Preliminary Results

Stack Overflow Question Quality

Ryan Tobin, James Westbrook

The predictability of question quality depends on our definition of question quality. As a preliminary step, we used logistic regression with 5-fold cross validation to compare our two definitions. With the score definition, we were able to achieve a cross-validation score of 91.1% (8.9% cross-validation classification error). With the answered/not answered definition, we were able to achieve a cross-validation score of 75% (25% cross-validation classification error). Note that in both cases, datapoints with zero score were removed as we determined them to have too low community activity to be identified as good or bad, thus we removed them from our prediction [1]. Positive/negative score appears to be far more predictable than whether the question will end up answered. This poor predictability paired with our hypothesis that score is a better indicator of a quality question suggests to us that "answered" is not a viable definition of a good question. On the other hand, score seems to be very predictable, and though this suggests that we may be underfitting the idea of a "good" question, the use of it in related works permits us to use it in our future analysis.

Using score, we can analyze the importance of each feature. We computed the Pearson Correlation Coefficient of each feature with score (not positive or negative score, the raw score). This is summarized in Figure 1:

Though none of these features have particularly strong correlation with score, code-to-text ratio has the best, followed by number of codeblocks. This means that including code in a question is the strongest indicator we have of a good question.

Finally, we performed three classification algorithms on our dataset. The results are summarized in Table 1

Table 1: Classification Score

Logistic Regression	SVC	XGBoost
91.067%	B	C

(SVC and XGBoost coming soon…)

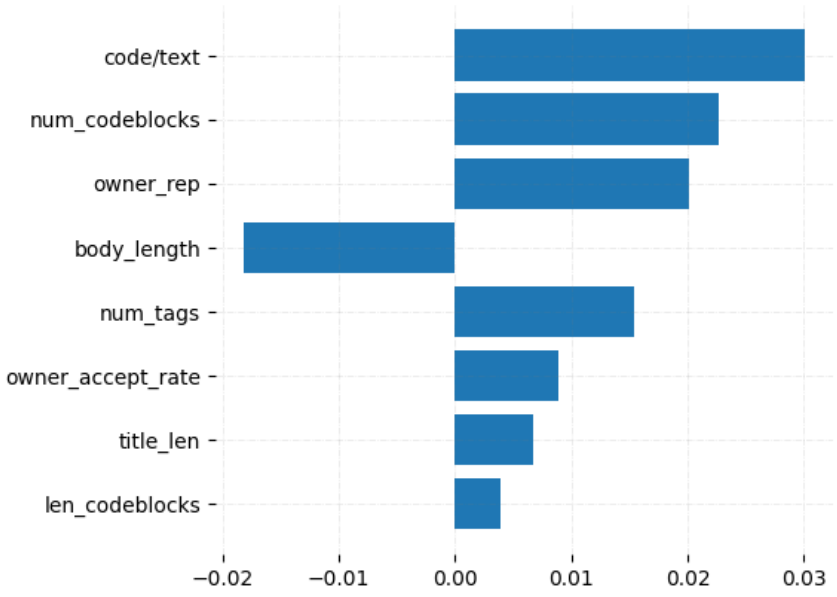

Figure 1: Pearson Correlation between Features and Score

References

1. Duijn M, Kucera A, Bacchelli A (2015) Quality questions need quality code: Classifying code fragments on stack overflow