

Christine Cai  
STA250 HW3  
WINTER 2014

Before doing any analysis, let us first explore and clean the data. I ignored the variable “PostId” as that should be unique to each post and should not have anything to do with a post being open or closed. One of my variables is the time between “OwnerCreationDate” and “PostCreationDate”. It seems intuitive that the longer someone has been a member of the StackOverflow community, the less likely their posts will be closed. There are some entries where “PostCreationDate” is earlier than “OwnerCreationDate”, let’s simply ignore these rows. Another variable is “ReputationAtPostCreation”; the higher a user’s reputation is, the less likely that his/her post will be closed. Since the lowest reputation is supposedly one, let’s also ignore the rows where “ReputationAtPostCreation” is less than one. Other variables include: “OwnerUserId”, “OwnerUndeletedAnswerCountAtPostTime”, and number of tags.

I started out exploring linear/quadratic discriminant analysis. However, these methods are infeasible for this assignment, as an assumption is the normality of the independent variables. Usually, the normality assumption can be slightly violated, and we would still get reasonable results. However, normality would be grossly violated here. For one thing, we have some discrete variables; it would be a shame to ignore them. For example, for a variable such as “OwnerUserId”, even though the entries are numeric, we cannot treat this as a numeric variable. It should be treated as a factor, or discrete variable. As for the numerical (continuous) variables, most of them are not even close to normal even after transformations. Therefore, I directed my attention to logistic regression.

I created a variable  $y$ , which is equal to 1 if the post is open, and 0 otherwise. This will serve as our dependent variable.