Journal of Intelligent Learning Systems and Applications, 2025, 17(1), 1-7
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Early Machine Learning Models: Biases Shown and How It Can Be Corrected

Nandini Guduru

Frisco ISD, Frisco, TX, USA
Email: nandini.guduru@gmail.com

How to cite this paper: Guduru, N. (2025) Early Machine Learning Models: Biases Shown and How It Can Be Corrected. *Journal of Intelligent Learning Systems and Applications*, **17**, 1-7.
https://doi.org/10.4236/jilsa.2025.171001

Received: October 23, 2024
Accepted: December 23, 2024
Published: December 26, 2024

Abstract

The purpose of this research paper is to explore how early Machine Learning models have shown a bias in the results where a bias should not be seen. A prime example is an ML model that favors male applicants over female applicants. While the model is supposed to take into consideration other aspects of the data, it tends to have a bias and skew the results one way or another. Therefore, in this paper, we will be exploring how this bias comes about and how it can be fixed. In this research, I have taken different case studies of real-world examples of these biases being shown. For example, an Amazon hiring application that favored male applicants or a loan application that favored western applicants is both studies that I will reference in this paper and explore the situation itself. In order to find out where the bias is coming from, I have constructed a machine learning model that will use a dataset found on Kaggle, and I will analyze the results of said ML model. The results that the research has yielded clarify the reason for said bias in the artificial intelligence models. The way the model was trained influences the way the results will play out. If the model is trained with a large amount of male applicant data over female applicant data, the model will favor male applicants. Therefore, when they are trained with new data, they are likely to accept applications that are male over female despite having equivalent parts. Later in the paper, I will dive deeper into the way that AI applications work and how they find biases and trends in order to classify things correctly. However, there is a fine line between classification and bias and making sure that it is rightfully corrected and tested is important in machine learning today.

Keywords

Machine Learning, Artificial Intelligence, Bias, Generative AI, Training Data, Testing Data

1. Introduction

To provide an adequate background of the research, the fundamentals of AI must be known and understood. Artificial Intelligence is a branch of computer science that involves the process of teaching machines to learn from experience and perform human-like tasks. AI involves computer systems that work to perform complex tasks that only humans were historically able to do (things like decision-making, problem-solving, etc.). Furthermore, regarding AI, Machine Learning is a branch of computer science that falls under Artificial Intelligence which refers to the process of training and testing the data that will go into the Artificial Intelligence model [1]. The training data for your Machine learning model refers to the hundreds/thousands of pieces of data that are labeled accordingly. Using this data, the model finds patterns and trends within the training data and learns from what it sees. After the model gets familiar with the training data, that is where the use of the testing data will come into play. The model will use what it's learned from the training data to label the testing data accordingly. Through this, we're able to see how artificial intelligence is a form of machinery that has the ability to learn. It can find patterns and learn from them, applying them to the testing data that it is given [1].

Being an individual studying machine learning and artificial intelligence, I know the importance of training and testing data in order to understand the fundamental processes of machine learning. When referring to testing data, the main purpose is to evaluate the performance of a trained model. When you take a portion of your data and withhold it exclusively for testing, it is easy to assess how well the model generalizes to unseen data [1].

AI in the workplace is a concept that's been being implemented for a while, but it's a concept that isn't understood as well as it should be. Many people use AI in day-to-day life, which they don't even realize is AI. Things like Facial Recognition, Spam Email filtering, and Google translate are all instances of AI being used in everyday life where their importance is overlooked. Furthermore, these instances of AI are used in different career fields and impact people in ways many common people are not aware of. For example, artificial intelligence and machine learning models are used in healthcare settings. One popular example is through Generative AI.

Generative AI is a type of Artificial Intelligence that can create new content and new ideas to form stories, conversations, images, etc., with its consumers. Examples of where Generative AI has been most effective are through Siri, Google Assistant, Alexa, etc. When people use Siri or Alexa, they are using a Generative AI that is able to create ideas to hold a conversation. This form of AI is also used in medicine, where it is used to streamline lots of data and commit more time to patient care. Generative AI can increase practitioners' productivity and allow better care to be delivered to patients. While this form of AI has been proven to protect HIPPA rights, some pose a valuable argument that it cannot be fully guaranteed that HIPPA rights will be protected within this artificial intelligence model.

Similar to this, there are many examples of how artificial intelligence is being used in the fields of finance. The most prominent way that it's being used is an AI that filters through people's financial histories to find out if they are suitable for a specific loan. However, biases can also be seen here, where different AI models have favored some applications over others on a basis other than financial history. When given lots of data, artificial intelligence and different machine learning models tend to find patterns in things that are not supposed to have patterns. AI has advanced to where some can argue that it is doing its job too well [2]. Its job is to find patterns and trends with the training data that are fed into the model, but sometimes, models will find patterns in things that can be seen as biased. For example, things like gender or race shouldn't have any preference or pattern attached to them, as those should not be relevant when determining a loan. However, sometimes data like gender is useful. When AI is being used in the medical field, there are instances where having gender is important because there are some medical issues that are specific to certain genders. For this reason, it is important to include gender in training data when using medicine.

All of these examples open up different arguments about artificial intelligence and how biases can be seen in so many different models. What type of data to use is also an issue: in medicine, having gender is useful, while in the finance field, it won't do much good. Because of this, biases in AI are present, no matter how small they may be [3].

In this paper, I want to be able to solve this problem of bias in AI and explore how it can be corrected. Since AI is used in daily life, knowing how to correct and deal with biases it can portray is important to the progression of AI within society.

2. Case Study

When talking about Artificial Intelligence in different workplaces, one prime example of AI showing a bias in healthcare is through racial biases that have been observed in previous years. This is seen in a 2019 AI that showed racial prejudice in healthcare and was used in different US hospitals. This algorithm was designed to predict which patients needed extra medical care, and it was used by over 200 million people. Essentially, it would analyze their healthcare cost history and make the assumption that cost indicates a person's healthcare needs. Based on these factors, it would make decisions on the cost that patients would need [4]. This was flawed, however. This assumption between cost and healthcare needs didn't account for the differences in cost which black and white patients pay for healthcare. While this might seem straightforward (the more medical costs you have, the more medical needs you will have), there were lots of nuances to this in healthcare in 2019. During this time, "black patients were more likely to pay for active interventions like emergency hospital visits, despite showing signs of uncontrolled illness" (Denison). Because of this, the intelligence model had given black patients lower risk scores than white patient counterparts, were put on par with healthier white patients in terms of costs, and didn't qualify for extra care as much as white

patients with the same needs. So a really large issue with this algorithm was that it assigned many black patients the same level of risk as sicker white patients. And because of this, the model "reduced the number of black patients identified for extra care by more than half" (Obermeyer). Thai bias has been able to occur because the intelligence model uses health costs to directly reflect the level of health needs that patients are in need of. However, studies were shown which have evidence that "less money is spent on black patients who have the same level of need", and this creates a growing bias in the health algorithms (Obermeyer). So what's needed is to reformulate the algorithms so they no longer use health costs as the sole determiner of health needs. In today's day and age, healthcare systems have been relying on algorithms to identify patients with complex needs, give doctors and nurses more time with their patients, and provide better quality one-on-one patient care. So, with an algorithm there to take care of needs such as healthcare costs, the doctors can have more time freed up to develop this necessary doctor/patient bond. So, algorithms and intelligence models will continue to be used and implemented in healthcare for many years to come. So, being able to identify where biases are being seen and how to fix them is important for developments to be made in the Machine Learning field.

3. Building the Model

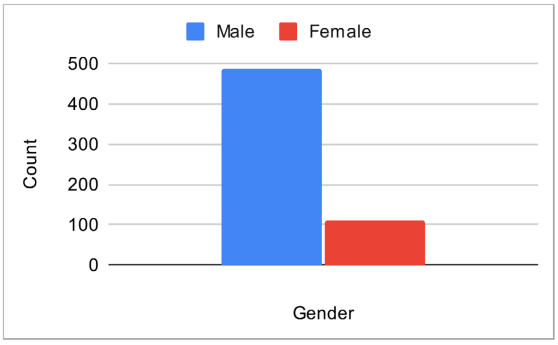
To understand the extent to which bias is present in AI, not only is it important to study cases where this is present but also to create your own case, which can be used to analyze and study at length. In order to explore bias in Machine Learning models, my mentor (Mr. Gabriel Ohaike) and I have created a Machine Learning model that would determine if an applicant was eligible for a loan. To construct the model, we used Python through the iTerm IDE. For this model, we've used a real-life dataset from Kaggle to provide training data for the model [5]. In theory, the model would analyze the financial history, employment status, income, etc. (all the things relating to financial and loan status) and determine if the applicant was eligible for this loan. As shown in **Figures 1(a)-(f)**, we can see the spread of the data and where different demographics lie. When observing bias, we are looking to pay the most attention to one factor, and in this case, my mentor and I have focused on the gender of the applicants. Referring back to **Figure 1(a)**, we can see how large the disparity is between the number of training data points that are male versus the number of training data points that are female. This goes back to the concept of training data showing the presence of bias. With one type of data being more represented than another in the training data, this is a pattern that the model will find and learn from when moving on to the testing data.

Once the model itself was run, we can see in **Figure 2** how the accuracy for the male data differed largely from the accuracy of the female data. With the female data, the accuracy in how the model predicted its availability for the loan was 16% less than that of the male data. Since the machine learning model was trained with a much larger amount of male data, the model found a pattern within the gender

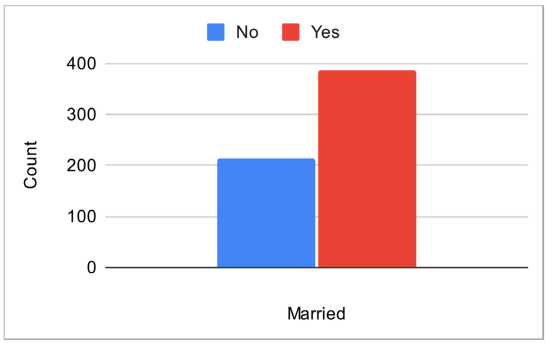
of the training data applicants. Since the data itself had a large number of male applicants, the larger number of data points in the training data that were labeled to accept the loan were male too. The model learned from this pattern and applied it to the testing data. And so, with the training data when presented with a male applicant versus a female applicant counterpart (meaning every other factor was identical), the model was more likely to accept the male applicant.

With this result in mind, my mentor and I ran the model again, but with one critical change to the training dataset we used. We removed the label gender from the data and ran the model with just the information on marriage, education, income, employment, etc. The results can be seen in **Figure 3**, where the disparity between male accuracy and female accuracy is closer. While before, the data was 80/64 male to female, the male has gone down, and the female has subsequently gone up a little—now 76/66. This model definitely runs better than the previous model, as the disparity of bias is less than what was observed before. However, there is still some obvious bias apparent as the male accuracy is still 10 percent higher than the female accuracy. This is simply because there was more male data in the training dataset than female data. Since there's more male data in general, the male accuracy will be higher—as there is just more data to work with.

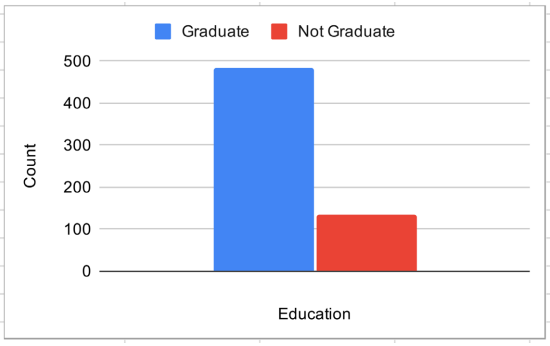
Taking this into account, we're able to observe the presence of something like gender, severely skew the results of a machine learning model, and bring bias into the equation. We were able to fix this problem to some degree by removing the label of gender altogether, but given that there were more male data than female data in the training dataset, there will be a preference for the male data.



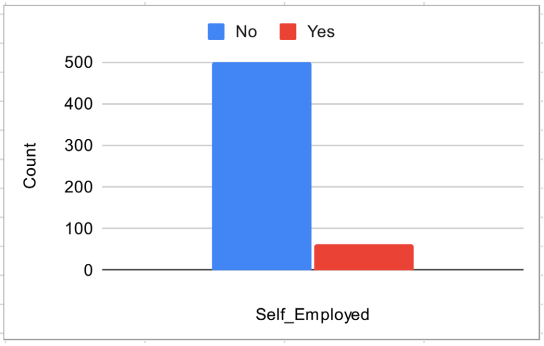
(a)



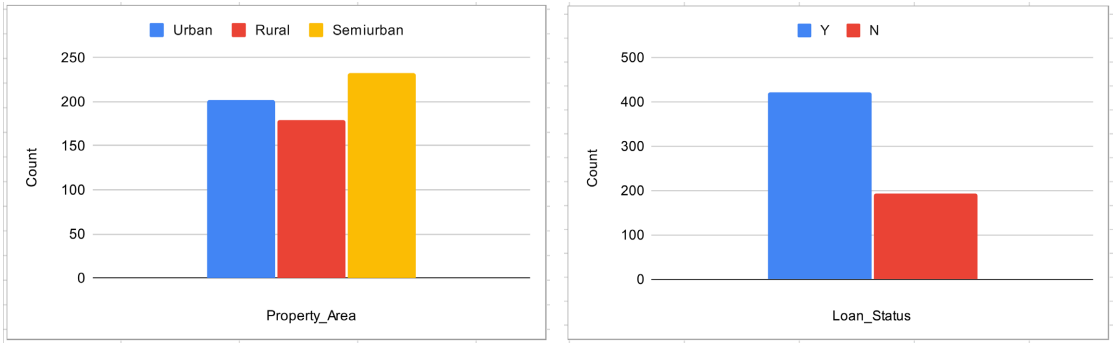
(b)



(c)



(d)



(e) (f)

Figure 1. (a) Distribution of gender; (b) Distribution of married; (c) Distribution of education; (d) Distribution of self-employed; (e) Distribution of property area; (f) Distribution of loan status.

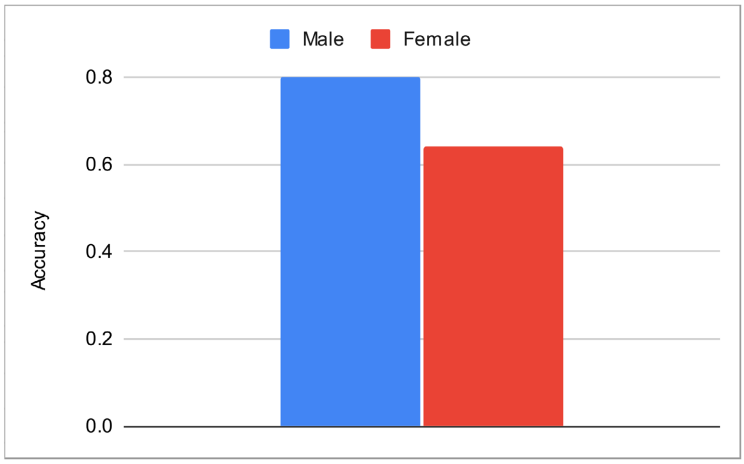


Figure 2. Male: TP: 40, FP: 10, FN: 0, TN: 0; Female: TP: 32, FP: 18, FN: 0, TN: 0.

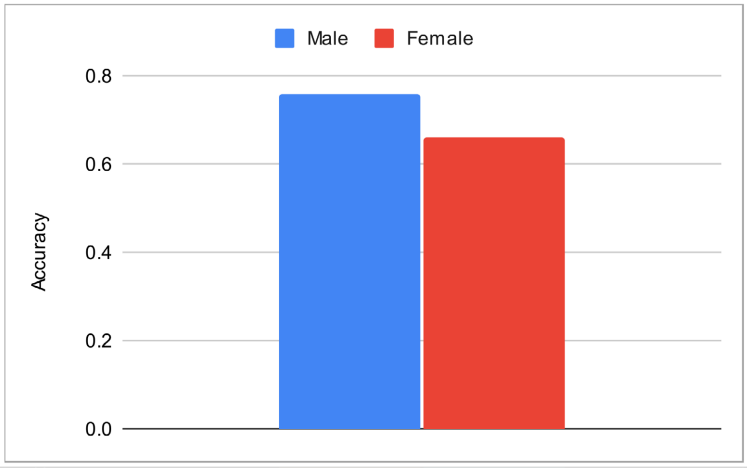


Figure 3. Male: TP: 38, FP: 12, FN: 0, TN: 0; Female: TP: 33, FFP: 17, FN: 0, TN: 0.

4. Data Analysis

With both the case study of AI in healthcare and the model we've created with the loan application, we're able to see how bias can be present in AI, bringing a

concern into the correct usage of AI. In the case study of healthcare, we explored how people were unfairly given treatment based on their gender; similarly, with the loan application model, we were able to see how some applicants were unfairly denied a loan based on their gender. When observing these two case studies, we were also able to see the common link between the biases that were shown in the results. The training data given to the model was skewed in a specific way, and this influenced how the model learned from the data. For example, the data fed into the loan application model consisted of a larger number of male applicants than female applicants. Because of this, the model illustrated a bias towards the male testing data over the female testing data. Because of this disparity in the training data, the model displayed a bias when it had run through the testing data.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Brown, S. (2021) Machine Learning, Explained. *MIT Sloan*, 21 April 2021. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

[2] Chapman University. Bias in AI. https://www.chapman.edu/ai/bias-in-ai.aspx

[3] Denison, G. (2023) 4 Shocking AI Bias Examples. *Prolific*, 24 October 2023. https://www.prolific.com/resources/shocking-ai-bias

[4] Obermeyer, Z. (2024) Dissection Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, **366**, 447-453. https://www.science.org/doi/full/10.1126/science.aax2342 https://doi.org/10.1126/science.aax2342

[5] Kumar, V. Loan Application Data. *Kaggle*. https://www.kaggle.com/datasets/vipin20/loan-application-data