Data Analysis

Group 4: Kaori Hirano, Alicia Nguyen, James Xia

2023-07-20

Methods Overview

1) Ridge & Lasso Regression: We are using lasso and ridge regression to choose the best predictors. Lasso will eliminate non-influential variables, while ridge will shrink them to a smaller value. These models will allow us to compare the influence of different variables on the outcome. Our tuning parameters will be chosen by using cross validation on the lambda.
2) Single Decision Tree: This method will identify the most important predictors in a different way than ridge and lasso regression will by showing us a clearer ranking and relationship between variables. Variables at the top will be more important and the path of nodes a decision takes will show different relationships between variables at different values. The number of terminal nodes will be chosen by cross validation.

Application

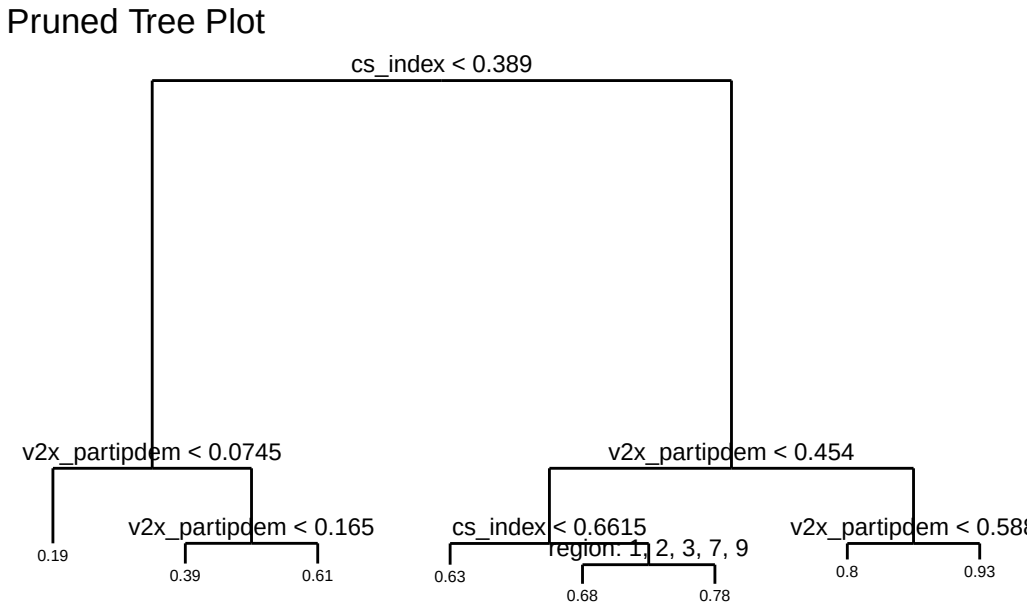
```
cs_full <- readRDS("/cloud/project/data/civil_society_factor")
# get only 2019 and remove civil war and coup because there are none in 2019
cs <- cs_full %>% subset(year == 2019) %>% select(-one_of("civil_war", "coup")) %>% drop_n
# 70, 30 test train split
set.seed(145)
train <- sample(c(TRUE, FALSE), nrow(cs), replace = TRUE, prob=c(.7,.3))
test <- (!train)
val <- test
```

```
# looks for region with most observations to set as base level table(cs$region)
cs$region <- relevel(cs$region, ref = 4)
#Sub Saharan Africa, was used as the baseline because it has the most observations.
```

```r
# create x and y for glmnet
set.seed(129)
x <- model.matrix(cspart ~ csrepress+v2x_partipdem+edu+corr+cs_index+social_support+
                    choices+gen+region, data = cs)[, -1]
y <- cs$cspart
# do cross validation
cv_r <- cv.glmnet(x[train,], y[train], alpha = 0, lambda = 10^seq(10, -2, length = 100))
# saving optimal lambda
bestlam_r <- cv_r$lambda.min
# calculating MSE
ridge_pred <- predict(cv_r, s = bestlam_r,
newx = x[test, ])
ridge_mse <- mean((ridge_pred - y[test])^2)
# find how to fit ridge model and import here
ridge_mod <- glmnet(x, y, alpha = 0, lambda = bestlam_r)
coef_r <-coef(ridge_mod)
```

```r
set.seed(18)
# do cross validation
cv_l <- cv.glmnet(x[train,], y[train], alpha = 1,
lambda = 10^seq(10, -2, length = 100))
# saving optimal lambda
bestlam_l <- cv_l$lambda.min
# calculating MSE
lasso_pred <- predict(cv_l, s = bestlam_l,
newx = x[test, ])
lasso_mse <- mean((lasso_pred - y[test])^2)
# coefficients that matter
lasso_mod <- glmnet(x, y, lambda = bestlam_l)
coef_l <-coef(lasso_mod)
```

```r
set.seed(247) # makes tree with training data
tree_train <- tree(cspart ~ . -year -country_name, cs,
    subset = train)
cv_train <- cv.tree(tree_train)
# uses 8 because that's where it levels off on the graph
prune_train <- prune.tree(tree_train, best = 8)
plot_tree(prune_train) +
  labs(title = "Pruned Tree Plot")
```

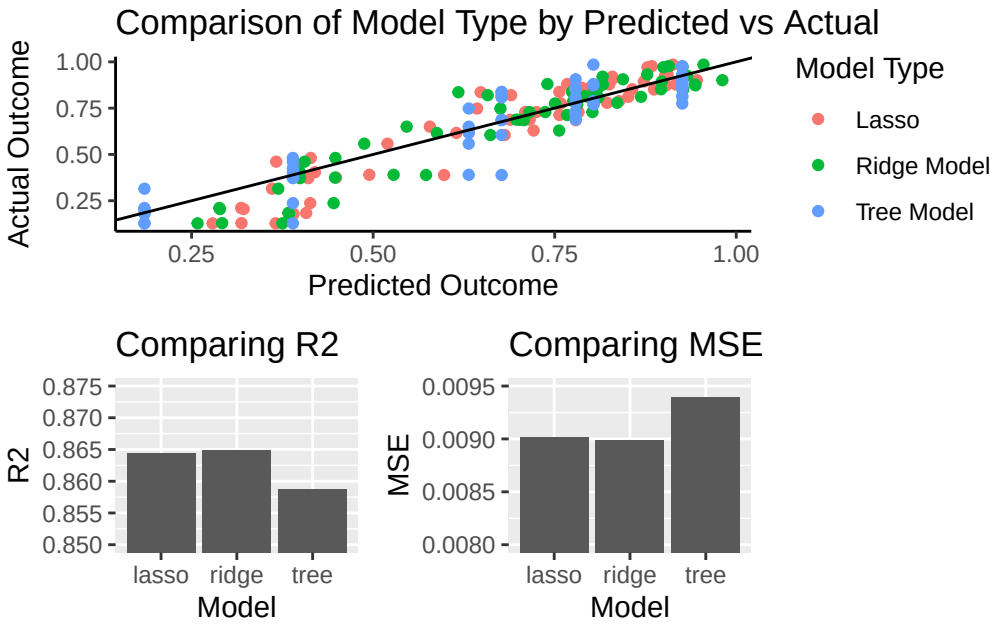



```
tree_pred_tuned <- predict(prune_train, cs[test,],
    type = "vector")
y_test <- y[test]
# gets mse for pruned (which is same as full tree)
mse <- mean((tree_pred_tuned - y_test)^2)
```

Visualizing

```
# putting together data of predicted, actual, and model type
dataplot <- data.frame(true_value = c(y[test], y[test], y[test]))
dataplot$model_type <- c(rep("Lasso", length(lasso_pred)), rep("Ridge Model",
            length(ridge_pred)), rep("Tree Model", length(tree_pred_tuned)))
dataplot$predictions <- c(lasso_pred, ridge_pred, tree_pred_tuned)
#plotting predicted vs actual by model type
compare <- ggplot(dataplot, aes(x = predictions, y = true_value, color = model_type)) +
  geom_point() + geom_abline(intercept = 0, slope = 1) +
  labs(x = "Predicted Outcome", y = "Actual Outcome",
       title = 'Comparison of Model Type by Predicted vs Actual',
       color = 'Model Type') + theme_classic()
# function giving us R2
r2 <- function(predicted, y) {
  #find SST and SSE
```

```
  sst <- sum((y - mean(y))^2)
  sse <- sum((predicted - y)^2)
  #find R-Squared
  rsq <- 1 - sse/sst }
# setting up values for graph
name=c("tree","lasso","ridge")
mse_all=c(mse, lasso_mse, ridge_mse)
value=c(r2(tree_pred_tuned, y[test]), r2(lasso_pred, y[test]), r2(ridge_pred, y[test]))
compare_data=tibble(name,mse_all,value)
p1=ggplot(compare_data, aes(x=name, y=value))+
  geom_col()+ coord_cartesian(ylim=c(0.85,0.875))+
  labs(x="Model",y="R2",title = "Comparing R2")
p2=ggplot(compare_data, aes(x=name, y=mse_all))+
  geom_col()+ coord_cartesian(ylim=c(0.0080,0.0095))+
  labs(x="Model",y="MSE",title = "Comparing MSE")
compare / (p1+p2)
```



These bar graphs help visualize the differences in MSE and R2 between models. The scatterplot shows the differences in predicted vs actual values by model by the distance to the line.

Discussion

From our regression models, we learned the most important predictors when predicting civil s

From the decision tree, we are able to see a breakdown of the most important predictors in civil society participation in 2019. The number of nodes as 8 was chosen by cross validation. The most important predictors according to the tree model are civil society index, participation in democracy, and after that region the county is located in/around, specifically being in regions 1,2,3,7, and 9. These initial modelling attempts tell us that the most important predictors of civil society participation in 2019 are civil society index, region, and participation in democracy, which were seen across all three models, and social support which was also selected as important in the regression models only. The model that did the best was ridge regression, with an MSE of .0089, followed by lasso regression, MSE .0090, then the tree with an MSE of .0093. Lasso and ridge regression had the highest R2 values with about 86% of the change in civil society participation being explained by the models (ridge = 86.8%, lasso = 86.7%), while the decision tree explained 85.8%. All of these methods had very similar MSEs and R2s, with the lasso and ridge regression models slightly outperforming the tree method on both of the metrics. These findings mostly fit with our hypothesis in that we expected regional variation and civil society index to have an influential relationship with civil society participation, while we were suprised by how important participation in democracy was and that social support had a negative relationship.