Predicting Code Readability Score

YoungWoo Song

Level of interest 8/10

Description of the project:

Code readability is an important factor in software development, as it influences the ease of maintaining and understanding code. However, assessing readability can be subjective and inconsistent across different developers. This project aims to build a machine learning model that predicts the readability score of a code snippet based on several data instances, including comments, variable names, indentation, line length, etc. By automatically identifying code that may need refactoring for better readability, the model can help developers improve code quality and maintainability. The project could also extend to providing automated code review processes by suggesting improvements based on the predicted readability score in the future.

What features the dataset might include:

The dataset for this project would include several features related to code quality and structure. These features may consist of the number of comments in the code, average line length, code complexity (using metrics like Cyclomatic complexity), the use of meaningful variable names, and adherence to style guides (such as proper indentation or consistent formatting).

Example dataset instance

Number of comments	Average line length	Cyclomatic complexity	Variable name quality	Style guide adherence	Target – readability score	etc
11	48	6	High	90%	0.75	…

Cyclomatic complexity – software metric used to measure the complexity of a program.

Readability Score – continuous value ranging from 0 to 1.

How and from where would the dataset be gathered and labeled:

The dataset can be gathered from open-source code repositories such as GitHub, where a variety of code snippets and projects are available. Additionally, submissions from coding competitions (e.g., Codeforces or LeetCode) can provide more diverse examples of coding styles for different projects and situations. To label the dataset, predefined readability metrics will be used to compute a readability score for each code snippet. Tools that calculate Cyclomatic complexity, Halstead metrics, and other code quality indicators will be employed for this purpose. The project can efficiently create a large dataset suitable for training and testing machine learning models using these tools. The main effort will involve extracting and preprocessing these features to ensure consistency across different code sources. This project will be focusing on python only. Later, it will be aiming for generalizing all the readability scores for various programming languages.