What makes a great chess player?

DS-GA 1001 Capstone Project (Date:20-12-2022)
Team AlphaZero
Shikhar Rastogi (sr6644) Chitvan Goyal (cg4174)

Introduction

Chess is one of the oldest known board games still played to this day. It is believed to have originated in northern India or eastern Iran in the 6th century. Being an abstract strategy game, it involves a lot of calculation, patience and creativity to become a good chess player. Since the pandemic and focus on chess in popular media (Queen's Gambit on Netflix), more and more people have started taking to chess. According to this 2022 UN report [1], there are more than 600 million people playing chess regularly!

Competitive chess is governed by FIDE (International Chess Federation) [2] which was founded in Paris, in 1924. They define the rules, conduct competitions, and most importantly - calculate the ELO ratings of each player and award titles such as Grandmaster or International Master to players. The ELO rating is a positive decimal number used to quantify the relative skill level of chess players. The Grandmaster (GM) title is the highest possible title usually awarded to people who cross the ELO rating of 2500, International Master (IM) title is next and is given to people who cross 2400 and so on, combined with some other criteria.

Given the abstract and mental nature of the game (as compared to baseball or basketball), chess data and analytics were never considered necessary in the past. However, with the chess boom and rising viewerships, professional chess players now compete for millions of dollars in online and in-person events. This has led to more chess analysts thinking increasingly about data and insights powered by newer and more advanced chess computers and engines. This project extends the same line of thought and aims to answer questions such as - what factors can have an impact on a chess player's future ratings, whether they can become a professional chess player and what this peak rating will be.

Dataset

Our dataset was collected from the official FIDE website[3] which is the only official source of chess ELO ratings. We scraped the monthly ratings from the website, starting from the earliest available month - January 2001 up to the latest, November 2022. However, as noted above, chess data was not considered very important earlier, and a lot of the data from 2001-2010 is inconsistent and even missing. While combining the data from each month, if data was missing, it is denoted as a NaN value. These rating values are either imputed using the previous/next valid rating value ('ffill'/'bfill') of the player for the purpose of this project. This is because using the median/mean would have an adverse effect for a time series and could destroy the trends of ratings growth of a chess player. Rows which have the birth year missing/impossible (any value <1900) are dropped, and if the number of games is missing, we assume they did not play that month and impute with the value of 0. All of this data is combined into a single dataframe which contains the FIDE ID as the row index, and the ratings and number of games played each month by that player in the columns. In all, after combining all data and removing the redundant rows, there were 392056 rows and 535 columns. There were two files available each month on the FIDE website described below -

Metadata

1. FIDE ID - This is the FIDE Identification number (FIN) assigned by FIDE as a unique identifier to each player recognized by them.
2. Name - Name of the person
3. Fed - Name of the local FIDE federation. This refers to the country the player is registered as a chess player in, stored as a 3 letter code. For example: 'USA', 'IND' etc.
4. Sex - This column takes two values - M if the player is male, and F is the player is female.

5. Tit - This refers to if the player has a title awarded by FIDE. It takes the values ‘GM’, ‘WGM’, ‘WFM’, ‘WIM’, ‘IM’, ‘FM’, ‘CM’.
6. B-day - This refers to the year the player was born. If the birth year is not present, it is marked as ‘0000’ in the file.
7. Flag - This marks whether the player is still active, i.e. they have played at least one rated game in the past year. This can take the value ‘i’ if the player is inactive, or ‘wi’ if a female player is ‘inactive’. The value is blank for male active players, and ‘w’ for female active players.

Monthly Ratings Data

1. ID number - This is the FIDE Identification number (FIN) assigned by FIDE as a unique identifier to each player recognized by them. (Serves as the joining key for us between the two files)
2. Name - Name of the person
3. Titl - This refers to if the player has a title awarded by FIDE. It takes the values ‘GM’, ‘WGM’, ‘WFM’, ‘WIM’, ‘IM’, ‘FM’, ‘CM’.
4. Fed - Name of the local FIDE federation. This refers to the country the player is registered as a chess player in.
5. {Month}{Year} - ELO rating of the player in the month of {Month} and the year of {Year}
6. Games - Number of official games played in the month of {Month} and year of {Year} by the player
7. Born - Year of birth
8. Flag - This marks whether the player is still active, i.e. they have played at least one rated game in the past year. Same values as above.

Inference: Factors affecting the rating of a chess player

Here we investigate whether the FIDE ELO rating of a chess player in November 2022 is affected by the number of games they played as a child i.e. before the age of 12.

Q1: Is a chess player’s rating higher if they played more games of chess in their childhood before the age of 12?

Our null and alternative hypothesis are defined as follows -

H_0: There is no effect on a player’s rating by the number of games they played in their childhood before the age of 12.
H_1: Players who played more games of chess in their childhood before the age of 12 have a higher current rating in November 2022 than those who played lesser games.

Approach

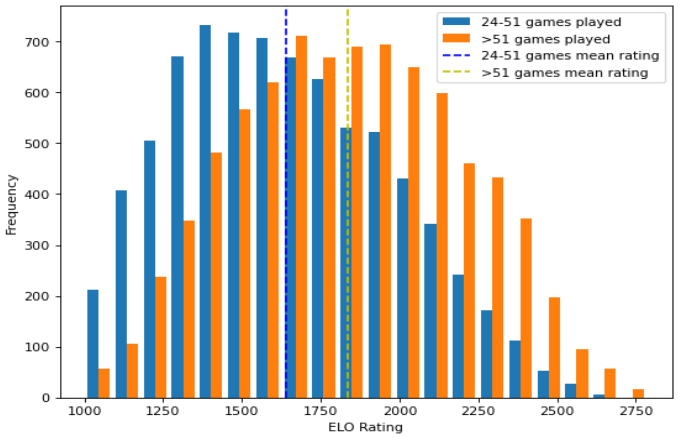
1) From the dataset, we first calculate the cumulative number of games played by each player before the age of 12.
2) We filter out any players whose year of birth is greater than 2005 - as they would still be under the age of 18 in 2022, and we typically want to investigate whether a chess player’s rating once they have matured, is affected by the number of games played in their childhood. We also filter out players whose year of birth is less than 1992, as data for their childhood years till the year 2000 is missing.
3) Next, we do a quartile split and pick players from the top two quartiles for our hypothesis test as the two samples.
4) We run a single tailed student’s independent t-test (after appropriate power analysis) to compare the November 2022 ratings of players in these two groups to test whether the rating of people who played more is higher with an alpha value of 0.005.

We choose an independent t-test as ELO rating is a continuous mathematical quantity which in fact has a direct relationship to probabilities. Next, we see that the sample size is fairly large and there were ~8000 players in each of the quartiles, and so we can assume that the means would distribute normally due to the central limit theorem. The variances in each of the samples were fairly similar as well, with the ratio of the larger variance to the smaller variance being 1.078. All these conditions make it favorable for us to

use the single tailed student's t-test. We decide to compare only the top 2 quartiles here, and the results are even more drastic if we compare with the lower quartiles.

Results and Conclusion

First from the power analysis for independent samples in a single-tailed t-test, for an alpha level of 0.05, with expected effect size of 0.5, and desired power level of 0.9 or 90%, we get the sample size needed as around 120 in each sample. This works for us as we have well over 8000 data points in each of the samples.



From the single-tailed t-test, we find that the test statistic is -35.29 and the p-value is $6.44*10^{-263}$. The Cohen's d effect size comes out to be 0.563.

As the p-value is much lesser than our defined alpha level of 0.005, we can safely reject the null hypothesis and conclude that the current ratings for players who played more games in their childhood is higher than those who played comparatively lesser games. We also see that the effect size is greater than 0.5 which means there is a more than medium effect. We have successfully found one important factor which affects a chess player's rating!

Clustering/Classification: Will they go on to become a professional chess player?

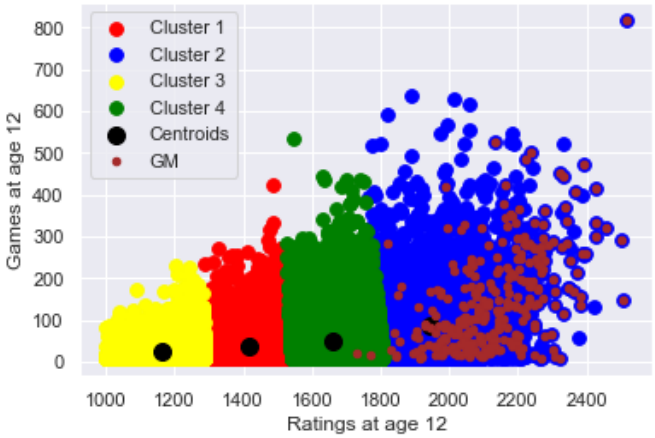
Q2.1: Clustering Analysis of the Relationship Between Childhood Chess Performance and Grandmaster Potential

Approach: From the data that is used to draw the inference that the player who played more games in childhood will have higher ratings when they grow up, Our analysis aims to investigate whether there is a relationship between the number of games played and ratings in childhood to the likelihood of becoming a Grandmaster (GM) or International Master (IM) in chess.

We performed Kmeans clustering to cluster the data based on their ratings and the number of games played. To decide the number of clusters we used the Elbow method where we infer if increasing the number of clusters is decreasing the sum of squared error from centroid. The number of optimal clustering is four in this case.

Results and Conclusion

Our analysis of the data reveals that 99.3% of players who eventually achieve the title of Grandmaster (GM) are part of cluster 4. This cluster consists of players who had high ratings at a young age, compared to players who had played a similar number of games but were not part of this cluster. These findings suggest that a player's potential to become a GM in the future can be discerned based on their performance at a young age. It is also evident that players who do not perform well in childhood may be less likely to achieve the GM title in the future.



Q2.2: Classification approach to predict if the player will go on to become professional player

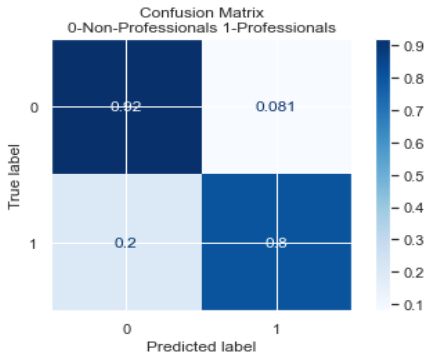
Approach:

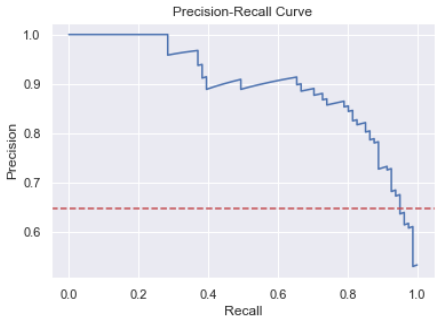
1) As per our inference in the previous question we filter out the players with age limit 12 and who were born between 1992 and 2005. We have filtered out people born after 2005 because their ratings will still not have been closer to their peak ratings.
2) We consider the monthly ratings of players in childhood, total games played till the age of 12 and Country name to predict whether the player will become a professional player in future.
3) We consider the player who achieved the Grandmaster or International master title later in life as a professional player.After sampling 20% players from the non professional group we have we have considered a total of 2607 observations.
4) On the sampled data we have used a histogram based gradient boosting algorithm to predict the player status in future.

Results and Conclusion

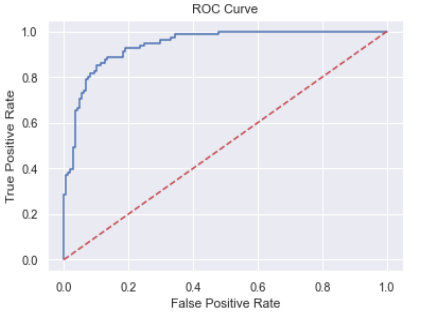
Based on the results, the classification model has achieved a total precision of 0.87, recall of 0.86 and f1 score of 0.86. In addition, when using cross validation the model has achieved a mean score of 0.83 which indicates that it has a good generalization performance. Area under PR/ROC curve is greater than 0.89, this suggests that there is good balance between precision and recall and the model is mostly able to identify professional and non-professional players.

Classification Report	Precision	Recall	F1-Score	Support
Non-Professional Players	0.89	0.92	0.91	148
Professional Players	0.84	0.8	0.82	81
Accuracy			0.88	229
Macro Average	0.87	0.86	0.86	229
Weighted Average	0.88	0.88	0.88	229





Area under PR Curve is: 0.8966913783435957



Area under ROC Curve (AUROC): 0.9421921921921922

Prediction -

Q3: Can we predict the ratings of professional chess players in November 2022 using their ratings from their childhood, while controlling for confounds such as the country they belong to, the number of games they played in their childhood and their year of birth?

In the previous sections, we have discovered a few good indicators to predict the trajectory of a chess player's career - starting from what impacts their ratings, to whether they turn professional or not. Now, we turn to these professional players and using their childhood statistics, we will predict their future ratings.

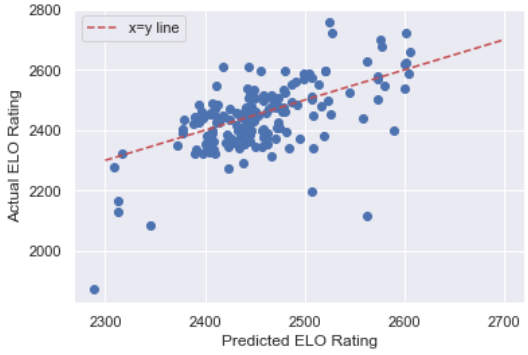
Approach:

1) In a similar way as the last two questions, we first filter out professional players (IM/GM) whose year of birth is greater than 2005 or lesser than 1992 for the same reasons as above. This is even more crucial for this question, as not filtering out very young players born after 2005 could be a potential data leakage. If we use their rating in October 2022 for example, to predict their November 2022 - they will almost always be the same, and this will not be a useful model.
2) Our most important features here are of course the childhood ratings for the players. For the months that we do not have ratings data, we impute the NaN values from the previous/next valid rating of the player, whichever is closer.
3) We then adjust for confounding variables by considering them as features for the model. These include cumulative games played till the age of 12, the country the player resides in (encoded as a one hot vector), and the year of birth of the player (also encoded as a one hot vector). We cannot treat the country and years of birth as a numeric variable given that it does not make sense to add/divide them.
4) After our filtering process, we are left with 858 data points, which we split into a train and test set of 686 and 172 respectively - test split of 20%. The training data is then fed into the model and cross validated to figure out the appropriate hyperparameters.

We impute using the neighboring ELO ratings for a player instead of using column wise or row wise median/mean as this is a time series dataset. Taking a row-wise mean could lead to many points which don't follow the trend and reduce the performance of the model. We pick the RandomForestRegressor model as it is interpretable and relatively fast to train.

Results and Conclusion

We find that the RMSE (Root Mean Square Error) value for our predictions for the ratings in November 2022 are 92.15. However, this is not very interpretable as it does not give us any idea about the magnitude of the errors. For that reason, we compute another metric called the Mean Absolute Percentage Error (MAPE), which is a mean of all the absolute percentage errors for each data point in the test set. Our MAPE value turned out to be around 2.74%. This means that on average, our predictions were off from the target by around 2.74%. We found that the approach of using the HistGradientBoosting based model decreased performance here and brought up the RMSE to close to 102.76, and hence chose the RF model.



The optimal model parameters which produced this result were found through cross validation as max_depth of 175, min_samples_split of 0.26 and n_estimators of 547. We find that the past ratings combined have a feature importance of around 0.88, and the country has a feature importance of around 0.10.

Here is a visualization of the regression results on the test set, and the tabular results:

Model Type	RMSE	MAPE
Random Forest Regressor	92.15	2.74%
Hist Gradient Boosting Regressor	102.76	3.20%

Conclusion -

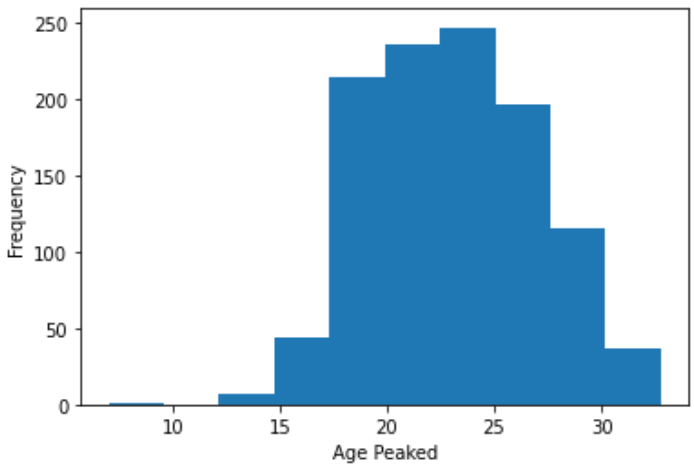
1. Our top conclusion which is supported by all three questions is that a chess player's future ratings and career trajectory are deeply influenced by their childhood practice - whether they play a lot of games, and whether they have a steady increase in rating during their childhood. This is clearly visible with the statistically significant t-test for cumulative games played under the age of 12, and then the career classification, future ratings prediction using past ratings, with a high feature importance of 0.88 for the prediction problem.
2. From questions 2 and 3, we also see that along with just the number of games played/results in these games, the country the player belongs to is an important factor which determines whether they succeed in their chess career or not. They had a feature importance of close to 0.10 for the prediction question. This could have several underlying reasons - such as availability of resources for chess, more tournaments, or a better chess culture in the country.

Assumptions and Limitations -

1. The biggest limitation to the project was not having a complete dataset to be able to make predictions. FIDE has not consolidated and published any ratings before the year of 2000, and started maintaining monthly ratings in a proper manner only around 2010. There are many holes in the data between 2000 and 2010, which led us to do a lot of interpolation. We were forced to filter out players born before 1992 as there was insufficient data about their childhood. So our models and results currently would only really be fit for children born in the late 90s or early 2000s. They would not generalize well to other generations.
2. One big assumption we made was that 'childhood' for a chess player is limited till age 12. This was partly done due to the limited data we had in terms of the date range data is available. If we assumed any younger, we would not have enough data points, and any older, then most predictions would not be as impressive and useful.
3. We were not able to adjust for all of the confounds as there was no way to measure them. The biggest example for this would be a chess player's socioeconomic status in their childhood. This data point could have a tremendous impact on their ratings or chess career especially if there is a sudden downturn in their economic status.
4. In an ideal world, this data could potentially be joined with census data to give us a lot more insight about how a player lives and trains. Indicators like per-capita GDP are too generic and entirely correlated with country and are not appropriate here.

Extra Credit

We attempt to find the age most professional players peak in their chess careers. For each professional player, we figure out at what age they achieved their highest ever rating, and create a histogram. We find that for most players this is around 22-23. We can see in the diagram that there is a peak at around 22-23. The mean for the peak rating age is 23.07.



References

1. World Chess Day Jul 22 United Nations. United Nations. Available at: https://www.un.org/en/observances/world-chess-day#:~:text=Under%20initiative%20of%20FIDE%2C%20July,around%20the%20world%20since%201966.(Accessed: December 20, 2022).
2. FIDE: https://www.fide.com
3. https://ratings.fide.com/download.phtml?period={0}
4. https://scikit-learn.org/stable/
5. https://docs.scipy.org/doc/scipy/