**DSA 5103 Reflection Paper**

Matthew J. Beattie (beat0000)

December 16, 2016

**Revision Process:** I opted to try to create a near-final analysis in my rough draft, and in doing so ran several iterations of models with a variety of parameters to tune my models. Because my paper is of importance to me personally, and because I would like to prepare it for publication, I sought feedback from a practicing data scientist as well as from Ms. Barfield, my partner. Upon receiving that feedback, I incorporated their feedback into my revision. Additionally, I made one substantial change to my analysis based upon the feedback from the practitioner.

**Feedback:** Ms. Barfield’s feedback centered on revisions to my write-up, and I found her feedback to be helpful in clarifying some of the points I was trying to make. She called out the need to be more clear in explaining some of the graphics I used in the beginning of the paper. Ms. Barfield validated my approach to the problem and confirmed that my application of the models was appropriate.

The practitioner’s feedback also called for clarification of my explanation of the effect of education on heroin use. In particular, I needed to better explain the correlation between education level and likelihood of use. More importantly, she pointed out that due to the very low number of heroin users in the study, my initial choice of 20% of the dataset for a training sample was too low, and that I should increase it.

**Application of Feedback to Analysis:** I redid my entire analysis using a training sample size of 50% of the dataset. When I did so, I found that my decision tree models, particularly the CART model, changed significantly. The unpruned CART model was greatly simplified, and in fact, I could’ve elected not to apply a penalty to the tree formation. I elected to retain the penalty to remove a variable from the outcome. Another way to accomplish the same effect would’ve been to have removed that variable from the dataset, as it was of no use in the other analyses either.

The impact of the larger training dataset was less on the AdaBoost analysis, which retained much of the important variables as the smaller set. However, there were some changes, such as the dropping of Librium use as a predictor in favor of tranquilizer use. Because I used the AdaBoost outcome as a way to design the logistic regression, this had a downstream impact on my final model. In the logistic model, the importance of early and frequent cocaine use became more important.

**Further Research and Improvement**: One of the common findings across the models was the importance of early drug use in identifying heroin users. However, while the logistic model hinted at an extreme impact of early general drug use due to the size of the model’s coefficients, those variables did not pass a statistical validity test. Further data engineering might isolate those variables in such a way as to make them valid. Additionally, I can correlate some of these findings with a study of non-survey data. There are other drug use datasets that I can research that focus on treatment data.

Finally, I did not find many published big data drug use studies, despite the ready availability of the data. I intend to continue my research along these lines and publish my findings. By doing so, I hope to contribute in whatever way I can to a reduction in drug deaths in the United States.