

Exercise

Create a prediction model

Section 2 Exercise 1

November 4, 2020



Create a prediction model

Time to complete

150 minutes

Introduction

Prediction is an important part of spatial data science. You can use prediction to forecast future values (for example, predicting tomorrow's air quality for a specified location), downscale information (for example, using voter turnout data at the county level to predict voter turnout at the census tract level), or fill in missing values in a dataset.

ArcGIS provides various prediction tools to help you complete these types of analyses. In this exercise, you will use the Forest-Based Classification And Regression tool, which uses an adaptation of Leo Breiman's random forest algorithm. This supervised machine learning algorithm allows you to use existing data to train models that may be useful for predictive analysis.

The tool creates many decision trees, called an ensemble or a forest, that are used for prediction. Each tree generates its own prediction and is used as part of a voting scheme to make final predictions. The strength of the forest-based method is in capturing commonalities of weak predictors (the trees) and combining them to create a powerful predictor (the forest). You will use this tool to train and evaluate a predictive model, modifying variables and parameters to improve the model performance.

Exercise scenario

After preparing and visualizing your data, you are ready to begin your predictive analysis. In this exercise, you will create models that predict voter turnout. These models will use explanatory variables, such as income and age, to predict the dependent variable, voter turnout.

You will use this model to downscale voter turnout from the county to the census tract level. This information will be used to organize a "Get Out the Vote" canvassing campaign. These campaigns are used to get supporters to vote on Election Day. This model will help identify local regions that are expected to have low voter turnout so that you know where to target your campaign.

Step 1: Download the exercise data files

In this step, you will download the exercise data files.

- a Open a new web browser tab or window.
- b Go to <https://bit.ly/sdspredict> and download the exercise data ZIP file.

Note: The complete URL to the exercise data file is <https://www.arcgis.com/home/item.html?id=6c177c0b07ca481698065354b958c8d9>.

- c Extract the files to a folder on your local computer, saving the files in a location that you will remember.

Step 2: Open an ArcGIS Pro project

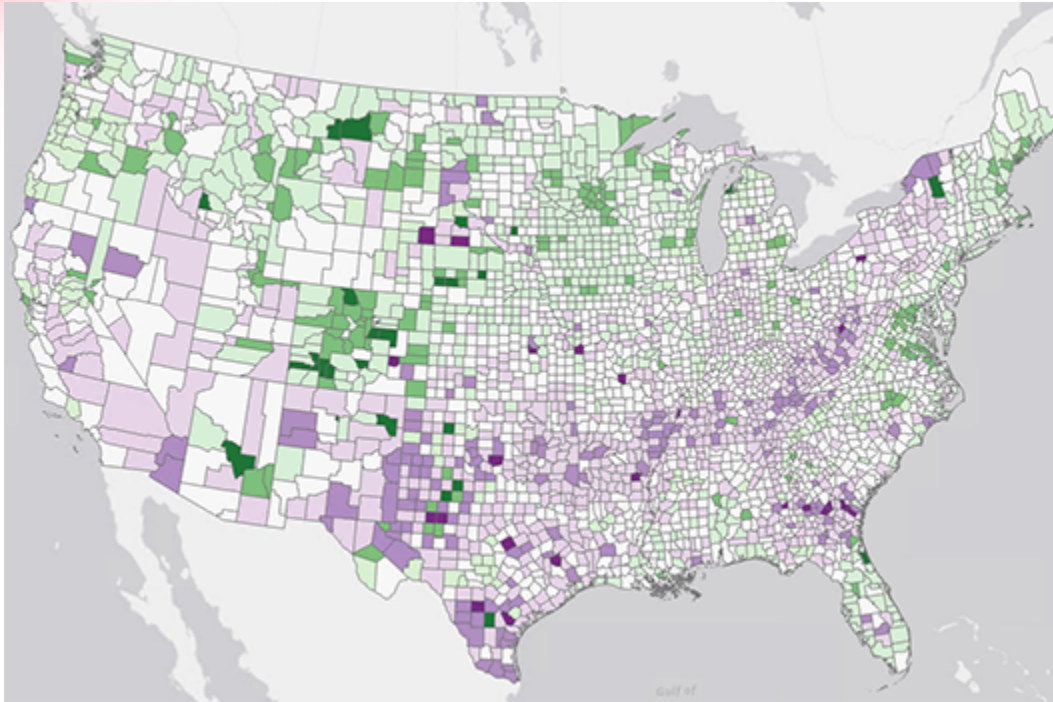
- a Start ArcGIS Pro.
- b If necessary, sign in using your course ArcGIS account user name and password.

Note: The Section 1 Exercise 1 PDF explains where to locate your course account information to sign in to ArcGIS Pro. If you have trouble signing in, please refer to the Common Questions list on the MOOC home page Help tab.

- c In the bottom-left corner of the ArcGIS Pro Start page, click Open Another Project.

Note: If you have configured ArcGIS Pro to start without a project template or with a default project, you will not see the Start page. On the Project tab, click Open, and then click Open Another Project.

- d In the Open Project dialog box, browse to the Prediction folder that you saved on your computer.
- e Click Prediction.aprx to select it, and then click OK.



A Prediction map tab opens to a gray basemap with a map layer that represents the 2016 election results for each county in the United States. Counties with a voter turnout value below the mean are purple, and counties with a voter turnout value above the mean are green.


Step 3: Create a prediction model

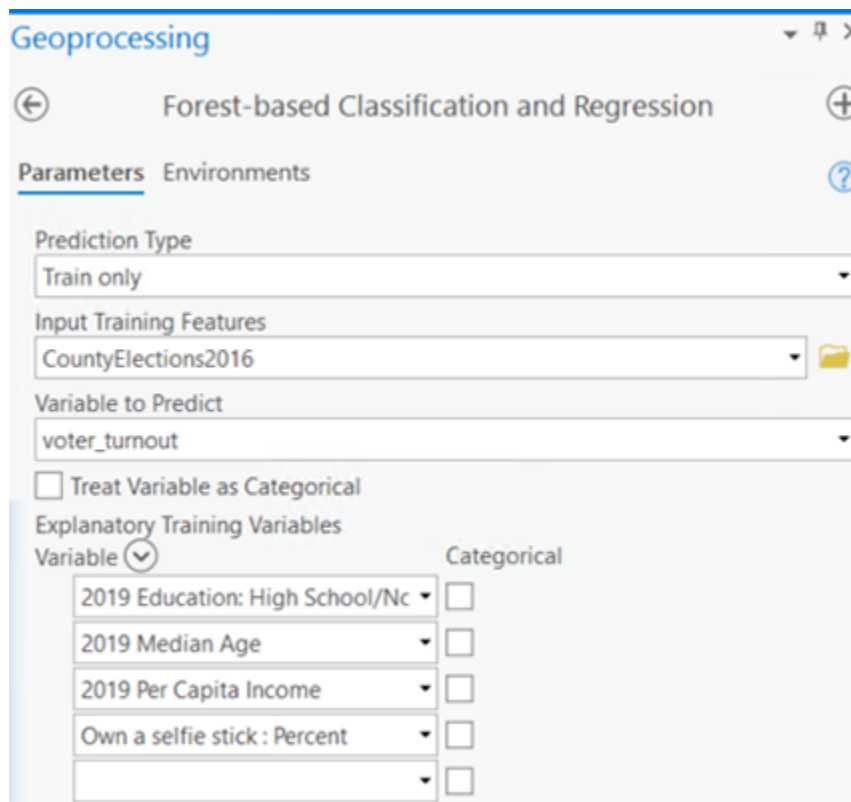
During the data engineering exercise, you enriched the 2016 election data with various demographic variables. During the data visualization exercise, you explored the relationship of these variables to voter turnout, identifying variables that have a strong relationship to voter turnout. You will use these variables in your first prediction model.

- a From the Analysis tab, in the Geoprocessing group, click Tools.
- b In the Geoprocessing pane, search for **Forest-Based Classification And Regression (Spatial Statistics Tools)**.
- c In the search results, click the Forest-Based Classification And Regression (Spatial Statistics Tools) tool.

Note: Be sure to use the Spatial Statistics tool and not the GeoAnalytics Desktop tool.

The Forest-Based Classification And Regression tool opens in the Geoprocessing pane.

- d In the Geoprocessing pane, set or confirm the following parameters:
 - Prediction Type: Train Only
 - Input Training Features: CountyElections2016
 - Variable To Predict: Voter_Turnout
- e Under Explanatory Training Variables, next to Variable, click the Add Many button .
- f In the variable window, check the box for the following variables:
 - 2019 Education: High School/No Diploma : Percent
 - 2019 Median Age
 - 2019 Per Capita Income
 - Own A Selfie Stick : Percent
- g In the bottom-right corner of the variable window, click Add.



The screenshot shows the Geoprocessing pane with the tool 'Forest-based Classification and Regression' selected. The 'Parameters' tab is active. The 'Prediction Type' is set to 'Train only'. The 'Input Training Features' is set to 'CountyElections2016'. The 'Variable to Predict' is set to 'voter_turnout'. The 'Treat Variable as Categorical' checkbox is unchecked. Under 'Explanatory Training Variables', the 'Variable' button is selected, and a list of variables is shown with checkboxes for each:

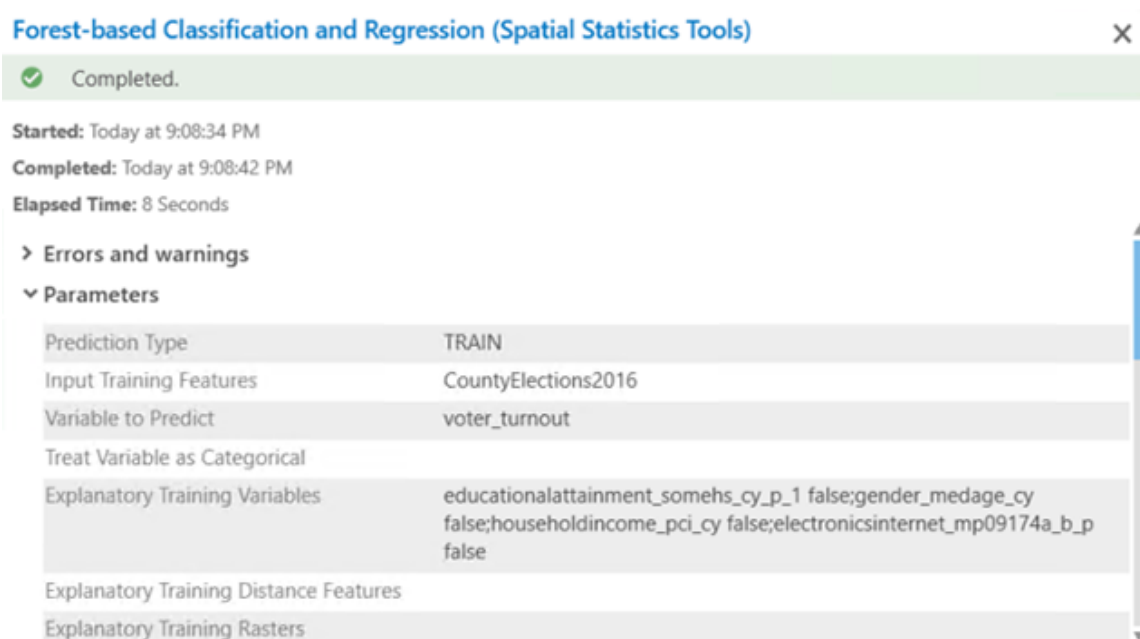
Variable	Categorical
2019 Education: High School/Nc	<input type="checkbox"/>
2019 Median Age	<input type="checkbox"/>
2019 Per Capita Income	<input type="checkbox"/>
Own a selfie stick : Percent	<input type="checkbox"/>
	<input type="checkbox"/>

- h In the Geoprocessing pane, click Run.

✓ Forest-based Classification and Regression completed.
[View Details](#) [Open History](#)

At the bottom of the Geoprocessing pane, you will see a message confirming that the tool completed. You did not specify in the tool to create an output. Instead, you will review the model's performance using the tool messages.

- i At the bottom of the Geoprocessing pane, in the green message box, click View Details.



The Forest-Based Classification And Regression (Spatial Statistics Tools) tool message window appears. Tool messages contain information like the parameters used to run the tool, how long the tool ran, and model performance diagnostics.

- j Scroll down and expand the Messages section.
- k In the Messages section, scroll down and locate the Training Data: Regression Diagnostics section.


```

----- Training Data: Regression Diagnostics -----
R-Squared                                0.909
p-value                                0.000
Standard Error                          0.005
*Predictions for the data used to train the model
compared to the observed categories for those features

---- Validation Data: Regression Diagnostics ----
R-Squared                                0.519
p-value                                0.000
Standard Error                          0.029
*Predictions for the test data (excluded from model
training) compared to the observed values for those
test features

```

Note: Each time that you run the Forest-Based Classification And Regression tool, you may get slightly different results due to the randomness introduced in the algorithm to prevent the model from overfitting to the training data.

By default, forest-based classification and regression reserves 10 percent of the data for validation. The model is trained without this random subset, and the tool returns an R-squared value measuring how well the model performs on the unseen data.

When a model is evaluated based on the training dataset rather than a validation dataset, it is common for estimates of performance to be overstated due to a concept called overfitting. Therefore, the validation R-squared value is a better indicator of model performance than the training R-squared value.

The model returned a validation R-squared value of 0.519, indicating that the model predicted the voter turnout value in the validation dataset with an accuracy of about 52 percent.

Next, you will review how important each explanatory variable was in generating a prediction.

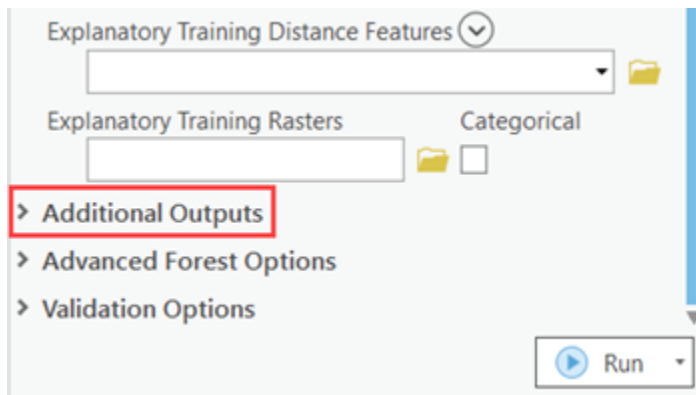
- i** Close the Forest-Based Classification And Regression tool message window.
- m** In the Contents pane, turn off the CountyElections2016 layer.



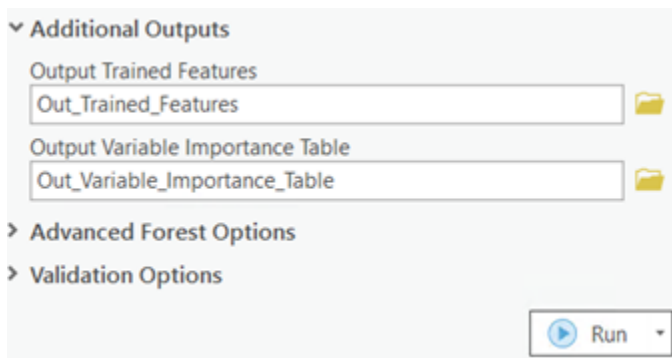
Throughout this exercise, you will rerun the Forest-Based Classification And Regression (Spatial Statistics Tools) geoprocessing tool using different variables and parameters. Read each step thoroughly, and carefully change the tool's parameters to match the instructions in this workbook.

Step 4: Explore variable importance

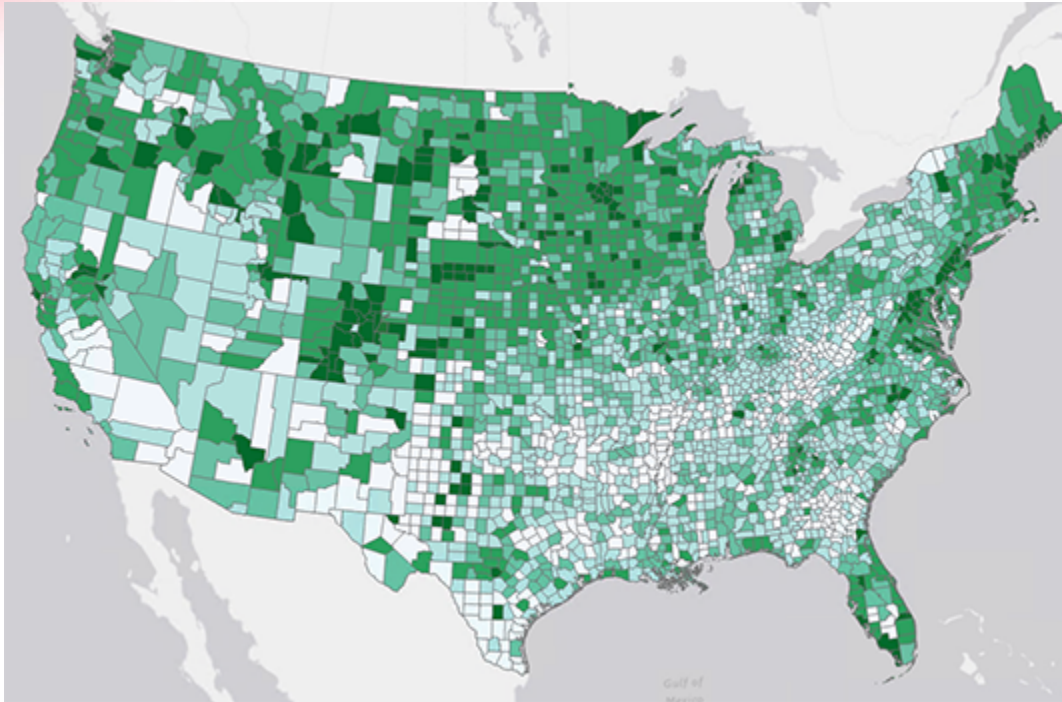
- a In the Geoprocessing pane, expand Additional Outputs, as specified in the following graphic.



- b Type the following parameters:
- Output Trained Features: **Out_Trained_Features**
 - Output Variable Importance Table: **Out_Variable_Importance_Table**



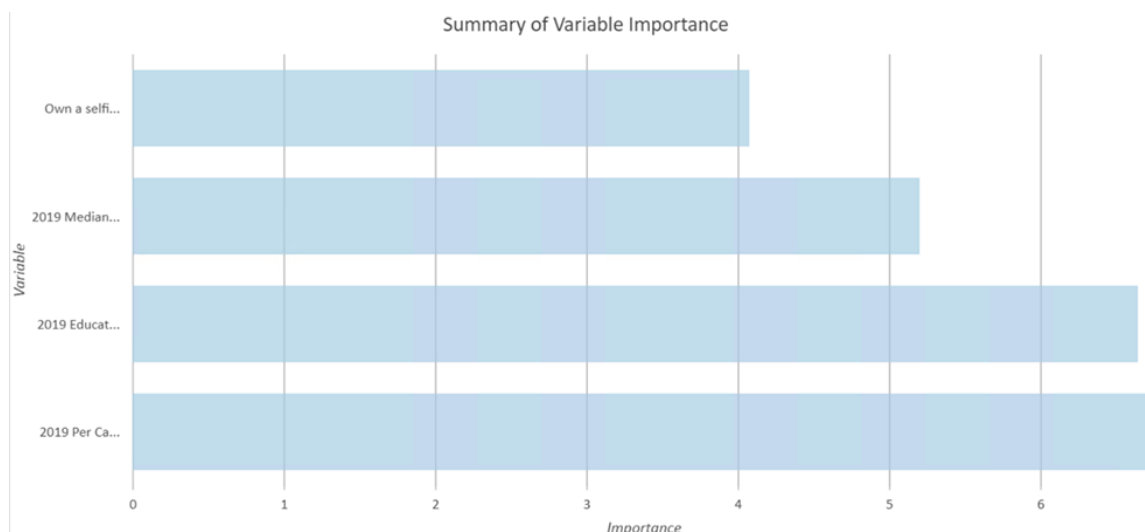
- c Click Run.



The Out_Trained_Features layer displays the predicted voter turnout for each county in the contiguous United States. A variable importance table and associated bar chart are added to the Contents pane and can be used to explore which variables were most important in this prediction.

- d** In the Contents pane, open the Summary Of Variable Importance chart.

Hint: Under Charts, right-click Summary Of Variable Importance and choose Open.



The 2019 Per Capita Income and 2019 Education: High School/No Diploma : Percent variables have the highest importance, meaning that they were the most useful in predicting voter turnout.

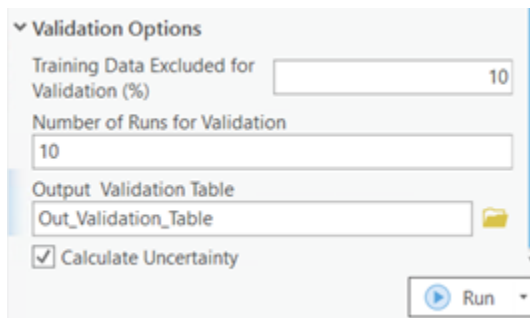
As previously mentioned, each time that you run the Forest-Based Classification And Regression tool, you may get slightly different results due to the randomness introduced in the algorithm to prevent the model from overfitting to the training data. To understand and account for this variability, you will use a parameter that allows the tool to create multiple models in one run. This output will allow you to explore the distribution of model performance.

Step 5: Examine model stability

- a Close the Summary Of Variable Importance chart.
- b At the bottom of the Chart Properties pane, click the Geoprocessing tab to return to that pane.
- c Near the bottom of the Geoprocessing pane, expand Validation Options.
- d For Number Of Runs For Validation, type **10**, and then click in the gray space of the tool so that the tool recognizes your update.

The Geoprocessing pane will refresh and the option for Output Validation Table will appear. This option only appears if the number of runs for validation specified is greater than 2. This table creates a histogram of the distribution of R-squared values for the model.

- e For Output Validation Table, type **Out_Validation_Table**
- f Under Output Validation Table, check the Calculate Uncertainty box.



Note: If you do not see the Calculate Uncertainty check box, scroll down in the Geoprocessing pane.

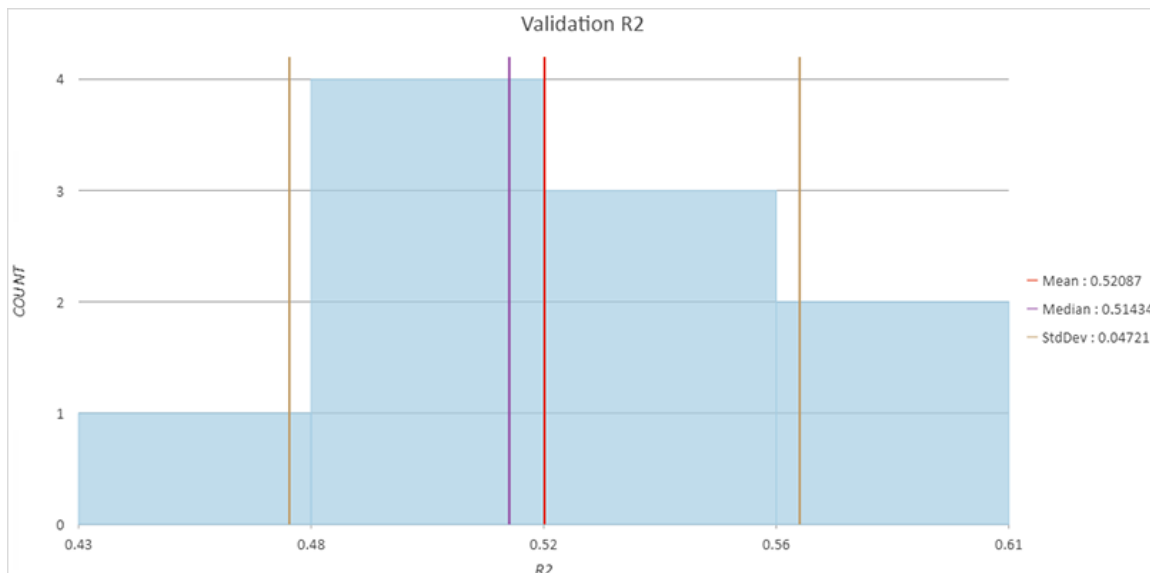
- g** Run the tool.
- h** At the bottom of the Geoprocessing pane, click View Details.
- i** In the tool message window, scroll down and expand Messages, and then scroll to the Validation Data: Regression Diagnostics section.

```

---- Validation Data: Regression Diagnostics ----
R-Squared                                0.527
p-value                                0.000
Standard Error                          0.028
*Predictions for the test data (excluded from model
training) compared to the observed values for those
test features
  
```

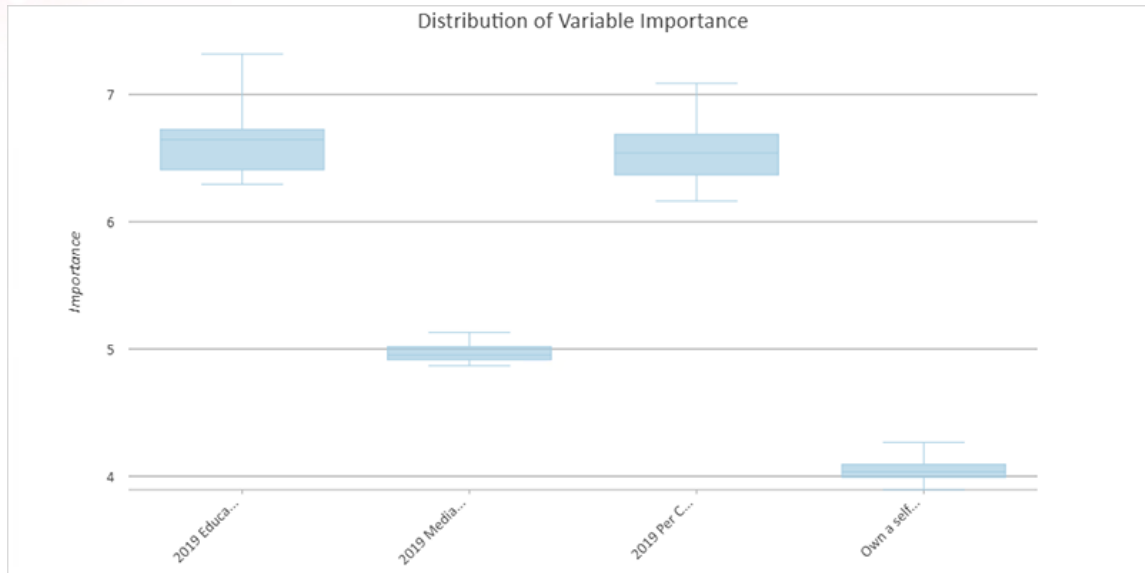
The tool trained 10 models with random subsets of validation data. The most representative R-squared value across the 10 runs is 0.527, corresponding to about 53 percent accuracy in prediction of the validation data. You can use a histogram to review the distribution of R-squared values returned over the 10 runs.

- j** Close the tool message window.
- k** In the Contents pane, open the Validation R2 chart.



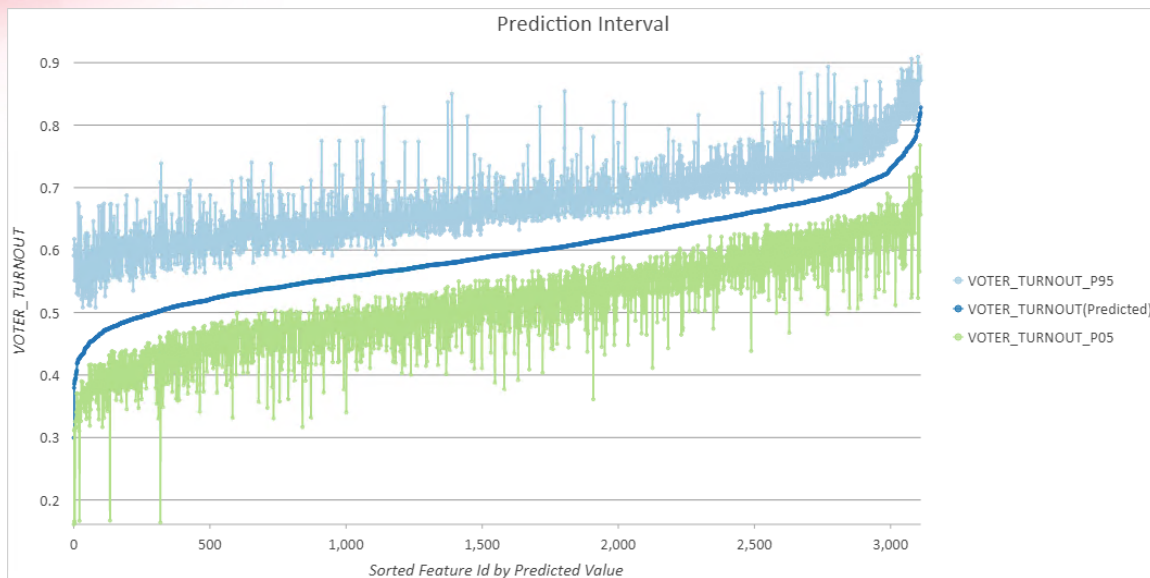
The histogram shows the variability in model performance by visualizing the distribution of R-squared values returned over the 10 runs. The mean R-squared value for the 10 runs of this model is 0.52.

- I** In the Contents pane, open the Distribution Of Variable Importance chart.



Instead of a bar chart, the variable importance is visualized using a box plot to show the distribution of importance across the 10 runs of the model. There is overlap in the distribution of importance of the 2019 Per Capita Income and 2019 Education: High School/No Diploma : Percent variables. In some runs of the model, Per Capita Income was more important, and in other runs, Education: High School/No Diploma was more important. Overall, both variables are strong candidates for your predictive model.

- m** In the Contents pane, under Out_Trained_Features, right-click Prediction Interval and choose Open.



The Prediction Interval chart displays the level of uncertainty for any given prediction value. By considering the range of prediction values returned by the individual trees in the forest, prediction intervals are generated indicating the range in which the true value is expected to fall. You can be 90 percent confident that new prediction values generated using the same explanatory variables would fall in this range. This chart can help you identify if the model is better at predicting some values than others. For example, if the confidence intervals were much larger for low voter turnout values, then you would know that the model is not as stable for predicting low voter turnout as it is for predicting high voter turnout. The prediction intervals in this model are fairly consistent, indicating that the model performance is relatively stable across all values.

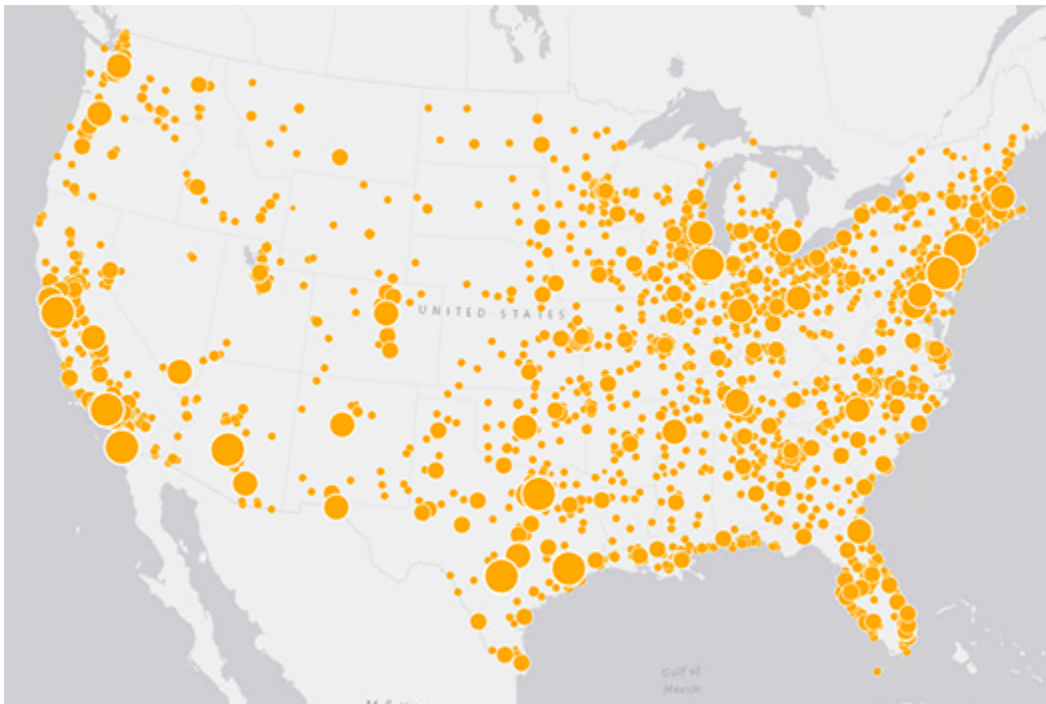
At this point, you have used the attributes in your dataset as the explanatory variables in your model. With the Forest-Based Classification And Regression tool, you can also calculate new variables based on distances to meaningful locations. In the next step, you will calculate new variables and assess their importance to the model.

n Close the charts and activate the Geoprocessing pane.

Step 6: Add distance variables to the model

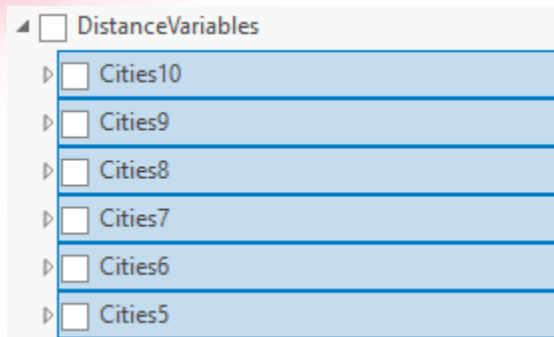
You want to incorporate each county's urban and rural characteristics into the model to determine if these variables improve voter turnout predictions. To represent urban and rural characteristics, you will calculate the distance between each county and cities of various sizes. The proximity to each of these cities will be used to represent the urban and rural characteristics, with more rural counties being farther from cities.

- a In the Contents pane, turn off the Out_Trained_Features layer.
- b Turn on and expand the DistanceVariables group layer, and then turn on the following layers:
 - Cities10
 - Cities9
 - Cities8
 - Cities7
 - Cities6
 - Cities5



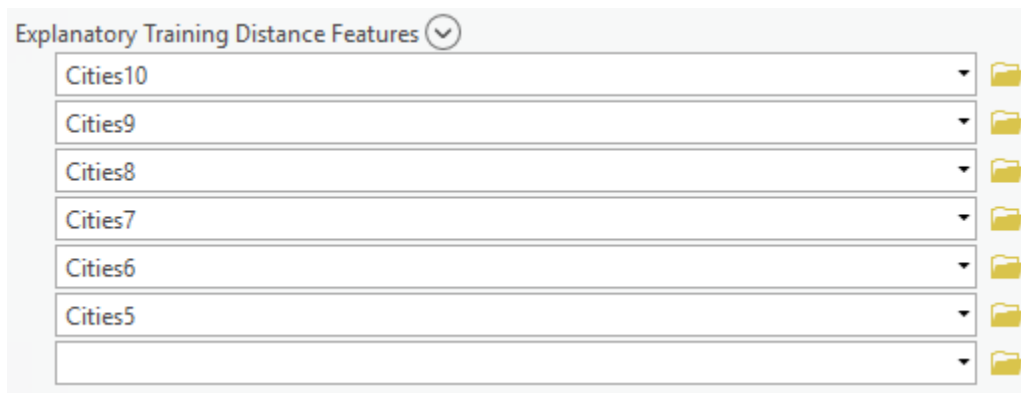
Each Cities layer in the DistanceVariables group layer represents a class of city size based on population. Cities10 represents cities with the largest populations, and Cities5 represents the cities with the smallest populations.

- c Turn off the DistanceVariables group layer and all the Cities layers.
- d Click Cities10 to select it (but do not check the box), and then press Shift and click Cities5.



The six distance variables are selected.

- e Drag the selected layers into the Geoprocessing pane, under Explanatory Training Distance Features.



- f Click Run.
- g In the Contents pane, right-click Out_Trained_Features and choose Attribute Table.
- h In the attribute table, scroll to the Cities attribute fields.

CITIES10	CITIES9	CITIES8	CITIES7	CITIES6	CITIES5
1007783.447616	467299.554413	12723.383042	91100.873325	0	427621.759861
819023.706977	566502.329637	4940.430496	20653.18596	0	422569.363134
1097517.892448	428772.323189	52174.578718	50363.593022	0	355994.640857
973335.016097	396462.1964	45547.712767	31097.869895	17657.98708	363624.639765
1068573.846971	261723.491894	35767.384935	53475.566964	16909.193194	413247.248244
1084712.54391	473636.975597	37790.391671	49118.422059	21940.985023	397675.474009

Note: When a city point is contained within a county, the distance will be zero.

The Forest-Based Classification And Regression tool calculates the distances from each county to the nearest city of each class (the closest class 5 city, the closest class 6 city, and so on). These distances are added to the Out_Trained_Features layer as separate attribute fields.

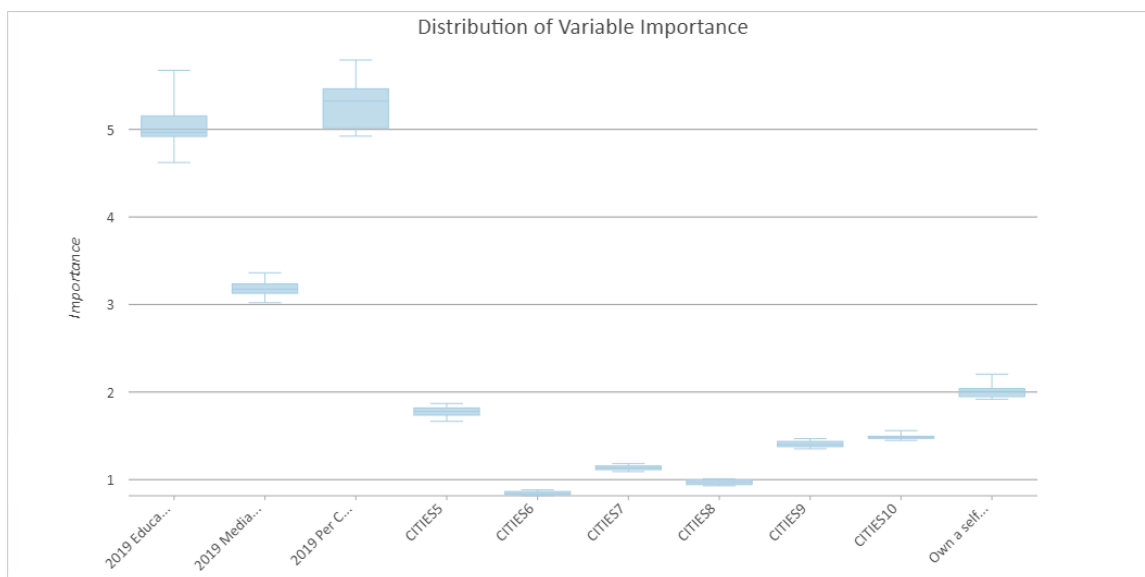
- i Close the attribute table.
- j At the bottom of the Geoprocessing pane, click View Details.
- k In the tool message window, expand Messages, and then scroll to the Validation Data: Regression Diagnostics section.

```

---- Validation Data: Regression Diagnostics ----
R-Squared                                0.614
p-value                                0.000
Standard Error                          0.025
*Predictions for the test data (excluded from model
training) compared to the observed values for those test
features
  
```

The model's R-squared value has increased to 0.614, which means that you are predicting with more than 61 percent accuracy based on the validation data. You will review the variable importance chart to see how influential each of these distance variables are to the model performance.

- l Close the tool message window.
- m In the Contents pane, open the Distribution Of Variable Importance chart.





The distance to the smallest cities (Cities5) and the distance to the largest cities (Cities10) are more important than the other distance variables. Overall, however, these variables were not as helpful as the income and education variables.

In the next step, you will add additional demographic variables in an attempt to make a more robust model.













- n** Close the Distribution Of Variable Importance chart.

Step 7: Create a prediction model with many variables

The Forest-Based Classification And Regression tool uses a random subset of the available explanatory variables in each decision tree. Commonalities in the predictions and variables used among all the trees in the forest are quantified in the variable importance diagnostic. In general, this means that you can test adding variables to the model without diminishing the model's predictive power. Variables that are useful result in higher variable importance scores, and variables that are not useful result in lower variable importance scores.

- a** Activate the Geoprocessing pane, and then under Explanatory Training Variables, click the Add Many button .
- b** In the bottom-left corner of the variable window, click the Toggle All Checkboxes button .
- c** Scroll to the top of the options, and then uncheck the box for the following variables:
- 2019 Education: High School/No Diploma : Percent
 - 2019 Median Age
 - 2019 Per Capita Income
 - County
 - Own A Selfie Stick : Percent
 - Shape_Area
 - Shape_Length
 - Voter_Turnout
- d** Click Add.

Explanatory Training Variables

Variable 	Categorical
2019 Education: High School/N 	<input type="checkbox"/>
2019 Median Age 	<input type="checkbox"/>
2019 Per Capita Income 	<input type="checkbox"/>
Own a selfie stick : Percent 	<input type="checkbox"/>
2019 Average Disposable Incon 	<input type="checkbox"/>
2019 Average Household Incon 	<input type="checkbox"/>
2019 Cash Contributions to Poli 	<input type="checkbox"/>
2019 Disposable Income \$1000 	<input type="checkbox"/>
2019 Disposable Income \$1500 	<input type="checkbox"/>
2019 Disposable Income \$1500 	<input type="checkbox"/>
2019 Disposable Income \$2000 	<input type="checkbox"/>

Multiple variables are now added.

- e** Run the tool.
- f** Review the R-squared value in the validation data regression diagnostics.

Hint: In the Forest-Based Classification And Regression tool message window, scroll to the Validation Data: Regression Diagnostics section.

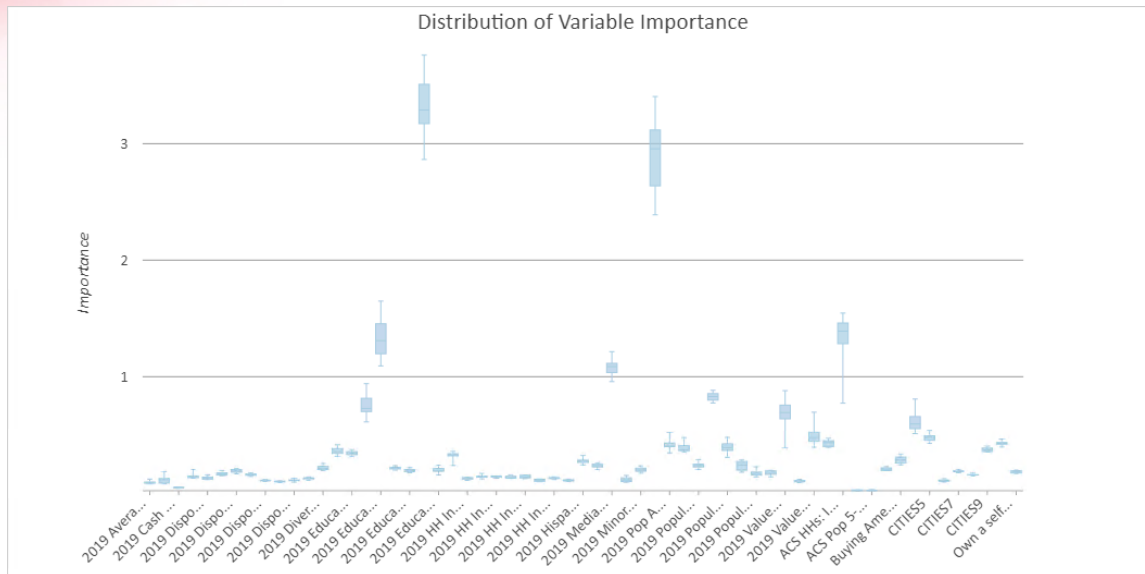
```

---- Validation Data: Regression Diagnostics ----
R-Squared                                0.707
p-value                                 0.000
Standard Error                           0.022
*Predictions for the test data (excluded from model
training) compared to the observed values for those
test features
  
```

The model's R-squared value has increased to 0.707, which means that you are now predicting with more than 70 percent accuracy based on the validation data.

- g** Close the tool message window.
- h** Review the Distribution Of Variable Importance chart.

Hint: Contents pane > Distribution Of Variable Importance chart



Note: You can zoom in to the chart to better see the distribution for a particular variable.

2019 Per Capita Income and 2019 Education: High School/No Diploma : Percent still have the highest variable importance in the model, but several new variables have contributed to the model and raised its performance. There are also several variables that may not be helping the model, represented by their low variable importance.

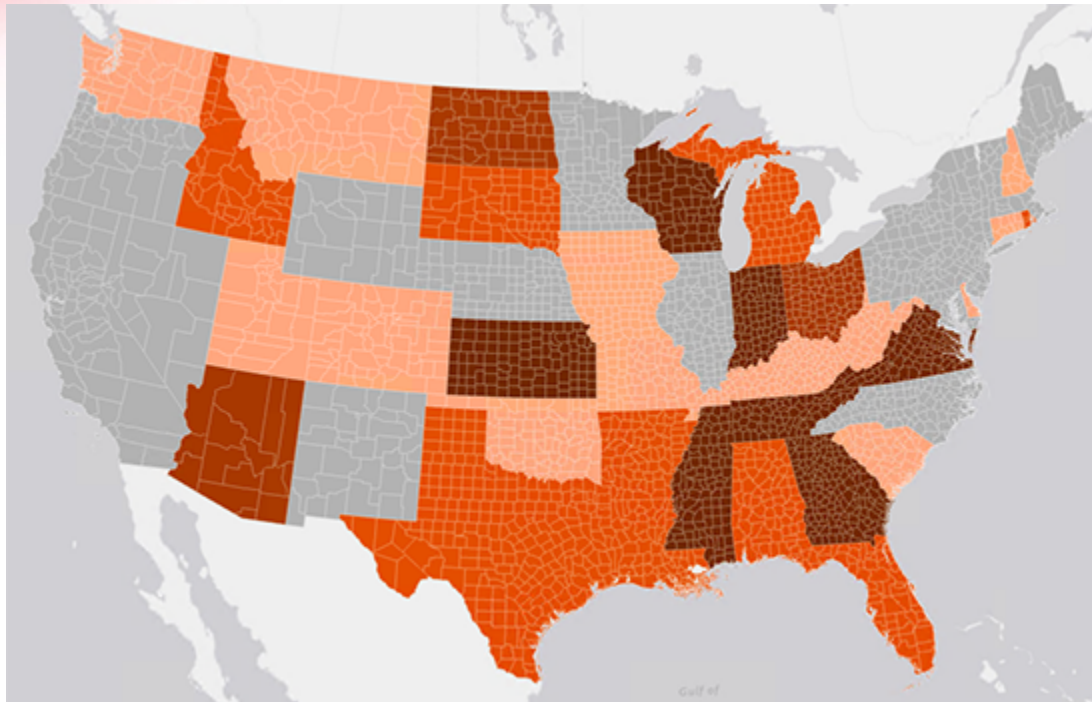
All of these variables represent continuous numerical values. In the next step, you will add a categorical variable to the model.

- i** Close the Distribution Of Variable Importance chart.

Step 8: Add categorical variables to the model

Research indicates that state voting requirement laws may affect voter turnout. You can add voting requirement laws to the model as a categorical variable to see if it helps predict voter turnout.

- a** In the Contents pane, turn off the Out_Trained_Features layer.
- b** Turn on and expand the CountyElections2016_VotingRequirements layer to view how the categories are symbolized on the map.

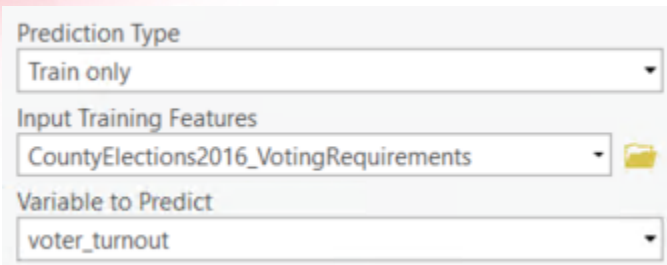


This layer includes all the attributes from the CountyElections2016 layer and an additional attribute that represents the following categories for state voting requirements:

- No document required
- ID without photo
- ID with photo
- Strict ID without photo
- Strict ID with photo

You will add this categorical variable to the model and assess the model performance.


- Turn off the CountyElections2016_VotingRequirements layer.
- In the Geoprocessing pane, for Input Training Features, choose CountyElections2016_VotingRequirements.

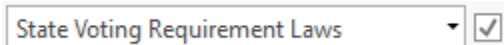


Prediction Type
Train only

Input Training Features
CountyElections2016_VotingRequirements

Variable to Predict
voter_turnout

- e Under Explanatory Training Variables, click the Add Many button .
- f In the variable window, check the State Voting Requirement Laws box.
- g Click Add.



State Voting Requirement Laws ☒

The State Voting Requirement Laws variable is added to the bottom of the explanatory training variables list and marked as a categorical variable.

- h Click Run.
- i Review the R-squared value in the validation data regression diagnostics.

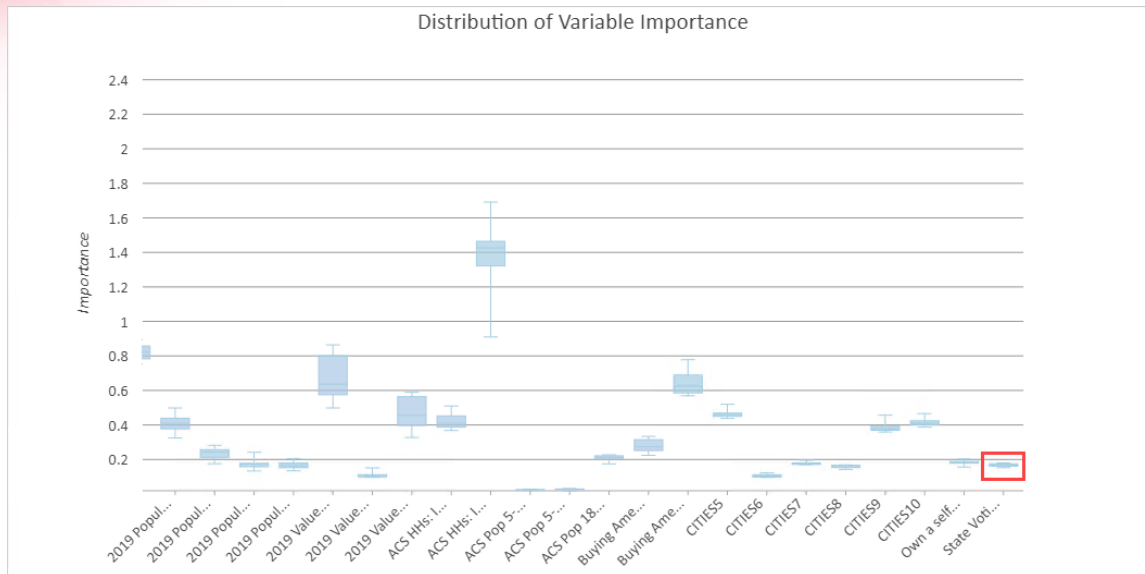
Hint: In the Forest-Based Classification And Regression tool message window, scroll to the Validation Data: Regression Diagnostics section.

```

---- Validation Data: Regression Diagnostics ----
R-Squared                                0.703
p-value                                0.000
Standard Error                          0.022
*Predictions for the test data (excluded from model training)
compared to the observed values for those test features
    
```

The R-squared value is about the same, with no improvements to the model. You can review the variable importance to determine how helpful state voting requirement laws are compared to the other variables.

- j Close the tool message window.
- k Open the Distribution Of Variable Importance chart.
- l Zoom to the bottom right of the chart to locate the State Voting Requirements variable.



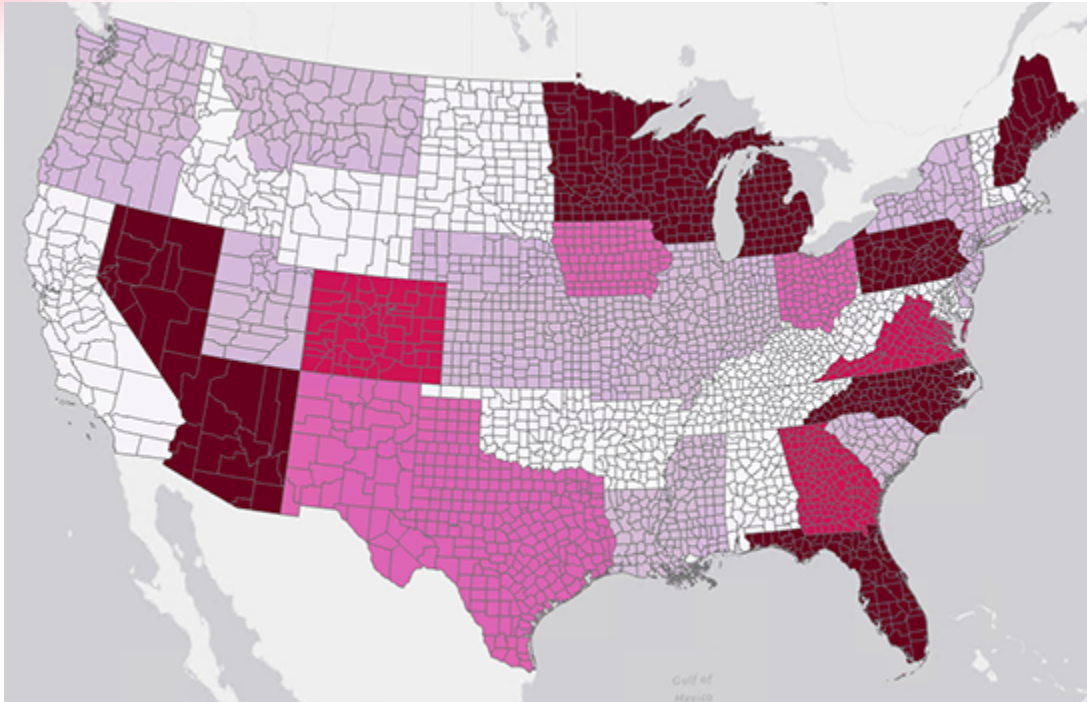
Although voting requirement laws are likely important in voter participation, they do not help the model very much—at least compared to the other explanatory variables. You will continue exploring state-related variables that can affect voter participation.

- m** Close the Distribution Of Variable Importance chart.

Step 9: Add a variable that represents election competitiveness

Additional research and literature associate voter turnout to how competitive each state is in the presidential election, due to a common perception that presidential parties will automatically win certain states. You will add this variable to the model and assess the model performance.

- a** In the Contents pane, turn off the Out_Trained_Features layer.
- b** Turn on and expand the CountyElections2016_PartyVotes layer.



This layer includes the attributes from the CountyElections2016 layer, the State Voting Requirement Laws attribute, and an additional attribute that measures the difference in election party votes as a percent. This attribute, State Percent Votes Difference, represents how competitive each state is in the presidential election, with a lower percent difference indicating a more competitive state.

Note: This ArcGIS Pro project includes an ArcGIS notebook, ElectionPartyVotes, that was used to calculate the State Percent Votes Difference.

- c Turn off the CountyElections2016_PartyVotes layer.
- d In the Geoprocessing pane, for Input Training Features, choose CountyElections2016_PartyVotes.

Prediction Type
Train only

Input Training Features
CountyElections2016_PartyVotes

Variable to Predict
voter_turnout

- e Under Explanatory Training Variables, click the Add Many button .

- f In the variable window, check the State Percent Votes Difference box, and then click Add.
- g Run the tool.

1. What is the validation R-squared value for this model?

Hint: In the Forest-Based Classification And Regression tool message window, scroll to the Validation Data: Regression Diagnostics section.

2. Compared to the other variables, is State Percentage Votes Difference a helpful variable in this model?

Hint: In the Contents pane, open the Distribution Of Variable Importance chart.

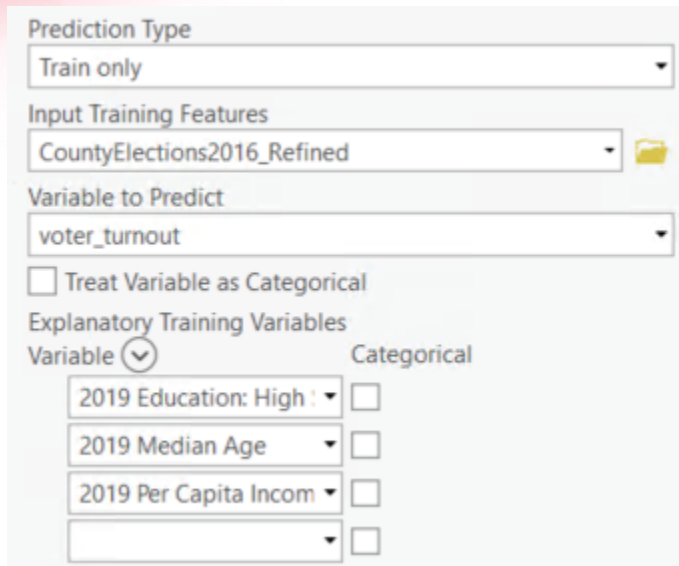
Identifying a "good" model is subjective and varies greatly based on the industry and how the model will be used. In many fields, including many of the social sciences, an R-squared value over 0.70 might be considered satisfactory for making a prediction. Before using this model to predict, you will simplify the model to only include the most important variables.

- h Close the Distribution Of Variable Importance chart and the tool message window.

Step 10: Refine the model

There are many different ways to select variables to include in a model. In this analysis, the most important variables were chosen by defining a threshold in the Variable Importance Table. The variables with importance above this threshold will be included in this refined version of the model.



- a In the Contents pane, turn off the Out_Trained_Features layer.
- b In the Geoprocessing pane, for Input Training Features, choose CountyElections2016_Refined.



The screenshot shows the ArcGIS Pro Model Builder interface. The 'Prediction Type' is set to 'Train only'. The 'Input Training Features' is 'CountyElections2016_Refined'. The 'Variable to Predict' is 'voter_turnout'. The 'Treat Variable as Categorical' checkbox is unchecked. Under 'Explanatory Training Variables', there is a table with two columns: 'Variable' and 'Categorical'. The table contains three rows: '2019 Education: High School/No Diploma : Percent', '2019 Median Age', and '2019 Per Capita Income'. Each row has an unchecked checkbox in the 'Categorical' column.

Variable	Categorical
2019 Education: High School/No Diploma : Percent	<input type="checkbox"/>
2019 Median Age	<input type="checkbox"/>
2019 Per Capita Income	<input type="checkbox"/>

2019 Education: High School/No Diploma : Percent, 2019 Median Age, and 2019 Per Capita Income are already listed under Explanatory Training Variables.

- c Under Explanatory Training Variables, click the Add Many button .
- d In the variable window, click the Toggle All Checkboxes button .
- e Uncheck the box for the following variables:
 - 2019 Education: High School/No Diploma : Percent
 - 2019 Median Age
 - 2019 Per Capita Income
 - 2019 Population Age 18+
 - County
 - FIPS
 - Shape_Area
 - Shape_Length
 - State
 - Voter_Turnout



You are unchecking the first three variables in the variable window to ensure that these variables are not listed under Explanatory Training Variables twice.

- f Click Add.
- g Scroll down and expand Advanced Forest Options.
- h For Number Of Trees, type **1000**.

▼ Advanced Forest Options

Number of Trees
1000

Minimum Leaf Size

Maximum Tree Depth

Data Available per Tree (%)
100

Number of Randomly Sampled Variables

Increasing the number of trees improves the chance that each variable will be used in a decision tree, resulting in a more accurate model prediction. Specifying the number of trees is a balance between the accuracy of the model and the processing time to generate the model.

- i Run the tool.

Note: Because of the increased number of trees, the tool may take a few minutes to run.

- j Review the geoprocessing tool message and various charts to answer the following questions.

3. What is the validation R-squared value for this model?

4. What is the mean R-squared value over the 10 runs of this model?

Hint: In the Contents pane, open the Validation R2 chart.

5. What are the two most important explanatory variables in this model?

The simplified model has approximately the same R-squared value, meaning that removing the variables with low importance did not compromise model performance. You will review additional model metrics to help you assess if the model requires any additional changes.

k Close the charts.

Step 11: Examine additional model metrics

- a** If necessary, from the Geoprocessing pane, reopen the Forest-Based Classification And Regression (Spatial Statistics Tools) tool message window.
- b** Under Messages, scroll to the Model Out Of Bag Errors section.

```
----- Model Out of Bag Errors -----
Number of Trees          500          1000
MSE                      0.002        0.002
% of variation explained  71.738      71.883
```

Model Out Of Bag Errors is another diagnostic that can help validate the model. The percentage of variation explained indicates the percent of variability in voter turnout that can be explained using this model. Model Out Of Bag Errors also shows how much performance is gained by increasing the number of trees in the model. If the percentage of variation explained significantly increases from the 500 to the 1000 column, you may want to increase the number of trees to improve model performance.

This model does not see a significant increase in percentage of variation explained, so you do not need to increase the number of trees.

- c** Scroll to the Explanatory Variable Range Diagnostics section.

----- Explanatory Variable Range Diagnostics -----						
Variable	Training		Validation		Training Share(a)	Validation Share(b)
	Minimum	Maximum	Minimum	Maximum		
2019 Median Age	23.30	61.80	24.80	58.00	1.00	0.86*
2019 Per Capita Income	11729.00	78564.00	9395.00	58856.00	0.97*	0.71*
2019 Education: Some College/No Degree : Percent	7.58	37.07	12.09	32.27	1.00	0.68*
2019 Education: Bachelor's Degree : Percent	2.59	45.16	2.96	41.21	1.00	0.90*
2019 Education: Associate's Degree : Percent	0.88	22.08	3.03	18.77	1.00	0.74*
2019 Education: Grad/Professional Degree : Percent	0.00	43.89	0.81	31.69	1.00	0.70*
2019 Pop Age 15+: Never Married : Percent	10.40	68.55	13.91	59.81	1.00	0.79*

Note: You may need to expand the window to better see this section.

The Explanatory Variable Range Diagnostics lists the range of values covered by each explanatory variable in the datasets used to train and validate the model. For example, median age values spanned from 23 to 61 in the dataset used to train the model and from 24 to 58 in the dataset used to validate the model.

The Validation Share indicates the percentage of overlap between the values used to train and the values used to validate. In this example, 86 percent of the median age values used to train the model were used to validate the model. A value over one indicates that the model predicted values outside the range of values in the training data. To minimize extrapolation, you will review this diagnostic as you predict voter turnout for census tracts.

For more information about these additional model metrics, see ArcGIS Pro Help: [How Forest-based Classification and Regression works > Output messages and diagnostics](#).

- d** Close the tool message window.
- e** In the Contents pane, turn off Out_Trained_Features.

Step 12: Predict values

You trained a model using the county data that you had available. You can use this model to predict voter turnout at the census tract level, which is much higher resolution and will allow you to get a sense of more detailed spatial patterns.

To predict voter turnout at the tract level, you need census tract data with explanatory variables that match the explanatory variables used to train the model. In this step, you will train the model using the county data and then apply that model to the same variables at the tract level and predict voter turnout.

- a** In the Contents pane, turn on the Tracts layer.

Note: Tracts that were not relevant to this analysis (for example, airports and national parks) were removed.

- b** Near the top of the Geoprocessing pane, update Prediction Type to Predict To Features.

Predict to features

Input Training Features
CountyElections2016_Refined

Variable to Predict
voter_turnout

- c Scroll down and update Input Prediction Features to Tracts.
- d For Output Predicted Features, type **Out_Predicted_Features**

Input Prediction Features
Tracts
























Output Predicted Features
Out_Predicted_Features

The prediction features must include the variables used to train the model, but the variables do not have to have the same name. You can use the Match Explanatory Variables to match the variables using their respective name.



- e Under Match Explanatory Variables, under Prediction, for the empty cells, click the down arrow and choose the matching Training variable name.

Prediction	Training
	2019 Education: High School/No Diploma : Percent
aggregationMethod	Median Age
2019 Median Age	Per Capita Income
2019 Per Capita Income	Diversity Index
2019 Education: < 9th Grade : Percent	Education: < 9th Grade : Percent
2019 Education: High School/No Diploma : Percent	Education: Associate's Degree : Percent
2019 Education: High School Diploma : Percent	


Your variables should match the following graphic that displays the completed list of matching explanatory variables.


Match Explanatory Variables Prediction 	Training
2019 Education: High School/N 	2019 Education: High School/No Dipl
2019 Median Age 	2019 Median Age
2019 Per Capita Income 	2019 Per Capita Income
2019 Diversity Index 	2019 Diversity Index
2019 Education: < 9th Grade : P 	2019 Education: < 9th Grade : Percent
2019 Education: Associate's Deg 	2019 Education: Associate's Degree : F
2019 Education: Bachelor's Deg 	2019 Education: Bachelor's Degree : Pe
2019 Education: GED : Percent 	2019 Education: GED : Percent
2019 Education: Grad/Professio 	2019 Education: Grad/Professional De
2019 Education: High School Di 	2019 Education: High School Diploma
2019 Education: Some College/ 	2019 Education: Some College/No De
2019 Pop Age 15+: Never Marri 	2019 Pop Age 15+: Never Married : Pe
2019 Population Age 65-69 : Pe 	2019 Population Age 65-69 : Percent
2019 Population Age 70-74 : Pe 	2019 Population Age 70-74 : Percent
2019 Population Age 75-79 : Pe 	2019 Population Age 75-79 : Percent
2019 Value: Checking/Savings/M 	2019 Value: Checking/Savings/Money
2019 Value: Stocks/Bonds/Mutu 	2019 Value: Stocks/Bonds/Mutual Func
ACS HHs: Inc Below Poverty Lev 	ACS HHs: Inc Below Poverty Level : Per
Buying American is not importa 	Buying American is not important to n
State Percent Votes Difference 	State Percent Votes Difference
State Voting Requirement Laws 	State Voting Requirement Laws
	

Note: Ensure that you match all the variables in the list.

-  Scroll down and, if necessary, expand Additional Outputs.
-  For Output Trained Features, delete Out_Trained_Features.

▼ **Additional Outputs**

Output Trained Features
 

⚠ Output Variable Importance Table
 

- h If necessary, expand Validation Options.
- i For Training Data Excluded For Validation, type **0**, and then click in the gray space of the tool so that the tool recognizes your update.

▼ **Validation Options**

Training Data Excluded for Validation (%)

Number of Runs for Validation

☒ Calculate Uncertainty

You are no longer assessing model performance, so you do not need to remove 10 percent of the training data for validation. Instead, you will use all training data to train the model so that the model can predict to the best of its ability.

- j Run the tool.

Note: The tool may take several minutes to run.

- k When the tool is complete, open the tool message window.

The R-squared value for validation data is no longer available because you are not using a validation subset. The Model Out Of Bag Errors uses training data to evaluate how well each tree in the model predicts, so you will focus on this metric.

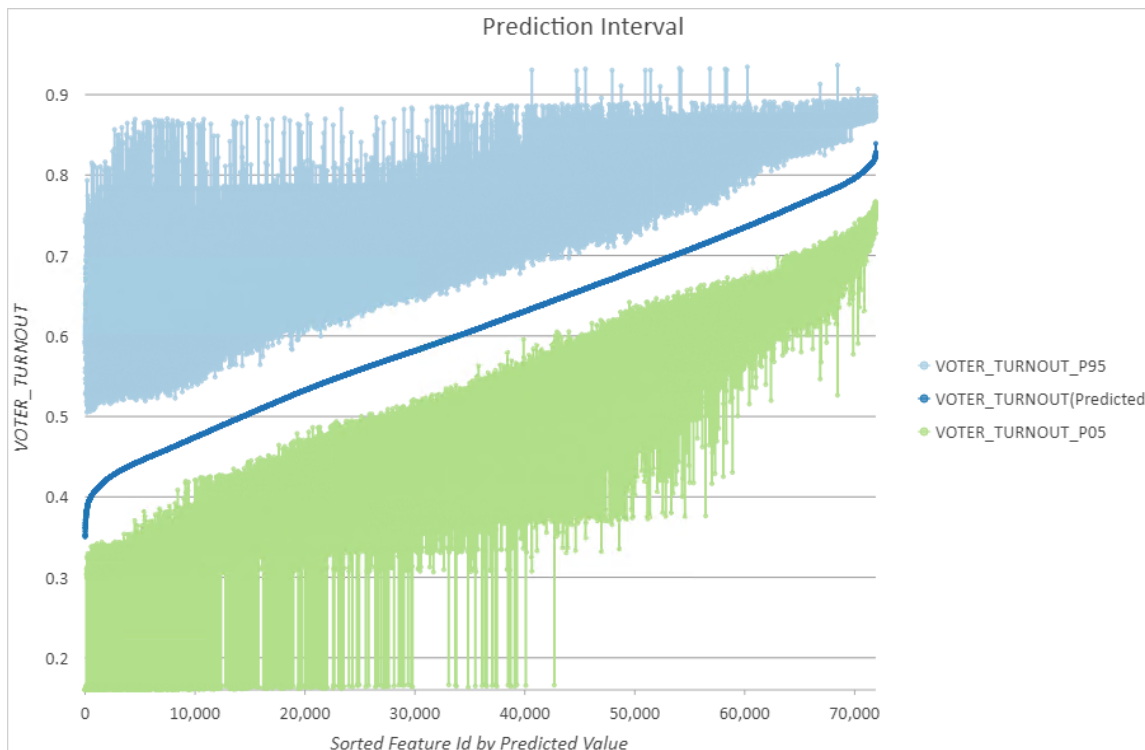
- l Scroll to the Model Out Of Bag Errors section.

```

----- Model Out of Bag Errors -----
Number of Trees           500           1000
MSE                       0.002           0.002
% of variation explained   71.322          71.574
  
```

The percentage of variation explained is still fairly high at 71 percent, and it does not vary greatly between 500 trees and 1000 trees. Because processing is not taking too long, you can keep 1000 as the Number Of Trees.

- m** Close the tool message window.
- n** In the Contents pane, under Out_Predicted_Features, right-click Prediction Interval and choose Open.



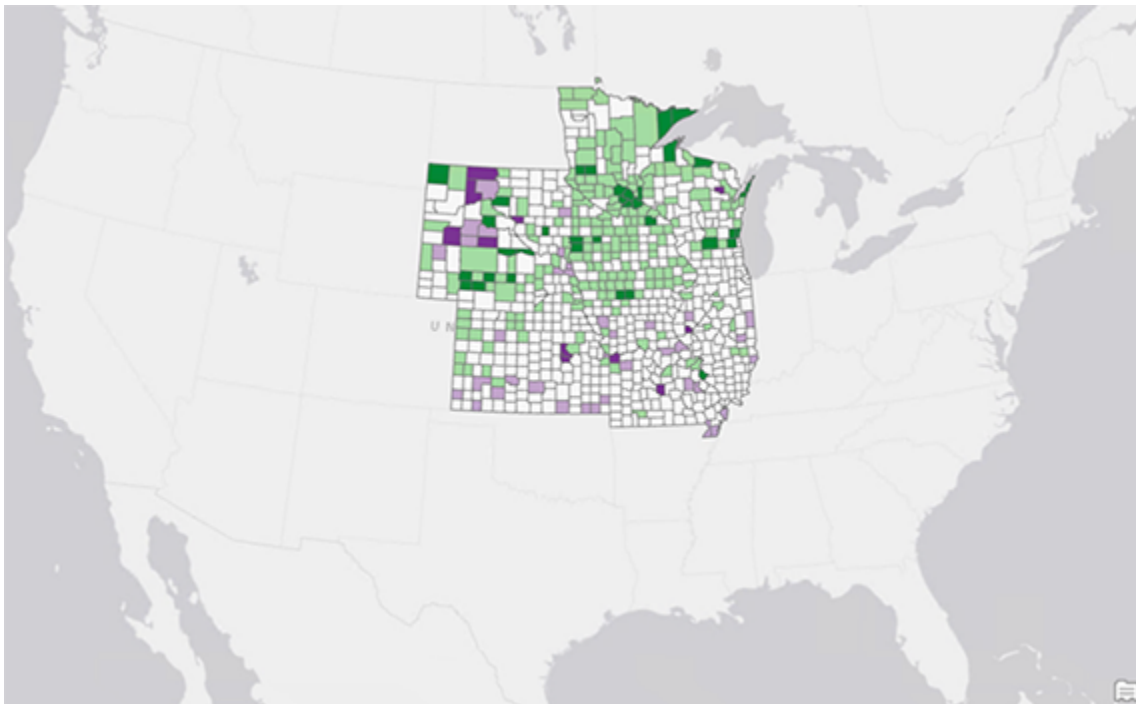
The confidence intervals are much larger for low voter turnout values than for high voter turnout values. This outcome indicates that the model is not as reliable for predicting low voter turnout as it is for predicting high voter turnout. Because the goal of your analysis is to identify areas with low voter turnout, this model is not reliable enough to meet your needs. The factors that drive voter turnout are likely very different from place to place, making it difficult to find a model that predicts well for the entire country. It is often good practice to reduce your study area and create more localized models. In the next step, you will attempt to model voter turnout in the state of Iowa.

- o** Close the Prediction Interval chart.
- p** In the Contents pane, turn off the Tracts layer.

Step 13: Change the scale of your analysis


Iowa is a competitive state, which means that voter turnout is important. You will perform a more localized analysis in the state of Iowa.

- a In the Contents pane, turn off the Out_Predicted_Features layer.
- b Turn on the CountyElections2016_NorthMidwest layer.




To localize your analysis, you will train your model using only upper Midwest county-level data and predict voter turnout at the Iowa tract level.

- c In the Geoprocessing pane, for Input Training Features, choose CountyElections2016_NorthMidwest.

Prediction Type
<input type="text" value="Predict to features"/>
Input Training Features
<input type="text" value="CountyElections2016_NorthMidwest"/> 
Variable to Predict
<input type="text" value="voter_turnout"/>

The Explanatory Training Variables will remain the same as the previous step.

- d For Input Prediction Features, choose Tracts_Iowa.

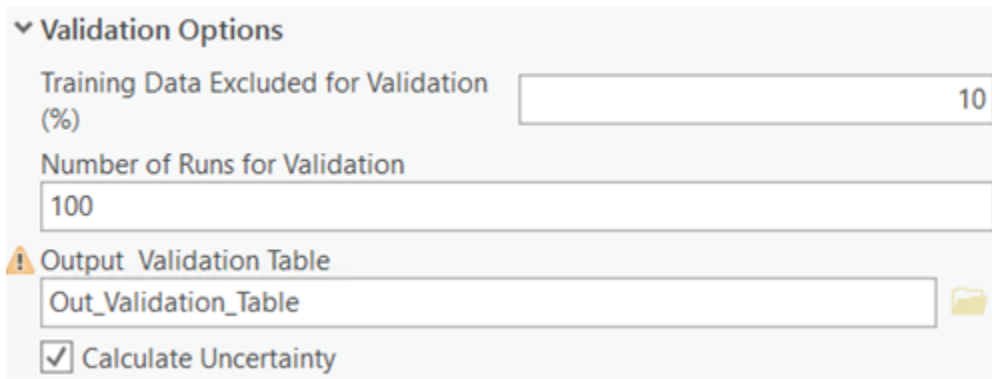


Input Prediction Features
Tracts_Iowa

Output Predicted Features
Out_Predicted_Features

- e Expand Validation Options, if necessary, and set the following parameters:

- Training Data Excluded For Validation: **10**
- Number Of Runs For Validation: **100**



Validation Options

Training Data Excluded for Validation (%) 10

Number of Runs for Validation 100

! Output Validation Table Out_Validation_Table

☒ Calculate Uncertainty

- f Run the tool.

Note: It may take several minutes for the tool to complete.

- g Review the Model Out Of Bag Errors.

Hint: Open the tool message window and scroll to the Model Out Of Bag Errors section.

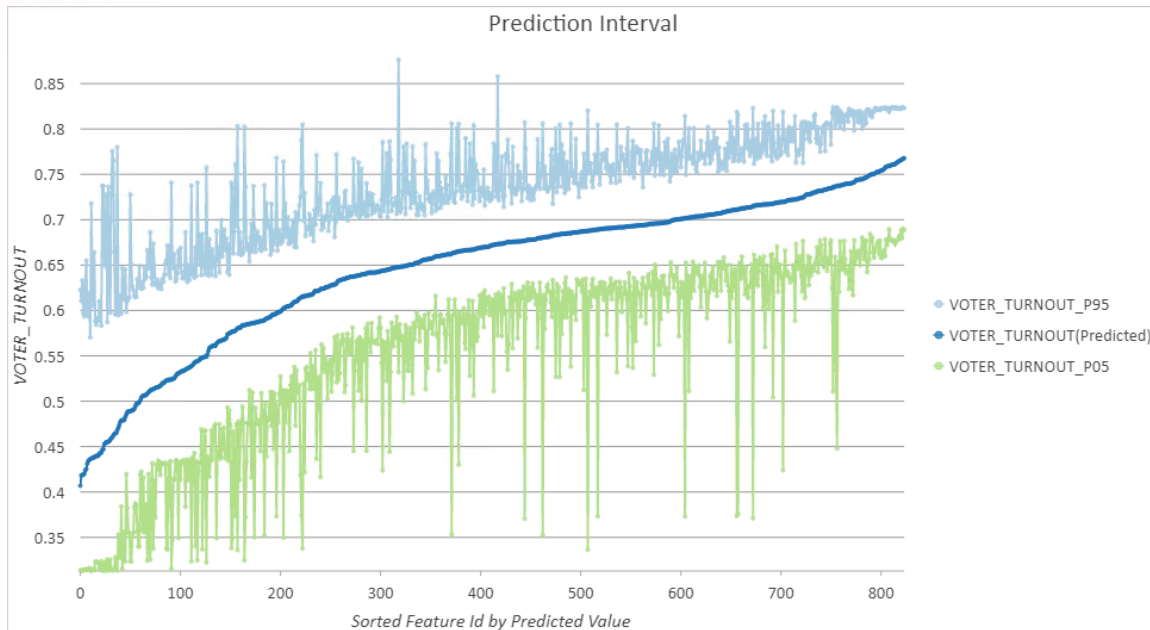
```
----- Model Out of Bag Errors -----
Number of Trees          500          1000
MSE                      0.002        0.002
% of variation explained  71.621     71.758
```

The percentage of variation explained is approximately 71 percent.

- h Close the tool's message window.

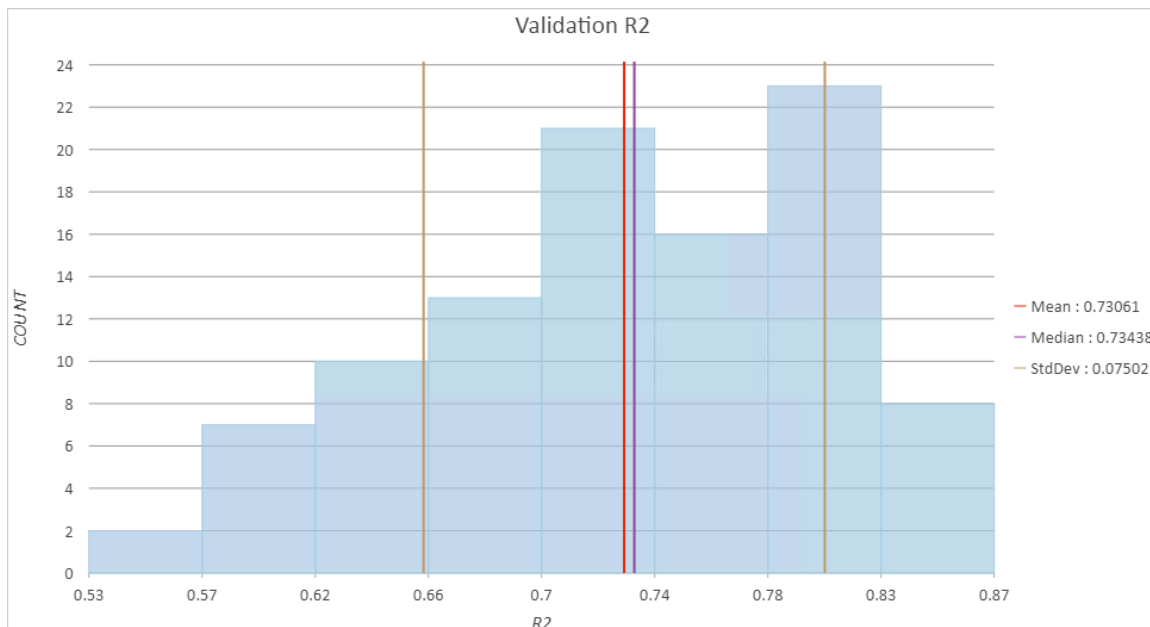
- i Open the Prediction Interval.

Hint: In the Contents pane, under Out_Predicted_Features, right-click Prediction Interval and choose Open.



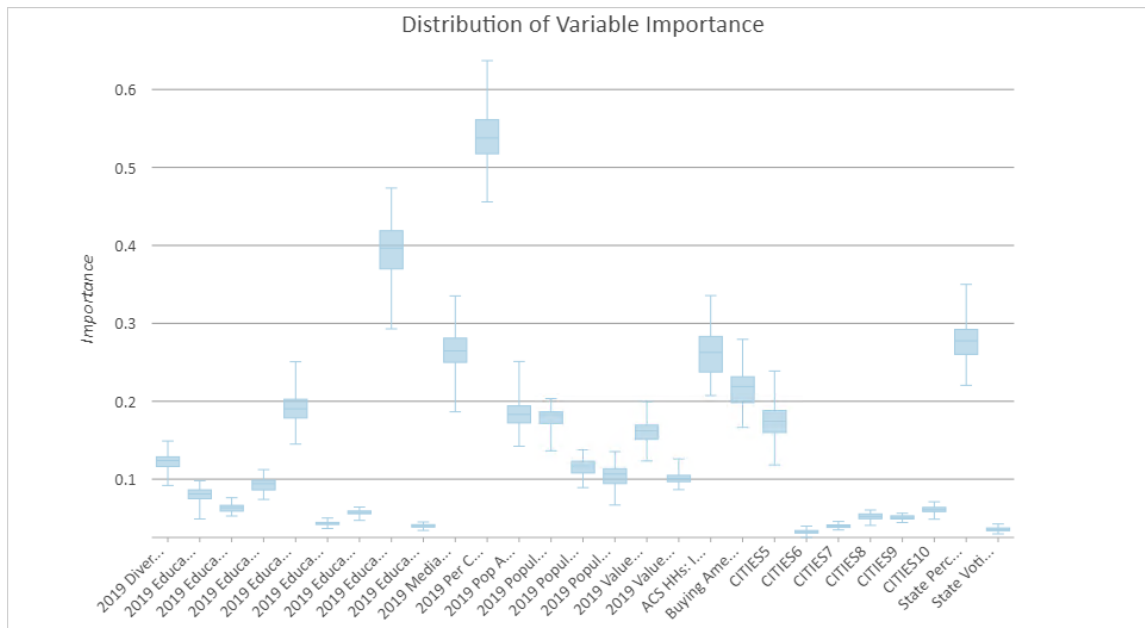
The confidence intervals are much smaller for low voter turnout values, indicating that the model has become more stable for predicting these values.

j Open the Validation R2 chart.



The distribution of R-squared values suggests that the model has stabilized R-squared values for the validation subset with a mean of approximately 0.73. You will review the variable importance chart to determine if there are any variables that you can remove to improve model performance.

- k** Open the Distribution Of Variable Importance chart.



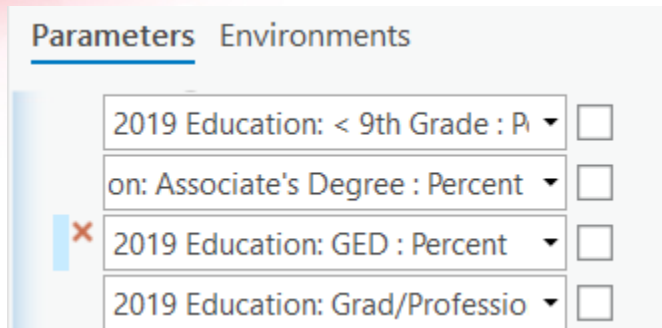
You will remove the variables of least importance and rerun the model to determine if these refinements improve model performance.

Note: Typically, this is an iterative process, where you remove some variables and then rerun and evaluate the model.

- l** Close all the charts, and then save the project.

Step 14: Refine the prediction

- a** In the Geoprocessing pane, under Explanatory Training Variables, click the 2019 Education: GED : Percent variable.
- b** To the left of the variable, click the red X that appears.



Parameters Environments













2019 Education: < 9th Grade : Percent	<input checked="" type="checkbox"/>
2019 Education: Associate's Degree : Percent	<input checked="" type="checkbox"/>
2019 Education: GED : Percent	<input type="checkbox"/>
2019 Education: Grad/Professional Degree : Percent	<input checked="" type="checkbox"/>

You have removed the 2019 Education: GED : Percent variable to refine the model. You will remove several other variables to further refine the model.

c Remove the following variables:

- 2019 Education: High School/No Diploma : Percent
- 2019 Diversity Index
- 2019 Education: Associate's Degree : Percent
- 2019 Education: Grad/Professional Degree : Percent
- 2019 Education: High School Diploma : Percent
- 2019 Education: Some College/No Degree : Percent
- 2019 Population Age 70-74 : Percent
- 2019 Population Age 75-79 : Percent
- 2019 Value: Stocks/Bonds/Mutual Funds : Average
- ACS HHs: Inc Below Poverty Level : Percent


Explanatory Training Variables





Variable 	Categorical
2019 Median Age 	<input type="checkbox"/>
2019 Per Capita Income 	<input type="checkbox"/>
2019 Education: < 9th Grade : Percent 	<input type="checkbox"/>
2019 Education: Bachelor's Degree : P 	<input type="checkbox"/>
2019 Pop Age 15+: Never Married : Pe 	<input type="checkbox"/>
2019 Population Age 65-69 : Percent 	<input type="checkbox"/>
2019 Value: Checking/Savings/Money 	<input type="checkbox"/>
Buying American is not important to r 	<input type="checkbox"/>
State Percent Votes Difference 	<input type="checkbox"/>
State Voting Requirement Laws 	<input checked="" type="checkbox"/>
	<input type="checkbox"/>

You are left with 10 explanatory training variables.

d Under Explanatory Training Distance Features, remove the following features:

- Cities10
- Cities9
- Cities7
- Cities6
- Cities5

Explanatory Training Distance Features 

Cities8 	
	

Cities8 is the only remaining explanatory training distance feature.

e Under Match Explanatory Variables, under Prediction, in the empty cell, click the down arrow and choose the matching Training variable name.

Match Explanatory Variables	
Prediction	Training
2019 Median Age	2019 Median Age
2019 Per Capita Income	2019 Per Capita Income
2019 Education: < 9th Grade : Percent	2019 Education: < 9th Grade : Percent
2019 Education: Bachelor's Degree : Percent	2019 Education: Bachelor's Degree : Percent
2019 Pop Age 15+: Never Married : Percent	2019 Pop Age 15+: Never Married : Percent
2019 Population Age 65-69 : Percent	2019 Population Age 65-69 : Percent
2019 Value: Checking/Savings/Money Mkt/CDs : Average	2019 Value: Checking/Savings/Money Mkt/CDs : Average
Buying American is not important to me : Percent	Buying American is not important to me : Percent
State Percent Votes Difference	State Percent Votes Difference
State Voting Requirement Laws	State Voting Requirement Laws

- f Run the tool.
- g When the tool is complete, open the tool message window.
- h Review the Model Out Of Bag Errors.

```

----- Model Out of Bag Errors -----
Number of Trees          500          1000
MSE                      0.002        0.002
% of variation explained  70.090    70.052
  
```

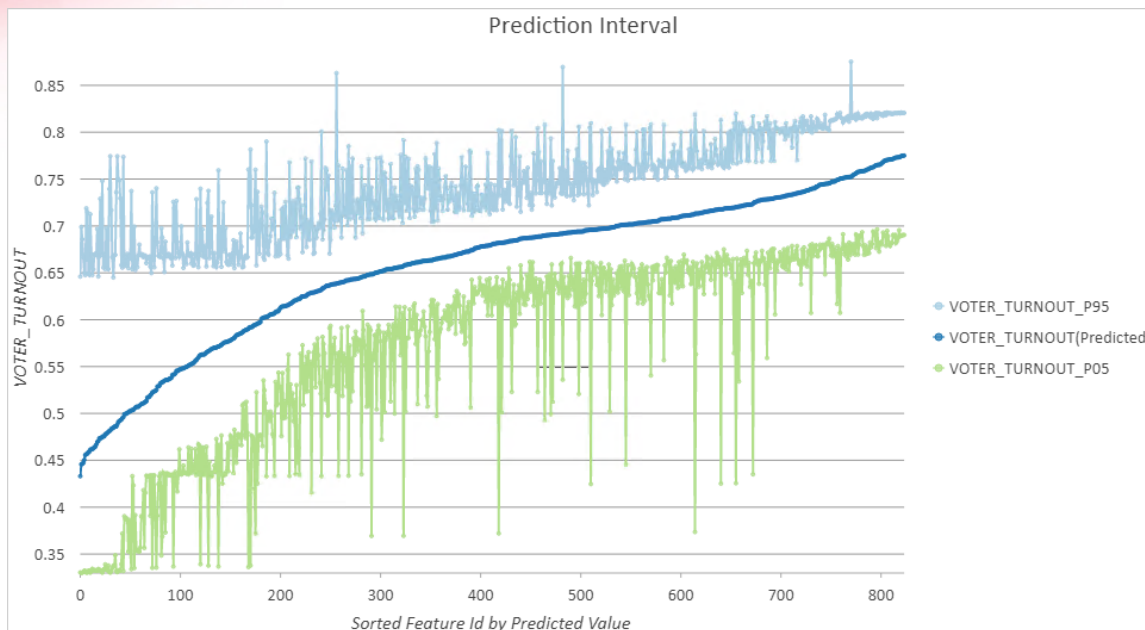
The percentage of variation explained is approximately 70 percent.

- i Review the Explanatory Variable Range Diagnostics.

----- Explanatory Variable Range Diagnostics -----									
Variable	Training		Validation		Prediction		Training	Validation	Prediction
	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Share(a)	Share(b)	Share(c)
2019 Median Age	25.00	55.90	29.30	54.40	18.10	59.40	1.00	0.81*	1.34+
2019 Per Capita Income	9395.00	46933.00	19303.00	39151.00	5132.00	67796.00	1.00	0.53*	1.67+
2019 Education: Bachelor's Degree : Percent	6.18	35.60	7.26	30.57	0.00	48.97	1.00	0.79*	1.66+
2019 Pop Age 15+: Never Married : Percent	10.40	59.81	14.54	46.54	11.43	99.96	1.00	0.65*	0.98+
2019 Population Age 65-69 : Percent	2.96	11.03	3.27	10.87	0.00	12.27	1.00	0.94*	1.52+

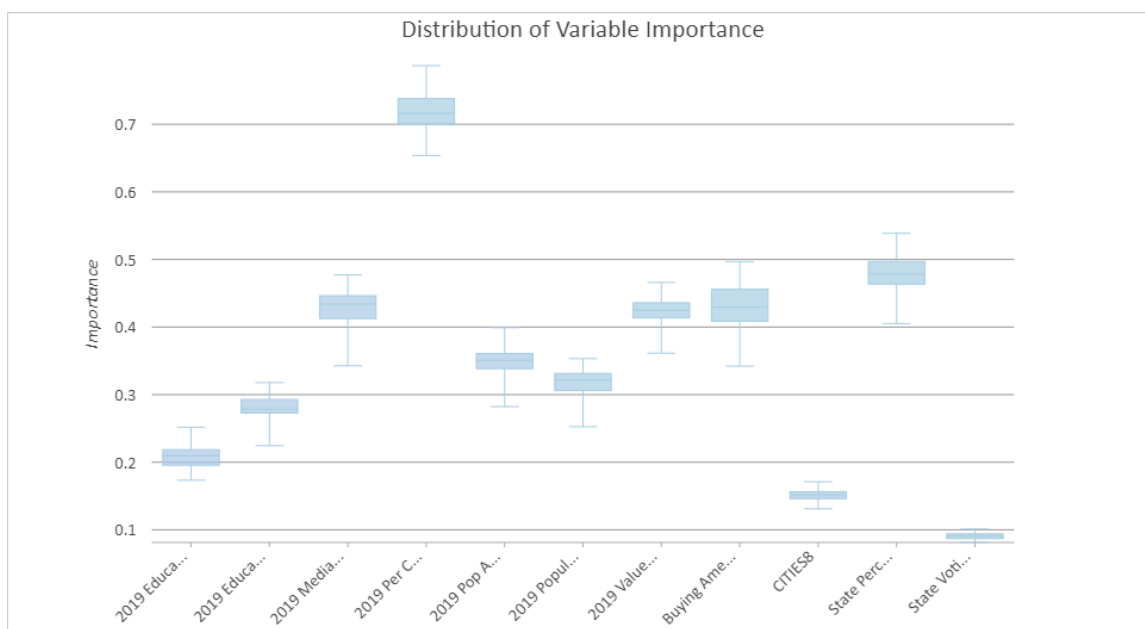
Some variables still have a larger prediction range (tract-level) than the training range (county-level). So, you will review the Prediction Interval to evaluate the model's stability.

- j Close the tool's message window.
- k Review the Prediction Interval.



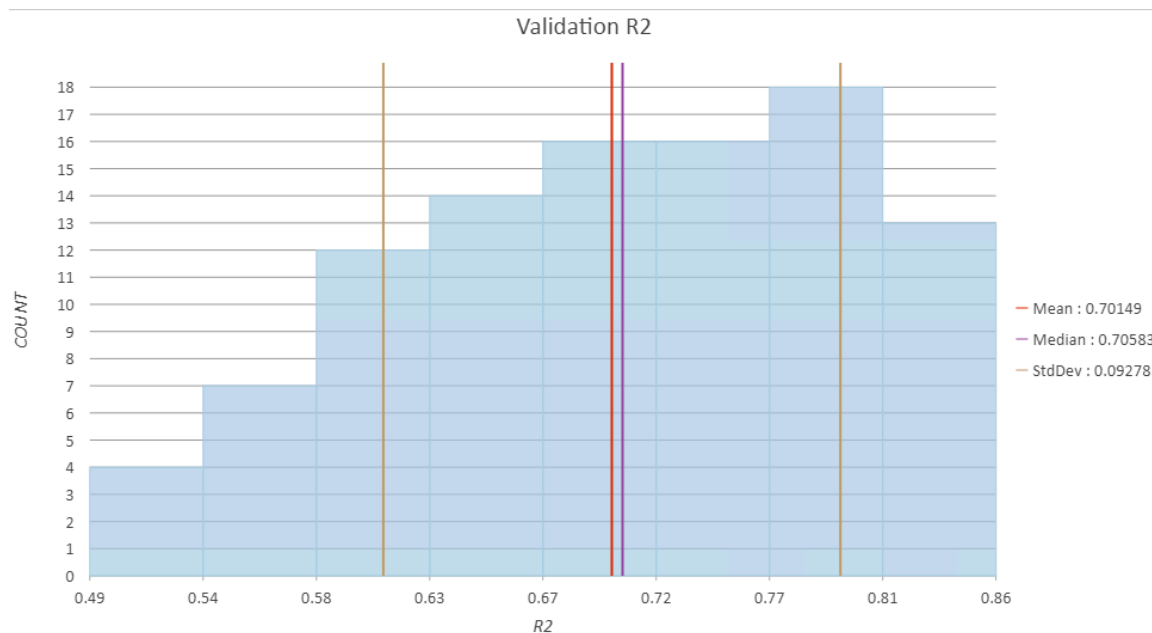
The confidence intervals are much smaller for low voter turnout values, indicating that the model is more reliable. You can also use the Distribution Of Variable Importance chart to evaluate the stability of the model.

I Open the Distribution Of Variable Importance chart.



You can evaluate the stability of a model by the length of the interquartile range of the explanatory variables. A large interquartile range (a long box) can indicate that the model is unstable because one variable can be more important in one run of the model compared to another. In this model, the interquartile ranges are fairly short, reconfirming that this model is stable.

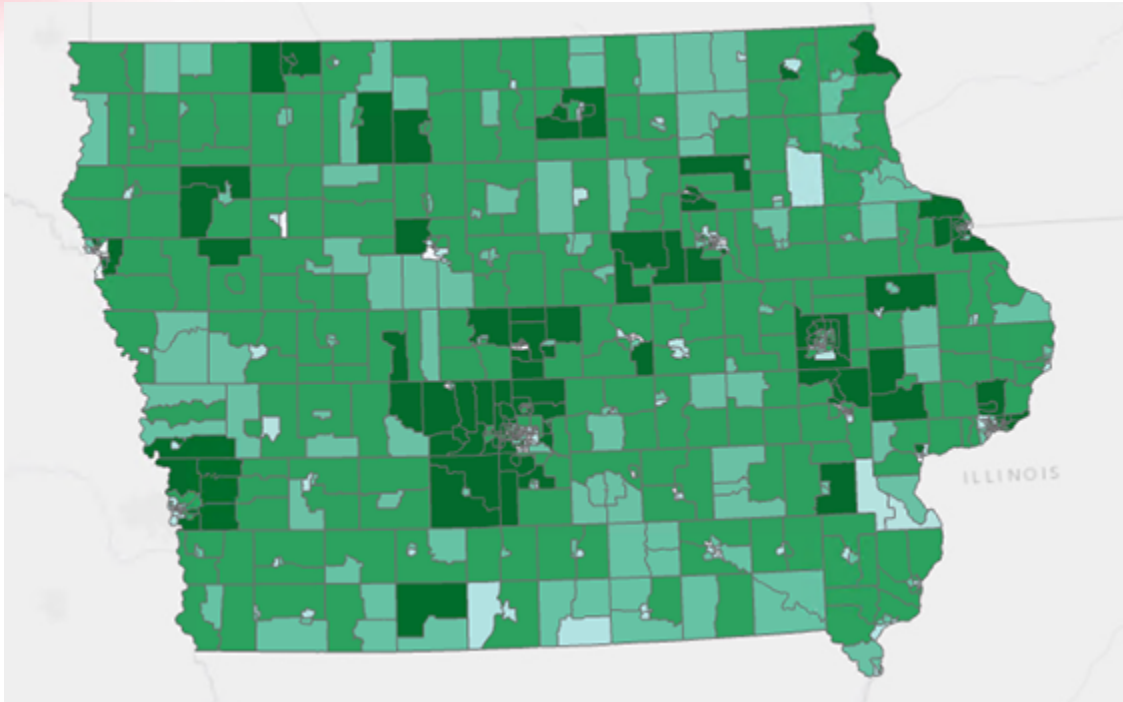
- m** Open the Validation R2 chart.



The distribution of R-squared values suggests that the model has stabilized R-squared values for the validation subset with a mean of approximately 0.70.

- n** Close the charts and the Chart Properties pane.
- o** In the Contents pane, turn off the CountyElections2016_NorthMidwest layer.
- p** Zoom to the Iowa bookmark.

Hint: Map tab > Navigate group > Bookmarks > Iowa bookmark



The Out_Predicted_Features layer contains the predicted voter turnout for Iowa. Exploring this map can help you identify the tracts with lower predicted voter turnout values and assess which regions of the state could be good locations for a campaign to get out the vote.

Overall, the model is performing well for this analysis question. You can continue modifying and improving the model (for example, changing the Training Data Excluded For Validation parameter to zero) or proceed with these results. Remember, your model will not be perfect. Your goal is to find a model that is useful for your objective, which, in this case, is a campaign to get out the vote.

- q If you would like to continue this analysis, proceed to the optional stretch goal; otherwise, save the project and exit ArcGIS Pro.

Stretch goal (optional)

The goal of prediction is to use the predicted values to make more informed decisions. If you would like to continue this analysis, you can apply the results of this prediction to organize a "Get Out the Vote" canvassing effort.

For this canvassing effort, you used the prediction model to identify a census tract in Florida that has low predicted voter turnout. You will assign 50 volunteers to a specific set of houses that they will visit to inform potential voters about an upcoming election.

Use the following high-level steps to continue this analysis:

1. Turn off the Out_Predicted_Features layer.
2. Zoom to the Naples, FL bookmark.
3. From the Prediction.gdb, add the AddressPointsCanvassing feature class to the map.
4. Use the Build Balanced Zones (Spatial Statistics Tools) tool to define 50 zones. One volunteer will be assigned to each zone.
5. Share your results in the Lesson Forum, using the **#stretch** hashtag in the posting title.

Note: The tool may complete with warnings after it runs. Open the View Details window to review the warnings and decide if these warnings will affect your analysis.

Use the Lesson Forum to post your questions and observations. Be sure to include the **#stretch** hashtag in the forum posting title.

When you are finished, save the project and exit ArcGIS Pro.

Answers to Exercise Questions

1. What is the validation R-squared value for this model?

Hint: In the Forest-Based Classification And Regression tool message window, scroll to the Validation Data: Regression Diagnostics section.

The R-squared value is approximately 0.74, meaning that the model predicted voter turnout in the validation data with an accuracy of about 74 percent.

2. Compared to the other variables, is State Percentage Votes Difference a helpful variable in this model?

Hint: In the Contents pane, open the Distribution Of Variable Importance chart.

Comparatively, the State Percent Votes Difference is one of the more helpful variables in this model. This outcome indicates that knowing how competitive each state is can help in predicting voter turnout.

3. What is the validation R-squared value for this model?

The R-squared value is approximately 0.74, meaning that the model predicted voter turnout in the validation data with an accuracy of about 74 percent.

4. What is the mean R-squared value over the 10 runs of this model?

Hint: In the Contents pane, open the Validation R2 chart.

The R-squared values for each run of the model range from 0.63 to 0.77, with a mean value of 0.722.

5. What are the two most important explanatory variables in this model?

The two most important variables are 2019 Education: High School/No Diploma : Percent and 2019 Per Capita Income.