MATH 5300 HW1

Chi Ma

Feb 8, 2018

**EXECUTIVE SUMMARY**

In this assignment, I am going to forecast the (CPI-U) with the historical data that can be accessed from BLS website by using both traditional and machine learning techniques. .

I will use multivariate linear regression model to forecast the CPI-U(%) numbers and CART model to predict whether the inflation will go up or down. Since I have big dataset from BLS and correct model selection, I believe both my models could give me desirable answers

For multivariate linear regression model, it will follow a 12 months previous/1 forecast and have maximum 5 independent variables. The CART model will also have no more than 5 predictors.

Only unadjust data from 2012-2017 dataset are selected. I will first retrieve, clean and store BLS data for both models, then create necessary dataset to run both models.

For multivariate linear regression, I will use ordinary least square method which is the one I feel most familiar with and classification tree for CART.

For output, the MSE for the multivariate linear regression model is 2.41. Due to the large magnitude of variable scale(usually over 240), MSE is low enough to reach my goal. The accuracy rate of the decision tree is about 66.1 percent, which is not an ideal output.

**METHODOLOGY/TECHNICAL APPENDIX**

For data management, all the CPI-U unadjust data from 2012-2017 is downloaded form BLS website. In data clean stage, Variables that only with an indent level of 1 and effect score over 0.1 are selected, which result in 11 variables and 72 data points each. Since the index are way larger than index changes, percentage change method rather than raw data is used to run models. Following with the 12 month previous/1 month predict rule, the dataset is left with 60 data points each.

*Example: first data point of independent variable is from Jan 2013, which is calculated based on the formula (Dec 2012 - Jan 2012) / Jan 2012 )*

Since OLS is implemented, no training and testing data separation will be performed. By running the 11 independent variables in preliminary test, 5 variables are selected and 6 are deleted. The five independent variables are Housing, Education and Communication, Energy, Recreation and Durables. The multivariate linear regression is runned with dependent variable y equals to CPI-U data (All.items) and independent variables x in percentage change fashion. All five independent variables are significant after inspecting the summary table, the OLS model also have a 0.89 R-squared. For the MSE, 2.41 is quite low consider all the CPI-U (All.items) data points are around 230 to 250. Five fold cross validation is performed and the cross validation predicted value graph is attached in the HTML and Rmd files.

For CART model, classification method is used since we want to predict the direction of inflation rate (CPI-U All.items data). Additional column “Direction” is added to the dataset and all the

rows are marked with proper direction values (“Up” and “Down”). By using package “rpart”, decision tree has been made and pruned with proper parameters(cp). Ten fold cross validation is performed and the results are both posted and graphed in the R files. A confusion matrix is also made in the end. The accuracy rate is only 61% at first, after the model is pruned, it reaches 66% which is still not a good result for my prediction.

All other details can be found in rmd files.