A Type-2 Fuzzy Logic Approach to Explainable AI for regulatory compliance, fair customer outcomes and market stability in the Global Financial Sector

Janet Adams
Business Banking
TSB Bank
London, UK

Hani Hagras
The Computational Intelligence Centre, School of Computer Science and Electronic Engineering
University of Essex, Colchester, UK

***Abstract*—The field of Artificial Intelligence (AI) is enjoying unprecedented success and is dramatically transforming the landscape of the financial services industry. However, there is a strong need to develop an accountability and explainability framework for AI in financial services, based on a risk-based assessment of appropriate explainability levels and techniques by use case and domain.**

This paper proposes a risk management framework for the implementation of AI in banking with consideration of explainability and outlines the implementation requirements to enable AI to achieve positive outcomes for financial institutions and the customers, markets and societies they serve. The work presents the evaluation of three algorithmic approaches (Neural Networks, Logistic Regression and Type 2 Fuzzy Logic with evolutionary optimisation) for nine banking use cases. We review the emerging regulatory and industry guidance on ethical and safe adoption of AI from key markets worldwide and compare leading AI explainability techniques.

We will show that the Type-2 Fuzzy Logic models deliver very good performance which is comparable to or lagging marginally behind the Neural Network models in terms of accuracy, but outperform all models for explainability, thus they are recommended as a suitable machine learning approach for use cases in financial services from an explainability perspective. This research is important for several reasons: (i) there is limited knowledge and understanding of the potential for Type-2 Fuzzy Logic as a highly adaptable, high performing, explainable AI technique; (ii) there is limited cross discipline understanding between financial services and AI expertise and this work aims to bridge that gap; (iii) regulatory thinking is evolving with limited guidance worldwide and this work aims to support that thinking; (iv) it is important that banks retain customer trust and maintain market stability as adoption of AI increases.

Keywords— Regulatory Compliance; Accountability and Explainability; Type-2 Fuzzy Logic; Neural Networks

I. INTRODUCTION

The World Economic Forum (WEF) in their August 2018 paper on AI in Financial Services [1] recognised that AI appears set to enjoy an unprecedented run of success and dramatically transform the landscape of the financial services industry. In their follow up report on deploying AI responsibly [2] the WEF observe that early AI adopters will reap rewards but risk customer backlash and regulatory alarm, however that "'trusted AI' can be a competitive differentiator". A survey published by the Bank of England (BoE) and the Financial Conduct Authority (FCA) in Oct '19 [3] found increasing use of Machine Learning (ML) in financial services with applications forecast to double in the next three years, but that firms' model validation and risk frameworks need to evolve in line with the complexity of AI. In the survey, firms report a need for regulatory guidance on use of AI, a sentiment echoed by the Treasury Select Committee report on IT banking failures [4] which includes a statement that "The important thing for the regulator is that these [models] cannot be a black box" [4] and states that firms should not deploy technologies such as AI if risks cannot be rigorously identified and mitigated, urging the Regulators to set clear guidance.

AI brings unprecedented opportunities as a technological wave to benefit consumers, markets and society. With these powerful tools, banks can provide more personalised and targeted products and services to consumers and businesses, enabling them to fulfil their goals and ambitions. Banks can provide faster, better and more streamlined products and services, smoothing customer friction points and enabling operations at lower costs, with a corollary benefit to society overall. With AI, banks can better protect customers and society from bad actors in fraud, money laundering, terrorist financing, and help to prevent other financial and humanitarian crimes. However, in order to be able to harness these technologies to realise such benefits, the industry requires a clear and simple articulation of the ethical and conduct requirements for safe AI adoption, along with a comprehensive explainability framework. This framework must recognise the need for 'sufficient explainability' as proposed in an FCA insight paper [5] on a use case specific basis, identifying high and low risk domains for AI, and clearly outlining the appropriate accompanying risk management and control frameworks.

However, the Age of AI heralds new risks to an industry that at its core is based on managing risk. New approaches to governance are urgently required, and another FCA insight paper [6] notes that "Boardrooms are going to have to learn to tackle some major issues emerging from AI – notably questions of ethics, accountability, transparency and liability'. Emerging or changing risks related to use of AI in financial services include but are not limited to:

• Loss of confidence in an industry that has re-built fragile trust since the most recent financial crisis of 2008 and can ill afford

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

any further major scandal.
• Operational risks (with accompanying financial loss) of deploying unprecedented algorithmic solutions at scale.
• Legal and compliance risks in deployment of algorithms and data in new ways, with limited understanding of AI across second and third lines of defence.
• Understanding and application of existing regulations to algorithms in early adoption could give rise to regulatory censure in future as un-thought through implications of AI under existing regulations crystallise in time.
• Reputational risks that may emerge as society's expectations of what constitutes acceptable use of data evolves; what is acceptable today may not be acceptable in a year's time.

In order to effectively implement AI in financial services, these risks and many more must be addressed and mitigated with accountable risk and control frameworks.

Since the 2008 financial crisis, banks have invested heavily in enhancing the culture and conduct at firms to promote market stability and ensure fair customer outcomes, with hundreds of billions of dollars having been levied in fines on the industry [7]. It is a natural evolution of this conduct capability to encompass the new ethical questions being posed in relation to AI. As part of the research for this paper, 20 ethical guideline publications or speeches were reviewed from multiple jurisdictions and a risk and control framework for ethical adoption of AI in financial services was developed, which is presented in section II of this paper. The overarching principle of the framework is that of accountability and explainability, supported by key pillars of transparency, robustness, human autonomy, fairness, ethics, and conduct, underpinned by strong governance, risk and controls. The framework endorses the use of existing controls and delegated authorities for AI based decisions that are clearly explainable by design, with proposed controls and escalation methods for decisions taken outside of recommended explainability boundaries. The framework contends that use of more complex models such as neural networks is not prohibited, but that the additional cost of compliance combined with the reduction of human empowerment that accompanies black box models, even with explainable techniques, must be considered and justifiable.

While there are a number of key features of AI that are essential for its successful adoption in financial services, including transparency (the ability to look within the model and see what data is in it and how it was designed) and robustness (including reliability, repeatability, and scalability), the keystone for safe implementation of AI is explainability. Explainable AI (XAI) was first coined as a phrase by DARPA who outlined in their 2016 paper [8] the approaches to explainability at the time and introduced the concept that for AI to be effectively adopted by users, first it must be able to explain itself. This topic is the subject of significant research focus across the globe giving consideration to model-specific and model-agnostic techniques, as well as to explainability applied to a global data set (i.e. the entire model) or to individual decision instances (local explainability).

In this paper, we will present a new Type-2 Fuzzy Logic approach for implementation of AI in banking. We will show that when optimised with evolutionary algorithms, Type-2 Fuzzy Logic delivers very good performance, comparable to or lagging marginally behind the Neural Networks in terms of accuracy, but outperforms all models for explainability and is recommended as a suitable machine learning approach for all automated decision making use case domains in financial services.

In Section III, we provide a brief overview of Type-2 Fuzzy Logic Systems (FLSs) while the topic of explainability is discussed in Section IV. Section V will present the experiments and results while the conclusions and future work are presented in Section VI.

II. Ethical and Regulatory considerations for AI in financial services

Regulators world-wide are considering regulation of AI, and draft ethical guidelines or speeches have been published in most major jurisdictions (Table 1 contains the full list of reviewed guidelines). Central to the regulation debate is explainability; to what extent do regulators require algorithms to be explainable, and in which contexts or domains is explainability particularly important. Accountability and explainability are inextricably intertwined; how can a firm or senior manager be accountable for something that cannot be explained?

TABLE I. Publications mined for Ethical AI Requirements

Ref.	Publications mined for Requirements for Ethical AI	
	Issuing Body	***Publication***
[25]	US Government / Senate	Algorithmic Accountability Act of 2019
[26]	Australian government	Artificial Intelligence: Australia's Ethics Framework Discussion Paper
[27]	Information Commissioners Office	GDPR
[28]	European Commission	Ethics Guidelines for Trustworthy AI
[29]	BaFin Federal Financial Supervisory Authority	Big data meets artificial intelligence – results of the consultation on BaFin's report
[30]	House of Lords	AI in the UK, Ready, Willing and Able?
[31]	French Government	For a meaningful artificial intelligence: Towards a French and European strategy
[32]	OECD	Recommendation of the Council on Artificial Intelligence
[33]	IEEE	Ethically Aligned Design (EAD1e), A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems
[34]	European Parliament	European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))
[35]	The Public Voice coalition, Brussels	Universal Guidelines for Artificial Intelligence
[36]	Bank of England (BoE)	Managing Machines: the governance of artificial intelligence

Ref.	Publications mined for Requirements for Ethical AI	
	Issuing Body	*Publication*
[37]	Securities and Exchange Commission (SEC)	Guidance Update: ROBO-ADVISERS
[38]	Government of China	Beijing AI Principles
[39]	Japanese government	AI Policy Japan
[40]	Indian government	National Institute for Transforming India. 2018. National strategy for artificial intelligence.
[41]	Monetary Authority of Singapore (MAS)	Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector
[42]	Office of the Privacy Commissioner for Personal Data	Ethical Accountability Framework for AI (Hong Kong)
[43]	AI Now	AI Now 2018 report
[44]	Financial Conduct Authority (FCA)	The future of regulation: AI for consumer good
[45]	De Nederlandsche Bank (DNB)	General principles for the use of Artificial Intelligence in the financial sector

From these publications, implementation requirements have been derived to ensure appropriate conduct, compliance and operational resilience. These requirements together form the framework depicted in Fig. 1. For each of the framework component areas, the level of explainability of the algorithms deployed has a direct impact on the organisation's ability to meet the requirement to satisfactory levels.

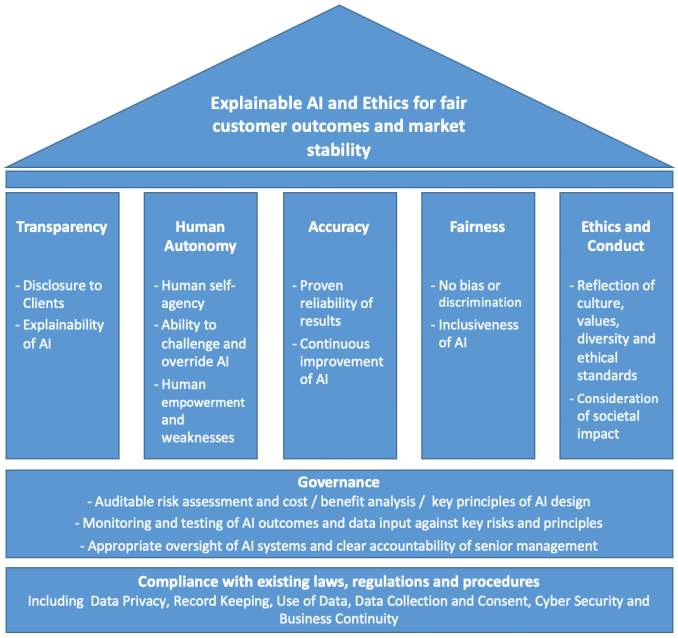

Fig. 1. Accountability and Explainability Framework

This paper proposes that explainability is the master key that will unlock AI benefits for the financial services industry, and that accountability is the single most important governance and ethics principle. Other key principles include:

- Compliance with all existing regulations is required.
- Existing delegated authority levels are sufficient for management of AI risk where the model is explainable by design (see section III) but decisions made using opaque models and interpretive explainability methods require more senior level sign off, and additional monitoring and controls.
- Education and upskilling are required across the full AI implementation lifecycle for governance and monitoring methods to be truly effective.
- The five pillars of Transparency, Human Autonomy, Robustness, Fairness, Ethics and Conduct, require translation and integration into the organisation's existing policies.
- The principle of human self-agency [9], respect, and non-abuse of information asymmetry required for retail AI use cases.
- "Where explainability is provided by means of an interpretive layer or model rather than as an intrinsic model property, it must be tested sufficiently well to ensure it gives accurate, unbiased outputs" [10].

III. Type-2 Fuzzy Logic and Explainability Methods

In financial services, there is a general lack of recognition of the power and potential of FLSs for human XAI [11]. Pioneered by Lofti Zadeh more than half a century ago, there has been in increase in research recently into the application of Type 2 Fuzzy logic to real-world problems, where handling uncertainty is key [12]. "FLSs attempt to mimic human thinking, although rather than trying to represent the brain's architecture as you would with a neural network, the focus is on how humans think in an approximate rather than precise way, creating a set of linguistic if-then rules to describe a given behavior in human-readable form" [11]. This ability of FLS to model uncertainty, learn from data, and achieve a good balance between prediction accuracy with explainability makes it an important XAI methodology [11].

Fuzzy Logic was developed with the objective of creating a rules-based logic that could deal with uncertainty and output values in the whole range of {0, 1} [13], which allows Fuzzy Logic to more accurately model real-life scenarios, where uncertainty is key. Fuzzy Logic generates a smoother boundary between the mapping of inputs and outputs, whereas a Boolean logic system can produce different outputs with very similar inputs.

Type 2 Fuzzy Logic is an extension of Type 1 Fuzzy Logic that allows more uncertainty levels to be handled in the membership functions of the fuzzy sets via the Footprint of Uncertainty (FOU) and the third dimension of the Type-2 Fuzzy set, which enables a smoother performance to be obtained than with Type-1[14].

This ability to handle high levels of uncertainty makes it ideal for use in many financial services use cases.

Research into Evolutionary Fuzzy Systems (EFS) has been progressing since the 1990s [15, 16]. The work reported in this paper is based on the Temenos XAI platform which employs EFSs that generate from a big number of inputs and huge data

sets, a small number of short IF-Then rules which can be easily understood, analysed and augmented by the business user.

IV. Explainability

Explainability is a topic of considerable research and a number of methods and approaches have been developed; these can be represented along two key axes as depicted in Fig. 2 [17]. These methods are described, and their advantages and limitations discussed below, considering model-specific and model-agnostic explainability as it applies to global datasets (i.e. the entire dataset or portfolio) and local data points (i.e. individual decisions). There is acknowledged to be some confusion about what explainability or interpretability means, [18], while Doshi-Velez and Kim also note that "there is little consensus on what interpretability in machine learning is and how to evaluate it for benchmarking" [19].

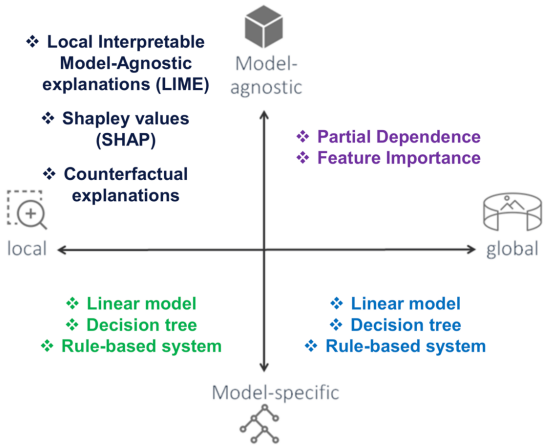

Fig. 2. Technical Solutions for Explainability [15]

A. Model-Specific Explainability

In model-specific explainability, a model is designed and developed in such a way that it is fully transparent and explainable by design. In other words, an additional explainability technique is not required to be overlaid on the model in order to fully explain its workings and outputs. In general, explainable models are simpler than non-explainable models and as such their performance in terms of accuracy can be relatively diminished. In the case of Type 2 Fuzzy Logic combined with evolutionary optimisation however, the performance can be seen to approach that of more complex opaque models, as demonstrated in the use case results. Explainable models include linear regression, decision trees, and rule-based systems. These models have different limitations and advantages as follows.

Linear models work well when there is a linear relationship between the inputs and outputs, however linear regression models are unable to perform well on non-linear problems.

Decision Trees (DTs) are easy to use and understand, handle different types of data well and work well when there are relatively few features for simple classification tasks. However due to their simplicity, DTs tend to have low accuracy.

Both the DTs and linear regression models can be seen as black box models for high-dimensional spaces.

B. Model-Agnostic Explainability

A mathematical explainability technique can be applied to the outputs of any AI model including very complex and opaque models, to provide an interpretation of the decision drivers.

One technique for model-agnostic explainability is Feature importance which is based on a simple concept; scramble/corrupt the data for one of the features in a model and observe the impact on the error rate of the model. This can be done for every feature in the data set one by one to observe how important each feature is to the prediction and rank them accordingly. However, this technique is computationally expensive and time-consuming, and has a significant limitation of working in a univariate way, i.e. exploring the importance of each feature individually and not accounting for how some features might interact with each other.

Partial dependency works well alongside feature importance and shows the marginal effect that one or two features have on the predicted outcome of a machine learning model [20], and whether the relationship between the target class is linear or more complex [21]. A partial dependency plot enables visualisation of the impact of ranges in the feature on the final model. However, this technique works only with two features at a time, and the assumption of independence is a significant limitation [21].

The goal of Local Interpretable Model-Agnostic Interpretation (LIME) "is to identify an interpretable model over the *interpretable representation* that is locally faithful to the classifier" [22]. This allows a local approximation of why a class was assigned to a particular data point but does not perform well with outliers.

Shapley explanations (SHAP), "are a class of additive, locally accurate feature contribution measures with long-standing theoretical support" [23, 24]. This works best in datasets with very few features, as the number of permutations of features rapidly becomes computationally expensive. Another limitation is that features cannot be accurately assumed to always be conditionally independent of each other.

The technique of counterfactual explanations considers how the model would behave if some features had different values [22]. However, repeating this process with each feature is computationally expensive.

A general limitation of all model-agnostic explainability techniques is that they entail running an additional model on top of an already complex model. The explainability technique will never be 100% accurate and so a layer of additional inaccuracy is introduced, and the output becomes one step further removed from reality. A second drawback of these approaches is that when deployed, some end users may need to understand how these models operate as well as understanding the underlying model if required to explain to regulators, customers or other stakeholders.

From the above it can be seen that model-specific explainability is the optimum approach, if sufficient accuracy can be achieved in the model. This is the approach which most easily addresses the requirements of the emerging regulatory and ethical guidance, and the cost of implementation for financial services firms is materially lower than the use of model-agnostic approaches. The additional costs of model-

agnostic approaches include computational resource, time and cost of additional analytics layer, potential time delays in producing explanations if requested, and the additional compliance and controls that must be in place to manage the introduction of an additional layer of error placed on top of a model that is already by definition not 100% accurate.

V. Experiments and Results

A. Use cases, data and methodology

Research was conducted on the Temenos XAI platform under a research licence kindly provided by Temenos. The Temenos XAI platform is a cloud based, API driven 'machine learning as a service' predictive analytics platform which enables Neural Networks (NNs), Logistic Regression (LR), and Fuzzy Logic (FL) models to be built, configured and run. The three models serve different purposes in the toolkit. The LR model provides a statistical baseline and in datasets where the relationship between the features is linear or broadly linear, will provide strong model results. The NN is an opaque model which has been shown in many use cases and research papers to achieve very high levels of accuracy and as such is often the algorithmic method of choice for use cases where the opacity of the model's workings does not pose a problem. Lastly, the FL model is a fuzzy rules based approach with built in explainability by design, combined with evolutionary optimisation employed to maximise the model interpretability and performance. The FL models enable full understanding of the factors influencing the predictive outcome, and full global and local explainability for each model and data instance output respectively.

As shown in Table II, we have explored nine use cases using publicly available data; four retail/consumer banking use cases and five wholesale/trading use cases. Binning was applied to selected features in the retail models, therefore eight models were developed for retail (one binned and one unbinned for each case). For each case, 10 versions of the LR, NN, and FL algorithms were executed.

TABLE II. Retail and Wholesale Banking Use cases developed

Use Cases	Use Cases and datasets		
	Use case	***No of features***	***Classification Goal***
Retail	Propensity to Buy (PTB)	21 features: 11 categorical, 9 continuous, 1 mixed	To predict if the client will subscribe to a term deposit
	Churn Modelling	11 features: 6 categorical, 5 continuous	To predict whether a customer will exit the bank
	Loan Default	21 features: 18 categorical, 3 continuous	To predict whether a customer will default on a loan
	Credit Card Default	24 features: 10 categorical, 14 continuous	To predict whether a customer will default on a credit card
Whole-sale	FX Price Change NZD:USD	35 features: 1 categorical, 34 continuous	To predict whether the cost of buying USD with NZD will go up or down
	FX Price Change USD:CHF	35 features: 1 categorical, 34 continuous	To predict whether the cost of buying CHF with USD will go up or down
	Gold:USD Price Change	35 features: 1 categorical, 34 continuous	To predict whether the cost of buying gold with USD will go up or down
	Bitcoin:USD Price Change	35 features: 1 categorical, 34 continuous	To predict whether the cost of buying BTC with USD will go up or down
	Nikkei Index Price Prediction	35 features: 1 categorical, 34 continuous	To predict whether the price of the Nikkei index will go up or down

B. Model results discussion – Retail use cases

On average across the 8 models the Neural Networks marginally outperformed the Logistic Regression and Fuzzy Logic models across all three measures of accuracy, recall, and precision (as shown in Table III) in both test and whole data sets. In the retail use cases, recall is a very important performance measure. The Fuzzy Logic models had the best recall in four out of eight of the models assessed for the four use cases, and the performance difference with the Neural Networks for recall across all models on average was marginal.

TABLE III. Retail Results – Averages across use cases

Model	**Retail Model Performance Results; average across all use cases**		
	Accuracy %	***Avg. Precision %***	***Avg. Recall %***
LR - Test	78.31375	70.040625	77.423125
NN - Test	79.010625	71.16525	79.083625
FL - Test	77.0495	69.871875	78.93125
LR - Whole	78.4205	69.843	76.89125
NN - Whole	79.25925	71.030625	78.680625
FL - Whole	77.185625	69.78475	78.63425

The Temenos XAI platform allows the user to display the most important rules derived from the dataset as shown below in Fig. 3 using Credit Card Default as an example.

Fig. 3. Top global rules for Credit Card Default use case

These rules are displayed in order of dominance (which encompasses the rule confidence and support). Frequency is defined as the number of times that each rule is fired, while Winning indicates the number of times that it has been the most important rule for determining the output class.

The rules show that the dominant predictor of whether a customer will default next month or not is the status of the bank account, rather than personal factors such as age, gender, education level, or marital status. The most dominant feature

predicting that the customer will default next month is the arrears status of previous months, which is present as a single or joint antecedent in all 7 of the top default rules. Likewise, for predicting that a customer will pay, an account status of 'up to date' is present in all of the 'Pay' rules in the top ten dominant rules. This global explainability allows any auditing and augmentation by the business user/auditors before the model deployment.

The platform also provides local explainability for each data instance, an example of which is shown below in Fig. 4.

The drivers in this instance all make sense to the user, and are very consistent with the global explanation. The top four drivers can be seen to be whether the account status was in arrears in previous months. Other drivers such as education being 'graduate' would require further analysis and investigation as this is counter intuitive, and therefore the rules underlying the drivers can be further considered and drilled into.

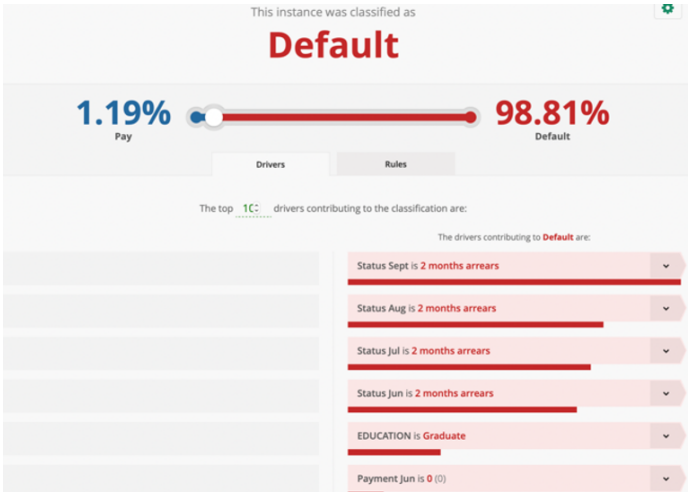

Fig. 4. Fuzzy Logic instance drivers for example of 'Default' classification in Credit Card model.

C. *Model results discussion – Wholesale use cases*

For the wholesale trading price predictions, precision was selected as the most important performance measure; the percentage of true positives identified over all positives identified will indicate how many winning trades can be placed (note that the trading strategy simply buys when the response variable is predicted to be one, i.e. price going up the following day, so true positives are the target of the model). For each of the models developed, ten versions each of LR, NN, and FL models were executed. The average results across all five use cases are tabulated below in Table IV.

The FL models were the best performing models overall from a precision perspective in the whole data sets, with the NN leading for test data precision. The Fuzzy Logic models significantly outperformed the other models on the whole data set across all three performance measures, with the Neural Networks leading on test data results more marginally.

TABLE IV. WHOLESALE RESULTS – AVERAGES ACROSS USE CASES

Model	Wholesale Model Performance Results; average across all use cases		
	Accuracy %	*Avg. Precision %*	*Avg. Recall %*
LR - Test	54.681	54.976	54.8632
NN - Test	56.3026	59.2652	55.9336
FL - Test	55.187	55.8492	55.3404
LR - Whole	56.8034	56.8808	56.8732
NN - Whole	54.2348	56.322	53.783
FL - Whole	60.1836	62.262	60.657

The explainability of the Fuzzy Logic approach has a distinct advantage in a trading scenario where the trader will want to know not just what it is the model is telling them, but why. The trader wants to know how strong the signal is and how broad based; whether it is based on a multitude of variables or just a few, and can use this information to inform the ultimate decision when balanced with idiosyncratic risk. The FL models provide a clear breakdown of rules and drivers for decisions; an example of which is shown in Fig. 5 below depicting the top four dominant rules for the NZD/USD price change use case.

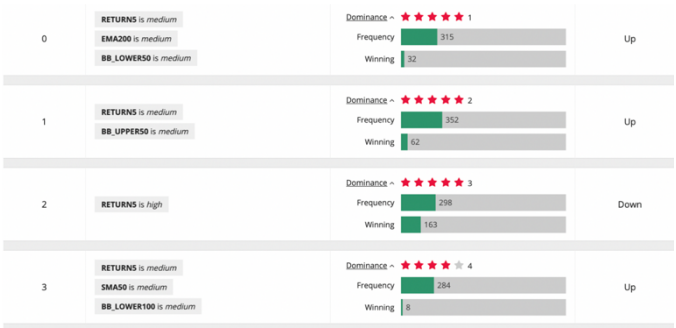

Fig. 5. Top global rules for NZD/USD use case

Here it can be seen that the most dominant rule, which presents in 315 (out of 1001) of the instances is that if the 5 day return is medium, the 200 day EMA (Exponential Moving Average) is medium, and the 50 day BB-lower (lower Bollinger Band) is medium then the price is likely to close up the following day. Local explainability can also be provided for each individual trading instance. This enables the trader to consider the human knowledge they have about idiosyncratic risk factors - such as economic data, political events, and information from the central banks of the two countries in the currency pair - to form an algorithmically empowered human decision on whether to trade or not.

D. *Model results discussion – Explainability and compliance with risk and control requirements*

The NNs and FL models and outputs were assessed against the proposed Accountability and Explainability framework, in order to assess the degree to which the requirements of the framework can be met by each of these two algorithmic approaches. The LR model was excluded from this analysis due to loss of explainability of the LR models once the number of features exceeds approximately 20.

The models were assessed against the sub-category level of the framework, with consideration to the individual

requirements underpinning these. The FL models outperform the NNs in respect of their ability to support the requirements of every sub-category of the framework, with the potential exception of model performance in respect of accuracy, precision and recall where both models have strengths and the NNs in general as seen in the research above, can be stronger in these measures. However, in all other categories the FL ease of explainability supports the compliance, governance, risk assessment, oversight, monitoring and controls, disclosure, explainability, autonomy, human empowerment, sustainability, scalability, conduct, culture, and societal impact requirements for effective, safe, and compliant implementation of AI in financial services to a significantly greater extent than a NN model. There were some differences between considerations for the retail and wholesale use cases, for example consumer protection and GPDR are not applicable in the wholesale cases and as such the requirement for explainability is reduced in some areas of the framework, but financial and prudential risks are amplified in the wholesale sector, and the importance of explainability to meet the requirements for monitoring and controls, fallback and robustness, to protect against 'herd and cascade' behaviours, and to ensure openness and transparency of culture, is increased.

VI. Conclusions and Future Work

In this paper, we have shown that Type-2 Fuzzy Logic has the capability to deliver strong performance in terms of recall, precision, and accuracy, and has a tremendous advantage over other algorithmic approaches in that it is capable of deriving human understandable interpretations of the factors underlying the decision making of the algorithm, both at global and local levels. Type-2 Fuzzy Logic with evolutionary optimisation should be explored as an algorithmic method of choice for the financial services industry, particularly for supervised learning problems.

We argue that firms should consider building AI capability from the outset with model-specific explainable methods which can be understood and safely adopted.

It is recommended that firms with ambitions to deploy AI at scale develop and rollout a global AI Accountability and Explainability Framework, such as that proposed in this paper.

A risk-based approach to explainability is proposed, as outlined in this paper. A reasonable starting benchmark should be that every customer matters and any significant customer-impacting decision being driven or informed by AI methods should be clearly explainable by design, with exception handling, human in the loop and risk controls in place for any deviations from this principle.

Issues of diversity in the workplace, particularly in data- and technology-focussed areas, must be addressed as a priority by financial services firms wishing to grow expertise in use of AI, to avoid group think and ensure that AI is adopted and harnessed in ways that benefit all of society.

The scope of this work was limited to publicly available data and future work would benefit from use of more extensive banking data sets, and testing of further financial services use cases for suitability of explanation technique. Future work could expand the range of algorithmic approaches evaluated to include other popular algorithmic methods such as decision trees and random forests, K Nearest neighbour, Support Vector Machines, K means clustering, and reinforcement learning. The scope could also be expanded to include regression as well as classification challenges.

Acknowlegement

The authors would like to acknowledge the support of Temenos for providing us with a research instance from the Temenos AI platform. We would like also to acknowledge the support of Adriano Koshiyama at UCL on explainability and Eva Lueckemeier of HSBC Thailand for work on the risk framework.

VII. References

[1] R. J. Waters et al, "The New Physics of Financial Services," World Economic Forum, 08/15/2018

[2] R. J. Waters et al, "Navigating Unchgarted Waters; A roadmap to responsible innovation with AI in financial services" World Economic Forum, 23/10/2019

[3] C. Jung, H. Mueller, S. Pedemonte, S. Plances, O. Thew, "Machine learning in UK financial services" BoE and FCA, 16/10/2019

[4] Committee staff, " IT Failures in the Financial Services Sector, House of Commons Treasury Committee, 28/10/2019 https://publications.parliament.uk/pa/cm201919/cmselect/cmtreasy/224/224.pdf

[5] K. Croxson, et al, "Explaining why the computer says 'no,'" FCA, 05/31/2019

[6] M. Faulk, "Artificial Intelligence in the boardroom," FCA, 08/01/2019 https://www.fca.org.uk/insight/artificial-intelligence-boardroom

[7] S. Chatterjee, "U.S., EU fines on banks' misconduct to top $400 billion by 2020" Reuters Business News, September 2017

[8] D. Gunning, "Explainable Artificial Intelligence," Darpa, 08/2016, https://www.darpa.mil/attachments/XAIIndustryDay_Final.pptx

[9] M. Taddeo and L. Floridi, "How AI can be a force for good," Science, Vol. 361, August 2018

[10] A. Weller, "Challenges for Transpar-ency," Arxiv:1708.01870, 07/2017, https://arxiv.org/pdf/1708.01870.pdf

[11] H. Hagras, "Towards Human Understandable Explainable AI", *IEEE Computers*, Vol.51, No.9, pp. 28-26, September 2018

[12] A. Shukla, S.K. Banshal, T. Seth, A. Aparna, R. John, P. Muhuri. A Bibliometric Overview of the Field of Type-2 Fuzzy Sets and Systems. IEEE Computational Intelligence Magazine. 15. 89 - 98. 10.1109/MCI.2019.2954669, January 2020

[13] H. Hagras, C. Wagner, "Introduction to Interval Type-2 Fuzzy Logic Controllers - Towards Better Uncertainty Handling in Real World Applications", *The IEEE Systems, Man and Cybernetics eNewsletter, Issue 27, June 2009*

[14] J. Mendel, Type-2 Fuzzy Sets and Systems: a Retrospective. *Informatik Spektrum* 38**,** 523–532 (2015). https://doi.org/10.1007/s00287-015-0927-4

[15] A. Starkey, H. Hagras, S. Shakya, G. Owusu, "A Multi-Objective Genetic Type-2 Fuzzy Logic Based System for Mobile Field Workforce Area Optimization" *Journal of Information Sceinces*, Vol. 333, pp. 390-411, September 2016

[16] A. S. Koshiyama et al, "Automatic synthesis of fuzzy systems: An evolutionary overview with a genetic programming perspective," Wiley Interdisciplinary Reviews, Vol.9, No. 2, May 2018

[17] A. Koshiyama and Z. Engin, "Algorithmic Impact Assessment: Fairness, Robustness and Explainability in Automated Decision-Making," Presentation (Open Access) 06/08/2019

[18] W. J. Murdoch "Interpretable machine learning: definitions, methods, and applications" arXiv 1901.04592, UC Berkeley, Allen Institute for Brain Science. January 2019

[19] F. Doshi-Velez, B. Kim, "A roadmap for a rigorous science of interpretability," arXiv:1702.08608, 2017

[20] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of statistics, 2001

[21] C. Molnar, "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable," 2019 https://christophm.github.io/interpretable-ml-book/

[22] M. T. Ribeiro, S. Singh, and C. Guestrin,"Model-agnostic interpretability of machine learning," arXiv, 1606.05386, 2016

[23] P. Hall, "On the Art and Science of Machine Learning Explanations." arXiv, 1810.02909 08/02/2019 https://arxiv.org/pdf/1810.02909.pdf

[24] M. L. Scott, Su-In Lee, "A Unified Approach to Interpreting Model Predictions." Advances in Neural Information Processing Systems, *31st Conference on Neural Information Processing Systems (NIPS 2017),, California, USA, December 2017*

[25] Wyden, "Algorithmic Accountability Act of 2019," Senate of the United States, 2019

[26] Dawson D et al, "Artificial Intelligence: Australia's Ethics Framework," Australian Government, 2019

[27] "EU General Data Protection Regulation," Information Commissioner's Office, 04/2016. http://www.privacy-regulation.eu/en/index.htm

[28] European Commission's High-Level Expert Group on AI, "Ethics guidelines for trustworthy AI," European Commission. August 2019

[29] J. Bartels, T. Deckers, "Big data meets artificial intelligence – results of the consultation on BaFin's report," BaFin, March 2019

[30] House of Lords Select Committee on Artificial Intelligence, "AI in the UK: ready, willing and able?" Authority of the House of Lords, Report of Session 2017–19

[31] C. Villani, "For a meaningful artificial intelligence: towards a French and European strategy," AI for Humanity, 2018

[32] OECD Principles on Artificial Intelligence, OECD, 2019 https://www.oecd.org/going-digital/ai/principles/

[33] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically Aligned Design: A Vision for Prioritising Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019

[34] European Parliament resolution, with recommendations to the Commission on Civil Law Rules on Robotics, 02/16/2017

[35] "Universal Guidelines for Artificial Intelligence," The Public Voice coalition, Brussels, October 2018

[36] J. Proudman, "Managing machines: the governance of artificial intelligence," FCA Conference on Governance in Banking, London, 06/04/2019

[37] "Guidance Update on Robo-Advisors," US Securities and Exchange Commission Division of Investment Management, Februray 2017

[38] "Beijing AI Principles," Beijing Academy of Artificial Intellignce, May 2019

[39] "Global AI Policy," The Future of Life Institute, webpage as of 08/26/2019; https://futureoflife.org/ai-policy-japan/?cn-reloaded=1

[40] A. Roy et al, "National Strategy for Artificial Intelligence," NITI Aayog, https://www.niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf?utm_source=hrintelligencer

[41] "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector," Monetary Authority of Singapore, 11/12/2018

[42] "Ethical Accountability Framework for Hong Kong, China" Privacy Commissioner for Personal Data, Hong Kong, 10/01/2018

[43] M. Whittaker et al, "AI Now Report 2018," AI Now, December 2018.

[44] C. Woolard, "The future of regulation: AI for consumer good," FCA speech, London, July 2019 https://www.fca.org.uk/news/speeches/future-regulation-ai-consumer-good

[45] J. van der Burgt, "General principles for the use of Artificial Intelligence in the financial sector," DeNederlandscheBank, 2019