Bang Nguyen

DATA-231: Applied Statistical Methods

How Gender, Team Size, and Commit History can Predict the Effectiveness of Software Engineering Teamwork in an Educational Setting

I. Introduction

Teamwork has been consistently recognized as an essential skill for software engineering practices. Several studies have been carried out to investigate the important factors associated with effective teamwork in software engineering. According to Verner et al. (2014), low motivation is found to be associated with the failure of software development team in a professional setting. Motivational factors such as whether the working environment is of great quality and whether the software engineer has a pleasant experience are highly related to the team's productivity (Verner et al., 2014). In an educational setting, the Software Engineering Teamwork Assessment and Prediction (SETAP) project, led by San Francisco State University, predicted teamwork performance based on students' activities on their final group project. The SETAP project focuses mainly on predicting whether the groups of students meet certain expectation on effective software engineering teamwork using a random forest classifier (Petkovic et al., 2016). According to Petkovic et al. (2016), the top predictors of teamwork success or failure are said to be intuitive, including the number of issues that needs an instructor's intervention and the average number of unique versions of code that were saved.

Using the same data from the SETAP project, this research paper, rather than creating a model with a powerful predicting ability, focuses on identifying the relevant factors associated with effective teamwork in software engineering classes. Teamwork performance will be assessed based on two criteria: process (how well the team applied best software engineering practices) and product (the quality of the finished product the team produced). Teamwork process and product are separately assessed, each in two phases – the design phase and the implementation phase. Through the use of multiple logistic regressions and relevant methods such as stepwise selection and cross validation, this paper aims to address the relationship between teamwork performance and the following factors: team size (number of team members), gender, and commit history (unique saved versions of the code).

II. Methods

The data used in this research is provided by the San Francisco State University in dedication to the SETAP project. As students in software engineering classes at San Francisco State University worked on their final group project, team activity measures such as number of people in a team, total number of

meeting hours, percentage of each gender in a team, etc. were collected. The team activity measures were aggregated through three main sources: the collection of the individual activity measures of each team member from weekly timecards, instructor observations, and software engineering tool usage logs. The collection of data was conducted for several consecutive semesters and achieved a sample of 74 observations (teams of students). This data, then, is not random but rather a census record that reflects the purpose of an observational study. In the final analysis, none of the observations were removed.

Like the SETAP project, this research paper is concerned with two response variables: the software engineering (SE) process grade and the software engineering (SE) product grade. Both of these variables have 2 levels: A and F, which indicate whether the team process/product is at expectation level or below expectation level, respectively. Out of 50 variables provided by the original data set, this research considered 8 explanatory variables: number of team members, percentage of female members in a team, gender of team leader, average meeting hours per member, average time spent on coding deliverables per member, average time spent on non-coding deliverables per member, average number of commits per member, and average number of unique messages associated with the commits. A commit is defined as a unique version of the code that is saved to the repository. Each commit has a message that briefly describes the changes have been made to the code.

With the same response variables that were collected at the end of the software engineering project, explanatory variables (except for demographic information that stays constant throughout the project such as number of team members, percentage of female members, gender of team leader) were collected separately in two phases – the design phase and the implementation phase. The design phase refers to when the team identified high level requirements, followed by more detailed requirements, and finally working on their first prototype. The implementation phase refers to the completion of their first prototype, followed by beta release and the final product.

Through analysis and modeling, three explanatory variables were recoded. First, the percentage of female was converted from a numerical variable to a binary one to indicates whether the team has at least one female team member. Second, the number of team members was recoded to whether a team has less or more than 4 members. Third, the average number of commits per team was converted into whether the team had made at least 20 commits on average. In summarizing the variables individually, proportions, and conditional proportions were calculated for categorical variables while means, medians, and standard deviation were calculated for quantitative variables. To investigate the relationship between the response variables and potential predictors, boxplots, scatterplots, and contingency tables were utilized. To investigate the relationship between categorical explanatory variables, two-sample proportion tests were conducted. Given the nature of the data and research questions, four logistic regression models

were fitted for each of the two assessments (SE product grade and SE process grade) in each of the two phases (design and implementation). In fine-tuning the models, two methods were at play: leave-one-out cross validation and step-wise selection based on AIC. AIC is an estimator of out-of-sample prediction error for logistic regression that is similar to adjusted R-squared in multiple linear regression (Tripathi, 2019). In leave-one-out cross validation, to calculate the accuracy rate of the model and at the same time avoids overfitting, a data set is split into a training set and a testing set n times (with n being the number of observations), where the training set has n – 1 observation, and the testing set has the remaining observation (Shaikh, 2018). During the modeling process, several interactions between gender and other factors were found to be significant (see appendix). However, the accuracy rate for these interaction models was lower than that for the non-interaction models. That said, the final models highlight the relationship between software engineering grade and the following factors without interaction: whether a team has at least one female member, whether the team has less or more than 4 members, and whether the team has made at least 20 commits.

III. Results

In the assessment of the software engineering teamwork process, 66.2% of the 74 teams receive grade A (at or above expectation) and 33.8% receive grade F (below expectation). In assessing teamwork based on the software product, 56.8% of the teams receive grade A (at or above expectation) and 43.2% receive grade F (below expectation). Table 1 and 2 showcase the demographic of the 74 teams from the sample based on the important variables.

Number of female members	Gender of team leader	Number of team members	SE Process Grade	SE Product Grade
None : 24	Female: 14	3-4: 16	A: 49	A: 42
At least one female: 50	Male: 60	5-7: 58	F: 25	F: 32

Table 1 - Summary statistics of demographic variables that stay constant throughout the two phases of the project

	Average number of commits - numeric	Average number of commits - binary	Percentage of unique messages
Design Phase	Median: 26.52 Mean: 30.09 Standard deviation: 22.74	Less than 20 commits: 23 At least 20 commits: 51	Median: 0.79 Mean: 0.69 Standard deviation: 0.26
Implementation Phase	Median: 34.90 Mean: 35.65 Standard deviation: 25.69	Less than 20 commits: 19 At least 20 commits: 55	Median: 0.85 Mean: 0.75 Standard deviation: 0.26

Table 2 - Summary statistics of variables that are different during each phase of the project

1. Gender, Number of team members, and Number of commits and SE Process Grade

The analysis has found that the significant factors related to SE Process Grade are the same for the design phase and the implementation phase, including whether the team has at least one female member, whether the team has 3-4 members or 5-7 members, and whether the team has at least 20 commits on average.

a. Exploratory Data Analysis

Analysis has found that teams without female members seem to be more likely to be below the expectation level in terms of SE Process grade. Specifically, 50% of the teams without a female member received an A in teamwork process while 74% of those with at least one female member received an A in SE process. In terms of the number of team members, teams with 3 to 4 students seem to be more likely to receive and F in SE process grade than teams with 5 to 7 students. Specifically, 62.5% of the teams with 3-4 were below expectation level, lower than their 5-7-members counterparts whose proportion of receiving an F is 25.9%. Interestingly, the number of female members per team and the number of members are significantly related to each other. The two-sample proportion tests have indicated that teams with 3 to 4 members are significantly less likely to have at least one female member than those with 5 to 7 members. That said, in the final model, only the number of female members per team is used in order to avoid multicollinearity.

Table 3 and 4 present the distribution of SE Process Grade based on the average number of commits during the design phase and the implementation phase, respectively. It can be seen that, for both the design phase and the implementation phase, the proportion of receiving an F in process grade for teams with at least 20 commits on average is lower than that for teams with less than 20 commits on average.

	Less than 20 commits	**At least 20 commits**
A	0.52	0.73
F	0.48	0.27

Table 3 - Contingency table of Whether the team has at least 20 commits on average during design phase and Process grade

	Less than 20 commits	**At least 20 commits**
A	0.47	0.73
F	0.53	0.27

Table 4 - Contingency table of Whether the team has at least 20 commits on average during implementation phase and Process grade

b. Modeling

The final logistic models that predict SE Process grade for the design phase and the implementation phase both use two significant predictors: number of female members and average number of commits. As mentioned earlier, although the number of team members is also related to SE process grade, it was removed from the final model due to multicollinearity with the number of female members. Table 5 and 6 present the results of

the final models that predict SE grade during the design phase and the implementation phase, respectively.

Coefficients and Odd Ratios (OR)					
Predictor	**Coefficient**	**Standard Error**	**OR**	**95% CI for OR**	**p-value**
Number of females in team - at least one female member	-1.06	0.5323	0.346	(0.119, 0.977)	0.047
Average number of commits - at least 20 commits	-0.901	0.5393	0.406	(0.138, 1.167)	0.095
Overall summary					
Number of observations	74				
Drop-in-deviance (Null deviance – Residual Deviance)	94.659 - 87.772 = 6.887				
p-value for drop-in-deviance test of overall fit	0.02				
Accuracy rate	74.3%				

Table 5 - Results of the final model that predict SE process grade during the design phase

Coefficients and Odd Ratios (OR)					
Predictor	**Coefficient**	**Standard Error**	**OR**	**95% CI for OR**	**p-value**
Number of females in team - at least one female member	-1.046	0.5357	0.351	(0.12, 0.998)	0.051
Average number of commits - at least 20 commits	-1.087	0.5393	0.337	(0.108, 1.022)	0.0556
Overall summary					
Number of observations	74				
Drop-in-deviance (Null deviance – Residual Deviance)	94.659 - 86.883 = 7.776				
p-value for drop-in-deviance test of overall fit	0.03				
Accuracy rate	73%				

Table 6 - Results of the final model that predict SE process grade during the implementation phase

Note: *Reference group for categorical variables:*

- SE Process Grade: A
- Number of females in team: None
- Average number of commits: < 20

Using the number of female members and the average number of commits to predict SE Process Grade, none of the observations were removed. The multiple logistic regression models have resulted in a residual deviance of 87.772 for the design phase and 86.883 for the implementation phase. The p-values for the drop-in-deviance tests indicates that during both the design and the implementation phase, the models using the number of female members and average number of commits are effective. For the design phase, 74.3% of the data were correctly predicted by the model. For the implementation phase, the accuracy is rate is about 73%. In assessing the conditions of a logistic regression model, the linearity condition is satisfied since both predictors are binary.

According to tables 5 and 6, both predictors are significant in the design phase and the implementation phase. During the design phase, the coefficient for whether a team has at least one female is -1.06, equivalent to an odd ratio of 0.346. This coefficient indicates that that the odds of receiving an F for those with at least one female member is 34.6% the odds of receiving an F for teams without a female member, holding commit category (< 20 or >= 20) constant. Similarly, the odds of receiving an F for teams with at least 20 commits on average is 40.6% that for teams with less than 20 commits on average, holding whether a team has at least one female member constant .

2. Gender, Number of team members, and Number of commits and SE Product Grade

a. Exploratory data analysis

In investigating the factors related to SE product grade, four explanatory variables demonstrate a potential relationship with SE product grade: whether the team has at least one female member, whether the team has 3-4 members or 5-7 members, the average number of commits, and the percentage of unique messages for commits. It should be noted that the average number of commits was treated as a numerical variable in predicting SE product grade.

The relationship that gender has with SE product grade is opposite to what we witnessed in SE process Grade. Specifically, teams without female are more likely to receive an A in SE product than teams with at least female, with proportions of receiving an A are 75% and 48% respectively. In terms of the number of team members, the relationship with SE product grade is less clear than that with SE process grade. However, it seems that teams with 5-7 members are less likely to be below expectation level in terms of SE product, with a proportion of having an F of 0.413 compared to that of 0.5 for teams with 3-4 members.

Figure 1 visualizes the distribution of SE product grade based on the average number of commits and the percentage of unique message for commits during the implementation phase. The figure has shown that the teams that receive an F tend to have lower average number of commits and lower

percentage of unique messages during the implementation phase. The relationship between these two variables with SE Product Grade is not clear during the design phase.

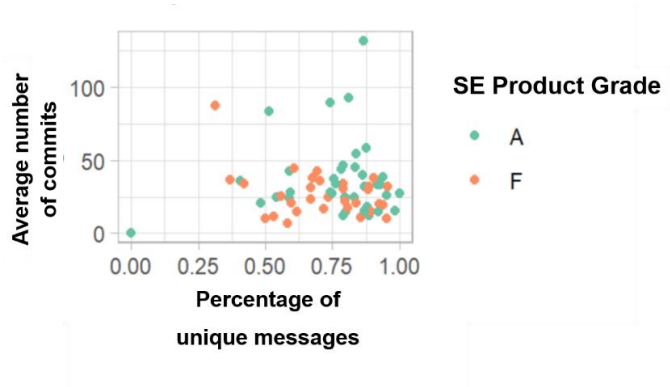

Figure 1 - Scatterplot of percentage of unique messages vs average number of commits during implementation phase, grouped by SE product grade

b. Modeling

The final models that predict SE product grade are exactly the same for the two phases. Specifically, the models for the two phases both only use whether the team has at least one female member as a predictor. The other predictors discussed in the exploratory data analysis are dropped due to either their insignificance in predicting SE product grade or their failure to meet the linearity condition (see appendix). As mentioned earlier, whether the team has at least one female member is a constant variable throughout the software engineering project (which included both the design phase and the implementation phase). Therefore, there is one common model that predicts SE product grade based on whether the team has at least one female member, which is displayed in table 7.

Coefficients and Odd Ratios (OR)					
Predictor	**Coefficient**	**Standard Error**	**OR**	**95% CI for OR**	**p-value**
Number of females in team - at least one female member	1.179	0.55	3.251	(1.152, 10.223)	0.032
Overall summary					
Number of observations	74				
Drop-in-deviance (Null deviance – Residual Deviance)	101.230 – 96.227 = 5.003				
p-value for drop-in-deviance test of overall fit	0.025				
Accuracy rate	59.5%				

Note: *Reference group for categorical variables:*

- SE Product Grade: A
- Number of females in team: None

Table 7 - Results of the final model that predicts SE product grade

According to table 7, the drop-in-deviance of the model is 5.003 and the relevant p-value is 0.03, indicating that the model is effective. In assessing the assumption of a logistic regression model, linearity is automatically satisfied because the predictor is a binary variable. The accuracy rate of the model is

about 59.5%, which is slightly better than that of the null model (56.8%). The coefficient of 1.179, equivalent to an odd ratio of 3.25, suggests that the odds of receiving an F in SE product grade is 3.25 times higher for teams with at least one female member than for all-male teams. This relationship is opposite to the relationship between gender and SE process grade where teams with at least one female member are less likely to receive an F in process grade.

IV. Discussion

The main findings of this research highlight the relationship between gender and whether the team is at/above or below expectation in terms of process grade and product grade. Specifically, results indicate that the odds of receiving an F in process grade is lower for teams with at least one female member than for teams without any. On the other hand, teams with at least one female are more likely to receive an F in product grade. This difference indicates that teams with at least one female member tend to better demonstrate best software engineering practices while the products of all-male teams tend to be rated higher. Additionally, for process grade, this research finds that teams of 5-7 students are more likely to receive an A than teams with 3-4 students. Last but not least, for process grade, an increase in the average number of commits is associated with a decrease in the odds of receiving an F. These relationships were not highlighted in the original SETAP project where the random forest classifier was used to best predict SE process and product grade using all of the 50 variables (Petkovic et al., 2016).

Except for the average number of commits per member, the other two predictors – whether the team has at least one female member and the number of team members – can be controlled by the instructor. Therefore, when assigning students in software engineering classes, the instructor can focus on these relationships in order to best promote teamwork. The generalizability of the results, however, is questionable due to the lack of random processes. A population that the results of this paper can be safely applied to is one of students in high-level software engineering classes in San Francisco State University. With thorough research, one can even extend these findings to students in software engineering classes among universities whose ranking of the computer science/software engineering department are of similar level to that of the San Francisco State University.

The strength of this research is the simplicity of the final models after multicollinearity among predictors were accounted for. One shortcoming of this research is the recoding of numerical variables into binary ones. For example, even though the results indicated that teams with at least 20 commits on average are more likely to receive an A in terms of process grade, this research is unable to identify the difference (or similarity) in the odds of receiving an A between teams with 20 commits on average and teams with 60 commits on average. Moreover, due to the nature of an observational study, there are confounding variables that are not addressed, such as whether the team members had taken a similar

software engineering class. If the team members have done so, they tend to make more commits to the code repository because experience from the other class has advised them to keep track of the code frequently, while also preparing the team members with the necessary practices that results in a higher chance to receive an A in process grade. Another limitation is the biased way in which SE process and SE product were graded. Even though the SE product and SE process were graded by external reviewers, such assessments are still based on subjective evaluation. For example, as Verner et al. suggested (2014), the assessment of effective teamwork can also be based on the personal experience of the team member, i.e., their sense of achievement when doing a good job in a project. Therefore, future research can look into other perspectives in assessing the effectiveness of software engineering teamwork. Finally, this paper highlights the relationship that effective teamwork has with aforementioned predictors without delving deeply into their interactions. Hence, another suggestion for future research is to consider such interactions, especially those between gender and other factors such as average number of commits and average meeting hours (see appendix).

References

Petkovic, D., Sosnick-Pérez, M., Okada, K., Todtenhoefer, R., Huang, S., Miglani, N., & Vigil, A. (2016). Using the random forest classifier to assess and predict student learning of Software Engineering Teamwork. *2016 IEEE Frontiers in Education Conference (FIE)*, 1-7.

San Francisco State University. (2016). Data for Software Engineering Teamwork Assessment in Education Setting Data Set [Data files]. Retrieved from https://archive.ics.uci.edu/ml/datasets/Data+for+Software+Engineering+Teamwork+Assessment+in+Education+Setting#

Shaikh, R. (2018, November 26). Cross Validation Explained: Evaluating estimator performance. Retrieved December 01, 2020, from https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

Tripathi, A. (2019, June 16). What is stepAIC in R? Retrieved December 01, 2020, from https://medium.com/@ashutosh.optimistic/what-is-stepaic-in-r-a65b71c9eeba

Verner, J., Babar, M., Cerpa, N., Hall, T., & Beecham, S. (2014). Factors that motivate software engineering teams: A four country empirical study. *Journal of Systems and Software*, *92*, 115-127. doi:10.1016/j.jss.2014.01.008