although as mentioned earlier some models (such as Neural Networks) require a large amount of input data, the data must be relevant to the model and there should be no redundancy within the input data. In a case in which the data consist of ten columns and a data analyst notes that two of these are highly correlated, one of these two columns may be discarded as it will not add value to the model.
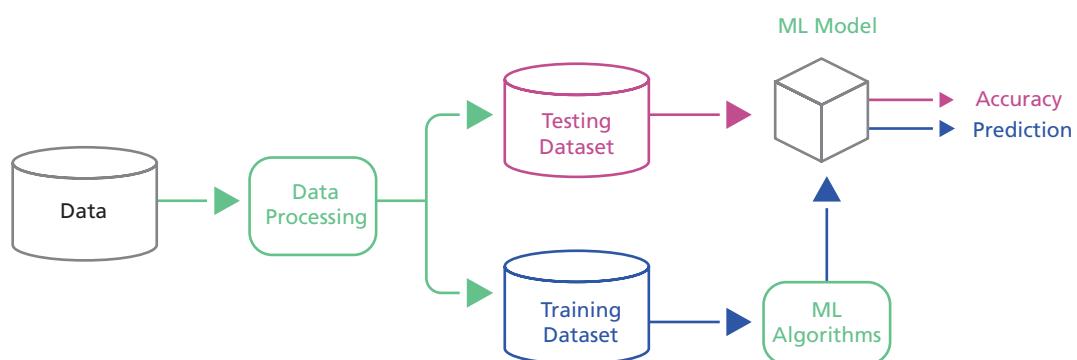
Artificial intelligence and machine learning algorithms are trained and tested in a similar way. The input dataset is usually split into a training and a testing set. The usual rule for splitting the data is 70% for training and 30% for testing (or 80% and 20% respectively), based on the Pareto Principle[9].

---

7. XGBoost. (n.d.). XGBoost. Retrieved August 20, 2020, from https://en.wikipedia.org/wiki/XGBoost.

8. Pre-built EDA tools are now available in data analysis libraries such as Pandas for Python. Additionally, libraries such as Pandas Profiling for Python can add a visual aspect to the EDA.

9. Box, G. E., & Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics, 28*(1), 11–18.

1.3 Model Training

Machine learning algorithms establish statistical/mathematical models that can make inferences. The inputs of machine learning algorithms are the training data, also known as predictors or independent variables, and the outputs of machine learning algorithms are the responses, also known as predictions or dependent variables. The inputs and outputs of machine learning algorithms can be defined as either quantitative (numerical) or qualitative (categorical). A numerical output corresponds to regression problems, such as the future price of a stock. A categorical output corresponds to classification problems, such as whether a client will fail to repay a loan.

In practice, the raw data need to be processed into meaningful data of good quality. Next, the qualifying data are divided into a training dataset and a testing dataset (see Figure 2.2). The training dataset is the input of the model built by machine learning algorithms, and the testing dataset is used to evaluate the performance of the model, such as the accuracy of its predictions in response to a certain question.


Figure 2.2 Overview of Machine Learning Model
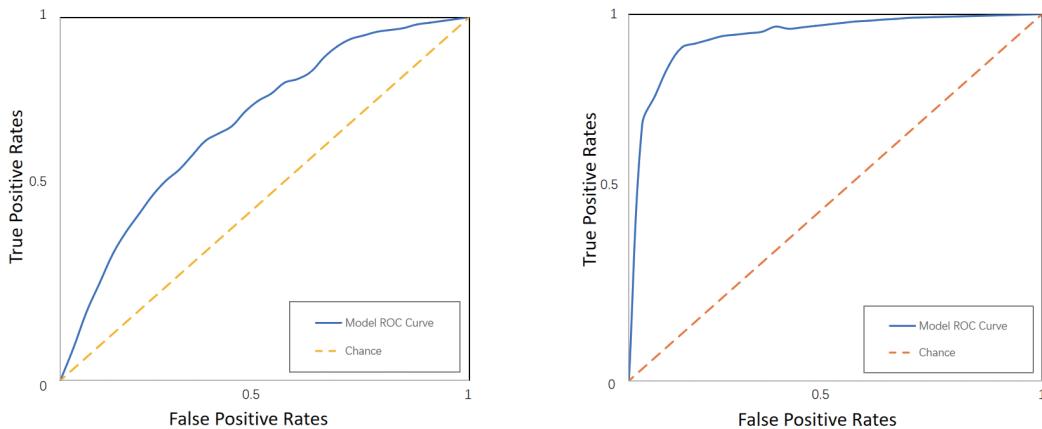
1.4 Model Assessment

In the banking industry, alternative credit scoring models provide answers to questions such as "will this individual or entity default in paying?" and "how much should be loaned to this individual or entity?". The quality of the answers resides in the model output and its interpretation. On the one hand, the model needs to be as accurate as possible to avoid the bank incurring losses; on the other hand, the model needs to align with the bank's working capacity and liquidity. For example, a financial lender employing a model that approves all of its customers for a loan may end up exceeding its limits. To avoid these kinds of situations, it is important not only to consider the accuracy of the model but also to align it with the lender's actual business operations.

To assess the quality of the model, one commonly used metric is Area Under the Receiver Operating Characteristic (ROC) Curve (see Figure 2.3). The ROC curve represents all the possible values of default probability generated by the model. It is plotted with a True Positive Rate (TPR) against a False Positive Rate (FPR), with TPR on the y-axis and FPR on the x-axis. The higher the Area Under the Curve, the better the model is at correctly predicting default. A machine learning model that rejects too many loan applicants may, for example, not allow the bank to deliver enough of their products. On the other hand, if the number of True Positives is large, the bank may not have enough staff to handle the cases individually.

In conclusion, an alternative credit scoring model needs to perform well both quantitatively and qualitatively. The right threshold needs to be determined by taking the perspectives of both data scientists and business managers into account.

Figure 2.3 Receiver Operating Characteristic curves. The ROC curve on the left is closer to the chance line (representing random output from the model) than the ROC curve on the right, indicating that the one on the right is the better performer.

1.5 Feature Importance

Once the input data have been explored and the model has been trained and tested, an additional step is to analyse the importance of the variables within the model before outputting any results. Analysing feature importance involves inspecting the variables and deciding whether a change in a variable (e.g. a change of distribution) would change the model output. Feature importance can be achieved by evaluating the variables using software tools such as the Partial Dependence Plots[10], ELI5[11], or SHAP[12] Python libraries for specific algorithms, such as Random Forests and Boosted Trees.

The reason for investigating feature importance is to further improve the model. The more important a feature is, the greater the amount of information it contains. Conversely, if a feature has low importance, it could be irrelevant for model training and there will be no loss of model accuracy even if it is discarded.

---

10. Partial dependence plots. (n.d.). Partial Dependence Plots. Retrieved July 13, 2020, from https://scikit-learn.org/stable/modules/partial_dependence.html.
11. ELI5. (n.d.). ELI5. Retrieved July 13, 2020, from https://eli5.readthedocs.io/en/latest/.
12. Christoph Molnar. (2020, July 6). SHAP (SHapley Additive exPlanations). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/shap.html.

By comparing the importance of all the features, a subset of features can be selected to replace the original training dataset. There are three advantages to applying feature selection. First, the training time is shortened. Second, reducing the number of features can simplify the learning model and improve model interpretability. Lastly, this process can effectively prevent the occurrence of overfitting and enhance the versatility of the model.

1.6 Model Interpretability

According to Miller (2017)[13], "Interpretability is the degree to which a human can understand the cause of a decision. The greater the interpretability of a machine learning model, the easier it is for humans to understand why a certain decision or prediction has been made." Model interpretability has proved a barrier to the adoption of machine learning for the financial industry. If a model is not highly interpretable, a bank may not be permitted to apply its insights to its business.

To help humans to interpret the outcomes of machine learning models, a number of model interpretation technologies have been developed. These technologies include SHAP[14], ELI5[15], LIME[16], Microsoft InterpretML[17], XAI — explainableAI[18], Alibi[19], TreeInterpreter[20], Skater[21], FairML[22] and fairness[23]. Among these, SHAP and LIME are both popular Python libraries for model interpretation with their own strengths and weaknesses. LIME is model-agnostic, meaning that it can be applied to any machine learning model and is very fast even for large datasets, but it lacks stability and consistency. SHAP has properties such as consistency and local accuracy, but it is very time-expensive, as it checks all of the possible combinations of variables. LIME is usually used for performance reasons, but if computation time is not an issue, SHAP is a preferable choice.

13. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.
14. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765–4774).
15. ELI5. (n.d.). ELI5's documentation, https://eli5.readthedocs.io/en/latest/
16. LIME. (n.d.). LIME. Retrieved August 5, 2020, from https://github.com/marcotcr/lime.
17. InterpretML. (n.d.). InterpretML. Retrieved August 5, 2020, from https://github.com/interpretml/interpret.
18. XAI — An eXplainability toolbox for machine learning. (n.d.). XAI — An EXplainability Toolbox for Machine Learning. Retrieved August 5, 2020, from https://github.com/EthicalML/xai.
19. Alibi. (n.d.). Alibi. Retrieved August 5, 2020, from https://github.com/SeldonIO/alibi.
20. TreeInterpreter. (n.d.). TreeInterpreter. Retrieved August 5, 2020, from https://github.com/andosa/treeinter preter.
21. Skater. (n.d.). Skater. Retrieved August 5, 2020, from https://github.com/oracle/Skater.
22. FairML: Auditing Black-Box Predictive Models. (n.d.). FairML: Auditing Black-Box Predictive Models. Retrieved August 5, 2020, from https://github.com/adebayoj/fairml.
23. Fairness. (n.d.). Fairness. https://github.com/algofairness/fairness-comparison.

Acknowledgements for contributions to this section:

Company	Contribution
Euler Hermes	Information on the issues of data quality, model assessment, and model output consideration
Nova Credit	Information on some commonly used machine learning algorithms

2. Automation of credit underwriting for alternative credit scoring

The previous section discussed various machine learning models for evaluating data and making predictions about creditworthiness. What though is the best way to collect and manage the alternative data that is to be input into the machine learning model, and also to manage the output of the model (e.g. how to present and visualise the credit scoring results generated by the model)? This section proposes a single online lending platform that integrates all three of these stages. The online lending platform would structure data from different channels, categorise data variables, run the machine learning models on the structured and categorised data, engage in continuous monitoring of the models' prediction results, and present the results in an explainable manner.

Banks can use the platform to automate the generation of credit scores for MSMEs that are applying for loans and monitor the credit status of their borrowers. This new capability facilitates faster responses by banks, enabling them to better control risk and limit exposure to problematic loans. This section describes the key components of an online lending platform for the automation of credit underwriting, and explains how the platform supports the alternative credit scoring workflow.

Conventional and alternative credit scoring models coexist, so the decision-making process of bank loan managers requires them to consider credit score results from different machine learning models as well as score results generated by conventional credit scoring models. This section also outlines strategies for combining conventional and alternative credit scoring, such as the champion-challenger approach, which can be applied to support loan managers when making the final loan application decision.

2.1 Online lending platform for the automation of credit underwriting

The conventional approach of financial lenders to processing loan applications is costly, because it normally relies on traditional personal relationship management with customers. The credit scoring process also requires substantial manual efforts to extract relevant data from financial statements, formulate financial ratios, and generate credit scores. Consequently, banks tend to focus on sizeable loans that offer them a reasonable return in revenue to cover the costs involved in the processing of the loan applications. The conventional approach is also demanding for MSMEs, which need to devote a great deal of time and effort to apply for bank loans under a system that is not set up to suit their requirements. There is a genuine need to simplify the MSME lending process, and fintech provides the ability to do this.

## 2.2 Workflow for Alternative Credit Scoring
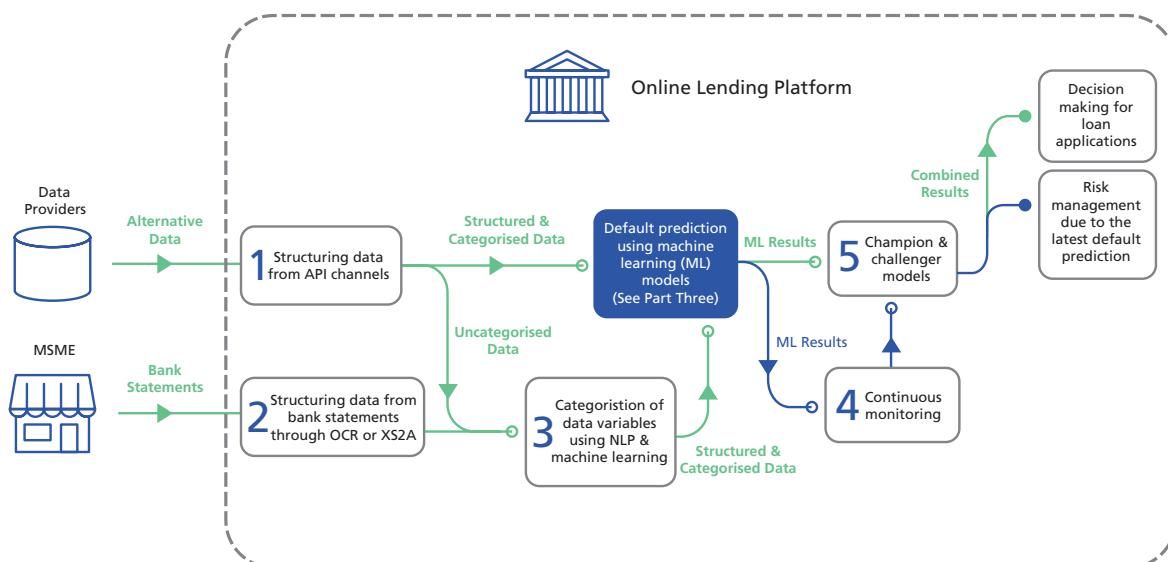
Figure 2.4 Workflow for Alternative Credit Scoring

To facilitate automation of the workflow for alternative credit analysis, an online lending platform is needed to manage the steps involved in the process (see Figure 2.4). These steps include the structuring and categorisation of data fields, analysis by machine learning, decision-making, and continuous monitoring. An online lending platform can achieve shorter turn-around times for loan approvals, which in stormy economic times can be critical in helping MSMEs to survive. It can also help lenders' operations become more cost-effective in the processing of loan applications.

2.2.1 Step ①: Structuring data from API channels

To streamline and automate the credit underwriting workflow, an online lending platform will collect alternative data related to the credit history of MSMEs directly from third-party data providers (with the consent of MSMEs). Straight-through transfer of this alternative data can be achieved by new open banking API interfaces that are being made possible by Open API initiatives. In compliance with the requirements of Open API initiatives, an online lending platform needs to maintain the status of the consent of MSMEs. For example, the platform should revoke the consent of an MSME if the validity period of the consent has expired or the MSME decides it wishes to revoke its consent.

A notable example of an Open API initiative is the revised Payment Services Directive (PSD2) that came into force in January 2018 in the European Union (EU). All regulated payment service providers in the EU need to comply with PSD2 and the Regulatory Technical Standards set out by the European Banking Authority. The requirements of Access to Account (XS2A) under PSD2 give financial institutions and regulated third parties access to the bank accounts of consumers.

Structuring and categorisation of input data from APIs and other direct access methods are critical pre-processing steps required before the structured data moves to the next step. Data fields coming from the API channel are pre-defined and well-structured.

2.2.2 Step ②: Structuring data from bank statements through OCR & XS2A

MSMEs can also upload their own bank statements and financial statement documents to the online lending platform. OCR technology can then be applied to locate, pull, and capture the data fields from the uploaded documents. Alternatively, MSMEs can authorise access to their bank statement information via access channels that comply with the requirements of XS2A.

2.2.3 Step ③: Categorisation of data variables using NLP & machine learning

The data fields captured by OCR and other direct access pipelines are unstructured and not pre-defined, and so need to be structured. The next step is to perform the categorisation of data fields by transaction text analysis. Natural Language Processing (NLP) technology is required to determine the meaning of data fields and categorise them into the data variables that are required by the machine learning model.
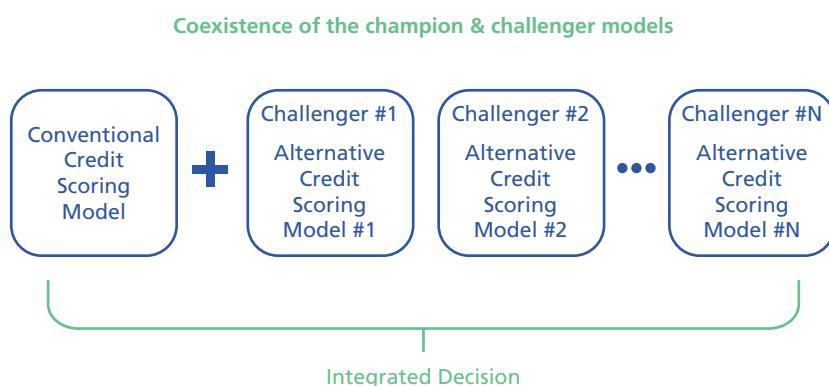
2.2.4 Step ④: Continuous monitoring

MSME business is often more volatile than that of established corporations, so the risk profile that formed the basis of a lenders' loan decision at the time of the loan application may change over time. The online lending platform can perform a reassessment of an MSME's creditworthiness based on any up-to-date alternative data received. Continuous monitoring of changes in an MSME's creditworthiness can help lenders to control and minimise their risk exposure. Compared with the conventional approach to credit scoring, another major benefit to lenders of deploying an online lending platform is that it provides them with the ability to perform continuous monitoring of the ongoing financial risk associated with the MSMEs in their lending portfolios. With continuous monitoring, an online lending platform can evaluate smaller loan credit lines more often and detect the following situations:

- Tendency for delinquency
- Change in risk profile
- Potential loan application fraud
- Signs of risky credit conditions

2.2.5 Step ⑤: Champion & challenger models
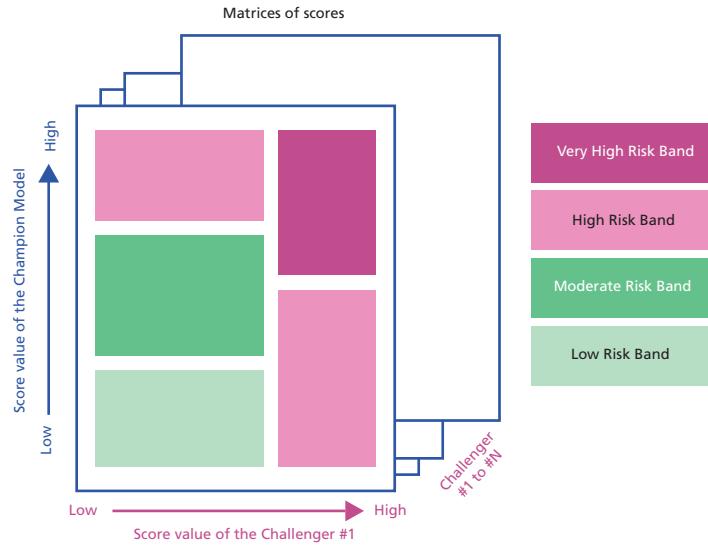
In the final step, the results of the alternative credit scoring using a machine learning model need to be combined with the results of the conventional credit scoring model. Using both conventional and alternative credit scoring models is a prudent strategy for financial lenders. This combining of results supports both the final decision-making for loan applications, and risk management for continuous monitoring. Due to the nature of machine learning algorithms, alternative credit scoring models require historical data and iterative fine-tuning to improve their accuracy. The insights generated by conventional models should, therefore, always be used as a basic reference for creditworthiness assessments. One common approach to managing the coexistence of conventional and alternative credit scoring is known as the champion–challenger approach.

Coexistence of the champion & challenger models
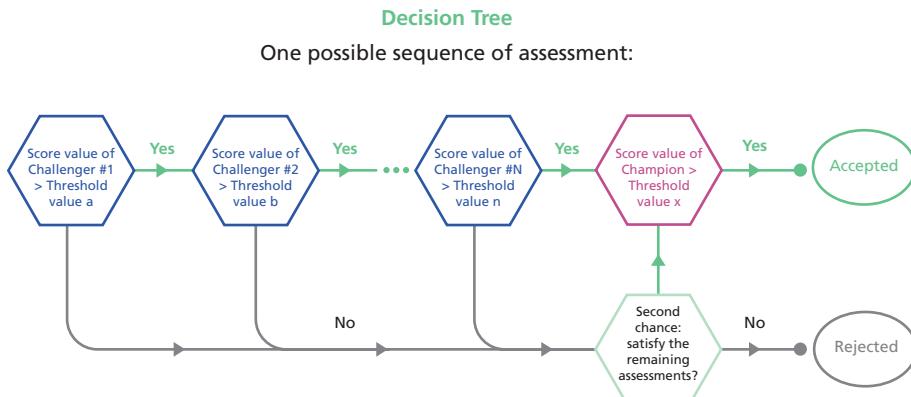
Figure 2.5 Coexistence of Conventional and Alternative Credit Scoring Models

The champion–challenger approach involves comparing the results of a conventional credit scoring model (champion) with the results of different alternative credit scoring models (challengers), as shown in Figure 2.5. For example, financial lenders can adopt this approach to compare credit score outputs by the existing champion with those by a number of challengers, which are dynamically created by adjusting different rule sets. However, reliable ways are required for comparing the effectiveness of champions and challengers, and measuring and combining the results.
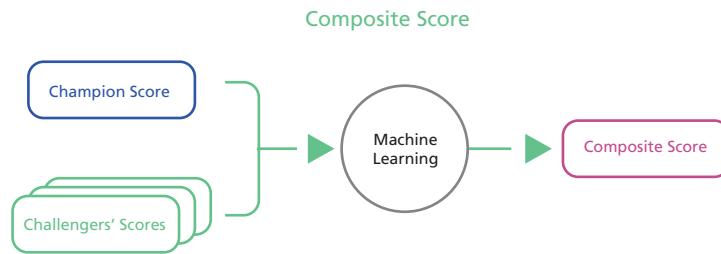
Matrices of scores

Figure 2.6 Score Matrices

Decision Tree

One possible sequence of assessment:

Figure 2.7 Decision Tree

Figure 2.8 Composite Score

To support decision making, the individual champion and challenger credit scores should be used to generate a combined scoring result. As illustrated in Figures 2.6, 2.7 and 2.8, the combined scoring result of the champion and challenger models can be visualised using three separate methods.

- **Score Matrices:**   the risk scores of the champion and individual challengers can be compared by a matrix representation in which different bands of risk level can be identified.

- **Decision tree:**   loan applications can go through different assessments using the champion and challenger models. The risk scores of the champion and individual challengers can then be assessed one-by-one or phase-by-phase in a specific sequence. Based on the sequence of assessment, loan applicants may be granted a second chance if some risk scores are unsatisfactory.

- **Composite score:**   decisions can also be made based on a composite score generated from multiple individual scores. Machine learning techniques such as logistic regression and stacking can be deployed if a composite risk score is the result of combining the scores of the champion and multiple challengers.

Human discretion is involved only when the final decision for a loan application is made. The advantage of this approach is its flexibility in considering the results generated by challengers using a wide range of alternative data (both transactional data and non-transactional data).
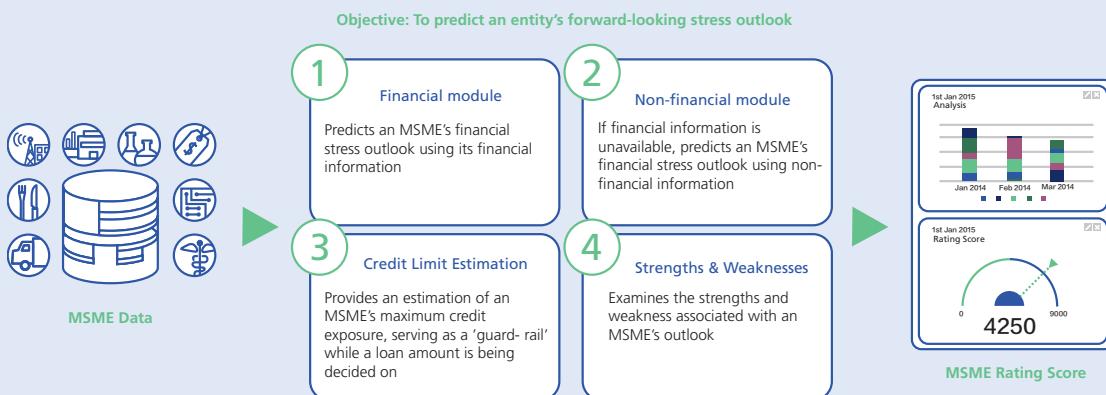
High-level machine learning-based credit scoring framework

A complete solution for credit scoring for MSME lending will integrate both conventional and alternative credit scoring, based on relevant information from various sources. This section describes two examples of High-level machine learning-based credit scoring framework.

Industry example: CRIF's machine learning based credit scoring framework
CRIF provides Business Information (BI) reports in various regions in South East Asia. A BI report provides a creditworthiness rating for an enterprise based on dimensions such as financial information, non-financial information, legal structure, industry, and management experience. All of these are dimensions that are considered by lenders in their credit scoring. The rating is based on a machine learning algorithm that predicts the forward-looking stress of the enterprise.

A CRIF BI Score for each dimension of the creditworthiness rating is a three-digit score ranging from 300 to 900. The scores of all the dimensions are further classified into 10 score tranches. As the scores increase, the financial outlook of the company is expected to be better. The score is also converted to five tranches which are used as a credit rating for the entity.

Objective: To predict an entity's forward-looking stress outlook

1 Financial module

Predicts an MSME's financial stress outlook using its financial information

2 Non-financial module

If financial information is unavailable, predicts an MSME's financial stress outlook using non-financial information

3 Credit Limit Estimation

Provides an estimation of an MSME's maximum credit exposure, serving as a 'guard- rail' while a loan amount is being decided on

4 Strengths & Weaknesses

Examines the strengths and weakness associated with an MSME's outlook

MSME Data

1st Jan 2015
Analysis

Jan 2014   Feb 2014   Mar 2014

1st Jan 2015
Rating Score

0   9000
4250

MSME Rating Score

Figure 2.9 CRIF's ML-based Credit Scoring framework

The credit scoring framework has the following major components:

- *Financial Module*
  This module uses financial information along with other non-financial information to score the entity and provide a rating. Financial information is sourced from the entity's financial statements (e.g. revenue, profits, assets, liabilities, and debts). Non-financial aspects includes data such as the entity's primary industry and the number of its employees. All of this information is fed to the machine learning model, which generates a score and a rating for each enterprise.

- *Non-financial Module*
  For a scenario in which all of an entity's financial information is not available, the non-financial module kicks in. The non-financial module uses a separate machine learning based algorithm to score such entities, based on non-financial information such as industry, property type, management information, customer information, and number of employees. All of this information is fed to the machine learning based algorithm, which generates a score and a rating for each enterprise. On top of the machine learning model, an expert scorecard is used to bridge the gap caused by the lack of detailed financial information.

- *Credit Limit Estimation*
  This refers to the estimation of the maximum loan amount that can be extended to the entity. This estimation is available to entities with a financial statement and is mostly driven by indicators of the company's financial performance over time, such as revenue and proportion of debt to equity, together with experts' views on other qualitative and industry factors. Along with the assessment of the risk indicated by the rating, lenders need to understand the amount of the enterprise's exposure to loans. To facilitate this, the Credit Limit Estimation module is used to assess the maximum loan that the enterprise can handle. Its current debt and liabilities can be subtracted from this to understand how large a loan can be provided to the enterprise.
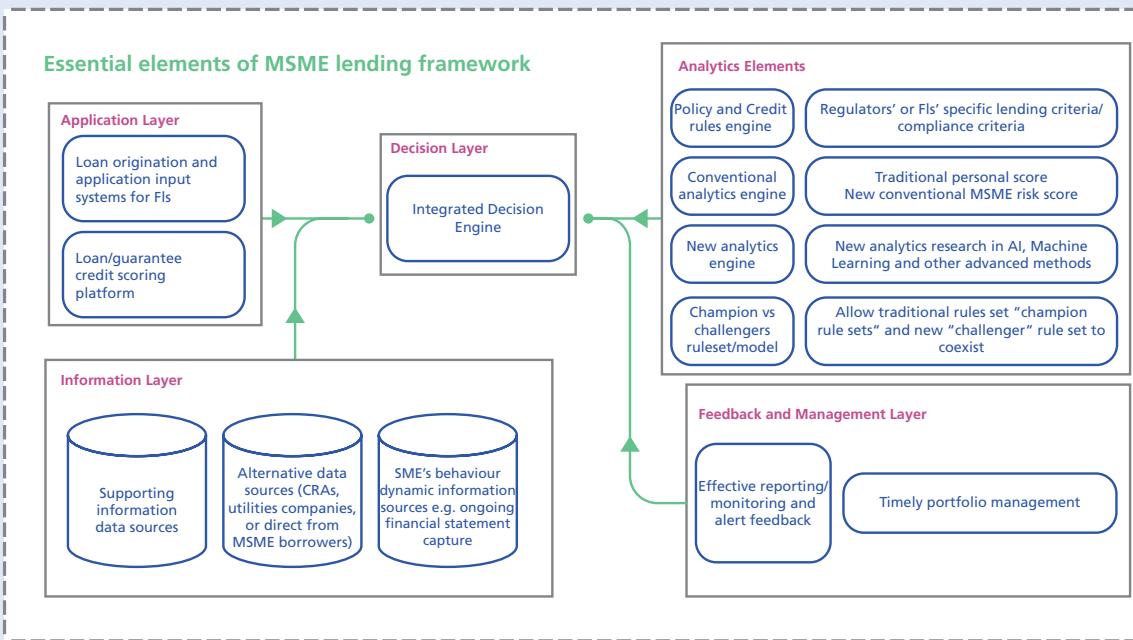
- *Strengths and Weaknesses Analysis*

  For the most significant drivers of a company's outlook, scoring factors are shared that broadly explain the areas where a business is running strong or has areas for improvement. Whereas the MSME rating and the credit limit estimation suffice as the two dimensions necessary for credit scoring, the credit scoring framework provides the reasoning for the rating. The details it provides include the top three strengths and top three weaknesses of the MSME, providing lenders with valuable insights.

Industry example: Nova Credit's MSME lending framework

Nova Credit's MSME lending framework is composed of five elements: an Application Layer, an Information Layer, a Decision Layer, Analytics Elements, and a Feedback and Management Layer.
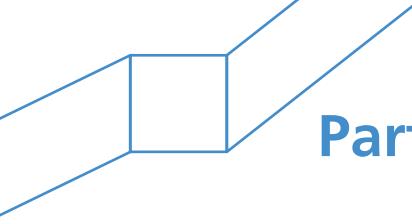

Figure 2.10 Nova Credit's SME lending framework

Alternative Credit Scoring of Micro-, Small and Medium-sized Enterprises

- *Application Layer*
  1. A loan origination system that is shared across multiple financial institutions or allows connections to financial institutions' current application systems
  2. Effective information sharing with secondary assessment entities
  3. Allows collaboration and flow management among all stakeholders and assessors
  4. Easy connection with various data sources

- *Information Layer*
  1. Supporting information data sources — A database, or a shared database that is either centralised or decentralised, and that provides traditional credit scoring information such as bank statements, financial statements, proof of assets, and valuation reports that are reviewable by all stakeholders.
  2. Alternative data sources — Other non-traditional financial information, such as a company's risk profile from CRAs, utility bills, trading information, and other business performance indicators that are reviewable by all stakeholders.
  3. SMEs' dynamic behavioural information sources — Sharable data sources that review SMEs' dynamic behaviour, such as ongoing financial statements analysis, market-related information, negative court data, and other ever-evolving information sources that are acquirable under current rules and regulations and are sharable in standard format amongst all FIs and credit assessors.

- *Decision Layer*
  A credit engine integrated with the other layers that:

  1. Is compatible with different data standards.
  2. Facilitates a champions/challengers rule set and coordinates the outputs from the other layers.
  3. Allows conventional analytics and flexible plug-ins that can perform alternative credit scoring.

- *Analytics Elements*
    1. Policy and credit rule engine — Most conventional credit engines have this function built in; its merit depends on auditability, usability, ease of maintenance, and the total cost of ownership.
    2. Conventional analytics engine — The combination of various retail credit scores, available both in-house and from CRAs, which contribute to insights about the creditworthiness of the loan applicant. Also supports product assignment, pricing, and other essential credit functions.
    3. New advanced analytics engine — Supports various new research analytics methods, such as AI and machine learning. Allows flexible plug-ins to enable both current analytics methods and new methods that may be developed in the future.
    4. Champions/Challengers rule set comparison review engine — A good way to compare the effectiveness of champions and challengers, measure the results, and enable quick ruleset manipulation and maintenance.

- *Feedback and Management Layer*
    1. Real-time dynamic reports available for every level of management, all at their fingertips.
    2. Data and access controls for different stakeholders throughout the risk cycle.

Acknowledgements for contributions to this section:

Company	Contribution
Nova Credit	Information on the Champion and Challenger approach and the high-level credit scoring framework
CRIF	Information on the high-level credit scoring framework
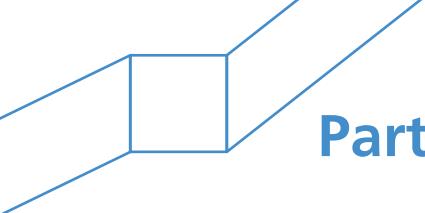
3. Privacy-enhancing technologies in sharing alternative credit data

Solving the challenge of data privacy is seen as "the last mile" in unleashing the full potential of machine learning, because data in the real world are owned and stored by different organisations. Many innovative applications will not be able to be developed if data sharing cannot be solved in a way that complies with the requirements of data privacy laws and regulations. Data privacy is also a key obstacle to developing alternative credit scoring, because alternative data include private and sensitive MSME information. Current data privacy laws and regulations, such as the EU's General Data Protection Regulation (GDPR) and Hong Kong's Personal Data Privacy Ordinance (PDPO), impose stringent requirements on disclosing data for commercial usage.

This section outlines use of various privacy-enhancing technologies (PETs)[24] to address the issue of data privacy for data sharing in financial services. These include differential privacy, zero-knowledge proofs, multi-party computing, homomorphic encryption, and federated analysis. PETs offer technical solutions that enable banks to access alternative data about MSMEs from third-party data providers without compromising data privacy regulations. They are unlocking the value of data sharing in support of the development of machine learning models and default prediction for credit scoring. PETs have been advancing rapidly in recent years, and the number of commercial implementations of PETs for financial applications has been growing. The outlook for using PETs to tackle the data privacy issues of alternative credit scoring is promising.

---

24. The next generation of data-sharing in financial services: Using privacy enhancing techniques to unlock new value. (2019, September 12). World Economic Forum. https://www.weforum.org/whitepapers/the-next-generation-of-data-sharing-in-financial-services-using-privacy-enhancing-techniques-to-unlock-new-value.

3.1 Differential privacy

One problem with traditional privacy protection approaches that involve removing, anonymising or obfuscating personal data is that outsiders can still recover personally identifiable information if they have access to other correlated datasets or side knowledge (e.g. privacy breaches due to reverse engineering). Differential privacy is a useful technique to tackle this problem. It randomly adds controlled "noise" to the individual data while preserving the statistical representativeness of the original dataset[25].

A potential use case would be a bank that wishes to employ a third-party data provider to extract recent credit profile information about an MSME, in a situation where that information also includes sensitive private information about the company. In this case, differential privacy could be used to protect the sensitive information. The data provider changes the information of the credit profile by randomly altering the content so that if another party queries the altered dataset, the query result cannot be used to infer much about the original content. The altered dataset is now differentially private. As long as the probability of altering is known, certain characteristics of the information can still be estimated. Naturally, the larger the probability, the better the privacy protection, but the altered dataset will become less meaningful; therefore, a balance between privacy and meaningfulness needs to be struck. This technique is useful in that it allows the calculation of aggregates from protected data and can generate good-enough results. The costs are relatively low because the technique can be integrated into existing data systems. It is also applicable to advanced machine learning models.

---

25.   Nissim, K., Steinke, T., Wood, A., Altman, M., Bembenek, A., Bun, M., ... & Vadhan, S. (2017, June). Differential privacy: A primer for a non-technical audience. In *Privacy Law Scholars* Conf. (Vol. 3).

3.2 Zero-knowledge proof
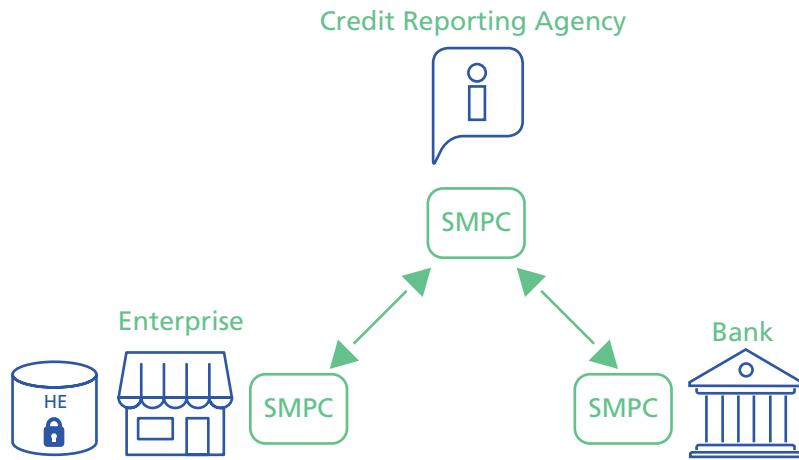
In some situations, users want to prove to another party that they own some private information to meet certain requirements, but they do not want to share or reveal that private information. Zero-knowledge proof (ZKP)[26] is a method by which this type of problem can be resolved. For example, the owner of an MSME who is applying for a loan may want to prove his or her personal financial health and demonstrate a certain level of repayment ability, but may not want to disclose any specific financial details. In this case, the lender can interact with the bank of the owner using the ZKP technique to verify whether the owner really does satisfy specific financial status requirements, without details of the owner's personal data being disclosed.

3.3 Secure multi-party computing and homomorphic encryption

Secure multi-party computing (SMPC)[27] allows multiple data owners to perform collaborative encryption calculations on a combined data set operated by a semi-trusted third party to extract the data value without revealing the original data of each data owner. Each of the participants can only see the analysed results relating to his/her own data, and not the other owners' private data. SMPC can be applied to a wide range of applications such as e-voting, e-auctions, and financial applications. Homomorphic encryption (HE)[28] is a kind of public-key encryption algorithm that supports a function for processing encrypted data directly. HE can perform certain operations under ciphertext, which is equivalent to processing data under plaintext. It has the benefit of preserving privacy and supporting remote computing services.

26. The idea of ZKP was conceived in 1989. Since then, this cryptographic concept has continued to evolve to take in new application areas, including authentication systems, end-to-end communication encryption, and privacy-preserving solutions on the blockchain. In real-world use cases, a type of cryptocurrency named Zcash mainly works with one ZKP scheme called zero-knowledge Succinct Non-interactive Argument of Knowledge (zk-SNARK). In 2017, Ethereum also adopted zk-SNARK proofs following its Byzantium update.

27. Yao, A. C. C. (1986, October). How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science* (sfcs 1986) (pp. 162–167). IEEE.

28. Rivest, R. L., Adleman, L., & Dertouzos, M. L. (1978). On data banks and privacy homomorphisms. *Foundations of Secure Computation*, *4*(11), 169–180.

Credit Reporting Agency

SMPC

Enterprise

Bank

HE

SMPC

SMPC

Figure 2.11 Combining secure multi-party computing and homomorphic encryption for credit scoring

SMPC and HE can be used together to support credit scoring without compromising privacy. For example, a bank can partner with a credit reporting agency and an enterprise to set up a network of computing nodes in support of SMPC, as shown in Figure 2.11. The credit reporting agency can perform credit scoring on the homomorphically encrypted data of the enterprise. The credit reporting party has no access to the enterprise's private information directly because the data is encrypted at all times, and only the required results are shared. SMPC allows the banks to receive the credit scores calculated by communicating with the enterprise and credit reporting agency. With the help of SMPC and HE, the bank can determine whether the enterprise satisfies the credit score requirement without having to access private information about the enterprise.

3.4 Federated Learning

When a commercial institution wants to perform machine learning analysis on large amounts of client data held across multiple datasets, the traditional first step is to combine them into a huge dataset on one server. However, this transfer may be prohibited by data privacy regulations if it violates certain data usage conditions. It can also be challenging to gain customers' consent for data sharing, especially if the data contain private information. Federated learning is one way to solve these issues. This privacy-enhancing technology integrates the use of distributed machine learning, SMPC, and homomorphic encryption to train a shared machine learning model using multiple datasets owned by different parties, without the need to combine all the datasets[29].

Federated learning involves training and updating multiple sub-models on local data samples and sharing the encrypted information of the sub-models between these local data nodes over time, to generate a federated machine learning model that is shared by all nodes. This approach is effective in avoiding legal risks regarding data privacy because no sensitive data are transmitted, as all data samples are stored physically separately. Open-sourced and commercial federated learning platforms are widely used in both academia and industry. Examples are Google's TensorFlow Federated framework[30] and WeBank's FATE library[31].

Federated learning can achieve privacy-preserving cross-border credit scoring[32]. Under the conventional credit scoring arrangement, data providers including banks and card issuers send a person's consumption activity to credit reporting agencies for scoring purposes. Banks can use the score provided by the credit agency to then assess the risk that a specific consumer will default on a loan.

29. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1–19.

30. TensorFlow Federated: Machine Learning on Decentralized Data. (n.d.). TensorFlow. Retrieved August 24, 2020, from https://www. tensorflow.org/federated.

31. FedAI. (n.d.). FedAI. Retrieved August 24, 2020, from https://www.fedai.org/.

32. https://blog.openmined.org/federated-credit-scoring/

3.5 Evaluating MSME credit ratings with privacy-enhancing technologies

Recent advances in privacy-enhancing technologies (PETs) are giving providers of alternative data the ability to share data for credit scoring. Federated analysis, homomorphic encryption, and differential privacy can facilitate data sharing for alternative credit scoring without compromising data privacy. Data sharing is essential for the development of alternative credit scoring because ML models rely on the supply of alternative data from different data sources.
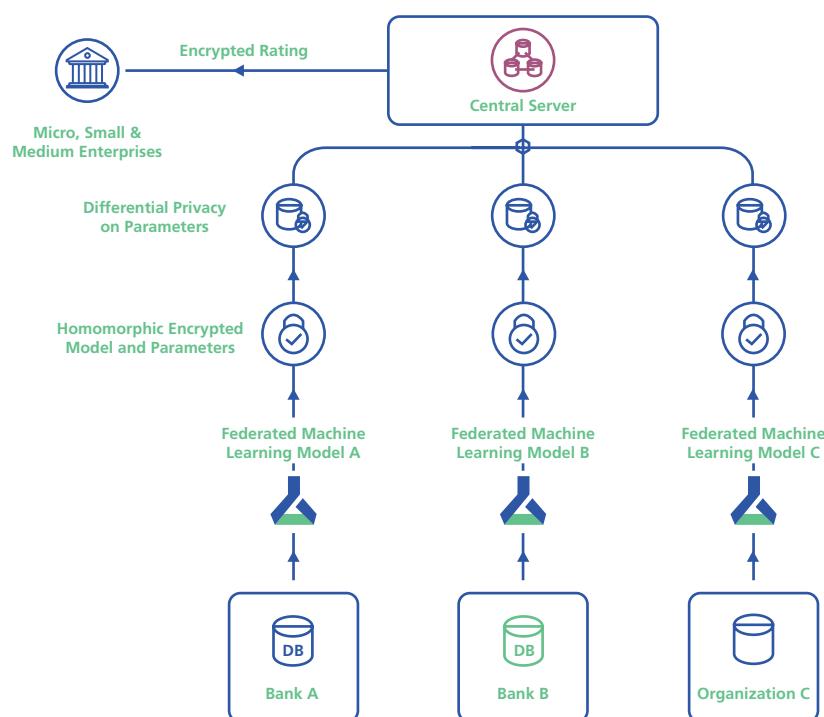

Figure 2.12 MSME Credit Rating Evaluation

Alternative Credit Scoring of Micro-, Small and Medium-sized Enterprises

During the MSME credit analysis process, federated learning enables different machine learning models to be aggregated, rather than different datasets. Training multiple federated learning models with their own datasets and then sending the trained models and parameters to the central server avoids the need to share the original data with untrusted parties. Homomorphic encryption can be used to encrypt some output information (i.e. models and parameters) so that the results can be transferred to the central server without the original data being disclosed to other parties. After the federated learning process in the central server is complete, the final rating will be returned to the MSME. In addition, differential privacy can further improve the privacy of individual MSMEs by preventing privacy leakage throughout the federated learning process.

The benefits of using PETs are obvious. They can break down data silos and mitigate the risk of violating data privacy regulations. PETs such as federated learning, homomorphic encryption, and differential privacy remain at an early stage in terms of their adoption for commercial applications. The main drawback of these technologies is their substantial operational overheads in terms of computation and communications resources. Continuous development efforts at the levels of design and implementation are needed before enterprises will adopt them widely.

Acknowledgements for contributions to this section:

Company	Contribution
CRIF	Information on privacy-enhancing technologies
Deloitte	Information on privacy-enhancing technologies

Part Three:

Technical evaluation of machine learning models

To demonstrate the technical feasibility of implementing alternative credit scoring in the banking industry, two sets of experiments were conducted. The first set of experiments was designed to evaluate the performance of different machine learning algorithms on MSME data. The second set of experiments was designed to explore the technical feasibility of the industry-specific alternative credit scoring framework that has been proposed as a basic reference for the industry.

For the technical evaluation of the performance of different machine learning algorithms on MSME data, the prediction results of nine selected machine learning algorithms were reported and compared. The machine learning algorithms were tested on a rich dataset containing bank account data of MSMEs in Japan. The experiments show that these machine learning algorithms can achieve acceptable predictive power in ascertaining the likelihood of MSME loan default.

As for the technical evaluation of the feasibility of the proposed *Industry-specific Alternative Credit Scoring Framework*, various experiments on a proof-of-concept (POC) implementation were conducted with three participating organisations in Hong Kong. These participants were a bank and two third-party data providers (a POS payment data provider and an Internet payment data provider). The framework outlines how the creditworthiness of MSMEs can be assessed based on the type of their transactional data and on the industry sector that they belong to. The experiments showed the technical feasibility of evaluating the monthly cashflow data of MSMEs' bank accounts along with transactional data from third-party data providers, as part of the proposed alternative credit scoring framework.

Part Three

1. Evaluating the performance of machine learning algorithms on MSME data

Machine learning algorithms for credit scoring and default prediction have developed rapidly in recent years. This section seeks to understand the pros and cons of different machine learning algorithms or models for alternative credit scoring. It does this by reviewing the relative performance in default prediction of nine of the latest machine learning algorithms (as described in Appendix B of this paper) on different datasets.

In collaboration with the CRD Association,[33] the nine machine learning algorithms were run on the CRD Association's MSME datasets to obtain comparative results. For this experimental work, the privacy of the MSME datasets was strictly protected. The machine learning algorithms were executed in the secure and controlled server environment of the CRD Association, and only information about the relative performance of the selected algorithms is reported in this paper.

1.1 Setup of the experiments

1.1.1 Original variables

The MSME datasets for the experiments described in this section contain data on more than 730,000 Japanese MSMEs, with nearly 3.6 million observations from the period 2010 to 2018. Around one percent of these MSMEs had loan defaults during this period. The datasets include yearly data, both financial and non-financial. As shown in Table 3.1, there are 20 independent variables in the CRD Association's original dataset. Each observation included the data fields of these independent variables of an MSME in a financial year. The experiments aimed to build a model that could predict whether a default would occur within a specific period. In this credit risk assessment analysis, the default observation period was seven years (from 2010 to 2016), and the default prediction period was two years (from 2017 and 2018). The objective was to examine and compare the accuracy of the selected machine learning algorithms' predictions of one-year and two-year default probabilities.

33. The Japan Credit Risk Database (CRD), currently managed by the CRD Association and based in Tokyo, Japan, was established as part of the Japanese government's SME financial inclusion efforts. The database collects SMEs' financial data and builds credit scoring models for SMEs, using conventional statistical models as well as machine learning — supported algorithms. To date, few databases exclusively dedicated to SMEs have achieved similar nation-wide coverage or the large volume required for robust credit risk modelling as Japan's CRD has. As of June 2020, the CRD had collected 28 million financial statements from around 3.9 million SME borrowers, benefiting the 171 institutions (government-affiliated or private financial institutions and credit guarantee corporations) that subscribe to the CRD's scoring services.

**Table 3.1 Independent variables of the
CRD Association testing dataset**

Label	Category	Note
OR01	Identification	Unique ID
OR02		Financial Year
OR03	Non-financial data	Business Type
OR04		Years in Business
OR05		Location Area
OR06		Number of Employees
OR07	Financial data	Cashflow from Operating Activities
OR08		Cashflow from Financing Activities
OR09		Cashflow from Investing Activities
OR10		Free Cashflow
OR11		End-of-year Cash balance
OR12		Total Assets
OR13		Total Liabilities
OR14		Total Equity
OR15		Total Revenue
OR16		Gross Profit
OR17		Net Profit
OR18		Short-term loans
OR19		Long-term loans
OR20	Default data	Delinquency data

1.1.2 Creation of the derived variables

Based on the original independent variables in the dataset, 24 new derived variables were introduced for analysis and modelling, as illustrated in Table 3.2. The derived variables were used to evaluate MSMEs in multiple dimensions, including those related to the entities' financial health, revenue strength, profitability, repayment ability, solvency credit condition etc.

**Table 3.2 Derived variables of the
CRD Association MSME financial dataset**

Derived Variables	Note
DV01	Equity Ratio = Total Equity/Total Assets
DV02	Equity to Debt Ratio = Total Equity/Total Liability
DV03	Asset Turnover = Total Revenue/Total Assets
DV04	Boolean1, 1 for Net Profit > 0; 0 otherwise
DV05	Boolean2, 1 for Cashflow from Operating Activities > 0; 0 otherwise
DV06	Boolean3, 1 for Cashflow from Operating Activities > Net Profit; 0 otherwise
DV07	Return on Assets1 = Gross Profit/Total Assets
DV08	Return on Assets2 = Net Profit/Total Assets
DV09	Cash Flow to Sales Ratio = Cashflow from Operating Activities/ Total Revenue
DV10	Gross Margin Ratio = Gross Profit/Total Revenue
DV11	Cash Ratio = Cash/Total Assets
DV12	Total Accruals to Total Assets = (Gross Profit – Cashflow from Operating Activities)/Total Assets
DV13	Gross Profit per Capita = Gross Profit/Number of Employees
DV14	Net Profit per Capita = Net Profit/Number of Employees
DV15	Assets to Short-term Debt Ratio + Total Assets/Short-term Liabilities
DV16	Cash Reserves Ratio = Cash/Short-term Liabilities
DV17	Fixed Assets to Fixed Liabilities and Total Equity = (Total Assets – Cash)/ (Total Equity + Total Liabilities – Short-term Debts)
DV18	Assets to Equity Ratio = Total Assets/Total Equity
DV19	Long-term Debt ratio = Long-term Debts/Total Assets
DV20	Short-term Debt ratio = Short-term Debts/Total Assets
DV21	Debt Dependency Ratio = (Short-term Debts + Long-term Debts)/ Total Assets
DV22	Debt Capacity Ratio = (Short-term Debts + Long-term Debts)/ (Cash + Total Assets)
DV23	Leverage = (Short-term Debts + Long-term Debts)/Total Revenue
DV24	Cash to Debt Ratio = Cash/(Short-term Debts + Long-term Debts)

1.1.3 Pre-processing of independent variables

The CRD Association's dataset contains no duplicated data or missing values. On average, each company had five years of historical records in the dataset. To predict the one-year and two-year probability of default, the time windows of the previous two years' historical records were used for each round of default prediction. Some variables were kept in all the windows of observation, based on the assumption that an enterprise is unlikely to change its business type or business model in a short time period. In addition to the financial data, some statistical information on cash-related variables and past default data were generated for analysis. The final analyses used 145 independent variables as variable candidate baselines for constructing a model, as shown in Table 3.3. For simplicity, the feature selection, variable operations and hyperparameter tuning of each machine algorithm are not discussed in detail.

Table 3.3 Overview of independent variable processing

	Category Name	Mark	#No of Variables
①	Original variables	Original variables in the CRD Association's dataset	20
②	Derived variables	New derived financial variables	24
③	Time window	Data of the time windows of two years' history of the independent variables, except company ID, financial year, business type, location and company's number of years in business	78 [= (20 + 24 − 5) * 2]
④	Statistical data	min, max, mean, standard and gap between the maximum and minimum value of cash-related variables	20 (= 5 * 4)
⑤	Default history	min, max and mean of default history data	3
		Total number of variables:	**145**

1.1.4 **Definition of the data groups**

In addition to the financial ratios based on the data in financial statements (used in conventional credit scoring), cashflow data were used for credit scoring in this experiment. Four data groups were created to evaluate the impacts of cashflow-related independent variables in different machine learning algorithms. Brief definitions of each data group can be found in Table 3.4.

- Group 1 is the CRD Association's original MSME dataset, with both financial and non-financial variables in Category ①.

- Group 2 focuses on the cashflow variables, which contain some cashflow-related independent variables selected from the original and derived variables.

- Group 3 covers all of the variables in Group 2 and Category ② to assess whether the cashflow and financial variables have an effect on credit scoring.

- Group 4 is the full dataset, which holds the most information on all five categories of variables: original variables, derived variables, time window, statistical data and default history data variables.
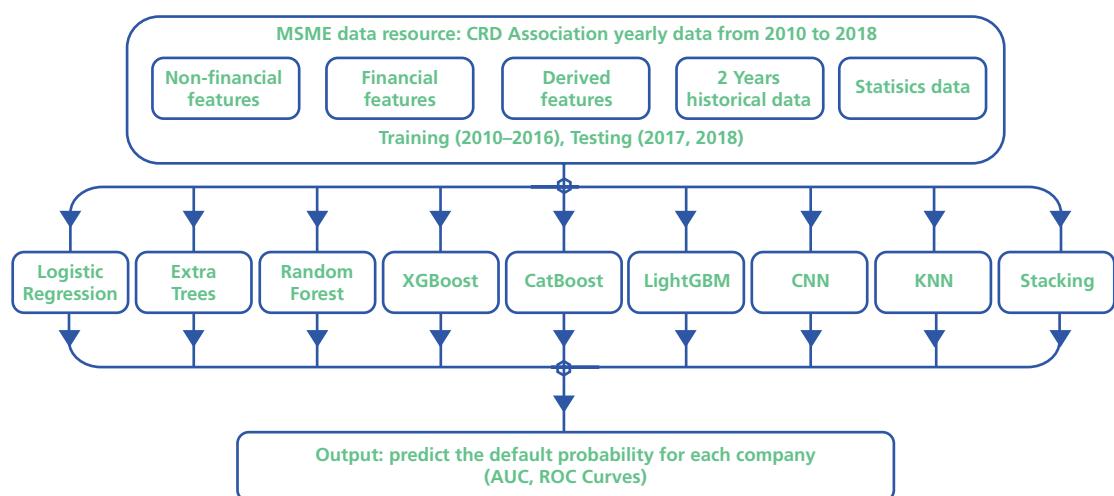
Table 3.4 Definition of data groups

Group	Definition	#No of Variables
Group 1 Original CRD dataset	CRD Association's original MSME dataset (Category ①)	20
Group 2 Cashflow	Cashflow-related variables (OR07–OR11, OR20, DV05, DV06, DV8, DV9, DV11, DV14, DV16, DV24)	14
Group 3 Cashflow + Derived variables	Cashflow-related variables and derived variables (OR07–OR11, OR20, Category ②)	30
Group 4 Full dataset	Dataset containing all five categories variables (Category ①+②+③+④+⑤)	145

1.1.5 Proposed methodology for evaluating different machine learning algorithms

Nine different machine learning algorithms were tested, and the accuracies of their default predictions were compared to evaluate their performance. As illustrated in Fig 3.1, the nine machine learning algorithms were Logistic Regression, Extra-Trees, Random Forest, XGBoost, CatBoost, LightGBM, Convolutional Neural Network (CNN), k-Nearest Neighbours (k-NN) and Stacking. Descriptions of these machine learning algorithms can be found in Appendix B of this paper. For each algorithm, the prediction model was trained independently with hyperparameter tuning.

All of the experiments were run on a Windows 10 server (Intel CPU with 16 cores and 64GB of RAM) in a secure environment on the premises of the CRD Association. The outputs of the experiments that were related to the performance of the machine learning algorithms were then exported for analysis. To examine the performance of the machine learning algorithms, the results of the experiments were represented as AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curves, as these are able to measure the discrimination power of a given model. Moreover, the AUC metric suits the unbalanced data scenario.


Figure 3.1 Proposed methodology for evaluating different machine learning algorithms
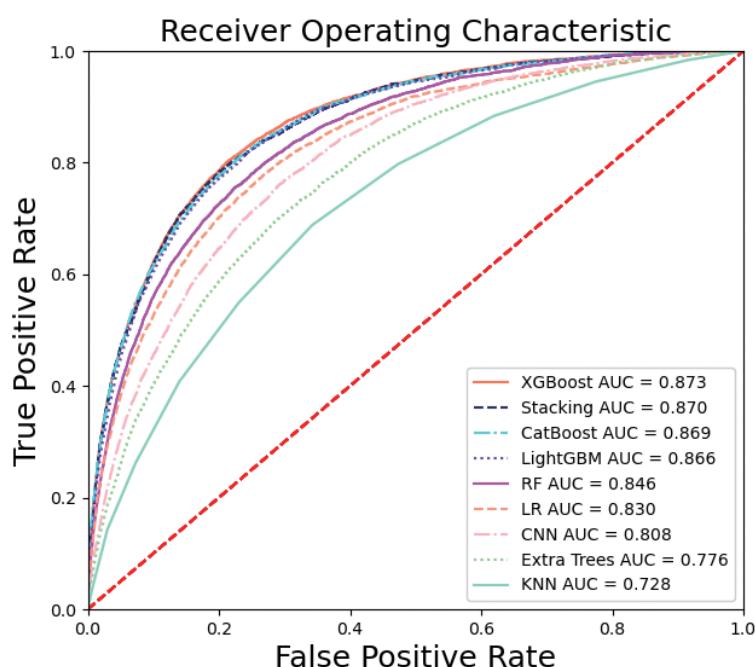
1.2 Results of the comparison of machine learning algorithms

1.2.1 AUCs of the machine learning algorithms

Figure 3.2 shows the ROC curves representing each machine learning algorithm's predictions of the one-year default probability for the full dataset (Group 4). There is no concrete threshold of AUC that signifies a functional model. Usually, the AUC is equal to 0.5 for a random model and approaches 1 as a model approaches perfection.

Generally, the boosting algorithms such as XGBoost, CatBoost and LightGBM showed greater predictive power than Logistic Regression (the traditional model for credit scoring), and their results were also better than conventional machine learning algorithms such as k-Nearest Neighbours. More specifically, the XGBoost reported the best AUC result among all of the models in this experiment.

As for the performance of the Stacking approach in this experiment, the first-level classifiers used k-NN, Extra-Trees and XGBoost and Logistic Regression was used as the stacker. The Stacking algorithm showed outstanding predictive capability (AUC = 0.870), much better than the other algorithms.


Figure 3.2 ROC curves of all algorithms on the full dataset
for predictions of one-year default probability

Table 3.5 shows the AUC details for all the algorithms' accuracy in predicting one-year default probability for the four data groups. The Group 2 dataset received the lowest AUC scores apart from the Group 1 dataset using KNN, as this group contains the least information and the smallest number of variables. The AUCs scores of the Group 3 dataset were slightly higher than those of Group 2, as Group 2 is a subset of Group 3. The increase in the number of derived financial ratio variables helped to improve the accuracy of prediction. Many models achieved better prediction accuracy using the Group 1 dataset, because Group 1 contains more information about the entities than either Group 2 or Group 3, including not only cash status, financial status and loan status, but also the company's non-financial information for the current financial year. The AUC scores of the Group 4 dataset had the best results among the four data groups because the Group 4 dataset contained the richest information, including historical trend data and statistical information, enabling the profiles of the companies to be built in multiple dimensions.

The only exception was CNN, which showed slightly atypical outcomes for the four datasets. It achieved the best AUC result on the Group 1 dataset, and the result on Group 4 was only the second-best. This outcome was mainly caused by the neural network's randomness and its loss function. First, there are various sources of randomness in the training of deep neural network models,[34] including but not limited to random parameter initialisation, random sampling of examples during training and random dropping of neurons. Second, the goal of the training of CNN is to find a summation of weights and biases that have low loss across training or validation sets (loss minimisation[35]). However, a decrease in loss does not necessarily lead to an increase in AUC accuracy.

34. Madhyastha, P., & Jain, R. (2019). On Model Stability as a Function of Random Seed. arXiv preprint arXiv:1909.10447.
35. Alpaydin, E. (2020). *Introduction to Machine Learning*. MIT Press.

**Table 3.5 The AUCs of all models on the four data groups
for predicting one-year default probability**

Prediction Period	One-year			
Data Group Model	Group 1 Original CRD	Group 2 Cashflow	Group 3 Cashflow + Derived variables	Group 4 Full dataset
KNN	0.5883	0.6427	0.6798	0.7276
Logistic Regression	0.8274	0.7762	0.7830	0.8303
Random Forest	0.8445	0.8278	0.8437	0.8457
CNN	0.8248	0.6419	0.7025	0.8080
Extra-Trees	0.6121	0.7044	0.7330	0.7765
LightGBM	0.8595	0.8400	0.8527	0.8665
CatBoost	0.8623	0.8428	0.8553	0.8694
XGBoost	0.8647	0.8431	0.8569	0.8730

Usually, banks are interested in one-year default probability. The default probabilities for two years or more are used as references, as predictions for a longer period may include many more uncertainties. To examine the prediction results for a two-year time window, another set of experiments was carried out, as shown in Table 3.6. In this round of experiments, all of the models adopted the hyperparameters trained from the one-year prediction data to make sure the two sets of experiments were run under the same configuration. The results showed that all of the AUC scores dropped slightly as the prediction period was extended. Moreover, the distribution of the AUCs for each model across the four data groups was very similar to that of the one-year prediction results. It is worth noting that Stacking achieved the best AUC result (AUC = 0.8424) when using LightGBM, Random Forest and XGBoost as the first-level classifiers and Logistic Regression as the stacker.

Table 3.6 The AUCs of all models on the four data groups for predicting two-year default probability

Prediction Period	Two years			
Data Group Model	Group 1 Original CRD	Group 2 Cashflow	Group 3 Cashflow + Derived variables	Group 4 Full dataset
KNN	0.5688	0.6198	0.6361	0.6998
Logistic Regression	0.7923	0.7610	0.7645	0.8026
Random Forest	0.8189	0.8064	0.8200	0.8224
CNN	0.7885	0.6875	0.7205	0.7530
Extra-Trees	0.6190	0.7051	0.7140	0.7612
LightGBM	0.8308	0.8181	0.8226	0.8385
CatBoost	0.8351	0.8220	0.8276	0.8444
XGBoost	0.8368	0.8201	0.8273	0.8444

1.2.2 Feature importance of machine learning algorithms

Selecting the most important feature variables is a key step in building a prediction model for many machine learning algorithms. Table 3.7 shows an overview distribution of the top 20 most important feature variables for the machine learning algorithms deployed in this experiment, except for the k-NN, CNN and Stacking algorithms. The results were generated by the feature importance library functions from the trained models. Determining feature importance is, however, not required for model building by the k-NN, CNN or Stacking algorithms. These three algorithms rely on nearest neighbours, neural networks and selected classifiers or stacker, respectively, for model development.

Table 3.7 Overview of the top 20 most important features by category

	Model Category Name	Logistic Regression	Random Forest	Extra-Trees	LightGBM	CatBoost	XGBoost
①	Original Variables	5	3	2	10	7	6
②	Derived Variables	3	12	4	9	10	10
③	Time Window	10	—	6	—	3	2
④	Statistical Data	2	5	7	1	—	2
⑤	Default History	—	—	1	—	—	—

As shown in Table 3.7, Logistic Regression takes more historical data into consideration when making a decision. The Bagging algorithm used in Random Forest and boosting algorithms like LightGBM, CatBoost and XGBoost focus more on information about the latest financial year. Extra-Trees collects information from all kinds of variables for decision-making. For a more in-depth examination, the top 10 most important features of the relevant algorithms can be found in Table 3.8.

Table 3.8 Top 10 most important features

Ranking	Logistic Regression	Random Forest	Extra-Trees	LightGBM	CatBoost	XGBoost
1	DV10②	DV24②	DV04②	OR04①	DV22②	DV24②
2	DV08③	DV22②	DV05②	DV24②	DV24②	OR14①
3	DV24②	DV24④	DV04③	DV14②	OR17①	DV22②
4	OR17①	DV11②	DV05③	OR09①	DV11②	OR03①
5	DV04②	DV24④	DV11④	OR14①	OR04①	DV02②
6	OR10①	DV01②	DV11④	DV08②	DV08②	DV11②
7	OR13①	OR14①	DV24②	DV03②	DV03②	DV13③
8	OR03①	DV21②	DV04③	OR08①	OR09①	OR17①
9	DV10③	OR17①	OR03①	DV10②	DV10②	DV08②
10	OR17③	DV02②	DV24④	OR03①	OR14①	OR05①

Table 3.8 shows that the same variable may have different importance in different credit scoring models. The financial variables Total Equity (OR14), Net Profit (OR17) and Cash to Debt Ratio (DV24) are among the top 10 most important features for almost all of the models. The non-financial variable Business Type (OR03) is one of the most critical independent variables for Logistic Regression, Extra-Trees, LightGBM and XGBoost.

Logistic Regression, LightGBM, CatBoost and XGBoost build the model using different kinds of cash, financial and non-financial variables. In contrast, three out of the ten most important features in the Random Forest model are related to Cash to Debt Ratio (DV24). Extra-Trees focuses on the time window of historical trends as well as the statistical status of a few derived financial variables such as the Boolean value representing Net Profit (DV04), the Boolean value representing Operating Cashflow (DV05), Cash Ratio (DV11) and Cash to Debt Ratio (DV24).

1.3 Insights gained from applying machine learning algorithms to MSMEs' financial data

The datasets of the CRD Association contain both financial and non-financial yearly records of MSMEs. The datasets contain 3.6 million observations from 730,000 MSME industries over a substantial period of time (2010–2018). Using this large quantity of observations, the following insights can be drawn from the experiments:

- The selected machine learning algorithms achieved a desirable level of prediction accuracy, demonstrating the effectiveness of performing default prediction by running the machine learning algorithms on MSME bank account information.

- Each machine learning algorithm in the experiments showed a different degree of predictive power for the credit scoring of MSMEs. A detailed analysis can be made by comparing the performance of model training and the accuracy of default prediction, as described in the next section.

- The predictive power of the machine learning models based on the transactional cashflow data of MSMEs (the Group 2 dataset) showed a reasonable level of accuracy. This demonstrates that transactional cashflow data are suitable for use as alternative data for credit scoring.

1.3.1 **Comparison of different machine learning algorithms**

Analysing the benefits and limitations of different machine learning algorithms is an essential step before selecting a machine learning algorithm to match a particular credit scoring scenario. The results of the experiments offer insights into how the machine learning algorithms compare with each other.

Some of the proposed credit scoring models showed strong predictive power in assessing the creditworthiness of MSMEs. In general, boosting algorithms showed stronger predictive capability than Bagging or traditional machine learning algorithms. More specifically, Logistic Regression was good at detecting multicollinearity among strongly correlated variables; therefore, it paid more attention to entities' historical data when building the model. Logistic Regression uses a maximum likelihood estimator, through which explanatory variables/features are limited to those that can be rationalised and interpreted by the banks. Although any model in general is likely to have better prediction results when more independent variables are involved, it is also important to understand the relative importance of the variables so that the selected model can be more intuitive and explicable.

KNN is one of the traditional machine learning algorithms; it depends on the majority votes of its k neighbours for default prediction in this experiment. It only takes a few seconds to build the model but requires a long time to make a decision based on the top k-nearest neighbours' votes. Random Forest constructs multiple trees in randomly selected subspaces of the independent variable space to overcome generalisation biases. It shows great stability when dealing with imbalanced data. The AUC of the Random Forest on Group 4 was 0.8457 for a 1% default ratio dataset.

The Boosting algorithms LightGBM, CatBoost and XGBoost showed generally comparable predictive power in the experiments. XGBoost had more predictive strength than the other techniques due to its regularisation capabilities (which avoid overfitting and bias), its handling of missing values, and its cross-validation. However, XGBoost takes a relatively long time to build the model (over five hours in the experiments). The computational time required for training should also be taken into consideration when selecting a machine learning algorithm.

CatBoost is good at handling situations involving categorical data and complex dependencies. LightGBM is designed to improve efficiency and scalability for high-dimension datasets. It achieved good accuracy in performance and required less than one minute to build the model in the experiments. LightGBM could, therefore, be adopted efficiently for certain demanding scenarios in which a new training model needs to be revised frequently based on incoming data.

As for the Stacking algorithm, all of the machine learning algorithms' prediction results were processed as its inputs. The algorithm took about 15 minutes to build one set of combinations of classifiers and stackers and required nearly 50 hours to identify the best combination from all potential combinations. An increase in the number of potential classifiers, stackers and levels would dramatically increase both the time and resources required for processing, but the improvement in accuracy may not be worth the data processing effort required. Compared with other machine learning models, Stacking also required more effort for model development and hyperparameter tuning.

1.3.2 Insights from transactional cashflow and non-cashflow data

The Group 2 dataset used for model development and default prediction included only the inflow and outflow details of the MSME transactional cashflow data. Transactional cashflow data are not usually regarded as conventional data for credit scoring. However, the results of the experiments show that almost all machine learning algorithms can generate reasonably good predictions using the Group 2 dataset (i.e. the cashflow-only data). The experiments also showed that better prediction results can be achieved if more relevant non-cashflow data are included in model training and model prediction.

It is worth pointing out that the performance of the models using machine learning algorithms could be further improved if the timeframe of the transactional cashflow data is reduced from yearly intervals to monthly or daily intervals. Moreover, the models could also be refined by categorising the transactional records of cashflow data according to the types of revenue and expenses they record, such as gains from investment and expenses for marketing. These changes would enable machine learning algorithms to identify patterns of MSMEs in financial distress within a shorter period. The above insights helped in the development of the generalised framework for alternative credit scoring proposed in the next section of this paper.

1.4 Case Study: Credit scoring of Japanese MSMEs

A joint research collaboration[36] by the CRD Association, the Bank of Japan and Resona Bank attempted to assess MSME credit risk via the bank account transaction information of MSME borrowers. The study used approximately 6,000 variables and applied two popular machine learning algorithms, Random Forest and XGBoost.

There were two initial challenges posed by the research: the large volume of data due to the extremely high frequency of monthly transactions recorded, and the very large number of features that could be included. While the former challenge can be tackled through technical means, the latter requires a serious exercise drawing on expertise in financial and credit risk analysis. This is because although machine learning can process big datasets and provide highly accurate predictions without really understanding the features it uses, credit risk modelling in the banking sector requires that the model features be comprehensible to and interpretable by lending and risk officers. Aiming for a non-black box prediction model, the research team attempted to organise the numerous available features into ten big comprehensible categories representing the purpose of the transactions, and used these categories as the foundation for interpreting the model results.

The study found that machine learning models using transactional data are highly accurate in predicting short-term default probability. Features related to monthly average cash balance are the best predictors of default. Features related to cash inflows from revenue and bank loans and those related to cash outflows such as variable costs and cost of goods sold are the next-best predicting features.

The model could serve as a credit scoring tool for MSMEs because the target level of accuracy has been achieved. More specifically, model accuracy, as measured by the accuracy ratio (AR), peaked at 0.707 (an equivalent AUC of 0.854) for the Random Forest algorithm and at 0.733 (an equivalent AUC of 0.867) for the XGBoost algorithm. These accuracy levels are generally acceptable to the conservative Japanese banking sector.

36.　三浦翔., 井實康.幸., & 竹川正.浩. (2019, June). 入出金情報を用いた信用リスク評価 ―機械学習による実証分析―. 日本銀行. https://www.boj.or.jp/research/wps_rev/wps_2019/wp19j04.htm/

The joint research project can be considered a successful attempt to take advantage of readily available resources (transactional data), field knowledge (financial expertise) and technology (machine learning) to solve a financial inclusion problem that concerns many governments and economies.

Acknowledgments for contributions to this section:

Contributor	Contributions
Dr Lan H. Nguyen, CRD Association and CRD Business Support Ltd.	The collaborative work in conducting the experiments described in this section.

2. Evaluating the technical feasibility of a proposed framework

The conventional approach to credit scoring extracts financial data from financial statements, whereas alternative credit scoring approaches evaluate the creditworthiness of MSMEs based on alternative data. The reference framework proposed in this paper uses two dimensions to categorise alternative data for credit scoring. The first one focuses on evaluating the transactional data of MSMEs, which includes the inflow and outflow of cash in MSMEs' bank accounts, as well as details of the revenue and expenses of MSMEs supplied by third-party data providers. The second one focuses on alternative data according to the industry sector that individual MSMEs belong to.

To explore the feasibility of the proposed reference framework, experiments for proof-of-concept implementations have been conducted. The scope of the experiments involved:
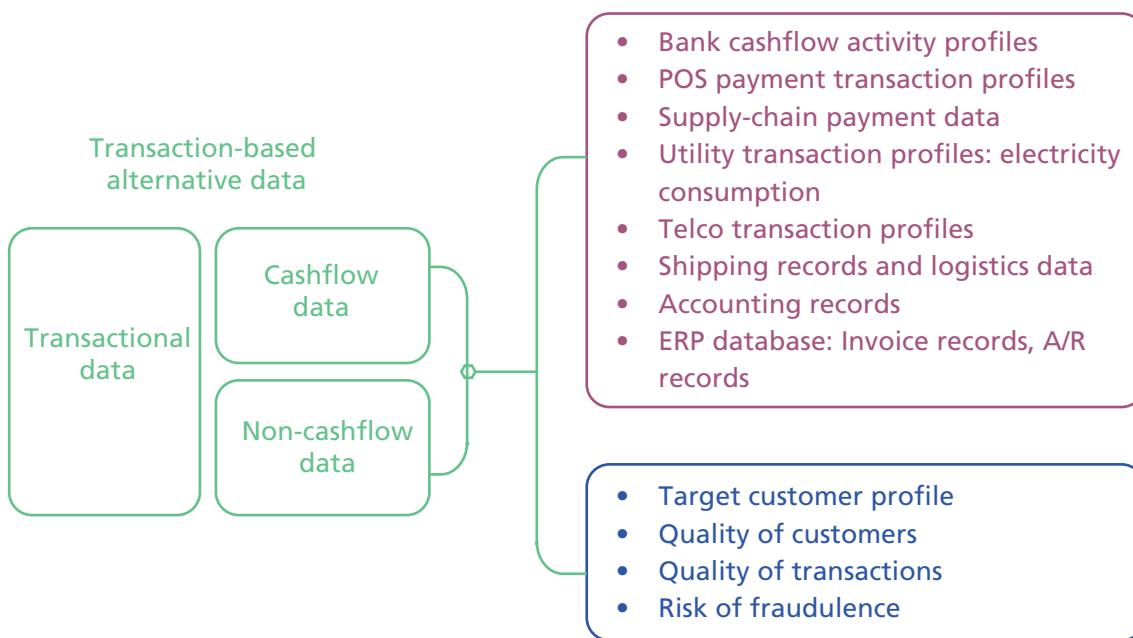
- Testing the effectiveness of running the machine learning algorithms on transactional cashflow data to build a transactional cashflow model.

- Testing the feasibility of building machine learning models by running the machine learning algorithms on the datasets of third-party data providers.

2.1 Proposed framework for alternative credit scoring

The simplest way to apply machine learning algorithms to transactional records is to convert all the features of the available transactional data into independent variables for the default prediction model. However, the complexity of machine learning models increases dramatically with the increase of the number of independent variables, as more alternative data are collected from data providers that banks acquire continuously. To manage this complexity, the independent variables of the credit scoring models should be categorised into groups, and a designated score for each group should be generated by the machine learning algorithms. This paper proposes a generalised framework called an *Industry-specific alternative credit scoring framework* based on segregation by data type and segregation by industry sector.

### 2.1.1 Segregation by type of transactional data

**Figure 3.3 Transaction-based alternative data
for alternative credit scoring**

If segregation by data type is carried out as shown in Figure 3.3, the framework will consist of two models: a transactional cashflow model and a transactional non-cashflow model.

Banks can build their transactional cashflow models for alternative credit scoring using their existing bank transactional data. This includes all cash inflows to and outflows from the bank accounts that receive revenue and settle expenses for MSMEs. Generally, banks possess large quantities of transactional data of credible quality, but only recently have banks realised that they could use this vast amount of data to their advantage.

Banks in different countries have started to use machine learning algorithms to analyse the cash movements in the bank accounts of loan applicants. The application of transactional data to credit scoring has recently become popular in Japan as an alternative to traditional financial data. Resona Bank, one of the four Japanese megabanks, recently introduced a credit line that only requires MSMEs to have had bank accounts at the bank for a certain period to be eligible for loan screening. The screening is mainly done by machine learning algorithms that analyse cash movements in the applicant's bank account.

As for transactional non-cashflow data, it represents data that can enrich the transactional cashflow data and can also be used to predict indirectly the creditworthiness of companies applying for loans. Transactional non-cashflow data are usually owned by third-party data providers that are involved in transactions for goods or services between end-customers and MSMEs. These data providers collect not only the cashflow-related information from the transactions but also other details related to the delivery of goods or services. For example, the profiles of end-customers and the quality of the business of MSMEs are typical non-cashflow transactional data.

Generally, any independent variables can be grouped as transactional non-cashflow data if they are not directly related to the inflow and outflow of cash. However, if there are too many independent variables in the transactional non-cashflow model to be manageable, it is advisable to group them into sub-models based on the type of transactional non-cashflow data. An advantage of having sub-models is that it makes the overall result of the transactional non-cashflow model more explicable because each sub-model carries a specific meaning, such as customer quality or transaction quality.

2.1.2 Segregation by industry sector

Companies in the same industry sector tend to have similar business operation models. Consequently, similar types of transactional data can be used to explore the status of their business operations and derive insights into their creditworthiness. The idea of developing sector-based or cluster-based models for credit scoring facilitates specialisation in analysis. This approach organises businesses of diverse natures into multiple clusters, with businesses in each cluster having similar traits. This allows the credit scoring process to focus on analysing businesses in the same cluster, and which have similar data points.

The industry-specific alternative credit scoring framework proposed in this section is a framework that can be applied to MSMEs in different industry sectors. The benefits of specialised alternative credit scoring by industry sector include the following:
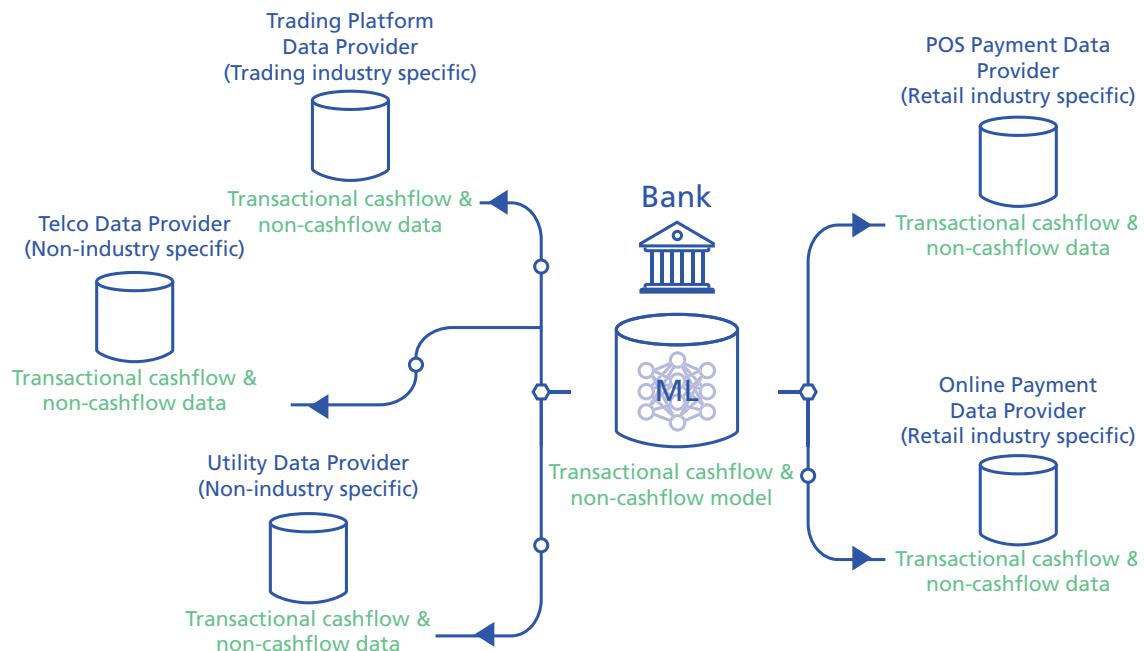
- Specialisation in default prediction — Industry sector peers share similar operational models and, therefore, similar independent variables for default prediction.

- Specialisation in monitoring — Better monitoring of credit risk can be achieved because valuable insights can be gained from monitoring the assessment results of industry sector peers, which are likely to have similar patterns of results during different economic environments.

- Specialisation in benchmarking — Performance results among industry sector peers are comparable, making ranking among peers possible.

2.1.3 Development of the Industry-specific Alternative Credit Scoring

In the transactional data of companies from different industry sectors, some of the data can be classified as industry-specific data. For example, the shipment records of goods exported to overseas buyers represent transactional data of MSMEs in the trading industry, and point of sale (POS) payment transaction records represent transactional data of MSMEs in the retail industry.

However, some transactional data can be classified as non–industry-specific data because MSMEs in any industry sector can generate these types of transactional data. Some examples include:

- historical cashflow records of MSMEs' bank accounts;

- bank account statements submitted by MSME applicants; and

- transactional records from telcos, utilities and other data providers.

Figure 3.4 Industry-specific alternative credit scoring framework

As shown in Figure 3.4, the industry-specific alternative credit scoring framework can be adopted by financial institutions by leveraging transactional cashflow and non-cashflow data from various data providers. Some of the data providers offer industry-specific data and some of them non-industry-specific data. Table 3.9 shows examples of transactional cashflow and non-cashflow data and their respective data sources for all industry sectors, the retail industry, and the trading industry.

Table 3.9 Examples of two-dimensional alternative data for industry-specific alternative credit scoring

Industry sector	Data source	Examples of transactional cashflow data	Examples of transactional non-cashflow data
All industry sectors	**Historical transactional cashflow records of MSMEs' bank accounts**	*For model training* • Cash inflow activities (revenue) • Cash outflow activities (expenditure) • Default history	N/A
	Bank account statements submitted by MSME applicants	*For default prediction* • Cash inflow activities (revenue) • Cash outflow activities (expenditure)	N/A
	Telco, utility or other data providers	*For model training & default prediction (based on expenditure data)* • Consumption history (utility metre reading, communication data, etc.) • Bill Amounts • Payment History • Defaults/late payments/dunning records	*For model training & default prediction* • Choices of product subscribed to (e.g. value-added services & packaged services) • Instalment plan requests • Address (district + building) • Business category/trade class
Retail Industry	**POS payment transactional records from third-party data providers**	*For default prediction (based on revenue data):* • Sales transaction amounts • Sales transaction frequency • Daily/monthly averages of sales transaction amounts • Payment types	*For model training & non-cashflow default prediction:* • Percentage of recurring customers • Number of problematic transactions (refunded, voided & reversed) • Transaction risk information
	Suppliers from third-party data providers	*For default prediction (based on expenditure data):* • Ordering transaction amounts • Ordering transaction frequency • Daily/monthly average of ordering transaction amounts • Repayment history	*For model training & non-cashflow default prediction:* • Statistics of recurring suppliers • Patterns of ordering • Number of problematic transactions (refunded, voided & reversed)
Trading Industry	**Shipment transactional records from third-party data providers**	*For default prediction (based on revenue data):* • Shipment amounts • Shipment frequency • Monthly/yearly average of shipment amounts • Time period for shipments • Trade finance history	*For model training & non-cashflow default prediction:* • Statistics of recurring clients • Import & export locations • Transaction risk information
	Logistic services from third-party data providers	*For default prediction (based on expenditure data)* • Logistic movement amounts • Logistic movement frequency • Daily/monthly average of logistic movement amounts • Cost of logistics services	*For model training & non-cashflow default prediction* • Patterns of utilisation in warehousing & transportation • Operational efficiency • Records of performance KPI

2.1.4 Adoption of the proposed framework

Deploying the industry-specific alternative credit scoring framework would involve the following steps for banks offering lending services.

Phase one adoption

1. Transactional cashflow model training based on historical cashflow records of MSMEs' bank accounts
   A transactional cashflow model for any industry sector would need to be developed by the bank, based on the historical cashflow data of its MSME clients. Model training based on machine learning algorithms would be generated by using the details of cash inflows and outflows in the MSMEs' accounts and their records of default or delinquency over a set period.

2. Alternative data collection
   When reviewing a new loan application, transactional cashflow data and non-cashflow data of the applicant would need to be collected from the relevant data providers. Upon receiving the consent of the applicant, the required data would be requested by the bank, namely:

   • bank account statements of the applicant uploaded online; and

   • transactional records from data providers that provide industry-specific and non-industry-specific data on the applicant through an API.

3. Applying the transactional cashflow model to review the loan application
   As the revenue and expenditure data from different data providers will at least partially resemble the loan applicant's cashflow patterns, default prediction can be generated based on the transactional cashflow model.

Phase two adoption

4. Transactional non-cashflow model training
    The non-cashflow data and the default and delinquency records collected on previous loan applicants by the various data providers over a set period would enable banks to develop a non-cashflow model.
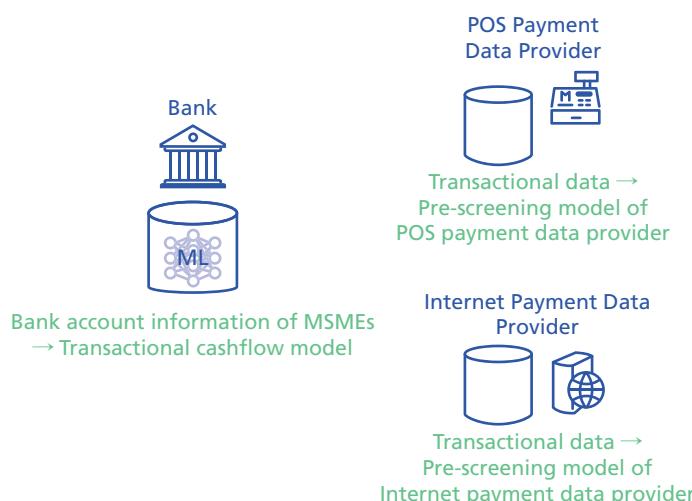
5. Applying both the transactional cashflow and non-cashflow models for credit scoring
    With the transactional cashflow and non-cashflow models in place, default prediction could be performed based on both transactional cashflow and non-cashflow data from the relevant data providers. Before the non-cashflow model is developed, the bank can engage the data providers to conduct pre-screening models using the data providers' own datasets, as described in Part One, Section 3.2.2 and Part Three, Section 2.2.1. The MSMEs in the whitelist generated by the pre-screening model are potential clients that can further be assessed by the bank.

2.2 Retail Alternative Credit Scoring (RACS): a demonstration

### 2.2.1 A RACS scenario for the proof-of-concept (POC)

Figure 3.5 A Retail Alternative Credit Scoring (RACS) scenario for the proof-of-concept

Retail Alternative Credit Scoring (RACS) is an industry-specific alternative credit scoring framework for the retail sector. Figure 3.5 shows the three participating organisations involved in the proof-of-concept (POC) implementation in a RACS scenario, namely a bank, a POS payment data provider, and an Internet payment data provider. This POC describes how the bank account information of participating MSMEs can be used to develop transactional cashflow models (see Part Three, Section 2.1.1), and how the transactional data of the participating data providers can be used to develop pre-screening models (see Part One, Section 3.2.2).

This RACS POC study has certain limitations:

- There was no sharing of data among the participating organisations. Only results related to the performance of the machine learning models are reported in this section. The machine learning models involved in this POC were developed based on the transactional data of each organisation's dataset, and the experiments took place at the premises of each organisation. Due to this limitation, only the transactional cashflow model could be developed, based on the transactional cashflow data of the bank. The transactional non-cashflow model could not be developed from this POC.

- As the datasets of the Internet payment and POS payment data providers do not contain MSME loan default information, information on service charge payment delinquency by MSMEs was used for its resemblance to loan default information. Delinquency of service charge payments became the prediction target of the POC experiments. The machine learning model developed based on the transactional data is called the pre-screening model, which is used to identify MSMEs with potential loan default problems.

- The transactional data held by the Internet payment and POS payment data providers represent only the inflow of cash (i.e. the revenue) of MSMEs. The data for RACS would be better-rounded if transactional data related to the outflow of cash (i.e. the expenses) of MSMEs was also included for assessment.

- Only specific variables of the datasets of the participating organisations were used for the POC experiments. The results of the experiments have therefore been used solely to demonstrate technical feasibility, and not to ascertain the best prediction accuracy that the datasets can generate.

The objectives of this RACS POC are to demonstrate that:

- Banks are able to use the historical transactional cashflow data of MSMEs' bank accounts to develop a transactional cashflow model for alternative credit scoring. The required transactional cashflow data is available from Internet payment and POS payment data providers.

    Transactional data (both cashflow and non-cashflow data) provided by Internet payment and POS payment data providers can be used to develop a pre-screening model that can be used to predict the probability of MSMEs developing potential financial problems. Banks can use the pre-screening model to white-list MSMEs with a low risk of financial problems, which can be pre-qualified for loan applications.

2.2.2 The transactional cashflow model developed by the bank

The bank's dataset
The historical monthly transactional cashflow data (inflows and outflows) of the MSMEs' accounts at the participating bank were used to develop the transactional cashflow model. The dataset of the participating bank for the experiments described in this section contained 74,000,000 monthly observations of more than 1,000 MSMEs for the period from October 2018 to July 2020. Each observation represents the monthly summarised data of the bank account of an MSME. As shown in Table 3.10, apart from the unique identification of each MSME, there are six original variables. The second variable is the indicator of the occurrence of delinquency in the month, and is used as the dependent variable for prediction. Based on the original variables, extra variables are derived using different statistical values of the original variables. Table 3.11 shows the derived variables of the bank's dataset.

Table 3.10 Original variables of the bank's dataset (monthly observations)

Original Variable	Description
1	Unique ID of MSME
2	Occurrence of delinquency
3	Total cash credit amount
4	Total cash debit amount
5	Number of credit transactions
6	Number of debit transactions
7	Number of years of bank relationship of client (or account)

Table 3.11 Derived variables of the bank's dataset

Derived Variable	Description	Number of Variables
1	Average credit/debit per transaction	2
2	Time window related variables (for different periods of the time window)	80
3	Statistical data for total cash credit amount, such as minimum, maximum, mean, standard variation, and range	15
4	Statistical data for total cash debit amount, such as minimum, maximum, mean, standard variation, and range	15
5	Statistical data for the number of credit transactions, such as minimum, maximum, mean, standard variation, and range	15
6	Statistical data for the number of debit transactions, such as minimum, maximum, mean, standard variation, and range	15
7	Statistical data for the number of years, such as minimum, maximum, mean, standard variation, and range	15

Development of the transactional cashflow model

The experiment aimed to build a transactional cashflow model that could predict whether a default target event will occur within a specific period. The target event is defined by the original variable #2 (occurrence of delinquency). For model training, the default observation period was 24 months (from October 2018 to September 2019). As for model prediction, the default prediction period was 10 months (from October 2019 to July 2020). Table 3.12 shows the AUC scores of the nine selected machine learning algorithms for different prediction periods (one month ahead, two months ahead, and three months ahead). When applying Stacking for prediction, the observation period (from October 2018 to September 2019) was separated into training and validation sets. The stacker of the Stacking model with the highest validation score was considered to be the final stacking combination. Only the test scores of the final stacking combinations are reported in Table 3.12.

Table 3.12 Experiment results of the transactional cashflow model

Prediction Period	One Month ahead	Two Months ahead	Three Months ahead
Feature / ML Algorithm	Original & Derived Features	Original & Derived Features	Original & Derived Features
Logistic Regression	0.9162	0.8098	0.7592
Random Forest	0.9249	0.8312	0.7606
Extra Trees	0.9198	0.8287	0.7727
LightGBM	0.9141	0.8187	0.7987
CatBoost	0.9339	0.8518	0.7641
XGBoost	0.9368	0.8436	0.7926
KNN	0.7783	0.7699	0.7681
CNN	0.8489	0.7770	0.7265
Stacking	0.9227	0.8103	0.7515

The results show that XGBoost, CatBoost, and Random Forest performed best in short-term predictions. The performance of KNN was undesirable. Accuracy in predicting default probability dropped as the prediction period increased. For example, the AUC scores of XGBoost are 0.9368, 0.8436, and 0.7926 for the periods of one month, two months, and three months respectively. The shorter the prediction period, the better was the model's prediction capability.

2.2.3 Model development by the POS payment data provider

Dataset of the POS payment data provider
The dataset of the POS payment data provider for the experiments contained nearly 1.55 million observations of more than 300 MSMEs for the period from April 2019 to April 2020. Each observation represents a POS payment transaction record of an MSME. As shown in Table 3.13, apart from the unique identification of each MSME, there are six original variables. Table 3.14 shows the derived variables of the dataset for building pre-screening machine learning models.

The POS payment data provider is a conglomerate offering not only POS payment services but also a wide variety of other services to MSMEs. Therefore, it was able to use the existence of late payment data for any service that MSMEs have engaged as resembling the default scenario of the MSMEs.

Table 3.13 Original variables of the POS payment data provider's dataset
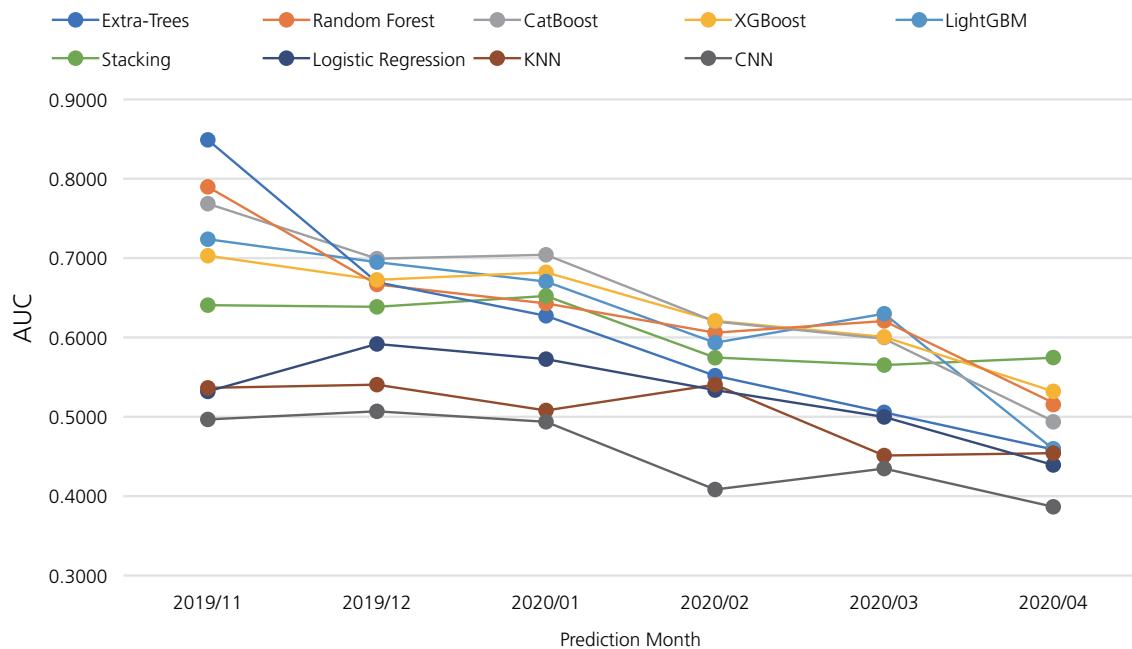
Original Variable	Description
1	Unique ID of MSME
2	Occurrence of late payment of any service charge (to resemble the default information)
3	Company type
4	Transaction date and time
5	Currency
6	Payment amount
7	Payment method, e.g. cash, Visa, Apple Pay, and Alipay etc.

Table 3.14 Derived variables of the POS payment data provider's dataset
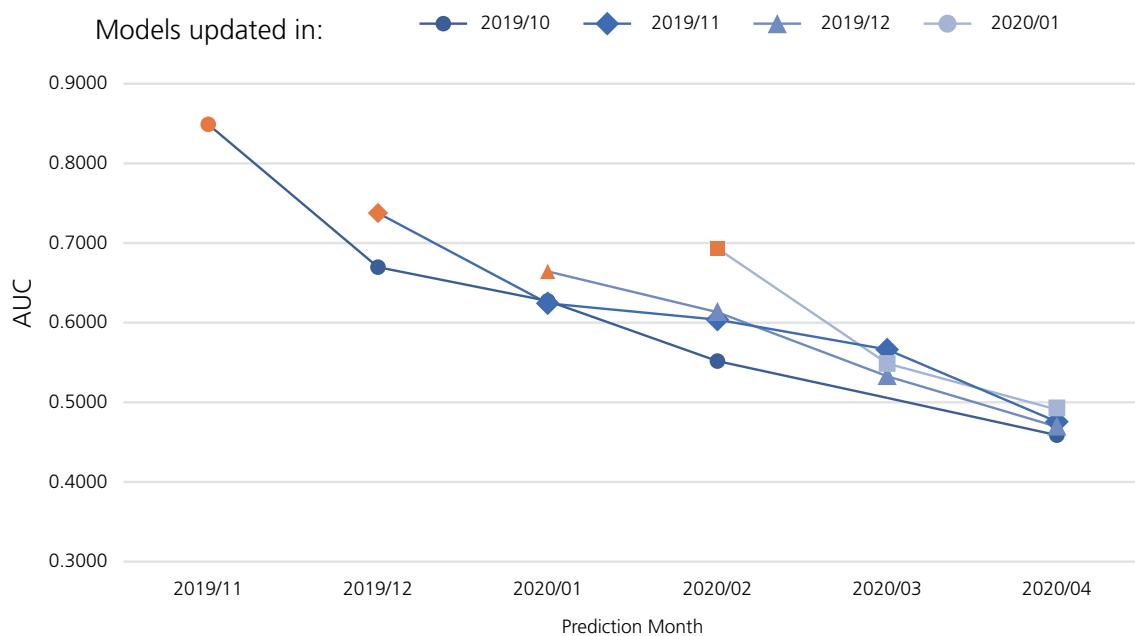
Derived Variable	Description	Number of Variables
1.	Monthly transaction time related variables	3
2.	Monthly transaction statistical data, such as min, max, mean, standard variation, summary	19
3.	Time window related variables, such as 14, 30, 60 days	10
4.	Payment method derived variables, such as total amount, total number and mean of each payment method	78
5.	Payment status derived variables, such as total amount and total number of each payment status	12

Development of the pre-screening models by the POS payment data provider


Figure 3.8 Performances of the pre-screening models by the POS payment data provider

The pre-screening models were developed using the monthly transaction data of the POS payment data provider from April 2019 to October 2019. In this experiment, monthly AUC scores were investigated in order to explore their tendencies and the prediction capabilities of the different algorithms. Figure 3.8 shows the AUC scores of the nine selected machine learning algorithms for predicting the occurrence of late service payments (i.e. those resembling loan defaults), based on the developed pre-screen models. Each line presents the results of a different machine learning algorithm. The figure shows that the predictive capabilities of the top five pre-screening models were reasonably good for the period from November 2019 to January 2020, but the accuracy of all the models dropped gradually for the whole period from November 2019 to April 2020. The top five machine learning algorithms were Extra-Trees, Random Forest, CatBoost, LightGBM, and XGBoost.


**Figure 3.9 Performance depending on updating of
the Extra-Trees pre-screening model**

Another set of experiments was carried out to compare the AUC scores of a specific pre-screening model for different re-training periods. Each line in Figure 3.9 presents the AUC scores of the Extra-Trees pre-screening model that was updated (i.e. underwent model re-training) in a specific month. The four lines show the prediction results of the model after updating in October 2019, November 2019, December 2019, and January 2020.

Figure 3.9 shows that the best prediction results were always achieved by the pre-screening model with the most up-to-date re-training. Predicting capability dropped as the prediction period increased. The more recent the model updating time, the better was the prediction accuracy.

2.2.4 Model development by the Internet payment data provider

Dataset of the Internet payment data provider
The dataset of the Internet payment data provider for the experiments contained over 10 million observations of more than one thousand MSMEs for the period from May 2019 to June 2020. Each observation represents an Internet payment transaction of an MSME. As shown in Table 3.15, apart from the unique identification of each MSME, there are six original variables.

As the datasets of the Internet payment data provider contained no information about MSME loan defaults, the delinquency of service charge payments by MSMEs was used instead as resembling default information. Table 3.16 describes the extra variables that were generated based on late service charge payments.

The entries in the Transactional Dataset 1 were first aggregated into a daily dataset which included the minimum, maximum, mean, and total of the number of the transactions of each MSME. These daily datasets were then processed to form a monthly aggregated dataset, which was then merged with the late service payment indicator as shown in Table 3.17.

Table 3.15 Original variables of the Internet payment data provider's dataset (Transactional Dataset 1)

Original Variable	Description
1	Unique ID of MSME
2	Transaction amount
3	Transaction status
4	Transaction currency
5	Payment type, e.g. Visa, Mastercard, Alipay, WeChat pay
6	Card type (such as debit, credit)
7	Mobile payment indicator

Table 3.16 Original variables of the Internet payment data provider's dataset 2 (monthly observation)
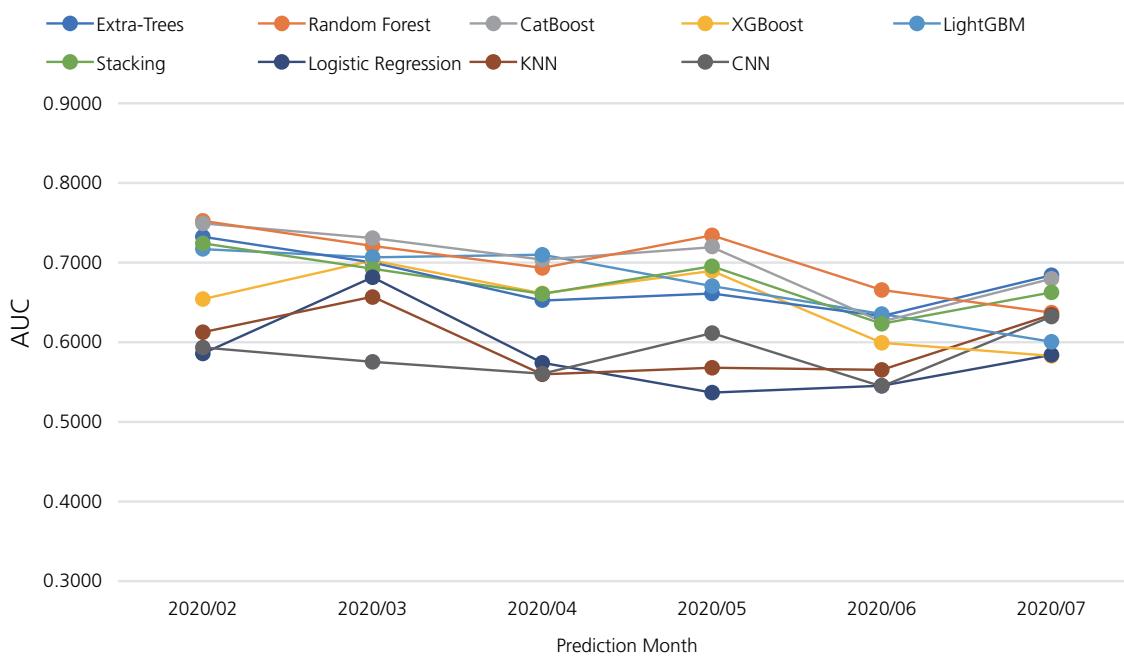
Original Variable	Description
1	Unique ID of MSME
2	Service charge payment due date (to resemble the default information)
3	Late service payment indicator

Table 3.17 Derived variables of the Internet payment data provider's dataset

Derived Variable	Description	Number of Variables
1	Time window related variables, such as 14, 30, 60 days.	30
2	Statistical data of monthly transaction amount and number per transaction status, transaction currency, payment type, card type, and mobile payment indicator, such as transaction amounts, occurring numbers, max, min, mean, etc.	128

*Development of the pre-screening model by the Internet payment data provider*

Figure 3.10 Performance of the pre-screening model by the Internet payment data provider

Alternative Credit Scoring of Micro-, Small and Medium-sized Enterprises

The pre-screening models were developed using the monthly transaction data of the internet payment data provider from May 2019 to June 2020. Figure 3.10 shows the AUC scores of the nine selected machine learning algorithms for predicting the occurrence of late payment (i.e. a resemblance of default), based on the developed pre-screen models. Each line presents the results of a different machine learning algorithm. The figure shows that the accuracy of all the models dropped gradually for the whole period from February 2020 to June 2020. The top five machine learning algorithms were Random Forest, CatBoost, Extra-Trees, Stacking, and LightGBM.

2.2.5 Insights gained from the POC

The POC described in this section demonstrates the technical feasibility of the Retail Alternative Credit Scoring (RACS) framework, consisting of the transactional cashflow and non-cashflow models for MSMEs in the retail industry.

The following specific insights arise from the experiments made by the bank:

- Banks can use MSMEs' bank account information to develop transactional cashflow models for the credit scoring of MSMEs.

- The transactional cashflow models developed by banks can achieve desirable results for short-term monthly predictions. For credit scoring using the developed transactional cashflow models, the banks can ask loan applicants to submit their monthly bank statements online, or ask for the monthly transactional data of the loan applicants' cash inflows and outflows to be sent to them from third-party data providers through API.

- Banks can use the transactional cashflow models to achieve continuous monitoring of the creditworthiness of MSMEs by using bank statement information or transactional data from third-party data providers.

The following specific insights arise from the experiments made by the POS payment and Internet payment data providers:

- Although transactional non-cashflow models could not be developed in the POC because both the POS payment and Internet payment data providers did not have MSME loan default information, effective pre-screening models could still be developed based on other MSME transactional data.

- Pre-screening models are helpful for making short-term predictions about problematic financial situations (based on late service payments, which resemble loan defaults). The accuracy of the pre-screening models dropped gradually as the prediction time period got longer. Better prediction capability could be achieved if the model is kept up to date with incoming transactional data.

- The transactional data of the POS payment and Internet payment data providers can be used to assess the creditworthiness of MSMEs. Depending on the business type of MSMEs, the transactional data from the relevant data providers can be used by the banks for credit scoring. Better overall credit scoring is anticipated if the outflow of cash (expenses) of the MSMEs is also contributed by the corresponding data providers.

As there was no data-sharing among the participating organisations, the credit scoring capability could not be enhanced by combining the transactional data about individual MSMEs from different third-party data providers. To tackle such restrictions on data-sharing among the participating data providers due to data privacy concerns, future work could apply privacy-enhancing technologies such as federated learning[37][38].

---

37. WeBank AI Group. (2018). Federated learning white paper V1. 0. WeBank, Shenzhen, China, White Paper, 9.

38. Openmined (2020). Federated learning for credit scoring. Retrieved October 9, 2020, from https://blog.openmined.org/federated-credit-scoring/.

In summary, the results of the experiments of this POC indicate that banks should further explore the feasibility of deploying the proposed Industry-specific alternative credit scoring framework as described in Part Three, Section 2.1 of this paper. In short, through collaboration with relevant third-party data providers, banks can first deploy the transactional cashflow models and pre-screening models. They can then develop transactional non-cashflow models as more transactional non-cashflow data of MSMEs is collected over time.

Acknowledgements for contributions to this section:

Company	Contributions
Bank of China (Hong Kong) Limited	The collaborative work involved in conducting the POC described in this section
HKT Limited	The collaborative work involved in conducting the POC described in this section
AsiaPay	The collaborative work involved in conducting the POC described in this section
CRIF	Experience-sharing of the deployment of the cashflow model
TransUnion Limited	Discussion of the use of the cashflow model
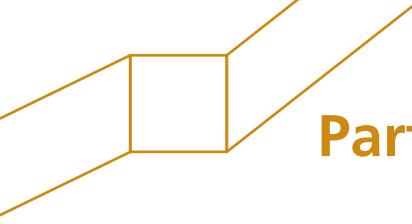
Part Four:

Roadmap ahead

This section lays out a suggested roadmap for the adoption of alternative credit scoring in Hong Kong. The roadmap includes the facilitation of data availability, continuous development of innovative models, and the setting up of a centralised data-sharing platform.

The lending industry faces both opportunities and challenges in applying fintech infrastructure to alternative credit scoring. To promote adoption by the banks, three critical issues need to be addressed. These are the availability of alternative data, continuous technological advances in the prediction models, and the need for a secure, centralised platform for data sharing and analysis. In tackling the problems related to these issues, there can sometimes appear to be more questions than answers. Nevertheless, the prospects for leveraging fintech so that banks can capture the MSME-financing market are promising.

Part Four

1 Facilitation of data availability

1.1 Continuous support by the government

A stable supply of alternative data is a key prerequisite for deploying alternative credit scoring, and continuous government support is critical to guarantee the availability of alternative data to banks. In supporting the use of fintech for credit risk management, the HKMA issued circulars on Credit Risk Management for Personal Lending Business on 9 May 2018 and 29 Aug 2019 which stated the principles for adopting new credit risk management techniques and practices enabled by fintech for personal lending business and small businesses. Clear guidance from the government regarding the usage and management of alternative data will effectively facilitate the sharing of alternative data between banks and non-banks.

- **New initiatives to promote the use of alternative data**

  New initiatives supported by the government could help to promote the use of alternative data, especially if they consider the needs of data security and data privacy vis-a-vis the benefits of fintech innovation. For example, specific uses of alternative data for credit scoring could be allowed under well-defined conditions.

- **Risk-based principles and guidelines in managing alternative data**

  Risk-based principles and guidelines relating to banks' management of alternative data for alternative credit scoring could help banks to ensure compliance in the following areas:

  o   Proper collection, processing and storage of alternative data from third-party data providers;

  o   The obtaining of legitimate consent from MSMEs by banks regarding the authorisation of data usage;

  o   Application of machine learning and AI in accordance with guidelines and principles issued, including data privacy regulations.

1.2 Infrastructure facilitation

APIs are one of the most important infrastructural components in helping to make alternative data available. To perform training and testing of the default prediction models, banks require a large volume of high-quality data on MSMEs covering a reasonably long period. APIs enable not only real-time data exchange between banks and non-banks but also data transfer between sharing parties across different geographical regions.

The OpenAPI Framework in Hong Kong was started in 2018 by the HKMA to facilitate the development and wider adoption of APIs by the banking sector. The framework is expected to provide detailed control measures for assessing customer data and transaction services. The OpenAPI framework could facilitate the use of alternative data for credit scoring if the following problems are tackled.

- **Access to Account** — Using the Access to Account (XS2A) requirement of PSD2 as a reference point, similar regulatory and technical supports could give banks authorised access to the bank accounts of MSMEs for specific purposes in a restricted way.

- **Consent management** — Allows MSMEs to have full control over how their alternative data are shared. Detailed rules need to be laid down to control the process of consent management.

- **Interoperability** — APIs need to be interoperable so that banks and third-party data providers can benefit from secure and controlled API services with minimal effort.

To enable scalable and efficient flows of data between banks and data providers, the HKMA plans to launch a new market development initiative called Commercial Data Interchange (CDI). CDI is a consent-based data sharing infrastructure with a standardised and secure technical interface. Banks and data providers can connect to the interoperable platform to share commercial data, and use this data to offer better products for MSME clients. The CDI infrastructure is designed to foster a vibrant and trusted data sharing ecosystem in the industry.

2 Continuous technical advances in modelling

A recent survey[39] published by the Hong Kong Institute for Monetary and Financial Research (HKIMR) reports that the banking industry has a positive attitude towards adopting AI, and that around 80% of survey respondents are planning to increase their investment in AI over the next five years. Alternative credit scoring is one area in which an investment in AI and machine learning will be most beneficial for banks, due to its ability to analyse a vast quantity of transactional records.

Key areas in the development of machine learning models have advanced rapidly in recent years. Such advances in technology will expedite the pickup rate among banks adopting AI and machine learning for alternative credit scoring. Some major areas that are being addressed or need to be addressed are as follows:

- **Model validation** — The question most frequently asked by users of machine learning algorithms is how to verify the results. Based on the limited data and resources on hand, banks need to test the validity of their prediction models and find out whether the results will achieve the credit scoring's objectives. Model validation offers a way to verify the accuracy and reliability of the selected models.

- **Model performance** — Accuracy and effectiveness are highly sought-after in the performance of prediction models, but they do not always come together. A complex model that is highly accurate may not be the best choice if it incurs exceptionally high costs in terms of the data required or the resources needed for execution. The prediction model design process should not only consider performance in terms of accuracy but also cost-effectiveness for deployment in real-life scenarios.

---

39   Hong Kong Institute for Monetary and Financial Research. (2020, August). Artificial Intelligence in Banking — The changing landscape in compliance and supervision. Hong Kong Academy of Finance. https://www.aof.org.hk/docs/default-source/hkimr/applied-research-report/airep.pdf
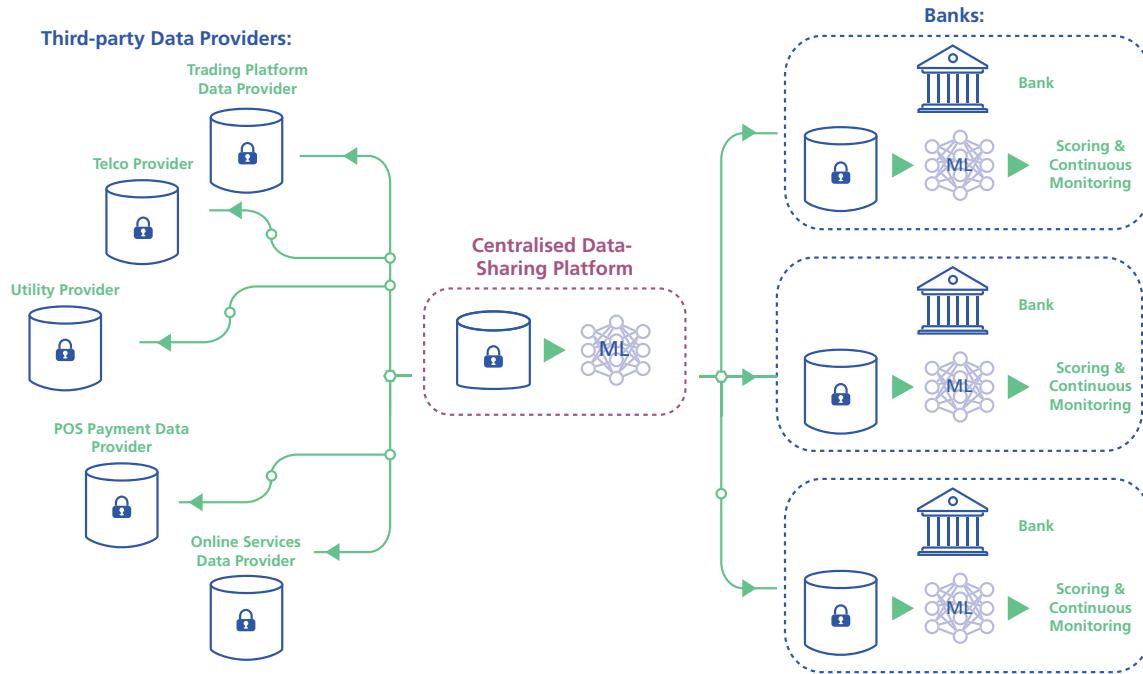
- **Privacy enhancement** — The validation and performance of prediction models are heavily reliant on the amount of alternative data that can be provided for model development by third-party data providers. However, many potential third-party data providers are restricted by data privacy regulations from sharing their clients' data with banks for credit analysis. Addressing this difficulty, privacy-enhancing technologies such as federated learning are rapidly maturing as ways to support model development without infringing on rules related to data privacy protection.

- **Model fairness** — The prevention of algorithmic bias and discrimination[40] is an important issue in deploying AI and machine learning in any fintech applications. Banks are expected to safeguard the fairness of the outcome of the credit scoring models. There is a need to put accountability mechanisms in place to ensure compliance with the relevant regulations and avoid algorithmic bias.

- **Model interpretability** — Interpretability is a weakness of machine learning algorithms in general because an explanation of the relative contributions of the specific independent variables to the outcome of the machine learning model is hard to describe or prove. Emerging techniques like interpretable machine learning[41] have become an increasingly important area of research, as they are helping to make the results of the model auditable and trustworthy.

---

40   Hong Kong Institute for Monetary and Financial Research. (2019, November). High-level Principles on Artificial Intelligence, https://www.hkma.gov.hk/media/eng/doc/key-information/guidelines-and-circular/2019/20191101e1.pdf

41   Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386.*

## 3   Centralised data-sharing platform for alternative credit scoring

Figure 4.1 Centralised data-sharing platform

This paper envisions a centralised data-sharing platform (Figure 4.1) that would enable data sharing between non-banks (third-party data providers) and banks, thus fostering the adoption of alternative credit scoring by the banking sector. By leveraging the platform, participating banks could supplement their own prediction models with extra insights into the creditworthiness of MSMEs, and the participating data providers could contribute their data about the creditworthiness of MSMEs without infringing data privacy regulations. In summary, with support from and endorsement by the government, a centralised data-sharing platform could offer the following positive elements for establishing a data sharing ecosystem.

- **Trusted environment** — The platform would create a safe environment for data sharing under an infrastructure supervised by the government, with emerging privacy-enhancing technologies such as federated learning used to tackle data privacy issues. Furthermore, comprehensive risk management guidelines could be enforced to safeguard the legitimate sharing and use of alternative data.

- **Controlled lab test environment for testing and fine-tuning** — Any newly created models require intensive training, testing and backtesting, and continuous calibration of the model details. A centralised data-sharing platform would help to create a controlled lab test environment where individual banks could develop their own models. The platform could also facilitate the sharing of useful datasets and software components so that all participating banks could benefit from shared technical know-how within the ecosystem.

- **Cost-effectiveness** — The centralised data-sharing platform could effectively reduce the potential investment of the banks during the initial phase of technology adoption. The platform would include pipelines for data collection, machine learning models and data privacy protection mechanisms, all of which are subject to modifications due to upgrades to technologies and changes in the relevant regulations.

- **Healthy ecosystem** — The platform would create a marketplace that would allow data and service providers and banks to connect their supply of and demand for alternative data and services. A reference standard for classifying the types of alternative data and categorising the feature names of the independent variables for machine learning models could drastically reduce the overheads involved in matching the alternative data fields to the independent variables for modelling. Furthermore, incentives could be granted to participating data and service providers to reward contributions of alternative data or services. An incentive scheme would facilitate healthy competition among data providers and value-added service providers.

Acknowledgements for contributions to this section:

The members of the Advisory Panel for the alternative credit scoring framework contributed significantly to the contents of this section. The Advisory Panel is split into working groups, and each working group went through three rounds of discussions in the period from April to July 2020.

Operational Considerations Of Alternative Credit Scoring

Authors:

Henry Cheng

Financial Services,

Director, PwC Hong Kong
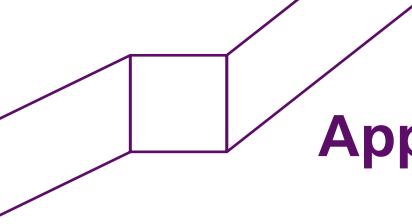
Gary Ng

Risk Assurance,

Partner, PwC Hong Kong

Section 1: Background

1.1 If history is any guide, we should expect a very long journey from the emergence of an innovative technology to its ultimate industry-wide adoption that benefits society. One of the most frequently shared stories about this phenomenon is that of the adoption of electricity. As factories had been designed based on the use of steam engines, for a very long time after the invention of electricity generators, manufacturers failed to see the need to switch to using electricity because it made no sense to use an electricity generator as though it were a steam engine! Things changed only when the manufacturers started to think in a fundamentally different manner, to the extent of being willing to revamp the design of factories, including the associated workflows, based on the ways in which electricity generators should be used. In the case of the deployment of alternative data for credit scoring, a long journey is unavoidable. However, we can definitely choose to embrace and enjoy the journey while doing our best to think fundamentally differently. We should aim for 'a change of factory design and associated workflows' rather than use alternative data (the electricity generator) as a substitute for conventional credit data (the steam engine).

1.2 This short paper was prepared in the spirit of accepting that fundamental changes in the way we handle data need to occur not only on the user level but also on a market-wide level. Moreover, in considering the various operational aspects that are crucial to the deployment of alternative credit scoring, we are convinced that all factors need to kick in and dovetail with each other in a holistic manner in order for fundamental changes to occur and reach a critical mass.

1.3 This paper discusses the obstacles we may expect to see in our journey from the emergence of new technologies that enable us to deal with new data types to the ultimate formation of an ecology in which all stakeholders benefit from the use of new data for, among other things, alternative credit scoring. The paper focuses on the operational aspects of these potential challenges, with a view to suggesting how different stakeholders can contribute to overcoming these challenges.

Appendix A

1.4 To facilitate the presentation of our analysis and thoughts, we use the following terms for situations where personal data are involved.

- Data subject — Individuals who own their data, e.g. personal, transactional and behavioural data.

- Data user — Entities which, in the course of their businesses, collect, process and use data from their customers (i.e. data subjects).

- Data processor — Entities that process personal data on behalf of a data user for activities such as processing data into formats that can be used in techniques such as machine learning and AI.

In this paper, we focus our discussion on handling personal data. For the handling of corporate data in alternative credit scoring, the risk considerations are similar, such as data confidentiality during data sharing and the need for consent management. As such, we can make reference to the key concepts discussed for handling personal data when it comes to sharing corporate data.
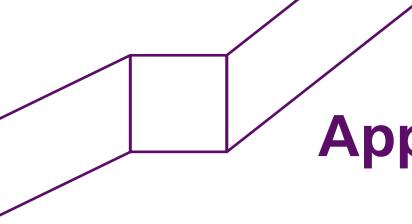
1.5 Undoubtedly, stakeholders will need to address a large number of operational issues throughout this long journey. This paper suggests a simple five-pillar framework that we hope can serve as a starting point to facilitate more industry discussion.

- Regulatory construct

- Government leadership

- Machine Learning and AI

- Consent management

- Ecology

Section 2: Regulatory construct

2.1 The first challenge that stakeholders face is the regulatory costs and risks of using alternative data, both being critical factors in determining whether and how alternative data can be used. Therefore, the first pillar is related to a fundamental question, i.e. how should the use of data be regulated?

2.2 This question is much trickier than it looks at first. One might think that a straightforward answer would be to appoint a Privacy Commissioner for Personal Data (PCPD). However, a PCPD may face several limitations.

- Scope — The definition of a breach of personal data privacy may be much narrower than the public would expect. Specifically, different people may have very different perceptions of (i) what constitutes personal data and (ii) what constitutes a breach. As a result, data users and data processors who do not properly collect, store, protect, process or use any data provided by data subjects could be subject to both legal risks (if the breach falls within the Personal Data (Privacy) Ordinance (PDPO) definition) and reputational risks (if the public perceives it as a breach).

- Resources — Regulators like the Hong Kong Monetary Authority (HKMA), the Securities and Futures Commission (SFC) and the Insurance Authority (IA) are empowered to govern the entry of regulated entities, set rules for them to follow and carry out supervisory work to ensure their compliance (and rectification in case of non-compliance). However, the PCPD operates in a rather different regime in which he or she oversees an activity (i.e. handling of personal data) in which anyone may get involved. As such, it may be difficult for the PCPD to come up with adequate resources to do the job in the manner of a typical regulator.

- Trust — Because breaches of personal data are not necessarily noticeable in a timely manner, some members of the public may have genuine concerns, which are in many cases well justified, as to (i) whether all data leakages/misuses can be identified in a timely manner, (ii) whether the identified leakages/misuses are disclosed in a timely manner to those affected and (iii) whether the identified leakages/misuses could have been prevented had stronger security measures been put in place.

- Rectification — Whereas some breaches may be to some degree rectifiable, misuse or leakage of personal data usually represents damage that cannot be undone (fines may not be able to address the concerns of those affected).

2.3 A further complication arises when it comes to the handling of personal data by regulated entities, e.g. banks, either in the process of conducting regulated activities or not. Several questions may need to be answered.

- Should the regulator or PCPD be primarily responsible for overseeing and enforcing the PDPO compliance of regulated entities?

- Should the regulator and/or PCPD be responsible for overseeing regulated entities' compliance with data privacy requirements imposed by other jurisdictions with extraterritorial applicability (e.g. the EU's General Data Protection Regulation (GDPR))? If so, how?

- Should the same standards be applied across different types of regulated entities (e.g. banks, telcos, airlines) and non-regulated entities (e.g. e-commerce operators)?

- If the same standards should be applied, how can this be achieved, as different businesses have different operating environments?

- Should entities which due to their business nature handle a lot of personal data but are not currently regulated by a specific industry regulator (e.g. e-commerce operators) be subject to a more robust regulatory regime?

2.4 The above limitations and complications make it very difficult, and most likely unjustifiable as a business case, for data users and data processors (e.g. banks, telcos, e-commerce operators) to work out a scheme with their customers for the collection, storage, processing and consumption of their data. To overcome this challenge, a two-pronged approach may be worth exploring.

- More transparent deliberations and collaborations among regulators, relevant government departments and PCPDs. The objective should not be to debate who is primarily responsible for ensuring that personal data are properly protected, but rather to contribute to personal data protection by sharing experiences gained from different industry settings.

- Development of a data exchange platform such that the handling and protection of data will be subject to the same standards, classifications and operating protocols. This data exchange platform can be achieved in two ways:

  i. a centralised platform, where there is a central owner to coordinate and manage the exchange of data; or

  ii. a decentralised platform, where no central owner exists but all parties involved in the data exchange and data modelling jointly establish a common set of standards and protocols.

- A centralised platform is the preferred model, as the central owner can regulate the stakeholders' onboarding and offboarding approaches, align technical standards and handle disputes between stakeholders, if any.

- It should be emphasised that public expectations of 'personal data protection' are always evolving with the emergence of new technologies, new data types, new incidents and new business models. Hence, a well-defined governance model is key to adapting to new challenges and consistently upholding a sufficiently high standard among all stakeholders so as to maintain the trust of the public.

2.5 Section 4 will further examine each type of data exchange platform and will look at the potential challenges in setting up a risk-free and balanced data exchange model. In any case, no matter which approach is taken, government leadership is of the utmost importance in achieving the intended results. This brings us to the second pillar below.
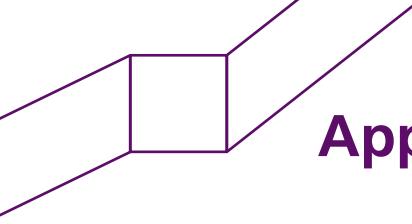
Section 3: Government leadership

3.1 It is obviously in the interest of the Hong Kong government to see the use of alternative data for, among other things, credit scoring of micro-, small and medium-sized enterprises (MSMEs). Not only does the success of these initiatives bring great strategic value to Hong Kong as an international financial centre, the adoption of financial technology, especially in the extraction of valuable credit information embedded in alternative data, is also a potentially key driving force behind financial inclusion. As such, the government has a considerable interest in investing its resources in this area and should also aim to use its leadership role to promote the use of alternative data. The key values of strong government leadership include the following.

- Tone from the top — With government endorsement, we believe it would be much easier for regulators and relevant government departments to collaborate on coming up with direction and guidance for this long journey. If a data exchange platform (centralised or decentralised) can be established as proposed in this paper, these are the key stakeholders who will take the lead in the development of the rules of the game.

- Synergy with government direction — As one of the biggest data users, the government has already taken a crucial step towards demonstrating an effort to balance the use of data (as a public good) and protection of privacy (e.g. DATA. GOV.HK). We believe that synergy can be achieved if the government is actively involved in the governance body overseeing the data exchange platform proposed above, and brings its experience with data management and plans for the use of public sector information in Hong Kong to the discussion.

- Public education — At present, some of the challenges faced by data users, data subjects and data processors stem from the perceptions of the members of the public. The government is in the best position to educate the public to understand how their data will be protected and assure them that they can retain control over how their data will be used, that there is a journey with occasional lessons learned and that we all ultimately benefit from putting alternative data to good use.

- Cross-border collaboration — In the context of regional cooperation that may involve the collection, processing and use of cross-border data, the government is in a position to lead or to endorse efforts in the discussion of cross-border arrangements and harmonisation of respective rules and regulations, particularly with respect to data privacy.

3.2 There are many ways the Hong Kong government can perform its leadership role. In any case, adaptation and flexibility will be important, as industry needs for leadership may vary significantly, depending on the development of technology, regulatory construct and public sentiment prevailing at the time. This paper does not aim to suggest specific forms of government leadership. Nevertheless, in Section 6, we briefly discuss the potential role of the government if a data exchange platform is established, as suggested in paragraph 2.4 above.

Section 4: Machine learning and AI

4.1 At the risk of oversimplification, this paper argues that the use of alternative data has the potential to be a game-changer. Alternative data contain useful information about the situation of data subjects; compared to conventional data, alternative data can be collected, analysed and used to trigger actions in a much more timely and frequent manner and, if properly designed, probably at a much lower cost.

4.2 Nevertheless, there has yet to be a treasure hunt to develop models that use alternative data for credit scoring. Machine learning and AI techniques seem to offer the best chance by far for the financial industry to crack the code. The following discussion considers two types of machine learning: traditional machine learning and federated learning. Under a traditional machine learning model, one central data processor collects data from various financial institutions and is fully responsible for training and developing the credit scoring model. This methodology, however, is typically associated with increased risk due to the cross-sharing of customer data across parties. Federated learning, in contrast, offers a safer alternative. Under this learning model, each data user trains a portion of the credit scoring model separately without cross-entity data exchange of raw data and contributes the intermediate results derived, resulting in a final machine learning model that reflects the combined efforts of all of the contributors.

4.3 The key to making a credit scoring model work using either machine learning method is a strong governance model, which we argue should take into consideration all of the factors outlined below.

- **Ownership** — When traditional machine learning is adopted, the owner of the trained model is clear. In the case of federated learning, however, ownership of the model is much less evident, creating the need for a strong governance structure to ensure a clear decision-making and issue resolution process.

- **Accountability** — Closely related to the previous point is the question of accountability. In traditional machine learning, accountability undoubtedly sits with the model owner. Under the federated approach, it is unclear which party will ultimately be accountable for the results produced by the credit scoring model. For example, if the model is found to deliver erroneous results that result in a negative impact on the customer, who will be held accountable, given that there is no central owner?

- **Bias** — A vast range of research has been conducted on the inherent biases present in products of machine learning arising from the types of training data fed to the models. As humans, prejudices are inevitable, but there are certain types of biases that may lead to decisions and actions that are discriminatory against specific groups of people (e.g. along racial or gender lines). If federated learning is to be adopted, a robust governance model must be established, and it should consider how to hold the parties responsible accountable if such a situation arises. Effective governance should also include regular assessment of the results delivered by the model so as to minimise the risk of such issues.
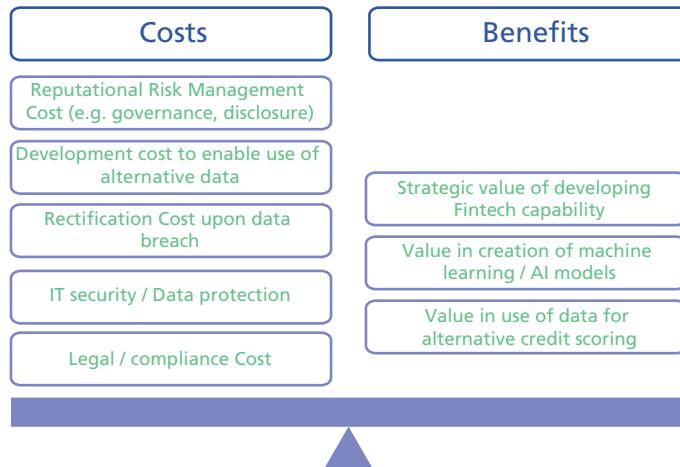
- **Data authenticity** — Building a credit scoring model requires the participation of a range of financial institutions, but how can the legitimacy of all of the data contributed be ensured? This is a particularly significant consideration for federated learning models, in which only the derived results, and not the raw data, are shared across entities. The illegitimacy of one single set of data will impact the authenticity of the entire model; it is, thus, crucial to define a sound legal and governance framework to prevent this from happening and hold dishonest parties accountable to the extent necessary. An additional consideration may be to establish a mechanism to detect and monitor such issues at each stage of credit model development, potentially through an independent assessment process.

- **Incentivisation** — In a data exchange platform where there is a clear central owner, the platform owner and the model owner will provide compensation to data processors and data users contributing data, thus incentivising them to participate in the venture. The incentive structure is not as clear-cut in a decentralised model; without a central owner, who is to offer compensation, and how will the compensation structure be determined? Moreover, considering the likely variance in the amount of data provided by each party (e.g. a global bank vs a growing payment platform), will the compensation structure take data size into account so as to ensure fairness across contributing members?

- **Data protection** — Last but not least comes the question of how machine learning can be adopted in a way that best protects customer privacy — a point briefly touched on in Section 4.2. The discussion here focuses on two areas.

  1) Privacy by design — The risks associated with data sharing are much larger when traditional machine learning is adopted, as opposed to federated learning. Whereas data need to be shared with the central owner by all entities in traditional machine learning, no sharing of raw data is required in a federated approach. At the point where information is shared between entities, it no longer contains any personally identifiable information (PII), thus posing no significant data privacy risk to data owners or data users. Furthermore, as the data are not held by a single party, data breaches can impact only one portion of the data at most, rather than the entire set of data. The heightened security and privacy associated with federated learning likely mean that more banks and organisations will be willing to participate as contributing parties.

2) Potential for reverse engineering — This is not to say, however, that federated learning is completely risk-free. Although the data may no longer contain PII, if only a very small sample size is used, it may be less difficult to connect an individual's transactional patterns with their identity. For example, if an individual holds accounts in two banks, their spending patterns and traits may be similar across both banks, making it possible to use reverse engineering techniques to identify individuals, even from the derived data. Given this possibility, institutions should define some preventative measures, for instance, setting thresholds for sample sizes for model training, removing any unique identifiers from their data or even exploring the use of synthetic data in place of real customer data.

4.4 Following from the above, a key issue to consider when it comes to machine learning is how to strike a balance between promoting machine learning on alternative data and protecting data privacy. The following paragraphs propose the adoption of a risk-based approach to guide the balancing act.

4.5 Broadly speaking, machine learning and AI techniques can be applied to alternative data under the constraints imposed by data privacy regulations through two directions.

- **Technological innovation** — As discussed above, one major challenge in using traditional machine learning is the risk of security breaches and data misuse resulting from having one centralised pool of data. In this respect, federated learning is a favourable alternative given the decentralisation of data; however, as has been explored, there are a number of factors necessary to ensure its success. These factors must be fleshed out and agreed on by all participating institutions to create a sound governance structure to oversee the implementation.

- **Process innovation** — There could be circumstances where technological innovation needs to be supplemented by process innovation, bearing in mind that the former is like the electricity generator, whereas the latter is like redesigning the factory. In this paper, we propose to adopt a risk-based approach to guiding process innovation.

4.6 Why risk-based? Although technology can lower the risk of data breaches and data misuse, 'zero risk' is impossible in the real world. In deciding whether to invest in using alternative data, potential data users and data processors must assess the business case. This includes (i) identification of the right risk management techniques that can mitigate the risk of data privacy breaches to a tolerable level, (ii) assessment of the initial and maintenance costs of putting in place such techniques and rectifying costs in case of breaches and (iii) evaluation of whether there is a business case, given the expected costs.

4.7 The challenge is a relatively uncertain element in the equation, i.e. what is the risk tolerance level? The severity of breaches that can be tolerated by regulators and the public is crucial to determining the 'degree' of data protection to be put in place by data users and data processors. Admittedly, it would be rather difficult for regulators and the public to explicitly accept anything other than a zero-tolerance approach, despite the fact that data breaches continue to happen occasionally. However, there can hardly be a business case if either the regulators or the public will not tolerate any breaches.

4.8 To resolve this apparent dilemma, analytics attempting to rank the level of risks (or severity of breaches) associated with different scenarios may be a good starting point. We believe that collaborative discussions and analytics covering a full range of risk scenarios with a view to ranking the relative tolerance level of these scenarios by regulators and the public will be very helpful. At present, some potential data users and data processors may perceive that the risk management cost is so high it outweighs the benefits of using alternative credit data. It is to be hoped that, with a more precise and certain understanding of the nature and severity of risks under different scenarios, data users and data processors can come up with cost-effective and proportionate risk-mitigation techniques, rendering investment in the use of alternative data justifiable.

Appendix A

Figure A.1 Costs and Benefits of data-sharing

4.9 Put another way, without sufficient granularity and transparency in terms of the risk implications of different data-breach scenarios, it would be difficult to identify the most practical techniques (e.g. through technological development, process management, or public education) to address the risks. In an effort to invite more discussion, we suggest below some dimensions that may be used in assessing the severity of the data privacy concerns of an incident.

- Data type subject to breach in the incident, e.g.

  o Personal data as defined under PDPO/GDPR

  o Transactional data (personal)

  o Transactional data (anonymous)

  o Behavioural data (personal)

  o Behavioural data (anonymous)

Alternative Credit Scoring of Micro-, Small and Medium-sized Enterprises

- Nature of the incident, e.g.

  o Repeated breaches of PDPO/GDPR (or no solution to prevent further breaches)

  o Isolated breach of PDPO/GDPR, with a solution to prevent further breaches

  o Breach of self-imposed standards

  o Potential breach (e.g. ambiguity over coverage of customer consent)

  o Alleged breach (e.g. media report alleging a breach, which may or may not be substantiated)

  o No breach, but does not meet public expectations

- Root cause of the incident, e.g.

  o Active and intentional, e.g. intentional misuse of data without consent

  o Active but unintentional, e.g. sending personal data to unintended recipients

  o Passive but due to poor controls, e.g. poor security design, inadvertently letting unauthorised persons gain access

  o Passive despite good controls, e.g. data hacked despite up-to-standard IT security arrangements

4.10 Let's use 'Data Type' as an example for the purpose of illustration. It can be safely assumed that most people would be much more concerned about sharing their transactional or behavioural data with personal tagging, compared to sharing them on an anonymous basis. With federated learning, only the derived, processed data is provided to other collaborators, such that each institution retains control of its data without disclosing any personally sensitive information. As applied to the case of credit scoring using alternative data, federated learning has taken into consideration 'privacy by design', maintaining data anonymity from the point where data is shared onwards externally. However, given that much personal data is still maintained within individual institutions, institutions must clearly outline the purpose of data collection and obtain explicit consent from the data subjects to avoid reputational risk. Furthermore, when conducting credit scoring for an individual applying for a loan, the data user should similarly obtain consent specifically for the collection and use of the individual's personal data. A practical and hopefully low-cost solution is for either the data exchange platform or loan portal to be designed such that data subjects can provide consent directly to data users and data processors.

Section 5: Data subject consent

5.1 As discussed in paragraph 4.10 above, an enhanced design that allows clear, flexible and granular consent by data subjects could be a cost-effective way to mitigate not only the legal but, more importantly, the reputational risks associated with the use of alternative data. This section explores some desirable features of consent arrangements.

5.2 The current practice of data subject consent is mostly bilateral between the data subject and the data user. In many cases, the data subject does not really have a choice if they need to obtain services from the data user. It can be observed that the current arrangement (Table A.1) has a few shortcomings, some of which are due to the following perceptions.

- Lack of choice — Data subjects feel that they have no choice if they are to use the service provided by data users.

- Too legalistic — The relevant clauses are drafted in legalistic terms embedded in a long terms and conditions section.

- Not transparent — Although, understandably, data users typically have to include terms seeking consent to pass data to a third party due to outsourcing arrangements, data subjects could be concerned that there is no transparency as to who the third parties are.

- Not flexible — Consent is normally given at the beginning of a business relationship. Data subjects may find it inconvenient if they want to review the terms or change their decisions.

- Unfair — Some data subjects believe that data users will use their data to generate revenues, but they are not fairly compensated or rewarded.

Table A.1 Current arrangement of consent management

Reference No	Process	Consent option given to data subjects (for illustrative purposes only)
1	Purpose of use	• For credit scoring only (Refer to Ref. 2) • For use in analytics and automated decision-making (e.g. behavioural profiling), including machine learning and federated learning (Refer to Ref. 4) • For receiving marketing information from the data user
2	Use of alternative data for credit scoring	• Allow the use of alternative data for their credit scoring • Do not allow the use of alternative data for credit scoring (Refer to Ref. 3) • Withdraw from the use of alternative data for credit scoring
3	Data elements to share for alternative credit scoring	Examples of alternative data to be shared with the data users (based on selection by data subject's selection). • Cashflow information (credits, debits, balances) • Credit card transactions • Loan payment information • Rent payment history • Payment record of telecommunication services • Property records • Social media rating/social sentiment

Appendix A

Reference No	Process	Consent option given to data subjects (for illustrative purposes only)
4	Location of data processing for the case of machine learning	• Data to be run on data subject's device only • Data to be run on data user's device only • Data to be run on systems of data users/data processors or on data exchange platform
5	Retention of data	• Data to be removed after credit scoring • Data to be retained for a certain agreed period for further use

5.3 A data exchange platform should be able to provide a standardised and user-friendly interface for data subjects to give, review and update their consent. Subject to acceptance by the public and the availability of technology, consent can be set granularly such that data subjects can decide how they want to allow their data to be used and for what purpose. The suggested structure of the data exchange platform demonstrates a more sophisticated way to obtain consent from data subjects, which is more advanced and stringent than the current common practices in obtaining consent. Below is an illustrative structure of a consent arrangement, which data subjects may choose for each data type suggested in paragraph 4.8 above.

5.4 The illustrative structure outlined in paragraph 5.3 above can address some of the shortcomings mentioned in paragraph 5.2. Sufficient granularity will give data owners a choice. A standardised interface can, to some extent, avoid things getting too legalistic. An online portal enables data subjects to change their decisions flexibly. In respect of transparency, ideally, it would be useful for data users to maintain updated records of third parties (e.g. due to outsourcing arrangements) that have processed or kept data obtained from them. Fairness is a rather philosophical point which is touched on briefly in Section 6.

5.5 Admittedly, the suggestions above could be too complicated for some data subjects. Nevertheless, one should not just look at the complexity as a standalone suggestion, but rather as an alternative to the current situation, as described in paragraph 5.2 above. Also, some measures could help data subjects on this journey. In particular, public education is very important, and government leadership is crucial. Data subjects need to be constantly reminded of the importance for them of understanding (i) why they need to provide data (i.e. otherwise data users will not be able to provide the level of services), (ii) that there are different types of data, of which some are sensitive and some are less so and (iii) that they are in control of their data and can decide granularly the extent to which their data can be used by others and in what ways. It should be stressed that public education is also about culture formation and encouraging the public to embrace new technology through a good understanding of its benefits. It is also about trust-building, giving the public confidence that the government, regulators and other stakeholders are working together as a team to ensure the soundness and fairness of the arrangement.

Section 6: Building an ecology

6.1 To conclude the thought process and suggestions shared in this paper, which is 'operationally' oriented, it is natural to touch on ecology, without which no operational arrangement can be sustainable. In the business world, one critical element for achieving ecology is the 'business case', as discussed in Section 4 above. As such, this paper attempts to explore whether there can be an institutional arrangement that may have a business case. In a sense, this is straightforward, as a business case is mainly about making the return worth the investment.

6.2 The earlier sections were mainly about different ways of keeping the investment costs reasonable and the associated risks tolerable. In this regard, we believe the two-pronged approach proposed in paragraph 2.4 above has its beauty. This approach is an effective way to ensure full integration of all regulatory requirements and expectations in respect of personal data protection into specific features, configurations and processing protocols of a credible data exchange platform. This is critical, as without a high degree of integration and operationalisation, data users and data processors can hardly overcome the challenges of (i) interpreting regulatory requirements, (ii) keeping a pulse on the expectations of the public and (iii) operationalising them effectively to the satisfaction of both the regulators and the public under a wide range of non-standardised arrangements. These challenges may render it not justifiable to invest in the use of alternative data, whose value will remain minimal until a critical mass of usage can be reached. Unfortunately, if major stakeholders are not willing to invest, the required critical mass can never be reached. To ensure public trust, it is recommended that a credible data exchange platform be sponsored by the government and operated by key financial infrastructure institutions in Hong Kong. This offers the best chance of representing the standards with which the regulators can be comfortable and members of the public can accept.

6.3 In addition, there is a need for governance of the data exchange and modelling platform. Whether the platform and modelling are operated in a centralised or decentralised manner, stakeholders will only trust and be willing to invest in or contribute to building up the ecosystem if they believe the end-to-end processes are well-governed. The interests of different stakeholders need to be balanced in the decision-making process, and common baseline criteria must be agreed on, from onboarding of new data users to data processor qualification and dispute management. To ensure the defined rules and standards are properly executed for such a collaboration, independent assessment of the data exchange platform and model training is required. This also enhances credibility and minimises the risks posed to data subjects. An independent assessor will likely be needed to perform regular checks on all parties involved, particularly on the data being used and shared by each participating institution and on the processes adopted within the platform. Such a process would ensure that no sensitive personal data are being shared across institutions and that the best interests of the data subjects are protected at all times.

6.4   The other side of the equation is much more uncertain, as putting alternative data to meaningful uses is, strictly speaking, still uncharted territory. A very positive sign is that many firms feel excited and optimistic about the value of putting alternative data to good use. Nevertheless, much innovation, creativity and perhaps luck will be needed to make their dreams come true. In its concluding remarks, this paper centres on two factors critical to success in creating an ecology for a data exchange platform: education and fairness.

6.5   Gaining the trust of customers is of paramount importance to the success of this endeavour. One likely challenge faced by data users and data processors is the public's fears regarding or scepticism towards data privacy, and this barrier is sure to be larger with federated learning, as it is still a relatively foreign concept. In this light, financial institutions should devote significant efforts to educating the public on the technology, its usage and what they are doing to best protect their customers' data privacy. Through a concerted effort, potentially with the government's involvement, the public should gradually come to understand how federated learning or other forms of machine learning work, and through this see the enormous possibilities and benefits alternative data can bring.

6.6   As data subjects, customers must also be able to see the value in being part of this endeavour. Currently, the reward for data subjects comes mainly in the form of the services to which they will have access (e.g. under a new credit scoring model, data subjects without a strong credit history may be more likely to receive a loan, given that the model considers a broader range of data). Apart from this, there is no established way for data users to offer any meaningful value (monetary or not) to data subjects providing consent to the use their data. A data exchange platform may be a good platform for trying out innovative ideas. One idea, albeit a rather wild one, is gamification using crypto-currencies. It is theoretically possible, although very challenging in terms of design, for potential data users to offer crypto-currencies to data subjects through a data exchange platform. As crypto-currencies do not have intrinsic value, this can be regarded as a points-scoring game at the initial stage. However, when with luck a healthily expanding ecology can be established, it is theoretically possible the crypto-currency will start to carry some intrinsic value. This is admittedly a very crude idea. It may serve to provoke more innovative ideas as to what types of arrangements can help develop an ecology for the use of alternative data in Hong Kong and the region.

Appendix B

Machine Learning Algorithms For Model Training And Default Prediction

Section 1: Ensemble learning techniques

An individual machine learning model may not be able to meet the demand for solving increasingly complicated problems, as the relationships between variables become more ambiguous and harder to detect. Some advanced machine learning techniques can achieve better predictions when faced with complicated problems. These techniques include ensemble learning and deep learning. Ensemble learning makes use of multiple machine learning algorithms to obtain a better predictive result, one that could not be obtained from any individual algorithm alone. Common ensemble learning techniques include bagging, boosting, and stacking.

The Bootstrap Aggregating (bagging) approach was designed to improve accuracy using an aggregated predictor[42]. Bootstrapping means randomly re-sampling several smaller datasets from the training dataset. Each small dataset trains a single weak classifier. Next, these single classifiers are aggregated by majority vote. The bootstrapping operation can reduce the impact of noise in the original dataset to further reduce the variance of the model, but cannot help with the bias of the model.

Boosting[43] in machine learning means converting a series of weak classifiers into strong classifiers. After training the previous classifier, a later classifier will increase the weight of the wrong sample in the previous training to pay extra attention to it in the next training. Boosting can significantly reduce bias but cannot help in reducing the variance of the model. The difference between bagging and boosting is that boosting uses the same dataset to train the weak classifiers, and its weak learners focus on adjusting the weight value of the misclassified data.

---

42.   Breiman, L. (1996). Bagging *predictors.Machine Learning,24*(2), 123–140.
43.   Schapire, R. E. (1990). The strength of weak *learnability.Machine Learning,5*(2), 197–227.
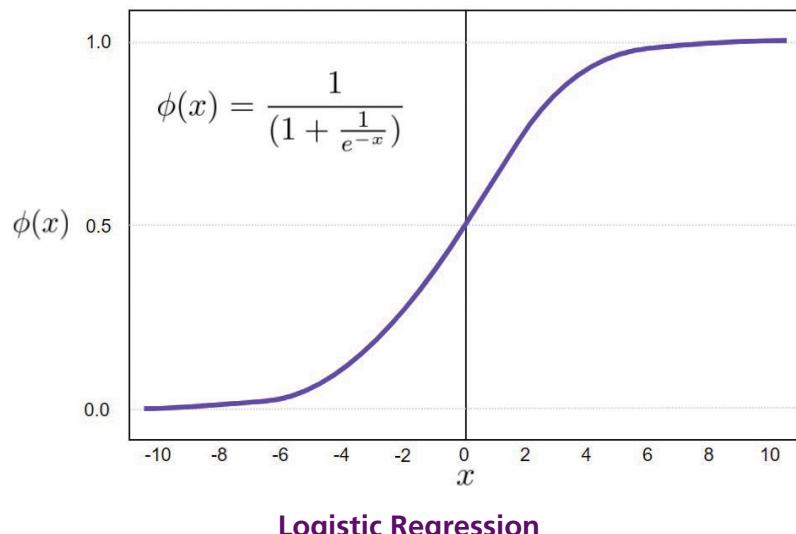
The Gradient Boosting Decision Tree (GBDT)[44] is one of the boosting ensemble learning techniques based on decision trees to achieve a better prediction by minimising a differential loss function in an iterative fashion. The advantage of the GBDT is that it deals with both regression and classification. The disadvantage is its inefficiency, because the trees are built sequentially and in an iterative fashion. Financial lenders usually adopt the GBDT technique in credit risk prediction for better performance. However, some have expressed concerns about using the model because its outcomes are not clearly explicable.

Stacking[45] is also one of the ensemble learning techniques used to obtain a better prediction by utilising several base (weak) learners. The meta learner can make a final prediction with the predictions of base learners as inputs. The disadvantage of stacking is that it is time-consuming compared with single-learner techniques.

Section 2: Common machine learning algorithms

2.1 Logistic Regression

$$\phi(x) = \frac{1}{(1 + \frac{1}{e^{-x}})}$$

Logistic Regression

---

44.   Breiman, L. (1997).Arcing the edge. Technical Report 486, Statistics Department, University of California at Berkeley.

45.   Wolpert, D. H. (1992). Stacked generalization.*Neural Networks,5*(2), 241–259.
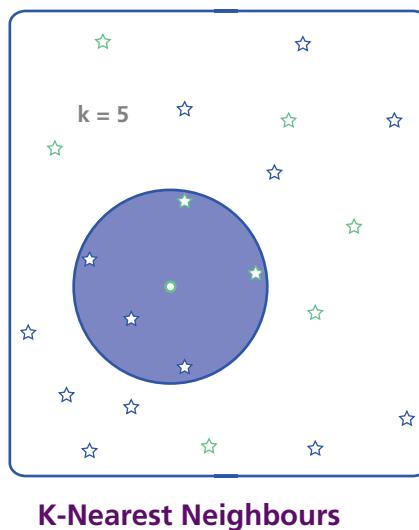
In statistics, regression is a fundamental technique for working out the relationship between a dependent variable and several independent variables. Linear regression[46] estimates the relationship with linear predictor functions. Logistic regression is a simple classification approach to estimate the relationship between a categorical dependent variable and several independent variables.

The logistic regression model[47] is commonly used by both mission-driven lenders and financial lenders for binary objective situations, typically appearing in the form of credit scoring, sales response models, and debt recovery models. It is intuitive, explicable, and faster than the other algorithms.

2.2 K-Nearest Neighbours (KNN)
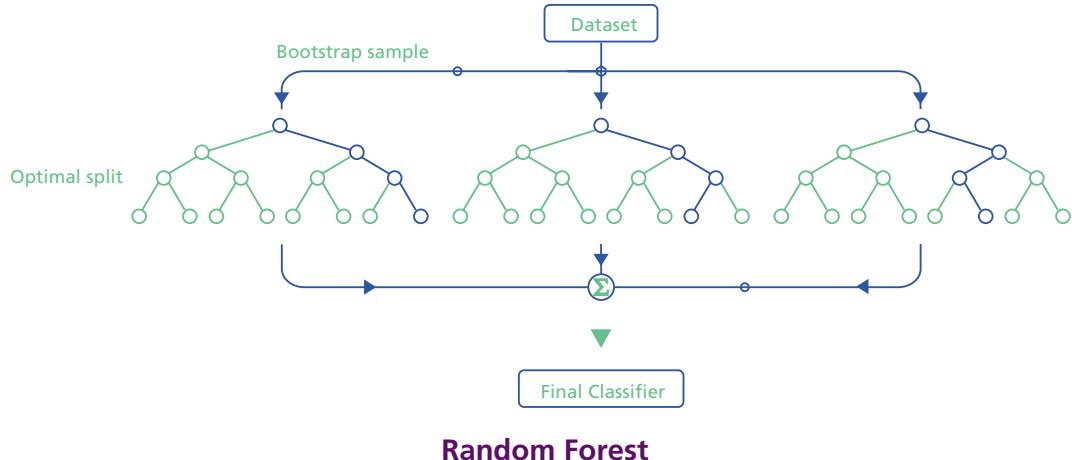
The Nearest Neighbours pattern classification[48] (KNN or K-NN) is used to identify the K number of the training data closest to the predicted target. The prediction result is based on the majority vote of its K neighbours (for classification) or the mean of the K neighbours' values (for regression).

Few examples of the practical application of KNN have been recorded. However, subject analysts have noted that some lenders have used KNN to improve the performance of their fraud detection and direct marketing activities.

K-Nearest Neighbours

46.   Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Harvard University Press.

47.   Cramer, J. S. (December 2002). The origins of Logistic Regression. Tinbergen Institute Working Paper, No. 2002-119/4.

48.   Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory,13(1)*, 21–27.
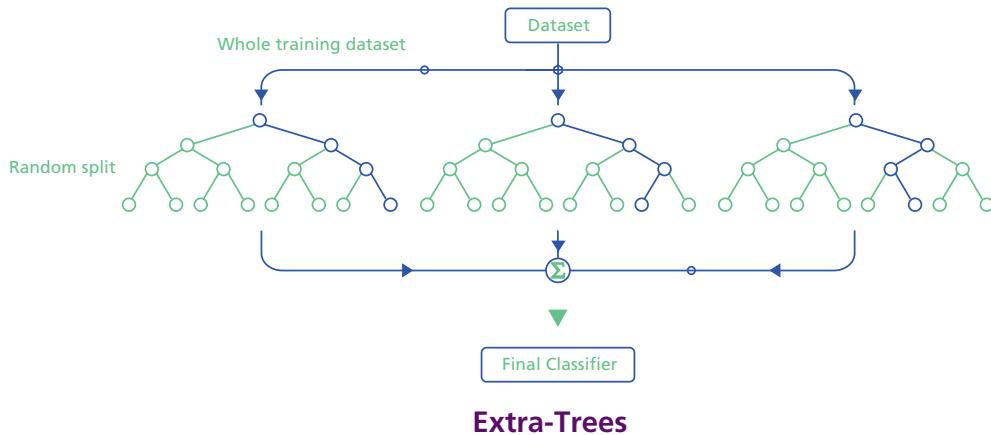
2.3 Random Forest

Random Forest

A Random Forest[49] adopts the bagging approach and a decision tree based method to deal with predictions for both regression and classification. The advances made by the introduction of tree-based methods were with respect to interpretation — in other words, the results of this machine learning model can be presented in an understandable way.

The Random Forest model bootstraps sub-trees from the training dataset to improve accuracy. Hence, a Random Forest is typically suitable for modelling a large number of predictors, such as in the quality assessment of scientific work[50].

2.4 Extra-Trees

Extra-Trees
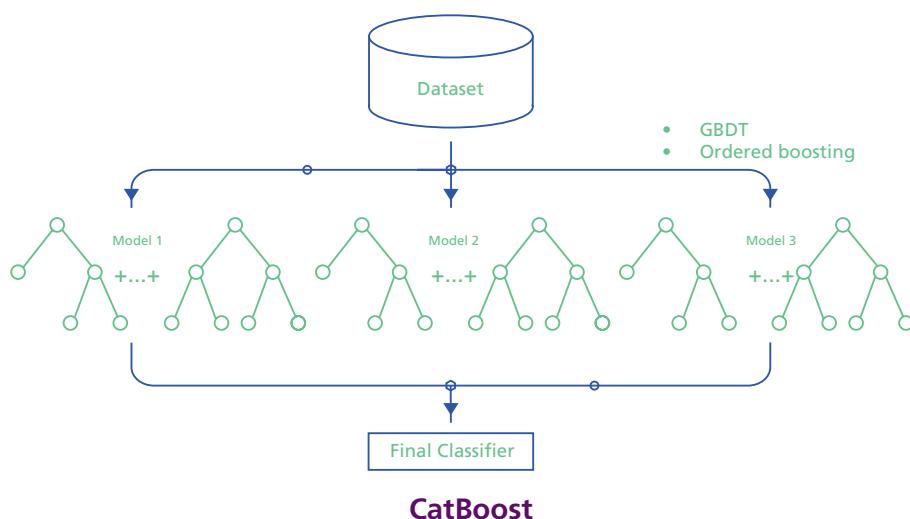
49.   Ho, T. K. (1995, August). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282). IEEE.

50.   James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). In *An Introduction to Statistical Learning: With Applications in R* (pp. 319–321). New York: Springer.

Extremely Randomised Trees (Extra-Trees) is a kind of decision tree learning. Unlike Random Forest, Extra-Trees can provide randomised choices of input variables and cut-points when splitting a tree node[51]. Besides, each tree in Extra-Trees is trained with the whole training set. However, the Random Forest method trains each individual model with its respective bootstrap sample. Higher randomisation levels of tree splitting can improve accuracy, which means lower variance.

2.5 CatBoost
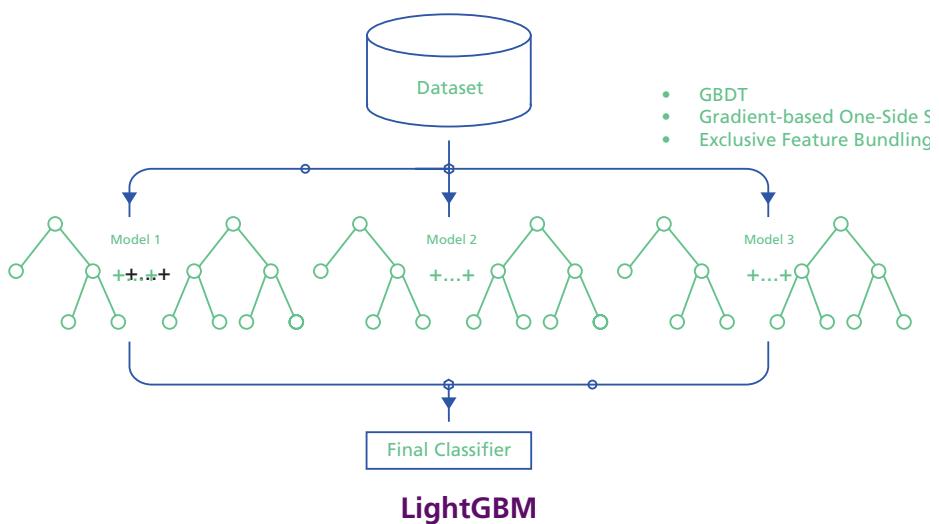
CatBoost

CatBoost[52,53] is an efficient gradient boosting machine learning algorithm (GBDT) that can use categorical variables directly and implements ordered boosting with categorical variables. Traditional gradient-boosting algorithms need to preprocess categorical data before building a decision tree via one-hot encoding, label encoding, hashing encoding, or target encoding techniques, which leads to large memory requirements, computational costs, and weak variables.

51.   Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63(1)*, 3-42.

52.   Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., CesaBianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 6638–6648. Curran Associates, Inc.

53.   Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support.arXiv preprint arXiv:1810.11363.

CatBoost successfully handles categorical variables during training rather than preprocessing during exploratory data analysis (EDA). It is one of the primary methods for data learning problems with heterogeneous variables, categorical data, and complex dependencies in Web search, recommendation systems, weather forecasting, medicine, industry, finance, sales prediction areas, and many others.

2.6 LightGBM

- GBDT
- Gradient-based One-Side S
- Exclusive Feature Bundling
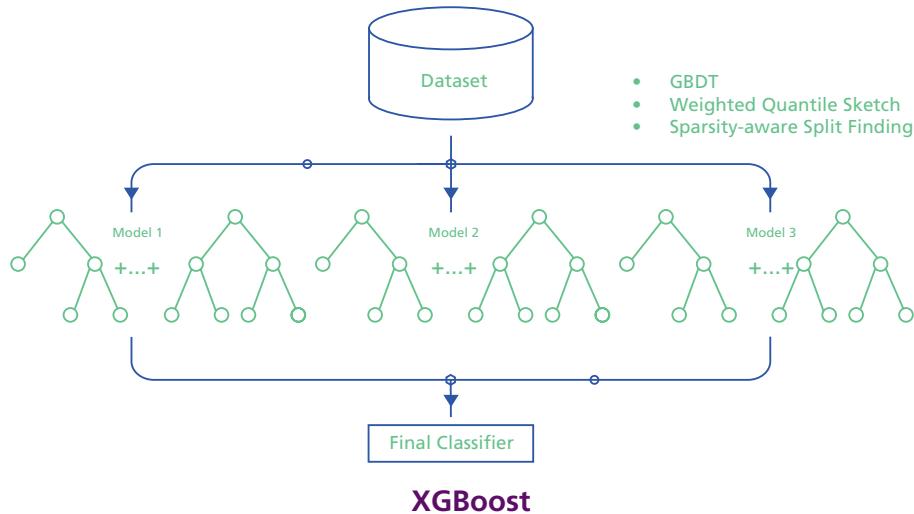
LightGBM

LightGBM is an improved GBDT that is designed to improve the efficiency and scalability of traditional gradient-boosting algorithms when the feature dimension is high and the dataset is extremely large[54]. It uses leaf-wise tree-based algorithms, whereas other GBDT algorithms work in a level-wise approach pattern. When growing on the same leaf in LightGBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm to get better accuracy. There are two novel techniques in LightGBM: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

By reducing both the number of data samples and variables, LightGBM speeds up the training process and achieves the same accuracy level as other GBDT algorithms. It can be used in many data science related fields, such as banking, insurance, manufacturing, and finance. For example, when making investment decisions, LightGBM can be used to forecast a company's revenues based on fundamental financial reports.

54.   Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146–3154).
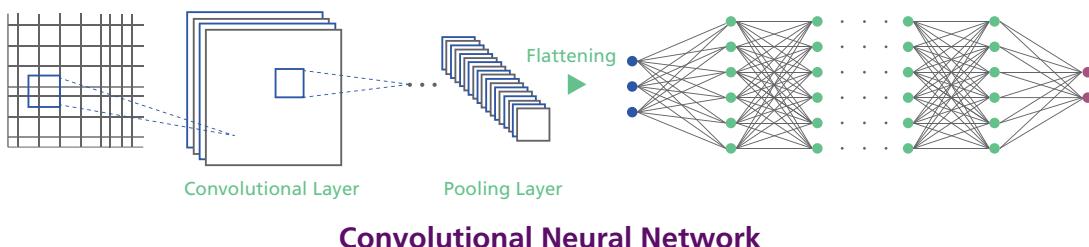
2.7 XGBoost

XGBoost

XGBoost stands for "Extreme Gradient Boosting", one of the ensemble learning approaches using gradient boosting. It is under the Gradient Boosting framework with optimisation and is designed to be highly efficient, flexible, and portable.
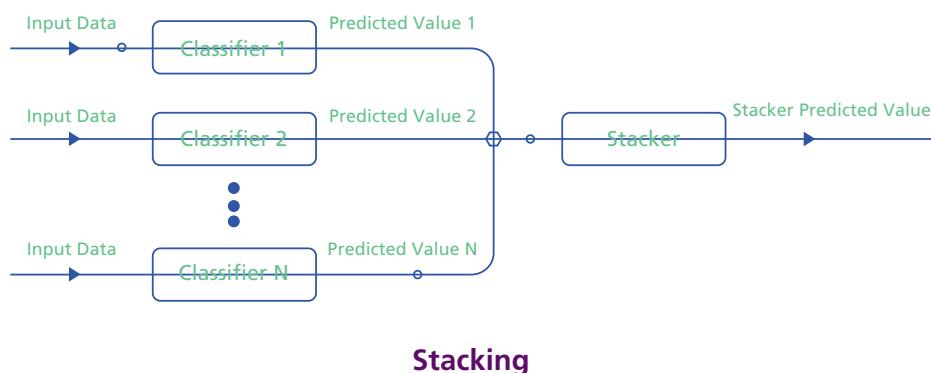
XGBoost provides parallel tree boosting (also known as GBDT, GBM), which solves data science problems in a highly effective and accurate way. XGBoost also provides a few advantages over traditional boosting algorithms, such as regularisation to reduce overfitting, parallel processing, high flexibility, handling missing value, a user-defined loss function, and built-in cross-validations. It is useful in data science problems in both research and industry.

2.8 Convolutional neural network (CNN)

Convolutional Neural Network

Alternative Credit Scoring of Micro-, Small and Medium-sized Enterprises
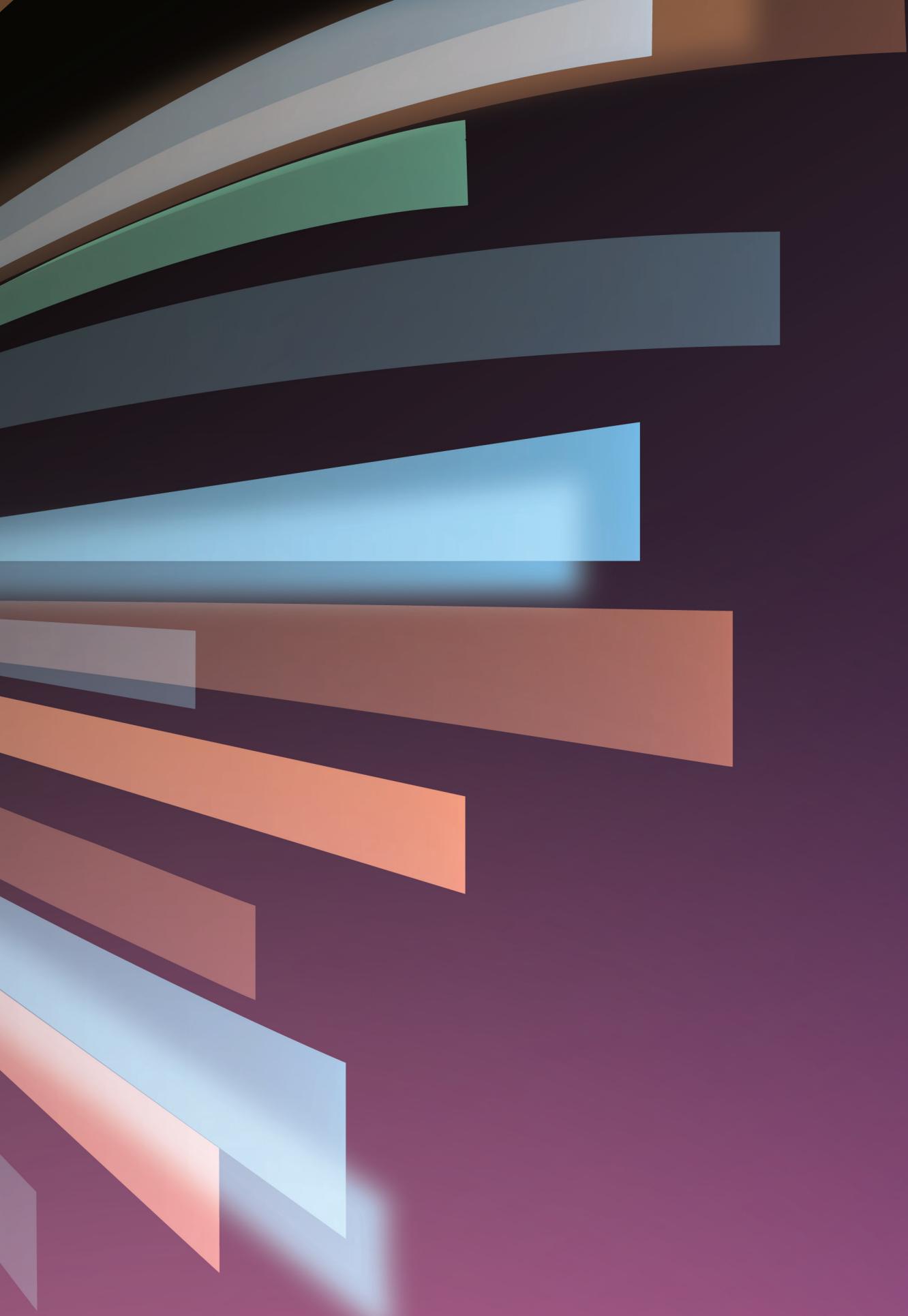
Deep learning is a machine learning technique that uses an artificial neural network (ANN). A convolutional neural network (CNN)[55,56] is a deep learning architecture. CNN uses convolution, detector, and pooling operations and consists of several layers, such as an input layer, multiple hidden layers, and an output layer.

2.9 Stacking

Stacking

Stacking is not an algorithm but an ensemble learning technique that combines multiple classification models via a stacker, i.e., meta-classifier. The first-level classifiers could be any of the models mentioned above. All of these models have their advantages and disadvantages, and the second-level stacker can try to learn advantages and discard disadvantages from these first-level classifiers to make a better prediction. Of course, stacking can use more levels to make a more precise model, but it will cost multiples of the time it takes to use a single classifier and the improvement in accuracy is quite limited. Hence, the stacking strategy is more common in data science competitions than in business practice.

55. Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology,160*(1), 106.

56. LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010, May). Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* (pp. 253–256). IEEE.