Analyzing Key Factors in Sudoku Puzzle Difficulty

Leonid Grekhov Ryan Tran Suresh Goud Mula Vishal Shinde
San José State University

November 2021

Abstract

Introduction

Sudoku, dating all the way back to the 19th century, is one of the world's most famous and popular logic-based puzzles. The objective in Sudoku is to complete a 9x9 grid of numbers, from one to nine, such that each row, column, and each of the nine 3x3 subgrids contain each digit exactly once. The puzzle begins in a partially completed state with pre-filled digits or "clues" as we call them in this paper. Sudoku puzzles each have a single unique solution, and the difficulty in finding such a solution varies wildly from absolutely trivial to seemingly impossible. In this paper, we will attempt to analyze and uncover the key factors that contribute to what makes or breaks the difficulty of Sudoku puzzles. We will be using a Kaggle dataset that provides us with three million samples, each with four features: puzzle, solution, clues, and difficulty. We are given the puzzle as a string of 81 characters representing the 81 possible grid cells. Initial clues are given as their respective numerical digit and unknown values to be discovered are represented with a period character. We are also provided with the solution of the puzzle, though in this case, we will not be utilizing it for our purposes. Another possibly useful piece of information is the number of initial clues, which we will be extensively analyzing for possible insights. Lastly, we are given the estimated difficulty of the puzzle. Sudoku difficulty is not an objective measure, since there are so many different rating systems. For this dataset, the difficulty was calculated based on the average depth of search trees over ten attempts. In this paper, we will utilize a variety of statistical methods from statistics, data mining, and machine learning to analyze this dataset for insights. More specifically, we will be taking a closer look at how the number of missing values, the values of the digits themselves, and the structure of such values affect the difficulty of the Sudoku puzzles.

Methods

Initially we will look at the distribution of empty space and the clue distribution. Our initial analysis has shown that the distribution of empty places center around 57 meaning that most puzzles have 24 clues given. The Average number of empty spaces per row is around 6 out of a possible 9 spaces per row. It is important to keep in mind that the location of the empty spaces also matters in relation to the filled spaces. The difficulty distribution is very skewed with most puzzles falling into the zero difficulty category. Generally, as the difficulty increases, there are less samples. This may make it difficult to find patterns for higher difficulties because there are very few samples to train on. We proceed to decrease the number of zero difficulty puzzles down from 1 million to 100,000 for a better distribution of puzzle difficulties. Through mapping difficulties to the number of clues we can conclude that there is no direct correlation between them, only that the most difficult puzzles have 23 clues but decreasing and increasing the number of clues generally decreases the difficulty. Furthermore, training linear regression and random forest models using clues to predict difficulty yielded R squared values of nearly zero, further indicating that clues alone is not a good predictor for difficulty. Our first attempt building a neural network yielded

insignificant results as finding correlation between the puzzle and difficulty proved to be hard with values of difficulty not being directly dependent on the number of clues. Converting clues to the value of one had a similar outcome. We then attempted a convolutional neural network with original values and also clues set to one. Results for these networks also failed to provide a positive result. Our current hypothesis is that our model could be too simple to capture the true correlation between the puzzle and the difficulty. More training and model building will be required to improve our results.

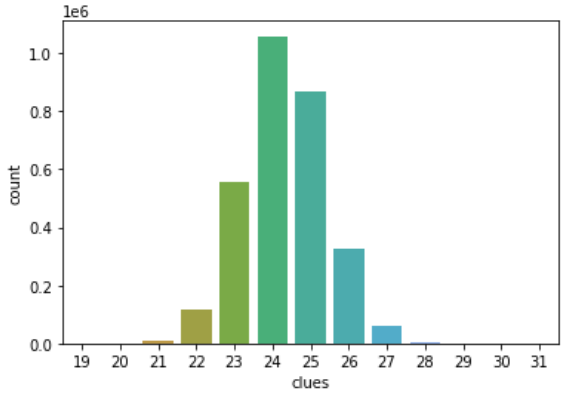

Figure 1: Clues Distribution

Comparisons

Example Analysis

Conclusions

References