Ethics, Fairness, Responsibility, and Privacy in Data Science (DATA 25900)
Andrew Goldblatt
5/16/2023

Bias in Artificial Intelligence and Machine Learning Models: Synthetic Data

Introduction

As artificial intelligence and machine learning models permeate into everyday life, all entity's actions and behavior are modeled and predicted. Since we are becoming increasingly reliant on machines to make decisions for us, we need to ensure that they are accurate and fair but these goals can be constrained at many different steps throughout the machine learning AI pipeline. Each process can be refined and tuned to create applicable and appropriate models and one critical piece is the data. Having accurate and usable training data will allow models to have meaningful results.

If the data input is not sufficient, model performance in the context of its use application can be jeopardized. Picking the appropriate machine learning model and tuning the parameters can quantitatively improve a model but does not inherently reduce bias that may be present within the training data. Synthetic data can be applied at this step to mitigate many of these issues. Fundamentally, synthetic data is statistically similar and representative of real data but contains edge cases that may not be present in real data, but still are technically plausible. Additionally, the sheer quantity of data allows models to be trained faster, critically important for large datasets and real-time decision-making applications (autonomous cars, for example). While the purpose of synthetic data is to maintain the structure of the original data, it can be modified to reduce biases that may be present in the original data or shift to better reflect a population. In training machine learning models, the training dataset should be similar to the test data, and with synthetic data, training data can conform to these goals if the desired structure and statistical properties of the population are known.

Generation of Synthetic Data

Synthetic data is being used for cutting-edge technologies and research including Waymo (Google's self-driving car project), National Institute of Health clinical research, and Amazon Alexa.[1] [2] Typically, synthetic data is generated with a Generative Adversarial Network (GAN) with a discriminator algorithm that confirms that the

[1] National Center for Advancing Translational Sciences. "N3C Data Overview," August 31, 2020. https://ncats.nih.gov/n3c/about/data-overview.
[2] Statice.ai. "Types of Synthetic Data and 4 Real-Life Examples (2022) - Statice," 2022.

generated data is different from the real data.[3] This is a common implementation of algorithms to generate synthetic data however entities and organizations have developed a variety of tools, algorithms, and packages for specific use applications. Synthetic data is a fairly new field so access to generation tools is publicly limited and/or proprietary. There are about 50 pre-seed/Series A startups that produce and/or sell synthetic data.[4] The tools that these companies have produced, however, do not all allow bias to be controlled in structured (tabular) or unstructured datasets. Accessible synthetic generation tools can be accessed through packages such as Synthetic Data Vault, CTGAN, or YData. In my experience, these tools were finicky and difficult to use - if working at all. It would require a highly skilled user to successfully generate data using one of these tools and implement it into a model with positive results. There are more accessible tools, like Gretel.AI which take care of some of the technical specifications and data engineering pipeline required to generate data. In this exploration of bias reduction for machine learning applications, I've used Gretel.AI's minority class booster on the Palmer Penguins dataset to see how synthetic data can be applied to improve fairness and accuracy in machine learning models.

Implications of Bias and Ethics of Bias Reduction

Any bias present in training datasets may have severe implications on human life (racial/ethnic/gender discrimination), social/economic decisions, and legal outcomes.[5] Currently, bias in AI/ML is occasionally regulated and evaluated by government agencies.[6] However, it is largely up to organizations to implement bias reduction standards. Any bias reduction should be evaluated critically to ensure there is not an over-reduction of bias and the bias reduction ultimately ends up doing more good than bad. Bias vs the ground truth should be carefully identified by skilled developers/engineers/social scientists and adjusted models should be compared to unadjusted models through a variety of quantitative and qualitative fairness metrics.

Bias Entering an AI/ML Pipeline

Bias enters an AI/ML pipeline through two main potential avenues: human bias and intrinsic data bias. When collecting data, humans may undersample, mislabel data,

---

[3] Saxena, Pawan. "Synthetic Data Generation Using Conditional-GAN - towards Data Science." Medium. Towards Data Science, August 12, 2021. https://towardsdatascience.com/synthetic-data-generation-using-conditional-gan-45f91542ec6b.
[4] Devaux, Elise. "[New] List of Synthetic Data Vendors— 2022 - Elise Devaux - Medium." Medium. Medium, October 6, 2022. https://elise-deux.medium.com/new-list-of-synthetic-data-vendors-2022-f06dbe91784.
[5] Harvard Business Review. "What Do We Do about the Biases in AI?," October 25, 2019. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai.
[6] JD Supra. "Legal Requirements for Mitigating Bias in AI Systems | JD Supra," 2022. https://www.jdsupra.com/legalnews/legal-requirements-for-mitigating-bias-3221861/.

and add confirmation bias.[7] Humans may execute an inappropriate/incorrect hypothesis test that leads to improper data for the model or they may conduct feature selection to exclude features that are related to the outcome variable or include sensitive variables that bias the outcome.[8] On the data side, training data may not include edge cases seen in the real data, the data may be manipulated so models are trained on incorrect data. Sometimes training data cannot be generalized to new data. Training data may also include inaccurate data.

Concept/Data Drift Bias

Model performance and embodied bias may be affected by concept drift/bias drift where upstream factors change relationships between dependent and independent and independent variables or property values have changed. If this change is identifiable, such as the ability to sample testing data as a tuning parameter for training data, the training data can be synthetically manipulated to be statistically representative of the testing data and the model can be re-trained on this more accurate data.

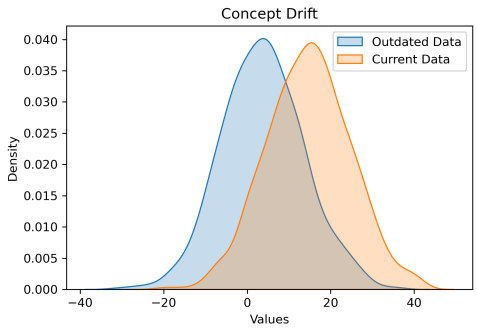

Figure 1. Distribution of training data for a variable for old training data vs new test data

Figure 1 shows a normal distribution of a singular variable across two different data sources. While they appear to have the same shape of distribution, it is pretty clear that they are not centered in the same space. This data is randomly generated by np.random.normal with two different means but can be representative of concept drift where population parameters no longer match training parameters, thus they are not statistically representative of each other. By taking samples of the training data and running t-tests, after some threshold, distributions are no longer statistically similar but the new testing data can become training data and the existing training data can be synthesized to resemble the testing data more closely. Here, the synthetic data generation model can be trained with the counterfactual information coming from the training set. Another application of using synthetic data to adjust for differences between

[7] Ydata. “Using Synthetic Data to Overcome Bias in Machine Learning.” Ydata.ai. YData, March 9, 2023. https://ydata.ai/resources/using-synthetic-data-to-overcome-bias-in-machine-learning#:~:text=Synthetic%20data%20offers%20a%20promising,a%20process%20called%20data%20balancing..

[8] Drew Roselli, Jeanna Matthews, and Nisha Talagala, “Managing Bias in AI,” Companion Proceedings of the 2019 World Wide Web Conference, May 13, 2019, https://doi.org/10.1145/3308560.3317590.

training data and testing data can be in out-of-sample predictions. In Professor Steven Levitt's class, Data Construction and Interpretation in Economic Applications (ECON 21300), at the University of Chicago, students were asked to apply OLS and/or Random Forest machine learning models to a housing dataset. The test set was identical to the training set except for a singular variable, home quality, which was set at a value for all homes in the test set that was higher than any value in the training set. Close inspection by an analyst or a bias identification algorithm could identify that these values between the datasets were statistically significant and economically relevant. Top-performing housing price prediction models used OLS as they were the best at generalizability. Synthetically accounting for this data drift by parameter tuning could mean that a more advanced ML model could perform as well/better than OLS models on this prediction model. While this model did not contain bias that pertained to racial/ethnic/gender discrimination, the model was still biased towards the training data, greatly affecting the performance, especially if this bias was not identified.

Training set size requirements

Sufficient training set sizes are necessary to minimize extrapolation. The sheer size is important but also the training data has to be relevant to the test set/model application. This size requirement is determined by the differentiability between classes and the complexity of features. The common machine learning dataset, MNIST digits/clothing, has 60,000 training images. Each MNIST image is 28x28 pixels and each pixel takes on a value between 0-255. Because of the complexity, an enormous amount of training data is needed to differentiate between letter/image classes. The Palmer Penguins dataset, on the other hand, only needs a few dozen to reach the accuracy plateau because of distinct differences in the species and limited training features. Figure 2 below shows the generalized learning curve theory that can be applied to models. By boosting the training size, and boosting minority classes, models can be trained on accurate and sufficient-sized data.

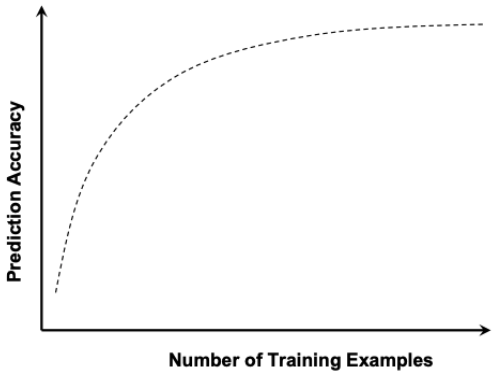

Figure 2: Learning Curves in Machine Learning (Graphic by Claudia Perlich, IBM)[9]

Bias and Bias Reduction: A Palmer Penguin Example

To understand the impacts of bias in data for ML/AI applications and to explore potential bias reduction techniques, I decided to work with the Palmer Penguin dataset curated by Alison Horst.[10]

Species	Beak Length (mm)	Beak Depth (mm)	Flipper Length (mm)	Body Mass (g)
1	47.5	14.0	212.0	4875.0
0	41.3	21.1	195.0	4400.0
0	39.6	17.7	186.0	3500.0
1	52.2	17.1	228.0	5400.0
1	44.0	13.6	208.0	4350.0

Figure 3: Example of Palmer Penguin train/test Data. Species 1- Gentoo, Species 0 - Adelie

This penguin dataset includes 3 different species, Gentoo, Adelie, and Chinstrap. For simplicity, Chinstrap penguins were omitted from the training and test set so there was only binary classification. The feature selection included Beak Length, Beak Depth, Flipper Length, and Body Mass. The classification algorithm was the Scikit-learn Multi-layer Perceptron classifier (MLP). Other classifiers, such as logistic regression classification or K-NN classification would likely perform similarly due to differentiability between the two penguin classes. The neural net layer specification was set to one hidden layer with 2 nodes. This made training the model quick and accurate. The features in this data were relatively simple and the training sets were small. Minor data cleaning was applied to the dataset including dropping columns such as island and bird sex, leaving numerical columns only. Despite omitting features that may correlate with the outcome variable, species, I wanted to push the limits of minority class imbalance and boost via synthetic data generation to supplement the real data. The flowchart below (figure 4) shows how the 5 different training datasets were constructed:

[9] Perlich, Claudia. “Learning Curves in Machine Learning,” 2009. https://dominoweb.draco.res.ibm.com/reports/rc24756.pdf.
[10] Horst AM, Hill AP, Gorman KB (2020). palmer penguins: Palmer. Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. Doi: 10.5281/zenodo.3960218.

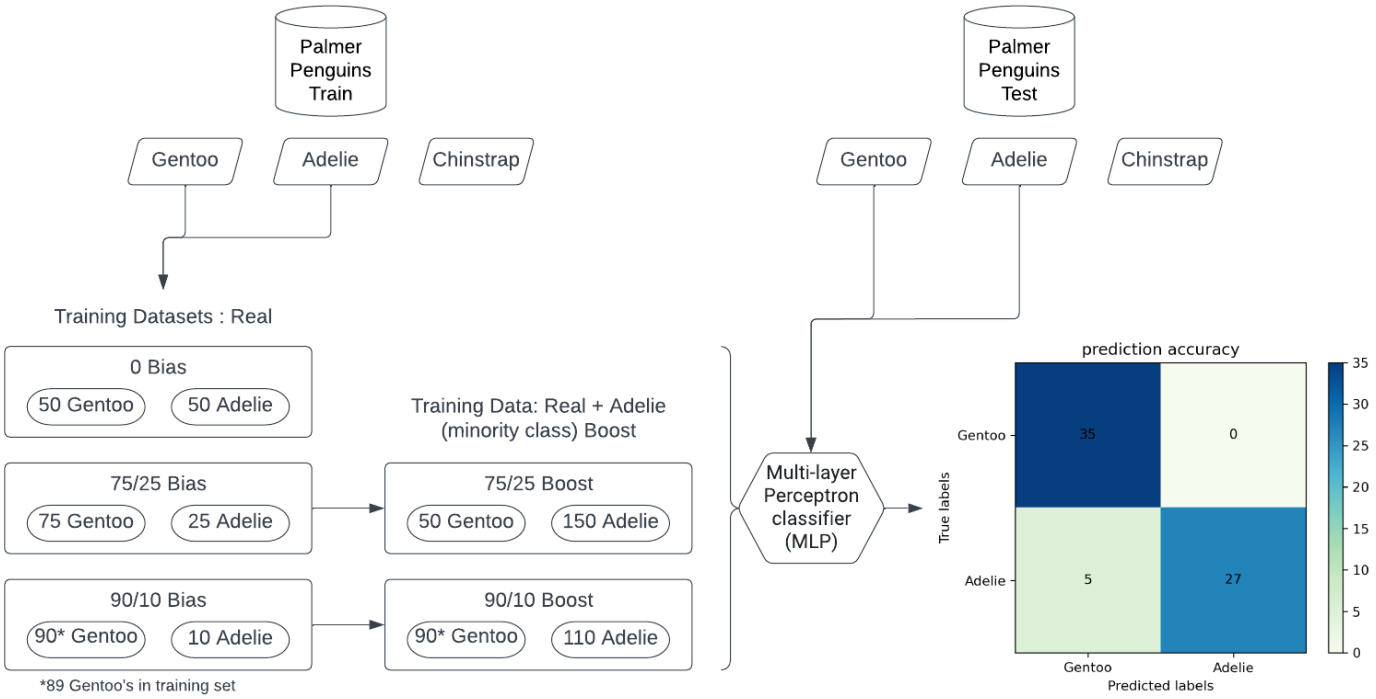

Figure 4. Flowchart of MLP Palmer Penguin classification for various biased and boosted datasets

From the training set, Gentoo and Adelie penguins were randomly selected using the .sample() method. For the 0 bias dataset, 50 Gentoo penguins were sampled, and 50 Adelie penguins and stored in a new data frame. For the 75/25 biased dataset, 75 gentoos were sampled and 25 Adelaide penguins were selected. For the 90/10 dataset, 89 Gentoos were randomly sampled along with 10 Adelie penguins. Since the training set was relatively small, there were not 90 Gentoo penguins in the dataset but 89 closely resembled the same proportion of bias in the dataset. These different biases were chosen to represent sampling bias or minority bias where a certain class/demographic is underrepresented within training data. The 75/25 and the 90/10 datasets were then exported into Gretel.AI's minority class boost algorithm which generated 100 more samples of the minority class and appended to the real data. These datasets were then all fit in the MLP classifier and used to predict species for the cleaned version of the test set.

Results:

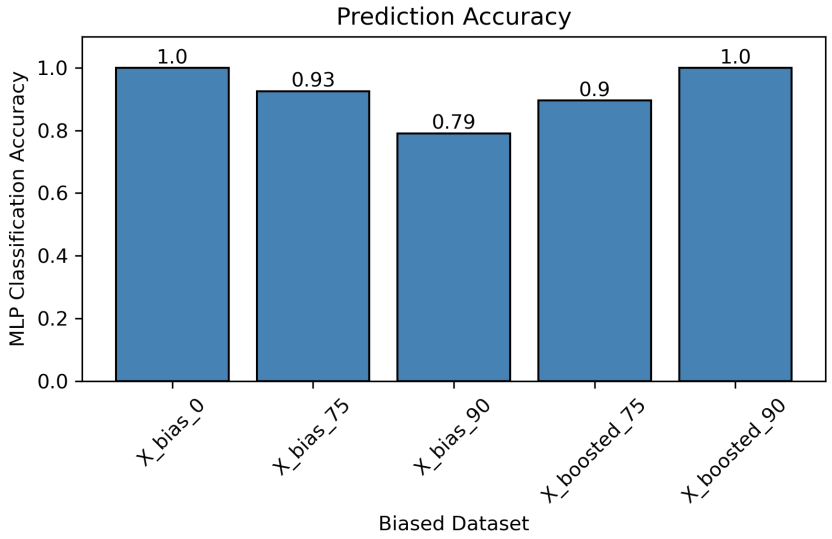

Figure 5: MLP classification for different biased datasets

The MLP selected for this classification task was a pretty powerful tool and with refinement with the parameter tuning, the model was able to predict species with 100% accuracy with the balanced 50/50 dataset (X_bias_0). The accuracy drops off as the dataset gets more biased. In the 75/25 dataset (X_bias_75), accuracy drops to 93% and drops to 79% for the 90/10 dataset (X_bias_90). With the minority class boosted datasets, the 75/25 (X_boosted_75) minority dataset was boosted to 75/125 and had an accuracy of 90%. This is lower than the accuracy of the non-boosted accuracy for the base dataset that the boost data was generated from. Finally, the boosted 90/10 dataset (X_boosted_90) had a mix of species with a 90/110 split after having the minority class boosted. The classification accuracy with this dataset was 100%, the same as the unbiased boosted dataset.

Discussion and Limitations:

From the biased but not boosted data results, we begin to see the quantitative impacts of biased data used for training. For these minority classes, there simply isn't enough data to be a good reference for the testing data. This could become even more problematic if minority classes are even more rare within the data. The boosted 75% dataset performed worse than the unboosted version of that data but this could be attributed to the sample of the data and the quality of the synthetic data generation. The boosted version of the 90/10 dataset had very high accuracy, performing much better than the biased, unboosted version of this dataset. From this test case example, synthetic data may be a viable method of improving model accuracy and reducing bias within data. To be more confident in its use for this application, results should be

simulated by resampling the biased data and re-generating the synthetic classes for the boosted datasets. Also, different algorithms and tuning parameters should be tested against accuracy and fairness metrics. The Palmer Penguin test example that I've conducted here serves as an indication that synthetic data as a tool for minority class boosting and bias reduction has potential application to real-life AI/ML implementations.