ФИНАЛЬНЫЙ ПРОЕКТ
ПРОГНОЗ ИСХОДА ФУТБОЛЬНОГО МАТЧА

5 команда

Суровцева Н. | Остапинский М. | Дудко Д. | Петренко Д.

Задача - предсказать исход футбольного матча

Kaggle

Использовать базу данных European Soccer Database

Выбрать 2-3 ключевые характеристики, которые влияют на исход матча

Подготовить дашборд для предсказания с возможностью выбрать играющие команды

Executive summary

SQL

- Сформировали исходные данные с помощью SQL-запросов;
- Создали из них 2 датафрейма в Jupyter Notebook:
  1) *Match (Country+League+Players);*
  2) Team attributes.

Python

- Провели предобработку данных (очистка + стандартизация);
- Определили ключевые характеристики:
  - *прессинг в обороне, риск. передач, удар по воротам, агрессивность обороны;*
- Использовали модель для решения задачи классификации - **Random Forest (точность на всей выборке ~60%, на тестовой выборке ~50%);**
- Настроили гиперпараметры Random Forest для увеличения точности и ускорения пересчета результата;
- Итог: **модель Random Forest превосходит линейную и логистическую регрессию**, **точность которых не превысила 40%;**

Tableau

- Подготовили фильтры (лига, команда, дома/на выезде) для пользовательского выбора;
- Вывели ключевые характеристики для 2 команд;
- Показали прогноз исхода матча (какая команда имеет больше шансов на победу).

Python: подготовка к работе датафрейма Match

Data cleaning, normalization, feature engineering

Предобработка данных

- Проверили на пропуски, дубликаты, аномалии;
- Привели столбцы к нужному типу данных;

Подготовка параметров модели

Кто с кем играет? История матчей:

- Объединили датафреймы Match и Team;
- Создали новый столбец Winner (результат для домашней команды,ключевой столбец для будущих моделей).

Теперь ***каждая строка*** показывает данные о каждом матче:

- Дата матча;
- 2 команды (домашняя и выездная);
- Исход матча для домашней команды.

Данные подходят для моделирования + удобны для использования в Tableau

Python: анализ датафрейма Team attributes

Выбор ключевых характеристик, влияющих на исход матча

Наиболее коррелирующие характеристики:

- ***defence Aggression***
  агрессивность в обороне
- ***defence Pressure***
  прессинг в обороне
- ***chance Creation Shooting***
  количество ударов по воротам при создании голевых моментов
- ***chance Creation Passing***
  рискованность передач при создании голевых моментов

Python: соединение датафреймов Match и Team attributes

анализ получившегося датафрейма

- Матчи и характеристики команд **свели вместе по совпадающим годам** (2010-2015). Таким образом устранили ситуации, при которых для матча использовались бы характеристики команды не того года;
- ***Feature Engineering:*** создали дополнительные столбцы по **статистике команд** дома и на выезде (выигрыш, проигрыш, ничья и их % от количества матчей).

Статистика дома

	home_team_name	home_win	home_lost	home_draw
0	1. FC Kaiserslautern	8	15	11
1	1. FC Köln	24	21	23
2	1. FC Nürnberg	28	29	20
3	1. FSV Mainz 05	46	34	23
4	AC Ajaccio	16	22	19
...	...	...	...	...
269	Xerez Club Deportivo	5	4	2
270	Zagłębie Lubin	24	17	19
271	Zawisza Bydgoszcz	5	9	2
272	Évian Thonon Gaillard FC	29	30	17
273	Śląsk Wrocław	37	18	26

Статистика на выезде

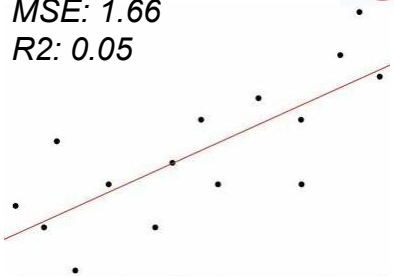
	away_team_name	away_win	away_lost	away_draw
0	1. FC Kaiserslautern	9	18	7.0
1	1. FC Köln	17	39	12.0
2	1. FC Nürnberg	18	38	19.0
3	1. FSV Mainz 05	28	41	31.0
4	AC Ajaccio	6	30	21.0
...	...	...	...	...
268	Xerez Club Deportivo	2	6	4.0
269	Zagłębie Lubin	15	29	17.0
270	Zawisza Bydgoszcz	5	9	4.0
271	Évian Thonon Gaillard FC	16	41	19.0
272	Śląsk Wrocław	25	29	30.0

Python: выбор подходящей модели

Linear Regression, Logistic Regression, Random Forest

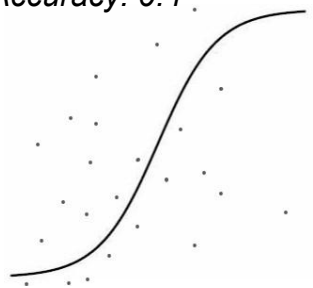
Линейная регрессия

MSE: 1.66
R2: 0.05



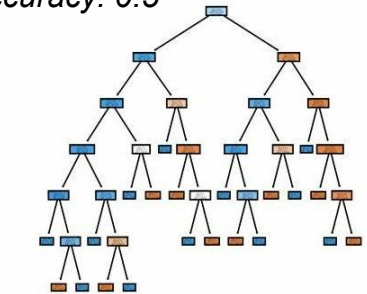
Логистическая регрессия

Accuracy: 0.4



Метод случайного леса

Accuracy: 0.5



- Линейная и логистическая регрессии не отражали логику данных (underfitting);
- Модель Random Forest подошла для задачи классификации. На полной выборке точность доходила до 60%, на тестовой выборке - немного более 50%;
- Анализ моделей позволил точнее выявить полезные характеристики (features) для моделирования. *Данные **Player attributes не добавили точности модели**, поэтому мы ограничились только командными характеристиками.*

Переходим в Tableau

Дашборд в Tableau доступен по ссылке