Предсказание результатов теннисных матчей для оптимизации спонсорских решений

Введение

В рамках данного проекта была разработана модель машинного обучения для прогнозирования результатов теннисных матчей с целью оптимизации спонсорских решений. Основная проблема, которую решает проект - традиционный подход компаний к спонсорству, когда фокус делается на уже известных спортсменах, что приводит к завышенным затратам.

Идея проекта заключается в выявлении "темных лошадок" - перспективных, но еще не раскрученных теннисистов, чья победа в будущих турнирах может принести значительно большую отдачу при меньших вложениях.

В работе использовались:

- Данные ATP Match Data (2000-2024 гг.): 66702 матча
- Данные ATP Players: 10 912 записей об игроках
- Погодные данные для матчей на открытом воздухе

Целевой метрикой качества модели была установлена точность прогноза не менее 80%.

Предобработка данных

Была проведена тщательная очистка и подготовка данных:

1. Удаление нерелевантных данных:
    - Удалены коэффициенты букмекеров и избыточные колонки
    - Удалены записи с незавершенными матчами
    - Удален дубликат в данных об игроках
2. Обработка пропусков:
    - Пропуски в ключевых колонках (WRank, LRank) удалены
    - Пропуски в счете заменены нулями
    - Пропуски в возрасте, весе и росте заполнены средними значениями
3. Создание раздельных датасетов для победителей и проигравших с префиксами "W_" и "L_"

Создание ключевых признаков

Были разработаны следующие важные признаки:

1. Процент выигранных сетов победителем - показывает доминирование в матче
2. Среднее количество очков за раунд и процент очков победителя - отражают качество игры
3. Нормализованная разница рейтингов - логарифм отношения рангов игроков
4. История игрока в локации - флаги "играл ли ранее в этом городе"
5. Погодные условия для матчей на открытом воздухе (температура, влажность, облачность)
6. Стандартные значения для матчей в закрытых помещениях

Модель и обучение

Для решения задачи классификации был выбран подход с использованием нейронной сети как наиболее подходящий для захвата сложных взаимосвязей в данных.

Архитектура финальной модели:

- 9-слойная полносвязная нейронная сеть
- 8 млн параметров
- Слои с dropout (0.2) для регуляризации
- Функция активации ReLU
- Оптимизатор Adam с learning rate 0.001
- Размер батча: 128
- Количество эпох: 500

Обучение проводилось на 95% данных, оставшиеся 5% использовались для тестирования. Была применена стандартизация числовых признаков и ранняя остановка при отсутствии улучшения качества.

Результаты и выводы

Метрики качества на тестовой выборке:

- Точность (Accuracy): 83.2%
- AUC-ROC: 0.91
- F1-мера: 0.83
- Precision: 0.83
- Recall: 0.83

Анализ важности признаков показал, что наиболее значимыми являются:

1. Нормализованная разница рейтингов игроков
2. Погодные условия для матчей на открытом воздухе
3. Возраст игроков
4. История личных встреч
5. Тип покрытия корта

Практическая ценность модели:

- Возможность выявлять перспективных игроков до их массовой известности
- Оптимизация бюджета на спонсорские контракты
- Повышение ROI за счет более ранних инвестиций в перспективных игроков

Основные ограничения:

- Отсутствие данных о текущей форме игроков (травмы, усталость)
- Неучет психологического состояния спортсменов
- Невозможность прогнозировать неожиданные события

Рекомендации по улучшению:

- Интеграция данных о текущей форме игроков
- Использование рекуррентных нейронных сетей для учета временных зависимостей
- Расширение датасета за счет данных о более ранних периодах

Заключение

Разработанная модель демонстрирует высокую точность прогнозирования результатов теннисных матчей (83.2%) и может быть эффективно использована

для оптимизации спонсорских решений. Проект показал, что комбинация традиционных статистических данных об игроках с внешними факторами (погода, тип покрытия) позволяет создавать точные прогнозы, которые могут стать основой для принятия бизнес-решений в сфере спортивного спонсорства.

Несмотря на некоторые ограничения, модель предоставляет ценные инсайты для компаний, желающих оптимизировать свои инвестиции в спорт. В будущем модель может быть улучшена за счет интеграции дополнительных источников данных и более сложных архитектур нейронных сетей.