LabMasters

szkolimy, analizujemy, doradzamy

Analiza danych i programowanie w Pythonie

prowadzący: Piotr Ćwiakowski

Egzamin

Treść zadania

Jan Nowak założył własną firmę produkującą telefony. Chce walczyć z dużymi firmami takimi jak Apple, Samsung itp. Nie wie, jak oszacować cenę telefonów komórkowych, które tworzy jego firma. Na tym konkurencyjnym rynku telefonów komórkowych nie można po prostu zakładać rzeczy. Aby rozwiązać ten problem, zbiera dane dotyczące sprzedaży telefonów komórkowych różnych firm. Jan chce znaleźć jakiś związek między funkcjami telefonu komórkowego (np. RAM, pamięć wewnętrzna itp.) i jego półką cenową. Chce zrozumieć co odróżnia modele telefonów komórkowych sprzedawanych w różnych przedziałach cenowych. Poniżej opis zbioru danych[1]:

Zmienna	Opis
id	ID
battery_power	Całkowita energia, którą bateria może przechowywać w jednym czasie [w mAh]
blue	Ma Bluetooth lub nie
clock_speed	szybkość, z jaką mikroprocesor wykonuje instrukcje
dual_sim	Ma obsługę dwóch kart SIM lub nie
fc	Przednia kamera mega piksele
four_g	Ma 4G lub nie
int_memory	Pamięć wewnętrzna w gigabajtach
m_dep	Grubość komórki w cm
mobile_wt	Waga telefonu komórkowego
n_cores	Liczba rdzeni procesora
pc	Mega piksele kamery podstawowej
px_height	Wysokość rozdzielczości pikseli
px_width	Szerokość rozdzielczości pikseli
ram	Pamięć RAM w megabajtach
sc_h	Wysokość ekranu w cm
sc_w	Szerokość ekranu w cm
talk_time	Maksymalny czas działania telefnupodczas rozmowy
three_g	Ma 3G lub nie
touch_screen	Ma ekran dotykowy lub nie
wifi	Ma wifi lub nie
price_range	Zmienna objaśniana: 0 (niska cena), 1 (średnia cena), 2 (wysoka cena), 3 (b. wysoka cena).

Twoim celem jest pomóc Janowi w poznaniu strategiu różnicowania cen na rynku telefonów komórkowych. W tym celu wykonaj eksploracyjną analizę danych (EDA - *Explanatory Data Analysis*) w celu rozpoznania informacji w zbiorze. Wykorzystaj:

- moduł **Pandas** i **Numpy**, aby statystyki opisowe każdej zmiennej,
- moduł **Pandas** aby rozpoznać zależności pomiędzy zmienną objaśnianą i objaśnianymi (spróbuj policzyć statystyki w podgrupach i tabele przestawne),
- zwizualizuj rozkład zmiennej objaśnianej, rozkłady empiryczne zmiennych objaśniających używając pakietu **matplotlib**

[1] Źródło: https://www.kaggle.com/iabhishekofficial/mobile-price-classification

- zwizualizuj zależności między zmiennymi objaśniającymi i objaśnianą. Zastanów się, które z nich będą dobrymy predyktorami zmiennej objaśnaniej - wykorzystaj pakiet **matplotlib**,
- poszukaj ciekawych wiuzalizacji w galerii pakietu **seaborn** - https://seaborn.pydata.org/examples/index.html i wykorzystaj zobrazowania ciekawych zależności - (także pomiędzy zmiennymi objaśniającymi).

Po wykonaniu pracy zbierz najważniejsze części kodu dotyczące wczytania, czyszczenia i przekształcanie danych oraz zamieść najciekawsze wnioski z eksploracyjnej analizy danych (najważniejsze wykresy, tabele, wnioski) i przedstaw je w formie raportu analitycznego przykładowo za pomocą jupyter notebooka. Gotowy raport proszę wysłać na adres: pcwiakowski@labmasters.pl zamieszczając plik *.ipynb* oraz *.html* i umówić się z prowadzącym na krótkie spotkanie (do 15 minut), na którym zostanie omówione (w luźnej rozmowie) rozwiązane zaproponowane przez uczestnika.

W razie pytań zapraszam do kontaktu na podany powyżej adres email.

Powodzenia!