UNIVERZITA KOMENSKÉHO V BRATISLAVE

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ODHADOVANIE CIEN MOBILNÝCH TELEFÓNOV

Projekt - Strojové učenie (2-INF-150)

Juraj Holas **10.1.2015**

ABSTRAKT

Projekt sa zameria na odhadovanie cien vybraných modelov mobilných telefónov (smartfónov) v slovenských internetových obchodoch. Skúmaná je minimálna a priemerná cena jednotlivých modelov skrz niekoľko desiatok obchodov, a jej vzťah vzhľadom ku technickým parametrom telefónu, dátumu jeho uvedenia na trh a počiatočnej predajnej cene.

Projekt je riešený pomocou metódy generalizovanej lineárnej regresie s využitím *k-fold* testovania. Vstupné parametre sú predspracované a prispôsobené tak, aby v čo najväčšej miere zodpovedali reálnemu svetu a zároveň boli prispôsobené pre túto metódu. Vďaka tomu sa nám podarilo získať odchýlky na úrovni cca 10% pre minimálnu cenu a 13% pre priemernú cenu pri dvojmesačnej predpovedi.

1. POPIS PROJEKTU

Cena nového mobilného telefónu závisí od mnohých faktorov, počínajúc jeho výbavou a končiac maržou konkrétneho predajcu. Projekt má za cieľ predpovedať vývoj cien telefónu v čase, konkrétne v období cca dvoch mesiacov dopredu. Zameriava sa pritom na čisto numerické parametre, historickú cenu, a aproximáciu niektorých empirických atribútov, ako je napr. výrobca zariadenia. Využitím analytických metód sa pokúsime získať z týchto dát čo najpresnejšie výsledky pre oba predpovedané atribúty.

2. DÁTA

2.1 ZDROJE DÁT

Pre tento projekt sme potrebovali dva druhy dát:

- **Historické údaje o cenách telefónov:** tieto údaje, konkrétne historickú minimálnu a priemernú cenu konkrétnych telefónov, poskytla spoločnosť MINET s.r.o., prevádzkovateľ portálu NajNákup.sk. Tieto dáta môžete nájsť v elektronickej prílohe dokumentu (*res/data.xlsx*, hárok *Ceny*).
- **Technické parametre telefónov:** tieto údaje sme čerpali z internetovej stránky www.gsmarena.com , kde sú verejne dostupné profily všetkých skúmaných modelov, vrátane podrobných technických parametrov. Dáta opäť nájdete v elektronickej prílohe (*res/data.xlsx*, hárok *Parametre*).

2.2 POPIS DÁT

Dáta boli zhromažďované pre vybranú vzorku 23 telefónov. Táto vzorka bola vybraná tak, aby zahŕňala pokiaľ možno čo najširšie spektrum telefónov: najlepšie telefóny z rokov 2014, 2013 a 2012, kompaktné telefóny z rokov 2014 a 2013, veľké telefóny – *phablety*, telefóny strednej triedy a telefóny nižšej triedy. Z každej skupiny boli vybrané 3 reprezentatívne modely telefónov, ktoré boli zaradené do výberu[1].

[1] Pre nízku predajnosť v dnešnej dobe, a teda zlú dostupnosť dát boli z kategórie najlepších telefónov 2012 vybrané namiesto troch iba dva modely.

Údaje o cenách od spoločnosti MINET sú v tabuľke vo forme záznamov, kde každý záznam prislúcha jednému dňu, a má podobu: *dátum–minimálna cena–priemerná cena–počet obchodov*. Posledný atribút – počet obchodov ponúkajúcich tovar sme v projekte nezohľadňovali i napriek tomu, že od neho závisí hľadaná priemerná cena mobilu. Pri predpovedi by sme však potom potrebovali toto číslo dopredu poznať, t.j. vedieť koľko obchodov bude daný tovar ponúkať o dva mesiace.

Pri technických parametroch telefónov sme zhromažďovali nasledovné údaje:

- **Uvedenie na svetový trh**
- **Počiatočná cena na Slovensku** – istým spôsobom aproximuje vplyv značky telefónu, nakoľko telefóny od niektorých výrobcov sú štandardne od začiatku drahšie ako od iných
- **Rozmery a váha telefónu**
- **Parametre displeja** – veľkosť, rozlíšenie a jemnosť
- **Verzia OS** – všetky telefóny boli s OS Android
- **Výkon** – frekvencia a počet jadier procesora, veľkosť RAM pamäte
- **Úložný priestor** a možnosť jeho rozšírenia pamäťovými kartami
- **Fotoaparát** – rozlíšenie fotografií a videa, rozlíšenie predného fotoaparátu
- **Verzia Bluetooth**
- **Batéria** – kapacita a výdrž v rôznych podmienkach
- **Skóre v testoch výkonnosti**

V tabuľke si môžete všimnúť červených polí – označujú údaje, ktoré neboli dostupné na stránke GSMArena.com. Väčšina týchto údajov bola doplnená z ďalších zdrojov na internete (najmä www.phonearena.com). Parameter *Batéria – Endurance rating*, ktorý je založený na testoch robených redaktormi portálu GSMArena.com, bol pri niekoľkých chýbajúcich položkách odhadnutý pomocou jednoduchej lineárnej regresie, ktorej vstupom boli ostatné parametre batérie.

2.3 PRÍPRAVA DÁT

Nakoľko počet parametrov telefónu pri zbere dát sa vyšplhal až ku číslu 29, bolo potrebné a vhodné túto množinu zmenšiť. Tým pádom sa jednak docieli väčšia efektívnosť algoritmu, nakoľko bude pracovať s menej dátami, ale aj vyššia presnosť, nakoľko došlo k vylúčeniu irelevantných či zlúčeniu závisiacich parametrov. Boli vykonané nasledovné zmeny:

Plocha prednej strany namiesto šírky a výšky telefónu – empirický pocit zákazníka o veľkosti telefónu nevyplýva samostatne zo šírky a výšky, ale skôr z celkovej plochy, ktorú telefón zaberá na dlani. Hrúbku sme do súčinu nezaradili, nakoľko tú ľudia vnímajú inak a podstatne citlivejšie ako ostatné dva rozmery. Navyše, vyrobiť o niekoľko milimetrov väčší či menší telefón stojí výrobcu minimálne náklady, zatiaľ čo vyrobiť o niekoľko milimetrov tenší telefón si žiada kompaktnejšie, a teda drahšie komponenty.

Výška displeja namiesto kompletného rozlíšenia – šírka a výška displeja v pixloch sa drží štandardného pomeru 16:9. Jedna hodnota je teda jednoducho lineárne závislá na druhej, preto je zbytočné zohľadňovať obe.

Vynechanie verzie OS – tento parameter má na cenu minimálny vplyv, nakoľko výrobcu stojí prvotná optimalizácia systému temer rovnaké peniaze pri ľubovoľnej verzii. Navyše všetky

skúmané telefóny mali OS vo verzii 4.*, čiže nie veľmi odlišné verzie, medzi ktorými väčšina radových používateľov ani nerozlišuje.

Vynechanie verzie Bluetooth – s jedinou výnimkou mali všetky skúmané telefóny rovnakú verziu, čiže pre algoritmus by tento parameter nemal takmer žiadnu výpovednú hodnotu.

Zmenšenie počtu parametrov batérie – v pôvodnej tabuľke sa nachádza až päť rôznych stĺpcov popisujúcich batériu. Rozhodol som sa ponechať dva – kapacitu ako technický údaj, a *Endurance rating* ako empirické porovnanie reálnej výdrže batérie.

Vypustenie údajov o testoch výkonnosti – tieto údaje silo závisia od parametrov procesora a RAM. Navyše u nich nastal problém so zlou dostupnosťou.

Po týchto zmenách ostalo v tabuľke 18 parametrov, ktoré hodnoverne popisujú každý model.

2.4 ROZDELENIE DÁT

Celkovo sme od spoločnosti MINET s.r.o. obdržali 2549 záznamov o cenách, ktoré sme následne obohatili o parametre jednotlivých telefónov. Záznamy o každom telefóne začínali väčšinou v druhej polovici roka 2013 alebo v roku 2014, s jednou výnimkou (HTC One X), kde bol prvý záznam už v apríli 2012. Posledný dátum záznamov bol 15. december 2014. Záznamy boli rozdelené tak, aby veľkosť testovacej množiny bola štvrtina až tretina veľkosti trénovacej množiny. Rozdeľovanie prebiehalo automaticky podľa nasledovného kľúča:

- Záznamy do 14.10.2014 sú použité pri trénovaní
- Záznamy od 15.10.2014 sú použité pri testovaní (testujú sa teda 2 mesiace záznamov do 15.12.2014)
- Táto hranica je posunutá na 5.12.2014 pri nových telefónoch, t.j. s dátumom uvedenia na svetový trh v septembri 2014 alebo neskôr (na slovenský trh dorazia spravidla asi o dva mesiace neskôr).

3. METÓDY A VÝSLEDKY

Pri riešení projektu sme postupovali iteratívnym spôsobom, pričom v každej fáze sa vykonávalo aj testovanie aktuálneho programu. Na základe tohto testovania bolo následne vždy stanovené ďalšie smerovanie projektu, či nastavenie kľúčových parametrov. Jednotlivé nasledujúce časti zachytávajú tieto iterácie v poradí, ako prebiehali.

3.1 PREDSPRACOVANIE DÁT

Prvým podstatným krokom bolo pripraviť údaje pre algoritmus lineárnej regresie. Vstupné dáta boli uložené v dvoch tabuľkách, a to vo forme záznamov:

- *model – zoznam záznamov o cenách*
- *model – dátum výroby – technické parametre*

Pre potreby algoritmu sme museli získať nezávislé záznamy obsahujúce vstupné atribúty a cieľové hodnoty. Pomocou programu v jazyku Java sme teda spomínané dve tabuľky spojili – rozdelili sme zoznamy cien na jednotlivé záznamy, a ku každému tomuto záznamu sme pripojili technické parametre daného modelu. Výsledkom bola jedna tabuľka záznamov v tvare:

- *dátum – dátum výroby – technické parametre – minimálna cena – priemerná cena*

Prvý atribút predstavuje dátum z tabuľky cien, t.j. v ktorý deň boli zaznamenané hodnoty minimálnej / priemernej ceny. Všimnime si tiež, že v zázname už nefiguruje model názov modelu telefónu – telefón je špecifikovaný iba jeho technickými parametrami a dátumom výroby. Vďaka tomu budeme môcť výsledný program použiť nie len na telefóny, ktoré boli zaradené do vybranej vzorky, ale aj ľubovoľné ďalšie.

Implementácia

Túto časť projektu zabezpečuje hlavná metóda triedy `MergeTables`, ktorú môžete nájsť v rovnomennom súbore v elektronickej prílohe v priečinku *src/*. Okrem samotného spojenia dvoch vstupných tabuliek prebieha v tejto metóde aj rozdelenie na trénovaciu a testovaciu množinu podľa kľúča uvedeného v časti 2.4. Na konci sa tieto množiny dát osobite uložia do dvoch výstupných súborov.

3.2 GENERALIZOVANÁ LINEÁRNA REGRESIA

Metódu regresie je pre tento problém vhodná z toho dôvodu, že cena zariadenia vo veľmi vysokej miere závisí práve od jeho parametrov a veku, a to zväčša pomerne jednoducho: čím lepšie parametre / novšie zariadenie, tým väčšia cena. Táto závislosť však nemusí byť iba lineárna, niektoré faktory v nej môžu mať aj iný charakter vývoja. Do množiny hypotéz boli teda zavedené tri typy funkcií:

- **mocninové funkcie do stupňa *d*:** $x_1^{d_1}.x_2^{d_2}.\ldots.x_n^{d_n}\ ,\ \sum d_i \leq d$
- **exponenciálne funkcie:** 2^{x_i}
- **logaritmické funkcie:** $\log_2 x_i$

Do množiny hypotéz som sa rozhodol začleniť aj exponenciálne a logaritmické funkcie vzhľadom na povahu technológií a ich postupu. Mnoho parametrov totiž postupom času rastie exponenciálne, ako je známe už z Moorovho zákona spred 50 rokov. Rozvoj týmito funkciami však v snahe vyhnúť sa preučeniu nie je aplikovaný na všetky vstupné parametre, ale iba na tie, u ktorých má význam – najmä teda parametre výkonu.

Implementácia

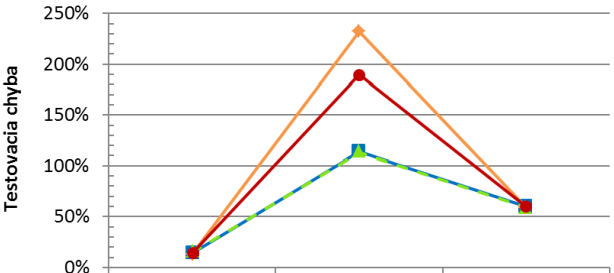
Celá funkcionalita regresie je implementovaná v jazyku Otave (priečinok *octave/* v prílohe). Samotné trénovanie je inicializované v *regres.m*, v časti „*Bez k-fold*“[2]. Trénovanie a testovanie ako také prebieha pomocou série vlastných funkcií v súbore *functions.m* .

Výsledky

Prvým výsledkom po spustení programu bolo zlyhanie Octave. Upozornilo to na chybu v koncepcii, nakoľko exponenciálnou funkciou boli rozvíjané aj parametre s pomerne vysokými hodnotami. Po umocnení na ne boli výsledkom čísla za hranicou dátových typov.

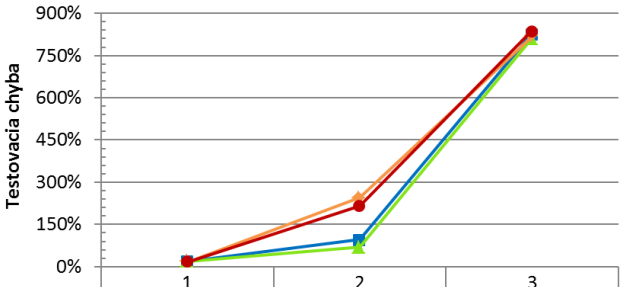
Algoritmus bol preto upravený tak, aby pred aplikovaním exponenciálneho rozvoja tieto hodnoty škáloval podľa vopred daných koeficientov. Tieto koeficienty boli počítané ako priemer hodnôt daného parametra vypočítaný z tabuľky technických parametrov. Po tejto úprave algoritmus úspešne prebehol, a vypočítal nasledujúce priemerné chyby:

[2] Vzhľadom na ďalší postup v programe je teraz daná časť nepoužívaná a zakomentovaná.



	d = 1	d = 2	d = 3
iba polynóm	14,331	113,83	59,675
polynóm + log.	15,026	113,69	59,205
polynóm + exp.	13,128	231,82	59,433
polynóm + exp. + log.	14,452	189,72	60,121

Graf 1 - Vplyv bázových funkcií na predpovedanie minimálnej ceny



	1	2	3
iba polynóm	17,136	96,026	826,76
polynóm + log.	17,686	68,977	810,41
polynóm + exp.	15,216	245,75	816,19
polynóm + exp. + log.	16,309	216,48	837,91

Graf 2 -Vplyv bázových funkcií na predpovedanie priemernej ceny

Na grafoch je jasne vidno, že polynomiálny rozvoj bázových funkcií rýchlo vedie k preučeniu, a teda nie je správnou cestou kam sa uberať. Pri ďalších testoch ostal teda parameter *d* rovný jednej, čo znamená že v množine hypotéz sme ponechali iba lineárne funkcie, prípadne s kombináciou s exponenciálnymi alebo logaritmom.

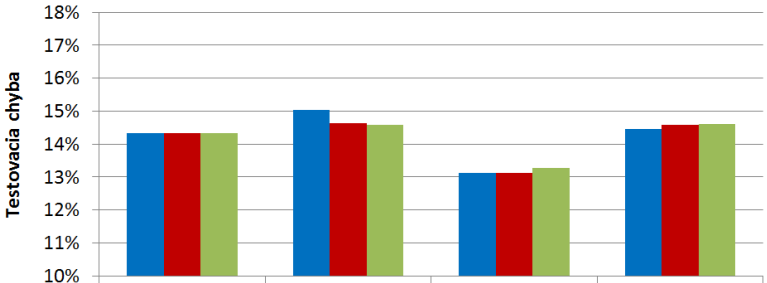
Druhým poznatkom týchto testov bolo, že rozdiel medzi výsledkami pri vyčlenení resp. zahrnutí exponenciálnych a logaritmických funkcií nie je veľmi veľký (najmä pri $d = 1$), čiže sa oplatí skúšať všetky štyri zobrazené alternatívy aj v ďalších testoch.

Automatické škálovanie

Ako bolo spomínané, pred aplikovaním exponenciálneho rozvoja sa dané hodnoty škálovali. Škálovacie koeficienty však boli predpočítané a napevno vsadené do algoritmu. Navyše spôsob, akým sa počítali, nereflektoval dáta úplne správne, nakoľko sa počítal ako priemer príslušného parametra z tabuľky technických parametrov, a teda ako priemer 23 hodnôt. Rôzne telefóny mali ale rôzny počet záznamov o cene, a teda aj iným počtom prispievali do trénovacej množiny.

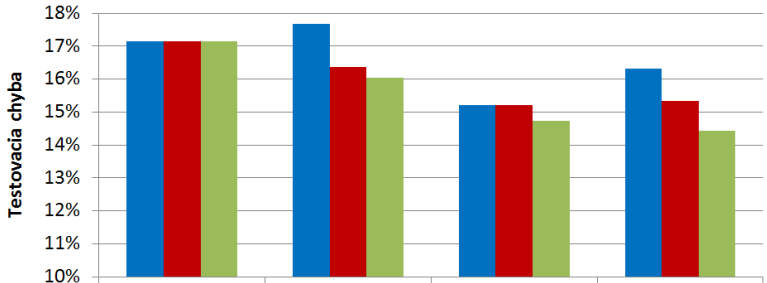
Ďalšie vylepšenie teda spočívalo v zavedení automatického počítania koeficientov pre toto škálovanie. Program vypočítal priemer pre každý stĺpec *trénovacích dát*, teda už priamo tých, ktoré vchádzali do regresného algoritmu. Pred exponenciálnym rozvojom potom škáloval podľa takto vypočítaných koeficientov. Pri testovaní sa používali rovnaké koeficienty, t.j. nepočítali sa nové z množiny testovacích dát. Implementačne je táto časť riešená funkciou `generateCoefs` v súbore *functions.m* . Rovnaký princíp som taktiež vyskúšal aj pre rozvoj logaritmom.

Výsledky



	iba polynóm	polynóm + log.	polynóm + exp.	polynóm + exp. + log.
Manuálne škálovanie pre exp.	14,331	15,026	13,128	14,452
Manuálne škálovanie pre exp. aj log.	14,331	14,629	13,128	14,592
Automatické škálovanie pre exp. aj log.	14,331	14,586	13,282	14,601

Graf 3 - Škálovanie pred nelineárnym rozvojom - minimálna cena



	iba polynóm	polynóm + log.	polynóm + exp.	polynóm + exp. + log.
Manuálne škálovanie pre exp.	17,136	17,686	15,216	16,309
Manuálne škálovanie pre exp. aj log.	17,136	16,358	15,216	15,335
Automatické škálovanie pre exp. aj log.	17,136	16,043	14,726	14,444

Graf 4 - Škálovanie pred nelineárnym rozvojom - priemerná cena

Zatiaľ čo výsledky testov pri predpovedaní minimálnej ceny sú takmer vyrovnané, pri predpovedi priemernej ceny je vidno zlepšenie úspešnosti algoritmu po takto vykonanom škálovaní. V ďalšom postupe preto bude toto automatické využité ako pre exponenciálny, tak pre logaritmický rozvoj[3].

3.3 ŠKÁLOVANIE TECHNICKÝCH PARAMETROV

Ďalšou snahou vylepšiť efektivitu programu bolo vniesť doň ďalšiu informáciu z reálneho sveta. Idea bola postavená na fakte, že za rovnakú cenu sa dá dnes kúpiť oveľa výkonnejšie zariadenie ako napr. pred rokom, na druhej strane rovnako výkonné zariadenie bude dnes stáť menej ako predtým, nakoľko úroveň technológií sa už medzitým posunula. Technické parametre telefónov sme teda chceli škálovať podľa nejakého **technologického štandardu**.

Voľba dobrého referenčného bodu – teda parametrov, ktoré budeme považovať za technologický štandard – je pre tento krok programu kľúčová. Rozvoj v tomto segmente je však veľmi dynamický, čo voľbu mierne skomplikovalo. Nakoniec som sa rozhodol pre použitie parametrov špičkových telefónov na trhu z rôznych časových období. Tie totiž pomerne verne reflektujú nie iba čistý technologický pokrok (napr. aký prototyp procesora sa vývojárom

[3] Pre zaujímavosť ostalo v zdrojovom kóde aj manuálne nastavovanie koeficientov, je však zakomentované a nepoužíva sa.

podarilo postaviť), ale aj jeho reálne využitie v komerčných produktoch. Navyše je pomerne jednoduché a jednoznačné vybrať z daného obdobia ten najlepší model (resp. niekoľko najlepších) spravidla sú to totiž vlajkové lode výrobcov, ktorými sa chvália najviac.

Dáta na škálovanie

K dvom spomínaným tabuľkám pribudla tretia, zhromažďujúca 26 najlepších telefónov na trhu v období od 2011 do súčasnosti. Tieto údaje môžete nájsť v elektronickej prílohe (*res/data.xlsx*, hárok *Top modely*). V tejto forme však ešte údaje nie sú prakticky použiteľné, potrebujeme ich transformovať do formy akéhosi technologického štandardu za dané obdobie. Ten budeme empiricky chápať ako odpoveď na otázku „*čo najlepšie vie trh aktuálne ponúknuť?*". Nie každý nový telefón však prekoná svojho predchodcu vo všetkých smeroch, preto sme museli posúvať lepšie hodnoty starších telefónov ďalej.

Pre každý parameter a časový údaj sme teda vybrali najlepšiu hodnotu z tých, ktoré sme pri danom parametri videli v minulosti. Napr. pre jemnosť displeja v októbri 2012 sme vybrali najväčšiu jemnosť, ktorá bola na trhu do tohto mesiaca (vrátane), v našom prípade 316ppi, a túto hodnotu sme použili ako referenčný bod, teda odpoveď na našu otázku: „*v októbri 2012 si môžem kúpiť telefón s jemnosťou displeja najviac 316ppi*". Takýmto spôsobom sme zostrojili celú referenčnú tabuľku. Pri väčšine parametrov sme pritom brali väčšiu hodnotu ako lepšiu, avšak napr. pri hmotnosti alebo hrúbke bola menšia hodnota považovaná za lepšiu. Hodnoty pre veľkosť zariadenia (šírka, výška) a uhlopriečku displeja sme s porovnávania vynechali, nakoľko sú skôr otázkou trendu či osobnej preferencie zákazníka, vo všeobecnosti však neplatí závislosť čím väčší (príp. menší) telefón, tým lepší.

Pri spomínanom vypĺňaní tabuľky sa však objavil problém: úroveň niektorých parametrov sa zhoršovala, namiesto toho aby sa zlepšovala. Najviac citeľný bol tento trend u hmotnosti zariadenia: kým vo februári 2011 bola referenčná hmotnosť na úrovni 116 gramov, pri dnešnom trende veľkých telefónov je priam nemožné tak ľahký telefón vyrobiť, priemerná hmotnosť sa pohybuje okolo 150g. Metódu pre výber referenčnej hodnoty pre daný čas a parameter sme teda mierne upravili – namiesto toho, aby sa vyberala najlepšia hodnota zo *všetkých* starších záznamov, sa bude vyberať iba zo záznamov *maximálne rok starých*. Týmto spôsobom sme v referenčnej tabuľke získali vierohodnejšie údaje.

Implementácia

Škálovanie sme zaradili do procesu ku spájaniu tabuliek. Java program teda načítal aj tretiu – referenčnú tabuľku, a každý záznam pred uložením do výstupného súboru škáluje podľa aktuálneho technologického štandardu. Tzn. že z referenčnej tabuľky sa vyberie čo najnovší riadok, avšak nie starší ako je dátum záznamu, a všetky hodnoty záznamu sa predelia hodnotami v riadku. Takto upravený – škálovaný záznam sa následne uloží do výstupného súboru.

Výsledky

Ihneď po implementácii tohto škálovania sa implementovalo aj *k-fold* testovanie, pričom testy prebiehali až po dokončení oboch. Výsledky teda nájdete v nasledujúcej časti.

3.4 *K-FOLD* TESTOVANIE

Posledná implementovaná technika na zlepšenie výsledkov je *k-fold* testovanie (resp. validácia) dát. Trénovacia množina T sa rozdelí na k podmnožín približne rovnakej veľkosti

$T_1, T_2, \ldots, T_k$, a k trénovaniami získame k vektorov na predpoveď, kde každý θ_i je výsledkom trénovania na množine $T \setminus T_i$. Pomocou neho potom predpovedáme výsledky na množine T_i a spočítame ich chybu err_i. Z vektorov $\theta_1, \theta_2, \ldots, \theta_k$ potom vyberieme ten, ktorý mal najmenšiu chybu.

Implementácia

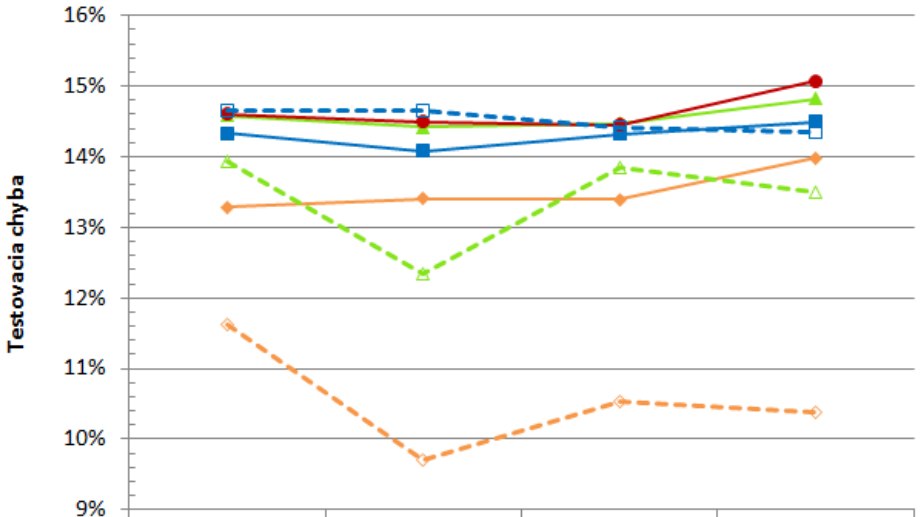
Táto časť algoritmu bola implementovaná v rámci programu v Octave. Daná časť kódu je na začiatku označená komentárom *„k-fold"*. Na začiatku programu sa dá voliteľne nastaviť parameter k, označený premennou `kMax`.

Výsledky

Posledné testovanie prebiehalo v niekoľkých variantoch podľa nastavenia parametrov spomínaných skôr v tomto projekte:

- bez/s *k-fold* testovaním, pri $k = 5$, $k = 15$ a $k = 25$
- bez/so škálovaním technických parametrov telefónov
- bez/s rozvojom exponenciálnymi a logaritmickými funkciami
- predpovedanie minimálnej/priemernej ceny

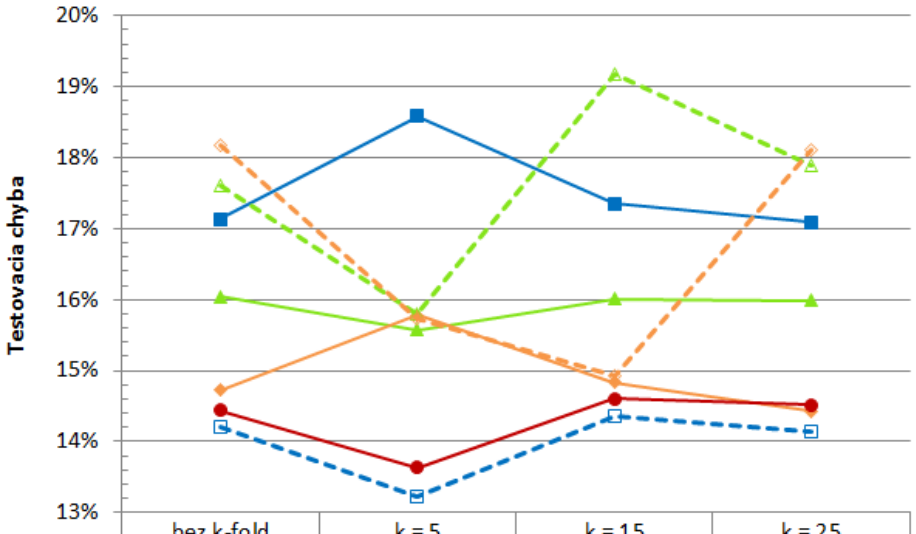
Výsledky si môžete pozrieť v nasledujúcich grafoch:



	bez k-fold	k = 5	k = 15	k = 25
bez škálovania - lin.	14,331	14,082	14,324	14,499
bez škálovania - log.	14,586	14,424	14,475	14,82
bez škálovania - exp.	13,282	13,402	13,397	13,985
bez škálovania - exp.+log.	14,601	14,497	14,451	15,073
so škálovaním - lin.	14,648	14,655	14,417	14,354
so škálovaním - log.	13,933	12,335	13,848	13,494
so škálovaním - exp.	11,628	9,7104	10,534	10,375
so škálovaním - exp.+log. (mimo mierky)	34,846	17,899	20,084	21,018

Graf 5 - Finálne výsledky - minimálna cena

Na grafe vidíme, že v prípade predpovede minimálnej ceny je jasne najlepším riešením škálovanie parametrov a použitie iba exponenciálneho rozvoja, ktorý má v priemere až o takmer 5% menšiu testovaciu chybu. Vidíme tiež, že použitie *k-fold* testovania sa oplatilo, a najlepšie výsledky dosahovalo pri $k = 5$. Celkovo najlepší výsledok je odchýlka na úrovni **9,71%**.



	bez k-fold	k = 5	k = 15	k = 25
bez škálovania - lin.	17,136	18,579	17,352	17,091
bez škálovania - log.	16,043	15,575	16,007	15,988
bez škálovania - exp.	14,726	15,784	14,837	14,431
bez škálovania - exp.+log.	14,444	13,63	14,606	14,521
so škálovaním - lin.	14,215	13,233	14,365	14,14
so škálovaním - log.	17,601	15,806	19,179	17,901
so škálovaním - exp.	18,185	15,731	14,921	18,116
so škálovaním - exp.+log. (mimo mierky)	70,297	31,452	32,071	31,507

Graf 6 - Finálne výsledky - priemerná cena

Už na prvý pohľad vidíme, že výsledky pri predpovedaní priemernej ceny sú o poznanie horšie (v priemere asi o 4%), a tiež menej stále. Pramení to z toho, že rozptyl priemernej ceny je vyšší ako pri minimálnej cene, nakoľko ho vždy ovplyvňuje nie iba jeden obchod, ale niekoľko desiatok. Aj v tomto prípade sme vďaka škálovaniu technických parametrov dosiali lepší výsledok, avšak tento krát iba o malý kúsok (0,3%). Najlepší výsledok dosiahol úrovne **13,23%**.

V oboch prípadoch môžeme vidieť, že použitie *k-fold* testovania pri parametri $k = 5$ buď nezmenilo, alebo zlepšilo výsledky (výnimkou je iba predikcia priem. ceny bez škálovania a ďalšieho rozvoja). Pri väčších hodnotách k však už odchýlka opäť stúpla. Za daných podmienok je teda práve $k = 5$ (prípadne o málo väčšie alebo menšie) ideálnou voľbou.

Za všimnutie tiež stoja veľmi zlé výsledky, ktoré sme získali pri použití rozvoja aj exponenciálnou aj logaritmickou funkciou po škálovaní technických parametrov. V tomto prípade pri predikcii oboch cieľových atribútov stúpla hodnota odchýlky dvoj- až trojnásobne. Pri škálovaní totiž došlo ku normalizácií hodnôt, ktoré tým pádom mohla vystihnúť aj jednoduchšia hypotéza. Ako môžeme pozorovať, pri zložitej hypotéze ako v tomto prípade nastane výrazné preučenie, a testovacia chyba značne stúpne. Toto tvrdenie podporuje aj fakt, že aj v ostatných variantoch testovania sa bez škálovania v priemere lepšie darilo mierne zložitejším hypotézam, zatiaľ čo po škálovaním boli skôr úspešnejšie tie jednoduchšie.

4. ZÁVER

Hlavným cieľom tohto projektu bolo získať algoritmus, ktorý by bol schopný čo najpresnejšie predpovedať cenu mobilného telefónu v slovenských internetových obchodoch, a to iba na základe jeho parametrov. Po implementovaní niekoľkých metód a techník na čele s generalizovanou lineárnou regresiou sa nám podarilo stlačiť chybovosť na úroveň **9,71%** pri predpovedaní minimálnej ceny, a **13,23%** pri predpovedaní priemernej ceny v horizonte dvoch mesiacov. Vzhľadom na tieto výsledky považujem projekt za úspešný, i keď ešte nie vhodný do praktického komerčného použitia. Pri dnešných cenách telefónov okolo 400€ totiž aj 10%-ná odchýlka spôsobí pomerne citeľný rozdiel 40€ na cene.

Pre vylepšenie úspešnosti existuje niekoľko variant. Koncepčne najjednoduchším riešením je rozšíriť trénovaciu množinu – pridať dáta o cenách a parametroch viac modelov, než iba 23 člennej vzorky použitej v tomto projekte. Ďalším, pomerne jednoduchým zlepšením by bolo citlivejšie nastavenie parametra pre *k-fold* testovanie. V rámci tohto projektu boli skúšané iba štyri možnosti s pomerne veľkým odstupom, avšak pri postupovaní po jednotkách by sme metódou horolezca možno dospeli ku ešte presnejším výsledkom. Mierne zložitejším riešením, ktoré by mohlo tiež priniesť lepšie výsledky, by bola implementácia lokálnej váhovej aproximácie do regresného algoritmu.