Bias in AI Project

xchs48
xchs48@durham.ac.uk

I. Project Proposal

A. Motivation

I would like to investigate how bias in artificial intelligence could have a large effect on someone's life and their future opportunities. It is in these circumstances where the introduction of artificial intelligent systems need to be the most fair and unprejudiced as its decision could be very important for that person's future. I also want to investigate an are in which it was likely that such an AI system could be implemented to reduce bias, but inadvertently entrench it.

For this I want a dataset with a clear cut advantaged and disadvantaged group who could be compared and shown to have different outcomes before an artificial intelligence had been implemented. I then want to show how these decisions could be passed to an AI, and cloud its ability to make unbiased decisions. Finally, I want to show how such an issue could be addressed to prevent the final model being biased.

B. Method

The first step will be to find a dataset on which a classification model could be implemented, modelling the decision an AI in an important position might make. Next, evidence of bias in the current decision making progress would have to be found which could influence the AI model. It would be best to find a dataset for a decision that is important to someone's life to illustrate the effect a biased AI could have. This could be in the fields of hiring, justice or finance for example.

After a dataset is found, a model should be created from the data in a naive way, not accounting for the bias found in the dataset. This model should then be investigated to see if the bias already present has manifested itself in the model, causing the artificial intelligence to be biased. If this is the case, this will show that implementing an artificially intelligent system in this field without accounting for bias could cause the system to create unfair decisions further down the line which could affect people's lives in an unfair way.

Finally, a new model should be created, accounting for the bias in the original dataset, to show how such an issue could be resolved. This is important as it demonstrates how a less biased system can be created from a biased one.

C. Implementation

I intend to use Python 3 in a Jupyter notebook for this project. Python allows the use of a large range of libraries for machine learning, and is a language that future systems are likely to be implemented using. A Jupyter notebook will be used for ease of displaying results, rerunning parts of the project and making the code easier to understand. I will use the Sklearn library as it is a commonly used library for machine learning that many people are familiar with that provides a wide range of useful functions for pre-processing of data.

The dataset I intend to use is german_credit_data from the list of suggested projects as the subject matter of whether a loan is deemed risky or not is important to those receiving the loan as not receiving the loan could have a large effect on their lives if they are unable to make a purchase. This dataset also has a clear advantaged and disadvantaged class of male and female which can be easily compared by if their loans were deemed risky or not. This can be easily converted into a statistic as success rate of loan application.

A common classification machine learning model such as linear SVC or SVM will be used to show how bias can enter commonly used algorithms.

TABLE I
MALE DATA

	Age	Credit Amount	Duration
Average	36.8	3448	21.6
Std	11.0	2900	12.4
Min	20	276	4
Max	75	15945	72

TABLE II
FEMALE DATA

	Age	Credit Amount	Duration
Average	32.8	2878	19.4
Std	11.7	2603	11.0
Min	19	250	4
Max	75	18,424	60

II. PROJECT PROGRESS

A. Dataset

The dataset chosen is:
https://www.kaggle.com/kabure/predicting-credit-risk-model-pipeline?scriptVersionId=7037624

The project paper chosen is:
https://www.kaggle.com/janiobachmann/german-credit-analysis-a-risk-perspective

B. Cleaning

To use the machine learning functions in sklearn it is necessary for the variables to be numbers. Label encoder from the sklearn preprocessing package was used to convert all strings to numbers so they could be fed in. These functions do not accept NA as a value, so all NA values were converted to 0 beforehand. This was all the cleaning necessary.

C. Group Statistics

It is interesting to note that female applicants ask for less money over a shorter duration than male applicants on average. Also of note is that the average age of a female applicant is less than a males. This may explain some bias as it is likely that younger applicants are in a lower paying job and are therefore less likely to be accepted for a loan.

D. Bias in Dataset

In the dataset male applicants had a 72.3% chance of being successful, while female applicants had a 64.8% chance, despite asking for on average less money. This suggests there is bias against females when applying for credit.

It is hard to describe where this bias comes from. My only explanation would be that the current human decision process is biased. This suggests this process might benefit from the assistance of a artificially intelligent system.

TABLE III
PERCENTAGE SUCCESS RATE BY SEX

Sex	Biased 1	Biased 2	Unbiased
Male	72.1	75.6	73.2
Female	53.4	71.4	72.8

TABLE IV
ACCURACY OF MODELS PERCENTAGE

Data	Biased 1	Biased 2	Unbiased
Training	76.0	73.0	74.6
Test	71.3	72.6	72.8

E. Base Algorithm

A good bad decision is needed from the model, so a classification algorithm is needed. I tested both SVC Linear and Naive Bayesian. SVC Linear produced more accurate results, so I decided to use it.

F. Results from Base Algorithms

(Results are in Tables 3 and 4)

The first model shows a large bias. The difference between male and female success rate being almost 20%. The second model is much improved, with only a 4.0 percentage difference. Accuracy is similar for both, and while not particularly high, is similar to that in the project paper.

G. Algorithm to Remove Bias

To attempt to combat this bias, I suggest that the ratio of successful applicants for male and female should be made equal. To do this, the proportion of good and bad decisions across the population should be found. This should mean the number of successful applicants is consistent, rather than being increased.

I found this solution to be inadequate when implemented, still producing an 8 - 11% difference between male and female applicants, with the male success rate in the high 60s and the female around 60. To try and improve this, I weighted the female successful results, adding each one to the training set twice. This seemed to solve the problem, with the results falling within 1% of each other.

H. Reduction in Bias

As can be seen from Table 3, the final algorithm has a sizeable reduction in bias compared to the previous models. With only an 0.4% difference between the two sexes, bias has been all but removed.

I. Accuracy

The accuracy of all algorithms was just above 70%. This is in line with the project paper, which recorded 0.7325. The accuracy values can be found in Table 4. The unbiased model produced similar accuracy scores to the other models, suggesting that the accuracy of the artificial intelligence had not been compromised.

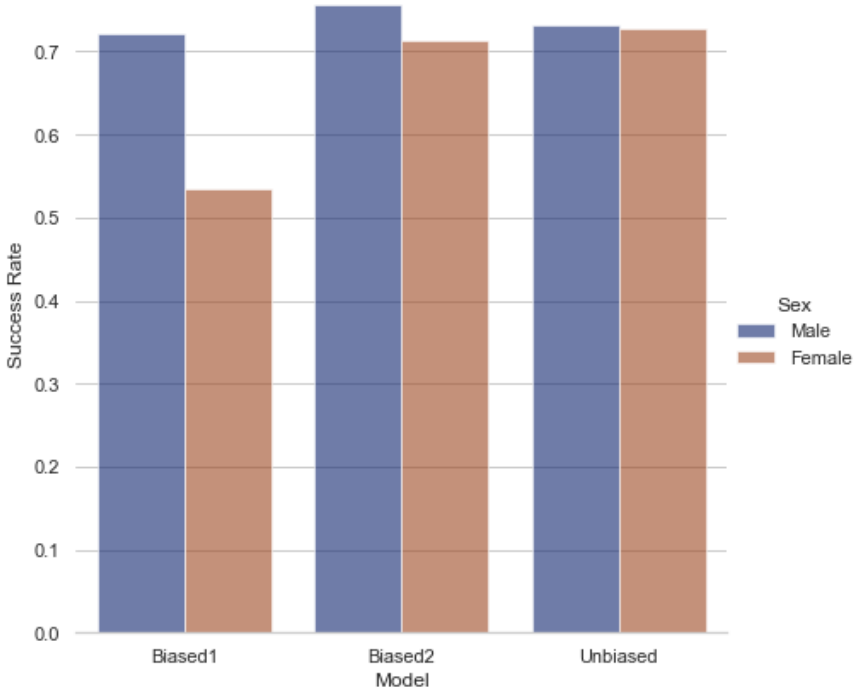

Fig. 1. Success Rate By Model

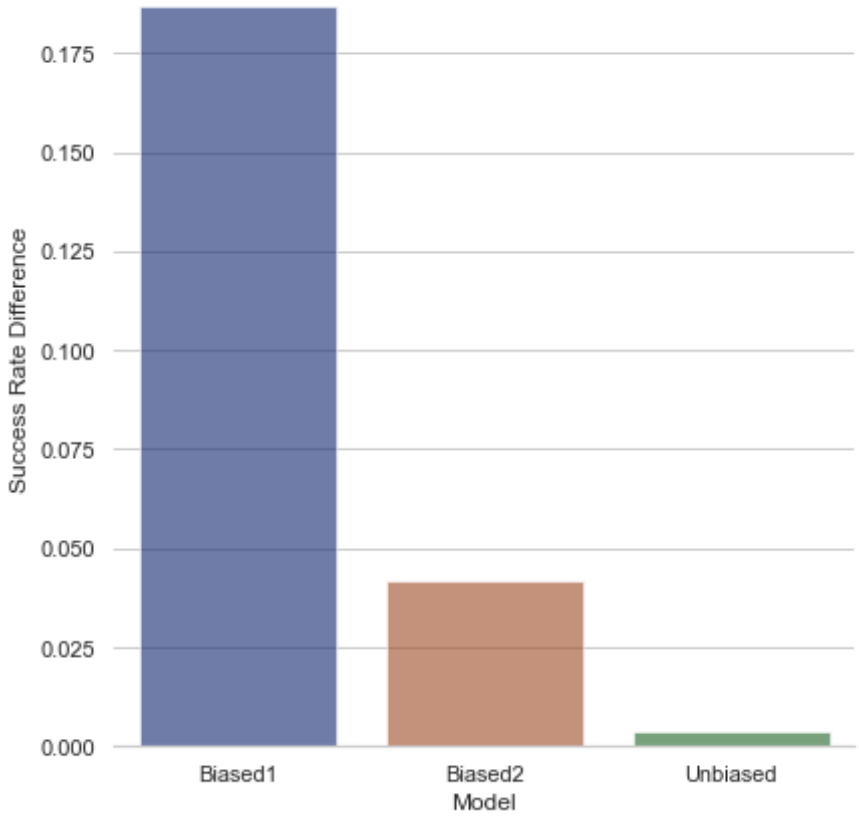

Fig. 2. Difference Between Male and Female Success Rate By Mode