Predicting Coding Levels of Stack Overflow Developers

Alan Wang

2024-12-13

Introduction

Project Motivation

Stack Overflow is an online forum where programmers and developers can post questions and answer those of others to gain advice or learn more about various technical challenges (Barua, Thomas, and Hassan (2014)). Questions from Stack Overflow users reflect a wide variety of programming levels, from software engineers to those who just program as a hobby, with the forum moderating hundreds of thousands of posts monthly (Barua, Thomas, and Hassan (2014); Allamanis and Sutton (2013)). Stack Overflow also releases an annual survey to developers that engage with the platform, and this case study looks into the data from the 2024 survey. This case study begins with an exploration of the study data and then an ordinal regression analysis to predict a survey respondent's coding proficiency level. With this analysis, this case study aims to use ordinal regression to identify characteristics of a developer that could be associated with their coding proficiency.

Data Overview

The data for this case study are from the 2024 Stack Overflow Annual Developer Survey. This survey was conducted in May 2024 and received responses from 65,437 developers. The survey focused on developers' basic information, educational background, career development, how much they engage with Stack Overflow, and stances on AI in software development and technology in general.

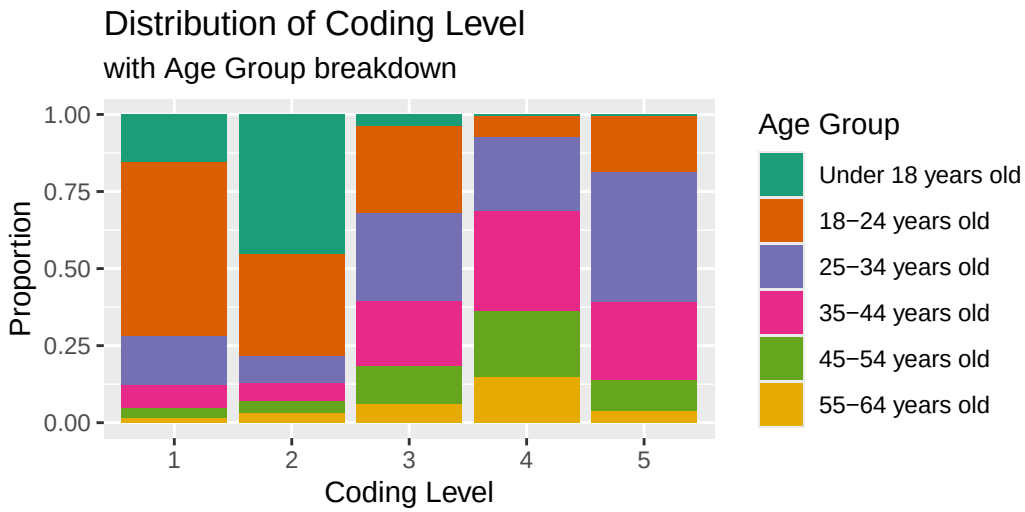
Relevant Variables

The following variables of interest were used in exploratory data analysis visualizations and/or the regression model as predictors.

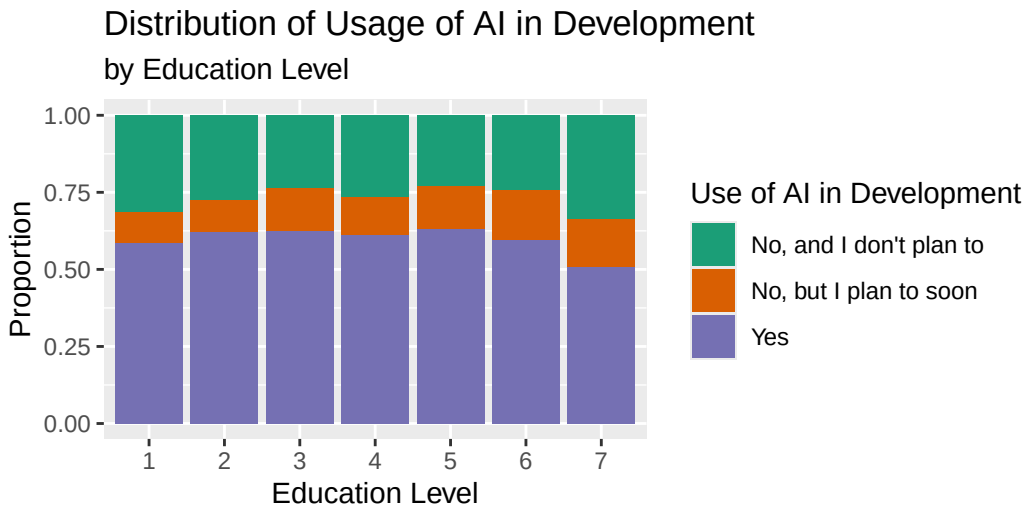
- Survey respondent's professional coding level on a scale of 1 to 5
- Survey respondent's highest education level on a scale of 1 of 7
- Survey respondent's age group on a scale of 1 to 7
- How frequently a survey respondent participates in the Stack Overflow community on a scale of 1 to 6
- How frequently a survey respondent visits Stack Overflow on a scale of 1 to 5
- How strongly a survey respondent considers themselves to be part of the Stack Overflow community on a scale from 1 to 5
- Survey respondent's use of AI in development on a scale of 1 to 3

See List 1 of Appendix for more descriptions of scales of these variables.

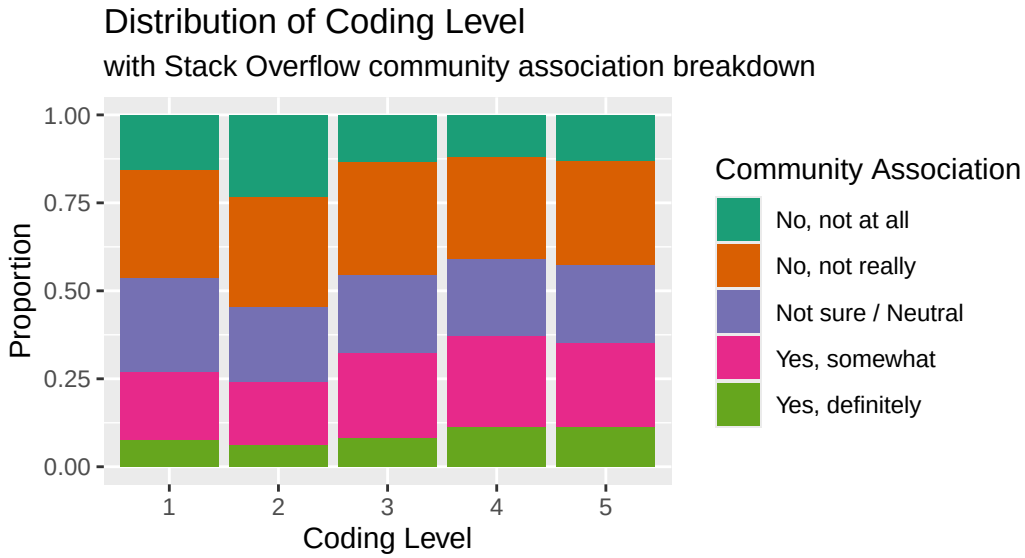
Exploratory Data Analysis



This visualization shows the proportion of respondents in each age group that are in each coding level group. It can be observed that for those who are learning to code, the majority are between 18 and 24 years old. Those who code as a hobby are primarily under 18 or between 18 and 24 years old. The respondents that stated that they code sometimes for academic or career purposes are primarily between the age of 18 and 44. Former developers tend to be older, with nearly 75% of these respondents being over the age of 35 and around 40% are over the age of 45. Those who currently are working as developers are primarily between the ages of 25 and 44 years old.



This visualization shows each of the seven education levels of the respondents and what proportions of them do or do not use AI in their development work. It can be observed from this visualization that the proportions of people who do use AI while coding are similar across all educational levels, with at least half of each group claiming that they do use AI. However, but the group of respondents that have professional degrees have the lowest proportion of people who use AI, at around 50%. This same group also has the largest proportion of people who do not currently use AI and do not plan on using it in the future. The proportion of people who do not currently use AI but plan to use it sometime soon is fairly consistent across all educational boards, with those that only have at most a secondary school education expressing less interest than those that have obtained more advanced educations.



This visualization shows the proportion of respondents' different association levels to the Stack Overflow community for each coding level. It can be observed that those who code as a hobby have the highest total proportion of respondents who do not feel as though they are part of the Stack Overflow community. On the other hand, it can be observed that those who are former developers have the highest total proportion of respondents who feel they are part of the Stack Overflow community. Both groups of former and current developers who filled out the survey have similar proportions of respondents who feel strongly about being part of the Stack Overflow community.

Research Question

There were multiple characteristics of Stack Overflow developers that were collected from the 2024 Stack Overflow Developer Survey. This case study aims to find which characteristics of a Stack Overflow developer are associated with their level of coding proficiency.

Methodology

Data Cleaning

Several of the relevant variables of interest were on scales that were not consistent with being lowest to highest, so these variables needed to be reordered.

The coding level variable needed to be reordered since the different coding levels were not in the order of coding proficiency in the original dataset. As an example, "Developer by profession" had a lower number than "Learning how to code".

Similarly, the original scale of education level was not in the order of least to most advanced. For example, "Primary/elementary school" had a higher number than "Master's degrees". When reordering the data, the categories "Some college/university study without earning a degree" and "Something else" were grouped into into one category that is in between Associate's degree and Bachelor's degree on the educational level scale. "Something else" most likely refers to bootcamps or trade school which were interpreted to be educations obtained that could be on par with or slightly more advanced than an Associate's degree, but certainly not more advanced than any other higher education degree. In addition, observations that had `NA` values were removed when creating exploratory data analysis (EDA) visualizations.

The age variable also needed to be reordered because the youngest age group, "Under 18 years old", was assigned to a number higher on the scale than any of the other age groups, so that age group was reassigned to be 1 and then the rest were shifted up one number on the scale. Observations for those respondents that said "Prefer not to answer" for their age group were also removed when creating EDA visualizations. The variables related to Stack Overflow engagement, which are participation frequency, visiting frequency, and community association,

were also reordered to have low-to-high scales. In addition, for community association specifically, respondents who said "Not sure" were grouped into the same category as those who said "Neutral".

The `as.factor()` function was applied to all variables of interest to ensure that the numerical values are treated as categories when fitting the model. See List 1 in the Appendix for more details on the reordering of the above variables.

Model Overview and Justification

An ordinal regression model was chosen to predict a respondent's coding level. This modeling approach was well-suited because the response variable is an ordered categorical variable on a scale from 1 to 5.

The process for selecting predictor variables for the model was primarily guided by their relevance to coding proficiency based on how they were defined in the survey dataset's data dictionary. While many variables were initially considered, not all were included final model. Some variables had to be excluded from the model to avoid potential multicollinearity between predictors. For example, age and years of coding experience were two potential predictor variables not included in the final model as predictors because of potential collinearity with education level. The final list of variables selected for the model is shown below:

Response Variable: Coding level

Predictor Variables: Education level, Stack Overflow participation frequency, Stack Overflow visiting frequency, Stack Overflow community association, Use of AI in development

Results

$$
\begin{aligned}
&\text{logit}(p_i) = \theta_j + 0.792 \times \mathbf{I}(\text{ed\_level}_i = 2) + 2.512 \times \mathbf{I}(\text{ed\_level}_i = 3) + 2.001 \times \mathbf{I}(\text{ed\_level}_i = 4) \\
&+ 2.906 \times \mathbf{I}(\text{ed\_level}_i = 5) + 2.775 \times \mathbf{I}(\text{ed\_level}_i = 6) + 1.727 \times \mathbf{I}(\text{ed\_level}_i = 7) \\
&+ 0.561 \times \mathbf{I}(\text{so\_part\_freq}_i = 2) + 0.433 \times \mathbf{I}(\text{so\_part\_freq}_i = 3) + 0.416 \times \mathbf{I}(\text{so\_part\_freq}_i = 4) \\
&+ 0.142 \times \mathbf{I}(\text{so\_part\_freq}_i = 5) + 0.083 \times \mathbf{I}(\text{so\_part\_freq}_i = 6) \\
&+ 0.546 \times \mathbf{I}(\text{so\_visit\_freq}_i = 2) + 0.832 \times \mathbf{I}(\text{so\_visit\_freq}_i = 3) \\
&+ 1.137 \times \mathbf{I}(\text{so\_visit\_freq}_i = 4) + 1.339 \times \mathbf{I}(\text{so\_visit\_freq}_i = 5) \\
&- 0.220 \times \mathbf{I}(\text{so\_comm}_i = 2) - 0.351 \times \mathbf{I}(\text{so\_comm}_i = 3) \\
&- 0.404 \times \mathbf{I}(\text{so\_comm}_i = 4) - 0.316 \times \mathbf{I}(\text{so\_comm}_i = 5) \\
&- 0.084 \times \mathbf{I}(\text{ai\_select}_i = 2) + 0.188 \times \mathbf{I}(\text{ai\_select}_i = 3) \\
&\text{where } p_i = P(\text{coding\_level}_i > j), j = 1, 2, 3, 4 \\
&\theta_j = j\text{th intercept (threshold between category } j \text{ and } j+1) \\
&\theta_1 = -0.008, \theta_2 = 0.851, \theta_3 = 1.815, \theta_4 = 1.991
\end{aligned}
$$

S.O. = Stack Overflow, C.L. = Coding Level

Term	Estimate	Standard Error	T-value	P-value	Coefficient Type
C.L. = 1|C.L. = 2	-0.008	0.087	-0.096	0.924	intercept
C.L. = 2|C.L. = 3	0.851	0.086	9.857	< 0.001	intercept
C.L. = 3|C.L. = 4	1.815	0.087	20.810	< 0.001	intercept
C.L. = 4|C.L. = 5	1.991	0.087	22.789	< 0.001	intercept
Educ. Level = 2	0.792	0.073	10.865	< 0.001	location
Educ. Level = 3	2.512	0.099	25.397	< 0.001	location
Educ. Level = 4	2.001	0.073	27.431	< 0.001	location
Educ. Level = 5	2.906	0.071	41.179	< 0.001	location
Educ. Level = 6	2.775	0.072	38.517	< 0.001	location
Educ. Level = 7	1.727	0.079	21.779	< 0.001	location
S.O. Part. Freq. = 2	0.561	0.029	19.050	< 0.001	location
S.O. Part. Freq. = 3	0.433	0.044	9.897	< 0.001	location
S.O. Part. Freq. = 4	0.416	0.066	6.261	< 0.001	location
S.O. Part. Freq. = 5	0.142	0.095	1.503	0.133	location
S.O. Part. Freq. = 6	0.083	0.134	0.619	0.536	location
S.O. Visiting = 2	0.546	0.048	11.284	< 0.001	location
S.O. Visiting = 3	0.832	0.048	17.266	< 0.001	location
S.O. Visiting = 4	1.137	0.051	22.169	< 0.001	location
S.O. Visiting = 5	1.339	0.060	22.414	< 0.001	location
S.O. Community Assoc. = 2	-0.220	0.047	-4.661	< 0.001	location
S.O. Community Assoc. = 3	-0.351	0.049	-7.182	< 0.001	location
S.O. Community Assoc. = 4	-0.404	0.049	-8.249	< 0.001	location
S.O. Community Assoc. = 5	-0.316	0.059	-5.364	< 0.001	location
Use of AI in Dev. = 2	-0.084	0.040	-2.077	0.038	location
Use of AI in Dev. = 3	0.188	0.029	6.460	< 0.001	location

When comparing educational backgrounds of survey respondents, a developer with a bachelor's degree has significantly greater odds of having the next higher coding level than a developer with primary education, while holding all else constant. Specifically, their odds are 18.284 times higher ($e^{2.906}$), which is the highest among all education levels relative to the baseline of just having primary education. In contrast, developers with professional degrees, such as a J.D., M.D., or Ph.D., only have 5.624 ($e^{1.727}$) times greater odds compared to those with primary education, holding all else constant.

At the $\alpha = 0.05$ significance level, there is statistically significant evidence to determine that increasing Stack Overflow visitation frequency is associated with higher coding levels, as it can be observed that the odds ratio increases for each level of visiting frequency. A developer who visit the platform multiple times a day has 3.815 ($e^{1.339}$) times greater odds of having the next higher coding level compared to a developer who visits less than once per month or monthly, holding all else constant, the highest among all visiting frequency levels.

In contrast, a potential inverse relationship between association with the Stack Overflow community and coding level is statistically significant at the $\alpha = 0.05$ significance level. The odds that a developer believes they are somewhat associated with the Stack Overflow community has the next higher coding level are 0.667 ($e^{-0.404}$) times that of a developer who believes they are not associated at all, holding all else constant, which is the lowest among all community association levels. All of the corresponding odds ratios are less than 1, suggesting that stronger associations with the Stack Overflow community may not be associated with higher coding levels.

In addition, when observing participation frequency, a developer who participates in Q&A on Stack Overflow less than once per month or monthly has odds of having the next higher coding level of 1.752 ($e^{0.561}$) times greater than a developer who has never participated at all, holding all else constant, which is the highest of all participation frequency levels. It can be observed that the odds that a developer has the next higher coding level in comparison to a developer who has never participated in Q&A on Stack Overflow decreases for each level of participation frequency from 2 to 4.

Finally, a developer uses AI in development only has 1.207 ($e^{0.188}$) times greater odds of having the next higher coding level compared to a developer who does not use AI in development and does not plan to use it, while holding all else constant. In comparison, the odds that a developer who does not currently use AI in development but plans to start using it soon has the next higher coding level are 0.919 ($e^{-0.084}$) times greater than a developer who does not use AI in development and does not plan to use it, which suggests that those who do not currently use AI but plan to may not have higher coding levels than those who do not plan to use AI at all.

Discussion

There is enough evidence to determine that education level, Stack Overflow visiting frequency, Stack Overflow community association, and the use of AI in development are statistically significant characteristics of a developer's coding level at the $\alpha = 0.05$ significance level. As noted in the `Results` section, the odds that a developer who has a bachelor's degree has the next higher coding level compared to a developer who has a primary education are higher than the odds ratios for both developers with master's and professional degrees, holding all else constant. This could potentially be attributed to variance in the types of degrees not represented in the dataset. Those with graduate degrees in humanities or natural sciences may

not have the same level of coding experience that those with bachelor's degree in quantitative sciences do.

It was also determined from the model that there was a potential inverse relationship between Stack Overflow community association and coding level, specifically that stronger community associations could be associated with lower coding levels. Because Stack Overflow is a website where developers typically ask questions for help on development work, current developers with a lot of coding experience may not necessarily need to use online forums for help. In addition, former developers may be coding less or not coding at all at the time of completing the survey, so they may not have to engage with Stack Overflow as much either.

In addition, there were a few limitations with the dataset, as the majority of variables were categorical, so the types of EDA visualizations that could be created were limited, and there were certain variables of interest that could not be included in the final model because of concerns of multicollinearity as mentioned previously in `Model Overview and Justification`. Other variables in the dataset, such as a respondent's compensation and what type of role they are currently in as a developer, unfortunately could also not be used for this case study because only a fraction of the respondents stated that they were currently employed when filling out the survey.

Future work for this case study could focus on only survey respondents who are current developers. In this case, a linear regression model could be able to include the variables used as predictors for the ordinal regression model in this case study, along with other variables such as job description, purchasing power of technology, and company size to predict a developer's compensation.

This case study's regression analysis could potentially be used by the Stack Overflow platform itself. The predictors in the regression model are pulled directly from the developer survey results, so Stack Overflow admin can use this model to predict actual Stack Overflow users' coding level, and allow for users to display a virtual badge representative of their coding level when they post or respond to questions. This can not only help other users connect with those who are of similar coding levels but also with those who are more experienced for advice and guidance.

References

Allamanis, Miltiadis, and Charles Sutton. 2013. "Why, When, and What: Analyzing Stack Overflow Questions by Topic, Type, and Code." In *2013 10th Working Conference on Mining Software Repositories (MSR)*, 53–56. IEEE.

Barua, Anton, Stephen W Thomas, and Ahmed E Hassan. 2014. "What Are Developers Talking about? An Analysis of Topics and Trends in Stack Overflow." *Empirical Software Engineering* 19: 619–54.

Appendix

List 1:

`coding_level` (originally `main_branch` in dataset): Respondent's coding level

- 1: I am learning to code (originally 2 in dataset)
- 2: I code primarily as a hobby (originally 4 in dataset)
- 3: I am not primarily a developer, but I write code sometimes as part of my work/studies
- 4: I used to be a developer by profession, but no longer am (originally 5 in dataset)
- 5: I am a developer by profession (originally 1 in dataset)

`ed_level`: Respondent's education level

- 1: Primary/elementary school (originally 4 in dataset)
- 2: Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.) (originally 6 in dataset)
- 3: Associate degree (A.A., A.S., etc.) (originally 1 in dataset)
- 4: Some college/university study without earning a degree + Something else (originally 7 and 8 respectively in dataset)
- 5: Bachelor's degree (B.A., B.S., B.Eng., etc.) (originally 2 in dataset)
- 6: Master's degree (M.A., M.S., M.Eng., MBA, etc.) (originally 3 in dataset)
- 7: Professional degree (JD, MD, Ph.D, Ed.D, etc.) (originally 5 in dataset)

`age`: Respondent's age group

- 1: Under 18 years old (originally 8 in dataset)
- 2: 18-24 years old (originally 1 in dataset)
- 3: 25-34 years old (originally 2 in dataset)
- 4: 35-44 years old (originally 3 in dataset)
- 5: 45-54 years old (originally 4 in dataset)
- 6: 55-64 years old (originally 5 in dataset)
- 7: 65 years or older (originally 6 in dataset)

`so_part_freq`: Respondent's participation frequency on Stack Overflow

- 1: I have never participated in Q&A on Stack Overflow (originally 4 in dataset)
- 2: Less than once per month or monthly (originally 5 in dataset)
- 3: A few times per month or weekly (originally 1 in dataset)
- 4: A few times per week (originally 2 in the dataset)
- 5: Daily or almost daily (originally 3 in the dataset)
- 6: Multiple times per day

`so_visit_freq`: Respondent's visitation frequency on Stack Overflow

- 1: Less than once per month or monthly (originally 4 in the dataset)

- 2: A few times per month or weekly (originally 1 in the dataset)
- 3: A few times per week (originally 2 in the dataset)
- 4: Daily or almost daily (originally 3 in the dataset)
- 5: Multiple times per day

`so_comm`: Respondent's association wit the Stack Overflow community

- 1: No, not at all (originally 2 in the dataset)
- 2: No, not really (originally 3 in the dataset)
- 3: Not sure / Neutral (originally 4 and 1 respectively in the dataset)
- 4: Yes, somewhat (originally 6 in the dataset)
- 5: Yes, definitely

`ai_select`: Respondent's use of AI in development

- 1: No, and I don't plan to
- 2: No, but I plan to soon
- 3: Yes