Final Report

My Dinh

12/8/2017

I. Introduction:

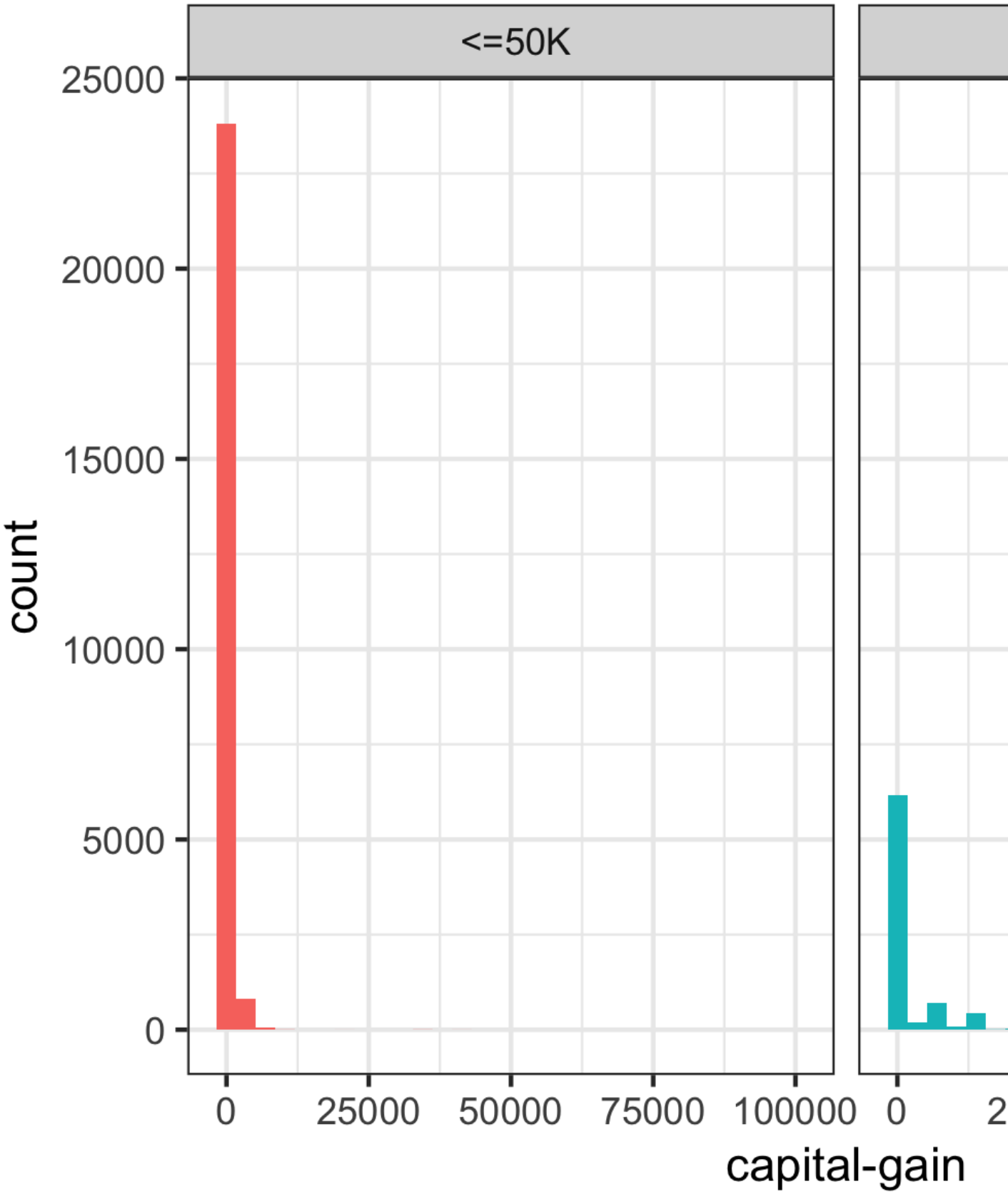
II. Data:

III. Data Processing/Feature Engineering:

Data Cleaning and Exploratory Analysis:

There appears to be a problem with imbalanced classes. We need to make necessary adjustments to adjust the probability threshold for imbalanced class.

At first glance, we look for a relationship between income and other continous variables:


<=50K
25000
20000
15000
10000
5000
0
count
0
25000
50000
75000
100000
0
capital-gain

- Capital loss: people with income over 50K tend to have higher capital loss overall, but with smaller maximum capital loss.
- Capital gain: people with income over 50K tend to have higher capital gains, with some extreme outliers with values over 100k. We may need to remove them during the cleaning process.
- Age: the age distribution of people whose income less than 50K tends to skewed to the left, indicating that they are of a younger age. Meanwhile, the age distribution of people with income higher than 50K is roughly normal, with higher mean and median.
- Education number: people with higher income tend to have higher education-numbers
- fnlwgt: both income groups roughly share the same distribution for the final weight

There is a total of 4262 missing values, for 2399 observations. Missing values are seen in the original dataset as '?' symbols. All the missing values belonged to three columns: occupation, nativecountry, and workclass, and multiple missing values were frequent for the same observation. As thus, it is very likely that the missing values are not at random (MNAR).

Feature Engineering:

IV. Model builing

A. Decision Tree:

B. Bagged Trees:

c. Random Forest