RISK MANAGEMENT MAGAZINE

Vol. 17, Issue 2
May – August 2022

EXCERPT

https://www.aifirm.it/rivista/progetto-editoriale/

Machine Learning for Credit risk: three successful Case Histories

Paolo Di Biasi, Rita Gnutti, Andrea Resti, Daniele Vergari

Machine Learning for Credit risk: three successful Case Histories

Paolo Di Biasi (Intesa Sanpaolo), Rita Gnutti (Intesa Sanpaolo), Andrea Resti (Università Bocconi and senior advisor CRIF), Daniele Vergari (CRIF[1])

Abstract

As the financial services landscape witnesses an unprecedented change, banks can use machine learning ("ML") to expand their databases through alternative sources providing unstructured and semi-structured information, such as transaction data and digital footprint data. However, ML algorithms also suffer from several potential shortcomings, as they may overfit sample data and prove unstable over time, they may quickly become obsolete and need re-estimation, and they may prove hard to interpret. This paper joins the debate on ML in banks by providing three case studies that highlight the benefits of machine learning, while showing how its drawbacks can be minimised: a rating model developed within the IRB framework, a challenger model used to validate a bank's main model for retail PDs, and an early warning system based on transaction data.

1. Foreword

The use of machine learning ("ML") models in banks has raised considerable interest and sparked a lively debate, among both scholars and practitioners. As the financial services landscape witnesses an unprecedented change (due, e.g., to the digitalisation of credit processes, open banking regulations, competition from non-bank players and more prescriptive regulations), banks can use ML to expand their databases through alternative sources providing unstructured and semi-structured information, such as transaction data and digital footprint data. Additionally, ML can manage multi-dimensional data by automatically selecting meaningful features and creating meta-variables that summarise the most relevant information, thereby improving the performance of internal scoring/rating systems. On the other hand, ML algorithms also suffer from several potential shortcomings: first, they may overfit sample data and prove unstable over time (meaning that they perform poorly when applied to new data); second, they may quickly become obsolete and need recalibration/re-estimation to keep attaining high performance standards; third, quality checks for unstructured data (e.g. natural language) may prove more complex to run than they are for structured data sources; finally, ML models may prove hard to interpret, meaning that – although they lead to overall correct results – the factors driving their outcomes may prove hard to pinpoint (or may vary sharply between individuals, making it hard for users to identify a consistent pattern).

Over the last 20 years, ML has become increasingly popular with both banks and supervisors. A bibliographic search for the keywords "machine learning" and "bank" (or "banking", or "supervision") finds 41 relevant articles and two book chapters published between 2000 and 2021 (Guerra and Castelli, 2021), with a significant acceleration after 2016 (see Figure 1)[2].

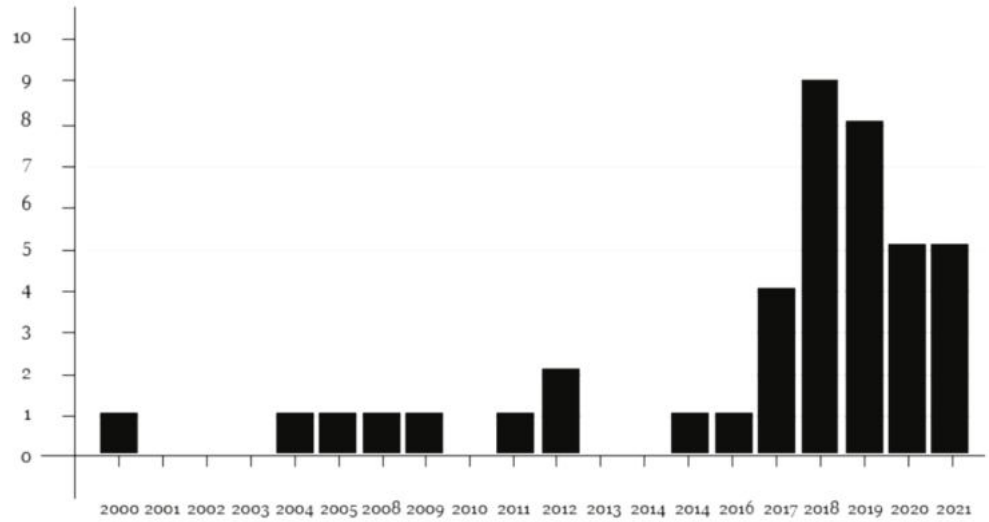

Figure 1 - Articles published in 2000-2021 referencing ML and banks/banking/supervision (source: Guerra and Castelli, 2021)

[1] This article draws on a paper (Di Biasi et al., 2022), written with Angelo Basile, Fiorella Bernabei, Cristina Caprara, Dario Cavarero, Mattia Marigliano, Roberta Ranaldi e Marco Vignolo (https://www.crif.it/ricerche-e-pubblicazioni/altre-risorse-e-ricerche/2022/marzo/l-applicazione-di-tecniche-di-ml-al-credit-risk-management-e-ai-modelli-irb/). We gratefully acknowledge comments from Simone Casellina (European Banking Authority), Francesco Cannata (Bank of Italy), as well as from various participants in a presentation to the EBA held on 17 March 2022.

[2] A similar survey, focusing on deep learning applications to banking and finance in 2014-2018, found 40 academic articles dealing mainly with stock market prediction and trading, while credit risk models only account for 12.5% of the published studies (Huang et al., 2020). Further surveys include (Leo et al., 2019) and (Rundo et al., 2019). The former looks at applications of ML in the management of banking risks (credit, market, operational and liquidity) and finds that ML usage doesn't appear commensurate with the importance of risk management and ML studies when considered separately; the latter looks at ML usage in the field of quantitative finance (including comparative studies about the effectiveness of ML-based systems), showing that innovative methods often outperform traditional approaches.

More recently, ML has been widely used by financial institutions worldwide. Out of 60 banks surveyed by the IIF in 2019[3], 25 were using ML models in a production setting (compared to 23 one year before), and another 27 were engaged in pilot projects (up from 12 in 2018). Credit scoring and early warning systems were by far the most common areas of application. As for 2020, there is evidence that investments in ML techniques by banks were not negatively impacted by Covid-19; indeed half of the UK-based institutions surveyed by the Bank of England (Bholat et al., 2020) expected an increase in the importance of ML and data science for future operations as a result of the pandemic.

Having implemented ML models in banks for several years, we would like to join the debate providing three case studies that highlight the benefits of machine learning, while showing how its drawbacks can be minimised. Table 1 provides a synopsis of the main characteristics of each project.

Table 1 – Synopsis of the three case studies discussed in this paper

Case study	*Benefits*	*Data and algorithms*	*Interpretability techniques*	*Main challenges*
A new IRB model aimed at retail SMEs	Provides customers with a full digital experience, focusing on high digitalisation standards and profiting from PSD2 and the GDPR. Deals with unprecedented conditions, e.g. due to Covid-19	Transaction data, credit cards, POS, and web sentiment. Decision trees, random forests and gradient boosting	Partial dependence plots, individual conditional expectation, LIME, and SHAP	Hiring new staff, investing in IT, reviewing regulations, deploying an ad-hoc validation framework
A challenger model to validate the IRB model for retail PDs	Performs initial validation to get supervisory clearance, and ongoing validation to promptly highlight any issues emerging from the bank's model	Innovative transaction data based on borrowers' current accounts. Decision trees with extreme gradient boosting ("XGB")	Feature importance analysis	Project timing and the choice of a trade-off between a full challenge and the need to ensure comparability
An early warning system based on transaction data	Ability to use transaction data to identify new information patterns	Individual transactions combined with pre-existing information. Random forests, XGB and neural networks	Feature importance analysis, SHAP, LIME and OptiLIME CRIF	Feature selection, model development and interpretability

In the remainder of this paper, each project will be described in detail: we start with two projects that relate to rating systems used for regulatory purposes, then we move to a "managerial" model focused on early warning systems.

2. Case 1: using ML for a new IRB model aimed at retail SMEs

The data available on SMEs (small and medium-sized enterprises) has recently experienced a sharp increase as new sources have emerged, providing value added to improve risk management models. Against this backdrop, a large bank decided to update its IRB model for the retail SME portfolio, comprising about 500,000 businesses with a turnover of up to €2.5 million and a credit exposure of less than €1 million. These are typically medium-sized limited liability companies, partnerships and sole traders/entrepreneurs. The bank launched a "Smart Lending" project, to provide retail SMEs with a full ("end-to-end") digital experience. Focusing on high digitalisation standards, the project aimed to profit from the evolution of the technological and regulatory framework (as PSD2 and GDPR, regulating data and customer protection, opened the door to the possibility of receiving new data for an all-round customer assessment). The rating model had to be available online and in real time, without easing risk management standards. ML algorithms provided a tool to improve the customer journey as well as rating performance, while also accounting for extraordinary events such as the Covid-19 pandemic.

An ML component can be introduced into credit risk models in two ways: by replacing previous models or by complementing them through new algorithms. In this case, ML was used to add new data sources to traditional ones, enhancing the breadth and accuracy of the pre-existing approaches. Accordingly, a new IRB model was developed that uses (see Figure 2):

- *traditional modules* to process information already used by the rating model, such as financial statements, external data (e.g., central credit registers like CRIF) and internal performance indicators;
- *modules using new data sources*, both internal and external to the bank, processed through either ML or traditional algorithms.

[3] See (Institute of International Finance, 2019).

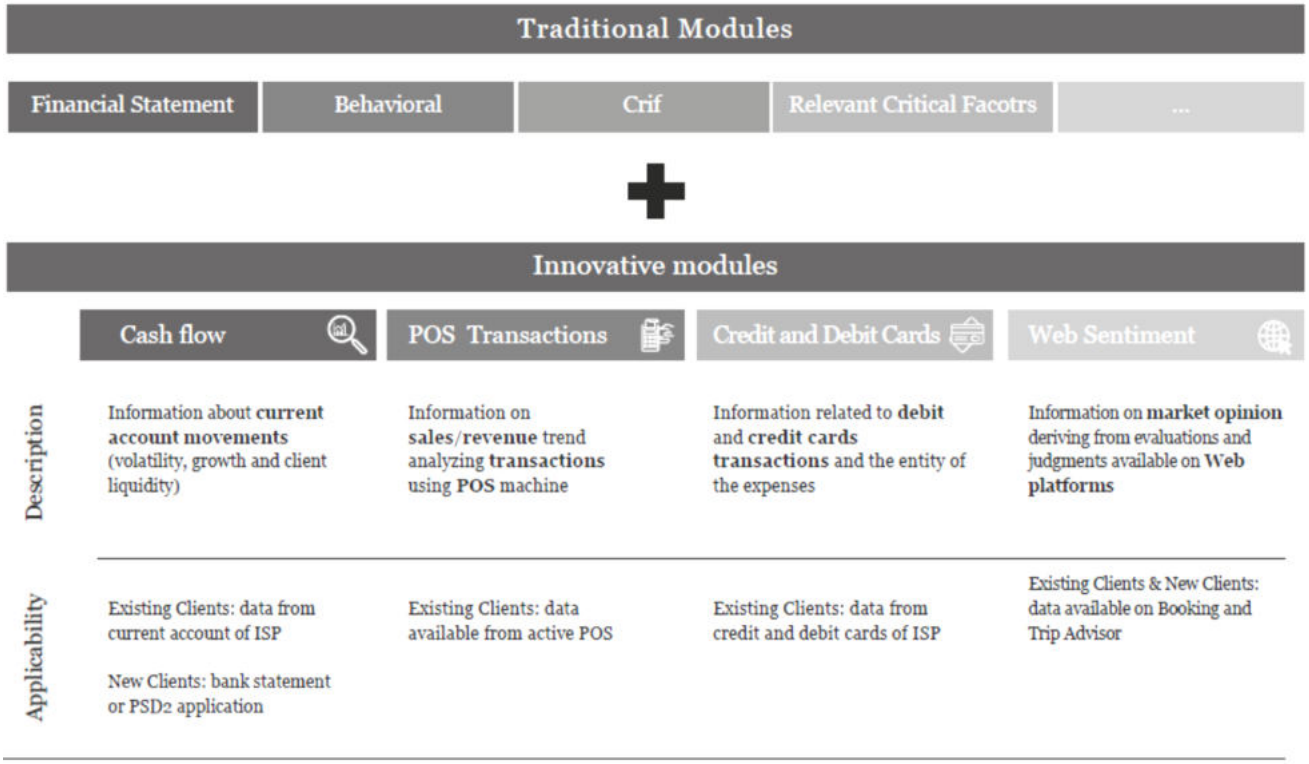

Figure 2 – Data sources of the new model for retail SMEs

New data sources include transaction data (from the bank's own current accounts and from other banks, available through PSD2), point-of-sale transaction data, payment card data, web sentiment data captured from web pages. Among these sources, transaction data is the most challenging, but also the best performing. It allows recent key information to be collected, and transforms the constraints imposed by PSD2 (forcing the bank to transfer information to third parties upon request) into an opportunity.

The algorithms selected for the ML component were[4]: decision trees, random forests and gradient boosting. Deep learning algorithms were not used, given their complexity: instead, a gradual approach was chosen, using algorithms that were more advanced than – while at the same time comparable to – logistic regression. As a benchmark to gauge their performance (and to improve interpretability) a traditional logistic regression was also run on the same input data.
Interpretability remained a key aspect throughout the estimation process. The main techniques deployed (besides traditional models used as a benchmark) were Partial Dependence Plots ("PDP"), Individual Conditional Expectation ("ICE"), LIME (Local Interpretable Model-agnostic Explanation, a technique that identifies the features that contribute most to an individual classification through a local approximation performed on slightly modified versions of the original observation[5]), and SHAP (SHapley Additive exPlanations, a relatively recent approach combining features from LIME and Shapley)[6].

The most effective methodology proved to be SHAP. It assigns a marginal contribution to each variable (feature) considering its possible interactions with other variables: for each combination of variables, the change in PD is measured, providing a basis to compute the relative weight of each feature. Figure 3 provides a sample SHAP plot, showing the range and impact of each variable by order of importance. Each point in the plot is a Shapley value (measured on the x axis) for a feature (as listed on the *y* axis, in decreasing order of importance) and an instance. The colours represent the value of a feature from low to high. Feature labels have been redacted for confidentiality reasons.

[4] See (European Banking Authority, 2020) for a brief discussion, and(Breeden, 2021) for a taxonomy of ML algorithms applied to credit risk.

[5] See (Ribeiro et al., 2016). An optimised version of LIME, tackling the instability problems that undermine its reliability, was proposed by (Visani et al., 2020): under this new approach ("OptiLIME CRIF"), stability is maximised for any chosen level of "adherence", i.e. similarity to the original ML model.

[6] Shapley values (a measure of how much each feature contributes to a prediction, based on a large number of comparisons between pairs of alternative feature sets); the Shapley value can "split" an individual prediction among all contributing features, providing a full explanation of why a given applicant has received a specific credit score. As noted in (Molnr, 2019), this can make it preferable in situations where the law provides customers with a "right to explanations". A thorough discussion of Shapley values is provided, e.g. in (Giudici and Raffinetti, 2021), which also introduces an extension of the original Shapley approach ("Shapley-Lorenz") based on Lorenz decompositions. LIME, Shapley and SHAP are known as *local* interpretation techniques, as they analyse how individual model predictions change when altering input data. They are mostly used to produce a visual representation that reflects the contribution of each feature (explanatory variable) to a single-point forecast, assigning a "weighted" importance to the characteristics that most affect the output generated by the model (e.g. default probability). Conversely, *global* interpretation techniques are aimed at understanding the relationship between each main feature (explanatory variable) of a model and its target variable in a more "traditional" way.

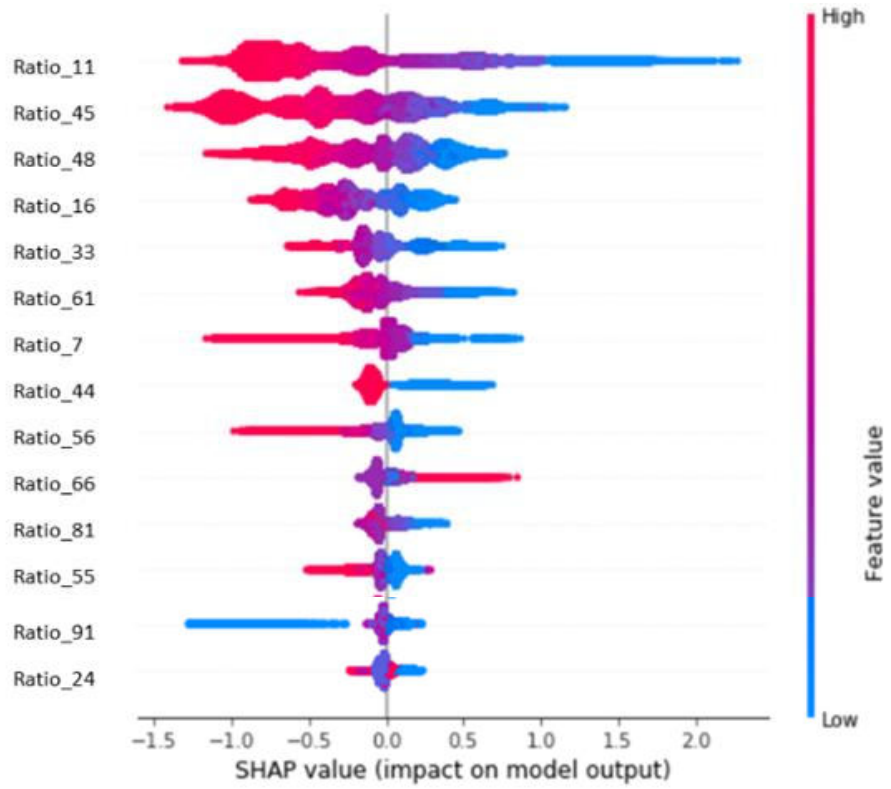

Figure 3 – Cash inflows for retail SMEs ("RS") estimated using transaction data

ML models showed the best performance (accuracy ratio) and a good out-of-sample stability. To assess overall performance, a benchmark model was created that used neither innovative sources, nor new algorithms. Keeping that model as the base case, marginal increases in the accuracy ratio were measured: new data sources processed through traditional methods led to a 5% increase; a further 5% increase emerged when the same sources were processed through ML. If ML had also been used to integrate the various modules in Figure 2, that would have led to a further 1% increase; that option, however, was not considered in the final model, as the improvements achieved in the previous steps were already satisfactory.

The new module using transaction data also proved very responsive in identifying changes in PD drivers following the Covid-19 outbreak. Indeed, the score based on transaction data remained stable in 2019, and quickly worsened during lockdown, capturing information that traditional default-forecasting models could not use in full. An example of its determinants is provided in Figure 4, showing cash inflows estimated from transaction data for different borrower groups (including industries that were deemed especially vulnerable to the drop in demand associated with lockdown regulations).

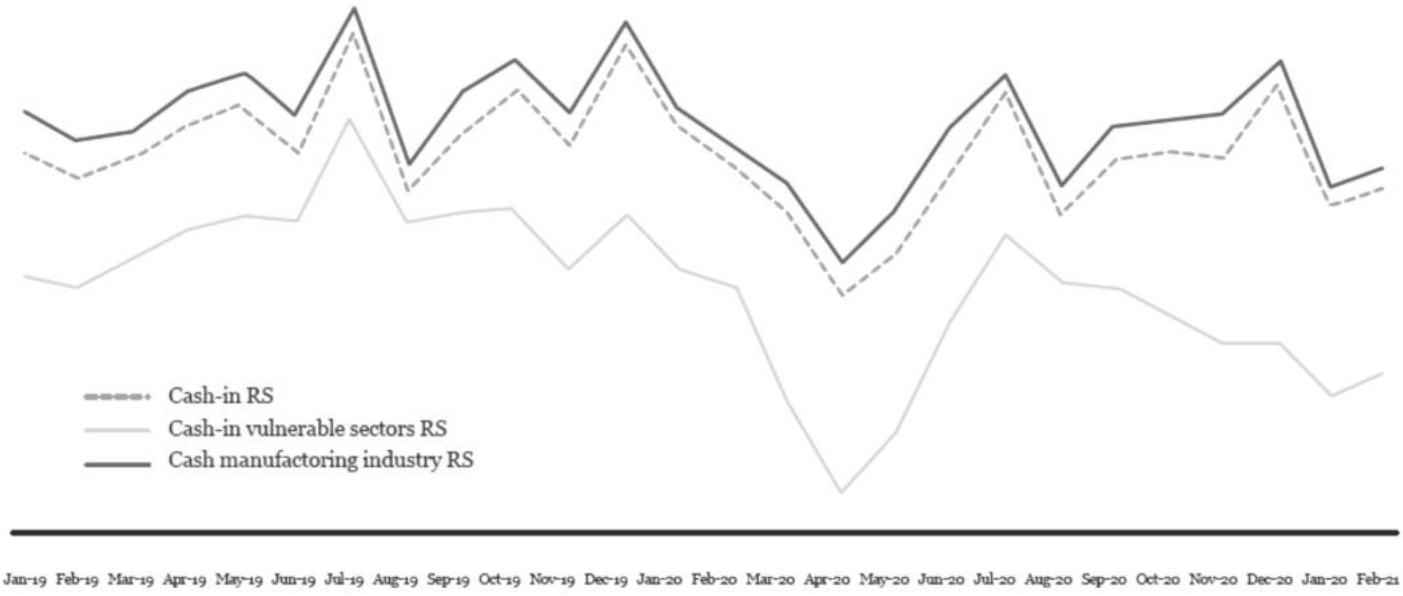

Figure 4 – Cash inflows for retail SMEs ("RS") estimated using transaction data

The ML module initially led to challenges that were (and to an extent are still being) addressed:

- the need to hire new experienced staff in a variety of fields: big data management and related programming languages, machine learning algorithms with a focus on transparency and interpretability, ability to provide adequate documentation, and knowledge of the underlying banking processes;
- a significant investment in IT, in both the estimation and implementation phases, in order to have an infrastructure that could assess machine learning algorithms on estimation samples and replicate them in a production environment;
- a review of the regulatory framework to check that the model was in line with the requirements imposed under the IRB approach. Special attention was paid to Article 179 a) of the CRR ("an institution's own estimates of the risk parameters PD, LGD, conversion factor and EL shall incorporate all relevant data, information and methods. […] The estimates shall be plausible and intuitive and shall be based on the material drivers of the respective risk parameters") and especially to the need to guarantee "plausible and intuitive" estimates. This meant that the bank had to be able to explain the model results, using the techniques (like LIME and SHAP) discussed above;
- finally, ML algorithms required an ad-hoc validation framework, taking into account their specificities (e.g., the optimization of hyperparameters[7]).

The new model allows for a fully digital lending process with ratings computed online, in real time and automatically, leading to greater efficiency and superior credit risk monitoring. The excellent performance of the new ML algorithms, using data retrieved in a fully automated way, allowed the rating to be produced without having to ask the borrower for financial statements/tax forms, which can weigh on the customer journey while becoming quickly obsolete. The model was validated by the ECB in May 2021 and is currently used to compute regulatory capital against credit risk.

3. Case 2: a challenger model to validate the regulatory model for retail PDs

A large European bank had recently changed its IRB model for estimating the PD of retail customers, in order to accommodate the EBA Guidelines on risk parameters and the new definition of default. Once the new model ("the bank's model") had been developed by the risk management department (using a combination of traditional and ML techniques), it had to be assessed by the validation unit as prescribed by the relevant regulations. The bank's pre-existing validation framework had to be reinforced by adding a "challenger" model, based on the same technology as the bank's model (including ML-based modules), to be used for *initial validation* (assessing alternatives to the bank's model obtained by changing/stressing some choices) and for *ongoing validation* (promptly highlighting any issues, including performance drops, in the bank's model).
The bank's model relies on the calculation of an "integrated score" incorporating several intermediate scores generated by specific modules (using, for example, personal data, CRIF's credit bureau data and scores, as well as other socio-demographic information). The challenger model had to focus on a few key modules that were considered especially relevant to the bank model's results. Namely:

- the *financial assets ("AFI")* module, which assesses the borrower's financial assets (including the current account balance) and represents a measure of the borrower's potential wealth;
- the *behavioural* module, focusing on the sub-module that evaluates a borrower's behaviour on the basis of that borrower's credit exposure and financial position with the bank;
- the *mortgage module,* which looks at the products owned by customers who also have a residential mortgage;
- the *cash flow* module, which uses current account data at a transaction level in order to assess the borrower's cash flow management (volatility, growth and liquidity levels) and to identify potential warnings, financial tensions and other income/expense flows. Here, both the bank's model and the challenger model used ML;

The challenger model experimented with many options, e.g. by changing variable categorisations and reducing cross-module correlations, testing different time frames for the indicators, applying different criteria when setting up the estimation sample, adding new indicators to the list of candidate variables. As concerns ML, the challenger model tested alternative solutions for hyperparameters, as well as simplified models based on different materiality thresholds.
By doing so, alternative results were generated for the four "challenged" modules, which were then combined (using the same methodology as in the bank's official model) with the results provided by "un-challenged" modules.

[7] Hyperparameters are parameters whose values control the structure and the learning process of an ML model, thereby determining the number and values of its final parameters. An example of hyperparameters could be the number of nodes and layers in a neural network (whereas its parameters are the weights used in the functions propagating information across nodes). An example of ML-based validation approaches is provided in Case 2 (§3).

The most interesting module was the cash flow module, where innovative transaction data was used, based on borrowers' current account movements.
The module is only applicable to customers using their current account as their "main" account, that is, for day-to-day transactions[8] (excluding accounts that are seldom used and do not correspond to the customers' real habits).
The input data was updated monthly and contained all the transactions taking place on a daily basis. The database aggregated both current accounts held with the bank and, with the explicit consent of the customer under PSD2, current accounts held with other banks[9].
The challenge activity was performed primarily through the tuning of the hyperparameters of the *extreme gradient boosting* XGB algorithm, the same one used in the bank's model[10].
To improve the comparability of the results, XGB was preferred to alternative techniques, such as random forests and deep neural networks.
The tuning activity focused on the following hyperparameters: learning rate, number of estimators (trees), maximum depth for each estimator, minimum child weight, column sample by tree, subsample, and gamma.
While most hyperparameters were already present in the bank's model, two of them (column sample by tree and subsample) were added by the validation unit. By sampling the features and the observations for any given tree, the risk of overfitting should be reduced.
A selection of the challenger models tested, as well as a comparison with the bank's model, is shown in Table 2[11]. Looking at the table, Challenger 5 ("Ch. 5") emerges as the challenger model with the highest overall accuracy, while Challenger 3 is the one with the best test sample accuracy. However, all challengers are characterized by a fairly stable performance in the test sample.

Table 2 - Challenging the hyperparameters (illustrative data)

			Challenger Model - candidates				
		Bank's Model	Ch1	Ch2	Ch3	Ch4	Ch5
Hyperparameters	Learning rate	0.1	0.1	0.1	0.1	0.1	0.1
	No. estimators	200	300	300	300	300	500
	Max. depth	5	5	5	5	5	5
	Min. child weight	10	10	5	5	5	5
	Column sample by tree	-	-	-	-	0.6	0.6
	Subsample	-	-	-	0.8	0.6	0.6
	Gamma	0.8	0.8	0.8	0.8	0.8	0.8
Results – AR (Gini Index)	Development sample	77.94%	80.06%	80.52%	80.63%	79.87%	82.83%
	Test sample	72.54%	72.69%	72.59%	72.78%	72.48%	72.46%
	Overall Sample	76.86%	78.58%	78.58%	79.06%	78.39%	80.76%

To gain a better insight into the underlying logic of the models, we computed the marginal contribution of each variable.
These scores quantify the relative importance of each variable when a model makes a prediction, allowing the most important ones to be identified.
The marginal contributions of all features can then be shown in decreasing order on a graph for each candidate model (Figure 5 provides an example).

[8] To be considered the "main" account, a current account has to meet at least one of these two criteria: i) it is used for receiving the customer's main source of income (salary, pension or welfare payments, self-employed income, housing rents or alimony); and/or ii) the customer receives money transfers or deposits cash on a regular basis.

[9] Keyword searches and tagging were performed on raw data, looking at different transaction types and their descriptions. All indicators were computed for a maximum of 12 months, merging all accounts to reflect the customer's situation as closely as possible. The list of candidate indicators was built by averaging, summing or detecting status changes over the last quarter, semester and year (e.g. the average salary was computed over the last 3, 6 and 12 months).

[10] XGB belongs to the category of ensemble models, as it is composed of a series of weak learners or decision trees, which are built upon in order to generate one final strong learner.

[11] For confidentiality reasons, the table only reports illustrative data.

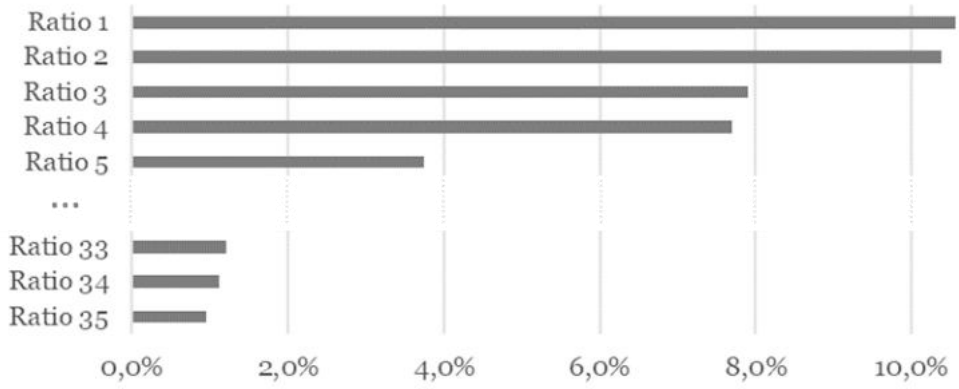

Figure 5 – Importance value graph of the first 35 indicators (illustrative data)

Based on such scores, five different cut-off values were considered: no cut-off, 0.7%, 0.8%, 0.9% and 1%.
For each cut-off strategy, the accuracy ratio ("AR") in the train sample, test sample and total sample were computed. Then, the average contributions to the AR of the variables included in each strategy were computed, and the strategy with the maximum average contribution was selected (see the illustrative example in Table 3).

Table 3 – Model selection (illustrative)

		#Variables				
		A	B	C	D	E
	Sample	90	40	35	30	28
AR	Train sample	80.63%	78.14%	77.87%	76.98%	76.64%
	Test sample	72.78%	71.59%	71.29%	70.53%	70.29%
	Total sample	79.06%	76.83%	76.56%	75.69%	75.37%
Avg. Contribution to AR of inserted variables	Train sample	0.05%	0.05%	0.15%	0.17%	0.18%
	Test sample	0.03%	0.05%	0.13%	0.12%	0.11%
	Total sample	0.05%	0.05%	0.14%	0.16%	0.15%

Criteria used to choose the final challenger model were: a balanced performance in both the test and the train sample, a lower number of variables, a set of features that were more intuitive and explainable to the model's users.

The final decision also considered the results obtained with alternative ML techniques (e.g., random forest), looking at common features and how their relative importance varied across models.

The challenger model enabled the validation unit to perform benchmarking of the bank's official model, looking at:

- the performance of all modules subject to challenge and their final accuracy;
- the different weight of the modules once they were combined into the integrated score;
- the change in the PD master scale when moving from the bank's model to the challenger model;
- the differences and special characteristics in the PD distributions (as shown in Figure 6).

The results of the challenger model were almost comparable to the bank's model: the differences in terms of rating were mostly within one notch; the increase in the accuracy ratio was almost immaterial and corroborated the choices made by the risk management department.

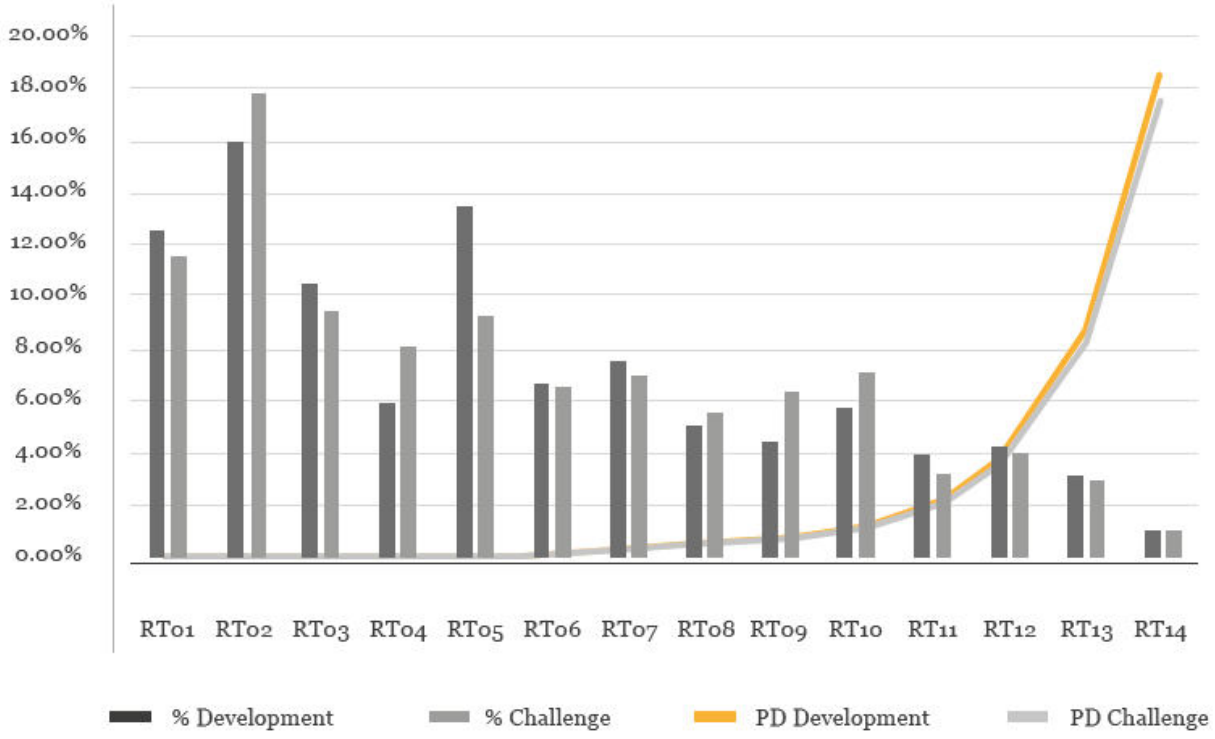

Figure 6 –Relative weight and average PD associated with different rating grades (bank's model vs. challenger model)

The main issues met when developing the challenger model were the following:

- project timing. Before the challenger model can be developed, the validation unit must learn about the bank's model and perform standard validation tests on it, in order to identify areas where the need for benchmarking is stronger. All this must take place in the period between the bank's "pre-application" to gain supervisory recognition of the new model and its final "application";
- the trade-off between the wish for a full challenge, which would completely overturn the framework adopted in the bank's model, and the need to ensure comparability of the results between the challenger model and the official one. As shown above, this quest for comparability means that the basic methodological choices made in the bank's model were kept unchanged in the challenge activity.

In short, the main issues faced were related to "how" the model had to be challenged through independent validation; we chose to act on the list of candidate variables and on the "judgmental choices" made by the risk management department when estimating the bank's model, taking into account only statistical evidence.

4. Case 3: using transaction data for early warning purposes

This case involves an Italian significant institution, a traditional retail bank serving both consumers and businesses and operating domestically. The project was part of a roadmap for the improvement of its early warning systems. The objective was two-fold: responding to supervisory requirements and adopting a cutting-edge early warning tool, with a view to improving risk management, including in a post-Covid-19 context.

The bank's pre-existing early warning model was based on a combination of different types of traditional data (e.g., socio-demographic information, internal and external credit-related signals, financial data, etc.). Traditional modelling approaches (including logistic regression) had proved very effective in discriminating risk and were generally accepted and understood by key users within the bank (e.g. credit analysts) and external stakeholders (e.g. supervisors).

However, due to technological and methodological advances, new transaction-level data was available and ready to use, which could be an important source of information if properly managed: thousands of new indicators could be developed for millions of customers. Where traditional approaches may not be able to fully exploit this data, one would expect machine learning models to identify new information patterns, including in the event of highly non-linear relationships between customer behaviour and credit risk.

The pre-existing solution only used behavioural information at an aggregated level, e.g. by looking at the number of transactions recorded in the last *n* days, at the number of days elapsed since the last transaction took place, or at the average credit amount used over a certain period of time. The bank was now interested in using "atomistic" transaction data that could potentially highlight

specific behaviours: for instance, instead of just considering the number or transactions, one may also look at the nature of the individual items (e.g., payment of instalments and taxes, purchase of goods or services, settlement of invoices, bank transfers from the parent company, etc.), and their mix and evolution over time.

The aim of the project was to enrich the pre-existing early warning solution through the development of several transaction-level modules (each one for a different customer segment) and their integration with the other components of the system (e.g. modules monitoring central credit registry data, etc.).

Leveraging more granular, powerful and faster-reacting transaction-level information, the new solution was expected to better support the bank in managing credit risk in a more challenging economic environment. The data for the ML module consisted of individual transactions for the entire customer base (both individuals and businesses, 24 months of daily transactions, more than 10 million transactions per month).

Each record included the sign, amount, and label ("description") accompanying each transaction.

The model development involved two steps. First, transactions were categorised and turned into customer-level indicators. Second, such indicators were merged with pre-existing information and the early warning system was developed.

Regarding the first step (developing customer-level indicators), transactions were categorised by means of machine learning algorithms integrated with NLP (Natural Language Processing) techniques.

These algorithms "learned" how to interpret current account transactions by reading their description, then assigned them to multiple classes of activities (for example: accounting services, legal fees, purchases of raw materials or services, interest payments, etc.). The algorithms predicted the most likely class of activity and assigned a "reliability" score to the categorisation in order to help identify (and potentially discard) weak outcomes. The process involved human supervision of the machine learning algorithms, aimed at introducing further calibrations if necessary. Once transactions were categorised, they were summarized into 20,000 customer-level indicators, to facilitate data handling, model governance and interpretability.

As far as the second step is concerned (developing the actual early warning system), ML methodologies were applied for both feature selection (i.e. to create a "short list" of relevant variables) and multivariate estimation[12]. The techniques used included three different model classes: random forests (RFs), extreme gradient boosting (XGB) and neural networks (NNs).

For each class, model estimation was performed through the following steps:

- hyperparameter definition (e.g. definition of the minimum number of items in a leaf) by means of a random search engine. This provided a higher configuration space for estimating models than a grid search (Bergstra and Bengio, 2012), thus achieving better results. Table 4 provides an example of how model performance can be affected by changing certain hyperparameters and leaving everything else unchanged;

Table 4 - Random Search Approach in a Random Forest Model

Random Search approach				
No. estimators	min. samples split	min. samples leaf	max. depth	Accuracy
200	**7**	**8**	**8**	**78.28%**
...	...	...	...	...
100	20	50	8	77.76%
...	...	...	...	
200	5	40	8	76.86%

- cross validation: in order to avoid overfitting, cross validation was performed through validation strategies ranging from 3-fold to over 10-fold. In the case shown in Table 4, a 5-fold cross validation strategy was applied to each configuration (in this case, over 1,000 different configurations of hyperparameters were tested). For each model configuration, the accuracy was defined by averaging the Area Under the Curve (AUC) of the 5 test samples defined in each model run;

[12] Although ML methodologies would also be applicable to the management of missing values, the latter were managed through traditional methodologies.

- performance evaluation: all alternative implementations within a class were assessed on the basis of different performance measures, including AUC, Gini index, confusion matrix, model reactivity, etc. (the average result for each cross-validation iteration was considered). Table 5 provides an illustrative example based on data for natural persons. Neural networks ("NNs") show the best performance (Gini index above 80%) and are the richest model in terms of the number of features and statistical units. However, random forests yield a higher TPCR (True Positive Classification Rate) while allowing greater model explainability (e.g., due to a lower number of variables).

Table 5 – Example of performance comparison for different model classes for natural persons

		GINI	TPCR	TNCR	Accuracy
RF	Development	75.2%	71.3%	**87.6%**	79.4%
	OOT	73.1%	72.0%	**85.0%**	78.5%
XGB	Development	74.6%	90.1%	66.5%	78.3%
	OOT	74.9%	89.9%	64.4%	77.1%
NNs	Development	**84.1%**	85.6%	83.3%	**84.4%**
	OOT	**82.3%**	84.8%	82.5%	**83.7%**

All the inputs to an ML model should be individually interpretable, as well as their impact on the target variable. One should, in principle, be able to explain all the determinants of each individual forecast. To achieve such a goal, both global and local interpretability techniques were used.

As concerns the former, feature importance analysis was used, generating visual outputs like in the example provided in Figure 7. It is worth noting that variables 6 and 25 would have been discarded by traditional short-listing methodologies, since their "information value" is below 0.01[13]: unlike linear models, ML models seem to capture relationships that are weak but relevant.

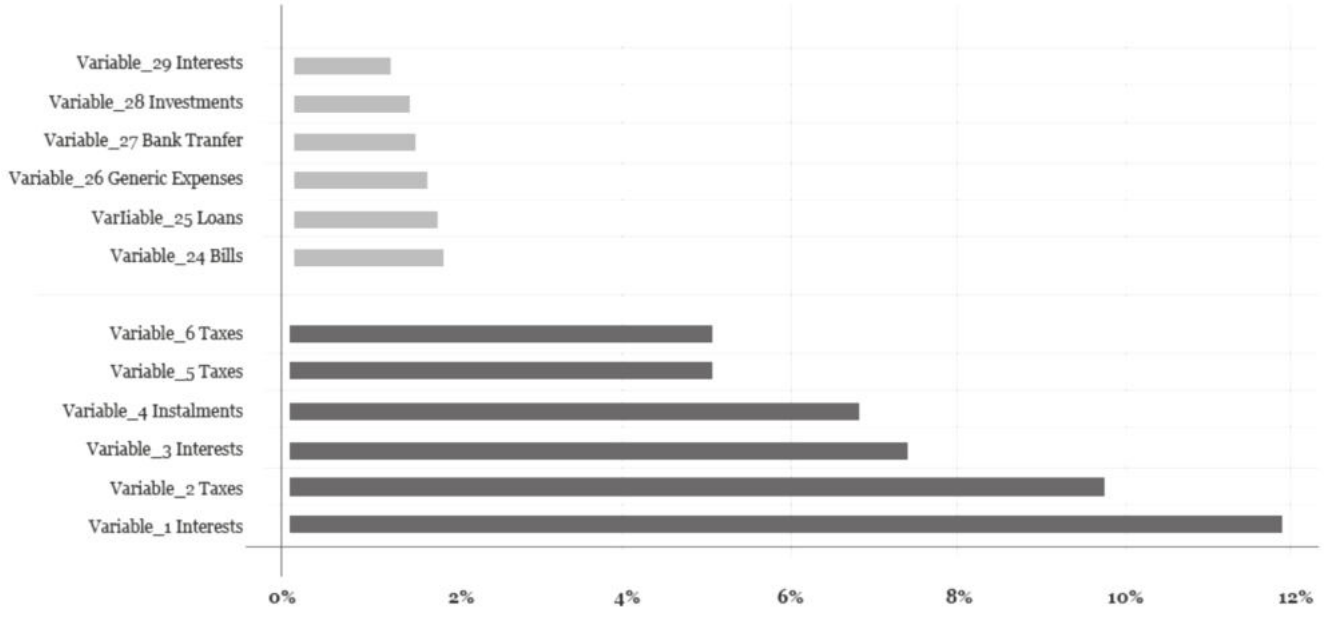

Figure 7 – Sample feature importance chart

[13] See, for example, (Zdravevski et al., 2011) for an introduction to information value and weight of evidence.

As for local interpretability techniques, Shapley values were used to identify variables with a stronger impact on PDs; LIME and OptiLIME-CRIF were also tested: both approaches led to very similar results.
The main benefits of the project can be summarised as follows:

- on the one hand, ML-based models increase accuracy. While such an increase is not dramatic (due to the fact that the bank already had in place a thoroughly tested system), it leads to a greater ability to classify high-risk customers, allowing pre-emptive risk management. On average, transaction data modules help increase overall performance, in terms of Gini index, by 6 percentage points, for both individuals and businesses;
- on the other hand, the bank has acquired a categorisation engine trained on proprietary data (and producing a reliability measure of each generated output), as well as a development lab to create, tune, test and deploy ML-based models.

While ML methodologies can reach outcomes that would otherwise be unattainable, they also raise additional challenges related to data management (i.e. the treatment of missing values and the choice of appropriate short-listing techniques), the governance of the overall modelling process and the interpretability of the final model.
In terms of *data management*, in order to skim the list of candidate variables (over 20,000) and make computations easier and faster, the following combination of traditional and ML steps was used:

- preliminary reduction: variables were excluded whenever they showed a high percentage of missing values, a high concentration of values on few levels or an extremely high correlation with other features;
- generation of new variables starting from those that had passed the first selection step (e.g. by taking the min, max, mean, trends, the standard deviation, etc.), leading to a new "long list" of about 2,000 variables;
- creation of a short list by identifying and removing the features that were not relevant enough in a multivariate model (low feature importance).

Careful feature selection (e.g. using multiple feature selection methods to avoid discarding relevant variables[14]) led to multiple advantages in terms of lower training time, lower risk of overfitting (as less redundant data means less noise, making biased decisions less likely) and greater performance.

Regarding *model development*, the hyperparameters' definition was optimised by means of a random search algorithm (testing several randomly selected combinations of hyperparameters) and cross-validation. Random search, as opposed to grid search, proved paramount in making the project manageable in terms of training time and costs[15].
Finally, as far as model *explainablility* is concerned, human oversight and interpretability were a major concern throughout the project, as they are key to gaining the trust of all stakeholders, ranging from internal functions to supervisors. Global and local interpretability techniques have helped address this goal in a satisfactory way.
Despite the presence of a well-established body of literature on machine learning, the project was carried out using a mixed approach, balancing the computational power of ML and the need for human expert control throughout the estimation process. For instance, missing value management and short-listing also leveraged traditional techniques, thus reducing the risk of counterintuitive solutions generated by ML alone.
The risk of relying on weak outcomes (which may perform badly on new data) was controlled through diversification: in both the modelling and the interpretability steps, several approaches were used, challenging one another. In the modelling step, three different classes were tested in order to select the most performing one (while understanding the power of each one); in the interpretability step, again, different approaches with a varying degree of complexity were adopted. The ultimate result is a compromise between total flexibility and the need to ensure an appropriate level of understanding by stakeholders.

5. Final remarks

ML is still expanding strongly in terms of methodological refinements and innovative applications but can be regarded as a well-established technology when it comes to its key characteristics, weaknesses, and strengths. Such a consensus, among researchers and practitioners, on what ML can and cannot do, should pave the way for a set of standards to be followed in the development, validation and supervision of ML-based models in banks.
ML is not a homogeneous body of results: indeed, it is a wide-ranging label used to encompass a set of techniques that are extremely diverse and should not be treated equally. This is especially true when it comes to interpretability and the risk of creating "black boxes" that are deployed without an appropriate level of awareness and oversight. While deep learning approaches are undoubtedly

[14] For instance, for natural persons, both random forest and gradient boosting-based selections were applied to discard 80% of variables (as opposed to discarding 85% by using random forest alone, and 90% if one were to look only at gradient boosting).
[15] Alternatives to random search include manual search (where several combinations of parameters are defined based on human judgement) and grid search (where different combinations of hyperparameters are chosen on a value grid, and parameters are then optimally chosen in a *neighbourhood*). While grid search requires the testing of every possible combination of hyperparameters, random search allows calibration of the number of search iterations on the basis of time/resource constraints. According to (Ribeiro et al., 2016), if the close-to-optimal region of hyperparameters occupies at least 5% of the grid surface, then random search with a fixed number of trials is highly likely to find it.

prone to such a risk, techniques like those used in our case studies (including, for example, decision trees, random forests, XGB) have reached full maturity and allow model developers to deal with transparency and avoid overfitting.

When it comes to transparency, however, the global and local interpretability techniques shown in our case histories should not be seen as a target, but rather as a means to facilitate the dialogue with model users. A continuous interaction with stakeholders (including the bank's business-oriented functions, its middle management and board of directors) is key to ensuring that all implications of a new algorithm are fully understood before it becomes part of an institution's risk management toolbox. Pilots, dashboards, and "explainers" should not be seen as mere sweeteners, given out to smoothen model acceptance, but rather as a fundamental step in model development, a recipe for greater robustness and a source of mutual enrichment for model engineers and users.

Finally, it should be borne in mind that slowing down innovation is not an option. While new ML-based models clearly involve risks and weaknesses that must be carefully addressed, inaction has its own costs and dangers. Discouraging the use of innovative models and data sources may result in banks competing with non-bank entities with one arm tied behind their back; increasing the gap between the models used for internal risk management purposes and those validated under the IRB approach may undermine the latter's credibility and, in the long run, prove detrimental to effective bank supervision.

References

- Bergstra, J., Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research 13, 281-305.
- Bholat, D., Gharbawi, M., Thew, O., 2020. The impact of Covid on machine learning and data science in UK banking. Bank of England Quarterly Bulletin.
- Breeden, J., 2021. A survey of machine learning in credit risk. JCR. https://doi.org/10.21314/JCR.2021.008
- Di Biasi, P., Gnutti, R., Resti, A., Vergari, D., Basile, A., Bernabei, F., Caprara, C., Cavarero, D., Marigliano, M., Ranaldi, R., Vignolo, M., 2022. Machine Learning for Credit Risk Management and IRB Models: Lessons from success Case Histories (A joint paper by Intesa Sanpaolo and CRIF). Intesa Sanpaolo - CRIF, Milano-Bologna.
- European Banking Authority, 2020. EBA Report on Big Data and Advanced Analytics (No. EBA/REP/2020/01). European Banking Authority, Paris.
- Giudici, P., Raffinetti, E., 2021. Shapley-Lorenz eXplainable Artificial Intelligence. Expert Systems with Applications 167, 114104. https://doi.org/10.1016/j.eswa.2020.114104
- Guerra, P., Castelli, M., 2021. Machine Learning Applied to Banking Supervision a Literature Review. Risks 9, 136. https://doi.org/10.3390/risks9070136
- Huang, J., Chai, J., Cho, S., 2020. Deep learning in finance and banking: A literature review and classification. Front. Bus. Res. China 14, 13. https://doi.org/10.1186/s11782-020-00082-6
- Institute of International Finance, 2019. Machine Learning in Credit Risk (Summary report). Washington D.C.
- Leo, M., Sharma, S., Maddulety, K., 2019. Machine Learning in Banking Risk Management: A Literature Review. Risks 7. https://doi.org/10.3390/risks7010029
- Molnar, C., 2019. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.
- Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.
- Rundo, F., Trenta, F., di Stallo, A.L., Battiato, S., 2019. Machine Learning for Quantitative Finance Applications: A Survey. Applied Sciences 9. https://doi.org/10.3390/app9245574
- Visani, G., Bagli, E., Chesani, F., 2020. OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms.
- Zdravevski, E., Lameski, P., Kulakov, A., 2011. Weight of evidence as a tool for attribute transformation in the preprocessing stage of supervised learning algorithms, in: The 2011 International Joint Conference on Neural Networks. pp. 181-188. https://doi.org/10.1109/IJCNN.2011.6033219