Report of CS340 Assignment 2: Bias Assessment

Yilai Chen 陈驿来 12013025

I. Identify potential bias (2 points)

Feature	Data types	Sample
race	Multi-categorical	Caucasian, African American, Asian, Hispanic, Other, Unknown
gender	Binary	Female, Male, Unknown/Invalid
age	Multi-categorical	30 years or younger, 30-60 years, Over 60 years
discharge_disposition_id	Multi-categorical	Discharged to Home, Other
admission_source_id	Multi-categorical	Referral, Emergency，Other
time_in_hospital	Continuous	numeric
medical_specialty	Multi-categorical	Other, Missing, Family/GeneralPractice, Cardiology
num_lab_procedures	Continuous	numeric
num_procedures	Continuous	numeric
num_medications	Continuous	numeric
primary_diagnosis	Multi-categorical	Diabetes, Musculoskeletal Issues, Respiratory Issues，Genitourinary Issues, Other
number_diagnoses	Continuous	numeric
max_glu_serum	Multi-categorical	None, Norm, >200, >300
A1Cresult	Multi-categorical	None, Norm, >7, >8
insulin	Multi-categorical	No, Up, Steady, Down
change	Binary	No, Ch
diabetesMed	Binary	No, Yes
medicare	Binary	FALSE, TRUE
medicaid	Binary	FALSE, TRUE

Feature	Data types	Sample
had_emergency	Binary	FALSE, TRUE
had_inpatient_days	Binary	FALSE, TRUE
had_outpatient_days	Binary	FALSE, TRUE
readmitted	Multi-categorical	NO, >30, <30
readmit_binary	Binary	0, 1
readmit_30_days	Binary	0, 1

Based on the provided sample sizes for each sensitive feature group:

1. **Race:**

   - Caucasian: 76099
   - African American: 19210
   - Asian: 641
   - Hispanic: 2037
   - Other: 1506
   - Unknown: 2273

   The dataset is imbalanced with respect to race, as there are significantly more samples for Caucasian individuals compared to other racial groups. This could potentially introduce bias if not properly addressed during dataset preprocessing and model training.

2. **Gender:**

   - Female: 54708
   - Male: 47055
   - Unknown: 3

   The dataset is fairly balanced with respect to gender, as the sample sizes for females and males are relatively close. However, the presence of a small number of samples with unknown gender may need to be handled appropriately in the analysis.

3. **Age:**

   - 30 years or younger: 2509
   - 30-60 years: 30716
   - Over 60 years: 68541

   The dataset is imbalanced with respect to age, as there are considerably more samples for individuals over 60 years compared to younger age groups. This imbalance could impact the performance of models trained on this data, especially if age-related factors are significant in the analysis.

II. Train your model to obtain the predicted result (2 points)

1. Explanatory Report on Using Random Forest Model

Reasons for Model Selection:

- **Interpretability**: The Random Forest model, by aggregating multiple decision trees, offers relatively strong interpretability. Each decision tree can be understood as a series of simple rules, making it easier to interpret and explain the model's predictions.
- **Model Expressiveness**: Random Forest is capable of effectively handling large-scale features and complex data relationships, providing high model expressiveness suitable for dealing with multiple features in medical datasets.
- **Training Time**: Compared to complex deep learning models, Random Forest typically has shorter training times, which is advantageous for handling large-scale datasets and frequent retraining needs.

Model Evaluation Results:

- You can find the code in "Random Forest.ipynb".
- On the training dataset, the Random Forest model achieved an accuracy of 100%.
- The model's accuracy on the test dataset is 65%, indicating good generalization ability to unseen data. The prediction results has been recorded as a new column called <readmit_30_days_pred> in "diabetic_preprocessed_test.csv".

III. Quantify fairness with given metrics (16 points)

Our sensitive feature race includes 5 categories (0-'African American',1-'Caucasian',2-'Asian', 3-'Hispanic',4-'Other')

1. Demographic Parity

- African American: Pr(Y'=1 | A=0) = 4217 / 9636 = 0.4376
- Caucasian: Pr(Y'=1 | A=1) = 16664 / 38023 = 0.4383
- Asian: Pr(Y'=1 | A=2) = 112 / 286 = 0.3916
- Hispanic: Pr(Y'=1 | A=3) = 398 / 1004 = 0.3964
- Unknown: Pr(Y'=1 | A=4) = 414 / 1171 = 0.3535

Synthesis to a Ratio	Value
Smallest Difference	0.0007
Largest Difference	0.0848
Smallest ratio	0.8065
Maximum ratio	0.9984

☐ **Description of the Fundamentals of Fairness**

Demographic Parity is a fundamental fairness principle that aims to ensure that predictive models do not discriminate against individuals based on sensitive attributes such as race, gender, or ethnicity. It requires that the distribution of predictions (Y') should be consistent across different sensitive groups (A). In other words, demographic parity promotes fairness by ensuring equal treatment in the prediction outcomes regardless of the sensitive attribute.

☐ **Comparison of Results under Different Measures:**

Smallest Difference: The smallest difference between the probabilities of predicting Y'=1 across different sensitive groups is very small (0.0007), indicating a high level of parity or fairness in the model's predictions.
Largest Difference: The largest difference is 0.0848, which suggests some degree of disparity in the predictions across sensitive groups, but not a substantial difference.
Smallest Ratio: The smallest ratio of probabilities (0.8065) suggesting a potential fairness issue where certain groups may be disproportionately affected by the model's predictions..
Maximum Ratio: The maximum ratio of probabilities (0.9984) implies a more balanced distribution of predictions across sensitive groups, indicating a higher level of fairness where predictions are more consistent regardless of the sensitive attribute.

2. Equalized Odds

- African American: FPR: Pr(Y'=1 | A=0, Y=0) = 3307 / 8593 = 0.3849
  TPR: Pr(Y'=1 | A=0, Y=1) = 910 / 1043 = 0.8725
  Ratio: 0.4412
- Caucasian: FPR: Pr(Y'=1 | A=1, Y=0) = 12968 / 33749 = 0.3843
  TPR: Pr(Y'=1 | A=1, Y=1) = 3696 / 4274 = 0.8648
  Ratio: 0.4444
- Asian: FPR: Pr(Y'=1 | A=2, Y=0) = 80 / 253 = 0.3162
  TPR: Pr(Y'=1 | A=2, Y=1) = 32 / 33 = 0.9697
  Ratio: 0.3261
- Hispanic: FPR: Pr(Y'=1 | A=3, Y=0) = 318 / 910 = 0.3495
  TPR: Pr(Y'=1 | A=3, Y=1) = 80 / 94 = 0.8511
  Ratio: 0.4105
- Unknown: FPR: Pr(Y'=1 | A=4, Y=0) = 337 / 1080 = 0.3120
  TPR: Pr(Y'=1 | A=4, Y=1) = 77 / 91 = 0.8462
  Ratio: 0.3666

Synthesis to a Ratio	Value
Smallest Difference	0.0032
Largest Difference	0.1184
Smallest ratio	0.7338
Maximum ratio	0.9928

☐ **Description of the Fundamentals of Fairness**

Equalized Odds focuses on the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome. Based on our calculations, synthesis are not so prefect. This implies that the model may not achieve equalized odds, as the probabilities are not consistent across all groups.

☐ **Comparison of Results under Different Measures:**

Comparing the results under Equalized Odds with those under Demographic Parity (as previously discussed), we observe differences in how fairness is assessed. While Demographic Parity focuses on the overall likelihood of positive outcomes being the same for different groups, Equalized Odds specifically looks at the accuracy of predictions within each group, considering both true positive rates and false positive rates. This highlights the nuanced nature of fairness assessment and the need to consider multiple metrics for a comprehensive evaluation.

Smallest Difference (0.0032): Possibly due to the large total number of samples, the African American and Caucasian data are very similar, indicating that the model fits very well.

Largest Difference (0.1184): Due to the large difference in the number of samples and some other potential reasons, there is a large gap in the Equalized Odds between the Unknown group and other groups.

Smallest Ratio (0.7338): The Smallest Ratio is moderate, indicating that the model is somewhat suitable.

Maximum Ratio (0.9928): The Maximum Ratio is very high, indicating that the model performs well among a high number of samples.

3. Equalized Opportunity

- African American: TPR: Pr(Y'=1 | A=0, Y=1) = 910 / 1043 = 0.8725
- Caucasian: TPR: Pr(Y'=1 | A=1, Y=1) = 3696 / 4274 = 0.8648
- Asian: TPR: Pr(Y'=1 | A=2, Y=1) = 32 / 33 = 0.9697
- Hispanic: TPR: Pr(Y'=1 | A=3, Y=1) = 80 / 94 = 0.8511
- Unknown: TPR: Pr(Y'=1 | A=4, Y=1) = 77 / 91 = 0.8462

Synthesis to a Ratio	Value
Smallest Difference	0.0049
Largest Difference	0.1235
Smallest ratio	0.8726
Maximum ratio	0.9942

☐ **Description of the Fundamentals of Fairness**

Equalized Opportunity (EO) is a fairness metric that assesses whether a predictive model provides equal chances of positive outcomes for both protected and unprotected groups. In other words, it evaluates whether the true positive rates (TPRs) are equal across different groups. The fundamental principle behind Equalized Opportunity is to ensure that individuals

from all demographic backgrounds have an equal opportunity to benefit from positive outcomes predicted by the model. It is essential for fairness because it prevents discrimination based on sensitive attributes such as race, gender, or ethnicity.

- [ ] **Comparison of Results under Different Measures**

  Smallest Difference (0.0049): The smallest difference represents the smallest variation in TPRs among different groups. In this case, the difference in TPRs between Hispanic and Unknown groups is the smallest, indicating a relatively balanced performance in predicting positive outcomes for these two groups.

  Largest Difference (0.1235): The largest difference indicates the highest disparity in TPRs among groups. In this scenario, the TPR difference between Asian and Unknown groups is the largest, suggesting potential imbalances in the model's predictions for these groups regarding positive outcomes.

  Smallest Ratio (0.8726): The smallest ratio reflects the smallest ratio of TPRs between protected and unprotected groups. A ratio close to 1 indicates a more equitable distribution of positive predictions across different demographic groups. Here, the smallest ratio is between Unknown and Asian groups, indicating a relatively fair distribution of positive outcomes between these groups.

  Maximum Ratio (0.9942): The maximum ratio signifies the highest ratio of TPRs between protected and unprotected groups. A lower ratio implies a more equitable distribution of positive predictions. In this analysis, the maximum ratio is observed between Hispanic and Unknown groups, suggesting a higher level of fairness in positive outcome predictions for these groups.

4. Conditional Statistical Parity

Since it is claimed that L= <age,gender>, we may set L = 1 as <30-60 years, Male>.

- African American: Pr(Y'=1 | L=1, A=0) = 674 / 1767 = 0.3814
- Caucasian: Pr(Y'=1 | L=1, A=1) = 1882 / 5253 = 0.3583
- Asian: Pr(Y'=1 | L=1, A=2) = 12 / 52 = 0.2308
- Hispanic: Pr(Y'=1 | L=1, A=3) = 72 / 220 = 0.3273
- Unknown: Pr(Y'=1 | L=1, A=4) = 33 / 180 = 0.1833

Synthesis to a Ratio <L = 1>	Value
Smallest Difference	0.0231
Largest Difference	0.1981
Smallest ratio	0.4806
Maximum ratio	0.9394

We may set L = 2 as <30-60 years, Female>.

- African American: Pr(Y'=1 | L=2, A=0) = 951 / 2405 = 0.3954
- Caucasian: Pr(Y'=1 | L=2, A=1) = 1809 / 4842 = 0.3736
- Asian: Pr(Y'=1 | L=2, A=2) = 8 / 26 = 0.3077
- Hispanic: Pr(Y'=1 | L=2, A=3) = 69 / 220 = 0.3136

- Unknown: Pr(Y'=1 | L=2, A=4) = 37 / 158 = 0.2342

Synthesis to a Ratio <L = 2>	Value
Smallest Difference	0.0218
Largest Difference	0.1612
Smallest ratio	0.5923
Maximum ratio	0.9449

We may set L = 3 as <30 years or younger, Male>.

- African American: Pr(Y'=1 | L=3, A=0) = 32 / 131 = 0.2443
- Caucasian: Pr(Y'=1 | L=3, A=1) = 71 / 283 = 0.2509
- Asian: Pr(Y'=1 | L=3, A=2) = Nan / Nan = Nan
- Hispanic: Pr(Y'=1 | L=3, A=3) = 8 / 17 = 0.4706
- Unknown: Pr(Y'=1 | L=3, A=4) = 1 / 7 = 0.1429

Synthesis to a Ratio <L = 2>	Value
Smallest Difference	0.0066
Largest Difference	0.3277
Smallest ratio	0.3037
Maximum ratio	0.9737

We may set L = 4 as <30 years or younger, Female>.

- African American: Pr(Y'=1 | L=4, A=0) = 82 / 249 = 0.3293
- Caucasian: Pr(Y'=1 | L=4, A=1) = 140 / 467 = 0.2998
- Asian: Pr(Y'=1 | L=4, A=2) = Nan / Nan = Nan
- Hispanic: Pr(Y'=1 | L=4, A=3) = 2 / 19 = 0.1053
- Unknown: Pr(Y'=1 | L=4, A=4) = 4 / 22 = 0.1818

Synthesis to a Ratio <L = 2>	Value
Smallest Difference	0.0295
Largest Difference	0.2240
Smallest ratio	0.3198
Maximum ratio	0.9104

We may set L = 5 as <Over 60 years, Male>.

- African American: Pr(Y'=1 | L=5, A=0) = 883 / 1850 = 0.4773
- Caucasian: Pr(Y'=1 | L=5, A=1) = 5890 / 12687 = 0.4643

- Asian: Pr(Y'=1 | L=5, A=2) = 48 / 99 = 0.4849
- Hispanic: Pr(Y'=1 | L=5, A=3) = 122 / 232 = 0.5259
- Unknown: Pr(Y'=1 | L=5, A=4) = 185 / 419 = 0.4415

Synthesis to a Ratio <L = 2>	Value
Smallest Difference	0.0076
Largest Difference	0.0844
Smallest ratio	0.8395
Maximum ratio	0.9843

We may set L = 6 as <Over 60 years, Female>.

- African American: Pr(Y'=1 | L=6, A=0) = 1595 / 3234 = 0.4932
- Caucasian: Pr(Y'=1 | L=6, A=1) = 6842 / 14491 = 0.4722
- Asian: Pr(Y'=1 | L=6, A=2) = 44 / 103 = 0.4272
- Hispanic: Pr(Y'=1 | L=6, A=3) = 125 / 296 = 0.4223
- Unknown: Pr(Y'=1 | L=6, A=4) = 154 / 385 = 0.4000

Synthesis to a Ratio <L = 2>	Value
Smallest Difference	0.0049
Largest Difference	0.0932
Smallest ratio	0.8110
Maximum ratio	0.9885

☐ **Description of the Fundamentals of Fairness**

Conditional Statistical Parity, as evaluated using the legitimate factors L = <age, gender>, focuses on ensuring fairness in prediction outcomes across protected and unprotected groups. In this context, legitimate factors other than sensitive attributes are considered to gauge whether individuals from different demographic backgrounds are treated equally in terms of positive outcome assignment.

☐ **Comparison of Results under Different Measures:**

Smallest Difference (0.0231): The smallest difference indicates that there is a relatively small variation in the probability of positive outcome assignment between the protected and unprotected groups based on the chosen legitimate factors. This suggests that, under the defined conditions, the model is relatively fair in terms of conditional statistical parity, as the differences in positive outcome probabilities are not substantial.

Largest Difference (0.1981): The largest difference signifies a notable disparity in the probability of positive outcome assignment between the protected and unprotected groups when considering the specified legitimate factors. This implies that there is room for improvement in ensuring fairness, as there are significant differences in positive outcome probabilities across different demographic groups under the given conditions.

Smallest Ratio (0.4806): The smallest ratio reflects a moderate level of fairness in terms of conditional statistical parity, with the positive outcome probabilities for the protected group being around 48.06% of those for the unprotected group. While this indicates some level of fairness, there are still disparities that need to be addressed to achieve a more equitable distribution of positive outcome probabilities across all groups.

Maximum Ratio (0.9394): The maximum ratio highlights a relatively high level of fairness, with the positive outcome probabilities for the protected group being approximately 93.94% of those for the unprotected group. This suggests that, under the specified legitimate factors, the model performs well in ensuring that individuals from different demographic backgrounds have comparable probabilities of positive outcome assignment.

5. Relationship between Fairness and Training Dataset Inequality:

The evaluation of fairness metrics such as Demographic Parity provides insights into how the model performs in terms of fairness across different sensitive groups. However, it is essential to recognize that fairness in predictions does not necessarily imply fairness in the training dataset. Disparities or biases present in the training dataset can influence the model's predictions and lead to fairness issues. Therefore, while assessing fairness, it's crucial to also address and mitigate biases in the training dataset to enhance overall fairness in the model's predictions.

Besides, it is important to highlight that fairness assessments are dynamic and require continual evaluation and improvement over time. Regular monitoring of fairness metrics, analyzing fairness disparities across groups, and incorporating feedback from impacted communities are essential steps in enhancing fairness in machine learning models. By addressing biases in the training dataset and continually evaluating fairness, organizations can work towards building more equitable and fair AI systems.