Individual Final Report
Xinchun Chen

Background Information:

My portion of work is to run the decision tree algorism on the NYC poverty dataset, also include some preprocessing works such as feature selection, instance selection and balance the target class. For our decision tree, our target is the *poverty status,* which has *1* indicates in poverty and *2* indicates not in poverty. The features contain both categorical and numerical variables, such as *age, pre- income, gender, education attainment, etc.* In order to reduce the overfitting problem and also for better visualization, I set the maximum depth of the decision tree to three and also use Gini as the cost function ($1 - \sum {p_i}^2$).

Details:

Preprocessing:

The first step is data reduction, the technique we are using is feature selection and instance selection. Our dataset has total number of 79 columns and around 70000 observations, data reduction could help us to have a reduced representation of the original data.

Features selection:

The goal is to drop irrelevant or redundant features to reduce the dimensionality of the dataset. Some of the variables are: *Household Unit ID*, *Age Category, Poverty Gap, Tax Unit, etc.* We also decided to drop the variables that have high correlation with the total income, since income and poverty status shared the same characteristic (low income directly decides whether the person is in poverty or not). If we include total income, the decision tree would pick income as criteria in every nodes, which will underestimate the impact of other variables.

Instance selection:

The dataset records 70000 individuals and are grouped by the household unit. The figure shows the first 7 observations of the dataset, the observations that have same SERIALNO indicates the individuals are belong in the same household. The Poverty Status of each individual is decided by the head of family (*Povunit_Rel = 1*), if the head of family is in poverty then rest of the members are in in poverty too regardless of other character. In the case, we only kept the head of family as our unit of measurement.

SERIALNO	AGEP	Povunit_Rel	NYCgov_Pov_Stat
39	51	1	2
55	60	1	2
55	52	2	2
55	26	4	2
55	20	4	2
55	20	4	2
69	39	1	2

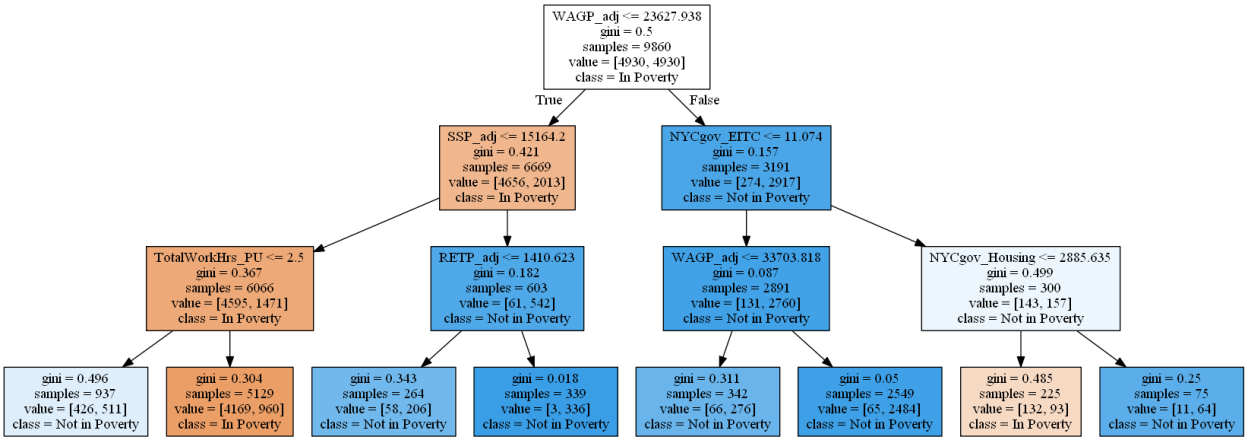
Balance the training set:

One of the potential issue of decision tree is imbalanced classes of the training set, if one class dominates the other class then the prediction would biased to the dominated class.

The original dataset is dominated by the non-poverty class, which is about 80% of the total observations. The way I am fixing the imbalanced problem is downsize the poverty class in the training set (EliteDataScience, 2017). I used *resample* function to randomly draw from the non-poverty class without replacement, and keep drawing until it matches the number of the poverty class. The new training set now would have equal distribution between poverty and non-poverty.

Build the tree:

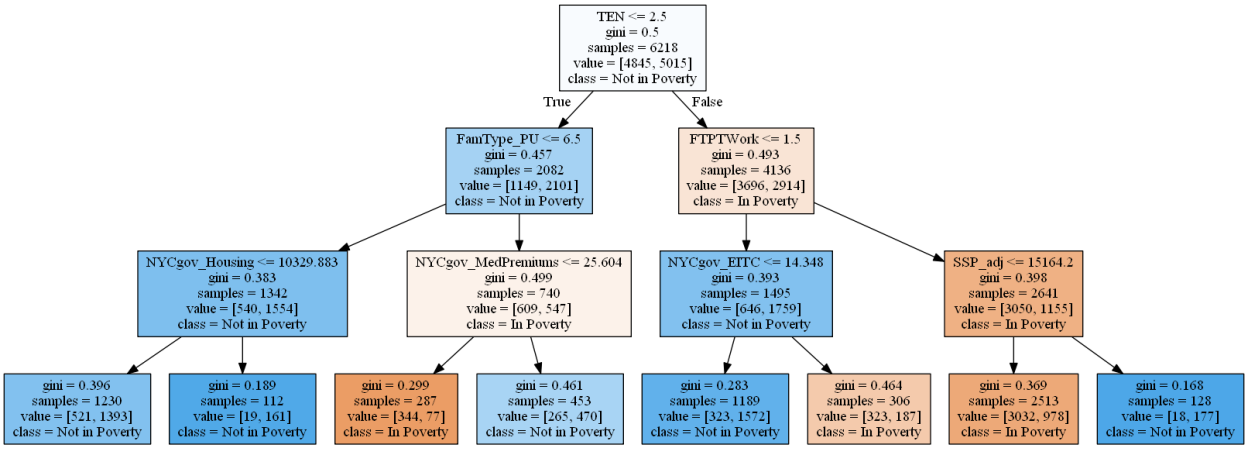
I am using Sklearn for the model building, I set the max_depth to 3 in order to avoid overfitting. The graph below is the decision tree. The accuracy for the training set is 83% and 79 % for the testing set.

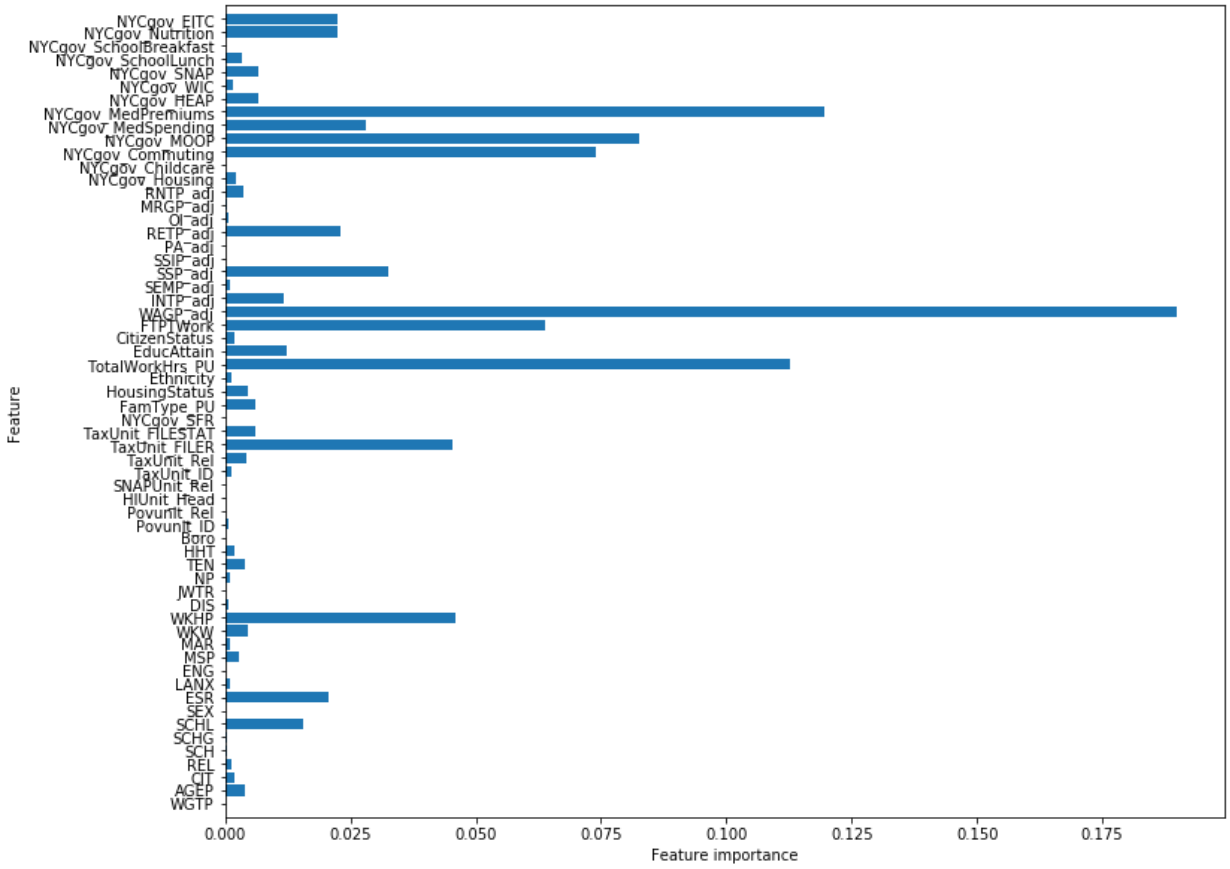


Since decision tree can produce unstable result because of the variation in the data, a random forest would improve the result.

Random Forest:

Random forest create many decision trees and combine all the unique trees to get a more robust result. For the random forest, I generated 1000 trees and used bootstrap sampling technique. The feature importance shows which attribute is more important to predict the outcome. The graphs below are one of the 1000 decision tree and also the importance of each features.





Finding:

The random forest shows the wage is most important feature that determine the poverty status, followed by medical insurance cost and working hour. *WAGP_adj* measures individual's wage in past 12 months, base on the official website of New York State (ny.gov), at the end of 2019, the minimum wage in NYC will increase to $15, which is roughly $28000 a year. Our decision tree shows a person that earns less than $23627 a year would more likely to falls into poverty, that means earning a minimum wage can guarantee a person stay out of poverty in most of the scenario.

Percentage of code from Internet: 36%