Chess ELO Prediction

Enes Koç – Ege Aydın – Arnisa Fazla

Furkan Kerim Çabaş – İlter Onat Korkmaz

Finding ELO

- Goal: To predict ELO rating of a chess player given a game
- Previous works: A Kaggle competition called “Finding ELO”
- Comparison with: The results of the first place in that competition

> “This competition challenges Kagglers to determine players' FIDE Elo ratings at the time a game is played, based solely on the moves in one game.
> - Do a player's moves reflect their absolute skill?
> - Does the opponent matter?
> - How closely does one game reflect intrinsic ability?
> - How well can an algorithm do?
> - Does computational horsepower increase accuracy? Let's find out!”

Dataset

- Kaggle Competition: Finding Elo (25k analyzed games) [1]
- Lichess: Standard rated games for September 2014 (1M games)

```
Data:
[Event "Rated Blitz game"]
[Result "1-0"]
[WhiteElo "1600"]
[BlackElo "1658"]
[ECO "A00"]
[Opening "Polish Opening: Czech Defense"]

SAN: 1. b4 e5 2. Bb2 d6 3. c3 Bf5 4. d3 Nf6 5. e4 Bg6 6. Be2
Be7...
UCI: b2b4 e7e5 c1b2 d7d6 c2c3 c8f5 d2d3 g8f6 e2e4 f5g6 f1e2
f8e7...
```

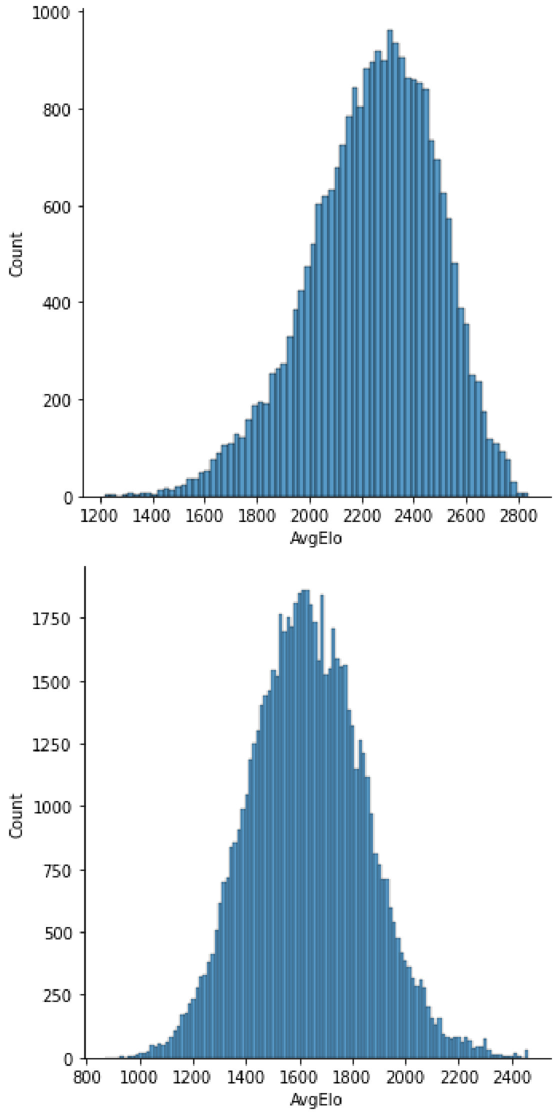


Figure 1: ELO distribution of Lichess Dataset (bottom) and of a subset of Kaggle Dataset (top).

Features - ECO

- ECO: Encyclopedia of Chess Openings A00-A99, B00-B99, … E00-E99
- One hot encoding is used for 13 opening type

A [edit]

- White first moves other than 1.e4, 1.d4 (A00–A39)
- 1.d4 without 1...d5, 1...Nf6 or 1...f5: Atypical replies to 1.d4 (A40–A44)
- 1.d4 Nf6 without 2.c4: Atypical replies to 1...Nf6 (A45–A49)
- 1.d4 Nf6 2.c4 without 2...e6, 2...g6: Atypical Indian systems (A50–A79)
- 1.d4 f5: Dutch Defence (A80–A99)

B [edit]

- 1.e4 without 1...c6, 1...c5, 1...e6, 1...e5 (B00–B09)
- 1.e4 c6: Caro–Kann Defence (B10–B19)
- 1.e4 c5: Sicilian Defence (B20–B99)

C [edit]

- 1.e4 e6: French Defence (C00–C19)
- 1.e4 e5: Double King Pawn games (C20–C99)

D [edit]

- 1.d4 d5: Double Queen Pawn games (D00–D69)
- 1.d4 Nf6 2.c4 g6 with 3...d5: Grünfeld Defence (D70–D99)

E [edit]

- 1.d4 Nf6 2.c4 e6: Indian systems with ...e6 (E00–E59)
- 1.d4 Nf6 2.c4 g6 without 3...d5: Indian systems with ...g6 (except Grünfeld) (E60–E99)

Figure 2: 13 main ECO interval.

Stockfish

- Stockfish engine is used to analyze 300k games with depth=8 and sometimes depth=10
- It took 160 hours in total to analyze and extract features
- For better results, 20+ depth should have been used
- It would take a few years to proceed all data with our CPU power

Depth d_1	$d_1 - 20$	$(d_1 - 20) \times \eta_{(20)}$	Strength (Elo)	95% conf. int.
20	0	0	2894	[2859, 2929]
19	-1	-66	2828	[2786, 2868]
18	-2	-133	2761	[2714, 2807]
17	-3	-199	2695	[2642, 2745]
16	-4	-265	2629	[2570, 2684]
15	-5	-331	2563	[2498, 2623]
14	-6	-398	2496	[2426, 2562]
13	-7	-464	2430	[2354, 2500]
12	-8	-530	2364	[2282, 2439]
11	-9	-596	2298	[2209, 2378]
10	-10	-663	2231	[2137, 2317]
9	-11	-729	2165	[2065, 2255]
8	-12	-795	2099	[1993, 2194]
7	-13	-861	2033	[1921, 2133]
6	-14	-928	1966	[1849, 2071]

Figure 3: Estimated strength of the engine at different search depths [2].

Features – Move Scores

- Stockfish chess engine

CP:	16,	71,	59,	211,	-8,
WDL:	0.524,	0.623,	0.598,	0.923,	0.487,
	White,	Black,	White,	Black,	White,

- Each game has different move count Series data
- Each series data is divided to 5 parts and their statistics such as min, max, mean, median, std, etc. are used as features

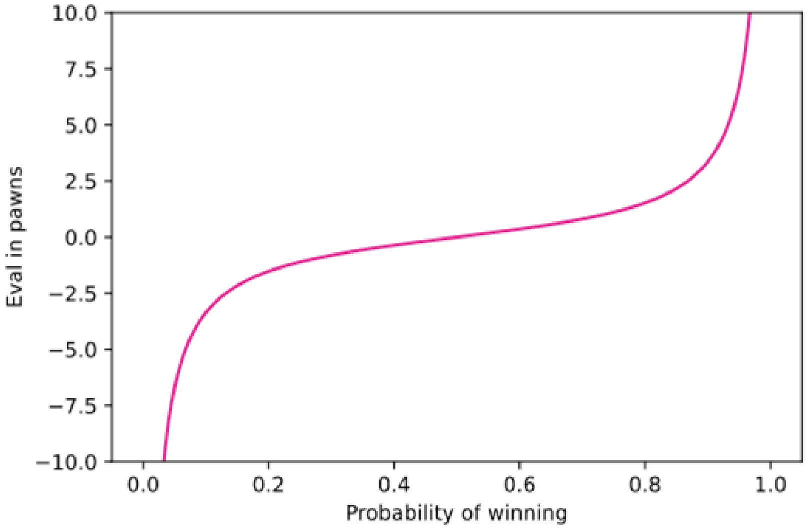


Figure 4: Centipawn evaluation versus winning probability. [3]

Features – Game Scores

- Game scores
  - `[queen_moved_at, queen_changed_at, promotion_count, total_checks, first_check_at]`
  - Normalized with respect to total number of moves
- Best Move and their score
  - First 5 best move for every board state
  - Series data

```
[0, 1, 0, 0, 0],
[0, 1, 0, 0, 0],
[0, 0, 0, 0, 0],
[1, 0, 0, 0, 0],
[1, 0, 0, 0, 0],
[0, 1, 0, 0, 0],
[0, 1, 0, 0, 0],
[0, 0, 0, 1, 0],
[0, 1, 0, 0, 0],
[0, 0, 0, 0, 0],
[1, 0, 0, 0, 0],
[1, 0, 0, 0, 0],
```

```
[   7,    23,    26,    27,    67],
[  15,    15,     5,    -3,   -18],
[  14,    14,    27,    46,    72],
[  44,     9,    -8,    -8,   -16],
[  13,    26,    43,    48,    57],
[  80,    42,    25,    12,     6],
[  77,   101,   151,   362,   443],
[  90,    63,    40,    39,    35],
[  43,    55,    66,    82,    95],
[  89,    75,    55,    45,    31],
[  36,    48,    55,    57,   112],
[  53,    50,    48,    23,    22],
[   7,    30,    34,    49,    62],
[  25,    17,  -113,  -137,  -140],
[  22,    50,    87,    93,    96],
[  99,    22,   -35,   -94,  -185],
```

Figure 5: Best moves (left) and best moves' scores (right).

Uniformization of the Data

- Non-uniform data causes underfitting at the edges
- A uniform subset of lichess data is used

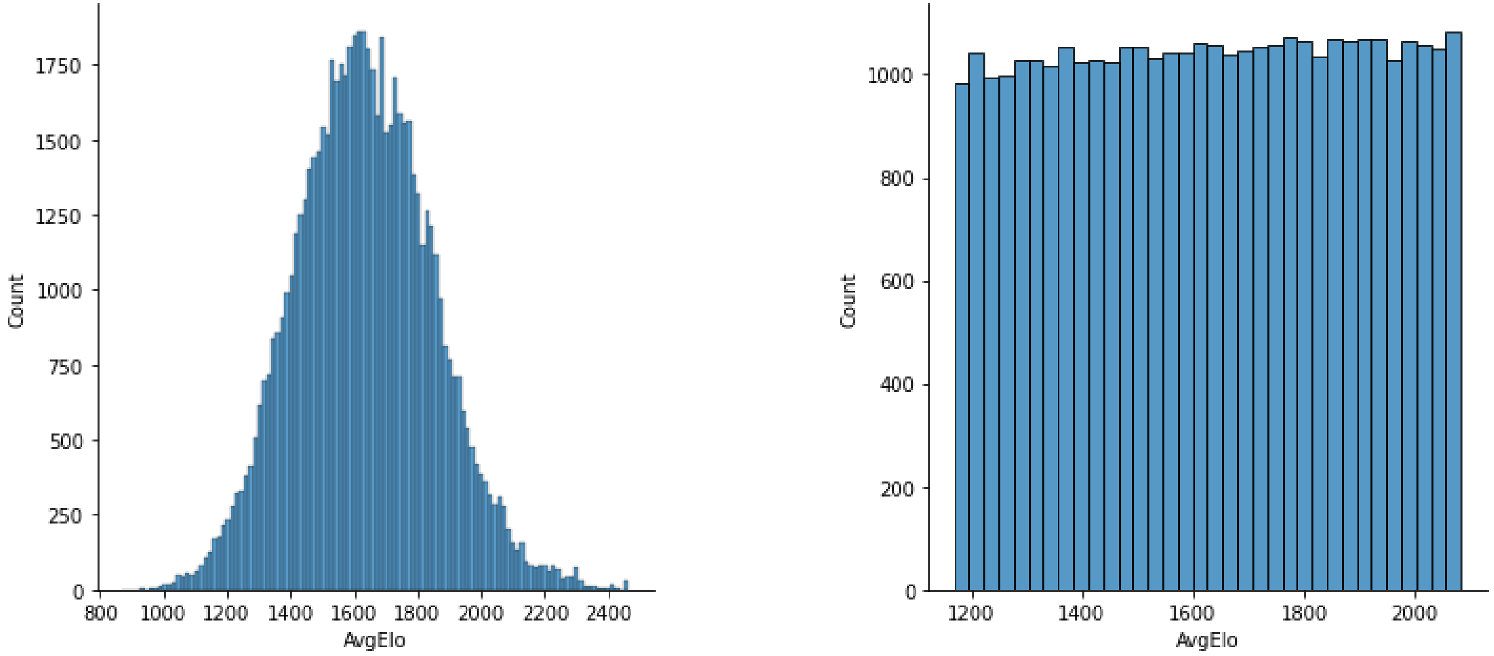


Figure 6: ELO distribution of randomly sampled lichess dataset
and uniformly picked lichess dataset

Principal Component Analysis

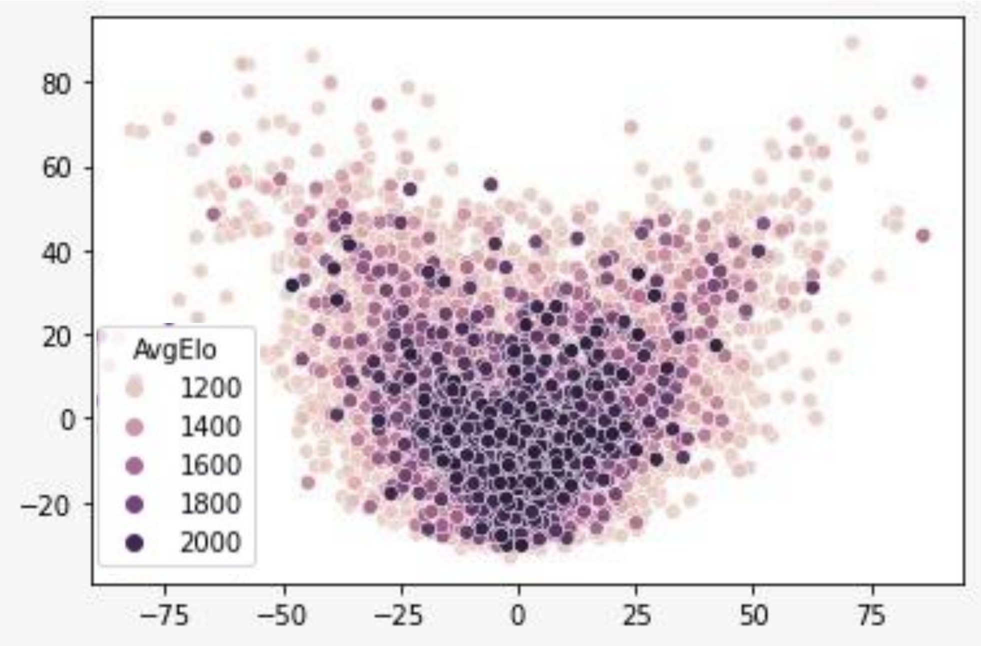


Figure 7: PCA plot of our feature set

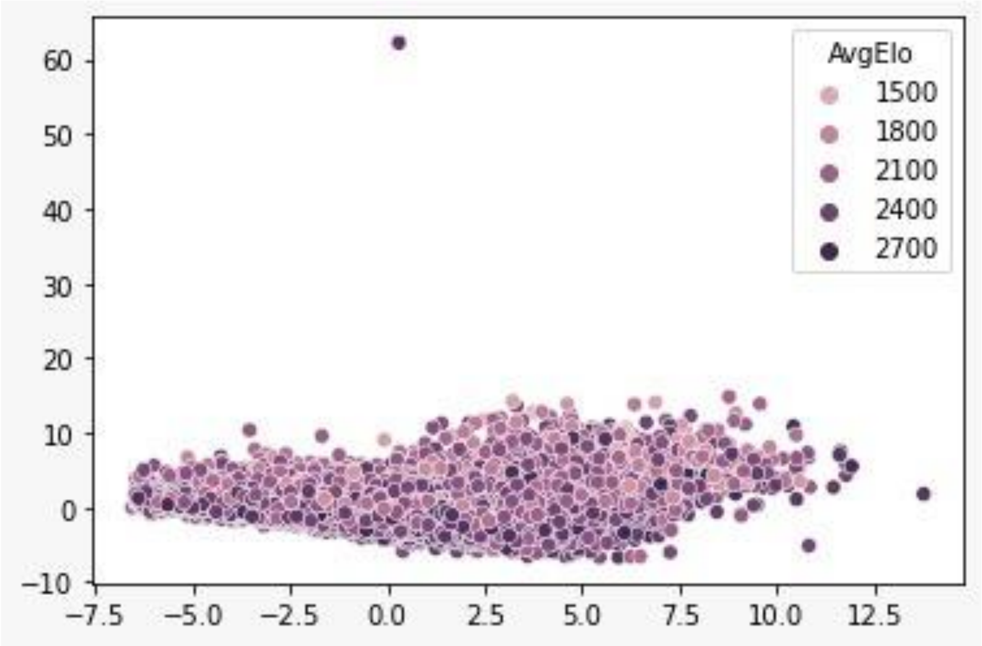


Figure 8: PCA plot of their feature set

Neural Networks

- 2800 Input, 2 Outputs: Average and Difference of ELOs
- 2 Hidden Layers: 400 and 28 Neurons
- L1 and L2 kernel regularization
- Dropout method
- Batch Normalization
- Activation Methods: Leaky Relu, Linear
- Loss: Mean Square Error (MSE)
- Optimizer: ADAM

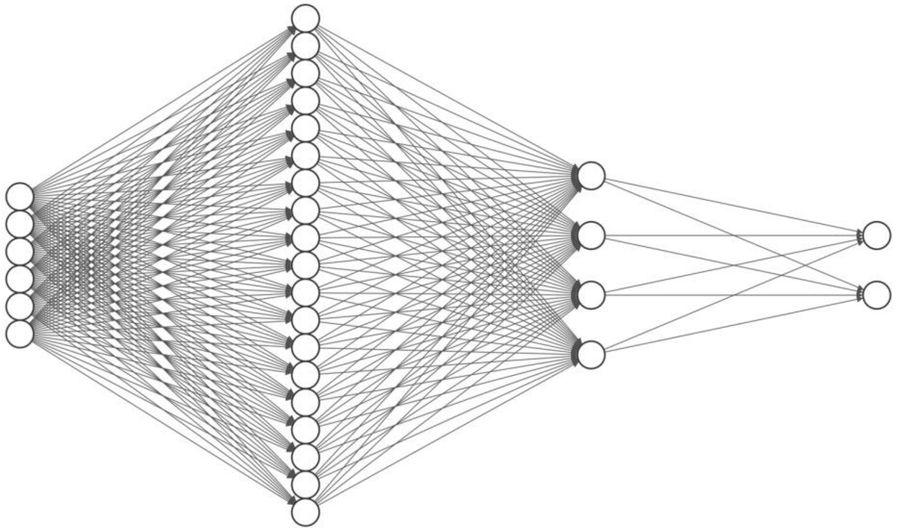


Figure 9: Neural Network Architecture.

Regression result

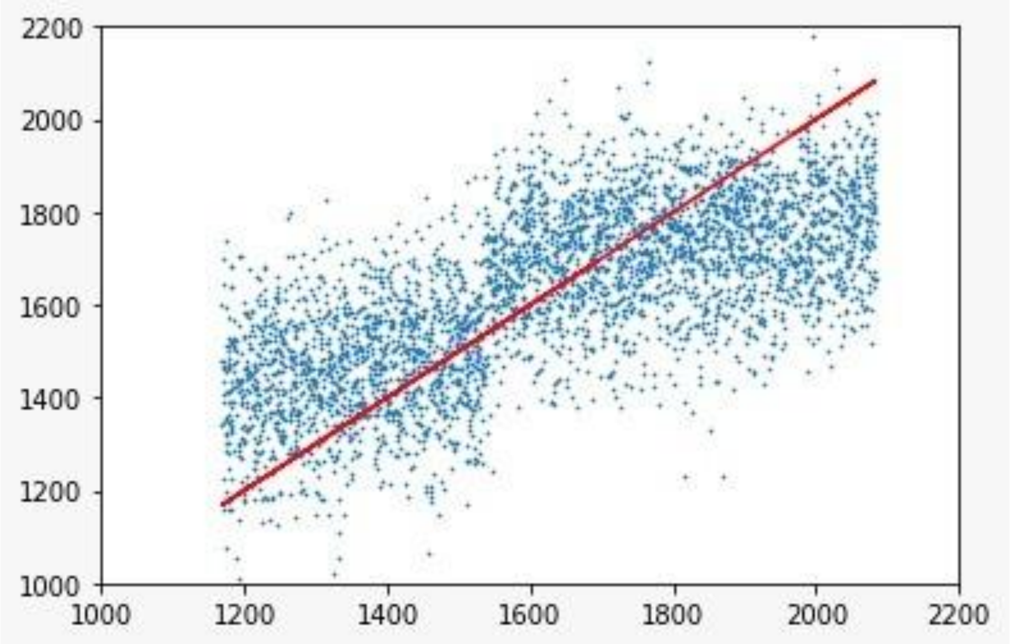


Figure 10: The predicted ELO versus original ELO scores.

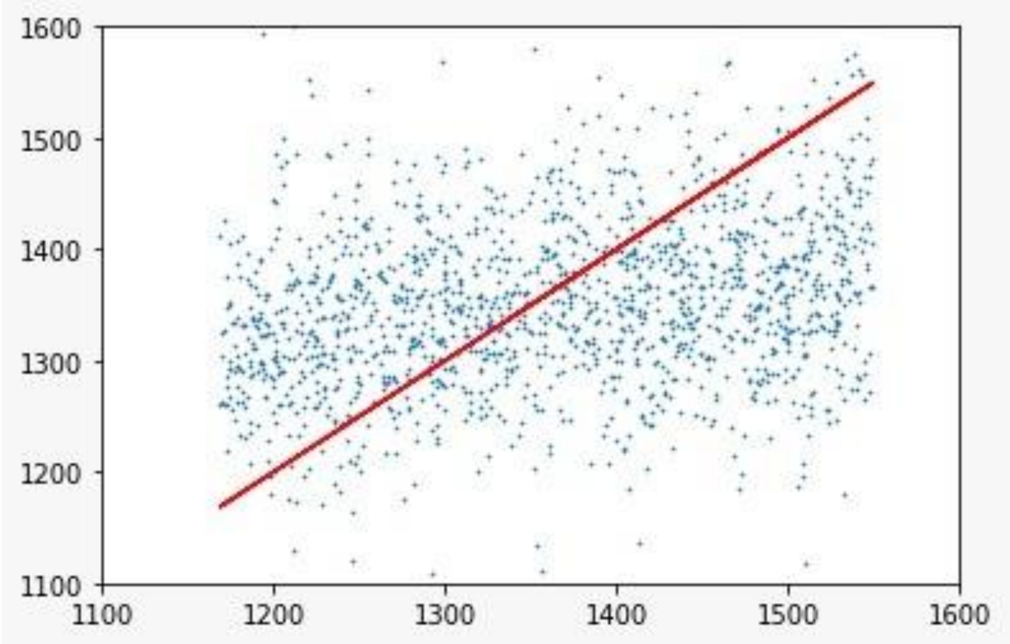
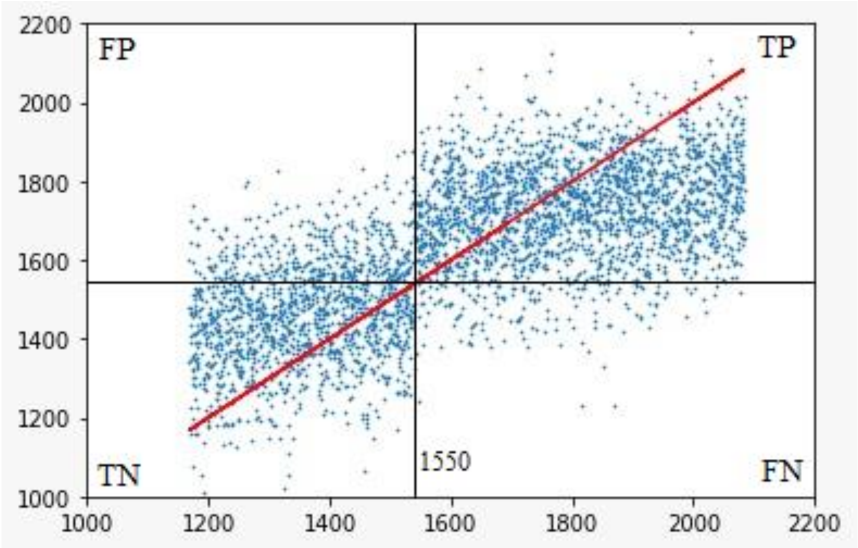


Figure 11: The predicted ELO versus original ELO scores that are lower than 1550.

Conclusion

- Not enough CPU power for a good Stockfish analysis
- Move scores are not well correlated with ELO
- Binary Classification:
  - The model discerned the ELO>1550 with ≈ 87% accuracy.



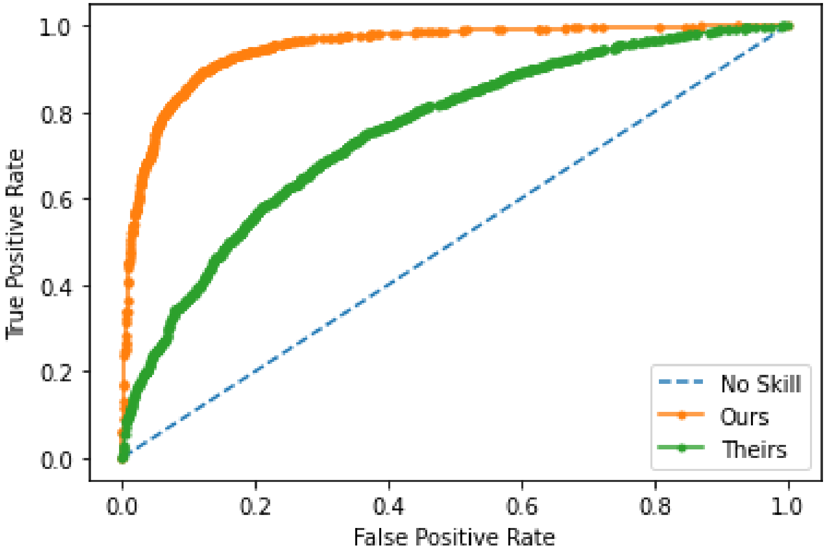
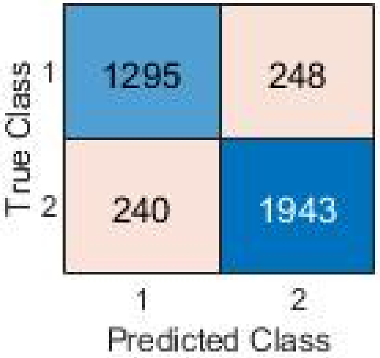


Figure 12: AUROC Curves and binary classification over regression

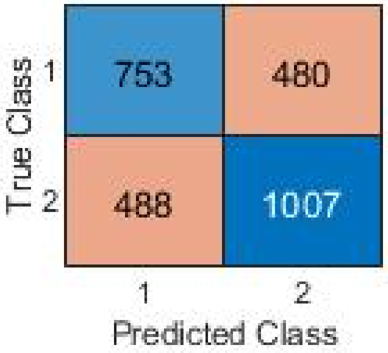
Conclusion

Our results:



- Precision = 1295 / (1295 + 248) ~= 0.839
- Recall = 1295 / (1295 + 240) ~= 0.844
- F1-Measure = $F = \frac{2PR}{P+R}$ ~= 0.844
- Accuracy = (1295 + 1946) / N = 0.87

Their results:



- Precision = 753 / (753 + 480) ~= 0.611
- Recall = 753 / (753 + 488) ~= 0.601
- F1-Measure = $F = \frac{2PR}{P+R}$ ~= 0.606
- Accuracy = (753 + 1007) / N = 0.645

References

[1] “Finding Elo.” Kaggle. https://www.kaggle.com/c/finding-elo. (accessed May 3, 2021).

[2] Ferreira, Diogo. (2013). The impact of search depth on chess playing strength. ICGA Journal. 36. 67-80. 10.3233/ICG-2013-36202.

[3] Crem. “Win-Draw-Loss evaluation.” Leela Chess Zero. Win-Draw-Loss evaluation - Leela Chess Zero (lczero.org) (accessed May 3, 2021).