IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 August 2022, accepted 26 August 2022, date of publication 29 August 2022, date of current version 20 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3202889

RESEARCH ARTICLE

Ethically Responsible Machine Learning in Fintech

MARYAN RIZINSKI[1,2], HRISTIJAN PESHOV[2], KOSTADIN MISHEV[2], LUBOMIR T. CHITKUSHEV[1], IRENA VODENSKA[3], AND DIMITAR TRAJANOV[1,2], (Member, IEEE)

[1]Department of Computer Science, Metropolitan College, Boston University, Boston, MA 02215, USA
[2]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000 Skopje, North Macedonia
[3]Laboratory for Interdisciplinary Finance and Economics (LIFE) Research, Administrative Sciences Department, Metropolitan College, Boston University, Boston, MA 02215, USA

Corresponding author: Maryan Rizinski (rizinski@bu.edu)

ABSTRACT Rapid technological developments in the last decade have contributed to using machine learning (ML) in various economic sectors. Financial institutions have embraced technology and have applied ML algorithms in trading, portfolio management, and investment advising. Large-scale automation capabilities and cost savings make the ML algorithms attractive for personal and corporate finance applications. Using ML applications in finance raises ethical issues that need to be carefully examined. We engage a group of experts in finance and ethics to evaluate the relationship between ethical principles of finance and ML. The paper compares the experts' findings with the results obtained using natural language processing (NLP) transformer models, given their ability to capture the semantic text similarity. The results reveal that the finance principles of integrity and fairness have the most significant relationships with ML ethics. The study includes a use case with SHapley Additive exPlanations (SHAP) and Microsoft Responsible AI Widgets explainability tools for error analysis and visualization of ML models. It analyzes credit card approval data and demonstrates that the explainability tools can address ethical issues in fintech, and improve transparency, thereby increasing the overall trustworthiness of ML models. The results show that both humans and machines could err in approving credit card requests despite using their best judgment based on the available information. Hence, human-machine collaboration could contribute to improved decision-making in finance. We propose a conceptual framework for addressing ethical challenges in fintech such as bias, discrimination, differential pricing, conflict of interest, and data protection.

INDEX TERMS Ethics, machine learning, explainability, finance, fintech, financial services.

I. INTRODUCTION

Machine learning (ML) systems have been implemented by financial institutions across various financial services. ML algorithms are applied to personal finance (through chatbots powered with natural language processing or personalized insights for wealth management), consumer finance (with the ability to prevent fraud in online payments), and corporate finance (by predicting, assessing, and reducing loan risks, improving loan underwriting, and prevention of money laundering) with aggregate potential for cost savings for financial institutions estimated at $447 billion by 2023 [1]. In particular, as one of the drivers for innovation in fintech, machine learning has been used in applications ranging from assessing individual credit risk [2] and defining criteria for lending [3] to designing credit scoring models [4] and optimizing asset management [5], [6] as well as predicting success of fintech projects using crowdfunding [7]. Empowered by machine learning, the fintech field holds promises for financial inclusion [8], [9].

Executives from 151 financial institutions from more than 30 countries identified AI as an essential business driver across the financial industry, as revealed in a comprehensive global survey on AI in financial services conducted jointly by the World Economic Forum and the Cambridge Centre for Alternative Finance (CCAF) at the University of Cambridge Judge Business School, supported by Ernst & Young (EY) and Invesco [10], [11]. According to the survey, 52% of the

The associate editor coordinating the review of this manuscript and approving it for publication was Yeliz Karaca.

respondents are currently implementing AI-enabled products and processes, with 77% expecting that AI will become essential to their business within two years. Similarly, according to a survey by Deloitte Insights, 70% of all financial services firms are using machine learning to predict cash flow events, fine-tune credit scores, and detect fraud [12].

Moreover, major international corporations even develop their own codes of ethics for AI, which set guidelines for the development of safe, robust, and explainable AI products that combine innovation with social responsibility [13], [14]. Similarly, professional organizations in the field of computer science require their members to take extraordinary care ''to identify and mitigate potential risks in machine learning systems'' [15].

While machine learning will undoubtedly have an impact on financial institutions, they are also expected to face a range of challenges. For example, more than 80% of the respondents in the World Economic Forum survey see data quality and access to data, as well as access to suitable talent, as major obstacles to implementing AI, whereas another set of problems is related to exacerbation of bias and regulatory uncertainty and complexity [11], [16]. Insufficient infrastructure to accommodate new AI technologies and inadequate data quality to test and validate AI outcomes, as well as lack of appropriately skilled staff, are identified as the main barriers of wider AI adoption according to Deloitte's Digital Banking Maturity 2020 global benchmarking study performed across 318 banks in 39 countries on 5 continents [17], [18].

While promising, machine learning in fintech comes with a set of ethical issues that need to be considered. These ethical aspects have not been analyzed as thoroughly as in the context of machine learning applications in fields such as healthcare, where the ethical challenges of ML have already been addressed [19], [20], [21], [22]. In particular, a comprehensive framework has been created in [23] for identifying ethical issues in machine learning healthcare applications throughout all stages of product development from conception to implementation, including supporting processes such as evaluation and oversight.

While there are ethical guidelines established for the traditional financial services industry [24], including codes of ethics for professional associations [25], these have not been sufficiently evaluated in the context of machine learning applications for fintech. In this paper, we aim to take the main ethical principles defined in traditional finance as a basis for our study and analyze how the principles are compromised in the field of ML applications in fintech. We propose solutions to these problems using readily available error analysis and visualization tools for explainability, assessment, and diagnosis of ML models. We show how this approach can solve a number of ethical issues of ML applications in fintech. This paper intends to provide a multidisciplinary framework for using ML in finance that is not only restricted to computer scientists, but also targets financial institutions, fintech companies, regulatory bodies and decision makers.

The rest of the paper is organized as follows. Section II systematizes a set of ethical principles relevant to finance, namely integrity, objectivity, competence, fairness, confidentiality, and diligence, and discusses their fundamental importance to the financial services industry. Section III presents the principles and goals of explainable machine learning. Following the discussion in Sections II and III, we proceed by mapping the relationship between finance and ML ethics in Section IV. We conduct an experiment with a group of experts in finance and ethics to manually annotate the mapping between the principles of finance and ML ethics. The results are compared with mappings performed using NLP transformers, which show an overlap with the expert annotations. The explanation of NLP methods is comprehensive to be accessible to the wider audience. The results show that integrity and fairness exhibit the strongest relationships with ML ethics. Section V focuses on the ethical problems of machine learning in fintech. We treat topics such as biased data, accuracy, transparency, discrimination, differential pricing, manipulated recommendations, conflict of interest, violations of code of conduct, insider trading, data protection, and lack of skilled staff and discuss their potential consequences. Section VI explains the state-of-the-art (SOTA) tools that are used for explainable ML. Section VII focuses on a use case scenario where an ML model is used for credit card approval. We show not only how the proposed tools can help understand the ML decision in a finance context but also that both humans and machines could make mistakes in approving credit card requests, thereby emphasizing the need for a human-machine collaborating to improve the decision-making process. In Section VIII, we propose a conceptual framework for addressing ethical challenges in fintech such as bias, discrimination, differential pricing, conflict of interest, and data protection. Section IX concludes the paper.

II. ETHICAL PRINCIPLES IN FINANCE

In this section, we review the traditional core principles of ethics in finance. Based on analysis of 11 financial services professional associations, the study in [26] has distilled seven basic principles found in their codes of conduct: integrity, objectivity, competence, fairness, confidentiality, professionalism and diligence as described in Table 1.

A. INTEGRITY AND OBJECTIVITY

Acting with integrity is one of the main principles that underpins codes of ethics in finance. Ethics is tied to moral self-governance, autonomy, trustworthiness and honesty. Integrity means to set consistent thinking and conduct, to have good conscience and to adhere to acting responsibly.

As defined in [26], objectivity is ground on the subordination of the interests of the financial professionals to the needs and interests of the clients. Two elements that represent threats to objectivity are perpetual bias and conflict of interest. Bias reduces the ability to have accurate perceptions about the surrounding world and leads to faulty beliefs. Conflict of interest appears in situations governed by

TABLE 1. A list of ethical principles in finance and their definitions.

Ethical principles in finance	
Principle	*Definition*
Integrity	Moral self-governance, autonomy, trustworthiness and honesty. Consistent thinking and conduct, good conscience and responsible acting.
Objectivity	Protecting and advancing the interests of clients. Maintaining trust and accurate perceptions. Avoiding bias and conflict of interests.
Competence	Rendering competent financial services to clients. Maintaining expertise through continuous education and professional experience in the workplace.
Fairness	Treating customers equitably, consistently applying the "Golder Rule", ensuring fair returns to everyone, balancing interests and avoiding disparate treatment.
Confidentiality	Handling client relationships with confidence, protecting and not divulging sensitive information, building and maintaining trust through sharing information.
Professionalism	Treating clients with courtesy and respect, establishing confidence, maintaining reputation and trust with clients and the general public.
Diligence	Providing services promptly and thoroughly, rendering services tailored to the customer needs with attention to detail and persistent focus, thorough review of support staff.

compensations when professionals advance their personal or institutional gains contrary to the position of trust and related duties that clients expect from professionals. Both factors adversely affect the objectivity, integrity and public trust in the financial industry.

B. COMPETENCE AND FAIRNESS

Professionals are obligated to maintain their competence through continued education and experience with the goal to prudently service clients and protect their interests. Financial products are becoming increasingly complex and clients face a challenge of assessing the expertise of professionals and whether they are acting in clients' best interest. The inherent information asymmetry may lead to conflict of interest such that professionals exploit their expertise to gain advantage at the expense of clients. Another issue may occur if professionals attempt to handle activities beyond their scope of expertise, which may contribute to conflict of interest in monetary compensation. Professionals have the obligation to give advice within the domain of their expertise and defer other services to outside experts.

The principle of fairness is an integral part of the codes of ethics in the financial industry. Fairness is broadly defined through three concepts: treating customers equitably, offering financial advice that professionals would be comfortable applying to their own portfolios (Golden Rule), and allocating fair returns to everyone [26]. With regards to the concept of equality, any disparate treatment requires an explanation and justification to the affected parties. The Golden Rule assists professionals with clarifying their actions based on the best understanding of their own interests. The third concept is related to the obligation to properly balance the interests of all parties affected by certain decisions.

C. CONFIDENTIALITY

Confidentiality is the obligation to hold client information in confidence. When seeking financial advice, clients may share sensitive information about their finances and financial goals such as family dynamics. Financial services professionals should not divulge personal information due to the relationship of trust. There are four reasons that show the need for confidentiality: personal autonomy, respect for relationship obligations, client vulnerability, and serving the common good [27]. Personal autonomy acknowledges that clients have jurisdiction over their own personal information, and it is important that professionals respect the obligations arising from trust relationships. Trust and intimacy are built through sharing of personal information. Confidentiality is needed as clients become vulnerable by sharing personal information. Professionals are obliged to act in the best interests of their clients. Finally, as noted in [27], a system that respects confidentiality works for the public interest as well.

D. PROFESSIONALISM AND DILIGENCE

The principle of professionalism has three requirements: treatment based on respect and consideration, duty of professionals to maintain their reputation, and improving the quality of service provided to the public [26]. Regarding the first requirement, professionals should not treat clients as mere means to achieve their own goals as such treatment hampers clients' autonomy. Treating clients with courtesy and respect is the basis for protecting the interests of clients and also for establishing trust. The second requirement is needed because the success of the financial services industry is grounded in the public trust. Without trust, it is much more difficult to establish confidence between professionals and clients. Finally, assisting clients with making better financial decisions contributes not only to their financial security but also to the societal wellbeing. The reputation of the financial industry improves when its practitioners work toward the common goal rather than focusing on personal success.

The ethical principle of diligence can be interpreted in three ways [26]. Firstly, through providing services promptly and thoroughly to meet clients' expectations. Failure to do so undermines the trust between the client and the professional. Secondly, professionals are required to render services with due care which means to act with attention to detail and persistent focus throughout the process of working with a client. For financial services professionals, this means to carefully examine the needs of each individual client and give financial advice tailored to the circumstances of that client. Lastly, due diligence extends the obligation for thorough review of support staff.

With the development and increased application of ML in finance, there is a need to establish a correspondence between the traditional finance and novel approaches to ML ethics, which are described in detail in the next section.

III. PRINCIPLES AND GOALS OF RESPONSIBLE MACHINE LEARNING

With the rapid technological advancements and increased usage of machine learning, there is a necessity for creating standards for ML ethics. One of the prominent organizations that has developed such standards is the Organisation for Economic Co-operation and Development (OECD).[1]

The OECD has made a strong contribution in defining public policy for AI. In 2019, the OECD adopted a set of principles on artificial intelligence to promote AI that is innovative, trustworthy and respects human rights as well as democratic values [30]. The principles were adopted by OECD member countries by approving the OECD Council Recommendation on Artificial Intelligence [31]. The OECD AI Principles is the first such intergovernmental standard on AI. Non-member countries beyond OECD have also adhered to the principles. While the OECD Recommendations are not legally binding, they are highly influential as they set international standards to help governments design national legislation.

The OECD Recommendation includes two substantive sections. The first defines five fundamental and complementary principles for the responsible stewardship of trustworthy AI. These five principles are: i) inclusive growth, sustainable development and well-being; ii) human-centred values and fairness; iii) transparency and explainability; iv) robustness, security and safety; and v) accountability [31].

The second section proposes specific steps to governments to implement national policies and international cooperation aligned with the five principles. This includes i) investing in AI research and development; ii) fostering a digital ecosystem for AI; iii) shaping an enabling policy environment for AI; iv) building human capacity and preparing for labour market transformation; and v) international co-operation for trustworthy AI [31].

In this section, we investigate how the OECD principles are mapped to the previously discussed ethical principles in finance. The purpose is to use this mapping to qualitatively evaluate the relationship between the goals of explainable machine learning and the ethical challenges in fintech from an ML perspective.

A. PRINCIPLES

The OECD [30], [31] defines the AI ethical principles (Table 2) as follows:

Inclusive growth, sustainable development and well-being. This principle states that AI should be developed and used to increase prosperity for all - individuals, society, and the planet. It recognizes the potential of AI to advance inclusive growth and sustainable development in areas such as education, health, transport, agriculture, environment, etc. Stewardship of trustworthy AI should be accompanied by addressing inequality, risk of divides due to disparities in technology access, and biases that may negatively impact vulnerable or underrepresented populations.

Human-centred values and fairness. Based on this principle, AI should be developed consistent with human-centred values, such as fundamental freedoms, equality, fairness, rule of law, social justice, data protection and privacy, as well as consumer rights and commercial fairness. The principle recognizes that certain AI applications may have negative implications such as deliberate or accidental infringement of human rights and human-centered values. Therefore, the development of AI systems should be aligned with these values including the possibility for humans to intervene and oversee such systems.

Transparency and explainability. Transparency defined in this principle has two aspects. The first one is to disclose if AI is being used in an application so that users are aware of it. The second is to enable people to understand how an AI system works so that they can make informed choices. Explainability means enabling people affected by the outcome of an AI system to understand the system's decision. To achieve explainability, the system should provide easy-to-understand information to people affected by an AI system's outcome so that they can challenge the outcome, if needed. An explanation may involve providing details on the determinant factors behind a specific outcome or decision, or explaining why similar circumstances generated a different outcome.

[1] The OECD is an intergovernmental organization with 38 member countries founded in 1961 to stimulate economic progress and world trade. The majority of OECD members are high-income economies with a very high Human Development Index (HDI), comprising 62% of the global nominal GDP ($49.6 trillion) [28]. The OECD is an official United Nations observer. Together with governments, policy makers and citizens, the OECD works on establishing evidence-based international standards and finding solutions to a range of social, economic and environmental challenges. A significant part of the OECD activities focuses on defining public policies and international standards [29].

Robustness, security and safety. This principle states that an AI system must be robust, secure and safe throughout its entire lifecycle, and that its function does not pose safety risks. AI systems should be traceable and provide means to assess datasets, processes and AI-based decisions. Traceability ensures that outcomes of AI systems can be analyzed, and can provide responses to user inquiries about the outcomes. In this context, AI actors should apply a systematic risk management in the AI system lifecycle to continuously address risks, including privacy, digital security, safety and bias.

Accountability. According to this principle, organisations and individuals developing, deploying or operating AI systems should be held accountable for the proper functioning of the systems and for respecting of the OECD AI principles. Accountability should be in line with the applicable regulatory frameworks. Documentation on decision-making processes during the AI system lifecycle and the actions taken need to be maintained and available for an audit.

B. GOALS

The goals [32] of the explainable artificial intelligence (Table 3) are as follows:

Trustworthiness. Trustworthiness represents the confidence that the model will act as intended, which is not easily quantifiable. Trustworthiness is necessary, but not sufficient property of explainability since not every trustworthy model can be considered explainable [33], [34], [35], [36], [37].

Causality. Causality refers to the goal of finding causal relationships among the data variables of a model. Explainable models might make the task of finding such relationships easier, but inferring causality requires a wide frame of prior knowledge. ML models can discover correlations in the data, however correlation does not imply causation. An explainable ML model could use the observed data to validate the results with causality inference techniques, or provide intuition of possible causal relationships [38], [39], [40], [41], [42].

Transferability. Transferability is the ability to understand a model (its assumptions and implementation) in order to facilitate model reuse in another problem. Lack of proper understanding can lead to incorrect assumptions and conclusions [33], [43], [44], [45], [36], [46], [47], [48].

Informativeness. Informativeness is related to the ability of ML models to give information about the problems being tackled and the decisions being made. While the main objective of ML models is to support decision making, the results obtained by ML models may not be the same as the decisions taken by a human. Therefore, distilling information about the inner-workings of ML models is an important goal for achieving explainability [33], [44], [45], [47], [49].

Confidence. ML models are expected to be reliable and the confidence in the model reliability is essential. The trustworthiness of model interpretations depends on whether a model is reliable. Thus, maintaining confidence in the working regime of a model is an important factor for assessing the usefulness of the model [33], [38], [46], [50], [51], [52], [53].

Fairness. Without explainability it is not possible to assess the fairness of ML models. Model explainability is achieved through visualization of the relations affecting the model results. Making the results visible helps avoid unfair use of the model results [33], [35], [38], [46], [54], [55], [56], [57].

Accessibility. Accessibility facilitates the involvement of end users in the process of developing, improving and monitoring ML models. Accessibility will ease the burden of non-technical or non-expert users when using AI systems and algorithms seemingly incomprehensible at first sight [36], [44], [45], [47], [48], [58].

Interactivity. Interactivity allows end users to assess and test explainable ML models. Interactivity can also serve as a tool for improving AI explainable models. This is relevant to fields in which end users need to have ability to interact with the models and to modify them [48], [58], [59], [60], [61].

Privacy awareness. The ability to assess privacy is one of the byproducts enabled by model explainability. ML models may have complex inner-workings, and not knowing how the model's results are represented internally may lead to a privacy breach. In addition, explaining the inner-relations of a trained model to non-authorized third parties may also compromise privacy [62].

To harness the potential of the novel approaches to ML ethics, we explore the correspondence between ML ethics and traditional principles of finance ethics.

IV. MAPPING BETWEEN FINANCE ETHICS AND ML ETHICS

The previous two sections make a broad overview of the principles of finance and ML ethics. While finance ethics is well established, ML ethics has witnessed an increased interest only recently due to the proliferation of ML-based solutions in finance. The contribution of this paper is in studying the relationship between finance and ML ethics with the goal of minimizing the adverse impact of ethical issues in fintech. The purpose of this study is to identify the most important criteria to consider when addressing ethical challenges in ML-based fintech applications. The results can help fintech companies in building products and services by considering the most relevant ethics principles.

A. MAPPING METHODOLOGY

To evaluate the relationship between finance and ML ethics, we conducted an experiment with a group of experts in finance and ethics to manually annotate the links between the ethics principles based on their definitions.

The group is composed of 8 experts from the academic community with expertise in finance and ethics who are also knowledgeable in machine learning. They are chosen carefully to ensure they tackle effectively the task of manually annotating the links between the ethics principles. Each of the experts received both the long and short definitions of ML ethics and finance ethics to assess the mapping between the principles. Each of the experts worked individually on the mapping. After the process was completed, the results were

TABLE 2. OECD principles of artificial intelligence.

OECD AI Principles	
Principle	***Definition***
Inclusive growth, sustainable development and well-being	Trustworthy AI should contribute to overall growth and prosperity for all – individuals, society, and planet – and advance global development objectives.
Human-centered values and fairness	AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and should include appropriate safeguards to ensure a fair and just society.
Transparency and explainability	Transparent and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.
Robustness, security and safety	AI systems must function in a robust, secure and safe way throughout their lifetimes, and potential risks should be continually assessed and managed.
Accountability	Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the OECD's values-based principles for AI.

TABLE 3. A list of goals in explainable machine learning and their definitions.

Goals of explainable machine learning	
Goal	***Definition***
Trustworthiness	Confidence that an ML model acts as intended when faced with a given problem.
Causality	Finding causal relationships among the data variables of an ML model.
Transferability	Ability to understand a model and reuse it in a different problem.
Informativeness	Extracting information from an ML model to understand the model decisions.
Confidence	Ability to assess the confidence in the model reliability.
Fairness	Visualizing the relations affecting the model results, highlighting bias, avoiding unfair or unethical use.
Accessibility	Involving end users in the development, improvement and monitoring of ML models; reducing complexity for non-technical or non-expert users.
Interactivity	Enabling end users to interact with ML models and adjust them to meet certain requirements.
Privacy awareness	Ability to assess privacy aspects of ML models and avoid privacy breaches.

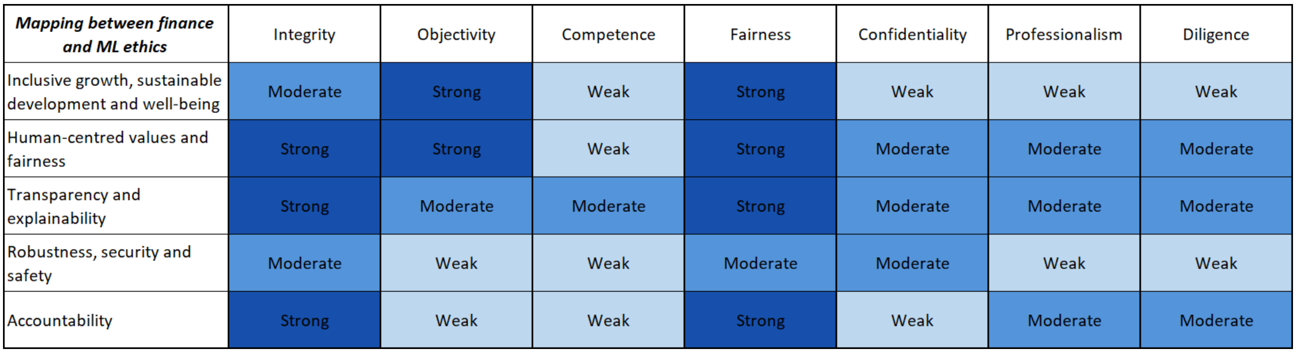
collected and reported in Fig. A.2 (column HA) based on majority voting. Our objective is to use expert knowledge to get insights into the ethical principles of finance which exhibit most influence on ML ethics and vice-versa.

We use P_f to denote a definition of an ethical principle in finance. Similarly, we use P_{ML} to denote a definition of an ethical principle in ML. The comparison consists of evaluating how strong the mapping is between P_f and P_{ML}. For a given pair (P_f, P_{ML}), the mapping can reveal weak, moderate or strong relationship depending on how much one principle is related to the other, which defines the strength of the link for that pair.

In order to make the actual comparison, the human annotators used the seven well-known financial principles defined according to common codes of ethics of financial organizations and institutions as explained in Section II [26]. In addition, they took the definitions of the five established AI principles adopted by OECD as a basis for ML ethics. Both the long and short definitions of finance ethics and ML ethics were used, i.e. all pairs (P_f, P_{ML}) were considered where P_f and P_{ML} can represent either long or short definitions. The long definitions of finance ethics are given in the respective paragraphs of Section II, whereas the long definitions of ML ethics are defined as per the OECD principles. The short definitions are obtained as summaries of the long definitions and can be found in Tables 1-2.

The results provided by the experts are presented on Fig. A.2 in the Appendix A. The abbreviation HA denotes columns with results obtained from human annotations. To enhance objectivity and improve decision making, we have assessed the experts' results using recent advancements in natural language processing (NLP) that led to substantial improvement in certain tasks, almost comparable with human performance.[2] One such task is semantic text similarity where SOTA results are obtained using NLP transformers. The columns denoted as LD and SD on Fig. A.2 refer to results obtained via the NLI-DistilRoBERTa-Base-v2

[2] The code and dataset for the comparison experiment can be found at: https://github.com/rizinski/Ethics-in-Finance-and-Machine-Learning/tree/main/transformers_notebook



Mapping between finance and ML ethics	Integrity	Objectivity	Competence	Fairness	Confidentiality	Professionalism	Diligence
Inclusive growth, sustainable development and well-being	Moderate	Strong	Weak	Strong	Weak	Weak	Weak
Human-centred values and fairness	Strong	Strong	Weak	Strong	Moderate	Moderate	Moderate
Transparency and explainability	Strong	Moderate	Moderate	Strong	Moderate	Moderate	Moderate
Robustness, security and safety	Moderate	Weak	Weak	Moderate	Moderate	Weak	Weak
Accountability	Strong	Weak	Weak	Strong	Weak	Moderate	Moderate

FIGURE 1. Mapping between finance and ML ethics. The intensity of the color represents the strength of the relationship.

transformer using long definitions and short definitions, respectively. The strength of the links is mapped with three color intensities to denote a strong, moderate and weak relationship between the corresponding ethical principles. While our focus relies on human expertise, we also demonstrate that the human-centric results are also aligned well with the transformer results. Although certain differences exist, the analysis shows that overall there are major overlaps in the two approaches. These overlaps can be helpful in enhancing objectivity by confirming the experts' mapping between finance and ML ethics.

Transformers are novel architectures in NLP for sentence encoding that employ techniques of attention to handle long-range dependencies in textual data, thereby solving challenges that were not possible with older models such as recurrent neural networks (RNNs) [63], [64], [65]. Transformers led to a breakthrough in NLP in recent years as they have demonstrated outstanding performance in a wide range of tasks such as machine translation [66], [67], [68], question answering [69], [70], [71], [72], sentiment analysis [73], [74], [75], [76], name entity recognition [77], [78], [79], [80], extractive and abstractive document summarization [81], [82], [83], [84], among others.

The main essence of transformers is that they can encode any text into a vector representation that can be then fed into a machine learning model for further analysis. One such application is assessing the semantic similarity between two texts such as two sentences or two paragraphs. We use the cosine similarity (i.e. normalized dot product) which is well suited to compute their semantic similarity for texts encoded into vectors [85]. The use of cosine similarity is viable because the NLP transformers are already pre-trained models, used for zero-shot learning. Thus, there is no need to split the dataset into a training and validation set.

The dataset (P_f, P_{ML}) for our experiment consists of two parts: (P_f^l, P_{ML}^l) for the long definitions, and (P_f^s, P_{ML}^s) for the short definitions. For each pair of ethical principles, we use transformers to convert the definitions into vector representations, and then calculate the cosine similarity between them. The calculations are repeated for both parts of the dataset. For determining the strength of the links, i.e. whether they reveal weak, moderate or strong relationship, we use the following approach. For each transformer, we calculate the 33.33% and 66.66% percentiles obtained from the set of cosine similarities for all pairs of principles for that transformer. Then, for each pair of principles, we check if the cosine similarity for that pair is less than the 33.33% percentile, less than 66.66% percentile, or higher than the 66.66% percentile. Depending on the comparison with these thresholds, the link for that pair is labeled as weak, moderate or strong, respectively. The reason for analyzing both the long and short definitions is to get insights into the links between the principles from two related perspectives with the goal of assessing the level of overlap between the two sets of results.

For the experiment, we used the NLI-DistilRoBERTa-Base-v2 model from Hugging Face [64], [86]. RoBERTa is chosen as it showed superior performance within the transformers analyzed in [76] on sentiment tasks in finance. Fig. A.2 presents the results obtained from the transformer experiment for both long and short definitions, and demonstrates overall alignment with the manually annotated mappings. In the following subsection, we discuss the overall insights obtained from the mappings.

B. INSIGHTS FROM THE FINANCE-ML ETHICS RELATIONSHIP

Fig. 1 unifies the experimental results of Fig. A.2 using the majority rule. We observe that integrity and fairness have a strong relationship with ML ethics across all finance principles. Similarly, human-centered values and fairness as well as transparency and explainability exhibit strongest relationship with finance ethics among all ML principles. Such conclusion is overall valid for both experimental approaches, i.e. with transformers and with human annotations. This comes at no surprise as integrity and fairness are essential principles in finance ethics. Therefore, the most important criteria for handling ethical challenges in ML-based fintech applications is to ensure integrity and fairness as well as transparency and explainability of the used ML systems while respecting human-centered values and accountability. Another general conclusion is that the computational results using transformers overall show a good agreement with the manually annotated mappings, validating the potential of transformers to achieve human-level performance in NLP tasks. Taking into account the ML ethics principles, the fintech companies could improve their products and services.

V. ETHICAL CHALLENGES IN FINTECH

In Sections II and III, we presented the ethical principles in finance and ML, while in Section IV, we established their relationship to develop a framework for responsible fintech applications. In reality, it is not trivial to apply the ethical principles in fintech. In this section, we identify 12 distinct challenges in applying the previously defined ethical principles.

As discussed in [87], finance ethics in general represents a subset of general ethics which is guided by norms such as truthfulness, honesty, integrity, respect for others, fairness and justice; however, while ethical norms guide human behavior in societal interactions, situations may arise in which the need to care for ourselves is in conflict with the need to care for others. As pointed out in [87], incompatibility arises from the fundamental assumption in the modern capitalist system that greediness drives profitability. Maximizing own interests in principal-agent relationships leads to numerous examples of ethical issues and violations of trust and loyalty. In this section, we review ethical challenges in fintech that arise from the perspective of machine learning, as summarized in Fig. 2.

A. BIAS, ACCURACY AND TRANSPARENCY

As financial decisions depend on data, it is essential to have representative data to train ML models when offering financial and investment services; if the data is not representative, there is a high likelihood that, in general, models will perform unsatisfactorily [88]. During data acquisition, it is crucial to understand both the source of the data and the data governing rules and regulations [89]. In this context, we face the challenge of selecting appropriate data sources. As the domain of data collection is vast, any news could theoretically have an impact on financial advice. Therefore, it is important to define impartial criteria to identify relevant data sources. The difficulties do not end here though. In fact, they are further intensified knowing that the collected data itself may be biased. Since ML models are typically designed to operate autonomously, it is difficult to check for bias unless data is verified manually by human intervention. However, this is not only a labor-intensive task but also practically infeasible given the potentially huge volume of data that needs to be checked. Collaboration with subject matter experts is essential when developing data and methods to avoid conclusions made on faulty assumptions using insufficient or biased data [90]. Ensuring proper data acquisition and use is necessary to avoid ethical problems. Even if certain information is publicly available, it can still pose ethical problems if used improperly [89].

Investors carefully select whom to trust with their investment decisions. There is a distinction between a brokerage and an investment advisory firm. Brokers engage in the business of effecting transactions in securities for the account of others, for which they receive compensation. When brokers recommend securities to their clients, they must ensure that the investment is ‘‘suitable’’ for the client. On the other hand, investment advisors advise others about investing in securities and receive compensation for the advice. When investment advisers recommend an investment to their clients, the investment needs to be in ‘‘the best interest’’ of the client. These differences are essential and create two different standards of conduct: i) suitability for brokers and ii) fiduciary (‘‘best interest of the customer’’) for investment advisers. Investors should know the difference, especially when the investment advice is selected based on an ML algorithm. It is challenging to understand whether the AI-based investment decision is made because it is ‘‘suitable’’ or in ‘‘the best interest’’ of the client. These questions are at the center of the Securities and Exchange Commission (SEC) regulatory discussion about the distinction between best interest and fiduciary duty and should be considered when developing ML-based investment algorithms [91].

If we assume that an ML model is fed with comprehensive and unbiased data collected from relevant and reliable sources, it may still be challenging to select an existing model or develop a new model having a level of accuracy that is sufficient for solving a particular problem at hand given specific client circumstances. One such example is predicting stock returns for optimizing investment decisions. Existing models may not entirely correspond to the requirements needed for tackling the problem domain and may have to be adjusted. Prior to improving these models, there may not be a clear indication that the adjustments would work well. Similarly, new models may need to be created to achieve an adequate fit with the problem domain. Modeling phenomena involving human beings often requires simplification that can mask risks for the sake of precision; over-reliance on probability and statistics can be a limiting factor as economics is a social and not a natural science [92]. Looking at this problem pragmatically, it is not only unclear whether the new models would accomplish better performance; development of such models may necessitate significant efforts that could delay rendering financial services with acceptable level of diligence, i.e. in a prompt and thorough manner. Consequently, the financial professional dilemma is deciding which model to choose and how to present its accuracy to the clients.

ML-driven solutions suffer from lack of transparency which can lead to numerous issues when applied in finance. This is particularly problematic in the case of deep learning models, which have become increasingly popular [93].

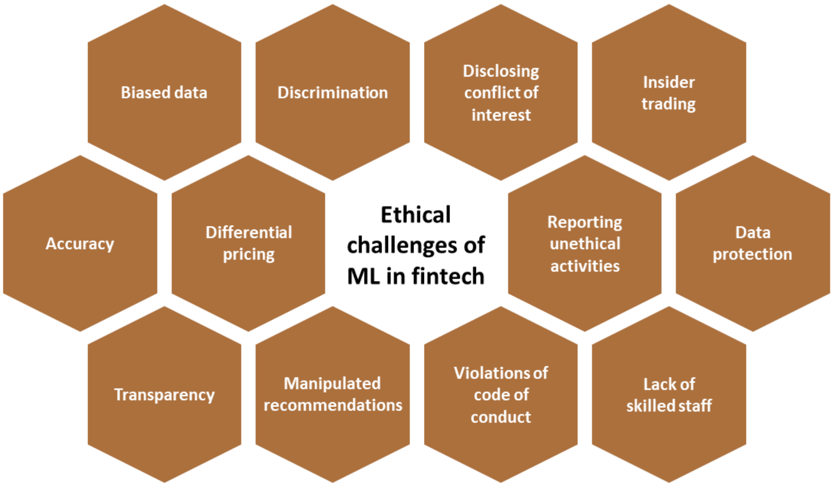

Biased data
Discrimination
Disclosing conflict of interest
Insider trading
Accuracy
Differential pricing
Ethical challenges of ML in fintech
Reporting unethical activities
Data protection
Transparency
Manipulated recommendations
Violations of code of conduct
Lack of skilled staff

FIGURE 2. **Ethical challenges of machine learning in fintech.**

These models use hidden layers for decision making, thus resembling a black box where the internal inner-workings cannot be precisely evaluated, even by experts. It is difficult to gauge such systems for precision, methodological validity or financial risk estimation. Moreover, it can cause regulatory compliance issues and even financial loss for businesses.

B. DISCRIMINATION, DIFFERENTIAL PRICING AND MANIPULATED RECOMMENDATIONS

Discrimination can lead to unfair practices for certain groups of the broader population. For example, access to financial services can be difficult for vulnerable groups of clients who have historically been subject to discrimination based on race, political or religious beliefs, sexual orientation, health problems or low income [94]. For example, a possible consequence could be discrimination in assessing credit risk and lending decisions. In this context, broader ethical issues can arise such as employment discrimination, discrimination based on physical or mental disabilities as well as service price discrimination [95]. The problem can be further deepened by introducing biased decisions in deep learning models based on the learning process, which may contribute to unethical decisions that cross the legislative framework [93]. Risks of discriminatory outcomes or perpetuation of existing socioeconomic disparities are also at the focus of the US Federal Trade Commission whose recommendations for businesses are based on consumer protection by using AI tools that are ''*transparent, explainable, fair and empirically sound, while fostering accountability*'' [96].

An inadequate policy of differential pricing can cause troubling ethical consequences if clients are charged with varying rates based on their race or any other characteristic that does not contribute to the value proposition of the rendered financial service [26]. Differential pricing in itself is not necessarily an ethical violation; for example, clients could be charged more if they are more demanding and thus require more effort by the provider of financial services. Similarly, low maintenance clients would be charged less since providing services for them is not as complex as in the case of high maintenance clients. If financial services professionals have more experience and education, this would constitute a fair component in defining the pricing structure of their services. However, factors that do not define the value proposition may violate ethical norms.

A fundamental premise that conditions clients' confidence in the advice they receive by ML models is the promise that models are designed ethically to maintain objectivity. ML models should be based on harnessing evidence available in collected data and applying state-of-the-art algorithms to ensure consistent and equitable treatment of clients. An objective model does not deliberately promote a set of input data at the expense of other input data. However, a model can be intentionally tailored to recommend a set of actions to specific clients without evidence that these recommendations are justified. A problem can arise if the model is injected with the ''right'' amount of bias to demonstrate that it has ''superior'' results, which would ultimately mislead clients [90].

C. DISCLOSING CONFLICT OF INTEREST, REPORTING UNETHICAL ACTIVITIES AND VIOLATIONS OF CODE OF CONDUCT

A significant part of the activities taken by financial services professionals is to render recommendations to their clients. Giving honest and accurate recommendations can be jeopardized if the professional has material conflict of interest that is disparate with the interests of the clients. When people receive advice regarding financial decisions, they need to know if their adviser has any interest that hinders their

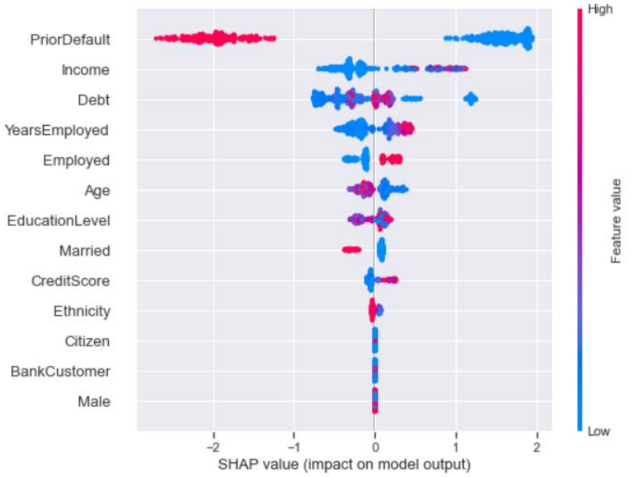


FIGURE 3. SHAP beeswarm plot explaining XGBoost on the credit card approval dataset.

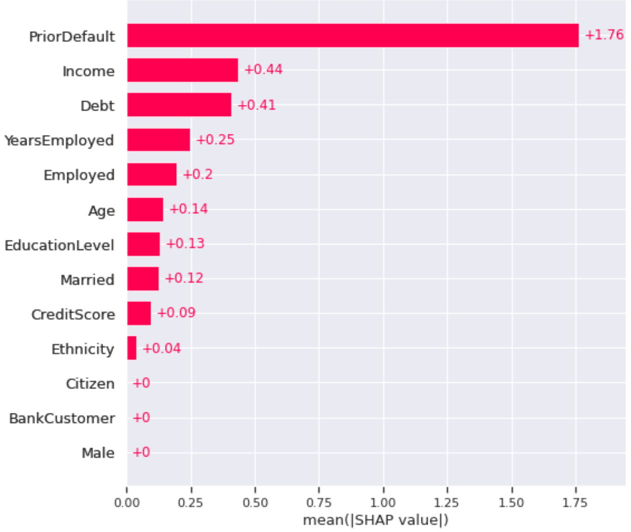


FIGURE 4. SHAP bar plot displaying global feature importance for XGBoost on the credit card approval dataset.

incentive to offer appropriate recommendations. Having such conflict of interest, but not disclosing it when dealing with clients, contradicts the principle of fairness and may lead to problems that can adversely affect the portfolio performance of the clients. ML model performance has been typically evaluated using single-valued metrics, but this does not provide insights into the distribution of errors across datasets or across model features. This can lead to masking potential conflict of interest unless the model is transparently evaluated for different components of the input data or different sets of features. If clients have access to the details of ML models in a transparent and explainable way, risks related to conflict of interest could be reduced.

Preserving ethics standards should be an industry-wide effort and merely not being involved in unethical activities is not sufficient to guarantee good ethical behavior [97]. If one possesses information about illegal or unethical actions by peers or other actors in the finance industry, it is their responsibility to report these activities up the hierarchy in the organizations and the respective regulatory bodies. Otherwise, financial professionals are subject to legal repercussions. However, this is not always possible as such reporting can irreversibly damage the relationship between the involved parties, so people tend not to report their peers or colleagues. The development of ML models without transparency and quality control could further contribute to unethical behavior because it will be impossible for companies to evaluate the objectivity of the ML models.

Organizations such as corporations, associations and institutions often develop codes of conduct to guide the behavior of their members. Violations of codes of conduct in the financial industry leads to improper dealings that harm investor interests and the market stability as a whole. There exist organizations that oversee financial experts' behavior to verify it is in accordance with established laws. For instance, in the U.S., the official regulatory agency that implements securities laws and establishes regulations for proper conduct is the SEC [87]. With the technological advancements and investment model implementations in the financial industry, it is becoming increasingly difficult to monitor compliance of ML algorithms with established regulations. This challenge could be overcome with appropriate training for regulators to be more technologically-savvy and competent in detecting violations of code of conduct induced by ML systems.

D. INSIDER TRADING, DATA PROTECTION AND LACK OF SKILLED STAFF

Insider trading is an unfair market practice in which market participants trade based on material non-public information to generate extraordinary gain [87]. Material information is defined as information that will change the investment behavior of a rational investor when they obtain the information. Hence, this information will affect the stock price behavior. Insider trading is harmful because it undermines the trust of investors in the financial markets, and creates an unfair trading environment. Given the damage that insider trading can create, the SEC prosecutes insider trading violations as one of its enforcement priorities. Deterring people from exploiting insider trading includes disqualification from acting in certain fiduciary positions for life or limited period of time, money fines, and prison sentences [98].

As fintech services are dominantly based on information technology, clients of financial services companies are more prone to breaches of privacy [89]. Throughout their work with customers, financial professionals using ML systems can gradually obtain a considerable amount of information about their customers such as personal data, financial data, online behavior and user preferences. Subject to the specific terms and conditions of use, this data can potentially be sold to third parties for profit. There is an entire array of parties interested in getting access to the data – ranging from advertisers, research agencies, political organizations (e.g. political campaigns), and law enforcement. As a result, this raises serious concerns about safeguarding privacy that can undermine the clients' confidence. At the same time, companies want to protect their clients' data to avoid lawsuits, which can be regarded as the primary deterrent for unethical practices [90].

Notwithstanding the advances achieved in the field of ML research, it is equally important to have skilled staff to implement or interpret the results obtained from machine learning [93]. A survey of the top 1000 firms in the U.S. on AI implementation in their firms found that their biggest concern in the implementation of AI was the readiness and ability of staff to understand and work with these new solutions [99]. Using ML models to render financial services requires professionals with the adequate skills. Therefore, it is essential to train staff, which requires additional investment of time and resources unless the work is outsourced to providers of ML solutions [93].

The recent example with the real-estate company Zillow demonstrates the difficulties faced when deploying ML models for productive use [100]. In an attempt make it convenient to sell homes while minimizing in-person interactions during the pandemic, the company created a new service called ''Zestimate'' which uses ML algorithms to estimate initial cash offers to home owners to purchase their listed properties. The service determined prices that are higher than the company could use to resell the properties after the necessary repairs. Thus, the company sold significantly fewer properties than initially expected due to the fast changes in the real estate market and lack of data. The financial consequence of using Zestimate were so profound that only eight months after launching the service Zillow decided to entirely close down that business. The company took a $304 million inventory write-down, leading to major stock price declines and plans to cut 2,000 jobs (25% of its staff). This case shows how critical it can be to use ML in an ever-changing market environment, especially when ML cannot leverage sufficient data to make accurate predictions.

VI. TOOLS FOR RESPONSIBLE MACHINE LEARNING

Recent advancements in explainable and responsible AI can help address challenges described in the previous section. As machine learning systems become ubiquitous, having a significant impact on society, there is an increased demand for ML models that can be explained. Since the inner-workings of ML models are difficult to asses given their resemblance to ''black boxes'', it is hard even for experts to interpret the predictions of the models. Therefore, it is essential to study the models thoroughly and interpret them well. ML model interpretability is currently an active research area with the goal of increasing model transparency [38], [101], [102]. Deploying ML models in practical applications has to be accompanied by a rigorous performance evaluation [103].

The model fairness problem is often related to selecting the right metric for benchmarking. In many cases, the benchmark is merely based on a single aggregate metric, such as accuracy, for the entire dataset [103]. However, this makes it difficult to understand how an ML model performs on various dataset partitions. The issue with such a single-valued metric is that even though most of the partitions of the input data may perform well by meeting the required benchmark, there could still exist non-negligible regions of data for which the model's predictions may render considerably different results. While the ML model may perform satisfactorily when averaged over the entire dataset, the discrepancies for certain regions can lead to ethical issues such as bias, inaccuracy, unfairness, discrimination, etc. Furthermore, using an aggregate metric makes it difficult to continuously monitor the model behavior when new data is collected.

To address this problem, the dataset is sliced into a one-dimensional or two-dimensional grid of input features and each cell of the grid is separately evaluated against the selected metric. For example, if the metric is the error rate, by analyzing the grid it is easy to visualize how the errors are distributed across various parts of the dataset. The visualization can be aided by heatmaps to color cells, e.g. using a darker color, if they exhibit higher inaccuracies [103]. This visualization technique emphasizes the problematic regions of data that suffer from model inconsistency and are difficult to evaluate by using an aggregate, single-number metric. Thus, by offering a deeper view of the model behavior, two goals are achieved: (i) ability to visually identify performance problems and (ii) gaining better insights, useful for performing model debugging. Both goals improve the interpretability of ML models and their responsible use.

To help the ML community accelerate model development, visualization dashboards for error analysis and explainability are developed as open source software. One notable example is the Responsible AI Widgets repository [104], which is a collection of model and data exploration and assessment user interfaces that enable better understanding of AI systems. Its purpose is to assist developers and stakeholders of AI systems in developing and monitoring AI more responsibly. The Responsible AI Widgets allow developers to interpret models by assessing errors and fairness issues [105], [106].

The explainers in Responsible AI Widgets are implemented based on SHAP (<u>SH</u>apley <u>A</u>dditive ex<u>P</u>lanations), which is considered state-of-the-art technique for ML explainability [107]. Its approach uses Shapley values from game theory to explain the output of ML models [108]. SHAP evaluates the contribution of each feature to the model predictions and assigns each feature an importance value, called a SHAP value. SHAP values are calculated for each feature across all samples of the dataset in order to assess the contribution of individual features to the model's output [109]. This means that a feature can have the same SHAP values for different samples, but their contributions toward the model's output can be different if the values of other features for those samples are different. SHAP values are also additive, thereby classifying SHAP as an additive feature importance technique. In other words, using SHAP values as a measure helps quantify marginal contributions of features to ML model predictions.

For the use case presented in the following section, we employ SHAP and Responsible AI Widgets to demonstrate the benefits of employing such tools for addressing ethical challenges in fintech. Our goal is to increase awareness within the financial industry regarding the possibilities of ML explainability. There are other tools available as open source software libraries that are aimed at ML model interpretability and explainability. While we do not cover all such tools in this paper, the interested readers may refer to AI Explainability 360 (AIX360) [102], [110], LIME [36], DeepLIFT [111], [112], [113] and What-if Tool (WIT) [114], [115].

The above tools focus on explaining ML models in the post-training phase. Specific aspects related to ML models' responsible use, such as model de-biasing, can be addressed even in the training phase. The work presented in [116] proposes an algorithm for flexibly fair representation learning by disentangling information from multiple sensitive attributes. They show that their flexible and fair variational encoder, which does not require the sensitive attributes for inference, is flexible with respect to downstream task labels and sensitive attributes. Since both training and post-training diagnosis are essential, ML experts and financial professionals should work together to perform the necessary due diligence throughout the entire lifecycle of ML applications in fintech. In addition, there are various ways of measuring ML model fairness across different regions of data that go beyond the typical approach of binning the data and assessing error rates for different data cohorts. A detailed interpretation of some of the most widely used fairness metrics and their mutual relationships is presented in [117].

A. LIMITATIONS AND CHALLENGES OF RESPONSIBLE ML

Despite advances in the explainable ML that resulted in its increased popularity, certain limitations of the explainable algorithms and tools need to be considered when applied in industrial applications [118].

First, due to the lack of general theories that allow quantitative analysis in complex areas such as healthcare or finance, statistical models and machine learning methods are introduced to solve those problems. Explaining the solutions to those complex problems, even with a perfect version of an explainable ML model, can be challenging because of the complex nature of the problem itself. Regardless of the potential success of such models, this implies that an ordinary user – not trained in the complex theory related to the area being modeled – could not understand nor interpret such a theory correctly and, hence, from their perspective, the results from the explainable ML would appear opaque and completely incomprehensible [119]. Second, the explainable models depend on the AI system and the available data that, in some cases, can be imperfect or limited. In this case, it is evident that the knowledge derived from such a system is further restricted in terms of the answers that can be offered and the number of explanations that can be provided [119].

Reference [120] developed a method to compute small adversarial perturbations (equivalent to creating adversarial instances for neural networks) that resulted in significant modifications to the feature importance of several explainable methods. Similar findings for the SHAP method were presented by [121]. However, in the case of real-world applications such as healthcare or finance, such adversarial perturbations and their influence on explainable ML approaches need to be further investigated [118]. The results from [122] suggest that the mathematical correctness that underpins SHAP is not sufficient alone; it should also be aligned with the specific use case and the human-centric understanding of SHAP's quality of explanations.

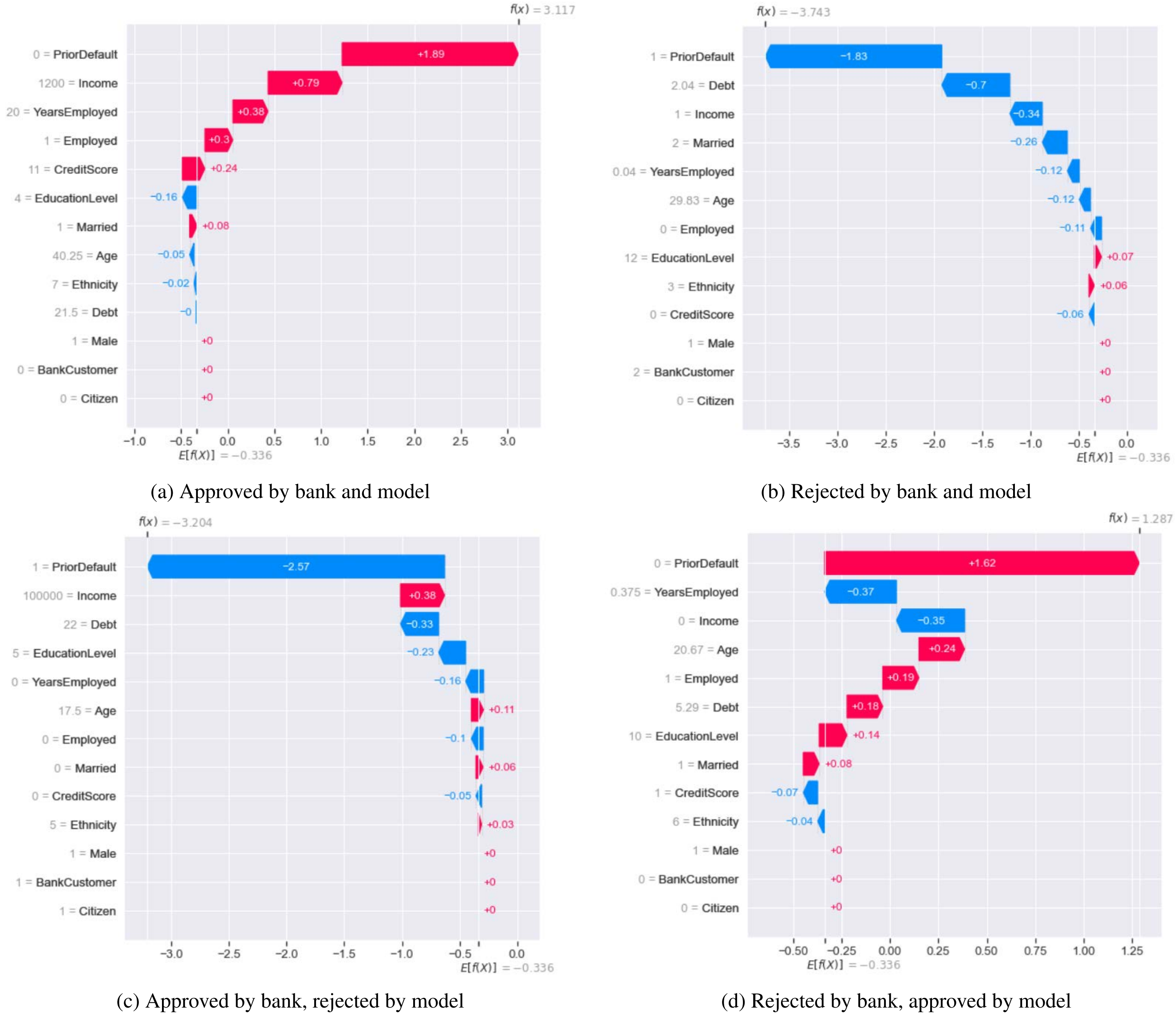
One of the possible solutions to some of the mentioned challenges is that the most powerful but opaque ML systems (e.g. deep learning) should not be preferred and applied by default, but a comparable alternative that is less powerful but inherently explainable can be employed. In general, the inherently explainable ML models should be adopted because of their transparency and explainability, while black-box models with model-agnostic explainability can be more difficult to defend under regulatory scrutiny [119]. Based on this, ML practitioners and financial professionals should be aware that responsible ML is applicable in settings where it can admit clear interpretation. This argument is also in line with the overarching understanding of the proposed ethical framework that human-machine collaboration is essential for addressing model explainability and transparency in general.

VII. USE CASE: DIAGNOSING CREDIT CARD APPROVAL PREDICTIONS WITH RESPONSIBLE ML

In the previous section, we presented tools for responsible ML that can address ethical challenges. Here we demonstrate a use case showing how to respect ethical principles while applying tools for responsible ML to address the challenges arising in fintech.

A. SETUP OF THE EXPERIMENT

In this section, we consider a fintech application for approving credit card requests based on machine learning predictions. We use the Credit Approval Dataset from the UCI Machine Learning Repository [123]. The dataset contains



(a) Approved by bank and model

(b) Rejected by bank and model

(c) Approved by bank, rejected by model

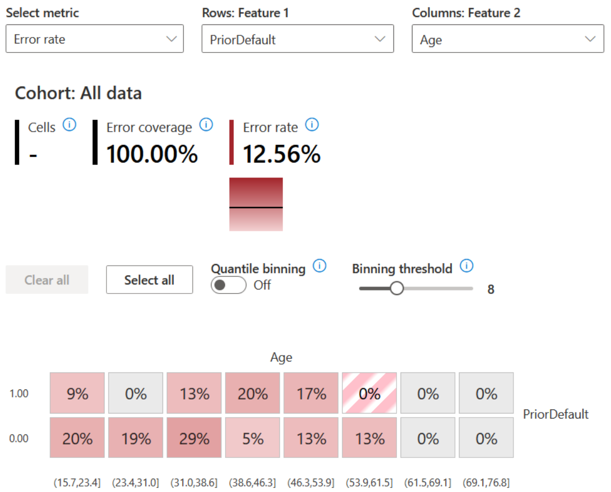
(d) Rejected by bank, approved by model

FIGURE 5. Local explainability with SHAP for predicting credit card approvals: (a) approved by human decision and model prediction, (b) rejected by human decision and model prediction, (c) approved by human decision and rejected by model, (d) rejected by human decision and approved by model.

690 instances and exhibits a mixture of features with categorical, integer and real values. The feature names and values of the dataset have been anonymized in order to protect the confidentiality of the data. However, a good overview of the probable features is given in [124] which lists 15 feature names (Gender, Age, Debt, Married, Bank Customer, Education Level, Ethnicity, Years Employed, Prior Default, Employed, Credit Score, Drivers License, Citizen, ZipCode, and Income) and one class (Approved). The dataset is well balanced: about 44.5% and 55.5% of the credit card requests are marked as approved and rejected respectively. Having a balanced dataset is a prerequisite for appropriate training of classifiers in supervised learning problems so that classifiers can perform effectively.

As a preparation for our experiments,[3] we perform standard data processing steps. First, we audit the dataset for missing values and find that less that 37 cases (5%) have one or more missing values. We impute the missing data with the mean value for each numerical feature. We replace missing categorical data with the most frequent value for each categorical feature. We then use a label encoder to convert all categorical values into numerical types, after which we remove the Drivers License and ZipCode features from the dataset as they are unlikely to have tangible impact on the predictive performance. Finally, we split the dataset into

[3] The code and dataset for the explainability analysis can be found at: https://github.com/rizinski/Ethics-in-Finance-and-Machine-Learning/tree/main/explainability_notebook



Select metric: Error rate

Rows: Feature 1: PriorDefault

Columns: Feature 2: Age

Cohort: All data

Cells: -

Error coverage: 100.00%

Error rate: 12.56%

Quantile binning: Off

Binning threshold: 8

	Age								
1.00	9%	0%	13%	20%	17%	0%	0%	0%	PriorDefault
0.00	20%	19%	29%	5%	13%	13%	0%	0%	
	(15.7,23.4]	(23.4,31.0]	(31.0,38.6]	(38.6,46.3]	(46.3,53.9]	(53.9,61.5]	(61.5,69.1]	(69.1,76.8]	

FIGURE 6. **Error heatmap for identifying regions in the dataset with higher errors. Age is on the horizontal axis and Prior Default is on the vertical axis, with 1 meaning that there was a prior default, and 0 meaning that there was no prior default.**

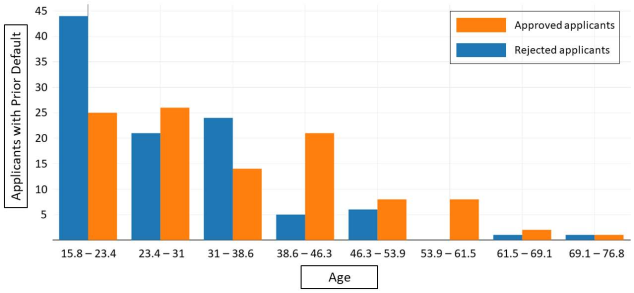


FIGURE 7. **Data cohort diagnosis for understanding error rates. Applicants rejected by the model are marked with blue, while approved applicants are marked with orange. Applicants of age 46.3 or above exhibit lower error rates.**

70% train and 30% test subsets. For the data processing steps, we use standard Python packages (numpy, pandas and scikit-learn).

Predicting whether a bank approves or rejects a client request for issuing a credit card is a binary classification problem. For this classification problem, we use the eXtreme Gradient Boosting (XGBoost) algorithm [125] to make predictions on the dataset. There are two reasons for selecting XGBoost among other classifiers. First of all, XGBoost has demonstrated its viability in industry-grade systems due to its computational speed, generalization capabilities, high predictive performance, and ability to work well on structured data in a variety of settings, including regression, classification and ranking problems. XGBoost is also a very popular tool among data scientists as it has won numerous Kaggle competitions [126]. Secondly, XGBoost is considered a black-box model. Despite being a high-performance algorithm, XGBoost's decisions are hard to interpret, which makes it suitable for our explainability analysis from a fintech perspective. Before selecting XGBoost, we compared it with another popular algorithm, Light Gradient Boosting Machine (LGBM), which resulted in a similar accuracy score on the same dataset. As part of this study, we also considered other approaches to create predictive models, including deep neural networks. However, our intention is not to present a comprehensive survey of all the possible models but rather to select an illustrative modeling example that can give helpful direction on applying ML explainability in the financial industry while considering ethics issues.

B. EXPLAINABILITY WITH SHAP

After training the XGBoost model on the dataset, we proceed with explainability analysis of the model using SHAP. The results are summarized in Figures 3-4. Both plots display the feature importance of the dataset, i.e. how each feature of the dataset impacts the model's output. As explained in the SHAP documentation, a single dot on each feature row in the beeswarm plot represents an explanation for a given instance of the dataset. The horizontal position of the dot is determined by the SHAP value of that feature, while dots are accumulated along each feature row to show density. Color is used to display the original value of a feature.

In Fig. 3 we observe that on average the feature *Prior Default* has dominant impact on decisions for approving or rejecting credit card requests. Customers without prior default are generally favored by the model, while customers who have defaulted on their credit card payments are generally disfavored. Another insight is that low income applicants may still be issued credit cards provided that they had no prior default. On the other hand, applicants with prior default are less likely to get an approval for the same income level, meaning that having a prior default is a very strong indicator for rejecting credit card requests. Prior Default is followed by Income and Debt as important features, while employment-related factors as well as age and education are less relevant when considering credit card applications.

The beeswarm plot in Fig. 3 shows the density distribution across all instances of the dataset. Fig. 4 shows a bar plot obtained from SHAP on the same dataset for the XGBoost model, showing the global feature importance for the overall model, defined as the mean absolute value for that feature across all samples.

The SHAP explainer is also able to create bar plots to describe local feature importance of individual instances of the dataset. Fig. 5 represents bar plots for local explainability of four instances, where each feature is represented by its SHAP value. Fig. 5a shows how a typical plot looks like for an instance of the dataset where both the human decision and XGBoost prediction approved the credit card request. We notice that most of the features exhibit SHAP values that contribute strongly in favor of the approval. This is not a surprise. The selected applicant did not have a prior default, has income, is currently employed, and has been employed for 20 years. The education level and age contribute negatively, but their impact on the final decision is negligible. An interesting fact is that ethnicity has a slight negative impact even though this contributes only insignificantly. In any case, the SHAP explainer overall accurately captured all relevant factors that contributed positively for the decision, thereby showing alignment between the model prediction and human decision.

Fig. 5b is similar to Fig. 5a, but in the other direction. Fig. 5b provides local explanations for an applicant where both the human decision and XGBoost prediction resulted in a rejection. The most important reason for rejecting the applicant's request for obtaining a credit card is that their track record involves a prior default. Another factor that contributed negatively is the scarce employment history: the applicant is currently not employed and has been in the workforce only for less than half a month. Fig. 5a and Fig. 5b show that the model predictions can be explained and are aligned well with human decisions when there are strong arguments for either approving a credit card request or declining it. As an illustration, a list of selected applicants for which both the human decision and model prediction coincided favorably is given in Fig. A.1a in the Appendix A. Similarly, Fig. A.1b displays data for applicants who were rejected by human decision and model prediction.

The instances which exhibit a mismatch between the human decision and XGBoost prediction are even more relevant for explainability. They involve interactions among feature values where there is no immediate explanation whether the human decision or model prediction is right. Such instances require further analysis to understand the reasons behind a decision or prediction. For example, Fig. 5c shows SHAP values for an applicant whose credit card request was approved by human decision, but the model rejected it. The applicant is labeled by index 101 in Fig. A.1c. One may get an immediate impression that the bank correctly approved this request based on the high income and despite disfavorable factors such as prior default, debt, and insufficient employment history. This is actually the applicant with the highest income in the dataset.

However, the bank may have also been biased in the decision: the savings could potentially be spent quickly, which invalidates the credit card approval. On the other hand, while the high income was positively rated by the SHAP explainer, XGBoost was very strongly negatively influenced by the prior default, which ultimately determined the model prediction. Based on this discussion and also by looking at the plot and data for this applicant, one may not be able to say with certainty whether the bank or the model made the right decision.

As a similar example, Fig. 5d shows a bar plot for local explainability of an applicant whose request was rejected by human decision and approved by the model prediction. The applicant is denoted with the index 166 on Fig. A.1d. One may say that the bank made an incorrect decision and was biased due the insufficient employment history and inexistent income. The SHAP explainer rated highly other factors such as education level, small debt, current employment, and not having a prior default, which ultimately led the model to an approval. At first sight, for this particular applicant, it seems like the model decided correctly, while the bank was biased. However, there is no strict indication whether this is actually true. One may also say the bank made the right decision, while the model was biased.

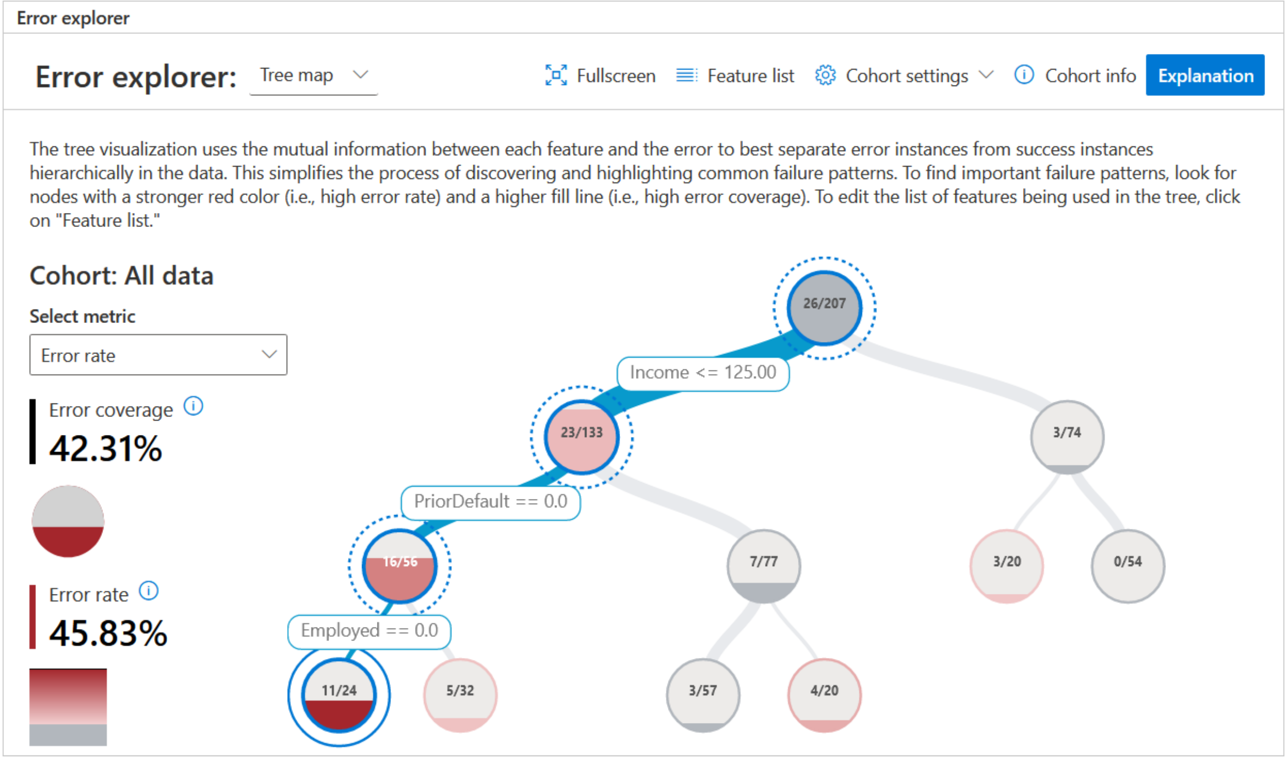
The examples on Fig. 5c and Fig. 5d show that both humans and machines could potentially make mistakes despite using their best judgement based on the available information. Therefore, it should be emphasized that the human-machine collaboration is important for making better decisions. When a machine is involved in the decision-making process, it makes the decisions more transparent. By involving ML models, human experts are able to return back to the applicant's case to make an additional review. This process could result in better decisions and reduce bias and mistakes.

C. EXPLAINABILITY WITH RESPONSIBLE AI WIDGETS

The Responsible AI Widgets developed by Microsoft provide a convenient visual dashboard for identifying cohorts of data that exhibit a higher error rate compared to overall (benchmark) error rate for the entire dataset. The error analysis in the dashboard can be performed by using two types of diagrams: i) error heatmaps obtained by selecting one or two features or ii) a binary decision tree that partitions the dataset into subgroups for discovering dominant error patterns.

Fig. 6 shows an error heatmap for two input features: Prior Default and Age. While various combinations of features can be selected, we wanted to see how errors are distributed for different age groups based on having a prior default or not, given that prior default is the most important feature for the dataset. As a two-dimensional grid, the heatmap partitions the dataset into different regions and visualizes how errors are distributed across the regions for these two features. The cells with higher errors are visualized with a darker red color, denoting a higher error disparity with the benchmark error rate. The analysis of the heatmap view depends on the understanding how feature importance may affect failure. The benchmark error rate for the dataset is 12.56%, but the heatmap reveals that some regions exhibit higher error rates than others. Credit card applicants of age 38.6 years or less and age above 53.9 are likely to suffer from model failures. However, applicants in the age range of 38.6-53.9 with prior default are more vulnerable than applicants who had no prior default.

The dashboard provides a data explorer which can be used to further analyze the cohort and uncover parts that are underrepresented. Fig. 7 shows how data is distributed across the feature Age. We notice an imbalance between the rejected applicants (marked blue) and approved applicants (marked orange) by the model. Most of the data is concentrated for applicants of age 46.3 or less, thereby explaining why this cohort is more susceptible to model errors. Since



Error explorer

Error explorer: Tree map

The tree visualization uses the mutual information between each feature and the error to best separate error instances from success instances hierarchically in the data. This simplifies the process of discovering and highlighting common failure patterns. To find important failure patterns, look for nodes with a stronger red color (i.e., high error rate) and a higher fill line (i.e., high error coverage). To edit the list of features being used in the tree, click on "Feature list."

Cohort: All data

Select metric: Error rate

Error coverage 42.31%

Error rate 45.83%

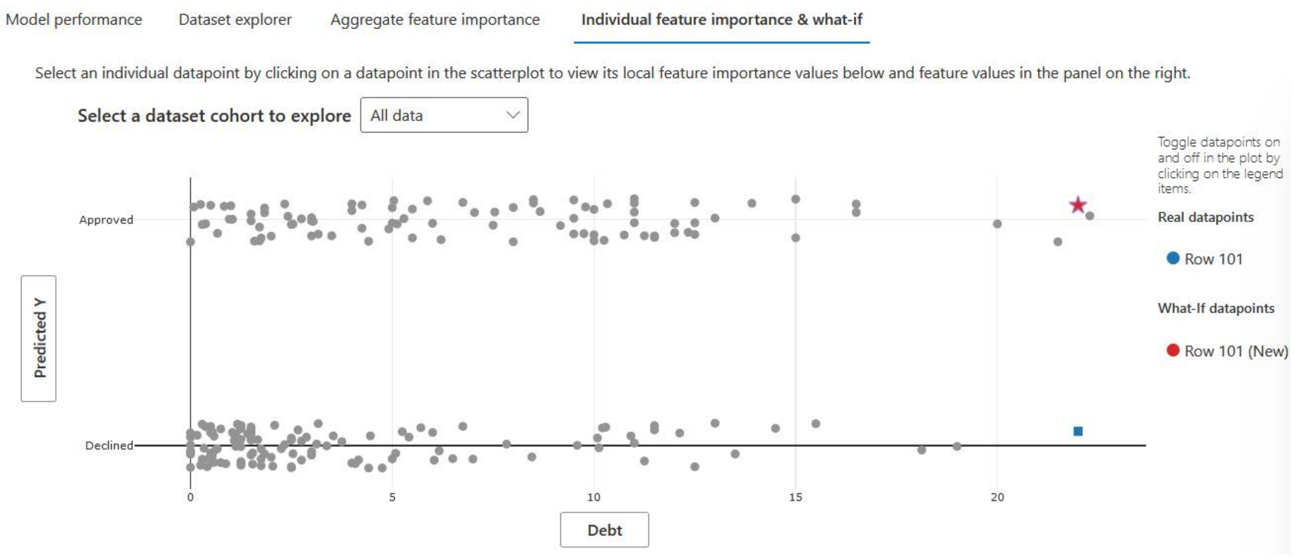
FIGURE 8. Decision tree approach for discovering patterns of data instances that are most prone to errors. We observe that the model makes more errors in credit card approvals for the groups of clients with income ≤ 125 who have had no prior defaults, and who are unemployed.

the cohort contains a smaller number of applicants aged 46.3 or more, the model exhibits lower error rates for this data region.

Error analysis can also be performed using a binary decision tree shown in Fig. 8. The decision tree separates error instances from success instances to simplify the discovery of common failure patterns. The decision tree consists of nodes and branches. The nodes with stronger red color indicate a higher error rate, while the thicker branches indicates a higher error coverage. Even though the overall error rate for the dataset is 12.56%, Fig. 8 shows that for lower income applicants who are not employed and have had no prior default the error rate can be as high as 45.83%, which is much higher than the benchmark.

An explanation why this cohort is so vulnerable to failures can be given based on the previous SHAP analysis. Namely, the SHAP explainer identified that Prior Default is the most important feature in the dataset, i.e. having no prior default is a strong contributor for approving a credit card request. On the other hand, being unemployed and having lower income has a negative impact and contributes to declining a request by the ML algorithm. As mentioned previously in the discussion of the SHAP beeswarm plot, low income applicants may still be issued credit cards provided that they had no prior default. It turns out that the positive effect of not having a prior default might be undone by lower income and unemployment, which makes the model more prone to erroneous decisions. For such applicants, the model may not render adequate predictions, meaning that a human expert may need to intervene to make the ultimate decision. A human may need to assess the situation by looking closely into other available factors and circumstances concerning the applicant in order to minimize the likelihood of failures.

It is also possible to perform a what-if analysis. The dashboard on Fig. 9 provides a way to select an instance, change values for some of the features, and see what would happen with the results once the values are changed. Going back to Fig. 5c, we can use the what-if tool to verify our conclusion that the model made a mistake not to approve the request for the applicant with index 101. We can see that if PriorDefault is changed to 0 (i.e. no prior default), while income is set to 0, then the model approves the applicant's request, as shown in Fig. 9, moving the blue square for instance 101 (declined) to the red star position (approved). Hence, the model would have approved the request if there was no prior default even if the applicant had no income. This result confirms that the model prediction is strongly biased by the Prior Default feature for this applicant. Conversely, the bank made the right decision to approve the request despite the prior default given that the income for the applicant is high (this is the applicant with highest income in the dataset).



Select an individual datapoint by clicking on a datapoint in the scatterplot to view its local feature importance values below and feature values in the panel on the right.

Select a dataset cohort to explore: All data

Toggle datapoints on and off in the plot by clicking on the legend items.

FIGURE 9. Using the what-if tool to understand the model and compare real data points with hypothetical ones obtained by modifying selected feature values. As an example, we change the prior default to no default, and the income status to no income for customer 101. These changes contribute to model-based decision change from "declined" to "approved" credit card application.

VIII. BENEFITS OF RESPONSIBLE ML IN ADDRESSING ETHICAL CHALLENGES IN FINTECH

In this section, we generalize our findings obtained from the previously presented use case scenario in which we demonstrated the application of tools for responsible ML in diagnosing credit card approval predictions. We utilize the ideas of error analysis, diagnosis and visualization techniques to give a conceptual framework for addressing the previously outlined ethical issues in fintech.

A. ADDRESSING BIASED DATA, ACCURACY AND TRANSPARENCY

The main ethical issues concerning bias are related to the use of representative data, defining impartial criteria to identify relevant data sources, and the ability to verify bias itself. With the use of tools for responsible ML, it is possible to check if the collected data is sufficiently rich with representative examples so that they are uniformly distributed across different cohorts of data. This makes it easy to check that one cohort is not favored over another within the dataset. As a result, the verification of impartial criteria and relevant data sources becomes viable. A related issue in this direction is checking the bias. While this activity will still involve human intervention, the tools are capable of making the process well defined and streamlined to minimize human intervention. Moreover, such tools automate the process, meaning that they can handle large volumes of data in a consolidated way. Another benefit is that they provide visualization dashboards that are easy to use and can help even the non-specialist to gauge the quality of data, its sufficiency and unbiasedness, thereby minimizing the need to consult with subject matter experts.

Concerning accuracy, the requirements consist of identifying the need to adjust or improve existing models, and verify whether better models are required for the particular problem. In both cases, the goal is to achieve accuracy that is commensurate with the problem domain at hand. Using the proposed tools, we can examine the entire dataset and obtain separate accuracies for different data regions rather than using an overall single-valued accuracy, which is an insufficient measure for accuracy in many use cases. The benefit of this approach is twofold. Firstly, it provides information whether a model performs well and demonstrates no discrepancies for various partitions of the data. In case the model does not perform satisfactorily, this approach gives insights into the problematic partitions, the reasons they fail, and offers a solution for adjusting or improving the existing model to meet the benchmark performance for the problematic regions. By observing the multi-dimensional aspects of accuracy, developers could obtain insights into the initial hypotheses to help them understand the most important features that contribute to the failures. Secondly, it will help developers compare performance among different ML models to identify the most suitable model. As a result, the analysis will ensure that the accuracy achieved is consistent across the dataset and that it is appropriate to the problem domain tackled by a given fintech service. This will enable faster due diligence of ML models, reduced efforts for model testing and evaluation, and increased customer confidence in the model optimality.

Finally, the problem with the lack of transparency could be effectively solved as a side effect of applying responsible ML tools since model transparency and interpretability are main priorities. Even when models with complex inner-workings are used such as deep learning models, an ML system can be easily and transparently gauged for precision with such a consolidated error analysis, thereby assessing fintech services and financial risks in a way that is methodologically sound. Furthermore, such approach will make the process transparent not only for finance professionals when rendering services to customers but also for financial regulators. This approach will give important compliance and risk management perspectives for ensuring fairness for financial institutions and individuals.

B. ADDRESSING DISCRIMINATION, DIFFERENTIAL PRICING AND MANIPULATED RECOMMENDATIONS

An essential requirement for preventing discrimination in the financial services industry is equitable treatment of various groups across the population and, in particular, eliminating unfair practices against groups that have historically been known for suffering from discrimination based on factors such as race, political or religious beliefs, sexual orientation, health problems or low income. One of the risks associated with discrimination is that ML models may inadvertently adjust the learning strategy to favor certain groups when making financial decisions. As a consequence, bias is further exacerbated which can lead to situations where certain groups may receive a better treatment in services such as credit scores, lending decisions, or price discounts compared to other groups. If an ML model is used, for example, to make a decision whether a person should be granted a loan or not, the positive or negative outcome of such a decision should be equitably and fairly justified. The process should be transparent and interpretable. This will be achieved if, upon customer request, providers of financial services are capable of providing proper documentation that justifies their decisions. The documentation implicitly assumes that the models should be verified upfront with error analysis as discussed in the previous subsection. The resulting effect is reducing discrimination, increasing customer trust, and providing auditable documentation for both legislators and regulatory bodies. Visualizing results for the applied ML models helps achieve these requirements to minimize or eliminate discrimination.

On the other hand, an ML model may learn a strategy for differential pricing based on factors that do not contribute to the value proposition of the rendered financial service. Increasing profits is one possible criterion for learning such a strategy. Differential pricing may give rise to ethical problems, through differential profit calculations. For example, different groups of customers may be charged varying rates based on factors such as race, ethnicity, religion, zip code, or basically anything else that does not in principle affect the intrinsic value of the financial service itself. The question is how to mitigate the problem. While it might be challenging for customers, especially in the online environment to verify the existence of differential pricing, the visualization capabilities of the responsible ML tools can serve as a safeguard against unfair pricing practices that may cause ethical consequences. Such methodology will enable legislators and regulators to audit pricing policies by examining ML models with visualization toolkits. In addition, financial services providers can keep logs of historical pricing for proving the consistency of pricing strategies with ethical requirements.

Along with bias, manipulated recommendations represent a related ethical problem that negatively impacts ML-based financial services. Manipulated recommendations refer to practices of tailoring ML models intentionally by purposefully injecting bias, hampering the objectivity of the model. Such bias hampers the objectivity of the model. For example, if the model is not objective, preference may be given to certain cohorts of input data at the expense of others. This can lead to giving biased advise to customers without sufficient evidence whether the actions are justified, possibly leading to financial loss. To prevent such consequences, it must be clear what dataset is used when making financial decisions to ensure that no specific subset of input data is preferred. In addition, as the dataset expands throughout time, dataset versioning is another important prerequisite for auditing purposes. Error analysis across various features of the dataset can identify problematic data regions that can contribute to inaccurate predictions. Transparently presenting the data sources and dataset itself together with the use of these tools can help recognize whether certain regions of the input data exhibit deviations from the rest of the data. This ensures interpretability of the justifications for the recommendations when rendering a specific financial service.

C. ADDRESSING CONFLICT OF INTEREST, REPORTING UNETHICAL ACTIVITIES AND PRESERVING CODES OF CONDUCT

Disclosing conflict of interest for financial services professionals means not only giving honest and accurate recommendations to their clients but also disclosing any interest that may contradict the interests of the clients. While there is a personal element involved in such dealings on the part of the financial services professionals, fintech services based on ML have the potential of making the advising process transparent. Performing error analysis of various regions of data and checking when and how certain regions fail against the chosen metrics, clients have an opportunity to verify if the analysis evidence matches the recommendations given by the fintech professional. The benefits are mutual for both sides as it increases trust and reduces risks related to conflict of interest.

Reporting unethical activities is another concern that is streamlined with responsible ML approaches. Responsible ML helps discover and report deficiencies in ML models that can lead to unethical behavior. Using error analysis, the responsibility for reporting unethical activities is no longer constrained only to an individual, but rather expands to ML teams that deploy models for use in fintech services. Having involved multiple professionals to work on the ML models reduces to a larger degree the risks associated with unethical behavior.

Lastly, transparency and explainability help regulatory agencies validate proper operation of fintech services based on ML models, thereby protecting interests of participants in the financial market and ensuring the codes of conduct are not violated.

D. ADDRESSING INSIDER TRADING, DATA PROTECTION AND LACK OF SKILLED STAFF

The tools for responsible ML could be used to develop monitoring systems to detect patterns that could alert regulators to investigate for potential fraud. The use of these tools could provide two major benefits in tackling the unfair practice of insider trading. Firstly, there is transparency of the input data used. This applies to both historical datasets used for training of ML models as well as live data streams fed into ML algorithms for the purpose of real-time deployment workflows. Secondly, with the use of such tools, the developers and users of the ML systems have the ability to continuously monitor the system performance, and to inspect potential risks of insider trading by AI-based monitoring of trading patterns or detecting unusual activities in securities trading. This allows monitoring of suspicious activities in real-time and efficiently detecting potential reasons for incorrect model behavior.

Data protection is a concern of systems based on information technology. Fintech is no exception. The additional challenge in fintech is to ensure data protection when personal data is analyzed using ML models for purposes of customer profiling. In those cases, ML models need to meet the regulatory framework of relevant data protection laws since customers may not understand how their data is being processed or may not be able to express their concerns or contest the decisions being made by ML algorithms [127], [128]. For example, articles 13-15 of Europe's new General Data Protection Regulation (GDPR) give rights to individuals to receive ''meaningful information about the logic involved'' in automated decision making, which is basically a right of individuals to obtain an explanation how ML models process their data [129].

In concordance with GDPR, responsible ML will actually help reduce personal data usage concerns and increase the trustworthiness of the decisions obtained by ML systems. In addition, the examples of handling personal data in fields such as healthcare can be used to improve data protection in the emerging field of fintech [20]. Personal data can be further protected using modern techniques such as differential privacy to prevent leakage of datasets used for ML model training [130].

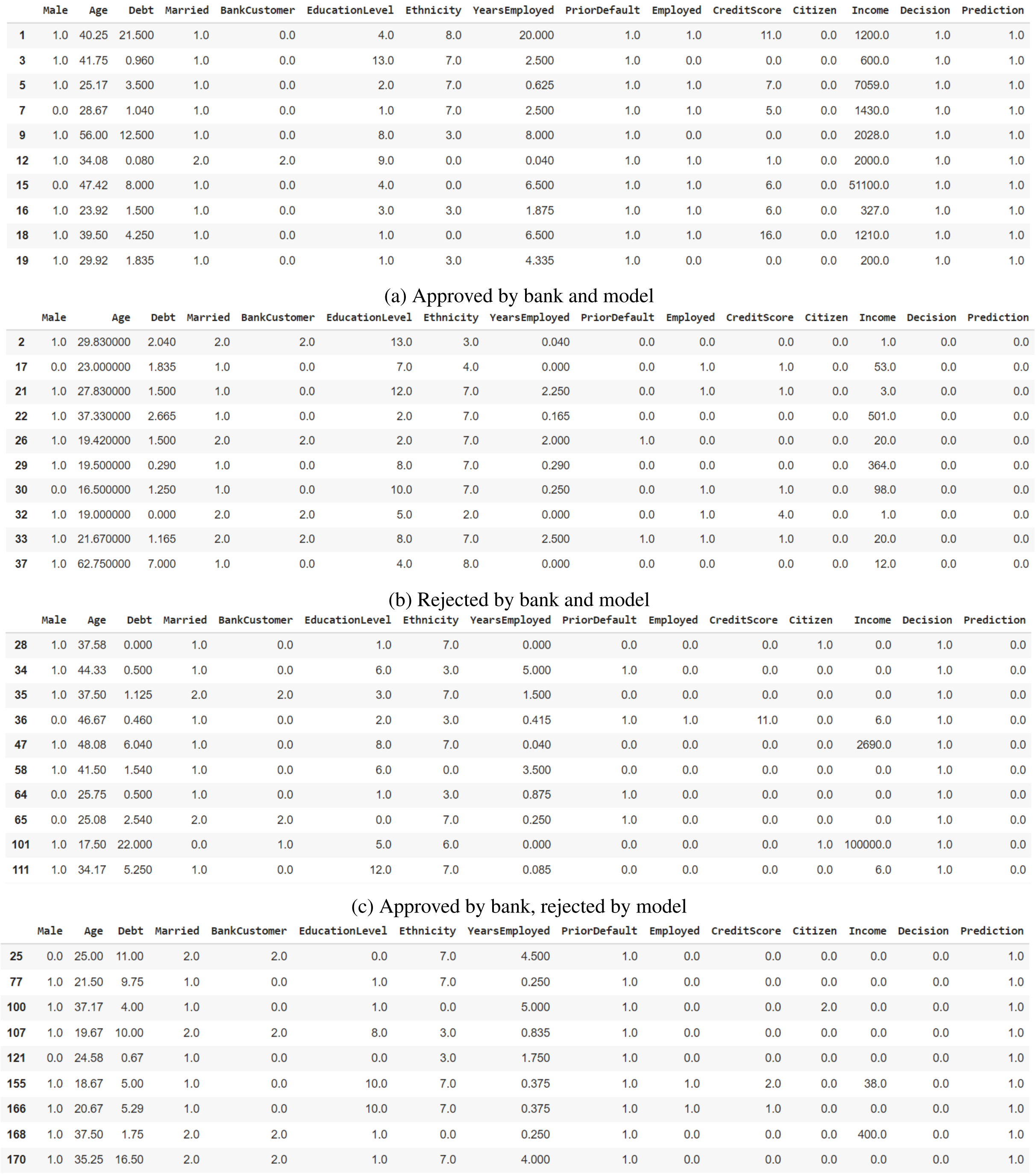
The explainability and interpretability of ML models will significantly aid specialists in the financial industry. ML experts are needed to develop the models and collaboratively work with finance specialists to help them analyze and use the results obtained by the models. In essence, financial services professionals will not need to know the sophisticated innerworkings of the ML models; they will have the correct error analysis and visualization toolset available to interpret results of ML models to take appropriate actions. Once the needed infrastructure for ML model analysis is in place, it will be very helpful for financial services companies to utilize the models for sophisticated financial analysis. As a result, the firms in the financial services industry will significantly increase the overall readiness and ability of their workforce to embrace machine learning.

IX. CONCLUSION

Machine learning is revolutionizing many economic sectors, including finance. Several global surveys with financial institutions reveal ample evidence that ML is poised to become the backbone of the financial industry in the near future. ML algorithms could enhance financial services and clients' value by harnessing the potential of large-scale automation that leads to significant cost savings. Despite the predicted positive impact for businesses, there is a range of ethical challenges in fintech that affect not only customers of fintech services but also financial institutions. To address these challenges, we performed mapping between ethical principles of finance and ethical principles of ML to reveal which traditional finance ethical principles have the most substantial correspondence to ML principles. The mapping outcome shows that traditional finance principles of integrity and fairness have the most significant overlap across ML ethics principles. Additionally, the ML ethics principles of human-centered values and fairness, as well as transparency and explainability, show the most considerable overlap with the traditional finance ethics principles. We study the correspondence of the conventional finance and ML ethics principles to merge the advantages of ML-based decision-making, such as cost and time savings, with the traditional finance decision-making based only on human-based criteria. This result confirms the importance of integrity and fairness as essential principles in finance ethics.

The paper presents a conceptual framework to identify and address challenges in financial decision-making such as bias, discrimination, differential pricing, conflict of interest, or data protection. The main objective of the mapping between finance and ML ethics is to identify the most critical criteria for handling ethical challenges in ML-based fintech applications. We rely on experts' opinions to evaluate the mapping between finance and ML ethics to assess these relationships. The application of the proposed framework is presented through a practical use case of creating an ML model for approving credit card requests. We showed how to develop a predictive model using state-of-the-art ML algorithms and explainable ML tools like SHAP and Microsoft Responsible AI Widgets. The application of explainability methods enhances model transparency and helps diagnose if models used in fintech settings suffer from inconsistencies that can cause ethical issues. Finally, we present a conceptual framework for using this approach to solve ethical challenges in ML applications for fintech.

APPENDIX A



	Male	Age	Debt	Married	BankCustomer	EducationLevel	Ethnicity	YearsEmployed	PriorDefault	Employed	CreditScore	Citizen	Income	Decision	Prediction
1	1.0	40.25	21.500	1.0	0.0	4.0	8.0	20.000	1.0	1.0	11.0	0.0	1200.0	1.0	1.0
3	1.0	41.75	0.960	1.0	0.0	13.0	7.0	2.500	1.0	0.0	0.0	0.0	600.0	1.0	1.0
5	1.0	25.17	3.500	1.0	0.0	2.0	7.0	0.625	1.0	1.0	7.0	0.0	7059.0	1.0	1.0
7	0.0	28.67	1.040	1.0	0.0	1.0	7.0	2.500	1.0	1.0	5.0	0.0	1430.0	1.0	1.0
9	1.0	56.00	12.500	1.0	0.0	8.0	3.0	8.000	1.0	0.0	0.0	0.0	2028.0	1.0	1.0
12	1.0	34.08	0.080	2.0	2.0	9.0	0.0	0.040	1.0	1.0	1.0	0.0	2000.0	1.0	1.0
15	0.0	47.42	8.000	1.0	0.0	4.0	0.0	6.500	1.0	1.0	6.0	0.0	51100.0	1.0	1.0
16	1.0	23.92	1.500	1.0	0.0	3.0	3.0	1.875	1.0	1.0	6.0	0.0	327.0	1.0	1.0
18	1.0	39.50	4.250	1.0	0.0	1.0	0.0	6.500	1.0	1.0	16.0	0.0	1210.0	1.0	1.0
19	1.0	29.92	1.835	1.0	0.0	1.0	3.0	4.335	1.0	0.0	0.0	0.0	200.0	1.0	1.0

(a) Approved by bank and model

	Male	Age	Debt	Married	BankCustomer	EducationLevel	Ethnicity	YearsEmployed	PriorDefault	Employed	CreditScore	Citizen	Income	Decision	Prediction
2	1.0	29.830000	2.040	2.0	2.0	13.0	3.0	0.040	0.0	0.0	0.0	0.0	1.0	0.0	0.0
17	0.0	23.000000	1.835	1.0	0.0	7.0	4.0	0.000	0.0	1.0	1.0	0.0	53.0	0.0	0.0
21	1.0	27.830000	1.500	1.0	0.0	12.0	7.0	2.250	0.0	1.0	1.0	0.0	3.0	0.0	0.0
22	1.0	37.330000	2.665	1.0	0.0	2.0	7.0	0.165	0.0	0.0	0.0	0.0	501.0	0.0	0.0
26	1.0	19.420000	1.500	2.0	2.0	2.0	7.0	2.000	1.0	0.0	0.0	0.0	20.0	0.0	0.0
29	1.0	19.500000	0.290	1.0	0.0	8.0	7.0	0.290	0.0	0.0	0.0	0.0	364.0	0.0	0.0
30	0.0	16.500000	1.250	1.0	0.0	10.0	7.0	0.250	0.0	1.0	1.0	0.0	98.0	0.0	0.0
32	1.0	19.000000	0.000	2.0	2.0	5.0	2.0	0.000	0.0	1.0	4.0	0.0	1.0	0.0	0.0
33	1.0	21.670000	1.165	2.0	2.0	8.0	7.0	2.500	1.0	1.0	1.0	0.0	20.0	0.0	0.0
37	1.0	62.750000	7.000	1.0	0.0	4.0	8.0	0.000	0.0	0.0	0.0	0.0	12.0	0.0	0.0

(b) Rejected by bank and model

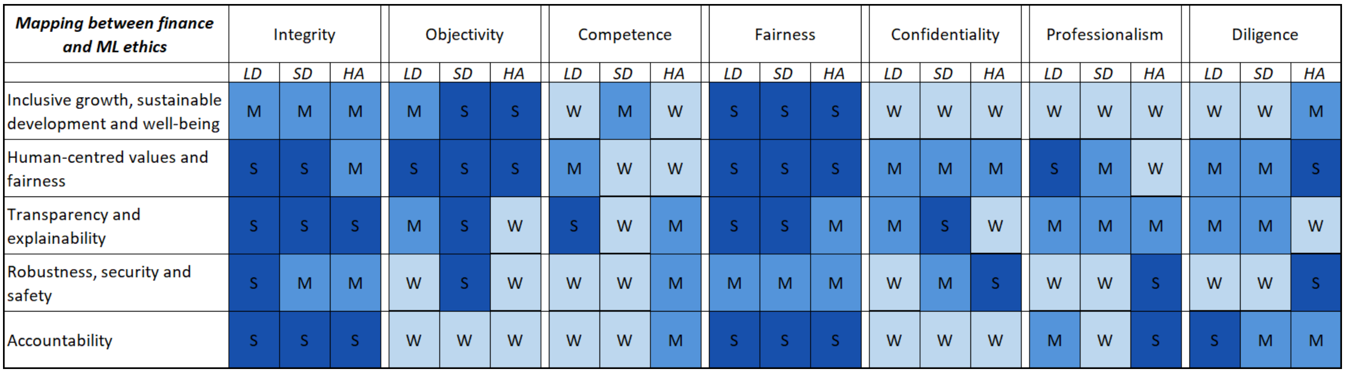
	Male	Age	Debt	Married	BankCustomer	EducationLevel	Ethnicity	YearsEmployed	PriorDefault	Employed	CreditScore	Citizen	Income	Decision	Prediction
28	1.0	37.58	0.000	1.0	0.0	1.0	7.0	0.000	0.0	0.0	0.0	1.0	0.0	1.0	0.0
34	1.0	44.33	0.500	1.0	0.0	6.0	3.0	5.000	1.0	0.0	0.0	0.0	0.0	1.0	0.0
35	1.0	37.50	1.125	2.0	2.0	3.0	7.0	1.500	0.0	0.0	0.0	0.0	0.0	1.0	0.0
36	0.0	46.67	0.460	1.0	0.0	2.0	3.0	0.415	1.0	1.0	11.0	0.0	6.0	1.0	0.0
47	1.0	48.08	6.040	1.0	0.0	8.0	7.0	0.040	0.0	0.0	0.0	0.0	2690.0	1.0	0.0
58	1.0	41.50	1.540	1.0	0.0	6.0	0.0	3.500	0.0	0.0	0.0	0.0	0.0	1.0	0.0
64	0.0	25.75	0.500	1.0	0.0	1.0	3.0	0.875	1.0	0.0	0.0	0.0	0.0	1.0	0.0
65	0.0	25.08	2.540	2.0	2.0	0.0	7.0	0.250	1.0	0.0	0.0	0.0	0.0	1.0	0.0
101	1.0	17.50	22.000	0.0	1.0	5.0	6.0	0.000	0.0	0.0	0.0	1.0	100000.0	1.0	0.0
111	1.0	34.17	5.250	1.0	0.0	12.0	7.0	0.085	0.0	0.0	0.0	0.0	6.0	1.0	0.0

(c) Approved by bank, rejected by model

	Male	Age	Debt	Married	BankCustomer	EducationLevel	Ethnicity	YearsEmployed	PriorDefault	Employed	CreditScore	Citizen	Income	Decision	Prediction
25	0.0	25.00	11.00	2.0	2.0	0.0	7.0	4.500	1.0	0.0	0.0	0.0	0.0	0.0	1.0
77	1.0	21.50	9.75	1.0	0.0	1.0	7.0	0.250	1.0	0.0	0.0	0.0	0.0	0.0	1.0
100	1.0	37.17	4.00	1.0	0.0	1.0	0.0	5.000	1.0	0.0	0.0	2.0	0.0	0.0	1.0
107	1.0	19.67	10.00	2.0	2.0	8.0	3.0	0.835	1.0	0.0	0.0	0.0	0.0	0.0	1.0
121	0.0	24.58	0.67	1.0	0.0	0.0	3.0	1.750	1.0	0.0	0.0	0.0	0.0	0.0	1.0
155	1.0	18.67	5.00	1.0	0.0	10.0	7.0	0.375	1.0	1.0	2.0	0.0	38.0	0.0	1.0
166	1.0	20.67	5.29	1.0	0.0	10.0	7.0	0.375	1.0	1.0	1.0	0.0	0.0	0.0	1.0
168	1.0	37.50	1.75	2.0	2.0	1.0	0.0	0.250	1.0	0.0	0.0	0.0	400.0	0.0	1.0
170	1.0	35.25	16.50	2.0	2.0	1.0	7.0	4.000	1.0	0.0	0.0	0.0	0.0	0.0	1.0

(d) Rejected by bank, approved by model

FIGURE A.1. Lists of applicants from the dataset: (a) approved by human decision and model prediction, (b) rejected by human decision and model prediction, (c) approved by human decision and rejected by model, (d) rejected by human decision and approved by model. Selected applications are used in Fig. 5 for our local explainability analysis with SHAP.



Mapping between finance and ML ethics	Integrity			Objectivity			Competence			Fairness			Confidentiality			Professionalism			Diligence		
	LD	*SD*	*HA*	*LD*	*SD*	*HA*	*LD*	*SD*	*HA*	*LD*	*SD*	*HA*	*LD*	*SD*	*HA*	*LD*	*SD*	*HA*	*LD*	*SD*	*HA*
Inclusive growth, sustainable development and well-being	M	M	M	M	S	S	W	M	W	S	S	S	W	W	W	W	W	W	W	W	M
Human-centred values and fairness	S	S	M	S	S	S	M	W	W	S	S	S	M	M	M	S	M	W	M	M	S
Transparency and explainability	S	S	S	M	S	W	S	W	M	S	S	M	M	S	W	M	M	M	M	M	W
Robustness, security and safety	S	M	M	W	S	W	W	W	M	M	M	M	W	M	S	W	W	S	W	W	S
Accountability	S	S	S	W	W	W	W	W	M	S	S	S	W	W	W	M	W	S	S	M	M

FIGURE A.2. We present the strength of the mapping between pairs of principles (traditional finance ethics and ML ethics) based on NLP methods using long (LD) and short (SD) definitions and human experts' assignments (HA). The intensity of the color represents the level of overlapping between the principles, which can be strong (S), moderate (M) or weak (W).

REFERENCES

[1] A. Phaneuf. (2020). *Artificial Intelligence in Financial Services: Applications and Benefits of AI in Finance*. Accessed: Mar. 6, 2021. [Online]. Available: https://www.insiderintelligence.com/insights/ai-in-finance/

[2] M. Bazarbash, *FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk*. Washington, DC, USA: International Monetary Fund (IMF), May 2019. [Online]. Available: https://www.imf.org/en/Publications/WP/Issues/2019/05/17/FinTech-in-Financial-Inclusion-Machine-Learning-Applications-in-Assessing-Credit-Risk-46883

[3] J. Jagtiani and C. Lemieux, "The roles of alternative data and machine learning in fintech lending: Evidence from the lending club consumer platform," *Financial Manage.*, vol. 48, no. 4, pp. 1009–1029, Dec. 2019.

[4] L. Gambacorta, Y. Huang, H. Qiu, and J. Wang, "How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm," Bank Int. Settlements (BIS), BIS Working Paper 834, Jan. 2019. [Online]. Available: https://www.bis.org/publ/work834.htm

[5] D. Snow, "Machine learning in asset management—Part 1: Portfolio construction—Trading strategies," *J. Financial Data Sci.*, vol. 2, no. 1, pp. 10–23, Jan. 2020.

[6] D. Snow, "Machine learning in asset management—Part 2: Portfolio construction—Weight optimization," *J. Financial Data Sci.*, vol. 2, no. 2, pp. 17–24, Apr. 2020.

[7] J.-Y. Yeh and C.-H. Chen, "A machine learning approach to predict the success of crowdfunding fintech project," *J. Enterprise Inf. Manag.*, Jul. 2020. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/JEIM-01-2019-0017/full/html, doi: 10.1108/JEIM-01-2019-0017.

[8] T. Philippon, "On fintech and financial inclusion," Nat. Bur. Econ. Res., Cambridge, MA, USA, Working Paper 26330, 2019. [Online]. Available: https://www.nber.org/papers/w26330, doi: 10.3386/w26330.

[9] D. Makina, "The potential of fintech in enabling financial inclusion," in *Extending Financial Inclusion in Africa*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 299–318.

[10] M. Colangelo. (2020). *Mass Adoption of AI in Financial Services Expected Within Two Years*. Accessed: Mar. 6, 2021. [Online]. Available: https://www.forbes.com/sites/cognitiveworld/2020/02/20/mass-adoption-of-ai-in-financial-services-expected-within-two-years/?sh=14720c307d71

[11] L. Ryll, M. E. Barton, B. Z. Zhang, R. J. McWaters, E. Schizas, R. Hao, K. Bear, M. Preziuso, E. Seger, R. Wardrop, P. R. Rau, P. Debata, P. Rowan, N. Adams, M. Gray, and N. Yerolemou, "Transforming paradigms: A global AI in financial services survey," *SSRN Electron. J.*, May 2020. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3532038

[12] N. Gokhale, A. Gajjari, R. Kaye, and D. Kuder. (2019). *AI Leaders in Financial Services: Common Traits of Frontrunners in the Artificial Intelligence Race*. Accessed: Mar. 6, 2021. [Online]. Available: https://www2.deloitte.com/content/dam/insights/us/articles/4687_traits-of-ai-frontrunners/DI_AI-leaders-in-financial-services.pdf

[13] BOSCH. (2020). *Code of Ethics for AI*. Accessed: Mar. 10, 2021. [Online]. Available: https://www.bosch.com/stories/ethical-guidelines-for-artificial-intelligence

[14] (2020). *We Have to Not Only Develop AI, But Build Trust in AI as Well*. Accessed: Mar. 10, 2021. [Online]. Available: https://www.bosch.com/stories/denners-view-artificial-intelligence-ethics

[15] Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct*. Accessed: Mar. 10, 2021. [Online]. Available: https://www.acm.org/code-of-ethics

[16] L. Ryll, M. E. Barton, and B. Z. Zhang. (2020). *AI Has Started a Financial Revolution—Here's How*. Accessed: Mar. 6, 2021. [Online]. Available: https://www.weforum.org/agenda/2020/02/how-ai-is-shaping-financial-services

[17] Deloitte. (2020). *Digital Banking Maturity 2020*. Accessed: Mar. 6, 2021. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/ce/Documents/financial-services/ce-digital-banking-maturity-2020.pdf

[18] Louis Columbus. (2020). *The State of AI Adoption in Financial Services*. Accessed: Mar. 6, 2021. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2020/10/31/the-state-of-ai-adoption-in-financial-services/

[19] D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care-addressing ethical challenges," *New England J. Med.*, vol. 378, no. 11, p. 981, Mar. 2018.

[20] E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: Addressing ethical challenges," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002689.

[21] A. M. Darcy, A. K. Louie, and L. W. Roberts, "Machine learning and the profession of medicine," *Jama*, vol. 315, no. 6, pp. 551–552, Feb. 2016.

[22] T. Grote and P. Berens, "On the ethics of algorithmic decision-making in healthcare," *J. Med. Ethics*, vol. 46, no. 3, pp. 205–211, Mar. 2020.

[23] D. S. Char, M. D. Abrámoff, and C. Feudtner, "Identifying ethical considerations for machine learning healthcare applications," *Amer. J. Bioethics*, vol. 20, no. 11, pp. 7–17, Nov. 2020.

[24] B. J. Bloch. (2019). *8 Ethical Guidelines for Brokers*. Accessed: Feb. 17, 2021. [Online]. Available: https://www.investopedia.com/articles/financialcareers/08/broker-ethics-tips.asp

[25] International Business Brokers Association. *Code of Ethics*. Accessed: Feb. 17, 2021. [Online]. Available: https://www.ibba.org/more-ibba/code-of-ethics

[26] J. A. Ragatz and R. F. Duska, "Financial codes of ethics," in *Finance Ethics: Critical Issues in Theory and Practice*, J. R. Boatright, Ed. Hoboken, NJ, USA: Wiley, 2010, ch. 16, pp. 297–323.

[27] S. Bok, "The limits of confidentiality," JSTOR, Hastings Center Rep., 1983, pp. 24–31. [Online]. Available: https://www.jstor.org/stable/3561549, doi: 10.2307/3561549.

[28] International Monetary Fund. (2018). *World Economic Outlook Database*. [Online]. Available: https://www.imf.org/en/Publications/SPROLLS/world-economic-outlook-databases

[29] Organisation for Economic Co-operation and Development (OECD). *About the OECD*. Accessed: Jul. 20, 2021. [Online]. Available: https://www.oecd.org/about/

[30] (2019). *OECD AI Principles Overview*. Accessed: Jun. 13, 2021. [Online]. Available: https://www.oecd.ai/ai-principles

[31] (2019). *OECD AI Principles*. Accessed: Jun. 13, 2021. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[32] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[33] Z. C. Lipton, "The Mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[34] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 210–215.

[35] D. Doran, S. Schulz, and T. R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," 2017, *arXiv:1710.00794*.

[36] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[37] M. Fox, D. Long, and D. Magazzeni, "Explainable planning," 2017, *arXiv:1709.10256*.

[38] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: Definitions, methods, and applications," 2019, *arXiv:1901.04592*.

[39] A. B. Tickle, R. Andrews, M. Golea, and J. Diederich, "The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks," *IEEE Trans. Neural Netw.*, vol. 9, no. 6, pp. 1057–1068, Nov. 1998.

[40] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," 2017, *arXiv:1705.08821*.

[41] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag, "Learning functional causal models with generative neural networks," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Cham, Switzerland: Springer, 2018, pp. 39–80. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-98131-4_3

[42] S. Athey and G. W. Imbens, "Machine learning methods for estimating heterogeneous causal effects," *Stat*, vol. 1050, no. 5, pp. 1–26, Jul. 2015.

[43] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 1721–1730.

[44] M. W. Craven, "Extracting comprehensible models from trained neural networks," Ph.D. dissertation, Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, 1996.

[45] A. Vellido, J. D. Martín-Guerrero, and P. J. Lisboa, "Making machine learning models interpretable," in *Proc. ESANN*, vol. 12, Apr. 2012, pp. 163–172.

[46] A. Theodorou, R. H. Wortham, and J. J. Bryson, "Designing and implementing transparency for real time inspection of autonomous robots," *Connection Sci.*, vol. 29, no. 3, pp. 230–241, Jul. 2017.

[47] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" 2017, *arXiv:1712.09923*.

[48] A. Chander, R. Srinivasan, S. Chelian, J. Wang, and K. Uchino, "Working with beliefs: AI transparency in the enterprise," in *Proc. IUI Workshops*, Jan. 2018, pp. 1–4.

[49] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[50] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 2017, *arXiv:1708.08296*.

[51] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019.

[52] V. Schetinin, J. E. Fieldsend, D. Partridge, T. J. Coats, W. J. Krzanowski, R. M. Everson, T. C. Bailey, and A. Hernandez, "Confident interpretation of Bayesian decision tree ensembles for clinical applications," *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 3, pp. 312–319, May 2007.

[53] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, "Learning how to explain neural networks: PatternNet and patternattribution," 2017, *arXiv:1705.05598*.

[54] C. Wadsworth, F. Vera, and C. Piech, "Achieving fairness through adversarial learning: An application to recidivism prediction," 2018, *arXiv:1807.00199*.

[55] B. Green, "'Fair' risk assessments: A precarious approach for criminal justice reform," in *Proc. 5th Workshop Fairness, Accountability, Transparency Mach. Learn.*, 2018, pp. 1–5.

[56] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, Jun. 2017.

[57] M. P. Kim, O. Reingold, and G. N. Rothblum, "Fairness through computationally-bounded awareness," 2018, *arXiv:1803.03239*.

[58] M. Harbers, K. van den Bosch, and J.-J. Meyer, "Design and evaluation of explainable BDI agents," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Aug. 2010, pp. 125–132.

[59] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," in *Proc. AAAI*, vol. 17, 2017, pp. 4762–4763.

[60] M. A. Neerincx, J. van der Waa, F. Kaptein, and J. van Diggelen, "Using perceptual and cognitive explanations for enhanced human-agent team performance," in *Proc. Int. Conf. Eng. Psychol. Cogn. Ergonom.* Cham, Switzerland: Springer, 2018, pp. 204–214. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-91122-9_18

[61] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 5686–5697.

[62] L. Edwards and M. Veale, "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for," *Duke Law Tech. Rev.*, vol. 16, p. 18, 2017.

[63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[64] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.

[65] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Oct. 2021, pp. 296–310.

[66] H. Xu, B. Van Durme, and K. Murray, "BERT, mBERT, or BiBERT? A study on contextualized embeddings for neural machine translation," 2021, *arXiv:2109.04588*.

[67] S. Mehta, M. Ghazvininejad, S. Iyer, L. Zettlemoyer, and H. Hajishirzi, "DeLighT: Deep and light-weight transformer," 2020, *arXiv:2008.00623*.

[68] M. Rikters, M. Pinnis, and R. Krišlauks, "Training and adapting multilingual NMT for less-resourced and morphologically rich languages," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1–8.

[69] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, and W. Y. Wang, "FinQA: A dataset of numerical reasoning over financial data," 2021, *arXiv:2109.00122*.

[70] Z. Abbasiantaeb and S. Momtazi, "Text-based question answering from information retrieval and deep neural network perspectives: A survey," *Wires Data Mining Knowl. Discovery*, vol. 11, no. 6, Nov. 2021, Art. no. e1412.

[71] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019, *arXiv:1910.10683*.

[72] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: Deep contextualized entity representations with entity-aware self-attention," 2020, *arXiv:2010.01057*.

[73] J. Dai, H. Yan, T. Sun, P. Liu, and X. Qiu, "Does syntax matter? A strong baseline for aspect-based sentiment analysis with roberta," 2021, *arXiv:2104.04986*.

[74] L. Mathew and V. R. Bindu, "Efficient classification techniques in sentiment analysis using transformers," in *Proc. Int. Conf. Innov. Comput. Commun.* Singapore: Springer, 2022, pp. 849–862. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-16-2594-7_69

[75] A. Singh and G. Jain, "Sentiment analysis of news headlines using simple transformers," in *Proc. Asian Conf. Innov. Technol. (ASIANCON)*, Aug. 2021, pp. 1–6.

[76] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: From lexicons to transformers," *IEEE Access*, vol. 8, pp. 131662–131682, 2020.

[77] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Mar. 2020.

[78] N. Praechanya and O. Sornil, "Improving Thai named entity recognition performance using BERT transformer on deep networks," in *Proc. 6th Int. Conf. Mach. Learn. Technol.*, Apr. 2021, pp. 177–183.

[79] M.-S. Huang, P.-T. Lai, P.-Y. Lin, Y.-T. You, R. T.-H. Tsai, and W.-L. Hsu, "Biomedical named entity recognition and linking datasets: Survey and our recent development," *Briefings Bioinf.*, vol. 21, no. 6, pp. 2219–2238, Dec. 2020.

[80] N. Jofche, K. Mishev, R. Stojanov, M. Jovanovik, and D. Trajanov, "PharmKE: Knowledge extraction platform for pharmaceutical texts using transfer learning," 2021, *arXiv:2102.13139*.

[81] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113679.

[82] W. Guan, I. Smetannikov, and M. Tianxing, "Survey on automatic text summarization and transformer models applicability," in *Proc. Int. Conf. Control, Robot. Intell. Syst.*, Oct. 2020, pp. 176–184.

[83] A. Gupta, D. Chugh, D. Anjum, and R. Katarya, "Automated news summarization using transformers," 2021, *arXiv:2108.01064*.

[84] A. A. Syed, F. L. Gaol, and T. Matsuo, "A survey of the state-of-the-art models in neural abstractive text summarization," *IEEE Access*, vol. 9, pp. 13248–13265, 2021.

[85] F. Benedetti, D. Beneventano, S. Bergamaschi, and G. Simonini, "Computing inter-document similarity with context semantic analysis," *Inf. Syst.*, vol. 80, pp. 136–147, Feb. 2019.

[86] Hugging Face. (2020). *NLI Distil-Roberta Base V2*. Accessed: Jun. 25, 2021. [Online]. Available: https://huggingface.co/sentence-transformers/nli-distilroberta-base-v2

[87] A. J. Shetty. *Ethics in Finance*. Accessed: Feb. 17, 2021. [Online]. Available: https://www.encyclopedia.com/finance/finance-and-accounting-magazines/ethics-finance

[88] J. Grus, *Data Science From Scratch: First Principles With Python*. Sebastopol, CA, USA: O'Reilly Media, 2019.

[89] M. Zook, S. Barocas, D. Boyd, K. Crawford, E. Keller, S. P. Gangadharan, A. Goodman, R. Hollander, B. A. Koenig, J. Metcalf, A. Narayanan, A. Nelson, and F. Pasquale, "Ten simple rules for responsible big data research," *PLOS Comput. Biol.*, vol. 13, no. 3, Mar. 2017, Art. no. e1005399, doi: 10.1371/journal.pcbi.1005399.

[90] B. Franks, *97 Things About Ethics Everyone in Data Science Should Know*. Sebastopol, CA, USA: O'Reilly Media, 2020.

[91] J. Clayton. (2019). *Regulation Best Interest and the Investment Adviser Fiduciary Duty: Two Strong Standards That Protect and Provide Choice for Main Street Investors*. Accessed: Dec. 7, 2021. [Online]. Available: https://www.sec.gov/news/speech/clayton-regulation-best-interest-investment-adviser-fiduciary-duty

[92] E. Derman and P. Wilmott, "The financial modelers' manifesto," *SSRN Electron. J.*, 2009, Art. no. 1324878. [Online]. Available: https://ssrn.com/abstract=1324878, doi: 10.2139/ssrn.1324878.

[93] S. Aziz and M. Dowling, "Machine learning and AI for risk management," in *Disrupting Finance*. Cham, Switzerland: Palgrave Pivot, 2019, pp. 33–50.

[94] N. J. King and P. W. Jessen, "Profiling the mobile customer—Privacy concerns when behavioural advertisers target mobile phones—Part I," *Comput. Law Secur. Rev.*, vol. 26, no. 5, pp. 455–478, Sep. 2010.

[95] N. J. King and P. W. Jessen, "Profiling the mobile customer—Is industry self-regulation adequate to protect consumer privacy when behavioural advertisers target mobile phones?—Part II," *Comput. Law Secur. Rev.*, vol. 26, no. 6, pp. 595–612, Nov. 2010.

[96] A. Smith. (2020). *Using Artificial Intelligence and Algorithms*. Accessed: Jul. 15, 2021. [Online]. Available: https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms

[97] B. Brian. *Ethics in Finance: How it Affects Professionals*. Accessed: Feb. 17, 2021. [Online]. Available: https://www.investopedia.com/articles/financialcareers/09/professional-standards-ethics.asp

[98] T. Frankel, "Insider trading," *SMU Law Rev.*, vol. 71, p. 783, 2018.

[99] H. J. Wilson, P. Daugherty, and N. Bianzino, "The jobs that artificial intelligence will create," *MIT Sloan Manag. Rev.*, vol. 58, no. 4, p. 14, 2017.

[100] R. Metz. (2021). *Zillow's Home-Buying Debacle Shows How Hard it is to Use AI to Value Real Estate*. Accessed: Dec. 7, 2021. [Online]. Available: https://edition.cnn.com/2021/11/09/tech/zillow-ibuying-home-zestimate/index.html

[101] D. V. Carvalho, M. E. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019.

[102] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. Vera Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," 2019, *arXiv:1909.03012*.

[103] B. Nushi. (2021). *Responsible Machine Learning With Error Analysis*. Accessed: Mar. 6, 2021. [Online]. Available: https://techcommunity.microsoft.com/t5/azure-ai/responsible-machine-learning-with-error-analysis/ba-p/2141774

[104] Microsoft. (2021). *Responsible AI Widgets*. Accessed: Mar. 6, 2021. [Online]. Available: https://github.com/microsoft/responsible-ai-widgets

[105] B. Nushi, E. Kamar, and E. Horvitz, "Towards accountable AI: Hybrid human-machine analyses for characterizing system failure," in *Proc. AAAI Conf. Hum. Comput. Crowdsourcing*, 2018, vol. 6, no. 1, pp. 126–135.

[106] S. Singla, B. Nushi, S. Shah, E. Kamar, and E. Horvitz, "Understanding failures of deep networks via robust feature extraction," 2020, *arXiv:2012.01750*.

[107] S. Mazzanti. (2020). *Shap Values Explained Exactly How You Wished Someone Explained to You*. Accessed: Jul. 5, 2021. [Online]. Available: https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30

[108] S. Lundberg. (2018). *SHapley Additive Explanations*. Accessed: Mar. 6, 2021. [Online]. Available: https://github.com/slundberg/shap

[109] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.

[110] IBM Research. (2017). *AI Explainability 360*. Accessed: Mar. 6, 2021. [Online]. Available: https://github.com/Trusted-AI/AIX360

[111] A. Shrikumar. (2017). *DeepLIFT: Deep Learning Important Features*. Accessed: Mar. 6, 2021. [Online]. Available: https://github.com/Trusted-AI/AIX360

[112] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[113] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016, *arXiv:1605.01713*.

[114] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 56–65, Aug. 2019.

[115] Google Research. (2019). *What-if Tool*. Accessed: Mar. 6, 2021. [Online]. Available: https://github.com/pair-code/what-if-tool

[116] E. Creager, D. Madras, J. H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 1436–1445.

[117] P. Garg, J. Villasenor, and V. Foggo, "Fairness metrics: A comparative analysis," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 3662–3666.

[118] S. N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H. Chen, X. Liu, and Z. He, "Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 7, pp. 1173–1185, Jul. 2020.

[119] F. Emmert-Streib, O. Yli-Harja, and M. Dehmer, "Explainable artificial intelligence and machine learning: A reality rooted perspective," *Wires Data Mining Knowl. Discovery*, vol. 10, no. 6, Nov. 2020, Art. no. e1368.

[120] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3681–3688.

[121] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," 2018, *arXiv:1806.08049*.

[122] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with shapley-value-based explanations as feature importance measures," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5491–5500.

[123] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[124] R. Kuhn. (2015). *Analysis of Credit Approval Data*. Accessed: Sep. 3, 2021. [Online]. Available: http://ryankuhn.net/CreditAnalysis/articles/FinalProject.html

[125] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, and H. Cho, "XGBoost: Extreme gradient boosting," *R Package Version* vol. 1, no. 4, pp. 1–4, Aug. 2015.

[126] A. Ye. (2020). *XGBoost, LightGBM, and Other Kaggle Competition Favorites*. Accessed: Sep. 10, 2021. [Online]. Available: https://towardsdatascience.com/xgboost-lightgbm-and-other-kaggle-competition-favorites-6212e8b0e835

[127] C. Kuner, D. J. B. Svantesson, F. H. Cate, O. Lynskey, and C. Millard, ''Machine learning with personal data: Is data protection law smart enough to meet the challenge?'' *Int. Data Privacy Law*, vol. 7, no. 1, pp. 1–2, Feb. 2017.

[128] D. Kamarinou, C. Millard, and J. Singh, ''Machine learning with personal data,'' School Law, Queen Mary Univ. London, London, U.K., Res. Paper 247, 2016.

[129] A. Selbst and J. Powles, '''Meaningful information' and the right to explanation,'' in *Proc. Conf. Fairness, Accountability Transparency*, Jan. 2018, p. 48.

[130] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, ''Deep learning with differential privacy,'' in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2016, pp. 308–318.

MARYAN RIZINSKI received the B.S. and M.S. degrees in electrical engineering and information technologies from Ss. Cyril and Methodius University, Skopje, where he is currently pursuing the Ph.D. degree in computer science. He is also an Engineering Manager at Bosch, with over ten years of industry experience leading globally-distributed software engineering teams. His expertise spans multiple aspects of the software project lifecycle management, from planning, requirement gathering, and analysis, estimations to driving delivery, rollout, and troubleshooting for international customers. Throughout his professional career, he has managed the implementation of the Internet of Things (IoT) and fiber-optics infrastructure projects and has been mentoring and consulting startup IT companies. He is also a Lecturer of computer science at the Metropolitan College, Boston University, where he is teaching and facilitating networking and data science classes. His Ph.D. research focuses on novel approaches for using machine learning (ML) and natural language processing (NLP) in the financial industry and other related areas. His research aims to enable more accurate decision-making and address fundamental problems of improving the explainability of deep-learning models and studying ML-related ethical challenges in finance applications. His research interests include computer networking, wireless communications, and new internet and the IoT architectures.

HRISTIJAN PESHOV is currently pursuing the Bachelor of Science degree in software engineering and information systems with the Faculty of Computer Science and Engineering, Saints Cyril and Methodius University, Skopje. He is in his last year of studies while he also works as a Software Engineer. His research interests include data science, machine learning, natural language processing, and network analysis.

KOSTADIN MISHEV received the bachelor's degree in informatics and computer engineering and the master's degree in computer networks and e-technologies degree from Saints Cyril and Methodius University, Skopje, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree. He is also a Teaching and Research Assistant with the Faculty of Computer Science and Engineering, Saints Cyril and Methodius University. His research interests include data science, natural language processing, semantic web, web technologies, and computer networks.

LUBOMIR T. CHITKUSHEV received the Dipl.Ing. degree in electrical engineering from the University of Belgrade, the M.Sc. degree in biomedical engineering from the Medical College of Virginia, VCU, and the Ph.D. degree in biomedical engineering from Boston University. He is currently an Associate Professor of computer science with the Metropolitan College, Boston University, where he works as the Director of health informatics and health sciences and the Associate Dean of academic affairs. He is also the Founder of the Health Informatics Program, Boston University, and a Founding Member of the RINA Laboratory, Boston University, where recursive inter-network architecture (RINA) was introduced as efficient, scalable, and secure approach to internet architecture. He is also a Co-Founder and the Associate Director of the Center for Reliable Information Systems and Cyber Security (RISCS), Boston University, which coordinates research on reliable and secure computational systems and infrastructure and information assurance education. His research interests include modeling of complex systems, computer network security and architecture, and biomedical and health informatics. He has served as the Principal Investigator for Boston University on research grants awarded by the European Commission, EU, the National Security Agency, USA, and the U.S. Department of Justice. He has also served as a Reviewer for the U.S. National Science Foundation.

IRENA VODENSKA received the B.S. degree in computer information systems from the University of Belgrade, the Master of Arts (M.A.) degree in economics, the M.B.A. degree from the Owen Graduate School of Management, Vanderbilt University, and the Ph.D. degree in econophysics (statistical finance) from Boston University. She is currently a Professor of finance, the Director of finance programs, and the Chair of the Administrative Sciences Department, Metropolitan College, Boston University. She conducts theoretical and applied interdisciplinary research using quantitative approaches for modeling interdependences of financial networks, banking system dynamics, and global financial crises. She also studies the effects of media on financial markets, corporations, financial institutions, and related global economic systems. She uses neural networks and deep learning methodologies for natural language processing to text mine important factors affecting corporate performance, environmental, social, and governance (ESG) corporate reporting, and global economic trends, primarily related to climate change. She teaches investment analysis and portfolio management, international finance and trade, financial regulation and ethics, and derivatives securities and markets at Boston University. Her research interests include network theory and complexity science in macroeconomics. She is also a Chartered Financial Analyst (CFA) Charter Holder. As a Principal Investigator (PI) with Boston University, she has won interdisciplinary research grants awarded by the European Commission, EU, U.S. Army Research Office, and the National Science Foundation, USA.

DIMITAR TRAJANOV (Member, IEEE) received the Ph.D. degree. He is currently a Visiting Research Professor at Boston University and the Head of the Department of Information Systems and Network Technologies, Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje. He is also the Leader of the Regional Social Innovation Hub established, in 2013, as a co-operation between UNDP and the Faculty of Computer Science and Engineering. From March 2011 to September 2015, he was the Founding Dean of the Faculty of Computer Science and Engineering, and in his tenure, the Faculty has become the largest technical Faculty in Macedonia. He has been involved in more than 70 research and industry projects, of which in more than 40 projects as a Project Leader. He is the author of more than 170 journal and conference papers and seven books. His research interests include data science, machine learning, NLP, FinTech, semantic web, open data, sharing economy, social innovation, e-commerce, entrepreneurship, technology for development, mobile development, and climate change. From 2012 to 2015, he was a member of the National Committee for Innovation and Entrepreneurship, where he worked on strategies and legislation to encourage innovation and entrepreneurship in the Republic of Macedonia through inspiring the use of technology for economic development.

• • •