Responsible draft Data Science: Homework 2

Aditya Dixit (ad5157)

Thursday 13^{th} April, 2023

1 Racial disparities in predictive policing

1.1 Part A

- Historical Bias: If racial discrepancies are present in the historical data used to train the machine learning system, the algorithm may unintentionally pick up on and reinforce these biases. For instance, even if drug use is not racially distinct, the algorithm may forecast more drug-related policing activity in Black communities where data indicates there have historically been more drug-related arrests.

- Confounding variables A biased model may result from the machine learning system's use of features that are linked with race. The model might produce predictions that unfairly target particular ethnic groups, for instance, if the system include variables like neighborhood income or educational attainment, which can be associated with race due to historical and systemic causes.

- Algorithmic Bias: The machine learning system's algorithms themselves may induce bias. For instance, some algorithms may give more weight to some features that are disproportionately related to certain racial groups, which can result in predictions that are prejudiced. Furthermore, a weak model may overfit to training data, collecting and exacerbating biases already present.

1.2 Part B

- Fairness-aware Pre-processing: To eliminate any racial discrepancies, one might pre-process the data before feeding it to the machine learning system. To guarantee that each ethnic group is fairly represented, this may entail re-sampling methods or re-weighting situations. In order to lessen the possibility of the model picking up racial biases, one might also eliminate or alter variables that have a strong correlation with race.

- Fairness-aware Learning Fairness limitations could be incorporated right into the machine learning system's learning process. This might entail creating a loss function that penalizes the model for making predictions that are racially unequal while also optimizing the model's accuracy. This would encourage the model to generate predictions that are correct and equitable for all racial groupings.

1.3 Part C

- Ecological fallacy Lum and Isaac's study is based on neighborhood-level aggregated data, which might not adequately reflect individual-level arrest and drug use trends. This could result in an ecological fallacy, in which generalizations formed from data that do not necessarily apply to specific individuals.

- Alternative Explanations There may be additional explanations for the observed racial differences in drug arrests that were not taken into account in Lum Isaac's analysis. For instance, disparities in drug-related policing initiatives may be influenced by other variables such as the frequency of drug sales or drug-related violence, which may not be directly related to the racial makeup of the communities. It is challenging to conclusively link the observed differences to racial bias in policing operations without accounting for these variables.

2 Data science lifecycle

2.1 Part A

The applicant groups identified as "female non-binary" and "other race" may be adversely affected by Alex's imputation method. This is due to the fact that their mean experience is higher than the dataset's average mean experience. Alex would be underestimating the experience of these groups by substituting missing values with the overall mean value, which could result in lower rankings when compared to the "male" and "white" categories.

2.2 Part B

For white male candidates, substitute 5.70 for NULL. Replace NULL with 5.66 for male candidates of other races. Replace NULL with 7.40 for female and non-binary white applicants. For female and non-binary candidates of other races, substitute 7.91 for NULL.

2.3 Part C

Historical societal biases and disparities are referred to as pre-existing bias. This might be the underrepresentation of women, non-binary people, and people of other races in the technology sector, like in the hiring example from MegaSoft. By consistently underestimating the expertise of these underrepresented groups, the imputation method outlined in (a) can further reinforce these prejudices, decreasing their chances of being ranked higher and being employed. Emergent bias is the term used to describe new biases that appear during the creation or use of a machine learning model. As the imputation approach in (a) may not fully account for the differences in experience among demographic groups, favoring those and producing new biases, this technical bias generated by the procedure can result in emergent biases in the model's ranking.

3 Randomized response

3.1 Part A

Let A be the truthful answer and B be the untruthful answer.

P(M(A) = yes) = 0.5 (coin is tails and truth is yes) + 0.5 (coin is heads) P(M(B) = yes) = 0.5 (coin is tails and truth is no) + 0.5 (coin is heads)

P(M(A) = no) = 0.5 (coin is tails and truth is no) P(M(B) = no) = 0.5 (coin is tails and truth is yes)

We can now calculate the ratios:

P(M(A) = yes) / P(M(B) = yes) = (0.5 + 0.5) / (0.5) = 2 P(M(A) = no) / P(M(B) = no) = (0.5) / (0.5 + 0.5) = 0.5

Both ratios are equal to the maximum ratio, which is eˆ, where is the privacy parameter. Thus, = ln(2) 0.693.

3.2 Part B

P(M(A) = yes) = 1/2 (D1 ¡ 4 and truth is yes) + 1/2 (D1 ¿= 4 and D2 ¡ 3) P(M(B) = yes) = 1/2 (D1 ¡ 4 and truth is no) + 1/2 (D1 ¿= 4 and D2 ¡ 3)

P(M(A) = no) = 1/2 (D1 ¡ 4 and truth is no) + 1/2 (D1 ¿= 4 and D2 ¿= 3) P(M(B) = no) = 1/2 (D1 ¡ 4 and truth is yes) + 1/2 (D1 ¿= 4 and D2 ¿= 3)

The ratios are:

P(M(A) = yes) / P(M(B) = yes) = (1/2 + 1/6) / (1/2) = 4/3 P(M(A) = no) / P(M(B) = no) = (1/2 + 1/6) / (1/2) = 4/3

Thus, = ln(4/3) 0.287.

3.3 Part C

P(M(A) = yes) = 1/2 (D1 ¡ 4 and truth is yes) + 1/2 (D1 ¿= 4 and D2 ¡ 3) P(M(B) = yes) = 1/2 (D1 ¡ 4 and truth is no) + 1/2 (D1 ¿= 4 and D2 ¡ 3)

P(M(A) = no) = 1/2 (D1 ¡ 4 and truth is no) + 1/2 (D1 ¿= 4 and D2 ¿= 3) P(M(B) = no) = 1/2 (D1 ¡ 4 and truth is yes) + 1/2 (D1 ¿= 4 and D2 ¿= 3)

The new ratios are:

P(M(A))

4 Classification association rules

4.1 Part A

To find the CARs with support ¿= 3 and confidence ¿= 0.6, let's first analyze the data and calculate support and confidence for each rule:

Total number of tuples (rows) = 32

CAR: sex=F → loan=no Support: (6+3)/32 = 9/32 = 0.28125 Confidence: 9/16 = 0.5625

CAR: sex=F, edu=HS → loan=no Support: 6/32 = 0.1875 Confidence: 6/6 = 1

CAR: sex=F, edu=BS → loan=no Support: 3/32 = 0.09375 Confidence: 3/9 = 0.3333

CAR: sex=F, edu=MS → loan=yes Support: 2/32 = 0.0625 Confidence: 2/4 = 0.5

CAR: sex=M → loan=yes Support: (3+3+4)/32 = 10/32 = 0.3125 Confidence: 10/16 = 0.625

CAR: sex=M, edu=HS → loan=yes Support: 3/32 = 0.09375 Confidence: 3/6 = 0.5

CAR: sex=M, edu=BS → loan=yes Support: 4/32 = 0.125 Confidence: 4/7 = 0.5714

CAR: sex=M, edu=MS → loan=yes Support: 3/32 = 0.09375 Confidence: 3/4 = 0.75

From the list above, we can see that only the following CARs meet the criteria (support = 3 and confidence = 0.6):

CAR: sex=F, edu=HS → loan=no (Support: 0.1875, Confidence: 1) CAR: sex=M → loan=yes (Support: 0.3125, Confidence: 0.625) CAR: sex=M, edu=MS → loan=yes (Support: 0.09375, Confidence: 0.75) To allocate portions of the privacy budget, we'll use sequential composition, as it allows us to divide the privacy budget among multiple analyses. Let's assume the total privacy budget is .

For sequential composition, we can split the total privacy budget into smaller portions (1, 2, 3) for each frequent itemset:

= 1 + 2 + 3

Since the itemsets have different support and confidence values, we can allocate a larger portion of the privacy budget to the itemsets with higher confidence and lower support, as these are more sensitive:

1 for CAR: sex=F, edu=HS → loan=no 2 for CAR: sex=M → loan=yes 3 for CAR: sex=M, edu=MS → loan=yes Considering the support and confidence values, we can allocate the privacy budget as follows:

1 = 0.3 * (302 = 0.4 * (403 = 0.3 * (30This allocation of the privacy budget will provide good utility for the frequent itemsets, as it considers the sensitivity of each itemset based on support and confidence values. Sequential composition allows for controlled allocation of privacy budget, ensuring better data protection while maintaining

4.2 Part B

The itemsets "sex=F, loan=no" and "sex=M, loan=yes" are both often used. We wish to allocate a larger privacy budget to the itemset with more support because it will be more beneficial in terms of utility. However, because both item sets are supported equally, we can divide the privacy budget evenly among them. Sequential composition entails that the privacy budget is split between the inquiries, and the sum of the privacy costs for each query represents the overall privacy cost. The privacy cost is the sum of the privacy costs for all queries in a parallel composition if the queries are run on separate subsets of the data.

In our case, since we have only two itemsets and our overall privacy budget is =1, we can use sequential composition and divide the privacy budget equally between the two itemsets:

(sex=F, loan=no) = 0.5 (sex=M, loan=yes) = 0.5 This allocation satisfies the privacy budget constraint and achieves good utility, as both itemsets are given equal importance.