What types of projects do Scratch users create?

Wojciech Marczuk

W.P.Marczuk@student.tudelft.nl

About Scratch

Block-based programming language. Popular as an introductory language

Many new projects every day. Projects differ, but there is no classification process.

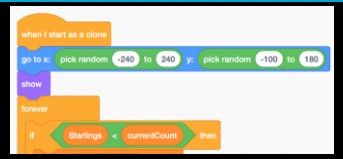

Figure 1: Example Scratch code

Related Work

A. Funke & K. Geldreich analyse gender differences in scratch projects created by children [1]. The projects created by children are classified into 6 different categories: Game, Animation, Interactive Art, Music and Dance, Story, Video Sensing.

Grassl et al. also look into scratch projects and analyse gender differences, but instead of categories for project types, they extract common terms generated automatically through an LDA model [2].

Introduction

RQ: How do different types of features relate to the project type?

Sub-question 1: Which pair of features is best to predict projects for each project type?

Sub-question 2: Which type of project is most accurately identified with the extracted features?

Is there a correlation between these features:

- Cyclomatic complexity
- Loop count
- "Say" block percentage
- Variable count

with the project types:

- Game
- Animation
- Interactive Art

Methodology

For my project I chose three categories since with manual classification, they are the two most distinct categories, which also show up in the explore tab of the scratch website [3].

Machine learning performed on dataset of 107 projects. This dataset was created by choosing projects at random from an existing dataset of scraped scratch projects [4]. A One VS Rest (OvR) approach was used.

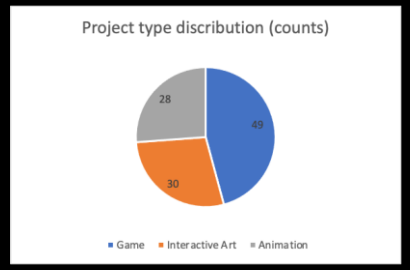

Figure 2: Distribution of projects by category in the dataset

Machine learning models used:

- Support Vector Machine
- Decision Tree
- Random Forest

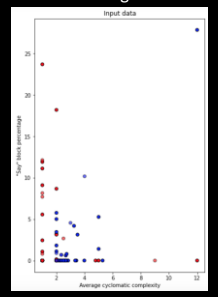

Figure 3: Input data example

Results

Example showing decision bounds for different machine learning models shown in figure 4. Blue points represent games, and red points represent all other projects.

Figures 5, 6, and 7 show the average performance per machine learning model for identifying each type of project.

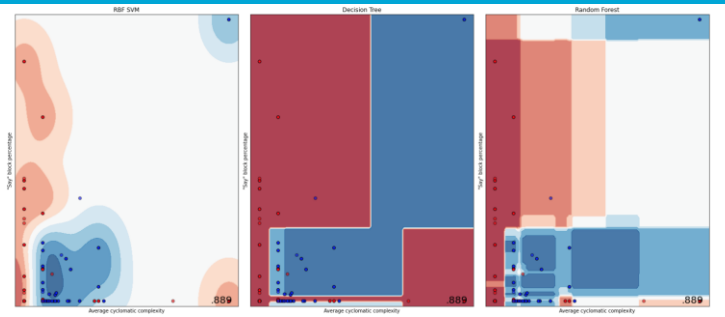

Figure 4: Classification of game projects by cyclomatic complexity and "say" blocks

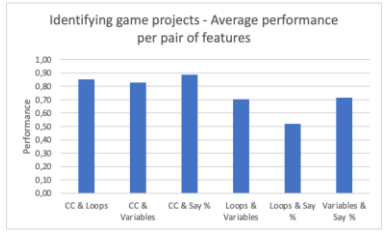

Figure 5: Performance for game identification

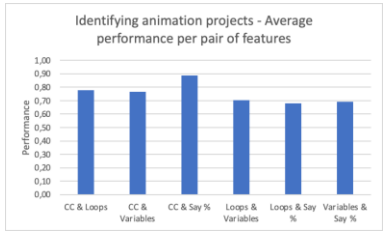

Figure 6: Performance for animation identification

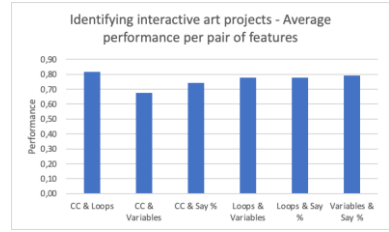

Figure 7: Performance for interactive art identification

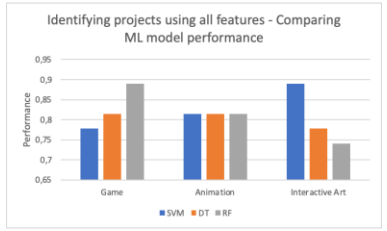

Figure 8: Performance for identification using all features

Figure 8 shows the performance identifying each project type using all features.
For identifying games, the top pair of features was cyclomatic complexity and the "say" block percentage. For identifying animation projects, the top pair of features was cyclomatic complexity and the "say" block percentage. For identifying interactive art projects, the top pair of features was cyclomatic complexity and the number of loops.

When using all features for classification: for games, random forest performed best. For animation projects, all models performed equally. For interactive art, support vector machine performed best. Since the results vary for each project type, it is not possible to decide on a best performing machine learning model for classification. The results are also similar enough.

Conclusion

Cyclomatic complexity was present in each top scoring pair of features. This is likely due to complexity differences within the chosen types of projects.

Since there was no clear answer to the question of what project is easiest to identify, it is possible that each type of project is equally difficult to identify, or a better choice of features could be chosen.

[1] Alexandra Funke and Katharina Geldreich. "Gender Differences in Scratch Programs of Primary School Children". In: Proceedings of the 12th Workshop on Primary and Secondary Computing Education. WiPSCE '17. Nijmegen, Netherlands: Association for Computing Machinery, 2017, 57a64. isbn: 9781450354288. doi: 10.1145/3137065.3137067. url: https://doi.org/10.1145/3137065.3137067.

[2] Isabella Graßl, Katharina Geldreich, and Gordon Fraser. "Data-Driven Analysis of Gender Differences and Similarities in Scratch Programs". In: Proceedings of the 16th Workshop in Primary and Secondary Computing Education. WiPSCE '21. Virtual Event, Germany: Association for Computing Machinery, 2021. isbn: 9781450385718. doi: 10.1145/3481312.3481345. url: https://doi.org/10.1145/3481312.3481345

[3] Scratch. Scratch statistics. https://scratch.mit.edu/statistics/, 2023

[4] Efthimia Aivaloglou, Felienne Hermans, Jesus Moreno-Leon, and Gregorio Robles. A dataset of scratch programs: Scraped, shaped and scored. In 2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR), pages 511–514, 2017.

TUDelft