Tytuł raportu

Badanie zależności liniowej oceny czekolady przez krytyków w zależności od jej składu

Wstęp

Wybrałem ten temat, ponieważ:

- bardzo lubię czekoladę
- chciałem dowiedzieć się, zawartość jakich składników ma największy wpływ na ocenę czekolady
- chiciałem sprawdzić, czy ocena czekolady jest funkcją (liniową) składu

Opis danych - struktura zbiorów, opis zmiennych, pochodzenie

Zbiór danych zawiera następujące zmienne:

- ref - unikalny dla każdej firmy numer
- company - nazwa firmy
- company_location - lokalizacja firmy
- review_date - data recenzji czekolady
- country_of_bean_origin - kraj pochodzenia ziaren kakao
- specific_bean_origin_or_bar_name - dokładniejsza lokalizacja pochodzenia ziaren kakao
- cocoa_percent - zawartość procentowa kakao
- rating - ocena
- counts_of_ingredients - ilość składników
- beans - czy czekolada zawiera ziarna kakao?
- cocoa_butter - czy czekolada zawiera masło kakaowe?
- vanilla - czy czekolda zawiera wanilię?
- lecithin - czy czekolda zawiera lecytynę?
- salt - czy czekolda zawiera sól
- sugar - czy czekolda zawiera cukier
- sweetener_without_sugar - czy czekolda zawiera słodzik
- first_taste - smak czekolady przy pierwszej próbie
- second_taste - smak czekolady przy drugiej próbie
- third_taste - smak czekolady przy trzeciej próbie
- fourth_taste - smak czekolady przy czwartej próbie

Dane pochodzą z następującej strony: https://www.kaggle.com/soroushghaderi/chocolate-bar-2020

Opis procesu przygotowywania danych do analizy - kolejne kroki

W pierwszej kolejności odrzucam zmienne:

- ref
- company
- company_location
- review_date
- country_of_bean_origin
- specific_bean_origin_or_bar_name
- first_taste
- second_taste
- third_taste
- fourth_taste

Celem tego zabiegu jest uniezależnienie modelu od zmiennych niezaliczających się do składu czekolady. Zmienne takie, jak nazwa firmy, oraz data recenzji nie powinny mieć wpływu na wynik.

Następnie sprawdzam ilość unikalnych wartości, w celu wyeliminowania zbędnych zmiennych.

```
cocoa_percent: [ 42.   46.   50.   53.   55.   56.   57.   58.   60.   60.5  61.   62.
  63.   64.   65.   66.   67.   68.   69.   70.   71.   71.5  72.   72.5
  73.   73.5  74.   75.   76.   77.   78.   79.   80.   81.   82.   83.
  84.   85.   86.   87.   88.   89.   90.   91.   99.  100. ]
rating: [1.   1.5  1.75 2.   2.25 2.5  2.6  2.75 3.   3.25 3.5  3.75 4.  ]
counts_of_ingredients: [1 2 3 4 5 6]
beans: ['have_bean']
cocoa_butter: ['have_cocoa_butter' 'have_not_cocoa_butter']
vanilla: ['have_not_vanila' 'have_vanila']
lecithin: ['have_lecithin' 'have_not_lecithin']
salt: ['have_not_salt' 'have_salt']
sugar: ['have_not_sugar' 'have_sugar']
sweetener_without_sugar: ['have_not_sweetener_without_sugar'
'have_sweetener_without_sugar']
```

Na podstawie powyższych wyników postanowiłem, że: - należy usunąć zmienną beans, ponieważ posiada tylko jedną unikalną wartość - zmienne coco_butter, vanilla, lecithin, ssalt, sugar i sweetener_without_sugar to tak na prawdę zmienne logiczne i powinny zostać na takowe zamienione - zmienna cocoa_percent może zostać znormalizowana

Analiza danych - przyjęte założenia, krótki opis metod i obranej metodologii analizy

Zbadana została korelacja liniowa między zmiennymi - zmienna counts_of_ingridients wykazała dużą korelację ze zmiennymi vanilla, lecithin i cocoa_butter, co może oznaczać, że uwzględnia ona zawartość tych składników - zmienne sugar, oraz sweetener_without_sugar

wykazują duzą (ok. -0.96) korelację ujemną

Postanowiłem w związku z tym usunąć zmienną counts_of_ingridients, oraz zostawić na razie zmienne zawierające informacje o cukrze, oraz słodziku, ponieważ zastąpienie cukru za pomocą słodzika może mieć wpływ na ocenę, ponieważ słodziki często smakują inaczej (np. stewia).

Modelowanie danych - przyjęte założenia, krótki opis metod i obranej metodologii budowania modeli

- Wykorzystany został model regresji liniowej.
- dane po przygotowaniu wykazały słabe rezultaty
- w następnej kolejności wystąpła próba sprawdzenia, czy zaokrąglenie w dół wartości oceny polepszy model
- zmienne zawartości cukru i słodzika były mocno (ujemnie) skorelowane, w związku z czym wyznaczyłem na ich podstawie zmienną mówiącą, czy czekolada zawiera słodzik lub cukier.
- ostatecznie usunięta zmienna ta została usunięta, by sprawdzić jaki wpływ będzie to miało na model

Rezultaty, wnioski i ich dyskusja

- Składniki nie wykazały znacznej korelacji ze zmienną *rating*, najwyższy wynik (warotść bezwzględna) uzyskała zmienna vanilla (ok. -0.1).
- Model regresji liniowej nie był w stanie strafnie przewidzieć oceny krytyka na podstawie składu czekolady.
- Zmiany w modelu pogarszały jedynie jego skuteczność.