МИНОБРНАУКИ РОССИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Южный Федеральный Университет»
Институт высоких технологий и пьезотехники

Кафедра прикладной информатики и инноватики

Направление: 09.03.03 "Прикладная информатика"

Отчет по дисциплине

«Большие данные»

Проект: «Сбор, предобработка и анализ данных о шахматах»

Выполнили студенты 3 курса 2_ВТ-09.03.03.01-о3 группы:

Руденко А.Д.

Соколов А.Д.

Ростов-на-Дону – 2024

Оглавление

Постановка задачи

Шахматы на сегодняшний день являются широко распространенным видом досуга, а также популярным и признанным видом спорта. Также, игра имеет широкую базу для анализа возможных событий.
Целью нашего проекта является визуализация данных и их последующий анализ, а также обучение модели для возможности предсказания исхода партии.
Данный проект будет полезен как шахматистам любителям, так и опытным игрокам.

Описание датасета

Нами был взят датасет из следующего источника:
https://www.kaggle.com/datasets/datasnaek/chess
Он представляет собой более 20 тысяч записей о различных партиях, сыгранных на сайте lichess.org в течение года.
Датасет содержит различные данные, такие как количество ходов, сами ходы, дебюты, время, рейтинг игроков и пр.

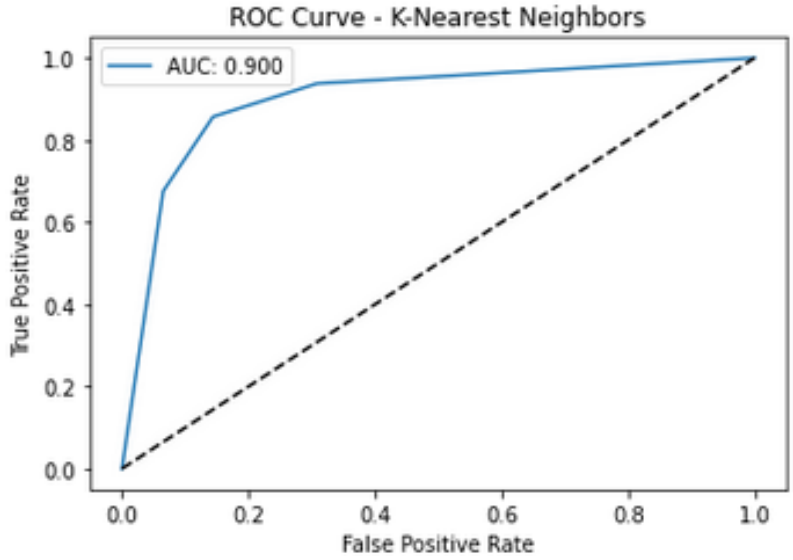
#	Column	Non-Null Count	Dtype
0	id	20058 non-null	object
1	rated	20058 non-null	bool
2	created_at	20058 non-null	float64
3	last_move_at	20058 non-null	float64
4	turns	20058 non-null	int64
5	victory_status	20058 non-null	object
6	winner	20058 non-null	object
7	increment_code	20058 non-null	object
8	white_id	20058 non-null	object
9	white_rating	20058 non-null	int64
10	black_id	20058 non-null	object
11	black_rating	20058 non-null	int64
12	moves	20058 non-null	object
13	opening_eco	20058 non-null	object
14	opening_name	20058 non-null	object
15	opening_ply	20058 non-null	int64

Ход работы

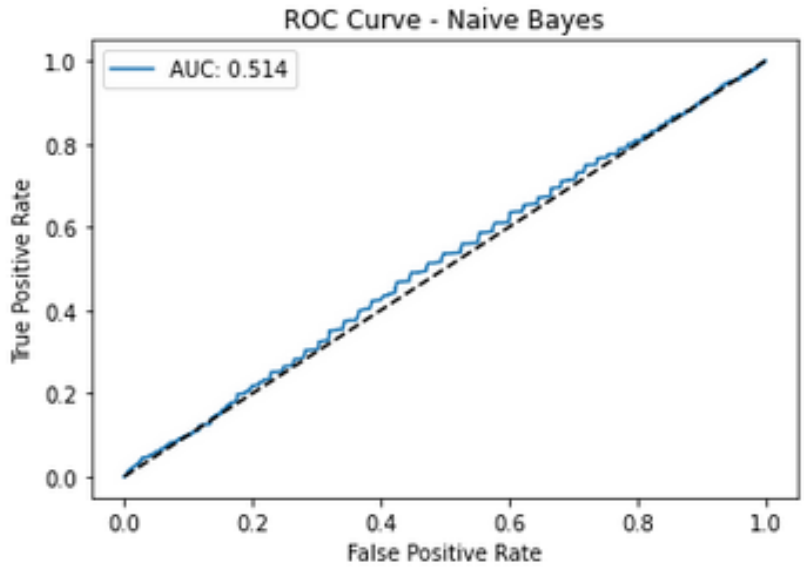
Гипотеза

Для достижения поставленной цели необходимо проверить гипотезу о зависимости исхода партии от различных параметров: дебюта, рейтинга игроков, длительности партии и др.
Для этого необходимо, помимо анализа визуализированных данных, натренировать модель на основе различных методов обучения.
Путем тестов было определено, что самым эффективным является метод дерева решений.

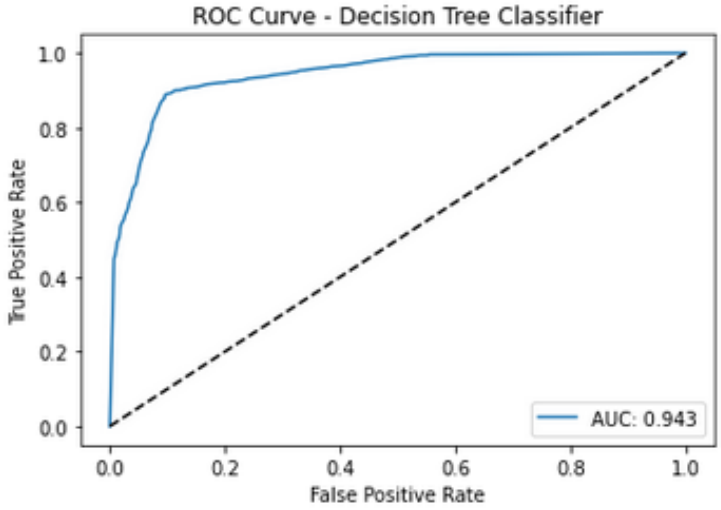
```
AUC: 0.8999610470951371
```




```
AUC: 0.514462732162612
```

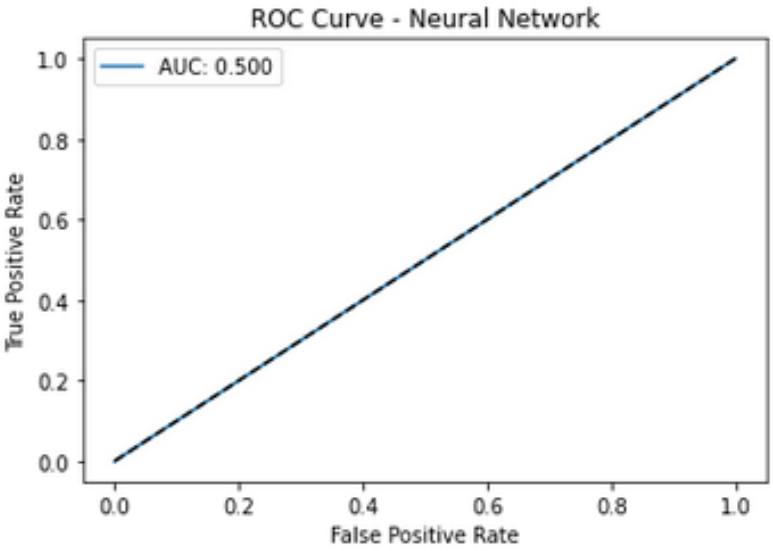


```
AUC: 0.9426458186353728
```



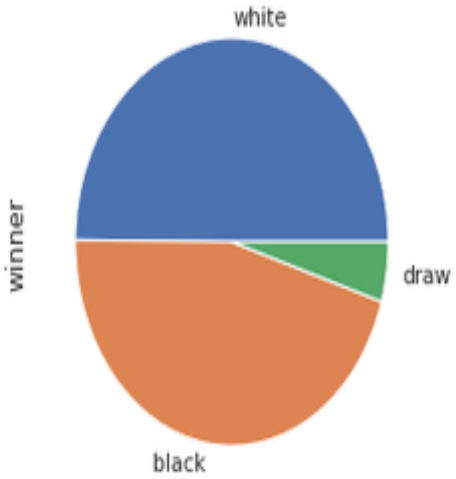
Метод дерева решений был выбран как самый точный (0.94)

```
AUC: 0.4999780065154476
```

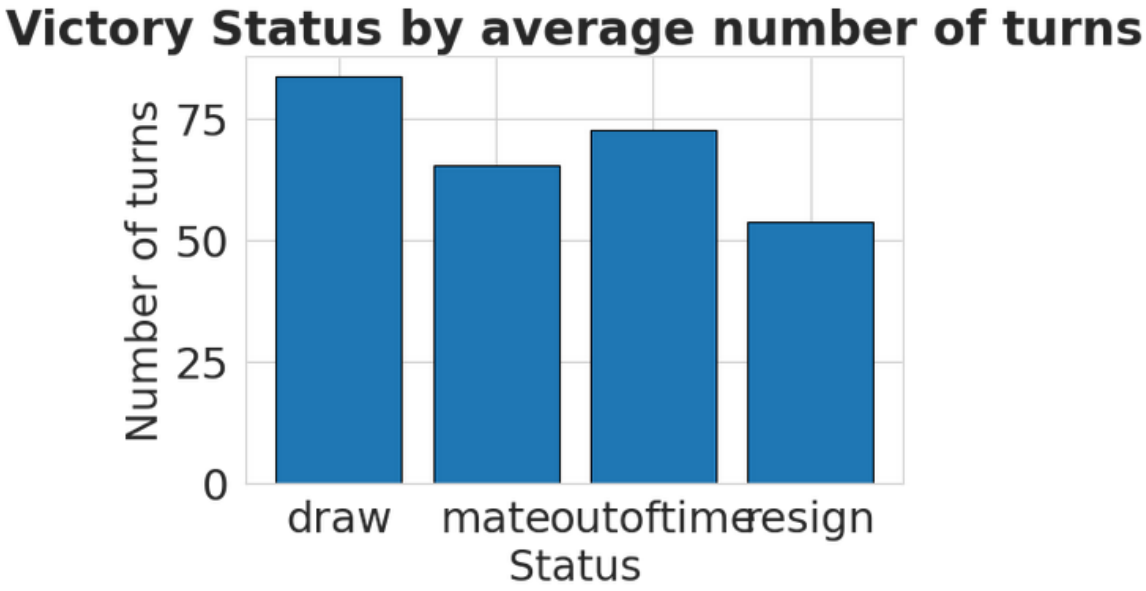


Визуализация

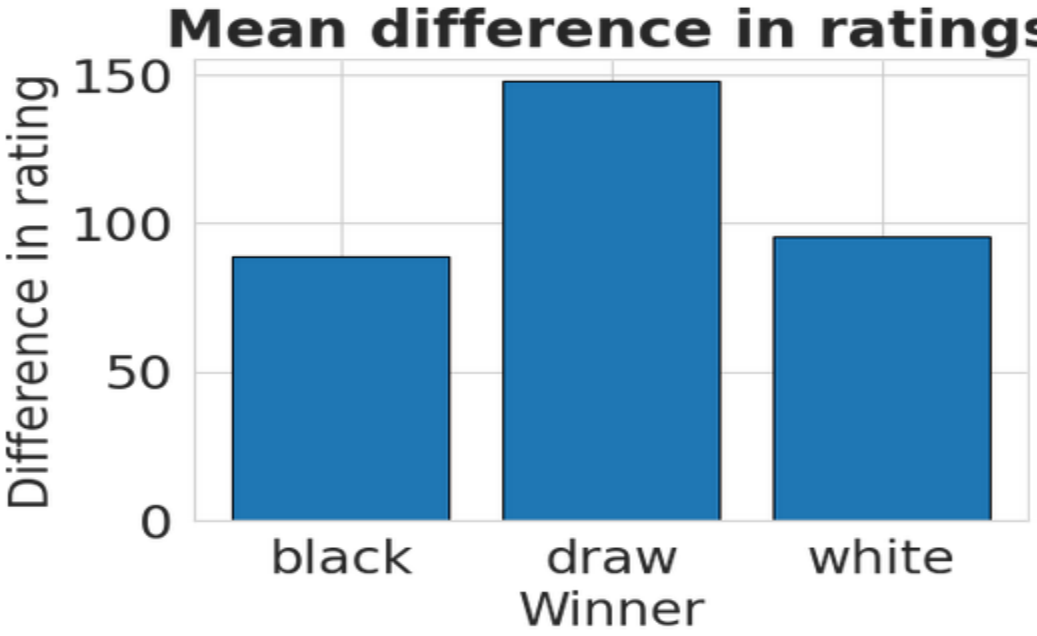
Во время работы были составлены визуализации различного характера и направленности.
Например, данные о результатах партии и количестве ходов были использованы при создании следующих визуализаций:



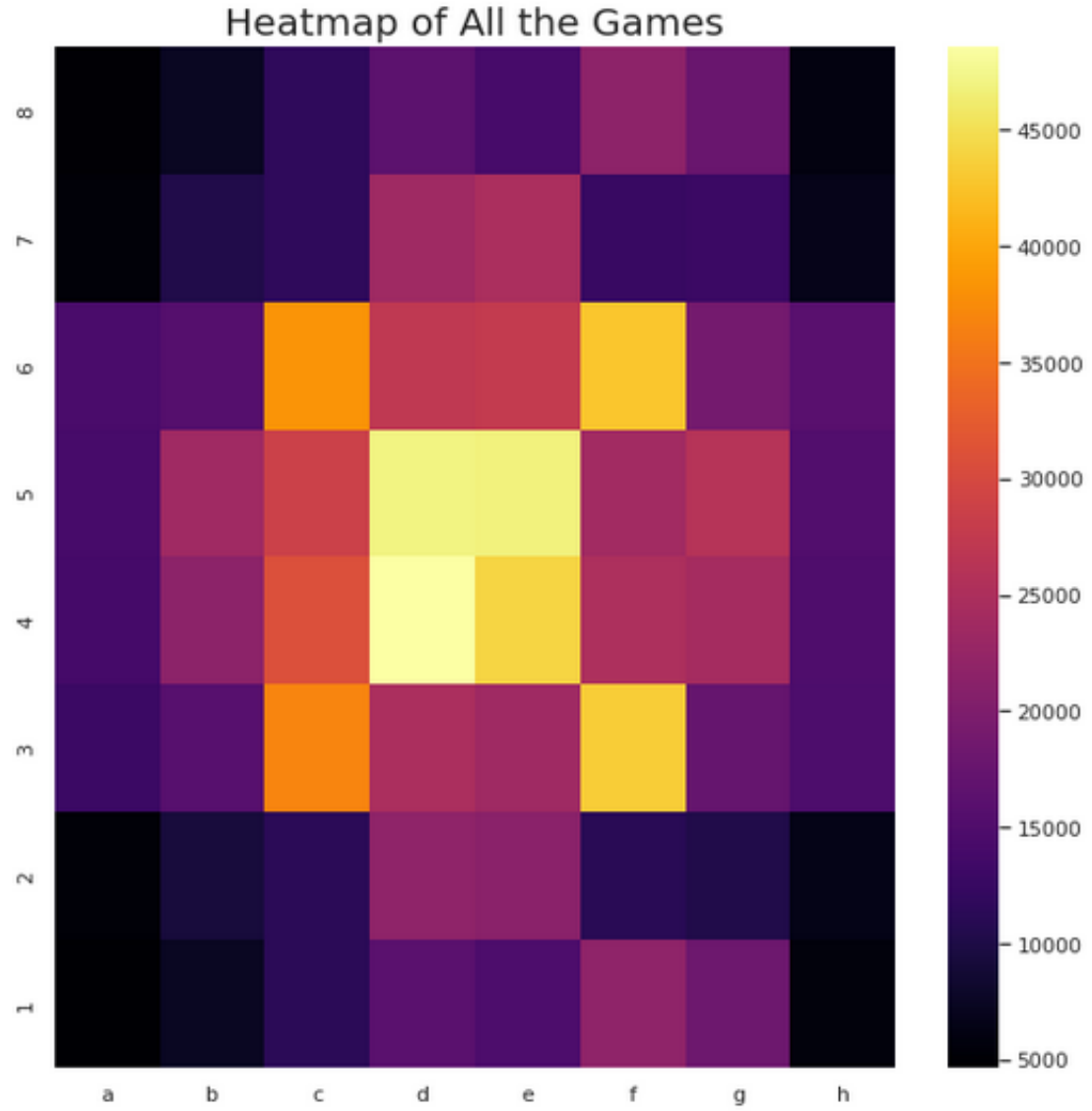
Общее соотношение побед белых, черных и ничей.



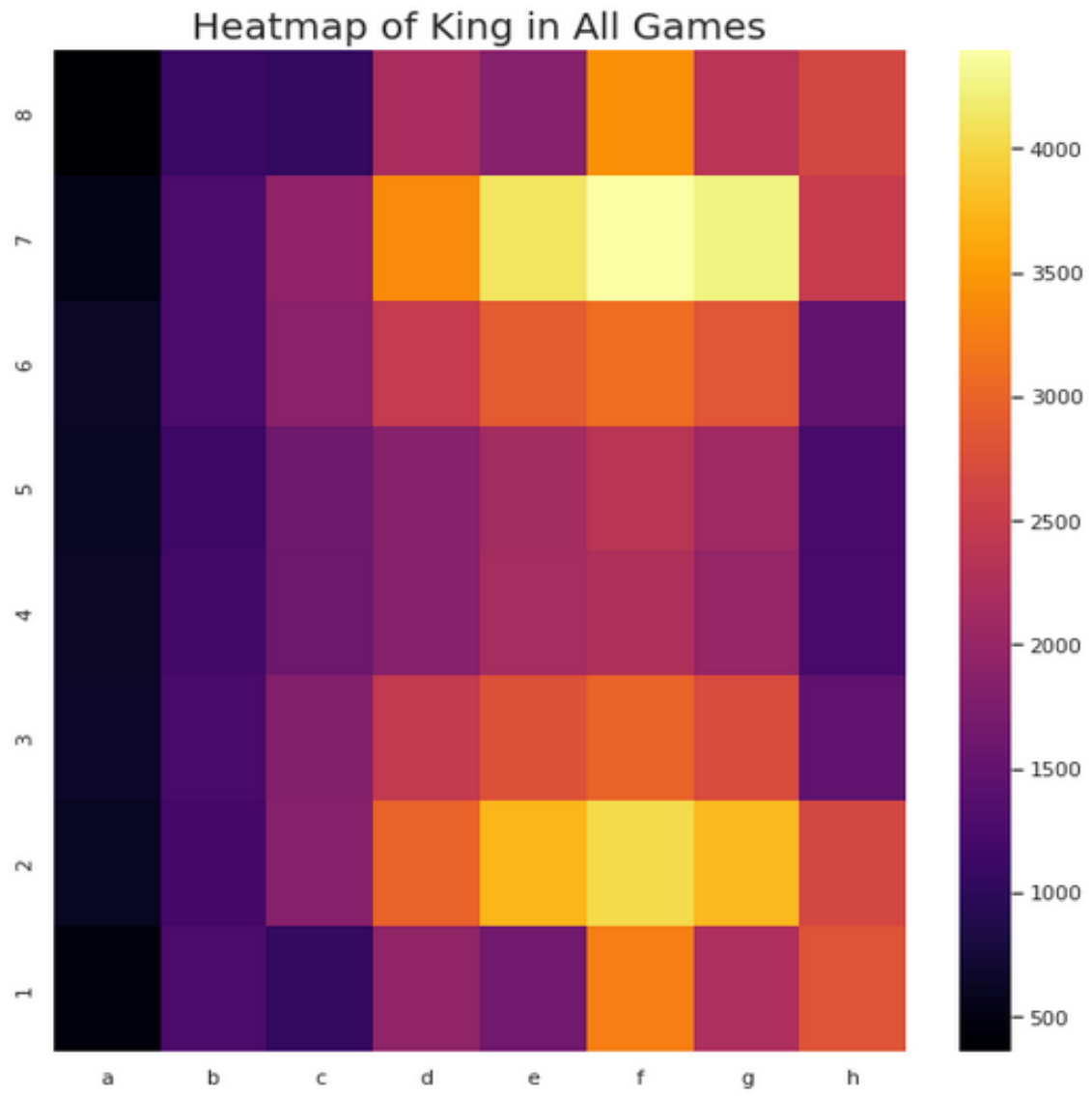
Склонность к тому или иному исходу партии в зависимости от числа ходов.



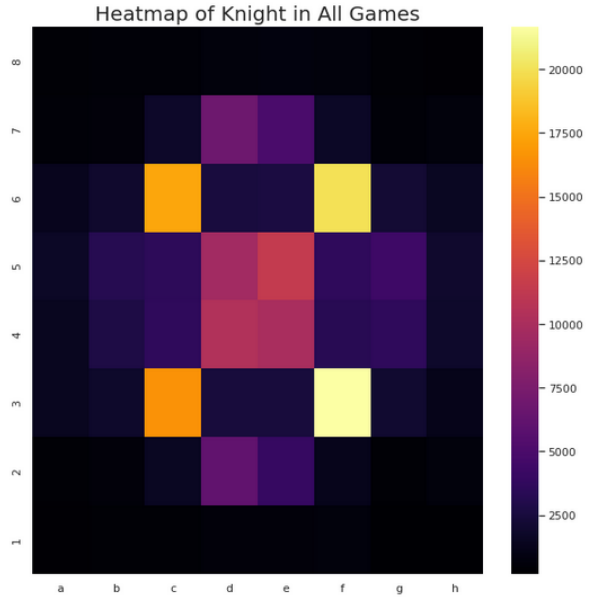
Влияние разницы в рейтинге двух игроков на исход партии.
Помимо этого, были составлены тепловые карты шахматного поля. Они отображают частоту выбора той или иной клетки поля в зависимости от фигуры. Анализ тепловой карты позволит точнее понять, какой ход в теории может оказаться наиболее вероятным или эффективным.



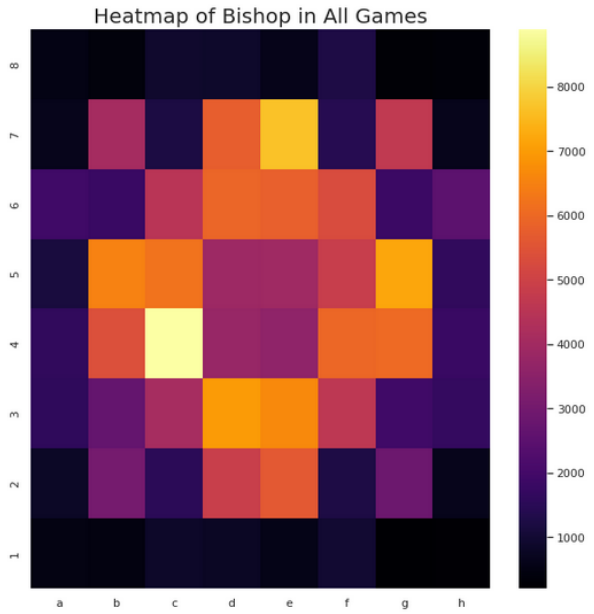
Общая тепловая карта



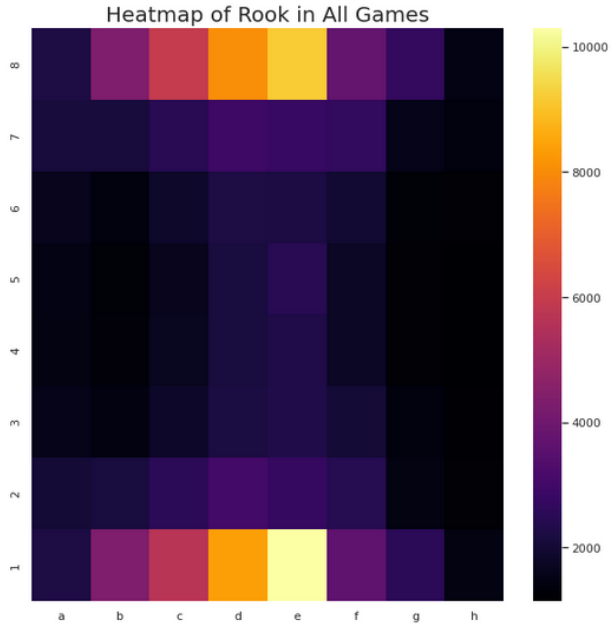
Тепловая карта королевской фигуры



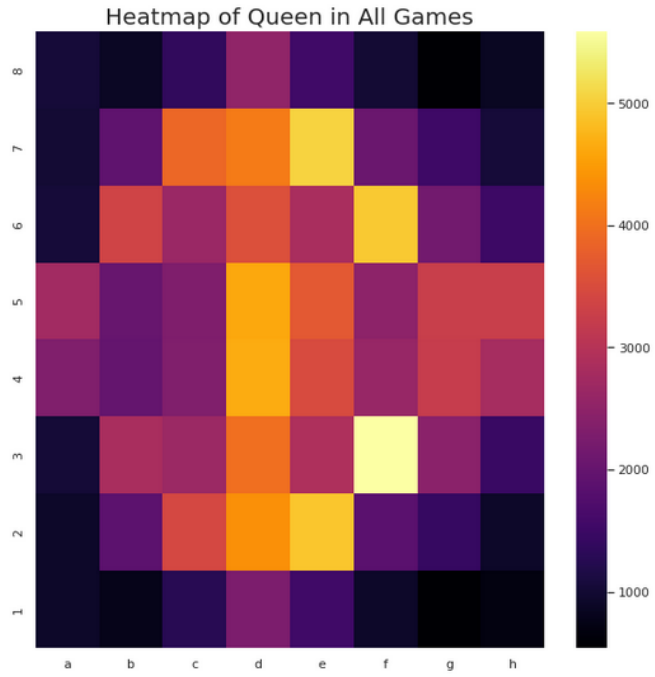
Тепловая карта коня



Тепловая карта слона



Тепловая карта ладьи

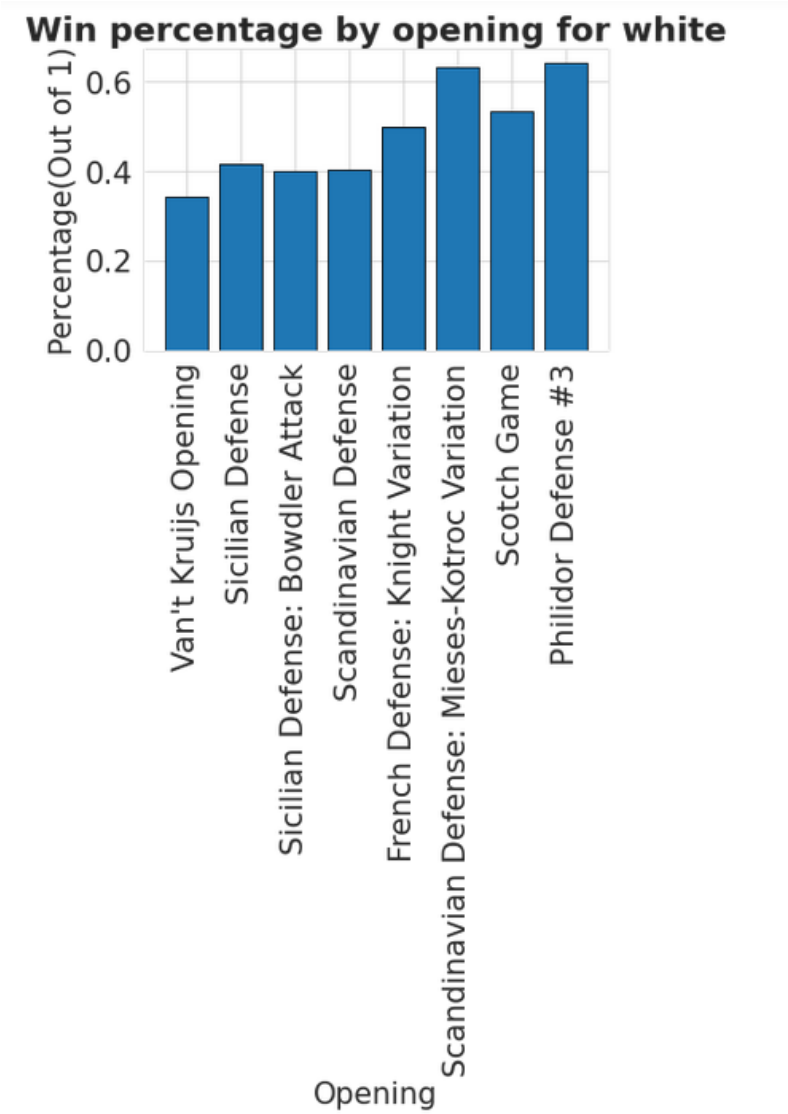


Тепловая карта ферзя

Работая с данными визуализациями, можно продумать превентивные контрмеры для различных ходов соперника.

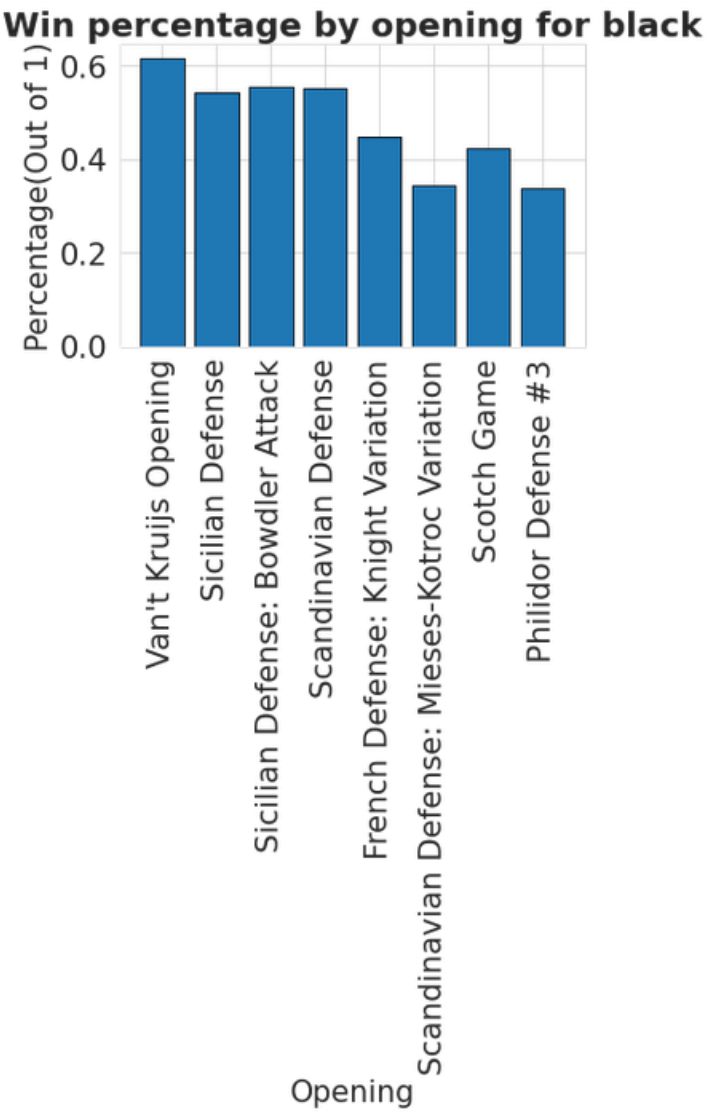
Например, совокупность всех тепловых карт говорит о том, что центр шахматного поля является важнейшим местом, контроль над которым дает оперативный простор.

Помимо прочего, был проведен анализ и визуализация популярности и эффективности дебютов за обе стороны.



	opening_name	winner	wins
2095	Scandinavian Defense: Mieses-Kotroc Variation	white	164
2190	Sicilian Defense	white	149
2113	Scotch Game	white	145
653	French Defense: Knight Variation	white	135
1451	Philidor Defense #3	white	127

Анализ дебютов за белых



	opening_name	winner	wins
2647	Van't Kruijs Opening	black	226
2189	Sicilian Defense	black	194
2220	Sicilian Defense: Bowdler Attack	black	164
2064	Scandinavian Defense	black	123
654	French Defense: Knight Variation	black	121

Анализ дебютов за черных
Данные визуализации позволяют понять вероятность применения и успеха того или иного дебюта в зависимости от выбранной стороны.

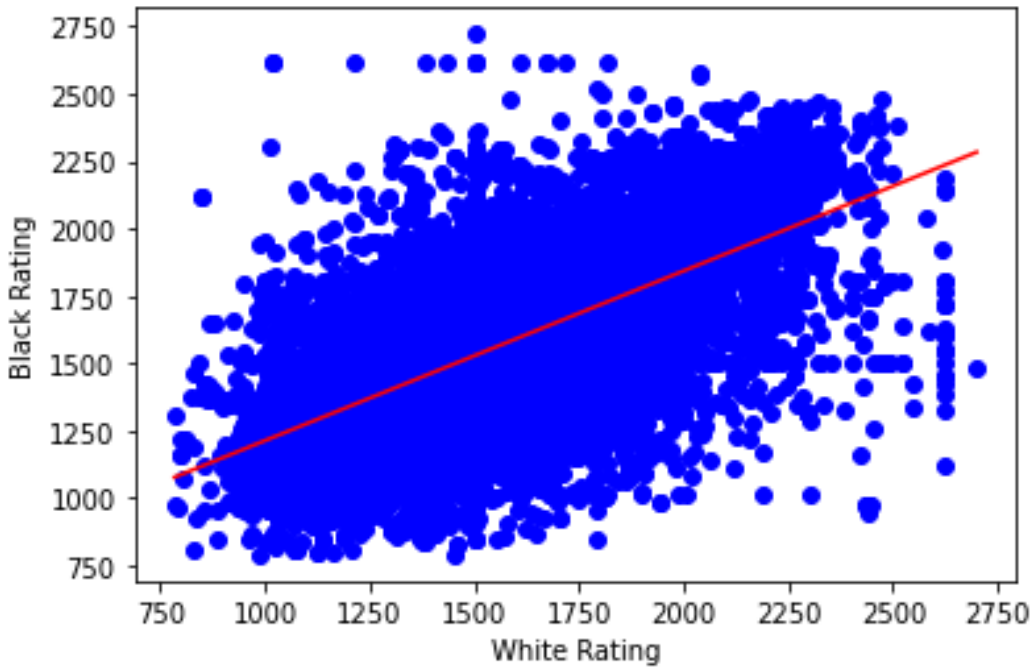
Код

Сравнение с другими решениями датасета

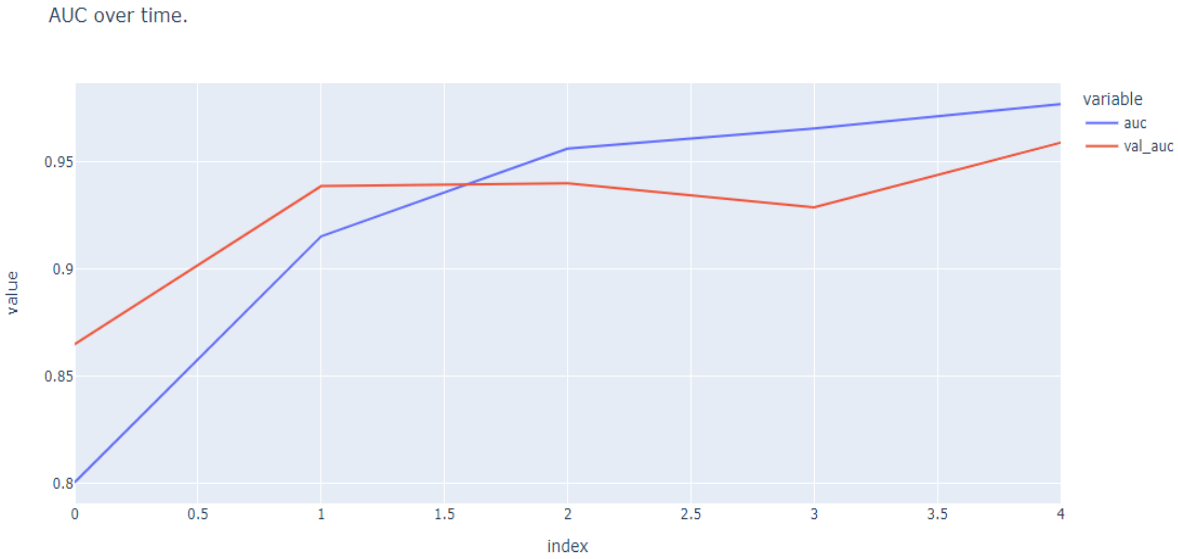
Для сравнения было взято два похожих решения.
Источники:
1. https://www.kaggle.com/code/ashish13898/linear-regression-of-predicting-rating-of-white
2. https://www.kaggle.com/code/vaishnavrathod50/chess-winner-prediction-by-rnn

В первом решении датасета выделяется следующая визуализация:

```
Mean Absolute Error of Model: 163.35
Accuracy of the Model (R^2): 0.40
```



Как мы можем наблюдать, модель обладает низкой точностью (0.40), поэтому данное решение нельзя назвать эффективным.



Второе решение датасета имеет довольно высокую точность (0.95+), поэтому его применение будет достаточно эффективным.

Выводы

Анализ и визуализация данных дают понять о зависимости исхода партии от дебюта, расположения фигур и прочих параметров.
Гипотеза подтверждена.

- Проект имеет ценность как для любителей, так и гроссмейстеров
- Возможность использовать МЕТА в рядовых партиях (Most Effective Tactic Available)
- Широкие возможности для различного анализа
- Метод дерева решений позволяет получить высокую точность прогнозирования