Homework 3

Yunjeon Lee(yl7143)

May 7, 2022

1 Problem 1

1.1 (a)

(1) Pre-existing bias

If the system was trained on historical employment data and some jobs were predominantly male in the data, then the system will recommend these jobs more to male candidates. Since the system is trained through the data, female candidates might be less exposed to those job postings. This means pre-exising bias toward jobs harmed the female candidates and this is the point for gender disparity in this problem.

(2) Technical bias

If the system is designed to use the attribute "gender" as a determining factor to decide whether this job ad is exposed to the person or not, then this can leads to gender disparity. For example, let's assume that there is a male candidate who is competitive for the caregiver position. This male candidate is kind and has health knowledge which is an important qualification for caregiver job position. However, since this candidate is male, the system predicts that the caregiver job is not suitable for this candidate and does not expose that job posting to this candidate. Even though this male candidate is well-qualified for the position, he can not get this job position because of his gender. This can be the point of gender disparity for this system.

(3) Emergent bias

The system is trained by historical employment data, and if the historical employment data includes many of male candidates are hired to the high-paying jobs, then the system will suggest many of high-paying jobs to male candidates and not recommend those jobs for female candidates. Since most of the high-paying jobs are mostly exposed to male candidates, male candidates will apply more compared to the female candidates. This leads to male candidates are more hired to the high-paying jobs than female candidates. Then, this new data might be used for updating and training the system. Within this updated data, the system would recommend high-paying jobs to male candidates. In other words, the historical data affects the system continuously and this is the point that feedback loop occurs.

1.2 (b)

(i) Under Unary QII condition, this intervention increases the number of times job openings in STEM are shown to African Americans. The Unary QII intervenes 1 feature at a time, and the feature "job experience" is the only feature that is intervened in this problem.

(ii) Under Marginal QII condition, this intervention might fail to increase the number of times job openings in STEM are shown to African Americans. The Marginal QII takes all sets of features and find a pair that influences the prediction the most. Accordingly, under marginal QII, a certain set of features might influence results the most. And this means intervening only the "job experience" feature can fail to increase the number of times job opening in STEM are shown to African Americans, because a specific set of features(for example, job experience and education) is the effective factors set and a set: job experience is not the set that works for increasing the number of times job openings in STEM are shown to African Americans.

2 Problem 2. AI Ethics: Global Perspectives

Lecture: The Intersection of AI and Consumer Protection

The lecturer talks about consumer data protection in the field of digital marketing. In this situation, the stakeholders are consumers, providers(companies), and advertisers.

For the consumers, they can be benefited by personalized advertisements, because advertisements fit well with their tastes which means they can get information about products that they really need and willing to buy. However, the problem is that consumers might not know how their personal data(originally collected for personalized advertisements) is used in other ways. In addition, if the transparency is low, which means consumers do not know which data is collected and how their personal data is used, it is really the severe problem considering consumers' privacy. Furthermore, considering the way how advertisers usually do advertise, some consumers might not be benefited by advertisements. Many advertisers divide consumers into several groups(look-alike groups) considering their characteristics(consumer profiling process), and give advertisements based on these groups. This means some advertisements are only exposed to certain groups. In this case, if there is a person who really need item A but that person is in the group that cannot get the item A advertisements, then this means the consumer cannot access item information(advertisement) even though the consumer really need to buy item A. For this case, providers(companies) are also adversely affected, because companies cannot sell item A to specific consumers although those consumers are willing to buy that item.

For the providers(companies), they can be benefited in that their sales can be increased through personalized advertisements. This is because advertisements are exposed to consumers who are likely to buy the products.

Besides stakeholders, there is an issue regarding data protection. The personal data is collected through cookies, pixels, third-party cookies, and web beacon, and saved in the platform or companies (not consumers' personal devices). This means the way how data is stored and utilized is decided by companies. And it is probable that the collected data is used in other purposes other than personalized advertisements while consumers do not know how their personal data is used.

In order to resolve these issues, related law for restricting data collection and processing should be made and work.(consumer data protection regulation in the EU and the California consumer data act are promising) This law should include requiring consent from consumers, restricting sales of personal data, and providing consumers the right to deny data collection and using for other purposes. In the process of getting consent from customers, consumers can have an opportunity to think about how their personal data is collected, saved, and used. Also, this process can give alerts to consumers that their personal data is exposed to others.

3 Problem 3

3.1 (a)

I used log loss for the SGBClassifier. SGBClassifier uses hinge loss by default, but we need to calculate output probabilities, so I used log loss. (Codes are in the notebook)

3.2 (b)

- Confusion Matrix

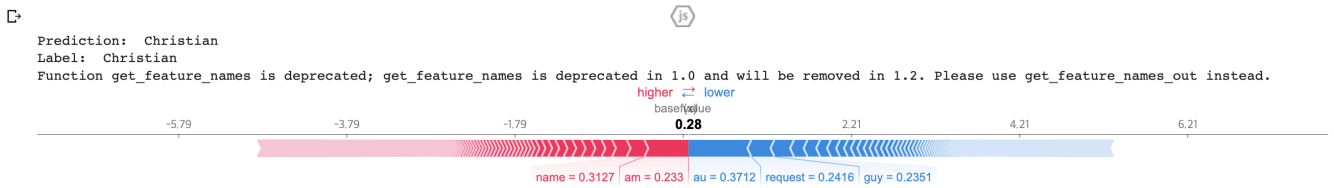
```
                          True Negative:  271
                          False Positive:  48
array([[271,  48],        False Negative:  6
       [  6, 392]])       True Positive:  392
```

correct Christian: 392
correct Atheist: 271
incorrect Christian: 48

incorrect Atheist: 6

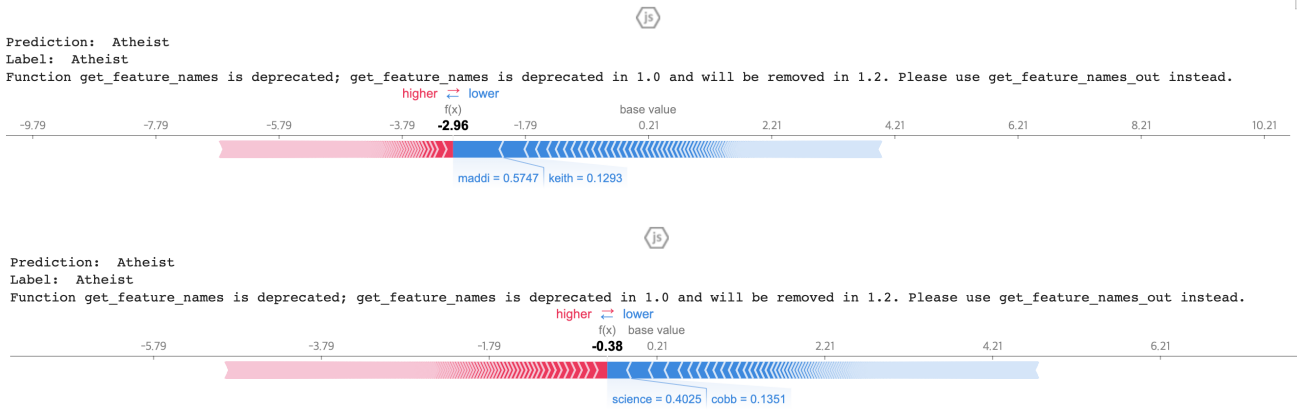
- 5 documents in test set

(1) Prediction: Christian, Label: Christian



Considering that the value is 0.28 which is positive, the model predicted as 'Christian'. Since the label was also 'Christian', the model classified correctly. The words 'name' and 'am' influences the model to classify the data as Christian.

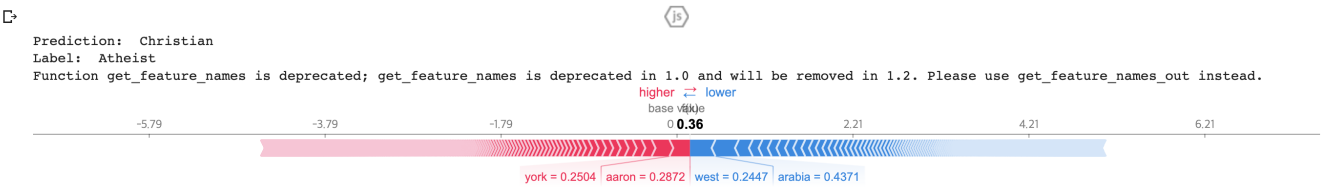
(2) Prediction: Atheist, Label: Atheist



For the first plot, considering that the value is -2.96 which is negative, the model predicted the data as 'Atheist'. Since the label was also 'Atheist', the model classified correctly. The words 'maddi' and 'keith' influences the model to classify the data as Atheist.

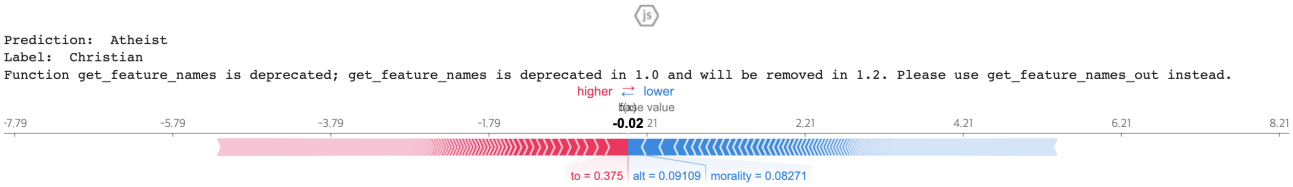
For the second plot, considering that the value is -0.38 which is negative, the model predicted the data as 'Atheist'. Since the label was also 'Atheist', the model classified correctly. The words 'science' and 'cobb' influences the model to classify the data as Atheist.

(3) Prediction: Christian, Label: Atheist



Considering that the value is 0.36 which is positive, the model predicted as 'Christian'. Since the label was 'Atheist', the model misclassified the data. The words 'york' and 'aaron' influenced the model to classify the data as Christian.

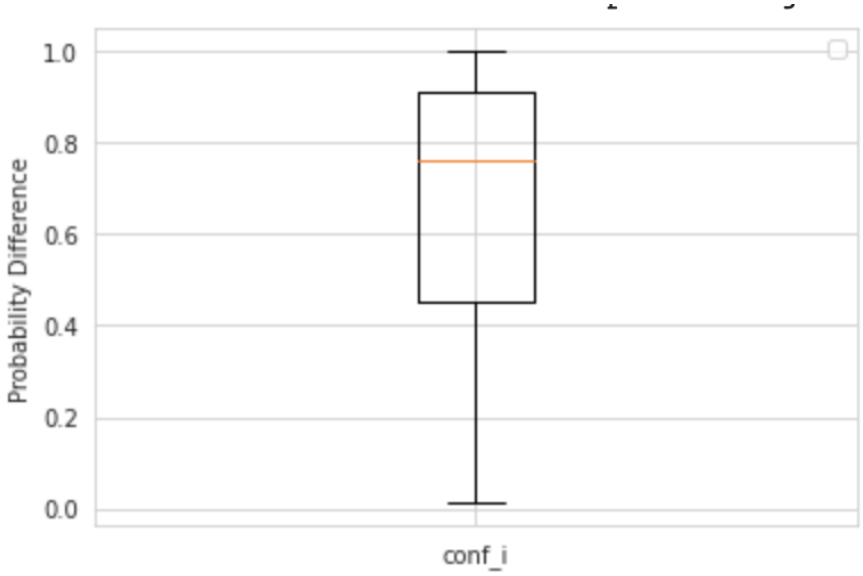
(4) Prediction: Atheist, Label: Christian



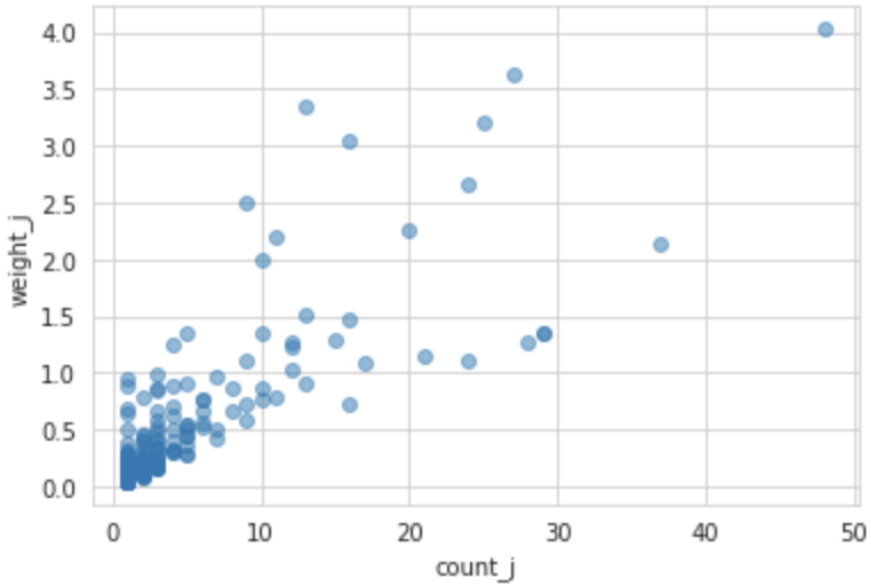
Considering that the value is -0.02 which is negative, the model predicted as 'Atheist'. Since the label was 'Christian', the model misclassified the data. The words 'alt' and 'morality' influenced the model to classify the data as Atheist.

3.3 (c)

- Accuracy: 0.9246861924686193
- the number of misclassified documents: 54
- distribution of errors : conf_i



Median value is around 0.75 and this means the difference between probabilities of two classes are around 0.75. In other words, probabilities of two classes are clear. For the max value, the difference is 1 which means one class has probability of 0 and the other class has the probability of 1, which gives clear probability. For the min value, the difference is 0 which means two classes have the same probability and cannot decide.



Based on the scatter plot, there is proportional relations between counts of misclassified and weights of misclassified. As the number of words count increases, the number of weights also increases. In other words, misclassfied words that has higher frequency has higher weights.

3.4 (d)

- Strategy
1. remove meaningless but frequent words such as am, are, prepositions and so on. (These words are included in stop_words)
2. Adjusting n-gram(which means using consecutive words)
3. Adjusting min_df (minimum document frequency)

Improved Accuracy: 0.9637377963737797

- Example that was misclassified before feature selection and correctly classified after feature selection

Considering that the value is 1.54 which is positive, the model predicted as 'Christian'. Since the label was 'Christian', the new model classified the data correctly. And the words such as 'usa lines', 'mailing list', 'let know' affected this prediction. In this example, since I changed ngram from unigram to bigram and trigram two consecutive words are used. However, if I check the prediction from the previous model, it gives 'Atheist' which is incorrect. Through above 3 strategies some of previously misclassified datas are classified correctly and the model has accuracy over 0.96.