AI explanibility and fairness in the Quantum realm

Vikas Sawant
er.sawant@gmail.com

Pooja Deshmukh
poojagunjal@gmail.com

Harshil Bhatt
bhattharshil307@gmail.com

Aman Jaswani
ajaswani11@gmail.com

Abstract

Financial services offered by most institutions are highly driven by different factors like level of engagement, roles, impact towards processes, customer engagement etc. This leads to inject couple of biases which severely affects the Quality of Service (QoS) metrics. With recent technology like AI/ML, finance industry can see many solutions which can help to improve the QoS. On this path, in the presence of terabytes of data and millions of per unit hour events, it is difficult to adopt traditional ML approaches. Deep Learning (DL) comes to the rescue but it is not very explainable to the regulators/users. On this pendulum, we need better solution proposal which uses highly explainable and efficient AI approach which meet both the metrics. In this paper we explore two use cases - one, to determine the fairness of the loan lending process, and second, to evaluate the understandability and observe the bias (if any), of quantum AI for stock projection.

1 Introduction

All major businesses, especially technology related ones, are being revolutionized with the help of new, state of the art Artificial Intelligence and Machine Learning techniques. The financial services sector is no stranger to these solutions and uses modeling approaches at every step these days. AI/ML is being used by all major financial institutions for dealing with a variety of business problems such as fraud detection, robo advisory, risk management, deciding mortgage/loan rates etc.

While the ML-backed solutions can be deemed to be more robust, there prevails a need to understand whether these models are fair and explainable. Fairness deals with the aspect that whether our models and predictions discriminate on the basis of some acquired parameters. These parameters can be associated (but not limited to) with age, race, gender, education background, social status, inherited traits, etc. The distinction between fairness and discrimination is an important one. Discrimination is the action taken based on bias, whereas fairness is a lack of bias. Explainability deals with the issue of how understandable a solution is to the end users or the regulators of the system. With neural networks taking over the standard machine learning models, there has been an aggravation in the explainable nature of these products. This boils down to the fact that such complex models fall under the ‘black-box’ category.

In this paper we explore both these aspects for systems built by financial institutions. First we explore how fair are the loan decision making systems that are used by banks to decide whether a person is likely to default on their payments or not. These systems take into account some acquired parameters such as age, gender, marital status and educational background in their predictions. We experiment upon whether changing these acquired parameters (one at a time) changes the output of the ML model or not to check whether the model makes its predictions in a fair and unbiased way or not. Secondly, we explore a fairly popular use case of the finance industry - stock projection. There are already various state of the art ML and DL models for the same. We wish to go a step further and dive into

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

the quantum aspect of computer sciences. We have created a quantum neural network for the same and investigated how it performs as compared to the existing work, and also, how explainable our solution is, and whether any bias is observed in our solution or not.

2 Loan Decision Making

There are enough number of events available with financial institutions to understand the behavior of loan default cases to prevent a company from letting into trouble. Suppose we generated a model from data and the model revealed that one of customer might be a defaulter soon, there should be a logical explanation to why our model predicted so. The model can predict who will be a defaulter from given data easily, however the model doesn't tell why the person has been classified so and what measures can be taken to prevent him from getting defaulted. To circumvent the possible consequence, we would want to ask the model "What if" he had planned his financials well or "What if" he looks for a greater number of income sources.

Adoption of DL based solutions on priority even in presence of other alternatives like boosting and random forest was thought to reduce reliance on subjective opinions by humans but here's a different distinct hurdle that we have to overcome when we adopt such sorts of AIs. Those explainable and interpretable reasons have become invisible in "black-box" models' outputs. Luckily it is evident from few research outcomes that it is now doable to obtain what-if explanations from model prediction with counterfactual explanations like "What if" hypothetical examples that we wanted to cast a question for the same input [3].

Use case example: Suppose a consumer applies for a loan and gets declined by a ML decision model. Counterfactual models work by designing a "digital quasi-twin" that is as close as possible to the profile of that user but would get a different decision. This helps identify variables that explain the decision of the model, and that can be validated by human experts and used to refine both new and classical models.

Counterfactuals are human-friendly explanations, because they are contrastive to the current instance and because they are selective, meaning they usually focus on a small number of feature changes. But counterfactuals suffer from the "Rashomon effect". Rashomon is a Japanese movie in which the murder of a Samurai is told by different people. Each of the stories explains the outcome equally well, but the stories contradict each other. The same can also happen with counterfactuals, since there are usually multiple different counterfactual explanations. Each counterfactual tells a different "story" of how a certain outcome was reached. One counterfactual might say to change feature A, the other counterfactual might say to leave A the same but change feature B, which is a contradiction. This issue of multiple truths can be addressed either by reporting all counterfactual explanations or by having a criterion to evaluate counterfactuals and select the best one [4].

For our experiment we took past data of whether a person has defaulted on their installment payments from [6]. This dataset consists of 30000 borrowers, also comprising the age, gender, marital status, and educational background of the person along with the various bill payments and the outcome variable depicting whether the person is credit worthy or not. We create a train-test-split and train an artificial neural network on this data. On the testing data (consisting of 6000 samples), we create gender flipped counterfactuals using DICE library for each sample and make our trained model predict on this data. We observed that in 228 cases (3.8%), the outputs were also flipped. Out of these, 118 consisting of 25 women and 93 men had originally been approved loans, however their counterfactuals with all same constraints but the gender had been disapproved for a loan application. Similarly, for the remaining 100 cases consisting of 86 women and 24 men that had been initially disapproved, their counterfactuals had been approved.

The results also show that while the overall proportion of males to females is about 3:2, the number of times a male application gets rejected for the exact same set of parameters otherwise is much higher. With the help of the three figures, we are exactly able to pinpoint the bias areas and corrective action can be taken on it hence. This indicates a slight bias towards men and against women in the process of loan decision making. A better graphical representation of our results is shown in Figure 1.

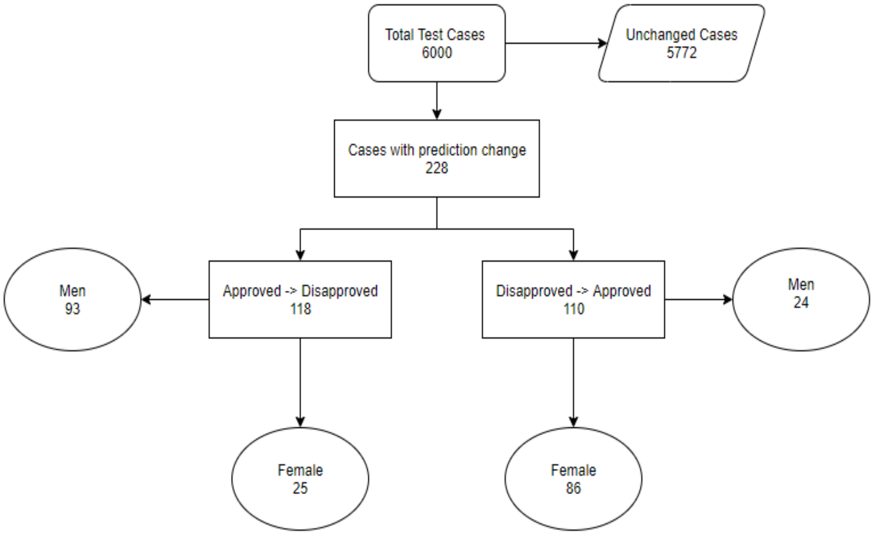

Figure 1: Counterfactual results

3 Stock Projection

To identify a lack of forward thinking toward ensuring that the system doesn't make aberrant predictions we propose quantum AI for stock projections. Just like how classical computers use 0 and 1 binary states to store information and carry out processing activities, quantum computers employ "qubits". A qubit can be in a complex-value-weighted combination of states called a superposition. A combination of "entangled" qubit states can be formed by combining multiple qubits, by putting them into linear combinations [2].

AI in finance is discussed with suggested ways to improve the metrics through quantum computing. Qiskit Aqua, Pennylane, GPU computation have been used to perform time-series analysis on stock market data, where the metric is to check for unwanted bias in datasets. Such state-of-the-art algorithms try to mitigate any bias. We observe there has always been a form of statistical discrimination, the discrimination becomes objectionable when it places certain privileged groups at systematic advantage and certain unprivileged groups at systematic disadvantage. Bias in training data, due to either prejudice in labels or under/over-sampling, hardware, yields models with unwanted bias. While quantum computing provides powerful computational tools, whether or not it can predict this type of events remains to be answered [5].

In our research, we employed qiskit.finance.data_providers to import the Google stock data for the past 5 years from Yahoo Finance. The data contains the Date, Open, High, Low, Close, Adj Close, and Volume features. Our aim is to predict whether the stock price for the following day will go higher or lower than the current day (based upon the average stock price for each day). We performed 3 experiments for the same. First, we used a traditional ML classifier - SVM with RBF Kernel for our task. Second, we created a quantum neural net using the pennylane framework. In the third, we created a variational circuit with the help of 2 frameworks - Qiskit and Aqua. This experiment was conducted with the help of IBM quantum back-ends.

Pennylane is a user friendly framework with two approaches to train variational circuit. First is simulator based which has a built simulation inside existing classical libraries and can leverage existing optimization and ML tools, great for small circuits, but not scalable. Second is hardware based but has no access to quantum information; only have measurements and expectation values. Qiskit is an open-source framework for working with noisy quantum computers at the level of pulses, circuits, and algorithms. Qiskit is made up elements that work together to enable quantum computing. This element is Aqua (Algorithms for Quantum computing Applications) which provides a library of cross-domain algorithms upon which domain-specific applications can be built [1].

Quantum algorithms employ the following sequence -

- Encoding input data into qubits
- Enabling superposition of qubits
- Applying algorithm to all states simultaneously
- Amplifying the probability of measuring the correct state
- Measuring the qubits

Our variational circuit consists of 3 ingredients -

1. Defining an initial state
2. A quantum circuit U(x; Θ)
3. Measurement of an observable as output - $\widehat{B}$

The expectation values $f(x; \Theta) = \left\langle 0|U^{\dagger}(x;\Theta)\widehat{B}U(x;\Theta)|0\right\rangle$ of one or more such circuits - possibly with some classical post-processing — define a scalar cost for a given task. The free parameters Θ of the circuit are tuned to optimize this cost function. The circuit diagram for the same is shown in Figure 2.

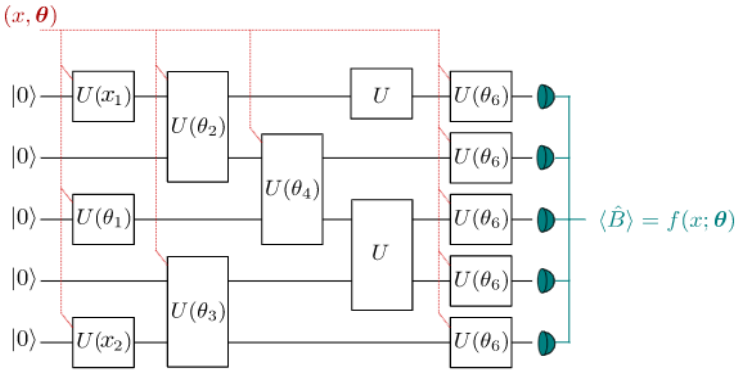

Figure 2: Circuit diagram

For all three experiments, we created and used the same test-train-split. The libraries and devices used for Quantum computing have been mentioned, along with the RMSE and Accuracy scores in Table 1. We observe that Quantum computing models do give lesser error and higher accuracy scores than traditional approaches, but explainability of such concepts is extremely difficult.

Table 1: Quantum vs Traditional Computing

	Quantum Computing		**Classical Computing**		
	Pennylane	**IBM Qiskit**	**ML Classifiers**		
Device	Strawberry Fields	IBM Vigo	CPU	GPU	TPU
Library	Pennylane	Variational Quantum Classifier	SVM with RBF Kernel		
RMSE	10%	11%	16%	14%	13%
Accuracy	0.9	0.88	0.84	0.86	0.88

4 Conclusion

In our research we tried to dive deep into the concepts of fairness and explainability with the help of two use cases. By creating counterfactuals in conjunction with neural networks for our loan decision making use case, we were able to introduce a new system wherein a check for fairness/bias can be introduced even while using a “black-box” DL algorithm. While the amount of bias that is created is dependent on the kind of data as well as the decisioning model employed, we can be sure that there is a simple test to check for this fairness/bias discrepancy that may be generated. While counterfactuals are usually employed in checking for outcome changes corresponding to narrow changes in the training parameters, it can also be used to take a corrective view of the data used by financial institutions for creating fair and unbiased models. The results from this use case could be leveraged in modifying data with any of the methods available in such a way that this test when run again, would yield minimal/negligible bias. Secondly, we have also demonstrated the performance of quantum algorithms on IBM QX hardware Vigo at backends, Strawberry Fields for a specific projection and demonstrate a comparison of quantum computing over classical computing. The data security is a challenge for banks to utilize devices outside premises taking QC to backseat for now. The multi-dimensional data modeling capacity of quantum computers may allow to find better patterns, with increasing accuracy yet the explainability remains a far-fetched reality such as cost of reading data, actual number of gates required, benchmarking classical and quantum algorithms. However, this does not mean that quantum computers can replace the classical system, they only compliment to it.

References

[1] Samuel Yen-Chi Chen, Chao-Han Huck Yang, Jun Qi, Pin-Yu Chen, Xiaoli Ma, and Hsi-Sheng Goan. Variational quantum circuits for deep reinforcement learning. *IEEE Access*, 8:141007–141024, 2020.

[2] Daniel J Egger, Claudio Gambella, Jakub Marecek, Scott McFaddin, Martin Mevissen, Rudy Raymond, Andrea Simonetto, Stefan Woerner, and Elena Yndurain. Quantum computing for finance: state of the art and future prospects. *arXiv preprint arXiv:2006.14510*, 2020.

[3] Christopher Molnar. Interpretable machine learning, Oct 2019.

[4] José Fernández da Ponte. Opening the black box of artificial intelligence in financial services – bbva |, Mar 2019.

[5] Kush Varshney. Introducing ai fairness 360, a step towards trusted ai - ibm research, Feb 2019.

[6] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.