# Table of contents

# Table of Figures

# Table of Tables

# Abstract

Here I will outline the actions performed in this thesis along with a brief description of the topic and results.

# Chapter 1: Introduction

This section will define the problem, background behind it, and the expected outcome as a result of the analytics performed

# Chapter 2: Research Design

A brief introduction into the planning behind the design and the impact of this design on the research.

## Primary Data

Primary data for this project will be established through an experimental process which will apply chess engines to chess mating puzzles with a fixed length. The engines will be tested for their depth efficiency at which they determine an endgame is a mating pattern, implying a higher possibility for depth of searching given a chess position.

## Problem Identification and Clarification

The problem set for this project is that of the incorporation of chess assessments into a chess opening position recommendation algorithms. Engines view problems as mathematical issues and do not assess with nuance like a human brain. Perhaps this straight to the point assessment can be deterministic in the development of a recommendation model?

## Research Objectives

Evaluate Chess Engines on Puzzles

## Validity Type

## Ethical Considerations

# Chapter 3: Literature Review

# Chapter 4: Methodology

## 4.1 Puzzle Analysis

A database of puzzles was obtained from Lichess Puzzle Database (Lichess, 2023). This is an open database published by Lichess to provide its users with access to all the puzzles they use on their site. It also includes summary statistics for the puzzles.

The Comma Separated Values file provided by Lichess.org was imported into a Python environment using Pandas where it was filtered to extract all the puzzles fitting a ‘Mate in X’ description. This provided a variety of different puzzles on which an analysis could be performed to establish the difference between engine horizon problem solving capabilities.

To provide sufficient testing for the engines, a range of mate lengths were chosen from the longest available. The database consisted of a range of ‘Mate in X’ problems ranging from 1 move to 10 moves. Five was chosen as the lowest value to provide a depth where the engines would need to execute a minimal search and 8 was chosen as the upper limit to allow for enough comparison in distribution for the selected engine parameters. Both populations for ‘Mate in 9’ and ‘Mate in 10’ puzzles would not have provided a large enough sample to compare distributions, both having 9 and 2 puzzles respectively, whereas puzzles with 5, 6, 7, and 8 moves have 3058, 2320, 503, and 183 puzzles in their samples.

A Python program using Python-Chess library (Python-Chess, 2023) was developed to implement an experimental procedure to test both Stockfish and Leela Chess Zero, two of the top open-source engines. In this program a custom function was developed to utilise the python-chess Universal Chess Interface (UCI) analyse function in an open ended analysis of each puzzle. This analysis allowed for the engines to access the puzzle and perform their search until a mate with the expected X value was found. Once found the function would cease its execution and return the parameters the engine had used to find its solution to the ‘Mate in X’ puzzle.

This approach was found to more granularity for analytics as opposed to the Python-Chess engine limit protocol. By providing defined limits for depth of search and time searching, the engine would be tested multiple times to see at which combination of parameters it could determine a solution. The results of this process showed multiple combinations of depth and time that provided solutions for each problem but did not allow the engine to approach the problem in it’s base state. The change to an indefinite analysis pending a solution allowed each tested engine to approach the puzzle and return an exact combination without intervention by the tester.

The results from each search were returned by the engine as an information dictionary as dictated by each of their UCI protocols. Although these Python dictionaries had different formatting, they both contained the parameters of interest for this experiment being ‘time’, ‘seldepth’, and ‘nodes’. The ‘time’ is defined as the time until a solution was reached, ‘seldepth’ is defined as the depth at which the accepted solution was selected from, and ‘nodes’ are the number of branches searched by an engine in its approach to finding a solution.

Once all the above parameters were extracted from the engines during the analysis process, each distribution was compared to its corresponding parameters for each categorical ‘Mate in X’ group to provide a one to one experiment.

All statistical tests performed were test with an alpha = 0.05. This is an established value proposed by Fisher as a means to suggest a cutoff point where confidence could be attributed to a result (Fisher, 1954). This value is used conventionally in many types of research and is generally considered that 95% is an acceptable confidence level (Thiese & Ronna, 2016).

As distributions were being compared, the correct statistical tests had to be selected to ensure an accurate comparison of data was being performed. The choice of test was primarily driven by whether the data was likely to conform to a normal distribution or not, which required its own statistical test. The Shapiro-Wilk test was used to assess whether the Null Hypothesis that each of the distributions for the individual variables split by category were normally distributed.

The Shapiro-Wilk-test tests for normality in distributions by comparing the assuming the distribution would adhere to a normal curve and generating a test statistic based on the deviation from normality (SHAPIRO & WILK, 1965). The hypothesis for the test in this case was for all distributions being normal given a P-Value of 0.05. All results for this test showed that the distributions were not normally distributed meaning the comparative tests would need to be non-parametric.

The Mann-Whitney U test is a non-parametric comparative test that is considered as powerful as the parametric t-test (MacFarland & Yates, 2016) which is one of the most common tests used in research journals (Yim, et al., 2010). Due to it’s usefulness as a comparative test it was implemented into this analysis to compare equivalent engine parameters between engines within their respective ‘Mate in X’ categories.

The results of the Mann-Whitney U test were checked to determine whether the distributions were similar, and boxplots were used to plot the display the parameter values for visual comparison of quartiles and median values. The values for ‘seldepth’ were plotted with a normal range y axis of values and the ‘time’ and ‘nodes’ parameters plotted using a Log range y axis to allow the data to be read with granularity with the skewed nature of the values.

## 4.2 Opening Analysis

## 4.3 Rating distribution investigation

Section 2 - Analysis of Openings

Excluded those that ECO code not joined on as Lichess ads some descriptions to non-descript openings

Chapter 5: Implementation

# Chapter 6: Results

## 6.1 Puzzle Analysis

Table - Result of Shapiro-Wilk test for Stockfish Parameter Normality

|  |  |  |  |
| --- | --- | --- | --- |
| **Mate in X** | **Time (Statistic, P-Value)** | **SelDepth**  **(Statistic, P-Value)** | **Nodes**  **(Statistic, P-Value)** |
| 5 | (0.02768, <0.001) | (0.76472, <0.001) | (0.01799, <0.001) |
| 6 | (0.13556, <0.001) | (0.17055, <0.001) | (0.14541, <0.001) |
| 7 | (0.37828, <0.001) | (0.69211, <0.001) | (0.39509, <0.001) |
| 8 | (0.62401, <0.001) | (0.90733, 0.0027) | (0.62904, <0.001) |

Table - Result of Shapiro-Wilk test for Leela Chess Zero Parameter Normality

|  |  |  |  |
| --- | --- | --- | --- |
| **Mate in X** | **Time (Statistic, P-Value)** | **SelDepth**  **(Statistic, P-Value)** | **Nodes**  **(Statistic, P-Value)** |
| 5 | (0.258067, <0.001) | (0.57965, <0.001) | (0.14757, <0.001) |
| 6 | (0.32733, <0.001) | (0. 60204, <0.001) | (0.14512, <0.001) |
| 7 | (0.36677, <0.001) | (0. 71260, <0.001) | (0.27385, <0.001) |
| 8 | (0.44483, <0.001) | (0.74047, 0.0027) | (0.19305, <0.001) |

Table - Result of Whitney-Mann U test for Parameter Comparison

|  |  |  |  |
| --- | --- | --- | --- |
| **Mate in X** | **Time (Statistic, P-Value)** | **SelDepth**  **(Statistic, P-Value)** | **Nodes**  **(Statistic, P-Value)** |
| 5 | (175223, <0.001) | (1730160.5, <0.001) | (4333424, <0.001) |
| 6 | (7321, <0.001) | (92111.5, <0.001) | (198887.5, <0.001) |
| 7 | (21446, <0.001) | (12390.5, <0.001) | (27410.5, <0.001) |
| 8 | (345, <0.001) | (552.5, 0.0065) | (1232, <0.001) |

The Shapiro-Wilk test uses the Null Hypothesis that a distribution follows a normal distribution. The results of the Shapiro-Wilk test displayed in Table 1 for Stockfish parameter normality shows a consistent result of P-value < 0.05 wherein the Null Hypothesis can be rejected such that the data is considered non-normal. Non-normal distributions are further seen in Table 2, where a P-value < 0.05 indicates the distribution of values obtained from the engine analysis by Leela Chess Zero can be compared with Stockfish values using a non-parametric test.

With the results seen in Table 1 and Table 2indicating non-normal distribution of values, the Mann-Whitney U test was used. The results can be seen in Table 3 where a collection of test statistics and P-values are displayed for each engine parameter collected between the two chess engines for each of the categorical ‘Mate in X’ values. In these results there is no indication of any similarity between engine results for each category, parameter combination with all P-Values being less than the significance cut-off of 0.05.

# Chapter 7: Discussion

# Chapter 8: Conclusion

# Appendix A: Workflow

# Appendix B: Interview Transcripts

# Appendix C: Data Permissions

# Appendix D: Consent Forms

# Reference List