Alfred Hurley
Mathematics
Class of 2019

Economists will often use econometrics to prove causation between events or trends in the economy. The two main hurdles when using econometrics are maintaining statistical truth, specifically ensuring that the results are widely applicable, and certifying that one is proving causation as opposed to correlation. A possible solution to these obstacles is machine learning.

Often times when economists have large datasets that they are using to either discover or prove causation with, they will resort to p-hacking their data to find mostly meaningless correlations between data points. The issue with p-hacking is that it creates work that is not reproducible and not true. These fallacies come about because the model that the economist is creating is over fit to the specific data that it has been fit modeled on, and thus not applicable to new data. A model is over fit when it starts taking into account irreducible error, something that will almost always be in a data set, but is random and thus not predictable. One of the reasons p-hacking will find correlation but not causation is because when given a data set with a huge amount of observations computers will almost always find some pattern in the data, this is where machine learning can help.

The simplest machine learning technique would be data splitting: splitting one's data into two halves, training the data on one half, and using the other half to estimate affects or causation, thereby reducing the frequency of meaningless correlation. The obvious issue with splitting one's data is that the sample size is drastically decreased, which has a negative effect on statistical significance. There is a simple solution to this problem: one can preform data splitting on one's data many times and simply take the average of the different models produced each time. By taking the average of all the different models, one reduces the meaningless correlation, as it should be averaged out, and maintains statistical significance because all of one's data is used. The technique of data splitting is effectively cross validation, the machine learning method where one runs a model many times on slightly varied test and training data sets to learn how good the model is.

Machine learning can provide economists with an alternative to current regression methods and thus give them the options to not hack their data into giving them meaningless or useless results.