The Methods to Select Variables of Consumption Behaviors

Hao LI

May 31, 2021

We have data on payment time, payment channel, and consumption categories. Payment times are the time of checking out online. We transform the continuous payment time into four categories: morning (6:00-12:00), afternoon (12:00-18:00), night (18:00-24:00), and midnight (0:00-6:00). The payment channel refers to the resource of money in the e-wallet. There are three channels: debit cards, credit cards, and online transfers from other e-wallets. Moreover, there are 127 consumption categories, which are products and services; for example, food, online game, lottery, bike-sharing, real estate, etc. For the payment time, payment channel, and consumption categories, we further derive candidate variables based on the proportion of monthly consumption amount or count. For instance, the mean of ratios of *amount:total consumption* in bike-sharing; the volatility of ratios of *count:total count* of payments in the afternoon.

We use two methods to select variables. First, we apply the logistic regression throughout all candidate variables in the way of *baseline data + one candidate*. That is, we combine the demographic data plus one candidate, then proceed with the logistic regression. Based on the statistical outputs, we select the variables within a p-value threshold. This method is effective when dealing with variables of sufficiently high statistical significance. Nonetheless, this method has a problem of collinearity. Since variables are selected independently, some may become insignificant when combined together with other variables.

Second, we use the lasso to help with the selection. Lasso can help remove redundant predictors at the same time as mitigating the problem of collinearity. But lasso has its weaknesses. The first is time-consuming because the rate of convergence of the lasso is slow for the sparse and high dimensional data. The second is that lasso tends to select one variable from a group of correlated ones (but this will depend on the value of penalty parameters chosen). Namely, lasso tends to select variables of the most statistical significance and may neglect others even if they are also statistically significant. We might not want this conduct because some consumptions are of great interest in the explanation of individual behaviors; for example, consumption in online games may indicate addictive consumption and is significant. But it may be neglected because another consumption is more significant than and correlated with online games. Therefore, we use a hybrid method of lasso and a variant of backward stepwise regression. Initially, we run lasso over baseline and all candidate variables. Lasso will pick the most significant ones. Then we put the selected variables aside and run lasso over other candidates. Lasso will further select significant variables in this round. Likewise, we put these variables aside and run lasso with the rest candidates. We repeat this process until all candidate variables are treated. Finally, we collect the significant variables from the processes and combine them together.

Compared with the first method, the second method is better because lasso helps mitigate the problem caused by collinearity. If solely with the first method, the variables are selected one by one, and we will find that some variables become insignificant while others become even more significant when combined together in logistic regression. On the contrary, the lasso can select candidate variables altogether. The weakness of lasso is that it tends to select the most significant variables and neglects some slightly less significant ones. Therefore, we use a mechanism similar to backward stepwise regression to let lasso choose significant variables in a dynamic process.