Empirical Project 3 | ECON 270

Group Members:

Shalin Luitel

Sweta Nanadakumar

Jahnavi Kalyan

Arnold Gyateng

Part 3: Out-of-sample validation

13. Load in the proj4_results.dta provided to you. Keep only the variables geoid kfr_pooled_p25 test training predictions_ols predictions_tree.

14. Merge the test dataset using the geoid variable.

15. Calculate the mean squared error for predictions_ols and predictions_tree out-of-sample. (Hint: Refer Table 2)

16. Which model did the best?

Bonus: Draw some graphs or maps to visualize your predictions.

Submit your answer to these questions and the dofile, by 11:59 p.m. on March 13th to receive credit.

Which model did the best?

In the summary table below, we see that among 2518 observations in OLS and 989 Decision Tree results, the mean predicted error is below in Decision Tree (0.000251) than in OLS regression (0.000873). The mean error in OLS regression is 2.47 times higher than the Decision Tree. This gives us high confidence to use the Decision forest model on our prediction.

```
. sum pred_error
```

Variable	Obs	Mean	Std. Dev.	Min	Max
pred_error	2,518	.0008732	.0015588	3.57e-11	.0296291

```
. sum pred_error1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
pred_error1	989	.0002514	.0009984	3.06e-11	.0279501