

Предсказание результатов шахматной
партии между профессионалами с
помощью алгоритмов машинного
обучения

Содержание

1	Введение	2
2	Обзор литературы	2
3	Работа	3
3.1	Выбор формата входных и выходных данных	3
3.2	Создание датасета	5
3.3	Выбор моделей	5
3.4	Оптимизация гиперпараметров	6
3.5	Тестирование	7
4	Выводы и заключение	9
5	Список литературы	10

1 Введение

Цель работы - научиться предсказывать вероятности каждого из трех возможных исходов (победа белых, ничья, победа черных) в шахматной партии, основываясь на актуальных данных об игроках. При этом исследование касается только игр, где хотя бы один из соперников - гроссмейстер, поскольку именно такие шахматы можно назвать профессиональными. В других случаях значительно возрастает роль случайности.

В шахматах активно используется система Эло рейтингов. Предполагается, что на основании рейтингов игроков можно с высокой точностью определять исход партий, однако в реальности качество таких предсказаний около 46%. Одной из задач работы будет обогнать подобные предсказания.

2 Обзор литературы

В предыдущих работах на тему предсказания спортивных результатах были получены модели с точностью предсказания от 50% до 75%. Для футбола качество получается выше, поскольку ничья в этой игре довольно маловероятна (в турнирах между сильными командами ничьих около 25%, в других случаях бывает меньше 20%).

В работах по шахматам авторы ставили целью предсказание результата партии при том допущении, что уже известно несколько первых ходов. Такой подход выглядит довольно бесполезным, ведь интересно узнать прогноз на будущую партию, а не предсказать заведомо известный результат уже прошедшей.

Целью этого исследования будет предсказание результатов с использованием только тех данных, какие есть еще до начала партии. Необходимо получить качество модели, сопоставимое с указанным выше.

3 Работа

3.1 Выбор формата входных и выходных данных

В качестве входных для модели использовались следующие показатели:

1. *Доля очков, набранных первым за все время противостояния двух игроков*
2. *Доля побед белых в противостоянии этих двух игроков*
3. *Доля ничьих в противостоянии этих двух игроков*
4. *Доля побед черных в противостоянии этих двух игроков*
5. *Ожидаемое количество очков, которые наберут в партии белые, исходя из рейтингов оппонентов*
6. *Ожидаемое количество очков, которые наберут в партии белые, исходя из перфомансов оппонентов*
7. *Ожидаемое количество очков, которые наберут в партии белые, исходя из максимальных за карьеры рейтингов оппонентов*
8. *Результат первого игрока в прошлой игре*
9. *Результат второго игрока во второй игре*
10. *Среднее количество очков, которое первый игрок набирает, играя за белых*
11. *Среднее количество очков, которое второй игрок набирает, играя за черных*
12. *Среднее количество очков, которое первый игрок набирает, играя против шахматистов с рейтингом ниже(выше) его*
13. *Среднее количество очков, которое второй игрок набирает, играя против шахматистов с рейтингом ниже(выше) его*
14. *Среднее количество очков, которое набирает первый игрок после такого же результата, какой был в его последней партии*
15. *Среднее количество очков, которое набирает второй игрок после такого же результата, какой был в его последней партии*

16. *Среднее количество очков, которое набирает первый*
17. *Среднее количество очков, которое набирает второй*
18. *Доля поражения первого игрока*
19. *Доля поражений второго игрока*
20. *Доля ничьих у первого игрока*
21. *Доля ничьих у второго игрока*
22. *Доля побед у первого игрока*
23. *Доля побед у второго игрока*
24. *Длина последней игры первого игрока*
25. *Длина последней игры второго игрока*
26. *Количество дней, прошедших с последней игры первого игрока*
27. *Количество дней, прошедших с последней игры второго игрока*
28. *Возраст первого игрока*
29. *Возраст второго игрока*
30. *Номер раунда турнира*
31. *Номер раунда / количество раундов*
32. *Среднее количество очков, которое первый игрок набирает в этом турнире*
33. *Среднее количество очков, которое второй игрок набирает в этом турнире*

Формат выходных данных был выбран следующим:

1. *Вероятность победы черных*
2. *Вероятность ничьей*
3. *Вероятность победы белых*

3.2 Создание датасета

В качестве материала для входных данных использовалась база партий Mega Database 2021 от Chessbase. Поскольку для обоих игроков в партии нужно знать прошлую статистику выступлений, для начала были отобраны все игры, начиная с 2010 года, в которых оба спортсмена имели эло рейтинг FIDE не меньше 2100. В самом датасете использовались партии, в которых эло рейтинг обоих игроков был не меньше 2300, при этом хотя бы один из них имел рейтинг больше 2500. Таким образом получаем партии гарантированно высокого качества.

Вся статистика, касающаяся последних партий игроков, обсчитывается на основании последних 10 партий для каждого шахматиста, если не указано другое.

Для выходных данных можно указать, что вероятность реального исхода равна 1, а других, соответственно, 0.

Также необходимо сбалансировать датасет - должно быть поровну примеров с каждым из исходов.

3.3 Выбор моделей

Для исследования были выбраны следующие алгоритмы машинного обучения, реализованные в модуле scikit-learn для python:

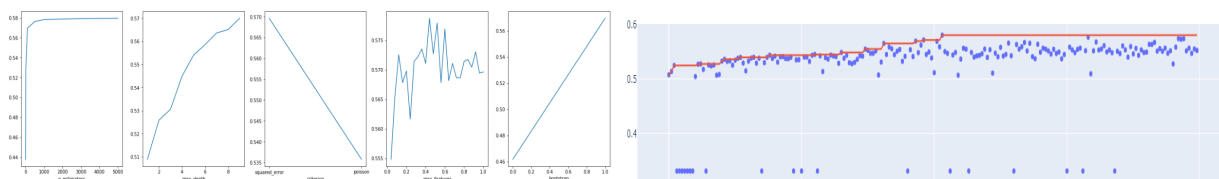
- Random Forest Regressor
- Multi-layer Perceptron regressor
- K-Neighbors Regressor
- Elastic Net (Linear regression)

3.4 Оптимизация гиперпараметров

Оптимизация гиперпараметров для каждой модели выполнялась двумя способами - с помощью простого последовательного перебора значений и с помощью модуля `optuna`. В обоих случаях задачей была максимизация точности. Исход считался угаданным, если модель посчитала его наиболее вероятным. Результаты представлены ниже.

RandomForestRegressor

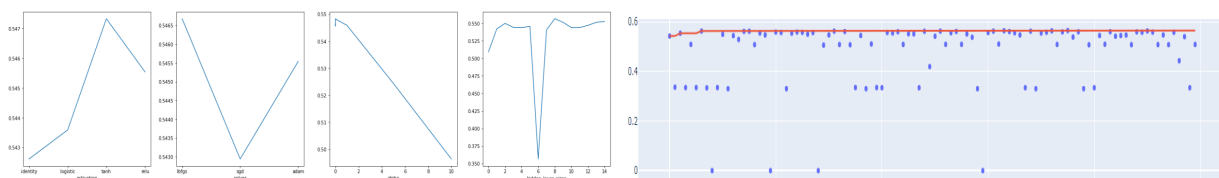
Первым методом мы подбирали такие параметры как количество деревьев, ограничение на максимальную глубину деревьев, критерий расщепления, число признаков для выбора расщепления и использование бутстрэпа. В случае с `optuna` добавились ограничение на число объектов в листьях и ограничение на минимальное число объектов, при котором выполняется расщепление. Качество моделей получилось равным 0.57 и 0.58 соответственно.



MLPRegressor

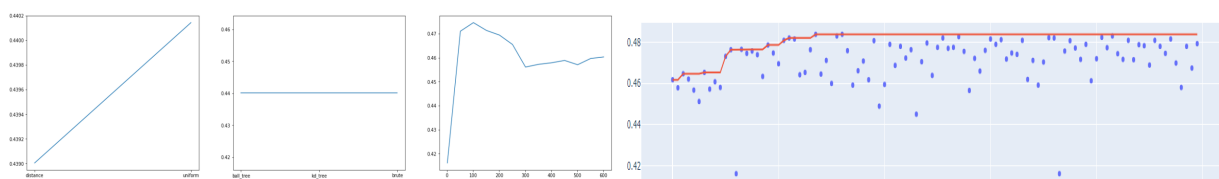
В переборе работа шла с функцией активации, решающей функцией, аргументом `alpha` и конфигурациями скрытых слоев*. Точность моделей после оптимизации составила 0.556 и 0.56 соответственно.

**Использовались следующие конфигурации скрытых слоев: (4), (8), (16), (32), (64), (100), (4, 4), (8, 8), (16, 16), (32, 32), (64, 64), (8, 8, 8), (16, 16, 16), (32, 32, 32)*



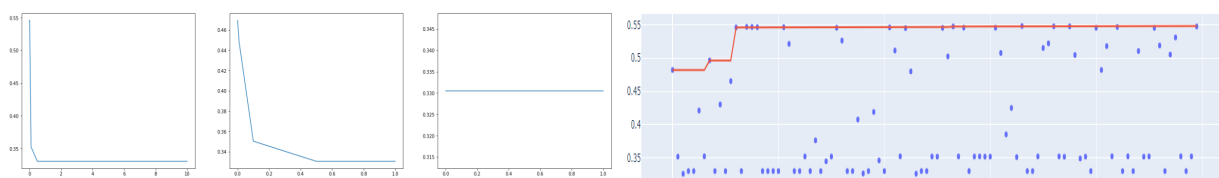
KNeighborsRegressor

Для этой модели перебирались функция весов, алгоритм решения и количество соседей. Качество моделей после оптимизации равно 0.475 и 0.484.



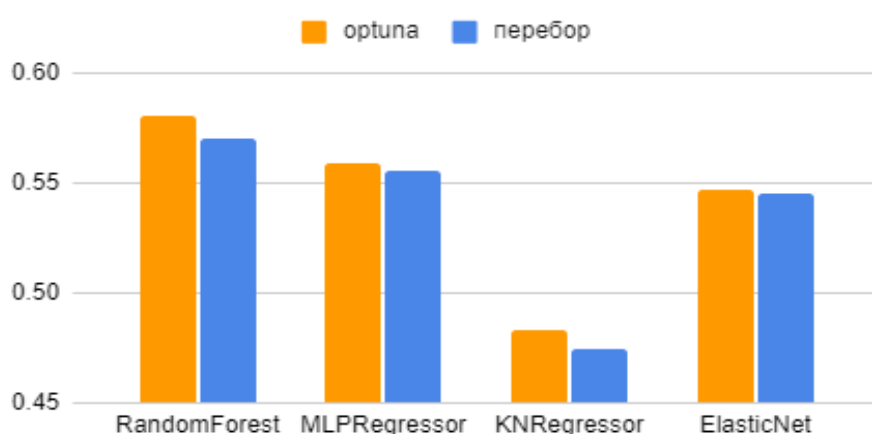
ElasticNet

В переборе использовались параметр регуляризации, аргумент alpha и параметр positive, касающийся ограничения на знаки коэффициентов модели. Точность после оптимизации составила 0.545 и 0.548.



3.5 Тестирование

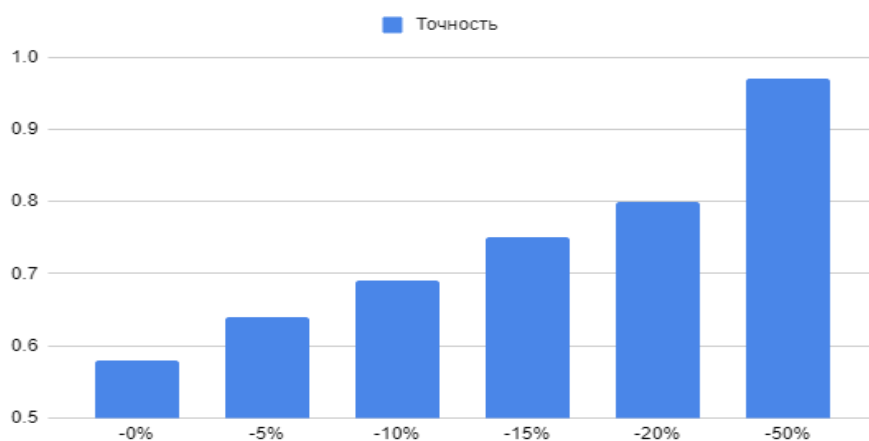
Таким образом, имеем следующие результаты:



Наилучшее качество было получено при использовании случайного леса и подборе гиперпараметров с помощью optuna - 58%, то есть с такой вероятностью случится исход, указанный моделью как наиболее вероятный. Рассмотрим статистику по всем исходам в отдельности. Можно видеть, что самая низкая точность у предсказания ничьих, но и она около 50%. Кроме того, поскольку мы предсказываем вероятность каждого исхода, важно посмотреть,



насколько вероятность реального исхода отклоняется от вероятности исхода, предсказанного моделью. Следующий график показывает точность предска-



зания при допустимом отклонении вероятности настоящего исхода от наиболее вероятного исхода. Можно увидеть, что даже если допустить отклонение вероятности на 10%, точность повышается до 70%.

4 Выводы и заключение

Мы научились предсказывать результат шахматной партии с точностью 58% при использовании алгоритмов машинного обучения. Наше качество получилось явно выше того, которое можно получить путем предсказаний, основанных на системе рейтингов. Также удалось обогнать в точности некоторые прошлые работы в этой сфере. При этом использовались лишь данные, известные до начала рассматриваемой игры, что является ключевым отличием от предыдущих исследований по шахматам.

Таким образом, цель работы была достигнута.

5 Список литературы

- [1] Hector Apolo Rosales Pulido; Predicting the Outcome of a Chess Game by Statistical and Machine Learning techniques. - 2016
- [2] Corentin HERBINET; Predicting Football Results Using Machine Learning Techniques. - 2018
- [3] Rafal Drezewski, Grzegorz Wator; Chess as Sequential Data in a Chess Match Outcome Prediction Using Deep Learning with Various Chessboard Representations. - 2021
- [4] Roland Shum; Neural Networks Football Result Prediction. - 2020
- [5] Модуль scikit-learn