Daphne Ozkan

Professor Bhatt

Responsible Data Science

26 February 2025

Homework 1

Problem 1

(a)

- A: Accuracy (ACC)
    - Judges, defendants, and the general public would benefit from accuracy because it ensures that the tool's predictions align with actual re-offending rates. For example, if the model predicts high risk for individuals who are actually at high risk of re-offending, the decision-makers can take appropriate action to protect society.
- B: Positive predictive value (PPV)
    - Victims of crime and the general public benefit from PPV, as it ensures that when a person is flagged as high risk, they indeed pose a significant risk to society, reducing the chances of false negatives (letting dangerous individuals free).
- C: False positive rate (FPR)
    - Defendants, especially those from marginalized communities, would benefit from minimizing the false positive rate, which ensures that innocent individuals are not unfairly punished or incarcerated based on an inaccurate risk prediction.
- D: False negative rate (FNR)

    - Judges and victims of crime benefit from a reduced false negative rate, which ensures that individuals who are likely to reoffend are properly flagged and given appropriate interventions.
- E: Statistical parity (SP)
    - Marginalized groups, particularly racial minorities, would benefit from optimizing for statistical parity, as it ensures that the model's outcomes are balanced across different groups (e.g., black and white defendants). This would help mitigate the racial disparities in the COMPAS tool's predictions.

(b)

- A: Pre-existing bias
    - Example:

      Pre-existing bias can arise if historical data used to train Prophecy reflects societal or organizational biases. For example, if TechCorp's past hiring decisions favored male candidates or individuals from certain universities, Prophecy might learn to prioritize these attributes. This could result in a tool that ranks male candidates or graduates from particular institutions more highly, even if they are not more qualified.
    - Stakeholder harmed:

      Female candidates, minority candidates, or candidates from non-prestigious universities may be harmed by this bias. If the tool reinforces biases that already exist in the company’s historical hiring practices, these groups might be unfairly disadvantaged, limiting their chances of being considered for interviews.

- Intervention:

  To mitigate pre-existing bias, TechCorp should audit the historical hiring data for bias and de-bias the training set. For example, they can remove gender, race, or other potentially biased features from the data used to train Prophecy.

  Alternatively, they could employ fairness constraints during model training to ensure that the predictions do not unfairly disadvantage certain groups.

- B: Technical Bias
  - Example:

    Technical bias may emerge if the features selected by Prophecy are not equally applicable across all candidates or unfairly prioritize certain types of experiences over others. For instance, if the tool uses factors like the number of publications in prestigious journals as a primary criterion, candidates from less traditional backgrounds (such as those from underrepresented groups or those who took non-academic career paths) may be unfairly overlooked.

  - Stakeholder harmed:

    Candidates from non-traditional educational or career backgrounds, including those from underrepresented groups or self-taught candidates, could be harmed by technical bias. This type of bias could lead to an unfair disadvantage for candidates who do not have the conventional credentials but may still be highly qualified for the role.

  - Intervention:

    TechCorp should ensure that Prophecy incorporates a broader set of diverse and relevant features, which represent a variety of qualifications and experiences

beyond traditional educational credentials. They can also ensure that the model does not overly rely on specific characteristics that are more prevalent in one group and less relevant to the actual job performance. Additionally, using fairness-enhancing interventions, such as reweighting features to avoid penalizing candidates from non-traditional backgrounds, can help.

- C: Emergent Bias
    - Example:

      Emergent bias may occur after Prophecy is deployed and begins to rank applicants based on its predictions. Over time, the tool could develop a feedback loop where it favors certain types of candidates that it has already been exposed to in the past. For example, if the tool continuously ranks male applicants higher, the system might learn to perpetuate this trend, even if new candidates who don't fit this pattern are equally qualified.

    - Stakeholder harmed:

      Female candidates, minority candidates, or candidates from less represented backgrounds may suffer from emergent bias. If the tool continually favors certain types of applicants, these groups may be unfairly marginalized, which could hinder their chances of getting an interview or hired.

    - Intervention:

      To prevent emergent bias, Prophecy should undergo regular audits to detect any feedback loops or unintended consequences of the model's predictions. It's essential to continuously retrain the model with updated data to account for evolving candidate profiles and avoid reinforcing outdated patterns. TechCorp

could also use counterfactual fairness techniques, ensuring that the ranking results would not change based on the presence of sensitive features like gender or race.

Problem 2

(a)

The False Positive Rate (FPR) and False Negative Rate (FNR) relate to the confusion matrix as follows:

- FPR = FP / (FP + TN), so FP = FPR * (FP + TN)
- FNR = FN / (FN + TP), so FN = FNR * (FN + TP)

Suppose we have Group A and Group B, with different base rates for the outcome of interest. Let pA = 0.8 be the probability that members of Group A have a positive outcome, and pB = 0.5 be the probability that members of Group B have a positive outcome. Assume group A has 100 observations and group B has 80 observations

For Group A:

We are given that group A has 100 total observations, and 80 of them are positive (based on the base rate pA = 0.8). So, the number of negative observations in Group A is 100 - 80 = 20.

Using the False Negative Rate (FNR):

- FNR = 0.75, so FN / (FN + TP) = 0.75
- This means FN = 0.75 * (FN + TP).

Breaking this down, we have:

- (FN + TP) represents the total number of positive instances in Group A, which we know is 80.
- So, FN = 0.75 * 80 = 60, and thus TP = 80 - 60 = 20.

Now, using the False Positive Rate (FPR):

- FPR = 0.4, so FP / (FP + TN) = 0.4
- The total number of negative observations in Group A is 20, so FP + TN = 20.
- FP = 0.4 * 20 = 8, and thus TN = 20 - 8 = 12.

So for Group A:

- TP = 20
- FN = 60
- FP = 8
- TN = 12

Now calculating these values for Group B:

- The total number of observations for Group B is 80, and 40 of those observations are positive (based on the base rate pB = 0.5)
- So, the number of negative observations in Group B is 80 - 40 = 40

Using the False Negative Rate (FNR):

- FNR = 0.75, so FN / (FN + TP) = 0.75
- This means FN = 0.75 * (FN + TP)

Breaking this down, we have:

- FN + TP is the total number of positive instances in Group B, which is 40.
- So, FN = 0.75 * 40 = 30, and thus TP = 40 - 30 = 10

Now using False Positive Rate (FPR):

- FPR = 0.4, so FP / (FP + TN) = 0.4
- The total number of negative observations in Group B is 40, so FP + TN = 40
- FP = 0.4 * 40 = 16, and thus TN = 40 - 16 = 24

So for Group B:

- TP = 10
- FN = 30
- FP = 16
- TN = 24

(b)

Accuracy (ACC) = (TP + TN) / (P + N)

Calculating accuracy for Group A:

From part (a), we already know the confusion matrix for Group A:

- TP = 20
- FN = 60
- FP = 8

- TN = 12

The total number of positive observations (P) in Group A is:

P = TP + FN = 20 + 60 = 80

The total number of negative observations (N) in Group A is:

N = FP + TN = 8 + 12 = 20

Now, we can calculate the accuracy:

$$ACC_A = \frac{TP + TN}{P + N} = \frac{20 + 12}{80 + 20} = \frac{32}{100} = 0.32$$

So, the accuracy rate for Group A is 32%.

Now calculating the accuracy for Group B:

From the part (a), we know the confusion matrix for Group B is:

- TP = 10
- FN = 30
- FP = 16
- TN = 24

The total number of positive observations (P) in Group B is:

P = TP + FN = 10 + 30 = 40

The total number of negative observations (N) in Group B is:

N = FP + TN = 16 + 24 = 40

Now, we can calculate the accuracy:

$$ACC_B = \frac{TP + TN}{P + N} = \frac{10 + 24}{40 + 40} = \frac{34}{80} = 0.425$$

So, the accuracy rate for Group B is 42.5%

Comparing the accuracy between the two groups, we have:

- Group A accuracy: 32%
- Group B accuracy: 42.5%

Therefore Group B has a better accuracy than Group A.

(c)

Calculating Positive Predictive Value (PPV) for Group A:

Positive predictive value (PPV) = TP / PP

From the parts (a) and (b), we know that the confusion matrix for Group A is:

- TP = 20
- FP = 8

So, the total Predicted Positives (PP) for Group A is:

PP = TP + FP = 20 + 8 = 28

Now, we can calculate the PPV for Group A:

$$PPV_A = \frac{TP}{PP} = \frac{20}{28} = 0.714285714$$

So, the PPV for Group A is approximately 71.43%.

Now, calculating the PPV for Group B:

So, the total Predicted Positives (PP) for Group B is:

PP = TP + FP = 10 + 16 = 26

Now, we can calculate the PPV for Group B:

$$PPV_B = \frac{TP}{PP} = \frac{10}{26} = 0.384615385$$

So, the PPV for Group B is approximately 38.46%.

Comparing the predicted positive values for both groups, Group A has a better PPV (71.43%) than Group B (38.46%).

(d)

The Chouldechova paper explains that fairness impossibility results occur when there are different base rates of positive outcomes across groups. Specifically, the paper shows how no classifier can simultaneously achieve the following for two groups with different base rates: (i) Equal Positive Predictive Value (PPV), (ii) Equal False Positive Rates (FPR) and (iii)Equal False Negative Rates (FNR).

In the case of this problem, we are given that Group A has a base rate of 0.8 (80% of the group has a positive outcome), and Group B has a base rate of 0.5 (50% of the group has a positive

outcome). Since the two groups have different base rates for positive outcomes, it's impossible to achieve equal values for PPV, FPR, and FNR across both groups.

This is impossible because:

(1) Positive Predictive Value (PPV) depends on the number of True Positives (TP) and the total number of Predicted Positives (PP) (which is TP + FP). Since the base rates for each group are different, the distribution of True Positives and False Positives will vary between the two groups. This makes it impossible to achieve the same PPV for both groups.

(2) The False Positive Rate (FPR) is the proportion of Negative observations that are incorrectly predicted as positive (FP / (FP + TN)). With different base rates, the total number of Negative instances (which is N = FP + TN) differs between the two groups. Thus, achieving equal FPR is not possible, since the number of False Positives and True Negatives is different in each group.

(3) The False Negative Rate (FNR) is the proportion of Positive observations that are incorrectly predicted as negative (FN / (FN + TP)). Again, with different base rates, the number of Positive instances in each group is different, so the FNR for each group will not be equal, making it impossible to achieve the same FNR across both groups.

Therefore this example shows the fairness impossibility result discussed in Chouldechova's paper the two groups' different base rate (0.8 for Group A and 0.5 for Group B) make it impossible to achieve equal PPV, FPR, and FNR for both groups simultaneously.

Problem 3: Memo on “AI for Whom?” lecture

In the lecture “AI for Whom?”, Danya Glabau addresses critical issues surrounding artificial intelligence (AI) in recruiting and healthcare. By analyzing real-world examples, Glabau highlights how AI systems, despite their promises of fairness and efficiency, often fail to serve marginalized groups. Her discussion explores two main applications of AI: AI in recruiting and AI in healthcare. In recruiting, AI systems are used to assess job candidates by automating resume screening and interview analysis. The promise of these systems is to eliminate human biases and improve the efficiency of hiring processes. However, these systems often perpetuate gender and racial biases that already exist in hiring practices. One example of this is Amazon’s recruitment tool, which, trained on resumes from predominantly male applicants, ended up penalizing female candidates, especially those from women’s colleges. In healthcare, Glabau explains how AI is used to automate processes such as benefits allocation, health risk assessment, and appointment scheduling. While AI is supposed to reduce bias and improve efficiency, it can inadvertently reinforce existing disparities. For instance, AI tools designed to predict health risks based on cost data tend to disadvantage Black patients by underestimating their health needs. Similarly, automated scheduling systems at healthcare facilities often overbook Black patients, assuming they are less likely to show up for appointments.

The primary beneficiaries of AI in recruiting, as highlighted by Glabau, are companies and organizations that save time and money by automating hiring processes. Ideally, these systems would help identify the most qualified candidates. However, in reality, women, Black people, and other underrepresented groups are harmed. In Amazon's case, women were systematically overlooked due to the tool's preference for male-dominated language in resumes. In healthcare, the beneficiaries are healthcare providers who gain efficiency from AI-driven

systems that reduce fraud, automate scheduling, and improve resource allocation. Unfortunately, Black patients, low-income individuals, and chronically ill patients are most negatively impacted. These groups face discrimination and unequal access to care, as AI systems do not account for systemic inequalities and often prioritize efficiency over equity.

The AI systems discussed in the lecture exhibit disparate impact and disparate treatment. In recruitment, AI tools that learned from biased historical data result in the disparate treatment of women and minorities. Amazon's recruitment system penalized resumes from women's colleges and gave preference to male applicants, reinforcing sexist biases. In healthcare, disparate impact is evident when AI systems allocate fewer resources or care to Black patients, despite similar health risk scores as their white counterparts. This occurs because the systems were trained on biased data that correlates race with lower healthcare spending. Furthermore, these systems are an example of emergent bias: as they evolve, they perpetuate existing societal inequalities. For instance, the AI recruiting tool used by Amazon and other companies failed to recognize the value of diversity in the workforce, missing out on talented women and minorities. Ultimately, while AI has the potential to improve efficiency, it is crucial that these systems are designed and implemented with fairness in mind to avoid exacerbating existing inequalities.

Problem 4: Report for Google Colab Code Results

Part (a): Baseline Random Forest Model

In this section, we trained a baseline Random Forest model on the "Diabetes Hospital" dataset. The goal was to evaluate the model's performance on various metrics and assess any disparities across different groups, using gender as the sensitive attribute. The data was mostly preprocessed for us in the given code by one-hot encoding categorical features and handling missing values. I

also removed the gender-related columns after one-hot encoding to prevent the model from using this information directly. After that, I split the data into training and test sets, with 80% of the data used for training and 20% for testing, as specified in the instructions. The baseline Random Forest model was initialized with a default of one estimator (n_estimators =1) and was trained on the preprocessed data. The model's predictions were then evaluated on the test set using multiple performance metrics, including accuracy, precision, recall, false negative rate (FNR), and false positive rate (FPR). The model's performance was evaluated using Fairlearn's MetricFrame, which allowed for the calculation of fairness metrics across different groups. Below are the overall metrics and performance metrics by gender:

Overall Performance Metrics:

- Accuracy: 0.531434
- Precision: 0.488959
- Recall: 0.498392
- FNR: 0.501608
- FPR: 0.440617

Performance Metrics By Group:

- For females (gender = 0):
    - Accuracy: 0.526198
    - Precision: 0.460000
    - Recall: 0.468193
    - FNR: 0.531807
    - FPR: 0.428571

- For males (gender = 1):
    - Accuracy: 0.535558
    - Precision: 0.509982
    - Recall: 0.520370
    - FNR: 0.479630
    - FPR: 0.450751

Analysis of Results:

- Accuracy:
    - The overall accuracy of the model is 53.1%, which indicates that the model performs moderately well at predicting hospital readmission, but there is significant room for improvement. When broken down by gender, the accuracy for females is 52.6%, and for males it is 53.6%. While there is a slight difference in accuracy between genders, it is not substantial. This suggests that, in terms of pure prediction accuracy, the model does not show a significant bias favoring either gender.
- Precision:
    - Precision measures the proportion of positive predictions that are actually correct. The model shows a precision of 48.9% overall, with males showing slightly better precision (50.9%) than females (46.0%). This could indicate that, for female patients, the model is more likely to make false positive predictions, which might lead to over-treatment or unnecessary interventions.
- Recall:

    - Recall measures the proportion of actual positive cases that are correctly identified by the model. The overall recall of 49.8% shows that the model is somewhat effective at identifying patients who will be readmitted. However, again, there is a slight gender-biased discrepancy, with males having a higher recall (52%) compared to females (46.8%). This suggests that the model is slightly better at identifying male patients at risk of readmission but may be missing a higher proportion of female patients who also face the risk.

- False Negative Rate (FNR):
    - The FNR rate is higher for females (53.2%) compared to males (48.0%), which indicates that the model is more likely to incorrectly predict that female patients will not be readmitted, even though they actually will. This could be a significant issue in a healthcare setting, where failing to identify at-risk patients could lead to inadequate care and potentially harm patients.
- False Positive Rate (FPR):
    - The FPR for females is 42.9%, while it is 45.1% for males. While the model performs better with respect to FPR for females, both genders exhibit relatively high false positive rates. This means that the model is prone to incorrectly predicting that patients will be readmitted, leading to unnecessary interventions, moreso for male patients.

Part (b) : Hyperparameter Tuning: Impact on Performance

In this part, I tuned the hyperparameters n_estimators and max_depth of the Random Forest model to evaluate their impact on performance. We used n_estimators = 1000 and max_depth=10, which were chosen based on their ability to maximize accuracy in experiments.

After training the Random Forest model with these hyperparameters, I evaluated its performance using several metrics, including accuracy, precision, recall, false negative rate (FNR), and false positive rate (FPR). My analysis of the results are below:

- Accuracy: The tuned Random Forest achieved an accuracy of 0.622790, which was an improvement over the baseline model (with n_estimators = 1) that achieved an accuracy of 0.526198. This demonstrates the positive effect of increasing the number of estimators and adjusting the model depth to capture more complex patterns in the data.
- Precision: The precision for the tuned model was 0.611940, indicating a moderate balance between correctly predicted positives and false positives. In comparison, the baseline model had a precision of 0.460000, reflecting a significant improvement with the tuned hyperparameters.
- Recall: The recall for the tuned model was 0.483837, showing that the model was better at correctly identifying positive cases compared to the baseline (0.468193).
- FNR: The false negative rate decreased from 0.531807 in the baseline model to 0.476930 with the tuned model, indicating that the tuned model reduces the likelihood of missing positive cases.
- FPR: The false positive rate also improved, with the tuned model achieving a rate of 0.252893 compared to the baseline's 0.428571, which suggests that the model's sensitivity to false positives was reduced.

These results suggest that tuning the Random Forest model's hyperparameters, particularly increasing the number of estimators and adjusting the tree depth, significantly improved the model's accuracy and ability to correctly classify both positive and negative cases. However,

while the model showed improvement in overall metrics, it is important to consider fairness metrics as well, especially in sensitive areas such as healthcare.

Part (c): Impact of Alpha on Fairness and Accuracy

In this part of the code, we experimented with the Adversarial Fairness Classifier (AFC) from Fairlearn to investigate the impact of the alpha parameter on fairness and accuracy. The alpha parameter in the AFC controls the tradeoff between fairness and accuracy, with lower values prioritizing accuracy, and higher values focusing more on fairness. We tested four different values of alpha: 0.0, 0.3, 0.7, and 1.0, and used 10 different random seeds for retraining the classifier.

We first computed performance metrics for each combination of alpha and random seed, focusing on accuracy, precision, recall, false negative rate (FNR), and false positive rate (FPR). Boxplots were generated to visualize the variation of these metrics across different values of alpha. The results indicated that as alpha increased, the tradeoff shifted towards fairness, typically at the expense of accuracy.

The results were as follows:

- For accuracy, we observed that alpha=0.0 (which emphasizes accuracy) led to the highest performance, while values of alpha=1.0 (emphasizing fairness) resulted in lower accuracy. This aligns with the tradeoff where increasing fairness can reduce model performance.

- Precision and recall displayed similar patterns, with higher alpha values leading to increased fairness, but at a slight cost to precision and recall. The model with alpha=0.0 achieved the best precision and recall.
- FNR and FPR showed more variation across the alpha values. As expected, higher alpha values (especially alpha=1.0) tended to reduce both FNR and FPR, demonstrating a better balance between false positives and false negatives.

Additionally, we compared these results with the performance metrics of the baseline random forest model (from part b) and the tuned random forest model (using n_estimators = 1000 and max_depth = 10). The baseline and tuned random forest models had higher accuracy but worse fairness metrics compared to the Adversarial Fairness Classifier with higher alpha values.

In conclusion, the choice of alpha in the AFC involves a tradeoff between accuracy and fairness. A higher alpha (favoring fairness) resulted in lower accuracy but improved fairness metrics, while a lower alpha achieved higher accuracy but compromised fairness. These findings highlight the importance of selecting the right alpha value based on the specific needs of the application, balancing fairness and performance as necessary.

Part (d): Results from ThresholdOptimizer and Adversarial Fairness Classifier

In this part of the experiment, I utilized the ThresholdOptimizer post-processing model with the equalized_odds constraint to mitigate the unfairness observed in the baseline random forest model. My choice of the equalized_odds constraint was based on the importance of balancing both false positive rate (FPR) and false negative rate (FNR) across sensitive groups, which, in

this case, was gender. In healthcare, for example, ensuring that both FPR and FNR are balanced across different demographic groups is critical, as both false positives (e.g., misdiagnosing a healthy individual) and false negatives (e.g., failing to identify a high-risk patient) can lead to serious consequences. The equalized_odds constraint ensures that both error rates are similar across the groups, thus aiming for fairness while not sacrificing accuracy completely. I evaluated the performance of the ThresholdOptimizer across 10 random splits and compared its results to the un-tuned (baseline) random forest and the tuned random forest model, focusing on key fairness and performance metrics such as accuracy, precision, recall, FNR, and FPR.

From the boxplots that were generated, I observed that ThresholdOptimizer led to a reduction in FNR, especially for alpha values 0.3, 0.7, and 1.0. However, there was a trade-off with FPR, which increased for some of these alpha values. This shows that as the ThresholdOptimizer worked to reduce FNR by equalizing false negative rates, it did so at the cost of increasing false positive rates in some cases. For the baseline random forest, the FPR and FNR were both significantly higher compared to the ThresholdOptimizer results, especially for the alpha values 0.0 and 0.3, where ThresholdOptimizer performed better in minimizing disparity.

Looking at accuracy and precision, the performance of ThresholdOptimizer remained consistent across different alpha values, with only slight variations. For the tuned random forest model, accuracy was notably improved, but its FPR and FNR were not as balanced as with ThresholdOptimizer, reflecting the trade-off between fairness and accuracy. This was consistent across the boxplots and line plots presented. When comparing ThresholdOptimizer to the Adversarial Fairness Classifier, it appears that while both models attempted to address fairness issues, ThresholdOptimizer was more effective at minimizing FNR at the cost of a slightly

higher FPR, whereas the adversarial fairness classifier resulted in lower FPR but higher FNR in some cases.