Ladislav Gaža
Utorok 18:00

Data science programming

Práca s jazykom R

Úvod

Vybraný dataset je o testovaní študentov. Možné ho je nájsť na odkaze https://www.kaggle.com/sonukumari47/students-performance-in-exams. Jedná sa o výsledky študentov z viacerých skúšok. Motiváciou pre vybranie tohto konkrétneho datasetu bola zvedavosť ako sa jedincom darí pri jednotlivých príznakoch ako napríklad pripravenosť na test alebo pohlavie s tým, že významným faktorom bola lineárna korelácia výsledkov testov z jednotlivých oblastí.

Dáta

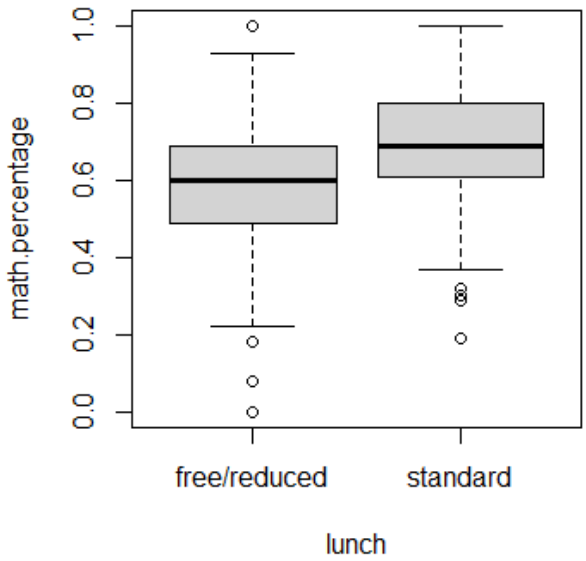
Základny opis dát

Dáta obsahujú 1000 záznamov s 9 atribútami. Prvý atribút obsahuje len číslo záznamu, a teda môže byť v kľude odstránený. Medzi hlavné atribúty patria percentuálne výsledky testov z matematiky, písania a čítania. Tie majú hodnotu od 0 do 1 s presnosťou na 3 desatinné miesta. Jedná sa teda o dátový typ double. Zvyšné atribúty informujú o rasovej skupine, úrovni vzdelania rodičov a pohlaví študenta. Atribúty su typu character. Posledné 2 atribúty určujú, či má študent normálny obed (standard) alebo redukovaný obed (free/reduced) a či sa človek zúčastnil prípravného kurzu na test alebo nie. Tie sú síce typu char ale je pre nás lepšie zmeniť ich na double s hodnotami 1 a 0 pre ďalšie spracovanie.

Podrobnejšia analýza dát

Ďalej môžeme preskúmať spojitosti a rozdelenie výsledkov na základe rozličných faktorov. Ako napríklad priemerné výsledky testov podľa pohlavia študenta, kde muži mali lepšie výsledky v matematike a ženy v písaní a čítaní. Taktiež u mužov bol najúspešnejší práve matematický test s priemerom 0,687 a u žien test z písania s priemerom 0,73. Následne sa vieme pozrieť na priemerne výsledky testov na základe prípravného kurzu na test. Študenti účastniaci sa kurzu majú priemerne výsledky z testov väčšie o 0,077. Najväčší rozdiel sa nachádza pri porovnávaní

výsledkov testov na základe typu obeda. Rozdiel medzi redukovaným a normálnym obedom je 0,086 s lepšími výsledkami pre študentov so štandardným obedom. Rozdelenie je lepšie viditeľné na krabicovom grafe.



Preto sa v ďalšej časti s hypotézami a regresným modelom pozrieme na to ako sú výsledky testov ovplyvňované práve druhom obeda.

Hypotéza a regresný model

Stanovenie Hypotézy

H0: Výsledky testu z matematiky závislé od výsledkov testu písania a čítania nezávisia od obedu.
H1: Výsledky testu z matematiky závislé od výsledkov testu písania a čítania závisia aj od obedu.

Ukážka regresného modelu

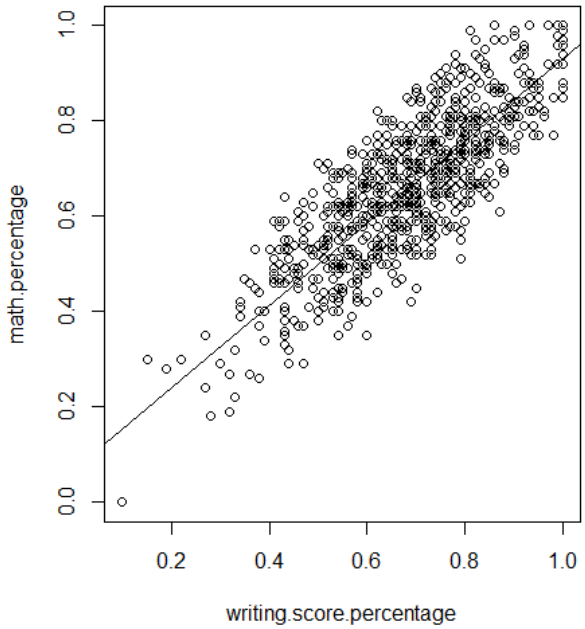
Na prácu s predikovaním údajov a počítaní odchyliek využívame knižnicu *caret*, ktorá obsahuje funkciu *predict* odhadujúcu hodnoty na základe predom vytvoreného modelu. Taktiež sú v knižnici funkcie na výpočet:

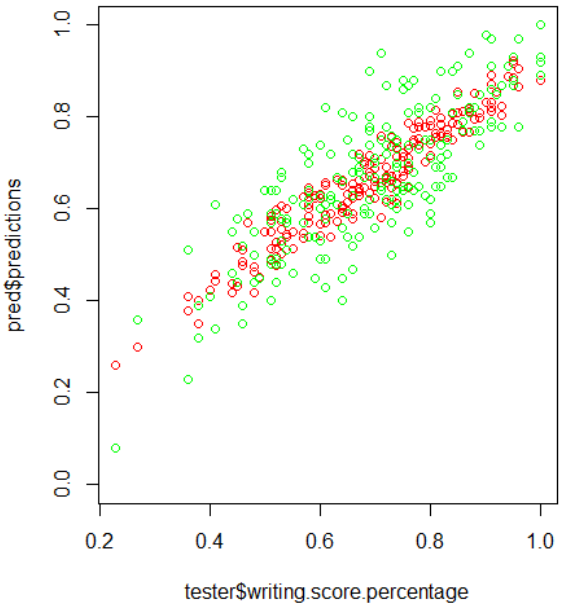
R-squared (R2) - umocnená koreláciu odhadovaných výsledkov a reálnych údajov, pri čom väčšie číslo znamená presnejší model

Root Mean Square Error (RMSE) - merajúcu priemerný rozdiel predikovaných údajov od pozorovaných, menšie RMSE znamená lepší model

Mean Absolute Error (MAE) - alternatíva RMSE ktorá meria priemerný absolútny rozdiel a teda je menej citlivá na outlierov, taktiež menšia hodnota značí lepší model

Po vytvorení modelu môžeme zakresliť do grafu regresnú krivku, ktorá zobrazuje náš model. Po predikovaní hodnôt vieme taktiež zobraziť pozície reálnych a odhadovaných hodnôt na jednom grafe, kde červené sú hodnoty vytvorené na základe modelu a zelené hodnoty odmerané v realite.





Regresný model a k-fold cross validácia

Pre k-fold cross validáciu je najprv potrebné rozdeliť celý dataset do k skupín, v našom prípade 5 o rovnakej veľkosti. Následne pre každú skupinu vytvárame model zo zvyšných dát (vynechána je 1 z 5 vytvorených skupín vždy po poradí) a predikujeme hodnoty aktuálnej skupiny. Pre každú skupinu počítame R2, RMSE a MAE podľa ktorých vieme určovať kvalitu modelov. Pre rýchlejšie spracovanie sme vytvorili funkciu ktorá je mapovaná na dáta.

Záver

	Bez obedu	S obedom	Rozdiel
R2 interval	0.609 - 0.714	0.642 - 0.734	
R2 priemer	0.675	0.700	0.025
RMSE interval	0.084 - 0.090	0.080 - 0.088	
RMSE priemer	0.087	0.083	-0.004
MAE interval	0.068 - 0.074	0.065 - 0.072	
MAE priemer	0.071	0.068	-0.003

Podľa tabuľky vidíme, že hodnota R2 je vyššia pre modely s obedom. Čím vyššie hodnoty tým lepšie. Tiež vidíme, že hodnoty pre RMSE a MAE sú nižšie pre študentov s obedom. Tu platí čím nižšia hodnota tým lepšie.

Výsledné hodnoty ukazujú lepšie výsledky, respektíve robustnejšie modely pre študentov s obedom. Nevieme s istotou povedať či sú tieto hodnoty štatisticky signifikantné na rozdiel od nuly, pretože by sme museli vykonať štatistické testy. Avšak pre účely tohto zadania môžeme zamietnuť nulovú hypotézu H0 v prospech alternatívnej hypotézy H1. Robustnejšie modely nám ukazujú, že výsledky testu z matematiky závisia aj od obedu.