Języki programowania Python i R

Raport z modułu II

Justyna Mackoś

1. **Tytuł**

   Przewidywanie wyniku gry Dota2 na podstawie wybranych na początku rozgrywki bohaterów.

2. **Motywacja, cele**

   Pomimo, że nie grałam konkretnie w tę grę, której dotyczą dane (Dota2), to miałam okazję spróbować innych gier typu MOBA (Multiplayer Online Battle Arena) i jestem ciekawa, czy sam wybór bohaterów ma duży wpływ na wynik rozgrywki. Pewnym jest, że niektóre postacie lepiej współpracują z innymi, co może się przełożyć na większą szansę wygranej. Dodatkowo podczas wyboru bohaterów tworzy się tak zwane „kontry", czyli wybiera postać, której umiejętności (np. czary) blokują przeciwnika, bądź nie mogą być zablokowane przez gracza drużyny przeciwnej – co daje przewagę. Oczywiście sam wynik rozgrywki zależy przede wszystkim od umiejętności graczy, ale mam nadzieję, że na podstawie analizy przebiegu gier uda się z pewnym prawdopodobieństwem określić która drużyna wygra, nie znając pozostałych zmiennych.

3. **Opis danych**

   Dane zostały pozyskane od:
   Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
   Dokładny link do bazy: https://archive.ics.uci.edu/ml/datasets/Dota2+Games+Results

   Łącznie w bazie znajduje się 102942 rekordów. Wszystkie zostały pozyskane w trakcie 2-godzinnego zapisu (rozgrywki odbywały się równolegle) w dniu 13.08.2016.
   Dane zostały podzielone przez udostępniającego na dwa pliki – zbiór uczący i testowy w stosunku 90% – 10%. Na późniejszym etapie zostały one przez mnie połączone w jeden plik i podzielone w kodzie.
   Każdy rekord składa się z 117 atrybutów oznaczających kolejno:

   3.1 Kto wygrał grę - wartość 1 lub -1 oznaczającą drużynę.
   3.2 Lokalizacja gry – liczba całkowita oznaczająca region gry np. Europa wschodnia, Chiny – każdy region w zależności od wielkości może się składać z kilku podregionów
   (np. Rosja to numery od 181 do 188).
   3.3 Mapa gry – liczba całkowita
   3.4 Typ tworzenia drużyn – liczba całkowita (np. gra samemu z losowym doborem drużyny, albo z drużyną dobraną przed rozpoczęciem rozgrywki)
   3.5 113 atrybutów opisujących kolejnych bohaterów oznaczonych jako
   - 1 gdy dana postać została wybrana przez pierwszą drużynę
   - -1 gdy postać należała do drużyny przeciwnej
   - 0 gdy bohater nie został wybrany do tego meczu

Pliki (nie mojego autorstwa) opisujące jacy bohaterowie/regiony/tryby kryją się pod ich ID.

Plik tłumaczący tryby rozgrywki:
https://github.com/kronusme/dota2-api/blob/master/data/mods.json
Plik łączący ID bohatera z jego imieniem:
https://github.com/kronusme/dota2-api/blob/master/data/heroes.json
Plik opisujący regiony:
https://github.com/kronusme/dota2-api/blob/master/data/regions.json

4. Opis procesu przygotowywania danych do analizy

Po sprawdzeniu w Pythonie wszystkie dane były już typu liczb całkowitych int64, więc nie była wymagana zmiana na zmienną kategoryczną.

Kolumny nie miały swoich nazw, więc przyjęły pewne automatyczne nazwy – próby wpisania ręcznie do pliku .csv w Excelu powodowały błędy i w rezultacie powstawała tylko jedna kolumna. Próby wpisania czegokolwiek do pliku csv użytego na zajęciach i ponownego zapisania również powodowały zmianę odczytywania kolumn z działających wcześniej kilku, na jeden długi string. Problem udało się rozwiązać wpisując nazwy kolumn do pliku .csv używając notatnika, zamiast Excela.

Do analizy nie powinny być brane pod uwagę dane w których rozgrywka była inna niż 5 graczy na 5 graczy, oraz te, w których pojawiają się nieznane wartości. Zbadałam, że:

- W trybie tworzenia drużyn nie pojawia się wartość 8, czyli gra 1vs1
- W trybie tworzenia drużyn nie pojawia się wartość -1, czyli „Invalid”
- W wyborze mapy nie pojawia się wartość 0 czyli „Unknown”
- Przy wyborze mapy nie pojawia się wartość 21, czyli mapa 1vs1

5. Analiza danych

Podczas analizy danych odkryłam, że jeden z bohaterów o numerze ID 108 nie został ani razu wybrany podczas wszystkich gier. Taka sytuacja jest bardzo mało prawdopodobna, ponieważ w zbiorze testowym jest ok. 90 tys. rekordów, a każda gra polega na wyborze 10 (różnych) bohaterów. Dlatego też było to zaskakujące, że podczas wybierania prawie milion razy ze zbioru zaledwie 113 postaci jakaś postać została całkowicie ominięta. Gdyby tą analizę przeprowadzić wtedy, kiedy te dane zostały pobrane, można by ten fakt zgłosić do twórców gry, zwracając uwagę, że jedna z ich postaci prawdopodobnie nie spełnia wymagań użytkowników (możliwe, że użytkownicy doświadczalnie zauważyli, że wybierając tego bohatera częściej przegrywają i dlatego tego nie robili).
Po rozwinięciu listy okazało się, że jeszcze jeden bohater (o ID 24) również nie został wybrany ani razu. Co ciekawe w pliku łączącym ID z bohaterem nie ma podanego imienia do bohatera nr 24.
Zdecydowałam się połączyć zbiory podzielone przez udostępniającego, żeby móc sprawdzić czy te wartości (i inne) nie pojawiają się też w zbiorze testowym.

Podane wcześniej ID dwóch bohaterów również nie pojawiły się w drugim zbiorze, więc nie będzie to generowało możliwych problemów.

6. Modelowanie danych

Wykorzystałam metodę k-najbliższych sąsiadów, która polega na przewidywaniu wyniku na podstawie znanych rekordów, które są zakwalifikowane jako najbliższe otoczenie badanego rekordu. Odległość, która określa czy rekordy są „najbliższymi sąsiadami", czy też nie może być liczona na różne sposoby (ja będę używać odległości Manhattan i Euklidesa).

Badałam zależności tylko między wybranymi bohaterami i wynikiem meczu. Kwestie regionu i typu gry wyłączyłam z budowania modelu. Podzieliłam dane na zbiór treningowy i testowy w proporcji 80%-20%.
Pierwsza próba dla liczby sąsiadów równej 5 wyszło niezadowalająco. Skuteczność wynosiła 53%, a macierz pomyłek miała postać:

		Wynik testu	
		Negatywny	Pozytywny
Wynik rzeczywisty	Negatywny	4679	5034
	Pozytywny	4662	6214

Próbowałam więc dalej zmieniając parametry.
W następnej próbie użyłam 32 najbliższych sąsiadów. Wynik się polepszył do 55% skuteczności:

		Wynik testu	
		Negatywny	Pozytywny
Wynik rzeczywisty	Negatywny	4867	4846
	Pozytywny	4444	6432

Dla 64 najbliższych sąsiadów udało się uzyskać skuteczność 56%:

		Wynik testu	
		Negatywny	Pozytywny
Wynik rzeczywisty	Negatywny	4438	5275
	Pozytywny	3826	7050

Na tym etapie odczułam potrzebę odtwarzania dźwięku po wykonaniu się bloku, gdyż jedna instrukcja potrafiła już trwać po kilka minut.

Dla 100 najbliższych sąsiadów wynik wyniósł 56,5 %, a macierz pomyłek prezentowała się w ten sposób:

		Wynik testu	
		Negatywny	Pozytywny
Wynik rzeczywisty	Negatywny	4180	5533
	Pozytywny	3415	7461

Zamiast zwiększać ilość sąsiadów zdecydowałam się spróbować z inną definicją odległości. Ustawiłam algorytm, żeby używał teraz odległości Manhattan, zamiast Euklidesowskiej.

Dla 5 sąsiadów:
53% skuteczności (tak jak przy odległości Euklidesa)

		Wynik testu	
		Negatywny	Pozytywny
Wynik rzeczywisty	Negatywny	4664	5049
	Pozytywny	4683	6193

Widząc, że wyniki są podobne przeskoczyłam od razu do 100 sąsiadów.
Udało się uzyskać niecałe 57% skuteczności, a macierz pomyłek wyglądała następująco:

		Wynik testu	
		Negatywny	Pozytywny
Wynik rzeczywisty	Negatywny	4125	5588
	Pozytywny	3334	7542

Powtórzyłam pomiar dla 200 najbliższych sąsiadów, udało się uzyskać trochę ponad 57% skuteczności (57.3%)

		Wynik testu	
		Negatywny	Pozytywny
Wynik rzeczywisty	Negatywny	3717	5996
	Pozytywny	2803	8073

Zwiększam ilość do 500 najbliższych sąsiadów, skoro póki co ilość sąsiadów sprawiała, że skuteczność rosła.
Skuteczność wynosiła 57,4 %

		Wynik testu	
		Negatywny	Pozytywny
Wynik rzeczywisty	Negatywny	3178	6535
	Pozytywny	2231	8645

Ostatnią próbą było przeprowadzenie metody dla 1000 najbliższych sąsiadów i wykazała się ona najlepszym rezultatem, czyli skutecznością 57,7%

		Wynik testu	
		Negatywny	Pozytywny
Wynik rzeczywisty	Negatywny	2737	6976
	Pozytywny	1734	9142

Ciekawym jest, że coraz więcej wyników jest zwracane jako wynik „pozytywny" – wygrana jednej drużyny, co zwiększa zarówno wyniki rzeczywiście pozytywne, jak i fałszywie-pozytywne.

Nadal nieusatysfakcjonowana przeprowadziłam jeszcze jedną próbę dla 200 najbliższych sąsiadów, ale z wszystkimi zmiennymi, nie tylko tymi pochodzącymi od wyboru bohatera. Wyniki jednak były gorsze, skuteczność 56%, więc na tym zakończyłam ten proces.

7. Rezultaty, wnioski i ich dyskusja

Tak naprawdę zagadnienie, które jest tutaj badane jest zależne od wielu czynników, które nie były zawarte w bazie danych – choćby ilość zagranych przez gracza meczy jako konkretna postać, albo ilość zagranych meczy w ogóle mogłyby lepiej przybliżyć wynik.
Nie spodziewałam się modelu o wysokiej skuteczności, ponieważ z założenia gra powinna być dobrze zbilansowana, a co za tym idzie od wyboru postaci nie powinien zależeć przebieg meczu. Jak się jednak okazało można z niemal 58% skutecznością określić która drużyna wygra już na starcie. Dla zwykłego modelu analizy danych jest to z pewnością niesatysfakcjonujący wynik, jednak w tym konkretnie zagadnieniu można mówić o pewnym sukcesie.