Kathryn Wantlin
Professor Chetty
Econ 1152
May 1, 2019
Project 4

From Trees to Forests: Assessing the Effectiveness of Predictive Models

In this study, we again see an are approaching the issue of mobility. With data we have already collected, we are able to ascertain certain patterns in the data and can verify their validity using other similarly collected samples, looking to see if fit remains strong out-of-sample. With such patterns and models, it would then be possible for us to ascertain factors with predictive power and forecast future mobility outcomes for people. This would be extremely for researchers, and there are a variety of ways to go about crafting the models. The ways we will consider are made through a multivariate linear regression, decision trees, and a random forest. Some of these methods may be more clean fits for our given data, while others may be stronger able to adapt out-of-sample. Here, we will investigate these results using a set of predictors ascertained from census data to consider how strongly and well predictive models based on them function compared to "actual" data values.

We did all our analysis at the county level and focused on a set of 10 predictive count variables: foreign born, owned homes, veterans, completed high school, born in state, married, below poverty, female, two or more races, and income above $75,000. After rescaling the counts into rates, we get the following summary statistics and regression output when each is regressed with krf_pooled_p25.

Variable	Obs	Mean	Std. Dev.	Min	Max
foreignborn	1,259	.0491593	.059473	.0005855	.511422
veteran	1,259	.0779055	.0222443	.0126233	.1841397
hscompletion	1,259	.1949153	.0470982	.059508	.3367716
borninst	1,226	.6760978	.1416899	.1861592	.9319226
married	1,226	.4165317	.0514083	.211188	.5929132
belowpoverty	1,259	.1707352	.0778767	.0383638	.6044627
female	1,259	.5023253	.0211075	.336619	.5550877
twoormorerac	1,259	.024386	.0353413	7.06e-15	.5771152
owneroccup~d	1,259	.6018381	.0845995	.179935	.8429752
inc_75plus	1,226	.0590448	.0323233	.010567	.2450678
kfr_poole~25	1,259	.4132139	.0517072	.2128646	.6140298

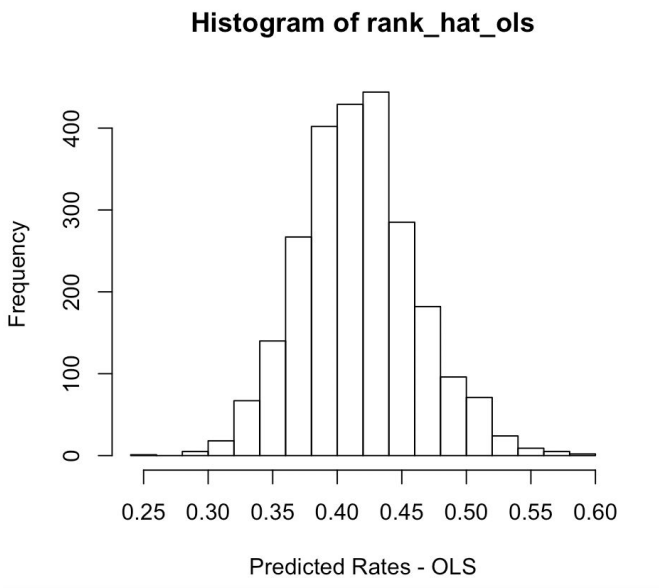
Linear regression

Number of obs	1,226
F(10, 1215)	83.04
Prob > F	0.0000
R-squared	0.4291
Root MSE	.03691

kfr_pooled~25	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
foreignborn	.0853427	.0292669	2.92	0.004	.0279234	.142762
veteran	-.5478925	.0815504	-6.72	0.000	-.7078877	-.3878973
hscompletion	-.0032334	.0321838	-0.10	0.920	-.0663754	.0599086
borninst	.0348121	.0111389	3.13	0.002	.0129584	.0566658
married	.4248638	.0297939	14.26	0.000	.3664106	.4833171
belowpoverty	-.3290586	.0341207	-9.64	0.000	-.3960007	-.2621165
female	-.2884339	.0586633	-4.92	0.000	-.4035265	-.1733414
twoormorerac	.079832	.0415643	1.92	0.055	-.0017137	.1613777
owneroccupied	-.051057	.0183833	-2.78	0.006	-.0871236	-.0149905
inc_75plus	-.0469922	.0614504	-0.76	0.445	-.1675529	.0735685
_cons	.4854351	.0398172	12.19	0.000	.4073169	.5635533

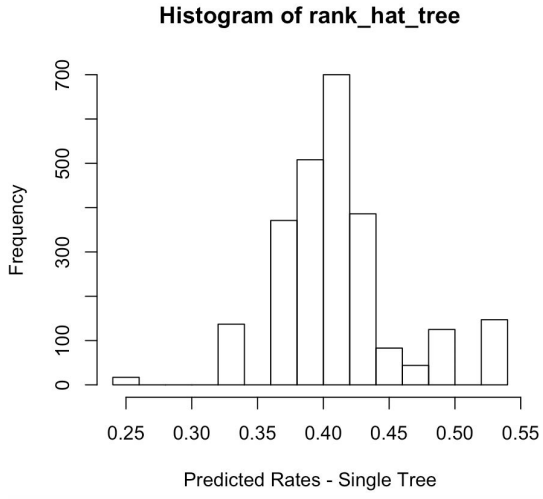
The regression demonstrates that the strongest relationships hold between krf_pooled_p25 and the following variables: veteran, married, belowpoverty, and female. For example, having more families below poverty was negatively correlated (coefficient of -.3290586) with income mobility, perhaps since more affluent communities provide some aggregate effect or influence on children and also reflect the quality or even diversity of opportunities and exposure afforded a child in that county. Our linear regression also predicts krf_pooled_p25 in-sample quite well, since we see that the average prediction error is only -.0000953 and the standard deviation is .0190741, both quite low though for these percentage rate predictors.

We can also run a linear regression of krf_pooled_p25 on the full predictor set at once, giving us some coefficients. We can, for example, interpret that since the coefficient for rate of high school completion is .1316, that for each one-unit difference in the predictive variable rate of high school completion, the predicted value of krf_pooled_p25 increases by .1316, fairly substantial here. We also get predictions of krf_pooled_p25 from this, distributed as in the below histogram. The average predicted krf_pooled_p25 is .4166, and we can see other common predictions from the distribution.

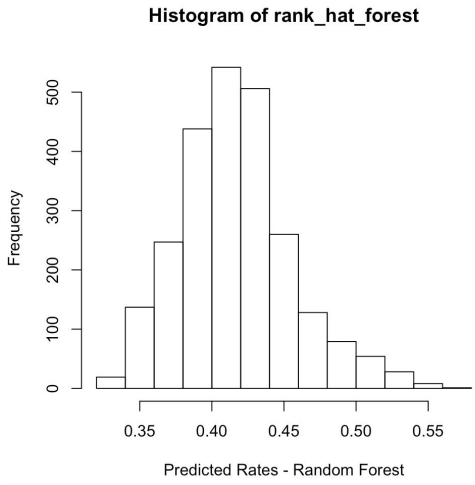


Another method we could use is to implement a decision tree on the full predictor set using 10 fold cross-validation to select the optimal tree size. The first split occurs at P_57>=14.75, which means that the first split occurs where the percent of Adults That Report Fair or Poor Health (Persons 18 Years and Over) is above or equal to 14.75%, and the branch that is indeed above or equal to this value is further analyzed in successive steps. The first split is often an important predictor or correlate of the outcome. The first split is always the most important split, because if there is one very strong predictor in the data, most of the random samples will use this predictor in the top split. That is, at every branch, the best split that minimizes error and fits best is chosen, and the first split is the split where we consider the entire body of data. Intuitively, if we chose a suboptimal split, every split after that would be theoretically suboptimal and could be improved upon. Also, if we put everything into its own region or “leaf”, then the in-sample prediction error would be 0; however, this predictive model would do very poorly out-of-sample, so instead note here that we use cross-validation to select a

smaller tree instead of just using as many splits as possible, which would have over-fit the model to our in-sample. Below we see a histogram of the distribution of predictions of kfr_pooled_p25 from this decision tree. Note that the average here for kfr_pooled_p25 is .4143.



Finally, we implement a random forest with at least 1000 bootstrap samples and obtain predictions. The histogram distribution is again below, and the mean predicted kfr_pooled is .4166.

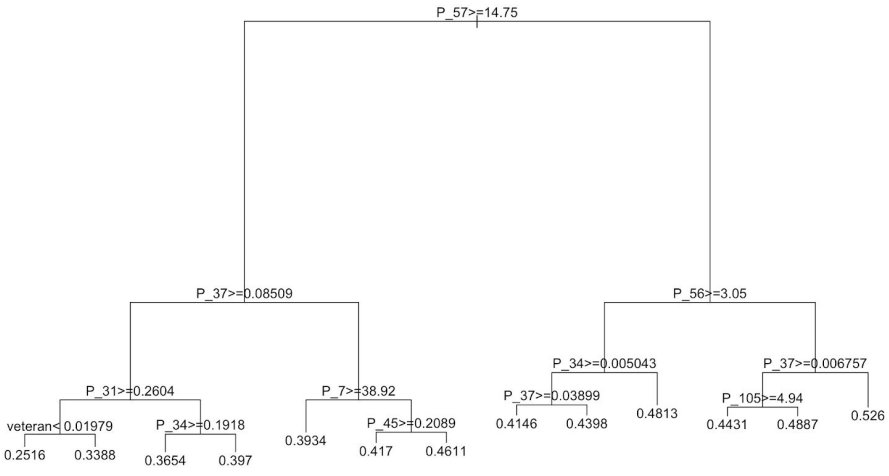


The the mean squared error for my results on for the linear, tree, and forest fits in -sample are .0002938, .0007504, and .0001973, meaning the forest was the best fit and the tree was the worst, because though the differences were quite small, the mean squared error for the forest was the smallest, meaning its predictions were closest to the actual kfr_pooled_p25. I would predict due that since the trees in the forest are uncorrelated with each other, it will be better than the decision tree out-of-sample and also better than the linear fit out-of-sample since it splits the data into groups that are increasingly similar to each other, and then attempts to reduce error at each branch.

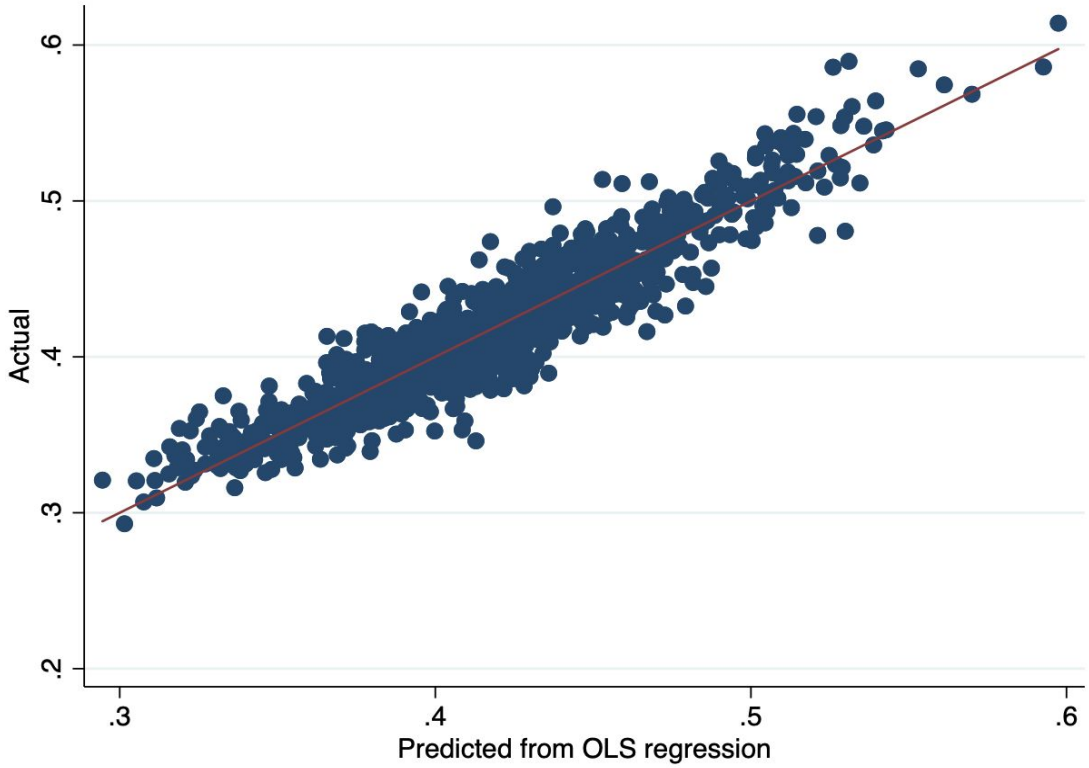
The the mean squared error for my results on for the linear, tree, and forest fits in -sample are .0004307, .0010045, and .000245, meaning the forest was the best fit and the tree was the worst, because though the differences were quite small, the mean squared error for the forest was the smallest, meaning its predictions were closest to the actual kfr_pooled_p25.

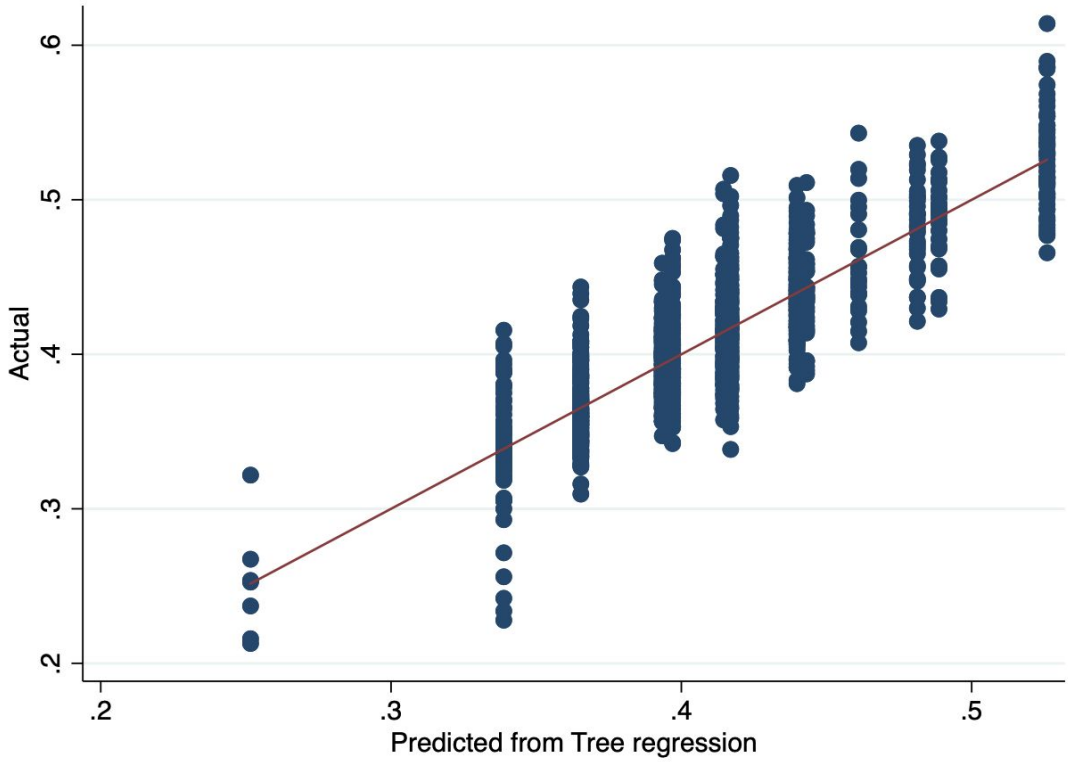
Summary

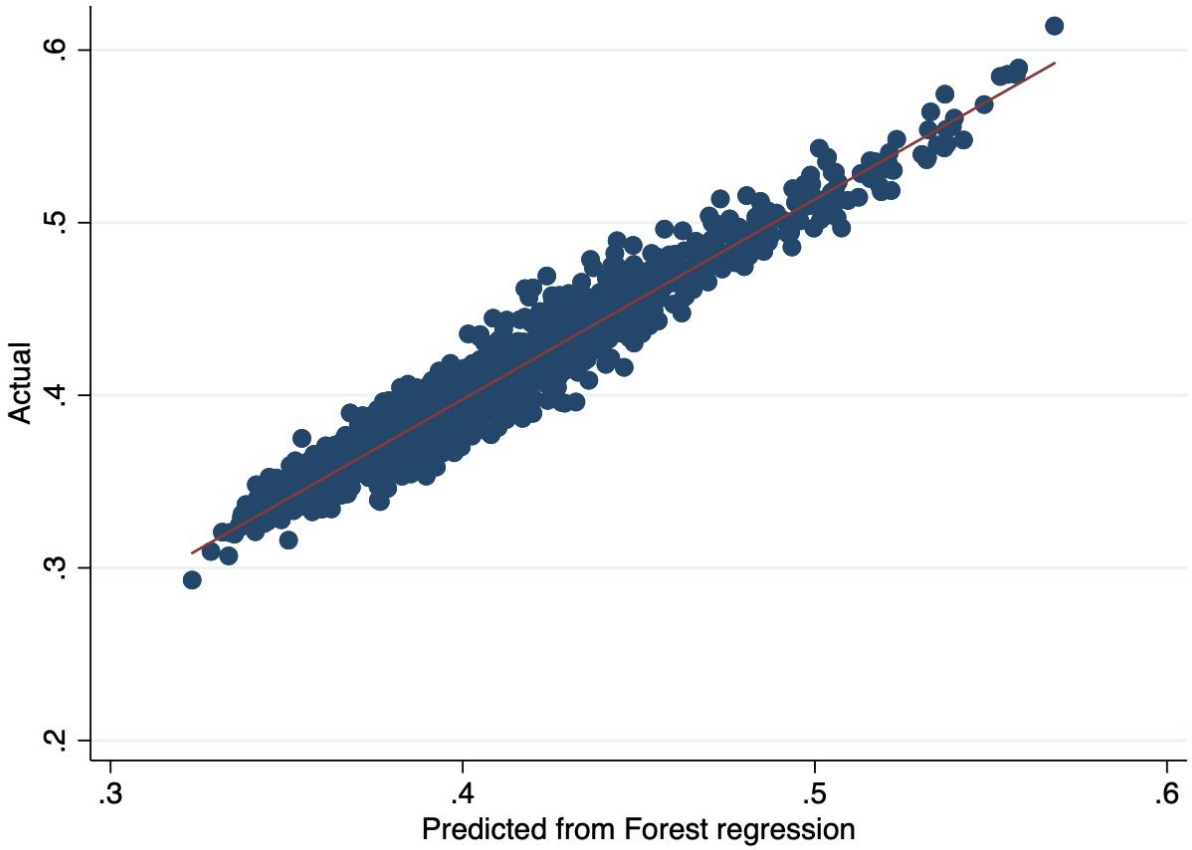
In this study, we considered multivariate linear regression, decision trees, and a random forest for predicting data. Some of these methods may be more clean fits for our given data, while others may be stronger able to adapt out-of-sample. We investigated the results using a set of predictors to consider how well predictive models based on them predicted mobility based on adult income compared to actual data values of this metric. Some metrics were more strongly correlated with this kfr_pooled_p25; for example, having more families below poverty was negatively correlated (coefficient of -.3290586) with income mobility. Changes in each predictor also had varying impacts on the predicted kfr_pooled_p25 values. For example, for each one-unit difference in the predictive variable rate of high school completion, the predicted value of krf_pooled_p25 increases by .1316. We then proceeded to run our models, and to visualize the decision tree, we can refer to the below diagram.



Overall though, we found that the random forest was the best model, both in and our sample since the error was most minimized for it. Out of sample, we can compare the predictions with actuals scatterplots, seeing that the forest is indeed the best since its best fit line most closely approximates 45 degree angle line (y=x). (See diagrams on next page.)


Actual
.2
.3
.4
.5
.6
.3
.4
.5
.6
Predicted from OLS regression


Actual
.2
.3
.4
.5
.6
.2
.3
.4
.5
Predicted from Tree regression


Actual
.2
.3
.4
.5
.6
.3
.4
.5
.6
Predicted from Forest regression