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. OwnerUserId is ignored due to the large number of factor levels involved, and it wouldn't make sense to include it as a continuous variable. Also, as Karen Ng pointed out, ReputationAtPostCreation should carry similar information as OwnerUserId (in this prediction context). I also ignored "PostClosedDate" as this cannot help predict if a post will be closed (in real data).

One of my variables is the time between OwnerCreationDat 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. I looked at the number of words in Title; the summary statistics for open vs closed are very similar. Therefore, the number of words in Title would not make a good predictor and is not used. The number of words in BodyMarkdown, however, is used. Other variables include: OwnerUndeletedAnswerCountAtPostTime, 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 dis-

crete variables; it would be a shame to ignore them. 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 serves as our dependent variable.