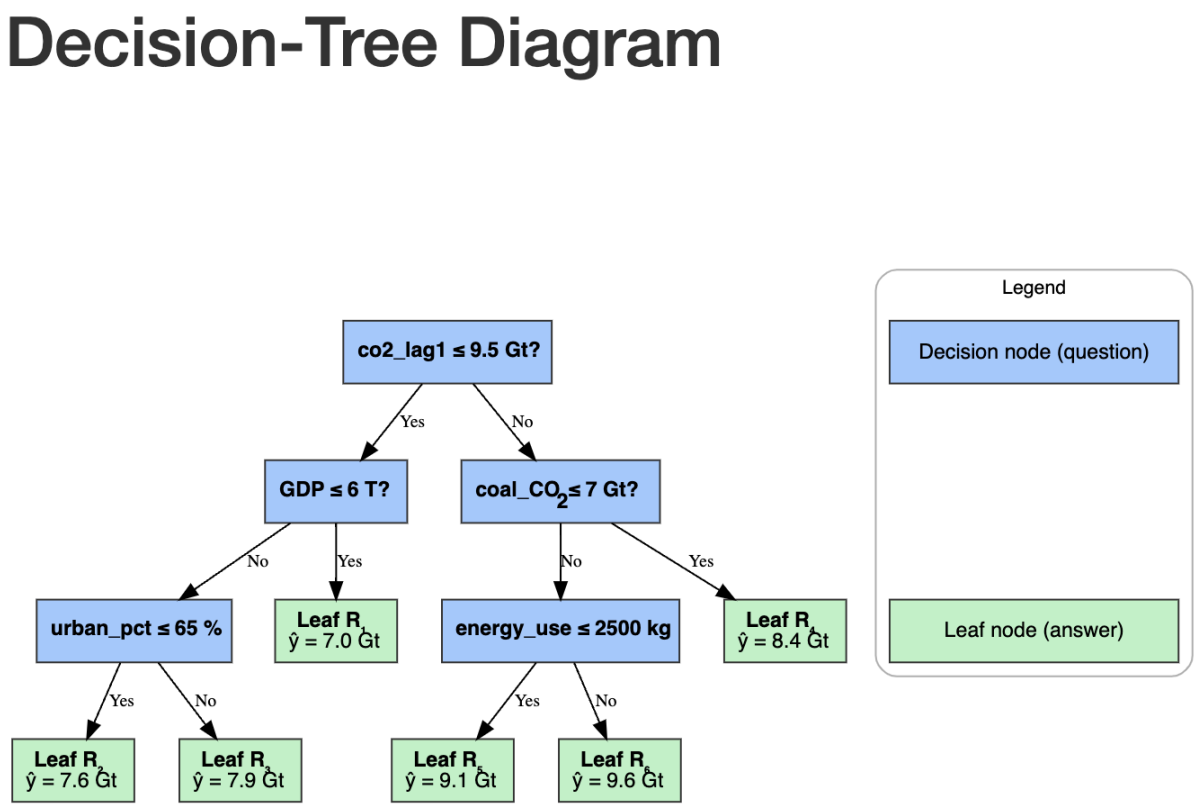
Gradient-Boosted Decision Trees for Annual CO_2 Forecasting in China

1. What is a Gradient-Boosted Decision Tree?

Diagram:



First, what is a decision tree? A decision tree splits the data into rectangular regions by asking yes or no questions. For example, is the GDP <= 5 trillion? If yes, branch left; otherwise, go right. Then, in each terminal node, otherwise known as a "leaf", it predicts a constant. For example, the mean CO_2 emissions for that specific path. There are many strengths associated with this model, the biggest being its ability to capture non-linear trends and interpretability, and handling of mixed units. However, there are also associated weaknesses; a single tree is **unstable** (small data changes can lead to big decision tree structural changes), and **under-fits** a smooth relationship, while too many trees can lead to **over-fitting (**excessively high complexity). However, gradient boosting aims to mitigate overfitting by using regularization (learning rate[1], regularization terms[2], pruning[3]), early stopping (monitor performance on a validation set), and controlling model complexity (tree depth[4], minimum samples per leaf, subsampling).

---

[1] Gradient boosting uses a learning rate (a multiplier between 0 and 1) to scale the contribution of each weak learner (usually a decision tree). A lower learning rate can help prevent overfitting by reducing the impact of each individual tree's prediction.
[2] Some gradient boosting algorithms (like XGBoost) incorporate regularization terms in the loss function, penalizing complex models and preventing them from fitting the training data too closley.
[3] Pruning involves removing branches or nodes in the decision tress that don't contribute significantly to the prediction, reducing complexity and preventing overfitting
[4] Limiting the maximum depth of the decision tress can prevent them from becoming too complex and overfitting

Tree depth: Controls the depth of the individual trees. Typical values range from a depth of 3–8, but it is not uncommon to see a tree depth of 1 (J. Friedman, Hastie, and Tibshirani 2001). Smaller depth trees, such as decision stumps, are computationally efficient (but require more trees); however, higher depth trees allow the algorithm to capture unique interactions but also increase the risk of over-fitting.

Now, moving on, what is boosting? Since we have a "gradient-boosted" decision tree model. Furthemore, we should think of boosting as sequentially connecting the weak learners[5] in a decision tree model to improve classification. For example, suppose we have a single blue rectangle in the diagram above. By themselves, these decision trees are quite useless and don't have powerful prediction capabilities. However, the idea is that the following weak learner is trained on the mistakes of the previous one, continuing in a sequence until it reaches a point where the collection of weak learners provides a powerful forecasting model. Now, the term gradient comes from the fact that we are using the gradient of the loss function to determine the direction of the process that the weak learners follow to correct the mistakes of the previous learner in the sequence. Notice, the model also includes a learning rate, implying the next learner in the process for GBDT modeling is weighted using a scalar.

GBDT works well for CO_2 forecasting because:

- Non-linear relationship in CO_2 can be split up by decision trees at those thresholds
- GBDT requires non-scaling (We have mixed units & scalars {dollars, %, TWh})
- Interaction effects, higher-level splits capture interactions automatically
- Modest-sized data (~60 annual observations), shallow trees + shrinkage reduces over-fitting; you can cross-validate for the optimal depth.

2. Predictors used for China's CO_2 Model

(Predictor Summary used in GDBT):

Predictor (column)	**What it is**	**Unit**	**Why include it**
co2_lag1	CO_2 emissions in year *t–1* (first-order lag)	million tonnes CO_2	Captures strong inertia in national emissions; most single-year forecasts are heavily anchored to last year's level.
co2_lag2	CO_2 emissions in year *t–2* (second-order lag)	million tonnes CO_2	Adds "momentum" information and helps the trees learn turning points or multi-year cycles that one lag alone can miss.
gdp	Real Gross Domestic Product (PPP-adjusted, 2011 US$)	billions of 2011 international $	Economic scale is a primary driver of energy demand and industrial activity, strong historical correlation with emissions.

[5] Deicision-tree that contains only one split, i.e. the diagram above but contains only one blue and two green rectangles.

population	Total resident population	millions of people	Larger populations generally consume more energy, which disentangles scale effects from per-capita efficiency changes.
primary_energy_consumption	Total primary energy used (all fuels)	terawatt-hours (TWh)	Direct physical measure of energy demand underlying CO_2; lets the model recognise decoupling or re-coupling between energy and GDP.
coal_consumption	Coal consumption (primary energy basis)	TWh (or Mt coal equi.)	Coal is China's dominant, most carbon-intensive fuel; variations here strongly leverage total emissions.
oil_consumption	Oil consumption (primary energy basis)	TWh (or Mt oil equi.)	Important for transport & industrial sectors; picks up shifts from coal to liquid fuels or vice versa.
gas_consumption	Natural-gas consumption (primary energy basis)	TWh (or bcm)	Gas is lower-carbon than coal/oil; tracks fuel-switching policies and winter heating demand spikes.
urban_pct	Urban population share	percent of total population	Urbanisation alters energy-use patterns (high-density housing, transit, industrial clustering) and has proved a significant long-run driver of Chinese emissions.
gdp_growth	One-year growth rate of real GDP	decimal fraction (e.g., 0.05 = 5 %)	Captures acceleration or slowdown effects—rapid growth years often yield higher marginal emissions even if the absolute GDP level is known.
pop_growth	One-year population growth rate	decimal fraction	Flags demographic surges or slowdowns that affect future demand without waiting for the level variable to change appreciably.
energy_growth	One-year growth rate of primary-energy consumption	decimal fraction	Adds a rate-of-change signal that can improve short-horizon forecasts when absolute energy demand is trending quickly up or down.

year	Calendar year encoded numerically or as time index	year integer	Provides the model an explicit time trend; sometimes helpful for capturing smooth structural change not explained by the other predictors. *(Include only if tests improve fit.)*

Note: including both primary_energy_consumption & energy_use_kg_oil_per_cap allows the model to determine whether a CO2 jump came from sheer scale (new steel mills) or higher intensity (inefficient equipment).

3. Mathematical formulation for annual CO_2 forecasts

Below is the **mathematical formulation** of China's annual CO2 forecasting model using gradient-boosted decision trees (GBDT). Let:

- y_t be the observed CO_2 emission in year *t.*
- $x_t = (x_{t,1}, \ldots, x_{t,p})$ be the *p*-dimensional predictor vector in year t, implying $x_t = (co2\_lag1,\ co2\_lag2,\ gdp, primary\_energy\_consumption,\ Energy\_use\_kg\_oil\_per\_cap, \ldots)$

You choose the **squared-error loss:**

$$L\big(y,\ F(\mathbf{x})\big) \ = \ \frac{1}{2}\big(y - F(\mathbf{x})\big)^2$$

as your training objective. (trained using the squared-error {least-squares} loss function)

Additive model of regression trees

Your prediction function is built as a sum of M regression trees:

$$\hat{y}_t \ = \ F_M(\mathbf{x}_t) \ = \ \sum_{m=1}^{M} \nu\, h_m(\mathbf{x}_t)\,.$$

Where:

- $h_m(\mathbf{x})$ is the m-th CART tree (leafwise constant function)
- $\nu \in (0, 1]$ is the **learning rate** (or shrinkage) that scales each tree's contribution

Greedy stagewise fitting

Starting from a constant base model:

$$F_0(x) \ = \ \arg\min_{\gamma} \sum_{t=1}^{T} L(y_t, \gamma) \ = \ \frac{1}{T}\sum_{t=1}^{T} y_t$$

We iterate for m = 1, . . . , M:

1. Compute residuals (negative gradients)

$$r_t^{(m)} \;=\; -\,\frac{\partial L\big(y_t,\,F(\mathbf{x}_t)\big)}{\partial F(\mathbf{x}_t)}\bigg|_{F=F_{m-1}} \;=\; y_t - F_{m-1}(\mathbf{x}_t).$$

2. Fit a regression tree $h_m(\mathbf{x})$ to the residuals $\{r_t^{(m)}\}$ by least squares.

3. Compute the optimal leaf values $\gamma_{m,j}$ in each region $R_{m,j}$ of the tree h_m:

$$\gamma_{m,j} \;=\; \arg\min_{\gamma} \sum_{\mathbf{x}_t \in R_{m,j}} L\big(y_t,\,F_{m-1}(\mathbf{x}_t) + \nu\gamma\big) \;=\; \frac{\sum\limits_{\mathbf{x}_t \in R_{m,j}} r_t^{(m)}}{\sum\limits_{\mathbf{x}_t \in R_{m,j}} 1}\,.$$

4. Update the ensemble:

$$F_m(\mathbf{x}) \;=\; F_{m-1}(\mathbf{x}) \;+\; \nu\, h_m(\mathbf{x})\,.$$

Final Prediction

After M rounds, the model predicts:

$$\hat{y}_t \;=\; F_M(\mathbf{x}_t) \;=\; F_0 \;+\; \nu \sum_{m=1}^{M} h_m(\mathbf{x}_t)\,.$$

Interpretation via SHAP

To interpret feature importance and marginal effects, you compute **SHAP values** ϕ_i for each predictor $x_{t,i}$, satisfying:

$$F_M(\mathbf{x}_t) \;=\; \phi_0 \;+\; \sum_{i=1}^{p} \phi_i(\mathbf{x}_t)\,.$$

Where ϕ_i quantifies how much $x_{t,i}$ contributes to pushing the prediction away from the base value $\phi_0 = F_0$

In CO2 context

- x_t includes lagged CO2, GDP, primary energy consumption, and new additional variables like energy_use_kg_oil_per_cap and urban_pct
- Shrinkage ν, tree depth, and number of rounds M are hyperparameters tuned to minimize out-of-sample RMSE
- The ensemble captures both autoregressive persistence (via lagged CO2) and non-linear interactions (e.g., diminishing returns to energy efficiency, urbanization thresholds).

In summary, your **mathematical GBDT model** is the additive expansion of small regression trees, each correcting the residuals of the prior iteration, trained to minimize the squared error of annual CO2 forecasts given your chosen predictors.

4. Application to China's CO_2 data

5. Takeaways

References: Just links for now

https://www.youtube.com/watch?v=en2bmeB4QUo (gradient boosting modeling)

https://www.youtube.com/watch?v=zs6yHVtxyv8 (decision trees)

file:///Users/henry/Desktop/University%20of%20Toronto/STA457/Group%20Project/1013203451.pdf

https://arxiv.org/pdf/1603.02754

file:///Users/henry/Desktop/University%20of%20Toronto/STA457/Group%20Project/sustainability-13-12302.pdf