Data Bias Detection in Machine Learning

Harshaprabha N Shetty*, Sony Asampalli*, Prabhu Vara Prasad Bonam*

* Responsible AI COE, Accenture Technology , Bengaluru, India

DOI: 10.29322/IJSRP.12.10.2022.p13071
http://dx.doi.org/10.29322/IJSRP.12.10.2022.p13071

Paper Received Date: 14th September 2022
Paper Acceptance Date: 15th October 2022
Paper Publication Date: 21st October 2022

Abstract- The paper talks about the various Bias detection methods using statistical measures. The methods are applied on binary as well as multinomial data.

Index Terms- bias, detection, binary, multi-class, detection, hypothesis

I. Introduction

Bias is defined as 'AI Bias or Machine Learning Bias where model produces results with prejudices due to various erroneous assumptions in the model building process'. We have biases due to data and due to the algorithms. The entire paper is divided into 6 sections- 1) Background and Related work 2) Data Bias 3) Types of Data Biases 4) Detection of data biases 5) Results and Discussion 6) Future work. The paper explains about the bias due to data and how can we detect them using various statistical measures. This helps the machine learning scientists to give an accurate result. Debiased data gives trust to the users in the society.

Background and related work

Background

There is a need for detection of bias in the data as we have seen that the results are not accurate enough. However, there are many measures available in various tools used by various organizations. Most of the tools in the market have used these measures for binary type of data.

Related Work

The existing literature identifies the bias in the data by converting multi-level data into binary data. Sorelle A. Friedler [2] have compared multiple fairness measures and found correlation between them. They also found that these measures are sensitive to fluctuations in the datasets. AI360 has ratio and difference versions of metrics.[1]. The paper on 'Fairness Measures for Machine Learning in Finance' [3] concentrates on the detection measures for data bias. They explain about synergy between bias and legal considerations in the finance industry.

All the above papers and other existing literature talk about the metrics on the datasets where only binary levels are used for calculation of data bias detection.

In this paper we have applied the measures on multiple levels of data on different datasets.

data bias

We have been using various Machine Learning (ML) and Artificial Intelligence (AI) methods for various business problems. The success of these applications has made the scientists to explore in social justice, hiring processes etc. The increasing use of ML/AI in these domains have induced bias in the reports related to gender, race etc. This has created huge impact on some of the existing models and users had to ban those models due to biases. The main source of these biases are the Data used to train the ML/AI models. As the old saying goes 'Garbage in, garbage out'. The bias is due to structural characteristics of the data used in the analysis of business problems.

type of data biases

As we have noticed from various domains, the data collected from people through responses of people, feedback from people, sampling taken from the other data sources, data generated from human for some studies, data generated from social media, or curated data for analysis etc. and so on will definitely are biased data. We should be aware of the different types of biases which may harm the precious data.

Bias due to the survey

The survey data takes input from the people for a specific purpose. Only small portion of the targeted group will contribute to giving the responses. This is highly biased data, and the outcome of that purpose will not give a good result or help in making better decisions.

Bias due to the drift in the system

The system has collected data while starting any process; but after certain years the data might have changed. If we do not update the system with the latest data, then again, we will have bias in the outcomes of our models.

Bias due to the exclusion of critical variables

The data collection and preparation stage are very critical for the business. The data profiling stage needs to be done by people who are experts in that domain. And there are cases where they may miss out business critical features. This may lead to bias in the outcomes of our models.

Bias due to the selection of data from the publicly available data or confirmation bias

The data collection from the social media where it has been posted by the humans are incomplete. We confirm that the available of the social media are correct all the time. The curated data is biased. This curated data can give improper messages in the outcome of our analysis.

Bias due to the selection of data from recommender systems or selection bias

In retail when we select the items as an output from the recommender systems and the users will not be able to get the full set of data. The selected data may not represent the data which we think. This induces bias in the data. We will miss out the items and the analysis will not be accurate. The decision taken by the analytics will mislead the business groups.

Historical Bias

Sometimes we carry with us the old beliefs and apply that on data. For example: In Indian context, while telling a story about the cook we usually refer that as a female. Other examples like nurse in a hospital is referred as a female. When referring an earning member in the family is most of the times it is referred as man than a woman.

Survivorship Bias

It is a human tendency to favor the winner than the looser. While collecting data we tend to include the characteristics of the winner or rich or poor or looser – any one category based on the situation. This induces bias in the data collected for the purpose. We will get biased results which will focus only on that particular group in the society.

Availability Bias or outlier Bias

This is a very rare case but quite common practice by most of the humans to go behind the rare new invention without doing much analysis. This may affect our business in the long run. The data may not represent the entire population due to availability of few such cases. There is a high possibility that we are biased towards that minimal information.

On the other hand, we can have cases where the average data is available for making some decisions with the data. But this is biased as the outliers are hidden when we have average data values. Always better to use raw values than the average values to get the better picture or for better decision-making process.

Methodology

Outline of the process

The analysis which we are performing while detecting the bias is classification problem. The paper talks about binary classification on target variable being studied.

detection of data biases

Data is most important part in our day-today life while making business decisions. Identifying the biases in our data is crucial part in our analysis. This can help us to take the next steps to mitigate the biases and which helps the business to get a meaningful interpretation.

There are various methods to identify the biases in the data. For the purpose of writing this paper we have considered the Adult.csv, German credit.csv and Comapass.csv from public repository.

We will study these methods in the below sub sections.

Disparity Impact

Disparity Impact is a measure of discrimination in the data [5]. If any outcome is seen to a greater or lesser extent between populations, then there is disparity.

Disparity is based on the four fifth rule which states that if the selection rate for a certain group is less than 80 percent of that of the group with the highest selection rate, there is adverse impact on that group.

Step 1- Find the selection rate for each group. For each group, divide the number of each level selected by the total of all levels.

step 2: Determine the group that is most favored and the group that is least favored. For positive personnel actions, the most favored group has the highest rate. For negative personnel actions, the most favored group has the lowest rate.

Step 3: Calculate the impact ratio analysis for each group. This compares the favorable group selection rate with the selection rates of all other groups.

Step 4: Determine whether the result is less than 80%.

If result that is less than 80% then is considered evidence of adverse impact.

Fisher's test

Fisher's exact test is useful for categorical data that result from classifying objects in two different ways; it is used to examine the significance of the association between the two kinds of classification.

Use Fisher's exact test when you have two nominal variables. You want to know whether the proportions for one variable are different among values of the other variable.

Step 1- H_0: (null hypothesis) The two variables are independent.
H_1: (alternative hypothesis) The two variables are *not* independent.

Step 2- if $p < 0.05$ then we reject H0

Chi-squared test of independence

The Chi-square test of independence checks whether two variables are likely to be related or not. We have counts for two categorical or nominal variables. We also have an idea that the two variables are not related,

The basic idea in calculating the test statistic is to compare actual and expected values, given the row and column totals that we have in the data. First, we calculate the difference from actual and expected for each combination of levels.

H0: Two or multiple variables are independent
H1: Two or multiple variables are dependent
The test statistic is lower than the Chi-square value. You fail to reject the hypothesis of independence.

Class imbalance (CI)

CI is under-representation of the disadvantaged group in the dataset. We want the differences to lie in *(-1; +1).*

We define, $CI = \frac{(n_a - n_d)}{n} \in (-1, +1)$ (1)

Ex: Assume a dataset of Adult Census on column gender binary class, 21790 are Male and 10771 are Female, CI = 0.336. There exists Bias.

Assume the same data for multiclass Race between difference levels,

Class imbalance (CI)	Amer-Indian-Eskimo	Bias
Asian-Pac-Islander	0.539259	Yes
Black	0.818923	Yes
White	-0.068729	Yes
Other	0.977886	Yes

Difference in positive proportions in observed labels(DPL)

Let $q_a = \frac{n_a^{(1)}}{n_a}$ be the ratio of type 1 for the advantaged class and $q_d = \frac{n_d^{(1)}}{n_d}$ be the same for the disadvantaged class.

$DPL = q_a - q_d$, and $DPL = \frac{q_a - q_d}{q_a + q_d} \epsilon(-1, +1)$

Ex: Assume a dataset of Adult Census on column gender binary class

Assume the same data for multiclass column Race having different levels,

Difference in positive proportions in observed labels(DPL)	Amer-Indian-Eskimo	Bias
Asian-Pac-Islander	0.0143	Yes
Black	0.0196	Yes
White	0.0296	Yes
Other	0.2330	Yes

Kullback and Leibler (1951) Divergence (KL)

We compare the probability distribution of the advantaged class (P_a) with that of the disadvantaged class (P_d), using KL divergence, i.e., relative entropy (Kullback in fact preferred the term "discrimination information")

$$KL(P_a; P_d) = \sum_y P_a(y) \log\left[\frac{P_a(y)}{P_d(y)}\right] \geq 0$$

Ex: Assume a dataset of Adult Census on column gender binary class

TABLE IX. Gender	TABLE X. Kullback and Leibler (1951) Divergence (KL) TABLE XI.	TABLE XII. Bias
TABLE XIII. Male vs Female	TABLE XIV. 0.14130207520131896 TABLE XV.	TABLE XVI. Yes

Assume the same data for multiclass column Race having different levels,

TABLE I. Gender	TABLE II. Difference in positive proportions in observed labels (DPL) TABLE III.	TABLE IV. Bias
TABLE V. Male vs Female	TABLE VI. 0.196 TABLE VII.	TABLE VIII. Yes

Kullback and Leibler (1951) Divergence (KL)	Amer-Indian-Eskimo	Bias
Asian-Pac-Islander	0.06808070552347882	Yes
Black	0.0003099579683429939	Yes
White	0.060724415110749067	Yes
Other	0.00307520634056992	Yes

Jensen-Shannon divergence (JS)

If the distribution of the combined classes is P, then,

$$JS(P_a;\ P_d; P) = \frac{1}{2}[KL(Pa;\ P) + KL(P_d;\ P)]\ \geq 0$$

Ex: Assume a dataset of Adult Census on column gender binary class

TABLE XVII. Gender	TABLE XVIII. Jensen-Shannon divergence (JS) TABLE XIX.	TABLE XX. Bias
TABLE XXI. Male vs Female	TABLE XXII. 0.0303	TABLE XXIII. Yes

Assume the same data for multiclass column Race having different levels,

Lp norm (LP)
Instead of the entropy differences in KL and JS, we may consider norm differences. For $P \geq 1$, we have,

$$Lp(Pa;\ Pd) = \left[\sum_y |P_a(y) - P_d(y)|^p\right]^{1/p} \geq 0$$

Ex: Assume a dataset of Adult Census on column gender binary class

TABLE XXIV. Gen der	TABLE XXV. Lp norm (LP) TABLE XXVI.	TABLE XXVII. B ias
TABLE XXVIII. M ale vs Female	TABLE XXIX. 0.2 7856 TABLE XXX.	TABLE XXXI. Yes

Assume the same data for multiclass column Race having different levels,

Lp norm (LP)	Amer-Indian-Eskimo	Bias
Asian-Pac-Islander	0.1498	Yes
Black	0.00812	Yes
White	0.1401	Yes
Other	0.0235	Yes

Total variation distance (TVD)
This is half the L1 distance:

$$TVD = \frac{1}{2} L1(P_a; P_d) \geq 0$$

this measure is non-negative
Ex: Assume a dataset of Adult Census on column gender binary class

TABLE XXXII. ender	TABLE XXXIII. Total variation distance (TVD) TABLE XXXIV.	TABLE XXXV. ias
TABLE XXXVI. ale vs Female	TABLE XXXVII. 0.196 97796466995765 TABLE XXXVIII.	TABLE XXXIX. es

Assume the same data for multiclass column Race having different levels,

Total variation distance (TVD)	Amer-Indian-Eskimo	Bias
Asian-Pac-Islander	0.1498	Yes
Black	0.0081	Yes
White	0.0235	Yes
Other	0.1401	Yes

Kolmogorov-Smirnov (KS), two-sample approximated version

$$KS = \max(|P_a - P_d|) \geq 0$$

Jensen-Shannon divergence (JS)	Amer-Indian-Eskimo	Bias
Asian-Pac-Islander	0.01860518388794966	Yes
Black	7.824046471884439e-05	Yes
White	0.0165547644387824	Yes
Other	0.0007425044248053946	Yes

It is possible to evaluate the KS statistical test from the following distance measure, where the null hypothesis is rejected at level

$$KS > C(\alpha)\sqrt{\frac{n_a + n_d}{n_a . n_d}}$$

The value of $C(\alpha)$ is given by $C(\alpha) = \sqrt{-\ln\left(\frac{\alpha}{2}\right) . \frac{1}{2}}$

Ex: Assume a dataset of Adult Census on column gender binary class

TABLE XL. G ender	TABLE XLI. Kol mogorov-Smirnov (KS) TABLE XLII.	TABLE XLIII. ias
TABLE XLIV. M ale vs Female	TABLE XLV. 0.19 6 TABLE XLVI.	TABLE XLVII. es

Assume the same data for multiclass column Race having different levels,

Kolmogorov-Smirnov (KS)	Amer-Indian-Eskimo	Bias
Asian-Pac-Islander	0.1498	Yes
Black	0.0081	Yes
White	0.0235	Yes
Other	0.1401	Yes

.Conditional Demographic Disparity in Labels (CDDL)
The metric asks the following question: Is the disadvantaged class a bigger proportion of the rejected outcomes than the proportion of accepted outcomes for the same class? We note that just this

question alone would lead to an answer to whether demographic disparity exists (DD),

$$D = \frac{No\ \ of\ rejected\ applicants\ from\ the\ protect\ facet}{Total\ rejected\ applicants} = \frac{n_d^{(0)}}{n_d}$$

$$A = \frac{No\ \ of\ accepted\ applicants\ from\ the\ protect\ facet}{Total\ accepted\ applicants} = \frac{n_a^{(1)}}{n_a}$$

If $D > A$, then demographic disparity (DD) exists. CDDL arises when demographic disparity exists on average across all strata of the sample on a user-supplied attribute. We will subgroup the sample and compute DD for each subgroup, and then compute the count-weighted average of DD. The function is as follows:

$$CDDL = \frac{1}{n}\sum n_i \, . DD_i$$

Ex: Assume a dataset of Adult Census on column gender binary class

Gender	Conditional Demographic Disparity in Labels (CDDL)	*Bias*
Female	0.2573300573300573	Yes
Male	0.2573300573300573	Yes

Assume the same data for multiclass column Race having different levels,

Race	Conditional Demographic Disparity in Labels (CDDL)	*Bias*
Amer-Indian-Eskimo	0.0146	Yes
Asian-Pac-Islander	0.0445	Yes
Black	0.1889	Yes
Other	0.2025	Yes
White	1.0367	Yes

Figures and Tables

Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

Table Type Styles

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

Sample of a Table footnote. (*Table footnote*)

Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

II. Results and Discussion

The results Shows that we want to make the world equitable to entire society.

future work

There is a possibility to explore other measures when the target feature in the model supports multi-level classification.

Acknowledgment

We would like to thank the team who has helped in development of the Data Bias detection Tool.

References

[1] AI Fairness 360 - Resources (mybluemix.net)

[2] Sorelle A. Friedler, Haverford College, Carlos Scheidegger, University of Arizona, Suresh Venkatasubramanian, University of Utah, Sonam Choudhary, University of Utah, Evan P. Hamilton, Haverford College k Derek Roth, Haverford College, A comparative study of fairness-enhancing interventions in machine learning, arXiv:1802.04422v1 [stat.ML] 13 Feb 2018.

[3] Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardtt, Jared Katzman,Krishnaram Kenthapadi, Pedro Larroy , Pinar Yilmaz, Bilal Zafar, "Fairness Measures for Machine Learning in Finance"

[4] Understanding Data Bias. Types and sources of data bias | by Prabhakar Krishnamurthy | Towards Data Science

[5] Kyle E. Brink,Jeffrey L. Crenshaw,Personnel Board of Jefferson County, 'Adverse Impact:What is it?How do you calculate it?'

Authors

First Author – Harshaprabha N Shetty, Responsible AI COE Accenture Technology, Bengaluru, India, h.n.shetty@accenture.com

Second Author – Sony Asampalli, Responsible AI COE, Accenture Technology, Hyderabad, India, sony.asampalli@accenture.com

Third Author – Prabhu Vara Prasad Bonam, Responsible AI COE, Accenture Technology , Hyderabad, India h.n.shetty@accenture.com