Project Proposal & Progress

skbh77
Student – Computer Science
Durham University
Durham – Country Durham
skbh77@durham.ac.uk

***Abstract*—Analysing a census dataset and how bias can alter the results of a classification model on it. Then applying a pre-processing de-biasing method to remove any disparate impact in the data. Then retraining the model identify any improvements to fairness and the fairness/accuracy trade-off.**

I. Project Proposal

A. Introduction

In today's world, there is an increasing use of machine learning to evaluate the financial situation of people. With Harvard law school already looking into the possible implications of these algorithms [1]. Given that these decisions are likely to have a large impact on people's lives it is crucial these decisions are made based on evidence rather than inbuilt biases towards sensitive attributes such as race or gender. Additionally, for the predictions to be viable to the finance companies they must still be accurate, it is this balance that will be explored in order to identify a model which satisfies both accuracy and fairness. It is because of this the project will work within the context of the financial sector and loan underwriting.

B. Motivation

The motivation for this project is to highlight the impact bias could have on people's lives, and especially minority groups who would be affected significantly by any bias. If a model used commercially to decide eligibility for certain financial products included bias it would have a significant impact. Creating an unbiased model would highlight the differences in how set groups could be treated as well as showing that with the removal of bias the outcomes do change.

C. Plan

The dataset the models will be trained on is information pulled from the US 1994 census [2]. The aim of this project is to create two separate machine learning models. Firstly, a biased predictor that will make predictions without considering bias towards any sensitive attributes. This model will be used to make comparisons against to identify any improvements in bias removal as well as checking for loss of accuracy. Secondly a non-discriminative machine learning model will be created. This model can then be compared to the first and conclusions drawn about its effectiveness. With the conclusions of these models the impact of bias predictors within the financial sector can be evaluated, along with the need to make sure the models fairly evaluate cases.

D. Tasks

The project can be broken down into a few clear, concrete tasks. All of the programming involved in these tasks will be completed in Python 3.9 using Jupyter Notebooks.

1) Problem Definition

For this we will state what the classification problem will involve, and how it will relate to the proposed project.

2) Data Cleaning

This involves transforming the data into a format which is suitable for the models to learn from, which involves removing any empty cells and encoding the data into numerical data rather than categorical so it can be used by the models. This cleaning is done using a module called Pandas, which will allow for column dropping and refactoring.

3) Biased Model Creation

The aim of this task is to create a model which will be biased. This will be done using the 'sklearn' library in python which provides access to multiple classifiers. Once the model is trained it can be tried on the test set in order to determine the base accuracy as well as to identify any bias in the results.

4) Removal of Bias

At this stage the data will be altered to remove any bias in the trained model. The model can be retrained and tested again and compared to the previous model's results to see if there is improvement. This step will again use Pandas in order to manipulate the data easier, but the implementation of de-biasing will be done for scratch.

II. Project Progress

A. Problem Definition

As stated above, the first step is the identify the classification problem to be solved. For this project the problem will be, given information about a person, deciding whether they earn more or less than 50k/year. This is relevant and fits with the problem proposal as that classification could be used to determine eligibility for loans for example, so incorrect or bias classifications have a large impact.

B. Data Cleaning/Encoding

The data is the US 1994 census data [2], can contains 14 different attributes. It comes in two separate files, a training and test set but given we want to alter how we split the data later they will be combined. To make the data useable for the model it must be cleaned and encoded. To begin, any rows containing missing values are removed, in this set the data marked '?' is classified as missing so is removed. Additionally, columns defined to be not useful to the problem are removed, these are.

1. Capital-gain
2. Capital-loss
3. Final Weight
4. Education

Both capital loss and gain contain only a small number of cases that aren't 0 and so will be little use to the model. And final weight doesn't solely relate to one person, representing the number of occurrences of identical data in the census. Education is removed because there already exists a column with that information encoded.

To better analyse two sensitive attributes as one we create a new sensitive attribute called 'Sex_Race' which combines the columns Sex and Race eg 'Male White'
Additionally, multiple columns contain categorical data which cannot be used by a few sklearn classifiers. All categorical models will be encoded using one-hot encoding. These are:

1. Workclass
2. Marital-Status
3. Occupation
4. Relationship
5. Native_Country
6. Result
7. Sex_Race

C. Data Analysis

To better analyse the data, we will split the data by sex_race and analyse the differences in data. The results are visible in the table:

Sex_Race	Results & Averages
Male White	Size: 25667 Age: 39 Workclass: Private Education: 9 (Hs-Grad) Marital Status: Married-civ-spouse Occupation: Craft-repair Relationship Status: Husband Hours/Week: 43 Native Country: United-States Result: <=50k
Male Black	Size: 2050 Age: 38 Workclass: Private Education: 9 (Hs-Grad) Marital Status: Married-civ-spouse Occupation: Other-service Relationship Status: Husband Hours/Week: 40 Native Country: United-States Result: <=50k
Male Asian-Pac-Islander & Male Amer-Indian-Eskimo & Male Other (similar results so grouped)	Size: 833/260/218 Age: 39/36/35 Workclass: Private Education: 13 (Bachelors) Marital Status: Married-civ-spouse Occupation: Prof-Speciality Relationship Status: Husband Hours/Week: 40/43/42 Native Country: United-States Result: <=50k
Female White & Female Black & Female Amer-Indian-Eskimo (Grouped due to similar results)	Size: 11256/1978/158 Age: 37/38/36 Workclass: Private Education: 9 (Hs-Grad) Marital Status: Never-married Occupation: Adm-clerical Relationship Status: Not-in-family Hours/Week: 37/37/38 Native Country: United-States Result: <=50k
Female Other	Size: 122 Age: 32 Workclass: Private Education: 10 (Some-college) Marital Status: Never-married Occupation: Adm-clerical Relationship Status: Not-in-family Hours/Week: 37 Native Country: United-States Result: <=50k
Female Asian-Pac-Islander	Size: 418 Age: 36 Workclass: Private Education: 13 (Bachelors) Marital Status: Never-married Occupation: Adm-clerical Relationship Status: Not-in-family Hours/Week: 38 Native Country: United-States Result: <=50k

The observations to be taken from these results are that although similarities exist between some races in the same sex, some race groups (notably Asian-Pac-Islander) have a higher average level of education across both males and females. Notably, there is also a large difference in the marital and relationship status between the two genders, with men largely being married and husbands, where women are never married and are single.

D. Bias in Dataset

There does appear to be bias in the range of data collected. The data set is largely comprised of white males and females, with other race groups making up a much smaller proportion of the data. Additionally, out of the roughly 7600 entries in the training set provided that have a >50k outcome, 6944 of those are white individuals. This bias, if not addressed, will lead to the model being biased as the model will favour white individuals. This bias in the data set occurs as the natural result of the population distribution in America (where the census was conducted). As the country as a large proportion of white people, it is natural that this has transferred into the census. The lean towards white groups having a larger proportion of >50k results may come because of racial economic disparity in the US.

E. Convential Implementation

1) Description of Algorithm

For the classifier algorithm I will use sklearn's implementation of a 'Random Forest Classifier'. This is because the implementation provided by the module is easy to use and allows for use with another module 'GridSearchCV' to test out and identify the best hyper parameters. The decision was made after testing multiple different classifiers that are present in sklearn and comparing the accuracy scores of them. Additionally, random forest classifier has a low chance of overfitting and will work efficiently with the large data set, with a good degree of accuracy [3].

2) Data Training

The data is naively split using sklearns 'train_test_split' method, where the data is split like so:

- 70% Training Set
- 30% Test Set

Because of the use of GridSearchCV for finding the best hyper-parameters there is no need for a validation set.
To identify the accuracy and fairness with model was trained and run 10 times, the average results were as follows:

- Accuracy: 0.81
- Fairness: 0.9002

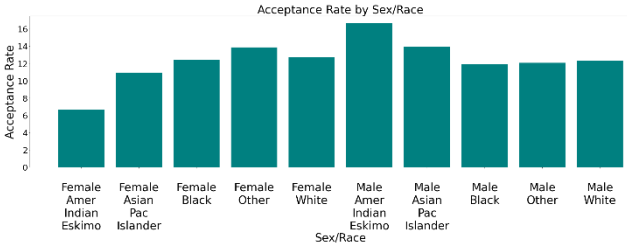

Figure 1 - Bias Model Naive Split

For the second splitting, the data will split in a way that ensure age diversity. For each unique age value in the data set, 70% of that data will be randomly chosen for the training set. This is done using sklearn's 'sample' method, with a fraction of 0.7 specified. The test set is then the other 30% of the data. The new sampling method, again using the sample random tree classifier model, performs like this:

- Accuracy: 0.905
- Fairness: 1.12

The new split leads to a slight increase in accuracy due to the model being trained on a more representative split of the dataset. Of note, the new split also leads to a shift in the bias towards the non-privileged groups. This is possibly because that with a more varied range of ages the percentage of accepting outcomes is greater by ~10% for all sex_race groups, leading to a move in bias.

Figure 2 - Bias Model Fair Split

However, the graph shows that the results lean more towards a larger acceptance for men than women.

The amount of bias in the model appears to be minimal and is more noticeable in the result graph than in the fairness number calculated. Although the results differ every time the model is retrained there seems to be a small bias towards men, with them getting a higher percentage acceptance (>50k) result than women. Any disparity between races differs between different attempts at training the model. The level of bias, although small, is worth resolving because of the nature of the classification. Even a small degree of bias is still a disparate impact on certain groups that would have an impact financially on people.

F. *Fair Implementation*

1) *Description of Algorithm*

To remove bias we will implement an algorithm designed to remove disparate impact [4]. The idea behind the algorithm is to remove any impact the sensitive attribute can have on the model's training. For each column the mean values for each different sex_race unique value is calculated, eg

Mean age where sex_race = 9

As well as the standard deviations. The new mean and standard deviation for that column is chosen to be the median of these values. Then for each individual value, it is moved to the same percentile of the new mean, eg

If the old value was 80% percentile of old distribution, the new value will be 80% percentile of new distribution.

Each column is then repaired, and the sensitive attribute column is removed from the training and testing sets. This way, the model will not bias towards any group as it never directly learns from that data. This method is useful because it allows us to use the same model as before but retrained, which will aid in comparing the effectiveness of algorithm in removing bias as well as identifying any decrease in accuracy.

2) *Analysis of algorithmic bias*

Using the same fairness scoring formula as before, the average fairness over 10 model trains is:

- Fairness: 0.9988

This shows a decrease in algorithmic bias, with a score near to the perfect score of 1. The results are below

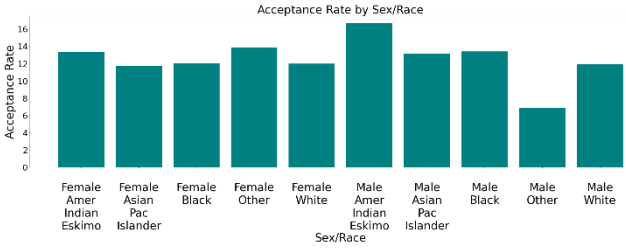

Figure 3 - Non-Bias Model

As seen from this, there is a more even spread among each sex_race group with the exception of 'male other' which could be due to the lack of this demographic in the original data set. This suggests that the sex_race sensitive attribute now plays no significant impact on the outcome of a prediction, which makes sense given the algorithm implemented removes it from the training data.

3) *Comparison to Paper*

The paper reports an initial accuracy of 0.82 and fairness of 0.92 for a decision tree model without any bias removing features, but only for the race attribute. When applying the same algorithm to the data the report shows a new accuracy of 0.83 and fairness 0.89. These results differ from our model for multiple reasons. Firstly, the bias model is trained only on the race attribute, so the unbiased model can only be compared on this attribute, rather than the combined race_sex attribute. Secondly, the paper does not state which classifier is used on the de-biased data, which would significantly affect the accuracy scoring especially. Additionally, from looking at the source code for the module used to de-bias the data, the numerical repairing is done by binning the data and using a categorical repairer [5]. This differs from the implementation used in this paper, which solely uses a numerical repairer. This change means the repaired columns have different values, but the implementation still has the same effect.

4) *Analysis of Accuracy*

A reduction in accuracy is present, the new model's accuracy is *0.691*. This was going to happen as a result of removing the senstive attribute from the test data, as it removes data the model can learn from. The reduction however is minimal and should have no significant impact on the model's ability to predict correct outcomes.

REFERENCES

[1] Yinan Liu and Talia Gillis. n.d., Machine Learning in the Underwriting of Consumer Loans. Available from: <https://projects.iq.harvard.edu/files/financialregulation/files/machine_learning_case_study.pdf>. [April 07, 2021].

[2] Barry Becker From. n.d., UCI Machine Learning Repository: Adult Data Set. Available from: <https://archive.ics.uci.edu/ml/datasets/adult>. [April 07, 2021].

[3] Towards Ai Team. (2021) Why Choose Random Forest and Not Decision Trees – Towards AI — The Best of Tech, Science, and Engineering. Retrieved April 12, 2021, from https://towardsai.net/p/machine-learning/why-choose-random-forest-and-not-decision-trees

[4] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. Proc. 21st ACM KDD (2015), 259–268.

[5] Sorelle (2021) BlackBoxAuditing/BlackBoxAuditing/repairers at master · algofairness/BlackBoxAuditing · GitHub. Retrieved April 13, 2021, from https://github.com/algofairness/BlackBoxAuditing/tree/master/BlackBoxAuditing/repairers