All Results

Introduction

This is a file of extra images and work from the project. Some images may not display unless downloaded.

Findings - Neural Network Interpretation Results

Finding #1: As privacy increases, the features that the private neural network deem important vary more than the non-private model, potentially showing why accuracy decreases.

Since neural networks can be intialized randomly, it is important to average them many runs of the same model. Each row is the interpretation output of a run. We see that in the nonprivate model, while there is still variation from run to run, the top 4 features are relatively still the top 4. As the model becomes more private, there is less consistency from run to run. This variation could potentially explain the drop in AUC or accuracy as the model becomes more private.


Non-private Models
Private Models of ε-2.06
Private Models of ε-0.27
Loan Purpose
Lien Status
Property Type
App. Race
App. Ethnicity
Agency
App. Income
# of times ranked top 5 feature
Rank
1
2
3
4
5

Findings - Logistic Coefficient Results

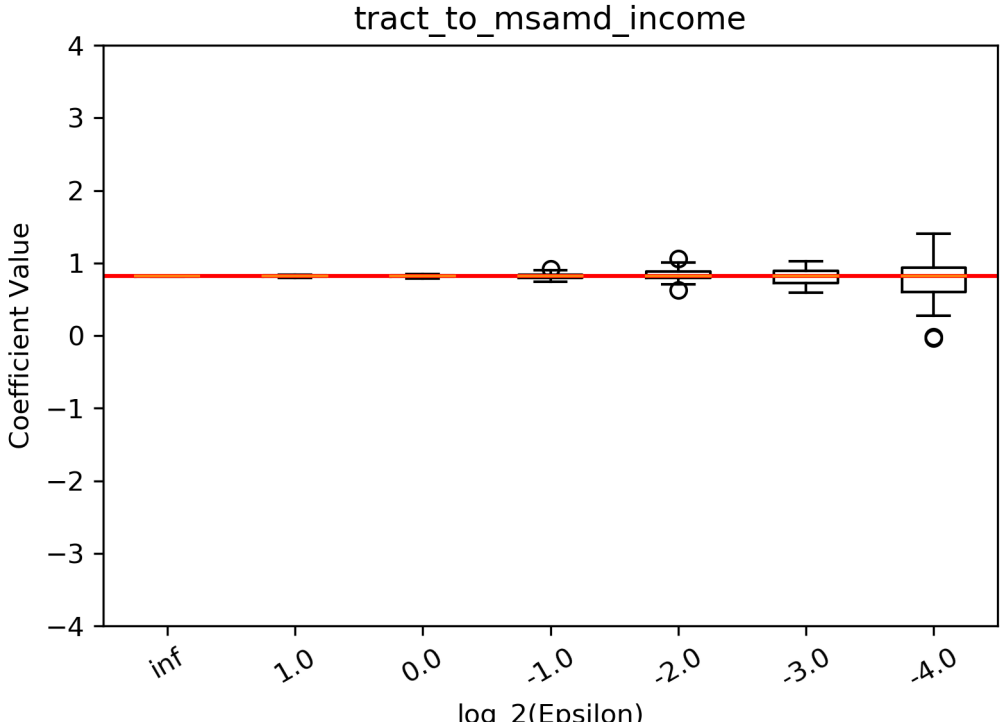
1. As the model becomes more private, the variation of coefficients *increases*.

From all of the boxplots, you can observe this happening.

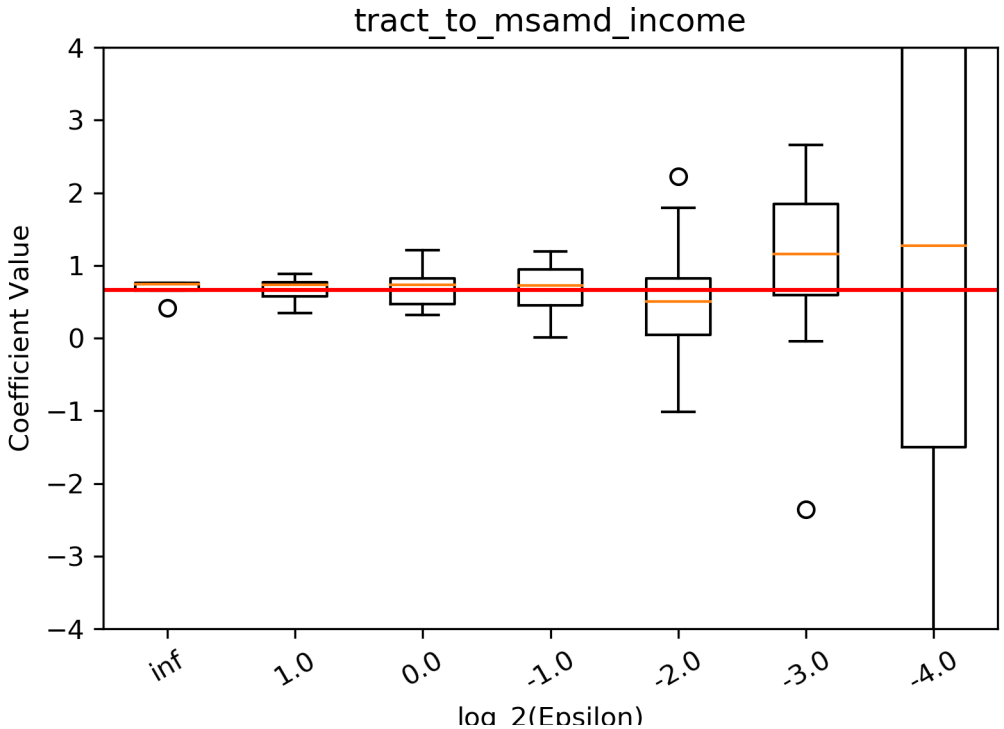
This is expected since the noise added to the models increase as epsilon decreases.

2. Compared to using the full dataset, using a smaller subset of the data in the model results in *larger* variation in coefficients and coefficient values.

For example, in the model using the full balanced acceptances and denials dataset, looking at the boxplot for tract_to_msamd_income shows that at the largest amount of privacy, the maximum the coefficient value reaches is 1.41, and the minimum is -0.03.



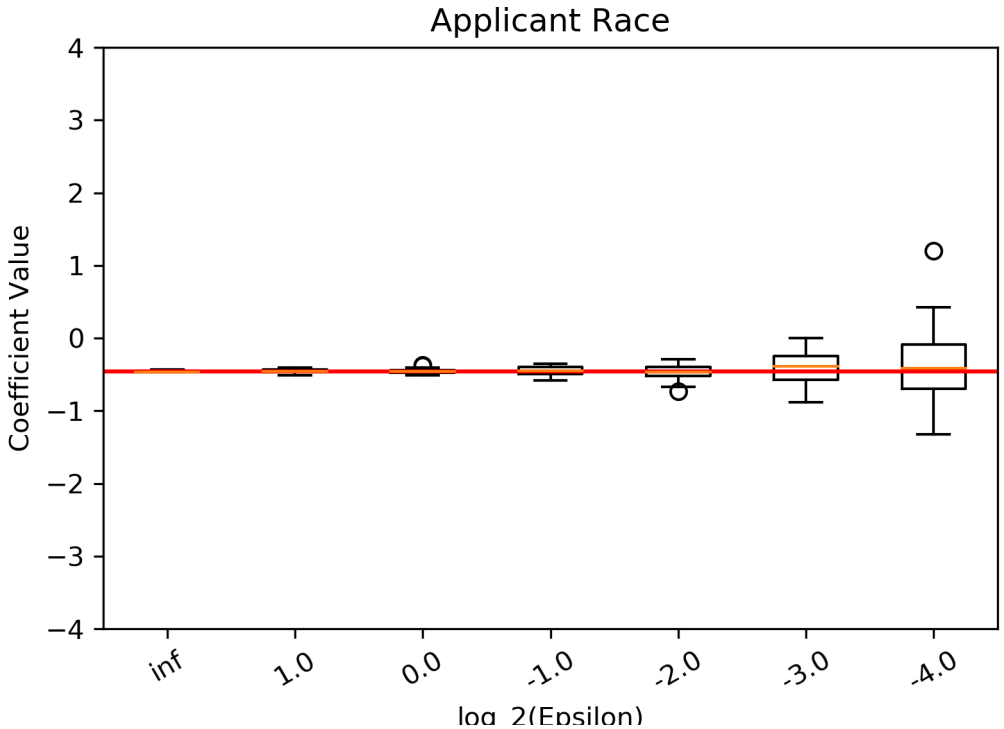
On the other hand, in the model using subsetted data, the boxplot for tract_to_msamd_income shows that at the largest amount of privacy, the maximum the coefficient value reaches is 8.39, and the minimum is -4.51.



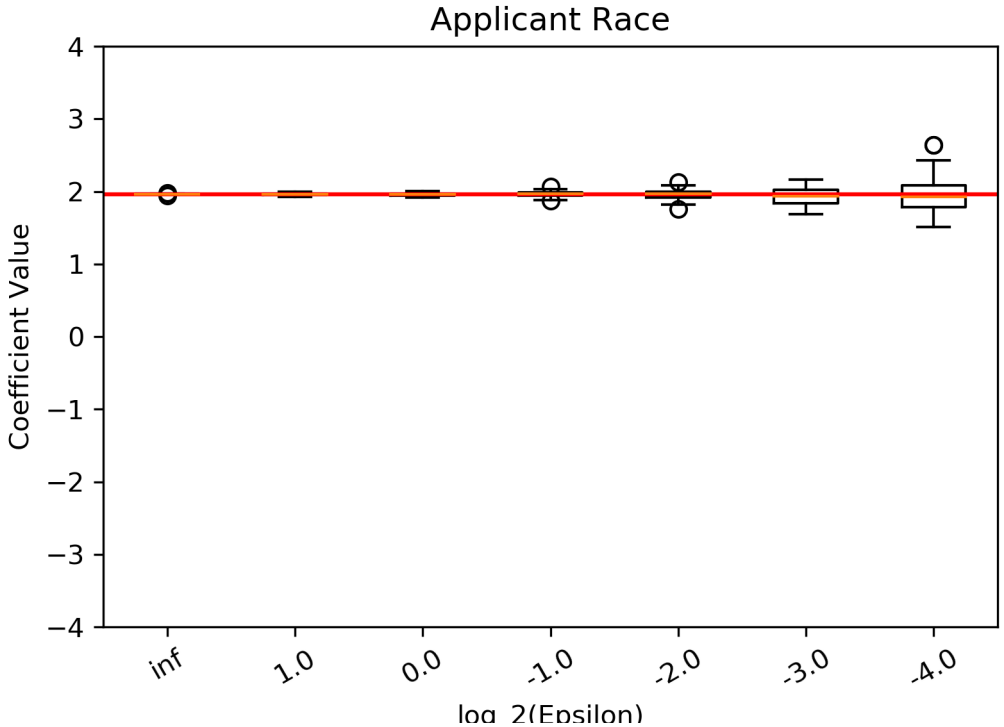
This could explain why you see larger changes in AUC and disparate imapact when using a subset of the dataset.

3. Balancing by acceptances, denials, and race *decreases* the median coefficient for applicant race (from 2 to -0.5)

Balanced by acceptances, denials, and race:

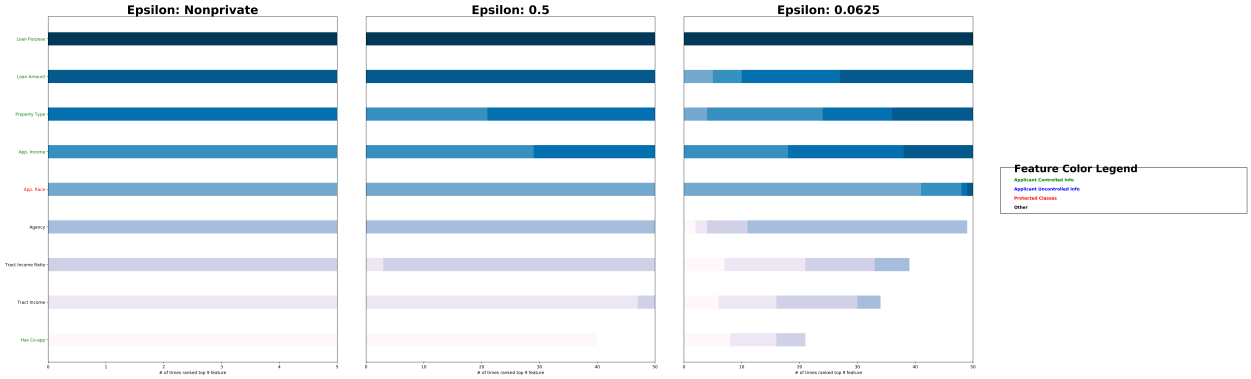


Balanced by acceptances and denials:



This could be why the model trained on this dataset is more fair than the model trained on only balanced acceptances and denials.

4. As the model becomes more private, the models tend to have the *same* top 7 rankings.



This could explain why the AUC and disparate impact is relatively unchanging for smaller epsilon.