We performed naive Bayes analysis, extracting into a dictionary the morphological stems of the words contained in the answer text. We managed the dataset to control the length of included words, inclusion of numerical values, etc. We additionally added some HTML tags manually, as the stem extraction removed these. We then trained a Bernoulli naive Bayes model to classify successful answers. Our accuracy was 62%. Top predictors included the following (in decreasing order of rank):

- HTML tags for paragraph separation
- HTML tags for code and fixed-width formatting (i.e., code)
- "You"
- Outside links ("href" from HTML link tags, "littp")
- Visual support: "img," "imgur," "plot," "png"

Because our data were relatively tractable to the implementation of naive Bayes, we had a very successful baseline to build from. The most valuable feature besides the score from naive Bayes for improving model performance was time elapsed since question post. Improvements beyond naive Bayes were marginal, as evidenced when plotting ROC, which shows universally high specificity and low sensitivity. The naive Bayes score therefore provides the bulk of predictive power, with only modest improvement in accuracy from inclusion of quantitative predictors and limited ability to improve on this model regardless of the modeling algorithm.

Nonetheless, we can extrapolate several successful strategies from the naive Bayes results, which include clear separation of guidance into paragraphs, providing concrete examples of good code rather than abstraction, and visual illustration, which is another means of providing examples. Time elapsed also provides a good predictor, and this indicates that prompt response is another factor that should be prioritized in online communication and in Stack Exchange in particular.

We hypothesized that new answers on old questions may not be as likely to have very many upvotes. Because visual inspection shows that the score decreases with time, we divide the data in 5000 quantiles and find the maximum score associated to each. This provides an upper bound on the way increased time elapsed in an answer decreases its associated utility. We implemented the estimated algorithms on the test set, we find no increase in predictive power with other features. When we implement multiple models including

- GDA
- Random forest
- Naive Bayes
- Logistic regression

Use of more advanced NLP methods to predict successful answers based on blocks of phrase as well as grammatical structure.

Further calibration of the model we have developed using these measures to supplement the naive Bayes scores.

We then trained a Bernoulli naive Bayes model to classify successful answers based on blocks of phrase as well as grammatical structure.

We performed naive Bayes analysis, extracting into a dictionary the morphological stems of the words contained in the answer text. We managed the dataset to control the length of included words, inclusion of numerical values, etc. We additionally added some HTML tags manually, as the stem extraction removed these. We then trained a Bernoulli naive Bayes model to classify successful answers. Our accuracy was 62%. Top predictors included the following (in decreasing order of rank):

- HTML tags for paragraph separation
- HTML tags for code and fixed-width formatting (i.e., code)
- "You"
- Outside links ("href" from HTML link tags, "littp")
- Visual support: "img," "imgur," "plot," "png"

Because our data were relatively tractable to the implementation of naive Bayes, we had a very successful baseline to build from. The most valuable feature besides the score from naive Bayes for improving model performance was time elapsed since question post. Improvements beyond naive Bayes were marginal, as evidenced when plotting ROC, which shows universally high specificity and low sensitivity. The naive Bayes score therefore provides the bulk of predictive power, with only modest improvement in accuracy from inclusion of quantitative predictors and limited ability to improve on this model regardless of the modeling algorithm.

Nonetheless, we can extrapolate several successful strategies from the naive Bayes results, which include clear separation of guidance into paragraphs, providing concrete examples of good code rather than abstraction, and visual illustration, which is another means of providing examples. Time elapsed also provides a good predictor, and this indicates that prompt response is another factor that should be prioritized in online communication and in Stack Exchange in particular.

### References

1. Elaine Allen and Jeff Seaman. Changing course: Ten years of tracking online education in the united states. . .

Kevin Chen (kchen42@stanford.edu), Andrew Slottje (slottje@stanford.edu), Nurbek Tazhimbetov (nurbek@stanford.edu)

Institute for Computational and Mathematical Engineering, Stanford University