Fairness in Machine Learning solutions: A far-fetched reality

Rishabh Agarwal Jain
Citi Research,
Bits Pilani, Dept. of Computer Science
rishabh.agarwaljain@citi.com

Sushrut Shendre
University of California, Irvine
sshendre@uci.edu

Vikas Sawant
Citi Research
vikas.sawant@citi.com

Jayant Sachdev
Citi Research
jayant.sachdev@citi.com

Abstract

With the advent of data science and machine learning, it is difficult to imagine any industry not using data-driven decisions to back their business models. While these techniques have proven to be extremely efficient, there have been enough cases to understand that these practices may not always be fair. While a model can do a great job in generalizing the trend by learning from the inherent data, it may, in fact, provide incorrect or somewhat biased trends as far as the subgroups in the data are considered. These subgroups are including and not limited to demographic attributes like race, gender, nationality, etc. As data scientists and machine learning practitioners, we must do our best to eliminate any such biases. In this paper, we summarize the problems that exist as far as fair ML practices are concerned and the tools, techniques that can be used to combat the same. While we try to present our research in order to cater to financial services scenarios, we have also cited some examples from other industries to further our understanding of the problem at hand.

1 Introduction

Analytics, Data Science, and Machine Learning have delved into all aspects of business and our lives. From recommender systems in Netflix to self-driven cars, technology and artificial intelligence have jumped into a host of domains which may have been impossible to imagine in the twentieth century. Companies, including major corporates and accelerating startups, rely heavily on data and machine learning to come up with solutions to adhere to their business requirements. The financial services sector is no stranger to these solutions and uses modeling approaches at every step these days. From customer acquisition, customer underwriting, deciding sanction limits and interest rates for loans, risk management, to fraud prevention and detection, there is a host of different algorithms and techniques that go in solving these problems. Above all, the economics of this efficiency play are compelling: a 10–25% jump in cost savings, possibly reaching 30–50% with robust cognitive automation. (Ashwin Yardi (1))

While the ML-backed solutions can be deemed to be more robust, there prevails a need to understand whether these models are fair, i.e., do our models and predictions discriminate on the basis of some acquired parameters. These acquired parameters are mostly to do with demographic characteristics of the entities, in this case, the customers. These characteristics can be associated with race, gender, education background, social status, inherited traits, etc. The Fair Housing Act and Equal Credit Opportunity Act defines Race, Color, National Origin, Religion, Sex, Familial Status, Disability,

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Marital Status, Age and Recipient of public assistance as the attributes based on which there should ideally be no bias. These are also called as **protected attributes** by people in the fairness research world. (Chen et al. (2))

Therefore, we need to understand whether there is any particular bias towards these variables and if yes, we need to cater to the same because we must realize that these decisions are directly affecting people's lives; and any discrimination or bias produced by our ML models is directly inflicting discrimination against actual people. In this paper, we attempt to gauge the fundamentals underlying fairness and bias. We look at some examples where unfair models have affected situations, both in the finance world and otherwise. We try to understand the different types of prevalent bias scenarios across business domains, data, and machine learning algorithms. Lastly, we summarize the methods that can help stop these unfair ML practices and the relevant tools we can incorporate in order to do so.

2 Fairness

Being fair means providing equal advantages and disadvantages to all concerned factions. If a system favours any particular party to an extent, then it is not a fair system and consequently, cannot be deemed ethical. The need for such a system was realised when various AI solutions, deployed across domains, started showing some form of bias or unfairness. Problems related to fairness and bias are plenty and go back in time, and several domains. One of the most common examples of models deemed unfair and in practice for a long time is that of COMPAS (Green (3)) which was used in courts to predict whether a criminal would likely commit future crimes. The results of the underlying algorithm showed that black origin defendants were often deemed to be more prone to criminal behavior in the future, although other parameters were the same. This example can clearly show how unfair practices in ML prevail because race is a parameter acquired by a person, rather than developing it; and the results clearly indicate discrimination based on this very ground.

A study by Suresh and Guttag (4) showed that image searching for CEOs resulted in a bias towards male CEOs because only 5% of the Fortune 500 CEOs were women. Again, past statistics should not have had to do much with the actual number of female CEOs and bias shown towards the male gender can be looked as a problem that fair machine learning modeling and practices should solve.

Coming to the banking and financial services sector, a study conducted in UC Berkeley by Bartlett et al. (5) shows us that credit lending was biased against Latinx/African-American borrowers. The statistics tell us that they were charged 7.9 and 3.6 basis points more respectively than other borrowers and were required to pay 765 million USD more of interest in aggregate per year. These examples provide us sufficient insights that although automated models were designed to improve the discrimination and fallacies of human decisioning, they have not been able to perform a great job at that. While one can argue that the overall decision making is far better than that involving human intervention, the problem at hand, in this case, fairness has still not been catered to as adequately and effectively as we would have liked.

These examples make us wonder how would we define fairness. There are several ways in which machine learning practitioners in diverse industries and fields have defined it. While these definitions use different metrics to put forth their idea of fairness, what is shared across all of them is the notion of eliminating discrimination when it comes to predicting results on live data.

Hardt et al. (6) define fairness as the condition when the predictor has equalized odds with respect to the classes and the outcome. For example, the true positive and false-positive rates for people across both fraudulent customers and good customers must be the same. While this serves as a superset of many definitions, there are several other ways to define fairness. Dwork et al. (7); Kusner et al. (8) define fairness as statistical parity which means that in our example, both fraudulent and good customers should have an equal chance of being assigned to a positive outcome. They also state that individuals having the same parameters should have the same outcome and not differentiated on the basis of any protected attributes. Easy enough to deduce, this will happen if all such protected attributes are not included in the decision-making process, and this was rightly stated by Grgic-Hlaca et al. (9). On similar lines, Berk et al. (10) have defined fairness to happen when the ratio of false negatives to false positives is same across protected attributes.

2.1 Techniques to induce Fairness

Having looked at the various definitions, we might conclude that fairness can fall under the following three broad categories: Individual fairness, where similar individuals are given similar predictions; Group fairness, where different groups are treated equally, and subgroup fairness where the ratios such as false-positive to false-negative and likewise are held equal across different subgroups (Kearns et al. (11, 12)).

Various methods or techniques can be used to ensure ,above discussed, levels of fairness in the system. These techniques to induce fairness in machine learning can be divided into three broad categories.

1. Pre-Processing methods: We remove bias or discrimination in data before starting the modeling phase of a data science pipeline. Ex: Using ***statistical calibration***: Leverage various statistical techniques to resample or reweigh data to reduce bias.
2. In-processing techniques: Techniques that address bias in the model training phase can be categorized into in-processing techniques. To do this, we can tweak the learning systems or the objective functions to suffice our requirements, such as subgroup fairness. Ex:

   Mathematical Regularization: Using fairness regularizer (a mathematical constraint to ensure fairness in the model) to existing ML algorithms (Kamishima et al. (13)).

   Calibrating the threshold: Calibrate the prediction probability threshold to maintain fair outcomes for all groups with protected and sensitive features
3. Post-processing techniques: n cases where an algorithm cannot be altered there, we can use some rules on top of the model we have built to ensure fairness (Bellamy et al. (14); D'Alessandro et al. (15)). Ex: ***Using surrogate models***: Wrap a fair algorithm around baseline ML algorithms already in use. (Mikulski (16))

3 Bias

Having talked about fairness and its definitions, we take a look at what bias means and its different scenarios. There are different types of bias that researchers in the field of fair ML have instated. According to Suresh and Guttag (4), historical bias exists even with the data generation process and cannot cease to exist even if there is perfect sampling and feature engineering and selection. Representation bias arises when defining and sampling from a population, and measurement bias arises when subsequently measuring features of interest. When it comes to model evaluation and tuning, solutions suffer from evaluation bias. Whenever there are flawed assumptions about the population, the model is affected by aggregation bias.

Then there is the Simpson's paradox which states that the inferences obtained from modeling the population as a whole can be partially different, completely different or even exact opposite to those obtained by modeling the subgroups of the population. (Pearl (17))

Yet another type of bias is the population bias which may arise when certain factions of the population demographics are inclined towards a certain type of behavior. In such a case, modeling the entire population can produce bias against the sects whose interests are not inclined towards that kind of behavior. Olteanu et al. (18)

Another type of bias is the sampling bias. While sampling, the best practices that need to be followed are to take random samples such that all demographics within the population are catered to. Failing to do that, not only we may get an erroneous model, but also the model would be naturally biased because the data corresponding to all segments was represented unfairly and given to the algorithm unfairly.

Another dimension to consider is time, in what is known as the temporal bias. A study was conducted on models built leveraging Twitter data, and it was observed that the discussion of a new topic begins by having hashtags about that particular topic, but with time people do not add hashtags but still continue to have discussions on the same. Therefore, extracting data based on only hashtags rules out the latter part of the actual discussions and sentiments in terms of the time period, and thus the results obtained from modeling machine learning algorithms on this type of data can lead to fallacies. Statistical bias can occur, for example, when let us say, the average/median of a particular variable is taken to generalize the whole population. The model can not only go wayward in its prediction accuracies but even if the prediction is correct, it may suffer from bias.

3.1 Implications of Bias in the finance industry

The finance industry is the backbone of the economy for every country. All aspects of the global economy rely upon an orderly process of finance. Capital markets provide the money to bolster business, and business offers the money to support individuals. Any Bias in the process can directly affect people's lives or industry at large. Hence, every decision made by a finance company has to be fair and free of biases. With the advances in computing power and robust AI industry, finance companies started replacing human intervention with ML systems to get rid of human prejudices. Moreover, it had an added advantage of cost reductions (Ashwin Yardi (1)). Later down the line, these systems started showing the traits of various AI biases.(Bartlett et al. (5) (2)

In banking and finance, customer behavior often changes with time. For example, people belonging to certain sects which were underprivileged in the past and are now improving in terms of finances. In this case, if the bank uses incomplete data with respect to the time period, we might not be able to produce efficient solutions for a particular section. (Olteanu et al. (18); Tufekci (19))

Cause-effect bias arises due to the incorrect assumptions about the causality of the event. For example, let us suppose there is a new credit card offer, and we observe that the customers using that product are spending more. Rather than calling this product as a success, it might be possible that the product was targeted to only customers who generally spend more.

Lastly, another incorrect and even unethical scenario is funding bias where the results and statistics of a study or a model in an organization are articulated in such a way to please the agencies funding the very organization. This can be extremely adverse if this agency is linked with a parameter that falls under the protected attribute list.(Mehrabi et al. (20))

These biases can lead to harmful discrimination against customers belonging to a specific class(es). Even if the effects are not adverse, it can result in a situation where they are given a neutral stance, whereas other classes gain a decisive advantage. This kind of discrimination, although indirect, can hamper their chances of progress and growth and would thus be deemed to be biased or unfair.

3.2 Supporting tools to combat bias in the industry

From where does the AI bias originate? Part of the problem is that ML models are typically outlined and trained solely for accuracy, as defined by the user, not for fairness or any other factor. Researchers must pay attention when building a model that it does not merely replicate human processes that were historically a root of biases; to build a fair model, they must consider fairness and ethical issues from the beginning (eet (21)).

There have been tremendous efforts to resolve bias and discrimination in machine learning models and comply with fairness. Researchers have introduced tools to find out the amount of fairness in a system. Aequitas is an example of such a toolkit. Aequitas enables data science and machine learning practitioners to test the models with respect to various bias and fairness metrics related to the different population subgroups. It can assist in stopping certain groups of people getting disadvantaged. A classic example where it can be used is the underwriting and loan sanction department. Having a tool such as Aequitas can ensure that there are constraints such that not only there is no hampering towards the business of the bank/financial institute, but also the loans are sanctioned in a non-discriminatory pattern to everyone (Saleiro et al. (22)).

Another similar toolkit is the AI Fairness 360 developed by IBM researchers and works towards facilitating a framework constrained with fairness metrics to evaluate algorithms. It even has an interactive module to let the users see the metrics and test the capabilities. The tool works with the help of many bias mitigation algorithms such as optimized preprocessing, disparate impact remover, and equalized odds postprocessing. (Bellamy et al. (14); Calmon et al. (23); Feldman et al. (24); Hardt et al. (6))

Then we have FairML, a toolbox written in python to quantify the significance of the inputs of the machine learning or AI models. It quantifies the model's relative predictive dependence on the inputs using a four input ranking algorithm. (Adebayo (25))

Lime is yet another open-source project that works across a host of machine learning and deep learning models. In addition to simple regression and binary classification, it can cater to multi-class classification, text classification, and image classification. It finds out which inputs have more weightage in the predictions and then tries to evaluate the nature of change in the prediction after removing these inputs, in the process, understanding whether some parameters have an extraordinary bias.

Last but not least, we have the What-If plugin built by Google, which has multiple demos and also uses an interactive UI to understand the black-box models. (Badr (26))

4 Conclusions

In this literature survey, we study about what fairness, bias, and discrimination are, and what adverse implications they can have. We look at the aforementioned topics in many dimensions: data, algorithms, domain i.e. individual/group/subgroup and at various steps in the machine learning model development lifecycle. We explained our research with a decent number of examples across all industries and tried to relate, how each aspect can affect the finance industry. Lastly, we looked at how we can tackle these prevailing problems and the tools which can/have been used in the industry. As said earlier, with the advent of machine learning and AI, though we are now able to make effective and rapid decisions like never before, the problems highlighted in the paper are of utmost importance and must be paid attention to. After incorporating them only, can we achieve the dream of having fast yet fair and reliable applications and truly serve the purpose of building automated solutions. Having looked at a comprehensive list of the problems at hand and the techniques that have been used to resolve them, we hope this study can inspire individuals and corporates to enhance their machine learning practices.

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of their respective employers/institutions .

References

[1] Nilesh Vaidya Ashwin Yardi, Jerome Buvat. Growth in the machine: How financial services can move intelligent automation from a cost play to a growth strategy. URL https://www.capgemini.com/gb-en/wp-content/uploads/sites/3/2018/07/Report-1.pdf.

[2] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geofry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 339–348, nov 2019. ISBN 9781450361255. doi: 10.1145/3287560.3287594. URL http://dx.doi.org/10.1145/3287560.3287594.

[3] Ben Green. "Fair" Risk Assessments: A Precarious Approach for Criminal Justice Reform. *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2018)*, 2018. URL https://scholar.harvard.edu/files/bgreen/files/18-fatml.pdf.

[4] Harini Suresh and John V. Guttag. A Framework for Understanding Unintended Consequences of Machine Learning. jan 2019. URL http://arxiv.org/abs/1901.10002.

[5] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-Lending Discrimination in the FinTech Era *. 2019. URL http://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf.

[6] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3323–3331, 2016. URL https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012. doi: 10.1145/2090236.2090255.

[8] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 2017-December, pages 4067–4077, 2017. URL https://papers.nips.cc/paper/6995-counterfactual-fairness.pdf.

[9] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, Adrian Weller, Nina Grgi-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016. doi: 10.2174/1381612821666150416100516. URL https://people.mpi-sws.org/~gummadi/papers/process_fairness.pdf.

[10] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods and Research*, mar 2018. ISSN 15528294. doi: 10.1177/0049124118782533. URL http://arxiv.org/abs/1703.09207.

[11] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *35th International Conference on Machine Learning, ICML 2018*, volume 6, pages 4008–4016, nov 2018. ISBN 9781510867963. URL http://arxiv.org/abs/1711.05144.

[12] Michael Kearns, Aaron Roth, Seth Neel, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 100–109, aug 2019. ISBN 9781450361255. doi: 10.1145/3287560.3287592. URL http://arxiv.org/abs/1808.08166.

[13] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7524 LNAI, pages 35–50. Springer, Berlin, Heidelberg, 2012. ISBN 9783642334856. doi: 10.1007/978-3-642-33486-3_3. URL http://link.springer.com/10.1007/978-3-642-33486-3{_}3.

[14] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. oct 2018. URL http://arxiv.org/abs/1810.01943.

[15] Brian D'Alessandro, Cathy O'Neil, and Tom Lagatta. Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. *Big Data*, 5(2):120–134, jul 2017. ISSN 2167647X. doi: 10.1089/big.2016.0048. URL http://dx.doi.org/10.1089/big.2016.0048.

[16] Bartosz Mikulski. Using a surrogate model to interpret a machine learning model. URL https://www.mikulskibartosz.name/using-a-surrogate-model-to-interpret-a-machine-learning-model/.

[17] Judea Pearl. The Sure-Thing Principle. *Journal of Causal Inference*, 2016. ISSN 2193-3677. doi: 10.1515/jci-2016-0005. URL https://ftp.cs.ucla.edu/pub/stat_ser/r466.pdf.

[18] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *SSRN Electronic Journal*, dec 2017. ISSN 1556-5068. doi: 10.2139/ssrn.2886526. URL https://www.ssrn.com/abstract=2886526.

[19] Zeynep Tufekci. Big Questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, pages 505–514, 2014. ISBN 9781577356578. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8062/8151.

[20] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. aug 2019. URL http://arxiv.org/abs/1908.09635.

[21] Reducing bias in ai models for credit and loan decisions. URL https://www.eetimes.eu/reducing-bias-in-ai-models-for-credit-and-loan-decisions/.

[22] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A Bias and Fairness Audit Toolkit. nov 2018. URL http://arxiv.org/abs/1811.05577.

[23] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 3993–4002, apr 2017. URL http://arxiv.org/abs/1704.03354.

[24] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 2015-Augus, pages 259–268, New York, New York, USA, 2015. ACM Press. ISBN 9781450336642. doi: 10.1145/2783258.2783311. URL http://dl.acm.org/citation.cfm?doid=2783258.2783311.

[25] Julius Adebayo. Fairml: Auditing black-box predictive models. URL https://github.com/adebayoj/fairml.

[26] Will Badr. Evaluating machine learning models fairness and bias. URL https://towardsdatascience.com/evaluating-machine-learning-models-fairness-and-bias-4ec82512f7c3.