**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 3](#_Toc170370446)

[Описание датасета 3](#_Toc170370447)

[Ход работы 4](#_Toc170370448)

[Гипотеза 4](#_Toc170370449)

[Визуализация 5](#_Toc170370450)

[Сравнение с другими решениями датасета 12](#_Toc170370451)

[Выводы 14](#_Toc170370452)

## 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.

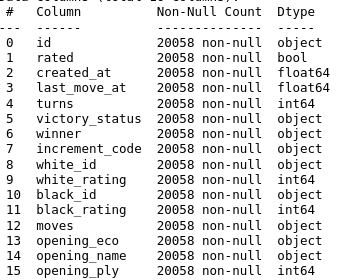
Данный проект будет полезен как шахматистам любителям, так и опытным игрокам.

## 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.



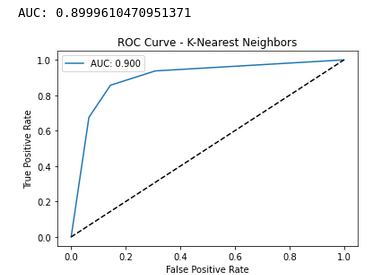
## Ход работы

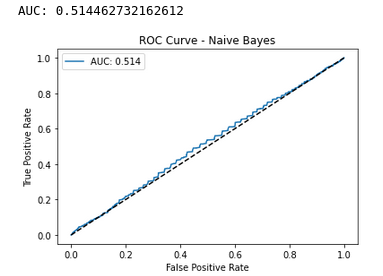
### Гипотеза

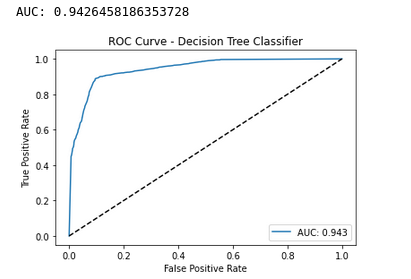
Для достижения поставленной цели необходимо проверить гипотезу о зависимости исхода партии от различных параметров: дебюта, рейтинга игроков, длительности партии и др.

Для этого необходимо, помимо анализа визуализированных данных, натренировать модель на основе различных методов обучения.

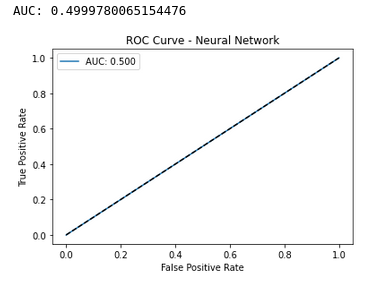
Путем тестов было определено, что самым эффективным является метод дерева решений.







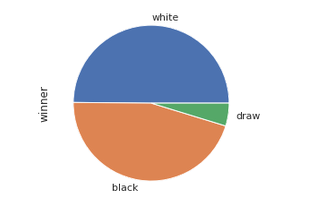
Метод дерева решений был выбран как самый точный (0.94)



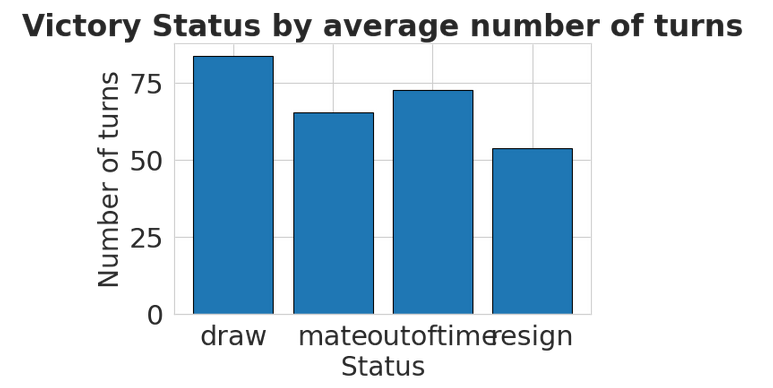
### Визуализация

Во время работы были составлены визуализации различного характера и направленности.

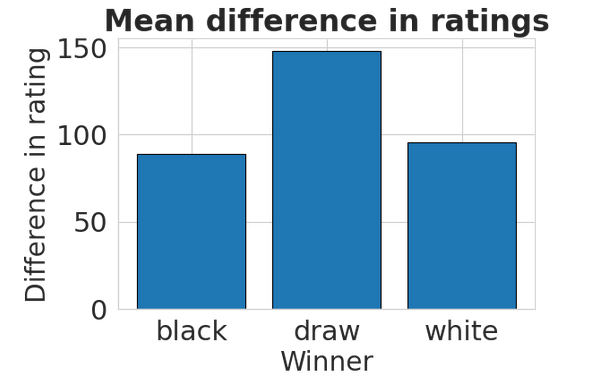
Например, данные о результатах партии и количестве ходов были использованы при создании следующих визуализаций:



Общее соотношение побед белых, черных и ничей.

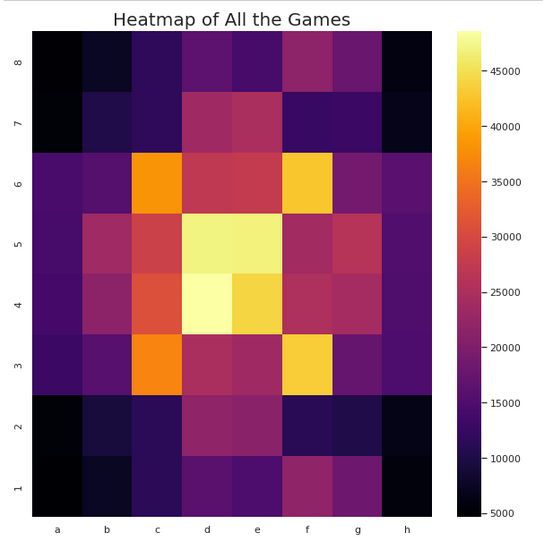


Склонность к тому или иному исходу партии в зависимости от числа ходов.

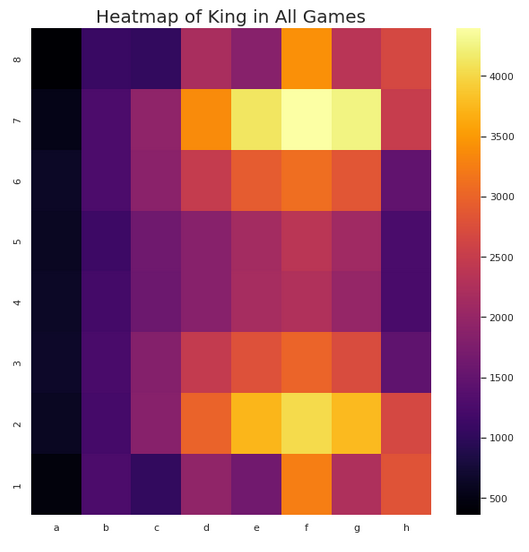


Влияние разницы в рейтинге двух игроков на исход партии.

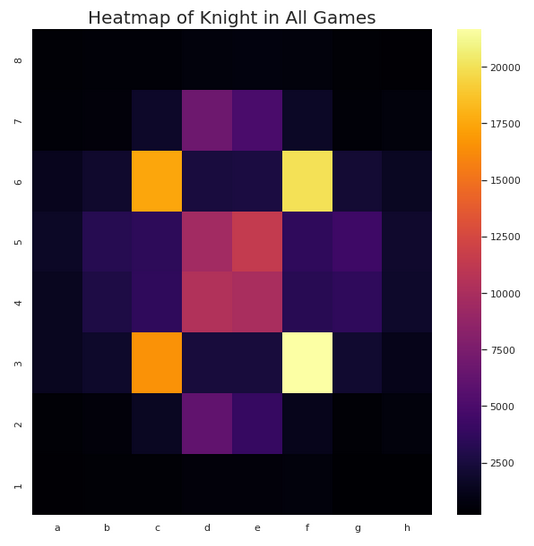
Помимо этого, были составлены тепловые карты шахматного поля. Они отображают частоту выбора той или иной клетки поля в зависимости от фигуры. Анализ тепловой карты позволит точнее понять, какой ход в теории может оказаться наиболее вероятным или эффективным.



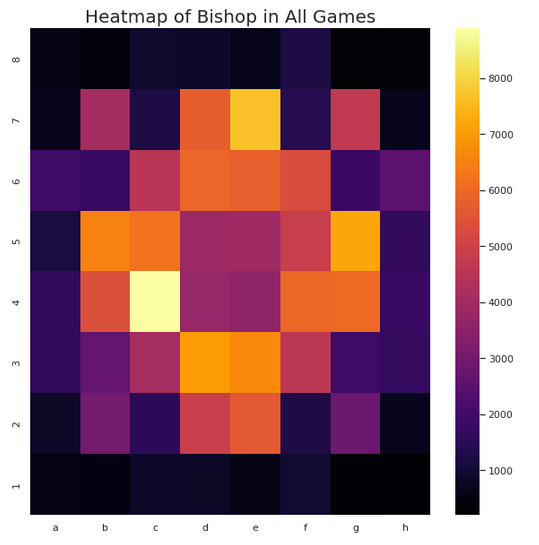
Общая тепловая карта



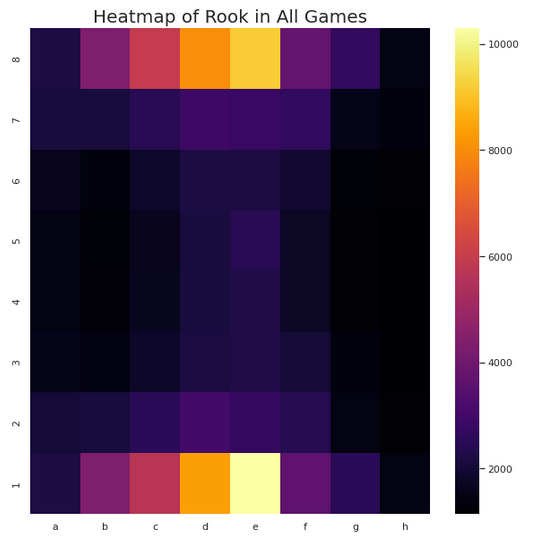
Тепловая карта королевской фигуры



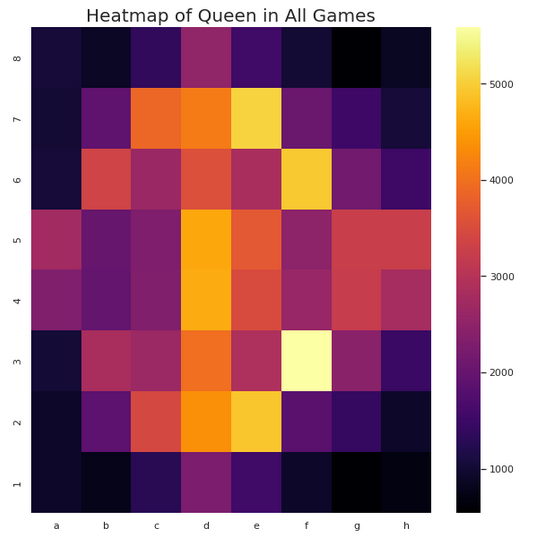
Тепловая карта коня



Тепловая карта слона



Тепловая карта ладьи

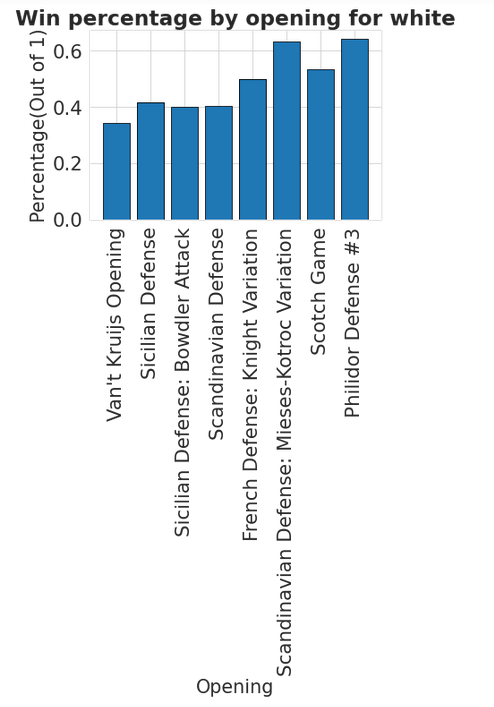


Тепловая карта ферзя

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

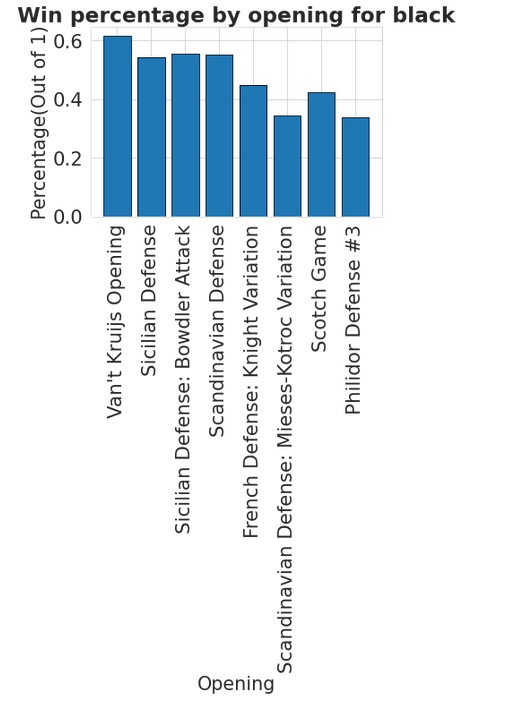
Например, совокупность всех тепловых карт говорит о том, что центр шахматного поля является важнейшим местом, контроль над которым дает оперативный простор.

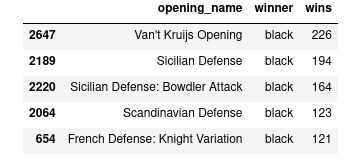
Помимо прочего, был проведен анализ и визуализация популярности и эффективности дебютов за обе стороны.





Анализ дебютов за белых





Анализ дебютов за черных

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

### Сравнение с другими решениями датасета

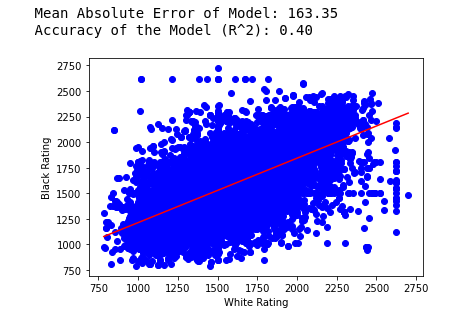
Для сравнения было взято два похожих решения.

Источники:

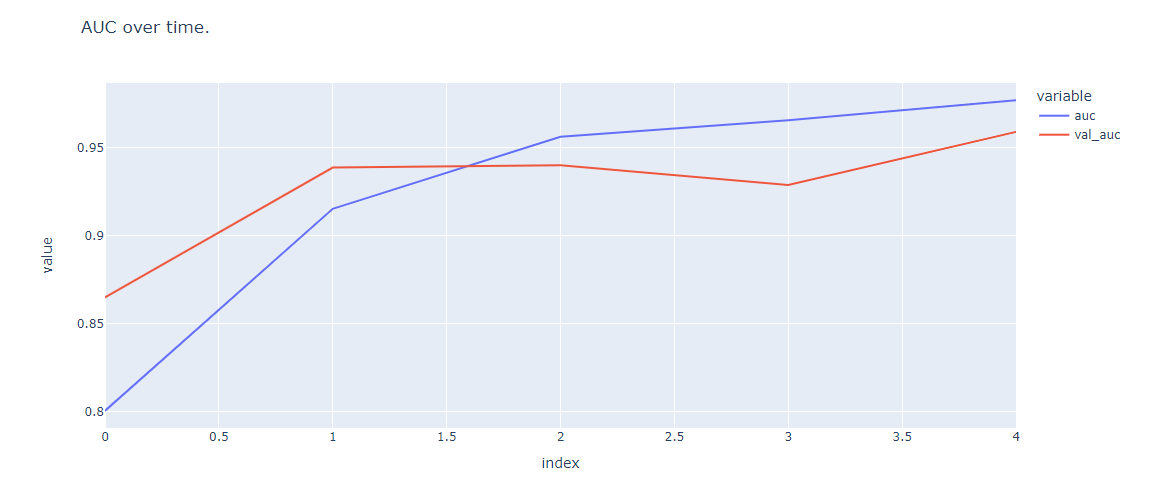
1. https://www.kaggle.com/code/ashish13898/linear-regression-of-predicting-rating-of-white

2. <https://www.kaggle.com/code/vaishnavrathod50/chess-winner-prediction-by-rnn>

В первом решении датасета выделяется следующая визуализация:



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 META в рядовых партиях (Most Effective Tactic Available)
* Широкие возможности для различного анализа
* Метод дерева решений позволяет получить высокую точность прогнозирования