STATISTICAL ANALYSIS PLAN

Data Overview

Stack Overflow serves as a central hub for computer programmers worldwide, offering a platform for both questions and answers as well as acting as a comprehensive resource for programming knowledge. Each year, Stack Overflow conducts an extensive online survey. The 2023 edition ran from May 8 to May 19, attracting participation from approximately 91,000 software engineers across 185 countries. Following rigorous privacy and data consent checks, 89,184 responses were deemed fit for analysis (Stack Overflow, 2023). This survey encapsulated 84 diverse variables spanning seven primary sections, including Basic Information, Education, Work and Career, Technology and Tech Culture, Stack Overflow Usage, Artificial Intelligence, and Professional Developer Insights. In our dataset, each row represents the response of one of the participants, while the columns store the answers to the questionnaire questions. The research questions we will be analyzing are as follows:

I. *How can we predict a developer's yearly compensation?*
II. *What elements shape a developer's job satisfaction?*

Modeling

We will now describe how we will address each of our research questions.

I. *How can we predict a developer's yearly compensation?*

The initial research inquiry pertains to a prediction model, specifically aiming to forecast a developer's annual compensation based on certain independent variables. For this question we aim to build a predictive multilinear regression model using the following variables:

Variable	***Type***	***Responses***
1. Age	Categorical	[18-24 years old, 25-34 years old, 35-44 years old, 45-54 years old, 55-64 years old, 65 years or older, prefer not to say, under 18 years old]
2. Type of work	Categorical	[Hybrid, In-person, Remote]
3. Years of coding experience not including formal education	Discrete	[Less than 1 year, numerical values in a range 1 to 50, More than 50 years]
4. Education level	Categorical	[Associate degree, bachelor's degree, master's degree, elementary school, professional degree, secondary school, some college, something else]
5. Years of working experience	Discrete	[0 to 50]
6. Number of programming languages each developer knew and worked with	Discrete	[0 to 51]

In terms of variable selection, we will be comparing the model that includes the interaction term with another model that excludes it. We will apply cross-validation to both models and then measure which one performs better using RMSE. Once we have the selected model, we will proceed with model assessment using the adjusted R-squared.

We are interested in how the relationship between years of coding experience and a developer's yearly compensation varies across different levels of education, so for this research question we will include an interaction term with educational level and years of coding experience. This will allow us to assess whether the impact of coding experience on compensation differs among developers with varying educational backgrounds.

II. What elements shape a developer's job satisfaction?

For our second question we will be using an ordinal regression. This question is related to an inference problem where we want to understand which variables impact job satisfaction. We will measure the job satisfaction metric on an ordinal scale (“Satisfied”, “Neutral", "Dissatisfied"). Specifically, we want to comprehend how predictors affect the likelihood of transitioning from one category to another. For this ordinal regression, it's essential to consider variables that are logically associated with job satisfaction.

We aim to build this ordinal regression model using the following variables:

Variable	***Type***	***Responses***
1. Age	Categorical	[18-24 years old, 25-34 years old, 35-44 years old, 45-54 years old, 55-64 years old, 65 years or older, prefer not to say, under 18 years old]
2. Type of work	Categorical	[Hybrid, In-person, Remote]
3. Organization size	Categorical	[2 to 9 employees,10 to 19 employees, 20 to 99 employees,100 to 499 employees, 500 to 999 employees,1,000 to 4,999 employees, 5,000 to 9,999 employees, 10,000 or more employees, I don’t know, just me]
4. Yearly compensation in USD	Continuous	[$1 to $74,351,432]
5. Usage and incorporation of Artificial Intelligence to the job	Categorical	[No and I don't plan to, No but I plan to soon, Yes]

For the selection of the dependent variable, we will convert the survey results for Job Satisfaction. "Strongly disagree" and "Disagree" will be categorized as Dissatisfied, "Neither agree nor disagree" will be considered as Neutral, and "Agree" and "Strongly agree" responses will fall into the Satisfied category. Regarding model assessment, we will evaluate the principle of the proportional odds assumption and make a comparison between multinomial and ordinal logistic regression.

We are interested in how the relationship between a developer’s yearly compensation and job satisfaction varies across different categories of age so we will include an interaction term with age and yearly compensation. In summary, including this interaction term with age and yearly compensation in our model will allow us to investigate how the relationship between yearly compensation and job satisfaction is modified or varies across different categories of age.

Potential challenges

To address the challenges identified in the EDA report, we will convert the 8 questions into numerical equivalents. Subsequently, we will use the average of each person's responses to determine which of the categories, namely "Satisfied," "Neutral," and "Dissatisfied," they fall into. In the case of missing data for the Job Satisfaction questions, we observed that when a person did not respond to one of the questions, they did not respond to any of the other 8 questions. Therefore, we have decided to remove these missing values.

We also observed non-numeric entries like 'less than 1 year' in the 'professional programmer experience' variable and we solve this by converting it to a numerical value of .5 year. Additionally, variable selection challenges arise due to few numeric variables, exemplified by age being categorized (e.g., 'under 18') rather than numerical, in that sense we have decided to get rid of those under 18 because our focus is on those who have already entered the workforce and are engaged in the industry. Another challenge we faced was to accurately plot the distribution of the target variables as well as their relationship with predictor variables. To make our plots insightful, we at times resorted to removing top 1% outliers from our compensation variable to show the distribution properly.

References

Stack Overflow. (2023). Stack Overflow Developer Survey 2023. Retrieved from https://insights.stackoverflow.com/survey