# **Global**

## Assumptions

## Model Development

## Advantages

## Disadvantages

# **Random Forest Regression**

## Assumptions

The set of chosen factors are in some respect, responsible for the output variable. Ensure data is available for it.

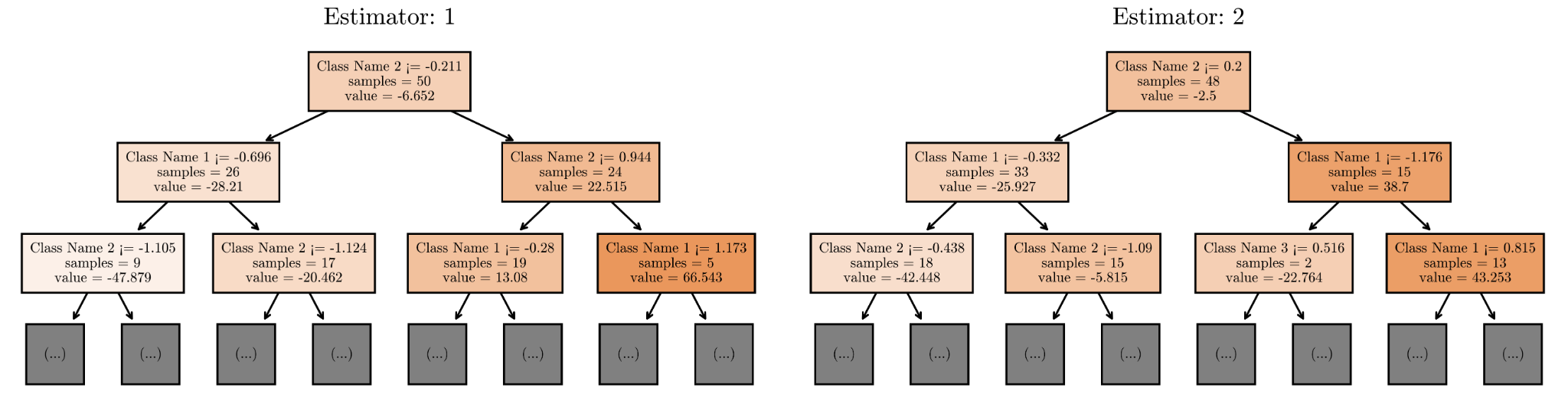
Choose a broad range of factors.

## Model Development

Investigating the effect of more than one factor lends itself to this method of data analysis.

This supervised machine learning model employs an ensemble of decision trees, that are trained off of bootstrapped data, and aggregates each of their outputs to make a prediction. At each decision tree, the data-set undergoes binary recursive partitioning.

The trees are engineered such that the highest order branch is, with respect to a random subset of distinct features, the split resulting in the lowest sum of deviations from each partition’s mean. This is computed across every possible binary split on the data, within the relevant subset of features. This process continues recursively in each branch, resulting in a ‘random tree’. The outputs of the individual decision trees are then averaged for the final ensemble prediction.



A grid search of typical values was employed to determine the best hyperparameters to maximise the (scoring metric). The data set was in an 80:20 split into two subsets: training data, and testing data (standard for industry). By employing 5 fold cross-validation, we were able to deduce the optimal number of trees, and optimal tree depth to be X and Y.

From the ranges XYZ to ZYX, XYZ was found to be the optimal value for the tree depth, and from the ranges XYZ to ZYX, XYZ was found to be the optimal number of trees.

This was done using 5 fold cross-validation.

basically I pick a range of max depths and num trees and it tests each combination and picks the best one)

Using a grid search to determine the best hyperparameters to maximise the scoring metric (yet to be determined) (basically I pick a range of max depths and num trees and it tests each combination and picks the best one)

Investigating the effect of more than one factor lends itself to this method of data analysis. ‘offer valuable insights into the key drivers’ of XYZ

Identify each factor and JUSTIFY each one with a short paragraph (neil or jakubs job)

the data-set undergoes binary recursive partitioning (decision tree)

Create bootstrap data then randomly select a subset of n variables. Compute sum of squared residuals then, the variable that does the best job separating the samples becomes the root node to the decision tree. For each branch, randomly select a subset of variables as candidates, and build the tree as usual. After repeating this process, with new bootstrapped data sets each time, this results in a wide variety of trees. New sample goes in all trees and it is classified from the aggregate. This is called bagging

Testing goodness:

Use a sample of data, run it through the trees for which it is an out of bag sample, then aggregate the results. Successful classifications imply good forest. Then do it for all the ‘out of bag’ samples. The proportion of out of bag samples correctly classified by the forest . This proportion is called the ‘Out-of-bag error’

Since this forest was done using subsets of n variables, change n, see how it affects the out of bag error. Choose the minimising value n, to reduce error.

For a data set with samples with missing data values, fill in the gaps using the average of other samples of the same class. Then, with the new data set we build a random forest. Then run all of the data down all of the trees. Keep track of similar samples using a ‘proximity matrix’. Row for each sample and column for each sample. Add a 1 in the column and row that end up at the same leaf node - matrix is diagonally symmetric. Repeat across all trees and divide by total number of trees. Use proximity values to make better guesses about the missing data. Using the proximity values as the weights, calculate the weighted frequencies for the data. REFER TO STATQUEST.

Hog cycle this with the new constructed data point. Run it until the missing values converge (do not change). 1- Prox matrix = distance matrix -> draw heatmap with it (or NDS plot)

“To calculate the relative importance of each factor that we included, we decided to use the feature importance values, produced by the forest. If we achieve a high R2 value, we can be confident that the relative importance of the factors (as determined by the forest) is also accurate.

”Mean decrease impurity” is an index often used for evaluating feature importance. As per [4], is often described as ”the total decrease in node impurity...averaged over all trees of the ensemble.” This relies on the concept of impurity, which for a regression tree can simply be defined as the variance of the partition (relative to the partition’s mean).[14] Other definitions of impurity do exist, however they serve the same purpose of measuring the homogeneity of the data with more diverse partitions receiving a larger value.”

This supervised machine learning model employs an ensemble of decision trees to make an accurate and robust prediction. Each decision tree is trained on a random subset of the dataset to reduce intercorrelation and utilises algorithms to determine optimal logical splits. Following the created splits, data can be categorised and a regression output is produced. The outputs of the individual decision trees are then averaged for the final ensemble prediction.

STATE num of trees, tree depth, features of data,

## 

## Advantages

The random forest model allows us to effectively combine multiple factors to make predictions. The algorithm is robust to skewed, non-linear data which makes it ideal for processing features such as population density.

Furthermore, the ensemble process minimises the risk of overfitting, a prevalent issue where a machine learning model learns the training data too closely and later fails to accurately predict unseen examples. CROSS VALIDATION

Overall, gradient boosting usually performs better than random forests but they're prone to overfitting. There ARE better models out there. We chose this because of simplicity to explain yadi nadiya waffle. Small number of trials

## Disadvantages

Can fall prey to biases in data. ”Our trained RuleFit model had a high specificity and a low sensitivity, as shown in Table 3.4.3. This means that our model has a fairly strong negativity bias. This could be improved in future iterations of the model by training it with more positive data.” - TC3 2022

However, a weakness of random forest regression is the fact that it is not easily interpretable since it consists of a large number of decision trees. It is difficult to pinpoint the exact decision process that the algorithm undergoes to arrive at its prediction. the lack of transparency in the black box nature of the Random Forest algorithm.

## Sensitivity Analysis

2023 3rd place did something interesting. Create a forest where the data set has one or more variables missing. Then compare it to the model created with all variables. Compare significance of each variable in both models, and comment on how there is low variability in significance, implying robustness of model.

# **Meyer Notes**

* Uses Sklearn’s python library RandomForestRegressor function
* Am using a grid search to determine the best hyperparameters to maximise the scoring metric (yet to be determined) (basically I pick a range of max depths and num trees and it tests each combination and picks the best one)
* Uses 5 fold cross-validation
* To calculate the relative importances of the variables we are using a Gini importance metric - Nice explanation here: https://stackoverflow.com/questions/15810339/how-are-feature-importances-in-randomforestclassifier-determined