High Dimensional Examination of Intergenerational Mobility with Random Forest and Gradient Boosting*

Cahn, Yisroel† Maasoumi, Esfandiar‡

September 19, 2025

Abstract

We propose measures of upward mobility based on the (conditional) probability of offspring outcome being at least as well as their parent(s). We estimate these probabilities using robust machine learning (ML) tree methods, which are capable of identifying key predictors of mobility while avoiding the curse of dimensionality inherent in kernel-based nonparametric models. ML tree methods offer two main advantages: (1) they are nonparametric, accounting for nonlinearities and relatively robust to misspecifications, and (2) they can rank features by overall predictive value. We find that parent income is the most important predictor of both absolute and relative income mobility. However, family wealth becomes relatively more important (while still not as important as parent income) for predicting large movements in income and income rank. Influence of parent age and other group characteristics are controlled for.

Keywords: Intergenerational Mobility, Income Distribution, Machine Learning, Random Forest, Gradient Boosting, Prediction

JEL Codes: D31, J31, J62

*We would like to thank the coeditors, three referees and participants at Stone Center for Research on Wealth Inequality Conference on New Methods to Measure Intergenerational Mobility for their helpful comments and suggestions.

†New York University, 19 West 4th Street, 603, New York, NY 10012.
Email: yisroel.cahn@nyu.edu

‡Emory University, Rich Building 324, Atlanta, GA 30322.
Email: emaasou@emory.edu

1 Introduction

Intergenerational income mobility (IGM) is a key measure of economic opportunity, influencing how children's outcomes are shaped by their parents' circumstances. Identifying which subgroups experience the lowest levels of IGM, and understanding the factors contributing to this inequality, is essential for designing targeted interventions that effectively support those most in need. This paper employs machine learning methods—specifically, Random Forest (RF) and Gradient Boosting (GB)—to examine IGM, enabling the ranking of contributing factors based on their predictive power.

The currently dominant approach to analyzing mobility is regression-based IGE analysis, which focuses on two key notions: predictability and dependence. Predictability refers to the extent to which offspring outcomes can be anticipated from parental states, while dependence examines how strongly offspring outcomes are linked to those of their parents. Both of these notions are largely examined by *the mean* of the conditional distribution of offspring outcomes given parental characteristics.

However, these two notions provide only a partial view of mobility, and through the lens of a conditional mean with less than stellar underpinnings as "measures of IGM". More complex and nuanced understandings emerge from broader welfarist perspectives, such as upward mobility and changes in both relative and absolute distribution of outcomes. These frameworks account not only for shifts in individual outcomes but also for the overall structure of social and economic mobility. For a more detailed discussion of these broader perspectives and broader measures of IGM, see Maasoumi (2020).

In this paper, we examine the probability of offspring outcomes being at least as good as their parents', in levels and/or ranks.[1] This is the (conditional) Probability of Upward Mobility (PUM), extensively examined in Bhattacharya and Mazumder (2011). This requires estimation of discrete dependent variable models (much like logit and probit and multinomial choice), as conditional probability models.

Random Forest and other tree-based machine learning methods are designed to be more robust than traditional parametric methods, particularly with respect to distributional assumptions and functional form restrictions. This is crucial given the extensive literature that questions the linearity of Intergenerational Elasticity (IGE) regressions and highlights heterogeneity across different population groups. One common approach to address nonlinearity is to use flexible mathematical functionals and polynomials to achieve the best fit for bivariate or low-dimensional relationships. However, high-dimensional methods like Random Forests offer the potential to uncover the underlying mechanisms

[1] While machine learning techniques such as Random Forest and Gradient Boosting can estimate continuous outcomes, the overall precision and accuracy of these models (discussed later in the paper) may be misleading if averaged. For instance, if ethnicity is a significant predictor of downward mobility but not upward mobility, averaging performance metrics across both outcomes can mask the model's true predictive capabilities for each scenario. That is why we prefer using probabilities, as they provide a more granular and interpretable measure of prediction across different outcomes.

of nonlinearity and heterogeneity, as they can reveal different effects at varying levels of income, wealth, education, age, and other covariates.

Random Forest is an ensemble method that combines multiple models through a technique called "bagging" (bootstrap aggregating), where regression trees are constructed using bootstrapped samples of the data. Each tree is built with different covariate combinations and identifies the ideal splits in covariate values. This model averaging approach is widely recognized in econometrics for its superior performance compared to model selection methods and its robustness to model misspecification. When paired with cross-validation and sample splitting techniques, Random Forest helps to avoid overfitting, ensuring more accurate predictions in test samples and enhancing counterfactual (predictive) analyses.

In contrast, *model selection* methods such as LASSO are biased toward variable elimination. [2] Additionally, ML tree methods accommodate nonlinear relationships between many predictors and outcome measures, and more importantly, nonlinear interactions among potential predictors of mobility. LASSO, in contrast, assumes linear relationships between predictors and outcome variables and does not allow interactions among predictors.

Our philosophy is to robustify findings to possibly restrictive choices of functional form and covariate controls. If findings corroborate simpler parametric models, that is extremely valuable and reassuring. If they do not, we are better informed as to reasons, be they nonlinearities, other covarites, or subgroup characteristics.

1.1 Measuring Mobility

As has been noted by many authors, for example see Chetty, Hendren, Kline, Saez, and Turner (2014), measuring intergenerational income mobility amounts to characterizing the joint distribution of parent and child income, which is comprised of (1) a dependence component, such as the *copula* (joint distribution of parent and child ranks), and (2) the *marginal* distributions of parent and child incomes. The marginal distributions hold the information on within-generation income distribution and states, while the copula determines the transmission mechanism and co-dependence between generations. Various measures of IGM are different functions of this joint distribution and may reflect subjective consideration and measurement of within as well as between states and ranks. There is no single measure that commands consensus, or accounts satisfactorily for all aspects of mobility.

Current literature has focused on measures of IGM that primarily focus on the "dependence" component, and "predictability." Common statistics that

[2] We will briefly refer to debiased ML methods which, in effect, withhold a priori important covariates from selection, as in Double ML algorithms. Our prior preference for RF type methods is based on superiority of model averaging over LASSO type "model selection," see Maasoumi et al. (2024). Unless a priori constrained, ML methods are designed to find the best fitting combination of covariates for an outcome. Partial effects of important covariates may be obscured and estimated with bias. We will also examine Gradient Boosting which is a means of furthering accuracy and speed of optimization in RF algorithm and generally.

summarize this aspect of the joint distribution of parent and child income are (i) intergenerational elasticities (IGE) of child income with respect to parent income, gleaned from log-log or (ii) rank-rank regressions, i.e., partial correlations, and (iii) rank transition probability matrices. Methods (ii) and (iii) reflect and depend on the copula, while IGE combines some features of both the copula and marginal distributions.

Methods that only summarize the copula and not the marginal distributions are useful for distinguishing changes in inequality from intergenerational mobility. However, it is important to note that doing so creates relative measures of IGM, i.e. measures that only depend on intergenerational ranks, and may be misleading depending on the normative question at hand.[3]

Our chosen measure of IGM, is the *Probability of Upward Mobility*: Let $M_{increase}(p)$, equal to 1 if the child stayed at or moved up p-percentiles compared to the parent rank, and 0 otherwise. Similarly, let $M_u(p)$ be 1 if the child has moved down p-percentiles from the parent, and 0 otherwise. Since a child whose parent is already in the top p-percentile cannot move up, the highest p-percentile should be eliminated from the sample to avoid ambiguity. The same holds true for children with parents in the lowest p-percentile. [4]

Formally, for parent-child pair i,

$$M_u(p)_i = \mathbb{1}\{F_c(w_c) \geq F_p(w_p) + p\},$$

where, w_c is the income of the child, w_p is the income of the parent, $F_c(\cdot)$ is the child's income distribution, and $F_p(\cdot)$ is the parent's income distribution. So for example, if $p = 0.2$ then $F_c(w_c)$ is the quantile rank at w_c of the child is being compared to $F_p(w_p) + 0.2$, the quantile rank of the parent plus 20 percentile points.

$$Prob(M_u(p)_i = 1) = P_{x_i}$$

We require flexible and robust models for P_{x_i} in comparison to logit, probit and other specific parametric distributions, and flexible functionals of the "features" in x_i. See below for ML solutions to this "classification"/estimation stage.

Our proposed Mobility Measure for a sample $i = 1, ..., n$, is

$$M_{society\_u}(p) = \frac{1}{n} \sum_{i=1}^{n} M_u(p)_i,$$

a measure between 0 and 1 of the portion of off springs that move up by at least p-percentile compared with their parents.

[3] An absolute measure of income mobility that measures mobility relative to some income level might be, for example, the percentage of children earning more than their parents or the average difference between child income and mean parent income.

[4] These measures are mappings of conditional quantiles by well known probability transforms that are employed to define copulas.

We also consider a second measure of IGM, of absolute mobility: this is based on an indicator/count that is equal to 1 if a child's conditional (inflation adjusted) income is greater than some amount, and 0 otherwise.

Formally, for parent-child pair i,

$$A_u(t)_i = \mathbb{1}\{w_c \geq w_p + t\},$$

where, t is the real increase in income of a child over their parent.

Therefore, we propose the absolute IG mobility as,

$$A_{society_u}(t) = \frac{1}{n}\sum_{i=1}^{n} A_u(t)_i,$$

a measure between 0 and 1 that reflects the portion of offspring incomes that increased by at least t over their parents.

1.2 Related literature

Hertz (2005) investigated transition probabilities to measure IGM by race. Using the PSID, Hertz estimated large racial deficits in the probability of leaving the bottom quartile for blacks. He also found evidence of greater downward mobility among blacks in the probability of leaving the top quartile. Hertz employed probit models, and found these racial differences could not be explained by parental income or education. Bhattacharya and Mazumder (2011) employed larger samples than are afforded by PSID, using the NLSY.

Transition probability measures the probability that a child is at or above the s-th quantile of F_c, conditional on the parent being between the s_1 and s_2 quantiles of $F_p(\cdot)$.

That is,

$$\theta(s,(s_1,s_2)) = \frac{Prob[F_c(w_c) \geq s, s_1 \leq F_p(w_p) \leq s_2]}{Prob[s_1 \leq F_p(w_p) \leq s_2]}$$

$\theta(s(s_1,s_2))$ is decomposable by levels of discrete and continuous characteristics X of both generations:

$$\begin{aligned}
&\theta(s(s_1,s_2)) \\
&= \int Prob[F_c(w_c) \geq s | s_1 \leq F_p(w_p) \leq s_2, X = x] d_F(x | s_1 \leq F_c(w_c) \leq s_2) \\
&:= \int \theta(x; s, (s_1,s_2)) dF(x | s_1 \leq F_c(w_c) \leq s_2)
\end{aligned}$$

where

$$\theta(x; s(s1,s2)) = Prob[F_c(w_c) \geq s | s_1 \leq F_p(w_p) \leq s_2, X = x]$$

[5]

[5] We prefer the Conditional Upward Mobility concept and measures due to their distinction

2 Estimation

The currently popular regression methods are focused on parametric model for conditional mean of offspring outcome, given parental income and other features. Many of these Intergenerational Mobility Elasticity (IGE) methods may be represented as follows (which could in turn undergo high dimensional ML extensions):

The age-adjusted rank-rank slope (i.e. (ii)) is obtained from the regression:

$$RankI_c = \gamma_0 + \gamma_1 RankI_p + \gamma_2 Age_c + \gamma_3 Age_c^2 + \gamma_2 Age_p + \gamma_3 Age_p^2 + \epsilon_c, \quad (1)$$

where $RankI_c$ is the rank of the child in the child's income distribution, $RankI_p$ is the rank of the parent in the parent's income distribution, Age_c is the child's age, and Age_p is the parent's age. The coefficient γ_1 is the rank-rank slope coefficient and unaffected by inequality differences in the marginal distributions.

Similarly, the age-adjusted IGE (i.e. (i)) is obtained from the following log-log regression:

$$\log(I_c) = \lambda_0 + \lambda_1 \log(I_p) + \lambda_2 Age_c + \lambda_3 Age_c^2 + \lambda_2 Age_p + \lambda_3 Age_p^2 + \epsilon_c, \quad (2)$$

where $\log(I_c)$ is the natural logarithm of the child's income and $\log(I_p)$ is the natural logarithm of the parent's income. The coefficient λ_1 is the IGE and can be interpreted as approximately the percent change in income of a child that is caused by a percent change income of the parent, all else being equal. The higher the IGE, the less mobility, and the higher predictive role of parental income (and other characteristics).[6]

Methods (i) and (ii) depend strongly on the functional form of the relationship between parent and child income. In particular, they may fail in representing the commonly observed nonlinearities, and do not characterize the heterogeneity in mobility by subgroups, e.g. mobility might be different for African-American parents and children. While this might be remedied by computing (i) or (ii) by subgroup, that would still not be informative as to which subgroups are most relevant in predicting mobility differently for the general population, see Maasoumi et al. (2024) for further criticisms of parametric regression IGEs.

Random Forest and Gradient Boosting machine learning methods estimate the conditional probabilities required in our proposed IGM measure. This can robustly determine which characteristics are the most important predictors of

and advantages over transitional probabilities, as propounded in Bhattacharya and Mazumder (2011) The latter point to challenges in statistical inference in nonparametric estimation of these measures, and provide large sample methods for their nonparametric setting. Our inferences can be based on Lechner and Okasa (2024) for the high dimensional Random Forest ML setting.

[6] Mazumder (2018) gives the example: "If, for example, the IGE is 0.2, then this would imply that approximately 20 percent of existing income gaps between families would be expected to persist to the next generation and that these gaps would largely disappear within three generations. In contrast an IGE of 0.6, paints a dramatically different picture of mobility, where income differences persist for 5 or 6 generations."

upward mobility. We find that family wealth is a relatively important predictor of large increases and decreases in IGM, suggesting that family wealth, not just income, is an important factor in determining upward mobility. We are able to identify groups for which wealth is especially important. Random Forests are also able to identify important "nodes" in parental age levels, as well as other covariates.

The importance of distinguishing between income and wealth was argued by Thomas Piketty in his popular book *Capital in the Twenty-first Century* (2017). Piketty asserts that, in the long term, the rate of return on capital (r) is greater than economic growth (g) and causes wealth to be concentrated among the rich. This is a generational argument. While Blume and Durlauf (2015) question the theoretical foundations of Piketty's claim, they allow that further theoretical and empirical analysis is warranted.

Using the Panel Survey of Income Dynamics (PSID), Pfeffer and Killewald (2017) find that parent and grandparent wealth are important predictors of wealth mobility and that most of the advantages of wealth arise early in a person's life, and not through bequests of wealth. Also using PSID data, Brady et al. (2020) obtain the additional finding that childhood wealth is less important than childhood income in predicting mobility.

Based on regressions and Intergenerational Income Elasticities (IGEs), Solon (1992) showed that using only a single year of parent income as a proxy for "lifetime income" can produce a downward bias in IGE estimates (also see Mazumder, 2005, Haider and Solon, 2006, Nybom and Stuhler, 2016, Nybom and Stuhler, 2016, and Bloise and Raitano, 2021). Using an average of several years of a parent's income, instead of just one, he suggested that there was significantly less mobility in the US than previously thought. Since then, using a measure of lifetime income is standard practice for measuring income mobility. See Mazumder (2018) for an excellent review of papers using PSID to measure income mobility in the US.

Many mobility studies also focus on the influence of occupation, education, race, gender, and geographical location, among other factors. Chetty, Hendren, Kline, and Saez (2014) and Chetty, Hendren, Kline, Saez, and Turner (2014) are influential studies because they used rich administrative data to measure IGM.

Machine learning and algorithmic methods are still not widely used in economics because their primary focus would appear to be good (conditional) predictions. Econometrics, however, is equally concerned with the estimation and inference of partial effects and policy evaluation based on certain features. (see Athey and Imbens, 2019).[7] Machine learning methods are being developed to address the need for "causal analysis" (see below for some references to double machine learning and automated debiased partial effect machine learning). Random Forest and appropriate regression tree methods are recognized as better

[7] Athey and Imbens, 2019 review and suggest new avenues for research that use machine learning methods and are relevant to economics problems. Certain ML methods, such a the naive LASSO may eliminate certain features or obtain highly biased estimates of partial effects. This is due to a kink at "zero" that favors elimination of many single features.

than LASSO in identifying feature effects, and good competitors for Debiased ML (DML). See Lechner and Okasa (2024) who examine Ordered Forests for flexible estimation for conditional choice probabilities while taking the ordering information into account. Ordered logit would be a parametric competitor. Our binary choice model here is a special case. [8]

Some notable uses of machine learning to estimate mobility include Bloise et al. (2021), Brunori and Neidhöfer (2021), Mullainathan and Spiess (2017), and Salas-Rojo and Rodríguez (2022).

Inference about "Average Treatment Effect" at the mean, or the mean of Treatment Effects, is also possible with DML and RF-type MLs, see Lechner and Okasa (2024), Wager and Athey (2018). Traditional econometric methods give the impression of being able to more powerfully, or cleanly, assess partial and policy effects for individual covariates and treatment effects. But this is an artifact of *objects defined by restrictive parametric models* (mostly linear additive ones), rather than real complex nonlinear and heterogenous effects.

There are several advantages to machine learning tree methods: (1) they are nonparametric and do not make strong functional form assumptions, (2) the models are validated and not subject to over-fitting concerns, and (3) give easily interpretable importance charts ranking predictors' importance, as well as indicating the contribution of groups within the sample.[9]

The RF estimator is based on the regression random forest algorithm as introduced by Breiman (2001). They make use of cumulative probability predictions based on binary indicators of categories to estimate the single choice probabilities of the particular category event, conditional on covariates. Partial effects of the covariates can be found by numerical derivative approximations for estimation of the mean marginal effects and marginal effects at mean as the typical quantities of interest. [10] Random Forest and Gradient Boosting methods rely on construction of multiple regression trees. These trees are built using diverse sets of predictive covariates, and splits in the ranges of covariate values, including factors like age, occupation, education levels, and wealth tiers. This is the source of RF ability to decompose outcomes by sources and groups. To mitigate overfitting issues commonly associated with ML, we apply cross-fitting to optimize tree selection.

Random Forest trees are "bagged," or averaged across numerous bootstrap resamples of the original data, while Gradient Boosting trees are refined with each iteration. These approaches are akin to optimal model averaging based on predictive criteria. While each individual regression tree may be weak at predicting mobility, averaging many randomly selected regression trees improves prediction.

[8] The situation that we consider for offspring outcomes is similar to other outcome variables measured on an ordered scale such as level of education defined by primary, secondary and tertiary education, or income coded into low, middle and high income level (or defined by parent level income percentiles). Other examples are survey outcomes on self assessed health status (bad, good, very good), level of life satisfaction and happiness, or political opinions (do not agree, agree, strongly agree), or grades and scores.

[9] See Section 4.

[10] the asymptotic results of Wager and Athey (2018) can be applied.

Section 3 describes the data used in this paper. Section 4 describes how we define and measure IGM, and how the Random Forest and Gradient Boosting algorithms work. It also discusses how to measure predictor importance. Further details on RF and GB can be found in the appendix along with illustrative examples. Section 5 analyzes the results. Section 6 concludes.

To avoid confusion between machine learning and econometric terminology, we note the following terminology: covariates or regressors are also called *features*. Data used in estimation is called the *training sample*, whereas data used to determine out-of-sample performance is called the *testing sample*. Estimating an unordered discrete dependent variable is known as *classification*, and when there are only two possible outcomes, it is called *binary classification*. *Hyperparameters* or *tuning parameters* are parameters whose values are set to control the learning algorithm, whereas endogenously determined parameters are usually called *weights*.

3 Data

This paper uses data from the Panel Study of Income Dynamics (PSID).[11] The PSID is a longitudinal data set of over 18,000 individuals in 5,000 households living in the US starting from 1968. It uses March Current Population Survey (CPS) weights to make the sample nationally representative and contains income, wealth, occupation, health, and demographic information.

This paper compares heads of households in 1991 with their offspring in 2017 between the ages of 25 and 64. The sample is made up of 1,085 parent-child pairs. To ensure the results are representative of "lifetime" income and not driven by transitory income phases, parents' incomes are averaged from 1985 to 1994 (annually) and children's incomes are averaged from 2009 to 2019 (biennially).[12] Only children present in the household in 1991 are used.[13]

Table 1 shows some descriptive statistics about the sample. While the parent and child samples share similar characteristics, children in the 2017 sample notably have higher levels of education than their parents.

Table2 displays family wealth in 1989, defined as all assets net of debt plus home equity,[14] and the correlation between family wealth, parent permanent income (income averaged over 1985-1994), and child income (averaged over 2009-2019). While wealth and income are correlated, they are not as highly correlated as one might expect.

The predictive variables/features used in the Random Forest and Gradient Boosting algorithms are: parent age, parent income averaged over three years, family wealth, and dummy variables for parent gender (female = 1), married,

[11] Available at https://psidonline.isr.umich.edu/.

[12] The PSID was conducted biennially after 1997.

[13] An earlier version of this paper only used incomes of parents and children averaged over only 2 years and have similar results.

[14] The PSID defines the assets as `S203:` value of farm/business, `S205:` value of checking/savings, `S209:` value of other real estate, `S211:` value of stocks, `S213:` value of vehicles, and `S215:` value of other assets, and defines `S207:` value of other debt.

in a geographical region,[15] in an industry, in an occupation,[16] race (White, African-American, and Other), and highest level of education completed (None, Some High School, High School, Some College, College, and Postgraduate).

4 ML Methods

4.1 Random Forest

Consider the *Random Forest* method as in (Breiman et al., 1984, Breiman, 2001). RF is an example of ensemble techniques which combine multiple models. Randomized model/regression averaging, over many feature combinations and their splits. Model averaging is known in econometrics for superior performance and robustness to model misspecification. RF can be used for classification (discrete outcomes), or for regression with a continuous dependent variable. The classification may be over K classes, possibly ordered. In this paper K = 2.

Consider N observations on covariates w, and an outcome response, for each of N observations: that is, (x_i, y_i) for $i = 1, 2, \ldots, N$, with $x_i = (x_{i1}, x_{i2}, \ldots, x_{iw})$. The idea is to partition the sample into M regions. The partitioning is done sequentially based on the covariates x_i passing a threshold. In this way, the space of all joint predictor values is partitioned into disjoint regions.

For example, consider the observation:

Obs.	Parent Age	Parent Income	South	Parent Black	$M_{increase}(20)$
x_1	35	100,000	No	No	1

A possible tree to classify $M_{increase}(20)$ could be:

[15] The four regions used are

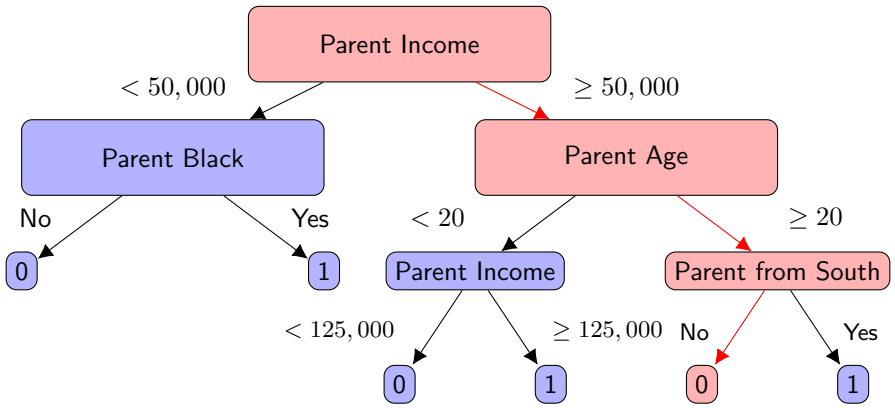
NORTHEAST: Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont

NORTH CENTRAL: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin

SOUTH: Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, Washington DC, West Virginia

WEST: Arizona, California, Colorado, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming.

[16] Occupation and industry dummy variables are from from 1970 Census of Population. Please refer to Appendix 2, Wave XIV documentation, for complete listings.



In this case, $T(x_1) = 0$ (bases on x_1 the decision tree predicts y_1 will be 0) but $y_1 = 1$, so the observation is misclassified (as indicated by the red arrows).

Because any single tree is likely a poor classifier, the Random Forest algorithm constructs many trees from bootstrap resamples and makes a prediction based on the "forest" (some decision rule taking into account all trees), as depicted in Figure 1. For example, a "majority" rule adopts the most commonly selected tree selections.

4.1.1 Random Forest Algorithm

The algorithm creates a prediction of the response variable by constructing *trees*, which can formally be expressed as

$$T(x; \Theta) = \sum_{m=1}^{M} \gamma_m \mathbb{1}(x \in R_m),$$

with parameters $\Theta = \{R_M, \gamma_m\}_1^M$ for $m = 1, 2, \ldots, M$. That is, $R_1, R_2, \ldots, R_M$ are disjoint partitions of the parameter space (terminal nodes).

For regression minimizing $\sum_{i=1}^{N}(y_i - T(x_i))^2$ obtains the best $\hat{\gamma}_m$ is $\hat{\gamma}_m = \text{ave}(y_i | x_i \in R_m)$. For classification (our case), $\hat{\gamma}_m$ is either 0 or 1.

The utilization of bagging and randomized tree and node selection helps address biases and overfitting issues inherent in traditional regression trees. Decision objectives guide the tree and node searches, incorporating criteria such as least mean prediction error squares, entropic measures, and related error variation metrics like "mean Gini" (refer to the details below). These methods collectively enhance the robustness and effectiveness of our predictive modeling framework.

There are two common ways to determine covariate importance in the overall prediction of the model with tree methods: (1) how much less accurate the prediction becomes if the variable is not included in the construction of the decision trees and (2) if that variable was given random values, how much would accuracy of the prediction decrease. These two ways of determining importance

usually rank covariates similarly. The first way can be standardized to sum to one (so it is easier to compare), and the second way can handle a misleading importance ranking if variables are highly correlated or if the model is overfitting the data.

Following a training step, models are tested against a hold-out sample. This testing step is crucial for reinforcing the out-of-sample predictive performance and avoiding over fitting, an important consideration in policy evaluation.[17]

4.2 Gradient Boosting

This approach is an adaptation of iterative log-likelihood optimization methods and steepest ascent algorithms in MLE. Using squared error to measure closeness, *Gradient Boosting* favors a tree at the mth iteration that is as close as possible to the negative gradient of the log-likelihood,

$$\bar{\Theta}_m = \arg\min_{\Theta} \sum_{i=1}^{N} (-g_{im} - f(x;\Theta))^2,$$

where $g_{im} = \frac{\partial L(y_i, f_m(x_i))}{\partial f_m(x_i)}$, $f(\cdot)$ is model (i.e., a tree), and the loss function (L) is the logistic loss function (defined below in the appendix).

4.2.1 Gradient Boosting Algorithm

The algorithm is depicted in Figure 2 where the algorithm starts with an initial tree and with each iteration it adds an improved tree. For more details on Gradient Boosting and an example, see the appendix.

4.2.2 Evaluating the Model: Accuracy and Precision

The accuracy of a machine learning algorithm is a measure of how well a classification approach performs on a given dataset. Accuracy is calculated as the ratio of correctly predicted instances to the total number of instances in the testing sample. This is familiar from discrete choice econometrics. The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}.$$

It can further be broken down into

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

where True Positives (TP) are instances that are correctly predicted as positive (i.e., $\sum_i \hat{f}(x_i) = 1$ when $y_i = 1$), True Negatives (TN) are instances that are correctly predicted as negative (i.e., $\sum_i \hat{f}(x_i) = 0$ when $y_i = 0$), False Positives

[17] See Appendix for additional details on the RF algorithm as well as an example.

(FP) are instances that are incorrectly predicted as positive (i.e., $\sum_i \hat{f}(x_i) = 1$ when $y_i = 0$), and False Negatives (FN) are instances that are incorrectly predicted as negative (i.e., $\sum_i \hat{f}(x_i) = 0$ when $y_i = 1$).

Accuracy can be a misleading metric, especially in cases where the event being predicted is rare. For example, if the model almost always predicts a 0 (i.e., the event does not occur) and only occasionally predicts a 1 (i.e., the event does occur), even if those rare predictions are incorrect, the model may still achieve a high accuracy score simply because the outcome is infrequently 1. This high accuracy is therefore an artifact of the event's rarity rather than the model's actual predictive power. To better assess performance in such cases, other metrics like precision, recall, or the F1 score provide a more reliable evaluation, as they account for the model's ability to correctly identify rare events.

Since it may be unlikely for a child to move far up in the income distribution relative to their parent, precision we also consider precision, defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{TN}}.$$

5 Results

Table 3 shows the age-adjusted intergenerational elasticities. Parent incomes are averaged over the 1985-1994 period, and child incomes are averaged over the 2009-2019 period. Chetty, Hendren, Kline, and Saez (2014) found an IGE of 0.34 using IRS data for children 29-32 (income averaged over two years) and parents (income averaged over 5 years) between the years 1996 and 2000. Using a similar sample with PSID data, Mazumder (2016) found an IGE of 0.28 to 0.33 depending on the sample. Both IGEs are similar to the age group 25-44 in this paper, 0.28.

Mazumder (2018) argues the primary reason for the low IGE of Chetty, Hendren, Kline, and Saez (2014) is the short averages of parent and child incomes, which are not good representations of lifetime incomes. In contrast, studies using PSID with parent and child incomes averaged over longer periods tend to have IGEs of 0.4 to 0.6.[18] Additionally, Chetty, Hendren, Kline, and Saez (2014) contends that rank slope coefficients are more robust to classical measurement error stemming from transitory income shocks and the lack of a lifetime income measure.[19]

Table 4 shows age-adjusted rank slope coefficients. Across all ages overall, a 10-percentile increase in the parents income rank is associated with a 4.1 percentile increase in the child's rank.

Table 5 presents the IGE and rank slope coefficients from regressions that include all relevant covariates. These linear models provide useful benchmarks

[18] Restricting our sample to only parents and children whose incomes are observed over (for example) a 20 year period would make the sample too small for meaningful predictions, particularly since this paper attempts to use as current data as possible.

[19] An earlier version of this paper used parent and child incomes averaged over a two-year period, but yielded similar results to the current study. This suggests that RF and GB, like rank slope coefficients, are robust to classical measurement error.

for comparing the results obtained from the machine learning models. Since the model is linear, it does not capture non-linear relationships or complex interaction effects between regressors, which are accounted for by RF and GB. For instance, the effect of age for a married female may differ significantly between the South and the West and may differ further for someone with a college education.

Despite the advantages of more complex models like RF and GB, a simple linear model still offers valuable utility due to its ease of interpretation. In this sense, RF and GB serve as complements to the IGE and rank slope coefficients, providing a more nuanced understanding, rather than complete substitutes.[20]

Table 6 shows the income quintile transition matrix.

Using the RF model, Figure 3 highlights the most influential features for upward mobility in the income distribution, based on the average decrease in "Gini impurity" — a measure used by RF models to determine feature importance. Intuitively, the average decrease in Gini impurity reflects how effectively a feature splits the data into distinct groups. The greater the decrease in impurity, the more significant the feature is in improving the model's classification accuracy.[21] In this context, wage consistently emerges as the most critical predictor of income movement. However, family wealth gains relative importance when predicting larger upward shifts in income rank.

Despite its utility, the average decrease in Gini impurity can sometimes be misleading, especially in datasets with features of different scales, highly correlated variables, or when the model exhibits low precision. Features with more categories or greater variability may naturally cause a larger decrease in Gini impurity, not because they are more informative, but due to their inherent structure. When model precision is low, it indicates a higher rate of false positives, suggesting that the model may be overfitting to noise rather than identifying true patterns. In such cases, even if a feature appears important based on Gini impurity, it may not genuinely contribute to meaningful predictions. This is particularly problematic in models predicting rare outcomes, such as a child moving up 40 percentile points above their parents' income rank. In such cases, the model may achieve high accuracy simply by predicting that nearly no one will achieve this level of upward mobility, but the precision remains low, rendering the average decrease in Gini impurity misleading.

To address these potential distortions, permutation tests are conducted as a robustness check. This approach is inspired by Shapley value decompositions (refer to Shorrocks, 2013), where each factor's predictive power is isolated by shuffling its values and observing the impact on model performance. Intuitively, permutation importance evaluates how much the model's accuracy deteriorates when a specific feature's information is removed, providing insight into the fea-

[20] RF and GB importance plots reveal the magnitude of a feature's impact, but not its direction since the direction is likely to change due to complex interactions with other variables. While counterfactual analyses, like the one shown in Table 7, can provide additional insights, simple linear interpretations still have value. These linear models tell the association assuming the relationship is linear, offering a straightforward understanding of feature influence.

[21] See the Appendix for a detailed example of how Gini impurity is calculated.

ture's unique contribution. Unlike Gini impurity, which can be influenced by feature characteristics and model precision issues, permutation tests offer a more reliable indication of feature relevance by examining how integral each variable is to the model's predictions.[22]

Figure 4, which presents a box plot illustrating the decrease in accuracy score from randomly shuffling the values of each feature 20 times, substantiates that wage is the most significant predictor of upward relative mobility, while family wealth gains prominence as the income gap between parent and child widens. Notably, in panel (d), for increases of 40 percentile points or more, wealth emerges as the most critical predictor.

Figure 5 shows similar results for relative *downward mobility* in the income distribution using RF models. However, accuracy and precision scores are lower, and in Figure 6, panels (c) and (d), randomly shuffling the values of each feature has almost no effect on accuracy. This is likely due to the smaller sample size (reduced by 30-40% respectively) and the rarity of such large movements in income rank.

Figures 7, 8, 9, and 10 use Gradient Boosting. The Gradient Boosting models generally have higher accuracy and precision than the Random Forest models. The results are, however, similar to the Random Forest findings.

The relative lack of contribution of many other covariates, generally, is notable. These covariates include ethnicity, education (below college), occupation type, employment status, and geographical regions. This finding does not necessarily imply that none of these factors are inconsequential. ML methods find other combinations of factors (wages and wealth here) that account for outcomes just as well. In other words, it is possible that these other factors work through, or are well proxied by the other factors.

It is important to note thats when both family wealth and parent income are included in these models, neither is redundant for prediction. For example, in panel (b) of Figure 8, by randomly assigning values for family wealth, the mean accuracy of the model roughly decreased between 0.01 and 0.06, with over half the permutation runs deceasing accuracy by over 0.03. If the inclusion of family wealth were redundant with respect to parent income, the accuracy by which the trees classify the data would be relatively unchanged (if values of wealth were randomly assigned). This is also in the spirit of Shapley decompositions.

5.1 Absolute Mobility Measure

Figures 11, 12, 13, and 14 use Random Forest to predict absolute mobility. As with relative mobility, *family wealth is relatively more important for larger increases*. Interestingly, these results are consistent only for the permutations regarding decreases in absolute mobility (and the precision in Figure 11 is lower than that in Figure 13). Thus, while downward relative mobility is more difficult to predict than upward relative mobility, downward absolute mobility is easier to predict than upward absolute mobility.

[22] See appendix for more details on determining feature importance with mean decrease in Gini impurity and permutations.

Figures 15, 16, 17, and 18 use Gradient Boosting to predict absolute mobility. Again, the Gradient Boosting models generally have higher accuracy and precision than the Random Forest models. The results are, however, similar to the Random Forest models.

5.2 Counterfactual Prediction Exercise

Table 7 displays predictions of upward mobility of seven representative parental profiles (predictions of $M_u(10)$ and $A_u(5000)$ with RF and GB).[23] All representative parental profiles are married, male, aged 40, employed in retail service jobs (i.e., work in the retail industry and their occupation is a service job), non-self-employed, and reside in the Northeast. By default, they are white, high school graduates, with an annual income of $25,000 and family wealth of $55,000. Variations in the profiles include: (1) being college-educated, (2) having only some high school education, (3) being black with no family wealth, (4) being white with no family wealth, (5) being black, (6) being white (default), and (7) earning a higher income of $40,000 per year.

Retail service workers were selected for demonstration purposes, as they are considered "blue collar" and may have limited upward IGM potential. An annual income of $25,000 places an individual in the bottom 40% of earners. If an individual earns $50,000 instead, as shown in Table 7 (7), he is not expected to see an increase in either relative or absolute IGM, likely due to "regression to the mean."(The only difference between (7) and (6) in Table 7 is that the individual in (7) has a higher income of $50,000 instead of $25,000.) This refers to the tendency for extreme values, such as unusually high or low earnings, to move closer to the average over time. In this context, if a parent's earnings are at an extreme, their children's earnings are likely to be closer to the overall average, rather than continuing at the extreme level.

If the individual's highest level of education is only some high school, $A_u(5000)$ is not predicted to increase. However, if the individual had college as their highest level of education, as shown in Table 7 (1) instead of (2), $A_u(5000)$ would be predicted to increase. While it is not clear what caused this immobility, which defies regression to the mean, a policymaker crafting policy would likely want target this individual.

RF predicts that Whites will experience increases in both absolute and relative IGM, while Blacks will not, as shown in Figures 7 (5) and (6). However, GB does not yield similar results, and the predictions are not robust.

From Figures 7 (3) and (4), having low wealth does not decrease the probability of either absolute or relative IGM for Blacks or Whites, except in the case of GB, where being Black is associated with a decreased probability of absolute IGM. However, the effect of wealth on upward mobility is nonlinear.

[23] We focus specifically on $M_u(10)$ and $A_u(5000)$ because higher levels of relative or absolute upward IGM generally yield predictions of "0." This is likely due to the fact that, at these higher levels, the relationship between IGM and the variables becomes more nonlinear. As a result, "stereotypical" profiles are less effective in predicting these rare events, where the dynamics of upward mobility are more complex and less predictable.

This means that the relationship between wealth and IGM may vary at different income levels. Small amounts of wealth might not significantly impact mobility, but as wealth increases, its influence could become more pronounced or take on a different form altogether interacting with other features.

6 Conclusion

Predictors of IGM have important policy implications. Machine learning methods such as Random Forest and Gradient Boosting are nonparametric, give a performance indicator based on out-of-sample prediction, and allow predictors to be ranked. Using such methods, this paper obtains a robust confirmation of parent income as the most important predictor of both relative and absolute IGM. Nevertheless, we find that family wealth is important as well, particularly for large changes in the income distribution. Immobility in the higher incomes classes is known to be high, and is given high implicit weights in Least Squares regression based studies of IGM, see Maasoumi et al. (2024).

Our results corroborate the findings in Brady et al. (2020), who state "The evidence mostly contradicts the prominent claim that childhood wealth is more important than childhood income. Indeed, the analyses mostly show that childhood income explains more of BW disadvantages and has larger standardized coefficients than childhood wealth." While Brady et al. (2020) employ standard regression-based methods, our use of RF and GB methods offers greater robustness to functional form misspecification.

The finding that wealth is nearly as important as parental income in predicting certain measures of IGM carries significant policy implications. While family wealth may not directly cause IGM, when both family income and wealth are included in empirical models, neither variable becomes redundant in predicting outcomes. This suggests that distinct causal mechanisms underlie the effects of wealth and income on IGM. Income is correlated with factors like education, occupational status, and industry, while wealth is linked to different factors, such as socioeconomic status and inheritance, and hence influence IGM through separate channels. As a result, policies aimed at promoting IGM—such as job training programs or investments in education—should consider both income and wealth in their design. These policies should factor in wealth disparities and incomes when determining which neighborhoods or communities to target for intervention.

If the relationship between wealth and IGM is indeed causal then policies, such as housing assistance for first time buyers, are important tools for both parents and offspring. Such policies would not only help individuals and families build wealth, but also foster long-term financial security and IGM, rather than temporal income policies that may only have short term consumption benefits. Given these potential long-term benefits, the causal role of wealth in IGM warrants further investigation and should be the focus of future research.

This paper also demonstrated the usefulness of ML tree methods (along with the measure of mobility proposed in this paper) to fashion highly personalized

policies (e.g., if someone is black, works in sales, is from the South, is age 37, etc., should they receive some intervention because their probability of upward mobility is low).[24] For example, in Table 7, service workers who are white, college-educated, work in retail, etc. may not require assistance if the goal is to increase their absolute upward mobility. However, the same individual may face challenges if their highest level of education is limited to some high school.

Further research could apply these methods and measures of IGM to predict health status, occupational status, or educational attainment. Additionally, both $M_{society\_u}(p)$ and $A_{society\_u}(t)$ can be calculated with the current generation's characteristics (and by subgroup). For example, it could be the case that $A_{society\_u}(5000)$ is smaller for the Black subgroup of the population than for White subgroup. This would mean that, on average, Black individuals have a lower probability of earning at least \$5,000 more than their parents than White individuals. These findings underscore the importance of considering subgroup differences when studying IGE and suggest directions for policy interventions aimed at reducing disparities.

[24] This is under the assumption that the features that predict mobility have not significantly changed in 25 years, as the model was trained on parents in 1991 and children in 2017.

References

Athey, S., & Imbens, G. (2019). Machine learning methods economists should know about.

Bhattacharya, D., & Mazumder, B. (2011). A nonparametric analysis of black–white differences in intergenerational income mobility in the united states. *Quantitative Economics*, *2*(3), 335–379.

Bloise, F., Brunori, P., & Piraino, P. (2021). Estimating intergenerational income mobility on sub-optimal data: A machine learning approach. *The Journal of Economic Inequality*, *19*(4), 643–665.

Bloise, F., & Raitano, M. (2021). Intergenerational earnings persistence in italy between actual father-son pairs accounting for lifecycle and attenuation bias. *Oxford Bulletin of Economics and Statistics*, *83*(1), 88–114.

Blume, L., & Durlauf, S. (2015). Capital in the twenty-first century: A review essay. *Journal of Political Economy*, *123*(4), 749–777.

Brady, D., Finnigan, R., Kohler, U., & Legewie, J. (2020). The inheritance of race revisited: Childhood wealth and income and black–white disadvantages in adult life chances. *Sociological Science*, *7*(25), 599–627.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth; Brooks.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brunori, P., & Neidhöfer, G. (2021). The evolution of inequality of opportunity in germany: A machine learning approach. *Review of Income and Wealth*, *67*(4), 900–927.

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1553–1623.

Chetty, R., Hendren, N., Kline, P., Saez, E., & Turner, N. (2014). Is the united states still a land of opportunity? recent trends in intergenerational mobility. *American Economic Review*, *104*(5), 141–47.

Haider, S., & Solon, G. (2006). Life-cycle variation in the association between current and lifetime earnings. *American Economic Review*, *96*(4), 1308–1320.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer New York Inc.

Hertz, T. (2005). *Rags, riches, and race: The intergenerational economic mobility of black and white families in the united states in unequal chances: Family background and economic success* (S. Bowles, H. Gintis, & M. Osborne Groves, Eds.). Princeton University Press.

Lechner, M., & Okasa, G. (2024). *Random Forest Estimation of the Ordered Choice Model* (Economics Working Paper Series No. 1908). University of St. Gallen, School of Economics and Political Science.

Maasoumi, E. (2020). On intergenerational mobility. *Emory University, Economics, Invited talk at HCEO conference, University of Chicago.*

Maasoumi, E., Wang, L., & Zhang, D. (2024). Generalized intergenerational mobility regressions. *working paper*.

Mazumder, B. (2005). Fortunate sons: New estimates of intergenerational mobility in the united states using social security earnings data. *The Review of Economics and Statistics*, *87*(2), 235–255.

Mazumder, B. (2016, August). Estimating the Intergenerational Elasticity and Rank Association in the United States: Overcoming the Current Limitations of Tax Data. In *Inequality: Causes and Consequences* (pp. 83–129, Vol. 43). Emerald Group Publishing Limited.

Mazumder, B. (2018). Intergenerational mobility in the united states: What we have learned from the psid. *The ANNALS of the American Academy of Political and Social Science*, *680*, 213–234.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Nybom, M., & Stuhler, J. (2016). Heterogeneous income profiles and lifecycle bias in intergenerational mobility estimation. *Journal of Human Resources*, *51*(1), 239–268.

Pfeffer, F. T., & Killewald, A. (2017). Generations of Advantage. Multigenerational Correlations in Family Wealth. *Social Forces*, *96*(4), 1411–1442.

Piketty, T. (2017). *Capital in the twenty-first century* (A. Goldhammer, Trans.). Belknap Press.

Salas-Rojo, P., & Rodríguez, J. G. (2022). Inheritances and wealth inequality: A machine learning approach. *The Journal of Economic Inequality*, *20*(1), 27–51.

Shorrocks, A. (2013). Decomposition procedures for distributional analysis: a unified framework based on the Shapley value. *The Journal of Economic Inequality*, *11*(1), 99–126.

Solon, G. (1992). Intergenerational income mobility in the united states. *The American Economic Review*, *82*(3), 393–408.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Tables

Table 1: Descriptive Statistics

	N=1085			
	Parent (1991)		Child (2017)	
	Mean or %	(Std. Dev.)	Mean or %	(Std. Dev.)
Demographics				
Age	39.2	(10.6)	34.2	(6.5)
Age group 25-34	40.3 %		57.7 %	
Age group 35-44	34.4 %		35 %	
Age group 45-54	12.5 %		6.5 %	
Age group 55-64	12.8 %		0.8 %	
White	77.8 %		75.4 %	
African American	21.3 %		22.3 %	
Other	0.8 %		2.3 %	
Male	79.4 %		71.5 %	
Female	20.7 %		28.5 %	
Education				
Less than high school	24.8 %		9.4 %	
High school	32.8 %		21.5 %	
Some college	18.5 %		29.2 %	
Bachelors	15.3 %		21.4 %	
Postgraduate	8.7 %		18.4 %	
Region				
Northeast	20.1 %		18.3 %	
North Central	27.7 %		25.6 %	
South	34 %		36.6 %	
West	16.2 %		18.9 %	
Other	1.5 %		0.5 %	
Income (2019 dollars)				
	1985-1994		2009-2019	
	51,164	(46,132)	51,875	(43,835)
Quantile 1 (lowest)	8,830	(5,843)	11,989	(5,783)
Quantile 2	26,730	(3,993)	28,807	(4,408)
Quantile 3	40,811	(4,595)	43,501	(4,277)
Quantile 4	61,013	(7,114)	60,551	(5,623)
Quantile 5 (highest)	118,993	(57,875)	114,814	(57,465)

Note: the table above shows descriptive statistics about demographics and incomes of the sample. Incomes of the parent generation are averaged over 1985 to 1994 annually and incomes of the child generation are averaged over 2009 to 2019 biennially.

Table 2: Family Wealth

	Mean	(Std. Dev.)
Family Wealth 1989 (in 2019 dollars)		
	197,977	(504,124)
Quantile 1 (lowest)	-13,803	(113,028)
Quantile 2	13,766	(8,382)
Quantile 3	56,885	(19,184)
Quantile 4	150,676	(45,161)
Quantile 5 (highest)	785,692	(904,522)
Correlation between Family Wealth 1989 and Parent Income 1985-1994		
0.25		
Correlation between Family Wealth 1989 and Child Income 2009-2019		
0.08		
Correlation between Parent Income 1985-1994 and Child Income 2009-2019		
0.39		

Note: the table above shows descriptive statistics about family wealth of the sample. Incomes of the parent generation are averaged over 1985 to 1994 annually and incomes of the child generation are averaged over 2009 to 2019 biennially.

Table 3: Inter-generational Elasticities

	IGE	(SE)	N
Overall	0.3^{***}	(0.05)	1085
By age (4 groups)			
(1) Age 25-34	0.24^{***}	(0.043)	491
(2) Age 35-44	0.34^{***}	(0.04)	321
(3) Age 45-54	0.33^{***}	(0.06)	140
(4) Age 55-64	0.4^{***}	(0.09)	133
By age (2 groups)			
(5) Age 25-44	0.28^{***}	(0.03)	812
(6) Age 45-64	0.36^{***}	(0.05)	273
Test of differences (p-values)			
(2) vs. (1)	0.29		
(3) vs. (1)	0.37		
(4) vs. (1)	0.19		
(5) vs. (6)	0.31		

Note: the table above shows results from regressing log(income) of the parent on log(income) of the child, controlling for ages of the parents and children. $p^{*} < 0.10$, $p^{**} < 0.05$, and $p^{***} < 0.01$ based on two-tailed tests. Grouped by parent's age. Incomes of the parent generation are averaged over 1985 to 1994 annually and incomes of the child generation are averaged over 2009 to 2019 biennially.

Table 4: Inter-generational Rank Correlations

	Rank Slope	(SE)	N
Overall	0.41^{***}	(0.04)	1085
By age (4 groups)			
(1) Age 25-34	0.46^{***}	(0.04)	491
(2) Age 35-44	0.33^{***}	(0.06)	321
(3) Age 45-54	0.45^{***}	(0.07)	140
(4) Age 55-64	0.47^{***}	(0.08)	133
By age (2 groups)			
(5) Age 25-44	0.39^{***}	(0.03)	812
(6) Age 45-64	0.45^{***}	(0.05)	273
Test of differences (p-values)			
(2) vs. (1)	0.16		
(3) vs. (1)	0.92		
(4) vs. (1)	0.96		
(5) vs. (6)	0.47		

Note: the table above shows results from regressing the rank of the parent in their income distribution on the rank of the child in their income distribution, controlling for ages of the parents and children. $p^{*} < 0.10$, $p^{**} < 0.05$, and $p^{***} < 0.01$ based on two-tailed tests. Grouped by parent's age. Incomes of the parent generation are averaged over 1985 to 1994 annually and incomes of the child generation are averaged over 2009 to 2019 biennially.

Table 5: IGE and Rank Slope with Regressor

	IGE	Rank Slope
$\log I_p$	0.20^{***}	-
	(0.05)	-
$RankI_c$	-	0.27^{***}
	-	(0.04)
Age_c	0.12^{**}	4.9^{***}
	(0.6)	(1.33)
Age_c^2	-0.001^{*}	-0.06^{***}
	(0.0007)	(0.018)
Age_p	-0.001	0.22
	(0.03)	(0.84)
Age_p^2	0.00	-0.002
	(0.00)	(0.01)
Family Wealth	-0.00000024^{***}	-0.00000655^{***}
	(0.00)	(0.00)
Parent Female	-0.13	-2.82
	(0.16)	(4.24)
Parent Married	-0.07	0.04
	(0.13)	(3.62)
Parent Black	-0.38^{***}	-10.85^{***}
	(0.1)	(2.83)
Parent Highest Eduction: College	0.01	4.98
	(0.12)	(3.11)
Parent Highest Eduction: Some College	-0.62^{***}	-20.9^{***}
	(0.11)	(2.98)
Parent Highest Eduction: High School	-0.09	-1.13
	(0.09)	(2.61)
Parent Highest Eduction: Some High School	-0.43^{***}	-16.21^{***}
	(0.12)	(3.55)
Parent from Northeast	-0.05	3
	(0.15)	(5.32)
Parent from Northcentral	-0.07	0.1
	(0.15)	(5.38)
Parent from South	-0.15	-3.26
	(0.15)	(5.12)
Parent from West	-0.2	-1.59
	(0.16)	(5.41)
Constant	6.9^{***}	-36.95
	(1.33)	(27.55)
R^2	0.32	0.38
N	1078	1078

Note: the table above shows results from regressing log(income) of the parent on log(income) of the child and regressing the rank of the parent in their income distribution on the rank of the child in their income distribution, controlling for all covariates. Industry and Occupation coefficients not shown. $p^{*} < 0.10$, $p^{**} < 0.05$, and $p^{***} < 0.01$ Incomes of the parent generation are averaged over 1985 to 1994 and incomes of the child are averaged over 2009 to 2019.

Table 6: Income Quintile Transition Matrix

Parent (1985-1994)		Child (2009-2019)					
		Lowest [<= $21k]	Quintile 2 [$21k-$37k]	Quintile 3 [$37k-$52k]	Quintile 4 [$52k-$72k]	Highest [>=$72k]	Total
Lowest	[<= $19k]	42	21	12	14	11	100
Quintile 2	[$19k-$34k]	22	29	20	18	11	100
Quintile 3	[$34k-$50k]	22	28	21	17	12	100
Quintile 4	[$50k-$76k]	10	16	30	25	19	100
Highest	[>=$76k]	4	7	17	24	48	100

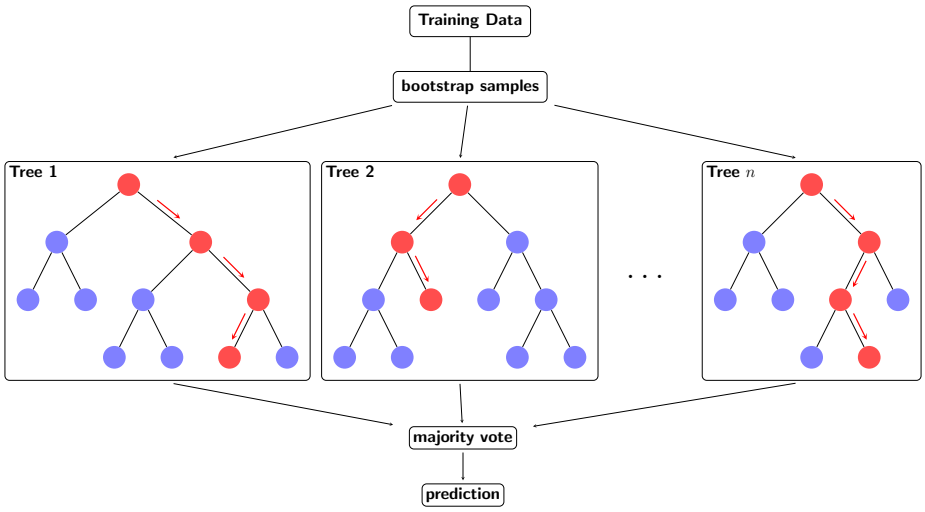
Note: The chart above shows the sample's income quintile transition matrix. In 2019 dollars. N=1085.
Incomes of the parent generation are averaged over 1985 to 1994 and incomes of the child are averaged over 2009 to 2019.

Table 7: Counterfactual Predictions of Upward Mobility

	Random Forest		Gradient Boosting	
	$M_u(10)$	$A_u(5000)$	$M_u(10)$	$A_u(5000)$
(1) College Educated	1	1	1	1
(2) Some High School	1	0	1	0
(3) Low Wealth Black	1	1	1	0
(4) Low Wealth White	1	1	1	1
(5) Black	0	0	1	0
(6) White	1	1	1	0
(7) Medium Income	0	0	0	0

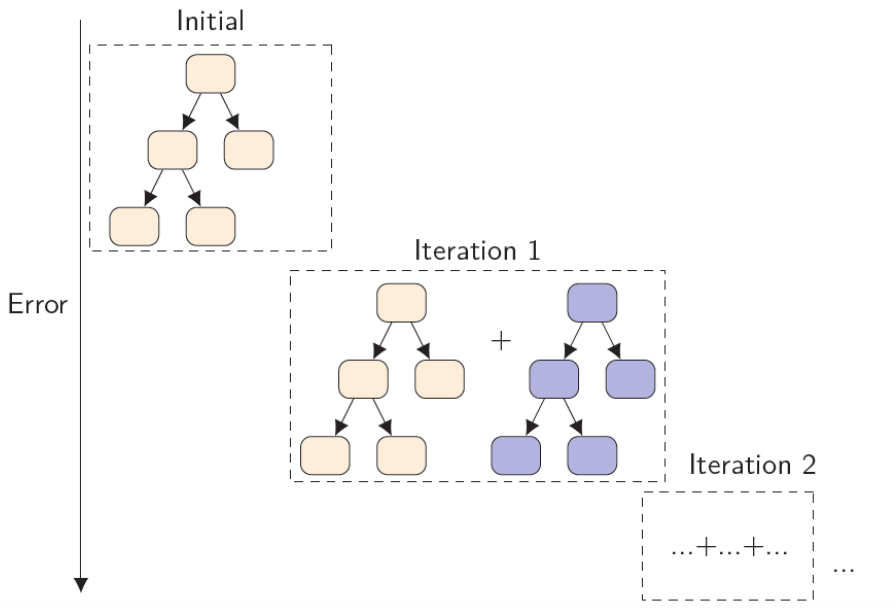
Note: The table presents a set of 7 representative parents used to predict IGM based on a variety of socio-economic and demographic features. These individuals are married, male, work in retail, service workers, age 40, are not self-employed, and are in the Northeast. By default, these individuals are white, high school educated, earn $25000 per year, and have $55,000 in family wealth. The specific cases that differ are: (1) college educated, (2) only attended some high school, (3) black with no family wealth, (4) white with no family wealth, (5) black, (6) white, and (7) earn $40,000 per year.

Figures



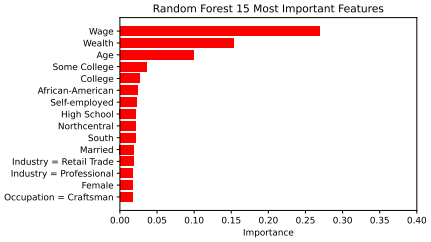
The RF algorithm draws bootstrap samples from the training data and construct trees. The prediction of the RF algorithm is what the majority of the trees predict.

Figure 1: Random Forest Algorithm

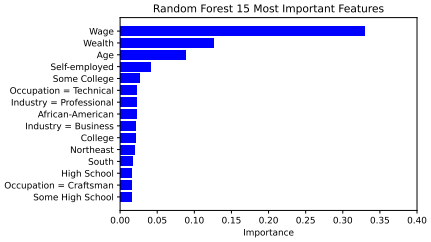


An initial regression tree is constructed, and with each iteration of the GB algorithm, a new tree is built to minimize the negative gradient of the loss function (the residuals). The weighted sum of these trees reduces the overall model prediction error.

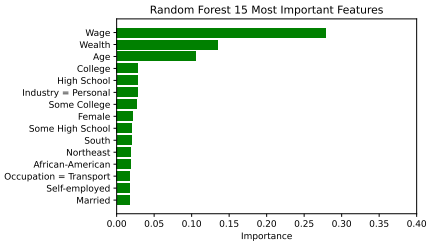
Figure 2: Gradient Boosting Algorithm



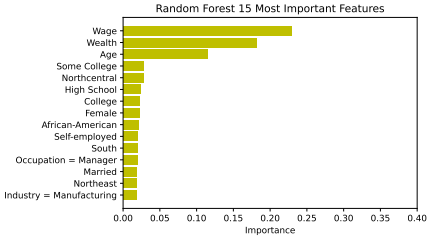
(a) Increase of 10 percentile or more.
Accuracy: 0.66
Precision: 0.44



(b) Increase of 20 percentile or more.
Accuracy: 0.75
Precision: 1



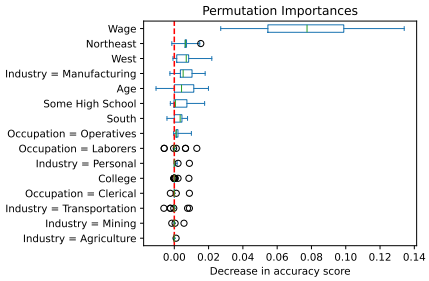
(c) Increase of 30 percentile or more.
Accuracy: 0.83
Precision: 0.67



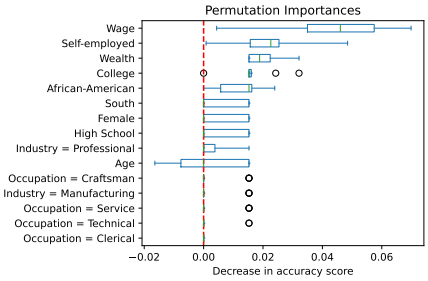
(d) Increase of 40 percentile or more.
Accuracy: 0.89
Precision: 72

The figures above show the importance plots of random forest models for predicting whether a child will be 10, 20, 30, or 40 percentile points higher than their parent. Feature importance is measured by the mean decrease in Gini impurity, which indicates how much a feature helps to make the model's predictions more accurate by creating purer, more distinct groups in the data after each split.

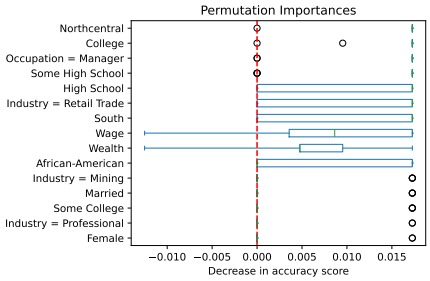
Figure 3: Random Forest, Most Important Features for Moving Up in the Income Distribution



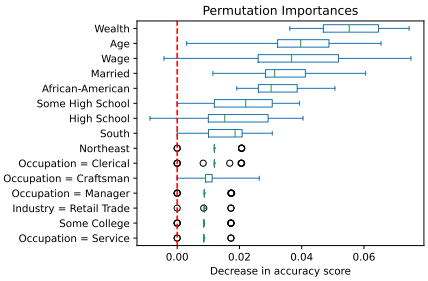
(a) Increase of 10 percentile or more.



(b) Increase of 20 percentile or more.



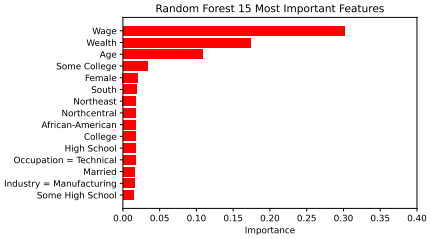
(c) Increase of 30 percentile or more.



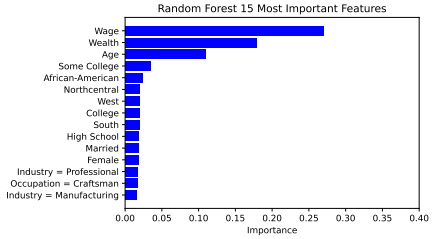
(d) Increase of 40 percentile or more.

The figures above show the box plots of feature importance from 20 permutations of random forest models for predicting whether a child will be 10, 20, 30, or 40 percentile points higher than their parent. In each permutation, one feature is randomly shuffled, and the resulting decrease in model accuracy is measured to assess its importance.

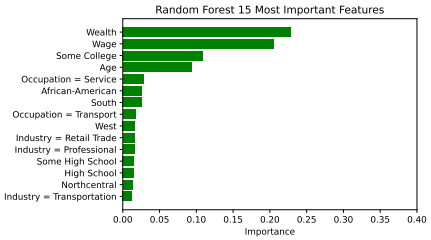
Figure 4: Random Forest, Permutations of Most the Important Features for Moving Up in the Income Distribution



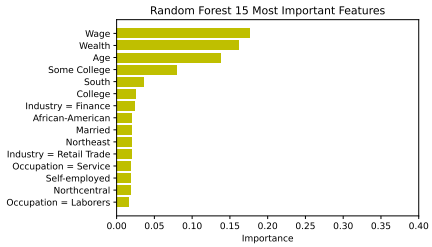
(a) Decrease of 10 percentile or more.
Accuracy: 0.63
Precision: 0.53



(b) Decrease of 20 percentile or more.
Accuracy: 0.61
Precision: 0.54



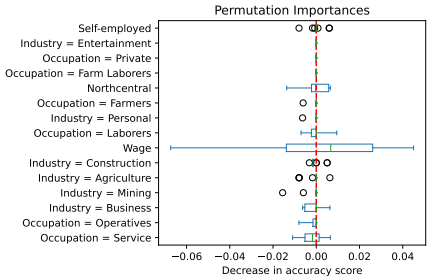
(c) Decrease of 30 percentile or more.
Accuracy: 0.79
Precision: 0



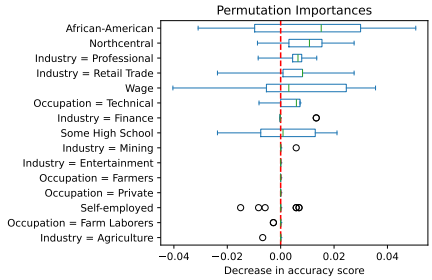
(d) Decrease of 40 percentile or more.
Accuracy: 0.83
Precision: 0

The figures above show the importance plots of random forest models for predicting whether a child will be 10, 20, 30, or 40 percentile points lower than their parent. Feature importance is measured by the mean decrease in Gini impurity, which indicates how much a feature helps to make the model's predictions more accurate by creating purer, more distinct groups in the data after each split.

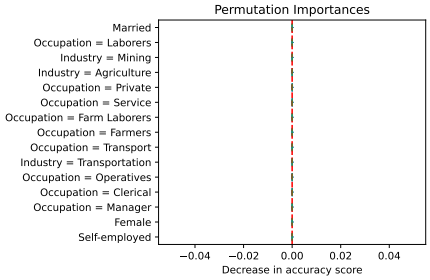
Figure 5: Random Forest, Most Important Features for Moving Down in the Income Distribution



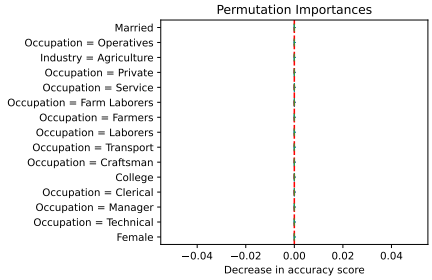
(a) Decrease of 10 percentile or more.



(b) Decrease of 20 percentile or more.



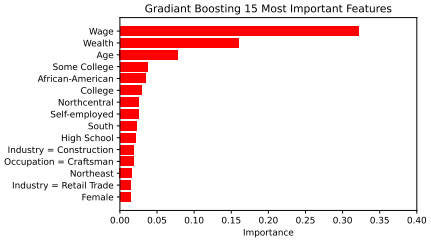
(c) Decrease of 30 percentile or more.



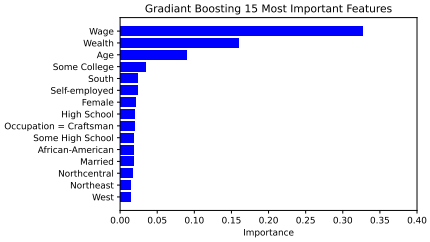
(d) Decrease of 40 percentile or more.

The figures above show the box plots of feature importance from 20 permutations of random forest models for predicting whether a child will be 10, 20, 30, or 40 percentile points lower than their parent. In each permutation, one feature is randomly shuffled, and the resulting decrease in model accuracy is measured to assess its importance.

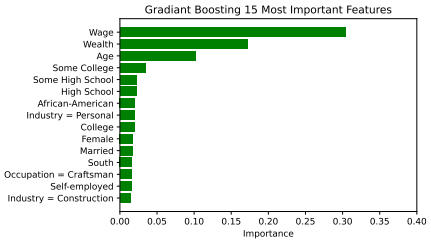
Figure 6: Random Forest, Permutations of the Most Important Features for Moving Down in the Income Distribution



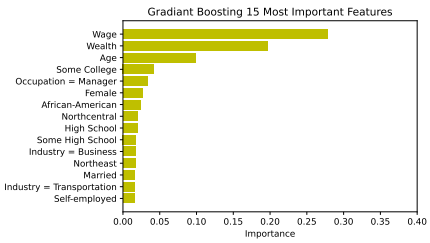
(a) Increase of 10 percentile or more.
Accuracy: 0.63
Precision: 0.39



(b) Increase of 20 percentile or more.
Accuracy: 0.76
Precision: 0.6



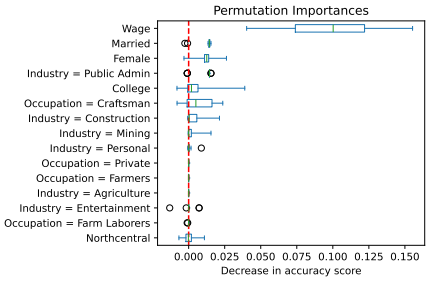
(c) Increase of 30 percentile or more.
Accuracy: 0.84
Precision: 0.6



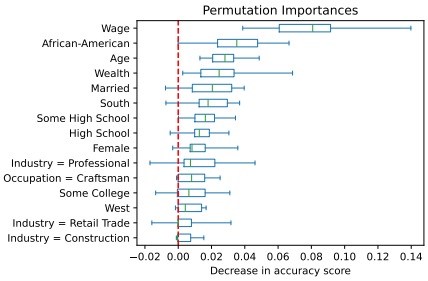
(d) Increase of 40 percentile or more.
Accuracy: 0.85
Precision: 0.5

The figures above show the importance plots of gradient boosting models for predicting whether a child will be 10, 20, 30, or 40 percentile points higher than their parent. Feature importance is measured by the mean decrease in Gini impurity, which indicates how much a feature helps to make the model's predictions more accurate by creating purer, more distinct groups in the data after each split.

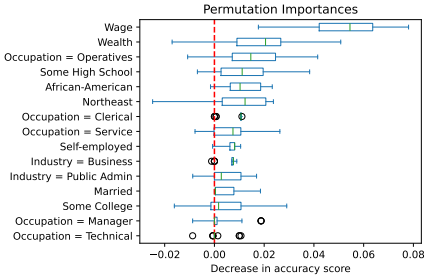
Figure 7: Gradient Boosting, Most Important Features for Moving Up in the Income Distribution



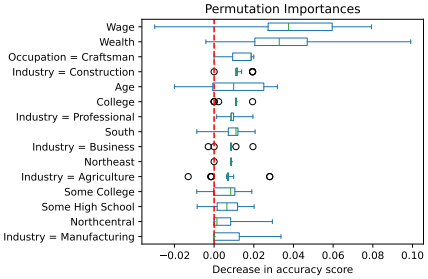
(a) Increase of 10 percentile or more.



(b) Increase of 20 percentile or more.



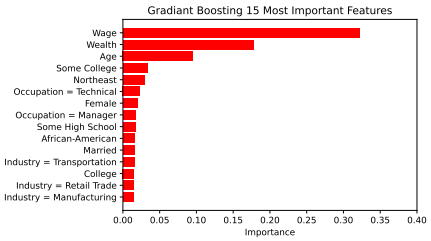
(c) Increase of 30 percentile or more.



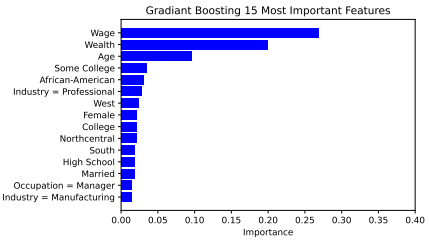
(d) Increase of 40 percentile or more.

The figures above show the box plots of feature importance from 20 permutations of gradient boosting models for predicting whether a child will be 10, 20, 30, or 40 percentile points higher than their parent. In each permutation, one feature is randomly shuffled, and the resulting decrease in model accuracy is measured to assess its importance.

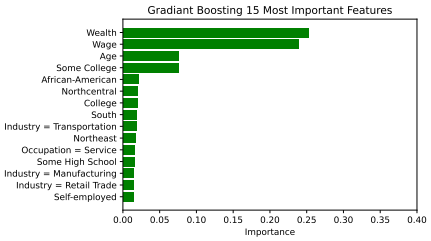
Figure 8: Gradient Boosting, Permutations of the Most Important Features for Moving Up in the Income Distribution



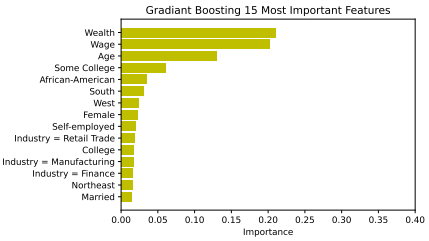
(a) Decrease of 10 percentile or more.
Accuracy: 0.64
Precision: 0.54



(b) Decrease of 20 percentile or more.
Accuracy: 0.64
Precision: 0.56



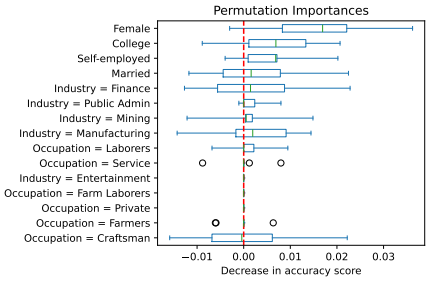
(c) Decrease of 30 percentile or more.
Accuracy: 0.76
Precision: 0.38



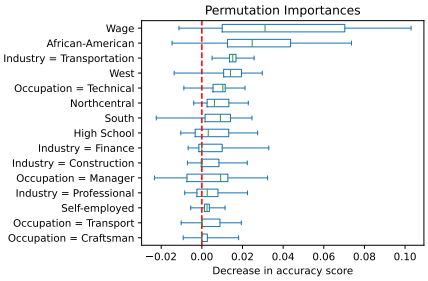
(d) Decrease of 40 percentile or more.
Accuracy: 0.81
Precision: 0.33

The figures above show the importance plots of gradient boosting models for predicting whether a child will be 10, 20, 30, or 40 percentile points lower than their parent. Feature importance is measured by the mean decrease in Gini impurity, which indicates how much a feature helps to make the model's predictions more accurate by creating purer, more distinct groups in the data after each split.

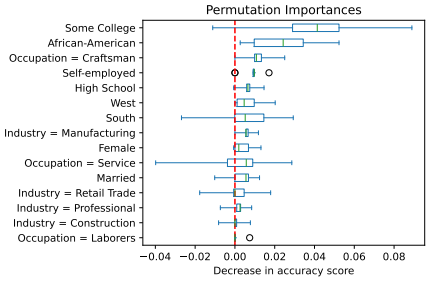
Figure 9: Gradient Boosting, Most Important Features for Moving Down in the Income Distribution



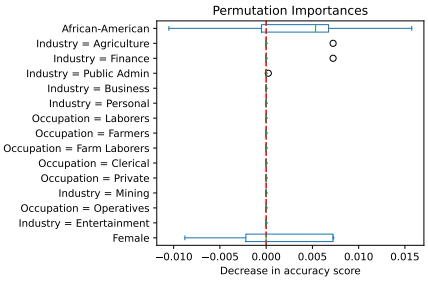
(a) Decrease of 10 percentile or more.



(b) Decrease of 20 percentile or more.



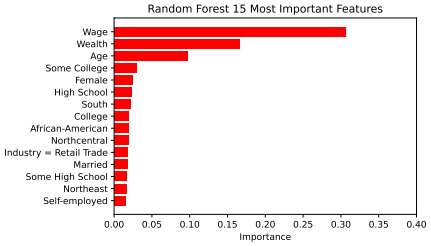
(c) Decrease of 30 percentile or more.



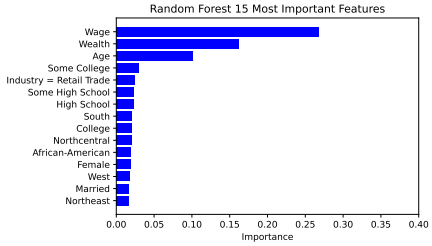
(d) Decrease of 40 percentile or more.

The figures above show the box plots of feature importance from 20 permutations of gradient boosting models for predicting whether a child will be 10, 20, 30, or 40 percentile points lower than their parent. In each permutation, one feature is randomly shuffled, and the resulting decrease in model accuracy is measured to assess its importance.

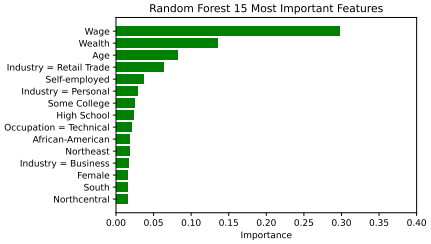
Figure 10: Gradient Boosting, Permutations of the Most Important Features for Moving Down in the Income Distribution



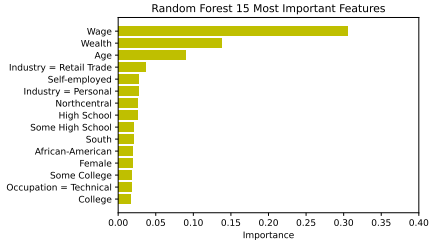
(a) Increase of $5,000 more than parent.
Accuracy: 0.68
Precision: 0.63



(b) Increase of $10,000 more than parent
Accuracy: 0.72
Precision: 0.6



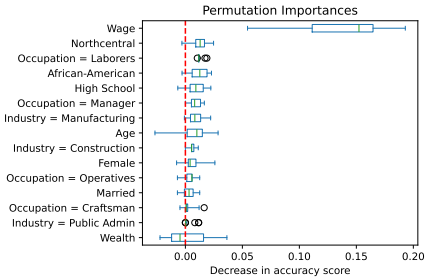
(c) Increase of $15,000 more than parent.
Accuracy: 0.74
Precision: 0.57



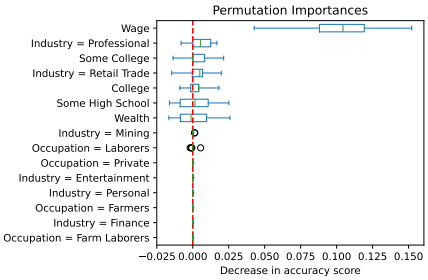
(d) Increase of $20,000 more than parent.
Accuracy: 0.78
Precision: 0.33

The figures above show the importance plots of random forest models for predicting whether a child will earn $5,000, $10,000, $15,000, or $20,000 more than their parent. Feature importance is measured by the mean decrease in Gini impurity, which indicates how much a feature helps to make the model's predictions more accurate by creating purer, more distinct groups in the data after each split.

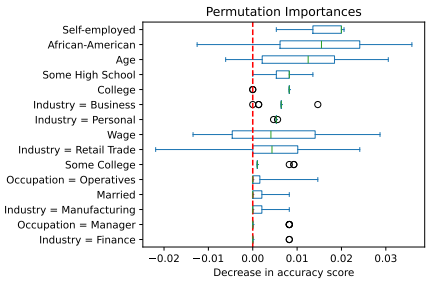
Figure 11: Random Forest, Most Important Features for Having a Higher Income than Parent



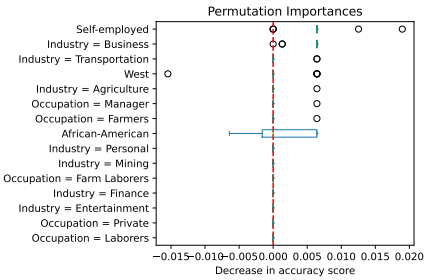
(a) Increase of $5,000 more than parent.



(b) Increase of $10,000 more than parent



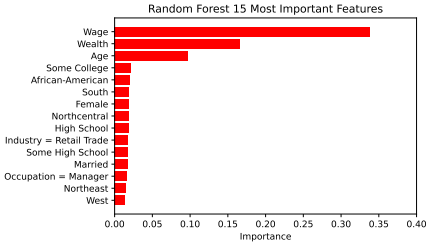
(c) Increase of $15,000 more than parent.



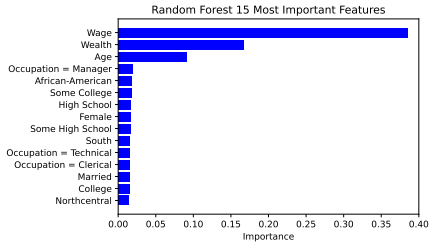
(d) Increase of $20,000 more than parent.

The figures above show the box plots of feature importance from 20 permutations of random forest models for predicting whether a child will earn $5,000, $10,000, $15,000, or $20,000 more than their parent. In each permutation, one feature is randomly shuffled, and the resulting decrease in model accuracy is measured to assess its importance.

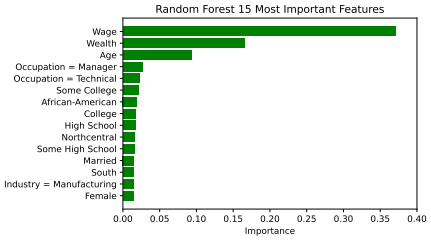
Figure 12: Random Forest, Permutations of the Most Important Features for Having a Higher Income than Parent



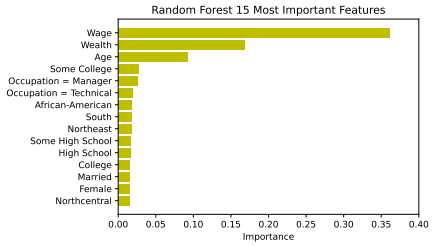
(a) Decrease of $5,000 less than parent.
Accuracy: 0.67
Precision: 0.65



(b) Decrease of $10,000 less than parent
Accuracy: 0.77
Precision: 0.74



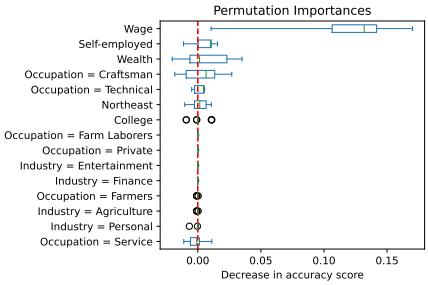
(c) Decrease of $15,000 less than parent.
Accuracy: 0.72
Precision: 0.6



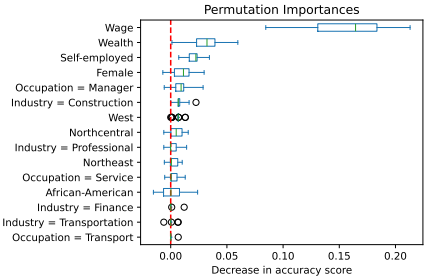
(d) Decrease of $20,000 less than parent.
Accuracy: 0.77
Precision: 0.68

The figures above show the importance plots of random forest models for predicting whether a child will earn $5,000, $10,000, $15,000, or $20,000 less than their parent. Feature importance is measured by the mean decrease in Gini impurity, which indicates how much a feature helps to make the model's predictions more accurate by creating purer, more distinct groups in the data after each split.

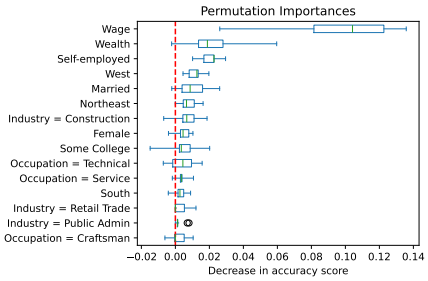
Figure 13: Random Forest, Most Important Features for Having a Lower Income than Parent



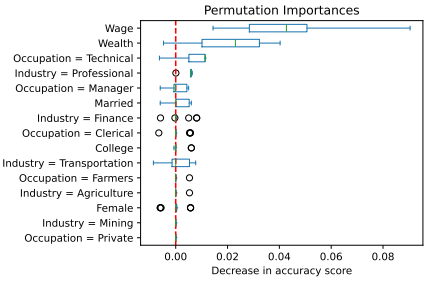
(a) Decrease of $5,000 less than parent.



(b) Decrease of $10,000 less than parent



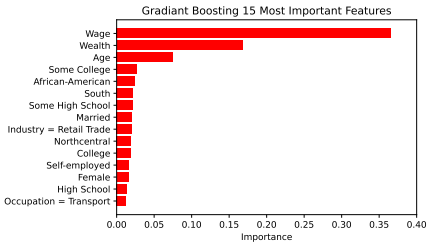
(c) Decrease of $15,000 less than parent.



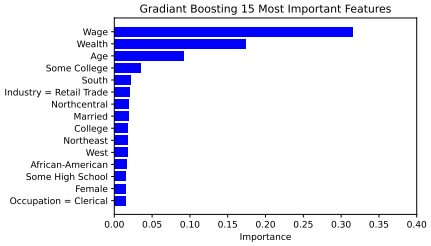
(d) Decrease of $20,000 less than parent.

The figures above show the box plots of feature importance from 20 permutations of random forest models for predicting whether a child will earn $5,000, $10,000, $15,000, or $20,000 less than their parent. In each permutation, one feature is randomly shuffled, and the resulting decrease in model accuracy is measured to assess its importance.

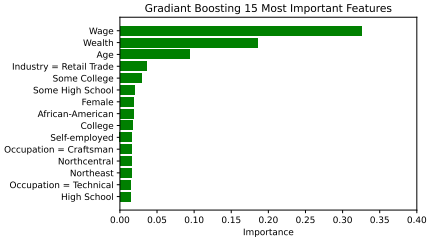
Figure 14: Random Forest, Permutations of the Most Important Features for Having a Lower Income than Parent



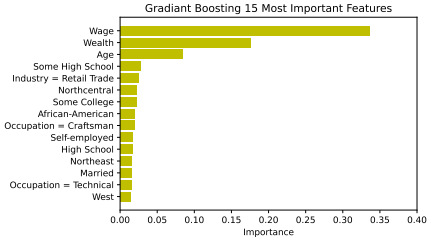
(a) Increase of $5,000 more than parent.
Accuracy: 0.68
Precision: 0.61



(b) Increase of $10,000 more than parent
Accuracy: 0.67
Precision: 0.48



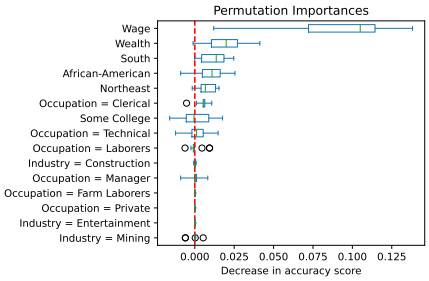
(c) Increase of $15,000 more than parent.
Accuracy: 0.71
Precision: 0.42



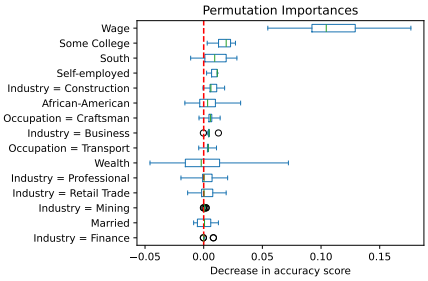
(d) Increase of $20,000 more than parent.
Accuracy: 0.78
Precision: 0.5

The figures above show the importance plots of gradient boosting models for predicting whether a child will earn $5,000, $10,000, $15,000, or $20,000 more than their parent. Feature importance is measured by the mean decrease in Gini impurity, which indicates how much a feature helps to make the model's predictions more accurate by creating purer, more distinct groups in the data after each split.

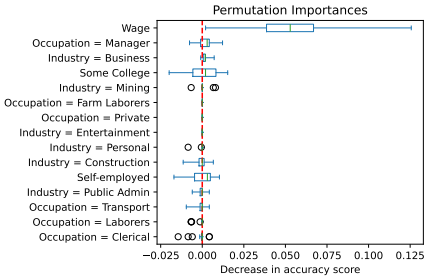
Figure 15: Gradient Boosting, Most Important Features for Having a Higher Income than Parent



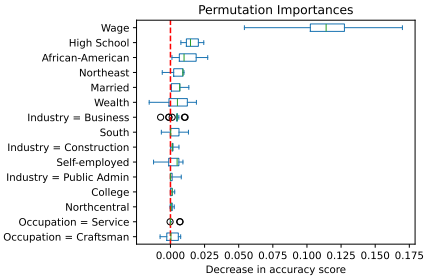
(a) Increase of $5,000 more than parent.



(b) Increase of $10,000 more than parent



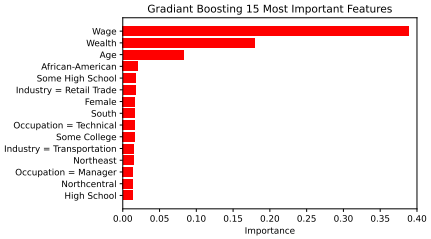
(c) Increase of $15,000 more than parent.



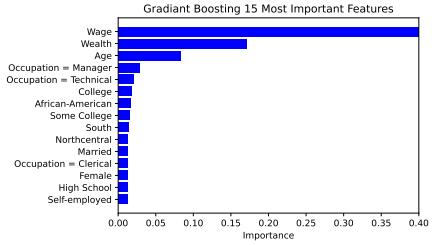
(d) Increase of $20,000 more than parent.

The figures above show the box plots of feature importance from 20 permutations of gradient boosting models for predicting whether a child will earn $5,000, $10,000, $15,000, or $20,000 more than their parent. In each permutation, one feature is randomly shuffled, and the resulting decrease in model accuracy is measured to assess its importance.

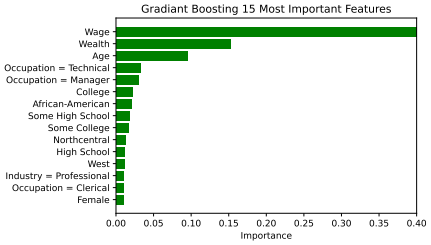
Figure 16: Gradient Boosting, Permutations of the Most Important Features for Having a Higher Income than Parent



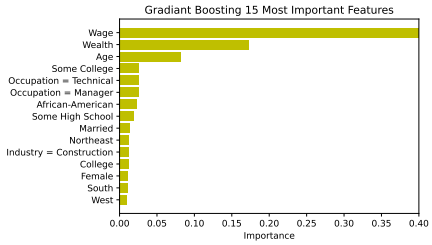
(a) Decrease of $5,000 less than parent.
Accuracy: 0.71
Precision: 0.69



(b) Decrease of $10,000 less than parent
Accuracy: 0.74
Precision: 0.67



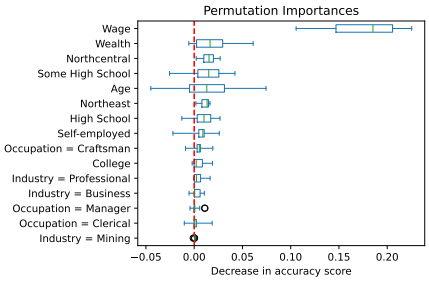
(c) Decrease of $15,000 less than parent.
Accuracy: 0.74
Precision: 0.62



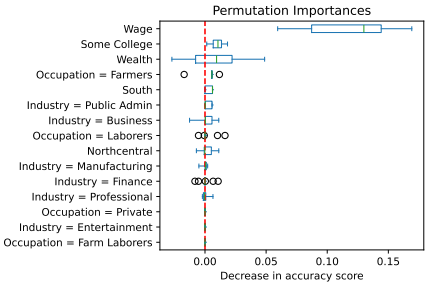
(d) Decrease of $20,000 less than parent.
Accuracy: 0.77
Precision: 0.62

The figures above show the importance plots of gradient boosting models for predicting whether a child will earn $5,000, $10,000, $15,000, or $20,000 less than their parent. Feature importance is measured by the mean decrease in Gini impurity, which indicates how much a feature helps to make the model's predictions more accurate by creating purer, more distinct groups in the data after each split.

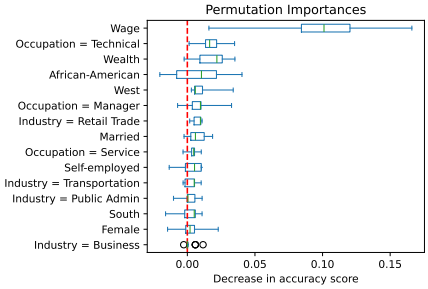
Figure 17: Gradient Boosting, Most Important Features for Having a Lower Income than Parent



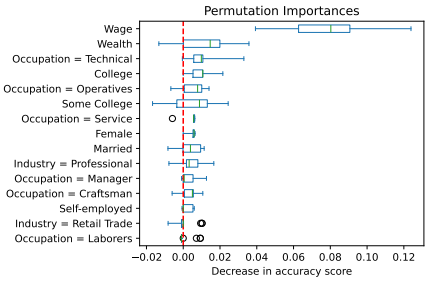
(a) Decrease of $5,000 less than parent.



(b) Decrease of $10,000 less than parent



(c) Decrease of $15,000 less than parent.



(d) Decrease of $20,000 less than parent.

The figures above show the box plots of feature importance from 20 permutations of gradient boosting models for predicting whether a child will earn $5,000, $10,000, $15,000, or $20,000 less than their parent. In each permutation, one feature is randomly shuffled, and the resulting decrease in model accuracy is measured to assess its importance.

Figure 18: Gradient Boosting, Permutations of the Most Important Features for Having a Lower Income than Parent

Appendix on Algorithms

A.1 Random Forest

A.1.1 Random Forest Algorithm

The algorithm using Random Forest for Classification is displayed below.

Algorithm 1 Random Forest for Classification

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample, $\mathbf{X}^*$ and Y^*, of size N from the training data.
   (b) Make a random forest decision tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      (i) Select m variables at random from the w variables.
      (ii) Pick the best variable/split-point among the m.
      (iii) Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$. Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{rf}^B(x) = majority\ vote\{\hat{C}_b(x)\}_1^B$ (i.e., the mode of the tree predictions).

For step (ii): to pick the best split point $c \in (-\infty, \infty)$, we can minimize some "impurity" measure (i.e., how poor these regions are at classifying the outcome variable).[25] Common node impurity measures include:

- Gini Index (no relation to Gini Inequality measure)
- Misclassification Error
- Cross-entropy

Consider the commonly used Gini index. In a node m, representing a region R_m with N_m observations, let

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} \mathbb{1}(y_i = k),$$

[25] For step (ii) when Random Forest is used for regression: for covariate h and threshold c the sum of in-sample squared errors is

$$Q(c,h) = \sum_{i:X_{ih} \leq c} (Y_i - \bar{Y}_{h,c,l})^2 + \sum_{i:X_{ih} > c} (Y_i - \bar{Y}_{h,c,r})^2$$

where l and r denote left and right and

$$\bar{Y}_{h,c,l} = \sum_{i:X_{ih} \leq c} Y_i / \sum_{i:X_{ih} \leq c} 1 \text{ and } \bar{Y}_{h,c,r} = \sum_{i:X_{ih} > c} Y_i / \sum_{i:X_{ih} > c} 1.$$

To split the sample, minimize $Q(c,h)$ over all covariates $h = 1, ...m$ and all thresholds $c \in (-\infty, \infty)$. Alternatively, an entropy-based criterion for sample splitting can be used. See Hastie et al. (2001) for details.

be the proportion of class k observations in node m. In this paper there are only two classes, a binary classification case. The *target* (what is being predicted) is a classification outcome that takes values $0, 1, 2, ..., K$. Then, the Gini index= $\sum_{k=0}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$, a variance/dispersion measure to be minimized.

The Gini index measures the probability of incorrectly labeling the observation based on the distribution of labels in the region (i.e., how often a randomly chosen element from the sample would be mislabeled if it were randomly assigned values of the m variables in the region). In that way, it measure how "impure" and non-homogeneous the observations are.

Notice that $\hat{p}_{mk} = M_{society\_increase}(p)$ if $y_i = M_{increase}(p)_i$. Therefore, the algorithm is explicitly constructing trees to minimum incorrectly labeling mobility.

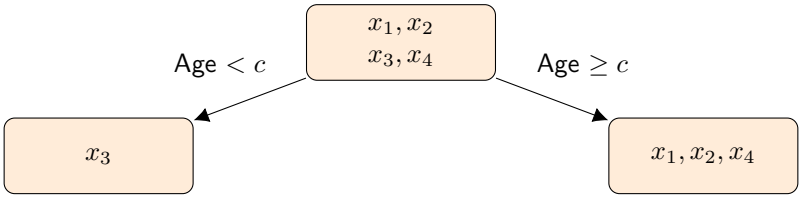
The weighted Gini index of a split is $p_L \text{Gini}_L + p_R \text{Gini}_R$, where p_L and Gini_L is the proportion of observations split to the left and the Gini index of the left node; p_R and Gini_R is the proportion of observations split to the right and the Gini index of the right node.

A.2.1 Random Forest Example

The following example demonstrates step (ii) of Algorithm 1 using the Gini index as a measure of node impurity. Consider the following data:

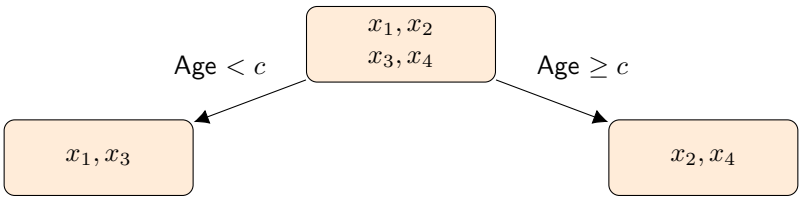
Observations	Parent Age	Parent Income	$M_{increase}(20)$
x_1	35	100,000	0
x_2	42	150,000	1
x_3	28	90,000	0
x_4	39	110,000	1

If Parent Age is selected as the first covariate to partition the data, the first split in a tree could be $c = 35$:



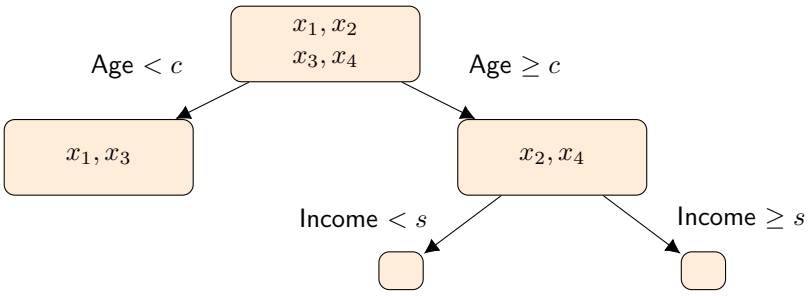
Then $\text{Gini}_L = 2 \cdot 0 \cdot 1 = 0$, $\text{Gini}_R = 2 \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{9}$, and Weighted Gini of Split $= \frac{1}{4} \cdot 0 + \frac{3}{4} \cdot \frac{4}{9} = \frac{1}{3}$.

If instead $c = 39$, then the split would be:



Then $\text{Gini}_L = 2 \cdot 0 \cdot 1 = 0$, $\text{Gini}_R = 2 \cdot 1 \cdot 0 = 0$, and Weighted Gini of Split $= \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0 = 0$.

Therefore, $c = 39$ is preferred because it has a lower weighted Gini index of the split. Another feature is considered and chosen instead of Parent Age if it has a lower weighted Gini index (the number of features considered is a hyperparameter). After the split, the process can be repeated. For example:



A.2 Pruning:

A tree can *over-fit* the data, meaning, predicting very well on the training sample but with poor out-of sample prediction. Therefore, trees are *pruned* (the depth of the tree is determined through cross validation).

The hyper-parameters of the random forest as defined by the "sklearn" package in Python are (i) `criterion`: splitting criterion (entropy or gini), (ii) `max_depth`: (n_{min}) is maximum depth of the tree, (iii) `max_features`: (m) is maximum number of features random forest considers to split a node (either the square root of the total number of features or the logarithm base 2 of the total number of features.), and (iv) `n_estimators`: (B) is the number of bootstraps.

The hyper-parameters are selected by k-fold cross validation.

A.3 Gradient Boosting

A.3.1 Gradient Boosting Algorithm

The algorithm for Gradient Boosting is displayed below:

Algorithm 2 Gradient Boosting for Classification

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^N L(y_i, \gamma)$.
2. For $m = 1$ to M:
   (a) For $i = 1, 2, \ldots, N$ compute:
   "pseudo residual" $r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$
   (b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, $j = 1, 2, \ldots, J_m$.
   (c) For $j = 1, 2, \ldots, J_m$ compute
   $\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$.
   (d) Update $f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} \mathbb{1}(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_M(x)$.

A.3.2 Example of Gradient Boosting

Consider the following data:

Observations	Parent Black	Parent Age	Parent Edu.	$M_{increase}(20)$
x_1	Yes	18	HS	1
x_2	No	87	College	1
x_3	No	44	HS	0

Step 1:

Initialize model with a constant value: $f_0(x) = \arg\min_\gamma \sum_{i=1}^n L(y_i, \gamma)$.
With a logistic loss function, the likelihood is

$$\begin{aligned}&\sum_{i=1}^N y_i \log(p) + (1 - y_i) \log(1 - p)\\ &= \sum_{i=1}^N -y_i\gamma + \log(1 + e^\gamma),\end{aligned}$$

where $p = \frac{e^\gamma}{1+e^\gamma}$ (i.e. $\gamma = \log\left(\frac{p}{1-p}\right)$).

So,

$$\frac{\partial \sum_{i=1}^N L(y_i, \gamma)}{\partial \gamma} = \sum_{i=1}^N -y_i + \underbrace{\frac{e^\gamma}{1 + e^\gamma}}_{p}.$$

With the example data, the *First Order Condition* is

$$(-1 + p_0) + (-1 + p_0) + (0 + p_0) = 0 \implies p_0 = \frac{2}{3}.$$

Therefore,

$$\gamma_0 = \log\left(\frac{\frac{2}{3}}{1 - \frac{2}{3}}\right) \approx 0.69.$$

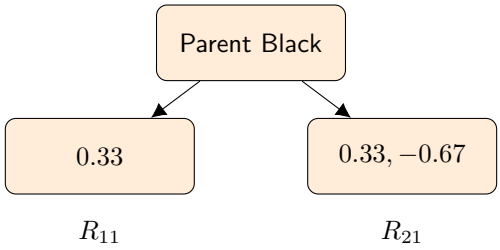
Step 2: for $m = 1$ to M:

(a) Compute $r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$ for $i = 1, 2, ..., n$

Hence

P. Black	P. Age	P. Edu.	$M_{increase}(20)$	pseudo residual
Yes	18	HS	1	1-0.67 = 0.33
No	87	College	1	1-0.67 = 0.33
No	44	HS	0	0-0.67 = -0.67

(b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, $j = 1, 2, \ldots, J_m$.



(c) For $j = 1, 2, \ldots, J_m$ compute

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

Using a 2nd order Taylor series approximation:

$$L(y_i, f_{m-1}(x_i) + \gamma) \approx$$
$$L(y_i, f_{m-1}(x_i)) + \frac{d}{df()} L(y_i, f_{m-1}(x_i))\gamma + \frac{1}{2}\frac{d^2}{df()^2} L(y_i, f_{m-1}(x_i))\gamma^2.$$

Therefore,

$$\frac{d}{d\gamma} L(y_i, f_{m-1}(x_i) + \gamma) \approx \frac{d}{df()} L(y_i, f_{m-1}(x_i)) + \frac{d^2}{df()^2} L(y_i, f_{m-1}(x_i))\gamma.$$

Setting $\frac{d}{d\gamma} L(y_i, f_{m-1}(x_i) + \gamma) = 0$, we have

$$\gamma = \frac{\text{residual}}{p(1-p)}.$$

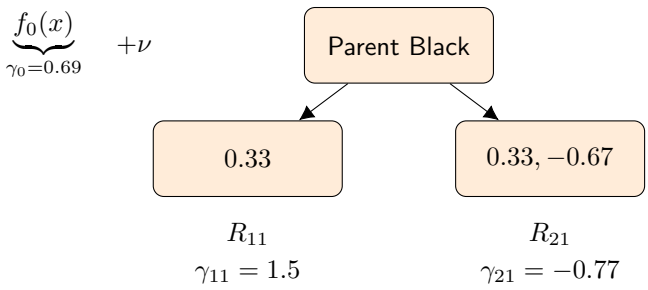
So,

$$\gamma_{11} = \frac{r_{11}}{\frac{e^{f_0(x)}}{1+e^{f_0(x)}}\left(1 - \frac{e^{f_0(x)}}{1+e^{f_0(x)}}\right)} = \frac{0.33}{0.67(1-0.67)} = 1.5$$

and

$$\gamma_{21} = \frac{r_{11} + r_{21}}{2p(1-p)} = \frac{0.33 + (-0.67)}{2(0.67)(1-0.67)} = -0.77.$$

(d) Update $f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} \mathbb{1}(x \in R_{jm})$.



Then the new predictions for the observations are:

x_1	Yes	18	HS	1

$$\implies f_1(x_1) = \underbrace{f_0(x_1)}_{0.69} + \underbrace{0.1}_{\nu\text{;learning rate}} \underbrace{1.5}_{\gamma_{11}} = \underbrace{0.85}_{\text{log odds ratio}}$$

x_2	No	87	College	1

$$\implies f_1(x_2) = \underbrace{f_0(x_2)}_{0.69} + \underbrace{0.1}_{\nu\text{;learning rate}} \underbrace{-0.77}_{\gamma_{21}} = \underbrace{0.613}_{\text{log odds ratio}}$$

Step 3:

Output $\hat{f}(x) = f_M(x)$.

A.4 Variable Importance and Permutation Importance

A.4.1 Feature Importance

Feature importance is calculated as the average decrease in the Gini impurity (how well the feature splits the data) across all trees. It is normalized so the sum across all features equals 1, making it easier to compare.

A.4.2 Permutation Importance

If some of the features are highly correlated or the algorithm over-fits the data, feature importance can be misleading. These concerns can easily be remedied by "permutating" the data. It works by training the algorithm on the original data and measuring its accuracy. Then it:

1. Reshuffles the values of one of the features while keeping the rest unchanged.

2. Uses the modified data set to predict the target variable and calculates the accuracy again. The drop in accuracy indicates the importance of the shuffled feature.

3. Repeat this process for all features.

Features with larger drops in model accuracy when shuffled are considered more important, as they have a greater impact on the model's prediction. This can be done several times, and boxplots of the results can be displayed.