Abstract

Regression analysis and tree based models are a widely used statistical methodology for fitting data and making a prediction. This capstone project is about studying factors affecting Georgia unemployment. The factors we considered are national unemployment, if there is a war going on, if there is a presidential or governor election, if the president, governor, house, or senate has a republican or democratic majority, the end of the year value for both the Dow and S&P 500. We fit regression and tree-based models. Then, we compared the prediction accuracy of these models using mean square error of the test data set.

Introduction

The unemployment rate in Georgia is something that affects everyone living in the state of Georgia. The unemployment rate is the percent of people who are collecting unemployment in the state of Georgia. This number is important since unemployment is something paid by the taxpayers, and collected by the people who have lost their jobs. It is a part of the economy that can either boost the overall quality of the economy, stall it from progressing, or make a negative impact on the economy. This being said the lower the percentage of unemployment in Georgia the better. The rate of unemployment is not
random, but rather is caused by certain predictors that affect Georgia’s unemployment.

What this project was about is to be able to predict Georgia’s unemployment while knowing the other predictors, and deciding what method is best used while predicting the unemployment rate in Georgia.

The predictors used in this project were each used due to one or more factors that could relate to unemployment. The predictors that used in this project are national unemployment, if there is a war going on, if there is a presidential or governor election, if the president, house, governor, or senate has a republican or democratic majority, the end of the year value for both the Dow and S&P 500. The national unemployment rate was used due to the fact that the Georgia unemployment rate is a subset of the national unemployment rate, and because of that they have a strong correlation. If there is a war going on is used because the economy can change into a heavier weapons and army supply depending on what the needs of the country are at that time, and that could lead to an influx of jobs and change unemployment in Georgia. If there is an election was used because the election campaigns could create jobs, or a politician may try to create some jobs so that they appear to have an impact on the unemployment rate in Georgia. The political party of the president, governor, house, and senate was used because depending on what political affiliation they are with could determine what type of physical policies they try to implement during their elected time. The end of the year value for the Dow and S&P 500 was used because depending on how the economy is could make an impact on the unemployment rate in Georgia. There is some overlap between some of these predictors, like the Dow and S&P 500, but they can affect the unemployment rate in Georgia differently depending on how many businesses in
Georgia are in each of the two stock market types, and thus it is necessary to include both of the two stock market types even though there may appear to be some overlay.

Result and Discussion

This project uses four main methods to use as predictors for Georgia’s unemployment. The first method is using linear regression. The next method discussed is using the best subsets regression. The third method used is using a decision tree to find the unemployment rate. The last method used was using a random forest method. Each of these methods have benefits and consequences using each one. The goal will be to find the method that has an outcome with the least error, and that method will be the most beneficial for using our predictors to determine the unemployment rate in Georgia. To find out what the error of each method is the software R was used. R is a system in which one is able to compute statistical computations and create different charts and graphs.

It is important to note that for binominal variables there is a 1 or a 0 instead of a number value. This is the case with quite a few predictors we used such as the political affiliation, if there is a war, or if there is an election year. For the case of if there is a war note that 0 indicates that there is not a war in that year while 1 indicates that there is a war during that year. It follows the same pattern for an election year that 0 means there is not an election while 1 means an election did take place during that year. The political affiliations are different in the case of it is not if an event occurred or not, but rather if they are affiliated with the Democrats or Republicans. In this case Democrats are marked as 1 while Republicans are marked as 0.
Part of the data is the stock market data, and that data is much larger in size than the rest of the data. The stock market is in the hundreds while the rest of the data is low numbers. To fix this normalize that stock market data, so it is easily comparable to the other data that is used. To normalize the stock market data create associated z-score values for each of the values in the stock market. To find the z-score for each value of the two stock market data the formula \( z = \frac{(X - \mu)}{\sigma} \) is used. \( X \) is the current value of the stock market that is being normalized, \( \mu \) is the mean of the stock market data, and \( \sigma \) is the standard deviation of the stock market data. This is used with each stock individually and not together. Such that the mean and standard deviation of Dow, and S&P will be two separate values. Observe that this does not change any of the values in a way that will alter how the data interacts with any of the data. Normalizing the data simply makes it easier to compare with the other data sets.

The goal of this project is to find out what method has the lowest error when attempting to predict the unemployment rate in Georgia. In order to compare the methods what needs to happen is training and test set of data needs to be created. A test set of data is a set of data that is set aside from the original set of data, so it is not used when fitting the models. Then once the data is separated and the method has been created, and is ready to test we compare the method to the test data to see how it compares by finding the mean squared error. The mean squared error is the expected value of correct output minus the original output squared. After all of our methods are finished our best method will be the method that has the least mean squared error.

The graphs that R generated, attached at the end of the project, gives information that makes it beneficial to change some of the aspects of the project. It is
important to notice is in Chart A where the graph of the natural log of Georgia’s unemployment being compared to the Dow produces a parabola shape, and for that reason a new predictor is considered, that is Dow squared. This is important because it appears to have a high relation with Georgia unemployment. The graphs will be referred to throughout this project.

The first method to predict Georgia’s unemployment used was linear regression. There are four main steps in order to find our linear regression model. The first step is to determine how many predictors are used to have the least error. Why this is possible to have a higher error with more predictors is because the predictors can be similar and throw off the data causing it to have a higher error than a linear regression model with less points. To accomplish this refer to Graph C where the critical points is mapped with the number of variables. It produces the fact that the lowest error is when we have five predictors. This means that when looking for variables we will take the five predictors that are most related when used together. The next step is to figure out these five predictors. To accomplish this a matrix is made with the natural log of Georgia’s unemployment and the data minus the training data. Next map the critical points and the different predictors to generate Graph D and observe that where five are shaded the predictors used are; federal rate of unemployment, the political affiliation of the senate majority, the political affiliation of Georgia’s governor, if the United States of America is involved in a war, and the last predictor is the square Dow. The next step using R the coefficients are generated giving the equation \( y = 0.14109928x_1 + 0.13020018x_2 + -0.32249755x_3 + -0.22908633x_4 + 0.0314239x_5 + 1.04354826 \). Where \( x_1 \) is the federal unemployment rate, \( x_2 \) is the senate majority political affiliation, \( x_3 \) is the political
affiliation of Georgia’s governor, \( x_4 \) is if the United States is involved in a war, and \( x_5 \) is the square Dow. The final step is to find the error of the linear regression model. Using the mean squared error we find that the error when using the linear regression model is 0.01151549. This error is low enough to the point that this method is quite useful in determining the unemployment rate of Georgia.

The second method used in this project to predict Georgia’s unemployment is best subsets regression. The cp plot is given on Chart C. Using the main predictors and interactions of predictors create the fit model. The examples that were tried while doing this project were the federal rate of unemployment + senate majorities’ political affiliation + Georgia’s governor’s political affiliation + if the United States is at war. While doing a wide verity of other best subsets fit models it occurs that they all have about the same standard error as each other. Since they all have roughly the same standard error we take the simplest one that is, federal rate of unemployment + senate majorities’ political affiliation + Georgia’s governor’s political affiliation + if the United States is at war. Now that the optimal fit has been identified using the fit function in R calculate the mean squared error. The mean squared error ends up to be 1.530059. Note that the mean squared error is much worse than the linear regression method above.

The last two methods are similar in the fact that one is just a more complicated version of each other. The first method we will talk about is decision trees. To create a decision tree in R the data is put into R and program it to output a decision tree with the data. Refer to Image A, and note the structure of the tree. The decision tree is made in such a way that depending on if the condition is met you follow that path of the decision tree. An example is if the federal rate is 6.5 then according to the decision tree the
6.5<7.05 then move down and left, then the next tier states the rate 6.5>6.05 hence we move down and to the right since the statement given is false. Thus the predicted state rate is 5.7. The error of the decision tree method is 1.070883.

The last method is the random forest method. It is a method that is made up of random trees looking at Image B we can notice how many trees produces the least error. To find the one with the least error compare the out of bag error and the test set error and find the one that has the least difference. The point at which the difference of errors is the least is at nine. Thus the optimal amount of trees is nine trees. Looking at the mean squared error we find that the error is 0.9835141. The error is the second best out of all of the methods we have looked at, but by far the best method is still using our linear regression model.

Throughout this project and all of the methods that were done it holds true that linear regression computed is the best method in using our predictors to determine Georgia’s unemployment. This is significant because of how much lower the error is compared to the other methods. The use of the different methods are important to make sure that the optimal method is being used for the data. The function used for the linear regression is \( y = 0.14109928x_1 + 0.13020018x_2 + -0.32249755x_3 + -0.22908633x_4 + 0.0314239x_5 + 1.04354826 \). Where \( x_1 \) is the federal unemployment rate, \( x_2 \) is the senate majority political affiliation, \( x_3 \) is the political affiliation of Georgia’s governor, \( x_4 \) is if the United States is involved in a ward, and \( x_5 \) is the square Dow. The mean squared error is 0.01151549 for the linear regression model, which is the best error found throughout the project. The random forest model using nine trees had a mean squared error of 0.9835141 which is the second lowest error found. The decision tree model had
a mean squared error of 1.070883 which is the third best error based upon the models. Lastly the model based upon best subsets using federal rate of unemployment + senate majorities’ political affiliation + Georgia’s governor’s political affiliation + if the United States is at war had an error of 1.530059 which is the worst error of them all.
Graph C

Graph D
Chart C
References


