Exercise 1: Create a prediction model

ℹ How can I print an exercise to PDF format?

> **Technical note**
>
> If you exit ArcGIS Pro during this exercise, you will have to restart the exercise from the beginning. It is best to complete the exercise all at once.

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.

Scenario

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

You will use this model to downscale voter turnout from the county level to the census tract level. This information will be used to organize a "Get Out the Vote" canvassing campaign. These campaigns encourage people 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.

Note: The exercises in this course include View Result links. Click these links to confirm that your results match what is expected.

Estimated completion time in minutes: 75 minutes

Expand all steps ▼ Collapse all steps ▲

- Step 1: Download the exercise data files

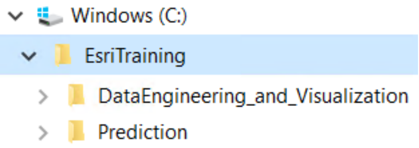
In this step, you will download the exercise data files.

a Open a new web browser tab or window.

b Go to https://links.esri.com/Section02/Data 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 the EsriTraining folder on your local computer.



Windows (C:)
EsriTraining
DataEngineering_and_Visualization
Prediction

*Step 1c***: Download the exercise data files.*

You downloaded and extracted the exercise data files that you will need to complete this section of the MOOC.

- Step 2: Open an ArcGIS Pro project

In this step, you will open the ArcGIS Pro project that you downloaded.

a Start ArcGIS Pro.

b If necessary, sign in using your course ArcGIS account username and password.

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

c On the Start page, near Recent Projects, click Open Another Project.

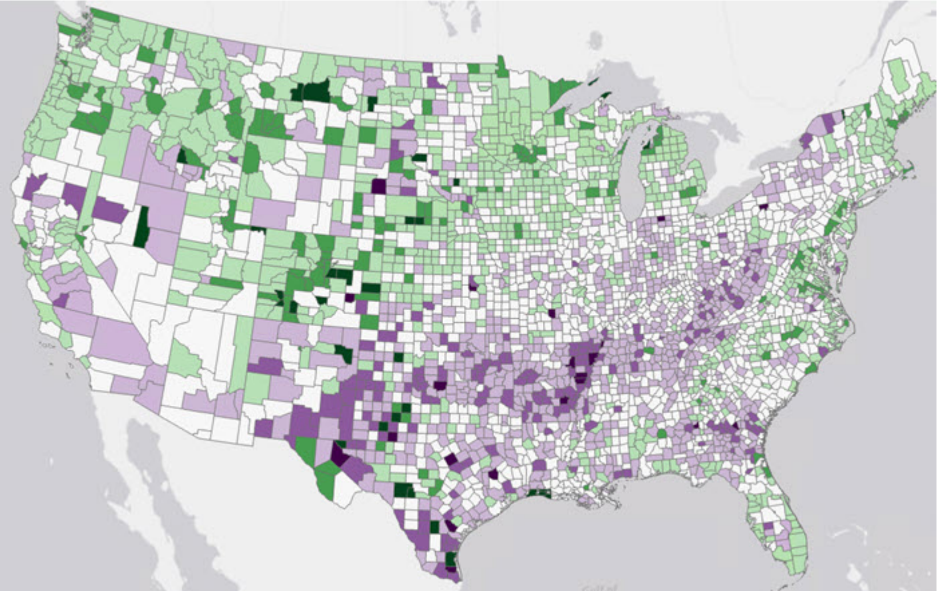
Note: If you have configured ArcGIS Pro to start without a project template or with a default project, you will not see the Start page. On the Project tab, click Open, and then click Open Another Project.

d In the Open Project dialog box, browse to the Prediction folder that you saved on your computer.

- Hint

..\EsriTraining\Prediction

e Click Prediction.aprx to select it, and then click OK.



*Step 2e***: Open an ArcGIS Pro project.*

A Prediction map tab opens to a gray basemap with a map layer that represents the 2020 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 previous data engineering exercise, you learned how to enrich 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.

In this step, you will create a prediction model using the Forest-based Classification and Regression tool.

a From the Analysis tab, in the Geoprocessing group, click Tools.

b In the Geoprocessing pane, search for **Forest-based Classification And Regression**.

c In the search results, click Forest-based Classification And Regression (Spatial Statistics Tools).

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: CountyElections2020
- Variable To Predict: Voter_Turnout_2020

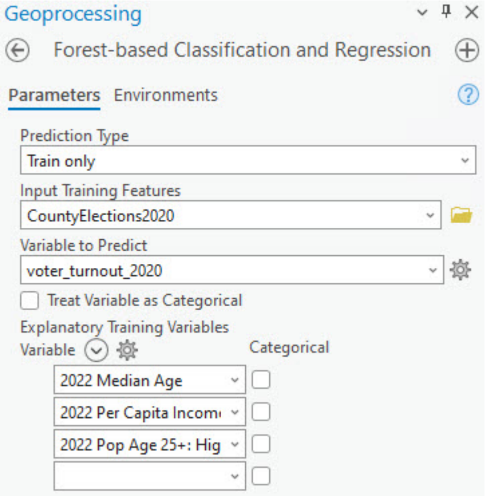
e Under Explanatory Training Variables, next to Variable, click the Add Many button ⌄.

f In the Variable window, check the boxes for the following variables:

- 2022 Median Age
- 2022 Per Capita Income

- 2022 Pop Age 25+: High School/No Diploma: Percent

g In the bottom-right corner of the Variable window, click Add.



Forest-based Classification and Regression

Prediction Type: Train only

Input Training Features: CountyElections2020

Variable to Predict: voter_turnout_2020

☐ Treat Variable as Categorical

Explanatory Training Variables

Variable	Categorical
2022 Median Age	☐
2022 Per Capita Incom	☐
2022 Pop Age 25+: Hig	☐
	☐

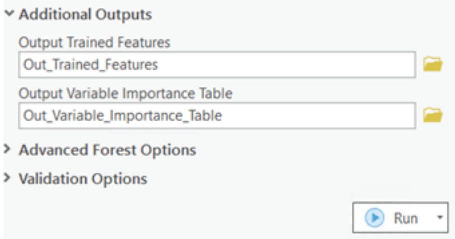
*Step 3g***: Create a prediction model.*

The selected variables are added to the tool's parameters.

h In the Geoprocessing pane, expand Additional Outputs.

i Under Additional Outputs, specify the following parameters:

- Output Trained Features: **Out_Trained_Features**
- Output Variable Importance Table: **Out_Variable_Importance_Table**



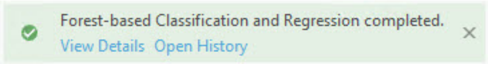
Additional Outputs

Output Trained Features: Out_Trained_Features

Output Variable Importance Table: Out_Variable_Importance_Table

*Step 3i***: Create a prediction model.*

j In the Geoprocessing pane, click Run.

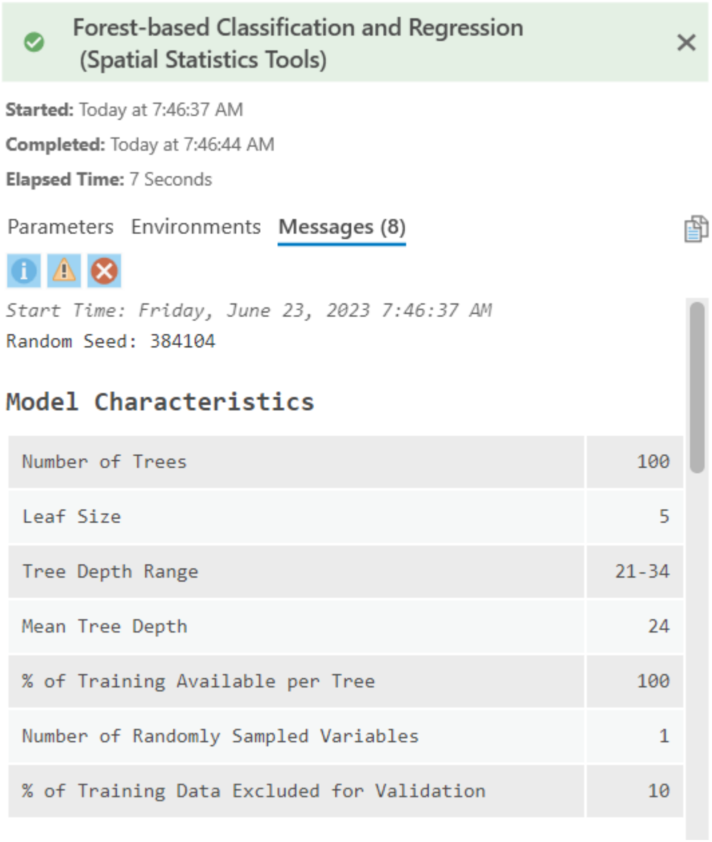


Forest-based Classification and Regression completed.

*Step 3j***: Create a prediction model.*

You created a prediction model by running the tool. At the bottom of the Geoprocessing pane, you will see a message confirming that the tool completed successfully. To understand how the model performed, you will review the model's performance metric using the tool messages.

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



Forest-based Classification and Regression (Spatial Statistics Tools)

Started: Today at 7:46:37 AM

Completed: Today at 7:46:44 AM

Elapsed Time: 7 Seconds

Start Time: Friday, June 23, 2023 7:46:37 AM

Random Seed: 384104

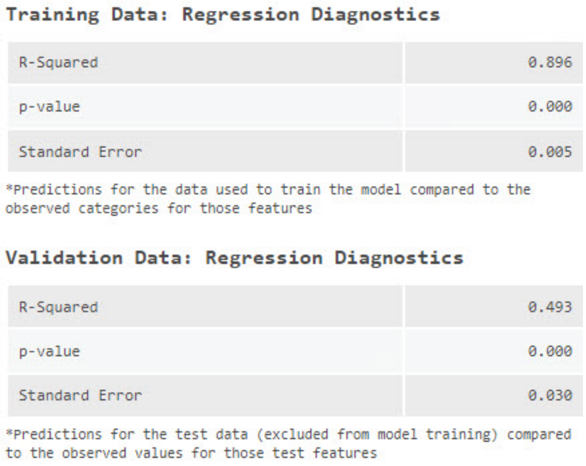
Model Characteristics

Number of Trees	100
Leaf Size	5
Tree Depth Range	21-34
Mean Tree Depth	24
% of Training Available per Tree	100
Number of Randomly Sampled Variables	1
% of Training Data Excluded for Validation	10

*Step 3k***: Create a prediction model.*

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

l On the Messages tab, scroll down and locate the Training Data: Regression Diagnostics section.



Training Data: Regression Diagnostics

R-Squared	0.896
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	0.493
p-value	0.000
Standard Error	0.030

*Predictions for the test data (excluded from model training) compared to the observed values for those test features

*Step 3l***: Create a prediction model.*

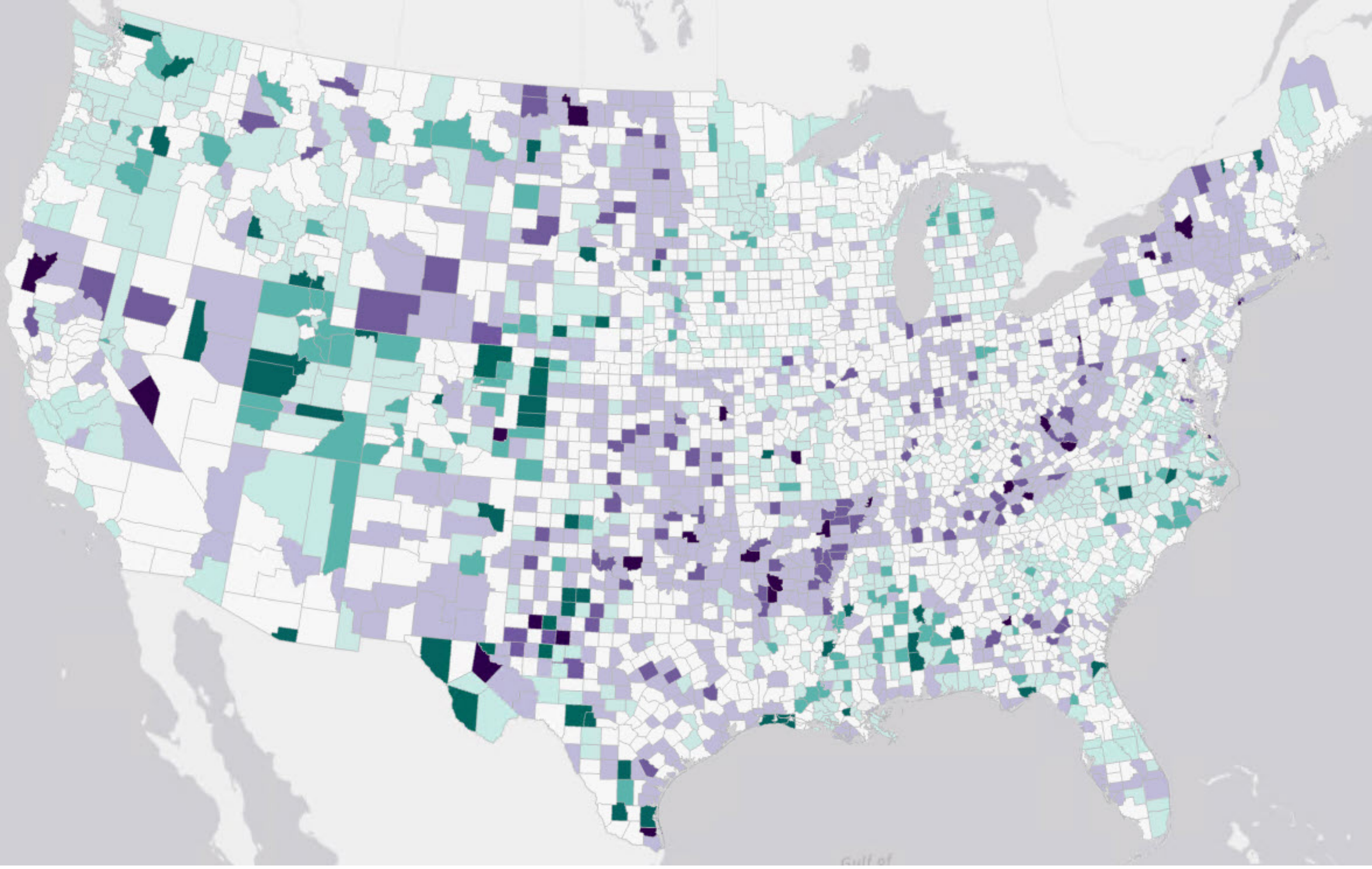
Note: Each time that you run the Forest-based Classification and Regression tool, you may get slightly different results due to the randomness introduced in the algorithm to prevent the model from overfitting to the training data.

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 that measures how well the model performed 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 0.493, indicating that the model predicted the voter turnout value in the validation dataset with an accuracy of about 49 percent.

m Close the Forest-based Classification And Regression tool message window.

n In the Contents pane, turn off the CountyElections2020 layer.



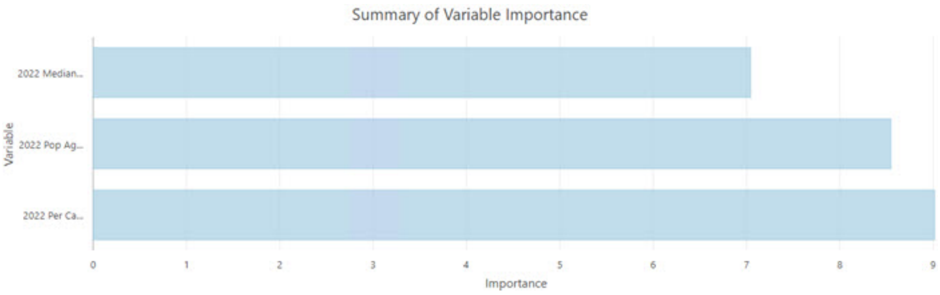
*Step 3n***: Create a prediction model.*

You will review how important each explanatory variable was in generating a prediction. 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.

o In the Contents pane, open the Summary Of Variable Importance chart.

- Hint

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



*Step 3o***: Create a prediction model.*

The 2022 Per Capita Income and 2022 Pop Age 25+: 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 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.

p Close the Summary Of Variable Importance chart.

To learn more about the Forest-based Classification and Regression (Spatial Statistics) tool, go to ArcGIS Pro Help: Forest-based Classification and Regression (Spatial Statistics).

Note: 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 exercise.

- Step 4: Examine model stability

In this step, you will review the prediction intervals in this model to see whether the model's performance is relatively stable across all values.

a Return to the Geoprocessing pane.

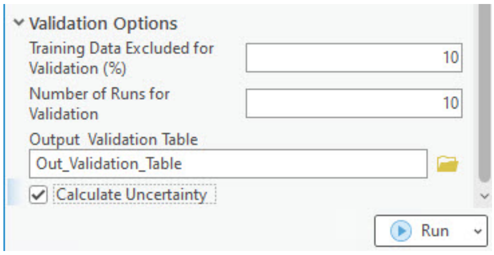
b Near the bottom of the Geoprocessing pane, expand Validation Options.

c For Number Of Runs For Validation, type **10**, and then click in the empty gray space of the Geoprocessing pane so that the tool recognizes your update.

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

d For Output Validation Table, type **Out_Validation_Table**.

e Under Output Validation Table, click the check box for Calculate Uncertainty.



Validation Options

Training Data Excluded for Validation (%): 10

Number of Runs for Validation: 10

Output Validation Table: Out_Validation_Table

☑ Calculate Uncertainty

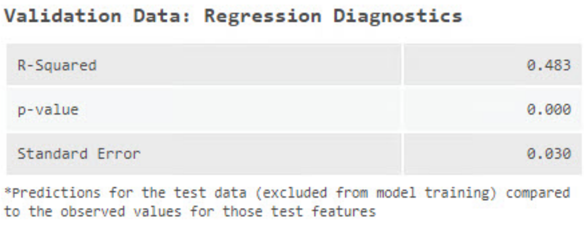
*Step 4e***: Examine model stability.*

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

f Run the tool.

g At the bottom of the Geoprocessing pane, click View Details.

h In the tool message window, from the Messages tab, scroll to the Validation Data: Regression Diagnostics section.



Validation Data: Regression Diagnostics

R-Squared	0.483
p-value	0.000
Standard Error	0.030

*Predictions for the test data (excluded from model training) compared to the observed values for those test features

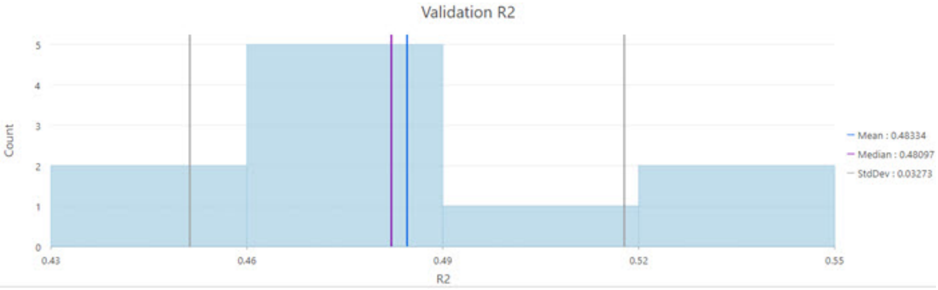
*Step 4h***: Examine model stability.*

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.

The tool trained 10 models with random subsets of validation data. In the example shown here, the most representative R-squared value across the 10 runs is 0.483, corresponding to about 48 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.

i Close the tool message window.

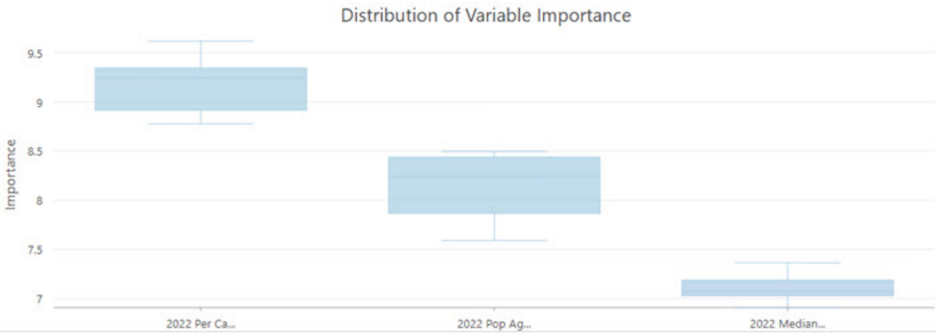
j In the Contents pane, in the Standalone Tables section, open the Validation R2 chart.



*Step 4j***: Examine model stability.*

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 0.483.

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

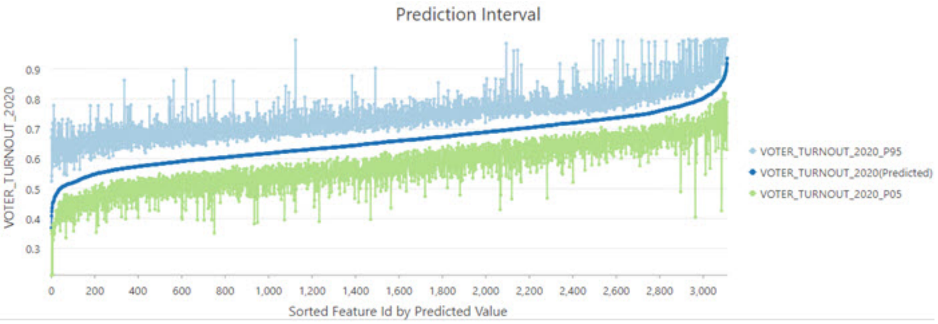


*Step 4k***: Examine model stability.*

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. In some runs of the model, Per Capita Income was more important, and in other runs, Pop Age 25+: High School/No Diploma: Percent was also important. Overall, both variables are strong candidates for your predictive model.

To learn more about the Distribution Of Variable Importance box plot, go to ArcGIS Pro Help: Output messages and diagnostics.

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



*Step 4l***: Examine model stability.*

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 whether 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.

To learn more about the Prediction Interval chart, go to ArcGIS Pro Help: Advanced forest options.

At this point, you have used the attributes in your dataset as the explanatory variables in your model.

m Close the charts and activate the Geoprocessing pane.

- Step 5: Add many variables to a prediction model

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, that 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.

In this step, you will add variables to your model to see if they enhance your model's performance for predicting voter turnout.

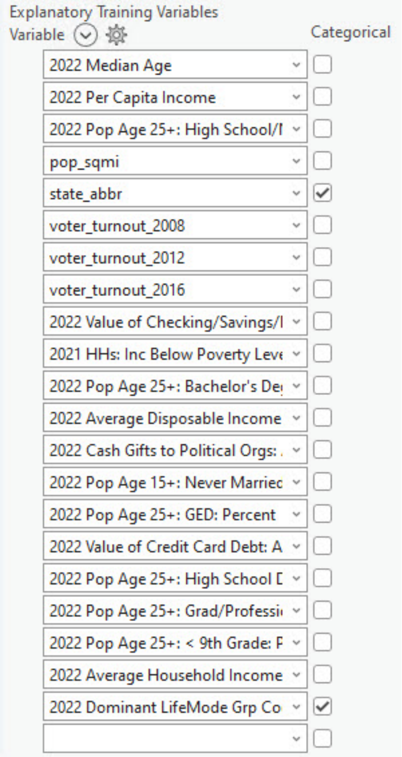
a In the Geoprocessing pane, under Explanatory Training Variables, click the Add Many button ⌄.

b Check the boxes for the following variables to add them to your model:

- Pop_Sqmi
- State_Abbr
- Voter_Turnout_2008
- Voter_Turnout_2012
- Voter_Turnout_2016

- 2022 Value Of Checking/Savings/Money Mkt/CDs: Average
- 2021 HHs: Inc Below Poverty Level (ACS 5-Yr): Percent
- 2022 Pop Age 25+: Bachelor's Degree: Percent
- 2022 Average Disposable Income
- 2022 Cash Gifts To Political Orgs: Average
- 2022 Pop Age 15+: Never Married: Percent
- 2022 Pop Age 25+: GED: Percent
- 2022 Value Of Credit Card Debt: Average
- 2022 Pop Age 25+: High School Diploma: Percent
- 2022 Pop Age 25+: Grad/Professional Degree: Percent
- 2022 Pop Age 25+: < 9th Grade: Percent
- 2022 Average Household Income
- 2022 Dominant LifeMode Grp Code

c Click Add.



Explanatory Training Variables

Variable	Categorical
2022 Median Age	□
2022 Per Capita Income	□
2022 Pop Age 25+: High School/I	□
pop_sqmi	□
state_abbr	☑
voter_turnout_2008	□
voter_turnout_2012	□
voter_turnout_2016	□
2022 Value of Checking/Savings/I	□
2021 HHs: Inc Below Poverty Leve	□
2022 Pop Age 25+: Bachelor's De	□
2022 Average Disposable Income	□
2022 Cash Gifts to Political Orgs: .	□
2022 Pop Age 15+: Never Marriec	□
2022 Pop Age 25+: GED: Percent	□
2022 Value of Credit Card Debt: A	□
2022 Pop Age 25+: High School D	□
2022 Pop Age 25+: Grad/Professi	□
2022 Pop Age 25+: < 9th Grade: F	□
2022 Average Household Income	□
2022 Dominant LifeMode Grp Co	☑
	□

*Step 5c***: Add many variables to a prediction model.*

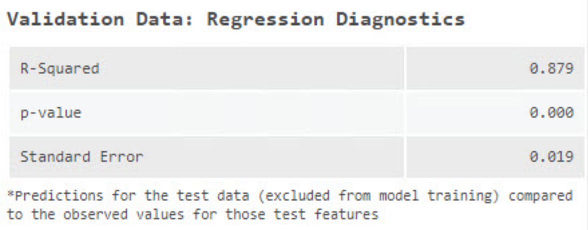
Many variables are now added as explanatory training variables. Voter turnout for 2020 is the variable that you are trying to understand, and the same variable should not be used as both the independent and dependent variables for this analysis. The Forest-based Classification and Regression tool can also use categorical variables, which are variables of a string field type instead of a numeric field type. The State_Abbr and the 2022 Dominant LifeMode Grp Code variables are both marked as categorical variables in the Explanatory Training Variables list.

d Run the tool.

e Review the R-squared value in the validation data regression diagnostics.

- Hint

In the Forest-based Classification And Regression tool message window, from the Messages tab, scroll to the Validation Data: Regression Diagnostics section.



Validation Data: Regression Diagnostics

R-Squared	0.879
p-value	0.000
Standard Error	0.019

*Predictions for the test data (excluded from model training) compared to the observed values for those test features

*Step 5e***: Add many variables to a prediction model.*

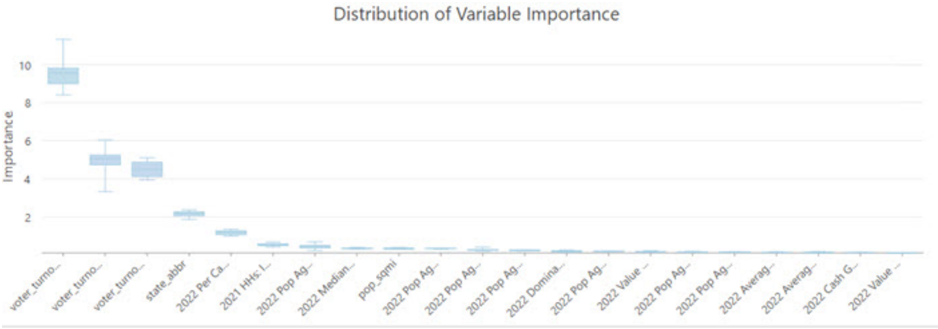
In this example, the model's R-squared value increased to 0.879, which means that it is now predicting with almost 88 percent accuracy based on the validation data. Remember: you may see slightly different results due to the randomness introduced in the algorithm to prevent the model from overfitting to the training data. The warning that shows in the tool message window is because the State_Abbr categorical variable contains more than 15 categories. The more categories that a variable contains, the more likely it is that the variable will dominate the model and lead to less effective prediction results. When you select variables for your model, consider this effect on results.

f Close the tool message window.

g Review the Distribution Of Variable Importance chart.

- Hint

Under Charts, right-click Distribution Of Variable Importance and choose Open.



*Step 5g***: Add many variables to a prediction model.*

Note: You can zoom in to the chart using the Zoom Mode button 🔍 or your mouse to better see the distribution for a particular variable. (If you do not see the Zoom Mode button, widen the chart view.)

The voter turnout variables 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.

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 distance variables and assess their importance to the model.

h Close the Distribution Of Variable Importance chart and activate the Geoprocessing pane.

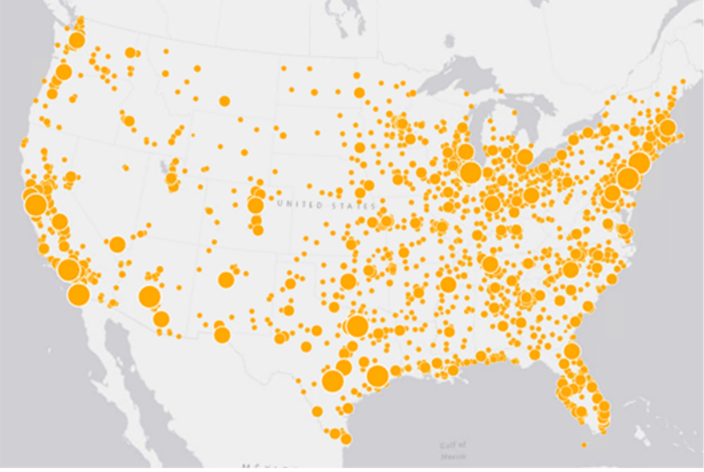
- Step 6: Add distance variables to the model

You want to incorporate each county's urban and rural characteristics into the model to determine whether 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

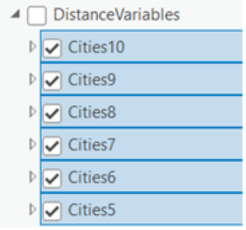


*Step 6b***: Add distance variables to the model.*

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.

d Click Cities10 to select it, and then press Shift and click Cities5.

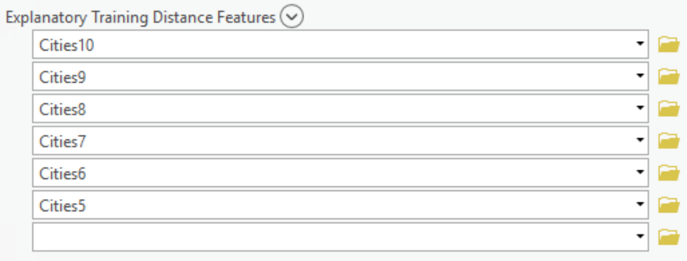


□ DistanceVariables
- ☑ Cities10
- ☑ Cities9
- ☑ Cities8
- ☑ Cities7
- ☑ Cities6
- ☑ Cities5

*Step 6d***: Add distance variables to the model.*

The six distance variables are selected.

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



Explanatory Training Distance Features
- Cities10
- Cities9
- Cities8
- Cities7
- Cities6
- Cities5

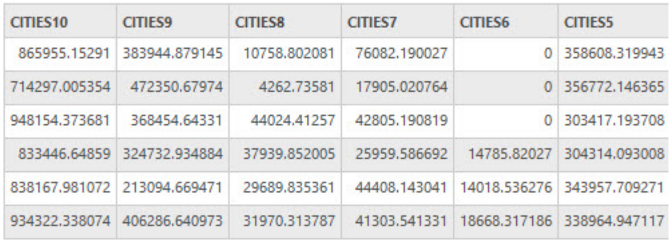
*Step 6e***: Add distance variables to the model.*

f Click Run.

Note: The tool may take a few minutes to 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
865955.15291	383944.879145	10758.802081	76082.190027	0	358608.319943
714297.005354	472350.67974	4262.73581	17905.020764	0	356772.146365
948154.373681	368454.64331	44024.41257	42805.190819	0	303417.193708
833446.64859	324732.934884	37939.852005	25959.586692	14785.82027	304314.093008
838167.981072	213094.669471	29689.835361	44408.143041	14018.536276	343957.709271
934322.338074	406286.640973	31970.313787	41303.541331	18668.317186	338964.947117

*Step 6h***: Add distance variables to the model.*

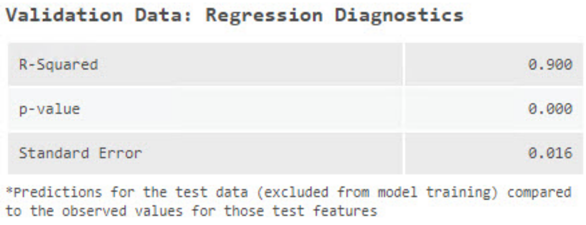
Note: When a city point is contained within a county, the distance will be zero.

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, from the Messages tab, scroll to the Validation Data: Regression Diagnostics section.



Validation Data: Regression Diagnostics

R-Squared	0.900
p-value	0.000
Standard Error	0.016

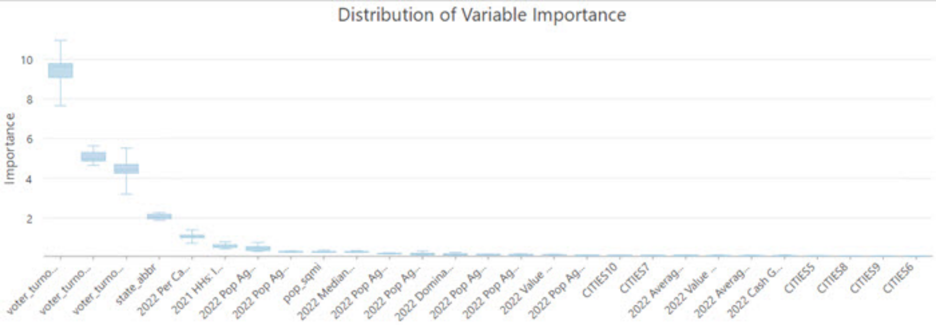
*Predictions for the test data (excluded from model training) compared to the observed values for those test features

*Step 6k***: Add distance variables to the model.*

In this example, the model's R-squared value increased to 0.900, which means that it is predicting with about 90 percent accuracy based on the validation data. You will review the variable importance chart to see how influential each distance variable is for the model performance. Remember that you may have slightly different results due to the randomness introduced into the algorithm.

l Close the tool message window.

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



*Step 6m***: Add distance variables to the model.*

The distance to Cities7 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 voter turnout variables.

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 that is greater than 0.70 might be considered satisfactory for making a prediction. Before using this model to predict, you will simplify the model to include only the most important variables.

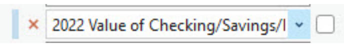
n Close the Distribution Of Variable Importance chart and activate the Geoprocessing pane.

- Step 7: Refine the model

Refining your model is typically an iterative process, where you remove some variables and then rerun and evaluate the model. There are many different ways to select variables to include in a model. In this analysis, the most important variables are chosen by defining a threshold in the variable importance table. In this step, you will refine the model to only include the variables with variable importance above the selected threshold.

a In the Contents pane, turn off the Out_Trained_Features layer.

b In the Geoprocessing pane, hover your mouse over the 2022 Value Of Checking/Savings/Money Mkt/CDs: Average variable.



× 2022 Value of Checking/Savings/I ▾ □

*Step 7b***: Refine the model.*

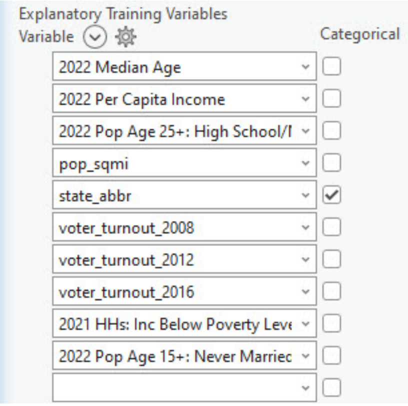
A red X appears next to the variable. When you click the red X, the variable is removed from the list.

c Click the red X to the left of the 2022 Value Of Checking/Savings/Money Mkt/CDs: Average variable.

d Click the red X for the following variables to remove them from the model:

- 2022 Pop Age 25+: Bachelor's Degree: Percent
- 2022 Average Disposable Income
- 2022 Cash Gifts To Political Orgs: Average

- 2022 Pop Age 25+: GED: Percent
- 2022 Value Of Credit Card Debt: Average
- 2022 Pop Age 25+: High School Diploma: Percent
- 2022 Pop Age 25+: Grad/Professional Degree: Percent
- 2022 Pop Age 25+: < 9th Grade: Percent
- 2022 Average Household Income
- 2022 Dominant LifeMode Grp Code



Explanatory Training Variables

Variable	Categorical
2022 Median Age	☐
2022 Per Capita Income	☐
2022 Pop Age 25+: High School/I	☐
pop_sqmi	☐
state_abbr	☑
voter_turnout_2008	☐
voter_turnout_2012	☐
voter_turnout_2016	☐
2021 HHs: Inc Below Poverty Leve	☐
2022 Pop Age 15+: Never Marriec	☐
	☐

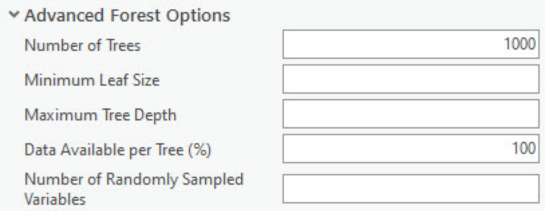
*Step 7d***: Refine the model.*

You refined the model by removing the explanatory training variables that had the least importance for the model. You will also remove all of the distance variables because of their lower variable importance scores.

e Under Explanatory Training Distance Features, remove all of the distance variables.

f Scroll down and expand Advanced Forest Options.

g 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	

*Step 7g***: Refine the model.*

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.

h Run the tool.

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

i Review the tool message window and various charts to answer the following questions.

Note: Remember that each time you run the Forest-based Classification and Regression tool, you may get slightly different results due to the randomness introduced into the algorithm.

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

- Answer

If the R-squared value is approximately 0.981, the model predicted voter turnout in the validation data with an accuracy of about 98 percent.

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

- Answer

If the R-squared values for each run of the model range from about 0.81 to 0.91, then the mean value is approximately 0.87.

- Hint

In the Contents pane, open the Validation R2 chart.

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

- Answer

The two most important variables are Voter_Turnout_2016 and Voter_Turnout_2012.

The simplified model has a higher R-squared value, meaning that removing the variables with low importance did not compromise model performance and enhanced the model's performance.

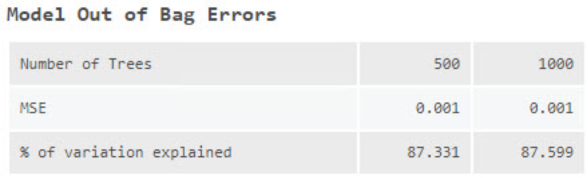
j Close the charts.

- Step 8: Examine additional model metrics

Next, you will review additional model metrics to help you assess whether the model requires any additional changes.

a If necessary, in the Geoprocessing pane, reopen the Forest-based Classification And Regression tool message window.

b On the Messages tab, review the Model Out Of Bag Errors section.



Model Out of Bag Errors

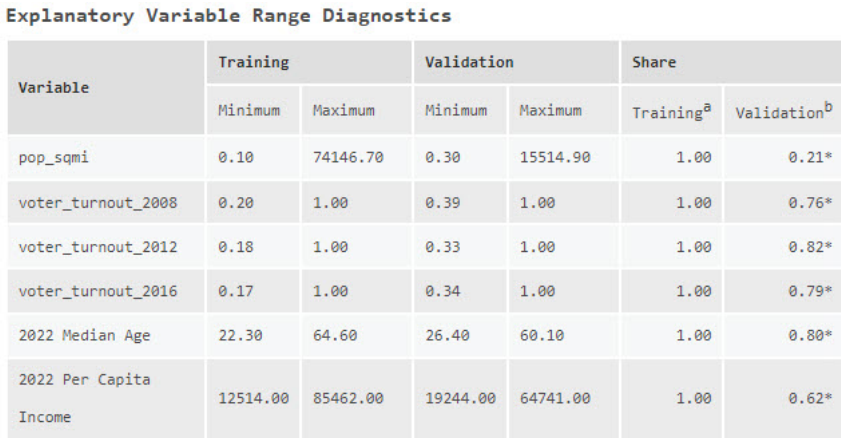
Number of Trees	500	1000
MSE	0.001	0.001
% of variation explained	87.331	87.599

*Step 8b***: Examine additional model metrics.*

Model Out Of Bag Errors is another diagnostic that can help validate the model. The percentage of variation explained indicates the percentage 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. Remember: due to the randomness built into algorithms, varied results are expected.

c Scroll to the Explanatory Variable Range Diagnostics section.



Explanatory Variable Range Diagnostics

Variable	Training		Validation		Share	
	Minimum	Maximum	Minimum	Maximum	Training[a]	Validation[b]
pop_sqmi	0.10	74146.70	0.30	15514.90	1.00	0.21*
voter_turnout_2008	0.20	1.00	0.39	1.00	1.00	0.76*
voter_turnout_2012	0.18	1.00	0.33	1.00	1.00	0.82*
voter_turnout_2016	0.17	1.00	0.34	1.00	1.00	0.79*
2022 Median Age	22.30	64.60	26.40	60.10	1.00	0.80*
2022 Per Capita Income	12514.00	85462.00	19244.00	64741.00	1.00	0.62*

*Step 8c***: Examine additional model metrics.*

Note: You may need to widen the window to see the full table in this section.

The Explanatory Variable Range Diagnostics section lists the range of values covered by each explanatory variable in the datasets that are used to train and validate the model. For example, median age values spanned from 22 to 64 in the dataset that is used to train the model and from 26 to 60 in the dataset that is used to validate the model.

The Share column indicates the percentage of overlap between the values that are used to train and the values that are used to validate. In this example, 80 percent of the median age values that are used to train the model were used to validate the model. A value that is greater than 1 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, go to ArcGIS Pro Help: Output messages and diagnostics.

d Close the tool message window.

e In the Contents pane, turn off the Out_Trained_Features layer.

- Step 9: Predict values

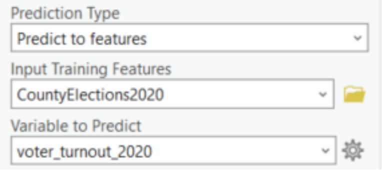
You trained a model using the county data that you had available. You can use this model to predict voter turnout at the census tract level, which is much higher resolution and will allow you to get a sense of more detailed spatial patterns.

To predict voter turnout at the tract level, you need census tract data with explanatory variables that match the explanatory variables that are 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 to 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, for Prediction Type, choose Predict To Features.

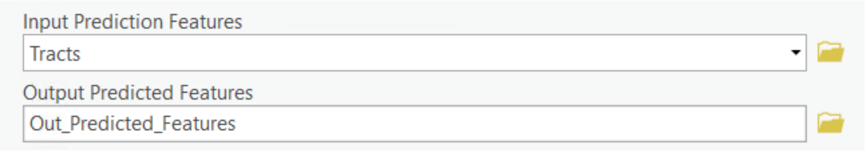


Prediction Type
Predict to features
Input Training Features
CountyElections2020
Variable to Predict
voter_turnout_2020

*Step 9b***: Predict values.*

c Scroll down the Geoprocessing pane and for Input Prediction Features, choose Tracts.

d For Output Predicted Features, type **Out_Predicted_Features**.



Input Prediction Features
Tracts
Output Predicted Features
Out_Predicted_Features

*Step 9d***: Predict values.*

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

e Under Match Explanatory Variables, under Prediction, for the empty cell next to 2022 Median Age, click the down arrow.

f Review the available variables in the Tracts layer and answer the question.

? Which variables used to train the model at the county level are not included in the Tracts layer?

- Answer

The Tracts layer does not include voter turnout data.

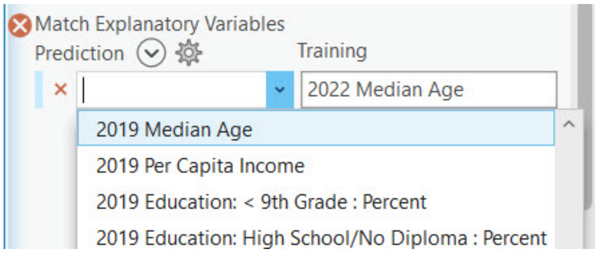
The Tracts layer was provided to you by a different analyst and did not go through the same data engineering steps that you completed. Because of that, you will delete the voter turnout variables from Explanatory Training Variables.

g Scroll up in the Geoprocessing pane.

h Under Explanatory Training Variables, remove the following variables:

- Voter_Turnout_2008
- Voter_Turnout_2012
- Voter_Turnout_2016

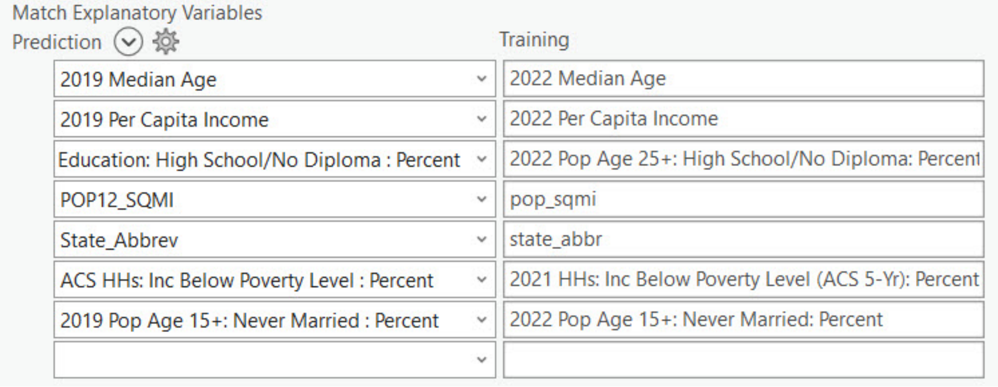
i Under Match Explanatory Variables, under Prediction, for the empty cells, click the down arrow and choose the matching training variable name.



Match Explanatory Variables
Prediction
Training
2022 Median Age
2019 Median Age
2019 Per Capita Income
2019 Education: < 9th Grade : Percent
2019 Education: High School/No Diploma : Percent

*Step 9i***: Predict values.*

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



Match Explanatory Variables

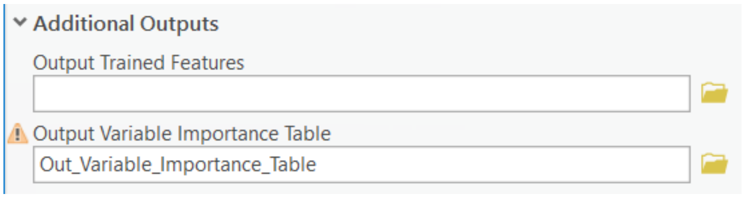
Prediction	Training
2019 Median Age	2022 Median Age
2019 Per Capita Income	2022 Per Capita Income
Education: High School/No Diploma : Percent	2022 Pop Age 25+: High School/No Diploma: Percent
POP12_SQMI	pop_sqmi
State_Abbrev	state_abbr
ACS HHs: Inc Below Poverty Level : Percent	2021 HHs: Inc Below Poverty Level (ACS 5-Yr): Percent
2019 Pop Age 15+: Never Married : Percent	2022 Pop Age 15+: Never Married: Percent

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

For the purpose of this analysis, using 2019 variables is acceptable because the difference between 2019 values and 2021 or 2022 values is minimal.

j Scroll down and, if necessary, expand Additional Outputs.

k For Output Trained Features, delete Out_Trained_Features.



Additional Outputs

Output Trained Features

Output Variable Importance Table

Out_Variable_Importance_Table

*Step 9k***: Predict values.*

l If necessary, expand Validation Options.

m For Training Data Excluded For Validation, confirm that it is still set to 10.

By removing the voter turnout variables, you changed the model. Because of this change, you will leave Training Data Excluded For Validation as 10 percent so you can assess model performance at the Tract level. If the variables had remained the same, you no longer need to assess the model's performance and you could change this parameter to 0 to use all the training data to train the model so that the model can predict to the best of its ability.

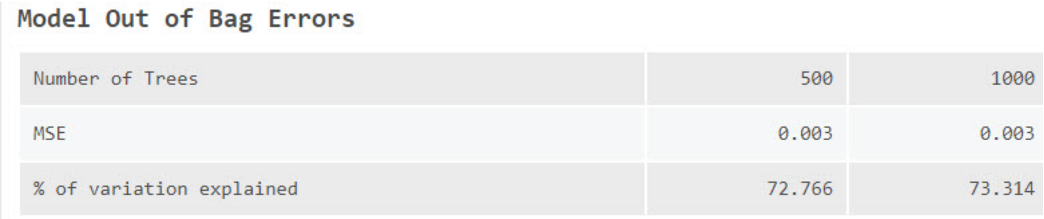
n Run the tool.

Note: The tool may take a few minutes to run.

o When the tool is complete, open the tool message window.

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.

p Scroll to the Model Out Of Bag Errors section.



Model Out of Bag Errors

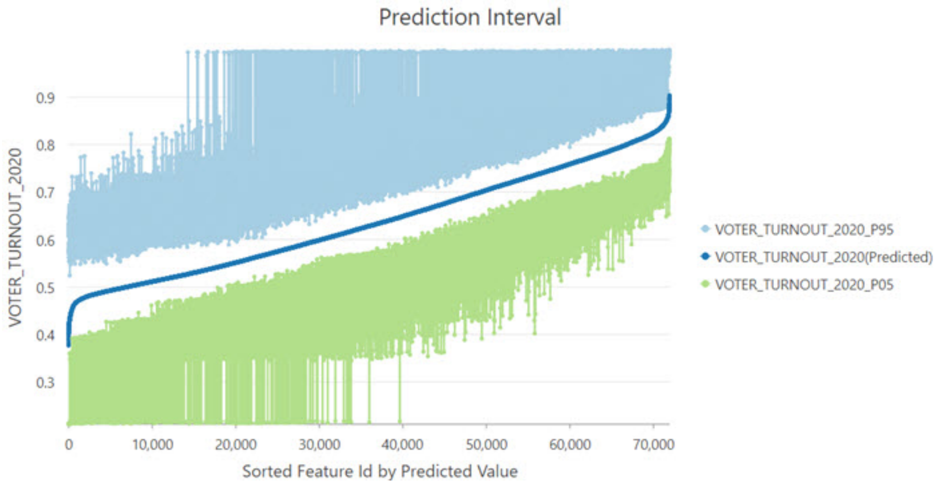
Number of Trees	500	1000
MSE	0.003	0.003
% of variation explained	72.766	73.314

*Step 9p***: Predict values.*

The percentage of variation explained was reduced, but it does not vary greatly between 500 trees and 1,000 trees. Because it does not vary much between 500 and 1000, there is no need to run the model again with more trees.

q Close the tool message window.

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



*Step 9r***: Predict values.*

The wide confidence intervals show that the model does not perform as well at the tract level as it did at the county level. In particular, it may be prone to overestimating turnout for low-turnout tracts. Because the goal of your analysis is to identify areas with low voter turnout, this model is not reliable enough to meet your needs.

When you change the scale of your analysis, you need to change or update your model. In this scenario, votes in the United States are mostly calculated at the county level and are not calculated at the tract level. You saw that tract-level prediction does not work nationwide. 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 a good practice to reduce your study area and create more localized models.

You can end your analysis here or you can continue modifying and improving the model by reducing your study area to create a more localized model. 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.

s If you would like to continue this analysis, proceed to the optional stretch goal; otherwise, close all the charts, save the project, and exit ArcGIS Pro.

- Step 10: Stretch goal (Optional)

For this stretch goal, you will use what you learned in this exercise to change the scale of your analysis and refine the model by identifying variables of least importance and adjusting the tool's parameters. To localize your analysis, you will train your model using only upper Midwest county-level data. Then, you will rerun the model to determine whether these refinements improved model performance for predicting voter turnout in the state of Iowa. Remember: refining your model is typically an iterative process, where you remove some variables and then rerun and evaluate the model.

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

1. In the Geoprocessing pane, set Input Training Features to CountyElections_IA_And_Neighbors.
2. Set Variable To Predict to Voter_Turnout.
3. Select the Explanatory Training Variables that you want to use for your model.
4. Add any Explanatory Training Distance Features that might affect the model.
5. For Input Prediction Features, choose Tracts_Iowa.
6. For Prediction, match the explanatory variables.
7. Run the model.
8. Review the tool message window and the charts to determine whether you improved the model's performance.

Note: If necessary, the project contains a bookmark that zooms to the state of Iowa.

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

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