STATS 415 Group 11 Project Proposal

Michael Williams, Nikhil Dayal, Samuel Tan, & Sarah Lim

Dataset: Demographic Variables and Sample Weights 2017-2018
https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO_J.htm

Two questions to explore:

1) Does the amount of children or education level have a greater impact on annual family income?
2) Are there any interaction variables that help us predict annual family income? To what extent do they help us?

Response variables:
INDFMIN2 - Annual family income

- We chose this variable because when doing some exploratory research on the data set we found many other variables had some kind of correlation with another single variable. Annual family income, however, did not have a high correlation with any individual predictor variable, thus we figured it would be interesting to explore how we can use combinations of other variables to predict this metric. This provides us with a difficult, but manageable, challenge.

At least 3 methods out of (1) KNN, (2) LDA/QDA, (3) lasso/ridge penalties, (4) curvature penalties, (5) svm, (6) trees, (6) bootstrap, and (7) cross-validation

Cross-Validation - While not a modeling method itself, we want to use CV to stress-test our results. While bootstrapping would also help to accomplish this, we were worried firstly, about data leakage, and secondly, about the amount of computing power it would require to generate a reasonable number of surrogate datasets on 7000+ observations with 30+ variables.

Svm - SVMs can be used for both binary and multiclass classification problems. They can naturally extend to multiple classes without the need for complex modifications. It can handle imbalanced datasets by giving appropriate weights to different classes. This is crucial in applications where one class might have significantly fewer instances than others, which is the case in this dataset. Lastly, SVM is very good in high-dimensional spaces.

Lasso/Ridge - Lasso leads to increased interpretability, as it sets many coefficients to zero (feature selection). Lasso can rapidly tell us which variables may be the most important. Lasso handles collinearity which is very prominent in this dataset.

Trees - While there is no obvious inherent benefit to using trees on this problem, we thought it would be fun to experiment with using trees on variables with many categories. THis is why we choose it as a fourth method.