Trusted AI©

Title Page

Solution: Detect and Remove Bias with GAN

Link to Video

List of team members: Pamela Gupta, Samantha Wigglesworth, Shivam Patil
Contact information for team’s point of contact:
Email: Pamela.Gupta@outsecure.com
Phone: (203) 816 -1826

Abstract

AI/ML algorithms offer promise for clinical decision making, however that potential has yet to be fully realized in healthcare. However it is one of the highest impact area when it comes to society at large.

There are several roadblocks for gaining the full benefit of AI and AI in healthcare. Bias can be introduced even before building ML algorithms and models, for example the actual data used for model training may be imbalanced and may introduce discriminatory biases towards specific groups of people but that requires a different mitigation and is not the focus for this challenge..

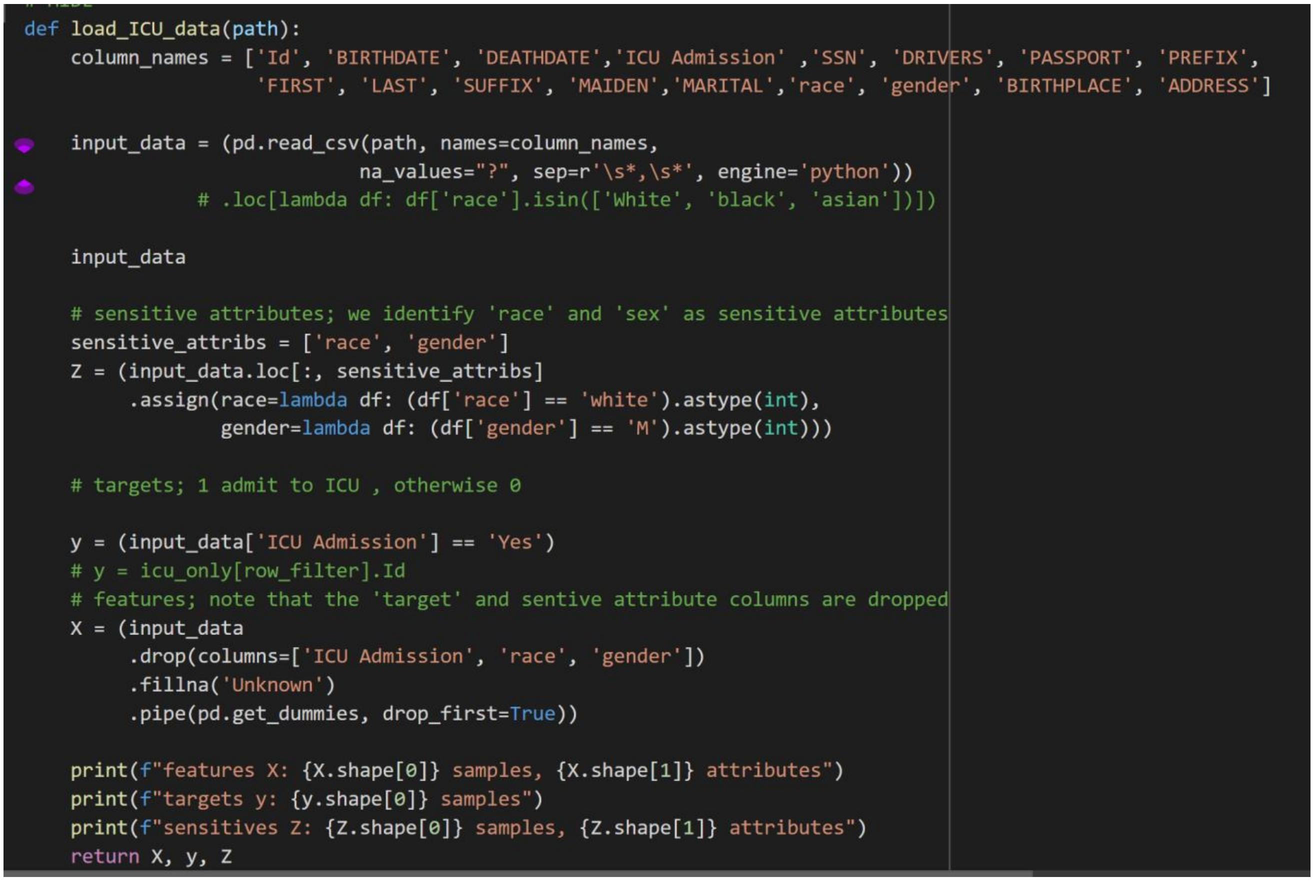
For this challenge we are focusing on Improving healthcare by detecting and mitigating bias in AI ML Algorithms. Our solution uses Generative Adversarial Networks or GANs, specifically for those models that use Gradient Descent.

Bias in machine learning models can rear up when attributes, especially sensitive attributes get correlated (even unintentionally) for e.g. black patients getting assigned the same level of risk as white patient even though they are more sick (Ziad et al 2019).

For our proof of concept, we trained a GAN using reinforcement learning (RL) by adding a penalty function term to the loss function, to minimize sequences with strong indication of user race and gender. It is important to note that even though data is the major source for bias, simple removing sensitive elements such as race and gender does not remove bias in the outcomes.

GitHub code (less than 1 page)
Link: https://github.com/pamegup/bias-detect-remove-tool

.



```
def load_ICU_data(path):
    column_names = ['Id', 'BIRTHDATE', 'DEATHDATE','ICU Admission' ,'SSN', 'DRIVERS', 'PASSPORT', 'PREFIX',
                    'FIRST', 'LAST', 'SUFFIX', 'MAIDEN','MARITAL','race', 'gender', 'BIRTHPLACE', 'ADDRESS']

    input_data = (pd.read_csv(path, names=column_names,
                              na_values="?", sep=r'\s*,\s*', engine='python'))
               # .loc[lambda df: df['race'].isin(['White', 'black', 'asian'])])

    input_data

    # sensitive attributes; we identify 'race' and 'sex' as sensitive attributes
    sensitive_attribs = ['race', 'gender']
    Z = (input_data.loc[:, sensitive_attribs]
         .assign(race=lambda df: (df['race'] == 'white').astype(int),
                 gender=lambda df: (df['gender'] == 'M').astype(int)))

    # targets; 1 admit to ICU , otherwise 0

    y = (input_data['ICU Admission'] == 'Yes')
    # y = icu_only[row_filter].Id
    # features; note that the 'target' and sentive attribute columns are dropped
    X = (input_data
         .drop(columns=['ICU Admission', 'race', 'gender'])
         .fillna('Unknown')
         .pipe(pd.get_dummies, drop_first=True))

    print(f"features X: {X.shape[0]} samples, {X.shape[1]} attributes")
    print(f"targets y: {y.shape[0]} samples")
    print(f"sensitives Z: {Z.shape[0]} samples, {Z.shape[1]} attributes")
    return X, y, Z
```

Methodology Overview

A Generative Adversarial Network, or GAN, is a specific type of machine learning where two neural networks, a generator and a discriminator, compete with each other in a zero-sum game. Typically, the generator creates fake data and the discriminator attempts to classify the data. Then, the discriminator will adjust either the generator or itself in an attempt to improve its accuracy. This process, also known as adversarial learning, continues until the discriminator is not able to distinguish between the real and the generated data.

Approach: Train a GAN by using an adversarial network for enforcing pivotal property on the predictive model. Pivotal property is used to ensure the model outcomes do not depend on unknown parameters, in this case race and gender.

Method:

Using Logistic Regression statistical model using three variables X, y and Z.

X is the input, which is patient data for COVID-19 patients

y is the prediction, admission into ICU; and

Z is the race and gender – sensitive attributes we want the model to learn to not consider when predicting admission to ICU.

Objective : Democratizing admission into ICU for COVID-19 patients without race or gender bias. In our model it is to maximize the ability to predict y, while minimizing the ability to predict Z, given an input X.

Thus, logistic regression is used in order to ensure that race and gender do not have an effect on the ICU admission.

Inspiration for this methodology –

We are using repurposed seminal work by Goodfellow et al where they demonstrated a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. improving data generation as a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to 1/2 everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples.

2017 NIPS paper "Learning to Pivot with Adversarial Networks" by Louppe et al.

Value Proposition

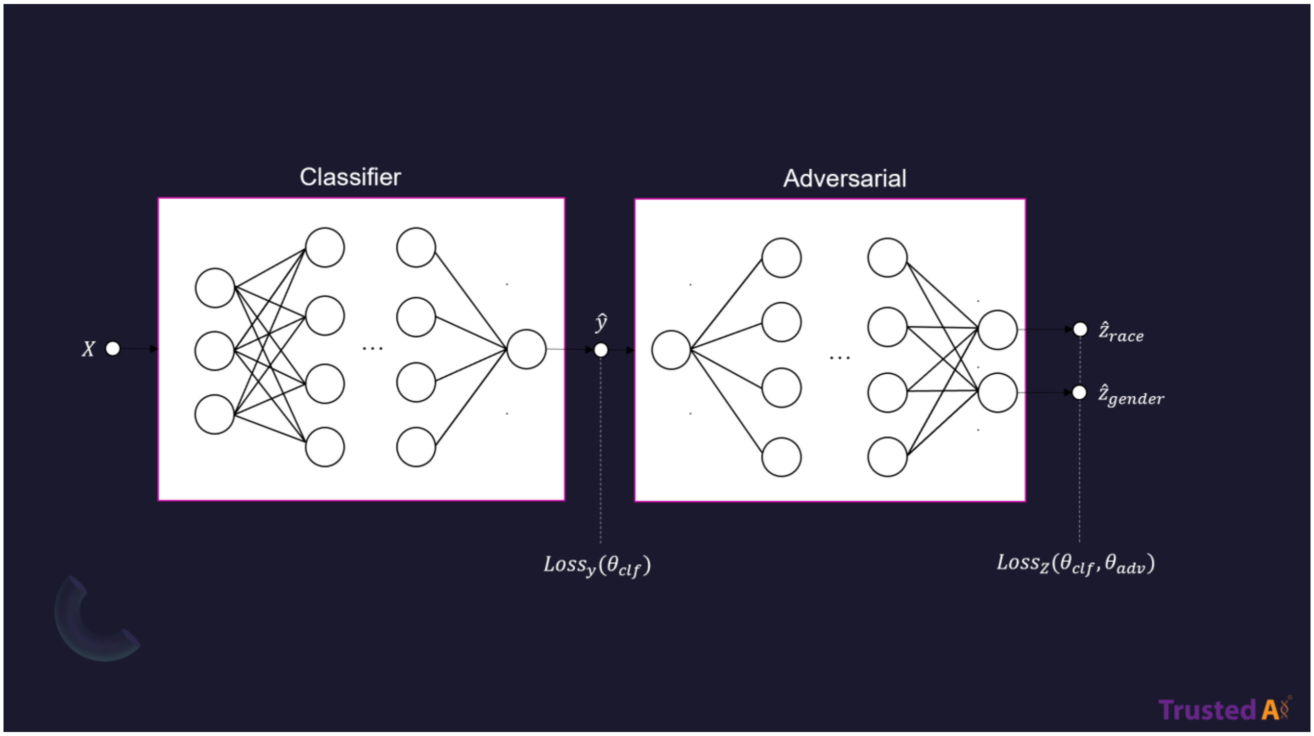
Our strategy is innovative and has two compelling advantages:

First, the fact that the tool approach is to reduce bias variables to unwanted or nuisance variables - Lets review the nature of the zero-sum game the classifier and adversarial are engaged in.

For the classifier the objective of twofold: make the best possible ICU admission recommendation whilst ensuring that race or sex cannot be derived from them.

So, it learns to minimize its own prediction losses while maximizing that of the adversarial

The objective during the game for the adversarial: predict race and sex based on the predictions of the classifier. The adversarial does not care about the prediction accuracy of the classifier. It is only concerned with minimizing its own prediction losses.



Secondly we bring expertise in data security, privacy, integrity, quality, regulatory compliance and governance in AI by design.

Healthcare Scenario

The tool would be used in a healthcare setting to build trust and confidence in the analysis of patient data. Here we have used it for ensuring ICU patients are considered free from bias attributes of race, gender and ethnicity by removing initial bias indicators for training the target state.
Intent is to reduce the unintended consequences of Bias in AI and the tendency to assume bias when referring to particular ethic groups or genders when recommending care programs or diagnosing healthcare concerns.
Our tool will address latent bias such as social and statistical bias over time by monitoring the model’s training and results based on the parameters it is trained upon and providing reported results.
Bias occurs in ML models at the early stages of data analysis and our model is designed to remove the identified social biases of gender, ethnicity and ICU referrals from COVID-19. Then targeted language model (LM) generates realistic but ethnicity and gender -oblivious outputs. We trained such debiased LMs with generative adversarial networks (GAN) through reinforcement learning (RL).
We will account for consistent evaluation and assessment of the algorithm over time by working out the algorithm’s runtime as a function of the problem-specific instance features and parameters and optimizing this by ensuring the model can run successfully within a given cap time and by analyzing performance over time.
By de-biasing we lower risk level in across all patients.
Internally engineers and ML experts with knowledge of the risks and ramifications of bias emerging in AI and data trained on models would be responsible for investigating cause and remediation options to intervene to ensure balance and reduction in latent biases. The tool can be implemented easily with access to GPU and local python libraries and notebooks.

Operational Requirements

Google Colab Pro - CPU details for Google Colab : 2vCPU @ 2.2GHz

13GB RAM

GPU: 1xTesla K80 , compute 3.7, having 2496 CUDA cores , 12GB GDDR5 VRAM

CPU: 1xsingle core hyper threaded Xeon Processors @2.3Ghz i.e(1 core, 2 threads)

Sustainability Plan

Our sustainability plan would involve the input of a multidisciplinary team with real world experience in a health care setting to ensure it is sustainable and performing optimally.

Our solution is industry and vertical agnostic and our aim is to create a library routine that any developer can provide input what bias is that they wish to detect and make it routine and build a model to enable them to de-bias their own data and it will reduce over cost and resource does not have to be spent to figure out their own solutions, routines will be efficient and no need to re-create the code.

The solution will make ICU admissions in hospitals fair and not based on race and gender, as hospitals are highly regulated organizations, under HIPPA compliance requirements and other regulations.

If deployed with an efficient runtime and deployed on GPUs on cloud platforms that are cost efficient the up keep costs will be minimal. this must be monitored and periodically evaluated.

Generalizability Plan

Using GANS is a powerful approach to address de-biasing algorithmic from 2 perspectives:
We are not only identifying the bias we are also helping the development of a targeted state. This is very useful from setting fairness goals at the outset of the project.

Our proof of concept is a solution to help make ICU admissions in hospitals fair and not based on race and gender, as hospitals are highly regulated organizations under HIPPA compliance requirements and this will require a particular data strategy.
The solution is industry and vertical agnostic and our aim is to create a library routine that any developer can provide input what bias is that they wish to detect and make it routine and build a model to enable them to de-bias their own data and it will reduce over cost and resource does not have to be spent to figure out their own solutions, routines will be efficient and no need to re-create the code.
The tool can be leveraged in any clinical discipline whereby there is a need to de-bias patient data in a healthcare setting, such as A&E, ICU, cardiology

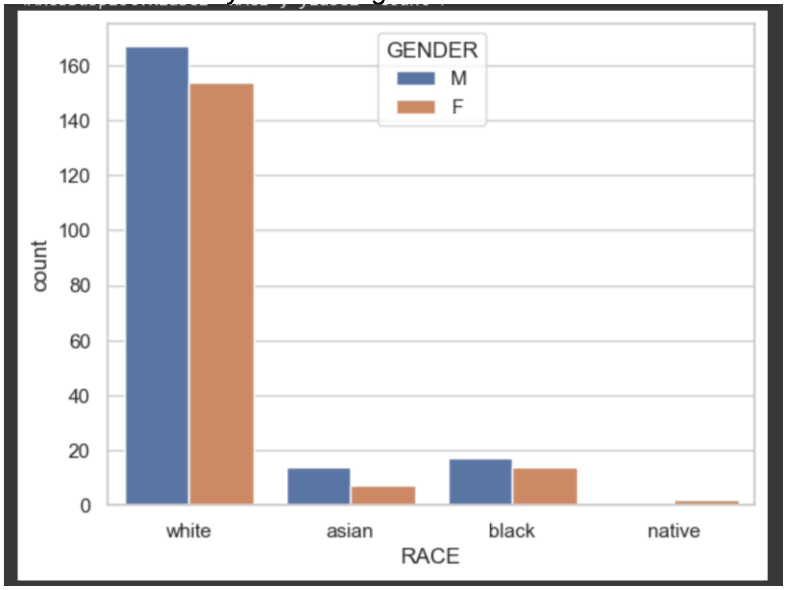
If deployed with an efficient runtime and deployed on GPUs on cloud platforms that are cost efficient the up keep costs will be minimal. this must be monitored and periodically evaluated.

We completed an initial analysis of the datasets in Seaborn for analysis of existence of bias and found correlation between gender and ethnicity and numbers of ICU cases, it did not suggest follow up investigations for other root causes.

Implementation Requirements

This would require a Data Scientist or ML Engineer with experience of training/ adjusting and maintaining GAN Algorithms to ensure controls and accuracy over time.

For this challenge we took the Synthia MASS 10k COVID-19 dataset, which does not contain any PII that we are exposing and is synthetic. We did analysis of the data and saw there were differences in ICU admissions by race and gender.



We decided to create an approach to ensure that race and gender were not bias factors in selection for admission to ICU.

Lessons Learned (less than 1 page)

The first challenge was to identify the most appropriate dataset that would provide the most complete data on race. gender, ethnicity bias. We initially analysed Columbia public data set for sample synthetic datasets.

Second challenge was designing a model agnostic bias detection tool that was simple to deploy across multiple settings.

References

www.towardsdatascience.com

Synthia COVID-19 dataset

What's the hardware spec for Google Colaboratory? - Stack Overflow

“Mitigating Bias in AI Using Debias-GAN”, WWT Artificial Intelligence Research & Development, October | 2021]

Algorithm runtime prediction: Methods & evaluation | Elsevier Enhanced Reader