Predicting the Success of American Counties

April 28th, 2019

Ruilian Xie and Nicholas Lourme

1 Project Overview

Our project seeks to predict the next successful American urban area. From the many variables that describe the quality of American life, we are seeking to identify those that are most predictive of success. In an era of rapid change characterized by automation, gentrification and the stagnation of wages and productivity, our research can help policymakers better distinguish important contributors to success from weaker ones. Our project can be considered a success if one of two criteria is met: either our model predict with greater than 50% accuracy whether a county will meet our definition of success, or we discover that a vraible is an unexpectedly large contributor to our definition of success. If the first criteria is met, then we will have created a model that can be useful for understanding how a county can grow in an equitable way. If the second criteria is met, we will have a further basis for studying how that particular variable contributes to the equitable growth of a county. The key to our research question is creating a narrow definition of "success". A cursory search of the literature reveals practically as many definitions of a successful city as there are cities. Some colloquially "successful" cities are characterized by low unemployment; others by high median wages or a strong concentrations of technology jobs. Still others are notable for their diversity and culture, for anchoring a large company, or even just for their nice weather. The point is that the most notable feature of a city, the feature that marks its as successful, changes depending on the city. Our project seeks to identify the factors common to each successful city in order to help policy planners better prioritize policy initiatives.

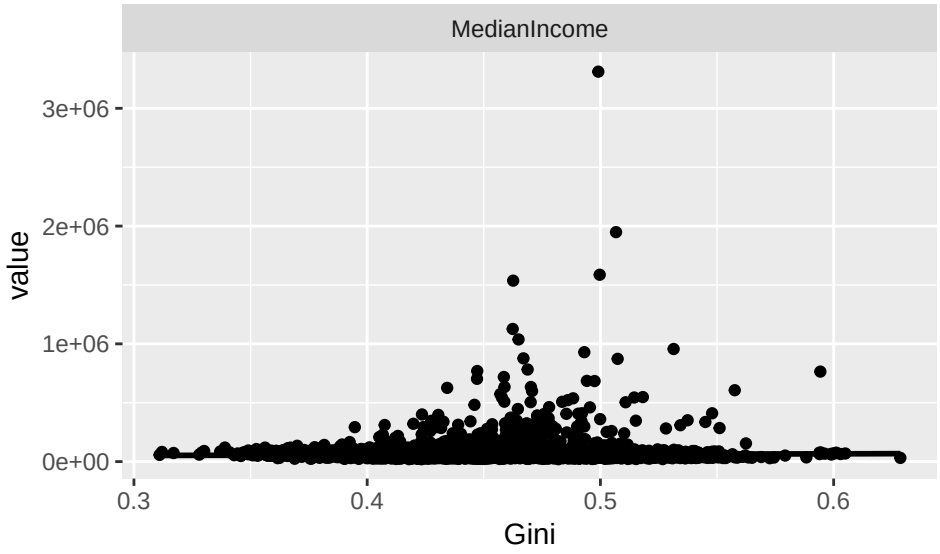
2 Methods and Tools

2.1 Methods and Assumptions:

We set up our project by pulling in variables from the United States census bureau. Those variables were: - Unemployment Rate - Median Income - Poverty Rate - Housing Costs - Mortgage Costs - Bachelor Rate - Population - Fertility We then computed the year-over-year change for each of those variables to create a new variable, the year-over-year percentage change for each measure. We had initially hoped to subset our dataset by metro area, but this reduced the number of datapoints available to us. We instead opted to divide our dataset by US county. We had also hoped to include average commute time, but this variable was heavily skewed by whether large metro areas lay within a single county or were divided into several. For example, the city of San Francisco, CA lies entirely within San Francisco county, while the city of Atlanta, GA is divided among four different counties. We thus ended up dropping the commute time variable. We ended up with a dataset that included 16 predictor variables for every US county and covered the years 2006 to 2017. We accounted for misisng values in the dataset by imputing the average for each county according to the Year and State. In other words, if Calhoun County, Alabama was missing a value in 2006 for fertility, we imputed the mean fertility for Alabama in 2006 for the missing value.

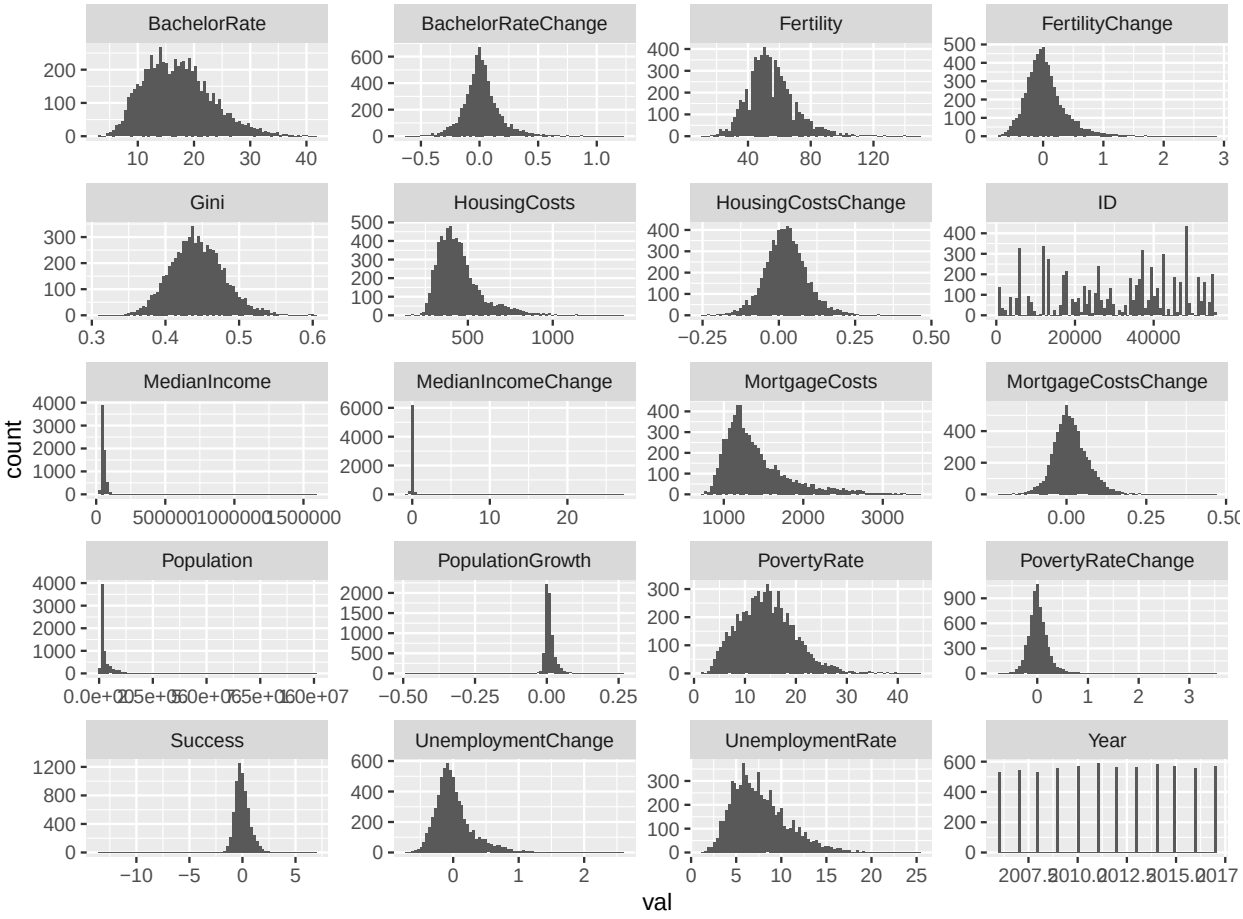
We defined a successful urban area as a place that combines population growth with a low gini coefficient. The gini coefficient is a measure of income inequality that ranges from 0 (most equal i.e. everyone has the same income) to 1 (least equal i.e. 1 person controls all the income). We focus on those two variables as markers of a successful area because we believe they capture the most important attributes of an urban area within the fewest variables. This also allows us to use other variables that could have been part of marking a successful city, such as median wage growth, crime rates etc. as predictors rather than outcomes. Sustained population growth is a proxy for, among other factors, the desirability of a particular city.

Simply put, we assume that people tend to move to more desirable areas. We use the gini coefficient as a way to capture some of the negative factors behind a growing population. For instance, gentrification of an area is marked by both an increase in population and an increase in the gini coefficient; screening out high gini values helps control for this. Another example is the sudden increase in population resulting from displacement due to natural disaster. The gini coefficient should also increase as the displaced have a lower income than those who previously lived in the area. The relationship between the gini coefficient and median income is shown below, where the majority of median income data points lie between .4 and .5 on the gini index.



2.2 Tools:

Our principal approach to this problem was to apply three different predictive models and see which of those yielded the most precise estimate. Once the basic preprocessing had been completed, we partitioned our data, with 70% of the data used for training and 30% for testing. We then normalized our data to prepare it for the classification techniques. Below are plots of the data pre- and post- normalization.



We note that the variables "BachelorRate", "Fertility", "FertilityChange", "HousingCosts", "MedianIncome", "MedianIncomeChange", "MortgageCosts", "Population", "PovertyRate", "PovertyRateChange", "UnemploymentRate" and "UnemploymentRateChange" are skewed. To deal with the skew, we will take the log of each.

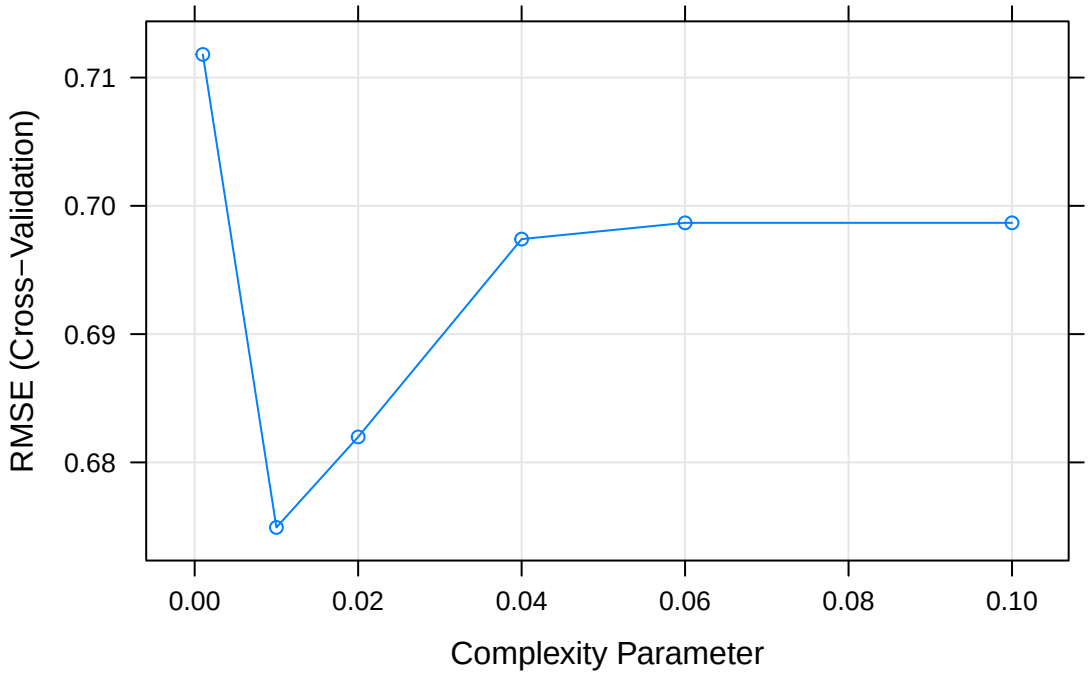
The three predictive models we chose to use were linear regression, regression tree and random forest. We partitioned the training data into 5 folds and ran each of the three models on the data. The packages we used for preprocessing the data included recipes and tidyverse, which were required to manipulate the data into a format that could be fed into the selected ML algorithms. We used the caret wrapper and ranger packages for the selected algorithms we chose to run. We chose those three models because this problem strove primarily to identify the main causes of a successful county. The random forest model was most useful for showing

the individual variables best correlated with success, while the regression tree and linear regression models gave us a sense of how well all of our variables jointly and individually correlated with success.

3 Results

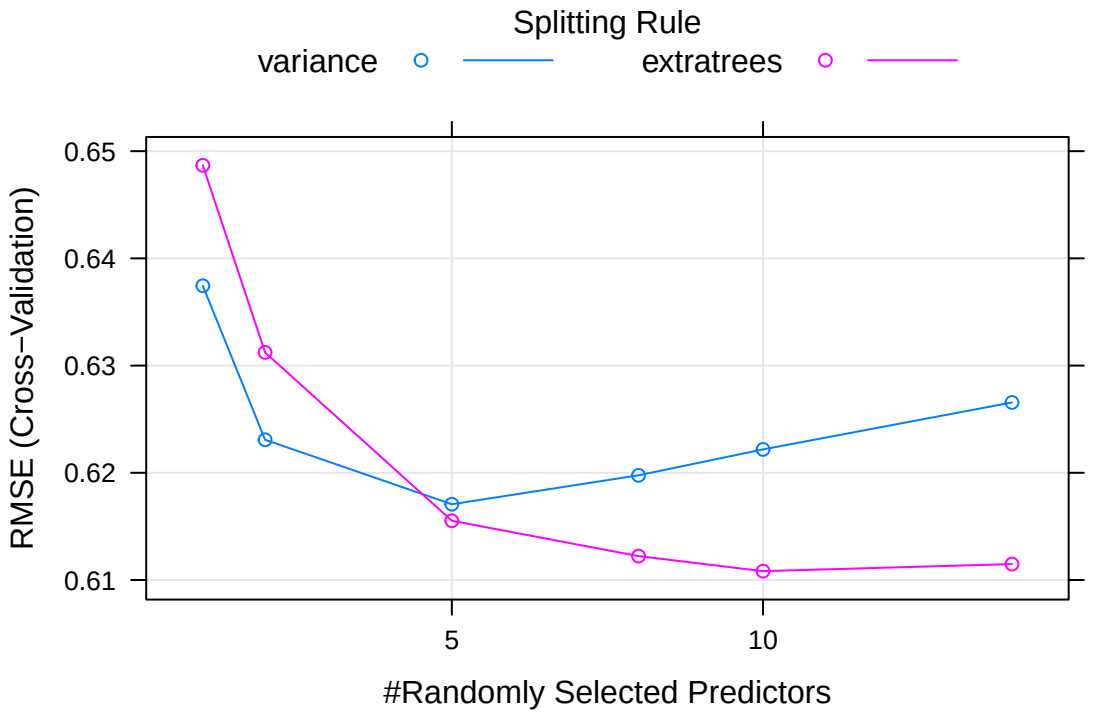
The results of our three models are below. For the linear regression, we can see that the model itself is significant ($p<.05$) and that it explains ~ 31.8% of the deviation. However, two variables, Unemployment Rate and Median Income, along with their respective lagged variables (Unemployment Rate Change and Median Income Change) are not significant ($P>.05$). This is unfortunate as we believed at the outset that those variables would have a strong relationship with our success variable.

The next model we ran was the regression tree model. We ran it on a tuning grid ranging from .001 to .1. The metric we used to evaluate the model was the Root Mean-Squared Error (RMSE). Unsurprisingly, the deepest tree was the most precise. However, even at that depth the model only yielded an adjusted R-squared of ~20%, which was less than the linear regression model.

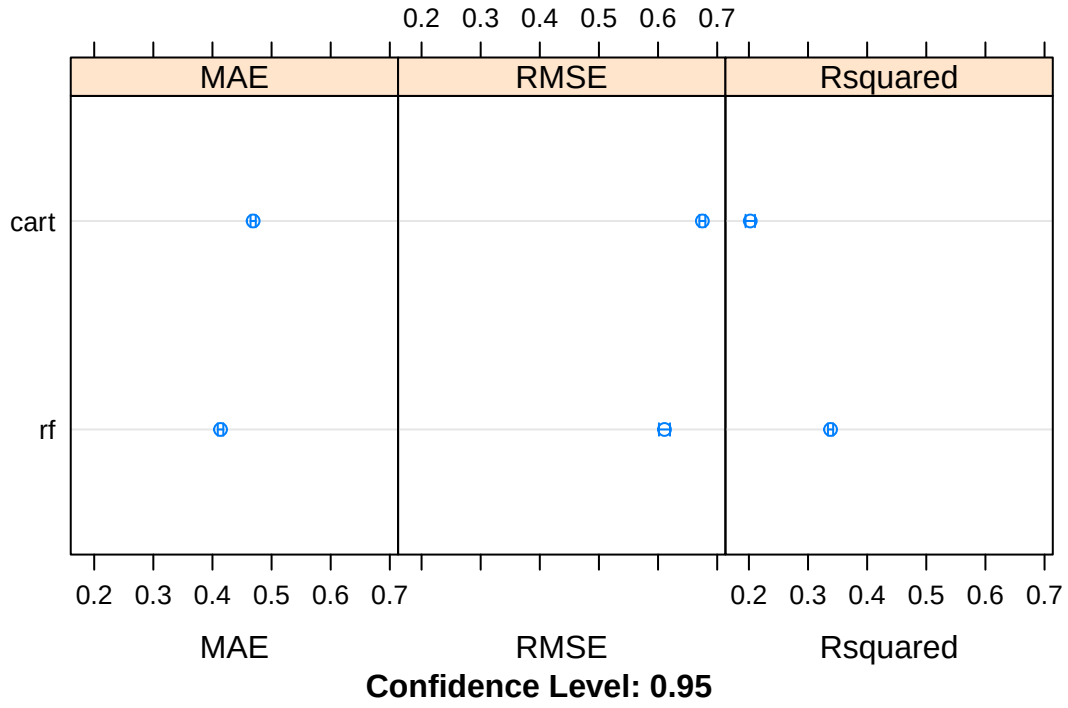


The third model we ran was the random forest model. This is computationally the most expensive model, but we expected it to return the most accurate predictor or success. As with the regression tree, we ran the model with RMSE as the accuracy metric.

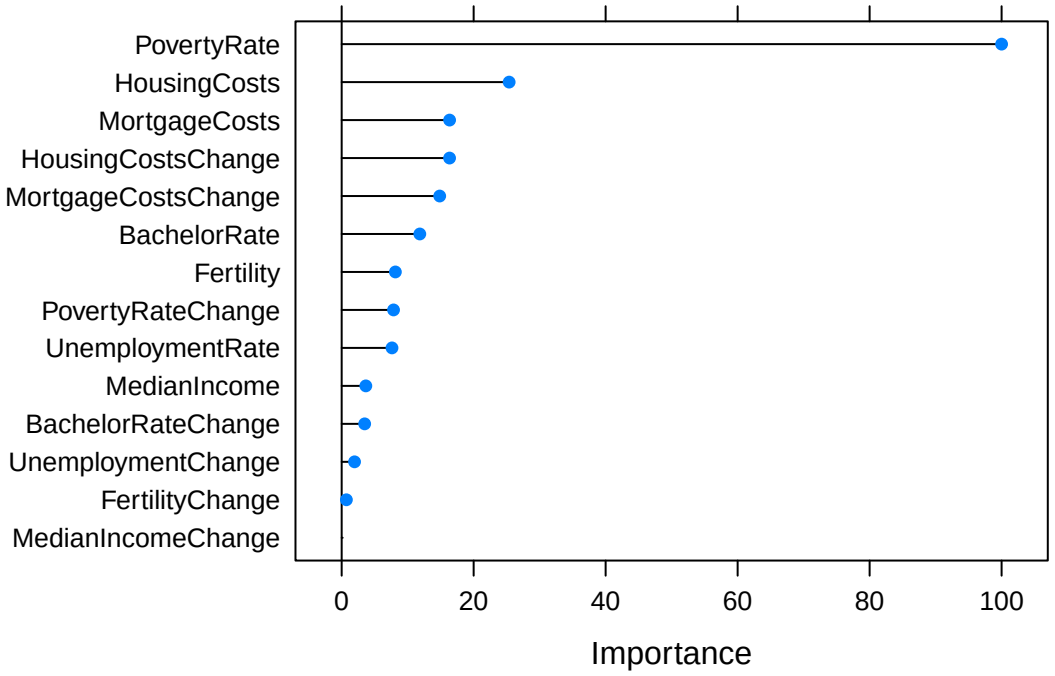
The most accurate set of tuning parameters were a 10-fold mtry, a splitrule = extratrees and a minimum node size = 1. Colloquially, this means that 10 variables were randomly sampled at each split, which themselves were classified as extratrees (usually the case for a regression problem). This yielded a RMSE of .611 and an adjusted R-squared of ~33.78%, which was more accurate than either the linear regression or regression tree models.



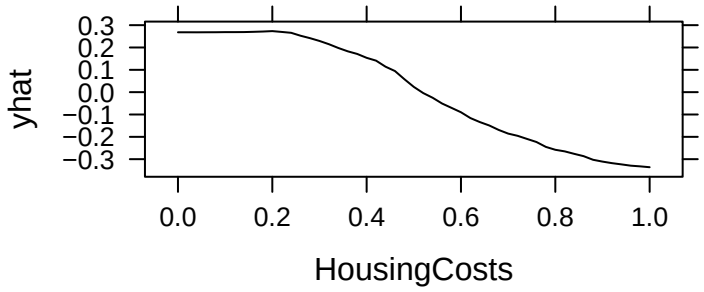
When we plot the Random Forest vs. the Regression Tree models against each other, we can see that the RMSE is better for the Random Forest model, and that the Random Forest explains more of the deviation. Thus, we will use the Random Forest model as the one to explore the various variable importance weights.

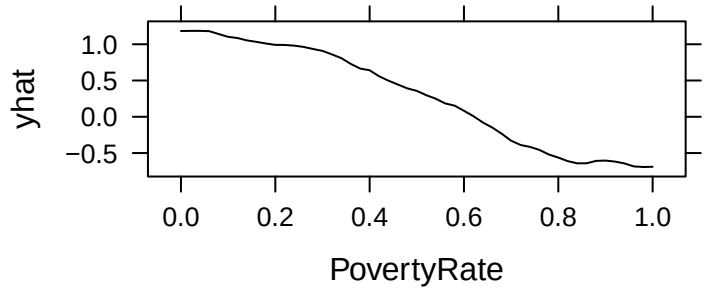


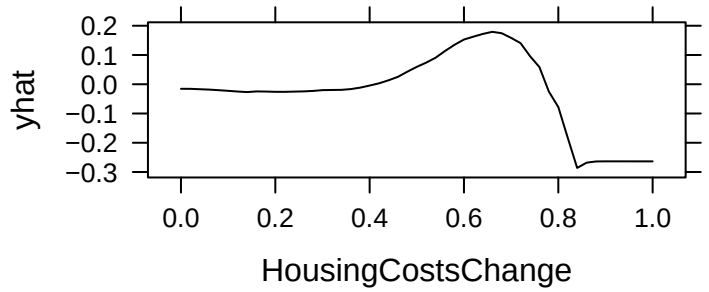
Below are the weights for the variable importance to the success of the models. We can see that the Poverty Rate has by far the most impact on the success of a county, followed by the general cost of housing, whether as an aggregate or measured by the cost of a mortgage. Somewhat surprisingly, incomes and unemployment have less of an effect. It is also noteworthy that the rate of change of poverty also has little effect on the success of a variable. This suggests that the absolute poverty level, rather than change in poverty, has the largest effect on success.

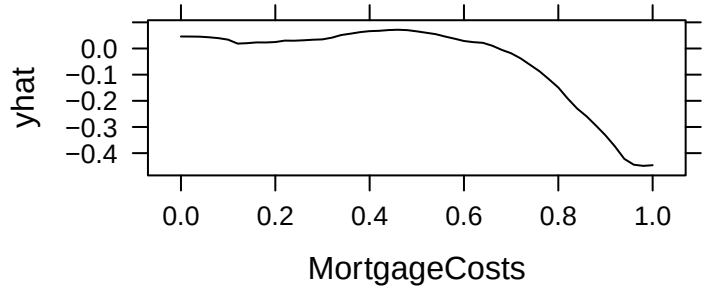


Here we plot the partial predictive plots for the variables with the greatest effect on success. As can be seen, each of those has an effect on our success variable. However, the effect of the poverty rate is noteworthy, as it feels counterintuitive that it has a greater effect than some of the other variables.









As those plots demonstrate, low levels of the respective measures are highly predictive of high success scores. We can guess that the reason our definition of success is so well correlated to low housing costs, mortgage costs and poverty rates is because there tend to be far more poor people than wealthier people. A high poverty rate translates to a far higher number of people having very little wealth, more so than a low poverty rate translates to more people having high wealth. In this sense, our analysis is useful as it provides a general guideline for policymakers looking to make an area more successful: reduce the poverty rate, or in other terms, focus on allocating resources to the poor as opposed to providing resources

(such as tax cuts, certain infrastructure investments etc.) to those who already have wealth above the median. Graphically, we can show the results of our analysis by showing actual and predicted values of "successful" counties for 2017. This demonstrates that, while our model only explains ~33% of the actual deviation, it does capture much of the regional concentrations of weath in the United States. It does mean that our analysis failed to meet our first criteria for success, i.e. that we are able to predict with greater than 50% accuracy whether a county is "successful" or not. However, the finding that the poverty rate plays such a strong role in determining the equitable development of a county is a potentially useful insight for policymakers, or for further research.

Appendix 1 Linear Model Results

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1598  -0.3435  -0.0419   0.2809   6.2971
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.32632    0.12868  18.078  < 2e-16 ***
## UnemploymentRate    0.09910    0.07509   1.320   0.1869
## MedianIncome        0.33550    0.35892   0.935   0.3499
## PovertyRate        -3.56856    0.10889 -32.773  < 2e-16 ***
## HousingCosts       -1.71786    0.11124 -15.444  < 2e-16 ***
```

```
## MortgageCosts       -0.23780    0.09601  -2.477   0.0133 *
## Fertility            0.79498    0.08745   9.091  < 2e-16 ***
## BachelorRate        -0.48990    0.07963  -6.152 8.10e-10 ***
## UnemploymentChange  -0.05406    0.07705  -0.702   0.4829
## MedianIncomeChange  -0.44111    0.41065  -1.074   0.2828
## PovertyRateChange    1.10480    0.12874   8.581  < 2e-16 ***
## HousingCostsChange   0.53105    0.08020   6.622 3.82e-11 ***
## MortgageCostsChange  0.63423    0.10396   6.101 1.11e-09 ***
## FertilityChange     -0.58994    0.08338  -7.076 1.64e-12 ***
## BachelorRateChange  -0.08035    0.08865  -0.906   0.3648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6175 on 6724 degrees of freedom
## Multiple R-squared:  0.3192, Adjusted R-squared:  0.3178
## F-statistic: 225.2 on 14 and 6724 DF,  p-value: < 2.2e-16
```

Appendix 2 Regression Tree Results

```
## CART
##
## 6739 samples
##   14 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1348, 1348, 1348, 1347, 1348
```

```
## Resampling results across tuning parameters:
##
##   cp     RMSE       Rsquared   MAE
##   0.001  0.7118015  0.1920790  0.5039547
##   0.010  0.6749292  0.2022748  0.4688326
##   0.020  0.6819849  0.1739488  0.4783530
##   0.040  0.6974102  0.1308050  0.4937275
##   0.060  0.6986747  0.1273269  0.4951598
##   0.100  0.6986747  0.1273269  0.4951598
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.01.
```

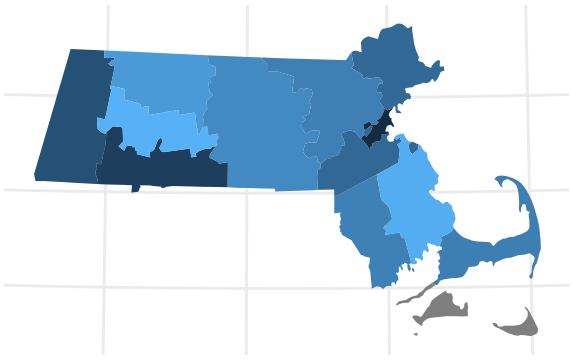
Appendix 3 Random Forest Results

```
## Random Forest
##
## 6739 samples
##   14 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1348, 1348, 1348, 1347, 1348
## Resampling results across tuning parameters:
##
##   mtry  splitrule   RMSE       Rsquared   MAE
##    1    variance    0.6374428  0.3086054  0.4402637
```

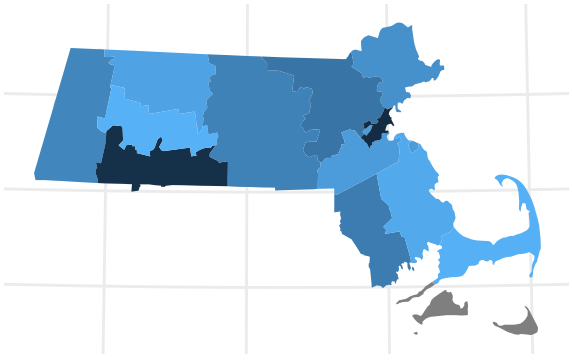
```
##    1    extratrees  0.6486755  0.3207167  0.4509304
##    2    variance    0.6230808  0.3223315  0.4255877
##    2    extratrees  0.6312278  0.3308910  0.4339148
##    5    variance    0.6170665  0.3215905  0.4165559
##    5    extratrees  0.6155270  0.3391271  0.4183533
##    8    variance    0.6197657  0.3140622  0.4168248
##    8    extratrees  0.6122303  0.3377200  0.4148376
##   10    variance    0.6221826  0.3091187  0.4182055
##   10    extratrees  0.6108277  0.3380756  0.4136756
##   14    variance    0.6265617  0.3007017  0.4204200
##   14    extratrees  0.6114856  0.3336065  0.4136040
##
## Tuning parameter 'min.node.size' was held constant at a value of 1
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were mtry = 10, splitrule =
##  extratrees and min.node.size = 1.
```

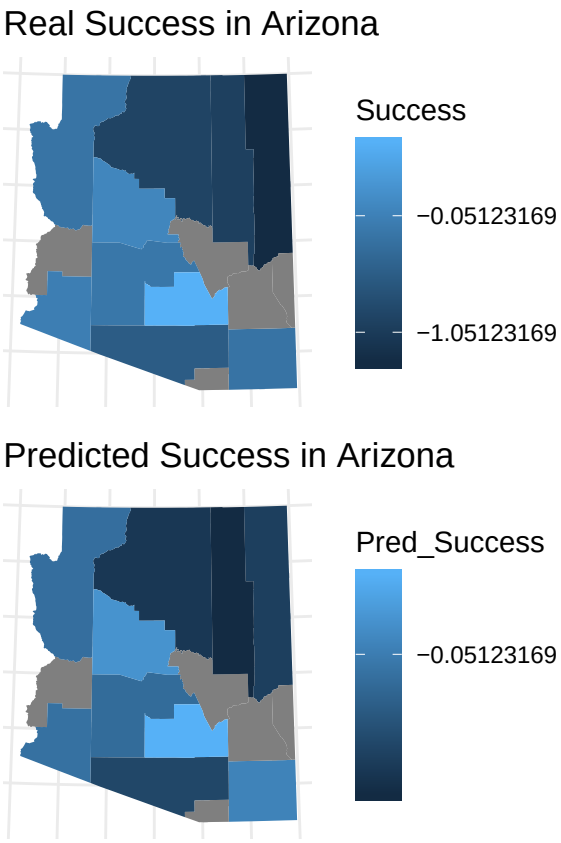
Appendix 4 Comparison between Reality and Prediction

Real Success in Massachusetts



Predicted Success in Massachusetts





4 Appendix 5 Sources

What Makes Urban Areas Around the World Successful; Benfield, Kaid; The Atlantic, April 2011: https://www.theatlantic.com/international/archive/2011/04/sustainable-cities-what-makes-urban-areas-around-the-world-successful/237668/

6 Examples of What Makes a Great Public Space, PBS Report, March 2016: https://www.pps.org/article/you-asked-we-answered-6-examples-of-what-makes-a-great-public-space

How to Quantify a Successful City; Beyer, Scott; Forbes, November 2015 https://www.forbes.com/sites/scottbeyer/2015/11/08/how-to-quantify-a-successful-city

The New Gilded Age; Sommellier, Estelle and Price, Mark; Economic Policy Institute, July 2018 https://www.epi.org/publication/the-new-gilded-age-income-inequality-in-the-u-s-by-state-metropolita