Прогнозирование победителя теннисных матчей

Автор: Захаров Марк

Стадии проекта

Данный проект состоит из следующих частей:

- Определение бизнес проблемы
- Исследование набора данных
- Подготовка набора данных
- Создание ML модели для предсказаний матчей
- Выводы и практическая польза

Определение проблемы, которую необходимо решить

Заказчиками могут быть компании по типу Rolex, дизайнеры и производители часов.

- Реклама крайне важна
- Спонсор спортивных мероприятий, включая теннис

Цель: повысить рентабельность инвестиций в рекламу за счет спонсирования перспективных, но еще не популярных игроков. Для этого нужно реализовать модель прогнозирования победителя теннисного матча.

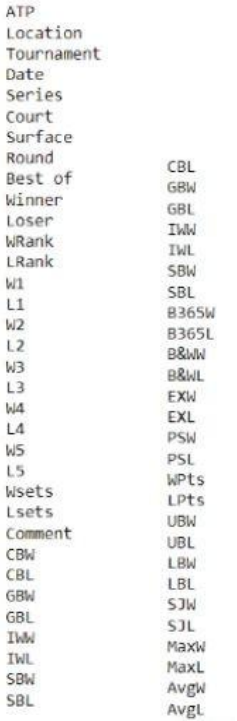
Исследование набора данных о теннисных игроках и их матчах

Информация о теннисный матчах с 2000 по 2024

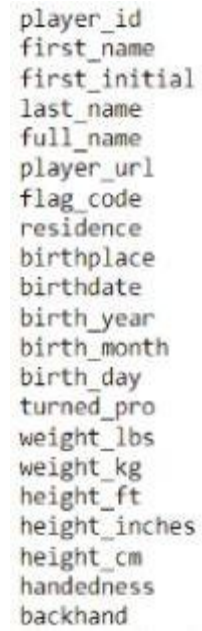
- Даты и места
- Результаты игры и сета: победитель, счет
- Характеристики игрока: вес, рост, страна, возраст и id
- Тип турнира
- Информация о корте: материал поверхности, крытый/открытый
- Коэффициенты букмекеров

Несколько дублирующихся строк

Множество значений NaN в коэффициентах букмекеров и параметрах игроков



Признаки матчей

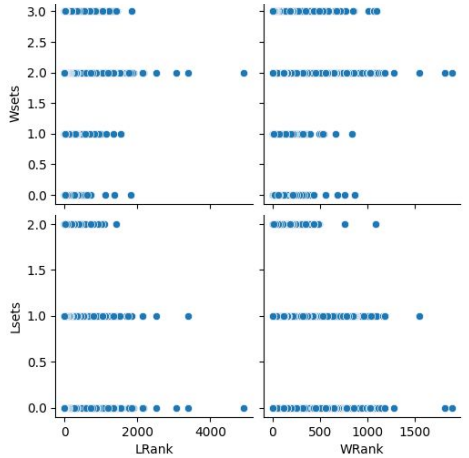


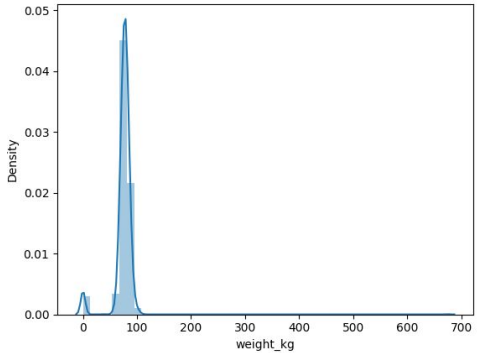
Признаки игроков

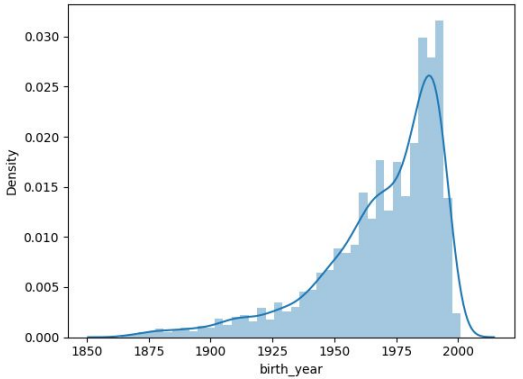
Исследование данных дает представление о корреляции характеристик с выигрышем в матче

Индикаторы победителя:

- высокий ранг
- молодой возраст
- вес близкий к 80 кг







При подготовке данных мы сначала выбрали признаки и очистили данные

Выбранные для работы признаки:

- Признаки из матчей: очки сетов, дата, место, тип корта, тип турнира
- Признаки игроков: ранг, возраст, вес, рост, id, страна

Очистка данных включила в себя:

- Заполнение пропущенных значений медианой или значениями по умолчанию
- Удаление дубликатов строк

	player_id	first_name	first_initial	last_name	full_name	player_url	flag_code	residence	birthplace	birthdate	...	birth
9999	v007	Jerome	J	Vanier	Vanier J	http://www.atpworldtour.com/en/players/jerome-...	FRA	NaN	Boulogne, France	19571102	...	
10000	v007	Jerome	J	Vanier	Vanier J	http://www.atpworldtour.com/en/players/jerome-...	FRA	NaN	Boulogne, France	19571102	...	

дубликаты строк

В процессе подготовки данных нам потребовалось ввести новые признаки и интегрировать внешние данные(API).

Произведенные признаки:

- день, месяц, и год
- нормализованная разница в рангах между игроками
- играл ли игрок на этом месте раньше
- процент побед игрока против соперника в прошлом

Внешние данные:

- погодные признаки: температура, скорость ветра, давление и т.п.

Month	Day	Year	W_over_L_rank_difference_normalized	W_played_here_before	L_played_here_before	Temperature	Windspeed_10m	Relativehumidity_2m	Surface_pressure	Precipitation	Cloudcover	Shortwave_radiation	player_1_h2h_winning_ratio
1	3	2000	0,200670695	ЛОЖЬ	ЛОЖЬ	15,05	27,75	68	1004,016667	0	40,5	0	0,5
5	1	2000	2,257122719	ЛОЖЬ	ЛОЖЬ	18,9666667	8,766666667	72,5	951,4833333	0,966666667	39	579,3333333	0,5
2	19	2001	0,427444015	ИСТИНА	ЛОЖЬ	23	0	50	1013,25	0	100	0	0,5
9	30	2002	-0,260283098	ИСТИНА	ЛОЖЬ	23	0	50	1013,25	0	100	0	1

пример строк новых признаков

В процессе подготовки данных мы закодировали(one-hot encoder) признаки и нашли наилучший масштабатор.

Закодированные признаки:

- Тип турнира
- Поверхность
- Тип корта

eries_International Gold	Series_Masters	Series_Masters 1000	Series_Masters Cup	Court_Indoor	Court_Outdoor	Surface_Carpet	Surface_Clay	Surface_Grass	Surface_Hard
0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	1	0	1	0	0
1	0	0	0	1	0	0	0	0	1

пример закодированных значений

Проверенные масштабаторы:

- МинМакс
- Стандартный - лучший

В процессе моделирования мы протестировали LogisticRegression and другие нейросети.

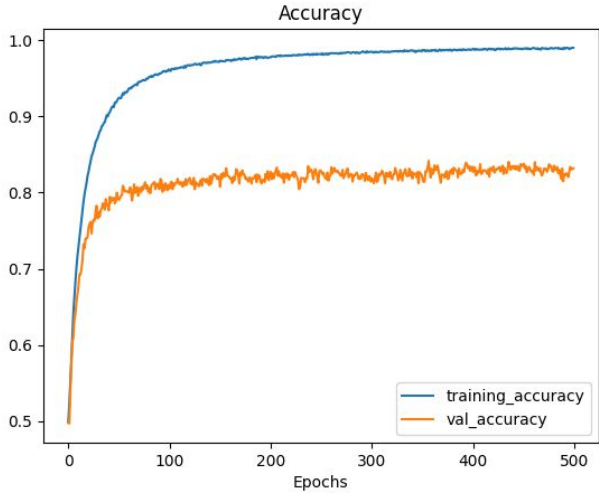
Базовая модель: LogisticRegression

Продвинутая модель: со слоями (dense + dropout(0.2))

Глубина, количество параметров и точность:

- 7-слойная сеть (4 Dense + 3 Dropout), 1.6 млн -> 75.55% точность
- 9-слойная сеть (5 Dense + 4 Dropout), 8 млн -> 80.27% точность (83.2% с масштабатором)

Финальная модель: 4-слоя (dense + dropout(0.2))
Нейросеть с 8 млн параметров и стандартным масштабатором

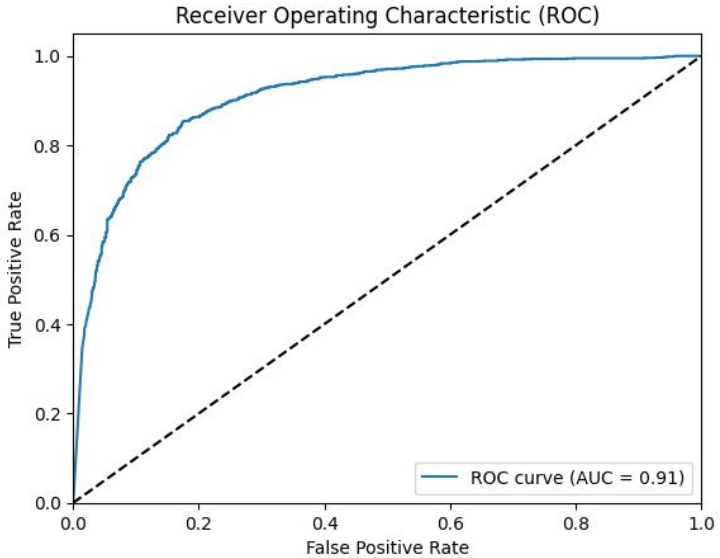


точность в процессе тренировки модели

Выводы и практическая польза

Площадь под roc кривой - 91%, итоговая точность - 83.2% означает, что натренирована модель достаточно хорошо, чтобы верить ее прогнозам в большинстве случаев.

Ценность для бизнеса: Rolex знает будущих победителей и может спонсировать недооцененных игроков



Спасибо за внимание!

Полезные ссылки:

- Репозиторий проекта - https://github.com/Get-My-Money/JMLC-project
- Оригинальный набор данных - https://data.world/tylerudite/atp-match-data