ŚCIĄGAWKA

WORKFLOW W UCZENIU MASZYNOWYM:

1. Pozyskanie danych
2. Zadanie pytania na które chcemy znać odpowiedź w oparciu o zebrane dane (chyba najtrudniejsza część procesu)
3. Przygotowanie danych (wczytanie je do pamięci, wyczyszczenie, doprowadzenie do postaci łatwej do zinterpretowania przez komputer – jednym słowem przetworzenie)
4. Wybór odpowiedniego algorytmu (jest determinowany przez problem, który chcemy rozwiązać, ale też przez liczbę dostępnych danych itd.)
5. Wytrenowanie tzw. modelu, czyli przygotowanie jednostki (np. obiektu) zdolnego odpowiedzieć na zadane w punkcie 2 pytanie – to tu jest szukany wzorzec.
6. Przetestowanie modelu. Jest to sprawdzenie dokładności z jaką nasz model „przewiduje przyszłość". Innymi słowy sprawdzenie skuteczności wzorca.
7. W razie potrzeby iteracja od punktu, który wg. naszej wiedzy przyniesie poprawę wyników.

RMSE - Root Mean Square Error (błąd średniokwadratowy) lub inaczej odpowiedź na pytanie, jak bardzo nasz model się pomylił

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n \text{ are predicted values}$$

$$y_1, y_2, \ldots, y_n \text{ are observed values}$$

$$n \text{ is the number of observations}$$

Błąd taki możemy wyznaczyć, porównując wartości przewidziane przez nasz model z wartościami rzeczywistymi ze zbioru testowego - Im mniejsze różnice, tym mniejszy błąd czyli tym lepszy model.

Przykład:

Price	Predicted
1900	2000
2000	2000
2100	2000

MSE = 1/3 * ((-100)*(-100)+ (0)*(0) + (100)*(100)) = 1/3 * (20000) = 6000
RMSE = 77.45
Zatem dla każdej wartości mylimy się średnio o 77.45 jednostki.
fajne źródła: https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e

BENCHMARK oraz **DONE DEFINITION** - done definition to moment, w którym uznamy, że nasz model jest już wystarczająco dobry. Benchmark to wartość, od której zaczniemy mierzenie. Przykładowo, w rodzinie Kowalskich ludzie zwykle są bardzo wysocy, od pokoleń Kowalscy osiągają powyżej 180cm wzrostu. Gdybyśmy chcieli przewidzieć, ile wzrostu dokładnie będzie miało kolejne dziecko Kowalskich, możemy przyjąć, że:

- Kowalscy średnio mają 170cm wzrostu
- istnieje wysokie prawdopodobieństwo, że następny Kowalski będzie miał 170cm wzrostu
- sprawdzamy, jak daleko 170cm leży od innych miar ze zbioru testowego (Ala Kowalska - 172cm, Mirek Kowalski 184cm, Kasia Kowalska 169cm)

MSE = ⅓ * ((2)*(2)+ (14)*(14) + (-1)*(-1)) = 1/3 * (4 + 196 + 1) = 67

RMSE = 8.18

Nasz benchmark (170cm) daje błąd równy 8.18cm. Wyznaczamy za done definition moment, w którym błąd będzie nie większy, niż 4 cm.

https://hackernoon.com/choosing-the-right-machine-learning-algorithm-68126944ce1f

CONFUSION MATRIX (MACIERZ BŁĘDU)

Jest podstawowym narzędziem stosowanym do *oceny jakości klasyfikacji*. Poniżej rozważymy przypadek klasyfikacji binarnej (dwie klasy).

1 – Positive (np.: fakt skorzystania z produktu przez Klienta, pacjent z potwierdzoną chorobą, pacjentka z potwierdzoną ciążą)

0 – Negative (np.: fakt nieskorzystania z produktu przez Klienta, pacjent z wykluczoną chorobą, pacjentka z wykluczoną ciążą)

Możliwe wyniki klasyfikacji

- *True-Positive (**TP** – prawdziwie pozytywna)*: przewidywanie pozytywne, faktycznie zaobserwowana klasa pozytywna (np. pozytywny wynik testu ciążowego i ciąża potwierdzona)
- *True-Negative (**TN** – prawdziwie negatywna)*: przewidywanie negatywne, faktycznie zaobserwowana klasa negatywna (np. negatywny wynik testu ciążowego i brak ciąży)
- *False-Positive (**FP** – fałszywie pozytywna)*: przewidywanie pozytywne, faktycznie zaobserwowana klasa negatywna (np. pozytywny wynik testu ciążowego, jednak faktyczny brak ciąży)
- *False-Negative (**FN** – fałszywie negatywna)*: przewidywanie negatywne, faktycznie zaobserwowana klasa pozytywna (np. negatywny wynik testu ciążowego, jednak ciąża potwierdzona)

Czułość i specyficzność jako miary „zasięgu”

- ***Czułość*** = ***TPR*** = ***True-Positive Rate***, miara „zasięgu / pokrycia / dotarcia” wskazująca w jakim procencie klasa faktycznie pozytywna została pokryta przewidywaniem pozytywnym (np. procent osób chorych, dla których test diagnostyczny wskazuje wynik pozytywny). TPR zapisujemy również jako
  TPR = P(pred = P | fakt = P) = P(pred = 1 | fakt = 1) = P(1 | 1)

- ***Specyficzność*** = ***TNR*** = ***True-Negative Rate***, miara „zasięgu / pokrycia / dotarcia” wskazująca w jakim procencie klasa faktycznie negatywna została pokryta przewidywaniem negatywnym (np. procent osób zdrowych, dla których test diagnostyczny wskazuje wynik negatywny). TNR zapisujemy również jako
  TNR = P(pred = N | fakt = N) = P(pred = 0 | fakt = 0) = P(0 | 0)

PPV i NPV jako miary precyzji

- ***Precyzja przewidywania pozytywnego*** = ***PPV*** = ***Positive Predictive Value***, miara precyzji wskazująca z jaką pewnością możemy ufać przewidywaniom pozytywnym, tzn. w jakim

procencie przewidywania pozytywne potwierdzają się stanem faktycznie pozytywnym (np. procent osób z pozytywnym wynikiem testu medycznego, u których następnie potwierdzono diagnozę). PPV można zapisać również jako
PPV = P(fakt = P | pred = P) = P(fakt = 1 | pred = 1)

- ***Precyzja przewidywania negatywnego*** = ***NPV*** = – ***Negative Predictive Value***, miara precyzji wskazująca z jaką pewnością możemy ufać przewidywaniom negatywnym, tzn. w jakim procencie przewidywania negatywne potwierdzają się stanem faktycznie negatywnym (np. procent osób z negatywnym wynikiem testu medycznego, u których następnie wykluczono chorobę). NPV można zapisać również jako
NPV = P(fakt = N | pred = N) = P(fakt = 0 | pred = 0)

Zależność pomiędzy miarami jakości klasyfikacji

- ***Czułość (TPR)*** vs ***Specyficzność (TNR)*** – teoretycznie miary niezależne, co dobrze obrazują powyższe schematy. W praktyce jednak zwiększanie czułości prowadzi często do zmniejszenia specyficzności.
- ***PPV i NPV*** vs ***Czułość (TPR) vs Specyficzność (TNR)*** – korzystając z twierdzenia Bayesa można łatwo wyznaczyć zależność pomiędzy miarami precyzji i miarami zasięgu

Przykład – do grupy 2000 osób skierowano komunikację marketingową zachęcającą do skorzystania z produktu. Spośród 2000 osób produkt zakupiło 600. Grupę 2000 podzielono losowo na dwie równoliczne części, każda po 1000 osób (w tym w każdej po 300 klientów, którzy skorzystali z produktu). Pierwszej grupie przydzielono rolę *„danych uczących”*, zaś drugiej rolę *„danych testowych”*. Wykorzystując dane uczące, dostępne charakterystyki klientów oraz informacje o fakcie zakupienia produktu (tzw. *target*), *przygotowano (wytrenowano / nauczono) klasyfikator* umożliwiający przewidywanie czy dany klient skorzysta z produktu. Oceny jakości klasyfikatora dokonano przy wykorzystaniu danych testowych (tzn. danych, które nie były używane w procesie uczenia). Wyniki oceny zaprezentowano w postaci poniższej macierzy błędów.

		Stan faktyczny	
		P	***N***
Przewidywanie	***P***	***250*** *True-Positive*	***100*** *False-Positive*
	N	***50*** *False-Negative*	***600*** *True-Negative*

<u>Wnioski</u>:

- *TP + FN + TN + FP = 250 + 50 + 600 + 100 =* ***1000*** – liczba klientów (baza, na której dokonano oceny)
- *P = TP + FN = 250 + 50 =* ***300*** – liczba klientów, którzy kupili produkt
- *N = TN + FP = 600 + 100 =* ***700*** – liczba klientów, którzy nie skorzystali z produktu
- *TP + TN = 250 + 600 =* ***850*** – liczba poprawnych klasyfikacji
- *FP + FN = 100 + 50 =* ***150*** – liczba błędnych klasyfikacji
- *ACC = (TP + TN) / (P + N) = 850 / 1000 =* ***85%*** – **jakość klasyfikacji**
- *ERR = (FP + FN) / (P + N) = 150 / 1000 =* ***15%*** – **poziom błędu**