Investigating Bias in Insurance Premium Prediction

Chuan Chen
cc6580@nyu.edu
Center for Data Science
New York University
New York, NY 10012

Yihang Zhang
yz2865@nyu.edu
Center for Data Science
New York University
New York, NY 10012

1 Background

As the importance of data collection and interpretation has been proven in the past decade, more and more companies, organizations, and government branches have invented and deployed the ADS(Automated Data System) to assist them to make decisions under various scenarios. However, the discussion of these ADS applications also arises, arguing such a technique is constantly bringing unfairness into production which exerts a negative impact on society from various angles.

The ADS focused in this project aims to use customers past medical expenses and their features such as sex, age, BMI, and etc. to predict their future expenses, ultimately to help insurance companies make decisions on premium charges. The purpose of this project is to investigate the methodology and implementation of the ADS and if there exists bias where the system favors one group over another unfairly, such as charging a particular group with the same feature values over another group.

2 Input and Output

This data is obtained from the Machine Learning course website (Spring 2017) from Professor Eric Suess at California State University. It has 1338 samples, containing 6 features, age, sex, BMI, number of children, smoker, region, and 1 target variable, medical expenses, as the output. Some of the data generalizations and visualizations are shown below.

variable	datatype	missing values
age	integer	0
sex	binary	0
bmi	float	0
children	integer	0
smoker	binary	0
region	category	0
expenses	float	0

The distribution of numerical features can be seen in figure 1. The distribution of numerical features can be seen in figure 2. The distribution of categorical features can be seen in figure 3. The pairwise correlation between each feature can shown in figure 4.

The relationship between each feature and the target variable can also investigated and can be seen in figure 5 and 6.

The output of this ADS is called `expenses`, which is the medical expenditure a particular client has spent in the past. It is a continuous and numerical feature, ranging from 1121 to 63770. According to these plots, it is indicated that expenses are correlated with features like age, BMI, sex, etc. Therefore, to interpret the relationships better, we convert the target feature `expenses` into a binary feature for further exploration, which would be introduced in more detail in a later section.

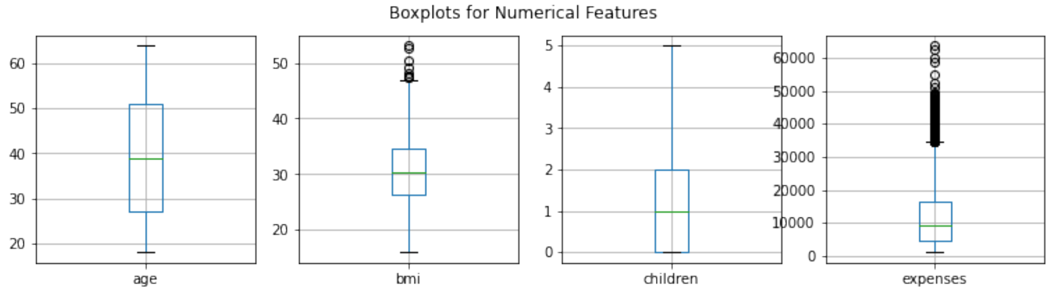


Figure 1: Boxplots for Numerical Features

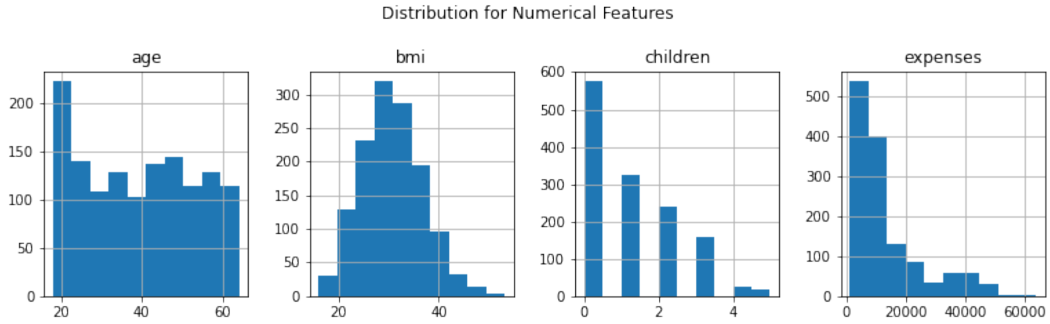


Figure 2: Distribution of Numerical Features

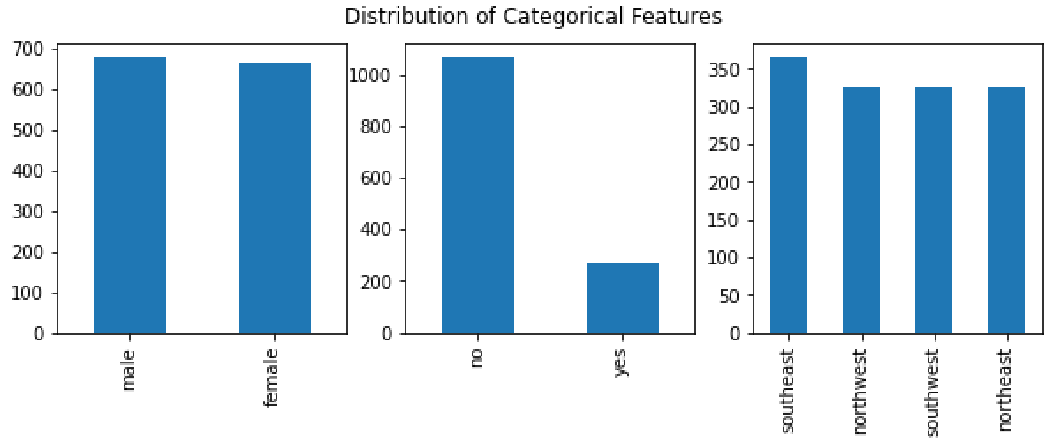


Figure 3: Distribution for Categorical Features

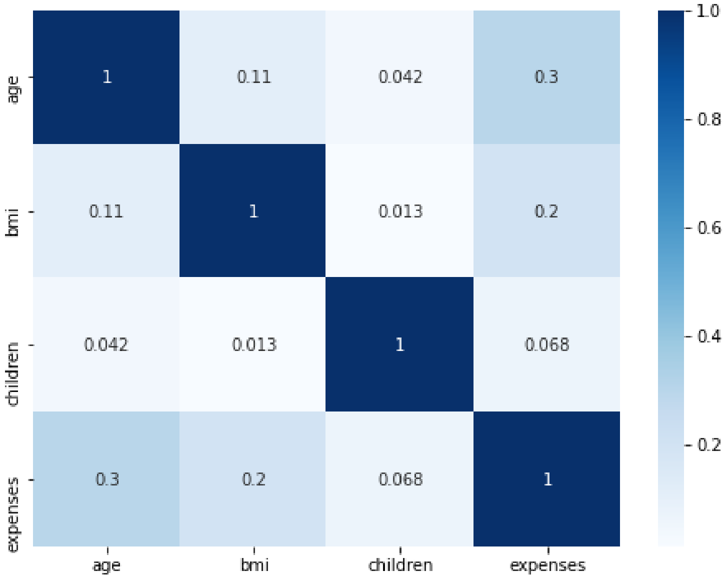


Figure 4: Pairwise Correlation Between Each Feature

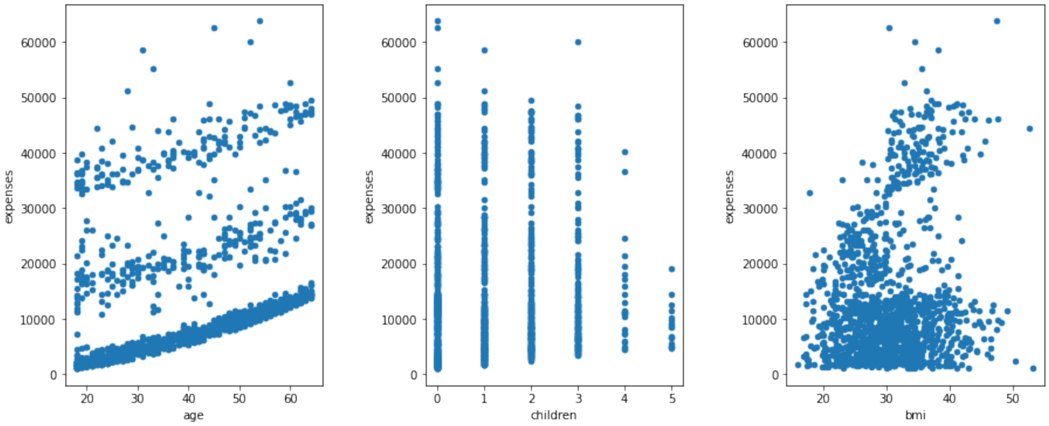


Figure 5: Correlation Between Numerical Features and the Target Variable

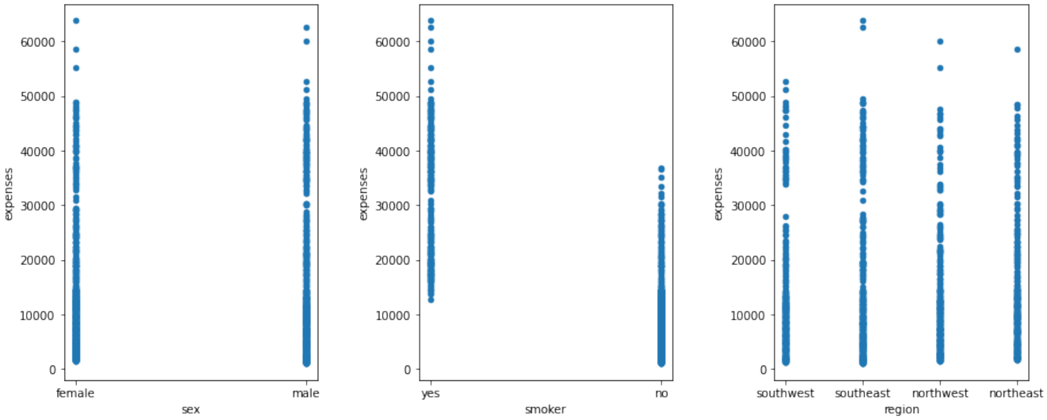


Figure 6: Correlation Between Categorical Features and the Target Variable

3 Implementation and Validation

This data has already been cleaned beforehand, so no additional data cleaning was performed. The designer of the ADS system removed the feature region during pre-processing. Since region did not have exhibit any missing or outlier data, it was unclear why the creator did so. But in order to reproduce the same ADS system, we removed this feature as well. To prepare the data for modeling, the categorical features were encoded as integers, including sex and smoker. Then, the data was split 75/25 into the train and test sets, and each feature was standardized to have a mean 0 and standard deviation of 1.

The ADS has implemented several different regressors, such as linear regression, polynomial regression, support vectors, decision tree, and random forest to predict the insurance premium of each individual in the dataset. Among these machine learning algorithms, the random forest was proved to be the one with the highest accuracy of 0.8969 on the test set and an RMSE of 4028.435 (lowest out of all the regressors). Thus, it was decided that the ADS should be implemented with random forest and is capable of predicting one's insurance premium with acceptable accuracy.

4 Outcomes

Although the ADS achieved relatively good accuracy, no other error metric was used by the designer to determine model performance and make model selections. Thus, we will investigate further into the model performance with various performance, fairness, and robustness metrics to help determine whether this ADS is fit to be deployed in the industry.

To help with investigation, we also added an additional "target" variable called binary_expense. The intuition for this variable stems from the decision part of the ADS. When stakeholders such as insurance companies are handed the predicted medical expenses of their potential clients, they will unavoidably make some decision on whether the expenses are considered high or low. Upon inspecting our full data, the mean of all recorded expenses is around \$13,270, and the median is around \$9,382. This suggests that we have some outliers with very high medical expenses present. Thus, we chose somewhat of a mid-point and categorized expenses that are greater than \$10,000 as {binary_expense:1}, which signifies high medical expenses and is also our unfavorable outcome (from the clients perspective), and expenses that are less than or equal to \$10,000 as {binary_expense:0}, which signifies low medical expenses and is also our unfavorable outcome (from the clients perspective). After training and predicting using the model, we also reintroduced the feature region into our dataset to investigate its relationship between the two target variables.

4.1 Sub-Populations Across Features

Before we can investigate the performance of the ADS across sub-populations, we must first define what they are. We examined the binary expense outcomes across different values for each of our features to determine which feature values are considered as more privileged (having low medical expenses/insurance premiums) than others.

Figure 7 shows the distribution of binary outcomes across all ages for both the training set (with true labels) and the test set (with predicted labels). We can see that around the average age of 39 there starts to exhibit a separation of high and low expense trends. Thus, we chose 39 as the threshold that separates our privileged and unprivileged age groups. A similar analysis was done for the other two numerical features BMI and children. Their analysis can be found in figure 16 and figure 17 in appendix A.

For categorical variables such as smoker, we plotted the distributions of binary outcomes and the box plot of numerical outcomes across groups. They are shown in figure 8 and 9. We can see that {smoker:yes} definitely have high expenses while {smoker:no} does not. Thus, we chose {smoker:yes} as our unprivileged group. Similar analysis was done for the other two categorical features sex and region. Their analysis can be found in appendix A.

In summary, our sub-population per feature are divided as shown in table 1:

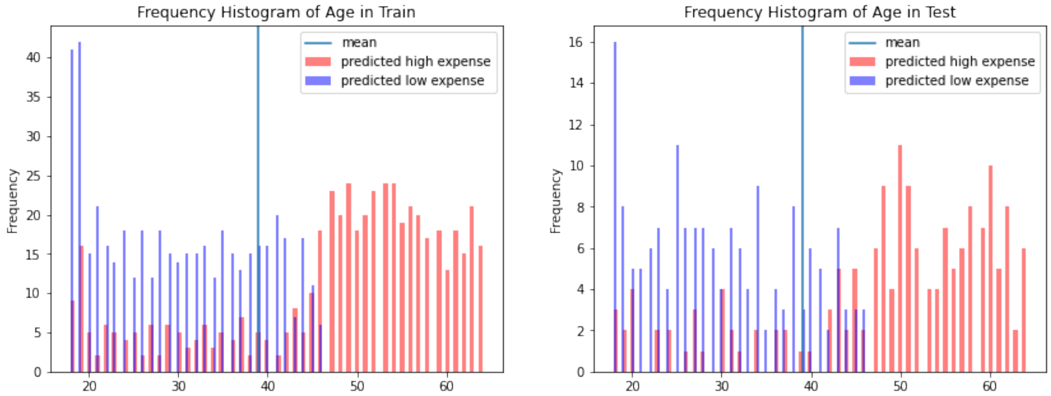


Figure 7: Distribution of Binary Outcome Across Age

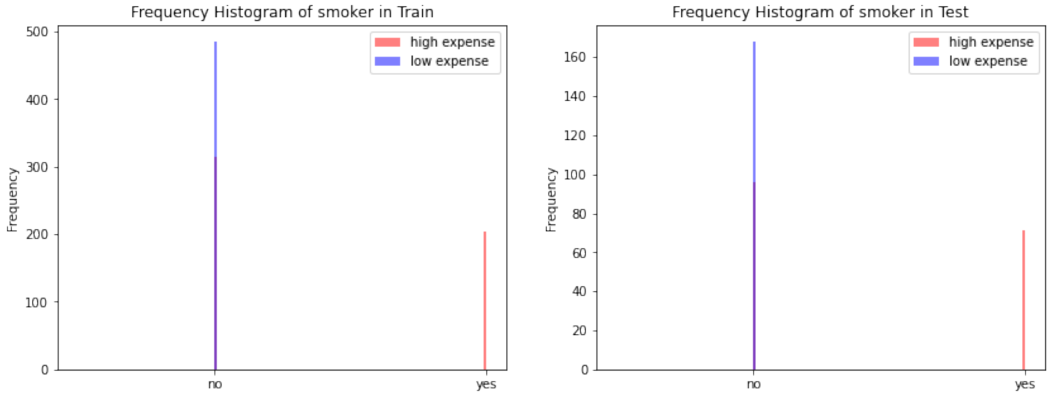


Figure 8: Distribution of Binary Outcome Across Smoker

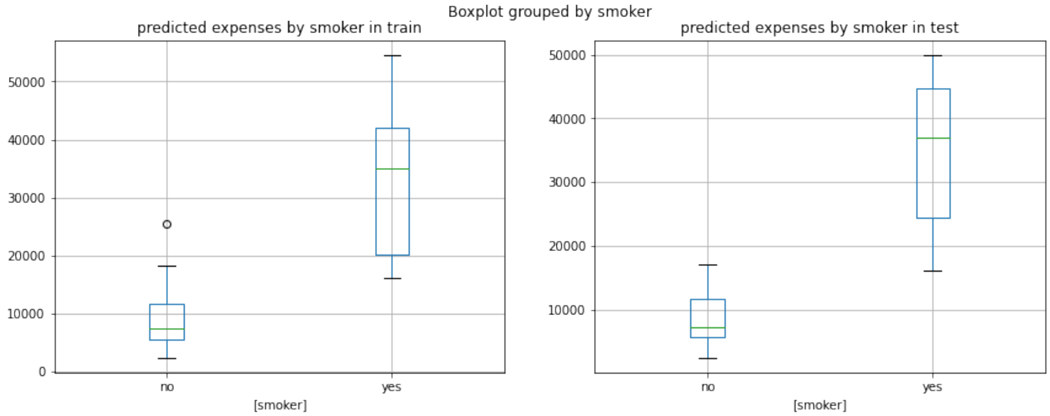


Figure 9: Box Plot of Numerical Outcome Across Age

feature	privileged (low expense)	unprivileged (high expense)
age	≤ 39	> 39
BMI	≤ 30	> 30
children	> 1	≤ 1
sex	male	female
smoker	no	yes
region	southwest	southeast, northwest, northeast

Table 1: Sub-Population Across Features

4.2 Accuracy Across Sub-Populations

Both numerical accuracy (r2_score computed using actual expenses and predicted_expenses) and binary accuracy (accuracy_score computed using actual binary_expenses and predicted_binary_expenses) used as accuracy measures. The overall accuracy of the dataset are shown in table 2. Interestingly, both accuracy measures of the test set are very similar, and even slightly better, than of the train set. This suggests that our model did not over-fit.

set	regression accuracy	binary accuracy
train	0.878	0.874
test	0.897	0.872

Table 2: Accuracy measures of the entire population

The same accuracy measures were computed for sub-populations across each feature. Their in-detailed reports can be seen in table 3 through 8 in appendix A. To better assess whether there is a difference in accuracy between privileged and unprivileged groups, the difference between their accuracy measures was computed and plotted in figure 10. A dot below the dotted line signifies that the unprivileged group has less accuracy than the privileged group. While a dot above the dotted line suggests otherwise. From the figure, we can see that 3 features deviate more from the dotted line than others, age, BMI, children. More specifically, we achieve less accuracy for the unprivileged age group (> 39), more accuracy for the unprivileged BMI group (> 30), and more accuracy for the unprivileged smoker group ({smoker: yes}). This suggests that unprivileged groups for bmi and smoker might exhibit more distinctive markers that lead to a more accurate prediction, while unprivileged groups for age exhibit less distinctive markers.

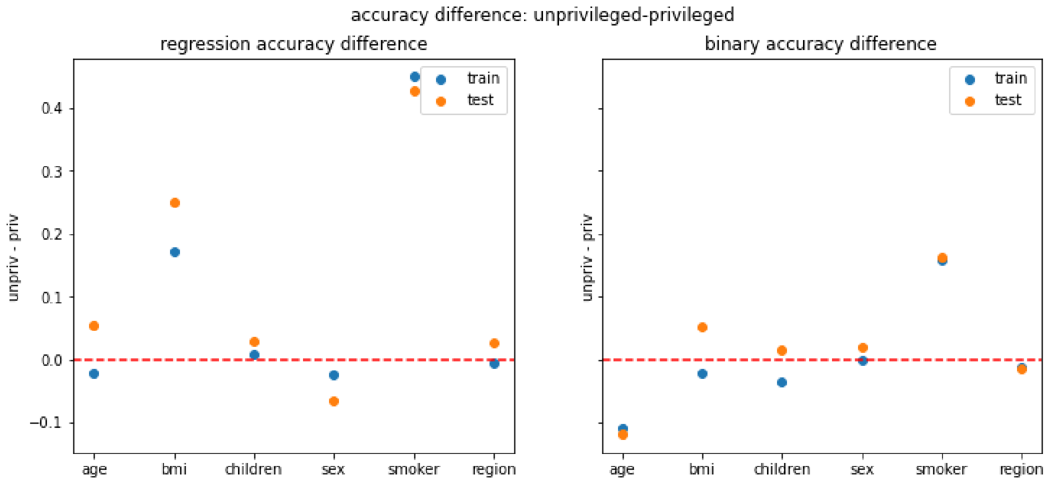


Figure 10: Difference of accuracy measures between groups

4.3 Misclassification Across Sub-Populations

Besides accuracy, which shows the model's ability to make predictions for a specific sub-population, misclassification rates made by the model were also assessed across the same sub-populations to determine whether or not it presents a bias. In addition to binary accuracy, 3 other metrics were used which are computed based on the binary results: selection rate, FNR, and FPR. Their in-detailed reports can be seen in figure 22 through 27 in appendix A. To better assess whether there is a difference in classification rates between privileged and unprivileged groups, the difference between their measures was computed and plotted in figure 11.

A dot below the dotted line signifies that the unprivileged group has less rate than the privileged group. While a dot above the dotted line suggests otherwise. For selection rate, we can see clearly that unprivileged groups of age and smoker were predicted to have less favorable outcomes (low

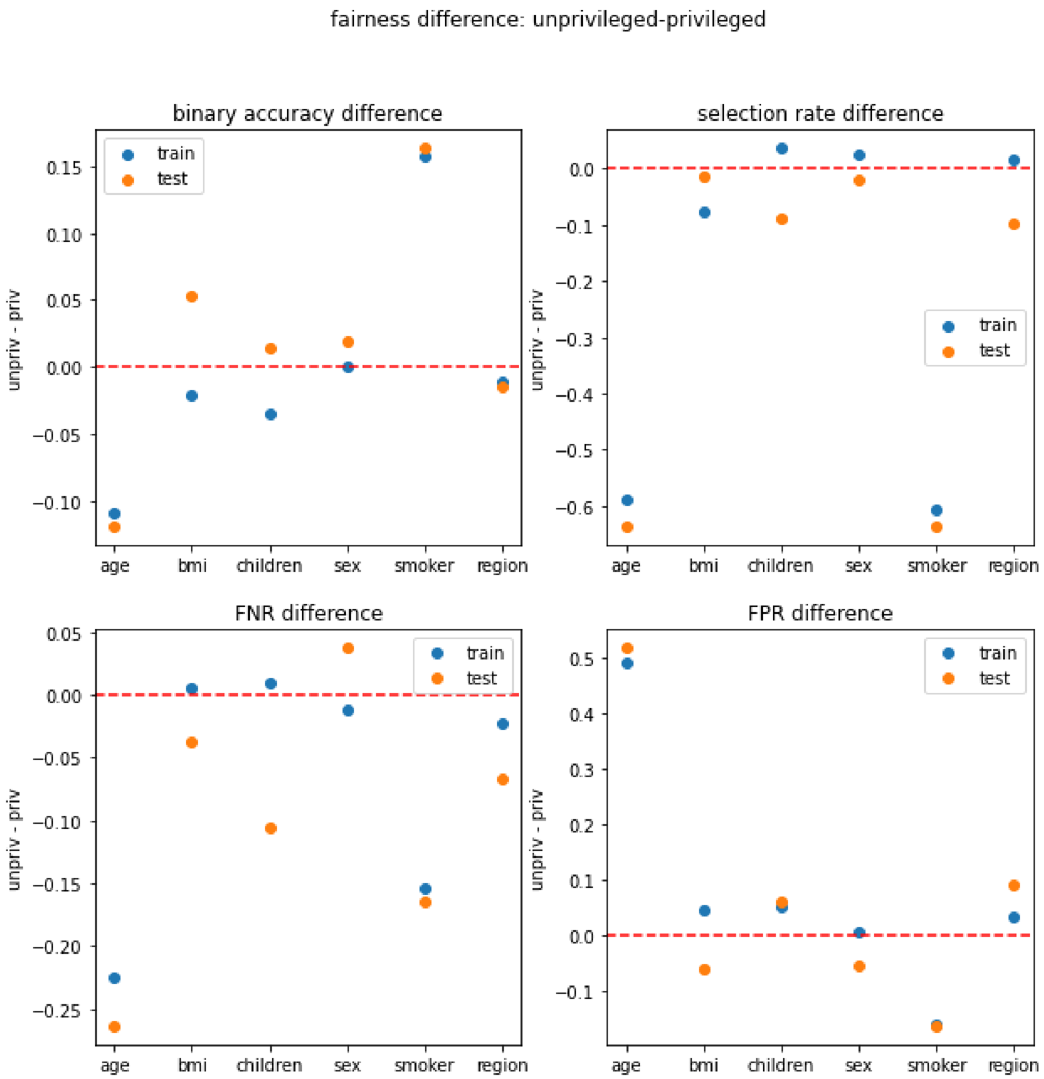


Figure 11: Difference in Misclassification Rate Across Sub-populations

expense). This is somewhat expected since these unprivileged groups had higher medical expenses in the training data which lead to higher medical expenses being predicted. The same unprivileged groups of `age` and `smoker` had less FNR rates than their privileged counterparts. This suggests that if a person who has high expenses were falsely predicted to have low expenses, he/she is likely to be young and/or not smoking, rather than old and/or smoking.

Surprisingly, only the unprivileged `age` group had significantly higher FPR than the privileged group, while the unprivileged smoker group had lower FPR than the privileged group. This means that older people are more likely to be classified as having high expenses when in reality they don't. Following the same logic, we would also expect to see a higher FPR for the unprivileged smoker group, where smokers are likely to be classified as having high expenses when in reality they don't. But upon closer inspection of the data, this lower FPR rate makes sense. Our model made no classification mistakes for the unprivileged smoker group. All smokers in the test set were predicted to have high expenses, and all of them did have high expenses. Thus, the ADS might have gotten lucky in the sense that its bias towards smokers was not validated with sufficient data. We conclude that the low FPR rate does not rule out the possibility of the ADS being biased towards smokers during deployment.

4.4 Statistical Fairness Across Sub-Populations

Besides the accuracy metrics, We also would like to explore the Statistical Disparity between all unprivileged groups and privileged groups. Statistical Disparity is helpful to examine whether there are favorable outcomes received by the privileged group. In order to grasp a comprehensive understanding of our data and model, the Statistical Fairness is checked both on the training set and the test set because this would give a better insight of whether pre-existing bias exists in the dataset and it has been reinforced by the ADS. To examine Statistical Fairness, we mainly focus on three metrics, including Mean Difference, Error Rate Difference and Disparate Impact. Mean Difference is calculated by mean label value on unprivileged instances - mean label value on privileged instances while Error Rate Difference is calculated by error rate of unprivileged instances minus error rate of privileged instances. The variation of these two metrics across all sub-populations is shown in figure 12. Disparate Impact is computed as the ratio of the rate of favorable outcomes for the unprivileged group to that of the privileged group and its variation across all sub-populations is shown in figure 13. The red dashed line in the figures represents the ideal value of each metric.

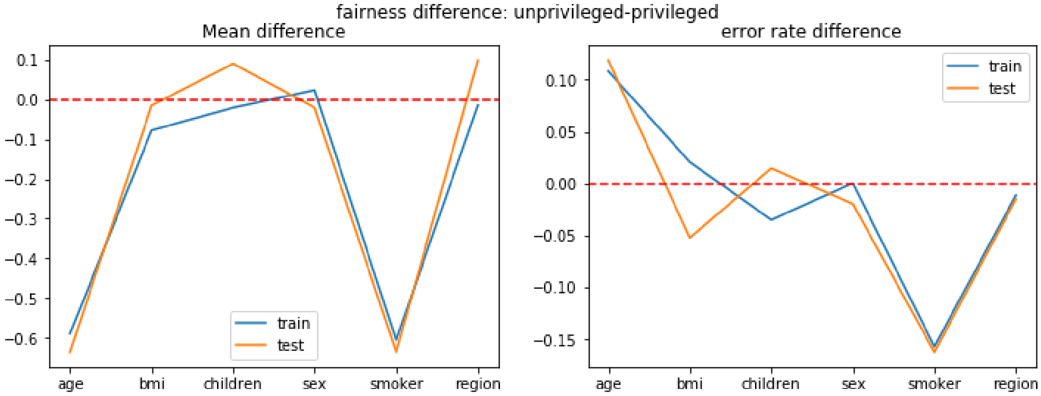


Figure 12: Variation of Mean Difference and Error rate Difference Across Sub-populations

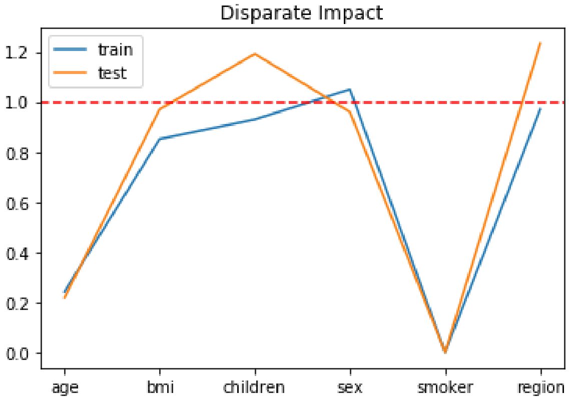


Figure 13: Variation of Disparate Impact Across Sub-populations

According to our observation, there are obvious biases existed within the sub-populations of `age` and `smoker` while minor bias existed within the sub-population of `bmi`. Comparing the Mean Difference and the Disparate Impact in these sub-populations on the training set and the test set, it is indicated that the ADS has enlarged the unfairness between the sub-populations of `age` and `smoker` since the Mean Difference and the Disparate Impact on the test set are lower than on the training set. In other words, the privileged groups in these two sub-populations are apparently receiving favorable outcomes against the unprivileged groups.

4.5 Stability

The stability of a model signifies whether or not the model is robust to uncertainties and noise. Thus, to assess whether or not this ADS is stable, we added random noise drawn from Gaussian distributions at different noise levels to observe the change in its predictions. Since it is not reasonable to add noise to categorical variables, we have only added noise to our numerical features `age, BMI, children`. In all figures, the similarity is measured as the difference between the predictions on the original test set and the predictions on the test set with added noise to a particular feature.

In figure 14, we can see that both regression and binary predictions show an almost linear decrease in similarity as noise level increases. This suggests that our predictions are somewhat sensitive to the noises in age. If the age of clients is not reported accurately, it may impact the accuracy of our prediction. However, in figure 15 we observe different behavior. As noise increase for `bmi`, the predicted numerical expenses becomes very different, but the predicted binary outcome stays the same. This suggests that while noisy `bmi` values may impact the exact predicted medical expense, it is unlikely to change whether a person is classified as having high expense or low expense. As noise increase for `children`, both the predicted numerical expenses and the predicted binary outcome stay very similar. This suggests that while noisy `children` values are unlikely predictions in any way, and our model produces stable results.

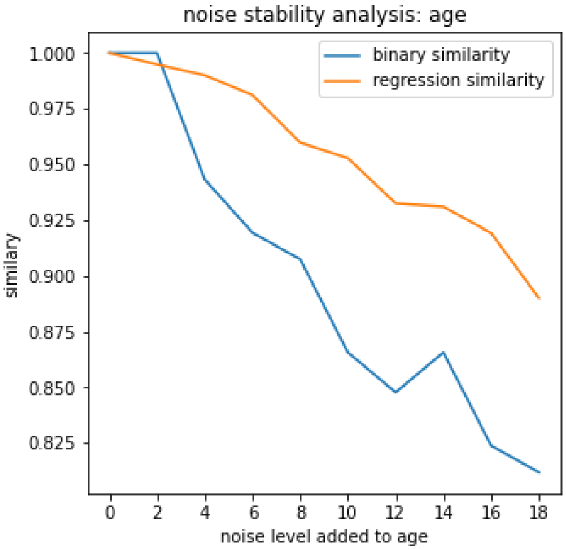


Figure 14: Stability in Age

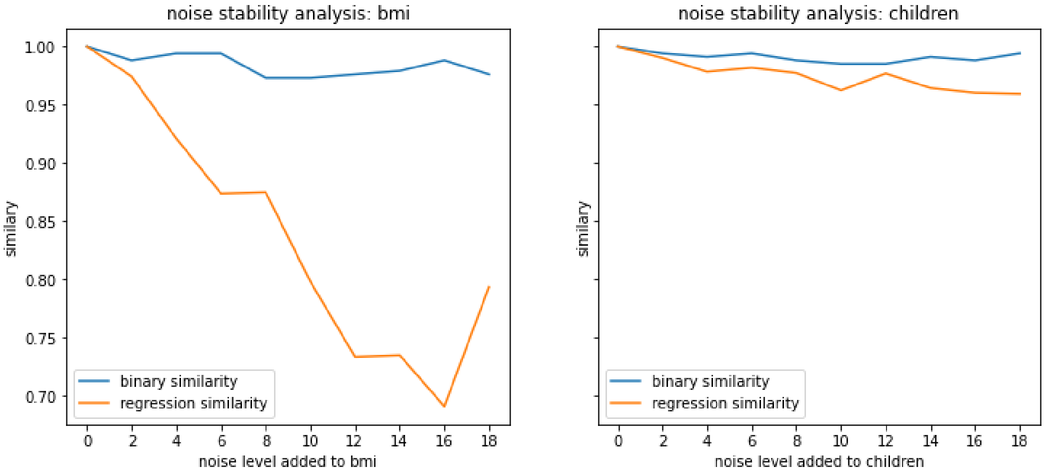


Figure 15: Stability in BMI and Children

5 Summary

The data used by this ADS is appropriate since we believe the data is real-world-reflective even if some biases have been observed in the data. Because it is truly the fact that elder people and smokers are more likely to have worse physical conditions and encounter more health issues. Therefore, the pre-existing biases observed in the data cannot be attributed to improper data collection. However, the fairness metrics indicate the privileged groups in these two sub-populations receive favorable outcomes against the unprivileged groups and the ADS has somehow worsened such discrimination. This would benefit the insurance companies since some smokers and elders with great physical conditions might also be charged a high insurance premium while impairing these applicants. In terms of the accuracy metrics, the overall accuracy of both the regression model and binary model is similarly high, around 0.88, while it is more important to pay attention to the misclassification metrics, FPR and FNR, across sub-populations. FPR gives insight into the rate of the low-premium applicants being misclassified as the high-premium group while FNR vice versa. Therefore, high FPR benefits the insurance companies and harms the applicants, while high FNR benefits the applicants and harms the insurance companies. From the perspective of stability, our metrics indicate that the ADS performance is quite stable on the variation of `age` and `bmi`.

Overall, we feel comfortable deploying this ADS in the industry. One reason is that the dataset itself does not seem to have issues in collecting methodology and no obvious bias has been blended in, which is capable of reflecting the real fact. Secondly, considering that all sorts of metrics, including accuracy, fairness, misclassification, and stability, are in good shape, we believe the model generally shows a perfect performance and reaches the stated goal. However, there is still some improvement we recommend for this ADS. Firstly, the original model did not implement parameter tuning for the random forest. It is very likely that the model could have a better performance in all aspects if a better parameter configuration is found. Secondly, techniques such as Disparate Impact Remover, Reject Option Classification could be applied to mitigate the observed unfairness in the ADS and get more balanced FPR and FNR across all sub-populations.

A Outcomes

A.1 Figures

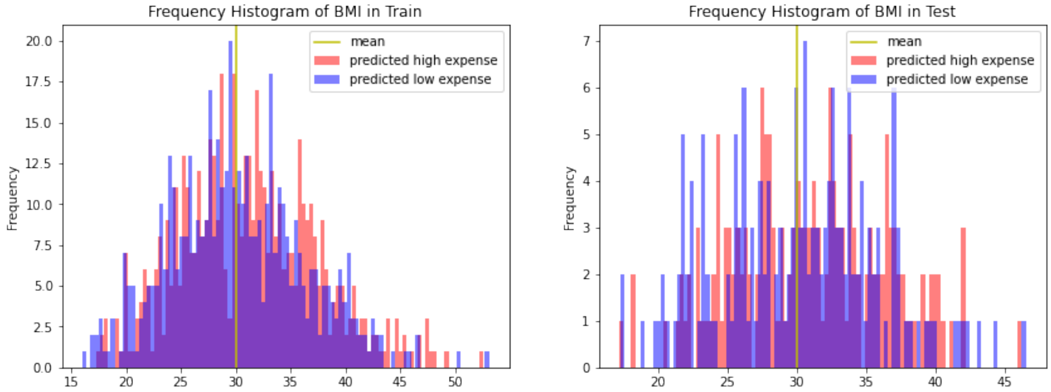


Figure 16: Distribution of Binary Outcome Across BMI

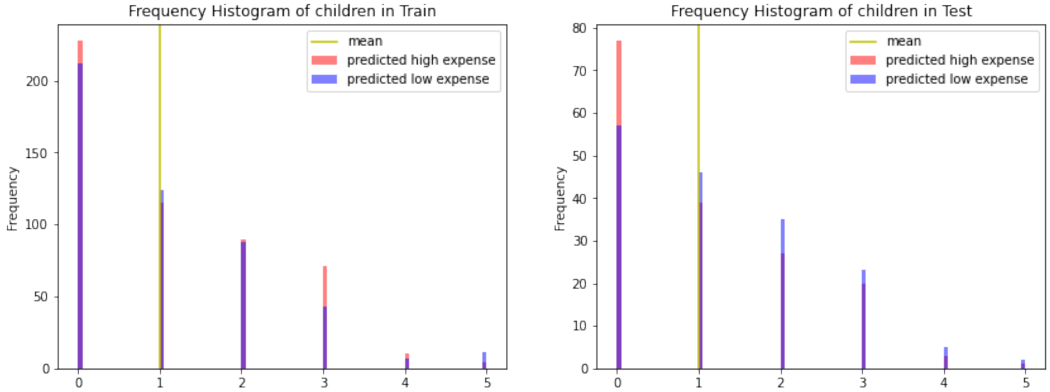


Figure 17: Distribution of Binary Outcome Across Children

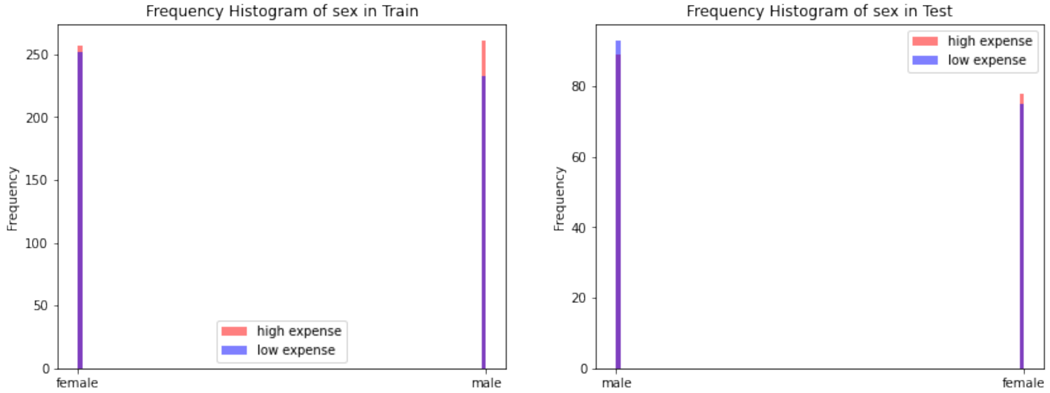


Figure 18: Distribution of Binary Outcome Across Sex

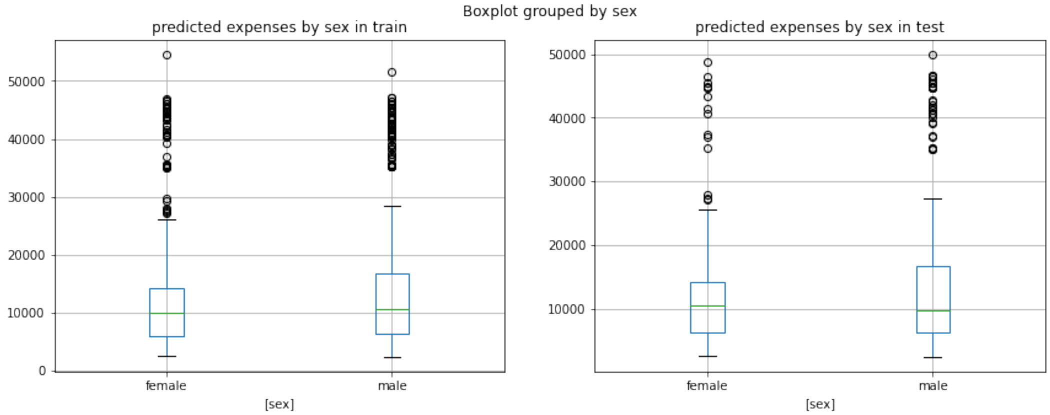


Figure 19: Box Plot of Numerical Outcome Across Age

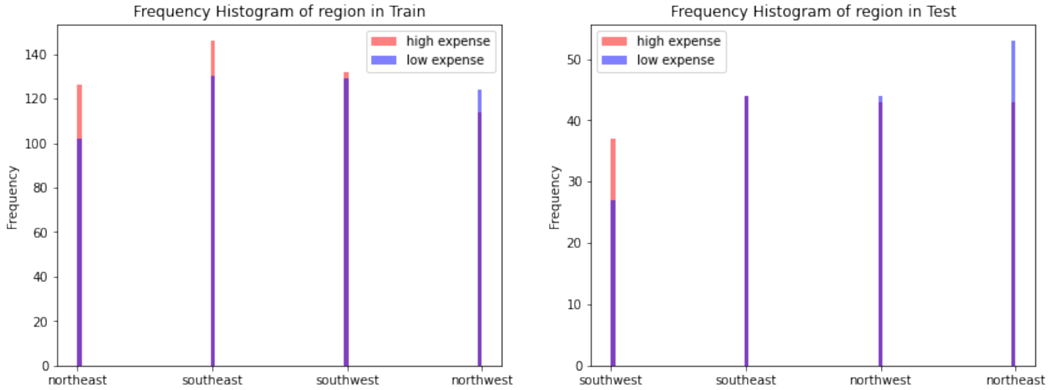


Figure 20: Distribution of Binary Outcome Across Region

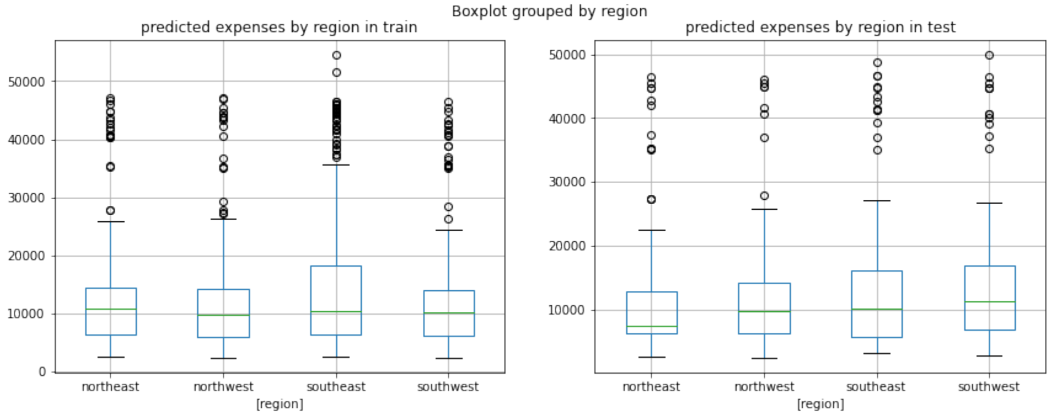
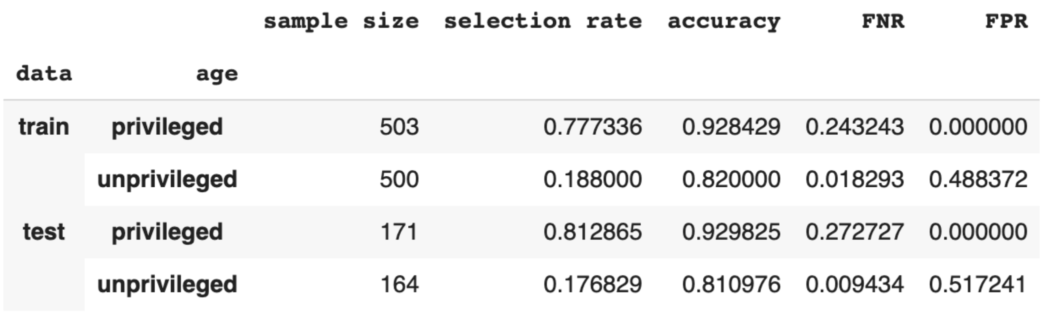
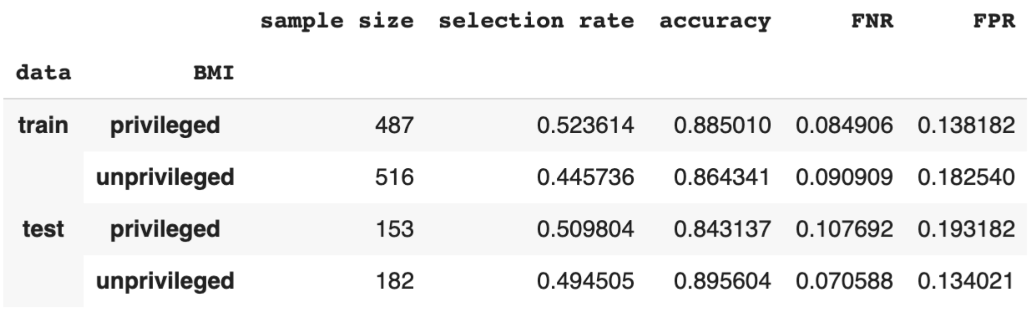


Figure 21: Box Plot of Numerical Outcome Across Region



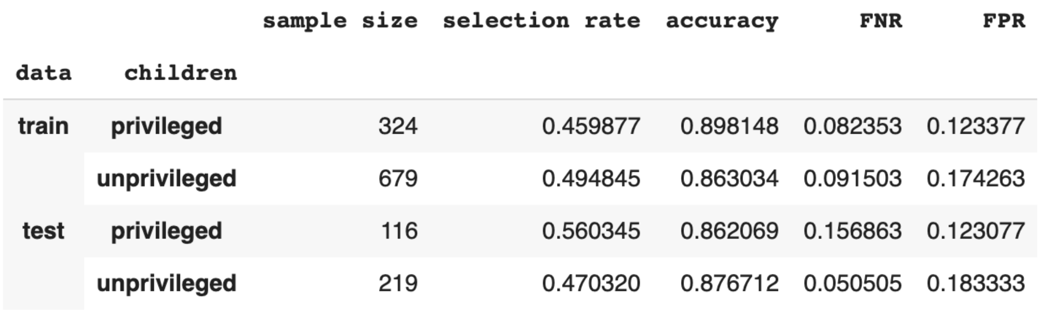
		sample size	selection rate	accuracy	FNR	FPR
data	age					
train	privileged	503	0.777336	0.928429	0.243243	0.000000
	unprivileged	500	0.188000	0.820000	0.018293	0.488372
test	privileged	171	0.812865	0.929825	0.272727	0.000000
	unprivileged	164	0.176829	0.810976	0.009434	0.517241

Figure 22: Misclassification Rate Across Age



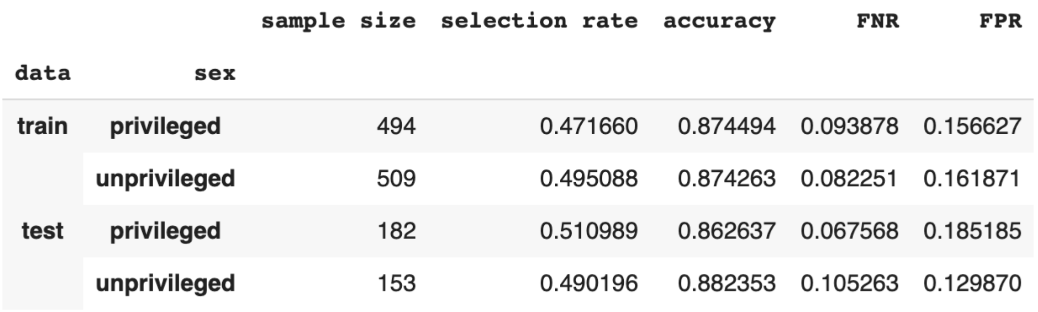
		sample size	selection rate	accuracy	FNR	FPR
data	BMI					
train	privileged	487	0.523614	0.885010	0.084906	0.138182
	unprivileged	516	0.445736	0.864341	0.090909	0.182540
test	privileged	153	0.509804	0.843137	0.107692	0.193182
	unprivileged	182	0.494505	0.895604	0.070588	0.134021

Figure 23: Misclassification Rate Across BMI



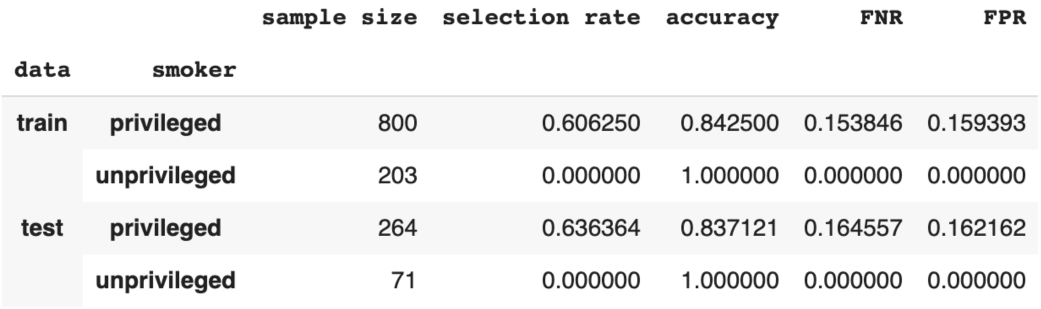
		sample size	selection rate	accuracy	FNR	FPR
data	children					
train	privileged	324	0.459877	0.898148	0.082353	0.123377
	unprivileged	679	0.494845	0.863034	0.091503	0.174263
test	privileged	116	0.560345	0.862069	0.156863	0.123077
	unprivileged	219	0.470320	0.876712	0.050505	0.183333

Figure 24: Misclassification Rate Across Children



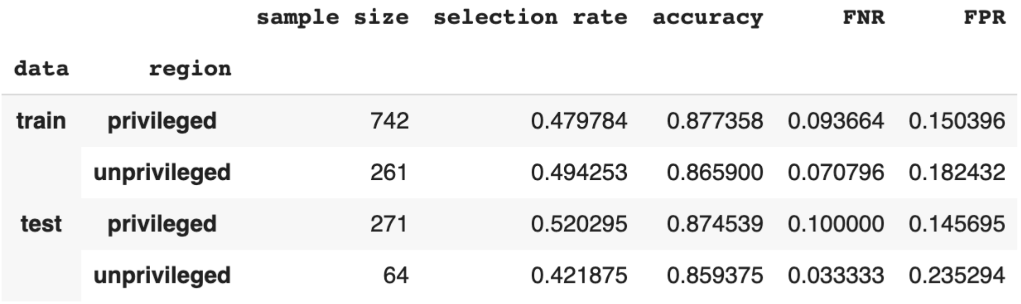
		sample size	selection rate	accuracy	FNR	FPR
data	sex					
train	privileged	494	0.471660	0.874494	0.093878	0.156627
	unprivileged	509	0.495088	0.874263	0.082251	0.161871
test	privileged	182	0.510989	0.862637	0.067568	0.185185
	unprivileged	153	0.490196	0.882353	0.105263	0.129870

Figure 25: Misclassification Rate Across Sex



		sample size	selection rate	accuracy	FNR	FPR
data	smoker					
train	privileged	800	0.606250	0.842500	0.153846	0.159393
	unprivileged	203	0.000000	1.000000	0.000000	0.000000
test	privileged	264	0.636364	0.837121	0.164557	0.162162
	unprivileged	71	0.000000	1.000000	0.000000	0.000000

Figure 26: Misclassification Rate Across Smoker



		sample size	selection rate	accuracy	FNR	FPR
data	region					
train	privileged	742	0.479784	0.877358	0.093664	0.150396
	unprivileged	261	0.494253	0.865900	0.070796	0.182432
test	privileged	271	0.520295	0.874539	0.100000	0.145695
	unprivileged	64	0.421875	0.859375	0.033333	0.235294

Figure 27: Misclassification Rate Across Region

A.2 Tables

group	regression accuracy for train	regression accuracy for test
privileged	0.881	0.854
unprivileged	0.860	0.909
	binary accuracy for train	binary accuracy for test
privileged	0.928	0.930
unprivileged	0.820	0.811

Table 3: Accuracy measures of the Age sub-population

group	regression accuracy for train	regression accuracy for test
privileged	0.741	0.682
unprivileged	0.913	0.932
	binary accuracy for train	binary accuracy for test
privileged	0.885	0.843
unprivileged	0.864	0.895

Table 4: Accuracy measures of the BMI sub-population

group	regression accuracy for train	regression accuracy for test
privileged	0.870	0.877
unprivileged	0.880	0.906
	binary accuracy for train	binary accuracy for test
privileged	0.898	0.862
unprivileged	0.830	0.877

Table 5: Accuracy measures of the Children sub-population

group	regression accuracy for train	regression accuracy for test
privileged	0.889	0.919
unprivileged	0.864	0.853
	binary accuracy for train	binary accuracy for test
privileged	0.874	0.863
unprivileged	0.874	0.882

Table 6: Accuracy measures of the Sex sub-population

group	regression accuracy for train	regression accuracy for test
privileged	0.482	0.458
unprivileged	0.932	0.885
	binary accuracy for train	binary accuracy for test
privileged	0.843	0.837
unprivileged	1.0	1.0

Table 7: Accuracy measures of the Smoker sub-population

group	regression accuracy for train	regression accuracy for test
privileged	0.879	0.890
unprivileged	0.873	0.917
	binary accuracy for train	binary accuracy for test
privileged	0.877	0.875
unprivileged	0.866	0.859

Table 8: Accuracy measures of the Region sub-population