Methods and Results

Bethany Bailey

May 9, 2018

1 Model

A major component of pension fund risk and return is the fund's asset allocation, specifically its allocation to equity. This equity allocation can thus be used as a proxy to explain a fund's risk behavior. Previous research has looked at how political composition, performance, and other factors effect risk behavior and tolerance in funds. In my model, I theorize that internal fund characteristics, such as size, funded ratio, and previous investment return, as well as state characteristics such as budget surplus and tax rate, can be used to predict the risk tolerance and thus the equity allocation of a fund. I will use machine learning algorithms, specifically a random forest model and neural network, to try to predict equity allocation given the above characteristics.

2 Data

The data for this study came from the Public Plans Data pension database. The data is produced and maintained by the Center for Retirement Research at Boston College in partnership with the Center for State and Local Government Excellence (SLGE) and the National Association of State Retirement Administrators (NASRA). It is available for direct download in .csv format at the Public Plans website. This database contains annualized details about 170 large pension plans in the United States (114 state and 56 large local plans), which covers 95% of public state and local plans in the U.S. The data covers the years 2001-2016. There are over 150 variables that include information topics including, but not limited to, funding, accounting assumptions and methods, allocation, returns, and fund characteristics. These data were collected from information available in the most recent Comprehensive Annual Financial Reports (CAFRs), actuarial valuations (AVs), and the Public Fund Survey and are updated regularly by the NASRA. Previous research using the Public Plans Data includes a study by Mohan and Zhange (2014) that looked at risk behavior in defined benefit pension plans. Additionally, researchers have looked at the funding status of pension plans using this data (Munnell et al. (2011)).

I joined this data with state-level data from the Correlates of State Policy Project at the Institute for Public Policy and Social Research (IPPSR) at Michigan State University. This dataset contains over 900 variables for each state from 1900-2016. The variables cover many topics, including, but not limited to, policy, demographics, economic and fiscal policy, eduction, election information, government, public opinion, partisanship, and ideology. You can download the data in excel, .csv, Stata, or R on the IPPSR website. The data has been collected, cleaned, and combined over the years by scholars and students who combined multiple smaller datasets. The specific variables I use in this model (budget surplus as a percent of state GSP and taxes as percent of state GSP) came from Carl Klarner's State Economic Data database published in 2013.

For this research, the data were downloaded from the above websites in .csv files. Each database was loaded into a pandas dataframe. The databases were then joined by year and state so that each fund/year combination had all the data. Then, different combinations of variables were extracted from the dataframe and cleaned of missing values, making sure that missing values occurred either random or not random in a way that could be dealt with, such as more recent or older years being subtracted from the data. In this specific case, our data ended up only covering 2001-2010 because the State Policy dataset did not have more recent years. These dataframes with different combinations of data were used in analysis as outlined below.

Table 1 shows the summary statistics for the variables of interest in this study. These variables are presented from 2001-2010 with all the missing values removed, but prior to normalizing the variables.

3 Methods

My first methodology was to compare two different predictive models: a neural network (specifically, the scikit-learn neural net regressor package in python) and a ran-

Table 1: Summary Statistics for Key Variables

Summary	Total	Inv. Return	Taxes	Total	Budget Surplus	Funded
statistics	Equity	(1 yr)	(% of GSP)	Membership	(%GSP)	Ratio
Mean	0.558178	0.044725	5.429465	140,687.5	-0.031150	0.855917
Std	0.105288	0.123225	1.221717	211,962	0.980134	0.186101
Min	0.000000	-0.307000	3.262894	2,627	-8.454143	0.191000
Max	0.813916	0.314000	17.753800	1,621,906	14.300940	1.973957
25%	0.508000	-0.050000	4.696317	19,090.25	-0.429372	0.750000
50%	0.573000	0.088000	5.369527	66,442.5	-0.055432	0.855000
75%	0.625000	0.138700	6.094493	161,896.5	0.325267	0.960000

dom forest model (using the scikit-learn decision tree regressor package in python) to predict my dependent variable, total equity (domestic and international). I theorized that budget and tax systems of a state, as well as fund size, funded ratio, and previous fund return, would effect the risk tolerance and thus the equity allocation of a fund. These variables mapped onto the theoretical concepts I was trying to test, and were chosen for the following reasons:

- The dependent variable we are trying to predict, *overall total equity allocation*, is a useful construct for risk due to asset allocation's place in modern portfolio theory and previous literature which has used allocation as a risk measure.

- *State budget surplus (% of Gross State Product)* is a measure of risk tolerance for the broader political and governmental sphere. It may directly (through greater funding of pensions) or indirectly (through political or economic policy) influence the risk behavior of a fund.

- *Funded Ratio* is a more direct measure of risk tolerance for a fund. The level of fundedness might cause investors to change their risk allocation in order to chase returns, or in order to avoid risk that might further diminish funds' abilities to meet their liabilities. This hypothesis comes from previous research that has asked whether investors are likely to engage in imprudent behavior when funds are low (Weller and Wenger (2009)). Others, such as Lucas and Zeldes (2009), have hypothesized the opposite, that funds engage is less risky behavior when funds are low. Though I was initially concerned that funded ratio might be highly correlated with investment return, a preliminary scatter plot and regression analysis indicated that it is not (R^2 of 0.004).

- Similarly, a low *one year investment return* might prompt an increase in risk behavior (chasing returns to make up for loss) or a decrease in risk taking (”cutting one’s losses”).

- *Taxes (% of Gross State Product)* represent two main concepts in this data: (1) conservativeness and (2) level of government fundedness.

- *Total membership* is a proxy for fund size.

Before I ran these models, I removed the observations from 2011-2016 (1,020 observations) because I did not have state data on these observations. Then, I removed additional missing values, which were missing at random. This took my data set from the original set of 2,692 observations of 16 years to 1,672 observations of 10 years to the final dataset of 1,502 observations. I then scaled all the variables other than year to have mean 0, standard deviation 1. This was necessary for my neural net application.

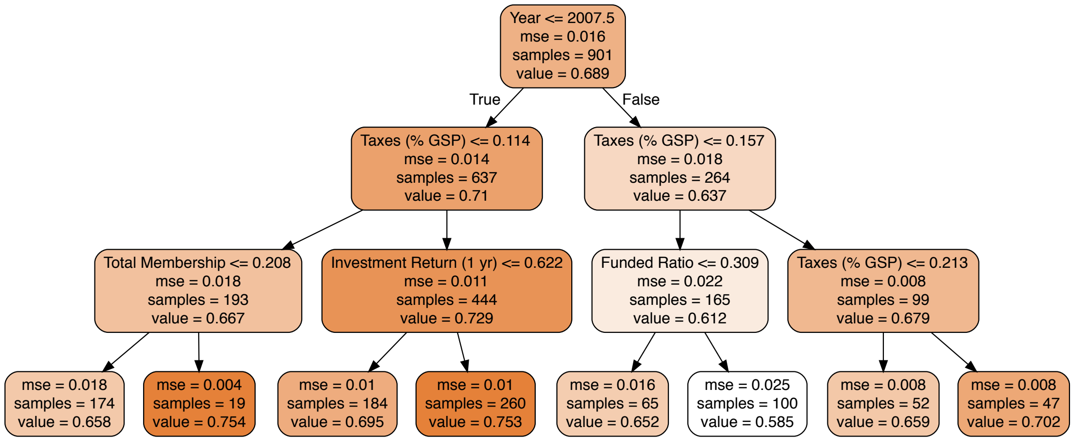
I also ran a simple linear regression on the normalized variables for interpretability of the effect of each variable. This regression did not yield significant results, so I will leave it out of my discussion of results.

4 Initial Results

I ran the random forest tree model using all the variables in my cleaned dataset (year, funded ratio, total membership, investment return (1 yr), budget surplus (% of GSP), and taxes (% of GSP). I divided the 1502 observations in my cleaned dataset randomly into a training and a testing set, and ran a tree with three layers and a minimum of ten samples per leaf. I reached these layer and leaf size parameters by systematically testing different combinations of each. This combination of parameters gives me the greatest predictive power.

The random forest tree generated is visualized below in Figure 1. The ”value” at each stage is the predicted value for Total Equity, and the "mse" is the mean squared error for that prediction.

Figure 1: Random Forest Model of Total Equity (MSE = 0.0159)



This model has a mean squared error of 0.0159, which at first glance seem very good. However, this gives a mean error rate of approximately 0.125, which is one standard deviation (0.13) away from the normalized mean equity allocation. Thus, this model is not very predictive.

Part of this error could be due to sensitivity to the training and testing set. Thus, in order to create a more predictive model, I also tried using a bagging regressor, which randomly draws many training sets from the data, estimates the trees for each set, and then finds the bagging decision tree as the average predicted x across all the trees. This performed better, with a mean squared error of 0.0137 and average error of 0.117. However, this model is still not very predictive, so I am currently in the process of adjusting my parameters and testing different variables to see if I can build a more predictive random forest model.

In addition to this RFM, I ran a multilayer perceptron neural network using scikit-learn's neural network MLPRegressor module in python. This neural net is a machine learning algorithm that creates layers of nonlinear functions of features. I found the best results using the the hyperbolic tan function ($tanh(v) = 2\sigma(2v) - 1$) as my activation function and stochastic gradient descent as my solver. To extend my data, I used k-fold cross validation, dividing the data randomly into four groups, holding out the kth datasets as my test sets, and training four different models on the additional k-1 datasets. My average mean squared error was 0.0167, which corresponds to an average error rate of approximately 0.129. Again, this error rate is too high to accurately represent the data; thus, I am currently analyzing different combinations of models and variables.

Table 2 shows the current results of the three different models.

Table 2: Model Error Rates

Model	Random Forest Model	Random Forest Model	Neural Network
Type	(one train/test split)	(Bootstrapped)	(4-folds)
MSE	0.015859	0.013706	0.016745
Error	0.125932	0.117071	0.129404

References

Lucas, Deborah J. and Stephen P. Zeldes, "How Should Public Pension Plans Invest?," *American Economic Review*, May 2009, *99* (2), 527–532.

Mohan, Nancy and Ting Zhange, "An analysis of risk-taking behavior for public defined benefit pension plans," *Journal of Banking and Finance*, March 2014, *40*, 403–419.

Munnell, Alicia, Jean-Pierre Aubrey, and Laura Quinby, "Public pension funding in practice," *Journal of Pension Economics and Finance*, April 2011, *10* (2), 247–268.

Weller, Christian E. and Jeffrey B. Wenger, "Prudent investors: the asset allocation of public pension plans," *Journal of Pension Economics and Finance*, October 2009, *8* (4), 501–525.