

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

First, we will look at the distribution of empty space and the number of clues. These distributions are intimately related because they are exactly inverses. If the number of empty spaces is high then the number of clues is low and vice versa. From Figure 1, we can see that the number of clues is somewhat normally distributed and centered at around 24 clues. The vast majority of clues are between 23 and 26.

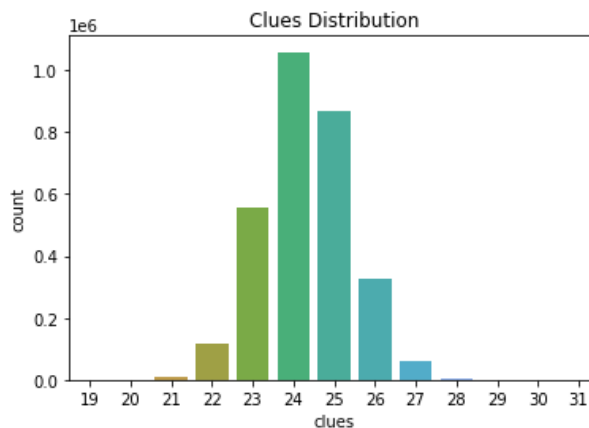


Figure 1: Clues Distribution

Next, we look at the distribution of difficulties. We

notice that the distribution is extremely right skewed due to there being so many samples of 0.0 difficulty. There are over one million samples of 0.0 difficulty, while the next most common difficulty of 1.0 had about 90,000 samples. We decided to reduce this discrepancy by randomly sampling 100,000 of the 1,000,000 samples of 0.0 difficulty. Though after the reduction, the difficulty distribution is still rather right skewed as seen in Figure 2.

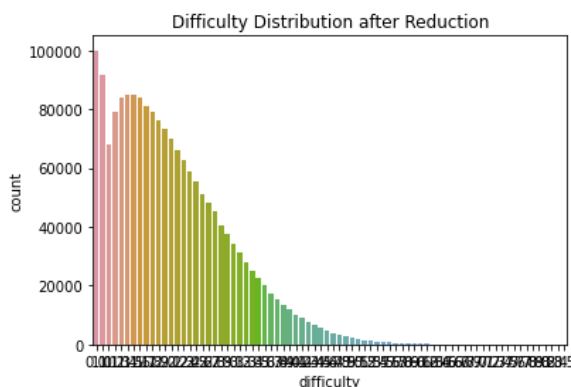


Figure 2: Difficulty Distribution after Sampling

Next, we attempt to uncover the relationship between the number of clues and difficulty. As seen in Figure 3, there is not a clear-cut relationship between the number of clues and difficulty. We somewhat expected the relationship to clearly indicate that more clues would result in lower difficulty, but the plot says otherwise. However, it can be noted that there is something to be said about the relationship between maximum difficulty and the number of clues. What's surprising is that this relationship is not monotonic. It seems the number of clues that results in the maximum difficulty is 23. Then, the maximum difficulty decreases as you either increase above or decrease below 23 clues. We then trained a linear regression model and a random forest model using the number of clues to try and predict difficulty. As expected, we were unable to fit such models and resulted in R squared values of nearly zero, indicating that the models were completely unable to fit the data. Thus, it is highly likely that the number of clues alone is not a good predictive independent

variable for the dependent difficulty variable.

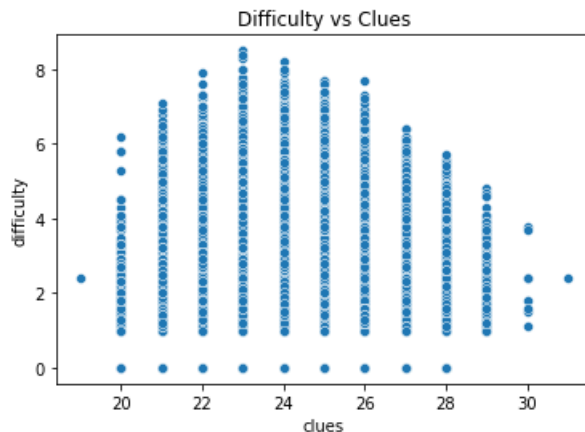


Figure 3: Difficulty vs Clues

Next, we try to use neural networks to find correlations between the puzzle itself and the difficulty. There are two approaches to using the puzzle data: numerical and binary. The numerical approach is to retain the actual numerical values of the digits in the puzzle. The binary approach is to convert all numerical values to the value 1 and all missing places to the value 0. We hypothesized that the numerical values may or may not be significant and that maybe simply the position in which digits existed may be the key factor. We attempted using vanilla neural networks as well as convolutional neural networks with both the numerical and binary approaches, but we came up empty with R squared values near zero once again. There could be many reasons why we did not achieve the results we desired. It could be that our models are simply not complex enough to capture the true relationship between the puzzle and difficulty. It could also be that it is simply not possible to reliably predict difficulty from the puzzles using data mining and machine learning techniques. This could very well be the case because from our exploratory data analysis, there was incredible overlap between difficulties with very similar looking puzzles and numbers of clues. Nevertheless, we will continue to explore different options for extracting further insights.

Comparisons

Example Analysis

Conclusions

References