Time feature processing

In the data exploratory analysis, we did a plot of average success over time, a similar plot as the literature suggested (We found that the time plot in the paper is incorrect, since it calculated time zone effect in the wrong direction. Therefore, we fixed this problem and created new plots). The time is measured by UTC time zone, and we found that submissions at different time may have different possibilities of successes. Furthermore, we found that the average number of comments or scores also differs depending on the submission times. Therefore, we decided to add a set of features related to the time to our models.

First, we regularized submission time into uniform UTC time zone. We obtained the raw time from the original data and did regular expression matchings to extract the information of date and time as submission time. For each submission time, we split it into two parts: one is the time slot during the day, the other one is the date of submission. For the first part, the 6 time slots are night (22-1), sleep time (2-7), morning (8-11), afternoon (12-17), and other time. The division of time slots are inspired by the previous plots. The time slot for each submission is decided by its hour of submission. For the second part, the date of submission contains the month and the day of the week. Notice that these are in UTC time zone, so we also added dummy variables to indicate the local time zone. There are 3 different time zones in the data set (UTC+07:00, UTC+08:00, and UTC+00:00). Therefore, there are 27 different categories of time feature. To avoid perfect collinearity, we drop one feature for each correlated set of features, so the total number of time features are 23.