В этом проекте предлагаются различные модели машинного обучения, которые предсказывают вероятности победы в матче для обоих игроков. Он включает в себя:: logistic regression, SVM with RBF kernel, random forest, XGBoost, 1-layer ANN.

Для этого проекта я использую файлы:
atp_matches_qual_chall_2022.csv
atp_matches_qual_chall_2023.csv
atp_matches_qual_chall_2024.csv
atp_matches_2022.csv
atp_matches_2023.csv
atp_matches_2024.csv

Все датасеты были взяты из https://github.com/JeffSackmann/tennis_atp. Эти датасеты содержат данные за матчи 2022, 2023, 2024 годов из турниров ATP, челленджеров, квалификаций (одиночный разряд). Файл https://github.com/JeffSackmann/tennis_atp/blob/master/matches_data_dictionary.txt содержит объяснение параметров датасетов. Также я использую файл odds.xlsx, он имеет ту же структуру, но 2 дополнительных столбца в конце, которые содержат коэффициенты букмекерской конторы (bet365) на линии открытия. Я использую этот файл для оценки моделей.

Файл features_create.ipynb содержит различные подходы к созданию признакового пространства. В результате я создал следующие признаки для игрока, используя информацию обо всех его предыдущих матчах:

1. Среднее количество эйсов за матч
2. Среднее количество двойных ошибок за матч
3. Среднее количество брейк поинтов за матч, совершенных противниками.
4. Среднее количество чужих брейк поинтов за матч, совершенных противниками, которые не были реализованы
5. Среднее количество брейк поинтов за матч
6. Среднее количество нереализованных брейк поинтов за матч
7. Процент побед на первой подаче
8. Процент побед на второй подаче
9. Процент побед на подачах противника
10. Процент выигранных матчей
11. Среднее количество выигранных очков за матч

Кроме того, я рассчитываю те же характеристики, но для предыдущих трех матчей игрока. Далее, я добавляю рейтинг игроков ATP и количество очков в рейтинге ATP. Таким образом, у нас уже есть 24 характеристики
После этого я добавляю 3 характеристики:

1. Корт с твердым покрытием: этот параметр отображает матч, сыгранный на корте с твердым покрытием.
2. Грунтовый корт, этот параметр отображает матч, сыгранный на грунтовом корте.

3. Травяной корт, этот параметр отображает матч, сыгранный на травяном корте.

После этого я добавлю еще один параметр

1. Одинаковая подающая рука

В конце я вычисляю первые 24 признака для победителя и проигравшего и вычитаю второй набор из первого и добавляю оставшиеся 4 признака (поверхность корта и одинаковость подающей руки). Он имеет соответствующий прогноз = 1. Я также вычитаю первый набор из второго набора и добавляю оставшиеся 4 параметра, он имеет соответствующий прогноз = 0. Я добавляю эти 2 строки в обучающий набор данных и делаю это для каждого существующего матча.
Это построение признакового пространства содержится в функциях ‘player_statistics2’ и ‘player_statistics_last’ и в последней ячейке файла features_create.ipynb.

Обучение моделей машинного обучения содержится в файле tennis_ANN.ipynb. Для обучения я использую файл odds.xlsx. Для каждой модели я подбираю параметры с помощью GridSearchCV. Наилучший результат показывает модель xgboost, точность = 0,68. Точность букмекерской конторы для этих данных составляет 0,592.
После этого, используя прогнозируемые коэффициенты, мы моделируем ситуацию, в которой мы ставим 1$ на каждый матч, в котором какой-либо прогнозируемый коэффициент ниже, чем соответствующий коэффициент, предлагаемый bet365.
ROI модели xgboost = 0.02.