Cardiovascular Death Rates in the United States

and an illustration in the importance of a good feature set...

Zachary Wimpee

November 9, 2020

A National Health Crisis

In the United States, heart disease is...

A National Health Crisis

In the United States, heart disease is...

- the cause of about **1 in every 4 deaths**.[1]

[1] CDC. URL: https://www.cdc.gov/heartdisease/facts.htm.

A National Health Crisis

In the United States, heart disease is...

- the cause of about **1 in every 4 deaths**.[1]
- the **leading cause of death**, with a mortality rate of 200 deaths per 100,000 population.[2]

[1] CDC,

[2] NCHS. URL: https://www.cdc.gov/nchs/fastats/heart-disease.htm.

A National Health Crisis

In the United States, heart disease is...

- the cause of about **1 in every 4 deaths**.[1]
- the **leading cause of death**, with a mortality rate of 200 deaths per 100,000 population.[2]
- the leading cause of death for almost **every population demographic**.

[1] CDC,

[2] NCHS,

Existing Approaches

There are a variety of resources and information available related to cardiovascular issues and how to prevent them, such as the *ABCS of Heart Health* guide from the Million Hearts initiative.[3]

[3] HHS. URL: https://millionhearts.hhs.gov/data-reports/factsheets/ABCS.html.

Existing Approaches

There are a variety of resources and information available related to cardiovascular issues and how to prevent them, such as the *ABCS of Heart Health* guide from the Million Hearts initiative.[3]

A. take aspirin as directed

[3] HHS,

Existing Approaches

There are a variety of resources and information available related to cardiovascular issues and how to prevent them, such as the *ABCS of Heart Health* guide from the Million Hearts initiative.[3]

A. take aspirin as directed

B. control blood pressure

[3] HHS,

Existing Approaches

There are a variety of resources and information available related to cardiovascular issues and how to prevent them, such as the *ABCS of Heart Health* guide from the Million Hearts initiative.[3]

A. take aspirin as directed

B. control blood pressure

C. manage cholesterol

[3] HHS,

Existing Approaches

There are a variety of resources and information available related to cardiovascular issues and how to prevent them, such as the *ABCS of Heart Health* guide from the Million Hearts initiative.[3]

A. take aspirin as directed

B. control blood pressure

C. manage cholesterol

S. don't smoke

[3] HHS,

Project Motivation

In the United States cardiovascular health issues are widespread.

Project Motivation

The *ABCS of Heart Health* and similar resources focus on an individual's lifestyle choices.

Project Motivation

Project Goal

By examining data on the **population level**, can a similar set of **recommendations** be made for large-scale systematic changes that would help **prevent increases in cardiovascular death rates**?

Project Outline

1. Compile dataset of cardiovascular death rates and population variables.

Project Outline

1. Compile dataset of cardiovascular death rates and population variables.
2. Create a model to predict changes in cardiovascular death rates for the populations.

Project Outline

1. Compile dataset of cardiovascular death rates and population variables.
2. Create a model to predict changes in cardiovascular death rates for the populations.
3. Determine which variables were most important in making the best predictions.

Project Outline

1. Compile dataset of cardiovascular death rates and population variables.
2. Create a model to predict changes in cardiovascular death rates for the populations.
3. Determine which variables were most important in making the best predictions.
4. Propose recommendations for preventing increases in cardiovascular death rates from these results.

Data Wrangling - CDC

2010-2018 cardiovascular mortality data by county obtained using the **CDC WONDER**[4] interface:

[4] WONDER. URL: https://wonder.cdc.gov/ucd-icd10.html.

Data Wrangling - CDC

2010-2018 cardiovascular mortality data by county obtained using the **CDC WONDER**[4] interface:

- county population

[4] WONDER,

Data Wrangling - CDC

2010-2018 cardiovascular mortality data by county obtained using the **CDC WONDER**[4] interface:

- county population
- cardiovascular related fatality count

[4] WONDER,

Data Wrangling - CDC

2010-2018 cardiovascular mortality data by county obtained using the **CDC WONDER**[4] interface:

- county population
- cardiovascular related fatality count
- estimated crude death rate (deaths per 100,000 population)

[4] WONDER,

Data Wrangling - Census

2010-2018 census data by county obtained using the **Census Data API**[5]:

[5] Census. URL: https://www.census.gov/data/developers/data-sets.html.

Data Wrangling - Census

2010-2018 census data by county obtained using the **Census Data API**[5]:

- educational attainment rates

[5] Census,

Data Wrangling - Census

2010-2018 census data by county obtained using the **Census Data API**[5]:

- educational attainment rates
- economic characteristic variables

[5] Census,

Data Wrangling - Census

2010-2018 census data by county obtained using the **Census Data API**[5]:

- educational attainment rates
- economic characteristic variables
- housing costs as a proportion of income

[5] Census,

Data Wrangling - Census

2010-2018 census data by county obtained using the **Census Data API**[5]:

- educational attainment rates
- economic characteristic variables
- housing costs as a proportion of income

[5] Census,

Data Wrangling - Census

2010-2018 census data by county obtained using the **Census Data API**[5]:

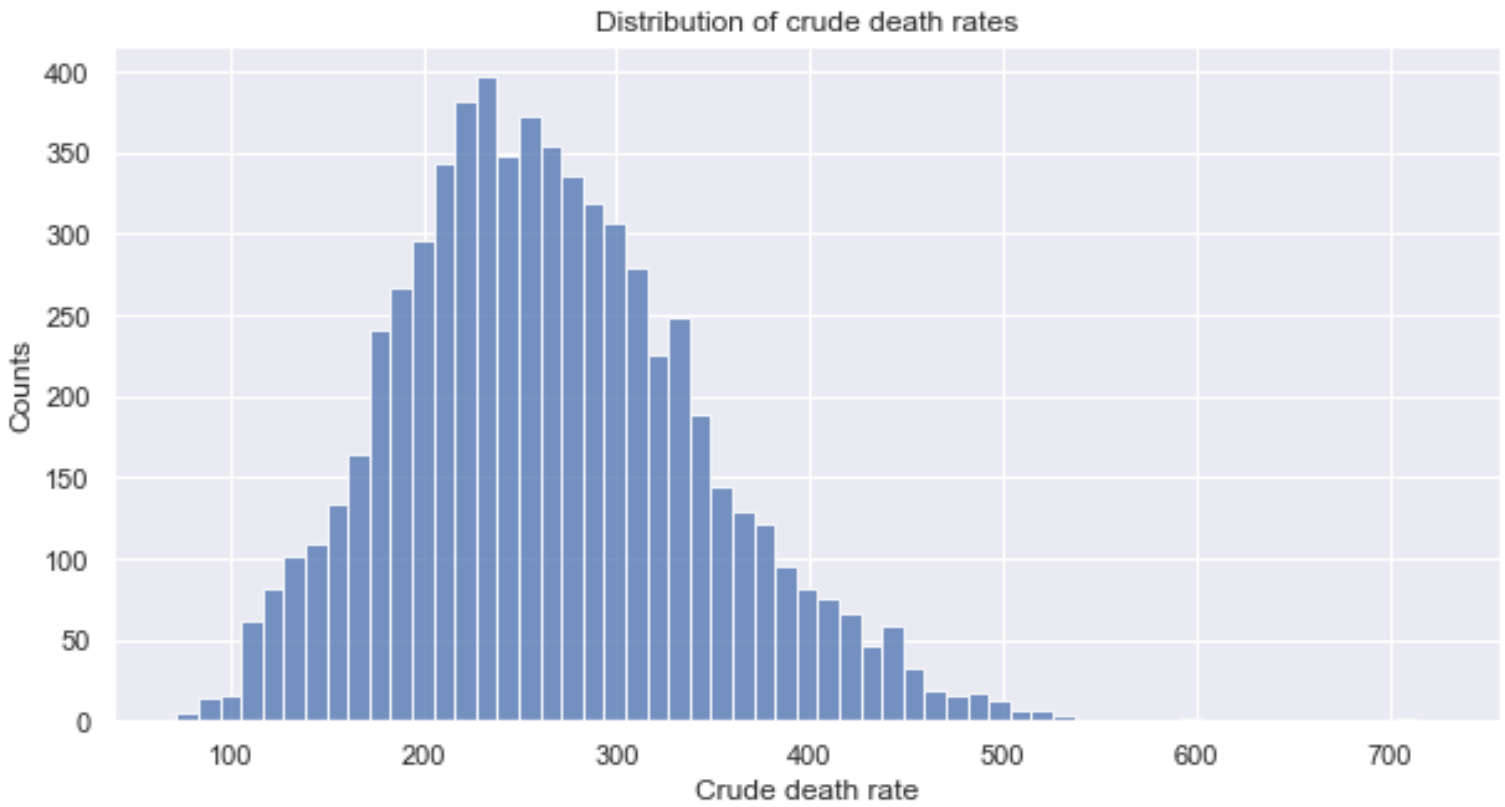
- educational attainment rates
- economic characteristic variables
- housing costs as a proportion of income

Both the current value and the change from the previous year's value were acquired for each of the census variables.

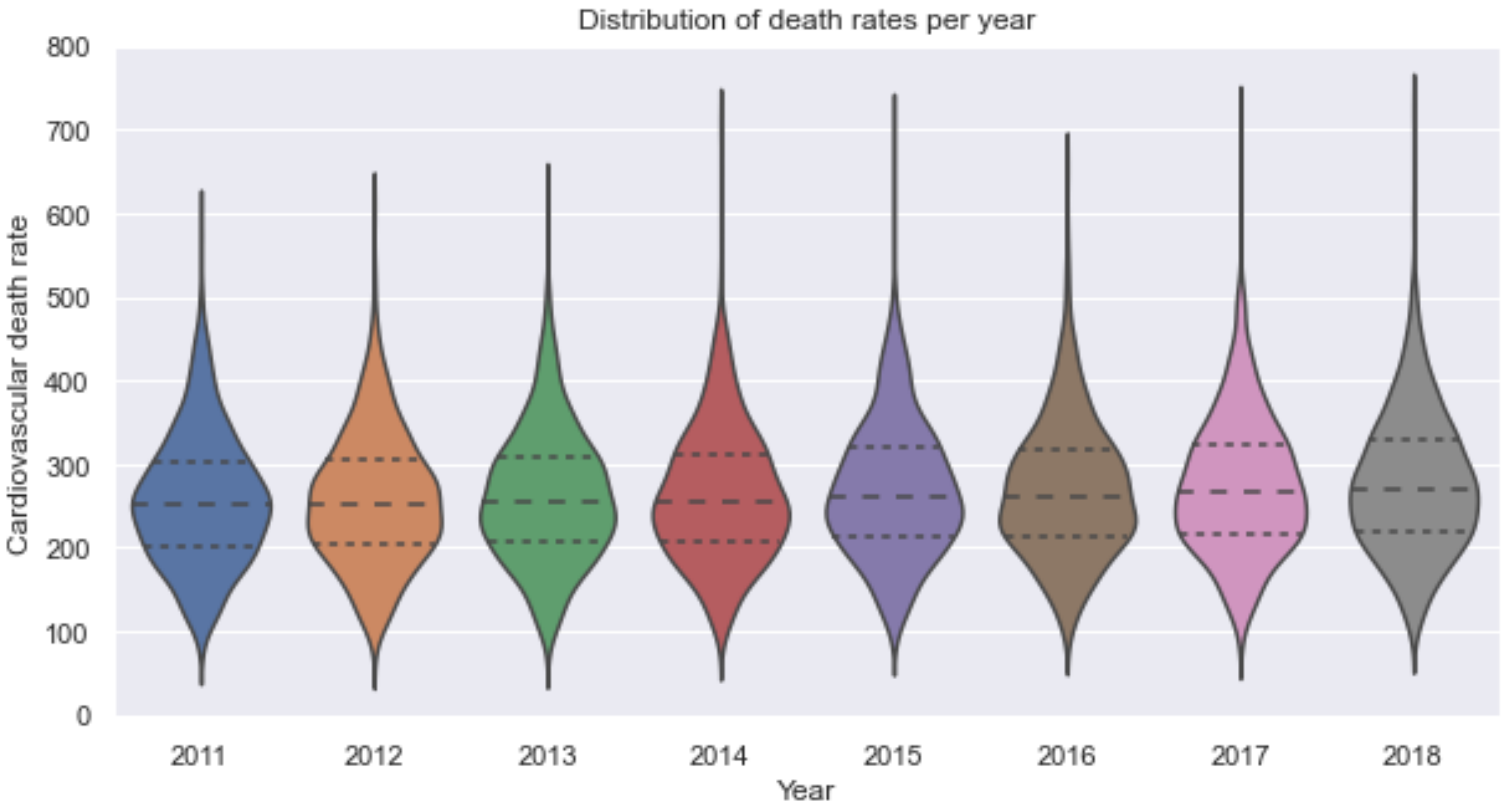
[5] Census,

Exploring the Data

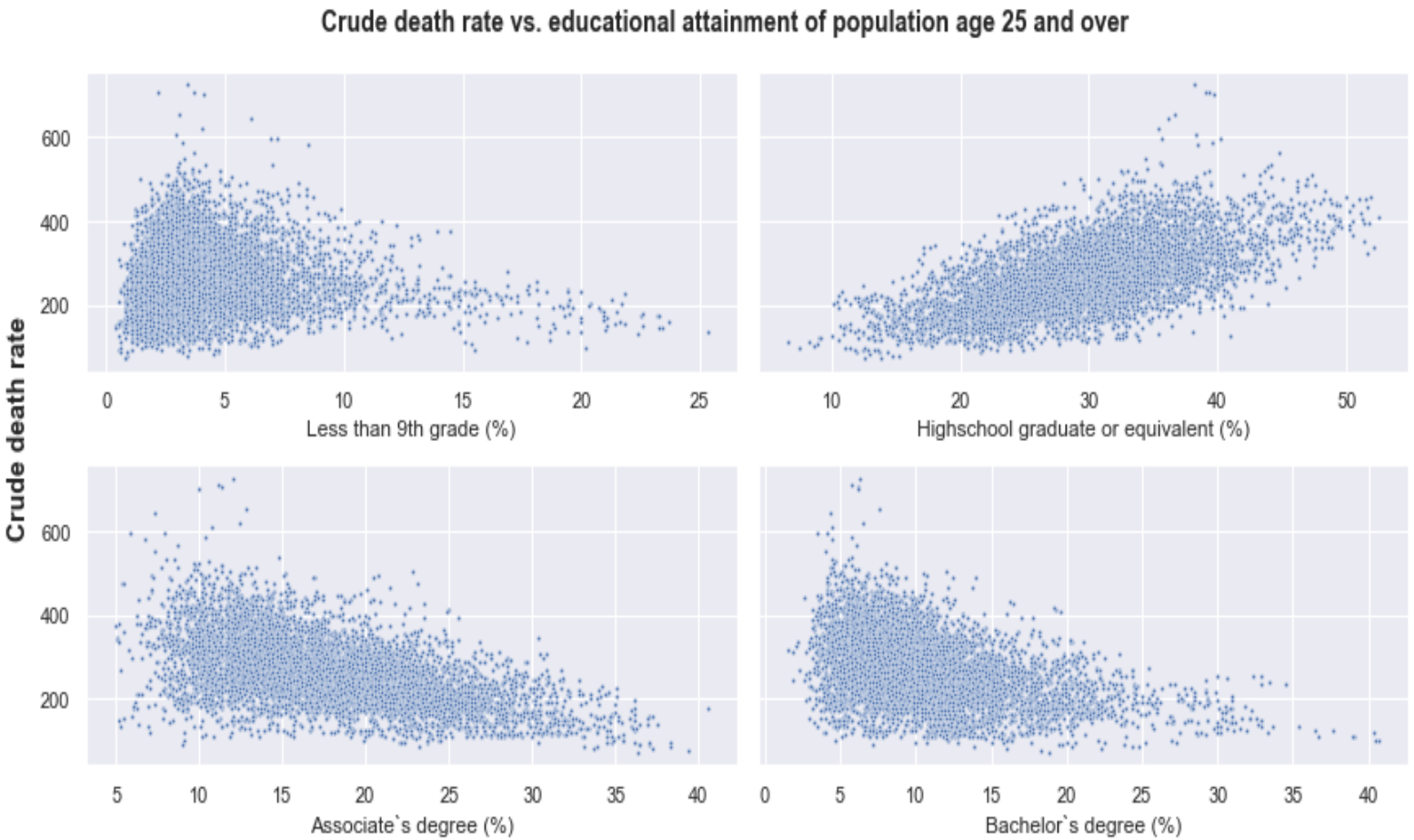
Exploring the Data



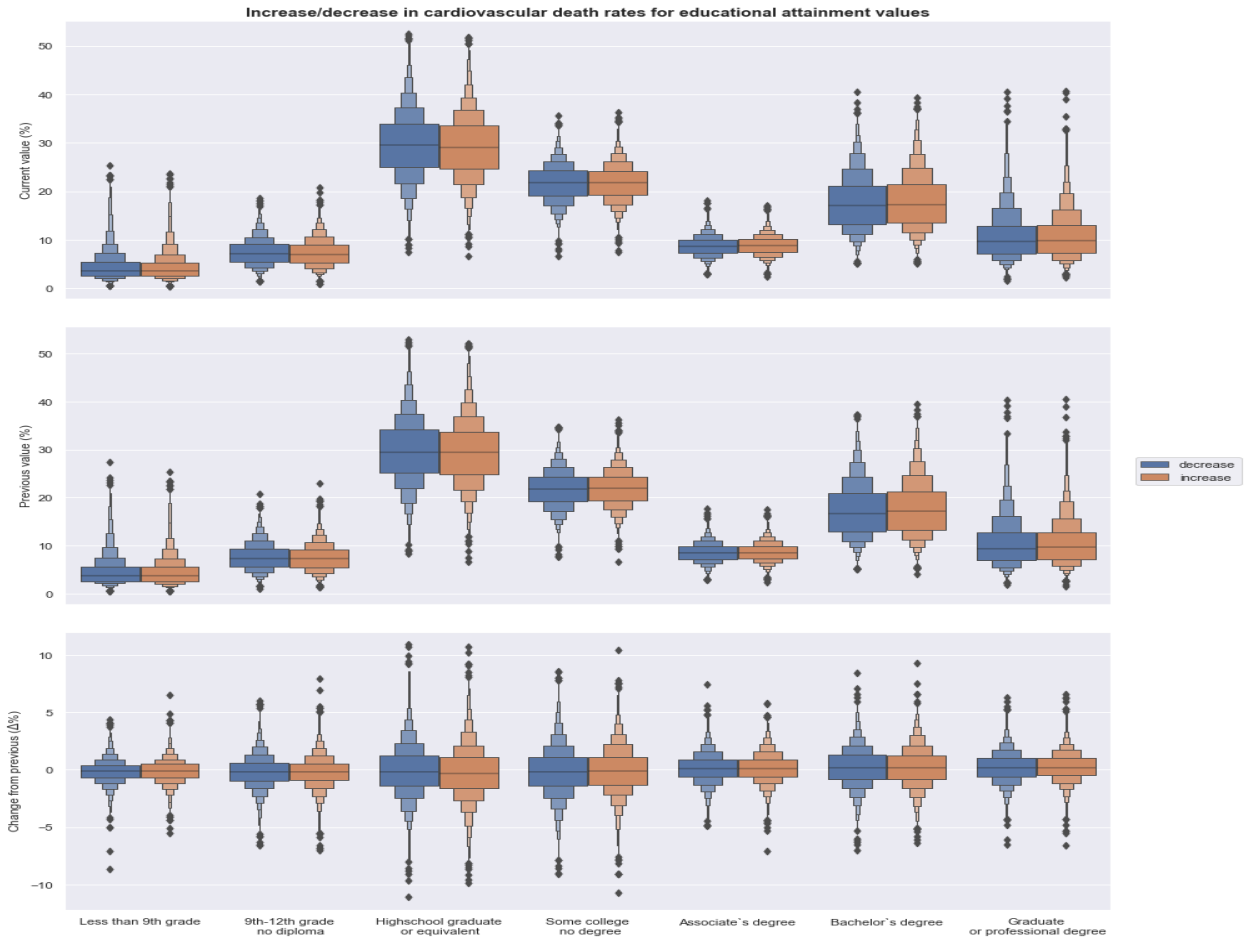
Exploring the Data



Searching for Visual Relationships



Searching for Visual Relationships



Refining the goal

- **Group** the samples by their change in death rate from the previous year.

Refining the goal

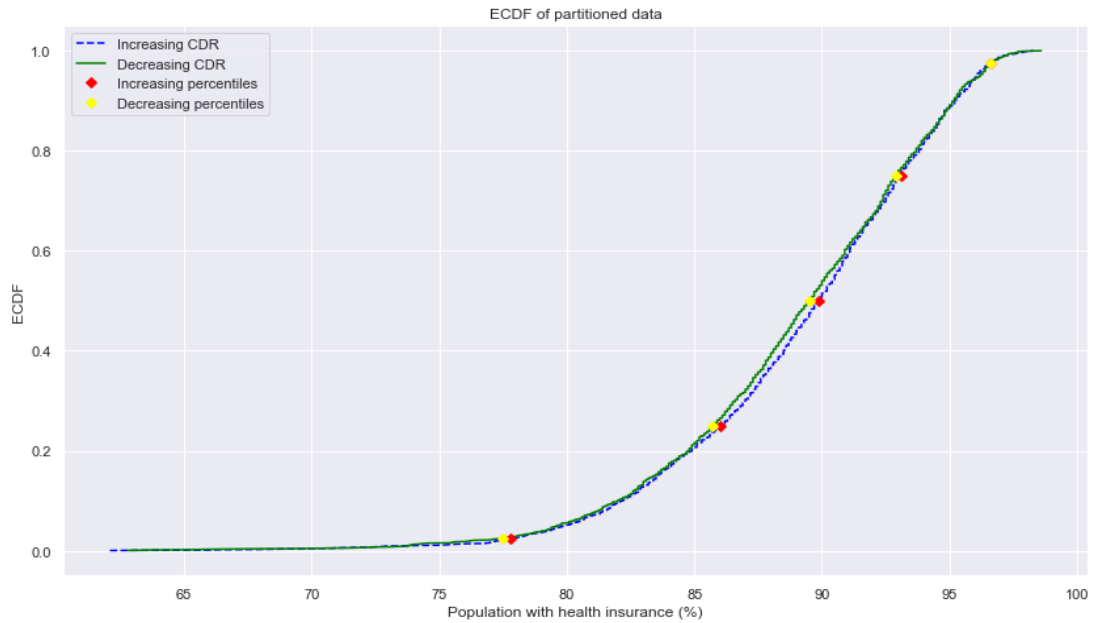
- **Group** the samples by their change in death rate from the previous year.
- Are the distributions of the census variables **different between the groups**?

Refining the goal

- **Group** the samples by their change in death rate from the previous year.
- Are the distributions of the census variables **different between the groups**?
- If so, can a **binary classification model** be made to predict which group each sample belongs to?

Statistical Analysis

With the samples partitioned by the sign change in death rate from the previous year, each census variable was checked for a significant difference between the mean value of each group.



Model Selection

The initial choice of model algorithm was a **Decision Tree**.

Pros

- Requires minimal pre-processing or data preparation
- Results are easy to interpret, which lends itself to the goal of providing recommendations

Cons

- Tend to be prone to overfitting training data

If there are issues with overfitting we can try using a **Random Forest**, which introduces elements of randomness that might reduce generalization error.

Decision Tree

Starting with the decision tree, there are several parameters that need to be tuned.

- max_depth

Decision Tree

Starting with the decision tree, there are several parameters that need to be tuned.

- max_depth
- min_samples_split

Decision Tree

Starting with the decision tree, there are several parameters that need to be tuned.

- max_depth
- min_samples_split
- min_samples_leaf

Decision Tree

Starting with the decision tree, there are several parameters that need to be tuned.

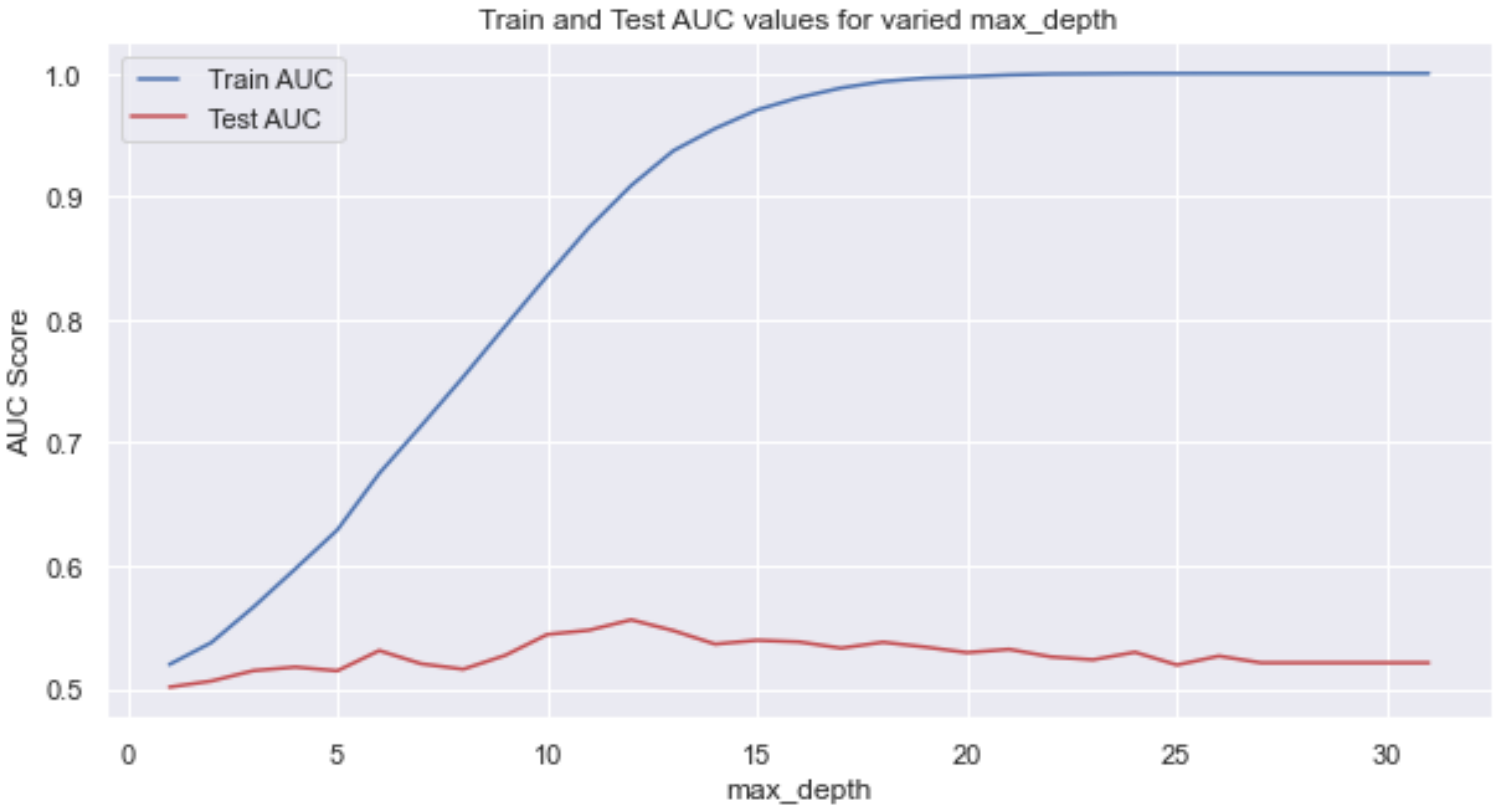
- max_depth
- min_samples_split
- min_samples_leaf
- max_features

Decision Tree

Starting with the decision tree, there are several parameters that need to be tuned.

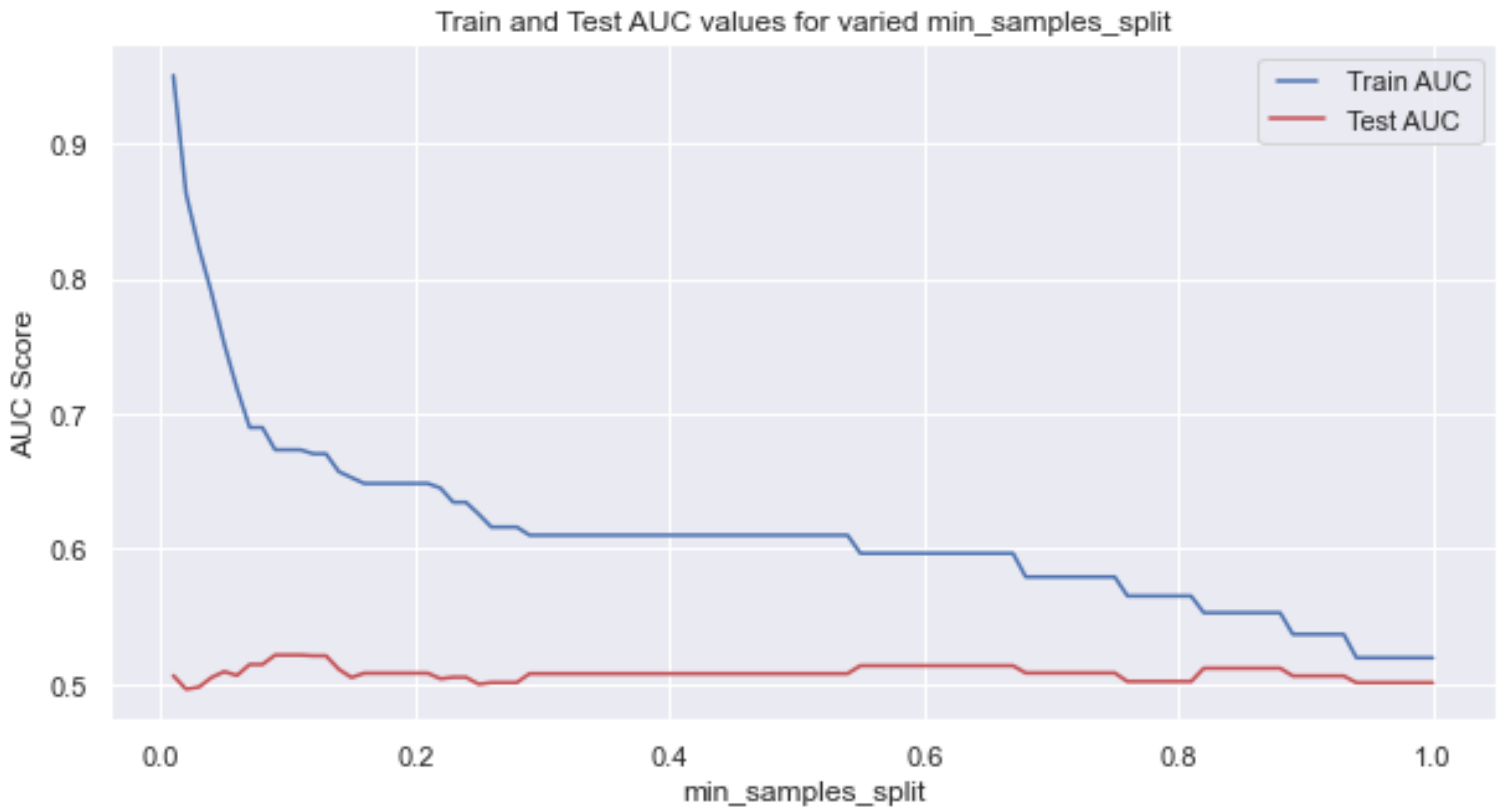
- max_depth
- min_samples_split
- min_samples_leaf
- max_features
- max_leaf_nodes

Decision Tree - Parameter Tuning



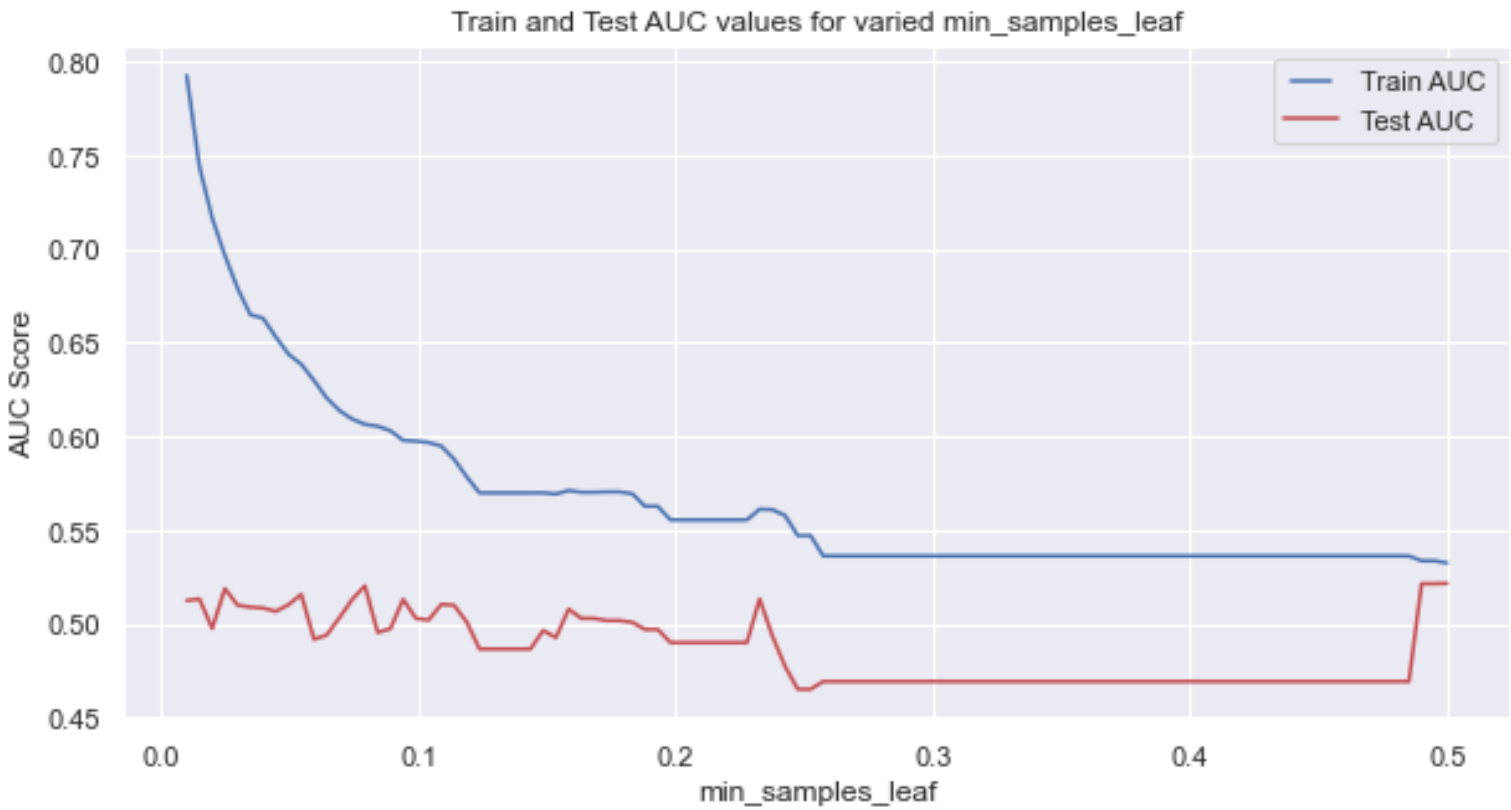
The model fails to generalize to unseen data regardless of hyperparameter values

Decision Tree - Parameter Tuning



The model fails to generalize to unseen data regardless of hyperparameter values

Decision Tree - Parameter Tuning



The model fails to generalize to unseen data regardless of hyperparameter values

Decision Tree - Parameter Tuning

The model fails to generalize to unseen data regardless of hyperparameter values

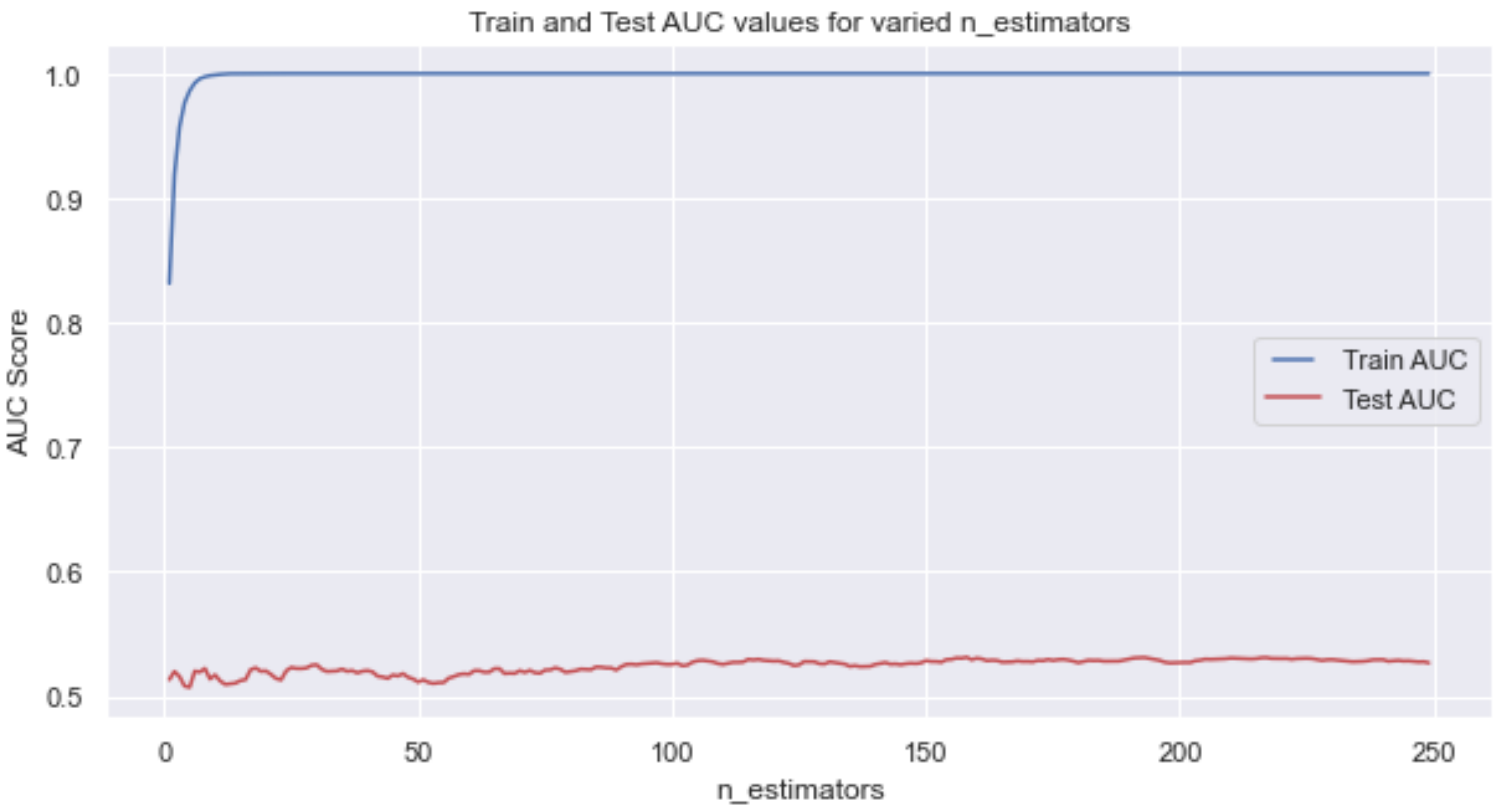
Decision Tree - Parameter Tuning

The model fails to generalize to unseen data regardless of hyperparameter values

Decision Tree - Results

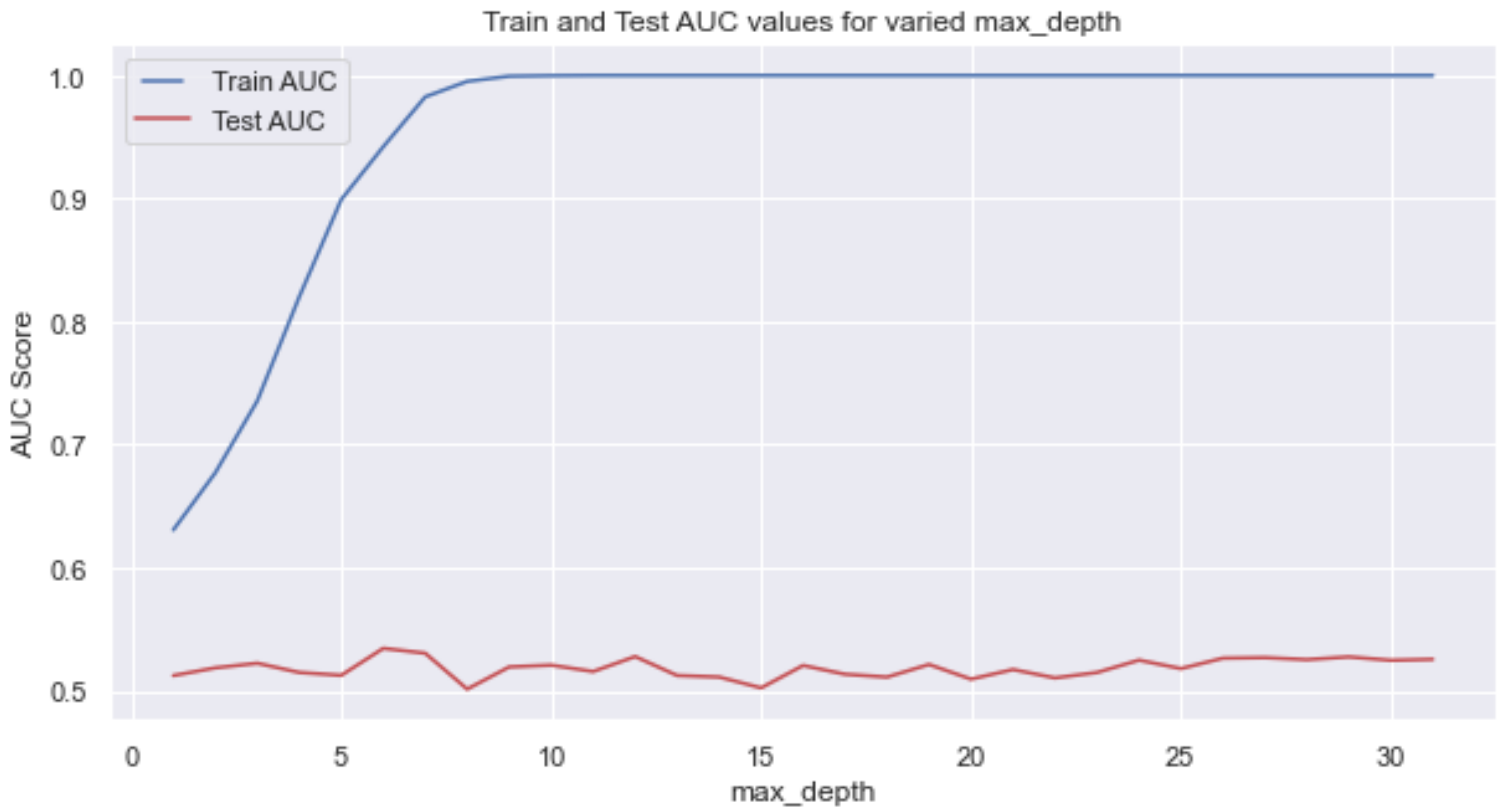
A grid search cross-validation of hyperparameter values was performed using scoring metric of *roc_auc*. A reduced feature set was also determined using the feature importance attribute of the best estimator returned by the grid search. The test set *roc_auc* scores for the full and reduced feature set were 0.495 and 0.526, respectively. **Are these results due to the limitations of using a decision tree, or is the issue related to the data set?**

Random Forest - Parameter Tuning



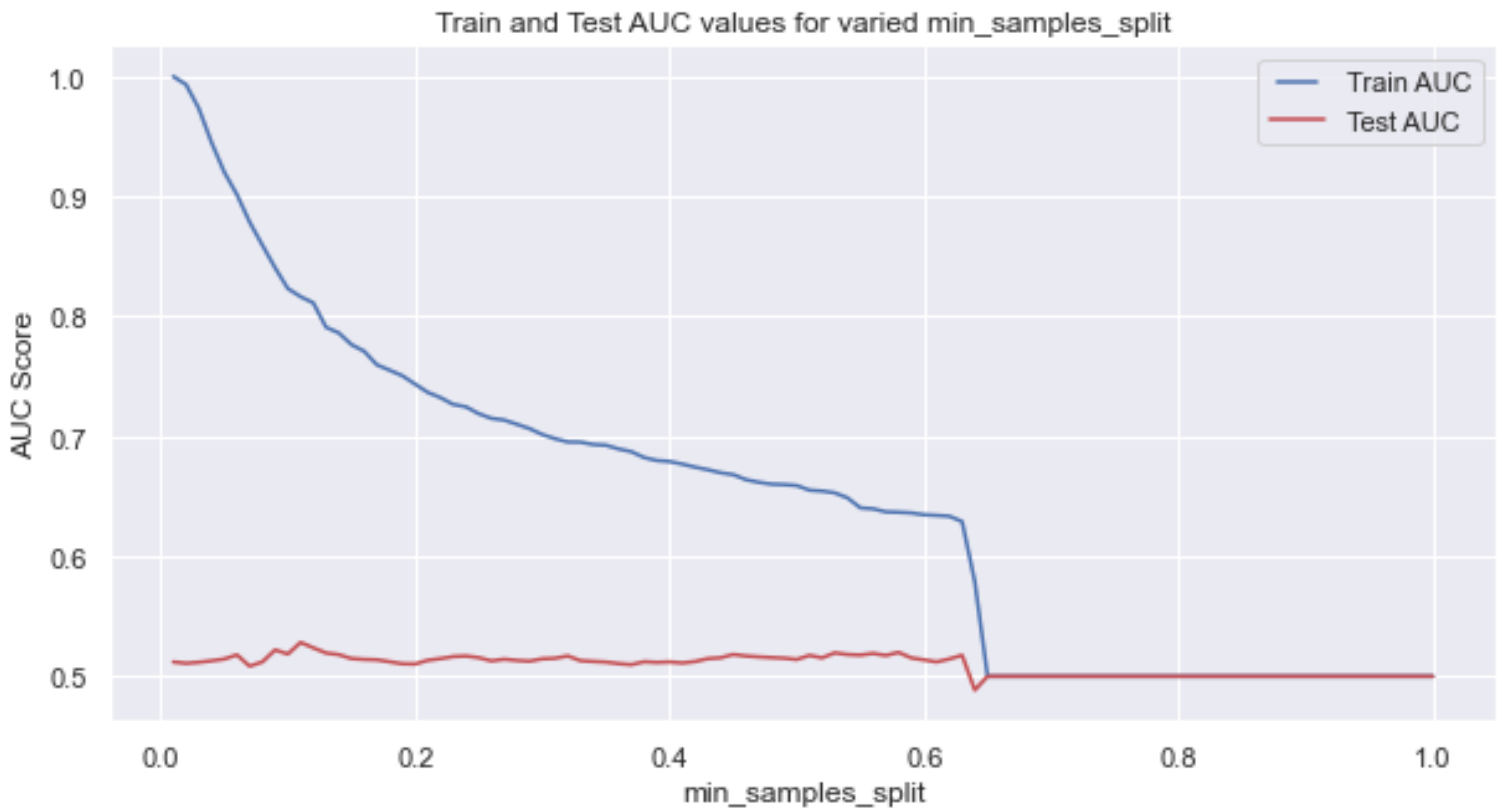
The issue does not appear to have been resolved using ensemble methods

Random Forest - Parameter Tuning



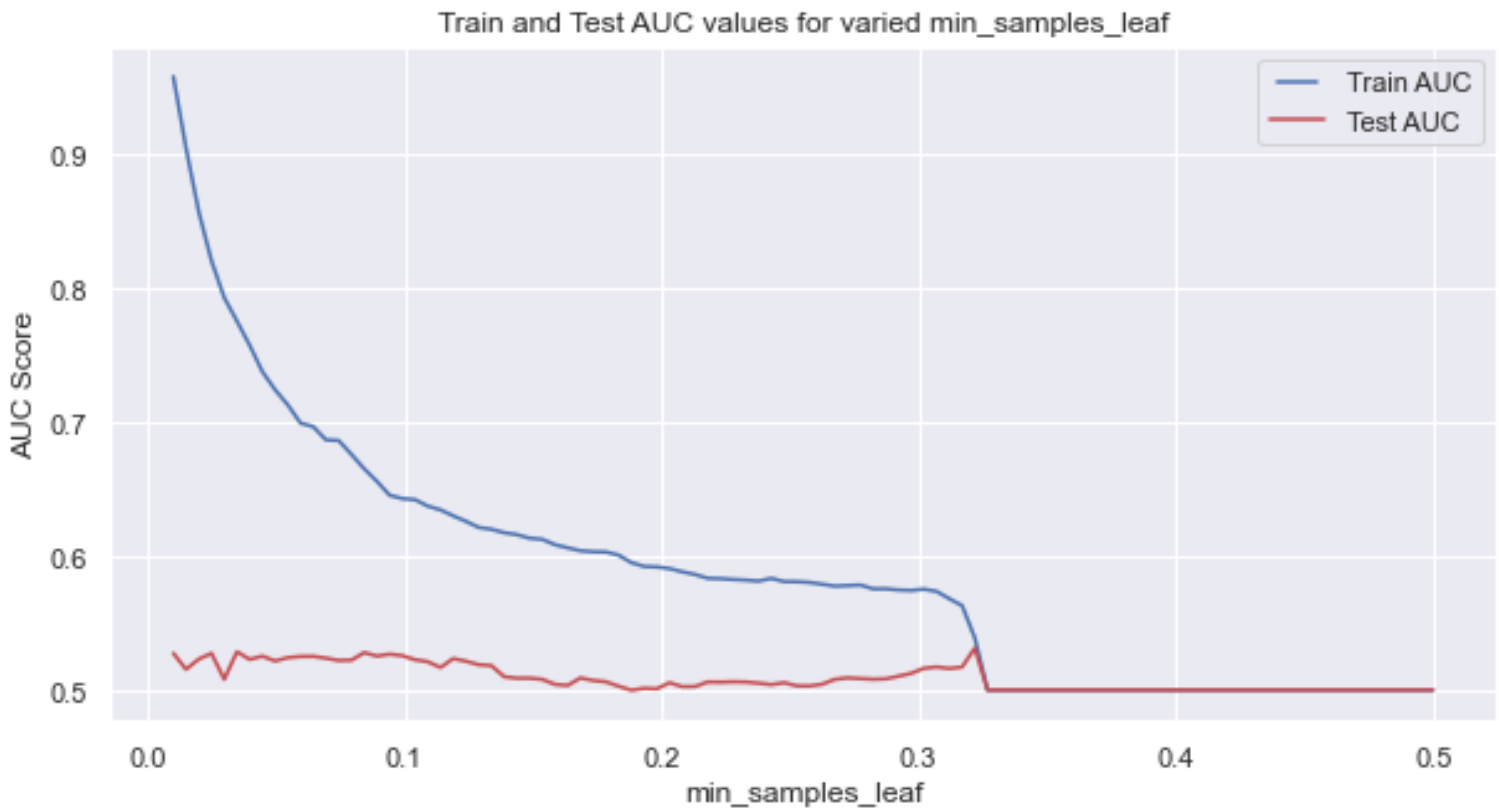
The issue does not appear to have been resolved using ensemble methods

Random Forest - Parameter Tuning



The issue does not appear to have been resolved using ensemble methods

Random Forest - Parameter Tuning



The issue does not appear to have been resolved using ensemble methods

Random Forest - Parameter Tuning

The issue does not appear to have been resolved using ensemble methods

Random Forest - Results

A grid search cross-validation of hyperparameter values was performed for the random forest, and a reduced feature set determined as was done for the decision tree. The test set *roc_auc* scores for the full and reduced feature set were 0.527 and 0.523, respectively. **Therefore it can be concluded that the failure to produce a model is likely due to an issue with the data itself.**

Conclusion

- Failure to produce a good model prevents any meaningful recommendations to be made.

Conclusion

- Failure to produce a good model prevents any meaningful recommendations to be made.
- Future work on this project will require careful selection of population variables.