Objavovanie znalostí

Martin Macej and Jozef Varga

Slovenská technická univerzita v Bratislavae
Fakulta informatiky a informačných technológií
Ilkovičova 2, 842 16 Bratislave

1 Analýza datasetu

1.1 Opis problému

Pri založení spoločnosti si častokrát jej zakladatelia kladú otázku, akú cenu výrobkov majú stanoviť, aby boli konkurencie schopný a zároveň aby dosiahli čo najvyšší zisk. Je teda potrebné zaradiť produkty do správnej cenovej kategórie.

Správne riešenie tohto problému dokáže následne dopomôcť firme so vstupom na trh a stať sa tak schopnou konkurencie už existujúcim a zabehnutým firmám.
My konkrétne sa budeme zaoberať zaradením mobilných telefónov do rôznych cenových kategórii. Toto zaradzovanie budeme predpovedať pre každé zariadenie na základe toho, akou hardvérovou konfiguráciou bude disponovať.

Na zadanie sme si chceli vybrať dataset [1] z domény, ktorá nám je srdcu blízka. Vybrali sme teda doménu IT a to konkrétne dataset obsahujúci údaje o hardvérovej konfigurácii 2000 mobilných zariadení.

Pri tomto datasete by sme sa radi zamerali na rôzne vzťahy medzi cenou a použitým hardvérom, ktorý obsahuje dané zariadenie. Napríklad porovnanie, či veľkosť pamäte RAM vplýva na cenovú kategóriu viac alebo menej ako počet jadier zariadenia.

1.2 Opis a charakteristiky dát

Vybraný dataset obsahuje 2000 riadkov a 21 stĺpcov (parametrov) vo formáte csv. Všetky hodnoty sú číselné a všetky záznamy sú kompletné. Hlavnou entitou sú mobilné zariadenia (1 zariadenie predstavuje 1 riadok súboru).
Najväčší rozptyl hodnôt je v stĺpci, ktorý opisuje veľkosť pamäte RAM a je na úrovni 1176643.606, kde minimálna hodnota je 256 a maximálna 3998 GB.

Najmenší rozpyl je 0,1817 a to pri stĺpci hovoriacom o tom, či mobilný telefón má podporu 3G, kde približne 76% má (hodnota 1) a ostatné nie (hodnota 0).

Všetky hodnoty máme ordinálne.

Pri analyzovaní hodnôt sme sa pozreli aj na koreláciu hodnôt. Všimli sme si zaujímavú skutočnosť a to takú, že jedinú významnú koreláciu má na cenovú kategóriu iba veľkosť pamäte RAM.

Ostatné atribúty na cenu nijako významne nevplývajú a tak sme sa rozhodli neskôr dopočítať v implementácii ďalšie stĺpce, ktoré by nám mohli pomôcť lepšie klasifikovať cenovú kategóriu.

Pridať nové stĺpce sme sa rozhodli pre atribúty, ktoré nie sú binárne, t.j. neobsahujú iba hodnoty 0 a 1 (dáta sú bližšie opísané v nasledujúcej časti Charakteristika dát).

	price_range
price_range	1
ram	0,917045736
px_width	0,165817502
px_height	0,148857555
int_memory	0,044434959
sc_w	0,038711272
pc	0,0335993
three_g	0,023611217
sc_h	0,022986073
fc	0,021998208
talk_time	0,021858871
blue	0,020572854
wifi	0,018784812
dual_sim	0,017444479
four_g	0,014771711
n_cores	0,004399275
m_dep	0,000853037
clock_speed	-0,006605691
mobile_wt	-0,030302171
touch_screen	-0,030411072

Obrázok 1 Korelácia dát

Ďalej pri analýze dát sme si vykreslili jednotlivé dáta pomocou pravidla Interquartile Range. Týmto krokom sme zistili, že väčšina stĺpcov neobsahuje žiadnych outlierov. Medzi stĺpce, ktoré obsahujú outlierov sa nám javili iba tie, ktoré reprezentujú rozlíšenie prednej kamery, výšku zariadenia a to, či mobilný telefón podporuje 3G.

Count predstavuje počet záznamov a True count nám hovorí o počte „problémových“ riadkov.

	fc	px_height	three_g
count	2000	2000	2000
True count	18	2	477
False count	1982	1998	1523

Obrázok 2 Interquartile range

Pre samokontrolu sme navyše vyskúšali aj výpočet z_score pre jednotlivé stĺpce. Ako najkritickejší sa nám javil stĺpec s rozlíšením prednej kamery.

Z týchto dôvodov by sme chceli odstrániť hodnoty, ktoré sa priveľmi odlišujú a tak použijeme odstránenie outlierov, aby sme dospeli k výslednej lepšej klasifikácii.

1.2.1 Charakteristika dát

battery_power – int64, celková energia, ktorú dokáže uložiť batéria v mAh, hodnoty sú od 501 po 1998 mAh, priemerná hodnota je 1238.5185, medián 1226 a smerodajná odchýlka je pomerne vysoká a to 439.418

blue – int64, obsahuje/neobsahuje bluethooth vyjadrené číslami 0 a 1, smerodajná odchýlka má hodnotu 0.5

clock_speed – float64, takt procesora od 0.5 do 3, priemerná hodnota je 1.522 a smerodajná odchýlka je na úrovni 0.816

dual_sim – int64, obsahuje/neobsahuje možnosť duálnej SIM karty, vyjadrené číselne 0 - neobsahuje, 1 – obsahuje

fc – int64, rozlíšenie prednej kamera, hodnoty sú od 0 po 19, priemerná hodnota je 4.3095, medián je 3 a štandardná odchýlka 4.341

four_g – int64, obsahuje/neobsahuje 4G, hodnoty 0 a 1

int_memory – int64, veľkosť internej pamäte v MB, hodnoty sú v rozmedzí od 2 po 64 GB, priemer je 32.0465 a štandardná odchýlka 18.1457

m_depth – float64, hĺbka telefónu v cm, od 0.1 cm po 1 cm, priemerná hĺbka je 0.50175 a štandardná odchýlka 0.2884

mobile_wt – int64, váha telefónu v gramoch, od 80 po 200 g, priemerná váha je 140.249 a štandardná odchýlka 35.4

n_cores – int64, počet jadier procesora, od 1 po 8, priemerne 4.5205, štandardná odchýlka je na úrovni 2.2878

pc – int64, rozlíšenie zadnej kamery, od 0 po 20, priemerné rozlíšenie telefónov je 9.9165 a štandardná odchýlka 6.0643

px_height – int64, výška rozlíšenia v px, od 0 po 1960, priemerne 645.108 a štandardná odchýlka je na úrovni 443.7808

px_width – int64, šírka rozlíšenia v px, od 500 po 1998, priemerne 1251.5155 a štandardná odchýlka je 432.2

ram – int64, veľkosť pamäte RAM v MB, od 256 po 3998, priemerne 2124.213, štandardná odchýlka je teda až 1084.732

sc_h – int64, výška mobilu v cm, od 5 po 19, priemerná výška je 12.3065 a štandardná odchýlka 4.2132

sc_w – int64, šírka mobilu v cm, od 0 po 18, priemerne 5.767, štandardná odchýlka je podobná ako pri výške mobilu a to 4.3564

talk_time – int64, dĺžka hovoru na jedno nabitie, od 2 po 20, priemerne 11.011 a štandardná odchýlka je 5.464

three_g – int64, obsahuje/neobsahuje 3G, hodnoty 0 alebo 1, priemer je 0.7616, takže väčšina mobilov podporuje 3G

touch_screen – int64, má/nemá dotykový displej, hodnoty takisto 0 alebo 1, priemer je na úrovni 0.503, z čoho vidíme že iba približne polovica mobilov z datasetu má dotykový displej

wifi – int64, má/nemá wifi modul, taktiež hodnoty 0 alebo 1, priemer podobný ako pri dotykovej obrazovke a to konkrétne 0.507

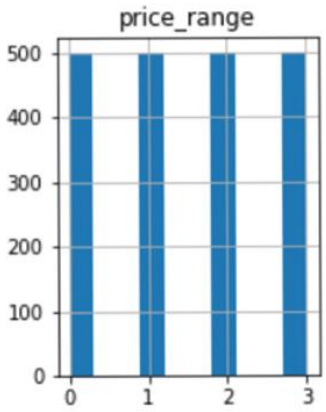
price_range – int64, charakterizuje cenovú kategóriu mobilného telefónu, hodnoty sú od 0 po 3, takže celkovo 4 cenové hladiny, priemer je na úrovni 1.5 a štandardná odchýlka 1.1183

Bližšie informácie o dátach a ich rozložení je v prílohách.

1.3 Definovanie úlohy

Dáta v tomto datasete máme v pláne použiť na zaradenie mobilného zariadenia do jednej zo štyroch cenových kategórií. Budeme teda riešiť úlohu klasifikácie. Pre zadané vstupné parametre zariadenia sa pokúsime predikovať, do akej cenovej kategórie by sme vedeli telefón s danou hardvérovou konfiguráciou zaradiť tak, aby nebol pod cenu a zároveň aby jeho cena nebolo premrštená.

Výhodou datasetu je, že má rovnomerné rozdelenie ako je vidno na obrázku 1, čo znamená že predpoveď cenovej kategórie nebude ovplyvnená prevahou niektorej kategórie.



Obrázok 3 Rozdelenie cenových kategórií

1.4 Scenár riešenia

Najprv si dáta očistíme a to tak, že odstránime outlierov aby nám neskôr pri zaradzovaní mobilného telefónu do cenovej kategórie neskresľovali klasifikáciu.

Plánujeme použiť 2 metódy: z_score a iqr.

Následne pridáme nové stĺpce, ktoré budú obsahovať medián hodnôt, horný a dolný kvantil, medián a priemer.

Vyberieme tie polia (hardvérové komponenty), ktoré majú najväčší podiel na zaradenie zariadenia do cenovej kategórie. Pre tento výber použijeme algoritmy correlation matrix a VIF (variance inflation factor) feature selection.

Ďalším krokom je už samotné klasifikovanie. Pre každý typ klasifikovania spravíme 4 klasifikácie (odstránenie outlierov iqr alebo z_score a feature selecture correlation_matrix alebo VIF).

Rozhodli sme sa použiť viacero typov klasifikátorov a porovnať ich, ktorý nám dá aké výsledky. Použijeme viacero klasifikátorov a to Decision tree, Nearest neighbors, Linear Discriminant Analysis, Svm sigmoid, Svm linear, Svm RBF a ďalšie.

Zdroje

1. Mobile Price Classification, https://www.kaggle.com/iabhishekofficial/mobile-price-classification#train.csv

Prílohy

Príloha A – výpis stĺpcov s ich počtom, popisom, dátovým typom a počtom unikátnych hodnôt

Columns	Count	Info	Type	Unique_value
battery_power	2000.0	Total energy a battery can store in one time measured in mAh	int64	1094.0
blue	2000.0	Has bluetooth or not	int64	2.0
clock_speed	2000.0	speed at which microprocessor executes instructions	float64	26.0
dual_sim	2000.0	Has dual sim support or not	int64	2.0
fc	2000.0	Front Camera mega pixels	int64	20.0
four_g	2000.0	Has 4G or not	int64	2.0
int_memory	2000.0	Internal Memory in Gigabytes	int64	63.0
m_dep	2000.0	Mobile Depth in cm	float64	10.0
mobile_wt	2000.0	Weight of mobile phone	int64	121.0
n_cores	2000.0	Number of cores of processor	int64	8.0
pc	2000.0	Primary Camera mega pixels	int64	21.0
px_height	2000.0	Pixel Resolution Height	int64	1137.0
px_width	2000.0	Pixel Resolution Width	int64	1109.0
ram	2000.0	Random Access Memory in Megabytes	int64	1562.0
sc_h	2000.0	Screen Height of mobile in cm	int64	15.0
sc_w	2000.0	Screen Width of mobile in cm	int64	19.0
talk_time	2000.0	longest time that a single battery charge will last when you are	int64	19.0
three_g	2000.0	Has 3G or not	int64	2.0
touch_screen	2000.0	Has touch screen or not	int64	2.0
wifi	2000.0	Has wifi or not	int64	2.0
price_range	2000.0	This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).	int64	4.0

Príloha B – analýza dát a ich hodnôt v datasete

Columns	Max	Mean	Median	Min	Quantile_25%	Quantile_50%	Quantile_75%	Standard_deviation	Variance
battery_power	1998.0	1238.5185	1226.0	501.0	851.75	1226.0	1615.25	439.4182060835312	193088.35983766866
blue	1.0	0.495	0.0	0.0	0.0	0.0	1.0	0.5001000400170029	0.25010005002500796
clock_speed	3.0	1.5222499999999983	1.5	0.5	0.7	1.5	2.2	0.8160042088950705	0.6658628689344699
dual_sim	1.0	0.5095	1.0	0.0	0.0	1.0	1.0	0.5000347661750046	0.25003476738369157
fc	19.0	4.3095	3.0	0.0	1.0	3.0	7.0	4.341443747983884	18.84813381690835
four_g	1.0	0.5215	1.0	0.0	0.0	1.0	1.0	0.4996624673623623	0.24966258129064378
int_memory	64.0	32.0465	32.0	2.0	16.0	32.0	48.0	18.145714955206863	329.26697123561803
m_dep	1.0	0.5017500000000017	0.5	0.1	0.2	0.5	0.8	0.2884155496235109	0.08318352926463188
mobile_wt	200.0	140.249	141.0	80.0	109.0	141.0	170.0	35.399654896388334	1253.1355667833905
n_cores	8.0	4.5205	4.0	1.0	3.0	4.0	7.0	2.287836718042658	5.234196848424202
pc	20.0	9.9165	10.0	0.0	5.0	10.0	15.0	6.064314941347797	36.77591570785413
px_height	1960.0	645.108	564.0	0.0	282.75	564.0	947.25	443.7808108064387	196941.40804002012
px_width	1998.0	1251.5155	1247.0	500.0	874.75	1247.0	1633.0	432.19944694633784	186796.36194072032
ram	3998.0	2124.213	2146.5	256.0	1207.5	2146.5	3064.5	1084.7320436099492	1176643.6064342167
sc_h	19.0	12.3065	12.0	5.0	9.0	12.0	16.0	4.2132450043563106	17.75143346673341
sc_w	18.0	5.767	5.0	0.0	2.0	5.0	9.0	4.356397605826395	18.978200100049946
talk_time	20.0	11.011	11.0	2.0	6.0	11.0	16.0	5.463955197766686	29.854806403201582
three_g	1.0	0.7615	1.0	0.0	1.0	1.0	1.0	0.4262729223187337	0.1817086043021532
touch_screen	1.0	0.503	1.0	0.0	0.0	1.0	1.0	0.5001160445626752	0.2501160580290157
wifi	1.0	0.507	1.0	0.0	0.0	1.0	1.0	0.5000760322381043	0.2500760380190055
price_range	3.0	1.5	1.5	0.0	0.75	1.5	2.25	1.118313602106461	1.2506253126563283

Príloha C - rozloženie dát pre každý stĺpec datasetu