GEVORG GHALACHYAN
PhD Student, Department of Mathematical Modeling in Economics,
Yerevan State University
https://orcid.org/0000-0002-0877-8138

MULTIDIMENSIONAL EVALUATION OF BINARY CLASSIFICATION MODELS WITH SOCIO-ECONOMIC METRICS

There is a rising interest towards ethical aspects of artificial intelligence. In this article we mention and discuss that a lot of decision support machine learning systems in spite of being highly accurate are not trained to reveal, measure and mitigate socio-economic bias of human nature. Analyzing the previous research on the topic we suggest our methodology - an evaluation of a binary classification model by five different criteria - accuracy, fairness, explainability, adversarial robustness, robustness to distribution shift. And furtherly we show and apply the methodology on eight different socio-demographic datasets and provide suggestions based on empirical analysis.

Keywords: *artificial intelligence, adversarial robustness, distribution shift, fairness, explainability*

JEL: C52, O31
DOI: 10.52174/1829-0280_2021_6_126

Introduction. With more and more AI-based solutions in social applications, such as jurisdiction, finance, etc. performance measures of models need to go beyond accuracy measures to assure its safety and to make it trustworthy. Some of them are distribution shift robustness, fairness w.r.t. protected groups interpretability and explainability, adversarial attacks robustness, which are recognized commonly ‘trustworthy machine learning.

From the business point of view, each decision-making system has different stakeholders. For example, banks give loans with an objective to have the lowest default rate, which will be reflected in accuracy metric. Also, governments/NGOs require the system to be unbiased towards some groups (gender, ethnicity, etc.) and the applicant requires an explanation of the application result (whether rejected or not) which will be reflected in fairness and explainability respectively.

In the context of the mentioned problem, we aim to empirically answer the following questions.

- Which couple of metrics has trade-offs and which couple can be improved to Pareto optimum?
- Which nature of models are good or bad for the different criteria?
- How to mitigation respective bias or defend from perturbations?
- How consistent can the patterns be across varied data?

Furtherly we discuss the suggested methodology, 4 different algorithms of machine learning, apply each of them on 8 different datasets with socio-economic attributes and measure trustworthiness of the models.

Literature review. The disparate impact over different occasions of human activities have been mentioned in the literature for ages but had its first appearance in Age Discrimination in Employment Act of 1967[45] in the United States signed by President Lyndon B. Johnson. It applied to pension standard and benefits by the employers, and required that the public must be aware of the age-related standards. However, the analysis was moved into quantifying form quite recently as many researchers suggest methodology on computation and mitigation of disparate impact. In this paper we discussed Reweighing [46] algorithm due to is simply form of application. Another methodology is Adversarial Debiasing[47] suggested by Stanford University and Google AI; point is to add logarithm of feature biases as a loss term.

Speaking of adversarial learning, this technique has evolved from fooling neural networks to misclassification to generating super-realistic images and face imaging. With their paper[48] on this topic, Goodfellow et al. first showed the vulnerability of AI models to perturbated samples which mostly comes from the complex non-linear nature of the models. We used a metric[49] to calculate impact of slight impact on the features to the model outcome.

The issue of interpreting inferential models has been around since the models itself were discovered, and the more complex models become the less explainable they seem to be. One-fits-all solutions are rarely available, or have trade-off with performance. We chose the model-agnostic linear explainability

[45] **Glenn, Jeremy, J., Little, Katelan, E.** (2014). A Study of the Age Discrimination in Employment Act of 1967", November 2014, GPSolo. 31 (6).

[46] **Feldman, Michael, Friedler, Sorelle A., Moeller, John, Scheidegger, Carlos,** and **Venkatasubramanian, Suresh** (2015).Certifying and Removing Disparate Impact.

[47] **Hu Zhang, Brian, Lemoine, Blake, Mitchell, Margaret** (2018). Mitigating Unwanted Biases with Adversarial Learning,

[48] **J. Goodfellow, Ian, Shlens, Jonathon, Szegedy, Christian (**2014). Explaining and Harnessing Adversarial Examples.

[49] **Chen, Jianbo, Jordan, Michael,** and **Wainwright, Martin.** HopSkipJumpAttack: A Query-Efficient Decision-Based Attack.

approach using LIME[50] model as we aim to compare different ML models and not to find the best explanation for a model.

Research methodology. For the analysis we implemented multi-dimensional analysis using different groups of metrics for classification task.

For predictive performance we considered two measures, accuracy and balanced accuracy. Accuracy of the binary classifier model processes the total number of correct predictions, and is so:

$$Accuracy \ = \frac{TP \ + TN}{TP \ + \ FP \ + TN \ + FN}$$

where TP is the number of True Positives, TN is the number of True Negatives, FP is the number of False Positives, and FN is the number of False Negatives.

Though balanced accuracy is a proportioned measure for accuracy, and we use it to score models that predict outcomes from data with imbalanced labels. Balanced accuracy can be defined as

$$Balanced\ Accuracy \ = \frac{1}{2}(\frac{TP}{TP \ + \ FP} + \frac{TN}{TN \ + \ FN})$$

Many other performance metrics are also often considered, based upon the task we solve, such as $F1\ score$, $AUC - ROC$, and $True\ Positive\ Rate$ (aka $sensitivity$); in this research we only consider accuracy and balanced accuracy.

Numerically quantified fairness is classified to following subgroups: Group versus Individual fairness. With Group fairness we analyze the relationship between different groups with respect to the sensitive/protected attributes: it measures the differences (ratios) of desired outcomes. We measure outcomes both from the data itself and, in our case, from model predictions. With Individual fairness we examine the rate of outcomes for similar groups of individuals which have been clustered along a number of the variables, indeed excluding sensitive/protected attributes. Here we consider a commonly used ratio group fairness metric, Disparate Impact[51]:

$$Disparate\ Impact \ = \frac{P(\hat{Y} = 1|D \ = \ unprivileged)}{P(\hat{Y} = 1|D \ = \ privileged)},$$

where $\hat{Y} = 1$ is the desirable outcome. The value ranges between 0 and 1 implying an absolute discrimination and absolute fairness respectively.

Furtherly, for the mitigation of the results we applied a popular pre-processing algorithm, Reweighting[52], which modifies sample weights for each protected attribute-label sub-group.

Measuring the interpretability of a model, we created local explanations for a randomly selected data sample (without replacement) using LIME, then

[50] **Tulio Ribeiro, Marco, Singh, Sameer** and **Guestrin, Carlos** (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1135–1144.

[51] **Feldman, Michael, A Friedler, Sorelle, Moeller, John.** Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, 259–268.

[52] **Kamiran, Faisal,** and **Calders, Toon** (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33, 1, 01 Oct., 1–33.

calculated the average faithfulness metric of the generated explanations. Faithfulness[53] of a data point was measured as the correlation between the importance value by LIME[54] to different variables in making a model prediction for that sample and the effect for each attribute on the prediction confidence.

From the point of view of ethical AI, we want to train a model that is as robust to adversarial attacks as possible, that is insignificant changes made to original features which lead to change of model outcome. The adversarial robustness of a model, as matter of fact, cannot be computed in a generic manner; it is, in fact, estimated with respect to specific and intentionally created adversarial examples. Here we generated a few adversarial examples using a model-agnostic algorithm, HopSkipJump[55], samples were later used to evaluate the Empirical Robustness, that is the average minimum perturbation which an attacker must apply for an attack to be a success.

Also, we evaluate the robustness of different models to attribute distribution shift, creating shifted datasets for each dataset by partitioning each into two parts, based on some attribute (ex. region).

We used four different classification algorithms: logistic regression (LR), random forests (RF), gradient boosting (GBC), and multilayer perceptron (MLP). Cross validation of 5 folds was used for the experiments. For each cross-validation split, we did the following:

1. The training dataset was used for building the four independent models. Categorical features are transformed to one-hot encoded, and feature standardization was done for numeric features by centering and scaling. The trained models were later evaluated based on the test data.
2. All models were further tested using the shifted dataset.
3. Bias mitigation algorithm was applied on the training data and the debiased dataset was then used to train the four models which were again tested on the test data.
4. The models learned in step 3 were also tested on the shifted dataset.

The estimates for all the metrics were calculated using the five splits. For faithfulness of explanations, LIME algorithm was used for generating local explanations and 50 random samples were feed to the model to output the mean faithfulness score. For empirical robustness, 20 random samples were used to generate adversarial samples. Three open-source toolkits, AIF360, AIX360, and ART, were used to evaluate model fairness/bias mitigation (disparate impact and reweighing), explainability (faithfulness), and adversarial robustness (empirical robustness), respectively.

[53] **Alvarez-Melis, David** and **Jaakkola, Tommi** (2018).Towards Robust Interpretability with Self-Explaining Neural Networks. In Advances in Neural Information Processing Systems. 7775–7784.

[54] **Ribeiro, Marco Tulio, Singh, Sameer** and **Guestrin, Carlos** (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1135–1144.

[55] **Chen, Jianbo, Jordan, Michael** and **Wainwright, Martin** (2020). HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. 1277–1294.

Data and Findings. For the analysis we used 8 diverse (by domain, size, sensitive attribute, etc.) dataset. In this section, we will discuss the datasets and provide with descriptive statistics.

HMDA[56] - First dataset contains the 2018 mortgage application data collected in the U.S. under the Home Mortgage Disclosure Act. It consists of 1,119,629 applications for single-family and principal-residence purchases. The outcome is to predict whether a mortgage is approved or not, and ethnicity was used as the protected attribute (non-Hispanic Whites vs non-Hispanic Blacks), other races are skipped for the consistency of experiment. State feature is used for dataset shift split. White approval rate was 94.4% historically and as for the Black it was 86.4%.

MEXICO[57] - Coming from the household survey (2016) in Mexico, next dataset includes demographic and poverty-level features for 70304 Mexican households. The purpose is to predict whether a family is poor or rich. For protected feature, we have age. We use the 'urban' variable to partition the dataset into the shift and base datasets: urban residents' data are used for training or testing the model while rural residents' data are used only for distribution shift tests. There are 53.2% young families in the base dataset, and 52.5% in the shift dataset. 39.5% of the young and 29.7% of the old are historically classified as poor.

ADULT[58] - The Adult dataset, which comes from the UCI ML repository, has demographic and financial data on 48841 individuals from the US Census Bureau database. We aim to predict if an individual's income exceeds $50K/year. We treat race as the protected attribute. We partitioned the data on the feature called 'native-country': 'United-States' residents are retained in the base dataset while the rest are - for distribution shift. The target variable is "true" for 25.2% (26.8% for 'whites', 13.8% for 'non-whites') of the training dataset and 19.2% (17.5% for 'whites', 23.1% for 'non-whites') of the shift data.

BANK[59] - The Bank Marketing dataset, again from the UCI repository, uses data of marketing campaigns by a Portuguese bank. The outcome is to predict which customers will subscribe to a term deposit. Shift distribution split is done by the month in which the latest contact was made with the customer. The sensitive feature is age (≥25 or not). 11.32% of the people in the training dataset subscribed to a term deposit (18.74% of the young people and 11.2% of the old) while 19.78% of the people in the shift dataset did not subscribe (57.2% of the young and 19.11% of the old).

[56] Home Mortgage Disclosure Act (HMDA) Snapshot National Loan Level Dataset, Federal Financial Institutions Examination Council (FFIEC), U.S. Government. 2018.

[57] **Noriega-Campero, A., Bakker, M.A., Garcia-Bulle, B.,** and **Pentland, A.** (2019). Active Fairness in Algorithmic Decision Making. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 77-83.

[58] **Kohavi, R.** (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996

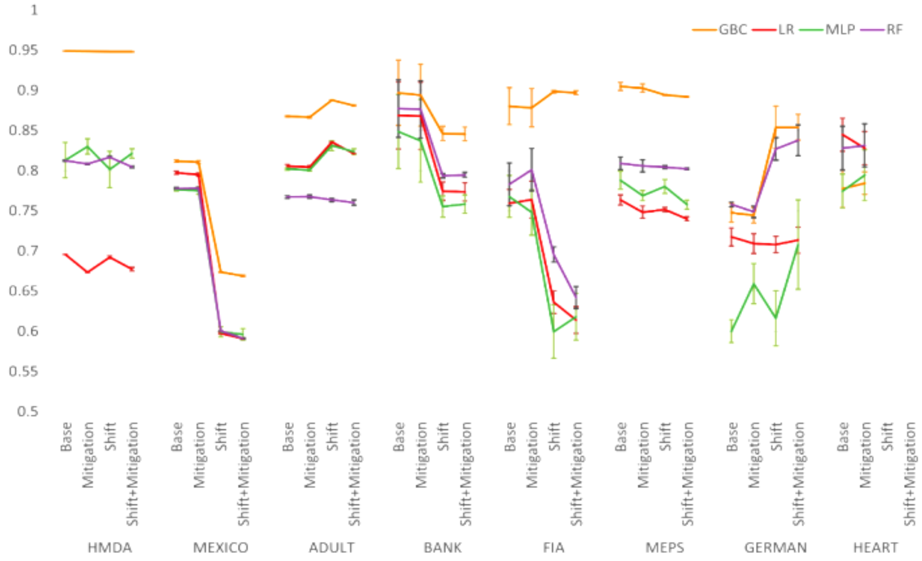
[59] **Dua D.,** and **Graff C.** (2017). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

FIA[60] - The Financial Inclusion in Africa (FIA) contains financial services and demographic data for 33611 individuals from 4 African countries: Tanzania, Kenya, Rwanda and Uganda. We used the publicly available 'train' component consisting of data for 23523 individuals. We aim to predict who is probable to have a bank account. The protected attribute considered is 'gender_of_respondent', and 'country' feature is used to split the dataset to test for distribution-shift. There are 41.6% males in the training dataset and 34.2% males in the shift dataset. Finally, 14.6% have bank accounts in the train dataset (19.6% of men, 11.2% of women), while the respective number for the shift data is 8.7% (11.6% of men, 6.9% of women)

MEPS[61] - The Medical Expenditure Panel Survey (MEPS) data is a set of annual surveys by the US Department of Health and Human Services. Every year, a new panel is started and interviewed for five rounds during the next two calendar years. The data consists of 2-year longitudinal - panel 19) as the base training/testing dataset (8137 records) and panel 20) as the shift dataset (8736 records). The goal is to predict patients that would have high second-year spendings, based on first-year demographic and health attributes. Race is used as the protected feature (64.2% White in training data and 67.8% White in shift data). High spending patients are 9% of people in the training set (10.26% of Whites and 6.77% of Blacks), and 10.1% in shift set (11.33% of Whites and 7.1% of Blacks).

GERMAN[62] - The German Credit dataset is the creditworthiness of over 1000 people. The aim is to predict which people will have good credit. We use 'gender' as the protected. The data is split using the 'foreign worker' attribute.

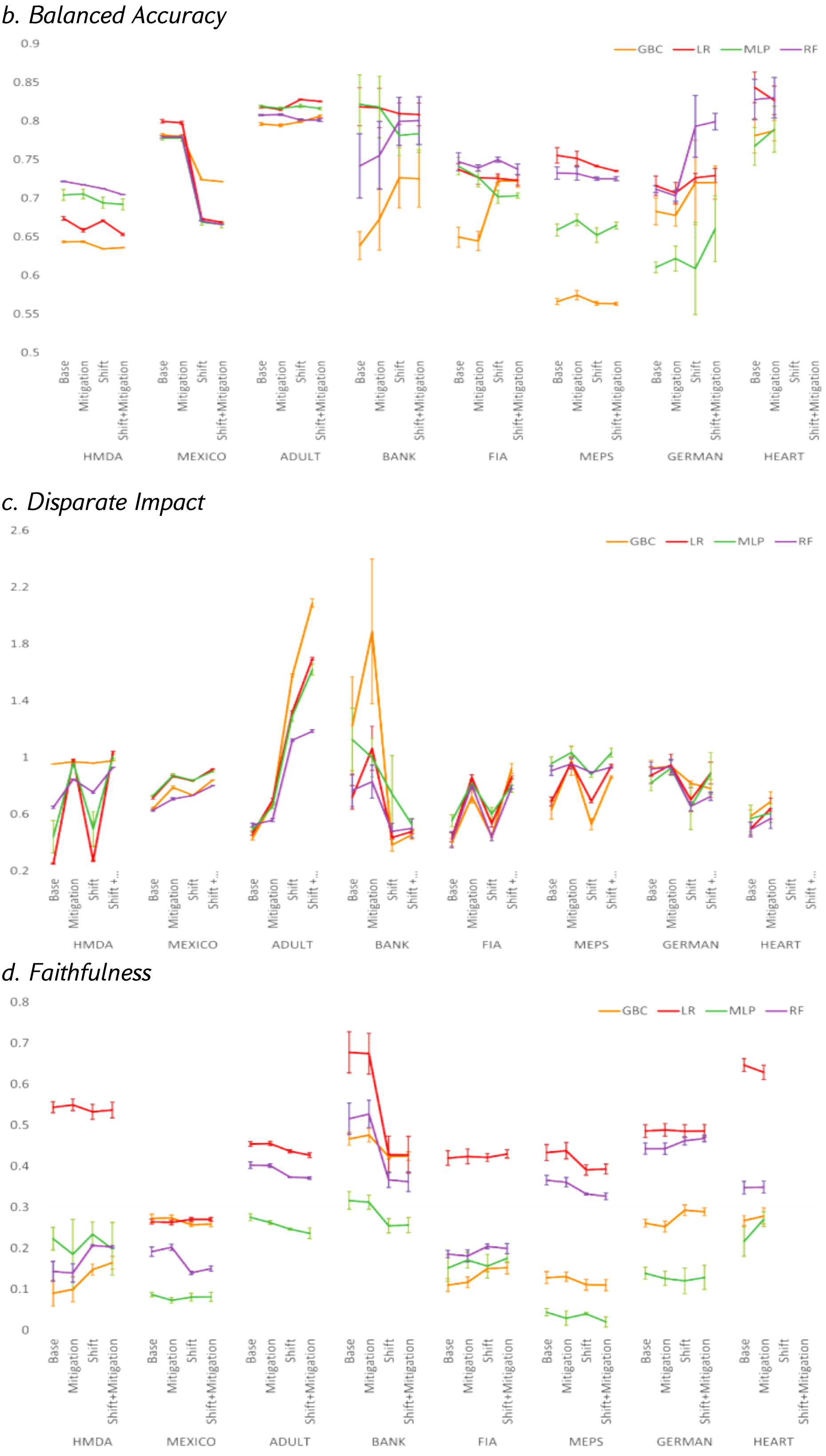
a. Accuracy



[60] **Zindi** (2019). Financial inclusion in Africa survey data. https://zindi.africa/competitions/financial-inclusion-in-africa

[61] Agency for Healthcare Research and Quality. Medical Expenditure Panel Survey (MEPS). 2018. https://www.ahrq.gov/data/meps.html

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



b. Balanced Accuracy

c. Disparate Impact

d. Faithfulness

e. Empirical Robustness

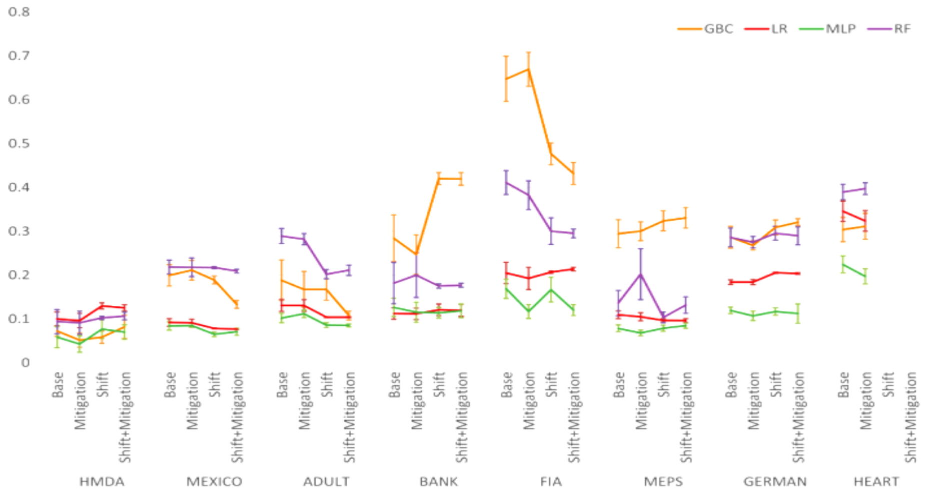


Figure 1. ***Performance metrics for trained models***

HEART[63] - The Cleveland Heart dataset is a small 304 record dataset. Considering the size of the dataset, we will not create a distribution-shift dataset. The protected feature is ‘age’ above or below the mean value (54.6 years): 53.6% have age above this value.

Figure 1a and 1b show us how well the models of different algorithms performed on the different datasets w.r.t. accuracy and balanced accuracy. While gradient boosting classifier (GBC) models mostly had the best accuracy for the datasets, they typically are the worst performers for the balanced accuracy. At the same time, multi-layered perceptron (MLP) models performed poorly in terms of accuracy, and had better balanced accuracy, mostly for the larger data. Random forest (RF) models are often not the best, but performed quite well on most of the datasets on all the metrics for the smaller datasets respective to its boosting origin. Logistic regression (LR) models are robust showing good balanced accuracy, but have relatively poor accuracy.

W.r.t fairness, all datasets have historical discrimination between the privileged/unprivileged groups as we defined the sensitive attributes. Yet this “unfairness” did not exhibit itself similarly in all the models (Figure 1c), e.g. the HMDA data, although the GBC was almost fair at 0.94, LR was unfair with DI of 0.25. In the case of BANK, two models were biased for young people and the other two were biased against older ones. Yet, in some other cases, e.g. ADULT, MEXICO, fairness measure was similar for all the models.

As mentioned before local explanations were generated by LIME and evaluated using faithfulness metric (Figure 1d). LR performed the best, describing its linear nature, while MLP as a black-box model were the worst, the same applies to the other non-linear and tree-based aggregation models.

[63] **Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, K., Guppy, S.,** and **Lee, V., Froelicher, S.** (1989), International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology 64, 304–310.

As to adversarial attacks (Figure 1e), GBC and RF were the most robust, and LR and MLP performed poorly. Such behavior comes from the training methodology. GBC and RF are voting classifiers, i.e. the result is concluded using different tree-based models and perturbated examples can not have direct effect on the results, while LR and MLP are single-model classifiers and often optimized with SDG optimizers which are easily compromised.

Bias was mostly mitigated using reweighing, as we see on the peaks in the second points of the line graphs. Yet the fairness achieved was dependent on the initial training data, RF was not able to benefit from mitigation in contrary to GBC, LR, and MLP (in the figures RF plot, violet lines, has smaller relative change). Generally bias mitigation comes with the lower accuracy of the models, the decrease is quite small for GBC and RF and more exhibited for MLP. Bias mitigation yielded a fair change in explainability (faithfulness metric). Likewise, adversarial robustness decreased through bias mitigation, sometimes quite enormously, yet there are some instances where it shows an increase.

The third point for each line graph shows that accuracy suffers as distribution of the data shifts. It varied between datasets, yet it was quite obvious (10-15%) for some models and datasets (FIA, Mexico, and Bank). GBC accuracy is quite robust to distribution shift.

Conclusion. This work presents the necessity of multivariate evaluation of machine learning models, and implements it for 5 different phenomena. We discovered that in spite of better predictive performance, cutting edge algorithms lacked ethical efficiency, such as explainability. Models of linear nature, in our example - logistic regression, provides better explainability and overall robustness, but have no predictive advancements. Moreover, we showed a minimal example of bias mitigation, reweighing, which has a tradeoff with accuracy that is more sever bias mitigation needs more “sacrifice” of model accuracy. Assumptions and hypothesis are supported with empirical analysis for eight datasets from different socio-economic situations.

On average reweighting results 31.4%±4.8% better disparate impact metric. It has non-significant impact on accuracy on ~40% of the cases from which it has significant impact on disparate impact on ~62% of the cases on non-shifted dataset. That means that ~25% of the models can be fairness-optimized with no trade-off with accuracy. The respective result is ~17% for the distribution shifted dataset.

For the future analysis we plan to discuss and implement some of the following. First of all, our goal will be to choose the most favorable model from the socio-economic point of view. For this we adopted more problem-oriented approach which is to create aggregations over a specific matter of issue. Then we plan on researching and implementing mitigation algorithms for different phenomena – explainability, robustness to distribution shift and adversarial attacks, in addition to fairness. While discovering 5 different phenomena at the same time we plan to reach the Pareto optimum and build the optimal frontier to make sure that each model is absolutely robust by one metric with respect to the other metrics.

References

1. Alvarez-Melis, David and Jaakkola, Tommi (2018). Towards Robust Interpretability with Self-Explaining Neural Networks. In Advances in Neural Information Processing Systems. 7775–7784.
2. Su, Dong, Zhang, Huan, Chen, Hongge, Yi, Jinfeng, Chen, Pin-Yu, and Gao, Yupeng (2018). Is robustness the cost of accuracy? - a comprehensive study on the robustness of 18 deep image classification models.
3. Goodfellow Ian J., Shlens, Jonathon, and Szegedy, Christian (2015). Explaining and Harnessing Adversarial Examples.
4. Chen, Jianbo, Jordan, Michael, and Wainwright, Martin (2020). HopSkipJumpAttack: A Query-Efficient Decision-Based Attack.
5. Feldman, Michael, Friedler, Sorelle A., Moeller, John, Scheidegger, Carlos, and Venkatasubramanian, Suresh (2015). Certifying and Removing Disparate Impact.
6. Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Natesan, Ramamurthy, K., Reimer, D., Olteanu, A., Piorkowski, D., Richards, J., Tsay, J., and Varshney, K. R. (2019). FactSheets: Increasing Trust in AI Services Through Supplier's Declarations of Conformity. IBM J. Res. Dev. 63, 4/5, 6.
7. https://aif360.readthedocs.io/en/latest/
8. https://aix360.readthedocs.io/en/latest/
9. https://adversarial-robustness-toolbox.readthedocs.io/en/latest/

ԳԵՎՈՐԳ ՂԱԼԱՉՅԱՆ

Երևանի պետական համալսարանի տնտեսագիտության մեջ
մաթեմատիկական մոդելավորման ամբիոնի ասպիրանտ

Սոցիալ-տնտեսական ցուցանիշների բազմաչափ գնահատումը բինար դասակարգման մոդելների միջոցով.– Արհեստական բանականության էթիկական բնույթի հարցերի նկատմամբ հետաքրքրությունն աճում է: Հոդվածում քննարկվում և շեշտադրվում է այն հանգամանքը, որ մեքենայական ուսուցման համակարգերը, լինելով շատ ճշգրիտ, չեն կառուցվում այն տրամաբանությամբ, որ կարողանան բացահայտել, գնահատել և մեղմել սոցիալ-տնտեսական բնութագրերով պայմանավորված շեղումները: Քննարկելով թեմայի վերաբերյալ նախորդ ուսումնասիրությունները՝ առաջարկել ենք որպես գնահատումների հիմք ունենալ հետևյալ հինգ չափանիշները՝ ճշգրտություն, արդարություն, բացատրելիություն, անշեղություն «հակառակորդ» նմուշների և անշեղություն տվյալների բաշխման փոփոխությունների նկատմամբ: Ցույց է տրված մեթոդաբանության գործնական կիրառությունը ութ տարբեր տվյալների վրա, կատարվել են փորձառական եզրահանգումներ:

Հիմնաբառեր. *արհեստական բանականություն, հակառակորդ օրինակներից կայունություն, բաշխման փոփոխություն, արդարություն, բացատրելիություն*
JEL: C52, O31
DOI: 10.52174/1829-0280_2021_6_126

ГЕВОРГ ГАЛАЧЯН
Аспирант кафедры математического моделирования в экономике Ереванского государственного университета

Многомерная оценка социально-экономических показателей с помощью моделей бинарной классификации.– В настоящее время наблюдается рост интереса к этическим аспектам искусственного интеллекта. В данной статье подчеркивается, что многие системы машинного обучения, несмотря на высокую точность, не в состоянии выявлять, измерять и смягчать отклонения, обусловленные социально-экономическими характеристиками. С опорой на анализ существующих исследований по данному вопросу автором предлагается собственная методология - оценка модели бинарной классификации по пяти различным критериям: точность, справедливость, объяснимость, состязательная надежность, устойчивость к сдвигу распределения. В статье также представлено применение указанной методологии к восьми различным наборам данных, а также приведены выводы, основанные на эмпирическом анализе.

Ключевые слова: *искусственный интеллект, противоборствующая устойчивость, сдвиг распределения, справедливость, объяснимость*
JEL: C52, O31
DOI: 10.52174/1829-0280_2021_6_126