Итоговый вывод о проделанной работе.

Мы проанализировали  признаки, относящиеся к факторам, которые влияют на победу в шахматных партиях.

На начальном этапе у нас было 140 792 строки с данными и 32 признака.

В ходе нашего анализа мы применяли следующие меры:

1. Удалили дубликаты: мы проверили данные на наличие дублирующихся записей и удалили их, чтобы избежать искажения результатов анализа.

2. Устранили пропуски.

3. Отфильтровали значения, чтобы не было искажения данных:

- отфильтровали нерейтинговые игры, чтобы избежать ошибок, связанных с некорректной игрой участников в партиях, не идущих на результат (по параметру Rated).

- отфильтровали значения fromPosition, чтобы не искажать данные играми, не начинающими с самого начала (по параметру Variant).

- отфильтровали игры, созданные или полученные через API, используемые для автоматизации или разработки приложений (по параметру Source).

- отфильтровали читерные партии из набора (cheat) и партии, где игроку засчитался проигрыш при отключении (timeout) (по параметру Game\_Status).

- отфильтровали данные по параметру Total\_Time\_Init.

- отфильтровали данные по параметру Total\_Time\_Increment.

- отфильтровали данные по параметру Total\_Time.

4. Удалили выбросы в данных:

- по параметрам White\_Rating и Black\_Rating.

- по параметрам White\_Mistake и Black\_Mistake.

- по параметрам  White\_Blunder и Black\_Blunder.

5. Прологарифмировали данные для уменьшения разброса значений:

- параметры White\_ACPL и Black\_ACPL.

- новый параметр Game\_Duration - продолжительность партии.

Так же мы исследовали сложные признаки:

- Clocks - информация о времени игроков по ходу партии в виде последовательного числового списка;

- Analysis - данные анализа партии от Lichess движка;

- Moves - список ходов, сделанных в партии – категориальный.

В результате анализа у нас появилось 28 новых признаков, которые нам понадобились для более детального анализа.

После проведенного анализа у нас осталось 112 481 строки.