Project Proposal for Introduction to Machine Learning

The Heterogeneity of Inequality of Opportunity in the US and their Roots

John Kim (ID: johndkim), Takahiro Minami (ID: tminami)
Xin Feng (ID: xinf), Yujiao Song (ID: songyujiao)

I. Research Motivation

Income inequality in the United States had reached its peak level in 50 years, and it is still growing (Ingraham, 2019). The Gini coefficient of the U.S. is 0.390 in 2015, which ranks 3rd among the OECD countries (OECD, 2020). Setting aside the scathing critique from the humanitarian and social justice sphere on severe inequality (Sen, 2000), there is a line of research that corroborates the negative impact of income inequality on economic growth (Herzer & Vollmer, 2012; Cingano, 2014). This negative relationship is conspicuous when income inequality is entailed by inequality of opportunity instead of individuals' efforts (Marrero & Rodríguez, 2013).

In light of this view, we adopted Roemer's (1998) conception of inequality of opportunity to differentiate the causes of income inequality by social circumstances and individuals' effort. We plan to assess the social circumstances that contribute to the current income inequality in the U.S. by utilizing the conditional inference tree method. With the structure of income inequality analyzed by this research, we then access the heterogeneity of inequality of opportunity across the U.S. by estimating the state-level Gini coefficient measuring income inequality associated with social circumstances.

Our analysis begins by illustrating the difference between the traditional non-machine learning methods and the machine learning method, justifying the use of the latter approach for this research. Then, we elucidate the conditional inference tree method and the data used for this project. Finally, we present our results of the structure of income inequality and estimate the opportunity-based Gini coefficient for each state in the U.S.

II. Theoretical Background

1. Inequality of Opportunity

John Roemer attributes an individual's outcome to two types of factors: effort (factors over which individuals have control) and circumstances (environmental factors beyond one's control, such as biological characteristics). He proposes that a study of inequality of opportunity should be a study of the circumstances variables, which is hard for individuals to compensate with effort.

Over the past decade, scholars debated the appropriate empirical method to estimate inequality of opportunity from different perspectives (Romos & Van de gaer, 2012). Among them, this research will take the ex-ante view, which regards inequality of opportunity as a divergence of average income between social groups (called classes) with different social circumstances. Suppose a social circumstance $C \in \{a, b\}$ (e.g. C is race and a, b are white or black) divides the society into two classes G_a and G_b. If there is no inequality of opportunity, the average income of G_a and G_b should be equal. In other words, we assume that the difference of average income between G_a and G_b is associated with the social circumstance C. Using this idea, we will divide the whole population into many classes based on various criteria of social circumstances. For each class, we will take the average of income, and compare with different classes. The difference between classes can be interpreted as inequality associated with the difference of social circumstances, which is nothing more than what we want to measure.

2. Estimation of inequality of opportunity and Advantages of Machine Learning

Although the concept itself is simple, classifying the population based on many social circumstances is not easy under the limited size of data. The more classifier variables we have, the fewer samples each class eventually has, which will decrease the precision of our estimation. Furthermore, the impacts of social circumstances on income level are non-linear and not independent of each other. For example, the research on India by Lefranc and Kundu (2019) showed that fathers' occupation is associated significantly with income heterogeneity among people in some castes, but not in others.

To deal with these problems, researchers have developed many models using various sets of circumstance variables in estimating the inequality of opportunity. However, no matter how cautious they are, they cannot avoid the arbitrariness in model selection and various sources of bias in estimation (Donni et al., 2015).

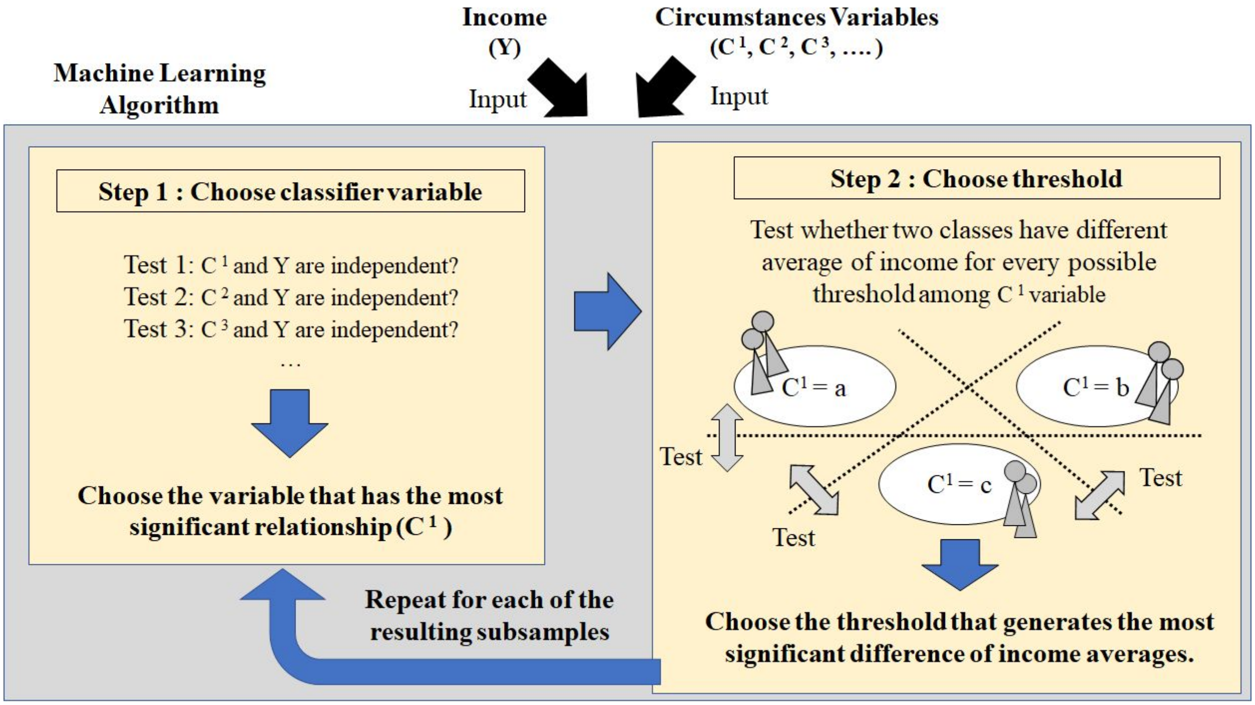
This research tries to overcome the critics of arbitrary model selection by adapting the conditional inference tree algorithm, which was suggested by Hothorn et al. (2006) and applied to estimate inequality of opportunity in EU countries by Brunori et al. (2018). This regression tree method has several advantages over traditional estimations. First, there is no assumption that the underlying relationship between the predictors and the dependent variable is linear or monotonic. This is suited for our research context. Second, the conditional inference tree will perform a sequence of hypothesis tests to prevent model overfitting, thus reduce the upward bias on the inequality of opportunity estimates. Third, the conditional inference tree is visually appealing and it can provide a more intuitive understanding of the structure of inequality of opportunities (Brunori et al., 2018).

III. Methodology

1. Conditional Inference Trees

The conditional inference tree algorithm conducts statistical tests to check whether each input variable is independent of income and chooses the one which has the strongest relationship with income (Step 1 in Figure 1). Then, the machine divides the population into two classes by all possible thresholds among the variable and tests whether the binary classes have significantly different income averages. Based on the p-values of these tests, the machine chooses the threshold which causes the biggest average discrepancy (Step 2 in Figure 1). This whole process is repeated for each of the resulting subgroups until all remaining input variables are independent of income in the subgroup. Through this process, the algorism builds the tree to structurize the factors associated with the income of each of the subgroups.

Figure 1: Mechanism of Conditional Inference Trees Algorithm


Income
(Y)
Circumstances Variables
(C¹, C², C³, ….)
Machine Learning
Algorithm
Input
Input
Step 1 : Choose classifier variable
Test 1: C¹ and Y are independent?
Test 2: C² and Y are independent?
Test 3: C³ and Y are independent?
…
Choose the variable that has the most significant relationship (C¹)
Step 2 : Choose threshold
Test whether two classes have different average of income for every possible threshold among C¹ variable
C¹ = a
C¹ = b
C¹ = c
Test
Test
Test
Choose the threshold that generates the most significant difference of income averages.
Repeat for each of the resulting subsamples

2. Cross-validation

The tree algorithm requires us to decide when the machine should stop repeating the process. To be more specific, we need to set the threshold number α, and the machine stops the process once the machine tests all remaining input values in Step 1 in Figure 1 and p-values generated in all these tests become larger than α. A lower α increases the possibility to miss some eligible classifiers while a higher α increases the possibility to misidentify meaningless variable as classifiers.

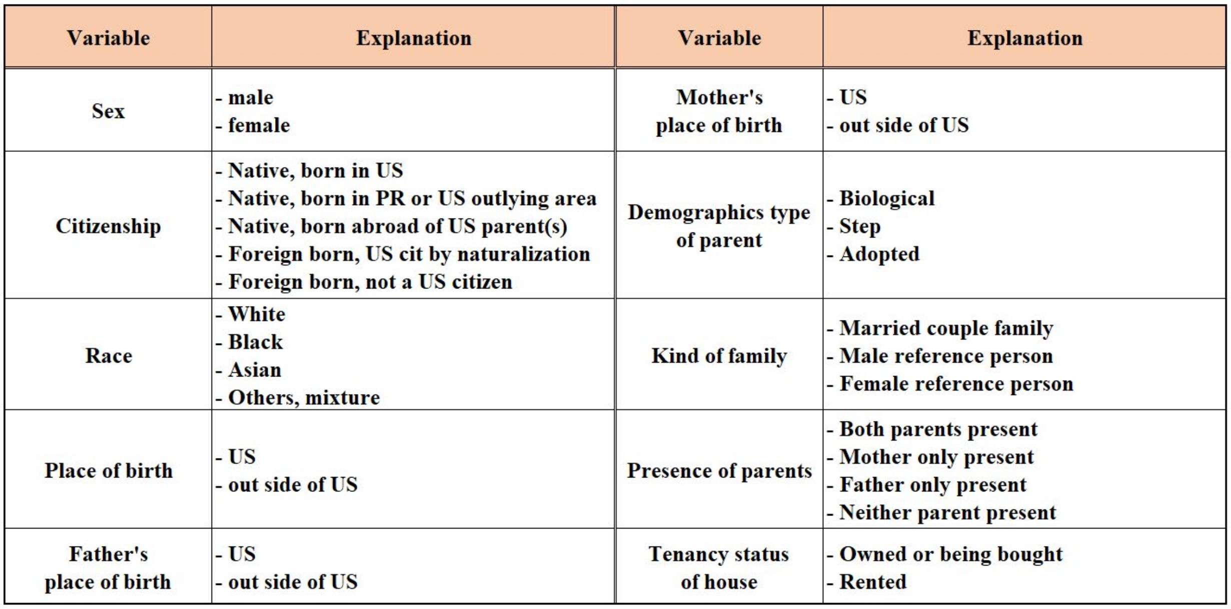
To deal with this trade-off between type 1 and type 2 errors, we will conduct cross-validation to choose optimal α. We will divide the original sample into 10, then calculate MSE using each subgroup based on the estimation from the other 9 subsamples. We will repeat this process for $\alpha = 0.01, 0.05,$ and 0.1, and choose the α which associates with the smallest MSE. Through the cross-validation, we can minimize the arbitrary choice of the model.

IV. Data

We use microdata from the Current Population Survey 2019 Annual Social and Economic (ASEC). ASEC is the nation-wide survey and contains many demographic and socio-economic data for subjective individuals and households. The anonymized microdata is provided to the public by U.S. Census Bureau, and the number of data records is 354,345, which is big enough to estimate on U.S. state basis. We limit our sample to individuals over 18.

Our outcome variable in the estimation is total household income. To avoid outliers from impacting our estimation heavily, we follow the method adopted by Brunori et al. (2018) and substitute all income larger than the 99.5th percentile of the state's income distribution with the 99.5th threshold. The input variables for social circumstances are listed on Table 1.

Table 1: List of Social Circumstance Variables

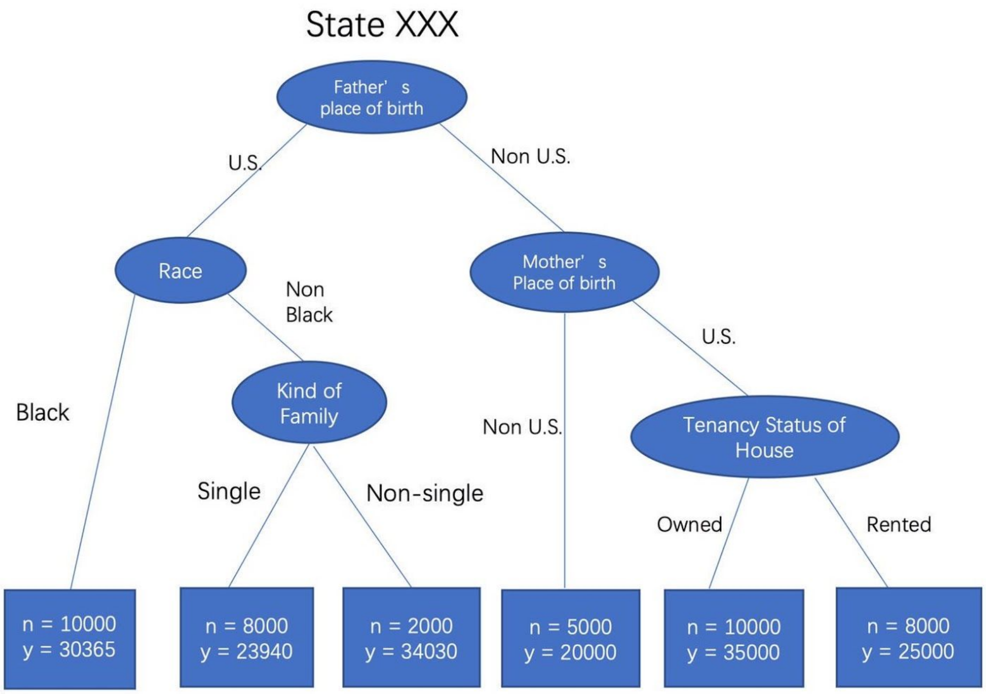


Variable	Explanation	Variable	Explanation
Sex	- male - female	Mother's place of birth	- US - out side of US
Citizenship	- Native, born in US - Native, born in PR or US outlying area - Native, born abroad of US parent(s) - Foreign born, US cit by naturalization - Foreign born, not a US citizen	Demographics type of parent	- Biological - Step - Adopted
Race	- White - Black - Asian - Others, mixture	Kind of family	- Married couple family - Male reference person - Female reference person
Place of birth	- US - out side of US	Presence of parents	- Both parents present - Mother only present - Father only present - Neither parent present
Father's place of birth	- US - out side of US	Tenancy status of house	- Owned or being bought - Rented

V. Expected Results

Our expected outcome will have 50 regression trees; one for each state. The tree is going to look like Figure 2 where all the root nodes represent different classes. The root nodes at the bottom represent classes within which no social circumstance variables don't generate income inequality (i.e. opportunity is equal). The income differences between these classes imply inequality of opportunity. The circles placed above the root nodes indicate splitting points, which are social circumstances playing a role in explaining the variation of income between the following subgroups. Looking at the structure of the splitting points, we can assess complex relationships between social circumstances and income. We are interested in whether we can find different structures across U.S. states.

Figure 2: Example of Expected Tree Representation



Note: y represents the average income of each root node. n is the number of people within each root node.
p-value in splitting points indicates the significance of each of these predictors.

In addition to visualizing the structure of inequality of opportunity, the conditional inference tree algorithm allows us to calculate opportunity-based Gini coefficient and measure the degree of inequality of opportunity numerically. First, we calculate the average income for each of root nodes, which is shown as y in Figure 2. Then, we replace income of individuals in each root node into the average of the root node. Using the replaced income level, we calculate the Gini coefficient. Since the divergence of income between root nodes are only associated with social circumstances, this Gini coefficient can capture the degree of inequality of opportunity.

VI. Additional Research

The research on section I to V is our main project. However, the research can show there are heterogeneous inequality of opportunity across U.S. states but doesn't give clear implication about why the heterogeneity happen. To dive into this question, we regress the opportunity-based Gini coefficient on potential drivers of inequality, including the degree of globalization, education system and taxation. We recognize that this simple regression is not enough to specify causal relationships, but we can still assess macroeconomic factors associated with state-level inequality of opportunity.

For globalization, scholars have long cautioned that it harms income distribution by forcing low-skilled workers into competition with cheap labor in emerging economies (Stiglitz, 2017; Bergh & Nilsson, 2010; Kentor, 2001). Literature that analyzed the impact of globalization on income inequality has assessed the level of globalization by using KOF Index of Globalization, Economic Freedom Index of the Fraser Institute (Bergh & Nilsson, 2010), foreign capital dependence, trade openness (Kentor, 2001), and financial globalization (Asteriou et al., 2014). On top of these indices, we can also take the industry structure of a state into account; a state with a higher proportion of manufacturing industry in the economy will be more heavily impacted by globalization.

Also, education is both a driver and a remedy of income inequality (Gregorio & Lee, 2002; Abdullah et al., 2015). For assessing the level of education for each state, we will consider education attainment, distribution of education (e.g., the percentage of people aged more than 18 years without a high school diploma), and government spending on public education.

Changes in taxation and transfers may also influence income inequality (Piketty & Saez, 2003). Progressive taxation on capital income and wage effectively alleviates income inequality, and more tax collected by the government can mean more expenditure on social welfare programs that promote redistribution. However, in the United States, the income tax rate for the wealthy has been decreasing to the point where the richest 400 families pay 1.2% points lower than the rate paid by the bottom half of American households (Saez & Zucman, 2019). Such a conservative tax system can be attributed to political corruption, which we could investigate further by looking into the cronyism of politicians and capitalists, initial real per capita GDP, the ratio of public employment to labor force, and such (Gupta et al, 2002).

Besides the factors mentioned above, there are some manifold potential drivers that contribute to income inequality, such as labor structure. Furthermore, the complex relationships between input variables make the analysis even more challenging, enhancing the need for a robust research technique like another machine learning method. However, even from our simple regression model, we anticipate discovering valuable insights about the macroeconomics factors associated with income inequality in the U.S.

References

Abdullah, A., Doucouliagos, H., & Manning, E. (2015). Does education reduce income inequality? A meta‐regression analysis. Journal of Economic Surveys, 29(2), 301-316.

Asteriou, D., Dimelis, S., & Moudatsou, A. (2014). Globalization and income inequality: A panel data econometric approach for the EU27 countries. Economic modelling, 36, 592-599.

Bergh, A., & Nilsson, T. (2010). Do liberalization and globalization increase income inequality?. European Journal of Political Economy, 26(4), 488-505.

Brunori, P., Hufe, P., & Mahler, D. G. (2018). The roots of inequality: estimating inequality of opportunity from regression trees. The World Bank.

Cingano, F. (2014). Trends in income inequality and its impact on economic growth. Paris: OECD.

Donni, P. L., Rodríguez, J. G., & Dias, P. R. (2015). Empirical definition of social types in the analysis of inequality of opportunity: a latent classes approach. Social Choice and Welfare, 44(3), 673-701.

Gregorio, J. D., & Lee, J. W. (2002). Education and income inequality: new evidence from cross‐country data. Review of Income and Wealth, 48(3), 395-416.

Gupta, S., Davoodi, H., & Alonso-Terme, R. (2002). Does corruption affect income inequality and poverty? Economics of Governance, 3(1), 23-45.

Herzer, D., & Vollmer, S. (2012). Inequality and growth: evidence from panel cointegration. The Journal of Economic Inequality, 10(4), 489-503.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3), 651-674.

Ingraham, C. (2019, October 9). For the first time in history, U.S. billionaires paid a lower tax rate than the working class last year. The Washington Post. Retrieved from: https://www.washingtonpost.com/business/2019/10/08/first-time-history-us-billionaires-paid-lower-tax-rate-than-working-class-last-year/

Kentor, J. (2001). The long term effects of globalization on income inequality, population growth, and economic development. Social Problems, 48(4), 435-455.

Lefranc, A., & Kundu, T. (2019). Inequality of Opportunity in Indian Society.

Marrero, G. A., & Rodríguez, J. G. (2013). Inequality of opportunity and growth. Journal of Development Economics, 104, 107-122.

OECD (2020), Income inequality (indicator). doi: 10.1787/459aa7f1-en (Accessed on 21 January 2020). https://data.oecd.org/inequality/income-inequality.htm.

Piketty, T., & Saez, E. (2003). Income inequality in the United States, 1913–1998. The Quarterly Journal of Economics, 118(1), 1-41.

Ramos, X., & Van de Gaer, D. (2012). Empirical approaches to inequality of opportunity: Principles, measures, and evidence.

Roemer, J. E. (1998). Equality of opportunity. Cambridge, MA: Harvard University Press.

Saez, E., & Zucman, G. (2019). The triumph of injustice: How the rich dodge taxes and how to make them pay. NY: WW Norton.

Sen, A. K. (2000). Social justice and distribution of income. In A.B. Atkinson & F. Bourguignon (Eds.), Handbooks in economics, 16, 59-86.

Stiglitz, J. E. (2017). The overselling of globalization. Business Economics, 52(3), 129-137.