Deep Science Publishing
https://doi.org/10.70593/978-93-49910-16-4_11

DeepScience
Open Access Books

Chapter 11: Building trust in automated systems through transparent credit risk evaluation models

11.1. Introduction

An algorithmic culture has emerged in which decisions affecting our daily life increasingly depend on automated systems such as machine learning. Developers of those systems strive for more accuracy, while at the same time demands for accountability increase. In many real-life applications, black box systems operate across domains like finance, health care, transportation, and criminal justice, often leaning towards more complex and less transparent machine learning (ML) models. Stakeholders responsible for an automating decision that can dramatically influence citizens' lives and regulate the operations of automated systems require information on why values are assigned and details of how the modelling decisions were made to build trust. Financial institutions are at the forefront of this trend and are slowly but creatively adopting these technologies to perform fundamentals such as financial audits, risk assessment, fraud detection, and customer scoring (Joshi & Patel, 2024; Patel & Shah, 2024; Sharma & Singh, 2024).

Credit assessments are necessary for financial institutions; they are essential in determining whether a financing request should be accepted or rejected. Generally, this task is done by risk experts who analyze data referring to loan applicants and manually produce credit risk reports. In practice, credit assessments can be easier and less prone to human error if they are automated with machine learning techniques. A primary objective throughout the process is building models that can estimate the probability of default of the applicant company, as well as highlighting which characteristics are responsible for this evaluation. Some of the models are black box systems that provide only a single value as output (the probability of a default), and thus no additional

information about the data are provided. Therefore, it is hard for a risk analyst to rely on the output estimated (Singh & Gupta, 2024; Verma & Mehta, 2024).

Recently, the transparency and interpretability of machine learning models has attracted increasing attention due to both societal and regulatory pressure across many domains. Among them, financial technology, and therefore credit risk scoring, has relevant implications on the economy as a whole and on people's lives. Financial institutions are subject to rigorous guidelines and regulations when addressing credit risk scoring. Transparency of the models is a crucial issue in that context. Despite rigorous requirements for interpretability and explainability, recent advancements in automated credit risk scoring tend to rely on black box algorithms. Consequently, the need arises for more transparent machine learning techniques with the ability to shed light on the credit risk score.

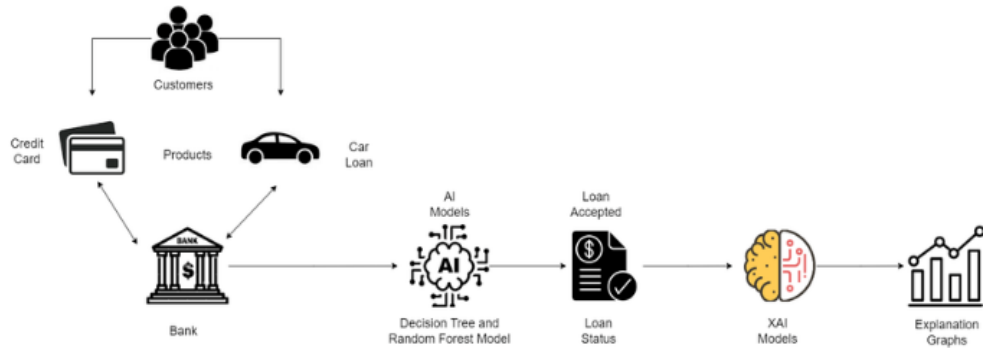

Fig 11.1: Credit risk evaluation architecture

11.1.1. Research design

The study follows an experimental (quantitative) research approach to investigate whether enhancing the transparency of credit risk evaluation models can positively affect trust in automated decisions and loan acceptance intentions. An experimental design was used in the study to manipulate the independent variable, the transparency of the credit assessment model. The model's transparency was either low or high. The trust in automated credit decisions was measured as a multi-dimensional construct that captures the level of epistemic, procedural, and interpersonal trust. The intention to accept the credit loan was operationalized using a behavioral intention-based construct. The data were collected using a scenario-based online survey targeting credit-seeking individuals. Three retained attributes, ease-of-use, accuracy, and trust, were used in the experiments. Furthermore, a set of perceived similarity scales and six control variables were used, including technology experience and payment attitude. The overall aim of the study is to examine how the transparency of automated credit risk evaluation models affects

individuals' trust in automated credit risk assessment and subsequent loan acceptance decisions. More specifically, the study aims to investigate, first, how presenting the output of the credit risk evaluation models in a text-based and graphical means affects trust in automated credit risk assessment and second, how individuals' trust in automated credit risk assessment affects their loan acceptance intention.

11.2. Background on Credit Risk Assessment

In section 11.2, a background on credit risk assessment is provided. It discusses the evaluation of credit risk which plays a pivotal role in a financial institution's risk management. The fund providers need to share the risk of defaulting through precise collateral deposit rules and fixed pricing rules. Default prediction models can help the decision-makers improve their operation through allocating inspection resources, automating the risk-scoring service, and improving the credit products. The transparent model establishment approach based on analytically ground global patterns with closed form is shown to solve the lack of transparency problem in credit scoring. Such transparent fiscal analysis is vital for legislators and regulators' understanding of a firm's fiscal status and the compliance of the adequate capital requirements for financed firms. Many financial rating agencies and large financial institutions supported by unsolved data sciences provide credit assessment services and charge a high price. However, the opaque models on easily governed data often induce market blindness, excitement, and even excessive complacency which worsen the financial stability when scaled. The social regulatory agencies who oversee the credit assessment are often neglected. The existing approaches are based on complex learning mechanisms with a higher share of risk prediction. Alternatively, reward-penalty based transparent models can be established in fiscal-risk-statistics representative zip files containing succinct analytical solutions, progressive horizon sampling, and a fixed set of market factors. According to the present disclosure, a data-driven transparent model establishment solution applied for the evaluation of a firm's credit risk through comprehensive fiscal performance analysis is detailed. The data-driven fiscal-variable-correlation discovery and market factor discovery embodied with fiscal-path-HDP HMM and reward-penalty model training driven by the optimal transaction price combined with controls are demonstrated. The ground fiscal-statistics-based evaluations of firm credit risk are provided through on-the-fly fiscal-statistics-distribution density estimates. The transparent model has a compact non-redundant structure with succinct analytical forms, and the fiscal variable valuation or model output is fixed with a closed form. Different from the existing opaque black-box models, the present model provides analytical grounds for legislators, regulators, issuers, and fund investors to interpret and explain the on-line model output. Moreover, comparative studies on the benefit comparison of opaque and transparent models in market data trading, prediction share, and robust

competitiveness are elaborated to analyze the economics implications of the present transparent models.

11.2.1. Historical Context

Trust in automated systems is an ever-growing necessity due to the historical lack of human involvement in key decisions such as loan acceptance or denial through credit risk scoring. In a time when limited human judgment in systems like bank lending is not new, it is easier to understand as this limited human involvement is coupled with an evaluation process that is able to explain the decision made. Unlike systems that can provide more understandable information about their decision, credit risk evaluation systems lack such a procedure. Therefore, evaluating the credit risk of a company based on financial historical information is a black-box process, and this has come at a cost. There is a clear need for the development of transparent models for credit risk scoring to avoid the impediments that such black-box systems can create in automated decision systems. However, this should not come at the cost of lowering the accuracy of such models. Automating the process of credit risk evaluation has received notable attention in the form of a credit risk scoring model, where the objective of the model is to classify whether the company is going to experience financial problems in a time horizon. With the emergence of machine learning, there is a chance to take advantage of increasingly versatile data. However, the establishment of a credit risk evaluation model is far more than just building a scoring model. The requirement rests on the black-box nature of most machine learning models. Current models, such as logistical regression and classification and regression trees, have mechanisms that can be understood to some degree. They are interpretable by design or, to put more formally, transparent. More powerful models, such as gradient-boosting trees and random forest classifiers, do not follow through and cannot explain the reasoning behind the black-box predictions.

11.2.2. Current Trends in Credit Risk Evaluation

Credit risk evaluation consists of deciding if financial facilities shall be granted to an applicant. The loan granting decision is based on a credit risk evaluation model, applied to the applicant's data to compute features such as probabilities of default. It is for the financial agent to assess the computed risk and to decide whether or not to grant the loan. The financial agent is responsible for the decision taken. Building such a credit risk evaluation model consists of assigning values to an accurate scorecard model where input features are risk indicators and splitting points are threshold values between model scores. The model's parameters can be learnt from historical data containing past application records and delinquency labels. Accurate models are based on supervised

machine learning methods. Models learnt using supervised methods together with data over a longer time period than the scorecard reference validity shall be considered uncontrollable black box models by the financial agent. In recent years, there has been a growing move towards the third generation of credit scoring models based on machine learning and unleashing the full potential of data . Most of the third generation commercial credit risk evaluation models are black box models. The need for interpretability and explainability of black box models has become stronger than ever. Finally, the problem of model credit risk evaluation refers to understanding a credit risk evaluation model and the risk level calculation associated with an application record. The model credit risk evaluation deals with explainable artificial intelligence for machine learning taken models representing credit risk evaluation processes.

After providing sufficient information about credit risk evaluation models, the next part of the paper deals with model credit risk evaluation. How to check if a machine intelligible credit risk evaluation model has been learnt has not been satisfactorily addressed in literature. A new approach for credit risk evaluation model credit risk evaluation is proposed based on the theory of forms of measure and mutual information. As a prologue, this proposal is preceded by discussing the meaning and attributes of a credit risk evaluation model. A credit risk evaluation model comprises a scorecard model and a model hyper-parameters' set making the scorecard model and the model hyper-parameters a model specification. It is for the model specification to be controlled so that the machine learnt model remains faithful to the specification.

11.3. Importance of Trust in Automated Systems

Trust is critical for the acceptance of decisions made or recommended by any automated system. Without trust, there is little belief that the recommendations made by the automated model will be accurate. Algorithmically produced credit score decisions will not be accepted by organizations until there is a high degree of trust in the credit score decision produced. Credit score decisions made by machine learning models must be understandable to help foster trust in automated credit risk evaluation models [3]. Due to a concern that the evaluation model was not transparent enough, there is a need to discover how ML-based credit risk evaluation models can be best communicated, explained, and understood. Addressing this need could decrease the barrier for adopting automated credit risk evaluation models while complying with the provisions of the General Data Protection Regulation (GDPR) across European Union (EU) nations and the United Kingdom (UK).

Trust is fundamental to relationships between parties. Whether trusting co-workers with a job, trusting friends with a secret, or trusting a lover with the heart, trust is essential for effective collaboration. The same principle applies equally to systems – be they

computerized or not. Trust is critical at all three levels of acceptance, delegation, and dependence for people's willingness to accept automated systems – to delegate tasks to them and depend on them – in industrial environments. Trust-based adoption is equally relevant to deciding whether to trust the performance of a newly developed intelligent system in a given scenario. Trust is essential for the widespread acceptance of all kinds of systems, in particular more intelligent ones. Trust in people or automatically generated systems change over time, whether learning from previous interactions, locally or socially. It evolves over the periods prior to, during, and after widespread acceptance.

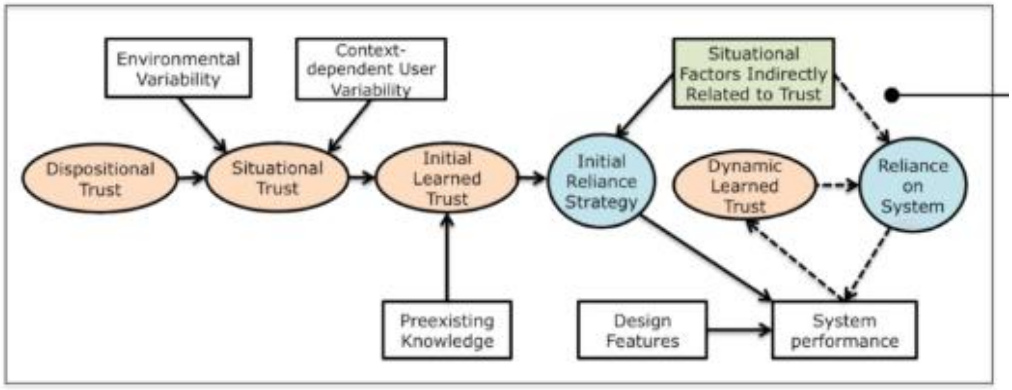

Fig 11.2: Trust in Automated Systems

11.3.1. Defining Trust in Automation

The National Institute of Standards and Technology (NIST) automates credit scoring models and risk management. To achieve those aims, clear expectations of transparency and more comprehensible automated credit scoring are defined. Currently, it is highly valuable to assess accuracy and/or explain credit scoring models. Modern machine learning-based anti-discrimination algorithms are considered particularly suited for the development of transparent credit scoring models, with regard to computational efficiency and model interpretability.

The concept of trust in automation (TiA) remains even more vaguely defined in real-world applications. This is especially true when it comes to contexts where decision support systems (DSS) automate high-stakes decisions that have severe implications for human well-being. For instance, in the criminal justice system, many states and localities use automated systems that estimate the probability of defendant recidivism to assist the risk assessment of pretrial detention. On the one hand, these systems are meant to provide just and unbiased determinations; on the other hand, they have sparked ethical

and legal controversies as the models are perceived as “black-boxes” that cannot generate trustworthy determinations.

Nevertheless, some relevant facets of TiA have been identified in previous research. Specifically, trust can be understood as a multi-faceted and hierarchical construct, and consumers’ trust is formed in a multi-stage process. The very first stage of trust building rests on human cognitive trust and is challenged by machine automation that deviates from expected norms or brings undesired outcomes. In this case, transparent credit risk evaluation models need to be developed for algorithmic systems in a way that meets consumer perceptions and concerns. Transparency-related criteria that enhance TiA under various circumstances in the task of automated credit risk evaluation may thus be defined alongside focusing on increasing model interpretability, visualizability, and comprehensibility.

11.3.2. Factors Influencing Trust

Trust in automated systems is a complex concept and understanding trustworthiness entails several aspects. Inconsistency between decisions made by human experts and automated systems decreases user trust as such discrepancies trigger concerns about the accuracy of models. Inconsistency between modeled features contributes to distrust as users' personal assistance knowledge may no longer be applicable. Transparency in the form of understandable explanations of the modeled feature space leads to increased user trust. Insufficient model explanations decrease user trust as well. Algorithmic design choices may also induce distrust. Reinforcement learning algorithms offering improved scalability, speed, and performance often result in less trust and explainability than Bayesian networks. Model performance, transparency features, and visualizations are amongst the most important criteria measuring users’ trust. Factors described by such as understanding of systems, model reliability, and consistency through assessment of models have particular consequences on trustworthiness in the context of automated credit risk evaluation tools.

11.4. Transparency in Automated Systems

Recent advances in the fields of science, machine learning, optimization, and data storage accelerators have pushed societies toward an increasingly automated and intelligent future. Algorithms are slowly taking over crucial roles that have been traditionally held by humans, such as hiring employees, judging investors, or identifying criminal behaviors. While the struggles of people who had their credit rejected can be generally understood, how a long-lived conversation left out one human can be more difficult to grasp. In addition, the opinion similarity between two people found by a

graph neural network can seem even more abstract. These realities beg the question: how to scrutinize opinions made by the industry in favor of full automation? What information is needed to ensure that the advantages of the allocation-based algorithms prevail?

One way of addressing these questions is to provide transparency to agents interacting with an algorithm, akin to how humans often share their reasoning processes. How will the customers be informed of the outcomes made by an automated credit scoring algorithm? What does the bank do when a person returns to seek explanations? Which analysis will likely result in unfair consequences such as disqualification? Transparency is not a universal requirement for an algorithm. Many algorithms are unintelligible due to the divergence in scale between human cognition and computation. By contrast, those agents striving for transparency have misaligned interests or simply too much information to concern individuals.

Recently established information theory notions such as maximum entropy and Kolmogorov complexity control the trade-off between the utility and explainability of a system. These theories were harnessed in both economics and computer science, leading to transparent platforms that reveal altruistic behavior and models with interpretable structures. While this line of work remains prosperous, risk evaluation is a highly demanding mission, where the allocation-based approach is only one of the partial solutions. Consequently, research is called upon that investigates types of credit risk evaluation models and the corresponding transparency requirements. Machine learning models, which have been popularly applied to the highest monetary stakes, are blamed for remaining primarily in the 'black box' form. Given the risks of unnecessarily high loan default or overly cautious credit loss, there have been regulations across the globe mandating reasoning behind any rejection decision policy. As a result, methods have emerged for explaining the predictions of a 'black box' model. Nonetheless, even these methods are only applicable to either regression-based score forecasts or 'gray box' models with algebraic structures.

11.4.1. The Role of Transparency

Espresso Mode and Turbo Mode are the two operational modes of the system. In the latter mode, where higher accuracy is of the utmost importance, both the dictionary and the eventually obtained model are larger and, consequently, less interpretable. Commonly accepted willingness to give up some transparency for the sake of a more accurate AI model is characterized as the transparency-explainability sustainable tradeoffs. However, trustworthiness decided on the basis of transparency requires both a confidential data infrastructure and system architecture in a tacit relation to one's experience of system operation. A transparent and intuitive operation of a complex AI

system is misleading as a proxy for a trustworthy operation of the system. Therefore, additional explainability is needed in order to regenerate trustworthiness decided on the basis of transparency.

Adaptive Explainable System (AES) is proposed as a high-level model illustrating how trustworthiness in transparent automated systems could evolve. AES has transparency that does not comprise confidentiality of data. But it's reasoned response to trust estimation based on transparency and adaptation of the level any undesired system operation was not re-evaluated could lead to trustworthiness that exceeds that in the case of the trust estimation based on transparency only. In the transparent credit risk evaluation model, AES was realized as generating self-history and self-explanation and providing higher-level transparency. It is integrated and applied onto the credit risk evaluation automated model. The adaptive explainable credit risk evaluation system demonstrated its adaptive explainability by providing self-explanation of the generated self-history and self-history's drivers and empowering the user with data transparency. It is an example of a public good with a high usability and transferability in a purposely designed way.

11.4.2. Measuring Transparency

Transparency measurements establish a standardized procedure for obtaining values for the transparency indicators presented in section 11.4.1. The measurements are as follows.

Default Indicators

Prior to making the model selection, information on the model applications is gathered. If the model is developed for auditing or regulatory explanations, there will be no requirements for model approvals. These default indicators will all take on a value of zero. Either before or after model selection, the model developers gather information on the data sources and create the details of the learning algorithms, both of which establish transparency for the model applications.

Actionability Indicators

Before the model selection or before creating the metrics for periodical target update, there were no routine actions to knock down the need of a new model. Therefore, these default indicators would take on a value of zero.

Ex-Ante Explainability Indicators

General definitions of features interpretability and use interpretability are adopted. These definitions are complemented by specifying their features and details for the access.

Features interpretability measures the complexity of the features in terms of their difficulty to understand as viewed in a micro scale. For global simplify attentiveness, definition of the influence functions is used with a local linear regression approximation for the target model. The maximum and minimum fractions of the smallest and largest values are used to define the numerical values. Global capabilities of a closed interval containing all values of the assigned risk ranges measure the degree of a global understanding of the prediction behaviours. For the implemented algorithms that make up the prediction capabilities, the criteria for comparing these code numbers and argument forms are defined respectively.

11.5. Credit Risk Evaluation Models

On April 24–25, 2023, a two-day symposium was held in the Faculty of Social Sciences at University of Antwerp, and the Port Authority's headquarters in Antwerp. The primary goal of the symposium was to bring together thought leaders in the fields of shipping, logistics, and big data to engage in discussions about the challenges and possibilities posed by the rapidly increasing digitalization of maritime and logistics operations. The symposium served as an exceptional platform for the exchange of ideas and expertise on the potential of big data in shipping and logistics among academia, government, and industry.

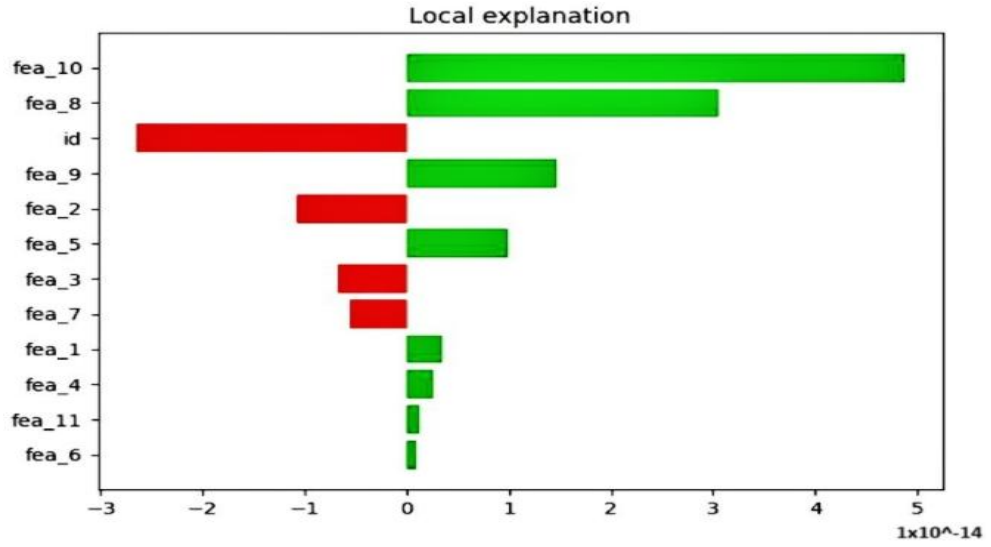

Fig : Credit Risk Assessment and Financial Decision Support Using Explainable Artificial Intelligence

Interpretable CREMs have broad applications in sensitive domains such as credit risk evaluation, personalized medicine, and law enforcement. Due to their well understood

predictive accuracy as well as their gained trust by credit rating experts over decades, conventional CREMs have been thoroughly studied and embraced widely by banks and on-line lending services. However, the rapid advance of new data-driven methods has started to challenge this status quo, introducing CREMs that have superior predictive accuracy . Unfortunately, many of these new methods have been termed black-box in terms of the way they treat and process input variables; and consequently, they may not trust, maintaining the potential risk of sending invalid credit grading decisions.

11.5.1. Traditional Models

The classic credit risk models rely on sophisticated statistical techniques that emerged in the mid-20th century. These statistical techniques include regression, discriminatory analysis, and tree-based models. Statistically oriented credit risk models convert the default risk into a score through a series of carefully designed numerical operations on underlying features.

The score is bounded, interpretable, and reveals risk, which is generally a continuous variable. Default is predicted based on a default threshold after normalizing the score to a default number. The credit risk evaluation procedure consists of three key components: feature transformation to predict credit risk; score generation based on selected features; and risk evaluation and decision based on the score. Feature transformation derives critical input variables to compute the score based on historical decisions and outcomes. A rigorous model construction and selection procedure, which consists of variable selection and statistical assessment of model explanatory power is then performed to ensure the model's prediction power.

11.5.2. Machine Learning Approaches

In finance, there are a myriad of applications of machine learning approaches. Applications include pre-approval, credit scoring, and credit underwriting for personal, business and mortgage loans . Machine learning can analyze the training data to automatically detect the non-linearities and interactions. This allows for credit decisions that are made faster and more accurately. For example, as an extension to the traditional credit scoring modelling, machine learning modelling that analyses available data to predict the credit scores of potential loan applicants is being used more often. In light of new regulations that require an explanation for credit decisions, the explainability of these models becomes an important item for the innovation of ML-based applications in finance. To understand the model and the lending policies better, various explainable AI (XAI) and explainable machine learning techniques are considered.

In recent years, the use of Machine Learning (ML) models in the financial services domain has grown significantly. In particular, ML methods are employed for the task of credit risk assessment. The introduction of new algorithm classifications provides on top of traditional scorecards a much broader and more positive palette. There are still many advantages that such models may bring. Yet, there was genuine concern regarding their acceptance and widespread use for credit scoring tasks. These concerns originate from their non-transparency and black-box characteristics. The incapacity to explain a prediction in a way that is understandable to industry experts and by this not fitting regulatory requirements are key obstacles preventing the general acceptance of ML approaches.

On the other side, a great demand for clarity and trust from regulators, auditors and consumers reveals the importance of these design principles. In this context, a framework is presented that enables the disclosure of the algorithmic structure of a model and its prediction. The explanation addresses algorithmic content (what was learned), instance-relation (how the known data and the prediction relate) and process-integrity (how the ML system is governed and monitored). An overview of techniques, software implementations and how they combine into a valuable toolbox is provided. Finally, the usability of the framework is illustrated within an example of credit scoring for consumer credit. As outlined, these specifications were in great discord with the workings of XGBoost.

11.6. Conclusion

In this chapter, a transparent credit risk evaluation model based on interpretable mortgage scores is presented. It draws on previous works that have analyzed the regulatory requirements with respect to transparency and auditability of machine learning models in credit risk evaluation or scoring as well as on the conditions for making the scorecard comprehensible by people not having a quantitative background. Providing scoring scores such as reasons for assignment to a scoring class may convince clients to trust a machine learning model. However, such arguments are strictly based on formal logic and do not mention the learning behavior of a model as a possible explanation for its predictions. Neural networks, deep learning, or surface complex models, for instance, may use tricker learning behavior, which is hidden from the clients' side. In addition, such explanations do not guarantee future reliability. As a rule of thumb, the more complex the underlying model, the more formal consequences may be inferred from a given explanation, the more discrepancies might arise between the model and the explanation. As this problem is particularly crucial in the financial sector, it is vital to deal with credit risk evaluation or scoring based on transparent and comprehensible machine learning models. If such models work with limited data,

reduced variance, and a large amount of feature variables, it becomes even more complicated.

The proposed model is based on an interpretable mortgage score and transparent rules automatically created out of the distributed mortgage data. The interpretable mortgage score considers deficits such as irregular income, unemployment, maturity of the mortgage, and personal situation of both partners and, if required, taxes between 18-68 months. Trust or non-trust in self-learning systems is a crucial topic as more and more self-learning models are (partly) standardly introduced into businesses. Cryptographic stock market trading or risk assessment of clients in financial institutions, for instance, can have fatal consequences. If it turns out that such systems behave poorly, clients will move to other providers, and trust in self-learning systems will decline. Nevertheless, simple rules, such as decision node rules in tree models, understanding their inner mechanisms requires comprehensive consideration and analysis of all possible scenarios. More importantly, the predictive power, resilience, and reliability of such models are crucial as overall trust is a combination of understanding reasons for trust and non-trust and other aspects such as performance, risk, and reliability.

11.6.1. Emerging Technologies

Digitalization is reshaping the economy, creating new opportunities for companies and society. However, it also opens new risks in the economic system which should not be underestimated. Indeed, the systemic risk of insolvencies and defaults, which was supposed to erase in the digital era, has appeared as an important challenge for central banks in different countries. Credit risk assessments are activities performed at financial institutions such as banks, insurance companies or investment funds to estimate the probability of a default of a company. The confidentiality surrounding the companies' business model leads to a difficult assessment for credit risk analysts who combine their knowledge and qualitative understanding with lots of imbalance and unstructured quantitative information. This process is both time and human resource consuming leading to approval delays and loss of opportunities. Providing screening tools that automatically assess the level of credit risk of companies and flag the most critical ones is therefore seen as a priority task for banks, consumers and credit companies. With the advancements in Artificial Intelligence, a new era for the automation of credit scoring emerged. In this context, the objective of this PhD thesis is to develop new automated models of credit risk evaluation compatible with the expertise dynamics. On the one hand, this work aims at building new modeling frameworks able to automatically assess the probabilities of defaults of companies using a bigger, more heterogeneous and more commercial dataset than before. Although the focus is primarily placed on the design of models, the key properties that such models should display and comply with are

identified. On the other hand, Seamless Decision Trees were proposed, a new hybrid machine learning model that improves existing applications by providing a rigorous modeling framework dedicated to capture the dynamics of the expertise over time.

References

Patel, J., & Shah, P. (2024). Predictive Analytics in Loan Default Prediction. Journal of Risk and Financial Management, 17(3), 112–126.

Singh, A., & Gupta, N. (2024). Cloud-Based Financial Platforms: A Comparative Analysis. Journal of Financial Technology, 12(4), 67–80.

Verma, K., & Mehta, A. (2024). Blockchain for Secure Financial Transactions: A Review. International Journal of Financial Technology, 6(1), 34–47.

Joshi, S., & Patel, D. (2024). AI in Credit Scoring: Enhancing Accuracy and Efficiency. Journal of Credit Risk Management, 11(2), 89–102.

Sharma, R., & Singh, J. (2024). Big Data in Investment Strategies: A Case Study Approach. Journal of Investment Analysis, 14(3), 123–137.